Eliminating Last-Mile Inefficiencies in the Trucking Industry

By
Kristian From
Bachelor in Military Studies, Logistics and Resource Management, Royal Norwegian Naval Academy, 2015

And

Katharina Mangan
Bachelor of Science, Kinesiology, The Pennsylvania State University, 2012

SUBMITTED TO THE SUPPLY CHAIN MANAGEMENT PROGRAM
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2020

Signature of Author: ____
Department of Supply Chain Management
May 8, 2020

Signature of Author: ____
Department of Supply Chain Management
May 8, 2020

Certified by: ____
Milena Janjevic
Research Engineer, MIT Humanitarian Supply Chain Lab
Capstone Advisor

Accepted by:
____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Eliminating Last-Mile Inefficiencies in the Trucking Industry

by
Kristian From
and
Katharina Mangan

Submitted to the Program in Supply Chain Management
on May 8, 2020 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Pilot Freight Services, traditionally a bulk cargo freight forwarder in the US, is in the process of expanding their business to provide last-mile delivery (LMD) services. This capstone project helps Pilot improve the performance of their LMD operations through higher visibility and elimination of efficiencies. First, an understanding of Pilot's current LMD operation is established. Next, a performance metric framework is defined, with two performance dimensions: (1) service level and (2) efficiency. Guided by the framework, the performance of Pilot's LMD operations is assessed by analyzing descriptive statistics. A visualization tool is built in Tableau, allowing Pilot to continuously assess their own performance. Finally, machine learning is used to identify parameters that affect performance and predict their impact. The parameters identified as having the biggest impact on stop time duration are: volume delivered, population density, quantity pieces delivered, stop number, time of day, and peak day. For drive time duration, the single most relevant factor is mileage. For each of the locations analyzed, coefficients are calculated and made available to Pilot's planners to predict stop and drive time based on the parameters. Planning accuracy, in terms of MAPE, is for stop time improved from about 85% to about 55%, and for drive time from about 45% to 25%. The insight provided by this capstone will allow Pilot to better understand and assess the performance of their LMD operations and help identify areas for improvement.

Capstone Advisor: Milena Janjevic
Title: Research Engineer, MIT Humanitarian Supply Chain Lab

ACKNOWLEDGEMENTS

We would like to sincerely thank Milena Janjevic, our advisor, for sharing her expertise and for her academic rigor that helped us explore various paths and decide the course of our Capstone.

Also, this research would not have been possible without the insights and resources from our sponsor. Their help and support throughout the Capstone ensured a project we believe will be of great use for them in the future. We would especially like to thank Brandon and Lance for their personal involvement and sharing of knowledge. Their insight has been vital to the completion of this project.

We are grateful to Pamela Siska for the generous dedication of her time to editing and helping to structure this capstone.

TABLE OF CONTENTS

List of Acronyms

Acronyms used in this capstone project:

B2B – Business to Business

B2C – Business to Consumer

CSCMP - Council of Supply Chain Management Professionals

CSV – Comma Separated Values

CVRPTW – Capacitated Vehicle Routing Problem with Time Windows

DC- Distribution Center

DT – Decision Tree

KPI – Key Performance Indicator

LMD – Last Mile Delivery

LR – Linear Regression

ML – Machine Learning

NP-hard - Non-deterministic Polynomial-time hardness

P-PAT – Pilot Performance Assessment Tool

RF – Random Forest

TSF – Travel Speed Factor

VRP – Vehicle Routing Problem

List of Tables

List of Figures

1 INTRODUCTION

1.1 Context

Pilot Freight Services, traditionally a bulk cargo freight forwarder in the US, is in the process of expanding their business to provide last-mile delivery (LMD) services. The company would like insight into any inefficiencies in their growing last-mile operations as they deliver products from distribution centers to customers' homes in five cities: Phoenix, Minneapolis, Chicago, Birmingham, and Portsmouth. In this capstone project, we worked together with Pilot to improve the transparency on the performance of their last-mile operations, identify inefficiencies and propose ways to improve their operations.

1.2 Motivation

1.2.1 Freight Transportation and Last Mile Delivery

According to the 30th Annual Council of Supply Chain Management Professionals (CSCMP) State of Logistics Report, the U.S. trucking market makes up about 70% of all transportation costs. Trucking costs are estimated at about 700 billion dollars, with independent, full truckload at about 296 billion and less-than-truckload is valued at about 72 billion (Ward et al. 2019). The consulting company Deloitte estimates that the amount of goods moved by freight will increase by 27 percent in 2027 (Deloitte University Press 2017). The increase is largely related to the increase in e-commerce and LMD to customers.

LMD refers to the last part of the supply chain when the product is delivered to the buyer: a customer (B2C) or another business (B2B). In recent years it has risen to be an important area of competition, and the main growth in LMD has been in B2C. Companies such as Amazon use shorter delivery times to distinguish themselves from traditional brick-and-mortar shops and gain market shares (Deloitte University Press 2017). As smaller companies are looking for efficient ways of distributing their products, the growth potential of LMD is further increased (Ward et al. 2019). As a result, more transport providers are moving into the LMD market to take advantage of the growth potential. One such company is Pilot Freight Services.

1.2.2 Pilot Freight Services

Pilot Freight Services has offices all over the world and provides full-range global transportation and logistics services (Pilot Freight Services 2019). Pilot's retail customers are increasingly seeking service for last-mile delivery. The company is expanding its business from primarily traditional shipping, B2B, into a more dynamic B2C model, where they have to handle direct delivery to customers. This shift is adding a new level of complexity to their route planning, as they are affected by constraints such as service level, shorter delivery time, and smaller deliveries to more and urban locations. Pilot is looking for ways to improve the efficiency of their current LMD operations.

Currently, Pilot is using Dispatch Track, a routing optimization software, to plan their daily delivery operations. Based on inputs from Pilot's operators, including customer orders and constraints such as driver service time, the program routes available vehicles to cover all orders for a given day. Dispatch Track can optimize these routes based on Pilot's preferences; for example, the routes can be optimized by overall driving time or by shortest distance between points (DispatchTrack 2019). As this is new and uncharted territory, it is still uncertain what factors these routes should be optimized for. The current routes are working, but there are inefficiencies that Pilot would like further insight into.

1.3 Research Problem

This project will help Pilot Freight Services remove inefficiencies within its operations by providing strategic insight to assist Pilot Freight Services identify wasted time and resources in their last-mile delivery operations. The goal is to help improve daily routing performance based on relevant performance indicators such as order completion rate and efficiency.

The scope of this project is to identify factors that are relevant in determining the performance of LMD routes. Operational locations within the scope of this project are Chicago, Phoenix, Minneapolis, Portsmouth, and Birmingham. This project will focus on operations while financial aspects such as profitability of routes will not be considered.

1.4 Hypothesis

We believe that Pilot's Freight Services LMD operations can be improved through higher visibility into the performance of their operations and elimination of inefficiencies. The project will identify major drivers in Pilot's current operations, analyze them and identify potentials for improvements. The research method that will be used is to first build an understanding of Pilot's current LMD operation. Based on this understanding, a performance assessment framework will be developed. In this framework, different dimensions of performance will be defined with appropriate key performance indicators (KPI) and levels of aggregation. Based on these performance dimensions, descriptive statistics will be used to analyze the performance of Pilot's current operations and identify inefficiencies. Lastly, using the insight from the analysis of Pilot's current operations, we will use machine learning (ML) to identify parameters that affect Pilot's LMD performance. The ML analysis will identify if, and how, parameters such as mileage and number of packages affect the performance of their LMD operations. The aim is to identify the relevant parameters and build prediction models that can help establish an understanding of what drives LMD performance. Before doing the analysis, a literature review is necessary to better understand LMD operations, the underlying operational challenge of vehicle routing and how ML can be used for predicting relevant factors.

2 Literature Review

2.1 Topic Overview

This capstone project will help Pilot Freight Services identify inefficiencies in their LMD operations. This will be achieved by first understanding Pilot's current operations. Second, establishing a performance framework for assessment. Third, doing the performance assessment using descriptive statistics. Lastly, we will use ML methods to identify variables that affect performance. To understand how to approach and solve this problem, we focus on three main topics: LMD, the Vehicle Routing Problem (VRP) and ML.

LMD provides context about the market this capstone is framed within, as we seek to understand how the LMD market works and why it is important. On an operational level, vehicle routing is one of the most important aspects affecting LMD operations performance. During our talks with Pilot, we learned that they use a tool called Dispatch Track to help with routing. Dispatch Track assists Pilot with daily route planning by solving the VRP. Understanding the company's current decision-making and routing practices is needed in order to identify ways to improve Pilot's operations. Once we understand Pilot's current operations and how it is performing, predictive ML methods will be used to understand how different variables affect performance.

2.2 Last-Mile Delivery

LMD is often also referred to as last-mile logistics, service, or transportation. The definition we chose is the one of Supply Chain Dive: "… last-mile deliveries encompass any movement of freight or products between a distribution center and the point at which the end consumer will receive it." (Lopez 2017). This definition focuses on freight between DC and customers, the business at the core of this capstone project. LMD is related to the last part of the supply chain: getting the product or service to the end customer or user. In recent years, it has developed into one of the main competitive advantages of growing companies such as Amazon (Boyer, Prud'homme & Wenming, 2009).

LMD is becoming a critical factor for companies to stay competitive as customers have grown to expect short delivery times without having to pay extra (Deloitte University Press 2017). The LMD market is growing rapidly and companies need to adapt (Ewedairo, Chhetri, and Jie 2018). In addition to meeting customer demand, LMD allows companies to customize their customer delivery experience as well as managing their own costs. Establishing efficient LMD operations is vital to companies, some estimates suggest that LMD makes up about 28% of total transportation costs (Goodman 2005). LMD is challenging because it often happens in urban environments where negative factors such as congestion, safety, and environment must be balanced against the need to deliver goods (Savelsbergh, Woensel, and Stewart 2016).

In order to optimize their LMD operations, companies can act on several levers: (1) strategic levers such as location of distribution centers, (2) tactical levers such as fleet size and night deliveries, and (3) operational levers such as vehicle routing (Olsson, Hellström, and Pålsson 2019). Improving LMD operations, strategic and tactical changes can often be large and challenging like moving or establishing new distribution centers (strategic) or changing the fleet size through hiring or lay-offs (tactical). For this capstone project we will focus on operational levers, as changes in this domain often can be done without any significant investments. By improving vehicle routing, we ensure that all drivers are utilized efficiently before we start considering tactical decision like changing fleet size. This is faster, easier and often more cost efficient (hiring drivers that are note utilized leads to waste of money). Therefore, in this capstone the focus is on operational levers, more specifically on vehicle routing.

2.3 Vehicle Routing Problem

There are many commercial vehicle routing software's available for companies that do LMD operations. Pilot uses Dispatch Track to help with creating efficient delivery routes. In essence, Dispatch Track helps Pilot solve the VRP. The VRP is a generalization of the Traveling Salesman Problem (TSP), first introduced by Dantzig and Roser in 1959 (Abidi, Hassine, and Mguis 2018). It is an optimization problem where one seeks to optimize routing to service a given number of customers with available vehicles (Baldacci, Battarra, and Vigo 2008). The VRP problem has many different variants depending on what goods are transported, level of service required, and characteristics of customers or delivery vehicles (Kumar and Panneerselvam 2012).

The VRP problem is considered a NP-hard (non-deterministic polynomial-time hardness) problem, meaning that it is hard to compute an optimal solution when the problem gets bigger (Jozefowiez, Semet, and Talbi 2008). As a result, heuristics have been developed to identify key factors to simplify the problem, and make it easier to solve (Chunhua 2010).

In our project, we focus on Capacitated VRP with Time Windows (CVRPTW). “Capacitated” means that the trucks used by Pilot have several constraints that affect how they can be routed. Examples of such constraints are driver service time and the vehicles maximum load capacity (Filipec, Skrlec, and Krajcar 1998). “Time window” refers to the fact that customers are available for delivery at certain times during the day. When ordering a product, the customer can choose a time window for when they want the good(s) to be delivered. The routing model must then plan routes that match with these time windows (Yulei and Jin 2019).

The object of the CVRPTW is to serve a number of customers within pre-defined time windows, without violating the capacity constraints of each vehicle (Kumar and Panneerselvam 2012). The optimization can be done for time, cost, or other factors based on what the user wants (Chunhua 2010). Approaches to solving the CVRPTW fall into the following categories: exact methods (trying to find exact, best solution), heuristics approaches (simplifying the problem to make it easier to compute), meta-heuristics (creating a solution that picks the best heuristics to simplify the problem), and hybrid methods (combining the three other methods) (Kumar and Panneerselvam 2012).

A lot of research has been done to identify relevant factors to consider when solving the different versions of VRP. In their study, Ewedairo, Chhetri, and Jie (2018) suggest splitting deliveries to increase the degree or freedom and make it easier to come up with optimal routing solutions. Boyer, Prud’homme, and Wenming (2009) suggest focusing on improving stop density and lengthen delivery windows to improve performance. VRPs can be solved either exactly (on small scale problems) or through using heuristics (more common on large problems). The route performance given by the model is highly dependent on the accuracy of the parameters used in the model. Parameters with poor accuracy, not able to capture the true value of effect on the route (like actual stop duration at a customer location) will make the VRP solution inaccurate (compared to the real route) and lead to poor planning. Especially in larger routing problems with many customers and locations, identifying appropriate parameters can be challenging, especially in complex last-mile urban settings. Here ML methods can be helpful to identify patterns and predict appropriate variable values.

2.4 Machine Learning Methods

To better understand how we might use ML to solve this problem, we must first discuss the concept of ML itself. ML is a combination of computer science and statistics, in the form of algorithms, that can be used to discover patterns in data and train a machine to predict certain outcomes or preferences (Güemes-Peña, Lõpez-Nozal, Marticorena-Sánchez and Maudes-Raedo. 2018). For LMD operations and vehicle routing, ML is capable of identifying previously unidentified patterns in large datasets and help streamline supply chains by improving parameters used in planning (Snoeck, Merchán, and Winkenbach 2020). ML techniques can be categorized into four main types: supervised, unsupervised, semi-supervised, and reinforcement learning.

Figure 2.1 shows the different categories of ML techniques and the type of data they require (Mohammed, Khan, and Bashier 2017). Although this project will be using supervised learning, it is necessary to understand the basic differences between each ML method in order to choose the best approach.

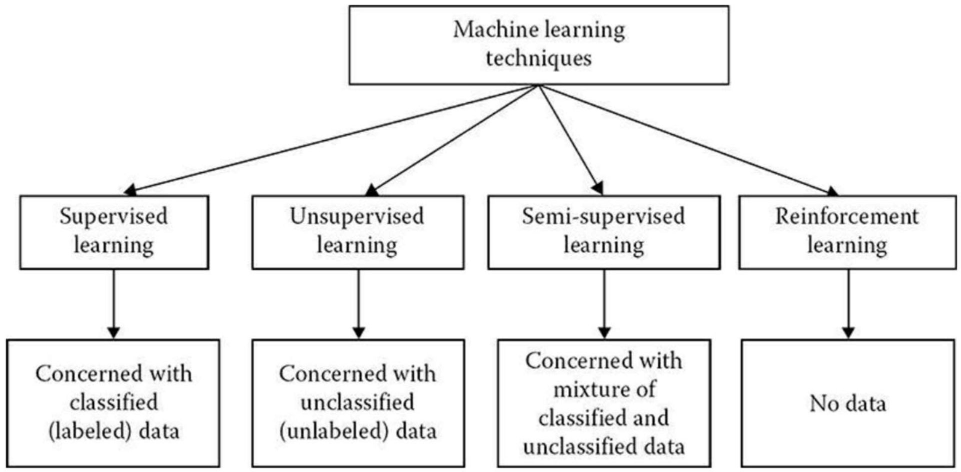

Figure 2.1 Machine learning techniques and required data types (Mohammed et al., 2017)

Supervised learning involves training data that is labeled and can be used to deduce a relationship or correspondence between two values. Imagine labels on an input axis (X) and an output axis (Y). The output axis (Y) explains the input from axis (X). Two main types of algorithms fall under supervised learning: regression and classification (Wuest, 2015). Due to the large amount of error in machines when completing these types of algorithms, these types of problems are "supervised" meaning that a human oversees the decision-making process. In classification problems, the machine is trained to place something into some class. An everyday example of classification with supervised learning is when a computer detects "spam" or "not spam" based on the content of an email (Knox, 2018). In regression problems, the machine is trained to predict a value such as price, weight, or size. An example of regression with supervised learning is when the price of a house is predicted based off of its lot size or some other factor (Kotsiantis, 2007).

Unsupervised learning involves unlabeled data with the objective of identifying hidden configurations in the data. Imagine data points on an input axis (X) and input axis (Y), where data points are unlabeled (Knox, 2018). In unsupervised learning, a machine may be able to recognize groups of data points or patterns and assign them a label without human oversight. The two main types of algorithms that fall under unsupervised learning are clustering and density estimation, or association. Clustering problems aim to discover inherent groupings in data, such as grouping customers by purchasing behavior. Association problems aim to discover rules that describe large portions of the data, such as "people who purchase X also tend to purchase Y" (Brownlee 2016).

Semi-supervised learning involves a situation in which both labeled and unlabeled data is being used. These types of problems have a large amount of input data (X) but only some of it is labeled (Y). An example of this type of problem is a Facebook photo album where only some of the images are labeled, but most are not (Kotsiantis, 2011).

Reinforcement learning uses positive or negative feedback from a user or its environment in order to make better decisions in the future. For example, whether users "likes" or "dislikes" a song is stored as positive or negative feedback in order to choose a better song match in the future based off of tempo, intensity, genre, or any number of other factors (Mohammed et al. 2017). Reinforcement learning will not be used during this project.

Figure 2.2 shows the four different categories of ML algorithms (classification, clustering, regression, and density estimation) based on whether or not the values in the range are finite and whether or not they are supervised. The red text illustrates common solutions to these problems that involve converting one problem into another. (Knox, 2018)

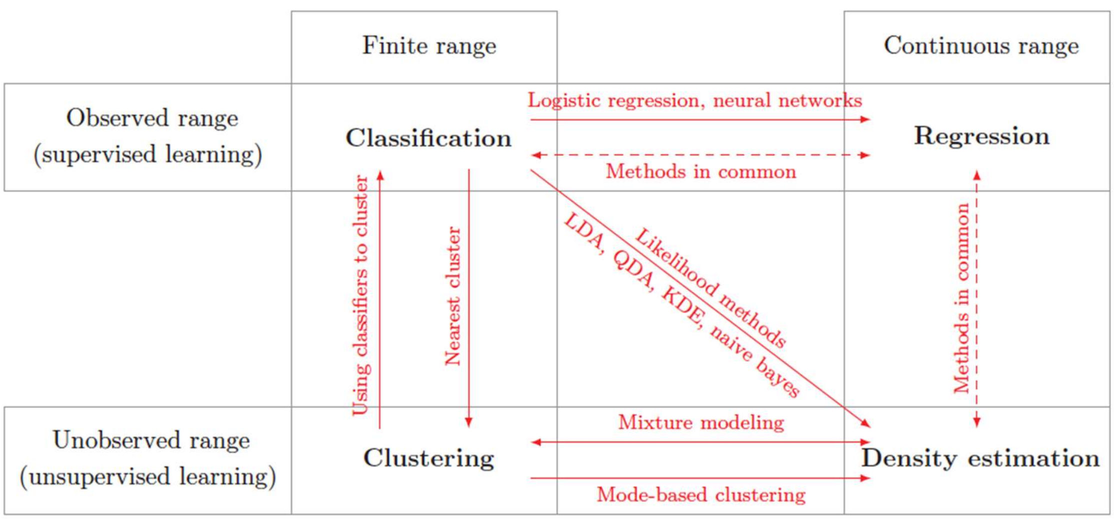

Figure 2.2 Machine learning algorithms by range and supervision (Knox, 2018)

For this capstone, ML will be used to identify and predict appropriate values for parameters used in Dispatch Track to do daily routing. For predictions supervised ML methods are appropriate. More specifically, the ML methods used in this capstone are Linear Regression (LR), Decision Tree (DT) regression and Random Forest (RF) regression. LR is a method used to estimate the relationship between two or more variables. It is linear when the relationship is best expressed by a linear function (basic format y = a + bx) (Tantawi 2019). DT is predictive modeling that classify instances by splitting the instances based on featured values in order to conclude on the value of the instance. Each node in the tree represents a sorting point and the branches the sorting values used to split the instances. Instances are classified starting at the root node and sorted based on their feature values (Chekole 2019). RF expands on DT by running multiple DT's with a random sub-set of the total dataset. This allows RF to explore more variations than the DT and reduce the risk of overfitting (Pedregosa et al. 2011).

Given the daily routing reports and data we have reviewed in Dispatch Track, supervised ML will be helpful in solving challenges with routing and handling times for Pilot Freight Services. Inaccuracy of handling time input could be the root cause of route planning inefficiency. Using ML to predict handling times with higher accuracy, will improve route planning leading to better customer service and more efficient operations.

2.5 Summary of Literature Review

Helping Pilot identify inefficiencies in their LMD operations, this capstone will focus on the operational level of their LMD operations by analyzing the performance of their daily routing. Pilot is currently using Dispatch Track to solve a CVRPTW version of the VRP. The performance of their routes is largely affected by the accuracy of the parameters they use for routing. For this project, we believe that prediction of appropriate values for parameters such as stop time and drive time, using supervised ML methods, will help Pilot improve their LMD operations. A better understanding of the underlying parameters will improve planning accuracy and create better understanding of how these parameters affect the performance of Pilot's current operations.

3 METHODOLOGY

This capstone project helps Pilot Freight Services identify inefficiencies and provides insight to improve their LMD operations. In the literature review, it was established that the underlying problem is a CVRPTW. The supervised ML methods LR, DT and RF have been found to be the most promising methods to approach the problem. In this chapter, the methodology of this capstone project is outlined. First, a brief overview of our method.

3.1 Methodology Overview

The major steps of the methodology is outlined in Figure 3.1.

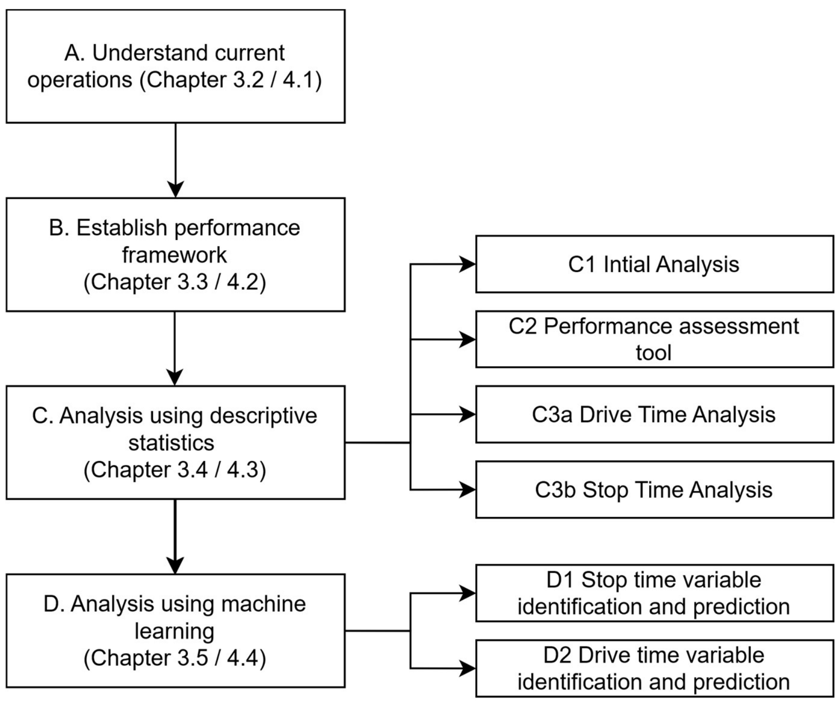

Figure 3.1 Major steps in research methodology

The first step towards improving Pilot's LMD operations was to understand their current operations. During a visit to Pilot's headquarters we met key personnel and got an introduction to their current LMD operations and their routing tool, Dispatch Track. Next, we defined a performance framework with appropriate performance dimensions. For each performance dimension, KPI's were defined to assess performance. Descriptive statistics was then used to assess the performance of Pilot's current LMD operations within the defined performance framework. Based on the insight from the analysis using descriptive statistics, we used ML methods to identify variables that affect performance and make predictions for improved planning accuracy.

3.2 Understanding Pilot's Current Operations

Knowing how Pilot currently runs their operation is vital to understanding how to improve their operations. A company visit to Pilot's headquarters provided an opportunity to learn how the company operates and meet key personnel. The visit served as a kick-off for the project, focusing on establishing an understanding for the company's history and culture, their way of operating and introduce us to their routing software Dispatch Track. Interviews and informal talks with Pilot personnel helped build cultural understanding and enabled cooperation throughout the project. The students then got access to Dispatch Track to explore the software and understand what information Pilot currently has access to. The reports provided by the system are divided into the subgroups Customers, Drivers, Routes, Orders and Items. Most reports are aggregated on a daily level, for example giving total miles driven for all routes in a location in a day.

Using the insight from learning about Pilot's current operations, we established the methodology for the capstone project. We decided that establishing a performance framework was necessary to measure the performance of their LMD operations. Using the framework, and descriptive statistics, the performance of Pilot's current operations can be assessed. The performance assessment is then used to identify inefficiencies and establish prediction methods using ML.

3.3 Establish Performance Framework

Using the insight from the company visit and understanding of Pilot's current operations, a performance framework is established. This framework starts with defining appropriate service dimensions. Within each service dimension, several KPI's are defined to identify and measure performance. Lastly, the appropriate aggregation level is defined for each KPI. The framework is developed using data and reports from Dispatch Track and requirements from Pilot managers. The performance framework and associated KPI's are presented in Section 4.2.

3.4 Analysis Using Descriptive Statistics

After building an understanding of Pilot's current operations and defining the relevant service dimensions and KPI's, the next step was to analyze the performance of Pilot's current LMD operations. The analysis was done using the programming language Python in Jupiter Notebook, with the associated Python libraries Pandas, Numpy, Seaborn, Scikit-learn, and Statsmodel. Tableau was used for visualization.

An initial analysis was done to get familiar with the data and better understand Pilot's operations. Next, we analyzed the performance of Pilot's current operations. Using the insight from this analysis, we did further analysis into the two main components of service time: stop and drive time. Stop time is the time spent at the customer location doing a delivery. Drive time is the time a driver spends driving between customer locations. Stop time is set by the planner for each individual stop. However, currently there is no clear method for determining the stop time. Stop time is normally set in multiples of 10 (between 10 and 40 minutes) and planners use whatever information they have available. Drive time is calculated by Dispatch Track based on the planned route. Planners can choose between farthest first or shortest distance, as well as rearrange stops. Based on this, Dispatch Track calculate the best routes. The only other way planners can adjust the routes is to change the driver speed by changing a Travel Speed Factor (TSF) in Dispatch Track.

3.4.1 Initial Analysis

An initial analysis using descriptive statistics was useful to better understand the information pulled from Dispatch Track. For our initial analysis, the Delivery Report proved to be the most useful. The Delivery Report describes individual stops on each route, providing information about customers, routes, planned and actual delivery times and packages. One Delivery Report can contain information about all routes in a location for a whole month, making it a good source of information for our initial analysis.

The analysis was done using data from August to November 2019. Pilot introduced several new initiatives, such as aiming to service a minimum of 10 stops per route, to improve performance after the summer. We decided to assess their performance from the start of these initiatives. Locations were analyzed separately to identify local differences. Cleaning was done in accordance with Figure 3.2.

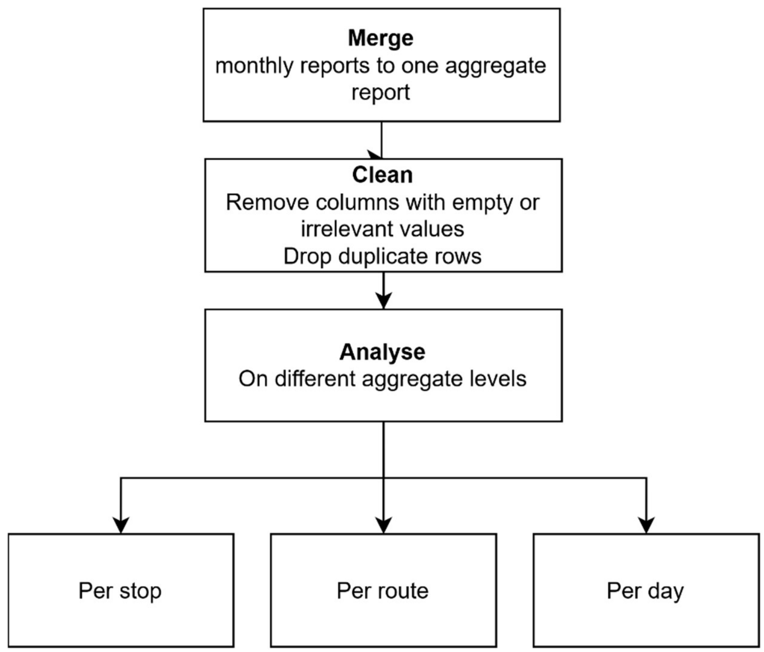

Figure 3.2 Cleaning for initial analysis

The findings from the initial analysis are presented in Chapter 4. After discussing the findings with Pilot, we agreed that the capstone would focus on establishing a framework for performance assessment and predicting parameters affecting performance. it was decided to take a dual approach to the further analysis. The first part would give Pilot insight into the performance of their LMD operations. The second part would improve planning accuracy through improving the underlying planning factors used in Dispatch Track.

3.4.2 Performance Assessment

3.4.2.1 Introduction

Before building the performance assessment tool, we made two important decisions. First, we decided to make the assessment tool reusable even after the capstone is concluded. Building a reusable tool means that Pilot can add new information and re-run the analysis. A challenge with this approach is that the data processing and analysis has to be designed in a way that is user-friendly and easy to repeat. The upside of this approach is that the tools and models built in this capstone can provide lasting value to the company.

Second, we decided to combine Python for data cleaning, merging and analysis, with Tableau for visualization of the final result. Doing the data manipulation in Python makes it possible to create a script that can be re-run, without any changes to the script, whenever new data is added. The python script creates a csv (comma-separated values) file that can be read by both Tableau and any spreadsheet software (such as MS Excel). Dispatch Track offer some visualization, but the information is highly segmented, and it is difficult to compare locations. In addition, Dispatch Track does not allow for manipulation of data or calculation of metrics that are not already available. Tableau allow for a user-friendly visualization of the assessment metrics, as well as creation of dashboards that can be designed to show whatever information is considered interesting. Humans are good at interpreting data visually, identifying discrepancies or inefficiencies is easier using visualization. As the Tableau file reads the CSV file created by the Python script, the visualization will update any time new information is added. In addition, Tableau allows for easy filtering of data, allowing Pilot to assess individual locations and desired time periods more easily. Making the cleaned data available as a CSV file improves user-friendliness since Pilot personnel can perform further analysis without any knowledge of programming in Python.

To support the reusability of the script and allow for additional reports to be added, a file structure was built for organizing reports, scripts and results. For visualization and detailed description of the file structure, see Appendix B: Capstone Analysis File Structure.

3.4.2.2 Data Selection and Cleaning

To assess the desired performance metrics, the following reports from Dispatch Track were used:

- Delivery Reports – provides information on each individual stop, such as volume delivered, customer information, scheduled and actual delivery times.
- On Time Reports – provides time status on orders relative to planned status (early, on time, delayed, scheduled or cancelled) per day and DC
- Order Completion Reports – provides information on the completion status of orders (in progress, completed, exceptions or cancelled) by date and DC.
- Route Time Reports - provides total scheduled and actual service time for all routes on a date.

Relevant reports are placed in their respective subfolder based on DC for data cleaning and processing. The major steps of the data processing is illustrated in Figure 3.3, see Appendix C: Data Processing for further details on data processing.

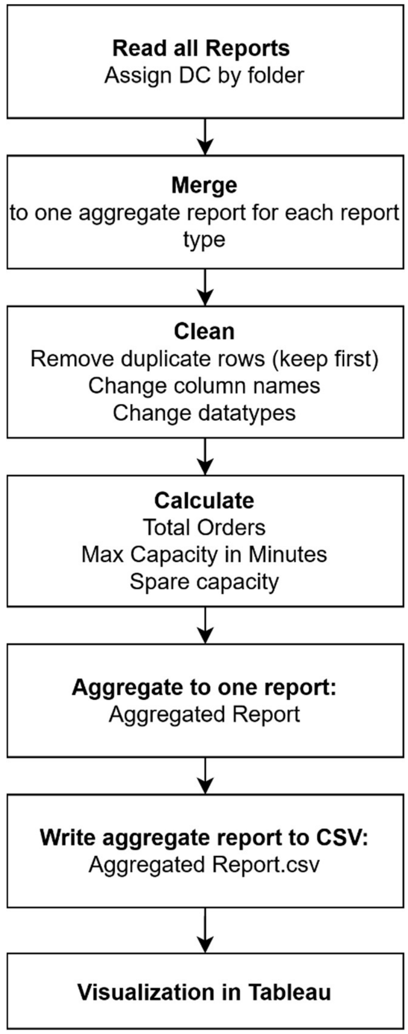

Figure 3.3 Performance Assessment Data Cleaning and Processing

The Tableau dashboards and associated findings are presented in Chapter 4. Next, we approach how to improve planning accuracy to improve operation performance.

3.4.3 Drive Time Analysis

3.4.3.1 Introduction

The analysis of drive time is based on comparing scheduled and actual drive time durations for different locations to calculate appropriate TSF values. Descriptive statistics is used to consider TSF over different variables and identify useful patterns or relationships. The entire analysis is done in one Python script. The result of the analysis is then used to guide the ML analysis and predictions.

3.4.3.2 Data selection and Cleaning

Dispatch Track does not contain a report with information on a per route level. The reports available provide either per stop information, or totals for all routes in a given day for each location. The stop information does not contain information about the load time, nor the linehaul time from the DC to the first customer and from the last customer to the DC. As a result, calculating driver time has to be done as a total drive time per location per day. This limits the level of detail the analysis is able to provide. At the same time, we believe the analysis will still be able to find trends and improve planning accuracy.

For TSF analysis, the following sources of information is used:

- Delivery Report - to find total stop time per day per location
- Route Time Report - to find total scheduled and actual service time per day per location
- Mileage Report - to find total mileage per day per location

For the stop time analysis data from January 2019 to February 2020 for all 5 locations was used. This allowed for sufficient data to build the complete Python script and an initial understanding of the performance of the prediction. Because the script is built to be reusable, Pilot can add or remove information as desired. Cleaning and processing of the reports is illustrated in Figure 3.4, see Appendix C: Data Processing for a more detailed description of the process.

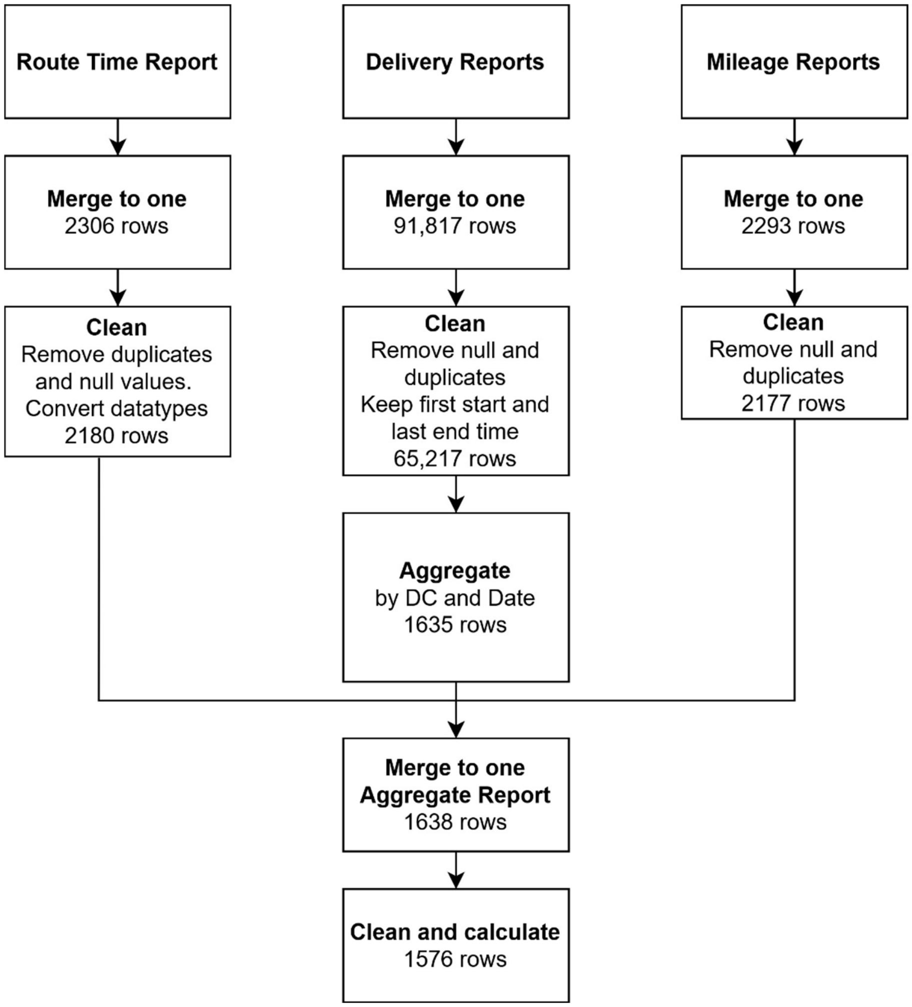

Figure 3.4 Data Cleaning and Processing for Drive Time Analysis

With merging, additional cleaning and calculations done, the next step in the analysis is to visually inspect the dataset to identify possible patterns. Using the visualization library Seaborn in Python (Seaborn.pydata.org 2020), the relationship between TSF and different variables are considered. The results of this analysis can be found in Chapter 4.3.3.

3.4.4 Stop Time Analysis

3.4.4.1 Introduction

The analysis of drive time is based on comparing scheduled and actual stop time durations for different locations. Descriptive statistics is used to consider stop time duration over different variables and identify useful patterns or relationships. The entire analysis is done in one Python script. The result of the analysis is then used to guide the ML analysis and predictions.

3.4.4.1 Data Selection and Processing

For the stop time analysis, the Delivery Report found in Dispatch Track provide all the necessary information, such as order information (e.g. scheduled and actual stop duration), customer information (e.g. zip code and address), and package information (e.g. package volume and quantity of packages). For stop time analysis data from January 2019 to February 2020 for all 5 locations was used. This allow for sufficient data to build a complete Python script and understanding of the performance of the prediction.

Additional information about a specific account, containing information about the stop location (stairs/no stairs, insurance yes/no, etc.) was considered in addition to the information available in the Delivery Report. Unfortunately, this additional information was only available for one customer account and a small portion of the overall dataset. During regression analysis, this data appeared to be of less relevance to predicting stop time duration. As a result, this information has not been included in the final analysis.

Cleaning and processing of the reports is illustrated in Figure 3.5, see Appendix C: Data Processing for a more detailed description of the process.

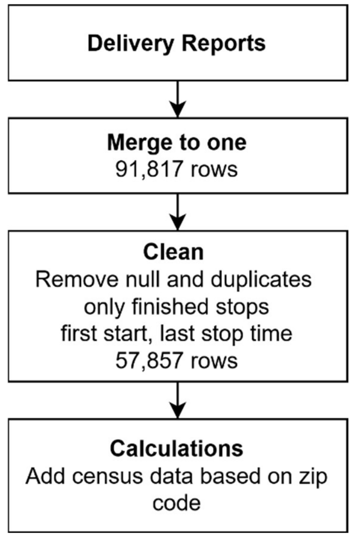

Figure 3.5 Data Cleaning and Processing for Stop Time Analysis

With data processing done, the next step in the analysis is to visually inspect the dataset to identify possible patterns and relationships between stop time and different parameters. The results of this analysis can be found in Chapter 4.3.4.

3.5 Analysis Using Machine Learning

Using the insight provided through assessing the performance of Pilot's current LMD operations with descriptive statistics, further analysis was done using ML methods. Descriptive statistics tell how the company is performing, but not what parameters affect performance and what values for these parameters should be used to improve planning. Since Dispatch Track is a 3rd party software, neither the students nor Pilot has access to changing the system directly. The only way to affect routing is to change the input parameters.

Because drive time and stop time is affected by different parameters, the ML analysis is done for each separately. For both analyses, the goal was to identify relevant parameters that affect stop or drive time. We also predicted values to improve planning accuracy compared to actual stop and drive duration.

3.5.1 Drive Time Analysis

For drive time analysis, the descriptive statistics indicate some strong linear relationships between drive time and certain parameters. LR was therefore used to do this analysis. Another benefit of using LR was that the results are easy and fairly intuitive to interpret and use. This increased the likelihood that Pilot will be able to reuse the analysis at a later time. The analysis is done in one Python script and the result is stored as a csv file with information necessary for planners to calculate a suggested TSF to use for any given route.

3.5.1.1 Setup for Drive Time Analysis

After the initial data processing, the only preparation of the drive time dataset required is to setup binary and one-hot-encoded variables where appropriate. Filtering and other manipulations done with the dataset as part of the analysis is presented in Chapter 4.4.1

3.5.1.2 Linear Regression

The Python library Statsmodels is the tool used to run the LR in this analysis ("Linear Regression — Statsmodels" 2019). This library allows for the use of LR as well as performance metrics such as p-values and adjusted R^2. LR is used to consider TSF against all the available parameters. Based on the first LRs, mileage, month and location appeared to be the most relevant. Therefore, further analysis filtered the dataset based on location and month.

The performance metrics used from the LR are adjusted R^2 (amount of variance explained), number of observations (more is preferred) and p-value (relevance of variable).

3.5.1.3 Testing Linear Regression Results

To test the result from the LR, initially in the Python script, a train and a test period is established. The LR analysis is run on the train period. Then the coefficients calculated is used to estimate TSF for each row in the test set by multiplying the coefficient with the total mileage. This TSF is used to calculate an estimated drive time. To compare the performance of the estimate to the scheduled time, the error for both are calculated by subtracting the actual duration. In addition to error from actual duration, the absolute error is calculated. To make comparison easier, the mean of absolute and actual errors for scheduled and estimated drive time, as well as the difference between the two, is calculated. This allows the user to see how much, if any, the estimated TSF improves planning accuracy for drive time.

3.5.2 Stop Time Analysis

3.5.2.1 Introduction

For stop time the descriptive analysis showed more complex relationships than for drive time. Identifying strong patterns or relationships appeared more difficult and complex. For this analysis we used LR (as with drive time), as well as DT and RF. LR could still be useful as it is more intuitive and easier to apply. However, because the relationships appeared to be more complex, DT and RF could be better suited to identifying appropriate relationships in the dataset. The analysis was used to consider features related each stop in our dataset and identify attributes that makes it possible to predict the stop duration. The analysis is done in Python. The result of the analysis is better understanding of what drives stop duration and how to better predict the stop duration.

3.5.2.2 Setup for Machine Learning

Using the understanding of the stop time dataset established through descriptive statistics analysis, the next step was to prepare the dataset for ML analysis. Using the parameters found relevant in through descriptive analysis, the dataset was filtered to a set that contained information about all these parameters. Next, one-hot-encoding and converting to binary variables is done where appropriate. The results of the filtering and manipulation done as part of the analysis is presented in Chapter 3.5.2.

3.5.2.3 Machine learning: Linear Regression

The method used to run and test linear regression to predict stop time is the same as described in Linear Regression3.5.1.2 Linear Regression and 3.5.1.3 Testing Linear Regression Results. The dataset is split on location and service level, and the variables used to predict actual stop times are quantity pieces, volume delivered, population density, stop number and peak day. The results of the analysis can be found in Chapter 4 RESULTS AND ANALYSIS.

3.5.2.4 Machine Learning: Decision Tree and Random Forest

To run DT and RF the Python library Scikit-learn is used (Pedregosa et al. 2011). Scikit-learn contains all the components needed to prepare, run and analyze the outcome of the analysis. In addition to Scikit-learn, the resources found on the website "Towards Data Science" was used to establish the framework and assessments metrics (Koehrsen 2018). For both DT and RF data from all 5 locations, starting on January 1. 2019 to February 28. 2020, was used to do the analysis. DT was useful to see that the analysis ran properly and build an understanding of the outcome of the analysis. When this understanding was established, moving on to RF helped to prevent overfitting and to further explore the dataset. Both classification and regression were tested, however only regression proved to provide usable results.

The method for running the analysis was the same for both DT and RF. First "Actual Duration" is defined as the label we want to predict. Then the dataset is randomly split into a train and a test label, where train makes up 80% of the whole dataset. A baseline prediction, using the scheduled duration, is calculated to enable comparison with the end result of the analysis. Then the DT and RF is run, with different variables, depths and number of trees (only relevant for RF). A prediction model is built by the algorithm, that is then used on the test portion of the dataset to evaluate the model's performance. The performance of the model is assessed by comparing the baseline error (difference between actual and scheduled stop time) with the error of the model (difference between actual and modeled stop time). To improve the prediction, the dataset is split on location and service level. The result of the analysis is presented in Chapter 4 RESULTS AND ANALYSIS.

3.6 Summary of Methodology

Helping Pilot improve their LMD operations, we started by establishing a good understanding of their current operations. Together with Pilot, performance dimensions, KPIs and appropriate aggregation levels were defined as a framework for the analysis. Once the performance framework was established, descriptive statistics was used to analyze the Performance of Pilot's current operations. The insight from this analysis was then used to further analyze Pilot's operations using the ML methods LR, DT and RF. The ML methods provided further insight into what parameters affect performance and how. In the next chapter the findings are presented and discussed.

4 RESULTS AND ANALYSIS

This capstone project is helping Pilot Freight Services improve their LMD operations through improved insights into own performance and better planning. After an introduction of the problem, review of relevant literature, and a description of the methodology, it is time to present and discuss the analysis and findings. The analysis is done in the context the performance framework established. The four steps of the analysis will be described: (1) Understanding Pilot's current operations, (2) establishing the performance framework, (3) Analysis using descriptive statistics, and (4) analysis using ML

4.1 Understanding Pilot's Current Operations

Knowing how Pilot currently runs their operations is vital to understanding how to improve their operations. When an order is made, the sales team places the order in the Dispatch Track system. The order contains information about the customer, the delivery place including time window, and information about the package. The time windows are a time period defining when the customer want the goods delivered. Time windows impose an important constraint in route planning. In addition, the customer chooses desired service level, such as door delivery of packages or full assembly of the product. The day before delivery, operators at Pilot use Dispatch Track to build the following day's routes. In addition to changing the orders provided by the sales team, the operators can affect constraints such as driver speed, total driver service time, and expected delivery unload time. A complete list of all the parameters the operators can manipulate in Dispatch Track, with description is provided in Appendix A: Table of planning operation parameters in Dispatch Track.

The operator then uses Dispatch Track's built-in function to build routes, based on the orders and the restrictions provided. Dispatch Track offers to optimize the route based on shortest time, shortest distance, farthest point first, or nearest point first. After the routes have been built, the operator can then make changes to the routes based on their own knowledge and evaluation of the routes built.

When the routes are complete, the operator locks the routes, indicating that they are finalized and ready to be sent to the drivers. The following day, the routes are sent to the driver's mobile devices. These devices are then used to keep track of orders and routes, as well as report progress back to the operators. All the information from the mobile devices is sent back to Dispatch Track. The operators can follow the routes and pull different reports from Dispatch Track to evaluate performance. However, the reporting tool in Dispatch Track is limited. The reports are very detailed, but they can be aggregated only for a period of a single day to a month, depending on the type of report. The lack of aggregation makes it difficult to evaluate performance over time.

From our initial observations, we can tell that Dispatch Track plays a crucial role in Pilot's LMD operations. The company organizes all their deliveries and routes in the system. As a result, the reports provided by the system give good insight into how their current operations work. The main issue is that Dispatch Track is not well set-up for performance assessment and evaluation. The system provides information on some performance metrics, such as on time delivery and order completion, but in separate reports. Aggregation and comparison of different locations is also difficult. Further, as all planning is done in Dispatch Track, using the routing software, the accuracy of the parameters used for routing highly affects planning accuracy. Currently, Pilot planners are using their experience and best guesses, and the process differs between locations.

4.2 Establishing Performance Framework

After building a good understanding of Pilot's current operations, the second step of the project is to establish a performance framework to assess the performance of current operations.

4.2.1.1 Service Dimensions

Together with Pilot, we defined two performance dimensions:

1. **Service Level** – this is the most important service dimension for Pilot and focus on the customer facing part of Pilot's LMD operations. Service level is related to the customer

experience and how the customers' needs are taken care of. For Pilot this include completing orders and delivering products within the delivery time window decided by the customer.

2. **Efficiency** – focus on how well the LMD operations are run internally in Pilot. Once appropriate service levels are achieved, Pilot strive to operate as cost-efficient as possible. This includes utilizing as much of the available driver time as efficiently as possible, serving as many stops as possible on any given route. Efficiency is important for the company to be profitable.

4.2.1.2 KPI's

Within the two performance dimensions, we defined KPI's that can be used to assess the company's performance within each dimension. Through discussions with personnel from Pilot, a few KPI's that can be evaluated through analysis of reports from Dispatch Track is identified and defined:

1. **Within Service Level:**
    a. **Order completion rate** – the number for orders that Pilot deliver for their customers. Order completion can be measured both as an absolute order (to assess total sales and size of operations) and as a ratio of total orders (to assess customer service and performance). Completing as many orders as possible is important both to be profitable and to provide sufficient customer service. Dispatch Track distinguish between orders that are completed, exceptions (most often not completed), in progress and not completed.
    b. **On time delivery** – for each order, a delivery time window is defined by the customer. Delivering within this time window is important both for customer satisfaction and for actually completing the order. Delayed orders are often associated with failed deliveries and higher costs associated with customer compensation (such as waived delivery fees). From Pilot's experience, early deliveries are rarely a problem as drivers can call ahead to the customer and make certain that the customer

is able to receive the order. On time delivery is measured as a rate of how many of the completed orders are delivered either early or within the customer time window.

2. **Within efficiency:**

   a. **Driver utilization** – measure how much of the total available driver service time Pilot utilize. Regulations limit the number of available hours to 14 hours per day. Pilot's policy is to stay within 12 hours per day per driver. Driver utilization is measured as a percentage rate of average driver time utilization compared to total available driver service time.

   b. **Spare available driver time** – Subtract the total used driver service time in a location on a given day, from the total available driver service time for the same location and day. Measure excess driver time capacity.

   c. **Excess driver capacity** – Measure the overcapacity of drivers for each location on a given date. This measure compares the total available driver service time to the total used driver service time for each location, to calculate the excess driver capacity in any location on a given date.

   d. **Daily excess stop capacity** – measure the potential number for additional stops that could be served in a location, by calculating the average time spent to service a stop (total service time for a location divided by the number of stops served). The total unused driver service time in the location is then divided by the average stop service time to estimate how many more stops could be served in that location for a given day. This KPI can be used to indicate growth potential for any location. KPI is given in hours.

   e. **Number of stops per route** – measure the number of stops or customers serviced by each route. As a general rule, Pilot aims to service an average of 10 stops per route. Servicing a higher number of stops per route is often associated with more efficient and profitable LMD operations. This information is directly available in Dispatch Track.

4.2.1.3 Aggregation level

For analysis, Pilot provided data access to 5 locations: Phoenix, Minneapolis, Chicago, Birmingham, and Portsmouth. Because of the difference in both order size and geographic location, all the KPI's are aggregated separately for each location. The performance assessment tools built for Pilot allow Pilot to aggregate performance over any time period they desire. The performance metrics are given as averages over the chosen time period. By default, the time period is given by month.

Analyzing the data on three different aggregation levels offer different perspectives. The per stop perspective provide detailed understanding of each stop. It makes it possible to compare performance with individual factors such as volume, customer, service type and type of goods delivered. These details are important to understand fluctuation in performance, but too detailed for an initial analysis seeking a more high-level understanding.

The per route perspective allow better understanding of the difference between planned and actual duration of each route. It makes it possible to compare planned and actual total drive times, and not only stop duration as on the per stop level. Information aggregated to a route level, allow assessment of both planned and actual overall route time, as well as total stop times and total drive times.

The per day perspective is helpful to better understand the overall effectiveness of Pilot's operations. By aggregating information per day, we are able to compare total number of packages, stops, and trucks to the available resources. Using these three aggregation levels Pilot's performance is analyzed.

4.3 Findings From Analysis Using Descriptive Statistics

4.3.1 Initial Analysis

Two key findings are identified in the initial analysis. The first finding is that Pilot appears to be underutilizing the available driver service time. Each driver has a total service time of 12 hours. Analyzing both actual average duration (Figure 4.1) of routes, it becomes clear that each route has somewhere between one and a half to four hours of unused service time.

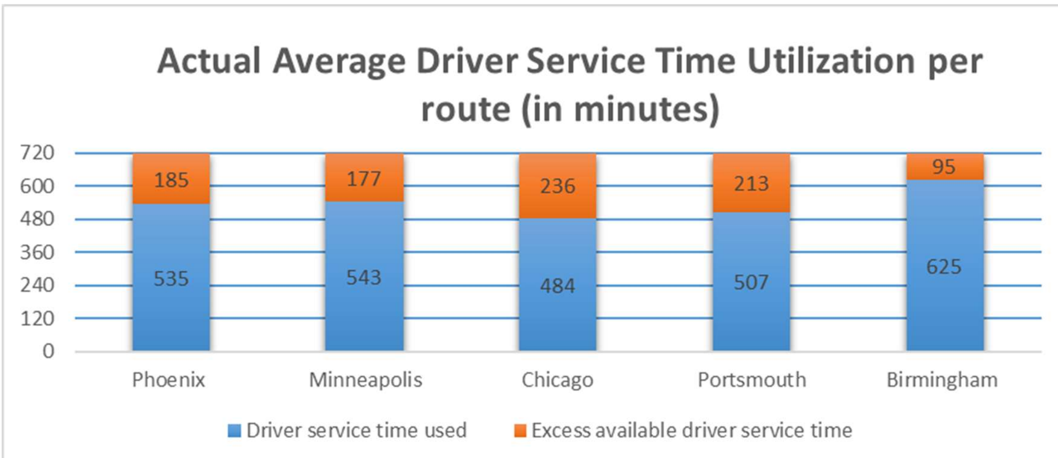

Figure 4.1 Key Finding 1: Actual average driver utilization per route (in minutes)

Combining this with the average numbers of routes per day per location (Table 4-1), we find that each location may be able to reduce the number of routes by 1 or more. Phoenix, as an example, has an average spare capacity of 185 minutes or a little over 3 hours. Multiplied by an average of 4.4 routes per day, we find that on average Phoenix has an excess capacity of 814 minutes. Again, one driver has a total service time of 720 minutes. This means that Phoenix even with one less driver per day, will have excess capacity.

	Routes per day		Stops per route	
Location	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**
Phoenix	4.5	1.2	8.8	2.9
Minneapolis	6.2	2	7.3	3.7
Chicago	11.3	4.3	7.6	2.6
Portsmouth	3	0.6	6.6	2.4
Birmingham	2.1	1	6.7	2.2

Table 4-1 Average routes per day and stops per route for each location

Perhaps even more interesting is that the excess capacity is even greater when analyzing the planned or scheduled driver utilization (Figure 4.2). This indicates that Pilot is not only not fully utilizing their drivers, they are not even planning to do so.

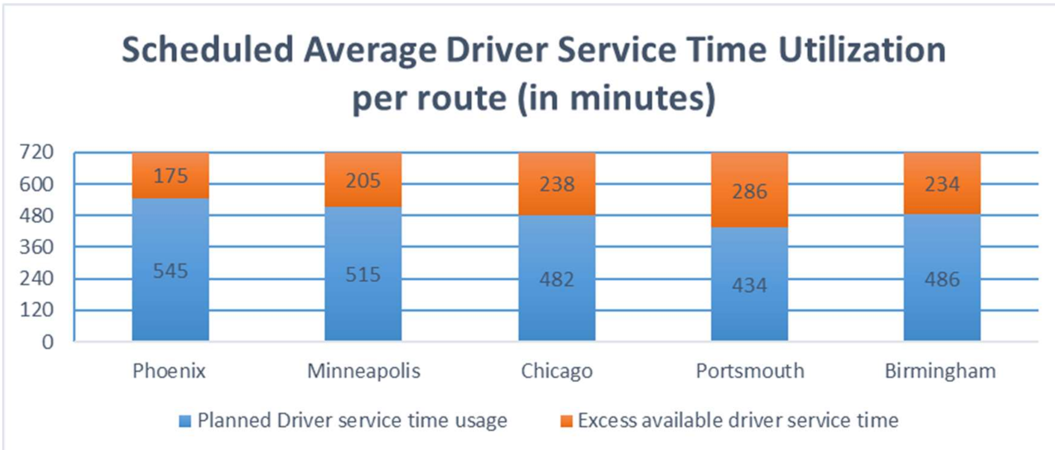

Figure 4.2 Key Finding 1: Scheduled driver/truck utilization per route

One explanation for why Pilot might not plan to use all of their driver's available service time may be that routes have reached the goal of 10 stops per route. To account for this, the average number of stops per route were analyzed. Table 4-1 rejects this idea, and reinforces the suspicion of underutilization of driver time, as none of the locations have reached Pilot's goal of 10 stops per route on average.

Another explanation may be that customer time windows are too constraining. Pilot cannot utilize the full 12-hour driver service time if all customer needs have to be serviced within a 10-hour time window. Through discussions with personnel in Pilot, we discovered that the delivery time windows were not absolute and could be moved by planners in dialogue with customers. This suggests that the constraint of time windows does not prevent Pilot from utilizing the full 12-hour driver service time.

Based on these findings, we concluded that Pilot is underutilizing their available driver service time. After discussing the findings with Pilot it becomes clear that part of the problem is lack of visibility into the performance of their LMD operations. The Dispatch Track system works well for planning, but management find it hard to assess and monitor performance as the reports available in the system is fragmented and hard to analyze. As a result, we decided to create a performance assessment tool to help managers in Pilot assess their performance. The performance assessment tool will give insight into Pilot's performance in the previously defined performance dimensions.

The second key finding is related to how accurate planned routes are compared to actual routes. This is analyzed by comparing the averages of scheduled and actual route duration for each location. The results are displayed in Figure 4.3.

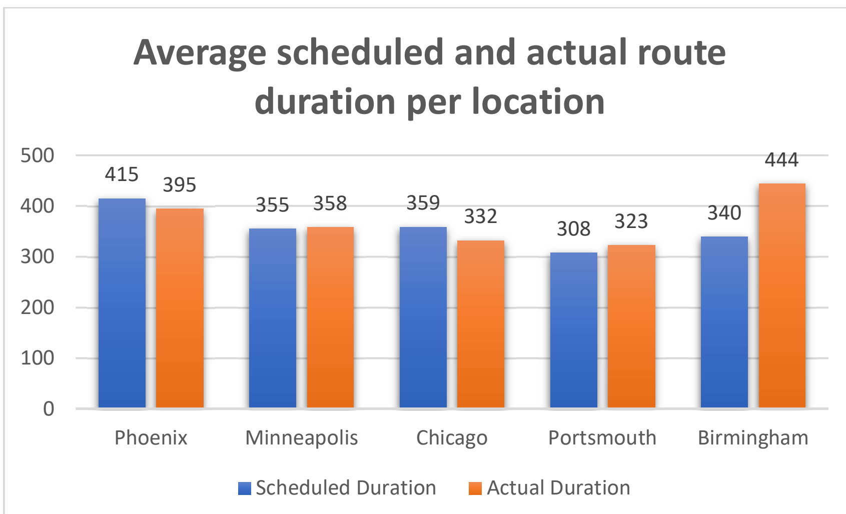

Figure 4.3 Key finding 2: Average Scheduled and Actual Route Duration per Location

When considering average total stop times and drive times, the differences become even clearer. In discussions with Pilot we find that more accurate planning factors would help improve efficiency and their LMD operations. As a result, we decide to do further analysis to identify parameters that affect Pilot's LMD operations and predict appropriate values to use for planning in Dispatch Track. Further analysis of routing in Dispatch Track reveal that route time is composed to two main components: stop time and drive time. In Dispatch Track drive time can only be adjusted by changing the TSF. Stop time can be set manually for each stop, making it possible to use more parameters to predict more accurate stop times. Improving stop and drive time predicting will help improve planning accuracy.

With the insight from the initial analysis, we are better equipped to do more analysis using descriptive statistics and ML. The initial analysis gives further insight into Pilot's current operations, and what factors may cause inefficiencies in their LMD operations. Next, we move into descriptive statistics analysis to assess the performance of Pilot's current operations.

4.3.2 Performance Assessment

The purpose of performance assessment is to give Pilot Freight Services insight into their current operations. The insight is provided in the framework of the performance dimensions service level and efficiency, with corresponding KPIs. The Pilot Performance Assessment Tool (P-PAT) is built as a re-usable tool to provide this insight. P-PAT gives this insight by merging and cleaning the relevant reports gathered from Dispatch Track, before visualizing the metrics in three dashboards in Tableau. The dashboards are interactive, allowing Pilot to filter on desired locations and time periods.

4.3.2.1 Service Level

In order to assess the service level performance dimension, two dashboards have been created: (1) On time delivery and (2) Order Completion.

The first dashboard is on time delivery (Figure 4.4). For each location, the total number of cancelled, delayed, early, on time and scheduled stops are displayed. In addition, a on time delivery rate for the desired filtered time period is provided. By providing both measures, Pilot can assess both absolute and relative on time delivery rate. For Pilot, both early and on time is accepted as "on time", as their drivers can call ahead to the customer and agree on an early delivery if possible. Looking at the on-time delivery rate for the period November 2019 to January 2020, we find that the on-time rate is between 80 and 90%. In addition, early delivery accounts for a significant portion of the on-time deliveries.

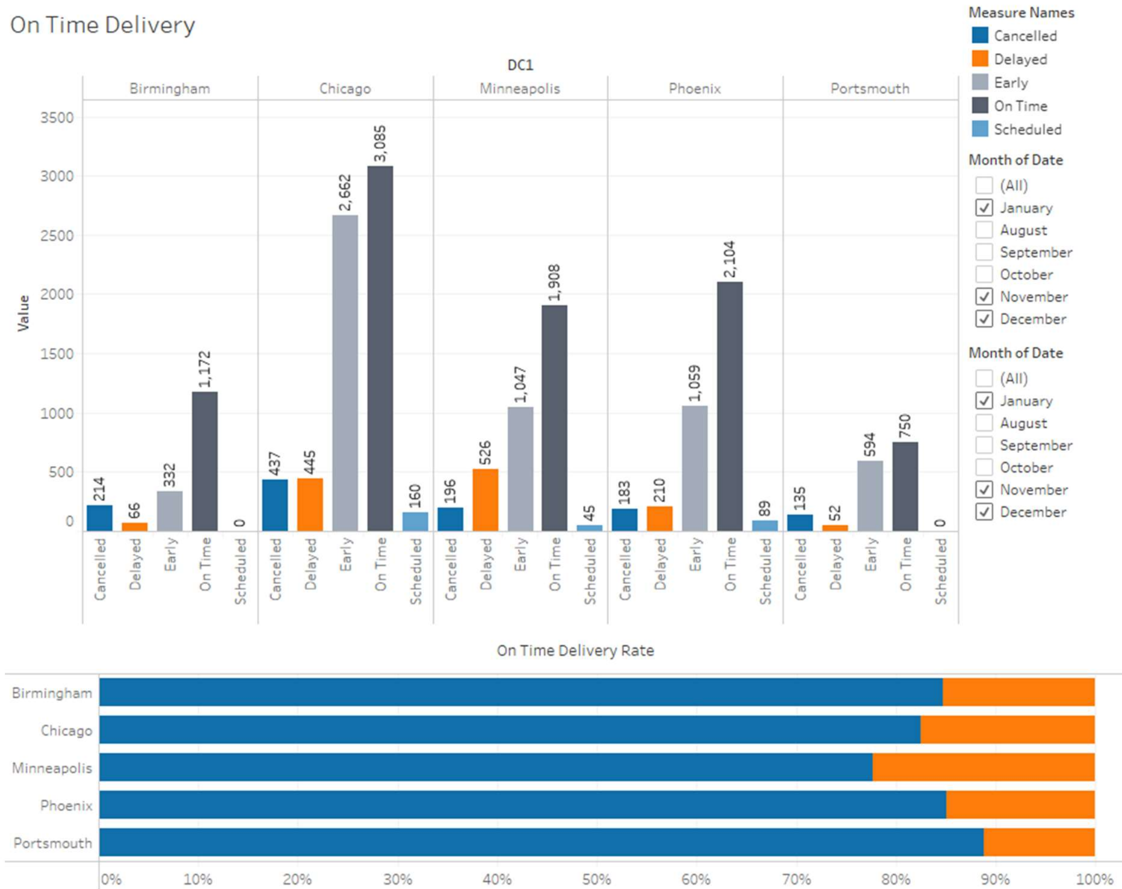

Figure 4.4 Service Level - On Time Delivery Dashboard in Tableau

In the order completion dashboard (Figure 4.5) number of completed, not completed, exceptions and in progress orders are displayed. An order completion rate bar chart is also provided, to show the percentage of completed orders compared to the total number of orders. The dashboard is designed with filters to allow for filtering by month. For the time period November 2019 to January 2020 Pilot's order completion rate is about 90% across all locations. The best performing location is Phoenix with a completion rate of about 92%. The worst is Birmingham with a rate of about 85%. Another interesting finding is that Chicago appear to have a lot of exceptions.

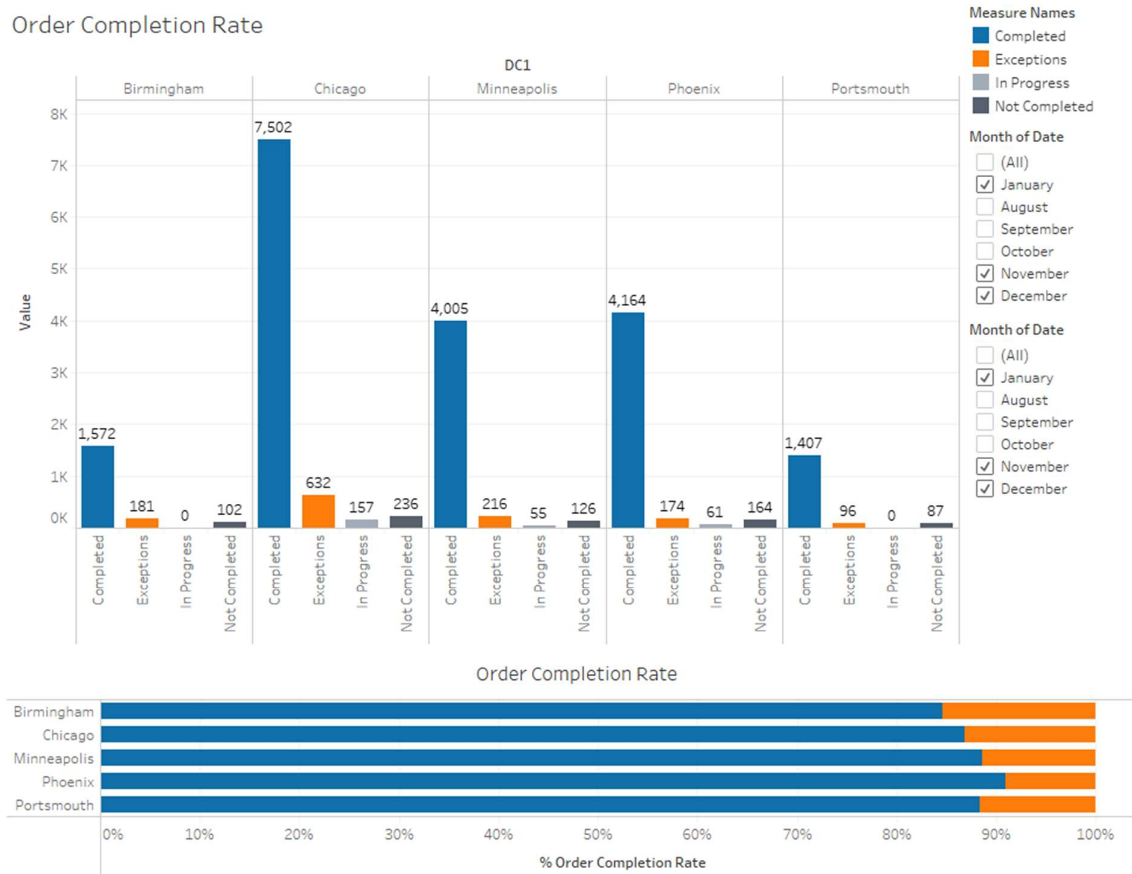

Figure 4.5 Service Level - Order Completion Rate dashboard from Tableau

4.3.2.2 Efficiency

The third and final dashboard address the service dimension efficiency (Figure 4.6). This is displayed using a driver utilization rate, showing the average of total utilized driver service time for a given time period compared to the average of total available driver time in the same time period. For the time period November 2019 to January 2020, the driver utilization rate varies greatly between locations. Chicago is by far the best location in terms of driver utilization with a rate of more than 90%. Birmingham is also performing fairly well with a rate of about 85%. Phoenix and Minneapolis has a rate of about 60%, while Portsmouth has a utilization rate of less than 50%.

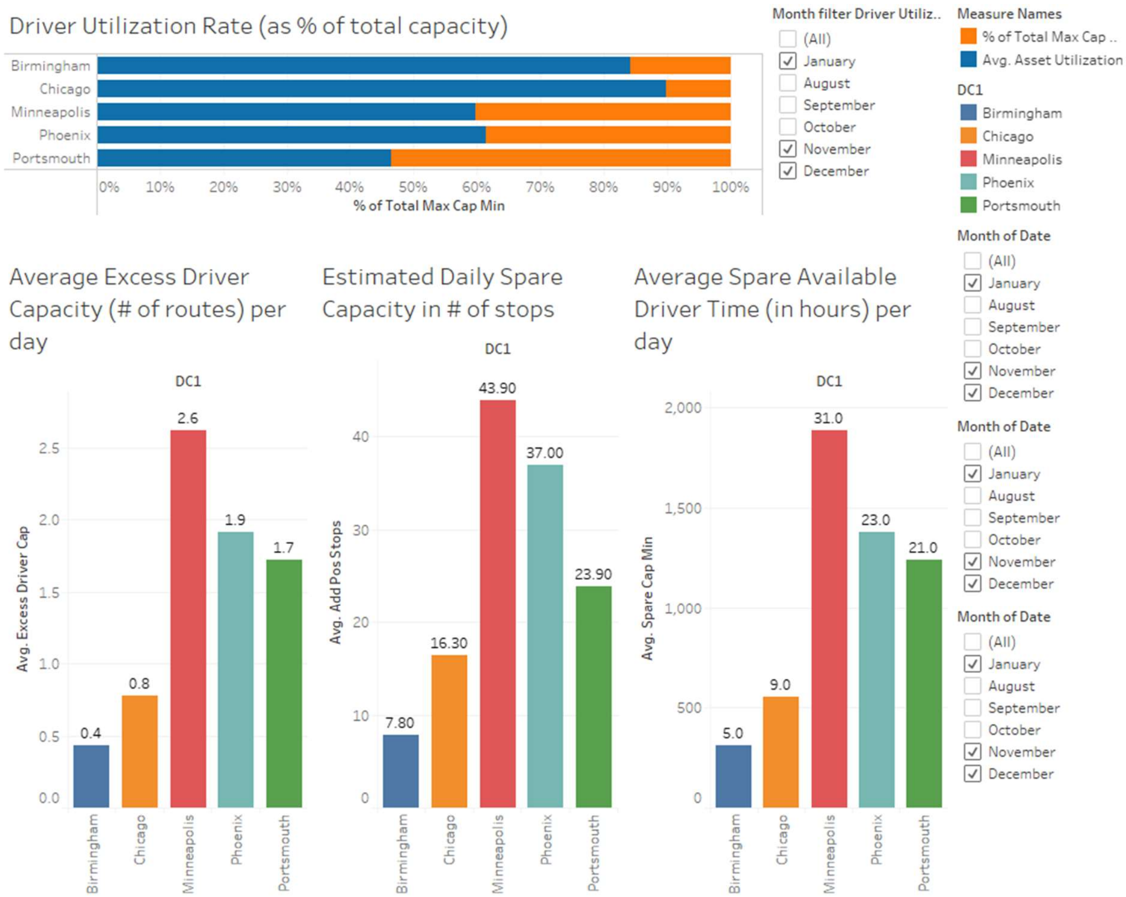

Figure 4.6 Efficiency Dashboard from Tableau

The excess driver capacity, daily spare capacity in terms of number of stops and spare available driver hours give some more context to the driver utilization rate. Again, we find that Minneapolis, Phoenix and Portsmouth seem to have significant underutilization of drivers. Phoenix and Portsmouth can reduce the average number of drivers each day by nearly two. For Minneapolis, the number is more than two. This finding is reinforced by the available drier time in hours chart. Spare capacity in number of stops indicate that several more stops can be served in all locations, from 8 in Birmingham to almost 44 in Minneapolis.

4.3.3 Drive Time Analysis – findings

Using insight from the initial analysis and the performance assessment, we use descriptive statistics to identify variables that affect drive time. As mentioned in the methodology, this is done by visually considering TSF over different variables. More specifically, the relationship between TSF and location, mileage, day of week and month is examined. Figure 4.7 show the distribution of calculated TSF for each location. In this graph, the 5 different DC locations is displayed separately. As one can see from the graph, there are large differences between the different locations, indicating that TSF should be calculated separately for each location. We also note that the majority of the calculated TSF are above 1, indicating that most scheduled drive durations are longer than actual durations.

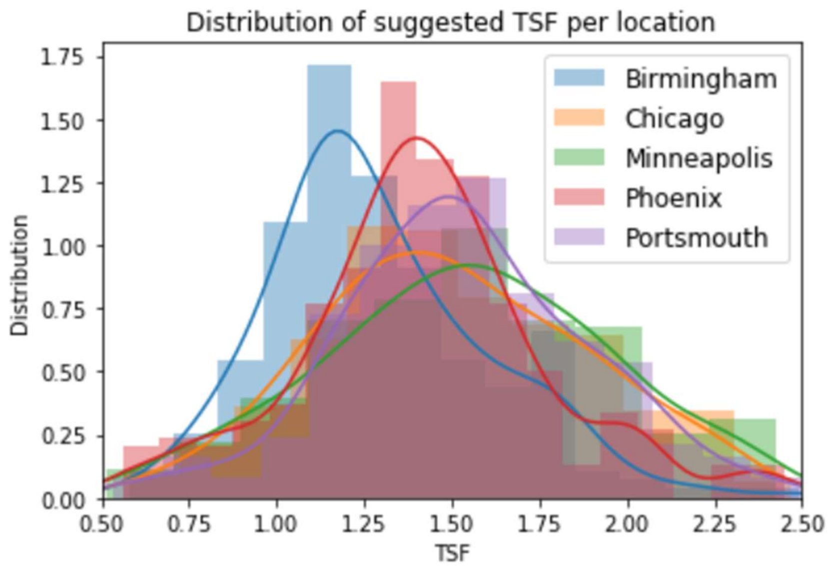

Figure 4.7 Distribution of suggested TSF per location

Figure 4.8 show TSF over Month where 1 = January and 12 = December. This graph indicates that in the period September to December, drivers are slower than the rest of the year. This might be connected to seasonal changes and adaption to winter conditions in several of the DCs.

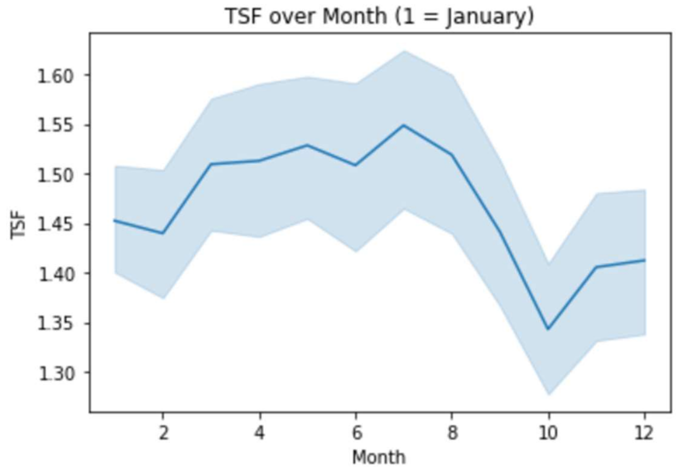

Figure 4.8 TSF over Month

Considering Mileage over TSF by location (Figure 4.9), we again see the difference between the different locations. The scatter plot also indicates that there is a relationship between mileage and TSF, where higher mileage could mean higher TSF. This would suggest that TSF could be predicted using mileage (miles).

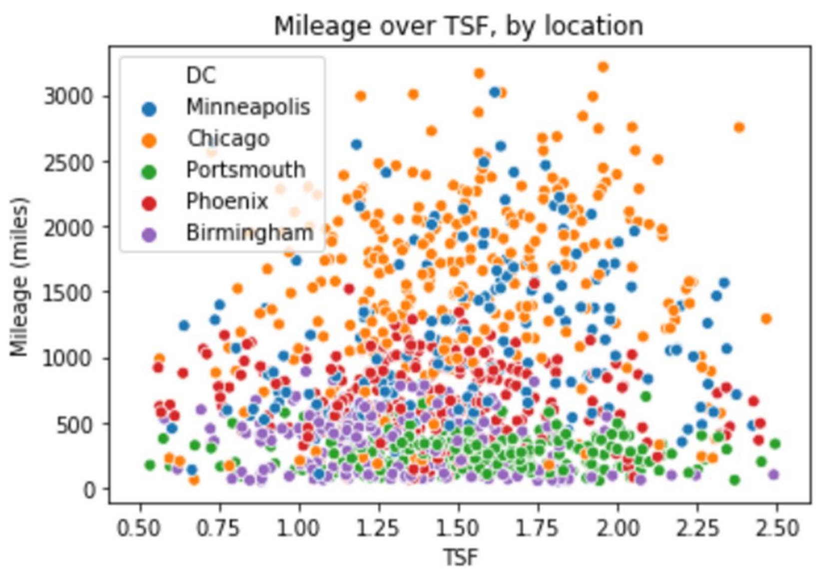

Figure 4.9 Mileage over TSF, by location

Based on the analysis of stop time using descriptive statistics, we decide to build a LR model to predict TSF using information about location, time of month/week/day and mileage.

4.3.4 Stop Time Analysis - Findings

For stop time analysis using descriptive statistics, we decided to try to identify relations between stop time and available parameters. Before starting the analysis with ML methods, an understanding of the dataset was established through graphic representation. The parameters considered were stop number, quantity delivered (in pieces), volume delivered, population density per square mile (by zip code), time of day, month, and distribution center (DC). The most noteworthy findings are presented in this chapter.

Considering scheduled and actual stop durations for all routes, show that both follow similar distributions. The distribution of scheduled durations follows a sawtooth pattern, this is likely due to planners using round numbers such as 10 or 20 minutes for a stop. The average scheduled time is 28.7 minutes, with a median of around 20 minutes, and the majority of outcomes between 20 and 30 minutes (25^{th} and 75^{th} quartile). The distribution of actual durations is smoother, with an average of 23.6 minutes, a median of 13 minutes and the majority of outcomes between 4 and 29 minutes (25^{th} and 75^{th} quartile). In both distributions there appears to be some right-side outliers that greatly affects the average. The outliers can be challenging when running the predictions, as they may heavily affect the outcomes. Based on the distributions show in Figure 4.10 we decided that reasonable minimum and maximum values for scheduled and actual duration is 5 and 60 minutes. That way we keep the majority of the distribution, while removing the most extreme outliers.

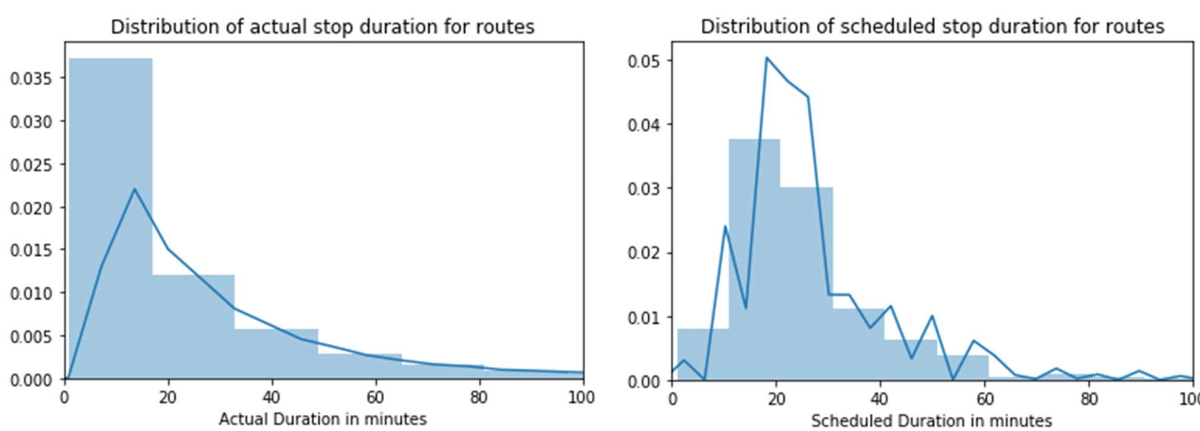

Figure 4.10 Distribution of scheduled and actual stop time duration for all routes

Figure 4.11 shows actual stop durations calculated for each location. The line within each box show the average value, while the box represents the quartiles (left side = 25th and right side = 75th quartile), while the lines going out from each box show outliers. The box plot shows large differences between locations, and further analysis will be done separately for each location.

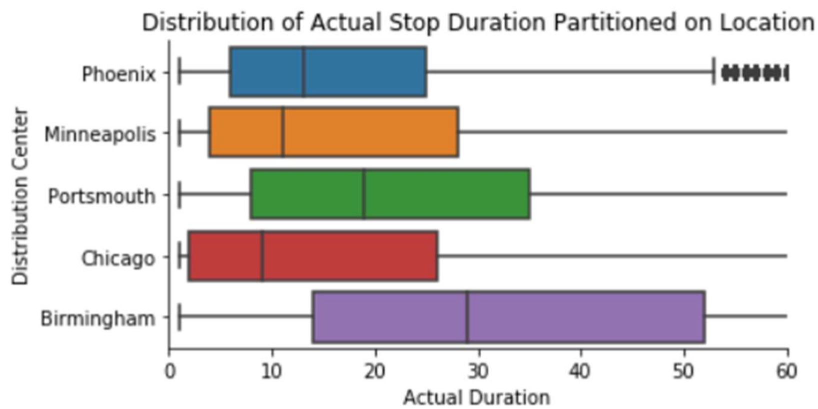

Figure 4.11 Actual stop duration partitioned on location

In addition to location, service level appears to be another variable where actual stop duration vary significantly. Pilot operate with two primary service levels: Mid-Tier and White Glove. White glove is considered the highest level of service. When examining the service level data, we found more than 20 different unique service types. To make the data more usable for analysis, the different service levels were grouped in either mid-tier, white glove or other (other account for about 15% of the total available data rows). Figure 4.12 show the significant difference between the service levels. As expected, Mid-Tier is associated with shorter stop duration than White Glove. The other category appears to have the widest variation of outcomes, an finding that is reasonable given the fact that it "collects" deviating service levels. Based on these findings, the further analysis needs to consider the different stop durations associated with each service level.

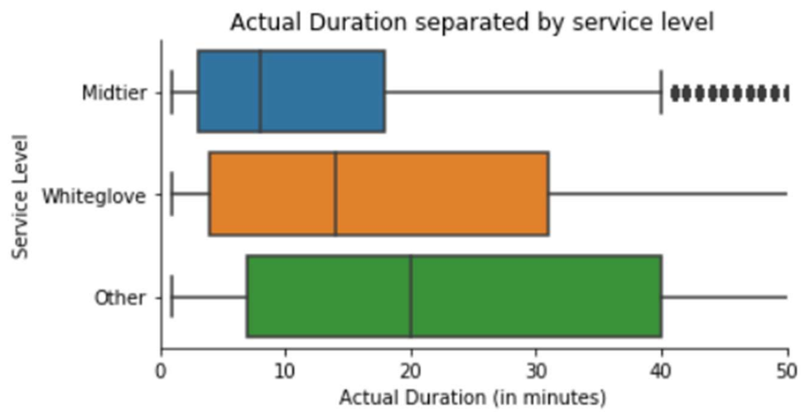

Figure 4.12 Actual Duration separated by service level

The stop number tells when a stop is served in a route. It will be closely connected to time of day for a delivery. Figure 4.14 show actual duration over time of day. The graph shows a clear peak in the morning hours (around 5-7), before the stop time is stable until late evening. Looking at actual duration over stop number (Stop#) (Figure 4.13) shows the same picture, where the early stops (likely to be early in the morning) are on average longer than stops later in the day. Figure 4.13 also show that there appears to be a strong variation in the stop duration for the later stops, indicating that it may not be a reliable predictor for these stops. Based on these findings, we find that either time of day or stop# should be included in the analysis.

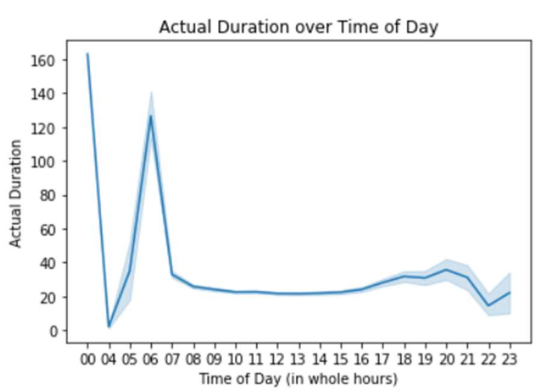

Figure 4.14 Actual Duration over Time of Day

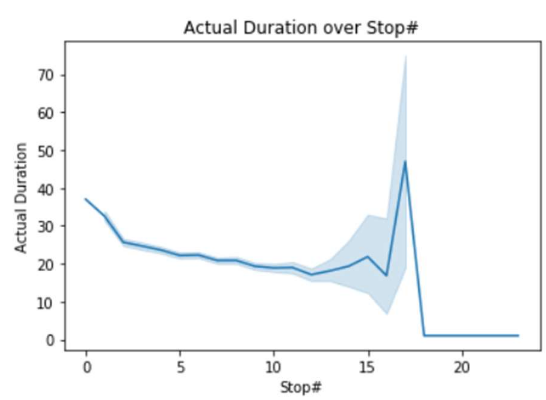

Figure 4.13 Actual Duration over Stop #

In addition to time of day and stop number, day of week appear to be of importance. Figure 4.15 show that Monday (day 0) and Sunday (6) appear to be "peak" days when actual stop duration is higher than the rest of the week.

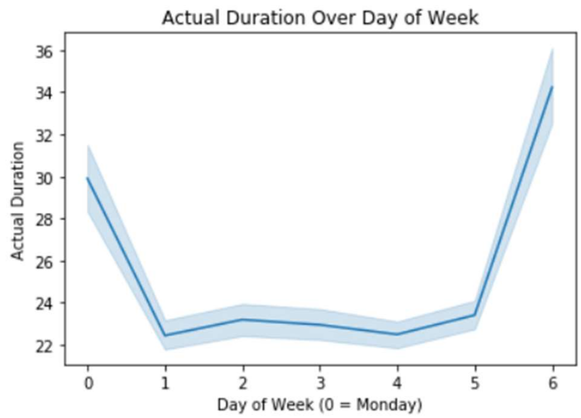

Figure 4.15 Actual Duration Over Day of Week

Quantity delivered describe the number of packages delivered at any stop. Figure 4.16 clearly indicate a linear relation between actual stop duration and quantity pieces delivered at a stop. Because of this graph, we decided to try LR in addition to DT and RF.

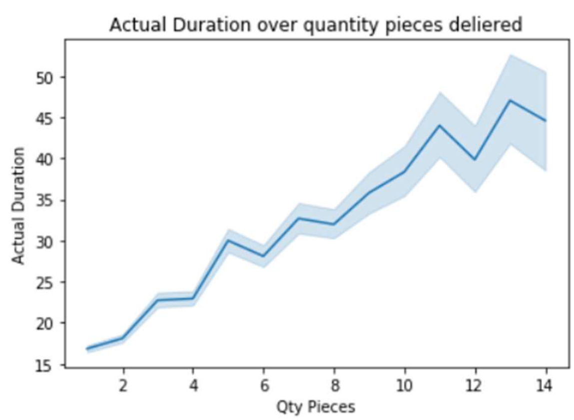

Figure 4.16 Actual Duration over Quantity Pieces Delivered

Viewing the data visually, it appears that further analysis should consider locations and service levels separately. Parameters that appear to be relevant for predicting actual stop duration appear to be quantity delivered, stop number, time of day and day of week. Additional parameters that will be considered are volume delivered, quantity picked up (returned goods) and population density and account.

4.4 Improve Planning Accuracy

4.4.1 Drive time Prediction

Predicting an adjusted TSF is done using LR. Because of the limited amount of data available, we decided to split the test and train data on different time periods. For testing we reserved the most recent month of data, while all other available data has been used for training the model. This was done to mirror the way Pilot will use the analysis in the future.

As suggested in the drive time analysis using descriptive statistics (Chapter 4.3.3), the LR model performs best when run on datasets that are separated on location and day of week. Running LR on the whole dataset, without any separation or filtering, with mileage as the only variable the adjusted R^2 comes out to 0.585. When separating the dataset on location, the adjusted R^2 improves to a range of 0.636 (Birmingham) to 0.830 (Chicago). Separating the dataset on day of week, the adjusted R^2 ends up in a range of 0.617 (Wednesday) to 0.782 (Thursday). When running LR on dataset that is separated on location and then day of week the adjusted R^2 for the train dataset improves even further. Table 4-2 show the result of the LR analysis on all locations and days. The coefficient represents the relationship between mileage and TSF. Using this coefficient, planners can estimate a TSF for any route by multiplying with the route total mileage.

Location	Day of Week	Coefficient	Adj R^2	Number of Observations
Birmingham	Wednesday	0.00284	0.848	7
Birmingham	Thursday	0.00223	0.794	20
Birmingham	Friday	0.00255	0.691	23
Birmingham	Saturday	0.00235	0.701	23
Birmingham	Sunday	0.00247	0.696	24
Birmingham	Monday	0.00255	0.704	26
Birmingham	Tuesday	0.00282	0.701	24
Chicago	Wednesday	0.00071	0.903	21
Chicago	Thursday	0.00074	0.908	29
Chicago	Friday	0.00071	0.888	30
Chicago	Saturday	0.00126	0.906	25
Chicago	Monday	0.00345	0.734	18
Chicago	Tuesday	0.00092	0.853	31
Minneapolis	Wednesday	0.00107	0.847	17
Minneapolis	Thursday	0.00123	0.917	22
Minneapolis	Friday	0.00095	0.812	28
Minneapolis	Saturday	0.00200	0.882	27

Minneapolis	Monday	0.00092	0.837	6
Minneapolis	Tuesday	0.00090	0.764	21
Phoenix	Wednesday	0.00127	0.856	21
Phoenix	Thursday	0.00132	0.911	25
Phoenix	Friday	0.00149	0.897	25
Phoenix	Saturday	0.00232	0.867	27
Phoenix	Monday	0.00370	0.645	21
Phoenix	Tuesday	0.00278	0.826	26
Portsmouth	Wednesday	0.00438	0.803	28
Portsmouth	Thursday	0.00589	0.852	4
Portsmouth	Friday	0.00388	0.836	28
Portsmouth	Saturday	0.00393	0.837	26
Portsmouth	Sunday	0.00429	0.869	25
Portsmouth	Monday	0.00454	0.841	25
Portsmouth	Tuesday	0.00587	0.850	5

Table 4-2 Result of LR on dataset separated on location and day of week

Further analysis and testing proved that there are significant differences between the months (as shown in Figure 4.8 TSF over Month). When testing the accuracy of the coefficients calculated on data for February 2020, a 3-month time period from August 1, 2019 to October 31, 2019 performs best. This strengthens the assumption than November, December, and January deviate from the other months in terms of TSF.

A challenge with the dataset, when splitting it on location, month, and day of week, is the low number of observations. To maintain the complete data set we tried to implement dummy variables for location and day of week. However, this does not appear to be able to capture the effect of these variables properly, creating much less accurate models. The reason is most likely that there are differences between the locations that makes the relationship between mileage and TSF different, such as infrastructure, climate etc. In addition, this approach makes the end result easier to apply for Pilot's planners.

Location	**Avg Sch Error**	**Avg Est Error**	**Difference**	**Sch MPE**	**Est MPE**	**Sch MAPE**	**Est MAPE**
Birmingham	201min	38min	163min	27%	10%	30%	21%
Chicago	881min	361min	520min	37.8%	23%	47%	35%
Minneapolis	364min	44min	319min	35.1%	26%	47%	45%
Phoenix	279min	-155min	434min	24%	-3%	32%	20%

Portsmouth	298min	111min	187min	60%	27%	60%	31%

Table 4-3 Test Performance of TSF prediction

Table 4-3 shows the performance of the estimated drive times calculated using an updated

TSF calculated using the relevant coefficient, compared to the scheduled drive times. The error is calculated by subtracting the actual drive time from the scheduled or estimated duration. Difference is estimated subtracted from scheduled. MPE and MAPE is calculated using these formulas:

Mean Percent Error: $$MPE = \frac{\sum_{t=1}^{n} \frac{e_t}{A_t}}{n}$$

Mean Absolute Percent Error: $$MAPE = \frac{\sum_{t=1}^{n} \frac{|e_t|}{A_t}}{n}$$

Equation 1 MPE and MAPE formulas

Where e = error, and A = actual stop duration. Table 4-3 show that the predicted drive times outperform the scheduled drive times. Still, the error is large in locations such as Chicago and Portsmouth. The negative value estimated error value for Phoenix further strengthens the fact that the estimated TSF are not able to capture all variance. Another challenge is the lack of observations. Using only 3 months of data and separating the dataset on location and day of week, leave very few observations to run the LR analysis. The model appears to be performing well, but variance will greatly impact the coefficients calculated. Pilot's planners should therefore use the coefficients as a helping tool and not as the perfect answer. As with our analysis, major outliers for estimated TSF should be carefully considered and possibly adjusted to more moderate TSF. Our findings and Figure 4.7, suggest that TSFs between 0.8 and 1.5 perform best.

The coefficients calculated using LR analysis on dataset separated on location and day of week can be used to calculate TSF that give drive times that are more accurate than Pilot's scheduled drive times. The coefficients are stored in a spreadsheet and can easily be multiplied with a routes total mileage to find the appropriate TSF. However, the calculated TSF is not able to capture all variance and the error of the estimates are still substantial. This is partly due to the small number of observations used to run the LR analysis. Next, we look into predicting stop time.

4.4.2 Stop Time Prediction

4.4.2.1 Filtering of dataset.

Table 4-4 shows the filtering operations and how they affect the overall dataset. The setup of the dataset causes a large challenge for the coming analysis: availability of usable data. The initial dataset contains 57,768 rows of data before filtering. After filtering, the usable dataset is reduced to only about 12,000 rows. 12,000 rows of data across 5 locations and 3 different service levels, will make it difficult to make good predictions with ML. The first reason making good predictions may be difficult is that irrelevant variations or "noise" in the dataset will have a larger impact on the prediction model. Second, the small dataset means that there are a lot of stops we are not capturing. Missing a lot of the stops increases the chance that any analysis may not capture all relevant features of the outcomes we are predicting.

Operation	Dataset Rows	Dataset Columns
Original dataset	57,768	10
Filtered dataset on 5 < Scheduled Duration < 60	53,648	10
Filtered dataset on 5 < Actual Duration < 60	33,863	10
Filtered dataset on Stop > 0	33,863	10
Filtered dataset on 0 < Volume Delivered < 200	12,276	10
Filtered dataset on 0 < Qty Pieces	12,051	10
Drop rows with null values	12,027	10

Table 4-4 Setup of dataset for Machine Learning

4.4.2.2 Decision Tree and Random Forest analysis

For stop time duration prediction the relationship between stop duration and the parameters are less clear than for TSF. Therefore, we decided to try DT and RF. DT is used to build the model and make initial predictions, before moving on to RF to improve accuracy and reduce risk of overfitting. Initially we tried both classification and regression, but quickly found that regression performs best. For classification we tried to bin the outcomes in buckets of stop-time duration and predict which one the duration would fall in. The challenge was that with classification, missing the bin is equally wrong regardless of with how much. Also, the bins did not appear to successfully capture how the parameters affect the outcomes. With regression we can predict continuous or numerical values, while classification is categorical (or discrete). From observing the dataset in Chapter 4.3.4 we observed that stop times vary what seem to be a continuous distribution, and this supports that observation.

Running RF analysis on the dataset we made three important findings. The first is that making accurate predictions with the current dataset is challenging, but the prediction perform better on datasets separated on location and service level. This supports the observations about the dataset done in Figure 4.11 Actual stop duration partitioned on location and Figure 4.12 Actual Duration separated by service level. When running RF on the whole dataset the baseline average error is 11 minutes, while the model's average error is 9.6 minutes. Running the model on dataset separated on location the average baseline error is in a range between 13.87 to 10.2 minutes, while the models average error ranges from 13.69 to 7.92 minutes. Table 4-5 show the performance of the RF model compared to the baseline (calculated using scheduled stop duration). For nearly all dataset, the RF model performs better than the baseline. However, the performance improvement is in most cases small and the number of available observations make the predictions unreliable.

Location	**Service Level**	**Avg Baseline Error**	**Avg Model Error (w/o scheduled)**	**Avg Model Error (w/ scheduled)**	**Observations (train/test)**
Chicago	Mid-Tier	9.59min	8.83min	8.11min	986 / 323
Chicago	White Glove	11.36min	10.74min	10.07min	842 / 281
Minneapolis	Mid-Tier	9.29min	9.04min	8.73min	142 / 48
Minneapolis	White Glove	10min	11.41min	10.57min	178 / 60
Phoenix	Mid-Tier	9.75min	7.37min	6.63min	1647 / 549
Phoenix	White Glove	10.8min	10.97min	9.56min	1576 / 526
Birmingham	Other	11.37min	10.63min	10.61min	1596 / 533
Portsmouth	Other	11.74min	9.82min	9.75min	2049 / 683

Table 4-5 Average baseline and model error for RF on dataset separated on location then service level.

The second finding from the RF analysis is that the scheduled duration used by Pilot seem to be capturing some information not captured by our dataset. When running RF with scheduled stop duration, the model outperforms the same model without stop duration information. Comparing the average model error with scheduled stop duration to the average model error without scheduled stop duration in Table 4-5 shows the difference between the two. In most cases there is only a slight improvement, yet the model with stop time duration consequently outperforms the model without. In our discussions with Pilot, they revealed that their planners adjust stop time based on their own experience and available information. In addition to experience, the planners have access to information about the products delivered. The product information is not available to our model. Such information is the type of information that might improve the accuracy of our model.

The third finding from our RF analysis is what parameters appear to be affecting stop time duration. Figure 4.17 shows the average importance for the parameters used in the RF model. Importance is found by using the feature importance feature built in to scikit-learn (Pedregosa et al. 2011). Volume delivered appear to be the most important factor, followed by population density (reflecting urban/rural environments). Quantity pieces delivered, time of day and stop number seem to be of medium importance, while peak day appear to be of less importance. This insight is helpful when running the LR model and can be helpful for Pilot's planners when trying to determine what information they should collect and consider when planning.

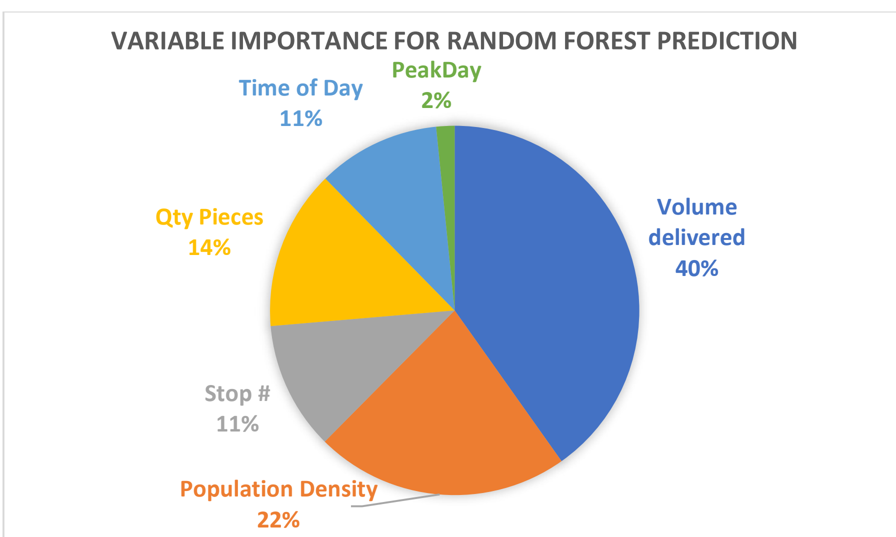

Figure 4.17 Variable importance for Random Forest Prediction

Running DT and RF we find that it is hard to predict stop time duration. The model is able to slightly outperform scheduled duration, but not by a lot. Second, we find that scheduled stop duration captures some information we do not have access to. As a result, models including scheduled duration outperforms models without. Lastly, the most important variables for predicting stop time duration are (in declining order): volume delivered, population density, quantity pieces delivered, stop number, time of day, and peak day. These findings can be used for building a LR model.

4.4.2.3 Linear Regression

Based on the findings from DT and RF, next we run LR to see if that method may perform better. As with DT and RF analysis, the LR model was trained on the whole dataset (except the most recent month), then location and service level separately, before combining the two. For the initial models, a period from January 1. 2019 to January 31. 2020 was used. The setup of the dataset is the same as for DT and RF. The variables chosen are the ones identified through the RF analysis to be of importance: volume delivered, population density, quantity pieces delivered, stop number, time of day, and peak day. Coefficients are calculated for each of these variables. We try to run the analysis with and without including intercepts, finding that excluding intercepts works best.

Again, the analysis run on dataset separated by location then service level performs best. LR run on the whole dataset achieved an adjusted R^2 of 0.718. Table 4-6 shows the result from the LR model built on datasets separated on location then service level. Again, number of observations is small. However, looking at the adjusted R^2 range between 0.653 and 0.8, indicating that the model is able to capture at least some of the variance in the dataset.

DC	Service Level	Coef Qty	Coef Stop	Coef Vol	Coef Dens	Coef Peak	Adj R^2	Num Obs
Phoenix	Midtier	0.5529	0.8966	0.1197	0.0010	0.6941	0.698	1066
Phoenix	Whiteglove	1.6991	1.2005	0.0954	0.0011	-1.4247	0.691	1056
Portsmouth	Other	2.5522	0.5305	0.0411	0.0008	3.9851	0.745	999
Birmingham	Other	3.0632	1.0423	0.0918	0.0014	2.5080	0.800	1059
Chicago	Midtier	0.3241	0.8998	0.1729	0.0003	1.2890	0.687	995
Chicago	Whiteglove	2.0971	1.2252	0.0865	0.0004	-1.1502	0.705	847
Minneapolis	Midtier	1.5152	1.3441	0.0778	0.0005	15.7733	0.684	106

Minneapolis	Whiteglove	1.0893	1.2591	0.1555	0.0004	26.7873	0.653	145

Table 4-6 Coefficients calculated with Linear Regression model for predicting stop time duration

To test the LR model, the coefficients are used to predict stop time for February 2020 across all locations, a period that was not included in training the dataset. The results are presented in Table 4-7, using the performance metrics described in Section 4.4.1 and the formulas shown in Equation 1. Using MPE and MAPE, LR appear to be performing very well. Estimated stop times are consistently more accurate than scheduled, reducing the error for several locations by half or more. However, looking at the average estimated and scheduled error, we can see that estimated stop times tend to be too short. For last mile delivery operations being late is much worse than being early. As a result, though estimated stop times are more accurate, they might not be worth using as the company may risk late deliveries.

Location	Sch Avg Error	Est Avg Error	Difference	Sch MPE	Est MPE	Sch MAPE	Est MAPE
Birmingham	-2.5min	-2.7min	0.2min	17%	14.8%	49.1%	48.1%
Chicago	4min	-3.3min	7.4min	60.5%	17%	80.3%	60.3%
Minneapolis	3.6min	-6.5min	10.1min	68.9%	0.9%	92%	60.6%
Phoenix	6.9min	-3.2min	10min	83.5%	14.8%	97.9%	59%
Portsmouth	7.1min	-0.5min	7.6min	72.9%	28.3%	85%	57.1%

Table 4-7 Test Performance of LR stop time prediction

The significantly lower error margins of estimated stop times the LR model may be useful to understand how different variables affect the stop time. One of the best features of LR, compared to many other ML methods such as RF, is that the results are easier to interpret and use for most people. The coefficients presented in Table 4-6 explain how stop time increases and decreases for each unit of the variable. For white glove deliveries in Chicago, adding one more package (Qty delivered +1) means that stop time duration should increase by about 2 minutes. This insight can be useful for planners when setting up routes as a support to validate stop times.

4.5 Summary of Analysis

Using our understanding of Pilot's current operations and insight from their managers, a performance framework was defined with two performance dimensions: (1) Service level and (2) Efficiency. Within service level 2 KPI's were defined: (1) order completion rate and (2) on time delivery. For efficiency 5 KPI's were defined: (1) driver utilization, (2) spare available drive time, (3) excess driver capacity, (4) daily excess stop capacity and (5) number of stops per route. For each KPI we also determined appropriate aggregation levels.

Using the defined performance framework, analysis of Pilot's current LMD performance was done using descriptive statistics. The preliminary findings reveal that Pilot appear to be lacking insight into the performance of their current operations. They also show a difference between scheduled and actual route durations, indicating that planning factors should be improved. The Pilot Performance Assessment Tool (P-PAT) use reports from Dispatch Track to show how Pilot's different locations perform in terms of their performance dimension. The analysis shows that for the time period November 2019 to January 2020, on time delivery rate is between 80-90%, order completion about 90% while driver utilization rate vary between 50 and 90%. Though we cannot explain why, this insight helps Pilot understand where they should target their effort and assess the performance of future initiatives.

For improving planning accuracy, the analysis was split between drive and stop time prediction. Using LR we were able to build a model that estimate more accurate drive times than Pilot's current scheduled durations. Improved drive time planning accuracy will help improve on time delivery through more precise routing, and improved efficiency through less uncertainty in how many stops can be served by a route. The findings reinforce the preliminary findings, suggesting that Pilot has significant additional capacity. This capacity can either be used to expand their business, or perhaps be reduced by reducing the number of drivers employed.

Stop times appear to be more difficult to predict than drive time. Using DT and RF we are not able to make any predictions for stop time duration that significantly outperforms the current scheduled durations set by Pilot's planners. The poor model performance is likely due partly to lack of sufficient data to run the analysis. Another reason may be that the dataset available does not capture all relevant information about a stop, package, or customer. For example, in our dataset we lack information about type of package delivered, information that is available to Pilot's planners.

The stop time analysis is, however, able to tell what variables are affecting stop time. The most important variables for predicting stop time duration are (in declining order): volume delivered, population density, quantity pieces delivered, stop number, time of day, and peak day. Using these parameters in a LR prediction model, we are able to build a model that slightly outperforms scheduled stop times. The accuracy of the LR model makes it more suited as a helping tool for planners, than a certain predictor, but can be helpful to understand the drivers of longer stop time durations.

The next chapter explains how the findings made in this capstone will help Pilot improve their last mile delivery operations.

5 DISCUSSION AND CONCLUSION

The objective of this capstone project is to help Pilot Freight Services improve the performance of their LMD operations. To achieve this, we established an understanding of Pilot's current LMD operation and a framework for assessing the performance of this operation. Based on the performance framework, descriptive statistics was used to analyze the performance of Pilot's current operations. Insight from the descriptive statistics analysis was then used in further analysis using ML. In this chapter we are now ready to present the main findings, discuss its relevance and suggest further research.

5.1 Summary of Main Findings

Through this capstone project we have established a performance framework for assessing the performance of Pilot's LMD operations. The two performance dimensions defined are (1) service level and (2) efficiency. In order to assess the company's performance within each dimension appropriate KPI's were defined: For service level: (1) Order completion Rate and (2) on time delivery. For efficiency: (1) Driver utilization, (2) spare available drive time, (3) excess driver capacity, (4) daily excess stop capacity, and (5) Number of stops per route. To allow for easy and continuous assessment of these performance dimensions, a performance assessment tool (P-PAT) was developed. P-PAT allow Pilot's managers to assess performance separately for each of their five locations, within the timeframe they desire. P-PAT has been built to allow for additional sites and reports to be added, making the tool reusable for Pilot.

Analyzing the performance of Pilot's operations from November 2019 to January 2020, we find large differences in the performance of the different locations. The analysis shows that on time delivery rate is between 80-90%, order completion about 90% while driver utilization rate vary between 50 and 90%.

Based on the insights provided from descriptive statistics analysis, ML was used to provide further insight. Predictive models were built to help planners calculate appropriate stop times and TSF when planning routes in Dispatch Track. In addition to establishing the prediction models, we identified parameters that affect both stop time and drive time. For both factors, there appear to be large differences between locations. Drive time appears to be heavily affected by mileage (higher mileage = higher average drive speed), day of week and month. Stop time is most heavily affected by volume delivered, followed by population density, quantity of packages delivered, stop number on route, time of day and peak day.

5.2 Relevance

For Pilot Freight Services, the performance assessment framework and P-PAT are the most useful results. This allows Pilot to assess the performance of their operations and explore how implemented policies affect the performance of their locations. The tool provides easy access to information that previously has been difficult and time consuming to find. We believe this tool will help Pilot not only to make more well-informed decisions, but also to help introduce a more data-driven decision-making process.

Further, the prediction models that have been developed will help Pilot's planners improve the accuracy of their routing in Dispatch Track. Improve planning accuracy will help improve both performance dimensions. The insight into relevant variables can be used to understand how operations are affected and adapt to this. For example, volume information for each delivery should be gathered to be able to use this variable for better planning.

5.3 Limitations

The major limitation of this capstone is the lack of relevant data. After data processing, we find that there is little information left to do the analysis. Therefore, the results found in this project should be seen as a support for planners and not as a replacement. The findings can be used to help indicate relevant variables and estimate approximate values that planners can use as a sanity check against what Dispatch Track gives them.

Another challenge is the fact that all the analysis has been subject to what information is available in Dispatch Track. Aggregation and detail levels have been dictated by what is available in the system. As a result, there might be details or information that we have not been able to capture. One type of information like this is SKU-categorization that might have been useful for predicting stop time. Had product categories been available in the system, we might have been able to identify patterns in how long different product groups take to deliver (e.g. beds or sofas slower than chairs).

A third limitation is the fact that the ML predictions barely outperformed existing scheduled estimates. This indicates that our analysis has not been able to capture all relevant variables. Further analysis using data that contains the relevant variables identified (such as delivery volume and number of packages delivered for stop time) for a longer period of time (preferably more than a year) should be done to validate our findings.

5.4 Further Research

Having established the performance framework and methods to assess the performance, it would be interesting to see research assessing how this affects the performance of Pilot's LMD operations in the future. Such research would help validate the findings and ensure that the project identified the correct service dimensions and KPIs.

As part of the further analysis, the performance assessment tool and analysis could also be expanded to cover other relevant performance metrics. An example of this could be predicting variables that are related to delayed or incomplete orders.

Considering the fact that the findings in this project was likely affected by lack of relevant data, further research could include gathering more data that captures the variables we have identified as relevant (for stop time: volume delivered, quantity pieces, day of week, population density) and running the same analysis with more relevant data.

As efficiency in LMD operations is important for many companies, research into how applicable the insight from analyzing Pilot's LMD operation is to other companies could be interesting. We find it likely that the performance metrics service level and efficiency would be relevant for all companies performing LMD operations.

REFERENCES

Abidi, Hamida, Khaled Hassine, and Fethi Mguis. 2018. "Genetic Algorithm for Solving a Dynamic Vehicle Routing Problem with Time Windows." In *2018 International Conference on High Performance Computing Simulation (HPCS)*, 782–88. https://doi.org/10.1109/HPCS.2018.00126.

Baldacci, Roberto, Maria Battarra, and Daniele Vigo. 2008. "Routing a Heterogeneous Fleet of Vehicles." In *The Vehicle Routing Problem: Latest Advances and New Challenges*, edited by Bruce Golden, S. Raghavan, and Edward Wasil, 3–27. Operations Research/Computer Science Interfaces. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-77778-8_1.

Boyer, Kenneth K., Andrea M. Prud'homme, and Wenming Chung. 2009. "The Last Mile Challenge: Evaluating the Effects of Customer Density and Delivery Window Patterns." *Journal of Business Logistics* 30 (1): 185–201. https://doi.org/10.1002/j.2158-1592.2009.tb00104.x.

Brownlee, Jason. 2016. "Supervised and Unsupervised Machine Learning Algorithms." *Machine Learning Mastery* (blog). March 15, 2016. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

Chekole, Assefa. 2019. "Application of Machine Learning Tools for Predicting Determinant Factors." *International Journal of Advanced Research in Computer Science* 10 (4): 40.

Chunhua, Tang. 2010. "An Improving Genetic Algorithm for Vehicle Routing Problem with Time Windows." *2010 International Conference on Intelligent Computation Technology and Automation, Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference On*, 603. https://doi.org/10.1109/ICICTA.2010.646.

Deloitte University Press. 2017. “The Future of Freight - How New Technology and New Thinking Can Transform How Goods Are Moved.” 2017. https://www2.deloitte.com/us/en/insights/focus/future-of-mobility/future-of-freight-simplifying-last-mile-logistics.html.

DispatchTrack. 2019. “DispatchTrack.” DispatchTrack. 2019. http://www.dispatchtrack.com.

Ewedairo, Kolawole, Prem Chhetri, and Ferry Jie. 2018. “Estimating Transportation Network Impedance to Last-Mile Delivery.” *The International Journal of Logistics Management* 29 (1): 110–30. https://doi.org/10.1108/IJLM-10-2016-0247.

Filipec, M., D. Skrlec, and S. Krajcar. 1998. “An Efficient Implementation of Genetic Algorithms for Constrained Vehicle Routing Problem.” In *SMC’98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, 3:2231–36 vol.3. https://doi.org/10.1109/ICSMC.1998.724987.

Goodman, Russell W. 2005. “Whatever You Call It, Just Don’t Think of Last-Mile Logistics, Last.” *Global Logistics & Supply Chain Strategies, 2005*, 6.

Güemes-Peña, D., C. Lõpez-Nozal, R. Marticorena-Sánchez, and J. Maudes-Raedo. 2018. “Emerging Topics in Mining Software Repositories.” *Progress in Artificial Intelligence* 7 (3): 237–47. https://doi.org/10.1007/s13748-018-0147-7.

Jozefowiez, Nicolas, Fr´ed´eric Semet, and El-Ghazali Talbi. 2008. “From Single-Objective to Multi-Objective Vehicle Routing Problems: Motivations, Case Studies, and Methods.” In *The Vehicle Routing Problem: Latest Advances and New Challenges*, edited by Bruce Golden, S. Raghavan, and Edward Wasil, 445–71. Operations Research/Computer Science Interfaces. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-77778-8_20.

Knox, Steven W. 2018. *Machine Learning*. 1st ed. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119439868.

Koehrsen, Will. 2018. "Random Forest in Python." Medium. January 17, 2018. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

Kotsiantis, S. B. 2007. "Supervised Machine Learning: A Review of Classification Techniques." *Informatica (03505596)* 31 (3): 249–68.

Kotsiantis, S.b. 2011. "Cascade Generalization with Reweighting Data for Handling Imbalanced Problems." *Computer Journal* 54 (9): 1547–59. https://doi.org/10.1093/comjnl/bxr016.

Kumar, Suresh Nanda, and Ramasamy Panneerselvam. 2012. "A Survey on the Vehicle Routing Problem and Its Variants." *Intelligent Information Management* 04 (03): 66–74. https://doi.org/10.4236/iim.2012.43010.

"Linear Regression — Statsmodels." 2019. 2019. https://www.statsmodels.org/stable/regression.html.

Lopez, Edwin. 2017. "Why Is the Last Mile so Inefficient?" Supply Chain Dive. May 22, 2017. https://www.supplychaindive.com/news/last-mile-spotlight-inefficient-perfect-delivery/443089/.

Mohammed, Mohssen, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. 2017. "Machine Learning: Algorithms and Applications." 2017. https://mit-library.skillport.com/skillportfe/main.action?assetid=RW$1172:_ss_book:117434#summary/BOOKS/RW$1172:_ss_book:117434.

Olsson, John, Daniel Hellström, and Henrik Pålsson. 2019. "Framework of Last Mile Logistics Research: A Systematic Review of the Literature." *Sustainability* 11 (24): 7131. https://doi.org/10.3390/su11247131.

Pedregosa et al. 2011. "Scikit-Learn: Machine Learning in Python." 3.2.4.3.2. Sklearn.Ensemble.RandomForestRegressor. 2011. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

Pilot Freight Services. 2019. “Pilot Freight Services | Global Shipping Transportation and Logistics Solutions.” Pilot Freight Services. 2019. https://www.pilotdelivers.com/.

Savelsbergh, Martin W. P., Tom Van Woensel, and Hamilton M. Stewart. 2016. “2 Trends and Advances in Technology Impacting City Logistics.” In .

Seaborn.pydata.org. 2020. “Seaborn: Statistical Data Visualization — Seaborn 0.10.0 Documentation.” Seaborn. 2020. https://seaborn.pydata.org/index.html.

Snoeck, André, Daniel Merchán, and Matthias Winkenbach. 2020. “Route Learning: A Machine Learning-Based Approach to Infer Constrained Customers in Delivery Routes.” *Transportation Research Procedia* 46: 229–36. https://doi.org/10.1016/j.trpro.2020.03.185.

Tantawi, Randa. 2019. “Linear Regression.” Linear Regression: BartonPlus. 2019. https://eds.a.ebscohost.com/eds/detail/detail?vid=15&sid=d1e26d76-4721-41b1-98c5-9f6d91dde140%40sessionmgr4006&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=90558376&db=ers.

Ward, Jeffrey, Yan Sun, Alberto Oca, Michael Zimmerman, Korhan Acar, and Balika Sonthalia. 2019. “AT Kearney State of Logistics Report 2019: Cresting the Hill.” 2019 State of Logistics Report. 2019. http://www.atkearney.com/transportation-travel/state-of-logistics-report.

Wuest, Thorsten. 2015. *Identifying Product and Process State Drivers in Manufacturing Systems Using Supervised Machine Learning*. Springer Theses. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17611-6.

Yulei, Yang, and Zhang Jin. 2019. “Vehicle Routing Problem with Soft Time Windows Based on Dynamic Demands.” In *2019 4th International Conference on Intelligent Transportation Engineering (ICITE)*, 222–26. https://doi.org/10.1109/ICITE.2019.8880257.

Appendix A: Table of planning operation parameters in Dispatch Track

Parameter	Description
Driver	Driver ID and Name
Time Window Duration (min)	Length of time window by default
Hours of Service (min)	The maximum # of minutes a driver can work (maximum length of route). Default setting: 12 hours.
Start Time	Start time for route
Load Time	Loading time at Pilot's Warehouse dock, where the drivers pick up the goods they are to deliver. Default: 90 minutes
Unload Time	Expected time to unload one order. Default 15 min.
Default Truck Capacity (cubic ft)	Maximum volume capacity of a truck
Default Truck Capacity Weight	Maximum weight capacity of a truck
Travel Speed Factor	Factor to allow adjustment in traveling time. Default is 1. A number higher than 1 indicates that the unit can travel faster than normal estimation, below 0 indicates slower.
Start Location	Start of route normally Pilot's warehouse or center of operation
End Location	End point for route. Normally the same as start location.

Table 0-1 Planning Operations Parameters in Dispatch Track

Appendix B: Capstone Analysis File Structure

To support the use of Python, a file structure was built for the necessary reports from Dispatch Track (visualized in Figure 0.1 Visualization of file structure built for capstone.). In the file structure, each relevant report used in the analysis is given its own folder. In each folder, there is a sub-folder for each of Pilot's locations. The reports are downloaded from Dispatch Track and stored in the respective folder. The Python scripts are written to read the reports in their respective folders, naming them appropriately, before merging them into aggregated reports. In addition to organizing the capstone, this allows easier sharing for the tools built. In most cases, Pilot can download the file structure, add new reports if wanted and re-running the Python scripts. As a result, Pilot can easily reuse and share the models built in this project.

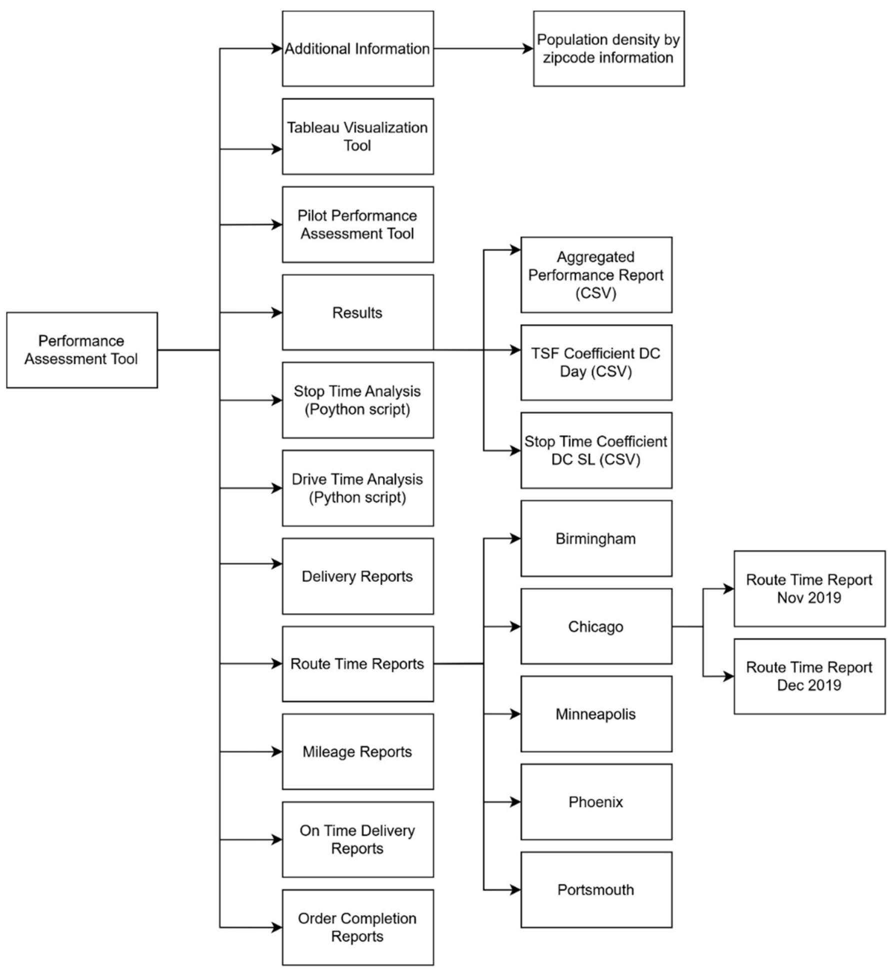

Figure 0.1 Visualization of file structure built for capstone.

Appendix C: Data Processing

Performance Assessment Data Processing

Once placed in its dedicated folder, the Python script can read the files, and assign a DC value to each row of data based on the location folder the report is placed in. Duplicates are removed from each report, keeping only the first entry, and the index reset. Column names are changed to improve readability and name equal information across the different reports the same. The format of the date column is cleaned and set to datetime to allow for merging. In the on time delivery report a "Total Deliveries" column is created by summing up the early, on time, delayed, scheduled and cancel columns. For the completed orders report a "Total Orders" column is created, summing up the in progress, completed, exceptions and not completed columns. The delivery report is cleaned to only contain one entry for each route, with the number of routes in each row. Then the route time report was merged with the delivery report and a "Max_Cap_Min" column was calculated multiplying max driver service time (720 minutes) with the number of routes. Further a "spare capacity" column is calculated by subtracting the actual duration of the route from the maximum available driver service time. This column is used to show driver utilization.

After data cleaning, all the reports are merged into one Aggregated Report containing all the information. Next the total duration of each route is divided by the number of stops on each route. The available additional service time is then divided by the average stop time, to give an indication of how many more stops could be served in that day. This is a very rough estimation and is only meant to give an indication of how many more customers could be served in any given day. Further excess driver capacity, giving an indication of how many drivers could be removed from a location in a given day, is calculated by dividing the spare capacity by 720 (a driver's maximum available service time). Finally, the script creates a csv file named "Aggregated Report.csv".

Next the Aggregated Report is read into Tableau, where three dashboards are created. Each dashboard represents one of the three performance metrics that Pilot want to assess: Order completion rate, on time delivery, and efficiency (asset utilization). The Tableau dashboards are built with filters to allow for assessing performance on a monthly basis, as well as looking at different locations. The tableau dashboards will automatically update when new information is added.

Drive Time Data Processing

Initially one aggregated report for all locations for each of the three reports is created. This is done by reading all reports located in the respective folders in the file structure and assigning a “DC“-column value based on the location folder the file is located in.

The aggregated Route time report has an initial shape of 2306 rows with 6 columns (Date, DC, Scheduled Time, Actual Time, Service Unit and Driver), containing information about the total scheduled and actual duration of all routes for a location in a day. The Service Unit and Driver columns are removed, as they are not relevant for the analysis. “Date” column is converted to datetime to allow for future merging and analysis. Null values are removed, as this report should not have any null values (in our dataset no rows are removed from this operation). Next duplicate rows are removed, keeping the first row (reducing dataset from 2306 to 2180 rows). Scheduled and Actual Time is converted to minutes and integer types to allow for further analysis.

The aggregate Delivery Report contains information about each stop. For TSF analysis only the Date, DC, Route#, Stop#, Scheduled Time and Actual Time columns are relevant, giving an initial shape of 91,817 rows and 6 columns. Scheduled and actual time, containing both start and stop time, are split into separate columns (TimeStart and TimeStop) and the old columns are deleted. Rows are then grouped by date, DC, route# and stop# to limit the number of entries to one line per stop. Several stops have more than one row to account for several packages in the same order. For each stop the first actual and scheduled TimeStart is kept, and the last stop entry for TimeStop. This reduces the dataset to 66,428 rows. Scheduled and actual times are converted to datetime, and scheduled and actual duration is calculated for each stop by subtracting TimeStart from TimeStop (times of day). Any duplicates are dropped, and null values removed leaving 65,217 rows. Lastly, in this part of the analysis we are using stop duration to calculate drive time by subtracting total stop time for a route from the total route time. Observing the stop duration, we find that there are several extreme outliers (less than 5 or greater than 60 minutes). To be able to run the analysis we adjust these values to be within 5 to 60 minutes (40 minutes for scheduled stops).

After cleaning the delivery reports, the report is grouped by Date and DC to aggregate for each day and location. Scheduled and actual stop duration is summed up, and the total number of routes per day is calculated. The final dataset contains1635 rows and 5 columns with information about total scheduled and actual stop time for each location and day.

The aggregated mileage report has an initial shape of 2293 rows and 5 columns (Date, DC, Mileage (miles), Service Unit and Driver), containing information about the total number of miles driven for a location and day. The Service Unit and Driver columns are removed, as they are not relevant for the analysis. "Date" column is converted to datetime to allow for future merging and analysis. Rows containing null values or duplicates are removed, keeping the first row for duplicates, reducing dataset from 2293 to 2177 rows.

Once all the separate reports are created and cleaned, they are all merged into one Aggregated Drive Report based on "Date" and "DC". Scheduled and actual drive time is calculated by subtracting stop duration from the total service time for each DC and day. Rows with scheduled or actual drive duration of less than 0, as well as routes with mileage less than 50, are removed. The removal of these values reduces the dataset from 1638 rows to 1576 rows. Scheduled and actual MPH is calculated by dividing "Mileage (miles)" by scheduled or actual drive time. In addition, a "TSF" column is calculated by dividing scheduled by actual drive time. The "TSF" column show what the TSF should have been for the scheduled drive time to be equal the actual time.

Stop Time Analysis Data Processing

The Python script reads all the Delivery Reports stored in the file structure, and assign "DC"-columns value based on what subfolder the file is stored in. The initial aggregated dataframe has the size of 91,817 rows and 54 columns. All columns with more than 80% null values are removed, as these columns will not contain sufficient information to run any analysis, reducing the number of columns to 32. All stop entries where either actual or scheduled stop duration is missing is not usable in the analysis. Therefore, all rows where actual or scheduled stop duration contains null values are removed, reducing the number of rows to 83,979. Next, irrelevant columns such as "unit number" and "Phone1" are removed, reducing the number of columns to 21. All duplicate rows are removed, keeping the first entry, reducing the number of rows to 81,514. Scheduled and actual time, containing both start and stop time, are split into separate columns ("TimeStart" and "TimeStop") and the old columns are deleted. Rows are then grouped by date, DC, route# and stop# to limit the number of entries to one line per stop. Several stops have more than one row to account for several packages in the same order. For each stop the first actual and scheduled TimeStart is kept, and the last stop entry for TimeStop. This reduces the dataset to 66,430 rows. Scheduled and actual duration is calculated by subtracting TimeStart from TimeStop and set to integer type. All rows where either scheduled or actual duration contains null values are removed, reducing the number of rows to 65,966. Finally, all rows where scheduled or actual duration has a value equal to or less than 0 is removed, giving a dataframe of 57,859 rows.

Once the dataframe has been cleaned, time of day (ToD) in whole hours and day of week (DoW) is calculated based on the "Date" column. Census data about population density per zip code from the latest US census is collected from www.census.gov. This data is added to each stop based on the zip code of each stop.