

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/344871967>

Predicting Package Delivery Time For Motorcycles In Nairobi

Thesis · August 2020

DOI: 10.13140/RG.2.2.27105.94567

CITATIONS

0

READS

1,328

1 author:



Joseph Magiya

KCA University

1 PUBLICATION 0 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Predicting Package Time Delivery For Motorcycles in Nairobi [View project](#)

**PREDICTING PACKAGE DELIVERY TIME FOR
MOTORCYCLES IN NAIROBI.**

BY


JOSEPH A. MAGIYA

**A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE AWARD OF MSC.
DATA ANALYTICS IN THE FACULTY OF COMPUTING
AND INFORMATION MANAGEMENT AT KCA
UNIVERSITY**

AUGUST, 2020

DECLARATION

I declare that this thesis is my original work and has not been presented for degree in any other university. No part of this project work may be reproduced without the prior written permission of the author and/or KCA University.

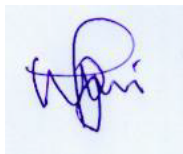
Sign:  Date: 22nd July 2020

JOSEPH ANYONA MAGIYA

Reg: 17/04987

This thesis has been submitted for examination with our approval as university

Supervisors



Sign: Date: ...28th July 2020.....

PROF. FELIX MUSAU,

KCA UNIVERSITY

DEDICATION

This project thesis is dedicated to my family and friends for their support during my studies.

TABLE OF CONTENTS

LIST OF FIGURES	viii
LIST OF ACRONYMS AND ABBREVIATIONS	xi
ABSTRACT.....	xii
CHAPTER ONE: Introduction	1
1.1 Background of the Study.....	1
1.2 Problem Statement	3
1.3 Research Objectives	5
1.3.1 Main Objective.....	5
1.3.2 Specific Objectives	5
1.4 Research Questions	5
1.5 Motivation of the Study.....	6
1.6 Significance of the Study	6
1.7 Scope of the Study.....	7
1.8 Organization of the Study	7
CHAPTER 2: LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Estimated Time of Arrival and E-Commerce Businesses.....	9
2.3 Predicting Estimated Time of Arrival – Previous Studies	11
2.3.1 Research on other modes of Transporting	11
2.3.2 Time Series Models	12
2.3.3 Data Driven Methods.....	12

2.4 Current ETD Prediction	14
2.5 Research Gaps	15
2.6 Conceptual Framework	16
CHAPTER 3 METHODOLOGY	17
3.1 Introduction	17
3.2 Research Design	17
3.3 Data Source	19
3.4 Data Processing and Modeling Procedure	20
3.5 Data Modeling and Analysis	21
3.6 Study Variables	22
3.6.1 Delivery Time	22
3.6.2 Placement Timestamp	22
3.6.3 Confirmation Timestamp	22
3.6.4 Weather	22
3.6.5 Pickup Location	23
3.6.6 Drop Off Location	23
3.6.7 Road Distance	23
3.6.8 Customer Type	23
3.6.9 Rider Details	23
3.6.10 Platform Type	24
CHAPTER 4 RESULTS	25

4.1	Sample Description and Delivery Times	25
4.1.1	Observations by Placement Day of Month	25
4.1.2	Observations by Placement Time (1 Hour Bins)	26
4.1.3	Observations by Placement Weekday	26
4.1.4	Observations by Customer Type	27
4.2.6	Observations by Platform Type	27
4.2.7	Observations by Distance (Km)	28
4.2.8	Observations by Pickup Location	28
4.2.9	Observations by Destination Location	29
4.3	Predicting the Estimated Time of Delivery	30
4.3.1	Data Profiling	30
4.3.2	Building A Model	55
4.3.3	Testing and Validating the Model	56
4.3.4	Significant Variables	59
CHAPTER 5 DISCUSSION OF RESULTS		63
CHAPTER 6 SUMMARY FINDINGS, CONCLUSION AND RECOMMENDATIONS		64
6.1	Introduction	64
6.2	Current ETD Prediction at Sendy Ltd.	64
6.4	Summary of Findings	64
6.5	Conclusion	65
6.6	Contributions	65

6.7 Recommendations	66
REFERENCES	67
APPENDICES	70
Appendix I: Budget	70
Appendix II: Research Schedule	71

LIST OF FIGURES

Figure 1: Conceptual Framework	16
Figure 2: Research Design.....	17
Figure 3: Data Processing and Modeling Procedure.....	20
Figure 4: Observations by Placement Day of Month.....	25
Figure 5 - Observations by Placement Time (1 Hour Bins)	26
Figure 6 - Observations by Placement Weekday	26
Figure 7 - Observations by Customer Type.....	27
Figure 8 - Observations by Platform Type	27
Figure 9 - Observations by Distance (Km)	28
Figure 10 - Observations by Pickup Location	28
Figure 11 - Observations by Pickup Location (Heat map)	29
Figure 12 - Observations by Destination Location	29
Figure 13 - Observations by Destination Location (Heat Map)	30
Figure 14 - Arrival at destination day of month statistics.....	31
Figure 15 - Arrival at destination day of month histogram	32
Figure 16 - Arrival at destination time statistics.....	32
Figure 17 - Arrival at destination time histogram.....	33
Figure 18 - Arrival at Destination Weekday Statistics	34
Figure 19 - Arrival at Destination Weekday Histogram.....	34
Figure 20-Arrival at pickup time statistics and histogram.....	35
Figure 21 - Confirmation time statistics and histogram.....	36
Figure 22 – Destination latitude statistics.....	36
Figure 23 – Destination latitude histogram.....	37
Figure 24 - Destination Longitude Statistics	37

Figure 25 - Destination Longitude Statistics	38
Figure 26 - Distance Statistics	39
Figure 27 - Distance (KM) histogram.....	39
Figure 28 Distribution of Personal or Business	40
Figure 29 - Statistics and distribution of pickup time.....	40
Figure 30 - Pickup latitude statistics	41
Figure 31 - Pickup latitude histogram.....	41
Figure 32 - Pickup longitude statistics.....	42
Figure 33 - Pickup longitude histogram.....	42
Figure 34 - Placement time statistics and histogram	43
Figure 35 - Platform Type statistics and distribution.....	43
Figure 36 - Precipitation in millimeters statistics	44
Figure 37 - Precipitation in millimeters histogram	44
Figure 38 - Temperature statistics.....	45
Figure 39 - Temperature histogram	45
Figure 40 - Time from pickup to arrival at destination statistics.....	46
Figure 41 - Time from pickup to arrival at destination histogram.....	46
Figure 42 - Rider's age statistics	47
Figure 43 - Rider's age histogram	47
Figure 44 - Rider's average rating statistics	48
Figure 45 - Rider's average rating histogram.....	48
Figure 46 - Number of orders a rider has completed statistics	49
Figure 47 - Histogram of number of orders a rider has completed.....	49
Figure 48 - Rider's number of ratings statistics	50
Figure 49 - Rider's number of ratings histograms.....	50

Figure 50 - Pearson r coefficient of correlation in orders data	53
Figure 51 - Pearson r coefficient of correlation in rider's data	54
Figure 52 - K-fold cross-validation.....	56
Figure 53 - XGBoost prediction histogram	57
Figure 54 - XGBoost difference statistics.....	58
Figure 55 - XGBoost absolute difference statistics	58
Figure 56 - XGBoost absolute difference distribution.....	59
Figure 57 - XGBoost feature importance bar graph	60
Figure 58 - XGBoost feature importance values	61
Figure 59 - First XGBoost tree	62
Figure 60 - Last XGBoost tree (700)	62
Figure 61: Budget	70
Figure 62: Research Schedule.....	71

LIST OF ACRONYMS AND ABBREVIATIONS

ETA	Estimated Time of Arrival
ETD	Estimated Time of Delivery
DSRP	Design Science Research Process
RMSE	Root Mean Square Error
BPT	Beijing Public Transport Holdings Limited
API	Application Programming Interface
GBM	Gradient Boosting Machines
ANN	Artificial Neural Networks
BAT	Bus arrival time
DSRP	Design Science Research Process
ETD	Estimated time of delivery
GPS	Global Positioning System
IS	Information systems
LQE	Linear quadratic estimation
SVM	Support vector machines
MPI	Message Passing Interface
SVR	Support vector regression
SGE	Sun Grid Engine

ABSTRACT

A great proportion of the transport and logistics applications today offer an Expected Time of Arrival (ETA). This serves as an attractive add-on feature for the users. Many on-demand transport and logistics companies are showing the predicted ETD to enhance customer experience. However very few logistics companies show the Estimated Time of Delivery (ETD). Predicting accurate estimated time of delivery is imperative to logistics providers as the user experience is heavily determined by the availability and accuracy of such information. This study used the Design Science Research Process (DSRP). The main purpose of the study is to accurately predict the estimated time of delivery of a package. Primary data was provided by Sendy Ltd. The data analysis step constitutes of the generation of descriptive and inferential statistics of the data provided. A model was then be developed to predict an accurate ETD. I aim to use Microsoft PowerBI to visualize the analysis and Python to create and test models that predict an accurate ETD. At the end of the study I expect to have developed a model that can predict an accurate ETD of a package.

CHAPTER ONE: Introduction

1.1 Background of the Study

Logistics is a word that involves cargo travelling by water, air and land. 90% of the world's logistics traffic is via ocean, vessel arrival certainty is key for integrated decision making and overall supply chain performance (Abhishek, 2008). There are numerous benefits – waiting time is shortened and owner's cost is reduced through asset optimization. Another novel use for the application of ETD prediction for aircraft. Predicting aircraft arrival time using machine learning methods while the aircraft is moving is also an interesting but challenging task.

There is a consistent effort for businesses to be the preferred choice in the hyper-competitive on-demand landscape. Convenience and speed being the major drivers of the economy, logistics companies need an edge to stay relevant whenever and wherever the customer demands (Dalman, 2007). Services such as package delivery, online hailing, car sharing (also known as carpooling) and car rental services are the main drivers of the rising on-demand transportation sector. To predict arrival time of a motorcycle at the pickup location, on-demand logistics applications enable users to track real-time movement of the vehicle on maps (giving an idea about areas with traffic and the route followed to reach the defined location).

In addition to informing the customer about the arrival time of the vehicle, providing an estimated time of arrival will also reduce the chances of a customer cancelling an order. (Oberoi, 2018).

The time between a service demand (placement of an order) and its delivery (delivery time) has a very important part to play in the on-demand ecosystem. On demand transportation and logistics bank a lot upon providing the users insight on their order or service. Providing such information to the user involves monitoring of a vehicle such as a truck, car, or motorcycle in real time which will enable an accurate estimation of arrival time (Oberoi, 2018). Estimate Time of arrival is defined as the timestamp (consisting of the date and time) that a package is

predicted to reach a specified location. The estimated time of delivery (ETD) is the timestamp of which a package is expected to be reach the destination. (Dushaj, 2018). The predicted ETD provides a huge distinction in the brand-customer relationship. Brand-customer relationship is the new currency in today's market. Customers quickly associate brands with the quality of service they provide. It's not surprising to hear someone say that application X is slow, expensive or cheap, etc.

Focusing on logistics and delivery of packages by motorcycles, the same thought, effort and resources need to be put into predicting a time, which is accurate, in arrival in which the rider will pick up the order. Furthermore, more thought needs to go to the prediction of the delivery time since customers also care about this.

1.2 Problem Statement

A study (Ijaz et al., 2018) pointed out that with the increased growth of the internet has resulted in the growth of electronic (online) business as an important trend in the economy (Ijaz et al., 2018). More and more people who shop are looking into screens instead of shopping centers, retailers, if they're to have a positive future in the market, they need to offer better customer experiences (Grosman 2018). Part of the customer experience is getting the goods on time. On-time delivery brings about customer satisfaction and enhanced efficiency. In a nutshell, on-time delivery only does good deeds for ecommerce companies and the customers.

With the introduction of e-commerce, business establishments have consistently undergone fast development and changes in recent years all over the world. The maintained growth shows that there is demand for the service. Beyond the start-up phase, businesses have been recording losses which might show the problems in the structure of the organization where the most crucial is logistics (Daganzo 2004).

Taxi application such as Uber, Taxify have developed their own models for predicting ETD using the best data scientists they can afford. The difference between moving a person from point A to B is very different from moving a parcel from point A to point B. Here's why:

1. Taxi applications' ETA is purely focused on transport while what this study is aiming to achieve is predicting ETD in logistics.
2. For Taxi applications most of the time the client who requested is the one who is being moved from point A to B while in deliveries the good is what is being moved from point A to B and usually to a different party.
3. For taxi applications the ETA is mainly consumed by the client, but for deliveries, both parties are interested in the ETD; the sender and the recipient.

4. For taxi applications the vehicle used is mainly cars while for delivery applications it can be a motorcycle, pickup truck, 3T Truck and even a trailer. For this study we're exclusively focusing on motorcycles.
5. Cars for taxis and motorcycles for deliveries move differently and are influenced differently by outside factors:
 - a. Rainfall has a high likelihood of causing a major delay in motorcycles since the riders can't ride in the rain while this might cause a slight delay for cars.
 - b. Traffic jam will slow down a car while the motorcycle can make its way through the traffic jam.
 - c. Cars can move faster in highways than motorcycles.
6. For deliveries since the riders aren't monitored in -person, they might be tempted to do other 'side-jobs' while fulfilling a delivery but for taxis the client is in the car monitoring the driver.

Logistics is very important if not the backbone of majority of the businesses in Nairobi and all other cities in the world. From delivering goods to delivering cheques to the bank, a huge portion of the business in Nairobi and around the world use logistics providers to solve their problems. As Kenya is a developing country, things are moving faster than ever before and time seems even less for everyone in business. This is especially true for businesses in Nairobi where there's stiff competition to serve the customer on time. Thinking about logistics and businesses, ETD is extremely important for the businesses to monitor and award the drivers and the customers' peace of mind. With this in mind, I have developed a model which can help logistics companies provide their clients an accurate predicted time for the arrival of their package.

Uber currently leverages its large team of over 6,000 employees to build complex machine learning models to solve this problem. Google is also a big player in this especially because

they have a Google Maps API which users can pay for using. The cost of the API is also not cheap, the cost for maps is \$7 for every 1,000 requests, additionally, the routes also can easily add up to \$70 per 1,000 requests.

1.3 Research Objectives

1.3.1 Main Objective

Building a predictive analytics model, given a historical dataset, to accurately predict the delivery time of motor cycle packages in Nairobi.

1.3.2 Specific Objectives

The research is directed by the below:

- i. To determine the significant variables for determining an estimate for the delivery time of a motorcycle order in Nairobi.
- ii. To build a machine learning model to predict the estimated time of delivery of a motorcycle order in Nairobi.
- iii. To test and validate the machine learning model built in ii above.

1.4 Research Questions

This study seeks to answer the following questions:

1. Which variables can significantly determine the estimated delivery time for a package delivery by a motorcycle in Nairobi?
2. Which machine learning methods can be used for building a model to predict the estimated delivery time of a motorcycle order in Nairobi.?
3. Which method of validation can be utilized to find out the model's accuracy for predicting the estimated delivery time of a motorcycle package in Nairobi.?

1.5 Motivation of the Study

There has been tremendous development of on-demand applications that offer services within the country and the whole world. With all these developments in the business side of the world, there has also been a change in customer expectations. Taxi applications have significantly improved the efficiency of the dispatching systems by being able to accurately predict how long a ride will take to arrive at the pickup location and even how long the trip will take. Customers who are used to getting this information find it unpleasant when they can't get this information on a delivery or logistics application. This leads to an increased cancellation rate since users will not place an order and if the ETD is too long, they'll opt to cancel the order and choose other means. Also, if the ETD reaches and the rider is not there, the customer will most likely cancel the order. The ETD is very important both for the customer and for the business. If the customer can know the ETD, he/she can plan better and will feel satisfied if the ETD is achieved. Also, the business can plan their dispatching better, if the business can know when a rider will finish his current order, they can allocate the order to the best driver better.

1.6 Significance of the Study

The study is expected to be of great value to the logistics applications in Nairobi, as they will get tremendous insight on the effectiveness of predicting and providing an accurate estimated time of delivery (ETD) to their customers. Logistics companies in other countries as well will benefit a lot from this study. By predicting and giving the consumers an accurate ETD, the logistics companies are able to reduce the number of cancelled orders, increase retention rates, improve customer satisfaction, improve dispatch and many more.

The study is of value to the customers who use on-demand logistics applications in Nairobi. In this study, I also aim to go into detail about what factors influence the Estimated Time of Arrival of a package. The study might also benefit customers from other cities as well, especially in the developing countries, as they can borrow some concepts from this study.

The study is also expected to be of value to scholars and researchers. The study has also added value to the existing body of knowledge and act as a useful resource for those who would be undertaking research on providing accurate estimated time of delivery. The study is also a basis for further research on the topic.

1.7 Scope of the Study

In the study, the researcher examines the variables that affect the Estimated Time of delivery on on-demand logistics applications for motorcycle orders in Nairobi county, with a focus on Sendy Ltd. The study consists of orders placed by customers in Nairobi. The choice of this study population is informed by the fact that most motorcycle orders placed are in Nairobi county. Nairobi is also growing rapidly in the number of businesses and the population as well. The findings of this study informs other growing areas across the world as Nairobi county is the perfect location to conduct the study.

1.8 Organization of the Study

The study is arranged in five chapters. The first chapter provides the background of the study, the problem statement, the research objectives, significance of the study and the scope of the study.

Chapter two covers the review of the literature which contains the empirical review to identify the knowledge gap. The chapter will dive into works done in previous studies, the research gaps and the conceptual framework

Chapter three covers the methodology adopted to reach and attain the goals of the study. The third chapter will describe the design of the research, the data source, the processing and modeling procedure, the data mining and modeling techniques used and the variables of the study.

Analysis, interpretations and presentation of the outcome of the study is laid out in the fourth chapter. The fifth chapter lays out the findings, conclusions, recommendations and contribution of this study.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

To develop the required methods and ways to accurately answer the research question laid out in section 1.4, a detailed literature review was carried out. Section 2.2 has provided a literature review on current methods used to forecast arrival times. Section 2.2 has provided a summary of the overview of the various techniques in machine learning and goes ahead to show the differences and similarities between them. A comparison in the various machine learning techniques will provide insight which will provide the opportunity to select the most promising ones for the current research. Discussions of past studies by other authors on the specific research objectives is covered in the chapter as well. Schematic illustrations that show the interactions between the variables is presented in the conceptual framework. The chapter concluded with gaps in other studies.

2.2 Estimated Time of Arrival and E-Commerce Businesses

The internet has increasingly become a key trading and business platform for selling and distributing products between businesses, consumers and organizations and even between consumers. This has transformed the way companies share information with business partners, communicate and also how they buy and sell. These changes have taken e-commerce to an entirely new level. Taking into account that a universally accepted definition for e-commerce doesn't exist, Kalakota and Whinston (Kalakota, Whinston 1997) defined e-commerce from different angles which were based on these perspectives:

- Communications - E-commerce is the provision of a product and/or service, payments or payments via computer networks and information.
- Online - E-commerce provides the capability of buying and selling products and information on the internet and other online services.
- Business process - E-commerce is the application of technology toward the automation of business transactions and workflows.

- Service - E-commerce addresses the desire of consumers, businesses and leadership in a business to reduce the cost of service and concurrently scaling up the speed of delivery and quality of merchandise.

All the above definitions are valid. What really matters is the perspective that you choose to view e-commerce from. For the purpose of this paper, I used the definition of e-commerce from a service perspective. From a service perspective, like it has been laid out above means that e-commerce is a tool that addresses the desire of consumers, businesses, firms and management to reduce the cost of service, while improving the quality of the goods and also increasing the speed at which they deliver their service. We'll narrow this down to the 'speed of delivery'.

Timely procurement and delivery of goods to the end consumer or to a business will definitely provide an improved experience to the customers. Delivery of goods and services will have a positive effect on the customers' retention because of good supply chain management (Farooq 2019). Onsite intelligence is needed by companies that want to achieve a high level of transparency, in addition to that, companies also need close collaboration with everyone involved in the product supply chain and stringent information requirements.

2.3 Predicting Estimated Time of Arrival – Previous Studies

Previous studies have been conducted on this topic. During the past decades, many studies have been conducted to predict travel times.

2.3.1 Research on other modes of Transporting

While this research paper is tailored toward motorcycles, looking at relevant research sheds some light into how predicting the estimated time of arrival was calculated.

In air crafts a recent study (Zhengyi Wang, 2020), firstly performed experiments on the dataset for selecting base learners by utilizing candidate machine learning models with nested cross validation. In the study (Zhengyi Wang, 2020), preprocessed the data by ensuring that all units are of the same unit, the study estimated the missing variables, the study excluded flights with long transit times (which accounted for less than 1% of the data), the study also excluded observations with missing pickup points. The study then did a comparative study on the preprocessed and un-preprocessed and proved that the proposed preprocessing steps were robust in dealing with outliers, missing points, and noises. It could greatly improve the accuracy of ETD prediction.

In a study focused on predicting the ETA for trucks (Van der Spoel et al., 2016), it was discovered that congestion, time of day/week/month/year and accidents are the most frequently mentioned factors in what influences travel time. In the research (Van der Spoel et al., 2016) which was focused on predicting the truck arrival time he proposed that these are the variables that should be used in the data set and the subsequent model.

In bus transportation a recent paper (Md Noor et al., 2020) it was concluded that one of the main issues for bus operators is that the ETA was not accurate and it deviated from actual ETA by too long, and this discouraged riders and so ridership is negatively affected. SVR, which is based on the SVM (Support Vector Machine) classifier model, was chosen for the research. SVR (Support Vector Regression) had less than 45 seconds of error with a low average RMSE

(Root Mean Square Error). The input features for the training of the model included distance, segments, weather, peak or non-peak hour were used to predict the travel duration for the segments. The study (Md Noor et al. (2020) further concluded that SVR may be a feasible model for ETA prediction however the model needs to be trained and tested vigorously with more data and features.

2.3.2 Time Series Models

Time series models construct the time series relationship of travel time or traffic state, and then current traffic data and/or past traffic data are used in the constructed models to predict travel times in the next time interval.

Kalman filters, which is a time series model was coined to predict travel times using Global Positioning System (GPS) information and probe vehicle data (Yang, 2005). Linear quadratic estimation (LQE) which is another term used to refer to Kalman filtering, is an algorithm that uses a series of observed measures over time. These measurements contain statistical noise and other inaccuracies, which in turn produces the estimates of unknown variables. The estimates of the unknown variables are likely to be more precise than those based on a single measurement which is achieved by estimating a joint probability distribution over the variables for each timeframe (Paul Zarchan et al. 2000)

2.3.3 Data Driven Methods

Artificial Neural Networks (ANN) is one of the data driven methods tackle the challenge of predicting the estimated time of arrival. ANN are systems in computing which are inspired by, but not are not similar to the neural networks that make up the animal brains. Like animals, artificial neural networks "learn" to perform tasks by considering examples. The artificial neural networks achieve this generally, without being programmed with any task-specific rules (Graupe, 2013).

Support Vector Regression (SVR) (Wu, 2005) are also data driven methods. SVM (support vector machines) and SVR (support vector regression) have become a more and more popular method for machine learning tasks which involve regression, novelty detection or classification. In particular, they exhibit good generalization performance on many real issues and the approach is properly motivated theoretically (Wu, 2005).

K-nearest-neighbor (k-NN) has also been used as a data driven model. In predicting Bus Arrival Time; Tao Liu, Jihui Ma, Wei Guan, Yue Song, Hu Niu (Liu et.al, 2012) used a modified k-nearest neighbor (k-NN) method. In the study Liu et.al, 2012, they integrated the cluster analysis and principal component analysis to predict the buss arrival time. Using historical GPS data, they then applied it to bus arrival time (BAT) prediction.

Once recent study, by Mohammed Elhenawy, Hao Chen and Hesham A. Rakha (Elhenawy, 2014) which was aimed at modeling and predicting the expected bounds they used a data clustering and genetic programming approach. They took this approach to estimate the lower, and upper bounds of travel times. The models obtained from the genetic programming approach were algebraic expressions which shed good light into spatiotemporal interactions. Because of the use of an algebraic equation, the computation time was very efficient which meant that this was suitable for real-time applications.

These techniques are implemented through direct and indirect procedures to predict travel times using different types of state variables. In data driven methods and model-based methods, travel time was directly used as the state variable to perform the prediction of travel times. Variables such as speed, traffic density, traffic flow and even occupancy in the vehicle were used in the indirect procedures the state variables to predict the traffic status which is used in calculating future travel time based on a transition function.

2.4 Current ETD Prediction

Looking specifically at motorcycle and how the estimated delivery time is calculated would shed a lot of light into this study. I looked at plenty of resources during my research, I focused on Uber and Google Maps API. It is highly likely that some of the logistics applications rely on Google Maps API to provide an accurate ETD to their customers.

In a talk by the Uber Engineering team to try and answer this, data scientist Sreeta Gorripathy, explained that ETA at Uber is calculated multiple times during the life of an order. The ETA calculation at Uber takes in account map data, routing, traffic and a machine learning model. The routing algorithm finds the best route and ETA of the route then the traffic data is added onto this. The machine learning model used at Uber is non-linear and non-parametric (where the interaction of the variables is not understood). The talk concludes with Sreeta Gorripathy saying that they are exploring gradient boosted decision trees, random forest, neural networks and kNN learning.

For applications that rely on Google Maps API, an article by Rohit Raj tries to answer how this is calculated. In the article a Google ex-engineer reveals that Google Maps ETAs are based on a number of variables which entirely depends on the data that has been collected in the given area. Some of the tricks that the Google ex-engineer revealed were that the legal maximum speed, the recommended speed and the speed according to the type of track is used in predicting the ETA. Additionally, the history of the average speed at those points at a specific time or even the entire day is added into the model (Rajdev, 2018).

2.5 Research Gaps

The studies done on other vehicle types shed a lot of light into how to approach the problem in this research paper. However, this research is dedicated to predicting the Estimated Arrival time for motorcycles.

Time series models that take into account traffic state at the current time or historic traffic makes plenty of sense for vehicles. Traffic does not heavily affect motorcycle riders as they can move in between cars. This will definitely slow them down to some extent, but this will not heavily change the estimated time of arrival of a package.

The Kalman filters are an interesting take on predicting travel times. This can be transformed into the project at hand; Predicting Package Delivery Time for Riders in Nairobi. For the Kalman filters to properly work, I will have to collect a lot of data on the rider's location. I see some challenges with this:

- A rider's phone can get disconnected and stop sending the GPS even when delivering a package which will leave gaps in the positions. I could filter these out, but this will exclude a lot of the trips.
- Riders can do other activities on the side instead of going to the pickup directly, picking up the package and going straight to the delivery point. This can heavily skew the method.

The data driven methods are a step in the right direction towards getting an accurate estimated time of arrival. The studies done on the data driven methods are also focused on buses, cars etc. This is the largest gap I see on these studies. Motorcycles are considered faster and for good reason; they can make their way through traffic.

A model needs to be developed solely focused on predicting package delivery time for riders in Nairobi.

2.6 Conceptual Framework

The conceptual framework below was chosen for this study. The conceptual framework shows the relationship between the two types of variables in this study; the dependent variables and the independent variables. The estimated time of delivery is the dependent variable in this study. A customer's previous order history, a rider's previous order history, weather information, dispatch pool and dispatch metrics are the independent variables.

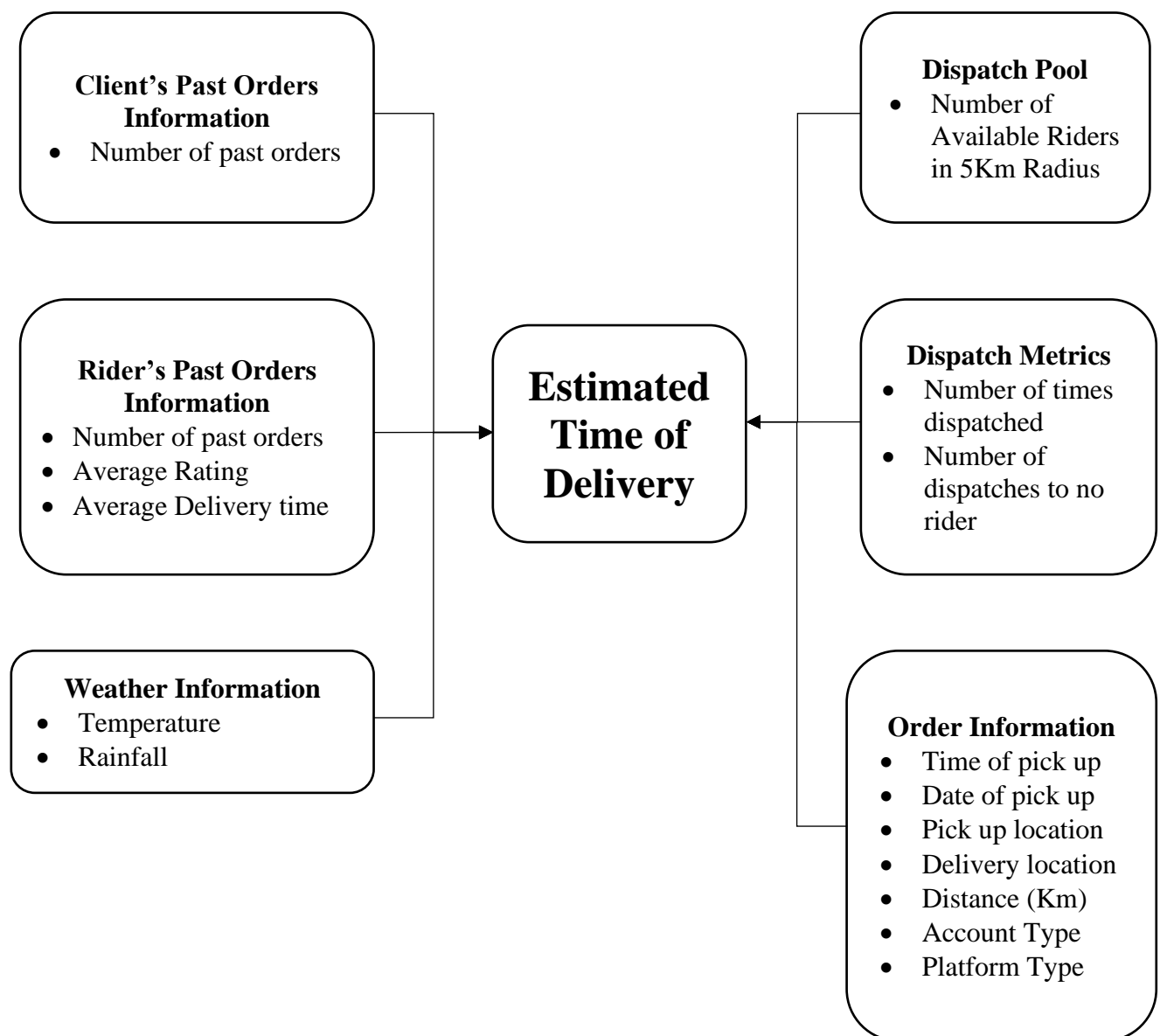


Figure 1: Conceptual Framework

CHAPTER 3 METHODOLOGY

3.1 Introduction

This chapter contains the methods that have been used in this study so as to attain the objectives of this study. This chapter has covered the framework of techniques, the data source, the data processing and modeling procedures, data mining and modeling techniques and the study variables.

3.2 Research Design

The framework of techniques and methods that have been used to conduct this study which combines different parts of the study to achieve the solution of the problem at hand is referred to as the research design.

The research design aims to answer the research questions mentioned before. The research design that fit this research perfectly was the Design Science Research Process (DSRP). According to Peffers et al., (2007) this research design is based on the Design Science Research Process (DSRP). This conceptual design science process model helps to produce and present high-quality research in information systems (IS).

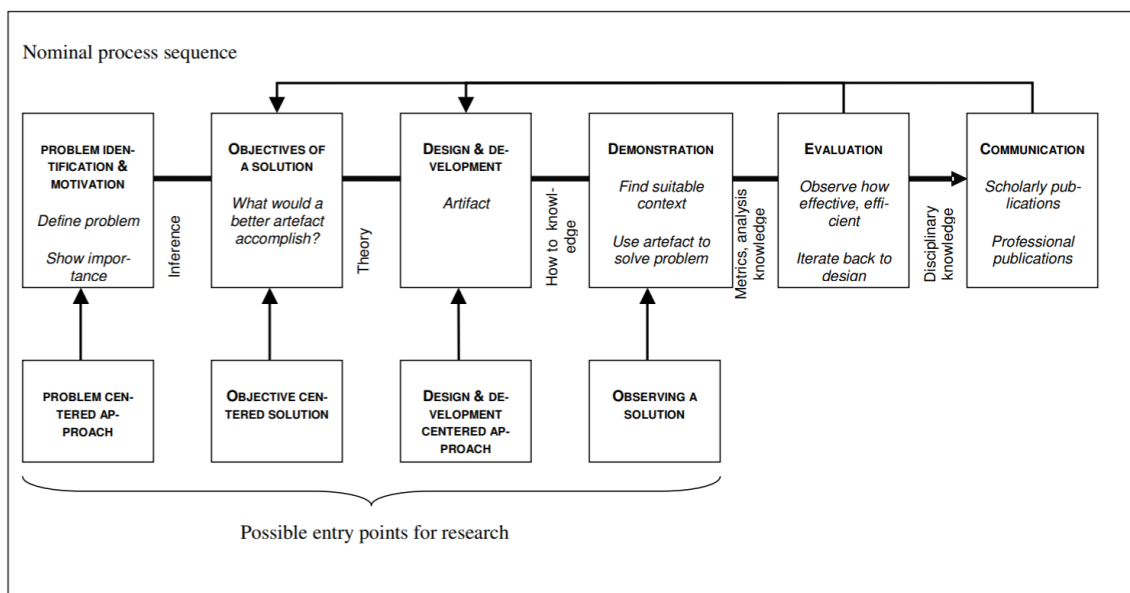


Figure 2: Research Design

DSRP takes into account principles, practices, and procedures that are necessary to conduct this study. This research design process includes six steps as shown in Figure 2 above. These steps are:

1. Problem identification and motivation – The study identified the problem of predicting motorcycle delivery in section 1.2 under the problem statement. The study has looked at previous literature in identifying the problem which is well laid out in section 1.2. Later in section 1.5 the motivation of the study was laid out in detail.
2. Definition of the objectives for a solution – As DSRP suggests, objectives of the study were laid out in section 1.3 where they were divided into two. The main objective of the study is to build a predictive analytics model, given a historical dataset, to accurately predict the delivery time of motor cycle packages in Nairobi. The specific objectives followed in section 1.3.2.
 - a. To determine the significant variables for determining an estimate for the delivery time of an order in Nairobi.
 - b. To investigate machine learning techniques that can be used for building a model to predict the estimated delivery time.
 - c. To develop a model to predict the estimated time of delivery.
 - d. To test and validate the model developed in iii above.
3. Design and development – The design and development step consisted of coming up with a research design and
4. Demonstration – The demonstration on the model that has been designed and developed for this study have well been laid out in the results in chapter 4. A summary of the model that has been designed for this study and its demonstration is in chapter 5.
5. Evaluation – The validation step of this study has also been conducted so as to evaluate the developed model. This is well laid out in chapter 4. Two methods were used to

evaluate which one is closer to the real-world scenario. A summary of the findings can be found in chapter 5.

6. Communication – This document is the method that I have decided to use to communicate the problem and the importance of solving the problem.

3.3 Data Source

I have partnered with Sendy Limited for this research. Sendy Ltd. is a crowd sourced courier marketplace which is focused on tackling on demand, last mile and hyper-local deliveries. Sendy Ltd. provides a mobile application and web platform that enables individuals and businesses users to connect with riders and drivers and request on-demand or scheduled courier services at any time, any day, 24/7.

The data was collected by querying Sendy's database then randomized and anonymized without my knowledge and I was sent a csv document of the data that I have used in this study.

3.4 Data Processing and Modeling Procedure

The researcher requested for a go ahead to obtain the data from the database and various representatives' partners and customers. Consent was be received via a letter for data collection which was obtained from the University. The data was processed as shown in the diagram below.

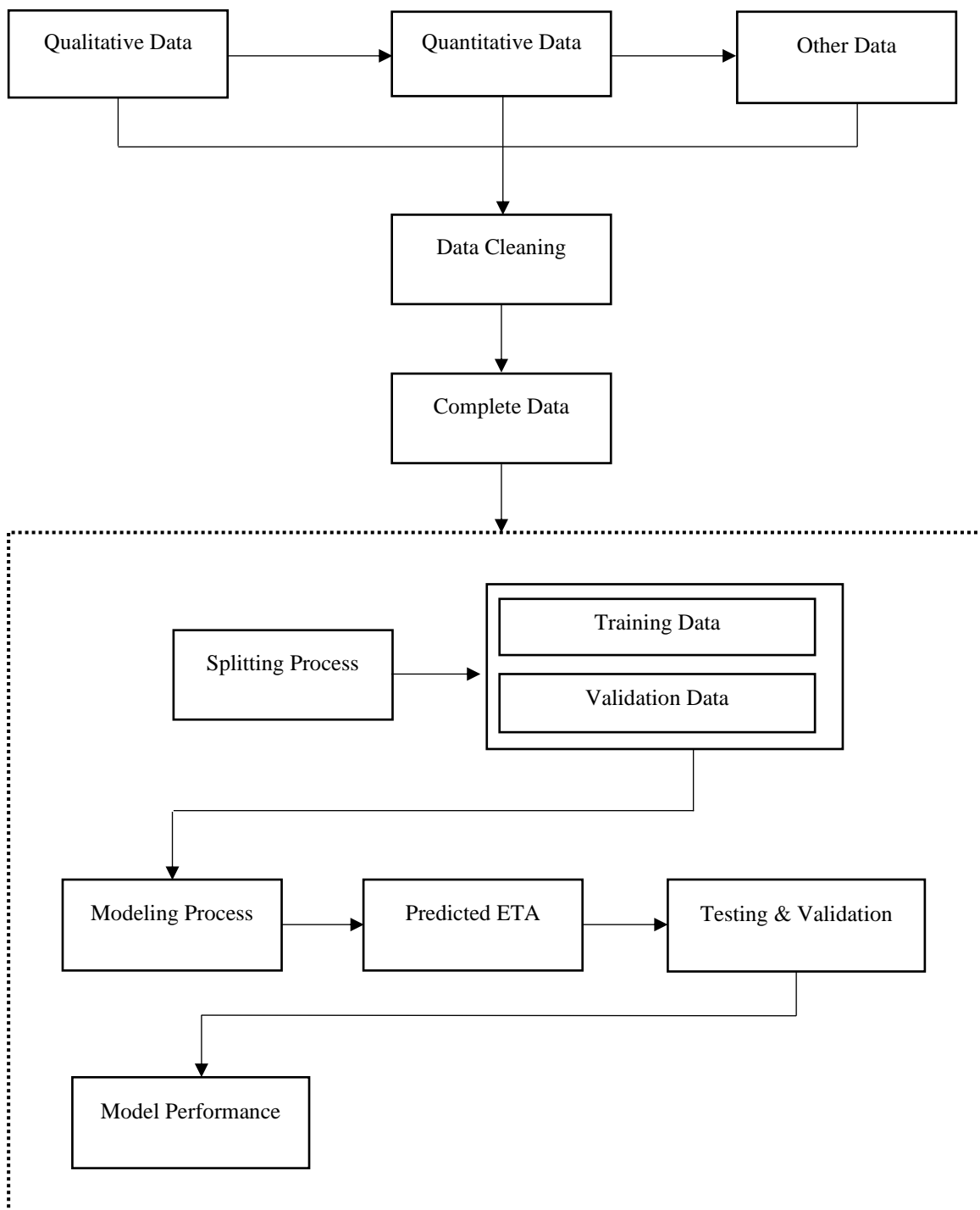


Figure 3: Data Processing and Modeling Procedure

Any observations that contain null values were removed. Thereafter, the data was divided into two. One division of the data – now referred to as the training data – consists of 80% of the data. The other division – now referred to as the testing data – consists of the 20% of the remaining observations. The training observations were used to assess the error of the final model chosen.

The dataset was randomly selected from 2016, 2017 and 2018. These modeling datasets were recursively split by selecting a random sample with no replacement so that the training set is 75% of the modeling dataset and the testing dataset made up 25%.

In order to equally assess all categories, the dataset included only one vendor type: motorcycles. The dataset also focused on orders that have only one pickup and one destination in essence point A to point B orders.

3.5 Data Modeling and Analysis

The data collected was analyzed using Microsoft PowerBI and Python on Jupyter Notebook. Creation of inferential statistics with the data provided is performed in the analysis of data, in addition to creation of the descriptive information. The data analysis involved generation of descriptive and inferential statistics. The descriptive statistics include frequency distribution tables and measures of central tendency (the mean), measures of variability (standard deviation) and measures of relative frequencies. The analyzed quantitative data was presented using tables, charts and graphs.

The algorithm used in predicting the ETD is XGBoost. XGBoost is a highly optimized distributed gradient boosting library. XGBoost was built and designed to be highly flexible, efficient and portable. Using machine learning algorithms under the Gradient Boosting framework, XGBoost is in a position to provides a parallel tree boosting (which is also referred to as GBM) which solve many data science problems in a fast and accurate way.

3.6 Study Variables

The conceptual framework in Chapter 2 illustrates the relationship between the variables both dependent and the independent in the study. In order to estimate an accurate Estimated time of delivery (ETD), several quantitative outcomes are to be measured. The significant variables in this study have been determined by previous studies done on the subject of predicting accurate ETD as discussed in the literature review. I have also included some of the variables that I thought might be significant, later on, we'll see how significant all the variables were to this model. I also had to work with the data that was provided by Sendy Ltd.

3.6.1 Delivery Time

The delivery time i.e. the time between when an order was placed and when the rider arrived at the destination is set to be the dependent variable.

3.6.2 Placement Timestamp

This is the date and time a customer has placed an order on the platform. This is a very important variable as there are peak times where the number of orders is significantly higher or lower than the other times. This variable was split into the day of the month, the weekday and the time.

3.6.3 Confirmation Timestamp

This is the time a rider confirmed an order after the order has been placed by the customer and dispatched by the system to the rider. It's important to note that the order was dispatched to many riders incrementally as time passes. This is a very important variable as there are peak times where the number of orders confirmed is significantly higher or lower than the other times. This variable was split into the day of the month, the weekday and the time.

3.6.4 Weather

Since the focus is purely in motorcycles, weather plays a huge role in predicting package arrival time. The assumption that drove about selecting the weather as a variable is the fact that during

the rainy days, it might take more time for the package to be delivered. The weather will include precipitation in millimeters and the temperature in degrees celcius.

3.6.5 Pickup Location

This is the longitude and latitude of where the package is to be picked up by the rider. A rider can only pick up the package after confirming the order.

3.6.6 Drop Off Location

This is the longitude and latitude of where the package is to be dropped off up by the rider. A rider can only drop off the package after confirming the order and picking up the same order.

3.6.7 Road Distance

The road distance in kilometers between the pickup location and the drop off location is vital to this study. This distance is automatically calculated using the google API which picks the best route to reach to the delivery location. The rider is usually shown this map after they have confirmed the order.

3.6.8 Customer Type

This refers to the customer who placed the order on the Sendy platform, the customer can either be a business or a peer user. There might be differences in the behavior of the two customers when comparing the delivery times.

3.6.9 Rider Details

Different riders have different speeds when it comes to logistics. We'll put this into consideration. Also, the platform allows for a customer to select a preferred rider, which makes this variable very important. The details of the rider include the number of orders he/she has done, how long he/she has been on the platform, the average rating and number of ratings.

3.6.10 Platform Type

Sendy is available on different platforms which are on web (via www.sendyit.com), an android application which is available on the play store, an iOS application which is available on the Apple play store and an API.

CHAPTER 4 RESULTS

4.1 Sample Description and Delivery Times

This chapter presents the main findings of the study. It begins by showing results of the summary descriptive statistics. The outcome of the methods was used to assess the prediction of the ETA.

4.1.1 Observations by Placement Day of Month

From the line chart below, it is clear that very few observations were recorded in the third week of the months.

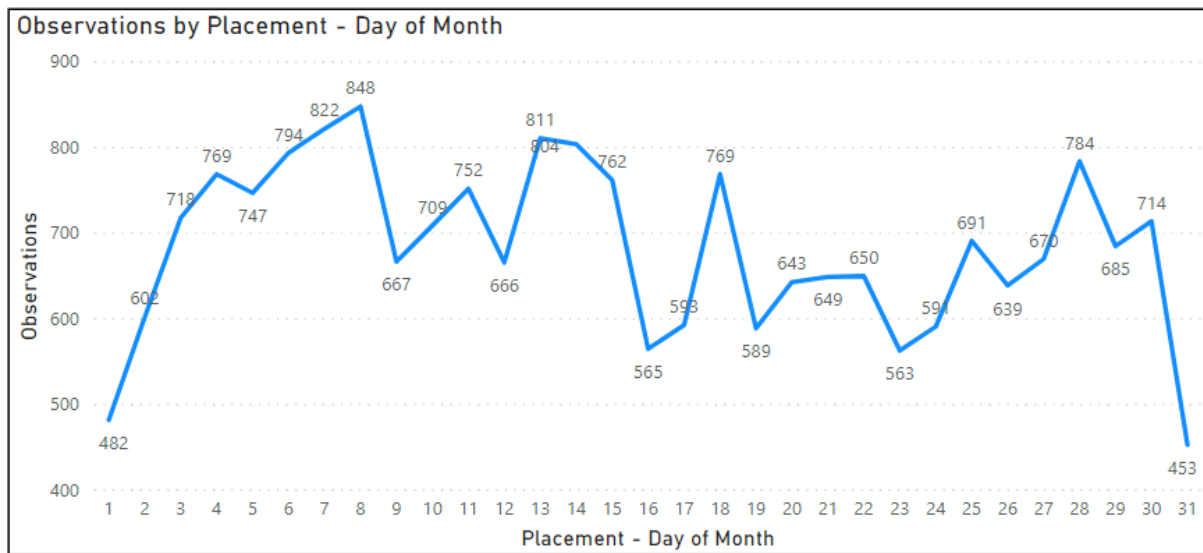


Figure 4: Observations by Placement Day of Month

4.1.2 Observations by Placement Time (1 Hour Bins)

Grouping the time to 1 hour such that orders within the same hour are grouped together. From the line chart below it seems that the observations start increasing during the working hours and fall after working hours.

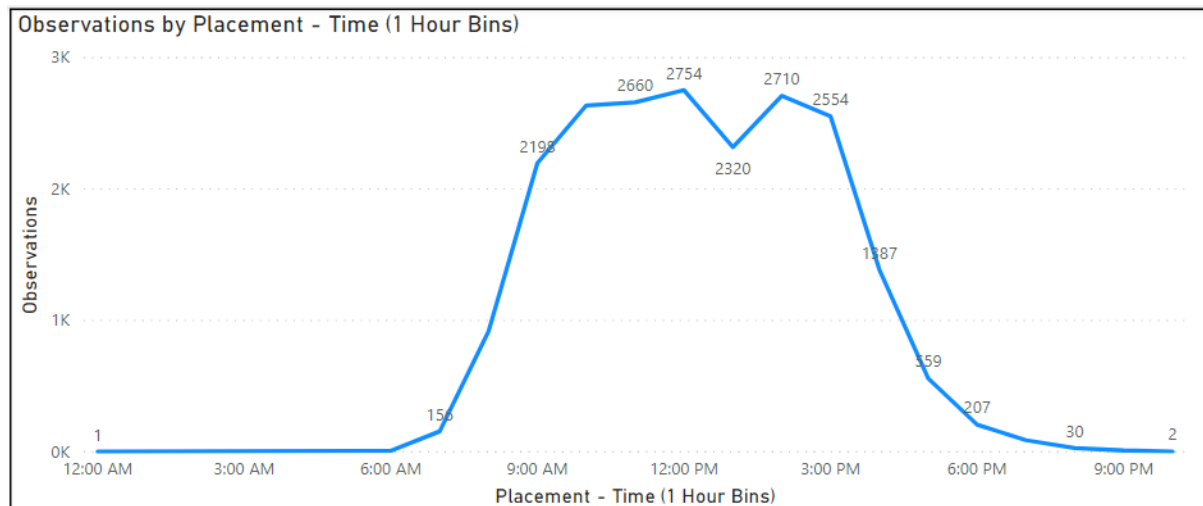


Figure 5 - Observations by Placement Time (1 Hour Bins)

4.1.3 Observations by Placement Weekday

Looking at the graph above, it makes sense to explore the days of the week that have the most observations. 1 corresponds to Monday while 7 is Sunday. Seems most orders are placed in the working days.

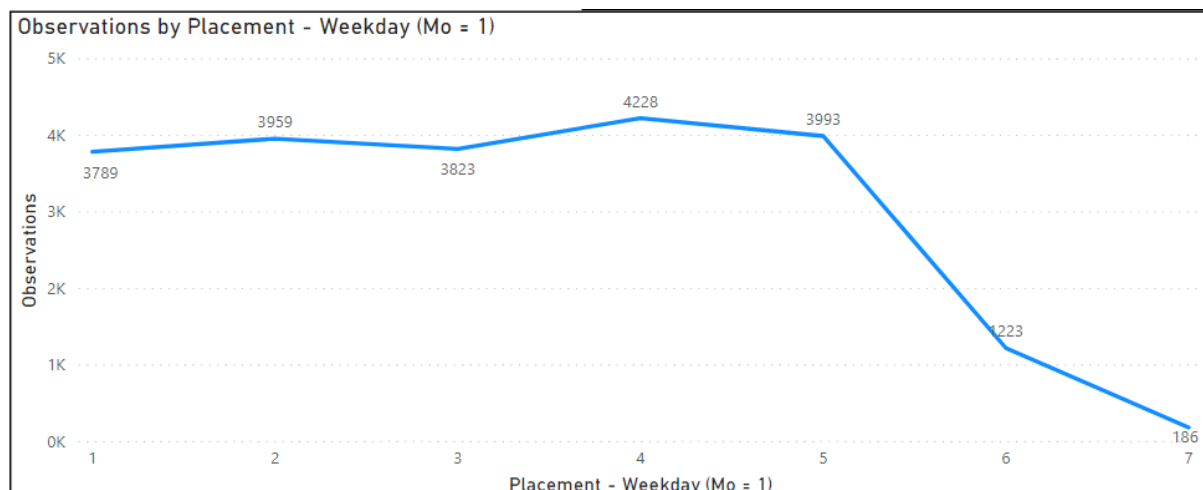


Figure 6 - Observations by Placement Weekday

4.1.4 Observations by Customer Type

The customer type in the data is divided into 2; personal and business. Keeping in mind the observations by placement day of month, day of week and time, the graph below affirms that most of the observations in the data are from Business.

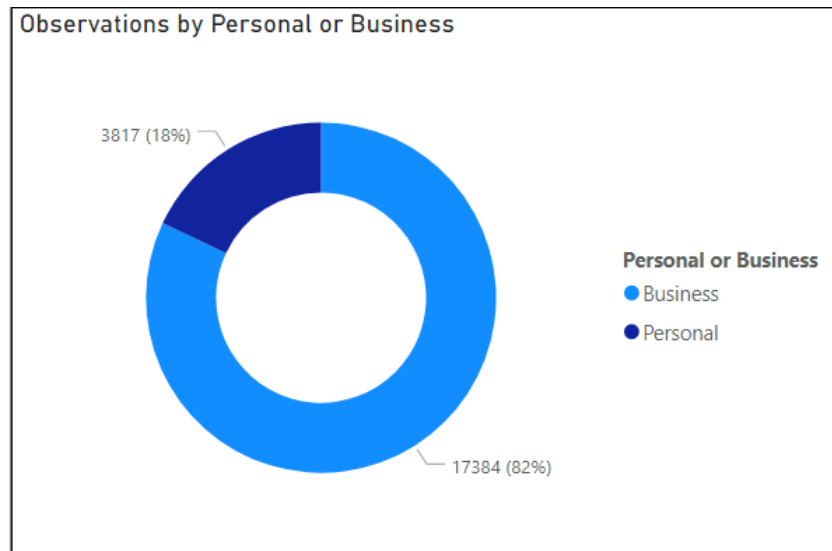


Figure 7 - Observations by Customer Type

4.2.6 Observations by Platform Type

There are 4 different platforms in the data. The platforms are numbered from 1 to 4. Platform 3 takes the largest share with 85% with platform 4 with only 0.09% of the orders.

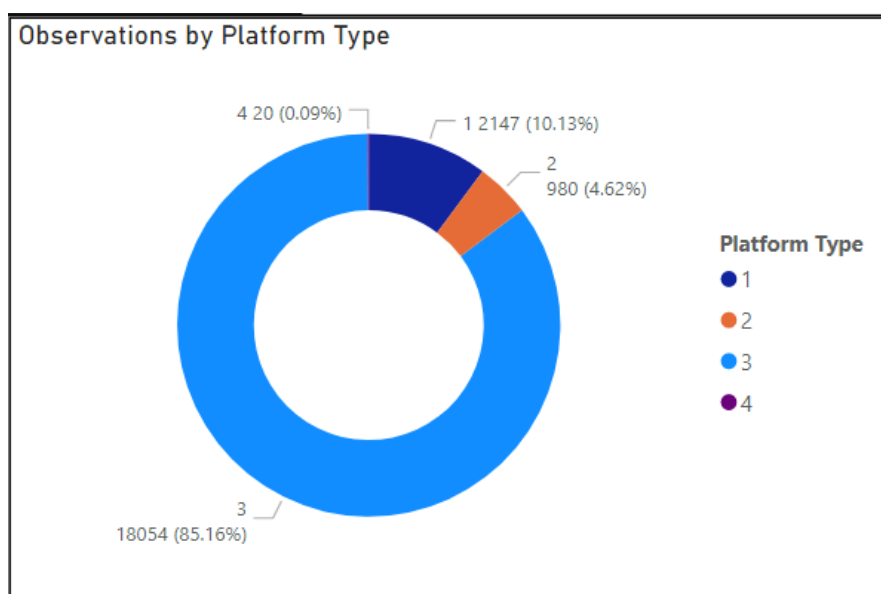


Figure 8 - Observations by Platform Type

4.2.7 Observations by Distance (Km)

The data provided only included bike orders, as expected, the distance wouldn't be very long. A larger proportion of the orders are below 10 kilometers between the pickup location and the destination.

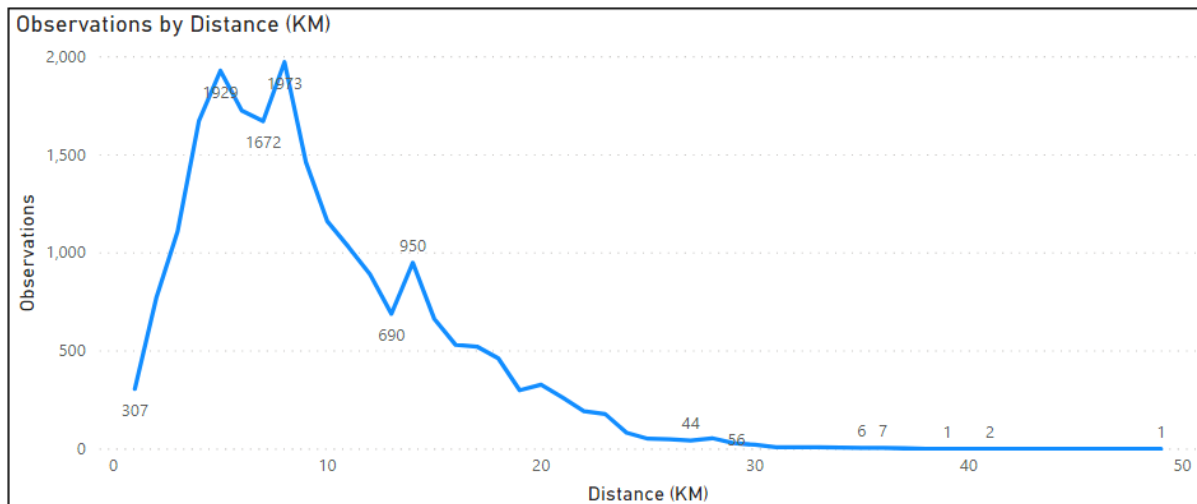


Figure 9 - Observations by Distance (Km)

4.2.8 Observations by Pickup Location

Using the latitude and longitude of every observation I can plot these on a map to see the distribution of orders. It is also going to be fairly easy to spot outliers by eye.

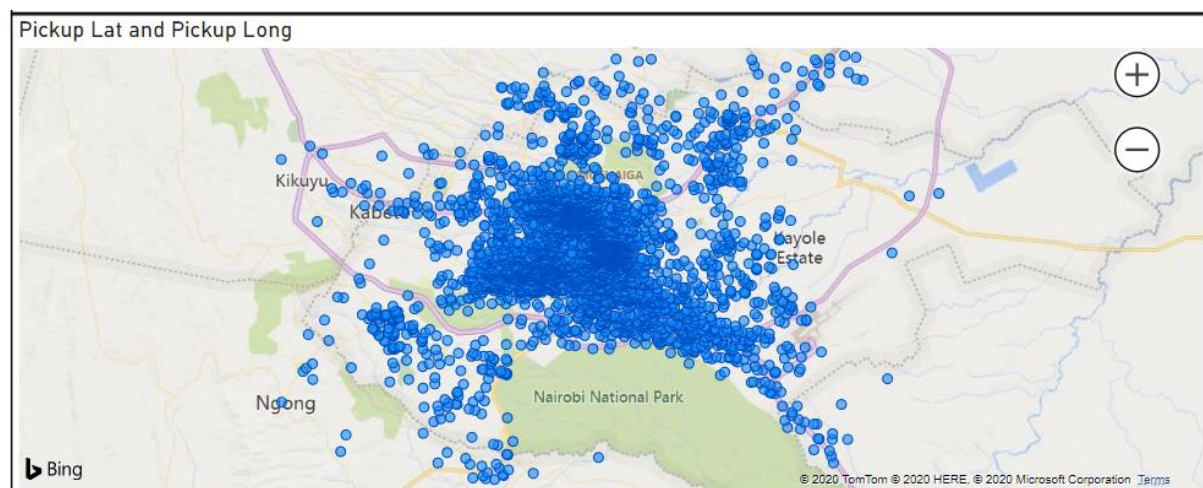


Figure 10 - Observations by Pickup Location

Plotting the latitude and longitude using a heat map, I get a picture of the distribution of the observations. I can get a picture of where the largest number of orders are to be picked up by the riders.

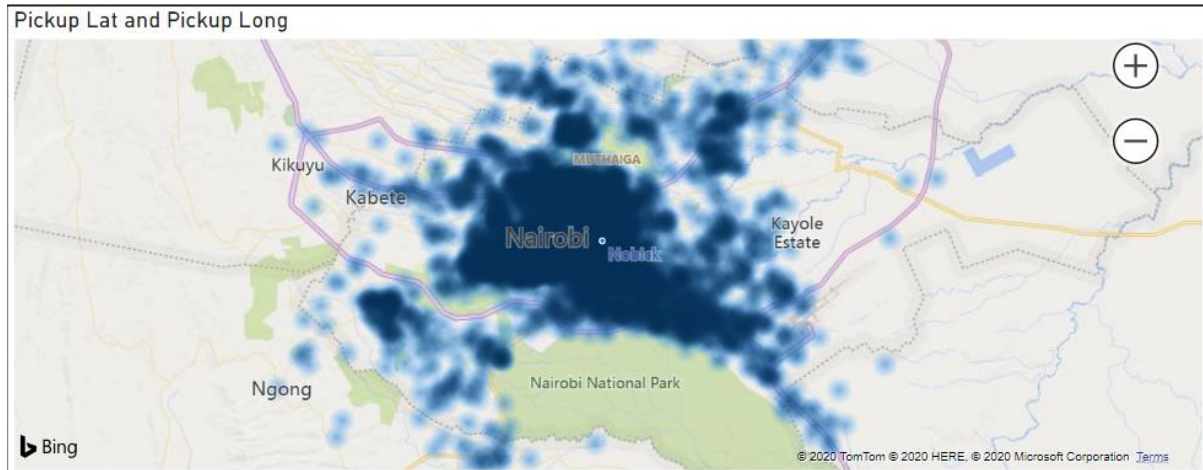


Figure 11 - Observations by Pickup Location (Heat map)

4.2.9 Observations by Destination Location

Plotting the destination of the deliveries on a map, I can get an idea of where the packages are dropped off. From the image below, it seems majority of the orders are concentrated in Nairobi.

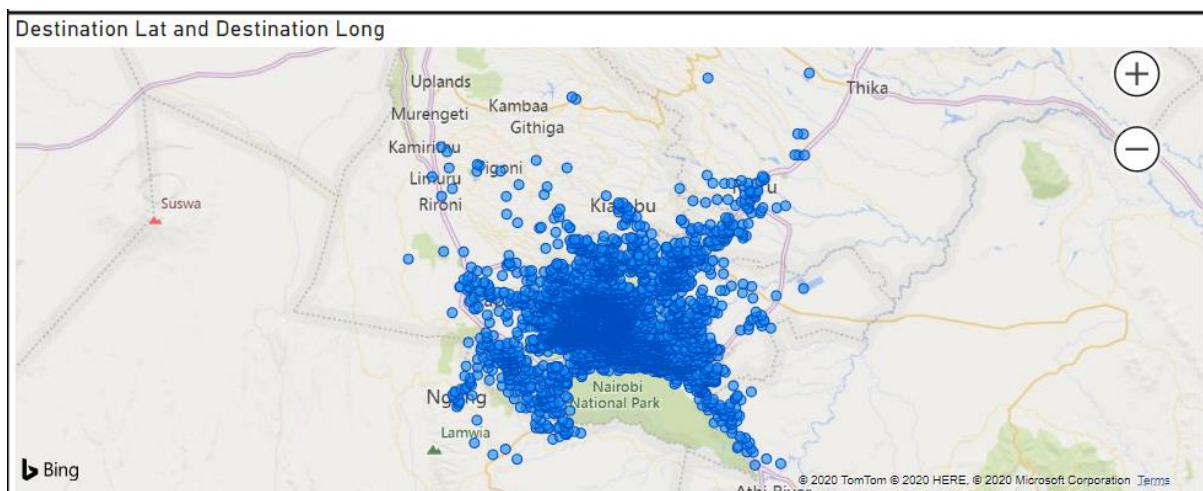


Figure 12 - Observations by Destination Location

Plotting the destination of the deliveries on a heat map, I can get an idea of where the packages are dropped off. Seems the data has a few going outside Nairobi.

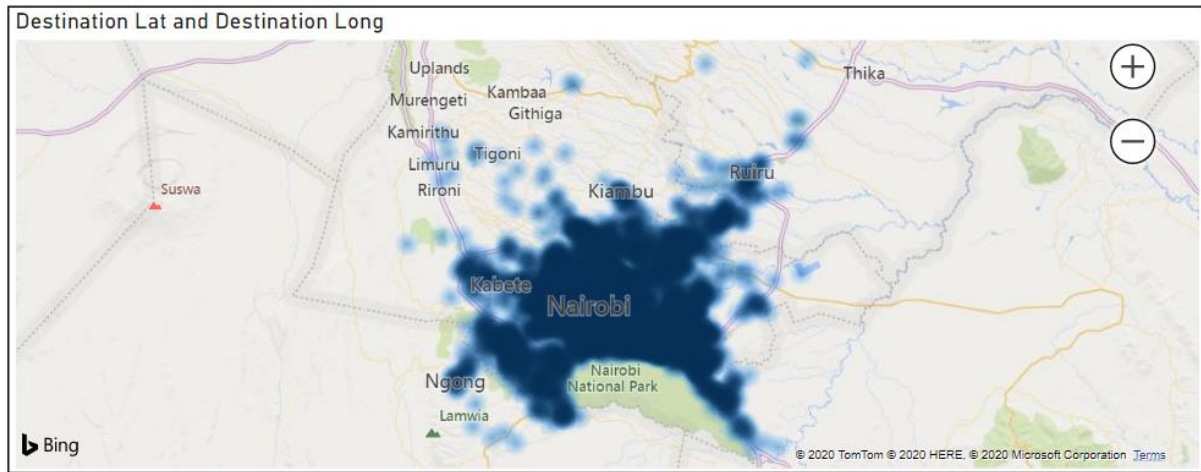


Figure 13 - Observations by Destination Location (Heat Map)

4.3 Predicting the Estimated Time of Delivery

The logistic regression model to predict the estimated time of delivery used the formula outlined in Chapter 3. First, the model was tested using part of the existing data set to test the ability and accuracy of the model in predicting accurate estimated times of delivery. The test for this was carried out by randomly splitting the data into training data and testing data. The training data took 80% of the existing data and testing was done using the remaining 20%. The entire data set consisted of 28,269 rows of data. I assigned the training data to 80% of the rows (22,615 observations) and I assigned the remaining 20% to test the accuracy of the models (5654 observations).

4.3.1 Data Profiling

Data profiling is monitoring and cleansing data. Data processing, analysis cannot happen without data profiling. Data profiling helps cover the basic information about the data provided for this study. Data profiling also helps in verifying that the information in the data matches the descriptions. For this data I used column profiling as the data profiling section. Column profiling scans through a table and counts the number of times each value shows up within each column. This method can be useful to find frequency distribution and patterns within a column of data.

4.3.1.1 Arrival at destination day of month

This refers to the day of the month, which took the values between 1 and 31, that the rider arrived at the destination. The column is represented as ‘Arrival at Destination – Day of Month’. The diagram below shows the statistics of the arrival at destination column.

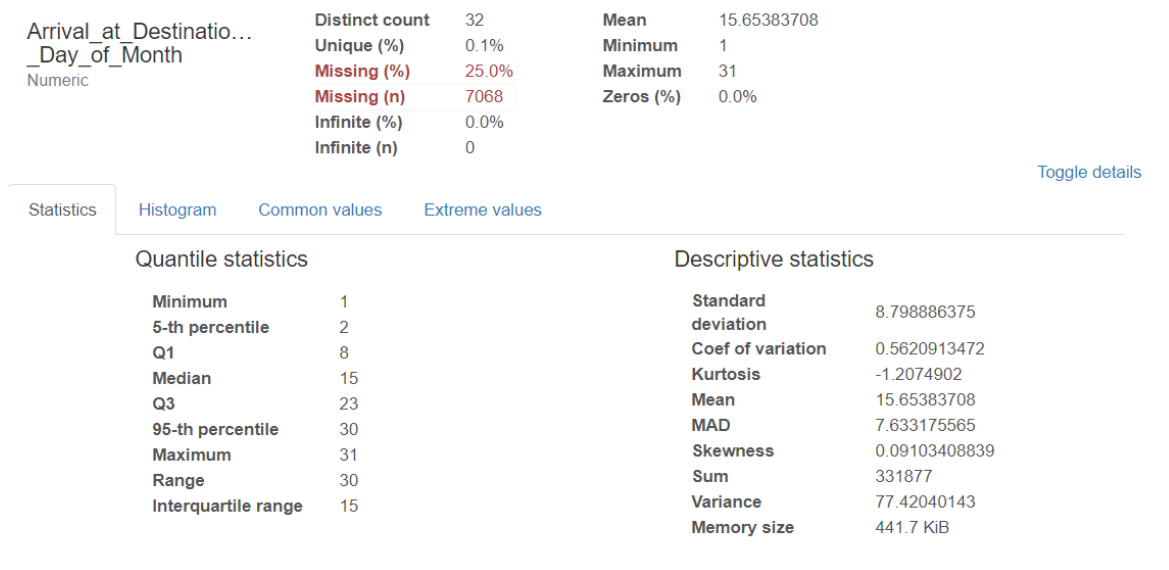


Figure 14 - Arrival at destination day of month statistics

The image below shows the histogram of the column.

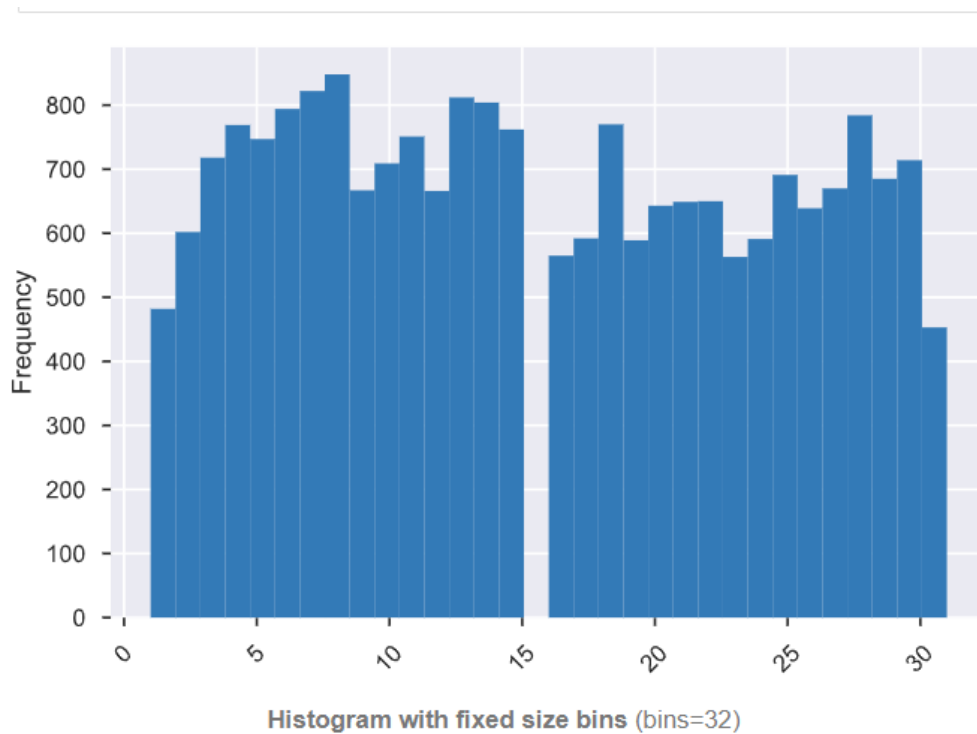


Figure 15 - Arrival at destination day of month histogram

4.3.1.2 Arrival at destination time

This column shows when the rider arrived at the destination. The column is represented as 'Arrival at Destination Time'. The image below shows the statistics of the column.

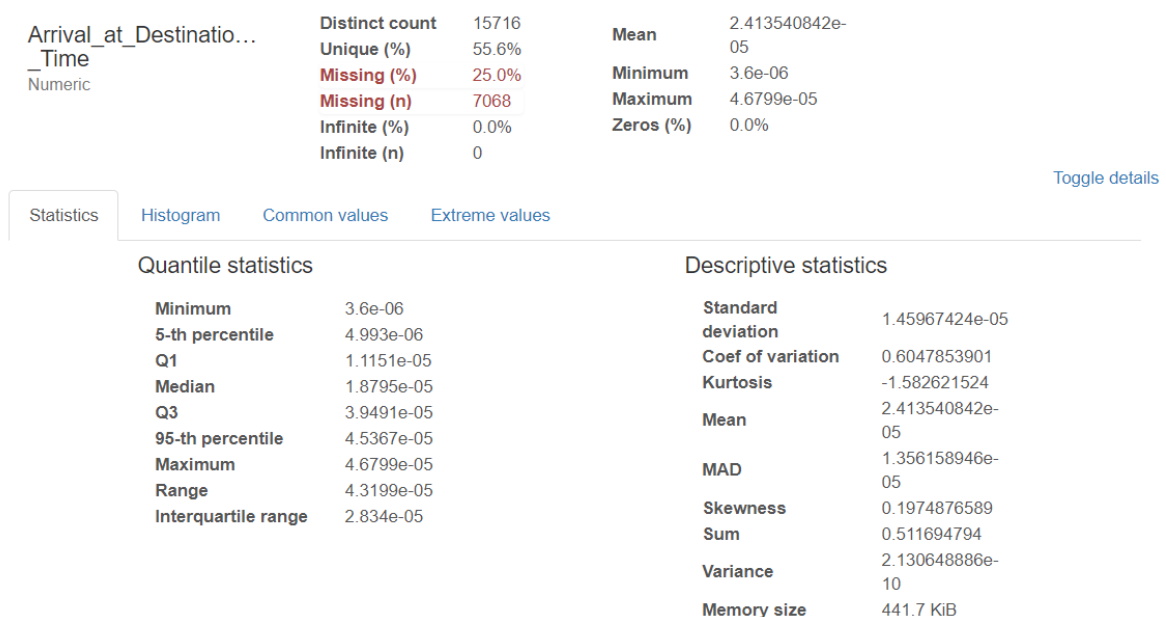


Figure 16 - Arrival at destination time statistics

A histogram shown below depicts the distribution of the arrival at destination time column.

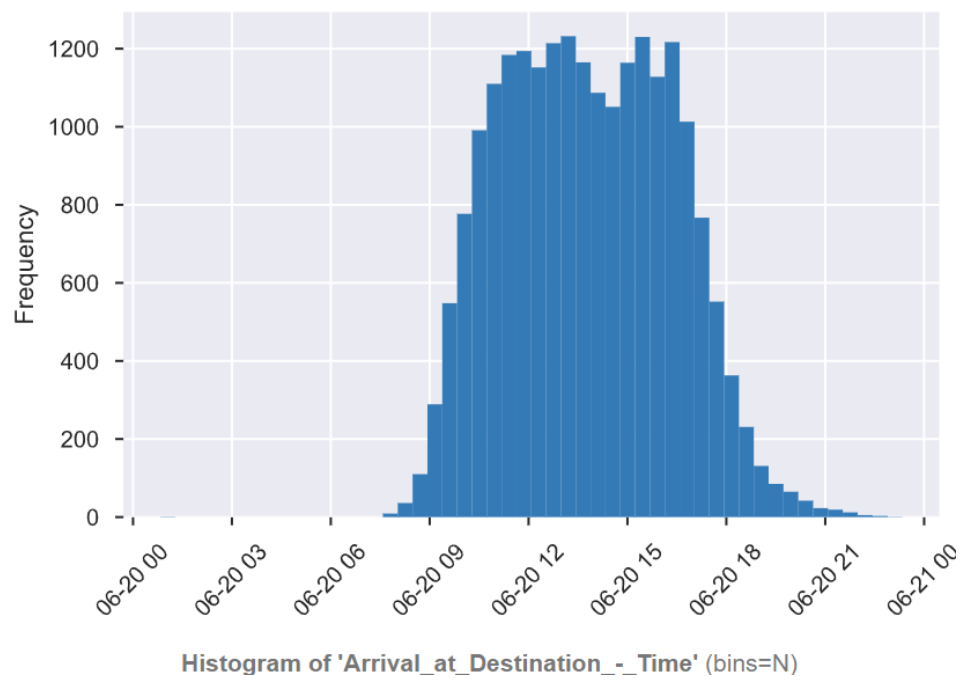


Figure 17 - Arrival at destination time histogram

4.3.1.3 Arrival at Destination Weekday

This is the day of the week that the rider arrived at the destination. The values for this variable range from 1 to 7. 1 is Monday, 2 is Tuesday, 3 represents Wednesday, 4 represents Thursday, 5 represents Friday, 6 represents Saturday and 7 represents Sunday.

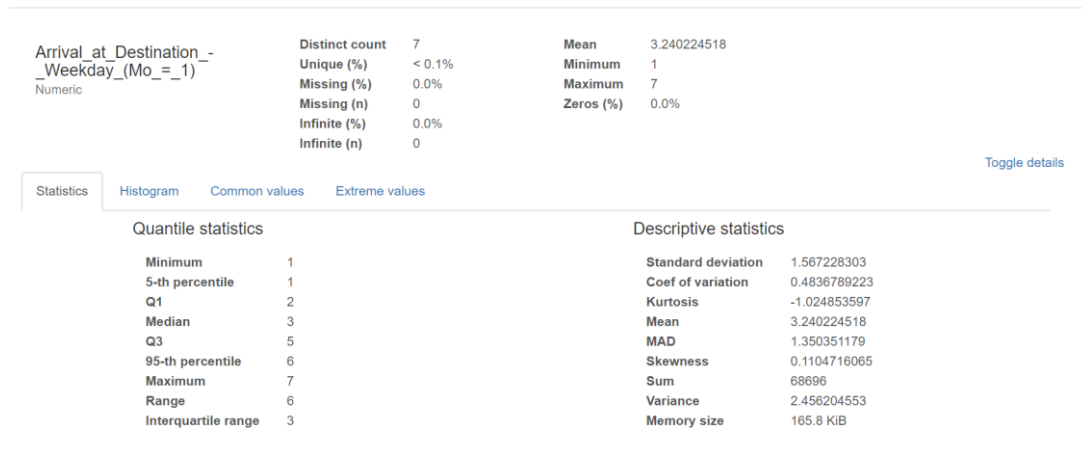


Figure 18 - Arrival at Destination Weekday Statistics

The diagram below shows the distribution of the variable.

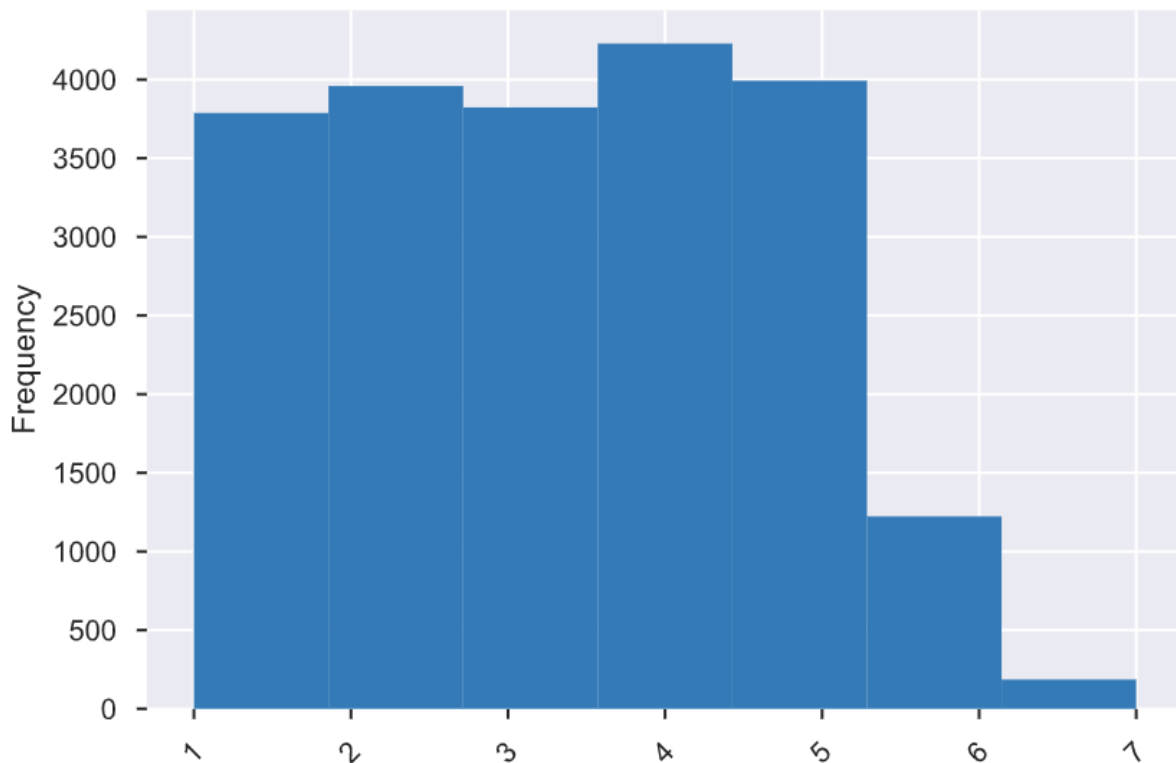


Figure 19 - Arrival at Destination Weekday Histogram

4.3.1.4 Arrival at Pickup Time

This refers to the time the rider arrived at the pickup. The column is represented as 'arrival at Pickup – Time' in the data. The diagram below shows the statistics and the distribution of the observations. The clear take away from this histogram below is that the orders start declining from 3pm, represented as 06-13-15 in the histogram below. Also, there are little to no orders between 9pm and 6am.

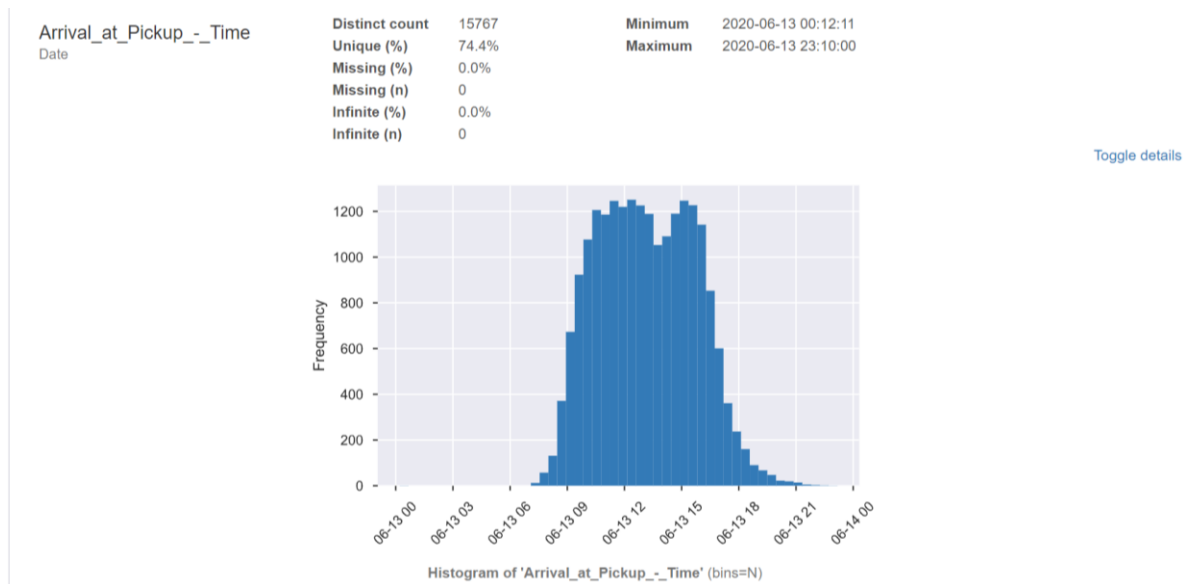


Figure 20-Arrival at pickup time statistics and histogram

4.3.1.5 Confirmation Time

This refers to the time when the rider confirmed the order. In the data this column is represented as ‘Confirmation – Time’.

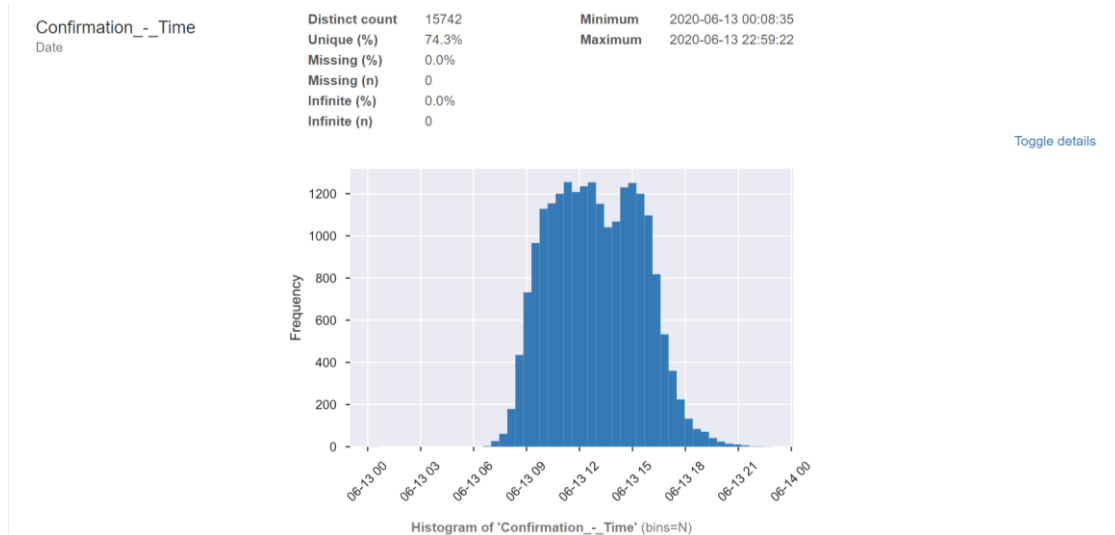


Figure 21 - Confirmation time statistics and histogram

4.3.1.6 Destination Latitude

This refers to the time when the rider confirmed the order. In the data this column is represented as ‘Destination Lat’.

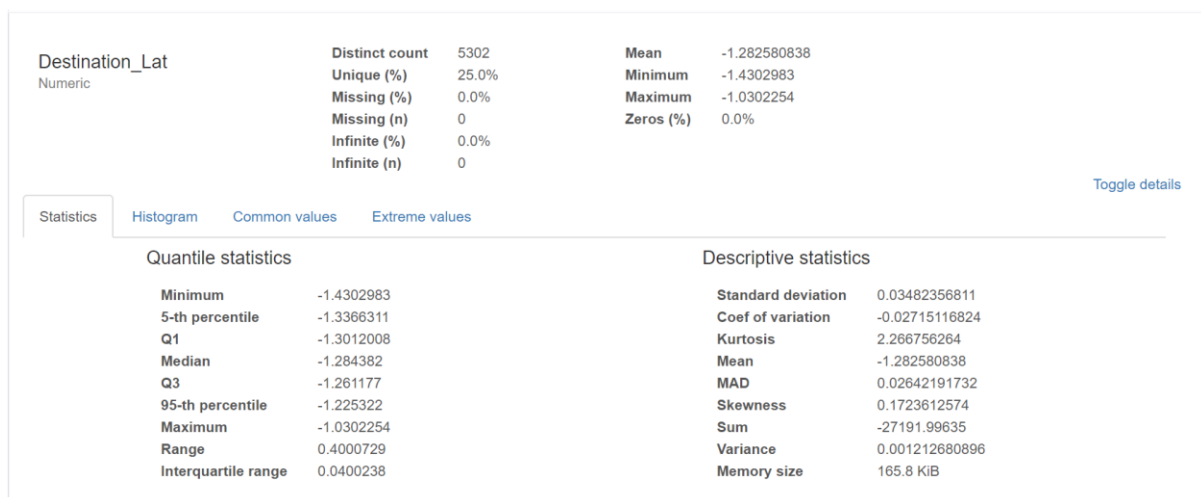


Figure 22 – Destination latitude statistics

The histogram presented below gives us an idea of where most orders go to. A clearer picture of this is shown above in section 4.2.10.

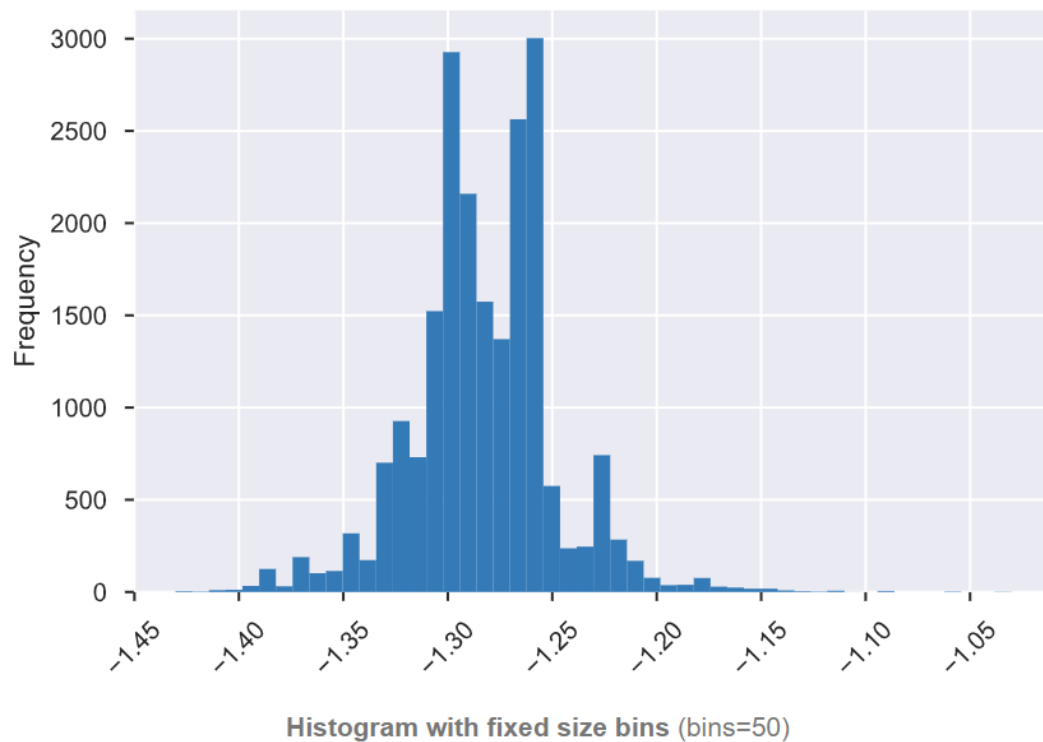


Figure 23 – Destination latitude histogram

4.3.1.7 Destination Longitude

This refers to the time when the rider confirmed the order. In the data this column is represented as ‘Destination Long’.

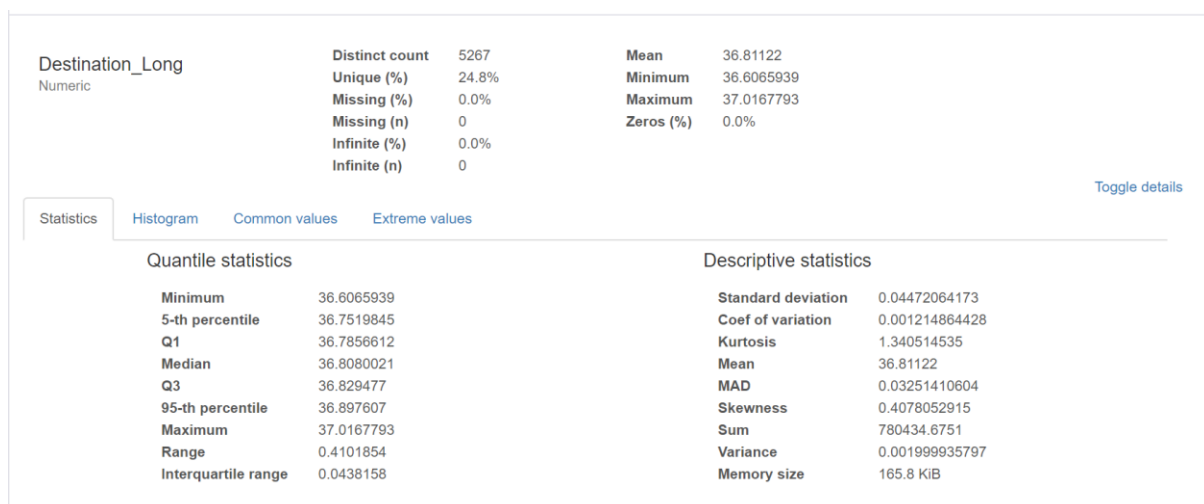


Figure 24 - Destination Longitude Statistics

The histogram presented below gives us an idea of where most orders go to. A clearer picture of this is shown above in section 4.2.10.

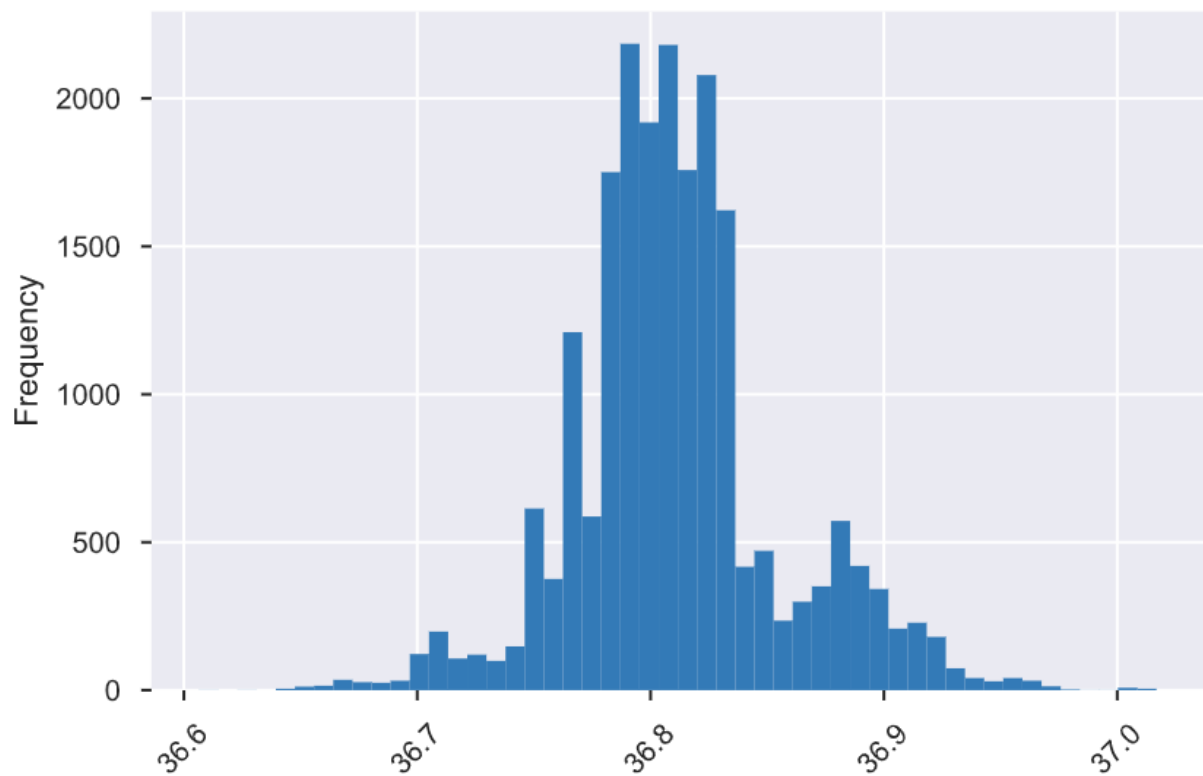


Figure 25 - Destination Longitude Statistics

4.3.1.7 Distance (Km)

This variable in the data set represents the road distance between pickup and destination. The best route is calculated by the Google API and shown to both the customer and the rider. The road distance is also used to calculate the price of the order and the partner amount. This is represented as 'Distance (KM)' in the data set.

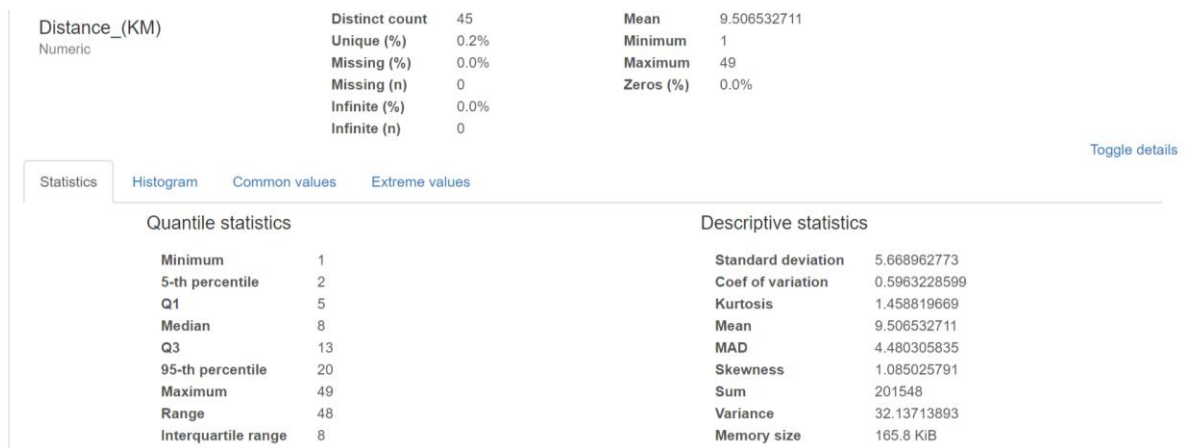


Figure 26 - Distance Statistics

The diagram below shows the distribution of the distance.

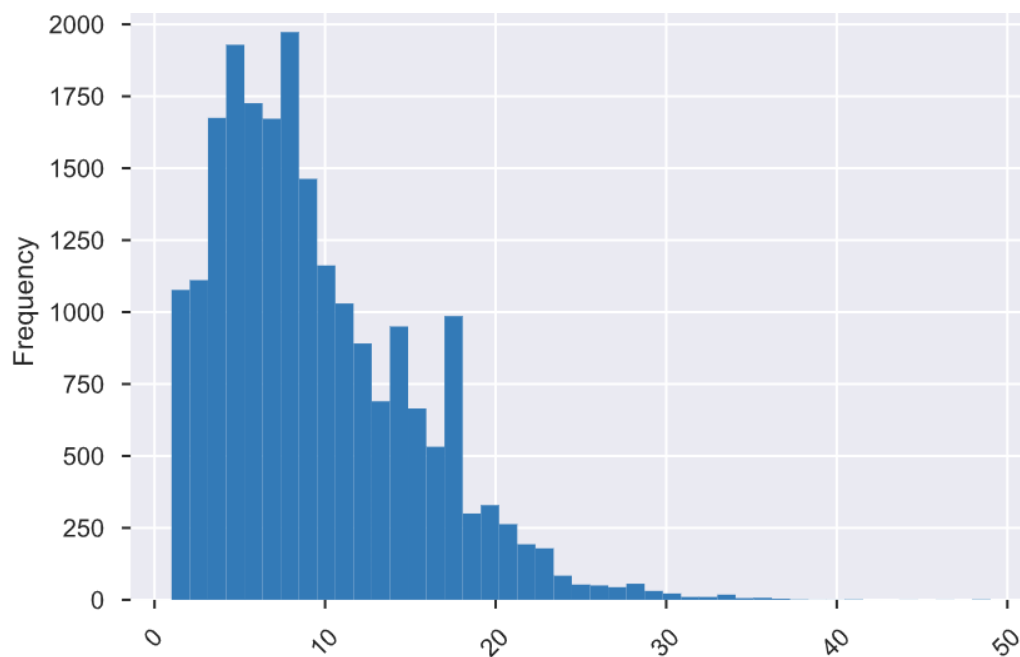


Figure 27 - Distance (KM) histogram

4.3.1.7 Personal or Business

Sendy has two types of clients; personal and businesses. This variable is represented as 'Personal or Business'.

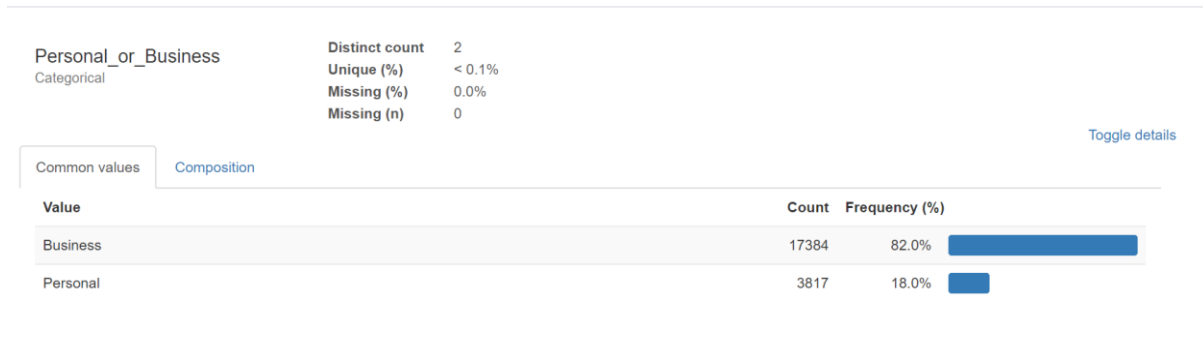


Figure 28 Distribution of Personal or Business

4.3.1.8 Pickup Time

This is the time the rider marked that he has picked the order. This variable is represented as 'Pickup – Time'

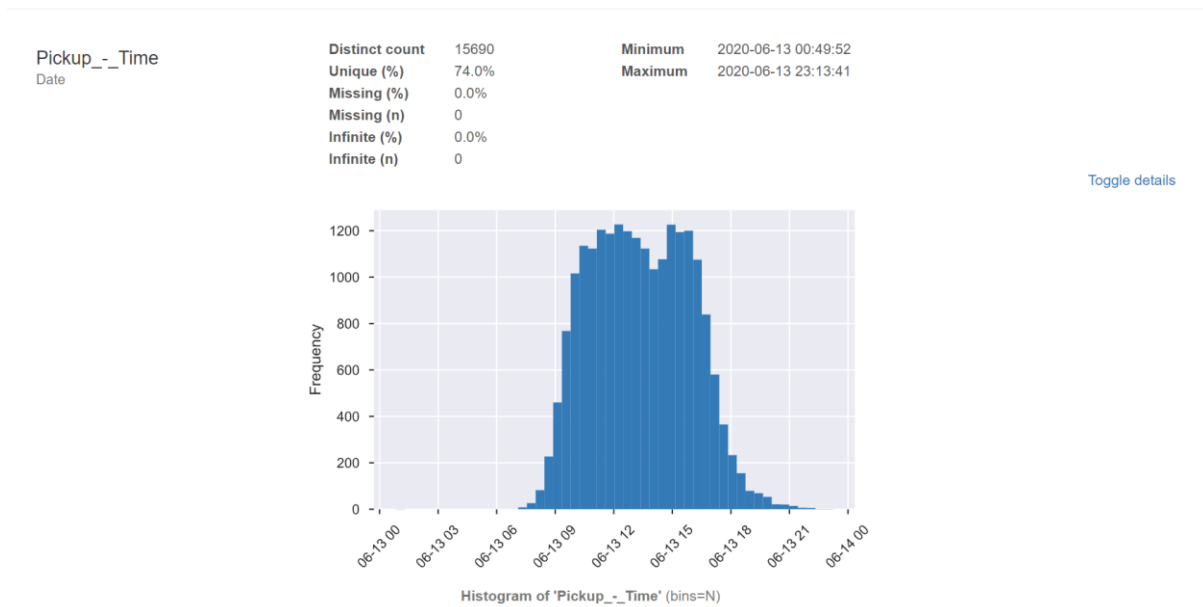


Figure 29 - Statistics and distribution of pickup time

4.3.1.8 Pickup Latitude

This is the latitude of the pickup location. This variable is represented as 'Pickup Lat' in the data.

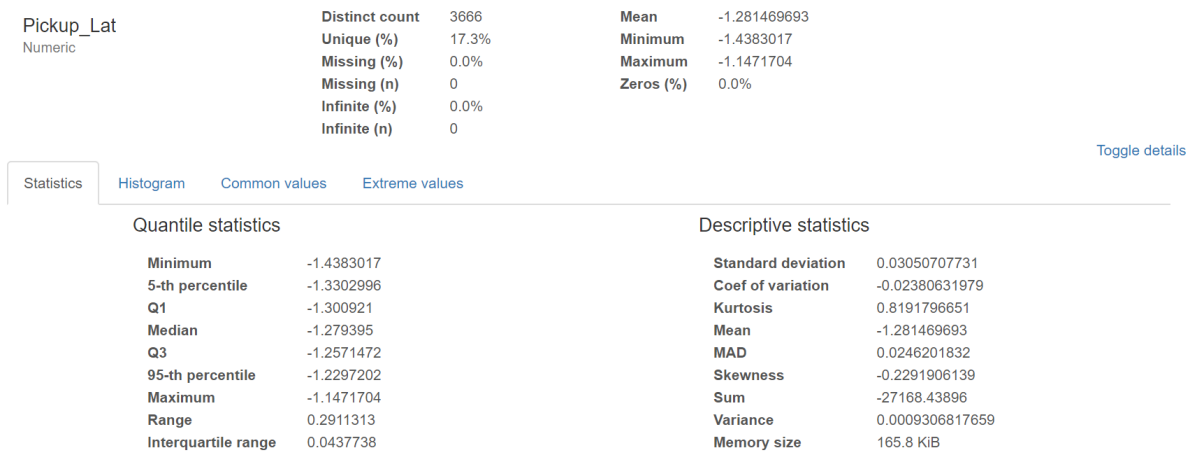


Figure 30 - Pickup latitude statistics

The distribution of orders and the pickup location is represented below.

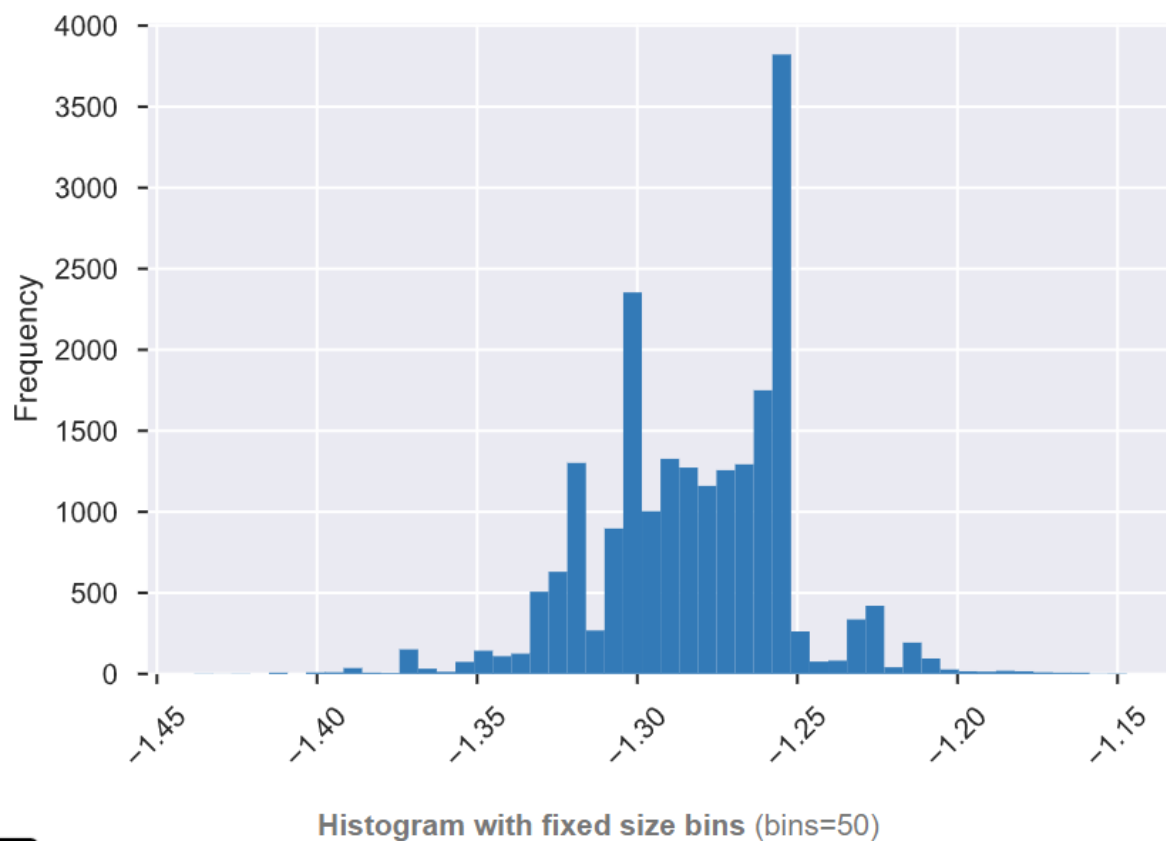


Figure 31 - Pickup latitude histogram

4.3.1.8 Pickup Longitude

This is the latitude of the pickup location. This variable is represented as 'Pickup Long' in the data.

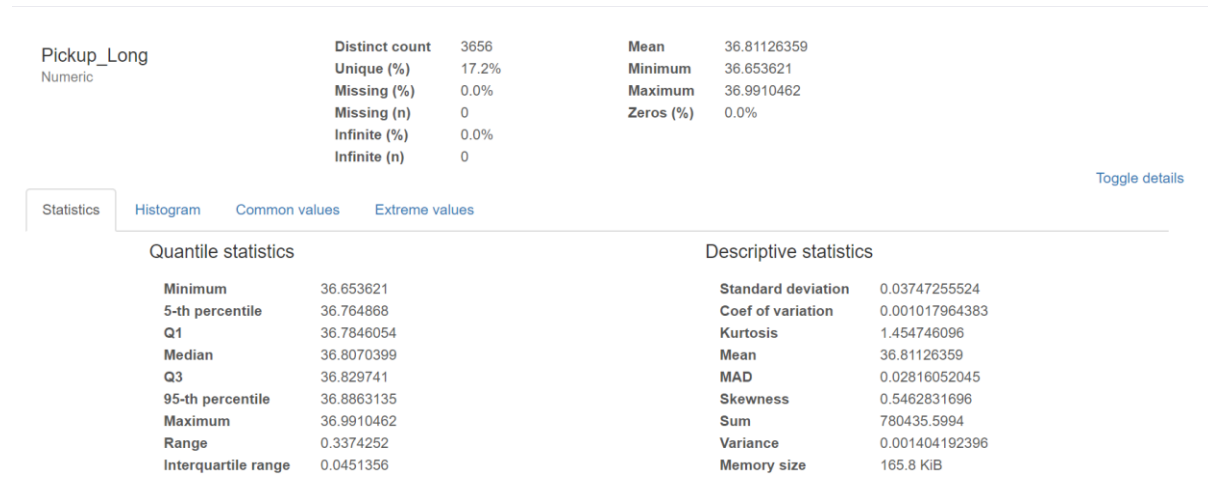


Figure 32 - Pickup longitude statistics

The distribution of orders and the pickup location is represented below.

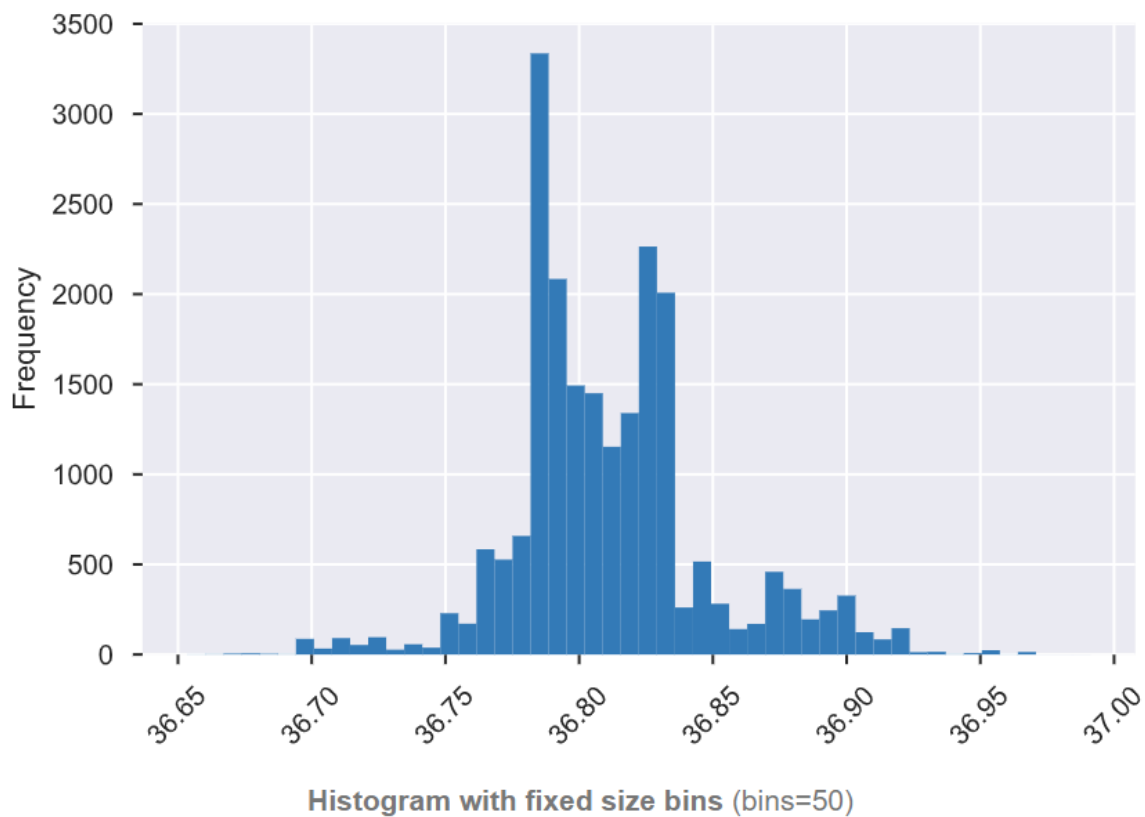


Figure 33 - Pickup longitude histogram

4.3.1.9 Placement Time

The image below shows the statistics and the distribution of the placement time. This would help answer the question when are orders in the system placed. This is represented as 'Placement – Time' in the data.

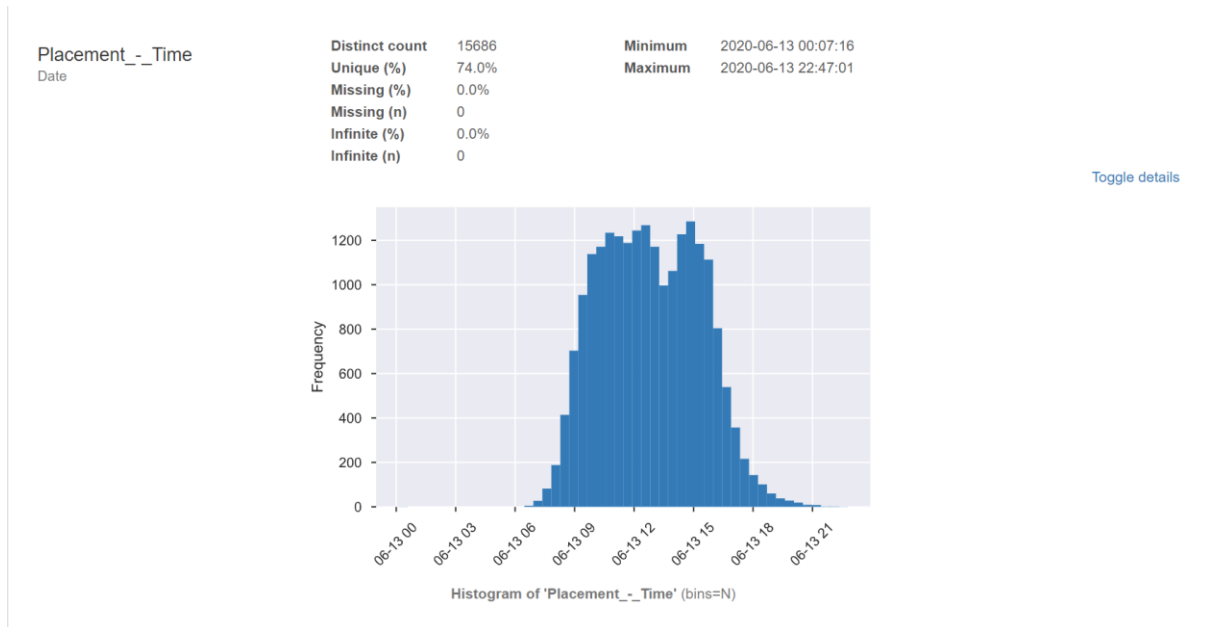


Figure 34 - Placement time statistics and histogram

4.3.1.10 Platform Type

Sendy offers the users different platforms to place order, these are; Web, API, Android application and an iOS application. The data has been anonymized so I can't know which is which. This variable is labelled 'Platform Type' in the data.

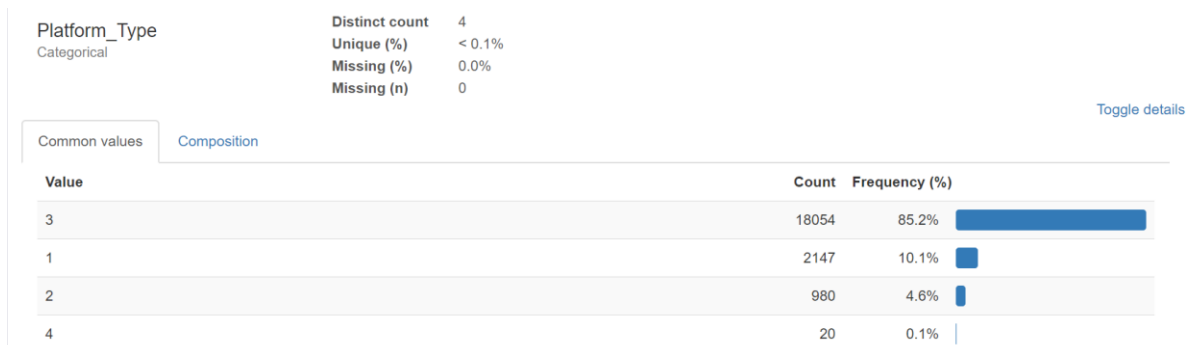


Figure 35 - Platform Type statistics and distribution

4.3.1.11 Precipitation in Millimeters

Our data is only comprised of motorcycles orders, with this in mind, I know that when there is heavy rainfall, motorcycles get affected negatively. This variable is represented as 'Precipitation in Millimeters' in the data set. The first thing that comes out from the diagram below is that 97.4% of this variable is missing from the observations. It seems that the data was recorded during a cold month.

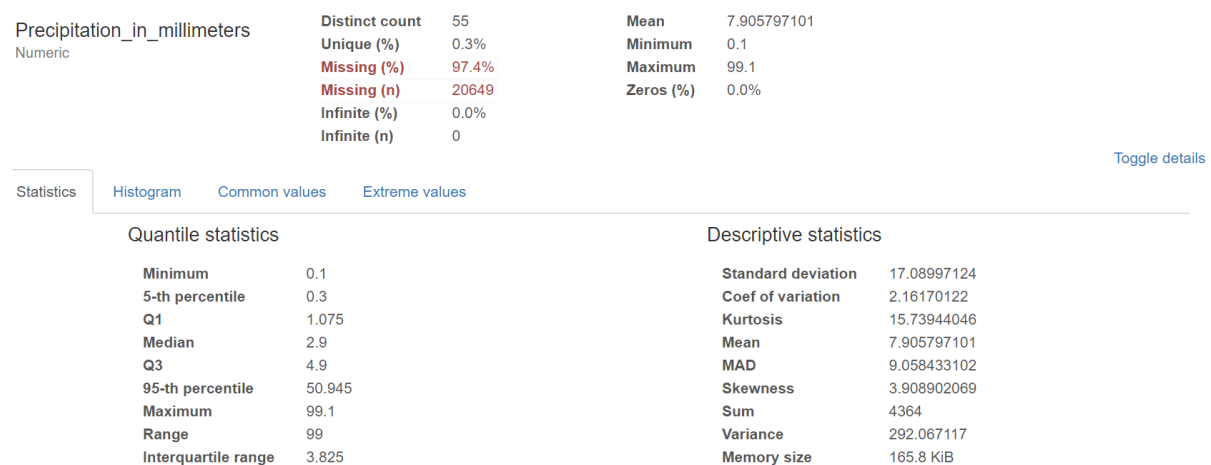


Figure 36 - Precipitation in millimeters statistics

From the histogram below, it seems that 2.6% of the observations had very little rainfall in ml.

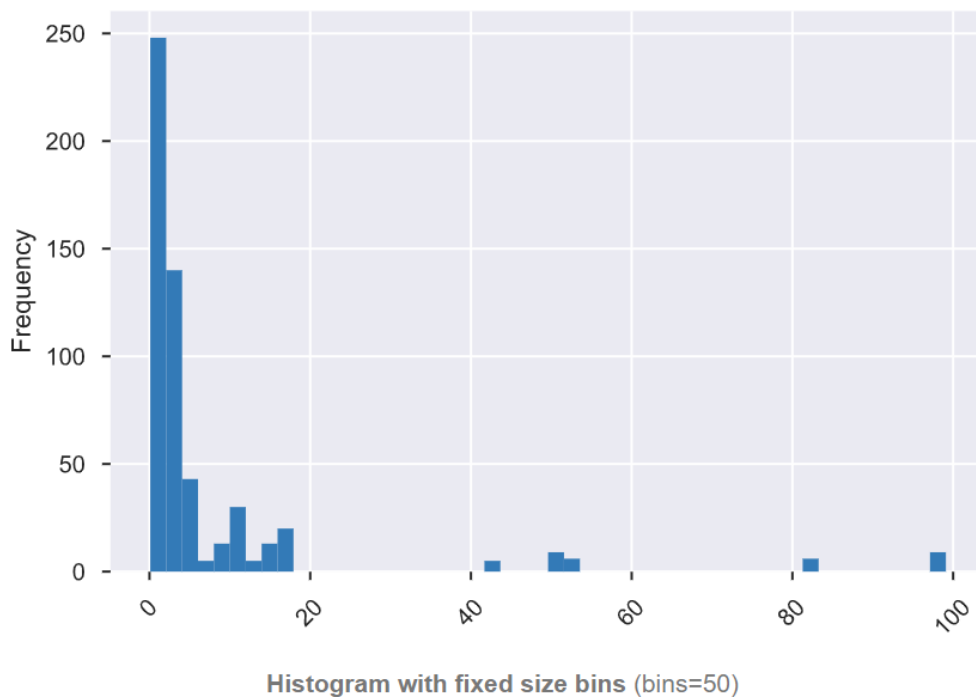


Figure 37 - Precipitation in millimeters histogram

4.3.1.13 Temperature

The temperature is recorded in degrees Celsius. From the image below, I can see that 20.6% of the provided data is missing this variable. This variable is represented as 'Temperature'.

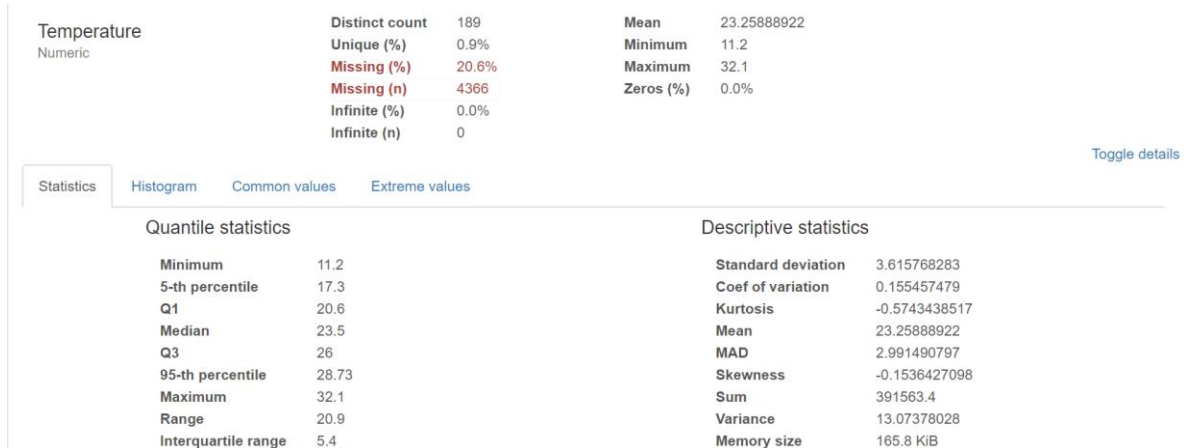


Figure 38 - Temperature statistics

The distribution of the data is shown below.

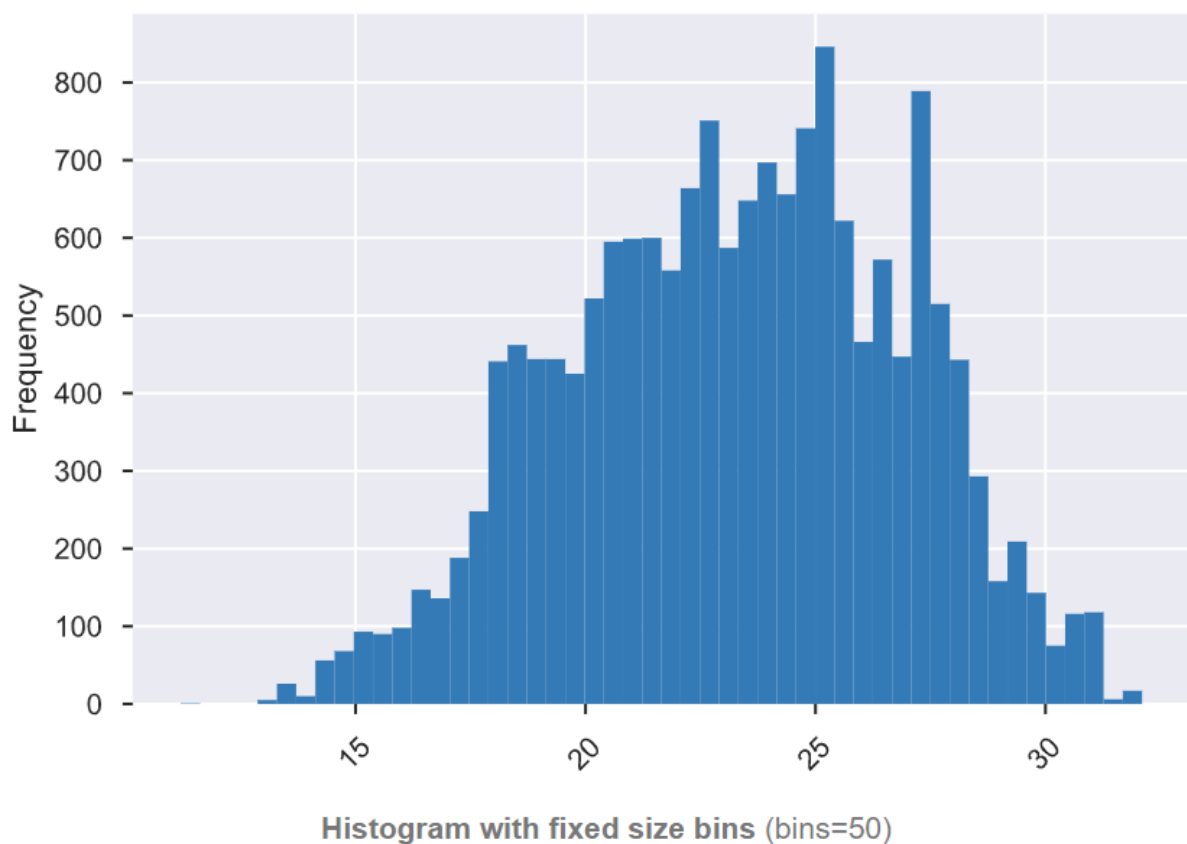


Figure 39 - Temperature histogram

4.3.1.14 Time from pickup to arrival at destination

The image below shows the time between when the rider picked up the order and when the rider arrived at the destination. This time is recorded in seconds.

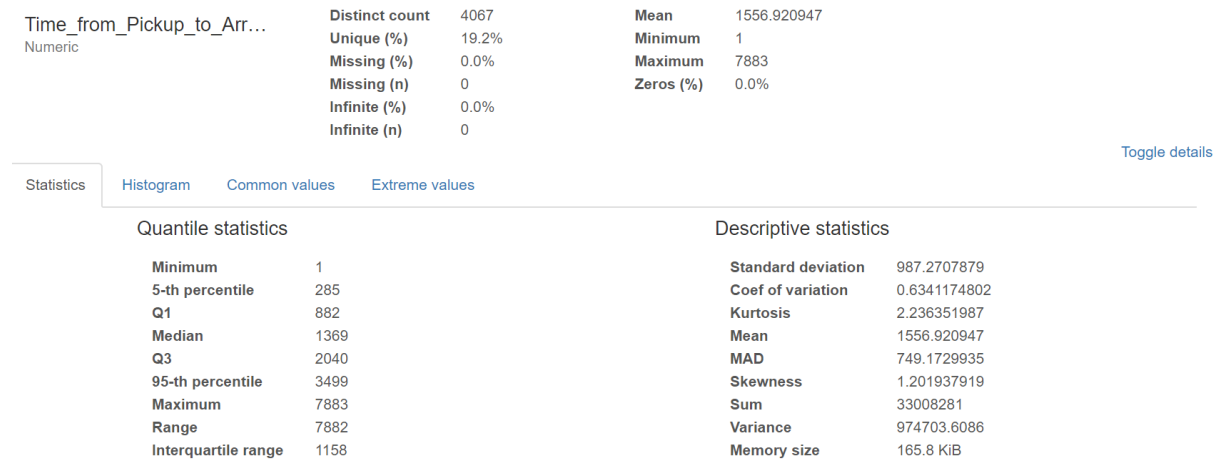


Figure 40 - Time from pickup to arrival at destination statistics

The histogram below shows us the distribution of the observations. What is surprising is that there are orders that the time is 0 seconds.

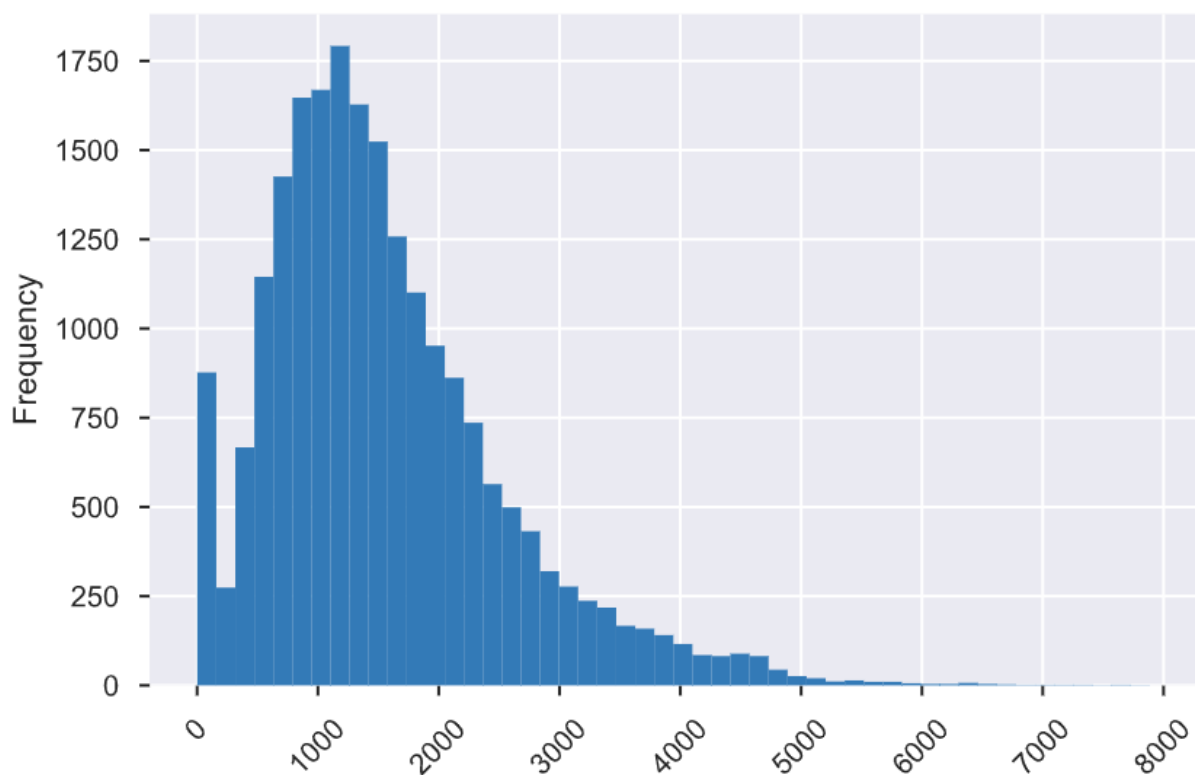


Figure 41 - Time from pickup to arrival at destination histogram

4.3.1.15 Rider Age

The 'riders_data.csv' contained information about the riders. Rider's age is the number of days since the rider made his/her first delivery on the Sendy platform. This variable is represented as 'Age' in the dataset.

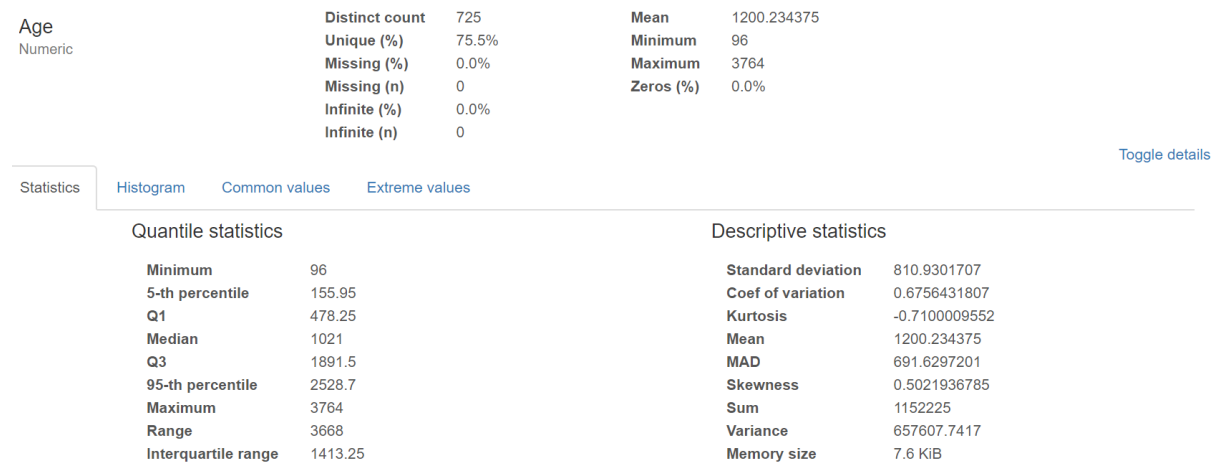


Figure 42 - Rider's age statistics

The histogram below shows the distribution of the riders by age.

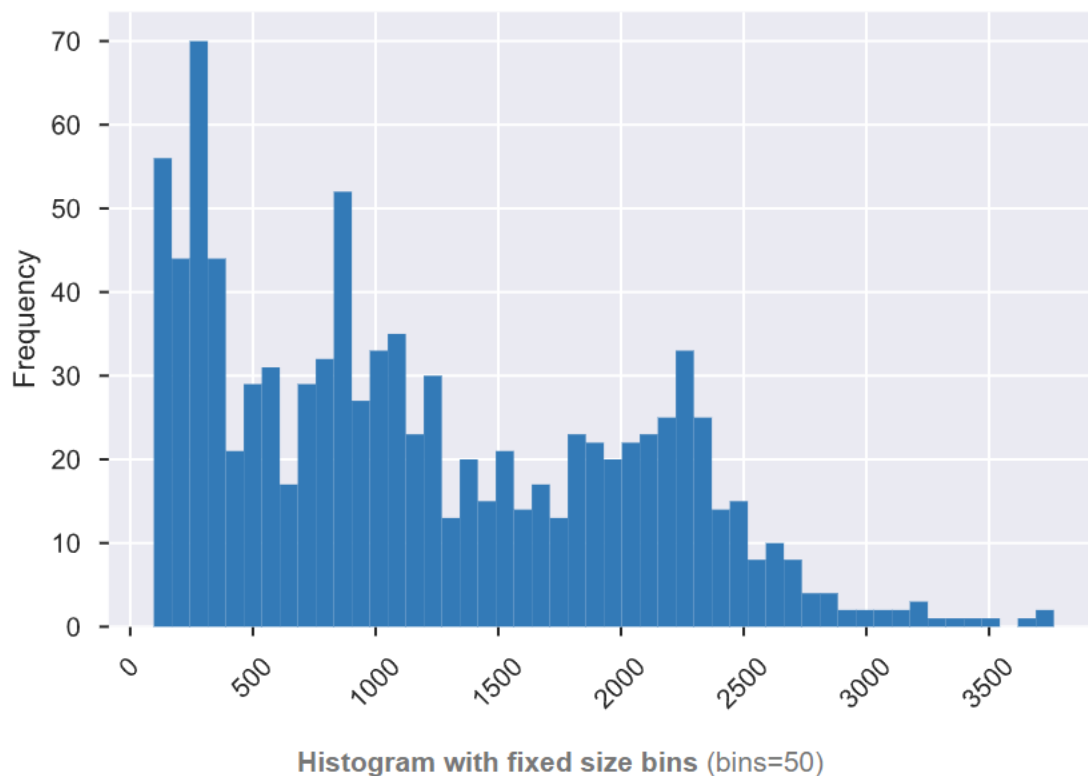


Figure 43 - Rider's age histogram

4.3.1.17 Rider Average Rating

This is the average of the ratings the rider received from previous client. I have to take note that customers have the choice of giving a rating or not. This variable is represented as 'Average Rating' in the dataset.

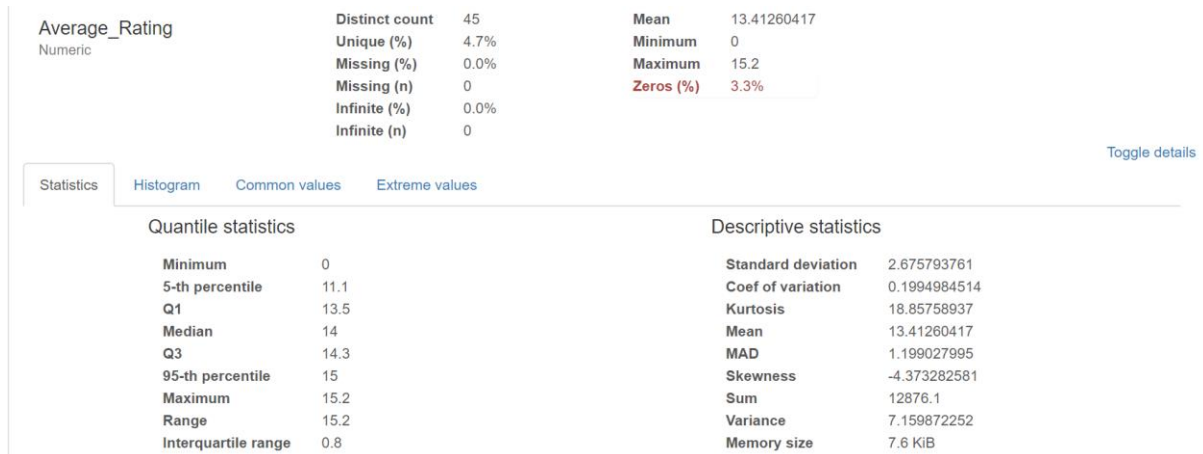


Figure 44 - Rider's average rating statistics

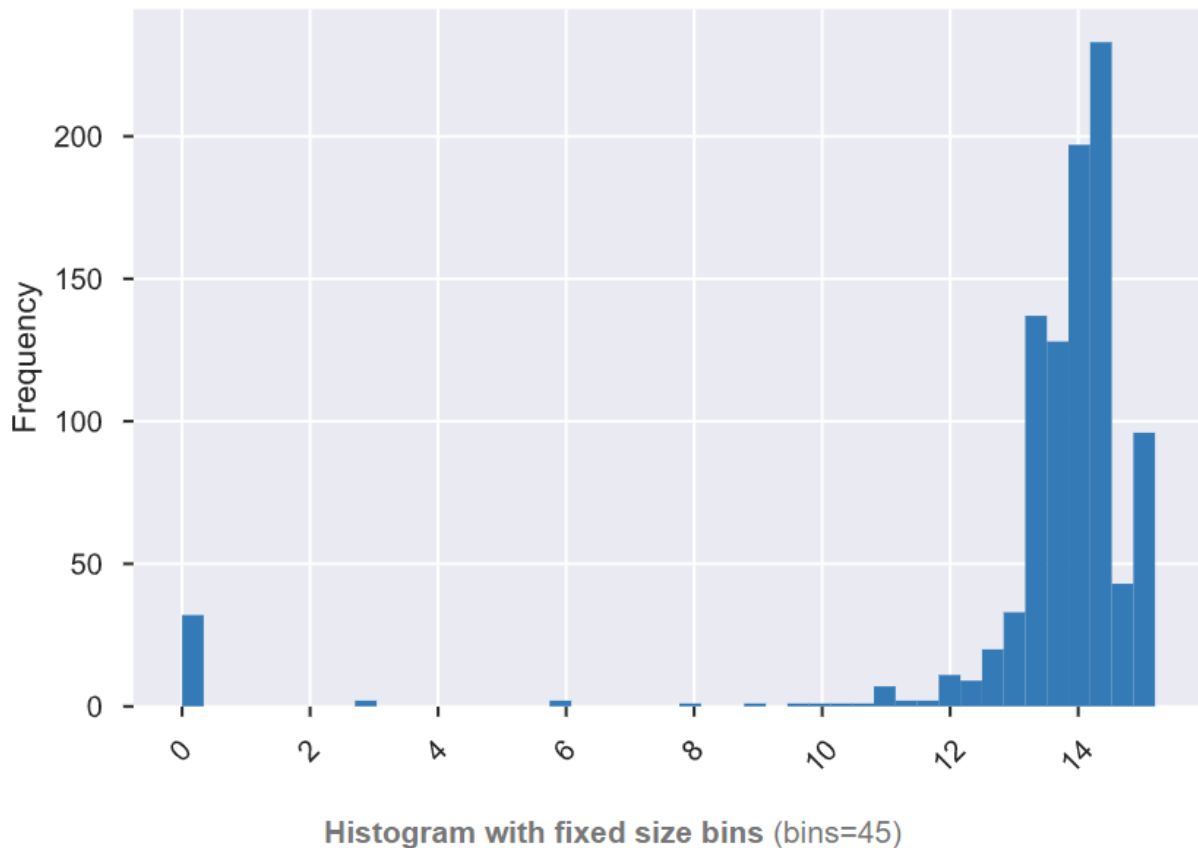


Figure 45 - Rider's average rating histogram

4.3.1.18 Rider's Number of Orders

This is the number of orders the rider has completed. This variable is represented as 'No of Orders' in the dataset.

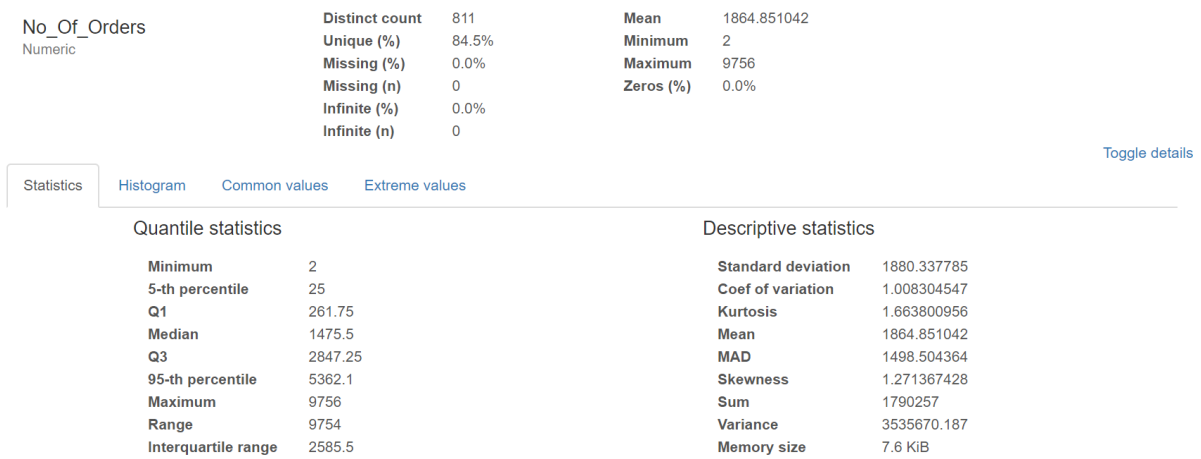


Figure 46 - Number of orders a rider has completed statistics

The distribution of this variable is shown below.

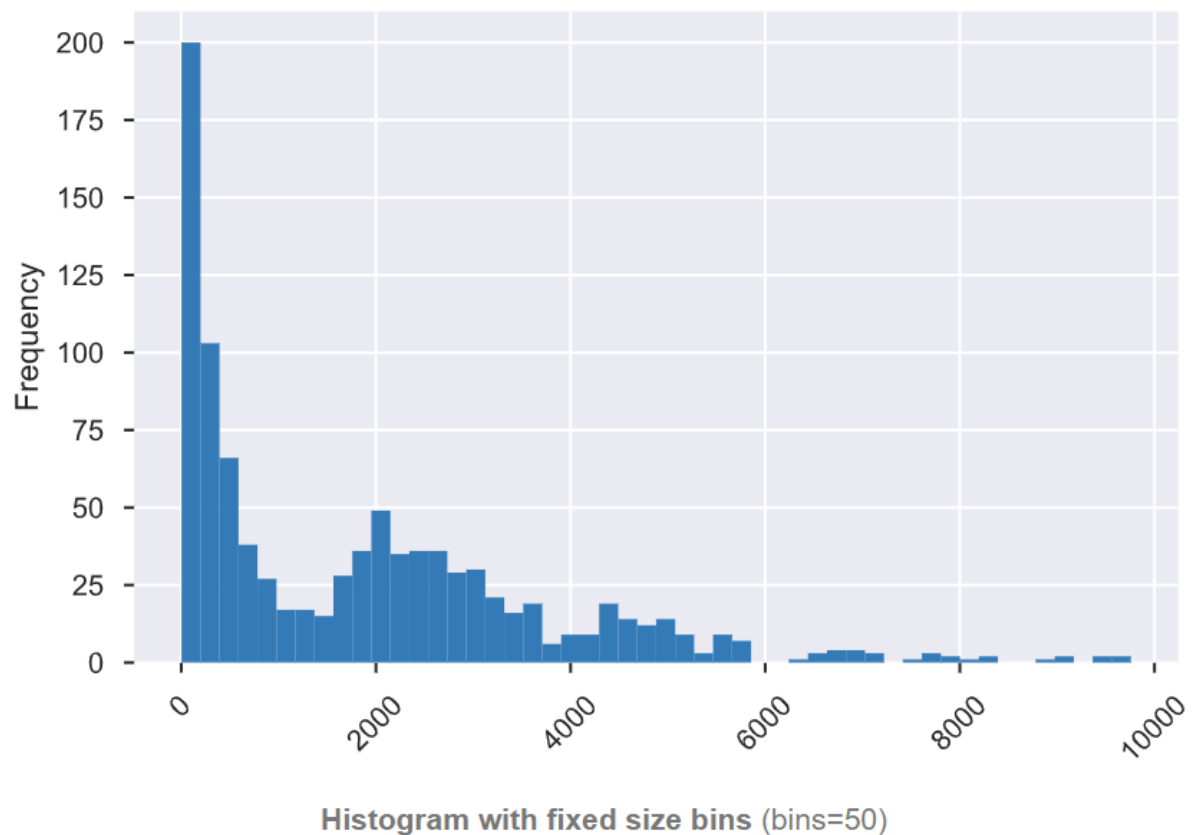


Figure 47 - Histogram of number of orders a rider has completed

4.3.1.19 Rider's number of rating

This variable shows the number of ratings the rider has received from the orders he has serviced. This variable is represented as 'No of Ratings'.

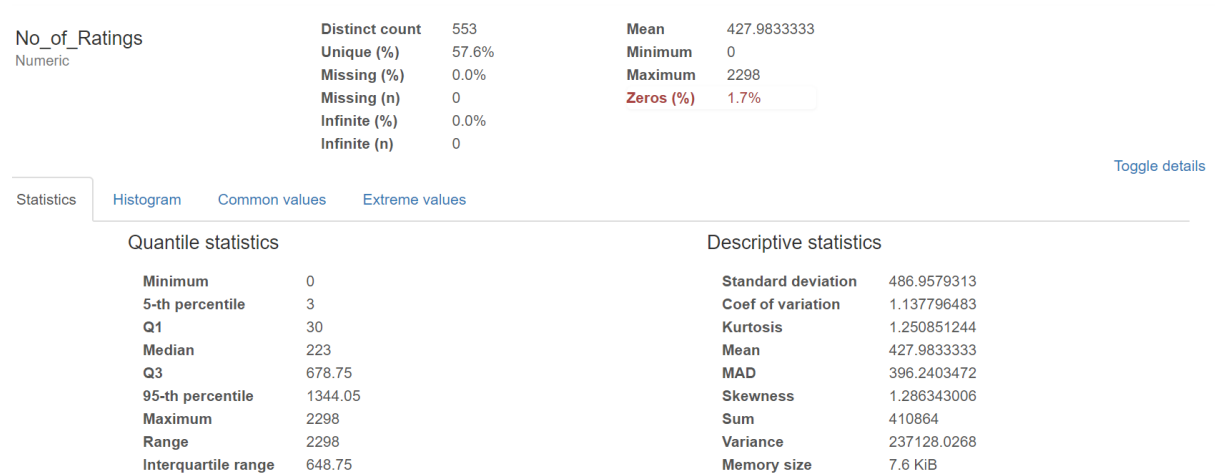


Figure 48 - Rider's number of ratings statistics

The distribution of the number of ratings per rider is shown below.

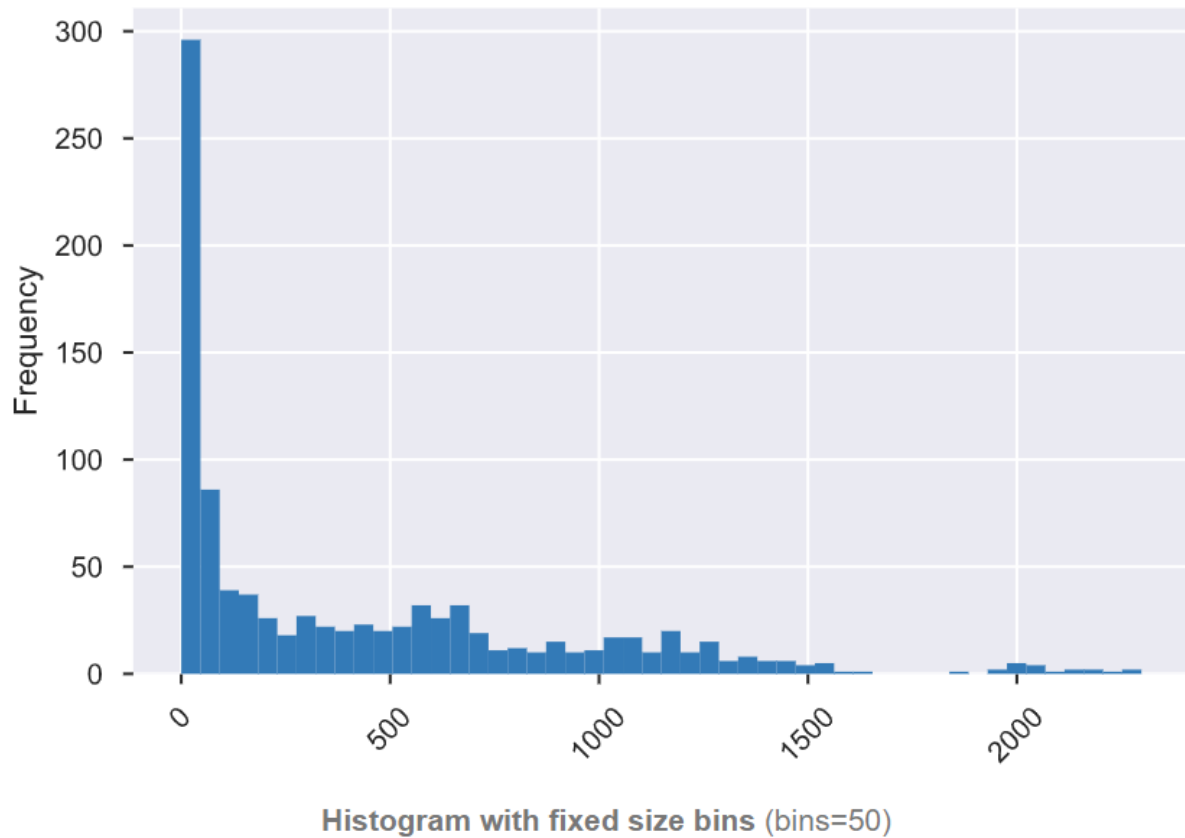


Figure 49 - Rider's number of ratings histograms

4.3.1.20 Other Variables

Some of the variables have a strong correlation (of 1) with some of the variables already discussed above, hence I didn't see the importance of showing them. Instead, I opted to list them here.

- Arrival at Pickup – Day of Month – This variable represents the day of week that the order was confirmed by the rider. This variable has a correlation of 1 with the 'Arrival at Destination – Day of Month)' variable which was discussed in 4.3.1.1.
- Arrival at Pickup – Weekday (Mo = 1) – This variable represents the day of week that the rider arrived at the picked up. This variable has a correlation of 1 with the 'Arrival at Destination – Weekday (Mo = 1)' variable discussed in section 4.3.1.3.
- Confirmation – Day of Month – This variable represents the day of week that the order was confirmed by the rider. This variable has a correlation of 1 with the 'Arrival at Destination – Day of Month)' variable which was discussed in 4.3.1.1.
- Confirmation – Weekday (Mo = 1) – This variable represents the day of week that the order was picked up. This variable has a correlation of 1 with the 'Arrival at Destination – Weekday (Mo = 1)' variable discussed in section 4.3.1.3.
- Pickup – Day of Month – This variable represents the day of week that the order was picked up. This variable has a correlation of 1 with the 'Arrival at Destination – Day of Month)' variable which was discussed in 4.3.1.1.
- Pickup – Weekday (Mo = 1) – This variable represents the day of week that the order was picked up. This variable has a correlation of 1 with the 'Arrival at Destination – Weekday (Mo = 1)' variable discussed in section 4.3.1.3.
- Placement – Day of Month – This is the day of the month the order was placed. This variable has a correlation of 1 with the 'Arrival at Destination – Day of Month)' variable which was discussed in 4.3.1.1.

- Placement – Weekday (Mo = 1) – This is the day of week the order was placed. This variable has a correlation of 1 with the ‘Arrival at Destination – Weekday (Mo = 1)’ variable discussed in section 4.3.1.3.
- Vehicle Type – This is the type of vehicle that serviced the order. This is a constant, as the data provided only includes motorcycle orders.
- Order No – This is the id of the order. This variable is unique in all observations in the orders data. It is represented as ‘Order No’.
- User ID – This is a unique identifier of the users who place an order in the data provided. This is represented as ‘User Id’ in the orders data.
- Rider ID – This is the rider id of the rider who serviced the order. This variable is seen in both the orders data set and the rider’s data set.

4.3.1.21 Correlations in the Data

The correlations in the data set were performed by using Pearson r coefficient of correlation. Pearson correlation coefficient is a measure of the strength of a linear association between two variables. The image below shows the results of the orders data. As described above, there are strong correlations in some of the variables.

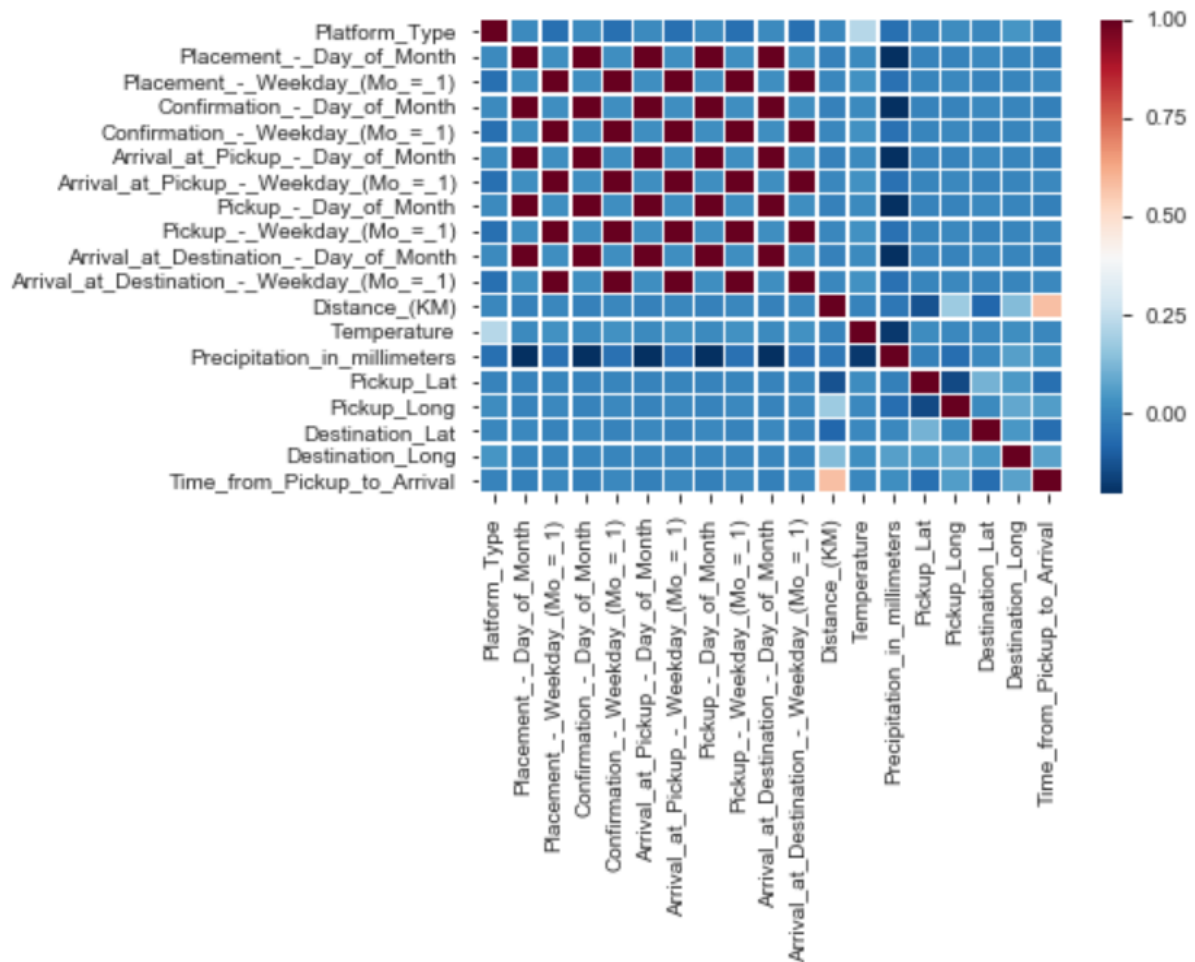


Figure 50 - Pearson r coefficient of correlation in orders data

Looking at the rider's data I also saw some correlation. The number of ratings, and the age both have a high correlation with the number of orders which makes perfect sense. Also, the age and the number of ratings and numbers of orders have a high correlation as shown in the image below. It is also important to note that the number of orders has a weak correlation (0.2) to the average rating.

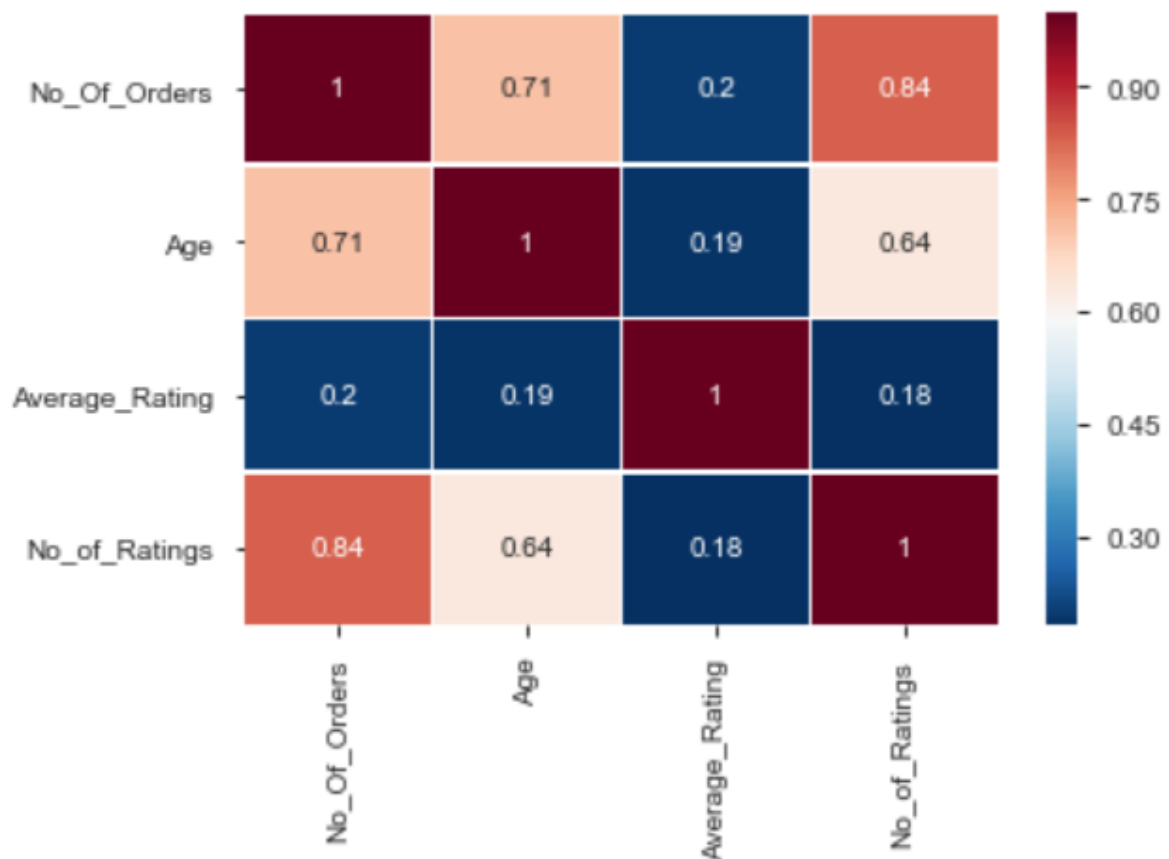


Figure 51 - Pearson r coefficient of correlation in rider's data

4.3.2 Building A Model

I used XGBoost to predict the time between pickup to arrival at destination. XGBoost is a highly optimized distributed gradient boosting library. XGBoost was built and designed to be highly flexible, efficient and portable. Using machine learning algorithms under the Gradient Boosting framework, XGBoost is in a position to provides a parallel tree boosting (which is also referred to as GBM) which solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. (xgboost.ai)

I then fit the model using the code below. In the code below, I've also shown the parameters I used after tuning the model.

```
best_xgb_model = xg.XGBRegressor(  
    bootstrap= False,  
    colsample_bytree= 0.8,  
    criterion= 'mse',  
    eta= 0.2,  
    learning_rate= 0.1,  
    max_depth= 2,  
    max_features= 15,  
    min_child_weight= 6,  
    min_samples_leaf= 3,  
    n_estimators= 700,  
    seed= 26,  
    subsample= 0.8)
```

4.3.3 Testing and Validating the Model

4.3.3.1 K-Fold Cross Validation

Validation of the model was done using the K-Fold cross validation technique. In the K-fold cross validation technique, $k-1$ folds are used for training. The remaining one ($k-1$) was used for testing and validating the model. This aspect is well depicted in the image below.

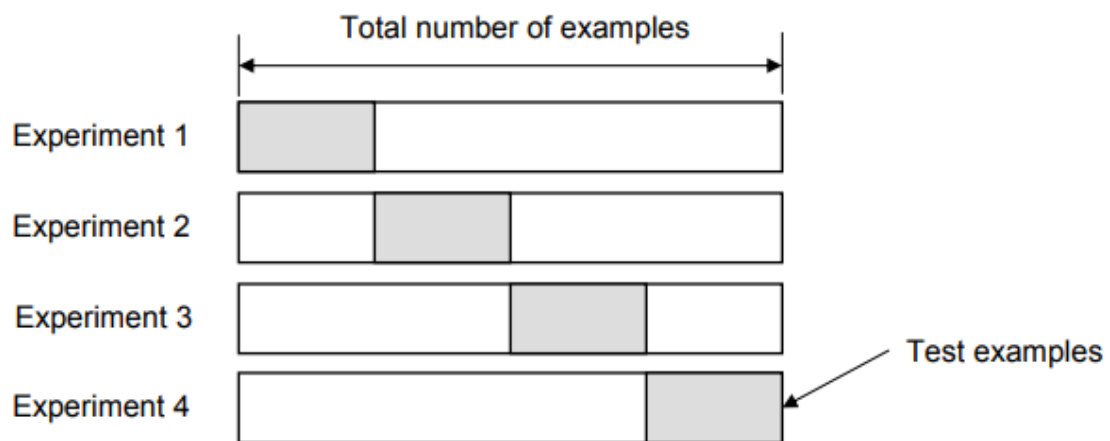


Figure 52 - K-fold cross-validation

The advantage is that entire data is utilized in training and testing. The error rate of the model is average of the error rate of each iteration. I tried different values of K and tested them for the accuracy. The results are shown below.

- 5 – 41.65%
- 10 – 41.98%
- 15 – 42.03%
- 20 – 42.21%
- 30 – 42.18%

With the decline increasing the folds to 30 I decided to use 20 folds

4.3.3.1 Hold Out Validation

I split the data into two to gain a better understanding the predictions. To achieve this, I used the holdout method using 25% of the data as the test set and 75% as the train set.

I then made predictions of the test set using the model described above. The accuracy score of this model is 0.0424 in fraction which translates to 4.24%. The accuracy score calculates the closeness of the predicted results to the actual result, either the fraction or the number of accurate predictions.

Taking a closer look at the predictions, I can calculate the time difference of the real time and the time the model forecasted. The difference is calculated by subtracting the prediction from the actual time. A negative value in the predictions means that the prediction gave a higher value than the actual. A positive value means that the prediction was less than the actual. A delivery that is late is worse than an early delivery. It would be interesting to look at the distribution of the late deliveries and early deliveries. From the predictions the model made, 34.6% of the predictions were late, while 65.4% of the predictions were early or exactly on time – meaning that if this was the estimated time of arrival that was communicated to the customer, the package would arrive earlier. The histogram below shows the distribution of the difference between the prediction and the actual time.

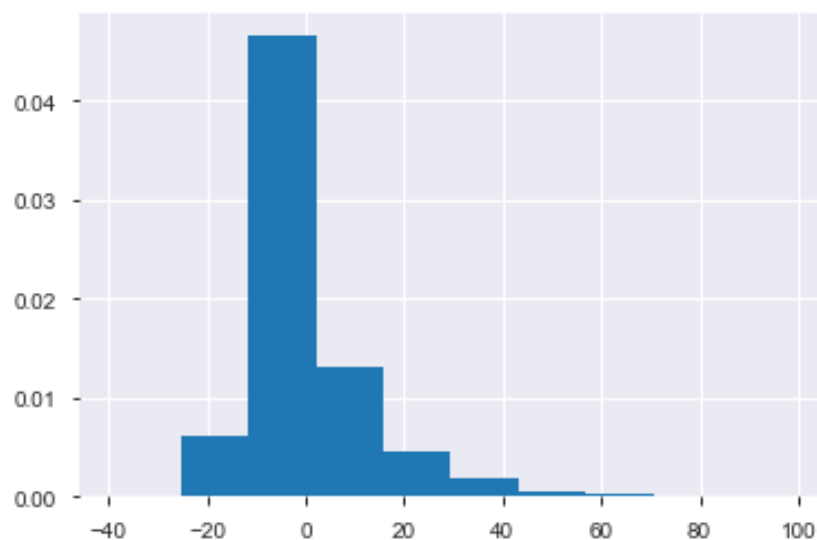


Figure 53 - XGBoost prediction histogram

Looking at the statistics of the difference this is what I get.

count	5004.000000
mean	0.192046
std	12.375258
min	-39.000000
25%	-7.000000
50%	-3.000000
75%	4.000000
max	98.000000

Figure 54 - XGBoost difference statistics

Looking at the absolute differences in the expected and predicted time, I get a much clearer picture. The mean difference is 8 minutes, which is not too much when it comes to package arrival. A histogram showing the difference gives us a better picture of how far the prediction was from the actual time that the delivery took to complete in the test set.

count	5004.000000
mean	8.44944
std	9.04304
min	0.000000
25%	3.000000
50%	6.000000
75%	10.000000
max	98.000000

Figure 55 - XGBoost absolute difference statistics

Looking at the absolute differences in the expected and predicted time, I get a much clearer picture. The mean difference is 8 minutes, which is not too much when it comes to package arrival. A histogram showing the absolute difference gives us a better picture of how far the prediction was from the actual.

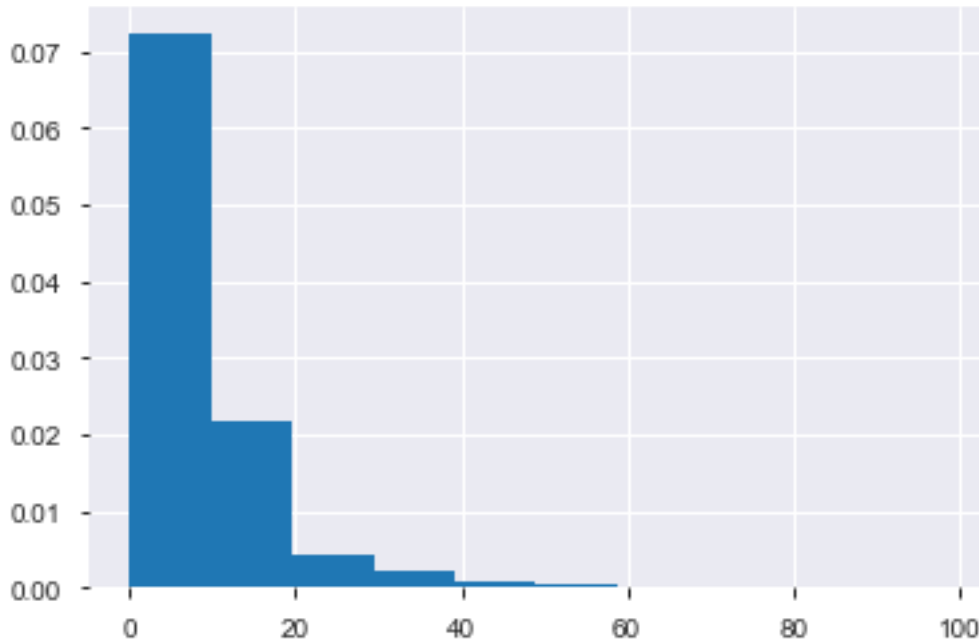


Figure 56 - XGBoost absolute difference distribution

4.3.4 Significant Variables

Feature importance refers to a class of methods for providing scores to the input features of a predictive model which in turn indicate the relative importance of each feature when making a prediction. Feature importance is paramount to gain a better understanding of the model. Having a look at the feature importance score provides insight into this XGBoost model. The feature importance diagram below shows us which features are the most important and least important to the model when making a prediction.

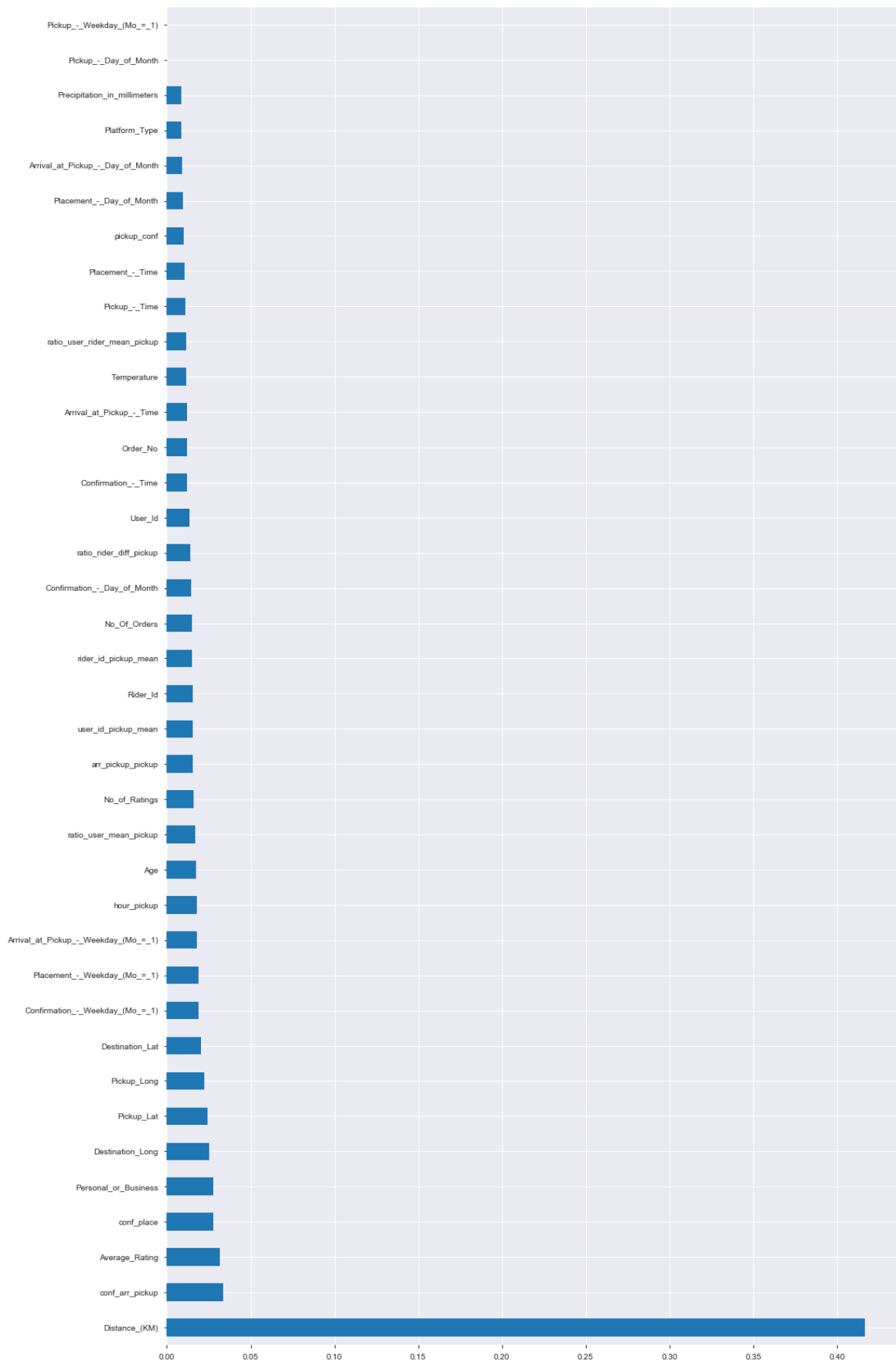


Figure 57 - XGBoost feature importance bar graph

The values that generate the graph above are shown in the image below.

Distance_(KM)	0.416513
conf_arr_pickup	0.033812
Average_Rating	0.031794
conf_place	0.027851
Personal_or_Business	0.027564
Destination_Long	0.025031
Pickup_Lat	0.024324
Pickup_Long	0.022544
Destination_Lat	0.020401
Confirmation_-_Weekday_(Mo=_1)	0.019035
Placement_-_Weekday_(Mo=_1)	0.018895
Arrival_at_Pickup_-_Weekday_(Mo=_1)	0.018097
hour_pickup	0.017889
Age	0.017456
ratio_user_mean_pickup	0.017049
No_of_Ratings	0.015723
arr_pickup_pickup	0.015622
user_id_pickup_mean	0.015545
Rider_Id	0.015387
rider_id_pickup_mean	0.015121
No_Of_Orders	0.014851
Confirmation_-_Day_of_Month	0.014316
ratio_rider_diff_pickup	0.013875
User_Id	0.013326
Confirmation_-_Time	0.012237
Order_No	0.011978
Arrival_at_Pickup_-_Time	0.011922
Temperature	0.011760
ratio_user_rider_mean_pickup	0.011514
Pickup_-_Time	0.011263
Placement_-_Time	0.010580
pickup_conf	0.010092
Placement_-_Day_of_Month	0.009802
Arrival_at_Pickup_-_Day_of_Month	0.009328
Platform_Type	0.008761
Precipitation_in_millimeters	0.008740
Pickup_-_Day_of_Month	0.000000
Pickup_-_Weekday_(Mo=_1)	0.000000
dtype: float32	

Figure 58 - XGBoost feature importance values

From the two images above, it seems that distance and the time between confirmation and arrival at pickup are the top two variables that the model used in calculating the prediction.

Plotting the first decision tree in the model, shows the below.

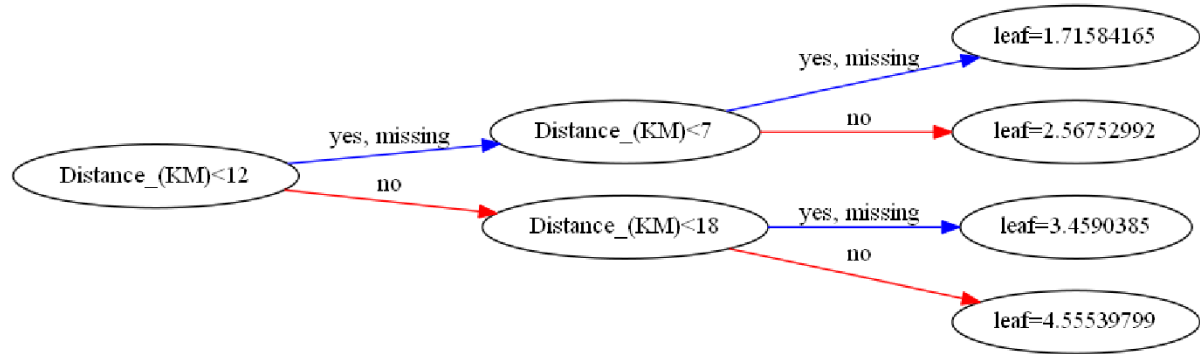


Figure 59 - First XGBoost tree

The image below shows the last decision tree of the model.

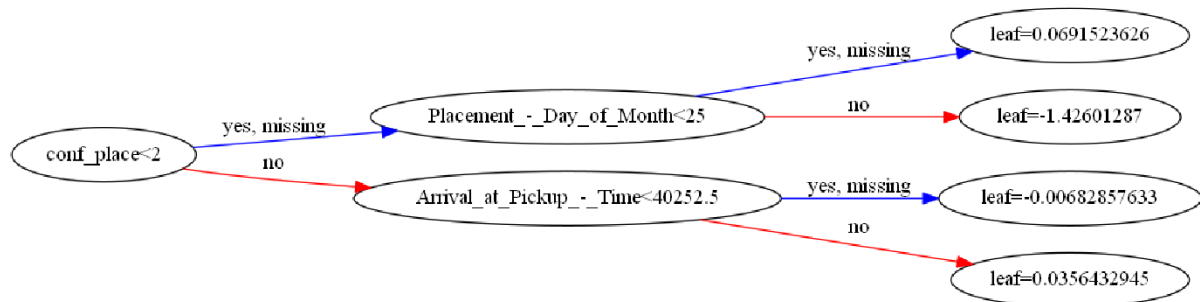


Figure 60 - Last XGBoost tree (700)

We can immediately spot the change between the starting point and at the ending point of the trees in the model. The first decision tree begins with distance while the second tree starts with the difference between placement and confirmation.

CHAPTER 5 DISCUSSION OF RESULTS

A machine learning model was built using XGBoost which aimed at predicting the estimated time of delivery of a motorcycle order in Nairobi. The model was tuned and the best parameters were used to obtain the predictions.

Looking at the significant variables that go into determining the ETD for motorcycle delivery in Kenya. The feature importance graph that was generated from the model in section 4.3.4 shows that the distance that a rider will have to cover from pickup point to the delivery location matters the most in predicting the ETD. This makes a lot of sense, an order that has a larger distance will definitely take a longer time than a short distance order. This is followed by the time between when the rider confirms the order and when he/she arrives at the pickup location. This might also make some sense if I assume that the time taken to arrive at the pickup location can tell something about the speed of the rider. A rider's average rating also came in showing that a riders rating greatly affects the estimated time of delivery.

Testing and validation of the model was done using 2 methods; the holdout method and k-Fold cross validation techniques. K-fold cross validation where k is 20 folds was used because this produced the best results as compared to 5, 15, 15 and 30. The accuracy of using 20 folds was 42.21%. The hold out method showed that the accuracy was very low, coming at only 4.24%. With the knowledge that the ETA for motorcycles isn't as rigid as what the validation techniques look at, it was important to look at the time difference between the predictions and the actual time of delivery. From the predictions the model made, 34.6% of the predictions were late, while 65.4% of the predictions were early or exactly on time – meaning that if this was the estimated time of arrival that was communicated to the customer, the package would arrive earlier.

CHAPTER 6 SUMMARY FINDINGS, CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter provides a look into the outcome and findings which were laid out in the previous chapter (Chapter 4). A review of the achievement of the objectives of the study has also been covered here. The parameters passed in the model can be found in section 4.3.2.

6.2 Current ETD Prediction at Sendy Ltd.

When conducting this study, Sendy Ltd.'s data and engineering team already had a model to predict the ETD of an order. Alerts had already been set up to monitor the duration of an order by the team. The model was very generalized and often gave a fixed number of 45 minutes regardless of the distance. The model didn't take into consideration other variables. Because of this, the orders could take as much as 45 minutes even for an order with a short distance without getting any action taken.

6.4 Summary of Findings

I conclude that the model developed in this study is highly effective in predicting an estimate of the time between when the rider picks up the order and when the rider arrived at the destination.

The research also looked at which variables are the most important, out of all variables the distance and the time between confirmation and arrival at pick up the prediction are the most important. On the other hand, however, the day of month and the day of week the order was picked up don't add any value to the model.

As per the objectives of the study a machine learning model based on XGBoost was developed and to determine an accurate estimated time of arrival.

Two validation techniques were used to determine the model's effectiveness. K-fold cross validation showed on average 42.01% accuracy which was all most 10 times higher compared

to the holdout method which averaged 4.24%. However, looking at the two accuracies of the validation techniques as compared to the predictions made, K-fold cross validation gave a number closer to the actual early predictions.

6.5 Conclusion

Looking at the outcome of the results reported in the fourth Chapter, it can be concluded that predicting the time it takes for an order to be delivered in Nairobi has been successful as enough variables have been analyzed. From this study it is noted that I can leverage more variables to get a better prediction for the order. This study has also highlighted the most important variables that go into predicting an accurate estimated time of arrival of motorcycle deliveries in Nairobi. This study has also developed a model that is easily replicable by the team at Sendy and anyone else who is interested in predicting the estimated delivery time of motorcycle deliveries in Nairobi. I believe the results of these findings greatly help in predicting the estimated time of delivery for any other city or area and other vehicle types.

6.6 Contributions

This study can easily contribute in the development of models that predict package delivery time across different vehicle sizes. Such models would be very useful for logistics industries across the world. This study will also help logistics companies in any other city or location to provide better ETAs to the customers. The key to this is optimizing the model for the specific problem.

This study constructed an advanced assessment model in predicting package delivery time and the model resulted in variables which have helped to identify the important values that determine an accurate estimated time of arrival. These values can contribute to how logistics companies approach predicting an accurate ETD.

Predicting an accurate estimate time of arrival for customer may support logistics companies in improving better customer satisfaction through setting the right expectations. The study was

optimized for motorcycle deliveries in Nairobi. Because of this, the important variables and parameters of the model might change. Customizing the variables and parameters of the model need to be well tailored to suit the needs of the logistic company using the model.

This study will also benefit customers of logistics applications in Nairobi who are keen to understand what goes into predicting an accurate estimated time of arrival. This study will also benefit customers of logistics applications in other countries and cities across the world.

This study has also added value to the existing research that has been done. Scholars and researchers can infer to this study as a resource in their research and knowledge in the area of predicting package delivery time.

6.7 Recommendations

After this study, the author would recommend that Sendy Ltd. adopt this model in predicting the delivery time of packages in Nairobi. This will definitely improve customer expectations in delivery time during an order. Getting an accurate delivery time prediction will also improve customer satisfaction through setting the right expectations. Extending the scope of this study to include other cities or town will also help Sendy Ltd as it expands. Furthermore, including other vehicle types might have a beneficial effect on Sendy Ltd.'s estimated time of arrival.

REFERENCES

- Abhishek D. (2018, February 08). Predicting Arrival time (ETA): Improving Customer's estimation. Retrieved June 12, 2019, from <https://jungleworks.com/predicting-accurate-arrival-time/>
- Rajdev, R. R. (2018, January 03). Google Maps & Its Estimated Time of Arrival (ETA) - Must Know. Retrieved August 22, 2020, from <https://www.techcoffees.com/google-maps-calculate-estimated-time/>
- Dahlman C. "Technology, globalization and international competitiveness: challenges for developing countries," in DESA. ed. Industrial development for the 21st century: sustainable development perspectives. New York: United Nations; 2007. pp. 29–83.
- Oberoi, A. (2018, May 30). On-Demand Services: Predicting Arrival Time (ETA) for Customers. Retrieved June 16, 2019, from <https://insights.daffodilsw.com/blog/on-demand-services-predicting-arrival-time-eta-for-customers>
- Dushaj, M. (2018, September 18). Estimated Time of Arrival vs. Estimated Time of Delivery Explained. Retrieved June 16, 2019, from <https://www.morethanshipping.com/estimated-time-of-arrival-vs-estimated-delivery-explained/>
- Ijaz, Muhammad & Rhee, Jongtae. (2018). Constituents and Consequences of Online-Shopping in Sustainable E-Business: An Experimental Study of Online-Shopping Malls. Sustainability.
- Md. Noor, Rafidah & Yik, Ng & Kolandaisamy, Raenu & Ahmedy, Ismail & Hossain, Mohammad Asif & Yau, Kok-Lim & Md Shah, Wahidah & Nandy, Tarak. (2020). Predict Arrival Time by Using Machine Learning Algorithm to Promote Utilization of Urban Smart Bus.

Wang, Zhengyi & LIANG, Man & Delahaye, Daniel. (2020). Automated data-driven prediction on aircraft Estimated Time of Arrival. *Journal of Air Transport Management*. 88. 10.1016/j.jairtraman.2020.101840.

Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and Samir Chatterjee. 2007. *A Design Science Research Methodology for Information Systems Research*.

Grosman, L. (2018, February 22). What the Amazon Effect Means for Retailers. Retrieved June 17, 2019, from <https://www.forbes.com/sites/forbescommunicationscouncil/2018/02/22/what-the-amazon-effect-means-for-retailers/#232e24042ded>

Van der Spoel, Sjoerd & Amrit, Chintan & Hillegersberg, Jos. (2016). Predictive Analytics for Truck Arrival Time Estimation: A Field Study at a European Distribution Center. *International Journal of Production Research*. In Press. 1-21. 10.1080/00207543.2015.1064183.

Daganzo, C. (2004). 29 E-Commerce and End Delivery Issues. Retrieved June 17, 2019, from <https://www.emeraldinsight.com/doi/abs/10.1108/9780080473222-029>

Kalakota, R., & Whinston, A. B. (1997). *Electronic Commerce: A Manager's Guide*. Addison-Wesley Professional.

Farooq, Q., Fu, P., Hao, Y., Jonathan, T., & Zhang, Y. (2019). A Review of Management and Importance of E-Commerce Implementation in Service Delivery of Private Express Enterprises of China. *SAGE Open*. <https://doi.org/10.1177/2158244018824194>

Paul Zarchan; Howard Musoff (2000). *Fundamentals of Kalman Filtering: A Practical Approach*. American Institute of Aeronautics and Astronautics, Incorporated. ISBN 978-1-56347-455-2.

Graupe, Daniel (2013). *Principles of Artificial Neural Networks*. World Scientific. pp.

- Yang, J.-S., 2005. Travel time prediction using the GPS test vehicle and Kalman filtering techniques. In: Proceedings of the 2005 American Control Conference, pp. 2128–2133.
- Elhenawy, M., Chen, H., & Rakha, H. A. (n.d.). *Transportation Research Part C: Emerging Technologies Volume 42 (May 2014)*.
- Wu, Chun-Hsin & Ho, Jan-Ming & Lee, D. (2005). Travel-Time Prediction with Support Vector Regression. *Intelligent Transportation Systems, IEEE Transactions on*. 5. 276 - 281. 10.1109/TITS.2004.837813.
- T. Liu, J. Ma, W. Guan, Y. Song and H. Niu, "Bus Arrival Time Prediction Based on the k-Nearest Neighbor Method," 2012 Fifth International Joint Conference on Computational Sciences and Optimization, Harbin, 2012, pp. 480-483.
- About XGBoost. (n.d.). Retrieved June 18, 2020, from <https://xgboost.ai/about>

APPENDICES

Appendix I: Budget

Budget Items	Cost (Ksh.)
Proposal development- Printing, stationery, internet costs	12,000
Data Collection - Internet Charges	5,000
Data Analysis and report - Printing, Stationery & Hardcover Binding	10,000
Transport Fuel to Campus, airtime	10,000
TOTAL	37,000

Figure 61: Budget

Appendix II: Research Schedule

Activity	Jan-July 2019	Aug-19	Sep-19	Oct-19	Nov 2019 - Feb 2020	Mar-Apr 2020	May-20	Jun-20	Jul-20	Aug-20
Proposal Draft and Writing										
Proposal Submission										
Proposal Presentation 1										
Addressing feedback from panel										
Data Preparation and analysis										
Model Building										
Report writing										
Addressing supervisor Feedback										
Project defense										
Addressing feedback from panel										
Final Report Submission										

Figure 62: Research Schedule

Appendix III: Sample Source Code for Regression Analysis

```
import pandas as pd
import pandas_profiling as pf
import seaborn as sb
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from catboost import CatBoostRegressor, Pool
from sklearn.metrics import accuracy_score
from sklearn import metrics
import numpy as np
from paramsearch import paramsearch
from itertools import product, chain
import xgboost as xg
from xgboost import plot_tree
from scipy.stats import randint
from sklearn.preprocessing import LabelEncoder
from sklearn.pipeline import Pipeline
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import RandomizedSearchCV
from sklearn import tree
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold

# Read & Load Data into Dataframes

orders_data = pd.read_csv('orders_data.csv')
riders_data = pd.read_csv('riders_data.csv')
data = pd.merge(orders_data, riders_data, how = 'left', on =
['Rider Id'])

# Change Data Types

orders_data['Placement - Time'] =
pd.to_datetime(orders_data['Placement - Time'])
orders_data['Confirmation - Time'] =
pd.to_datetime(orders_data['Confirmation - Time'])
orders_data['Arrival at Pickup - Time'] =
pd.to_datetime(orders_data['Arrival at Pickup - Time'])
orders_data['Pickup - Time'] =
pd.to_datetime(orders_data['Pickup - Time'])
orders_data['Arrival at Destination - Time'] =
pd.to_datetime(orders_data['Arrival at Destination - Time'])

# Orders Data Profiling

orders_data_profile = pf.ProfileReport(orders_data)
```

```

orders_data_profile.to_file('orders_data_profile.html')

# Orders Correlation
order_data_pearson_corr = orders_data.corr(method='pearson')
sb.heatmap(order_data_pearson_corr,
            xticklabels=order_data_pearson_corr.columns,
            yticklabels=order_data_pearson_corr.columns,
            cmap='RdBu_r',

            linewidth=0.5)

# Riders Data Profiling

riders_data_profile = pf.ProfileReport(riders_data)
riders_data_profile.to_file('riders_data_profile.html')

# Riders Correlation

riders_data_pearson_corr = riders_data.corr(method='pearson')
sb.heatmap(riders_data_pearson_corr,
            xticklabels=riders_data_pearson_corr.columns,
            yticklabels=riders_data_pearson_corr.columns,
            cmap='RdBu_r',
            annot=True,
            linewidth=0.5)

# Transformations

data['pickup_time'] = pd.to_datetime(data['Pickup - Time'])
data['arr_pickup_time'] = pd.to_datetime(data['Arrival at
Pickup - Time'])
data['confirmation_time'] = pd.to_datetime(data['Confirmation -
Time'])
data['placement_time'] = pd.to_datetime(data['Placement -
Time'])
data['dest_time'] = pd.to_datetime(data['Arrival at Destination
- Time'])

# Calculating New Variables

data['conf_place'] = (data['confirmation_time'] -
data['placement_time']).dt.seconds
data['conf_arr_pickup'] = (data['arr_pickup_time'] -
data['confirmation_time']).dt.seconds
data['arr_pickup_pickup'] = (data['pickup_time'] -
data['arr_pickup_time']).dt.seconds
data['pickup_conf'] = (data['pickup_time'] -
data['confirmation_time']).dt.seconds

```

```

data['user_id_pickup_mean'] = data['User
Id'].map(data.groupby('User
Id')['arr_pickup_pickup'].mean().to_dict())
data['rider_id_pickup_mean'] = data['Rider
Id'].map(data.groupby('Rider
Id')['arr_pickup_pickup'].mean().to_dict())

data['ratio_user_rider_mean_pickup'] =
data['user_id_pickup_mean'] / data['rider_id_pickup_mean']
data['ratio_user_mean_pickup'] = data['user_id_pickup_mean'] /
data['arr_pickup_pickup']
data['ratio_rider_diff_pickup'] = data['rider_id_pickup_mean']
/ data['arr_pickup_pickup']
data['hour_pickup'] = data['pickup_time'].dt.hour

data['Confirmation - Time'] = pd.to_timedelta(
data['Confirmation - Time'])
data['Confirmation - Time'] = data['Confirmation -
Time'].dt.total_seconds()

data['Pickup - Time'] = pd.to_timedelta( data['Pickup - Time'])
data['Pickup - Time'] = data['Pickup - Time'].dt.total_seconds()

data['Placement - Time'] = pd.to_timedelta( data['Placement -
Time'])
data['Placement - Time'] = data['Placement -
Time'].dt.total_seconds()

data['Arrival at Pickup - Time'] = pd.to_timedelta( data['Arrival
at Pickup - Time'])
data['Arrival at Pickup - Time'] = data['Arrival at Pickup -
Time'].dt.total_seconds()

data['Arrival at Destination - Time'] = pd.to_timedelta(
data['Arrival at Destination - Time'])
data['Arrival at Destination - Time'] = data['Arrival at
Destination - Time'].dt.total_seconds()

# Change time from pickup to arrival to minutes
data['Time from Pickup to Arrival'] = round(data['Time from
Pickup to Arrival']/60,0)

data.head(5)

# Data Preparation & Catering for Missing Variables

```

```
# Precipitation column is missing 97% of the values - knowing
Nairobi weather well, precipitation is seasonal. This data might
be during the dry season.
```

```
data['Precipitation in millimeters'].fillna(0, inplace=True)
```

```
# Use the average for the Temperature
```

```
data['Temperature'].fillna((data['Temperature'].mean()),
inplace=True)
```

```
# Only 20.5% of the observations are missing the temperature
values.
```

```
# therefore, I have decided to fill missing temperature values
with the mean
```

```
data['Temperature'].fillna((data['Temperature'].mean()),
inplace=True)
```

```
# Vehicle type is also the same through out, I might as well
drop it
```

```
data.drop('Vehicle Type', axis=1, inplace=True)
```

```
# The time from pickup to arrival has huge outliers.
```

```
# Some orders seem to have taken less than 5 minutes to complete.
```

```
# This is impossible, hence, I have decided to remove those
observations
```

```
sb.boxplot(x=data["Time from Pickup to Arrival"], orient="h")
```

```
under_5 = len(data[data['Time from Pickup to Arrival']<5]) #
Number of observations under 5 minutes
```

```
perc_under_5 = under_5 / len(data)
perc_under_5
```

```
data = data[data['Time from Pickup to Arrival']>5] # Filter out
orders completed under 5 minutes
```

```
# Split Train and Test
```

```
Target = ['Arrival at Destination - Day of Month',
          'Arrival at Destination - Weekday (Mo = 1)',
          'Arrival at Destination - Time'
          ]
```



```

data = data.drop(['Arrival at Destination - Day of Month',
                  'Arrival at Destination - Weekday (Mo = 1)',
                  'Arrival at Destination - Time',
                  'dest_time'], axis=1)

data.head(4)

# Encoding

class MultiColumnLabelEncoder:
    def __init__(self, columns = None):
        self.columns = columns # array of column names to encode

    def fit(self, X, y=None):
        return self # not relevant here

    def transform(self, X):
        '''
        Transforms columns of X specified in self.columns using
        LabelEncoder(). If no columns specified, transforms all
        columns in X.
        '''
        output = X.copy()
        if self.columns is not None:
            for col in self.columns:
                output[col] =
LabelEncoder().fit_transform(output[col])
        else:
            for colname, col in output.iteritems():
                output[colname] =
LabelEncoder().fit_transform(col)
            return output

    def fit_transform(self, X, y=None):
        return self.fit(X, y).transform(X)

data = MultiColumnLabelEncoder(columns = ['Personal or
Business',
                                           'User Id',
                                           'Rider Id',
                                           'Order No'

]).fit_transform(data)

data.head(2)

```

```

data.columns

#Split the labeled data frame into two sets to train then test
the models
y = data['Time from Pickup to Arrival']
X = data.drop(['Time from Pickup to Arrival'], axis=1)

# XGBOOST

data.columns = data.columns.str.replace(' ', '_')

#Split the labeled data frame into two sets to train then test
the models

y = data['Time_from_Pickup_to_Arrival']
X = data.drop(['Time_from_Pickup_to_Arrival'], axis=1)
X = X._get_numeric_data()

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=42)

X_train.shape, y_train.shape, X_test.shape, y_test.shape

# Fitting to Model
best_xgb_model = xg.XGBRegressor(silent=True,
                                bootstrap= False,
                                colsample_bytree= 0.8,
                                criterion= 'mse',
                                eta= 0.2,
                                gamma = 1,
                                learning_rate= 0.1,
                                max_depth= 2,
                                max_features= 15,
                                min_child_weight= 6,
                                min_samples_leaf= 3,
                                n_estimators= 700,
                                seed= 26,
                                subsample= 0.8)

# Kfold Cross Validation

kfold = StratifiedKFold(n_splits=20, random_state=7)
results = cross_val_score(best_xgb_model, X, y, cv=kfold)
print("Accuracy:   %.2f%%   (%.2f%%)" % (results.mean()*100,
results.std()*100))
y_pred = best_xgb_model.predict(X_test)

```

```

predictions = [round(value) for value in y_pred]

# Hold Out Validation Technique

mod=best_xgb_model.fit(X_train, y_train)
# make predictions for test data
y_pred = best_xgb_model.predict(X_test)
predictions = [round(value) for value in y_pred]
# evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))

# Feature Importance
feature_importances =
pd.Series(best_xgb_model.feature_importances_,
index=X_train.columns)
feature_importances = feature_importances.sort_values()
plt.subplots(figsize=(16,30))
index = X_train.columns
plt.barh(index, feature_importances)
plt.show()

feature_importances

# XGBoost Tree
xg.plot_tree(mod,num_trees=699, rankdir='LR')
plt.rcParams['figure.figsize'] = [500, 100]
plt.show()

# Interpreting Results with Actual
test_set = X_test.join(y_test, how='outer')
test_set.head()
test_set['predictions'] = predictions

test_set['predictions_diff'] =
test_set.Time_from_Pickup_to_Arrival - test_set.predictions
test_set['predictions_diff'].describe()

late =
test_set[test_set['predictions_diff']>0].count()["Order_No"]
early =
test_set[test_set['predictions_diff']<=0].count()["Order_No"]

per_late = late/len(test_set)
per_early = early/len(test_set)

```

```
"late", per_late*100, "early", per_early*100
```

```
get_ipython().run_line_magic('matplotlib', 'inline')
plt.hist(test_set['predictions_diff'], density=True)
plt.show()
```

```
test_set['predictions_diff+']                                =
abs(test_set['predictions_diff'])
test_set['predictions_diff+'].describe()
```

```
get_ipython().run_line_magic('matplotlib', 'inline')
plt.hist(test_set['predictions_diff+'], density=True)
plt.show()
```

```
over_2                                                       =
test_set[test_set['predictions_diff+']>2].count()["order_no"]
under_2                                                       =
test_set[test_set['predictions_diff+']<=2].count()["order_no"]
```