# Predicting Estimated Time of Arrival with Machine Learning

June 29, 2021


| Registration No | Names of Students |
|---|---|
| I07/44029/2017 | Waweru Kennedy |
| I07/43350/2017 | Kenneth Mwaura |
| I07/1416/2017 | Levis Akal |
| I07/1398/2017 | Meshack Ateya |
| I07/38345/2016 | Brian Mayuba |

School of Mathematics
University of Nairobi

*Submitted in partial fulfillment of the requirements for the award of the degree of*

**Bachelor of Science in Actuarial Science**

## Acknowledgements

We would like to express our profound and heartfelt appreciations to Dr. Timothy Kamanu and Dr. Davis Bundi, our lecturers and research supervisors for their guidance and patient they have shown us from the beginning to completion of this research by suggestions and guidelines provided throughout the planning and doing this project. We also thank our families for their utmost support and encouragement throughout this course. And above all, we give much gratitude to the Almighty God for His Grace and Mercy in seeing to it that we reach this far.

# Declaration

This project is our original work and it has not been presented elsewhere for a degree award.

| Name | Registration No | Signature | Date |
|---|---|---|---|
| Waweru Kennedy | I07/44029/2017 | _____ | _____ |
| Kenneth Mwaura | I07/43350/2017 | _____ | _____ |
| Levis Akal | I07/1416/2017 | _____ | _____ |
| Meshack Ateya | I07/1398/2017 | _____ | _____ |
| Brian Mayuba | I07/38345/2016 | _____ | _____ |

**Declaration by Supervisors**

This project has been submitted for examination with our approval as supervisors

| Supervisor | Faculty | Signature | Date |
|---|---|---|---|
| Dr Timothy Kamanu | School of Mathematics | _____ | _____ |
| Dr Davis Bundi | School of Mathematics | _____ | _____ |

# Contents

# Abstract

With the unprecedented growth in e-commerce in Kenya, transit time reliability has become a critical point in the logistics business as irregularities will lead to delays in delivery of products to customers. Our sponsoring company, Sendy links customers who have delivery needs with vetted transporters (from bikes to trucks), using a web and mobile application platform as well as an API. Customers select their vehicle of choice, get their price quote upfront and pay using various payment options. The system optimises the route and dispatches the order to the closest available drivers and riders (called Partners). However, Sendy is facing commercial pressure from its customers for a better estimation of its time to delivery reliability, which has become a key measurement of its operational performance. The goal of our project was to determine whether Machine Learning and predictive analytics can improve the estimated time of arrival for a shipment. Using Machine Learning computing, we developed a model capable of predicting the estimated time of arrival by training the algorithms on historical shipment data, and incorporating external sources of data related to the most impactful factors regarding schedule reliability (e.g. weather and rider ratings). We found that Machine Learning in this instance might be a partial answer to this problem. In addition it was found that utilizing appropriate features as inputs to the prediction models dramatically increased the performance of the algorithms.



**Keywords:** Estimated Time of Arrival E.T.A; Logistics; Sendy; Machine Learning Algorithms; Random Forest ; Regression; Support Vector Regression;

# 1 Introduction

## 1.1 Background of the study

Sendy is a Kenyan company. The company was founded in 2014 by Meshack Alloys together with three other co-founders – Malaika Judd, Evanson Biwott and Don Okoth. Sendy is a platform that connects customers with delivery needs with drivers. It does not own any delivery vehicle but provides technology that makes delivery of goods simpler, cheaper and transparent. Sendy operates in Kenya but there are plans to enter the Ugandan and Tanzanian markets. At Sendy, we believe it is more than delivering goods. We believe that delivery unlocks potential, powers business growth and opens up a world of new possibilities. Sendy is focusing on partnering with SMEs, corporates, and manufacturers to unlock new possibilities by powering their business growth [The Standard, 2018]. Sendy Ltd is a crowd sourced courier marketplace tackling last mile, on demand, and hyperlocal deliveries in Kenya. Sendy provides an app and web platform that enables individual and businesses to connect with Riders and Drivers and request on demand or scheduled courier services at any time, any day. To support their mission, Sendy crowdsources vehicles and their drivers (boda bodas, vans and pickup trucks, 3-ton trucks), extensively vet and trainer vehicles and then connect them to their GPS enabled tracking and dispatching system. The company uses an asset-free model, with an app that coordinates contract drivers who own their own vehicles, while confirming deliveries, creating performance metrics and managing payment. The business model helps to reduce inefficiencies in the market by removing friction costs from matching customers with providers, while in parallel improving the service offering by training drivers and riders. As a result, the company supports the development of a well-functioning logistics sector, which is essential to provide market access and drive economic development. Over the years, Sendy has managed to work with over 5,000 businesses and 50,000 individual customers because it is cost effective and efficient. The pricing is given upfront based on

the distance covered. The platform also enables users to track their deliveries in real time and all the packages are insured while in transit. By doing this, Sendy gives the users peace of mind as they conduct their businesses. Sendy's platform allows businesses to outsource their logistics to quality and affordable service. Looking at a company's operational expenditure, logistics cost is one of the largest components. For most firms, building an in-house fleet to deliver goods is capital intensive, time-consuming and wasteful. They will have either too many or too few vehicles on standby, limited means to track journeys and calculate fair prices. Outsourcing to informal transporters can be risky especially on high-value goods. The pricing is also not optimal because they factor in their inefficiencies such as idle time. By contrast, Sendy offers an easy-to-use service, lowers cost and increases transparency on the deliveries and the driver. It also provides peace of mind by connecting businesses with insurance providers to secure their goods while in transit. Companies using Sendy have managed to reduce up to 30 per cent of their logistics costs. [The Standard, 2018]

## 1.2 Data

The study uses the data taken by Sendy's delivery vehicles and motorcycles from the time an order is placed up to the time the order is delivered. The dataset provided by Sendy includes order details and rider metrics based on orders made on the Sendy platform [SendyIT, 2021]. The challenge is to predict the estimated time of arrival (E.T.A) for orders from pick-up to drop-off.

## 1.3 Problem Statement

The goal of setting up a logistics company such as Sendy is so that the investors can make profit. The investors look at the predicted delivery times and are able to estimate amount of profit. However, events such as fuel cost, driver shortage and retention, unfavorable government regulations and technology strategy and implementation make it difficult to predict delivery times. A good model will go a long way in ensuring better predictions of delivery times to help investors make sound financial decisions.

## 1.4 Research Questions

The purpose of this thesis is to propose and evaluate the ability of machine learning algorithms to predict the arrival time of bikes. Furthermore this thesis studies which factors influence the total travel time of the riders in a mission. The following research questions were considered in this work:

1. What are the variables that can significantly influence the total travel time of a motorbike in a mission and are optimal to be used as inputs to minimize the errors of the prediction model?
2. What proposed machine learning prediction model accomplished the best performance when a benchmark comparison of the ability of the models to determine the estimated time of arrival of an order was performed?

## 1.5 Objectives

A well-designed prediction model for the total travel time of motor bikes in a delivery mission is an essential ingredient to make decisions that can optimize planning. This can be achieved by utilizing the proposed prediction models to make better planning in order to increase productivity.

The main objective of our study was to improve resource management and order scheduling and therefore improve efficiency of the courier services.
Other objectives include:

- Improve reliability of the services.
- Enhance client communication.
- Improve the customer experience

## 1.6 Significance of the study

A well-designed prediction model for the total travel time of motorbikes in a transport mission is an essential ingredient to make decisions that can optimize planning. This can be achieved by utilizing the proposed prediction models to make better planning in order to increase productivity.

The purpose of this research is to propose and evaluate the ability of machine learning algorithms to predict the estimated time of arrival of delivery motorbikes. This research studies which factors influence the total travel time of the motorbikes in a mission within Nairobi and it's environs.

# 2 Literature Review

In this chapter a variety of proposed models based on past studies and their performance is presented. Moreover the input variables to the prediction models and also the technology used to collect the corresponding data are analyzed and discussed.

## 2.1 Machine Learning in solving business problems

The basis for the theoretical background and the application of Machine Learning in a business context was described in [Shmueli et al., 2017]. They described in details how Linear Regression, Neural Networks and Random Forest work and how they are applied in practice on different datasets.

## 2.2 Machine learning use cases in logistics

Machine learning helps understand where a package is in the entire logistics cycle. It allows supply chain professionals to track the location of goods during transportation. Also, it provides visibility into the conditions under which the package is being transported. With the help of sensors, retailers can monitor such parameters as humidity, vibration, temperature, etc.

Besides, Machine learning algorithms helps with real-time route optimization. It tracks weather and road conditions and gives recommendations on how to optimize the route and reduce driving time. This way, trucks can be diverted any time on their way when a more cost-effective route is possible.

Machine learning innovations and artificial intelligence has had a very significant implication for Dalsey Hillblons Lynn (DHL) company, a German logistics company providing international shipping and courier services. The company has improved machine learning to handle crucial operations such as: forecasting and managing spikes and drops in demand for key clients and multiple routes, process weather and other transport delays and systematically process orders in order to strategize an effective delivery method and most importantly reduce the

time taken to make those deliveries. This has in turn helped to solve logistics problems between the clientele, hence improved business relations. The adoption of machine learning at DHL has been so significant to their organization that they even published a report, in partnership with IBM Watson, highlighting the key applications of machine learning in the logistics space [DHL, IBM, 2018]. DHL has also developed a machine learning tool to forecast air freight transit time delays in order to enable proactive mitigation. By analysing a wide parameter of international data, the machine learning model has been able to predict if the average daily transit time for a given plane is expected to rise or fall, up to a week in advance.

On top of that, machine learning has been able to identify the crucial factors affecting shipment delays, also including temporal factors like day of departure or operational factors such as airline on time performance thus helping the company to plan ahead effectively, doing away with subjective guesswork, both time series analysis of information as well as technical analysis of data growing in importance.

## 2.3 Determination of estimated time of arrival

Researchers proposed Support Vector Machine (SVM) models [Cortes and Vapnik, 1995] such as Support Vector Regression (SVR) [Drucker et al., 1996] to make predictions for the arrival time of buses [Md. Noor et al., 2020]. Their model considered current segment, travel time of current segment and the largest travel time of next segment as inputs to their prediction models.

[KONSTANTINOU, 2019] develops a variety of machine learning approaches and benchmark their ability to predict arrival times. In particular Support Vector Regression, Artificial Neural Networks, Gradient Boosting, Random Forest and Stacked Generalization models were developed for the aforesaid task.

The study results verified that machine learning approaches have the ability to predict the arrival times of trucks. The Random Forest and Stacked General-

ization methods outperformed the other machine learning models in terms of Root Mean Square Error and Mean Absolute Percentage Error.

## 2.4 Summary of Literature review

The studies concluded that machine learning models in estimation of time of arrival perform well in terms of prediction accuracy. Most used methods for that purpose appear to be Artificial Neural Networks, Support Vector Regression and Random Forest. In particular most of the studies are focused in estimating the arrival time of urban buses. In this study we develop prediction models for the estimated time of arrival of delivery bikes operating within Nairobi and it's environs. We will explore linear regression, support vector regression and random forest regression algorithms in this project.

# 3 Research Methodology

This section describes the methodology adopted in the study. It consist the following subsections:

Our project focuses on estimating and predicting delivery times for logistics companies, that is, the time between the placement of the order and the receipt by the customer. This is in a bid to assist prospective customers get a more concise idea of how long their order may take to reach them and helps the company brainstorm ideas to reduce the time and hence improve customer satisfaction and delivery efficiency.

## 3.1 Sampling

In this study we took a total of 28,267 samples dataset to work with. Our data includes daily observations taken over the course of one month. We divided our data into two categories, one that we used in training our model and the other used in testing the model. 22,613 samples were used in training and 5,654 was used in testing, considering the several factors like problem statistics, data types dividing the data into the two categories is essential. Our training set was used to develop the model while the test set was used to evaluate the prediction ability of our model. In some of the columns in our dataset there were missing values which were imputed. Columns like weekday were converted to factor since they were numeric. The distance which we used was that between pickup location and drop off location.

## 3.2 Machine Learning

Machine learning (ML) methods are categorized according to the task they are solving. The commonly known types of ML include the supervised, unsupervised and reinforcement learning.

### 3.2.1 Supervised Learning

Supervised learning is the most widely used type of ML [Marsland, 2015]. It is a process where a computer program is trained using known example data. Since the output is also known, this process finds a connection in the form of rules, which relates input data to output data thus applying the learnt rules to the new data. This newly gained knowledge is now able to predict future input and output data. Some of the ML methods used to carry out the tasks under the supervised machine learning include: Support Vector Machines (SVM), Linear Regression, Discriminant Analysis, Support Vector Regression (SVR), Naive Bayes, Logistic Regression, Ensemble Methods etc.

### 3.2.2 Unsupervised learning

Unsupervised learning describes a system that is able to discover knowledge by itself. It identifies similarities between the inputs to categorize inputs by common patterns, for instance clustering, self-organizing maps, multidimensional scaling and non-linear dimension reduction [Marsland, 2015]. Methods used to perform this task include: K-Means, FP-Growth, Gaussian Mixture, Hierarchical, Mean Shift.

### 3.2.3 Reinforcement learning

Reinforcement learning is where the optimal solution is unknown to the system at the beginning of the learning phase and therefore must be determined iteratively. Sensible approaches are rewarded and wrong steps tend to be punished, thus the system finds its own solutions autonomously through directional rewards and punishment. Methods used to perform these tasks include: Q-Learning and Value-Iteration.

## 3.3 Support vector machines

Support Vector Machines (SVM) [Cortes and Vapnik, 1995] are machine learning algorithms that can be used for classification as well as for regression. This

section describes how the SVM algorithm works for classification problems. The basic idea of the SVM algorithm is to find the optimal hyperplane, often called maximal margin hyperplane, that maximizes the distance from the nearest data points on each side. A classification problem is said to be linearly separable if it is possible to find a hyperplane that separates the data into different classes. Moreover when the problem is linearly separable then two parallel hyperplanes that separate the classes and the distance between them is maximal can be selected. The region between those hyperplanes is called margin and points on the boundary of the margin are called support vectors. When the data are standardized and labeled as 1 and -1, for the first and second class respectively, those hyperplanes as well as the width of the margin can be described with the following equations.

$$First\ Parallel\ Hyperplane: \qquad Wx - b = 1 \tag{1}$$

$$Second\ Parallel\ Hyperlane: \qquad Wx - b = -1 \tag{2}$$

$$Maximal\ Margin\ Hyperplane: \qquad Wx - b = 0 \tag{3}$$

$$Margin\ Width: \frac{2}{||W||} \tag{4}$$

After receiving pairs of inputs and labels $(x_j, y_j) \forall j \in \{1, ..., M\}$ from the dataset the SVM algorithm is equivalent to solving the following linear optimization problem.

$$min \qquad \frac{1}{2}||W||^2 \tag{5}$$

Subject to:

$$y_i(Wx_i - b) \geq 1, \qquad \forall i \in \{1, ..., M\} \tag{6}$$

The optimization problem of the SVM algorithm has a quadratic objective function and linear constraints, hence it is a Quadratic Programming (QP) problem. Furthermore, many algorithms exist that can efficiently solve this type of optimization problem.

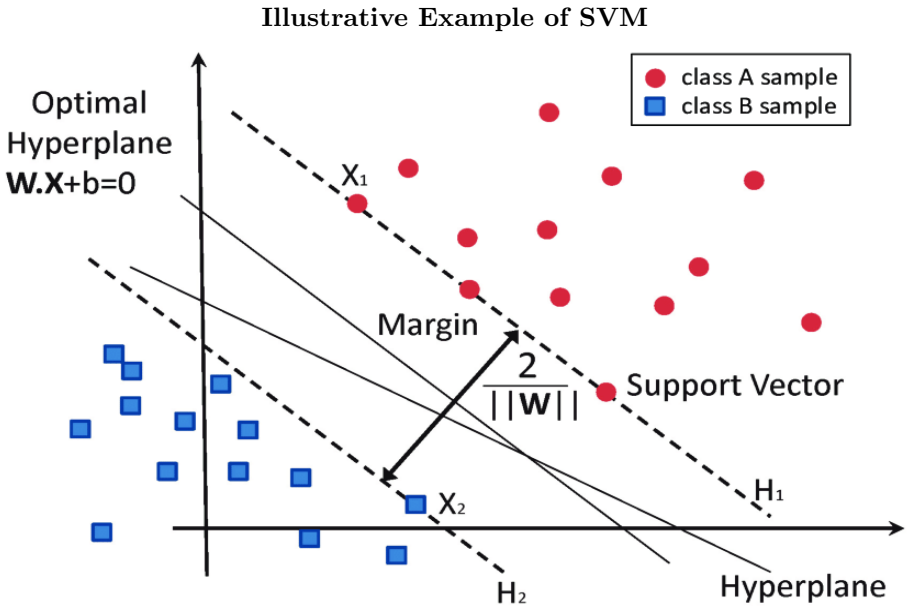



Figure 1: The Support Vector Machine algorithm.

Figure 1 illustrates how SVM algorithm works for classification problems. The maximal margin hyperplane is represented by the red line and the support vectors, the points in the dashed line are also presented.

The following section describes how a support vector machine algorithm can be modified in order to confront regression problems.

### 3.3.1 Support vector regression

This paragraph is focused on the Support Vector Regression (SVR) algorithm which is inspired by the SVM algorithm and is used to confront regression problems. The idea of SVR algorithm is to approximate the unknown real valued function f(x) with a higher dimensional hyperplane by using the kernel function $\phi(.)$. In other words we assume that f(x) can be expressed in the following way:

$$f(x) = w.\phi(x) + b \tag{7}$$

However since we deal with a regression problem instead of a classification problem we introduce an $\epsilon$ tube around our predicted function. Moreover in equation 7, we assumed that is feasible to approximate the function f(.) with precision $\epsilon$ but since this is not usually the case, slack variables $\xi_i$ and $\xi_i^*$ are introduced . After receiving pairs of inputs and target variables $(xi, yi) \forall i \in \{1, ..., l\}$ the SVR algorithm is then equivalent to solving the following optimization problem.

$$min \qquad \frac{1}{2}||w||^2 + C\sum_{i}^{l}(\xi_i + \xi_i^*) \tag{8}$$

Subject to:

$$\begin{aligned} y_i - w.\phi(x) - b &\leq \epsilon + \xi_i \qquad \forall i \in \{1, ..., l\} \\ w.\phi(x) + b - y_i &\leq \epsilon + \xi_i^* \qquad \forall i \in \{1, ..., l\} \\ \xi_i, \xi_i^* &\geq 0 \qquad \forall i \in \{1, ..., l\} \end{aligned} \tag{9}$$

The objective function of the above optimization problem is a trade of between two terms, the regularized term and the empirical error. Minimizing the regularized term, $||w||^2$, will make the function as flat as possible. The empirical error term, $C\sum_i^l(\xi_i + \xi_i^*$ penalizes the objective function if the prediction is outside of the $\epsilon$ prediction tube. This implies that the SVR model can perform regression by solving an optimization problem with two parameters, the regularization constant C and precision parameter $\epsilon$. Those parameters are controlled by the user and is often crucial to chose optimal hyperparameters to optimize the algorithms performance. For the scope of this thesis the optimization of the parameters is done by a trial and error grid-search.

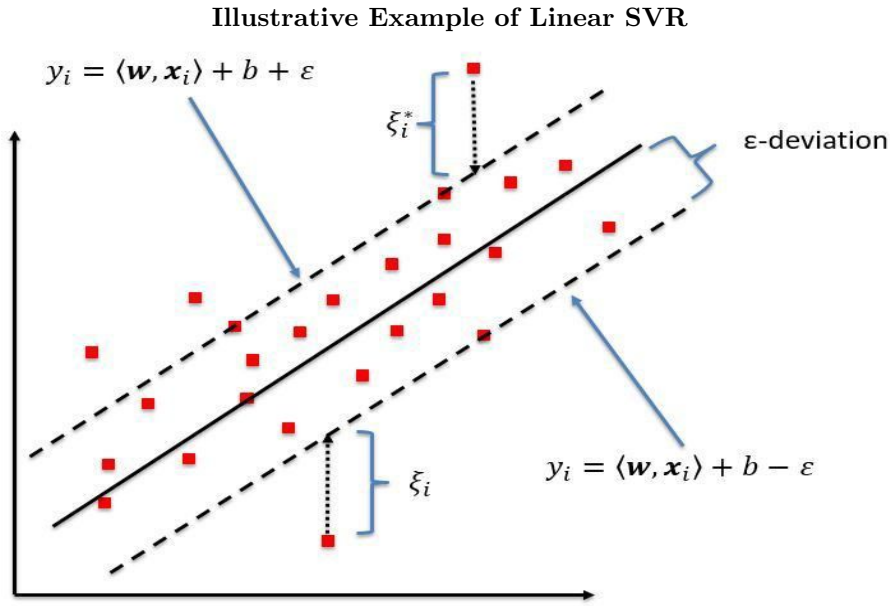



Figure 2: Points outside the $\epsilon$ region around the predicted hyperplane are penalized.

The SVR algorithm is illustrated in figure 2. In this study a Support Vector Regression algorithm is proposed to predict the arrival time of riders in a delivery mission within Nairobi.

## 3.4 Decision tree based models

A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch.

Its important to realize the partitioning of variables are done in a top-down, *greedy* fashion. This just means that a partition performed earlier in the tree will not change based on later partitions. The model begins with the entire data set, D, and searches every distinct value of every input variable to find the

predictor and split value that partitions the data into two regions ($R_1 and R_2$) such that the overall sums of squares error are minimized:

$$minimize\{SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2\} \tag{10}$$

Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. This process is continued until some stopping criterion is reached.

A random vector $X = (X_1, ..., X_d)$ is an array of random variables defined on the same probability space.

Assume training set of $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ drawn randomly from a (possibly unknown) probability distribution $(x_i, y_i) \sim (X, Y)$

**Goal**: Given an ensemble of (possibly weak) regression trees,

$$h = \{h_1(x), ..., h_k(x)\} \tag{11}$$

we build a regressor which predicts y from x based on the data set of examples D.

Each $h_k(x)$ is a decision tree and the ensemble is a random forest. We define the parameters of decision tree $h_k(x)$ to be:

$$\Theta_k = (\theta_{k1}, \theta_{k2}, ..., \theta_{kp}) \tag{12}$$

(these parameters include the structure of tree, which variables are split in which node, etc.)

$\Theta$ is the parameter that determines which random subset $D_\Theta$ of the full data vector D. We only choose a sub-collection of feature vectors $x = (x_1, ..., x_d)$ so $D_\Theta$ is a subset of

$$D = \{(x_1, y_1), ..., (x_n, y_n)\} = full\ training\ set \tag{13}$$

Thus parameter $\Theta_k$ for tree $k$ determines which subset of full training set D we choose for the tree $h_k(x) = h(x\Theta_k)$. Then the ensemble of classifiers (now an

RF) consists of trees, each of which sees a different subset of the data. In data mining (where dimension $d$ is very high) this situation is common (dimensionality reduction).

### 3.4.1 Random forest

Random forests are ensemble of decision trees that randomly learn multiple decision trees. The random forest method consists of a training step that constructs several decision trees, and a test step that classifies or predicts an outcome variable based on an input vector. Given a set of input variables and prediction $(x_i, y_i)$ the method randomly samples with replacement to create a number of decision trees, this technique is known as bagging (or bootstrap aggregation) [Breiman, 1996]. Moreover Random Forest applies a bagging method into the feature space as well. The size of the random predictor subset, denoted by *mtry*, is a tuning parameter of the model, though results are generally nearly optimal over a wide range of this parameter. This randomness guarantees that all of the weak learners, decision trees, are uncorrelated. In other words, this method creates a forest of random and independent decision trees. Therefore the variance of the trees is reduced by sacrificing some bias. Then, in the case of regression problems, the average prediction of all decision trees is taken as the final prediction.

Let $(x_i, y_i), \forall i \in \{1, ..., N\}$ be the training data. The predictor variable vector $x_i$ can be comprised of real valued and/or categorical variables. We will assume that the response $y_i$ is real-valued, as we are concerned herein with regression problems. In CART, the prediction of a tree given the new predictor variable vector X=x is

$$T(x, \theta) = \sum_{i=1}^{N} w_i(x, \theta) y_i \tag{14}$$

where $\theta$ represents the parameters (split points) defining how the tree is constructed and $w_i(x, \theta)$ are weights such that $w_i(x, \theta) > 0$ if the observation $x_i$ is in the same terminal node as x and $w_i(x, \theta) = 0$ otherwise. The weights are

normalized so that they sum to 1. Specifically, if $L(x, \theta)$ is the leaf (i.e., terminal node) in which x lands, then:

$$w_i(x, \theta) = \frac{I\{x_i \in L(x, \theta)\}}{\#\{j | x_j \in L(x, \theta)\}}. \tag{15}$$

In equation 15, $I\{x_i \in L(x, \theta)\} = 1$ if and only if $x_i$ is in the leaf $L(x, \theta)$ and the denominator is the total number of training points that are in this leaf.

Random Forests consists of a collection of CART predictors $T(x, \theta_k) \forall k \in \{1, ..., K\}$ where the parameters $\theta_k$ are independent, identically distributed random vectors that determine how a tree is constructed and K is the number of trees.

For RF the conditional mean $E(Y|X = x)$ is estimated as the average prediction over the K trees. Define

$$w_i(x) = \frac{1}{K} \sum_{k=1}^{K} w_i(x, \theta_k) \tag{16}$$

so $w_i$ is the average of the weights associated with the individual trees. Then the (deterministic) RF prediction for $E(Y|X = x)$ is

$$T(x) = \sum_{i=1}^{N} w_i(x) y_i. \tag{17}$$

Thus, the prediction is a weighted average over all observations and the weights depend on the covariate $X = x$. As shown in [Lin and Jeon, 2006] the weights $w_i(x)$ are largest for those i where the conditional distribution of Y given $X = x_i$ is most similar to the distribution of Y given $X = x$.

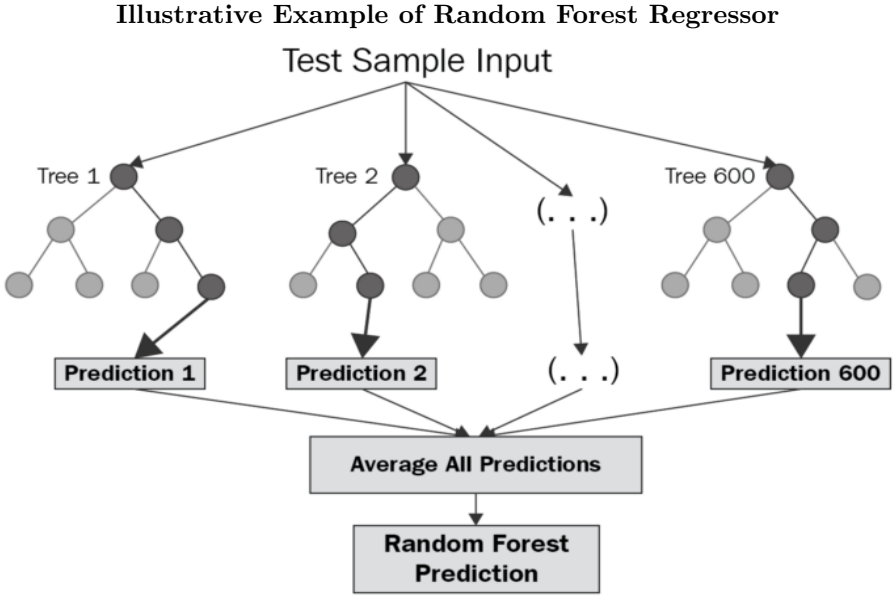



Figure 3: Random forest Regression With 600 Trees

An example of the Random Forest algorithm with 600 decision trees for a regression problem is presented in figure 3.

In this study a Random Forest algorithm for regression is proposed to predict the estimated time of time of a rider in a delivery mission within Nairobi and it's environs. A trial and error grid-search algorithm is utilised to optimise the parameters of the proposed random forest model.

# 4 Data Analysis

In order to answer the research question about how to use Machine Learning to improve prediction of expected delivery time by combining historical transit variability and data pertaining to external relevant factors, we perform the following procedures in the analysis. Note that in this report we use the bikes platform, which represent the most used platform for the company, as a base to build a model that will be a proof of concept for the company's future use on different units of carriage, which may include all pickups and trucks.

## 4.1 Toolkit

We were able to utilize the vast knowledge in the R and Python open-source communities for algorithm development, for instance we used the tidyverse package [Wickham et al., 2019] for data manipulation, the caret and scikit learn [Kuhn, 2008, Pedregosa et al., 2011] for model development and hyperparameter tuning.

## 4.2 Terminology and definitions

First, it is important to establish data science naming conventions that will be used throughout this report. In this project, we used supervised Machine Learning algorithms. To train these algorithms, we need as an input various samples that consist of features and a target variable. A sample is one row in the dataset and consists of all the information available about one delivery. A feature is the name for one specific type of information about this row. Since a row represents a delivery, a feature could be, for example, the average time a bike needs for a specific distance. The target variable is defined as the variable the model tries to predict. For our model, this is the time period in minutes from when the bike left from the pickup point to the time when the bike delivers the parcel at the point of destination.

## 4.3 Analyzing the historical data and selecting relevant external factors

Our partner company Sendy provided us with the following: 3 files, in the comma separated values (CSV) format covering a total of 28,267 deliveries that occurred within a period of one month in the year 2019. The CSV files comprises of a training dataset which contains 21,201 observations. The other CSV file contains an evaluation dataset 5,550 observations that is used to evaluate the accuracy of the model. The third dataset contains data pertaining to individual riders which include their experience and ratings as recommended by previous customers they served. For our project, the interesting part of the dataset consists of timestamps indicating when the order or parcel reached a milestone during the transport process. The important milestones for this project are, in chronological order.

- placement day of month
- placement weekday mo 1
- placement time
- confirmation day of month
- confirmation weekday mo 1
- confirmation time
- arrival at pickup day of month
- arrival at pickup weekday mo 1
- arrival at pickup time
- pickup day of month
- pickup weekday mo 1
- pickup time
- arrival at destination day of month

- arrival at destination weekday mo 1
- arrival at destination time

The full list of columns with the corresponding description can be found in Appendix B (1).

### 4.3.1 Selection of Relevant External Factors

- **New customer**

  The ups and downs of finding the precise block or location of a new customer can be time consuming

- **Weather**

  During rainy seasons the delivery of bikers can delay.

- **New Rider**

  A new rider can take longer to find routes in an area that he is new to leading to delays.

## 4.4 Exploratory data analysis (EDA)

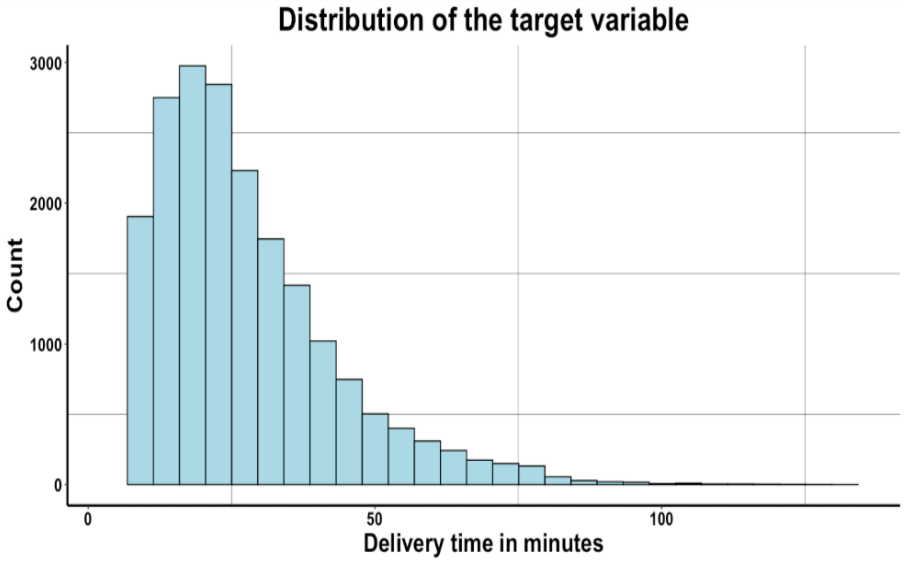
Exploratory data analysis (EDA) is the process of analysing data to gain further insight into the nature of the data, its patterns and relationships between the variables, before any formal statistical techniques are applied. We approach the data free of any pre-conceived assumptions or hypotheses. We first see the patterns in the data before we impose any views on it and fit models. In addition to discovering the underlying structure of the data and any relationships between variables, exploratory data analysis can also be used to:

- detect any errors (outliers or anomalies) in the data
- check the assumptions made by any models or statistical tests
- identify the most important/influential variables
- develop parsimonious models – that is models that explain the data with

the minimum number of variables necessary.

### 4.4.1 Distribution of the Target Variable

We aim to investigate the distribution of the target variable.




The times are positively skewed.

### 4.4.2 Handling Missing Values

Precipitation has 97.39635% missing values. We choose to drop the column. Temperature has 20.6 % missing values. We choose to keep the feature and carry on with the data cleaning.

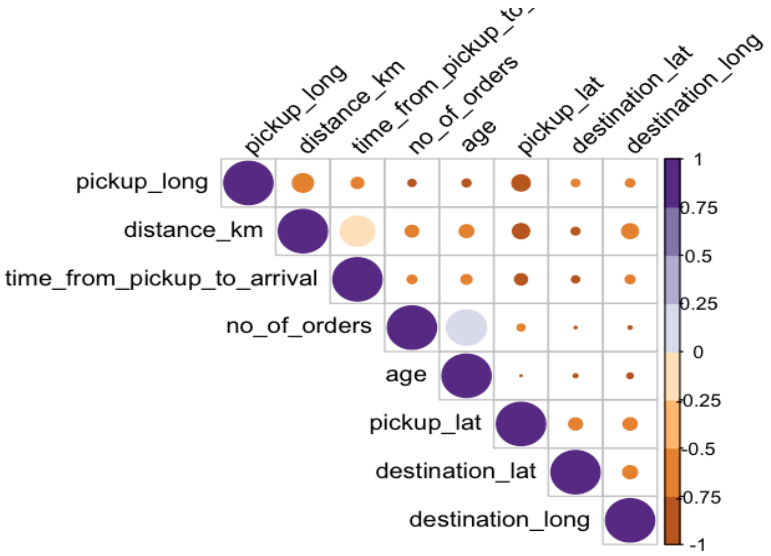
**Imputing missing temperature values** We use caret preprocessing feature to replace the missing values with the median of the temperature. This method is known as `medianImpute`

### 4.4.3 One-Hot encoding of categorical variables

Categorical columns as features need to be converted to numeric in order for them to be used by the machine learning algorithms. Just replacing the categories with a number may not be meaningful especially if there is no intrinsic ordering amongst the categories. So what we did instead is to convert the categorical variable with as many binary (1 or 0) variables as there are categories.

### 4.4.4 Correlation among features

We will use corrgram package to visualize and analyze the correlation matrix. In theory, the correlation between the independent variables should be zero. In practice, we expect and are okay with weak to no correlation between independent variables. We also expect that independent variables reflect a high correlation with the target variable.




From the above correlogram we noted that:

- No of orders and age are highly correlated. We had to eliminate one to avoid feature correlation.
- Distance has a strong correlation with time from pickup to arrival. The longer the distance the rider has to travel, the longer the estimated time of arrival.

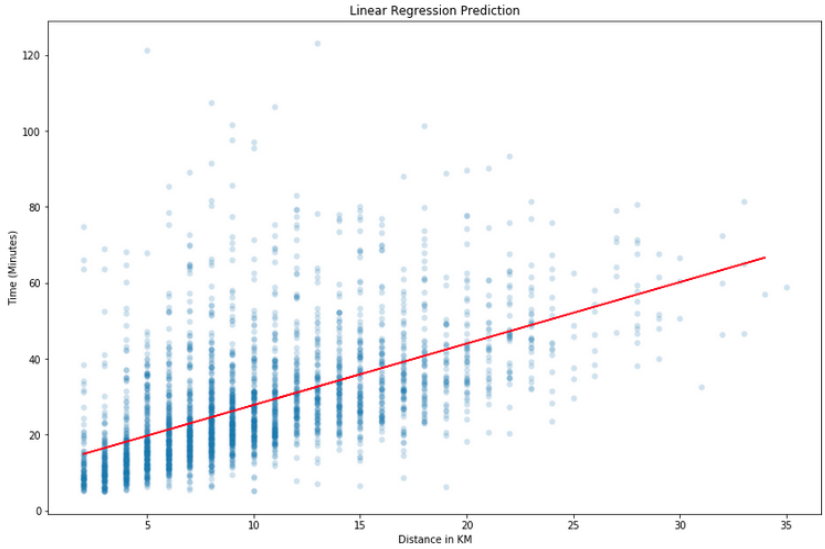
## 4.5 Selecting a Machine Learning algorithm

Building on the research from Fancello et al (2011) [Pani et al., 2015] and their Neural Network model, we experimented with Linear Regression, Support Vector Regression (S.V.R) and Random Forests to see if these algorithms perform well on our dataset. After comparing the results of these three algorithms, we decided to build the final prediction model with a Random Forest algorithm.

To test and validate the performance of the different algorithms, we split the data in a training set and a test or validation set. The training set consist of 75% of data and was used to train the models and select the best hyperparameters. The test set data was held out during the training phase and was used to validate the performance of our models. All algorithms were trained and tested on the same training and test set so that we could compare their performance. We used Python and R programming languages in the Jupyter notebook environment to write the code for our models. To create, train and use the final models for prediction there are three relevant phases in the final program. The first script is used to clean the data, remove missing values and transform all columns to the right format. The second script was used to visualize the spatial data on an interactive map. We achieved this by utilising the Shiny package that R provides. Users can interactively view the order pick-up locations and their corresponding drop-off locations on the map available on Appendix C (1). The third script is used for prediction in a production environment.

### 4.5.1 Linear Regression

Linear Regression establishes a relationship between dependent variable, Y, and one or more independent variables (X) using a best fit straight line (also known as regression line). We use the linear regression model as the benchmark model since we earlier indicated that distance has a positive correlation with time taken to arrive. We assume that delivery missions take an amount of time that is directly proportional to the distance.




$$y = 1.641x + 11.268 \tag{18}$$

```
Call:
lm(formula = time_from_pickup_to_arrival_minutes ~ distance_km,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-50.824  -7.707  -3.363   3.827 114.020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.26818    0.20760   54.28   <2e-16 ***
distance_km  1.64087    0.01856   88.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.98 on 15227 degrees of freedom
Multiple R-squared:  0.3392,    Adjusted R-squared:  0.3391
F-statistic:  7816 on 1 and 15227 DF,  p-value: < 2.2e-16
```

Figure 4: Linear Regression Results

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient $0 - 100\%$ scale.

R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots. R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

If your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value. Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors

in the model constant. Obviously, this type of information can be extremely valuable.

In our analysis, our R-Squared is 33.92% which is considered as low. However, the regression coefficients are significant so we go on with the analysis.

### 4.5.2 Support Vector Regression

The second proposed prediction model is a Support Vector Regression model implemented in the Scikit-learn library. The parameters of the model were tuned with a trial and error grid search. Table below shows the optimal parameters chosen by the grid search algorithm.

| Parameter Name | Value |
|---|---|
| Kernel | Linear |
| C | 5.311 |
| epsilon | 5.408 |
| Intercept | 9.463 |
| Coefficient | 1.607 |
| MAE | 8.14 Minutes |
| RMSE | 12.13 Minutes |
| Percentage within $\epsilon$ | 75.50 % |

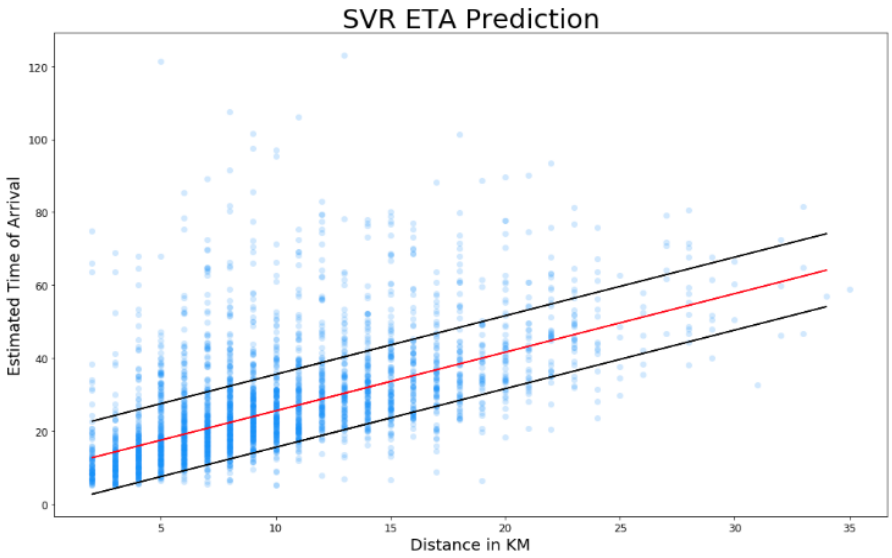
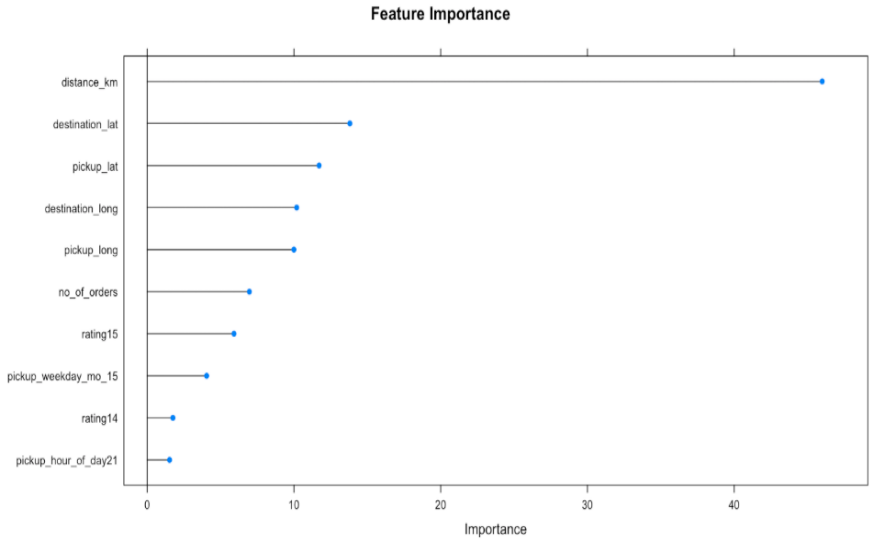



Figure 5: Support Vector Regressor with a Linear Kernel

### 4.5.3 Random Forest

This study presented a variable importance plot to show the prediction power of the main explanatory variables. The variable importance indicates the effect of an explanatory variable on the accuracy of a model. Therefore, when an explanatory variable improves the performance of a model, the importance of the variable increases.

Random forests are useful for feature selection in addition to being effective for classification and regression. One approach to dimensionality reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features. If an attribute is often selected as best split, it is most likely an informative feature to retain. A score calculated on the attribute usage statistics in the random forest tells us – relative to the other attributes – which are the most predictive attributes.




| Parameter Name | Value |
|---|---|
| mtry | 15 |
| MAE | 9.24 Minutes |
| RMSE | 12.81 Minutes |

# 5 Conclusion

The main objectives of the study were to design, develop and evaluate the ability of machine learning models to predict the estimated time of arrival of a delivery partner operating in delivery mission as well as to identify which features are optimal to be used as inputs to the predictions models.

A Random Forest algorithm is a useful technique on the prediction of estimated delivery times for Sendy. This is because it presented a variable importance plot to show the prediction power of the main explanatory variables, hence indicating the effect of an explanatory variable on the accuracy of the model. This project fits a random forest algorithm to the Sendy dataset, assesses the significance of predictor variables, accuracy of the fit and assesses the predictive potential of machine learning algorithms on logistics companies.

From the results, the model showed that distance was an extremely significant variable in estimating the time taken from pick-up to delivery. Other important variables included rider ratings and the number of orders the delivery partner has successfully completed. It can further be seen that a rider having a rating of 15 can significantly influence the estimated time of arrival, this informs us that a rider's ability to deliver on time is key to customer satisfaction.

## 5.1 Limitations of our Study

A limitation of the study is the unavailability of data related to important factors such as traffic flow information. One of the main assumptions of the study is that traffic greatly influences the arrival of delivery bikes but unfortunately no information about the traffic situation was available to us.

## 5.2 Recommendations For Future Research Studies

A study that repeats the experiments performed in the current study by including traffic flow information should be considered as a future work. Furthermore, collecting traffic flow data can be very hard and quite expensive, hence research for designing and developing models that can approximate traffic flows should be conducted.

# 6 References

## References


[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2):123–140.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

[DHL, IBM, 2018] DHL, IBM (2018). "dhl and ibm report cites benefits and potential of ai in logistics". https://www.logisticsmgmt.com/article/dhl_and_ibm_report_cites_benefits_and_potential_of_ai_in_logistics.

[Drucker et al., 1996] Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines.

[KONSTANTINOU, 2019] KONSTANTINOU, K. (2019). A comparative study of machine learning algorithms and their ability to determine estimated time of arrival.

[Kuhn, 2008] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

[Lin and Jeon, 2006] Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101:578–590.

[Marsland, 2015] Marsland, S. (2015). Machine learning: an algorithmic perspective. *Decision by committee: ensemble learning*, pages 267–280.

[Md. Noor et al., 2020] Md. Noor, R., Yik, N., Kolandaisamy, R., Ahmedy, I., Hossain, M. A., Yau, K.-L., Md Shah, W., and Nandy, T. (2020). Predict arrival time by using machine learning algorithm to promote utilization of urban smart bus.

[Pani et al., 2015] Pani, C., Vanelslander, T., Fancello, G., and Cannas, M. (2015). Prediction of late/early arrivals in container terminals - a qualitative approach. *European Journal of Transport and Infrastructure Research*, 15:536–550.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[SendyIT, 2021] SendyIT (2021). Https://www.sendyit.com — Sendy, end to end transport and logistics for business. https://www.sendyit.com.

[Shmueli et al., 2017] Shmueli, G., Bruce, P., Yahav, I., Patel, N., and Lichtendahl, K. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.

[The Standard, 2018] The Standard (2018). https://www.standardmedia.co.ke — Financial Standard, q&a: Sendy creates delivery opportunities for drivers. https://www.standardmedia.co.ke/business/article/2001290105/q-a-sendy-creates-delivery-opportunities-for-drivers.

[Wickham et al., 2019] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *J. Open Source Softw.*, 4(43):1686.

# Appendix A

The code used for the analysis of this project is publicly available on Github.

Github repo

# Appendix B

Description of each of the columns contained in the original dataset provided by Sendy.

Table 1: Features in the original dataset

| Column Name | Description |
|---|---|
| Order No | Unique number identifying the order |
| User Id | Unique number identifying the customer on a platform |
| Vehicle Type | For this competition limited to bikes, however in practice, Sendy service extends to trucks and vans |
| Platform Type | Platform used to place the order, there are 4 types |
| Personal or Business | Customer type |
| Placement Day of Month | i.e 1-31 |
| Placement Weekday | (Monday = 1) |
| Placement Time | Time of day the order was placed |
| Confirmation Day of Month | i.e 1-31 |
| Confirmation Weekday | (Monday = 1) |
| Confirmation Time | time of day the order was confirmed by a rider |
| Arrival at Pickup Day of Month | i.e 1-31 |
| Arrival at Pickup Weekday | (Monday = 1) |
| Arrival at Pickup Time | Time of day the rider arrived at the location to pick up the order - as marked by the rider through the Sendy application |
| Pickup Day of Month | i.e 1-31 |
| Pickup Weekday | (Monday = 1) |
| Pickup Time | Time of day the rider picked up the order - as marked by the rider through the Sendy application |
| Distance covered (KM) | The distance from Pickup to Destination |
| Temperature | Temperature at the time of order placement in Degrees Celsius (measured every three hours) |
| Precipitation in Millimeters | Precipitation at the time of order placement (measured every three hours) |
| Pickup Latitude and Longitude | Latitude and longitude of pick up location |
| Destination Latitude and Longitude | Latitude and longitude of delivery location |
| Rider ID | ID of the Rider who accepted the order |
| Time from Pickup to Arrival | Time in seconds between 'Pickup' and 'Arrival at Destination' - calculated from the columns for the purpose of facilitating the task |

# Appendix C

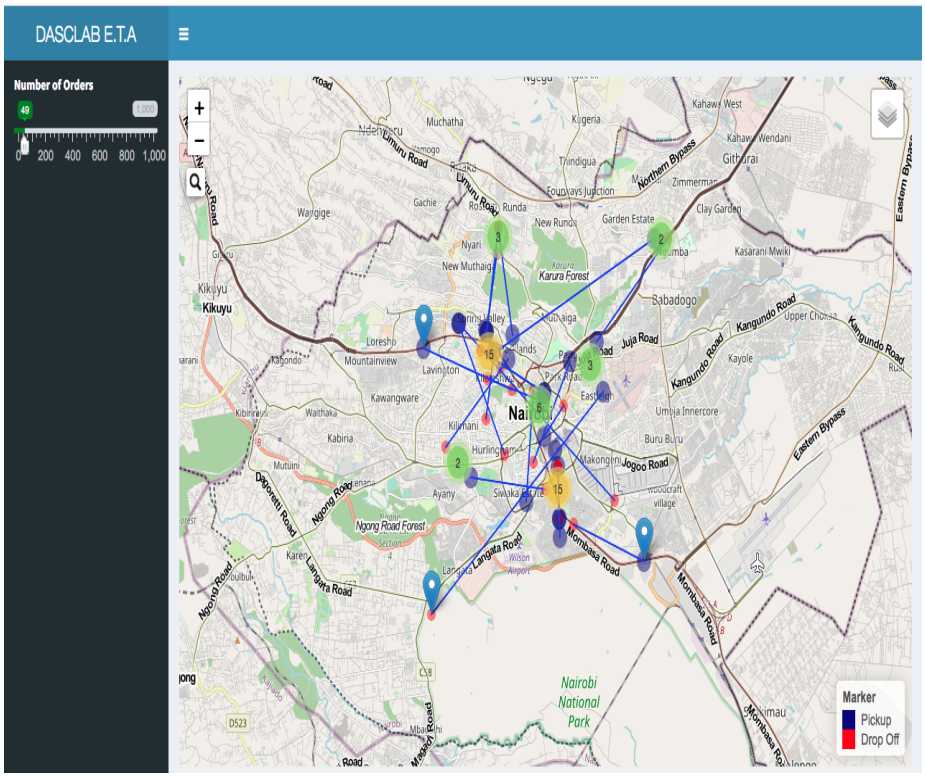



Figure 6: Map to interactively visualize where pickup and dropoffs are located

Visit the map using the link: DASCLAB E.T.A