Exploratory Data Analysis: Understanding CargoLedger Data sets

Timothy J. Palmer
University of Amsterdam
Amsterdam, The Netherlands
timothyjhn1@gmail.com

1 INTRODUCTION

This Exploratory Data Analysis (EDA) sets to provide a summary and understanding of the data sets provided by sponsor, CargoLedger B.V.,in efforts to support the Master's Thesis candidate subject at the University of Amsterdam. The digitization of logistical methods within the supply chain can collect, forecast and analyze a lot of data in real time. The more data that is securely shared between stakeholders, a more accurate and transparent supply chain will emerge. The sponsor aims to investigate the orchestration of this data to make interactions using sequential decision making techniques more efficient.Additionally, organizing data via blockchain can make sharing data easy while remaining secure. Understanding the importance of stakeholder and organizational data is essential to optimize and model decisions based on factors generated during the shipment of products. Data integration between stakeholders can make for more better forecasting and once modeled using machine learning techniques, can optimize the supply chain in real time. This data integration includes contextual information, stakeholder roles, the state of the cargo, etc. This EDA will focus on organizing and understanding data which will be used to explore the research question.

2 METHOD

For this section, we begin by analyzing the data provided by CargoLedger B.V. This data consists of a zip folder containing comma separated values (csv) files, and extendable markup language (xml) files from 3 logistics companies in 3 separate folders.The data set folder files range from 49 files to 563 files per folder respectively. The EDA is performed using Python 3.8.8. and runs on Jupyter Lab 3.3.0. In order to properly run the data files in a readable format for Python, the data had to be organized and cleaned before use. To load the data into the notebook, I needed to first merge the csv files into one using Microsoft Excel. Once merged, the now combined data file for logistics company Bosdaalen was loaded into Jupyter lab notebook.

2.1 Data Cleaning

The data was cleaned and organized by checking the different data types within the file, as stated in figure 1.

```
datetime              object
tripid                 int64
tuid                  object
description           object
company                int64
                      ...
goodspalletplaces    float64
goodscolli           float64
goodsproduct          object
goodsdescription      object
goodsreference        object
Length: 95, dtype: object
```

Figure 1: Data Types

There are several columns that include personal identifying information (PII) that were removed due to privacy concerns. This included information identifying driver details. Driver phone number, driver e-mail, driver id and co-driver information were all removed from the notebook. Proper identification of transport will be identified by company id and trip id.

2.2 Understanding the Columns

The columns detail several diferent aspects concerning the delivery of the products to the customers. There are a total of 83 columns, which were split into 3 different categories of importance. The first category are specified as the "trip information", this included column names including, "datetime", "tripid", "description" and "company". This type of data is given throughout the file and is reused several times. The second category is listed as customer identifiable information. This type of information comprised of the driver name, drive email and phone numbers. This data was removed, and usage will be detailed in a later summary in this report. The last category listed is "Customer/Product" information. This data details the specifics of the products being delivered, this includes the weight of the shipment, the product description and destination. Additionally, this category adds customer

information that is not as identifiable as the driver information. Further discussion on the use of this data will be discussed.

2.3 Missing Values

Throughout the data file there are several values that reflect as null within the notebook. columns containing vital and useful data such as "product id", "goods product" and "weight", have gaps of missing values. This leaves a lot of needed information in order to develop a model for sequential decision making. As seen in figure 2, the .isnull function is used to show the amount of null values within the data set. In an effort to fill the gap of missing values, a random data generator will be used for missing numerical values, using the value ranges from the data set. This will be explained in detail once implemented with approval from the sponsor.

```
datetime                0
tripid                  0
tuid                    0
description             0
company                 0
                      ...
goodsproduct          253
goodsdescription      389
goodsreference        320
trans_countrycode       6
datetime1               0
Length: 97, dtype: int64
```

Figure 2: Columns containing null values

3 INITIAL EXPLORATION

In order to determine what machine learning techniques will provide more accurate forecasting within the supply chain, having an understanding of the different correlations between the data variables is an important step. To begin the exploration, an initial view of the outliers within the data was conducted to avoid any inaccurate results. To this effect, the Interquartile range technique or IQR score technique was used.

3.1 Numerical Data Distribution

The histograms below display the frequency of occurrences in variables in an interval. In this study, is is important to identify the types of products, weight, date and time and location of delivery of goods. Below looks at the distribution of the man features of the data set and plots them.

There are some features that are shown not to be as relevant than others. For this the main focus is an analysis of data that have correlations to features that are relevant.

Figure 3: Numerical Data Distribution between features

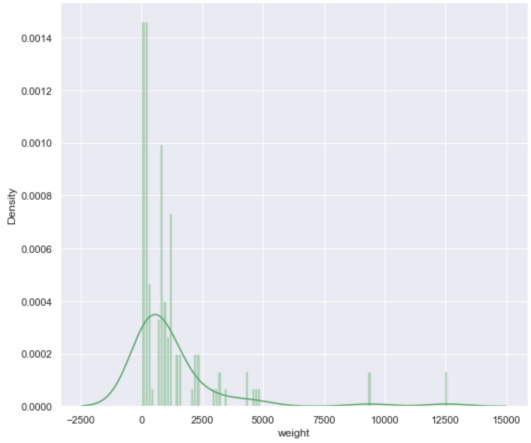

Figure 4: Weight Density Distribution

Relevant features as described above, include weight. A visualization of weight is shown in figure 4 below. A heat map is used to find dependent variables within the data set. The relationship between the features are shown below. There are many features that correlate to "weight" and "quantity", while others show relationships between "stop action" and "stop duration". One note to add is that there is still data that must be configured in order to be applied to the heat map. In a more detailed EDA the relationship between product and weight, as well as date and time with location will be

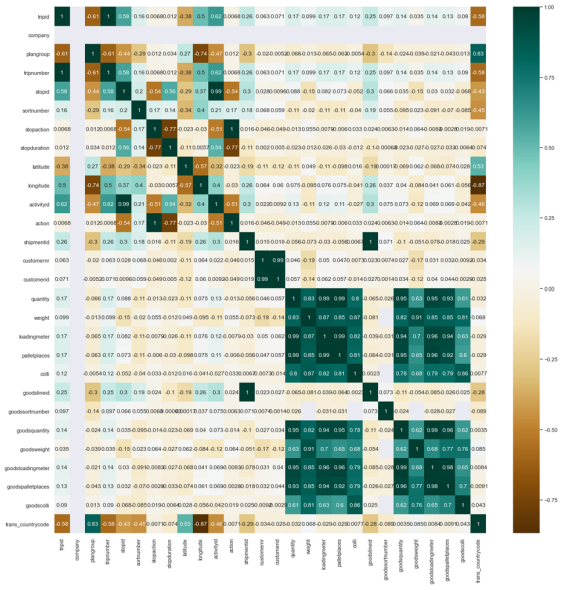

Figure 5: Heatmap showing feature relationships

included and have a higher chance of showing a high relation on this heat map.

3.2 Concerns

Main efforts to conduct a through EDA are underway according to the thesis project plan milestones. Retrieving the data from the sponsor was seamless, however cleaning and organizing the data for proper use is time consuming and will take a lot more resources than originally anticipated. The projected outcome of the final EDA will show correlations to features containing product data and possible delays with the transport of the products. The data provided does not show delays are stoppages within the cargo to provide this point. This will be analyzed in further detail in the final EDA report.

4 CONCLUSION

The data provided clearly shows transportation data between stakeholders for a particular date of operation. This data is useful in many ways however there is more data needed in order to have an effective EDA to preform models on. The main concept behind this data is to show how useful hidden or confidential stakeholder data will assist in the sequential decision making process regarding the transportation of the products. In the future, there will be a need for sensor data of the products, this includes the temperature, product origin, weather, and other stakeholder activities surrounding the supply chain.