Non-Technical Executive Summary:

In this project, our team set out to explore the shipping effectiveness of BigSupplyCo. Our goal was to summarise the shipping effectiveness of BigSupplyCo with the use of illustrative graphs and explore whether the customer segment, market region, product type, and payment method affected the shipping effectiveness.

Our main methodology was to combine the five datasets into one after eliminating any data columns that we were sure would not provide us with any useful info for determining correlation or were empty, like the product image and description. After this, we gradually cut down the dataset to exclude factors that would not be logically expected to bear any relationship with shipping effectiveness. Data columns for which the information contained were also present in another data column, e.g. Order Status and Payment Method columns. While doing this, we explored the relationship between different data columns by filtering the dataset.

We aimed, in the end, to train a classifier on the dataset, for instance, an SVM Classifier, that could then be used to identify future deliveries which were at risk of being delivered late. Unfortunately, this could not be achieved due to time constraints.

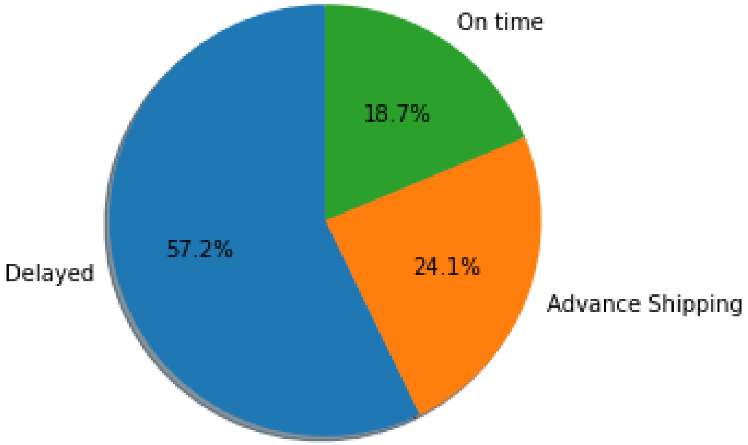


Figure 1: Orders by Delivery Status

Our key findings were that most BigSupplyCo's orders were delivered late, which held for different regions, products, and customer segment combinations. We have also failed to find a strong enough correlation between the Late Delivery Risk and any factors mentioned above.

Data Analysis Steps:

Step 1: Initial Adjustments

The first step was to import all five datasets and visually inspect them. Given that the datasets appeared very large and complex, we decided to make the following initial adjustments to cut non-existing data columns and make the data easier to inspect.

For the customer dataset:

- Dropped all columns containing personal customer information:
    - [Customer Email, Customer Fname, Customer Lname, Customer Password]
- Replaced the Country Label EE.UU. with USA (EE.UU. is Spanish for the USA)
- Set the index to customer-id

For the departments' dataset

- Set the index to Department ID

For the products and categories dataset

- Merged these two as the category dataset was quite small.
    - First renamed the Category Id columns to match then merged it on them
- Dropped the columns [Product Description, Product Image] as they were empty
- Dropped the [Product Status] column as all items appeared to be in stock

For the Orders Dataset

- Reset index to Order Id

- Removed [Unnamed: 0] column
- Renamed the column names to match those in other data frames for merging [Product Id], [Customer Id]

Step 2: Merging datasets

Merges in order:

- Merged orders and customers datasets on [Customer ID]
- Merged this data frame with departments dataset on [Department Id]
- Merged this data frame products dataset on [Product Id] (previously [Order Item Cardprod Id])

Step 3: Removing Excess and Duplicate Data

Cut the following data columns for being irrelevant for our analysis or because their information was embedded in another column:

- [Longitude, Latitude]: All department stores appeared to be very close to each other in inner Puerto Rico (**See figure x)**, and we had a Department Id column to represent them anyhow

- [Customer Country, Customer City, Customer State, Customer Street, Customer Zipcode] [Customer Country]: All customers are in Puerto Rico or the US. The orders do not ship to their location, so these are irrelevant.
- [Order City, Order State, Order Country, Order Zipcode]: We have decided to limit our geographical data to the Market data; otherwise, the set of data will be too specific, and we want to look at the geographical data from a higher level

- [Product Name, Department Name, Category Name]: Text-based descriptive columns will not aid the analysis, and we possess the numerical IDs

- [Category ID]: Department ID already contains information on Category ID as all stores sell exclusive categories of products, so we decided to limit our analysis to Department ID

- [Order Item Id]: We as will use product category to get info about the product, products

,[Customer Id]: We are only interested in the Customer Segment

- [Payment Method]: Order Status includes this information already (**See figure x)**

Step 4: Dealing With Cancelled Orders:

We have decided to drop all orders where Shipping was cancelled (Delivery Status] != Shipping cancelled)] as these orders are either fraudulent or cancelled. These are all marked as having Late Delivery Risk as null, even though their delay percentage is about the same as for the rest of the orders.

About 4.6% of orders appeared to have their Shipping cancelled

- Another approach would be to include these but fix the Late Delivery Risk column values, but the fraudulent and payment failure may have affected Shipping, so we decided to disregard these

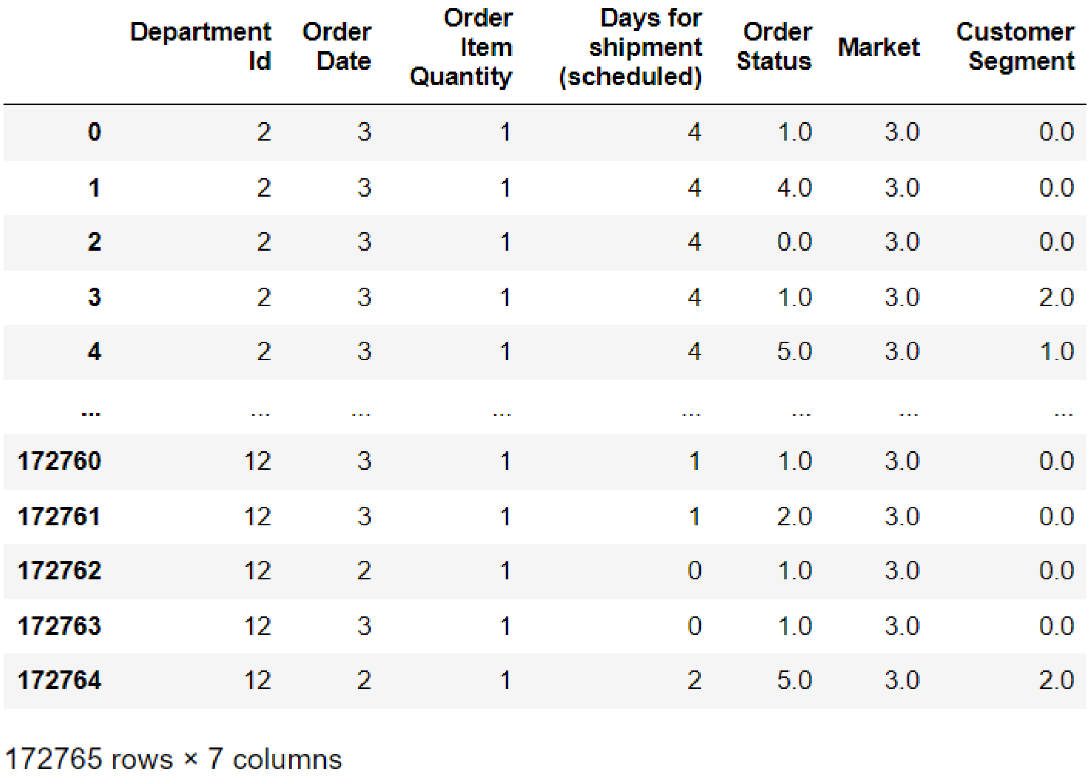
Step 5: Categorical Encoding

- Convert [Order dates] to categorical years: 1,2,3,4 for 2015,2016,2017,2018, as this would be better suited for a possible ML application, and we did not wish to delve into the month and day variation.

- Used Sci-kit’s ordinal encoder to encode [Order Status], [Market],[Order Region], [Delivery Status], and [Customer Segment] Info numerically

Step 6: Further cuts to the data

- Dropped [Days for Shipping (real)] and [Delivery Status] as we wanted to focus on a binary distinction between late and on-time deliveries
- Dropped [Order Region] and used Market to denote geographical location for simplicity



	Department Id	Order Date	Order Item Quantity	Days for shipment (scheduled)	Order Status	Market	Customer Segment
0	2	3	1	4	1.0	3.0	0.0
1	2	3	1	4	4.0	3.0	0.0
2	2	3	1	4	0.0	3.0	0.0
3	2	3	1	4	1.0	3.0	2.0
4	2	3	1	4	5.0	3.0	1.0
...	...	...	...	...	...	...	...
172760	12	3	1	1	1.0	3.0	0.0
172761	12	3	1	1	2.0	3.0	0.0
172762	12	2	1	0	1.0	3.0	0.0
172763	12	3	1	0	1.0	3.0	0.0
172764	12	2	1	2	5.0	3.0	2.0

172765 rows × 7 columns

Figure 1: Final Dataset Form

Step 7: Generate Visuals

Figures 2-7, generated by Matplotlib, show what percentage of the Late Deliveries by different factors. All charts are normalized. Meaning that they represent what percentage of the late deliveries each category would be responsible for had all categories been equally represented. (Eg. 10% of 1000 consumer orders and 10% of 9999 corporate orders will have an equal slice of the pie). See Figure 8 below for the meaning of each categorical ID for encoded variables, as well as total order counts for each category. Figures 9-12 graphically represent the distribution of total orders by some of these categories.

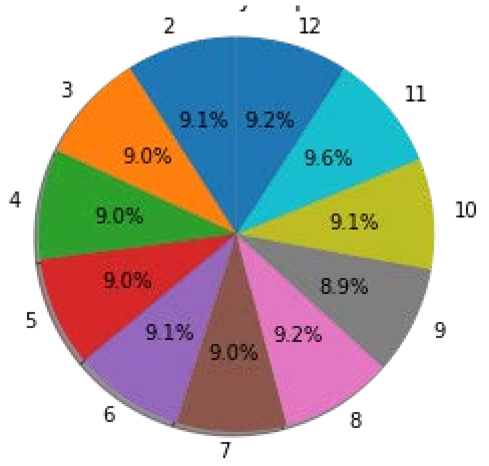


Figure 3: Late Deliveries by Department ID

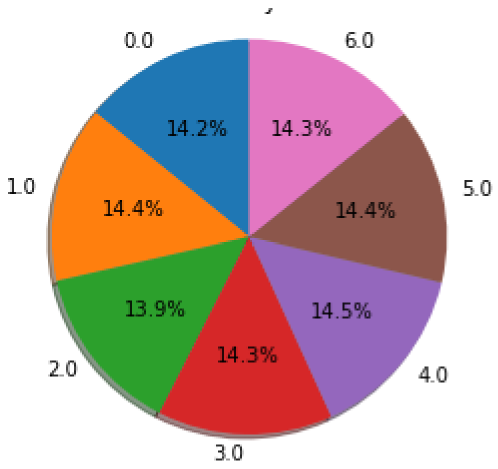


Figure 2: Late Deliveries by Order Status

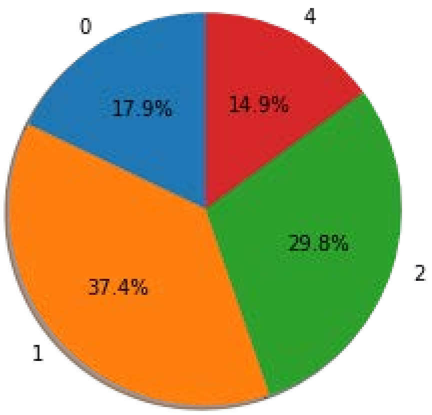


Figure 4: Late Deliveries by scheduled shipment days

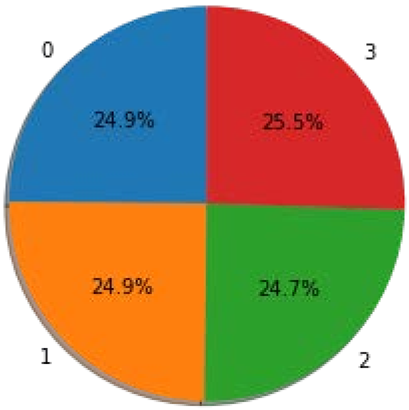


Figure 5: Late Deliveries by Order Year (-2015)

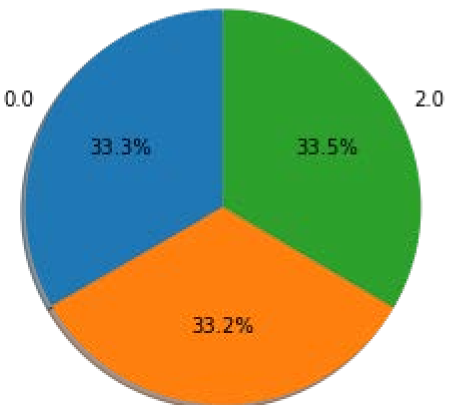


Figure 7: Late Deliveries by Customer Segment

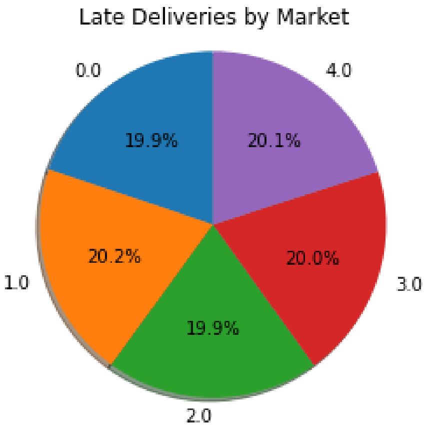
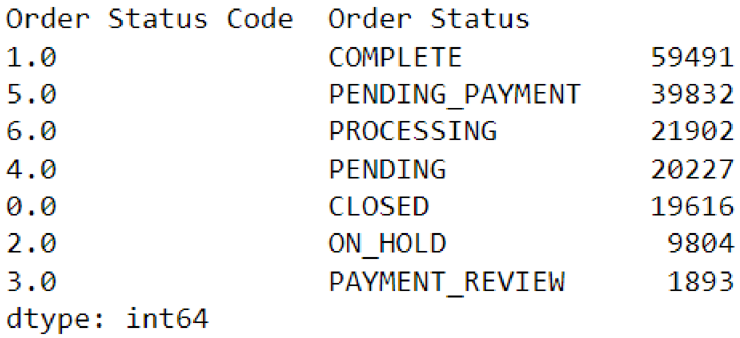


Figure 6: Late deliveries by Market

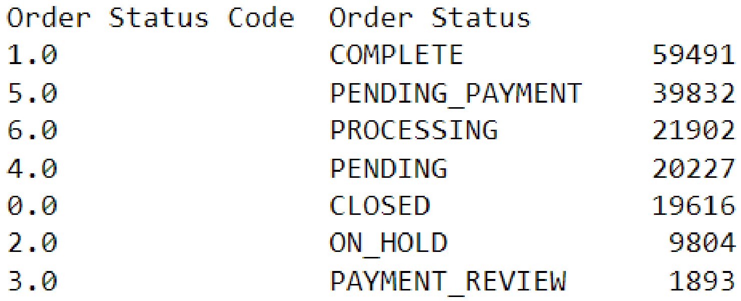
The above plots did not suggest any factor other than scheduled shipping days had any significant bearing on the shipping effectiveness.

However, it was not surprising that shipments scheduled to be delivered the earliest experience the most delays.

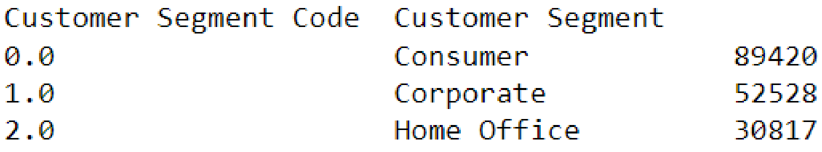
Figure 8: Categorical varaibles by Name and the total order counts associated



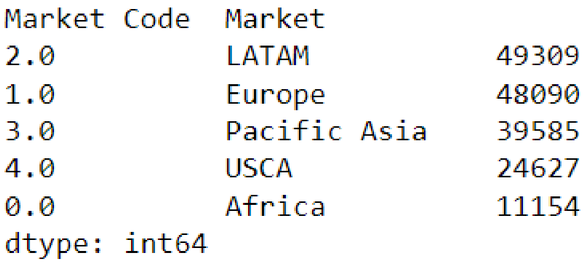
```
Order Status Code  Order Status
1.0                COMPLETE           59491
5.0                PENDING_PAYMENT    39832
6.0                PROCESSING         21902
4.0                PENDING            20227
0.0                CLOSED             19616
2.0                ON_HOLD             9804
3.0                PAYMENT_REVIEW      1893
dtype: int64
```



```
Order Status Code  Order Status
1.0                COMPLETE           59491
5.0                PENDING_PAYMENT    39832
6.0                PROCESSING         21902
4.0                PENDING            20227
0.0                CLOSED             19616
2.0                ON_HOLD             9804
3.0                PAYMENT_REVIEW      1893
```



```
Customer Segment Code  Customer Segment
0.0                    Consumer            89420
1.0                    Corporate           52528
2.0                    Home Office         30817
```



```
Market Code  Market
2.0          LATAM           49309
1.0          Europe          48090
3.0          Pacific Asia    39585
4.0          USCA            24627
0.0          Africa          11154
dtype: int64
```

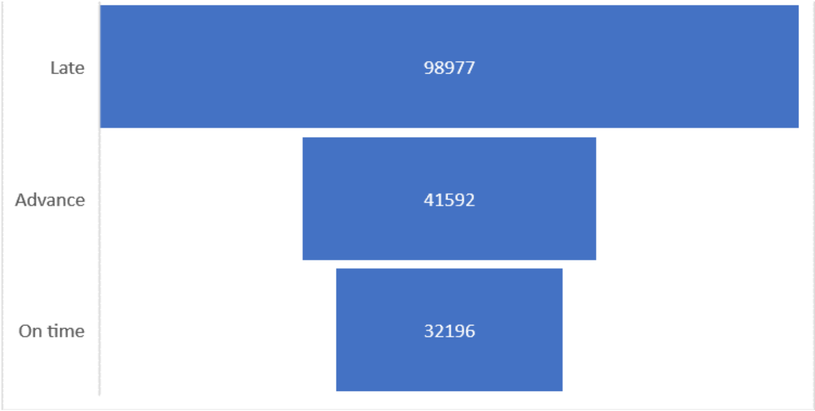


Figure 9: Orders by Delivery Status

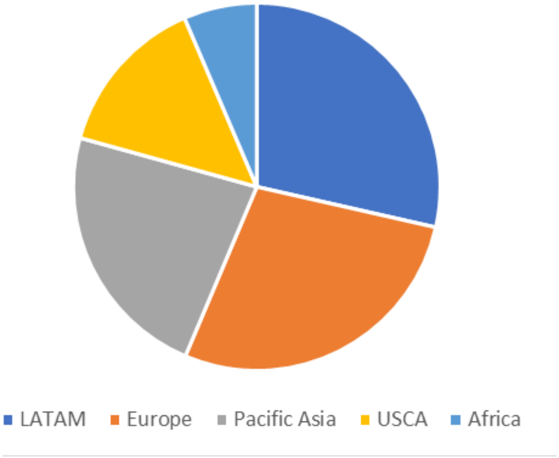


Figure 10: Orders by Market

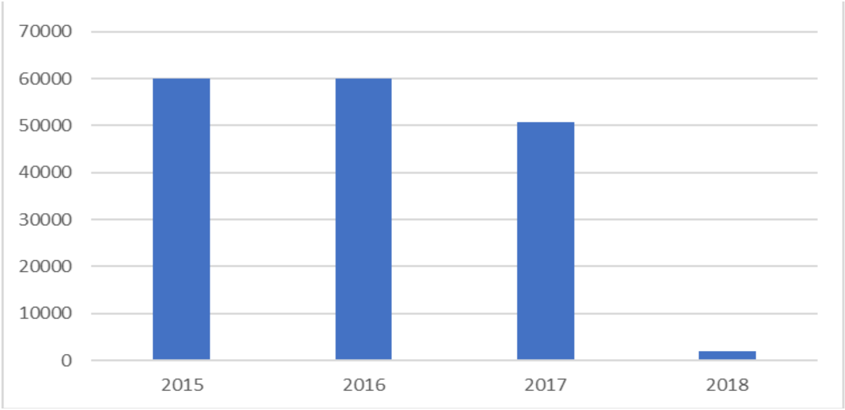


Figure 11: Orders by Year

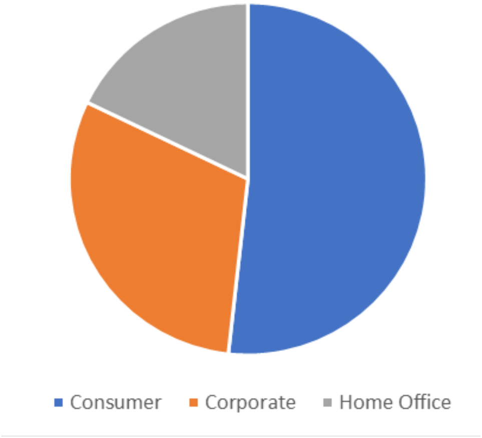


Figure 12:Orders by Customer Segment

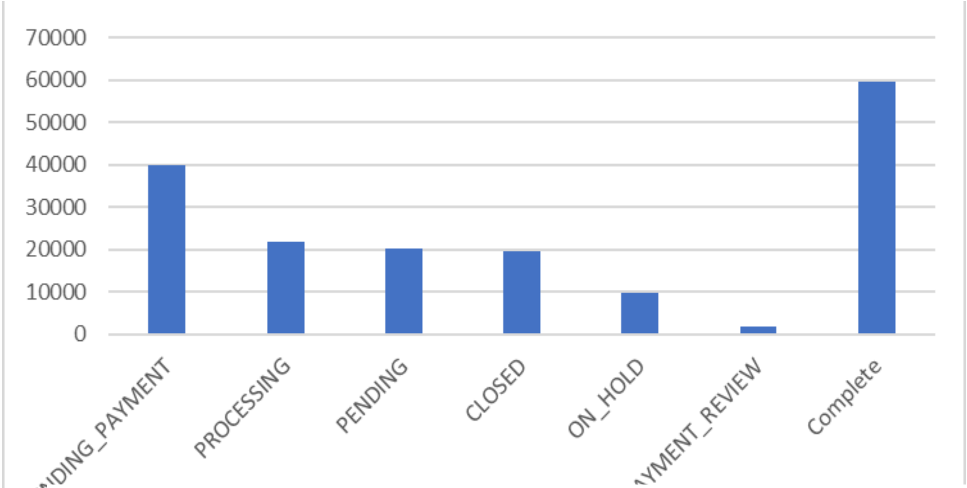


Figure 13: Orders by Status

We also used Pandas to generate correlation matrices (Figures 9-10) that helped visualize the relationship between Late Delivery Risk and a variety of factors

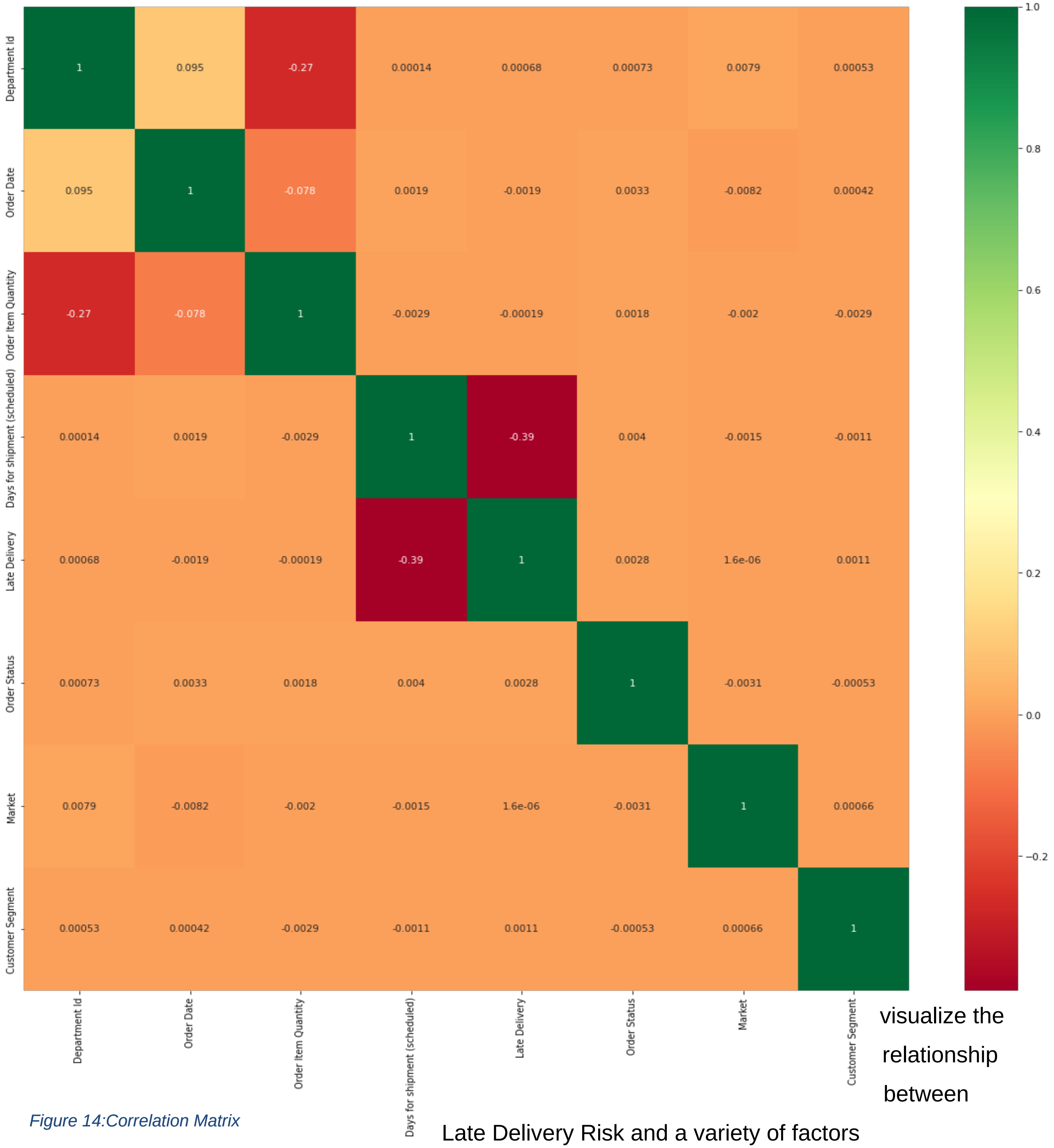


Figure 14:Correlation Matrix

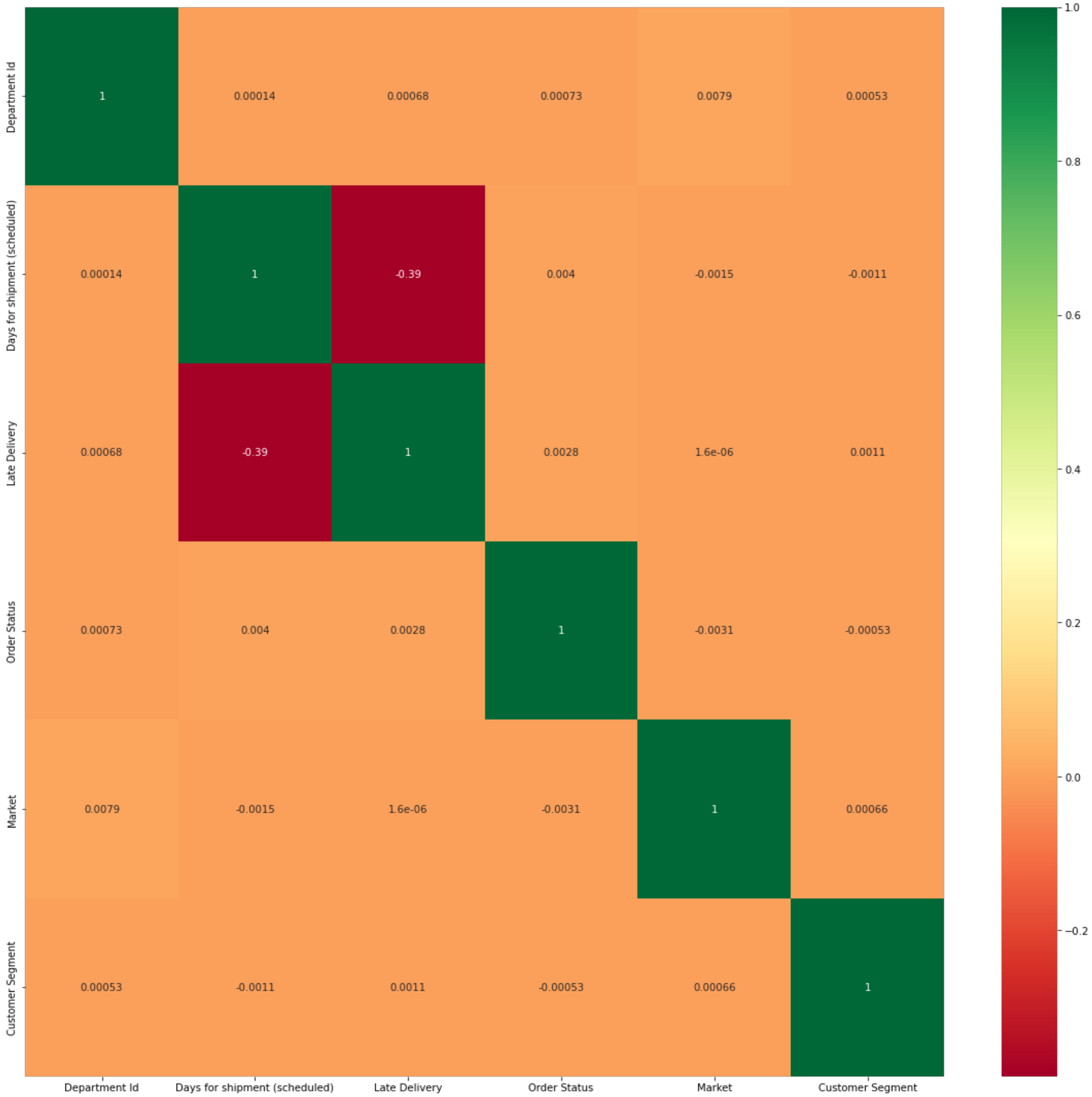


Figure 15: Correlation Matrix - Less features

The above matrices again reiterate that the scheduled ay for shipment is the only variable significantly correlated with Late Delivery Risk, but we would have already expected this.

Step 8: Prepare for ML

- Make Late Delivery Status the target label (Y), leaving the rest of the data as the X features
- Perform 80-20 train test split with Scikit
- Checked for any null data in the X and Y: none found

If we had not run out of time, we would like to build an SVM classifier to predict deliveries at risk of being late. However, we were unable to produce this.

Conclusion

Overall,57.2% of orders, regardless of any variable except the scheduled shipment days, were delayed, while about 24% of shipments were sent in advance and only 18.7& arrived on time. We would recommend the company shift resources from deliveries it knows will be delivered in advance and prioritize deliveries that are expected to be late. However, this requires an algorithm to predict which deliveries are likely to be late. The above analysis suggests this will be difficult as nothing other than scheduled shipping days appears to have a predictive power.

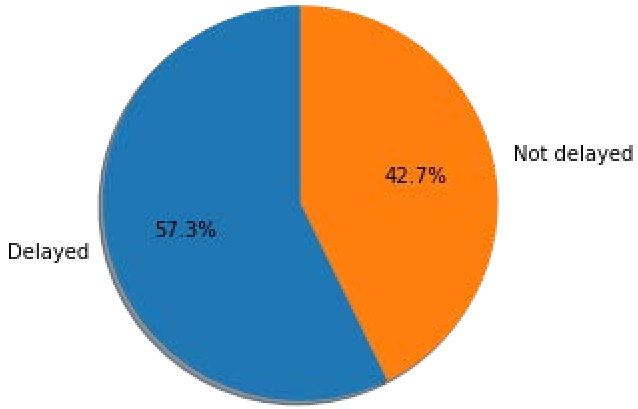
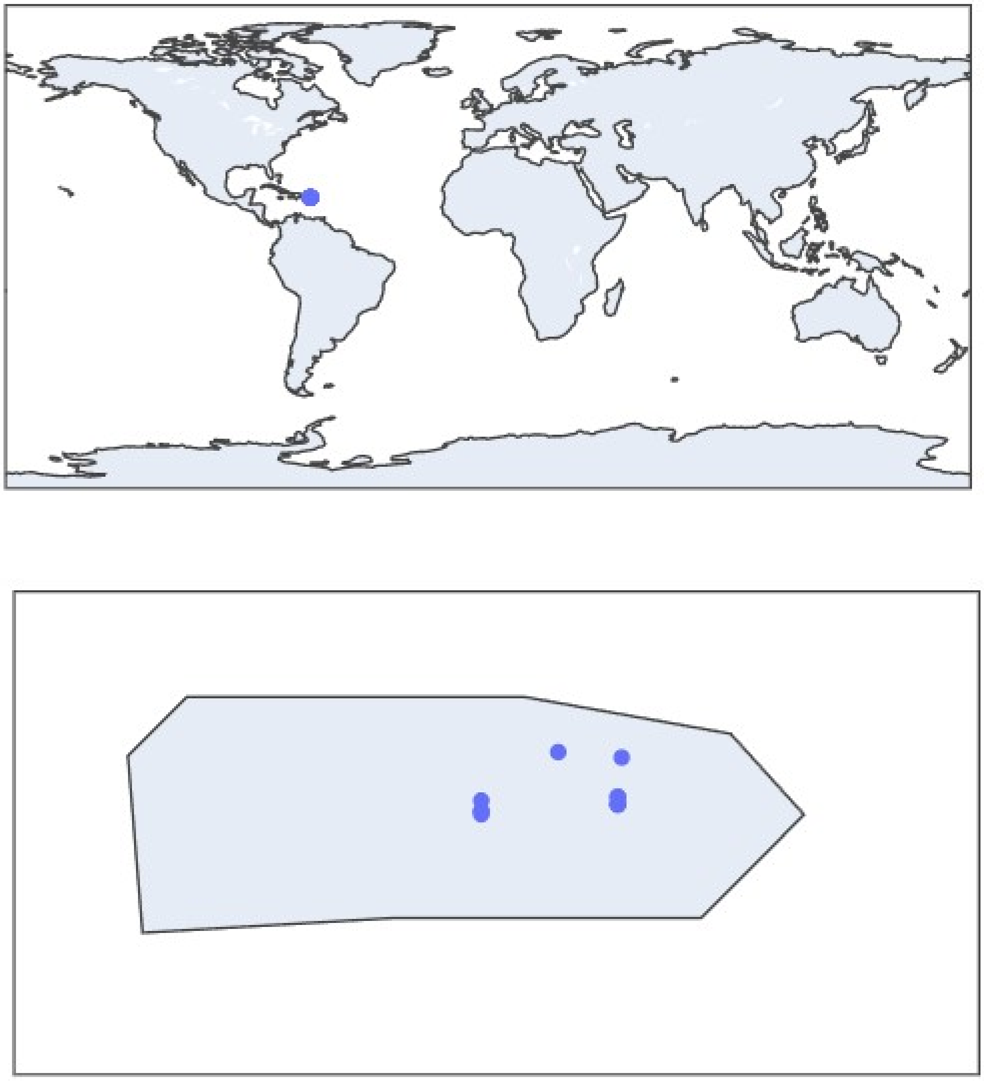


Figure 16: Delayed and Non Delayed Orders

Appendix

Figure 17: Location of Big Supply Co Stores



- We intended to colourmap this world map to denote average experienced shipping lengths per Market Region, but this could not be achieved in the time we had