Consent

We give consent for this to be used as a teaching resource.

Contents

Executive Summary

In a digital world, ecommerce has increasingly become more popular as well as customers' expectations. According to a survey 29% of shopper will discontinue their relationship with the retailer if given an inaccurate delivery time. Olist, an ecommerce platform, does not always predict an accurate delivery time. The average difference between the actual time of delivery and estimated time of delivery is 12 days.

This report will go through the data science process to reduce the error in estimating the delivery time. This is problem solving with data, getting the data, making the data fit for use, making the data confess and lastly storytelling with data. A more accurate delivery will be achieved by gaining a comprehensive understanding of the factors affecting delivery time.

After explorative data analysis, four features were considered significant income per capita, connection between cities, seasons and the lag time between the time of purchase and the time the courier receives the package. Given logistical intuition, distance was surprisingly not a significant factor that affects delivery time. Two models were analysed, Ordinary least squares and Random Forest using the four features. Both models found the connection between cities to be highly significant in predicting delivery time. Lag time increased accuracy slightly but Income and seasons had little to no effect.

In consideration to the findings of the analysis two main recommendations have been made. The first recommendation is allocating deliveries to third part logistic partners that are optimised for their local areas. Secondly to increase the connection between cities sellers should share their inventory. Olist will then be responsible for assigning the seller who has the highest connection between cities for the customers.

There are several limitations to this project. One of the most significant limitation was the lack of data for the route and distance each package took due to the use of external logistic companies. Secondly the size of the data and the transformation performed required heavy computation time. While accuracy of the predictions of this project is relatively high it can still be improved. Further studies in delivery network associated with traffic condition as well as exact customers & pickup locations would convey more information about planned or real-time routing process to which the delivery time is highly relevant.

Data Science Process

Problem Formulation

A sale is not complete until orders are shipped to the end customer and if deliveries are late it can be their last purchase (Victor, 2020). Of the 87% of online shoppers, survey 29% of shopper will discontinue their relationship with the retailer if given an inaccurate delivery time, even once (Edison, 2016). Price is less of an important factor than inaccurate delivery since two-third of them would rather pay more to get the goods delivered at their doorstep before the deadline. One way of building trust with a seller is ensuring orders are shipped within expected delivery date. More than half of the online shoppers believe accurate delivery would increase their trust with a brand. These shoppers have a fixed amount of days they are willing to wait for their orders. In a study by (Howen, 2014), it was found that 69% of consumers have lower probability of buying from the same seller again when their order is delivered for more than two days of the expected delivery date. Reasons for delays in delivery include, product customization, product handling (e.g. fragile items, high-value items). Infrastructure is another concern as delivery routes may always change and may not be the fastest. Unexpected and/or severe weather and/or traffic conditions can also be factors for orders to be delivered beyond the expected date.

Olist does not always predict an accurate delivery time. In the dataset provided by Olist, it was observed that the gap from between order purchase to estimated delivery, and order purchase to actual delivery to customer has a lot of variation as seen in **Figure 1.** This is calculated by getting the difference of actual delivery time from estimated delivery time in days.

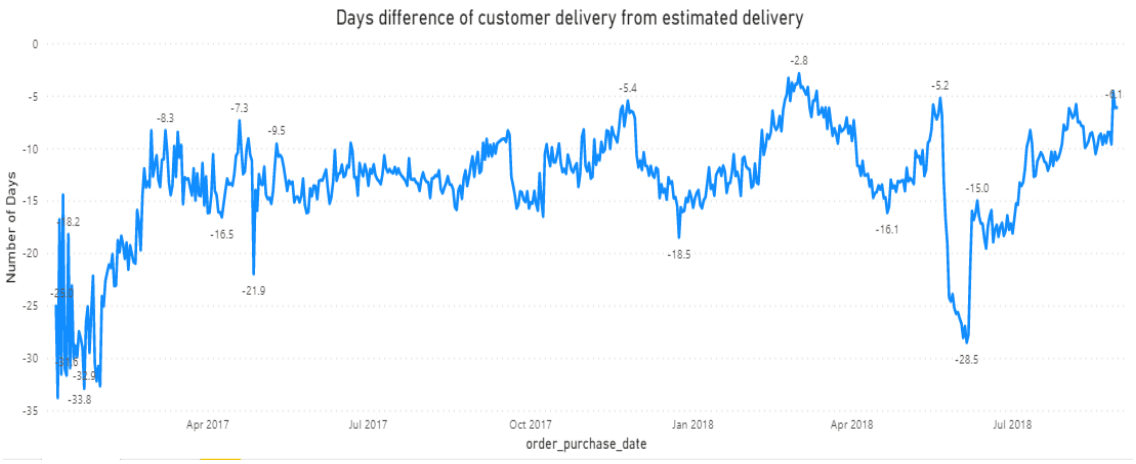

Figure 1 Days difference of customer delivery from estimated delivery

Getting the Data, I need

Olist Data

The data from Olist simply comes from Kaggle (Olist & Sionek, 2018). The ecommerce process can simply be split into three parts, first is the order creation, second is fulfillment, and third is delivery. Since the goal of the project is to be able to understand the factors that affect delivery time this is affected by the whole fulfillment process.

The main data required are order details, payment method, customer information (end point of the delivery process), seller information (starting point of the delivery process). Order details required are order placement date, fulfillment date (when the order is ready for delivery), hand-over to courier date, delivery date, order weight, order dimensions, product information (such as product type/categories since there could be a difference in the length of time required for handling various products), and order value (since there is potentially a different fulfillment and/or delivery process for high value goods). Payment method is also required since there will be variation in the processing of payments when it is made online, over the counter, cash on delivery (COD), and credit card on delivery (CCOD). Lastly, geolocation is also required to correctly pinpoint the customers and sellers (this was also made available in the Olist dataset).

Income Data

To get income data three different datasets have been collected and joined. The datasets contain the income per capita (RDPC), co-ordinates of the municipals and the boundaries of the municipals.

Data for the Income per capita collected from the Brazilian National Census was extracted through Kaggle which was sourced from the United Nations Development Programme (UNDP). The UNDP is a United Nations organisation focused on the development of underdeveloped and developing countries, the data covers the years of 1991, 2000 and 2010. This project only used the most recent data in 2010 only. The data came in the form of two csv files, one for the raw data and the other is meta data.

The census is conducted by the Brazilian government every 10 years. Every household in the population of Brazil is required to participate. Households either complete a questionnaire and mail it to a central location or an official visit the household and collect their information. A total of 232 indictors on the housing conditions, demographic, social and economic conditions of each county is collected.

The second dataset joined the geographic co-ordinate of each county and was collected from the World Bank, a global financial institution focused on the development of underdeveloped countries. This dataset in the form of a csv file contains the co-ordinates of each municipalities in Brazil. It was last updated in June 2017.

Lastly the boundaries of every municipal in Brazil was collected from the Humanitarian Data Exchange which was sourced from Database of Global Administrative Areas (GADM), a service provider for maps and spatial data. The data comes in the form of a shapefile with the boundaries as polygons. The boundary data used was updated in November 2019.

Making the data fit for use

The data gathered for this project is not readily usable. And each has to go through data cleaning processes which will be shown in detail in this section.

Olist Dataset

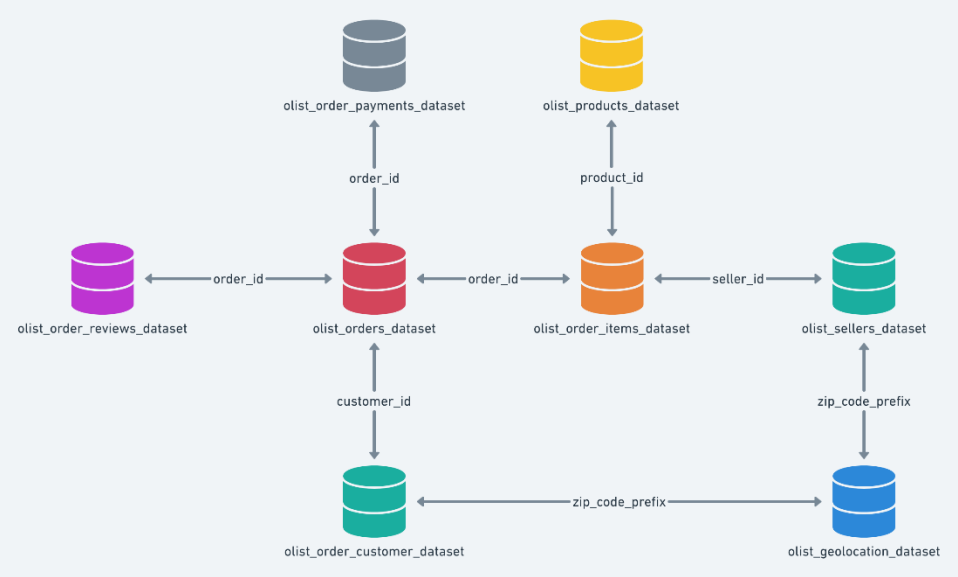

Figure 2 Olist E-Commerce Dataset Schema

The main dataset was gathered from the published dataset of Olist in Kaggle (Olist & Sionek, 2018). **Figure 2** above is the data schema and shows how each table are connected based on join keys.

There are 9 tables from the dataset namely, olist customer dataset, olist geolocation dataset, olist order items dataset, olist order payments dataset, olist order reviews dataset, olist orders dataset, olist products dataset, olist seller's dataset, and product category name translation. **Table 1** shows a summary of the contents of each table.

Table 1 Olist Database Summary

Table Name	Description	Number of Records	Primary Key(s)
olist customer dataset	Contains relevant customer information such as location and customer id	99,441	customer id, customer unique id
olist geolocation dataset	Provides the longitude and latitude for each postcode	1,000,163	geolocation zip code prefix

olist order items dataset	The item level breakdown of each order containing information about the product, seller, expected shipment date, product value, and the weighted shipping cost	112,650	order id, order item id, product id, seller id
olist order payments dataset	Contains details of customer payment on a per payment method level	103,886	order id
olist order reviews dataset	Contains customer review details	100,000	order id
olist orders dataset	The main table of the dataset that contains relevant order information, customer information, order status and timestamps	99,441	order id, customer id
olist products dataset	Detailed information of products including category, name, description, images, and dimensions	32,951	product id
olist sellers dataset	Contains seller location	3,095	seller id
product category name translation	Contains the translation of product categories to English	72	product category name

Since there is minimal complexity in the tables, it was opted to not use SQL tools and only Excel was used to join them. The figure below shows the joining process of each table including the join method and joining keys used. All the unique columns are adapted in the joined table.

Table 2 Joining Tables

From Table	Join Table	Join Statement	Join Key	Output Name
olist order items dataset	olist orders dataset	Left Join	order_id	Updated Table
Updated Table	olist customer dataset	Inner Join	customer_id	Updated Table
Updated Table	olist order payments dataset	Inner Join	order_id	Updated Table
Updated Table	olist order reviews dataset	Inner Join	order_id	Updated Table
Updated Table	olist products dataset	Inner Join	product_id	Updated Table
Updated Table	product category name translation	Inner Join	product_category_name	Updated Table
Updated Table	olist seller's dataset	Inner Join	seller_id	Updated Table
Updated Table	olist geolocation dataset	Inner Join	geolocation_zip_code_prefix	Updated Table

Joining Income data

Three different joins were necessary to join the Income per capita, municipal co-ordinates and boundaries and lastly the data from Olist. Exact joins were made to join RDPC and municipal co-ordinates and the other three joins were made by the closest and intersection of the co-ordinates. See **Figure 3** for specific details.

To join RDPC and the co-ordinates RDPC from the census data needed to be aggregated first as the project's interest was on the municipal level but the census contained data on the county level. Therefore, an average of the income per capita of the counties in the municipals was used. Each municipal is denoted by a seven-digit code in the census which was used to join the geographic co-ordinate from the world bank which shared the seven-digit municipal code. RDPC was joined with the Olist data by taking the RDPC closest to the customer's co-ordinates.

To join RDPC and boundary data for use in the explorative data analysis the RDPC was merged with the boundary data by joining the co-ordinates that lied in the boundary. There was a 1-1 matching, where every municipal corresponded to one boundary polygon.

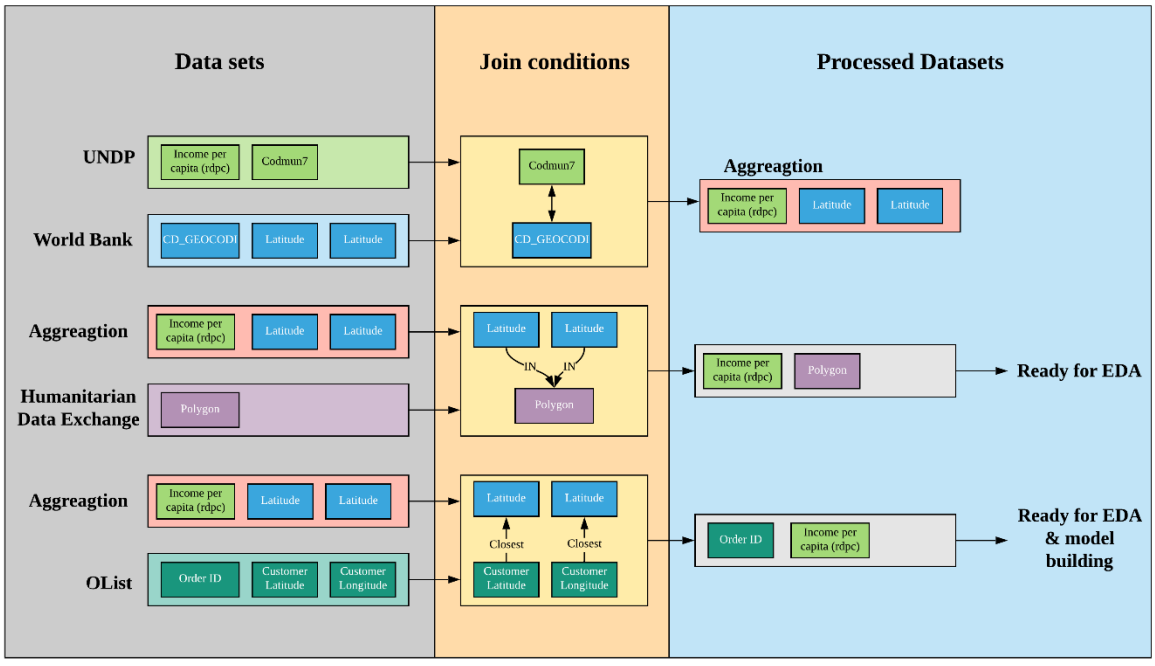

Figure 3 – Schema to join income data with Olist

Since the project is focused on delivery time, one of the challenges after generating the database is having the rows on a product level. Instead, the data needs to be on an order-level. The group wanted to see how all the data is related, this is the reason why the main database used for joining the tables is *olist_order_items_dataset*. It was found out that there was no need to have the observations on a product level, thus, everything is grouped by *order_id*. Given this, aggregations were performed on some columns.

Table 3 Column Aggregations

Column Name	**Aggregation**
order_item_id	Max of Values
payment_sequential	Max of Values
distance	Max of Values
payment_value	Sum of Values
product_weight_g	Sum of Values
product_length_cm	Sum of Values
product_height_cm	Sum of Values
product_width_cm	Sum of Values
freight_value	Sum of Values
price	Sum of Values
payment_installments	Sum of Values
product_category_name_english(multiple catagories in order)	Category with Max *total_delivery_time*
product_category_name_english (single category in order)	Unique Value

Variable Conversions

The following variables were converted into appropriate data types to ensure that python is able to read them properly

Table 4 Variable Conversions

Columns Name (Variable)	**Conversion**
payment_type	Categorical
seller_id	Categorical
customer_zip_code_prefix	Categorical
customer_city	Categorical
customer_state	Categorical
seller_zip_code_prefix	Categorical
seller_city	Categorical
seller_state	Categorical
product_category_name	Categorical
shipping_limit_date	Datetime
order_purchase_timestamp	Datetime
order_approved_at	Datetime
order_delivered_carrier_date	Datetime
order_delivered_customer_date	Datetime
order_estimated_delivery_date	Datetime

Calculated Columns

The data gathered were not sufficient to undertake a concrete study on factors that affect delivery time. Additional columns were calculated to be able to help understand what affects delivery time. A way to determine whether delivery is within estimated date or not is calculated using *promise_date*. To identify the total time required to prepare an order from customer purchase until it is handed over to a third-party logistics (3PL) partner *actual_lag_time*, and in order to determine the estimated time for this *estimate_lag_time* were calculated. To determine the effect of *seasons*, the months from *order_purchase_timestamp* were categorized into wet and dry. The skyway distance between customer and seller were used to calculate *distance*. The *product_volume* accounts for the size of an order in terms of dimensions. The *connection_between_*states, and *connection_between_*cities were derived to determine the degree of connection between a pair of states, and a pair of cities respectively. The *freight_ratio* was used to incorporate freight value and product volume. To account for the historical delivery time of each seller, *seller_delivery_PrevMean* is introduced. The time that consists only of the delivery component of the order delivery process is represented by *actual_delivery_time*. The response variable, which accounts for the complete order delivery process is represented by *total_delivery_time.* The calculations are shown in **Table 11** in **Appendix D**

Data Exploration

Distance Analysis

The first and the most trivial factor that could possibly affect the total delivery time is Distance since the speculation would be the further distance between a customer and a seller, the longer delivery time will take. When considering Distance, an order of a customer might have multiple products bought from multiple sellers and all the products are delivered to the customer at the same time. Thus, the distance between the customer and the sellers is referring to the distance between the customer to the furthest seller, which corresponds to the aggregation performed on the feature *distance* in **Table 3**. Additionally, the latitudes and longitudes of customers and/or sellers, whose prefix zip code, state, and city are the same, should have the same latitude and longitude as a result of anonymizing individual's location by Olist. Note that, the calculation of the distance could be referred to **Table 11** and the distance used this analysis is nothing but the skyway distance. Finally, while examining the relationship between distance and total delivery time, which is illustrated by **Figure 4**, the distance seems to be independent from the total delivery time due to the dispersion of total delivery time is likely to be equally along the distance axis and even with some points having distance above 3000 km, their total delivery time stays below 1000 hours.

Figure 4

In a nutshell, Distance is not a primary factor of the total delivery time in this context due to insufficient data for estimating the actual distance for every pair of customer and seller as well as the method for evaluating distance is not consistent with the exact distance, which leads to inaccuracy estimation.

Seasonality Analysis

One of the major reasons why Seasonality is taken into consideration is that this country has tropical climate and majority parts of Brazil has only 2 seasons (Climate in Brazil, 2020), which are wet and dry season. More importantly, those countries, whose climate is tropical, are usually have many rainy days during the wet season and raining weather is also a typical culprit of the traffic congestion and the driving speed in general is much slower compared to sunny weather.

The most straightforward approach to examine whether Seasonality affects the delivery time is to look at the distribution of total delivery time with regards to both seasons in 2017 since majority of orders in this dataset is in 2017. The distributions of total delivery time are illustrated as follows:

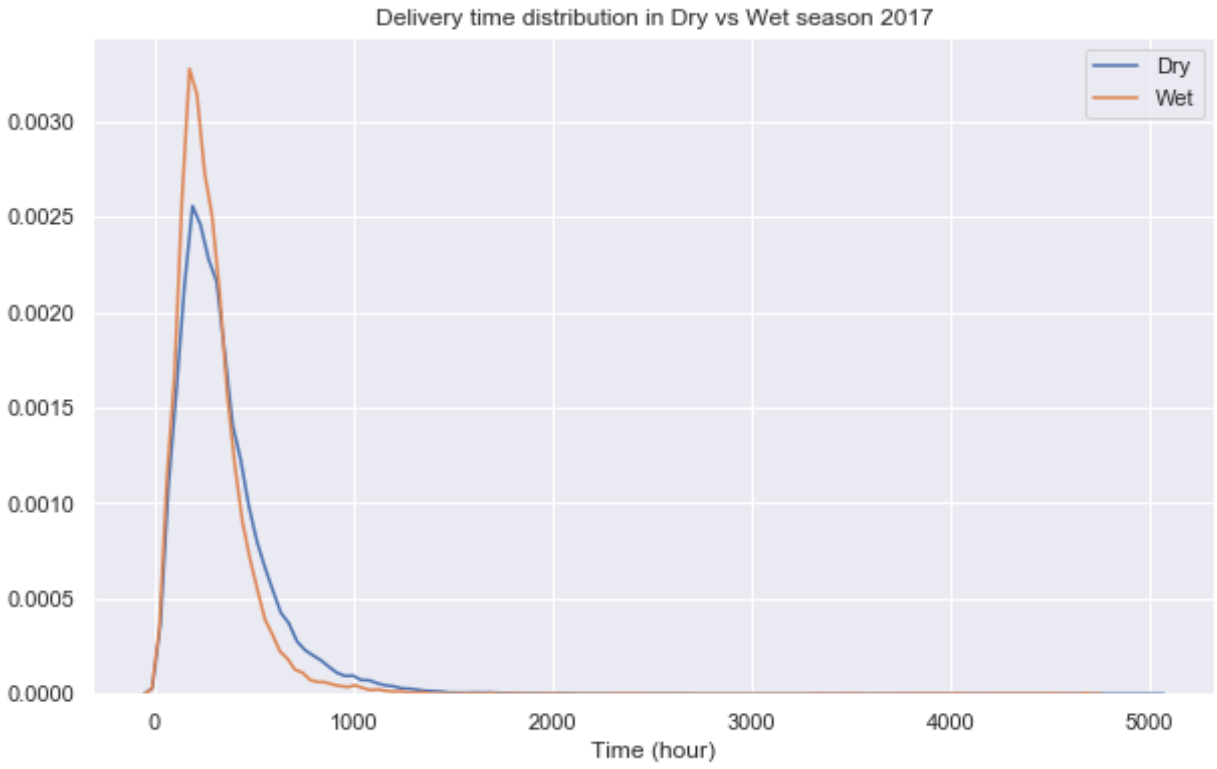

Figure 5

By having a quick glance at the **Figure 5**, it seems that those distributions are overlapping each other, in which one could conclude that those distributions could be from the same underlying distribution. However, to make sure that the total delivery time in both seasons is identical, i.e., the distribution of total delivery time in both seasons is analogous, (Permutation Tests, 2020) have been carried out to give a more persuasive evidence.

In this permutation test setting, the test statistic T is defined as:

$$T = \bar{x} - \bar{y}$$

where $\bar{x}$ and $\bar{y}$ is the sample mean of the total delivery time in Dry and Wet seasons, respectively.

What is more, the null hypothesis H_0 is that both Dry & Wet season have the same distribution w.r.t total delivery time and H_a is for the alternative case.

Finally, the *p-value* from the permutation test is 0.0 and **Figure 6** is the visualization result of the generated test statistics (blue) vs the observed test statistic (red). Hence, this is a strong evidence to show that the total delivery time distribution of Dry season is strongly distinct from that of Wet season. For more information, the procedure to conduct this test as well as the result is fully described in the section **Seasonality Analysis**, in the notebook **Data-Exploration.ipynb**.

Figure 6

To sum up, Seasonality is also a significant factor but it is not the main contributor of total delivery time since it could hardly be used to estimate the total delivery time, the further discussion is in Average Mean Seller Time.

Lag Time

After carefully considering and applying design thinking process to understand which factor is likely to affect the total delivery time, *actual_lag time* is derived based on given features in the dataset. Regarding the definition of lag time, it is the time interval of when customers successfully ordered a product (i.e. *order_purchase_timestamp* variable) and when sellers dispatch the product to a logistic partner (i.e. *order_delivered_carrier_date* variable).

Note that the lag time used in this section is referring to the variable *actual_lag_time* described in **Table 11**.

The following scatterplot shows how the total delivery time varies when the lag time changes.

Figure 7

Note that, the red line in **Figure 7** is the boundary to discriminate those orders, whose lag time is greater than 50% total delivery time, should stay below the red line and vice versa. Moreover, the number of orders having lag time more than 50% of the total delivery time is 16896 out of 77750 in total, which accounts for 21.7% of total number of orders in this dataset. Thus, this factor could be considered as a cause for the delay of total delivery time. However, the rest 78.3% orders have no issue with the lag time and. That means, lag time is not a root-cause factor for the delivery time.

Product Category Analysis

The next potential factor that could affect the total delivery is Product Category and the reason for considering this factor is that different product category would have different characteristics, which might affect the delivery time. For example, the total delivery time of office furniture products usually takes longer time compared to that of computer products and this could be due to the size of products, number of items per order as well as office furniture products are typically not ready-made (e.g. after a customer buy a furniture product, that product has to be assembled and sometimes if the materials for that product are not sufficient, sellers have to import those before they can manufacture it). To sum up, **Figure 8** shows that office furniture products have higher total delivery time compared to the rest and the lag time is used in this graph as an alternative way to explain for the manufacturing process time, which accounts for almost 50% total delivery time with regards to office furniture products.

Having said that, **Figure 8** is an only single value representing for total delivery time w.r.t product category in general, in which the variability for total delivery time for every product category has not been considered. Especially, if the variability (e.g. standard deviation or variance) of the total delivery time of a particular product category is high and the presence of outliers is significant, the mean/average total delivery time of a product category can now be a common representation for the majority of delivery time regarding that product category. **Figure 9** below shows the variability of total delivery for each product category.

Based on the visual result given in **Figure 9**, most product categories have very high variability in terms of total delivery time and many outliers. Therefore, it is intractable to recognize any pattern or relationship between product category and total delivery time.

In summary, product category seems not to be qualified as a factor of total delivery time.

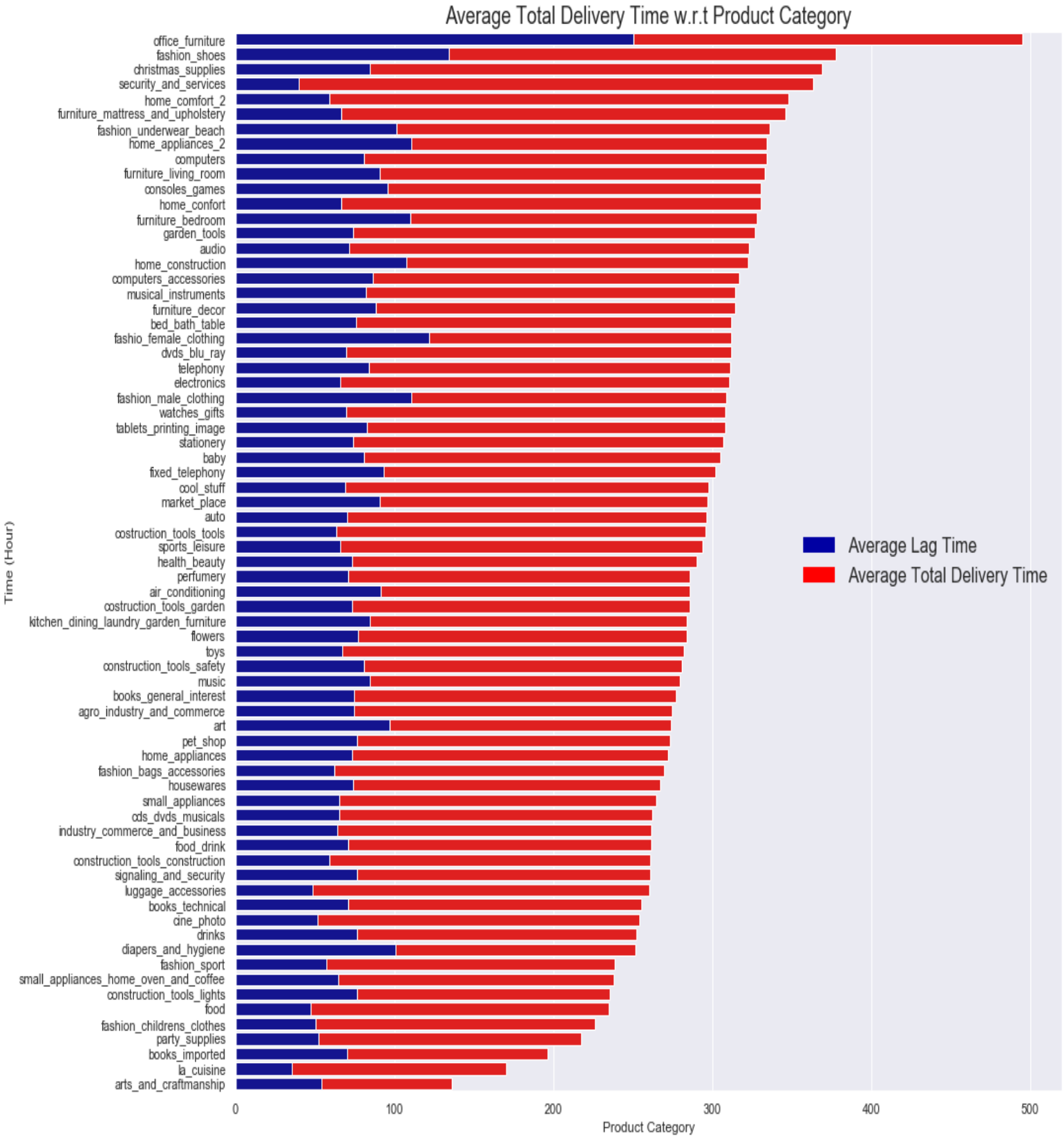
Average Total Delivery Time w.r.t Product Category
office_furniture
fashion_shoes
christmas_supplies
security_and_services
home_comfort_2
furniture_mattress_and_upholstery
fashion_underwear_beach
home_appliances_2
computers
furniture_living_room
consoles_games
home_confort
furniture_bedroom
garden_tools
audio
home_construction
computers_accessories
musical_instruments
furniture_decor
bed_bath_table
fashio_female_clothing
dvds_blu_ray
telephony
electronics
fashion_male_clothing
watches_gifts
tablets_printing_image
stationery
baby
fixed_telephony
cool_stuff
market_place
auto
costruction_tools_tools
sports_leisure
health_beauty
perfumery
air_conditioning
costruction_tools_garden
kitchen_dining_laundry_garden_furniture
flowers
toys
construction_tools_safety
music
books_general_interest
agro_industry_and_commerce
art
pet_shop
home_appliances
fashion_bags_accessories
housewares
small_appliances
cds_dvds_musicals
industry_commerce_and_business
food_drink
construction_tools_construction
signaling_and_security
luggage_accessories
books_technical
cine_photo
drinks
diapers_and_hygiene
fashion_sport
small_appliances_home_oven_and_coffee
construction_tools_lights
food
fashion_childrens_clothes
party_supplies
books_imported
la_cuisine
arts_and_craftmanship
Time (Hour)
0
100
200
300
400
500
Product Category
Average Lag Time
Average Total Delivery Time

Figure 8

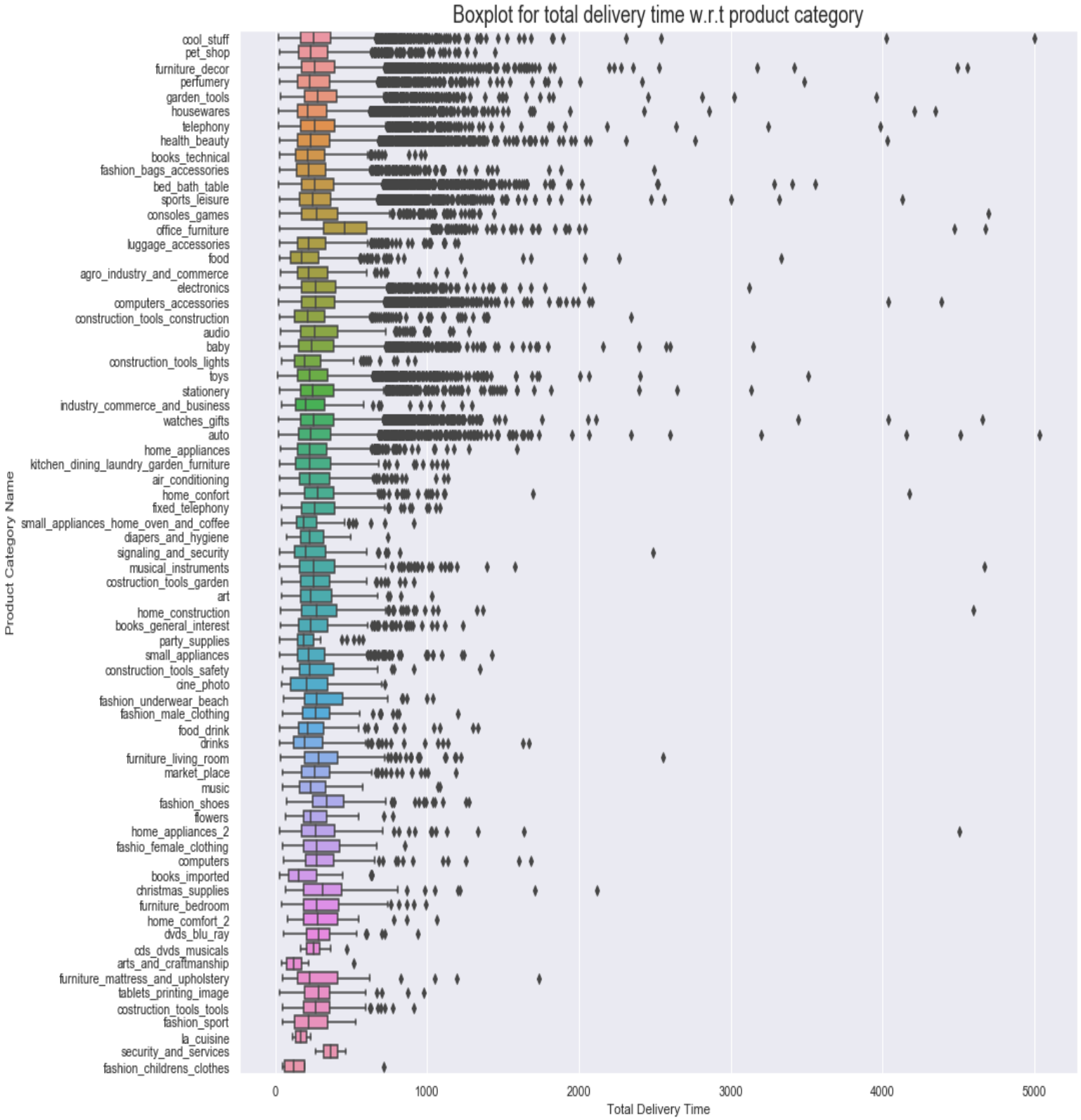
Boxplot for total delivery time w.r.t product category
Product Category Name
cool_stuff
pet_shop
furniture_decor
perfumery
garden_tools
housewares
telephony
health_beauty
books_technical
fashion_bags_accessories
bed_bath_table
sports_leisure
consoles_games
office_furniture
luggage_accessories
food
agro_industry_and_commerce
electronics
computers_accessories
construction_tools_construction
audio
baby
construction_tools_lights
toys
stationery
industry_commerce_and_business
watches_gifts
auto
home_appliances
kitchen_dining_laundry_garden_furniture
air_conditioning
home_confort
fixed_telephony
small_appliances_home_oven_and_coffee
diapers_and_hygiene
signaling_and_security
musical_instruments
costruction_tools_garden
art
home_construction
books_general_interest
party_supplies
small_appliances
construction_tools_safety
cine_photo
fashion_underwear_beach
fashion_male_clothing
food_drink
drinks
furniture_living_room
market_place
music
fashion_shoes
flowers
home_appliances_2
fashio_female_clothing
computers
books_imported
christmas_supplies
furniture_bedroom
home_comfort_2
dvds_blu_ray
cds_dvds_musicals
arts_and_craftmanship
furniture_mattress_and_upholstery
tablets_printing_image
costruction_tools_tools
fashion_sport
la_cuisine
security_and_services
fashion_childrens_clothes
0
1000
2000
3000
4000
5000
Total Delivery Time

Figure 9

Customers State and Sellers State Analysis

The following factor, which is a more general factor after considering all concrete factors given and derived from the dataset (i.e. seasonality, product category, lag time, etc.), is the average delivery time from one state to another state. Specifically, this project aims to consider the average delivery time from a state where a customer is to another state where a seller is. The following heatmap (**Figure 10**) shows the average delivery time between customers state and sellers state.

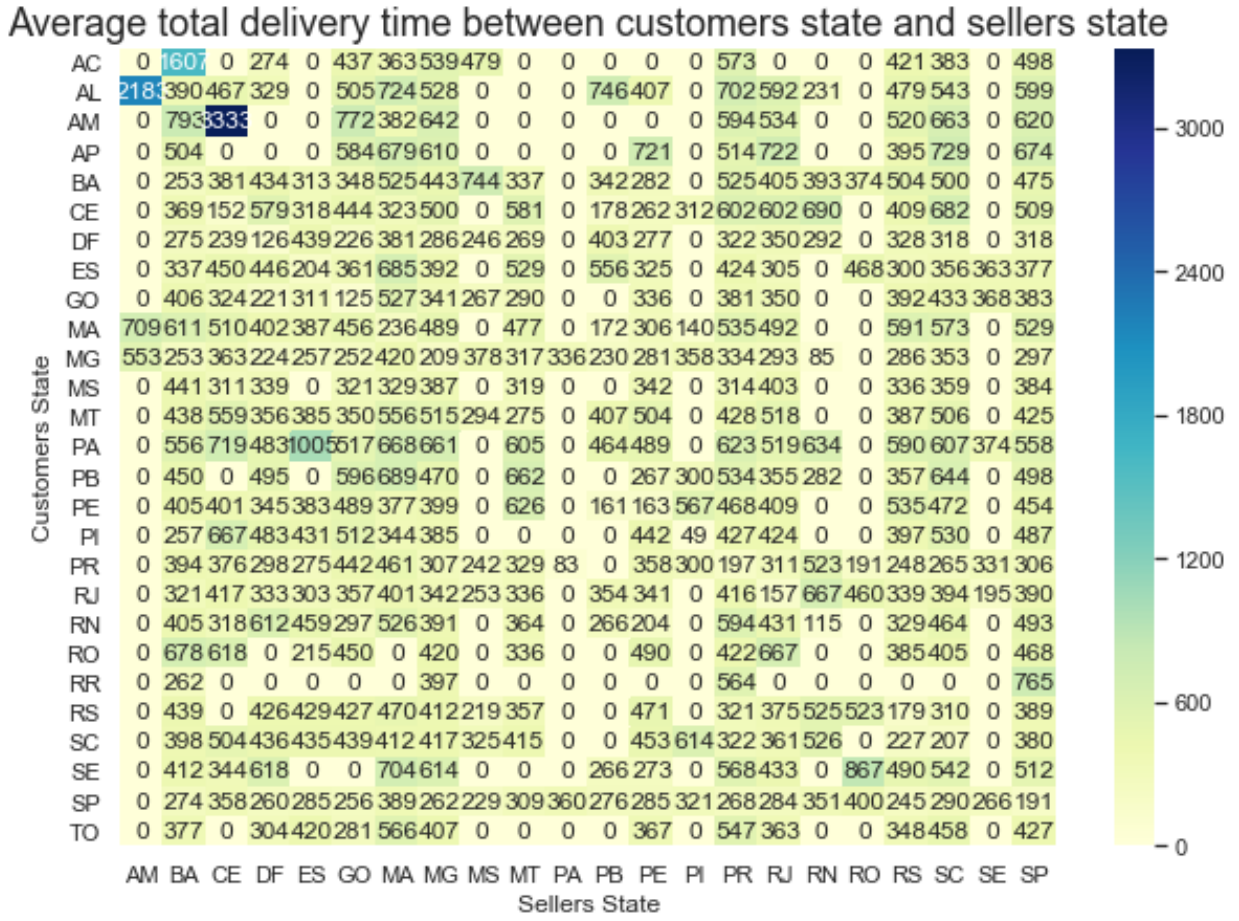

Figure 10

Note that, there are some pairs of states having 0 values in terms of average total delivery time, which states for no transaction has seen between consumers and suppliers between those states.

Having a quick look at **Figure 10**, it seems that there is not much difference in average delivery time between states because the colors for most cells are likely to be the same and only 3 cells at the top left corner, whose associated states are (AM, AL), (BA, AC), (CE, AM), could be treated as the outliers. Thus, the visual result from **Figure 10** could be misinterpreted due to the presence of the outliers could affect the pigmentation, which gives rise to hard recognition in terms of contrast of delivery time for every pair of states. Given that there are 3 outliers in average delivery time between states, **Figure 11** illustrates the result after 3 mentioning outliers are removed.

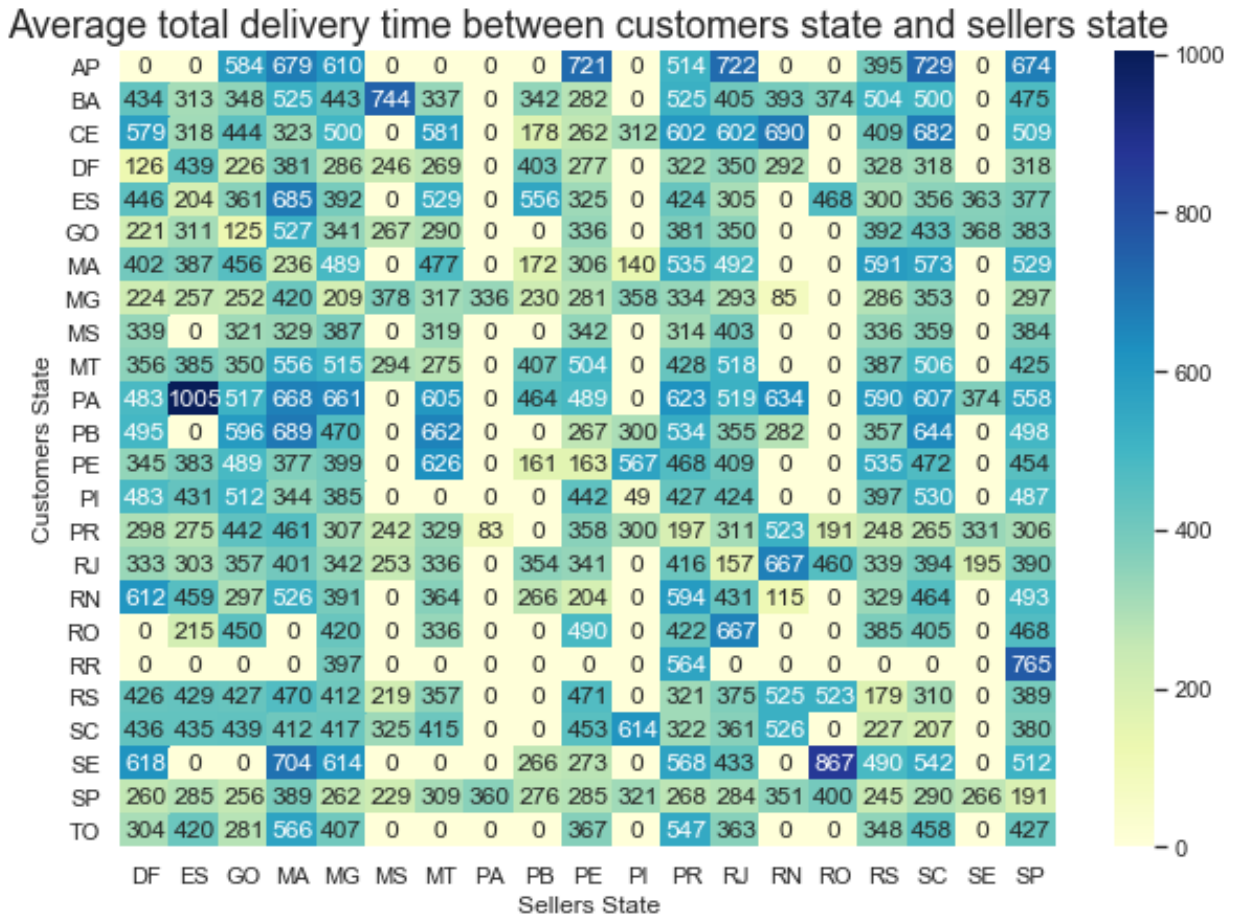

Figure 11

Regarding the result of **Figure 11**, the average delivery time between states are highly contrast, which means that different pair of states would result in different delivery time. Hence, this could be also a potential factor of total delivery time. However, one further step needs to be done before this could be concluded as a factor of total delivery time is by mapping each pair of buyers' state and sellers' state by a value called *connection_between_states* (the encoding process for *connection_between_states* is described in **Table 11**) and validating if there is any relationship between total delivery time and the feature *connection_between_states* via the scatterplot, which is shown in **Figure 12**.

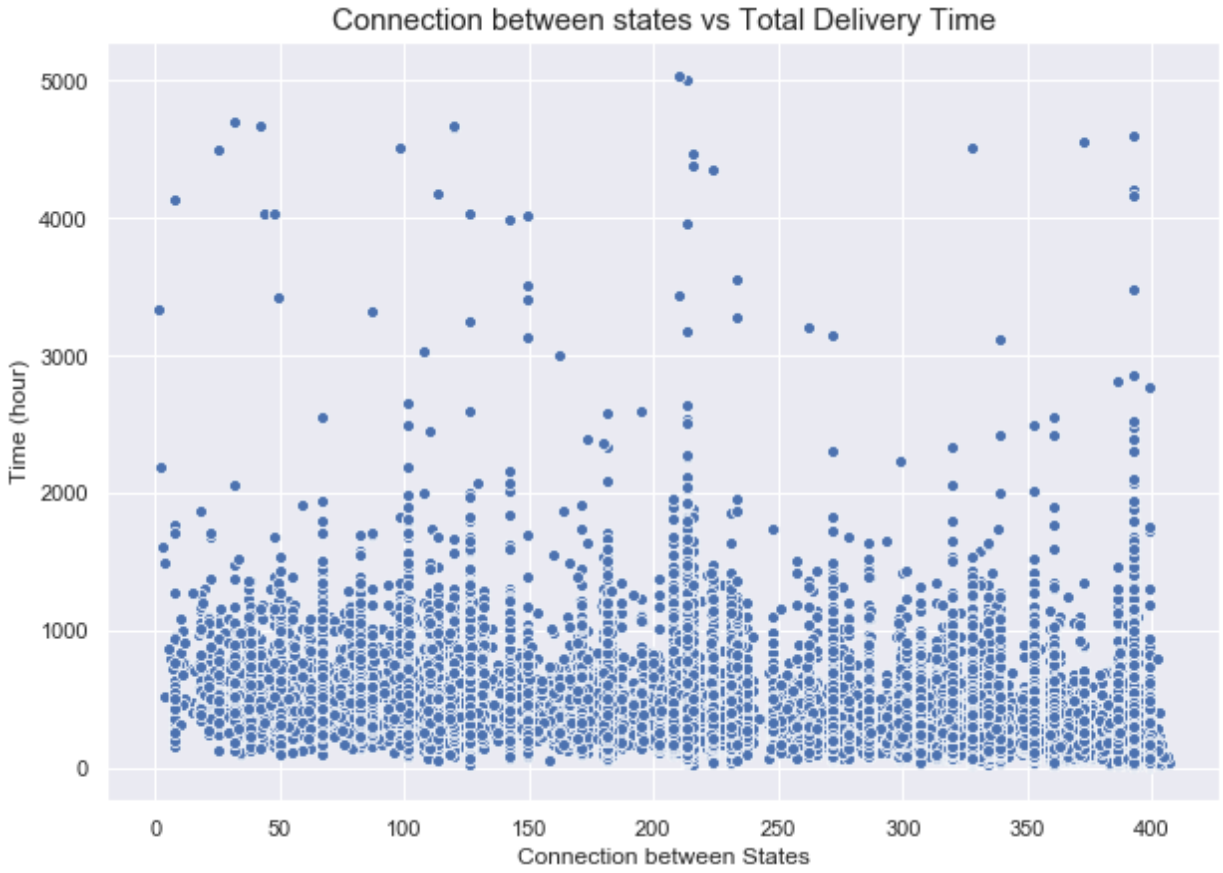

Figure 12

The result in **Figure 12** does not convey much information about the relationship between total delivery time and the predictor *connection_between_states* since the dispersion of total delivery time is likely to be equal irrespective of *connection_between_states* values.

In a nutshell, even the average delivery time varies for different pair of states, there is not much to tell about the correlation between the mentioning predictor and the target variable. Consequently, pair of customers' state and sellers' state is not a factor of total delivery time.

Customers City & Sellers City Analysis

Similarly, to what has been mentioned in **Customers State and Sellers State Analysis**, further exploration is carried out in the city scope. More specifically, the average delivery time for each pair of customers city and seller's city is computed and the result is used to derive another feature called *connection_between_cities* (the process of calculating the variable *connection_between_cities* is introduced in **Table 11**).

This time, pairs between cities are assumed to be dependent on average total delivery time (i.e. different pair of cities will have different average total delivery similar to pairs of states) and one of the major reasons for that assumption fact that heatmap could not be utilized to check the contrast in average total delivery time between cities due to the large amount of pairs of cities which results in computationally expensive and checking the correlation between the derived predictor *connection_between_cities* and the total delivery time is sufficient to tell whether pair of consumers' city and suppliers' city is a factor of total delivery time. **Figure 13** is the scatterplot between total delivery time vs *connection_between_cities*.

Figure 13

According to the result in **Figure 13**, it seems that the outliers and high variation in delivery time w.r.t *connection_between_cities*'s value is large make. It, therefore, is hard to visualize the negative correlation between total delivery time and *connection_between_cities.*

As a result of influences from outliers and high variation in total delivery given in **Figure 13**, one way to make the data more well-behaved is by taking the logarithm of total delivery time in order to reduce the incremental rate of total delivery time (i.e. reduce the effect from outliers) and linearize the relationship

between predictors and target variable. Another worth mentioning point is that, the distribution of total delivery time for this dataset is right-skewed, which could be visually explained by **Figure 14**. Hence, based on all conditions given, log transformation is applicable for total delivery time in the opinion of (Log transformation of target variable, 2019).

Figure 14

After applying log transformation for the total delivery time, the result given by **Figure 15** demonstrates a better visualization for the relationship between total delivery time and *connection_between_cities*. To put it more concretely, this relationship could be explained like the higher degree in connection between cities, the lower total delivery time should be.

In summary, pair of cities can be treated as the crucial factor for total delivery time. Even though this is a general factor, due to time constraint and lack of domain knowledge, this factor now is assumed as the mediator for other underlying factors.

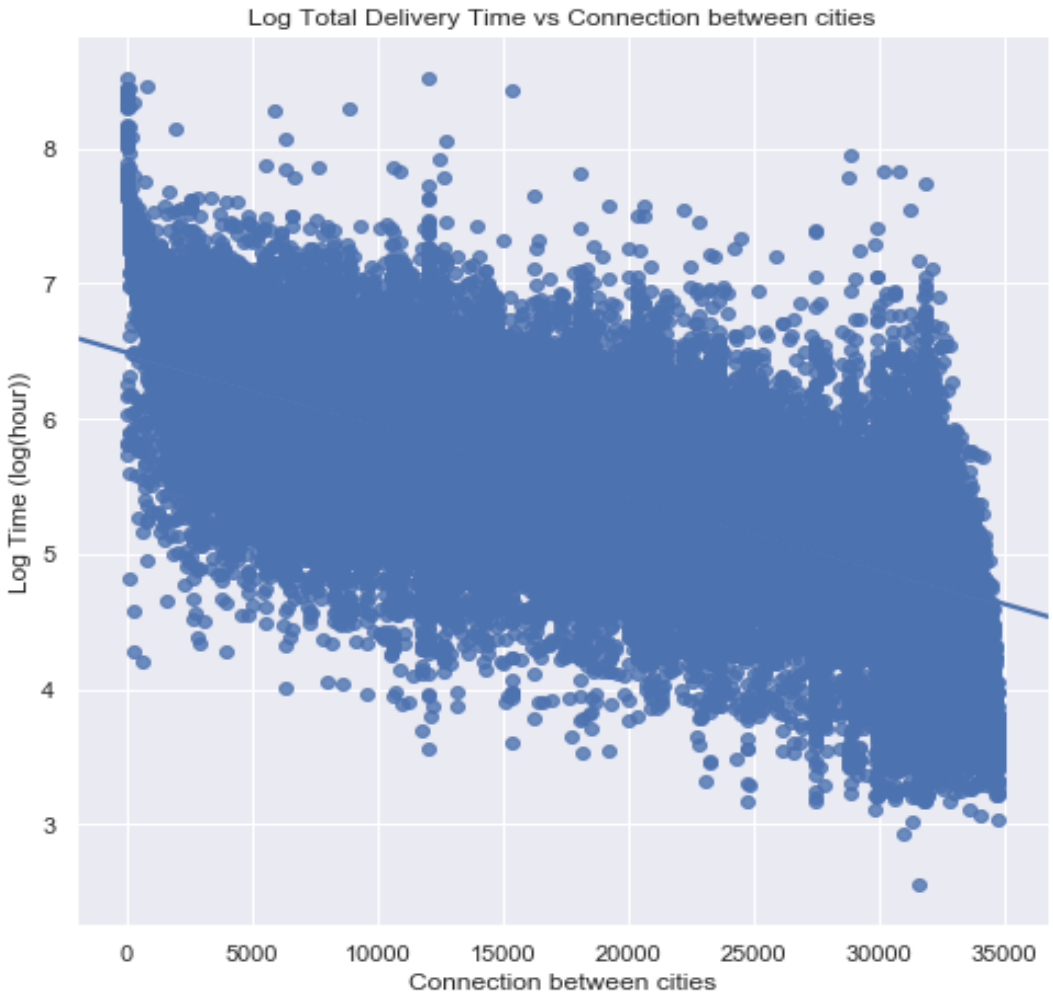
Log Total Delivery Time vs Connection between cities
Log Time (log(hour))
8
7
6
5
4
3
0
5000
10000
15000
20000
25000
30000
35000
Connection between cities

Figure 15

Payment Type

One of the potential reasons why deliveries are delayed is due to customer payment after orders are placed. There are 4 methods of payment; credit card, boleto, debit card, and voucher. Orders paid with credit card and debit card are not likely to have delays in the order preparation stage since all of these go through payment gateways that are directly linked to the marketplace/ecommerce website. However, the case is different for orders paid via boleto. This payment method is basically a bank slip which can be done via online banking, or offline through ATMs, banks, and other over the counter methods (Boleto, 2016). Based on article published by Tech in Brazil, over 20% of e-commerce transactions are transacted with Boleto Bancários during 2016 (General Practices for Accepting Boleto Bancário Payment in Brazil, 2016). Given this we explore how payment types affect delivery times.

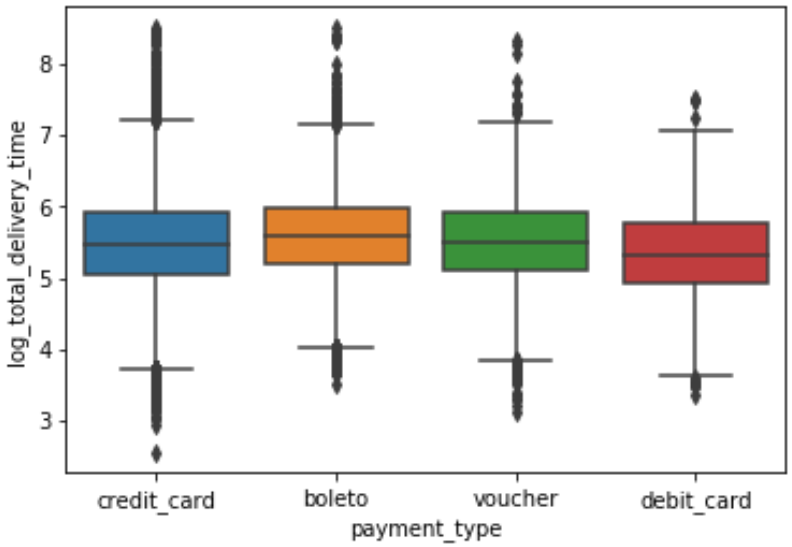

Figure 16 Boxplot Payment Type

The boxplot in **Figure 16** Boxplot Payment Type shows the *log of total_delivery_time* performance of each payment type. It does not show any significance difference among the payment types. Which means there is no significant gap between the processing time of payments done with Boleto as compared to credit card and others. In the dataset, payments comprise of 76% credit card and 19% boleto. Further exploration on these payment type's effect on log of total delivery time vs customer states were compared but still no significant difference in performance were found.

Product Volume and Product Weight

Fulfillment and shipping of bulky items is a challenge in the ecommerce industry. These items are more difficult to handle during fulfillment, especially during peak seasons (Scott & Lasher, 2018). On the delivery side, handling bulky items require proper scheduling. There are a lot of factors to consider when shipping bulky items, 1) stock availability to ensure that it is ready for pickup or fulfillment, 2) considerations for delivery locations that require special handling, 3) availability of the right personnel (e.g. items that require installation), 4) shipment schedule are different for bulky and small orders (BIG AND BULKY CHALLENGES IN AN OMNI-CHANNEL WORLD, 2016). In line with this knowledge, there is a potential that the delivery time is longer for bulkier orders.

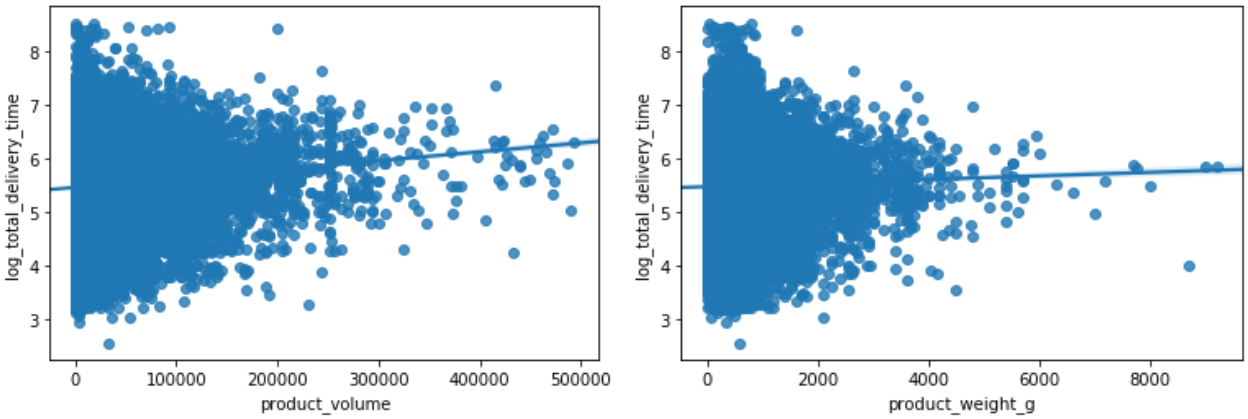

Figure 17 Scatterplots of Product Volume and Product Weight in Grams

In the analysis (**Figure 17)**, it is not clear how delivery time is affected by *product_volume* and *product_weight_g*. There is a very huge variation on delivery time for smaller items so not much information can be drawn from this. Also, it was initially assumed that bulkier items would have longer the delivery time, however, it does not seem to be the case here. It can be inferred that there seems to be no clear relationship for product weight and log of total delivery time. On the other hand, there seems to be an increasing relationship with product volume and log of total delivery time. However, that relationship is too weak to be considered in the model.

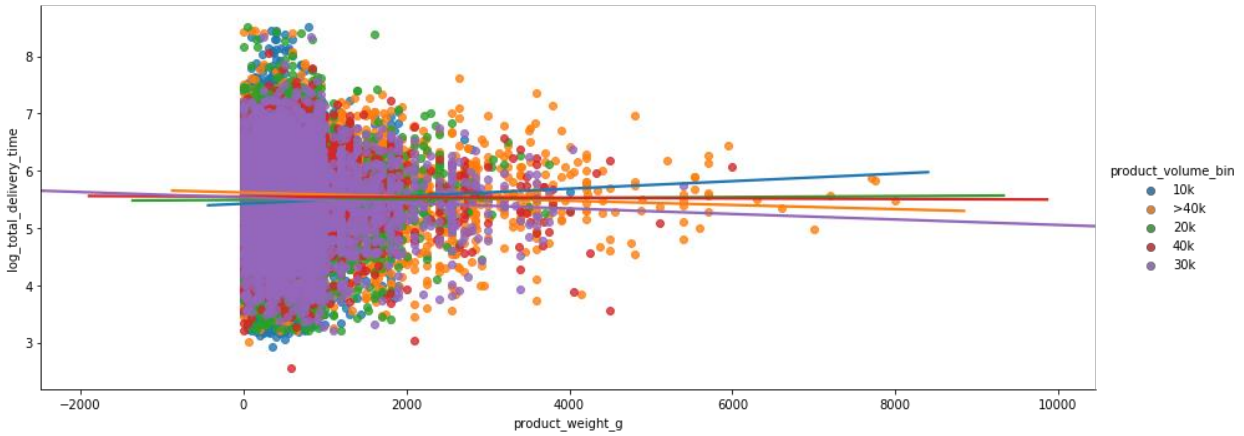

Figure 18 Interaction Effect of Product Volume and Product Weight in Grams

Additionally, the interaction of *product_volume* and *product_weight_g* (**Figure 18**) was investigated since these 2 features both describe the bulkiness of an order. A temporary variable was created to place product volume in bins, labelling them with the following conditions on *product_volume*; “10k” for observations <= 10,000, “20k” for observations <= 20,000 and >10,000, “30k” for observations <= 30,000 and >20,000, “40k” for observations <= 40,000 and >30,000, and “>40k” for observations >40,000. The interaction plot above also does not show any significant relationship with log of total delivery time. Hence, we do not include these features in the model.

Income per capita (RDPC)

There is a disparity of total delivery time between the different regions in Brazil. Shown in **Figure 19** the average total delivery time for each municipal is plotted at each of the customer's locations. The darker the colour the quicker the delivery was. The figure shows that deliveries were made to all regions in Brazil however most of the deliveries were concentrated along the southeast coast. The delivery time was also diverse the more southeast a customer lives the quicker their delivery was.

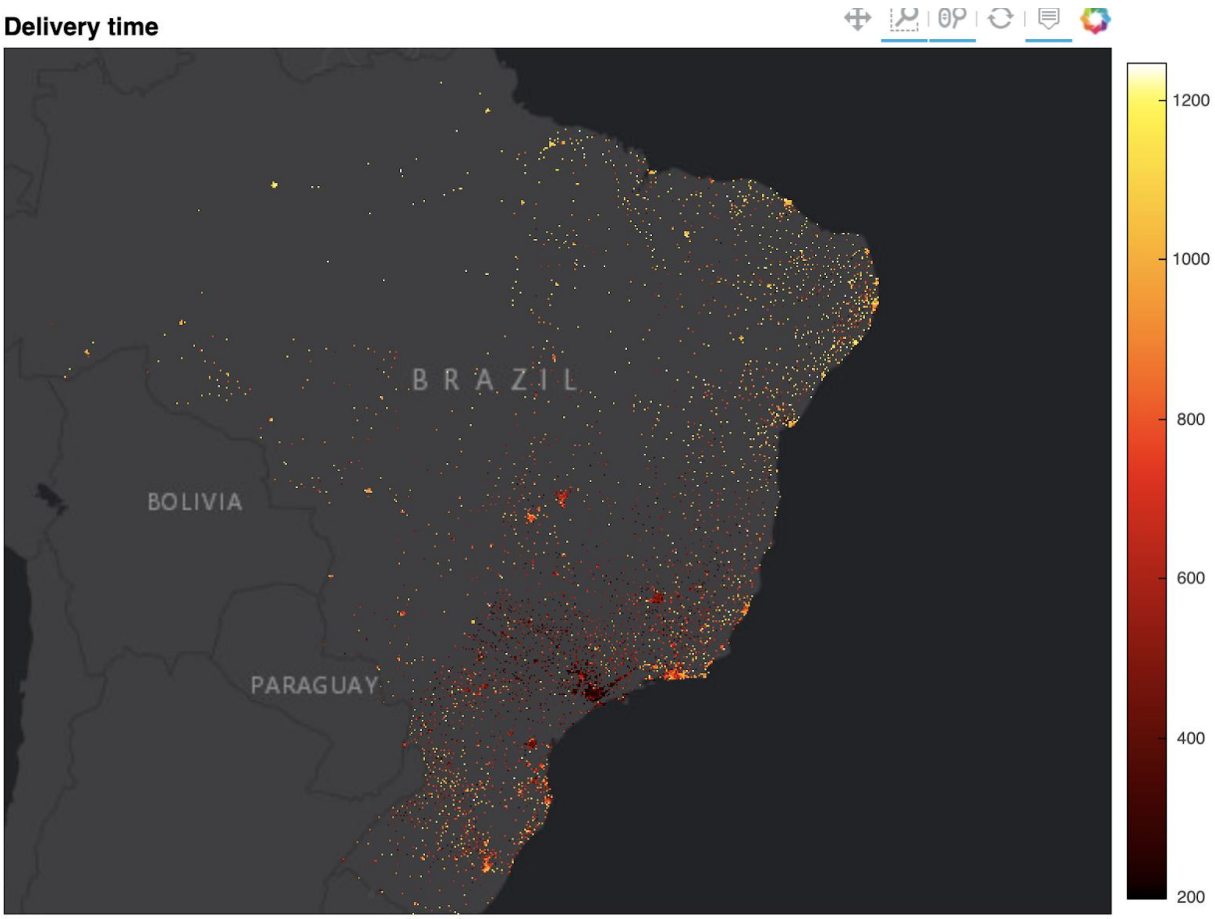

Figure 19 – Distribution of total delivery time (hours) in the different regions of Brazil

To attain a clearer analysis the data was broken up into the regional difference according to the Brazilian Institute of Geography and Statistics. The location of the different regions can be found in **Figure 35** in **Appendix A**. A box plot showing the distribution of total delivery time in the five regions is shown in **Figure 20** and summarized in **Table 5**.

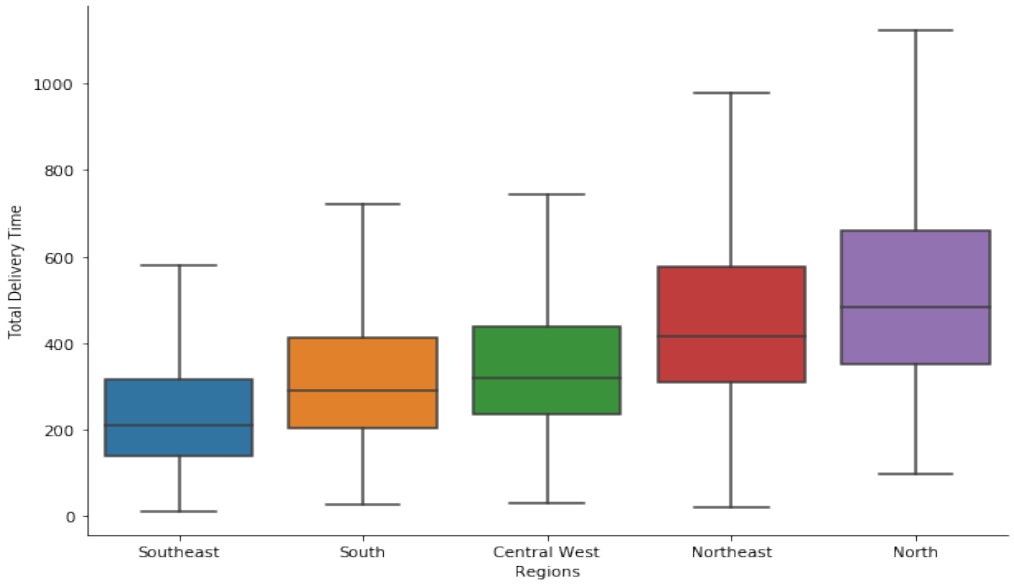

Figure 20 – Distribution of total delivery time of the regions of Brazil

The length of the boxes in **Figure 20** shows the ranges of the delivery time between the first and third quartile. The longer the box the larger the variation in delivery time. In **Figure 20** the sizes of the boxes are similar with the exception of the north and northeast region. These two regions have greater variation compared to the rest. The average values of each region can also be interpreted from **Figure 20** and **Table 5**. Each region has different average delivery times with the north region having the longest at 564 hours while the southeast had the lowest at 333.49 hours. From these figures it is clear regional difference exists for delivery time. Therefore, to increase the accuracy of predictions for delivery time regions should be considered.

Table 5: Summary statistics of total delivery time between regions of Brazil

	North	Northeast	Central West	South	Southeast
Count	1487	8894	5342	13546	25248
Mean	564.68	480.18	361.26	336.7	333.49
STD	330.47	294.35	198.17	202.4	237.39
Min	98.85	20.72	30.25	25.78	12.8
25%	366.57	309.07	235.98	203.9	190.97
50%	504.38	415.87	319.45	290.77	270.72
75%	686.24	576.04	438.88	411.12	402.36
Max	4695.22	4676.4	4345.45	4469.68	5031.08

Differential delivery time may exist across different regions due to varied development and condition of the transportation Infrastructure in Brazil. Infrastructure are facilities and systems that is used to service all levels of admirative regions such as highways, bridges and public transport (Daask, 2020). It is hypothesised that condition and density of infrastructure will determine how efficient trade routes are and how much volume it can support and therefore a good indicator to delivery time.

Development in infrastructure density and condition improves efficiency of trade routes by allowing transport methods to move faster, more frequent and in greater volume. Conditions and developments of infrastructure can come in many forms. Some examples are better condition of roads so trucks can travel faster and safer. More roads to allow for more direct routes so that trucks can travel a shorter distance. More lanes in the higher to allow for more vehicles and faster flowing traffic. Other forms of improvements can be establishing more airports so that goods can travel closer to their destinations by air. Larger trains networks can allow more direct travel as well as more connections between regions.

To be able to quantify the relationship infrastructure data is required. However, infrastructure is complex and difficult to attain, therefore an alternative source, income (RDPC), was investigate. (Queiroz & Gautam, 1992) suggested that there is an association between per capita income and the density and quality of road infrastructure. The most significant finding in their research was the relationship between density of paved roads (km/million inhabitants) to the size of the economy. They found low-income economies had a density of 170 or which only 40 are in good condition, middle-income with a density of 1,600 with 470 in good condition and up to 10,110 in high income economies with 8,550 in good condition. They concluded that there is a statistically significant relationship between road infrastructure and the economic development. This findings from the report gave enough confidence to explore income in this project.

Income per capita (RDPC) data for each county was collected from the national census. RDPC is defined as the nominal monthly household earnings, this is the average monthly earning of all members living in the same household. The data was then plotted in **Figure 21** for the different municipals in Brazil. There is a clear disparity of income between the different regions.

Figure 21 – Distribution of RDPC in Brazil

The north is poorer compared with mostly less than $500 a month compared to the wealthier south east side by the coast with income mostly above $1000. Looking at the summary statistics in **Table 6** the lowest monthly household earning was $96.25 (USD $19.33) for the Marajá Do Sena municipal which is located in the north east of Brazil. This level of income would be classified as under the UN poverty line of earning less than USD$1.90 a day or $57 a month (Kenton, 2020). Furthermore, the bottom 25% of income are all located in the northern religion and has an income per month of $281.12 (USD $57), this is also under the poverty line. The top third quarter correspondingly located in the southern region with an income above $650.62 (USD $131.31).

Table 6 – Summary statistics of income

RDPC			
Count	5565	($) 25%	281.12
Mean ($)	493.61	($) 50%	467.65
STD ($)	243.27	($) 75%	650.62
Min ($)	96.25	Max ($)	2043.74

Infrastructure grows around the economy, when the economy of a region grows more people migrate to the region increasing the need for more infrastructure. This theory can be validated by comparing the road network in **Figure 22** with the income distribution in **Figure 21**. In **Figure 22** the majority of the road network exists in the southeast, especially along the coast. In **Figure 21** the same regions described also show a greater level of income, mostly above $1000. In the northern region there is a sever lack of roads, this is partly due to the lost population density. Correspondingly RDPC is also lowest in this region mostly below $600.

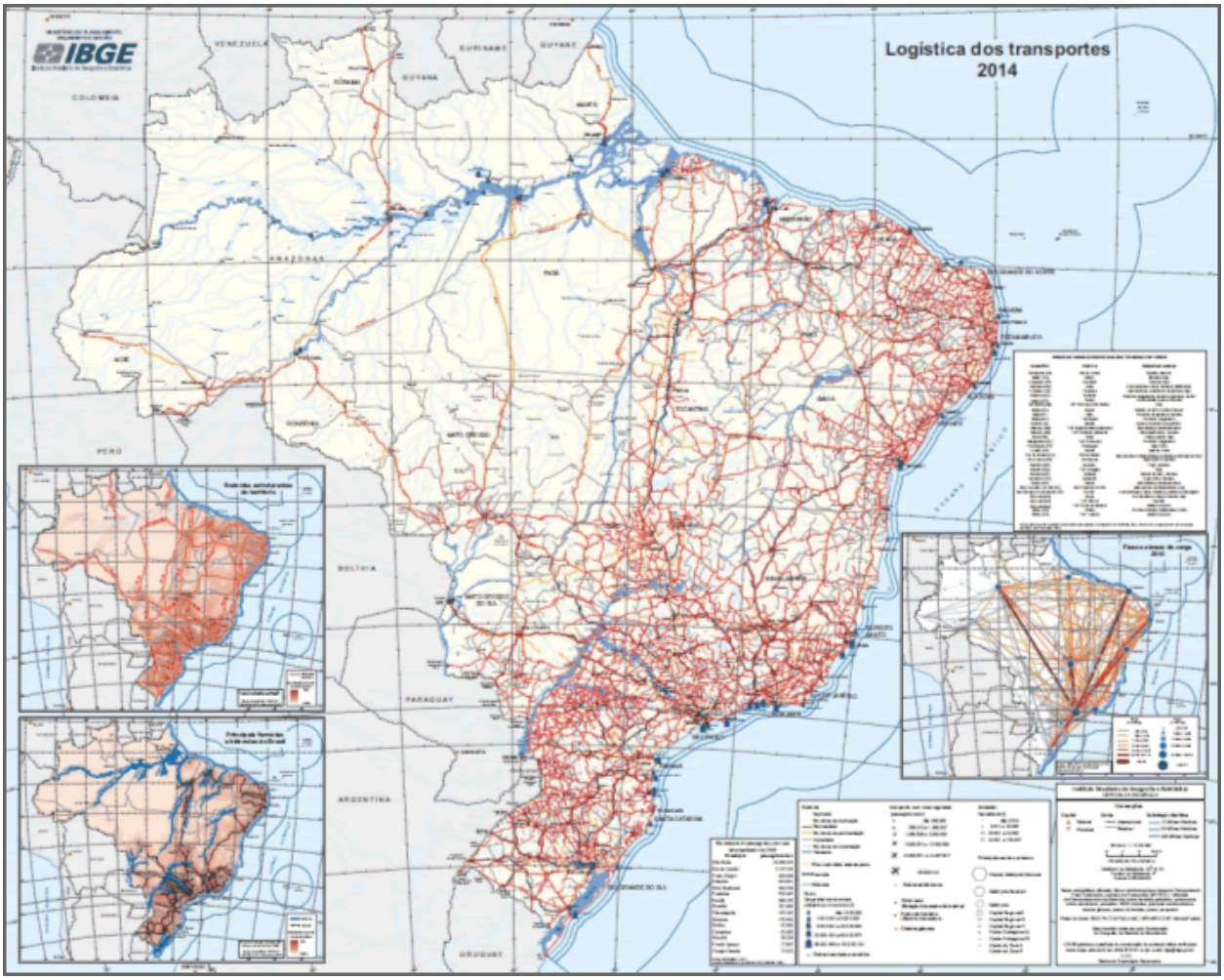

Figure 22 – Map of road network in Brazil (Prates, 2014)

Comparing the road network in **Figure 22** there is a clear relationship found between the density of road network and income per capita. The majority of the road network are located towards the wealthier and more populous south east region with little significant roads towards the north where the amazon forest

is. The same relationship is true for train networks shown in **Figure 23.** Most of the network are located in the same southern east region especially around Sao Paulo and Rio de Janeiro.

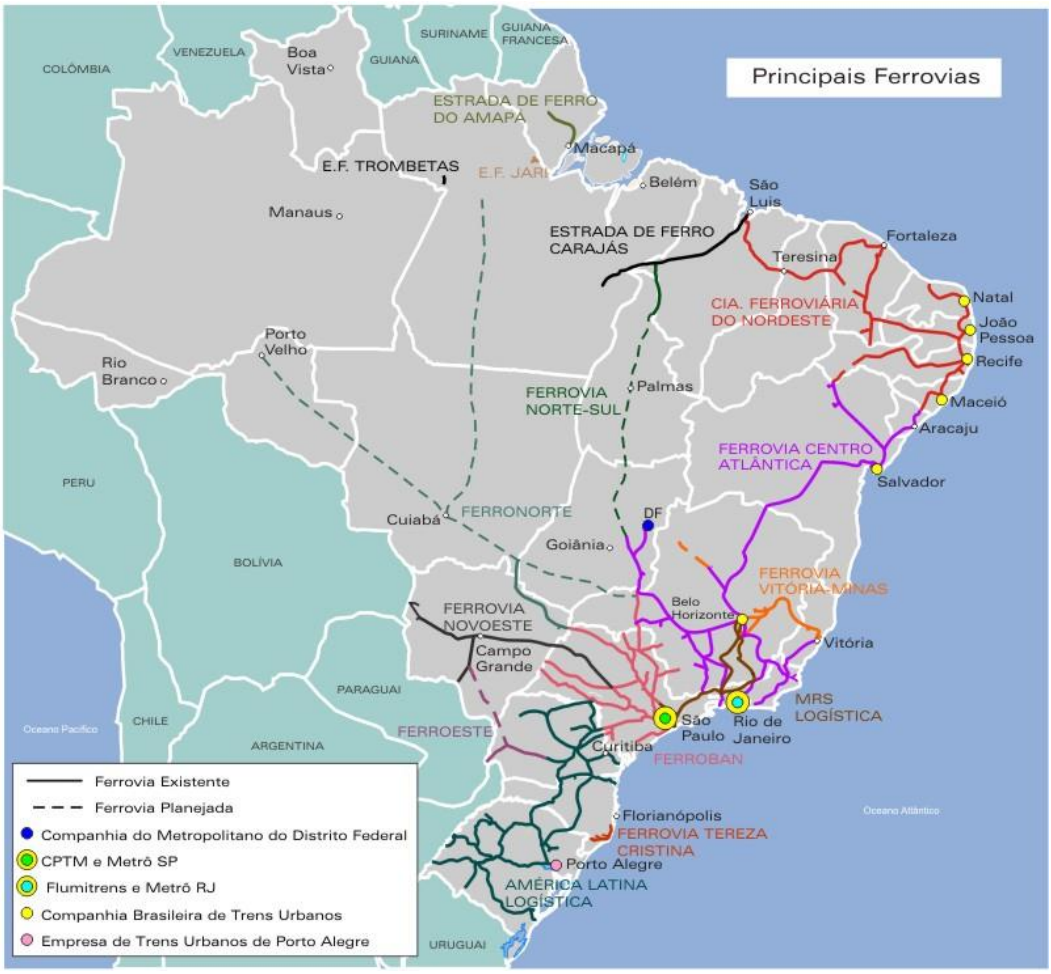

Figure 23 – Map of train network in Brazil

The above analysis has proved that income per capita is a good indicator of transport infrastructure. The delivery time can now be compared to the level of income. A plot has been made in **Figure 24** where average delivery time for each municipal has been plotted. The plot has been divided into three regions, the west, northeast and southeast shown in **Figure 25**, **Figure 26** and **Figure 27** respectively. The west region, a mid-income region, has in general longer delivery time, mostly above 800 hours, signified by the lighter yellow colour. The northeast region is poorer but has darker red colours which shows faster delivery times around 500 to 800 hours. The last region, south east, is the wealthiest region and does show the fastest delivery time around less than 400 hours with more darker red and black colours compared to **Figure 25** and **Figure 26**. The general trend follows the theory outlined above however the northeast region was expected to perform worse the east region but did not. Overall, across Brazil the general trends follow the theory the poorer the region the longer the delivery time.

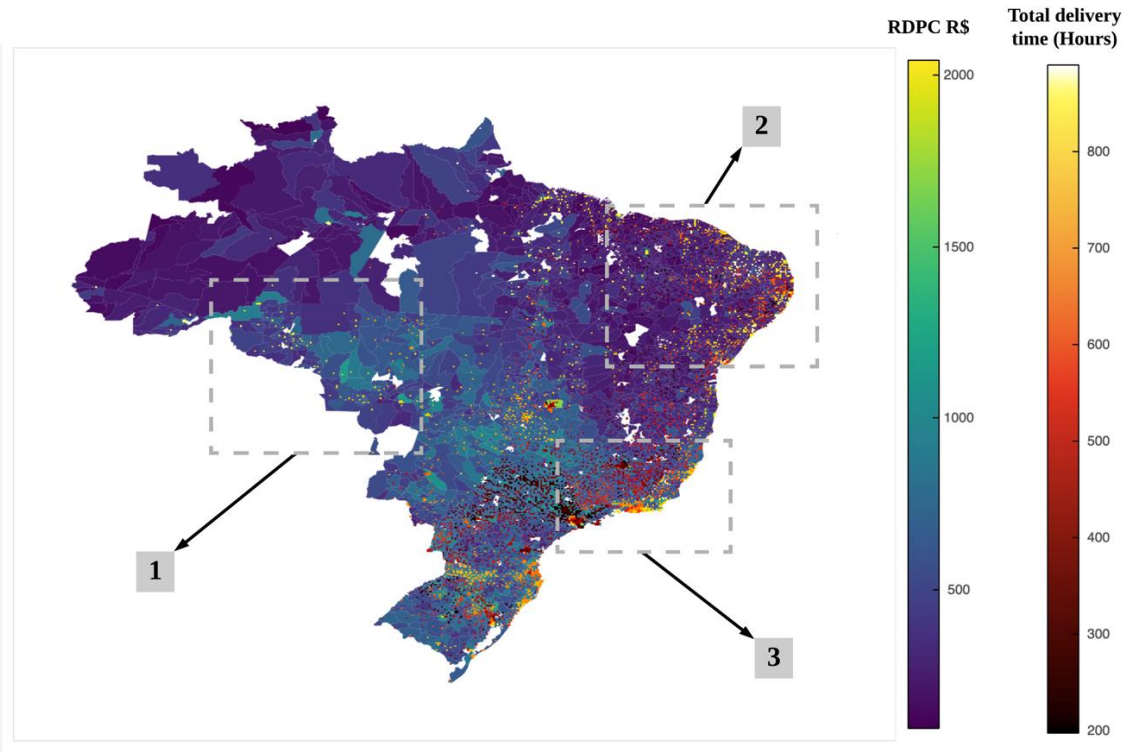

Figure 24 - Distribution of delivery time with income

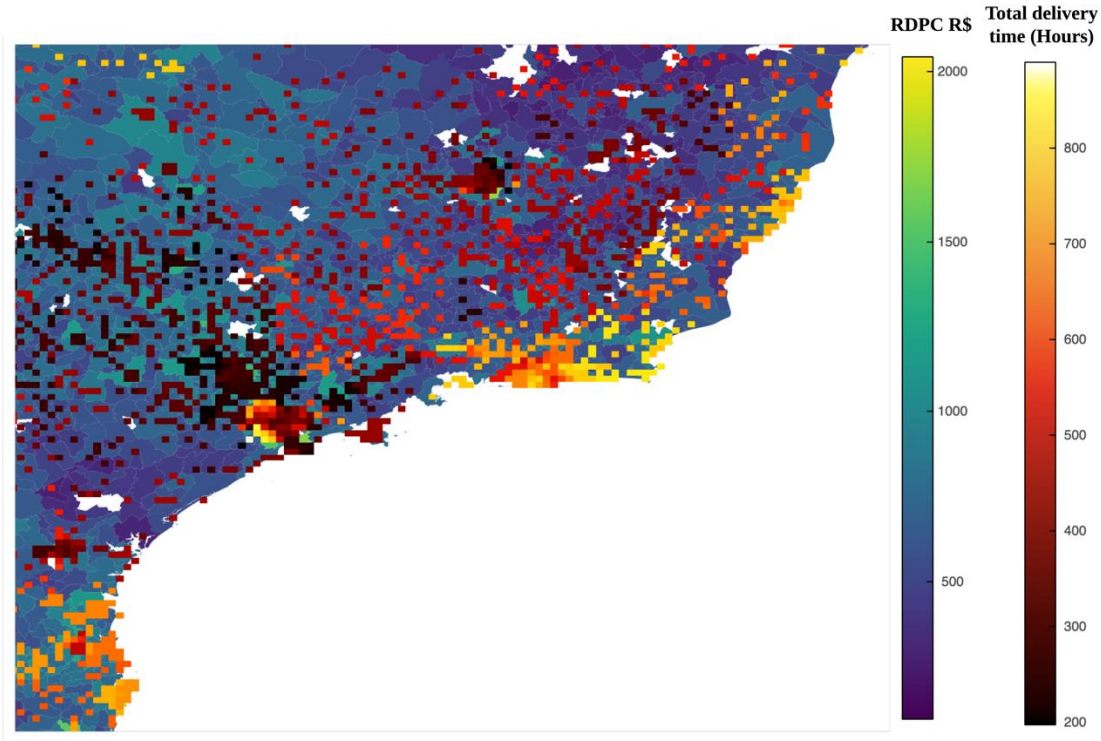

Figure 25 - RDPC of southeast region

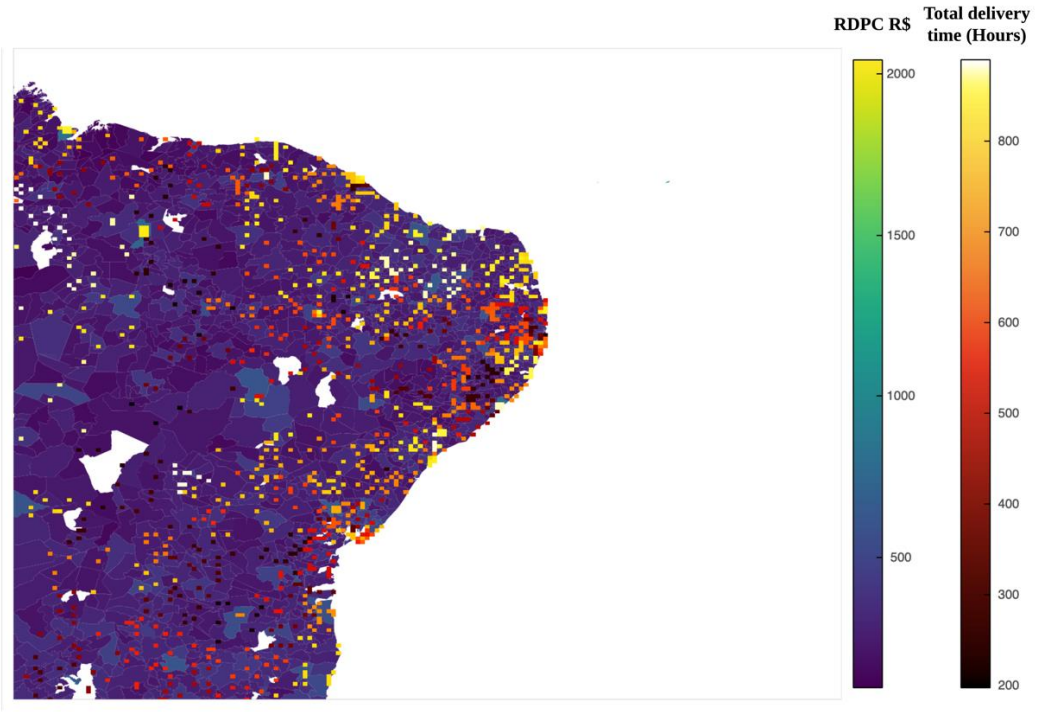

Figure 26 - RDPC of northeast Region

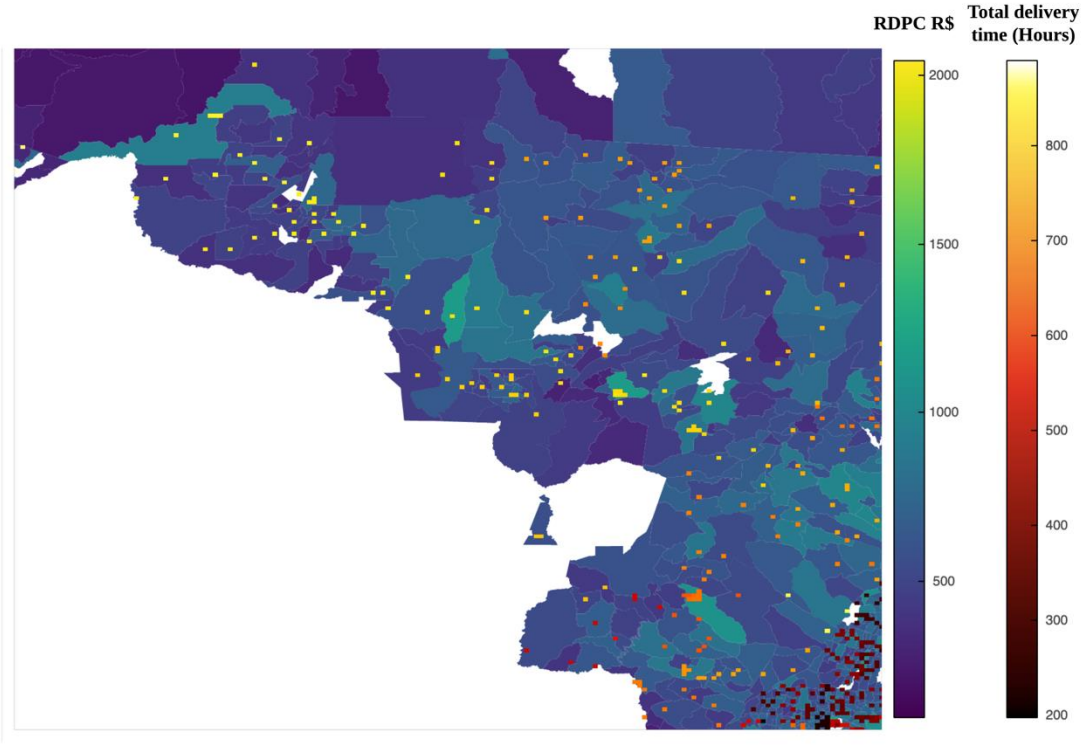

Figure 27 - RDPC of west region

Average Mean Seller Time

Total delivery time is the time taken by the seller to deliver to consumers. Each seller is based on particular city and deliver its product to a wide range of customer cities. Grouping by the *customer_city*, *seller_city, seller_id* and *order_purchase_timestamp* day we select the *total_delivery_time* count and its summation. The cumulative summation is worked out for their corresponding count and summation by the same group. The mean of these is calculated by dividing the cumulative summation via the count and subsequently, follow this procedure for each row by shifting the value to the next and grouping by the same attributes and taking the new mean seller time. We then came up with the rolling mean of seller that we can use as a time series plot.

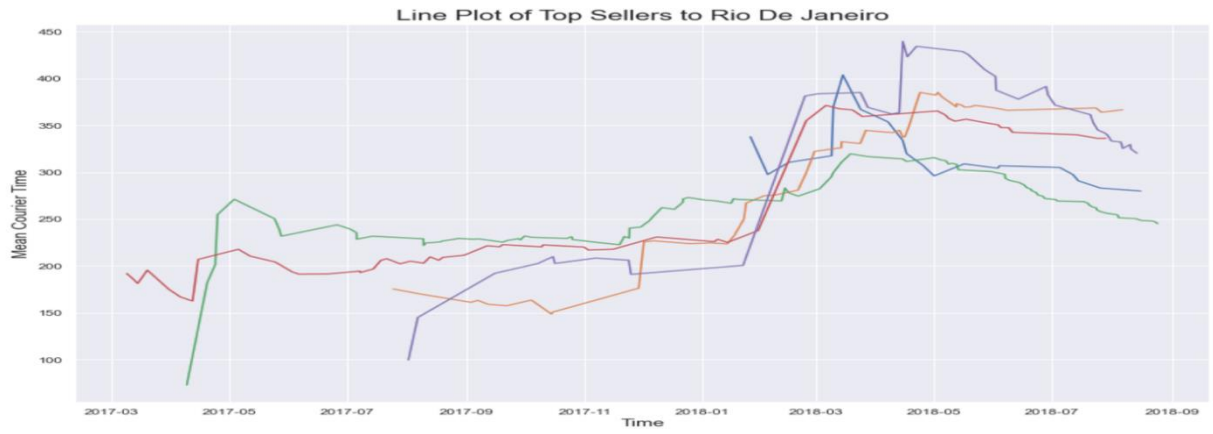

Figure 28: Historical delivery time of Top 5 Sellers in Sao Paulo to Rio de Janeiro

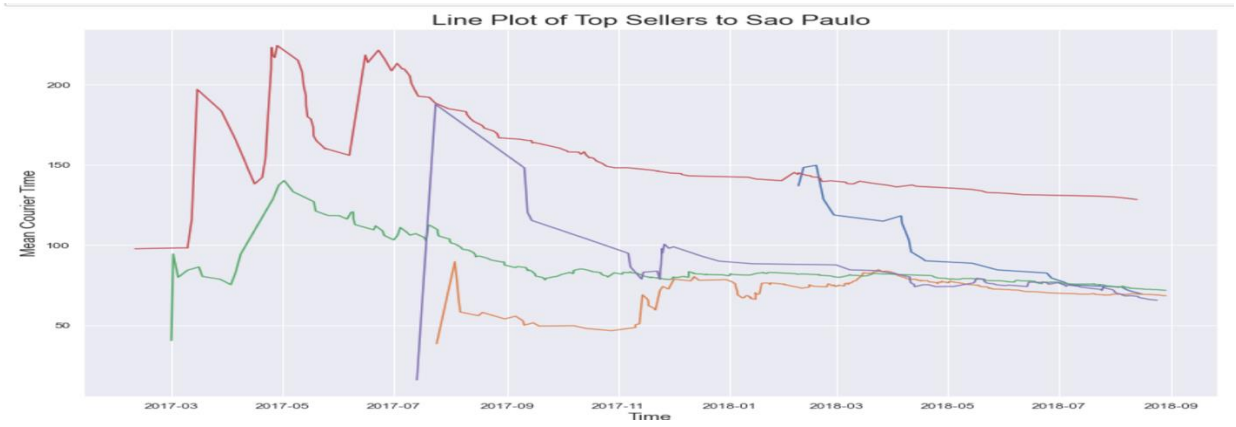

Figure 29: Historical delivery time of Top 5 Sellers in Sao Paulo to nearby area

The time series plot shows the average seller historical time of top 5 sellers based in Sao Paulo. The connection between Sao Paulo to Rio de Janeiro is greater than delivering in Sao Paulo area. Over time, performance of the seller delivering in nearby area improves **Figure 29**: Historical delivery time of Top 5 Sellers in Sao Paulo to nearby area, as mean courier time is getting lower i.e. downward trend. But in case of delivering to distant cities, the performance of sellers **Figure 28**: Historical delivery time of Top 5 Sellers in Sao Paulo to Rio de Janeiro disintegrates with the increase of average seller time as seen by the upward trend. But this time series is not enough to prove that is the case for every seller and its combination of cities since there are so many of them.

Freight Ratio

One of the most significant factors that affect online shopping is the shipping charges. In developing countries high volume of sales in some months play an important role in determining prices of commodities, specially the shipping cost. The freight ratio is calculated to see what percentage of each items contributes to this value corresponding to its volumetric weight. Different product has different shipping ratios and it varies from season to season. Normally, the freight ratio for denser products is higher than on lower and that is the case we see from the tree map for different product categories.

Figure 30: Freight Ratio of product categories in different season

One of the possible reasons for increased freight ratio is quicker delivery. The more you pay your shipping charges, the faster you can expect the delivery to arrive. From the scatter plot, we can see that even less weighted items take longer delivery time, thus freight ratio does not affect total delivery time.

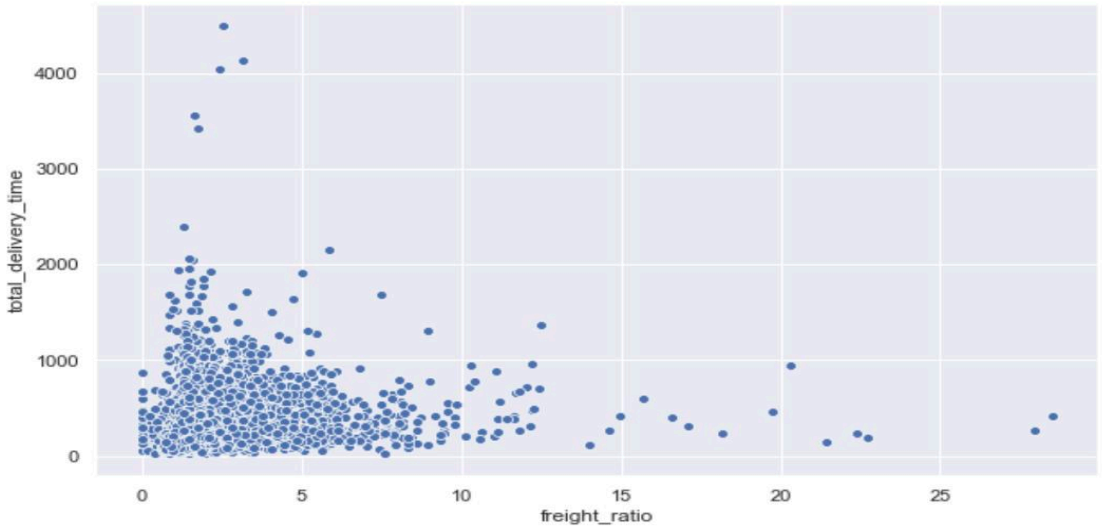

Figure 31: Freight Ratio vs Total Delivery Time

Making the data confess

Proposal Models

Multiple Linear Regression

Since the objective in this project is to predict the total delivery time, which is equivalent to solving a regression problem, and the evidence from **Figure 15** could be treated as a good starting point to incorporate more factors/features in order to build a multiple linear regression model. More importantly, as discussed in **Customers City & Sellers City Analysis**, the logarithm of the total delivery time tends to give a better result as well as reduce the influence from outliers; thus, the target variable for the multiple linear regression model in this section is the log of delivery time.

After selecting factors with high contribution and omitting nuisance factors, the multiple linear regression model could be formally described as follows:

$$\log(\mathbf{y}) = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Equation 1

Where,

$$\mathrm{X} = \begin{pmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \mathrm{x}_{13} & \mathrm{x}_{14} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \mathrm{x}_{23} & \mathrm{x}_{24} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & \mathrm{x}_{n3} & \mathrm{x}_{n4} \end{pmatrix}_{\mathrm{nx5}} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}_{5\mathrm{x}1}$$

and $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, 2, \dots, n$

are the matrix of data with n observations & 4 features including the intercept, vector of coefficients, and the error term for **Equation 1** respectively.

More precisely,

β_0 = the intercept

β_1 = the coefficient of the feature *connection_between_cities*

β_2 = the coefficient of the feature *actual_lag_time* or *estimate_lag_time*

β_3 = the coefficient of the feature *seasons*

β_4 = the coefficient of the feature *rdpc*

It is worth mentioning that, the parameter β_2 could be either represented for *actual_lag_time* or *estimate_lag_time* because delivery process could be segregated into many phases. For the first phase of delivery process, only the predictor *shipping_limit_date*, which is the deadline for sellers to dispatch the ordered products to logistic partners, is known; thus, the feature *estimate_lag_time* is evaluated based on *shipping_limit_date*. With regards to the second phase of delivery process, after sellers sent the ordered products to carriers, the predictor *order_delivered_carrier_date* is known; hence, it could be utilized to compute the value of *actual_lag_time*. In addition, the calculation process for both *actual_lag_time* and *estimate_lag_time* is described in **Table 11**. Consequently, **Equation 1** could be treated as the general formula for two sub multiple-linear-regression models of the first & the second phase of delivery process.

Since the goal is to estimate the total delivery time as accurate as possible, or equivalently to minimizing the difference between predicted total delivery time and exact total delivery time, Ordinary Least Squares method (OLS, 2020) is used to achieve the best-fitting line.

After applying OLS for fitting the models, **Equation 2** & **Equation 3** are the results of the first-phase delivery process & the second-phase delivery process models respectively (for further information see **Figure 36** & **Figure 37** in **Appendix B**).

$$\log\left(\widehat{\text{total\_delivery\_time}}\right) = 6.6427 - 5.148 \times 10^{-05}\text{X}_1 + 6.85 \times 10^{-06}\text{X}_2 - 0.1382\text{X}_3 - 0.0001X_4$$

Equation 2

where $X_2 =$ *estimate_lag_time*

$$\log\left(\widehat{\text{total\_delivery\_time}}\right) = 6.6411 - 4.76 \times 10^{-05}\text{X}_1 + 0.002\text{X}_2 - 0.1084\text{X}_3 - 0.0001X_4$$

Equation 3

where $X_2 =$actual lag time

Note that, the metrics for evaluating both models in this case are R Squared (Coefficient of determination, 2020) and Mean Absolute Error (MAE, 2020). For the reason of using R Squared as the indicator for comparing models, it tells how well the model is able to explain the data (e.g. 60% in R-squared indicates that the model is able to describe correctly 60% of the data). And for the MAE metric, it is used to express the mean difference between predicted values and actual values, so it is easy for interpretation standpoint as well as less sensitive to outliers.

According to the results from **Figure 36** & **Figure 37**, it is clear that the model incorporates the feature *actual_lag_time* have higher R-squared score in comparison with that of another model, which sort-of confirms that the more exact information can be collected, the better estimation should be. What is more, the MAE value from **Figure 37** tells that the absolute mean difference between the log of predicted delivery time and the log of actual delivery time is around 0.302, which is approximately to 90.9 hours or 3.78 days in terms of absolute average difference between predicted delivery time and actual delivery time.

For model training procedure, all records have been shuffled and then 70% of the dataset is used for training & the rest 30% is for testing.

After completing building the models, one final step is to make sure whether the assumptions for those models are valid. In this context, validating residual assumption & independent assumption between predictors for multiple linear regression model is considered. Specifically, for residual assumption, the residuals from those models must be normally distributed around mean 0 with variance σ^2. In addition, for independent assumption between variables or multicollinearity assumption, the (Peason Correlation Coefficient, 2020) between variables should not be highly correlated.

With regards to checking multicollinearity assumption, the Pearson correlation matrix for all combinations of variables using for those models is constructed as shown in **Figure 32**. It seems that the correlations between predictors are generally; to be specific, the highest correlation coefficient has seen so far is -0.29.

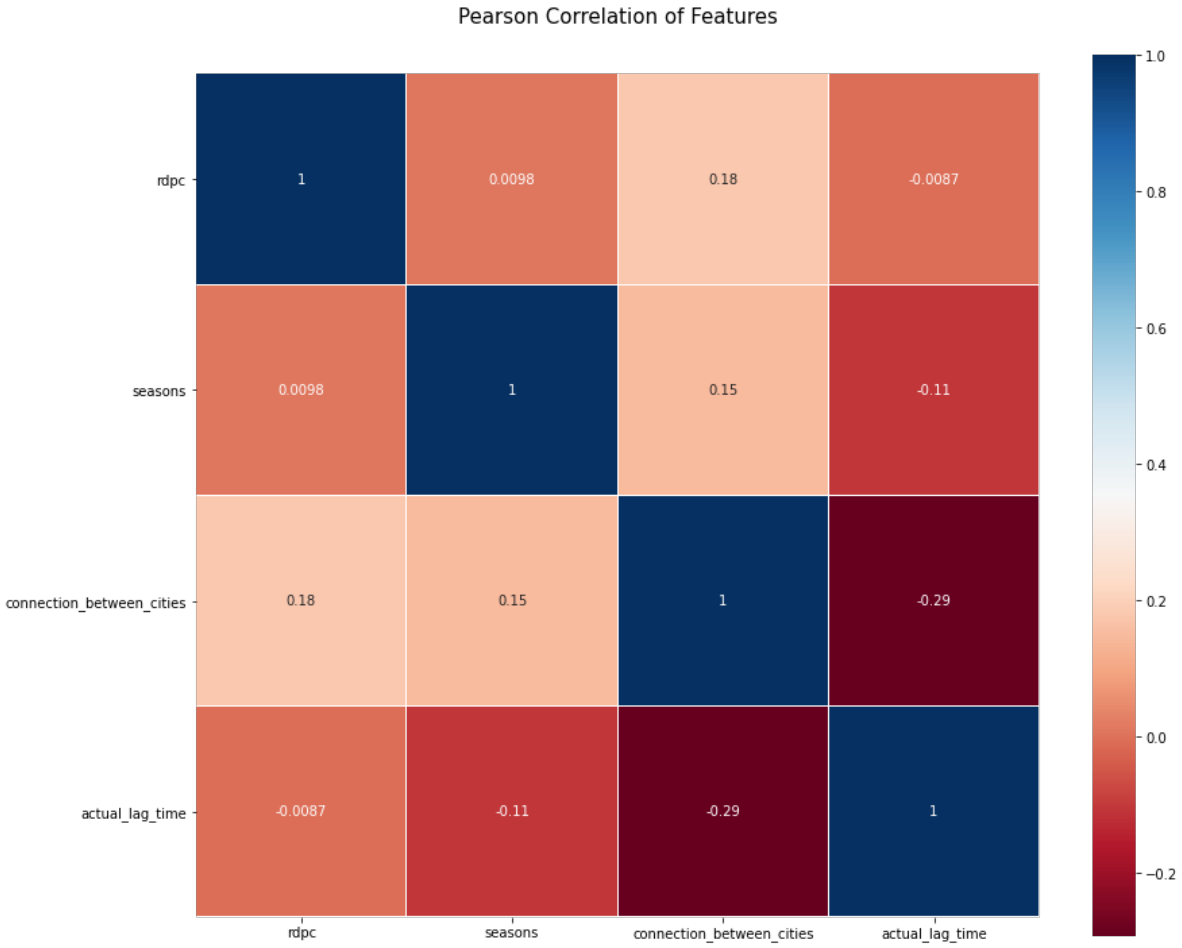

Figure 32

To validate the normality assumption of residuals, two approaches have been conducted. The most straightforward approach is to sketch the residual plot & residual distribution and conclude based on the visual results. Another approach is to carry out a hypothesis testing to verify whether the residuals from testing data of the second-phase delivery process model is from Normal distribution with mean 0 and variance σ^2. In detail, Kolmogorov-Smirnov test (K-S test, 2020) has been used in the following way:

- For each $\hat{\sigma}$ around the sample standard deviation of the residuals

* The reason for iterating $\hat{\sigma}$ around the sample standard deviation of the residuals is because of the Law of Large Number (LLN, 2020) as the size of the sample increases towards infinity, the sample standard deviation of the residuals approaches its population standard deviation.

1. H_0: the residuals are from $\mathcal{N}(0, \widehat{\sigma^2})$
   H_1: the residuals are not from $\mathcal{N}(0, \widehat{\sigma^2})$
2. Compute the observed K-S test statistic.
3. Generate many test statistics under H_0.
4. Compute the p-value based on the result from Step 2 & 3.

Based on the evidence from verifying visually via **Figure 33** and quantitatively via **Figure 34**, it is sufficient to say that the residuals are surely from $\mathcal{N}(0, \sigma^2)$, where σ is somewhere between 0.3 and 0.4.

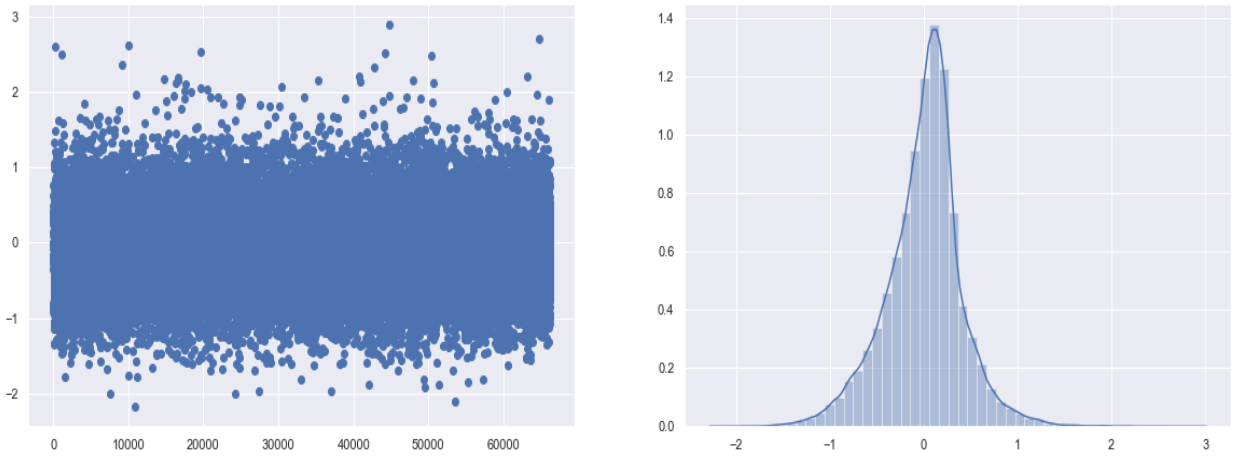

Figure 33. Residual plot (Left) & Distribution of residuals (Right) from performing prediction by 2nd-phase delivery process model on testing data.

```
Proposal sigma with p-value greater than 0.9 is
from 0.3075947371091472
to 0.4035685165055692
```

	proposed_sigma	p_value
26	0.307595	0.903
27	0.308966	0.912
28	0.310337	0.920
29	0.311708	0.922
30	0.313079	0.928
31	0.314450	0.922
32	0.315821	0.948
33	0.317192	0.941
34	0.318563	0.946
35	0.319934	0.950

Figure 34. Some proposal $\hat{\sigma}$ with p-value greater than 0.9 when applying K-S test

Random Forest

Random forest is another model analysed to help diagnose errors and biases of OLS and establishing stronger understanding of relationships between the dependent and independent variables. Random forest is a supervised ensemble machine learning algorithm that can be used for both classification and regression. Random forest operates by taking many samples of the data and feeding each one to a different decision tree. The predictions are then made as an average prediction all the decision trees for regression and the classification with the most votes for classification problems.

Random forest was chosen due to its ease of use, there is no need to scale the data or transformed, low computational time as jobs can be run in parallel and lastly most importantly it is robust to outliers as many trees are built on resampled data. Olist's dataset contain a significant number of outliers which may have skewed the results in OLS. (Kho, 2018)

The only assumption Random Forest has is that the sampling method is representative. The assumption is only satisfied when the sample taken for each decision tree reflects the entire dataset. This is important as each tree is making the decisions based on that sample. Thus, samples not representative will result in incorrect node splits and thus incorrect predictions. Sklearn bootstrap method involves iteratively resampling a dataset randomly with replacement of a specified size. As the sampling is random and the size of the data is big it is likely each sample will be representative of the population. (Brownlee, 2019)

The only pre-processing required for random forest is to split the data. The features dataset was split into a training and test dataset, a split of 75% and 25% respectively. The training dataset is used to train the random forest trees while the test dataset is used to evaluate the model's performance.

To train the model each decision tree is given a random sample of the training data. A decision criterion is used to split the nodes of the tree based on the values of a feature. In this regression problems the model searches for a value of a feature that results in the smallest mean squared error (MSE). The mean squared error is the average difference between the predicted value and the true value.

The importance of individual features can be outputted from the model. Each feature is scored based on the average reduction of variance for each node and scaled between 1 and 0 such that the sum of the scores is equal to 1. A feature with a score of 1 signifies the predictions are made entirely made from that feature while a score of 0 shows it had no weight on the prediction. The R^2 score is also an important evaluation of the model. It is the normalised MSE score and shows how well the model explain the variance of the data. A score of 1 shows a perfect model where all data points have been exactly predicted right while a score of 0 shows no datapoints have been predicted well.

A base model has been used to fit the Random Forest with parameters in **Table 9** of **Appendix B**. The R^2 on training data is 0.66 and 0.65 on the test data. These values are reasonably high, and the model can be

said to explain a significant portion of the variance in the data. As the R^2 is similar on the test dataset the model has not overfit. Looking at the feature importance score in **Table 7** the model shows that only two features out of the four features were considered in the model, the connection between cities and the actual lag time. The feature importance scores show overwhelmingly 87% of the prediction is based on the connection between cities and only 12% on the actual lag time. Therefore, the value of connection between cities will mostly determine how quick the delivery will be.

Table 7 – Feature importance scores before cross validation

Features	Feature Importance Score
Connection Between Cities	0.875024
Actual Lag Time	0.124976

To better fit the model cross validation has been utilised to tune the hyperparameters. Cross validation divides the training data into equal partitions. One of the partitions will be used to test the model's performance while the other partitions will be used to train the model. This is then repeated until all partitions have rotated such that they have all been used to train and test the model. To tune the parameters every combination of parameters is tested and the parameters that result in the best performing model is chosen.

After cross-validation was performed the best parameters has been chosen and detailed in **Table 10** of **Appendix B**. The R^2 score has improved from the base model to 0.69 for test dataset and 0.71 for the train dataset. This is an increase of approximately 5%. However more significantly the model now considers all the features as shown in **Table 8.**

The weighting on the predictions for the connection between cities has fallen by 4% to 0.83. Half of this weight has been transferred to actual lag time and the rest to income per capita and seasons. Despite the connection between cities still overwhelmingly the greatest weight in the predictions. The importance of income and seasons are very small at 1% or less.

Table 8 – Feature importance scores after cross validation

Features	Feature Importance Score
Connection Between Cities	0.83127846
Actual Lag Time	0.1484964
RDPC	0.01370604
Seasons	0.00651909

Relating back to explorative data analysis performed previously. The model reflects the strong relationship found between the connections between cities in **Figure 15**. Likewise, the lag time found in the EDA had a positive linear relationship although the relationship was not as strong as connections between cities. This is likely the reason why actual lag time was also weighted significantly less. However, it was surprising to see how poorly income was weighted in the model. In the EDA total delivery time was evidently longer in lower income areas and quicker in high income areas. An explanation of this may be attributed to the variation in total delivery time within the same regions. On average the correlation may be strong however individually, customers have bought items from different sellers, some sellers are faster than other some products are easier and quicker to ship than others.

Models performance

Both models with the incorporation of the feature *actual_lag_time* performed reasonably well with an R^2 score of 0.69 on the test dataset for the Random Forest model and 0.639 for the OLS model. The connection between cities were both the most significant factor with income and seasons less relevant.

The random forest model found that all four features income, lag time, connections between cities and seasons were considered. However, the connections between cities significantly more weight in the predictions than any other variables. Actual lag time had a small influence in the model while income and seasons had little to no impact on the model. The estimated delivery time is therefore mostly dependent on the connection between cities.

Recommendations

Solutions need to cater how to best optimize *connection_between_cities, estimate_lag_time* or *actual_lag_time, seasons,* and *rdpc.* The *connection_between_cities, estimate_lag_time* or *actual_lag_time* features can be considered into two parts of the ecommerce process, logistics and fulfillment, respectively. The goal would be to both minimize and ensure consistency of delivery time by focusing to increase *connection_between_cities* which is impacted by the city pairs of sellers and customers. Given this, the solution should be tailored to specific cities. There are two ways of doing this, first is explore the viability of assigning specific 3rd party logistics partners (3PLs), and second is to explore availability of inventory sharing among sellers. Some 3PLs perform better in some areas as compared to others (O'Byrne, 2017), thus, this could potentially increase consistency of delivery time, thus, being able to provide more accurate *estimate_delivery_time*. If the inventory of Olist's sellers is consolidated into one inventory management system (IMS), then it would be possible for sellers to share their resources with each other. As an example, Seller A is selling Product A to a customer and *connection_between_cities* is 10, but if Seller B, who has the same Product A, sold to the same customer *connection_between_cities* would have been 100. Therefore, delivery time would be lower with Seller B. If Olist can find a way for sellers to allow their inventory to be sold by other sellers, then inventory sharing would be possible by simply assigning inventory through the IMS. Once an order is placed in the system, this would enable Olist to assign specific sellers to customers that will increase the *connection_between_cities.* To optimize *estimate_lag_time* or *actual_lag_time* the solution should be focused in the pre-delivery process, which is inventory and order management, and fulfillment. In the setup of Olist, either Sellers or the Logistics partners are the ones responsible for fulfillment. One way to be able to ensure a high quality of fulfillment is to evaluate sellers and somehow penalize them for unnecessary delays on their end. This will ensure that sellers would always do their best to comply with the standards set by Olist. Another way to decrease fulfillment time is stock availability and ensuring that physical stocks match those in the system. E-commerce is fast paced, and orders are almost created instantaneously and can come from multiple selling platforms, these are why inventory management becomes critical in the fulfillment process. When orders are placed when there is no physical inventory left, then delays would occur while waiting for replenishment, or worst case, orders will be cancelled. To prevent this, ensuring proper inventory management is key and by doing so will ensure more stable fulfillment times, thus, having consistent *estimate_lag_time* or *actual_lag_time.* The *seasons, and rdpc* features are something that Olist does not have control over, therefore, no recommendations were required pertaining to these.

Limitations

One of the main challenges in this project was to be able to precisely determine the total distance between sellers and customers. The sellers of Olist does not necessarily fulfil and/or ship the products themselves. Most of those sellers follow either a drop-shipping or cross-docking model. To briefly explain these two concepts; drop-shipping requires the seller to be responsible for fulfillment and then hands them over to a 3PL for shipping, on the other hand, cross-docking only requires the seller to provide the products required, on demand, for an order to a 3PL for them to fulfil and ship. Given this, the starting point would more likely be from where the 3PL is situated. However, this information was not given in the available data. This required if the farthest seller be the starting point (see **Customers City & Sellers City Analysis**).

Another challenge is being able to determine the precise routes taken by the 3PLs. The data lacked information such as specified 3PL shipping an order, shipping method used (land, sea, air freight or any combination), routes usually taken by those 3PLs, routing/delivery schedule, and traffic data are some of the information required to clearly determine factors that can optimize delivery routes. This is the motivation behind utilizing skyway distance instead.

One of the problems we faced was the amount of computational time it took to transform the large dataset. Especially in computing the skyway distance using the haversine formula. Since pandas data frames are in memory and single server, their size is limited to the computational resource we had. Apart from this, to make better visualization we used libraries such as geoviews. Although these libraries are fairly easy to use but installing its dependencies can be tricky at times.

Future Studies

The features that were used in this project has only scratched the surface. The estimated delivery time might be better explained using features that involve traffic volume over some time interval (e.g. hours of day, week, month, etc.). Planned, and actual delivery routes may also improve the estimation. On another perspective, this may also be considered as a routing problem considering the delivery network, lane speed, average courier speed, and other metrics that may be able to explain interactions in land, sea, and air freight (depending on what is used for shipping). These metrics may be used to investigate the routes taken by 3PLs that can minimize the delivery time based on traffic conditions. More exact information on locations such as pick-up point, distribution centre address, and customer address may also help to identify which sellers would be the fastest to deliver certain products to customers.

Project Progression and Improvements from Feedback

The initial topic of the group is to predict the sales volume of Olist at any given month. The teaching team pointed out that it is not a necessity for predictions to be the focus of the project and that there may be other aspects of data science that can be explored instead. Additionally, it was pointed out that there is a lack of information and there were not enough features discussed to be able to successfully predict sales volume and that the topic itself was not very clear. The group responded by exploring other features that may be able to help predict sales, such as the gross regional product (GRP) of each state and seasonality. However, GRP posed as a challenge since it was highly correlated to *customer_state* variable. Apart from this, there were also other challenges faced by the group, therefore, it was decided that a change of approach is required. The group shifted to predicting *total_delivery_time* as a result.

With the new topic the teaching team wanted to understand how delivery time differs from each state and each city, and that the group should consider building multiple models if the delivery times were truly distinct for each. Instead of creating multiple models, the group continued to explore the data. During this exploration two new features were formulated to answer the problem. The assumption is location and delivery distance do impact delivery time; however, it was not known to the group where the starting point is. This is where Customers City & Sellers City Analysis, and Customers State and Sellers State Analysis were developed. Additionally, the group also assumes that the economic status of each area affects delivery time, thus, **Income per capita (RDPC)** was introduced.

The project pitch is "Analysing the factors affecting delivery time." The mindset during the project pitch is to simply understand what affects delivery time but the group has shifted to understanding how to create a better *estimate_delivery_time*. Given this, the group wanted to create a feature than can incorporate seller performance, thus, *average_mean_seller_time* was created (see **Average Mean Seller Time**).

After the presentation, feedback from peers mainly focused on the problem statement and the recommendation. The group's peers did not have a clear understanding of the problem statement as well as the purpose of showing that there is a 2% gap to meet the standard of number of orders shipped within estimated delivery time. This is addressed at the Problem Formulation section of the report, and the 2% gap statement was removed since it is not in line with understanding how to create a better *estimate_delivery_time*. The questions on recommendation include understanding why the income of an area or the distance between the sender and the receiver did not have a strong impact, and why are features relating to infrastructure and proximity not included. The distance conducted in this study is the skyway distance, which is not the actual travel distance, hence, it is hard to see the real contribution of distance on delivery time. Also, the exact locations of customers or sellers are unknown that is why estimation of the distance can only be based on relative location of prefix zip code. Further explanation is discussed in **Distance Analysis**. Other feedback included improving recommendations and empathizing with stakeholders of the delivery. These were discussed in **Recommendations** and **Limitations**.

References

BIG AND BULKY CHALLENGES IN AN OMNI-CHANNEL WORLD. (2016, September). Retrieved from FIDELITONE: https://www.fidelitone.com/blog/big-and-bulky-challenges-in-an-omni-channel-world

Boleto. (2016, March 7). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Boleto

Brownlee, J. (2019). *A Gentle Introduction to the Bootstrap Method*. Retrieved June 9, 2020, from https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

Climate in Brazil. (2020, April). Retrieved from https://www.climatestotravel.com/climate/brazil

Coefficient of determination. (2020, April). Retrieved from https://en.wikipedia.org/wiki/Coefficient_of_determination

Daask. (2020). *Infrastructure* . Retrieved June 8, 2020, from https://en.wikipedia.org/wiki/Infrastructure

Edison, N. (2016). *Fast delivery and premium packaging has a high impact on customer loyalty, according to recent data from Dotcom Distribution.* NJ: Dotcom Distribution. Retrieved from https://www.mhlnews.com/transportation-distribution/article/22051729/delivery-time-top-priority-for-online-shoppers

General Practices for Accepting Boleto Bancário Payment in Brazil. (2016, March 9). Retrieved from Tech in Brazil: https://techinbrazil.com/general-practices-for-accepting-boleto-bancario-payment-in-brazil

Howen, A. (2014, December). *The Impact of Late and Inaccurate Deliveries on Customer Loyalty*. Retrieved from website Magazine: https://www.websitemagazine.com/blog/the-impact-of-late-and-inaccurate-deliveries-on-customer-loyalty

Kenton, W. (2020). *International Poverty Line*. Retrieved June 7, 2020, from https://www.investopedia.com/terms/i/international-poverty-line.asp

Kho, J. (2018). *Why Random Forest is My Favorite Machine Learning Model*. Retrieved June 9, 2020, from https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

K-S test. (2020). Retrieved from https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

LLN. (2020). Retrieved from https://en.wikipedia.org/wiki/Law_of_large_numbers

Log transformation of target variable. (2019, August). Retrieved from https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion/103975

MAE. (2020). Retrieved from https://en.wikipedia.org/wiki/Mean_absolute_error

O'Byrne, R. (2017, August 10). *Logistics Outsourcing Secrets: A Serial Guide to Success with 3PLs*. Retrieved from Logistic Bureau: https://www.logisticsbureau.com/logistics-outsourcing-3pl-guide/

Olist, & Sionek, A. (2018, September). *Brazilian E-Commerce Public Dataset by Olist.* Retrieved from kaggle.com: https://doi.org/10.34740/KAGGLE/DSV/195341

OLS. (2020). Retrieved from https://en.wikipedia.org/wiki/Ordinary_least_squares

Peason Correlation Coefficient. (2020). Retrieved from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Permutation Tests. (2020). Retrieved from https://en.wikipedia.org/wiki/Resampling_(statistics)

Prates, P. I. (2014). *IBGE maps transportation infrastructure in Brazil*. Retrieved June 8, 2020, from https://mundogeo.com/en/2014/12/11/ibge-maps-transportation-infrastructure-in-brazil/

Queiroz, C., & Gautam, S. (1992). *Road Infrastructure and Economic Development.* New York: Working Papers.

Scott, H., & Lasher, C. (2018, May). *E-commerce Warehouses Adapt for Large and Bulky Item Handling*. Retrieved from Supply & Demand Chain Executive: https://www.sdcexec.com/warehousing/article/21002159/ecommerce-warehouses-adapt-for-large-and-bulky-item-handling

Touropia. (2020). *5 Most Beautiful Regions in Brazil*. Retrieved June 8, 2020, from https://www.touropia.com/regions-in-brazil-map/

Victor, R. (2020, February 4). *Hollingsworth*. Retrieved from How Late Deliveries Impact Customer Retention: https://www.hollingsworthllc.com/how-late-deliveries-impact-customer-retention/

Appendix

Appendix A

Figure 35 –Regional Locations according to Brazilian Institute of Geography and Statistics (Touropia, 2020)

Appendix B

Dep. Variable:	total_delivery_time_log	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.581
Method:	Least Squares	**F-statistic:**	2.301e+04
Date:	Sun, 07 Jun 2020	**Prob (F-statistic):**	0.00
Time:	14:09:27	**Log-Likelihood:**	-39461.
No. Observations:	66252	**AIC:**	7.893e+04
Df Residuals:	66247	**BIC:**	7.898e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.6427	0.005	1272.276	0.000	6.632	6.653
connection_between_cities	-5.148e-05	1.81e-07	-284.516	0.000	-5.18e-05	-5.11e-05
estimate_lag_time	6.851e-06	8.14e-07	8.419	0.000	5.26e-06	8.45e-06
seasons	-0.1382	0.004	-39.200	0.000	-0.145	-0.131
rdpc	-0.0001	3.8e-06	-27.758	0.000	-0.000	-9.81e-05

```
R-squared for testing data =  0.5814998053471087
Mean Absolute Error for testing data =  0.31896940827077314
```

Figure 36 – Summary statistics & Evaluation metrics for Equation 2

Dep. Variable:	total_delivery_time_log	**R-squared:**	0.639
Model:	OLS	**Adj. R-squared:**	0.639
Method:	Least Squares	**F-statistic:**	2.932e+04
Date:	Sun, 07 Jun 2020	**Prob (F-statistic):**	0.00
Time:	14:08:44	**Log-Likelihood:**	-34562.
No. Observations:	66252	**AIC:**	6.913e+04
Df Residuals:	66247	**BIC:**	6.918e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.4110	0.005	1198.897	0.000	6.401	6.421
connection_between_cities	-4.76e-05	1.72e-07	-276.449	0.000	-4.79e-05	-4.73e-05
actual_lag_time	0.0020	1.93e-05	103.163	0.000	0.002	0.002
seasons	-0.1084	0.003	-33.056	0.000	-0.115	-0.102
rdpc	-0.0001	3.53e-06	-33.183	0.000	-0.000	-0.000

```
R-squared for testing data = 0.6390405583271263
Mean Absolute Error for testing data = 0.30287715878618504
Mean Absolute Error for the difference between predicted delivery time vs actual delivery time:
 90.91371609676503
```

Figure 37 – Summary statistics & Evaluation metrics for Equation 3

Table 9 – Parameters of Random Forest Model without cross-validation

Parameter	Value
bootstrap	TRUE
ccp_alpha	0
criterion	MSE
max_depth	4
max_features	None
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_impurity_split	None

min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0
n_estimators	100
oob_score	FALSE
random_state	0
warm_start	FALSE

Table 10 - Parameters of Random Forest Model with cross-validation

Parameter	**Value**
bootstrap	TRUE
ccp_alpha	0
criterion	MSE
max_depth	8
max_features	auto
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	5
min_samples_split	2
min_weight_fraction_leaf	0
n_estimators	300
oob_score	FALSE
random_state	0
warm_start	FALSE

Appendix C

A. Names and sources of the data

Olist - https://www.kaggle.com/olistbr/brazilian-ecommerce

UNDP – https://www.kaggle.com/pauloeduneves/hdi-brazil-idh-brasil

The World Bank - https://datacatalog.worldbank.org/dataset/2010-brazil-municipalities-location/resource/64fc767a-524f-4c54-9502-ec99e7f4ca6e

Humanitarian data exchange - https://data.humdata.org/dataset/f5f0648e-f085-4c85-8242-26bf6c942f40

B. Libraries:

Pandas
GeoPandas
Numpy
Geoviews
Colorcet
Holoviews
Bokeh
DataShader
Statsmodels OLS
Matplotlib
Seaborn
Sklearn

a) Test, Train split
b) Metrics(R2)
c) Ensemble -> Random Forest

C. Tools

Python
Jupyter Notebook
Power BI

D. Project Repo: https://github.com/UQ-G12/DATA7001

Appendix D

Table 11 Calculated Columns

Calculated Column	Description	Calculation
promise_date	Determines whether delivery is within estimated delivery date	order_delivered_customer_date <= order_estimated_delivery_date
actual_lag_time	Total time from customer order purchase to order is handed over to courier	order_delivered_carrier_date - order_purchase_timestamp
estimate_lag_time	The interval between when a customer orders the products to when a seller must send those products to a carrier	shipping_limit_date - order_purchase_timestamp
seasons	Determines whether month is dry or wet season	Dry: Sept, Oct, Nov, Dec, Jan, Feb Wet: Mar, Apr, May, June, July, August
distance	Calculated skyway distance between a customer and a seller	Distance in km between two latitude/longitude points using the haversine formula
actual_delivery_time	Total time of delivery from order is handed over to courier until it reaches the customer	order_delivered_customer_date - order_delivered_carrier_date
total_delivery_time	Total time from customer purchase until order is delivered to them	order_delivered_customer_date - order_purchase_timestamp
product_volume	Combined Volume of the products in an order (cm^3)	*product_length_cm * product_height_cm * product_width_cm*
connection_between_states	Assigned value for each pair of states that determines the degree of connection between them	1. Group the dataset by customer_state and seller_state & perform mean aggregation on total delivery time column. 2. Sort the aggregated dataset by total delivery time in the descending order. 3. For every row, assign a value to connection_between_states field, which corresponds with the order of that row, starting from 1 to the total number of pairs of states. ⇨ The lower value for connection_between_states, the higher average delivery time and vice versa.

connection_between_cities	Assigned value for each pair of cities that determines the degree of connection between them	1. Group the dataset by customer_city and seller_city & perform mean aggregation on total delivery time column. 2. Sort the aggregated dataset by total delivery time in the descending order 3. For every row, assign a value to connection_between_cities field, which corresponds with the order of that row, starting from 1 to the total number of pairs of cities. ⇨ The lower value for connection_between_cities, the higher average delivery time and vice versa.
freight_ratio	Freight value to price ratio	Freight Value / Product Volumetric Weight
seller_delivery_PrevMean	Calculates the historical total delivery time performance of each seller	Rolling Mean of actual delivery time for each seller for each city combination