CIVE60008 - Transport Systems Coursework

Part 2 – Individual Tasks

Name: Mikaella Liveri

CID: 02032739

Table of contents

1. Introduction 1
2. Individual Tasks 1
2.1 Discussion of forecasting methods 1
2.1.1 Limitations 1
2.1.2 Improvements 3
2.2 Annual value of orders and valuable customers 4
2.2.1 Expected total value of orders per year 4
2.2.2 Identification of the most valuable customers 10
2.3 Determining set of 10 facility locations using the k-means algorithm 13
2.4 Inherent Limitations of the k-means algorithm 15
2.5 Facility location problem 17
2.6 Mathematical formulation of the facility location problem 17
2.7 Vehicle routing problem 19
2.8 Mathematical formulation of the vehicle routing problem 21
2.8.1 Reflection on overall demand and resolving for weekly demand to make business sense 23
2.8.2 Information needed to properly evaluate merits of plan 25
3. Conclusions 26
4. References 27

1. Introduction

This report analyses customer orders made from an online electronic goods retailer wanting to design a parcel last-mile distribution network that will serve customers in Rio De Janeiro, in Brazil. A dataset containing a randomly picked 5% subset of the overall activities in the region has been provided, which was used to extract specific customers and identify appropriate forecasting methods in the first part of the coursework. The second part which follows below, discusses the limitations of these forecasting methods, estimates the expected total value of orders per year and describes approaches for identifying valuable customers. The parcel last-mile distribution network is later designed after finding the optimal facility locations for the distribution network using the k-means algorithm, and implementing the facility location problem and the vehicle routing problem. Assumptions and limitations are discussed throughout.

2. Individual Tasks

2.1 Discussion of forecasting methods

2.1.1 Limitations

The forecasting method deemed as the most appropriate in Part 1 is the Holt Winter's Method. This is a time series forecasting method which forecasts seasonal data with trend b and level of demand a. The function is controlled by parameters α, β, and γ with the resulting model defined in Equation 2.1.

$$\hat{y}_{t+\tau} = (\hat{a}_t + \hat{b}_t t)F_{t+\tau-P} \tag{2.1}$$

Where,

$$\hat{a}_t = a\left(\frac{y_t}{\hat{F}_{t-P}}\right) + (1-\alpha)\left(\hat{a}_{t-1} + \hat{b}_{t-1}\right)$$

$$\hat{b}_t = \beta\left(\hat{a}_t - \hat{b}_{t-1}\right) + (1-\beta)\hat{b}_{t-1}$$

$$\hat{F}_t = \gamma\left(\frac{y_t}{\hat{a}_t}\right) + (1-\gamma)\hat{F}_{t-P}$$

Even though this method was found to perform the best for the fitting of electronic goods' demand, and it is widely used, it still has a number of limitations. Firstly, it assumes that the time series data follows a linear trend and seasonality. It is designed to handle data with regular, repeating seasonality. This would mean that when patterns are nonlinear or when the data exhibits sudden changes, it may not perform well. It will struggle to capture sudden changes since it relies on historical patterns, which is proved in Figure 2.1, generated in Task 1.3. As can be observed, there is an unexpected peak in the demand for electronic goods in December of 2017. Since this is a sudden change in demand, the model does not account for it. Even though we would be expecting a seasonality in the provided data given the nature of the products, it is not very well defined. For example, sales in electronic goods tend to go up over Christmas time where people spend more and buy gifts. It is observed that indeed over the period of December of 2017 the demand shoots up, comes instantly down, and shoots back up again.

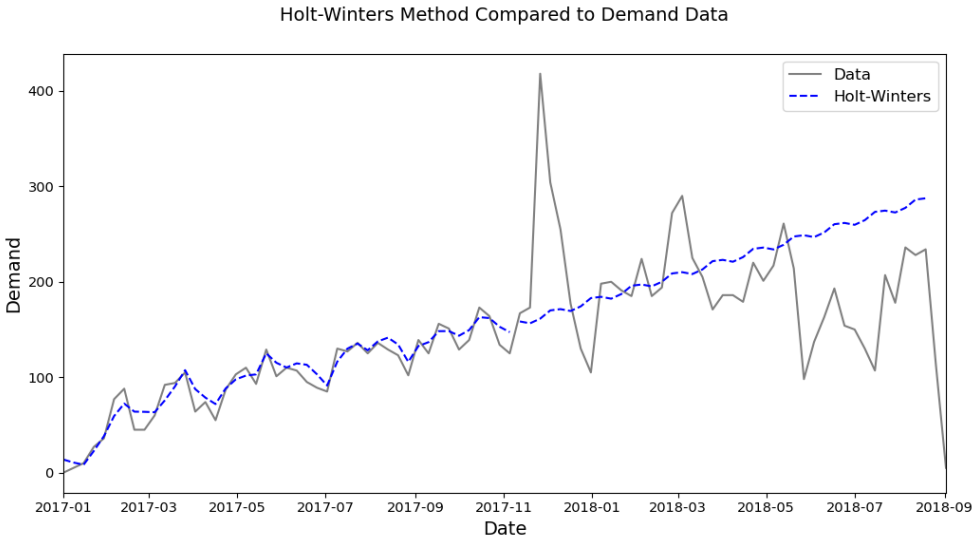


Figure 2.1: Holt-Winters method forecasting.

Additionally, even though Holt Winter's captures seasonality, it still assumes a linear trend. Even though Figure 2.1 displays that there seems to be an overall increase in the demand over the provided time series, the model would suggest that this increase is indefinite. It is very unlikely that the trend will keep increasing linearly in the long term, due to the nature of the product, but also due to unexpected events or even population stagnation. While it may predict well for the near future, it would be quite inaccurate for long term predictions where it may over-forecast. A way to improve this would be using the Holt-Winter's damped method. This can be done in Python by setting 'damped=TRUE' inside the 'ExponentialSmoothing' function. Figure 2.2 shows the difference in the prediction when this is performed. A parameter ϕ is introduced which 'dampens' the trend to a flat line at some point in the future. In fact, the forecast converges to $l_T + \phi b_T / (1 - \phi)$ as $h \to \infty$ for any value $0 < \phi < 1$, where l_T is an estimate of the level of the series, b_T is an estimate of the trend of the series, and h is the step ahead. This implies that long-run forecasts are constant, whereas short-run ones are trended (Hyndman & Athanasopoulos, 2018).

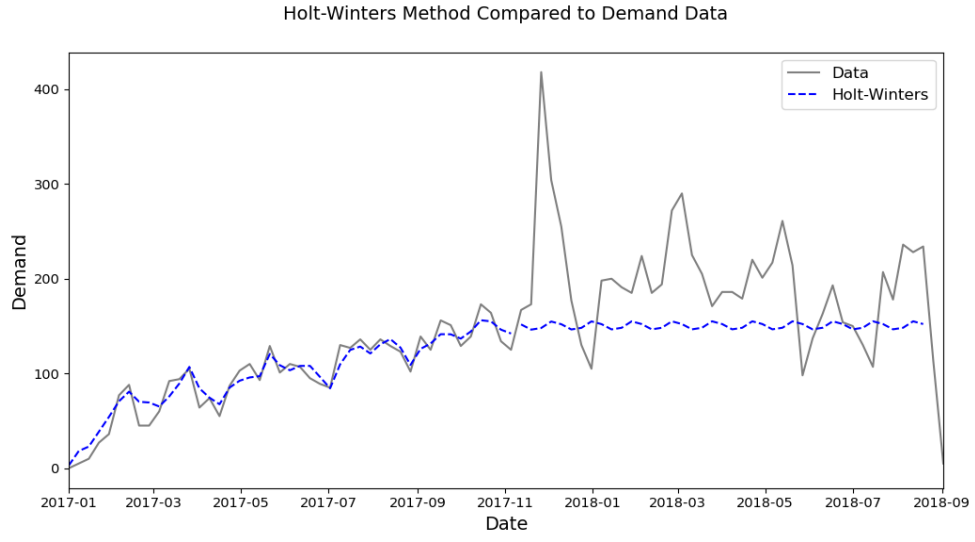


Figure 2.2: Holt-Winters damped method forecasting.

When it comes to limitations on how the Holt Winter's Method was fitted on the given dataset, the seasonality was tested for a period of 4 weeks. However, by observation of the trend from the figures produced, there seems to be cyclical movements, which are periodic events not tied to the calendar over a period greater than 2 years. More data would therefore be required to analyze this further. Additionally, the first 13 weeks were excluded from the data used for training the model. The demand over that period was very minimal and was not deemed as representative of the overall trend. The reason for this could be that during that period the electronic goods retailer was first being launched, or renovation works were being carried out, leading to a temporary shut of the online store. If it is the latter case, scheduled renovation works might occur at regular monthly intervals. This would be accounted as seasonality, and the fact that this data was excluded would mean that in the future it will not be anticipated nor predicted by the model.

Holt Winters is also very sensitive to the values of trend, seasonal components, and level and there is therefore difficulty in parameter tuning. While care was taken to select values that fitted the data well, any small changes to these values can lead to a significantly different forecasts of future demand. Finding the optimal values for smoothing parameters α, β and γ, requires manual tuning and was found to be quite challenging. While automated methods exist and could be of help, they are still imperfect.

2.1.2 Improvements

A number of potential improvements exist to create better forecasting models, some of which are explained below.

In the future, the demand could rapidly increase due to a rapid expansion of the electronic goods retailer, or even decrease for reasons such as high market competition. A larger dataset covering more than 2 years would make future predictions more reliable. Sector analysis can also be used to provide judgement on how well the company in the sector is expected to perform. This is an assessment of the financial and economic conditions of a given sector in the economy. It is based on the premise that certain sectors perform better during stages of the business cycle. A business cycle consists of expansions due to periods of economic growth and contractions due to periods of economic decline (Kenton, 2021).

Time series forecasting could also be performed with Prophet in Python instead. The Prophet library is an open-source library which is designed to make forecasts for univariate time series datasets. It automatically finds hyperparameters for the forecasting model and makes forecasts for data with seasonal structure and trends (Brownlee, 2020). Prophet is robust to missing data and shifts in the trend, while it handles outliers well in comparison to the Holt-Winter's Method which doesn't.

Another improvement would be to use advanced machine learning models like gradient boosting machines which would allow for complex patterns and non-linear relationships to be captured. It works by creating simpler (weak) prediction models in a sequence, where each of these models try to predict the error left over the previous model. A weak learner is a model that performs slightly better when compared to random predictions (Dash, n.d.).

Furthermore, hybrid models combining Holt-Winters with another forecasting method or a machine learning model could also be used. This is an approach that would leverage the strengths of the different techniques used, helping to minimize the impact of individual model limitations and provide predictions which are more robust.

Lastly, automatic model selection could be explored which involves techniques that automatically choose the best forecasting model given a dataset. An example is using Bayesian Optimization which is based on statistical modelling and intelligent exploration of the parameter space (Nag, 2021).

2.2 Annual value of orders and valuable customers

2.2.1 Expected total value of orders per year

The expected total value of orders per year is the value of the yearly orders. Therefore, information about prices and demand per year are needed. The overall method employed to get the expected total values of orders, was finding the total demand in a year and multiplying it by the ratio of price per demand. The dataset given covers 5% of the overall activities in the region and therefore the answer should be later multiplied by 20 to get the activities of the whole area.

In order to check whether calculating the mean price is an appropriate metric for calculating the price per demand, the distribution of prices in the sample was investigated. Figure 2.3 displays a box plot of the price data in the sample. It is observed that it is positively skewed, and hence finding the total average price would not give representative results.



Figure 2.3: Box plot of price data.

It is also observed that there is a significant number of outliers. The percentage of outliers is computed in Equation 2.2 below.

$$percentage_{outliers} = \frac{number\ of\ outliers}{number\ of\ price\ data} \times 100 = 7.40\% \tag{2.2}$$

Even though this is not a significant amount of the data, it is an important subset considering that the dataset itself is only a 5% subset of the overall region activity. Therefore, simply filtering out the outliers would not yield a representative result.

The price data was therefore divided into 4 subsets based on the price quartiles. The following assumptions were made:

- No inflation of prices in future years. The prices of electronic goods remain the same throughout all years for which the expected total value of orders is calculated.
- Holt Winter's method is an appropriate model for forecasting future demand within the subsets.

- The parameters used for the Holt Winter's method are the same as the ones used in Task 1.3.
- Mean price remains constant within each subset.

Table 2.1 shows the subsets of quartiles used, with the price ranges and the price per demand of each. The corresponding demands were extracted for each price range quartile, and 4 separate predictions were executed to find the forecasts of chosen years. Within each subset, the price per demand was multiplied by the total estimated demand to get the total value of orders.

Table 2.1: Price information for each subset.

Subset	Range	Price range [R$]	Price per demand [R$/unit of demand]
1	$min - Q1$	0.85 - 45.90	28.81
2	$Q1 - Q2$	45.90 - 84.90	63.24
3	$Q2 - Q3$	84.90 - 144.41	110.77
4	$Q3 - max$	144.41 – 4799.0	318.17

A decision had to be taken regarding the years for which the expected total value of orders should be estimated. Figure 2.4 shows a histogram of the value of orders per month from the dataset. It can be observed that there is very limited data within the year of 2016. This could be the period when the online electronic goods retailer was first being launched or a period of an unexpected event. It was therefore decided to exclude the 2016 data, and perform forecasts starting from the year of 2017. The last year the forecast was performed was for 2020, due to the unprecedented events which happened back then, explained further in the report.

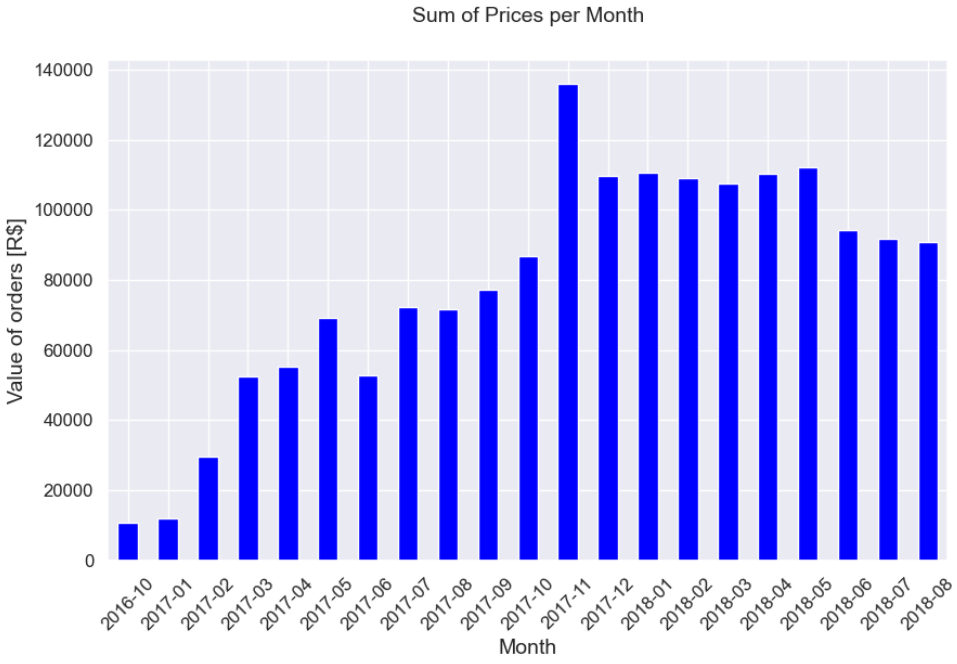


Figure 2.4: Value of orders per month.

In order to obtain the expected total value of orders per year from the region, it is first needed to predict the demand for the upcoming years of 2019 and 2020, as well as make a fit to the observed data. Since in Part 1.3, Holt Winter's method was deemed as the most appropriate, it is the forecasting method used here as well. The model was trained and fitted for all the dataset, before it predicted the demand until

01/01/2021, as seen in Figure 2.5. Then the expected demand values were extracted for each year considered, as showcased in Table 2.2.

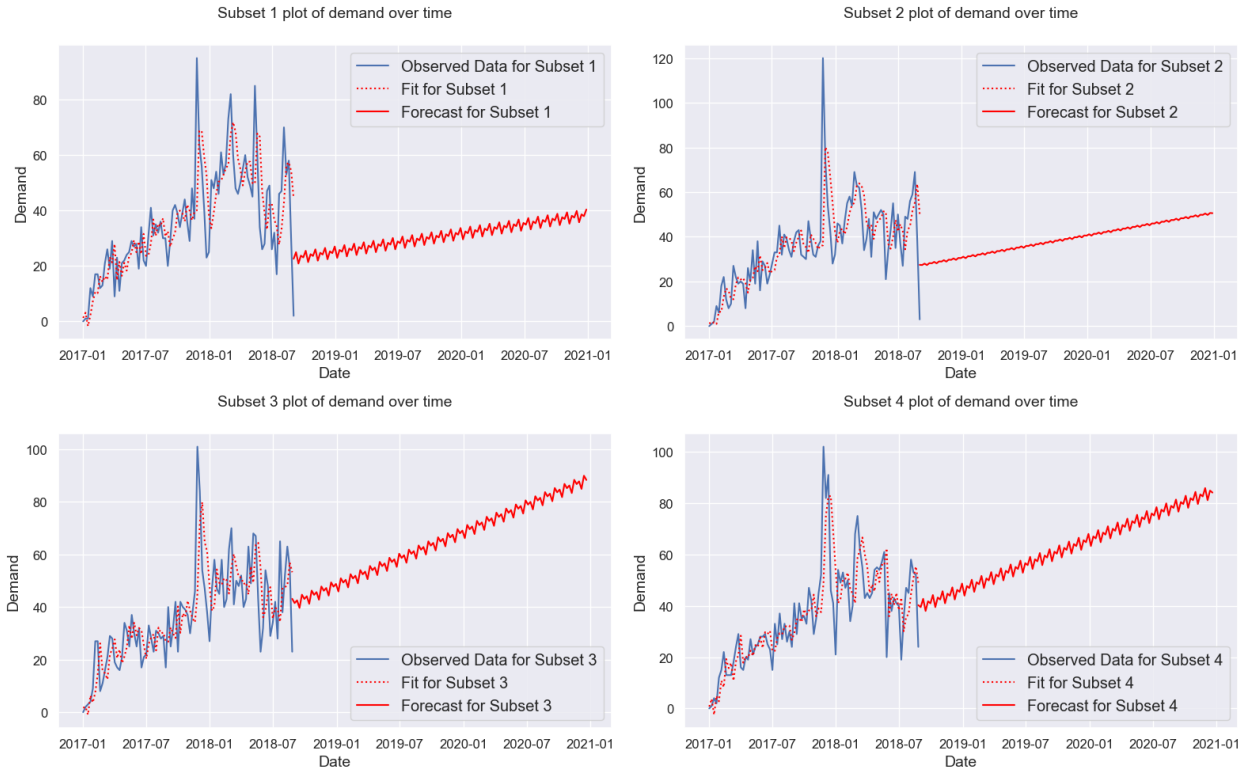


Figure 2.5: Forecasts of demand for each subset.

Table 2.2: Expected total value of orders per year from the orders in the sample.

Year	Subset	Total forecasted demand	Expected value of orders [R$]	Total value of orders [R$]
2017	1	1488.17	42877.39	805,191.69
	2	1543.76	97628.30	
	3	1530.01	169473.16	
	4	1556.46	495212.84	
2018	1	2155.012	62090.86	1,236,311.49
	2	2157.61	136448.83	
	3	2453.19	271730.83	
	4	2407.67	766040.98	
2019	1	1471.11	42385.83	1,415,868.17
	2	1843.84	116605.79	
	3	3021.60	334691.51	
	4	2898.44	922185.04	
2020	1	1818.05	52381.92	1,895,817.35
	2	2366.87	149682.97	
	3	4076.67	451556.77	
	4	3904.23	1242195.69	

Since it is asked to find the expected total value of orders per year from the entire region and not just the sample used which includes 5% of overall activities, the obtained answers are divided by 0.05. Table 2.3 lists the final expected total value of order per year from the entire region.

Table 2.3: Expected total value of order per year from the entire region.

Year	Expected total value of orders [R$]
2017	16,103,833.86
2018	24,726,229.71
2019	28,317,363.42
2020	37,916,346.98

In order to check whether the values reached are accurate, comparison with the actual value of orders over a year is performed. The only year with full data provided is 2017. The total observed value of orders was calculated as R$ 16,489,912.80.

Equation 2.3 below calculates the percentage difference between the observed and expected total value of orders of 2017. The percentage difference is very low, and hence the values reached can be deemed as reliable.

$$\%\, difference = \frac{abs(expected - observed)}{average} = 2.37\% \tag{2.3}$$

The approach explained above seems to perform well, however it did not account for the unprecedented events that happened in 2020 due to the COVID-19 pandemic outbreak. As explained, it is hard for models to predict for unexpected events. Figure 2.6 shows the volume of consumer electronics in Brazil from years 2018 to 2028 (Statista, n.d.).

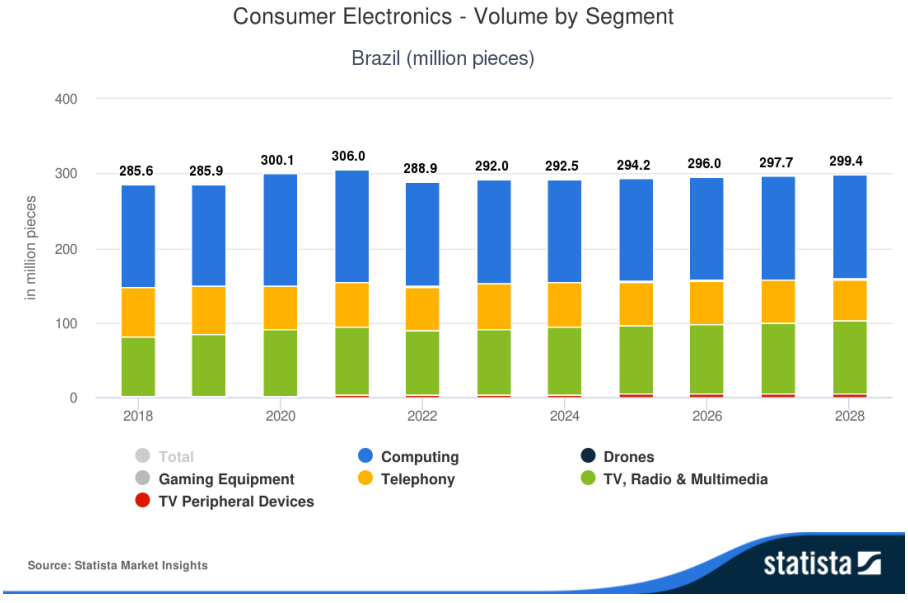


Figure 2.6: Volume of consumer electronics in Brazil (Statista, n.d.).

It is observed that there is an increase in the volume of consumer electronics from 2019 to 2020. This percentage increase is calculated as 5%. Therefore, another approach to calculate the expected total value of orders for 2020 is to account for this percentage increase. Figure 2.7 plots the subsets again, but this

time multiplying the demand of 2020 by 1.05 to account for this increase. Table 2.4 shows the corresponding expected total value of orders for 2020.

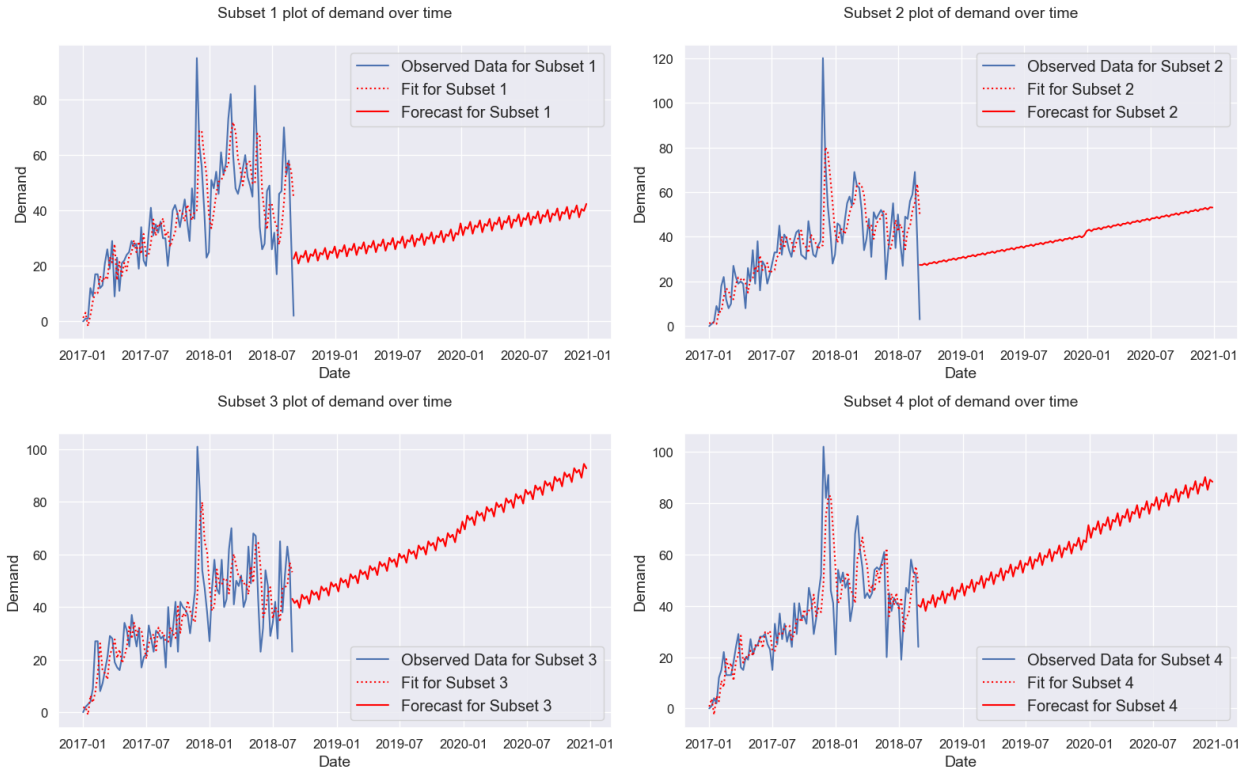


Figure 2.7: Forecasts of demand for each subset considering COVID-19 effects.

Table 2.4: Expected total value of orders in the region for 2020 accounting for the increase in demand due to the pandemic outbreak.

Year	Expected total value of orders from sample	Expected total value of orders from entire region
2020	R$ 1,990,608.22	R$ 39,812,164.33

It is interesting to consider the reasons for such an increase in demand. The restrictions during the pandemic made companies transition into remote working systems leading to an increase in demand for electronic equipment. Additionally, the increase in time spent at home and the need to stay connected rely on consumer electronics, which also leads to an increase in demand (Rebound Electronics, n.d.).

2.2.1.1 Limitations

Even though reasonable answers have been reached, there are a number of limitations in the approach taken. Firstly, the model used to predict the future demand is the Holt Winter's method which forecasts seasonal data considering both trend and level of demand. It displays a constant trend indefinitely into the future. This method could therefore over-forecast for longer forecast horizons (Hyndman & Athanasopoulos, 2018). This creates a variation in the result depending on the year chosen to perform the forecast. For years later into the future, the expected total value of orders increases indefinitely, which is not realistic in real life. A method to ensure the demand does not increase indefinitely into the future is to apply a linear regression model back into the initial model in order to minimize this demand inflation. This process would remove the linear trend from the original time-series data before the application of the Holt Winter's method, which is useful when there is a need to focus on the non-linear components of

a time series. A second method would be to use damped trend methods. A parameter is introduced that dampens the trend to a flat line sometime in the future. This could work quite well here if it is assumed that the initial increase is due to the launch of the retail store and that at some point sales become more stable. Figure 2.8 shows the subset forecasts with the damped trend method, and Table 2.5 the corresponding expected total value of orders per year from the entire region.

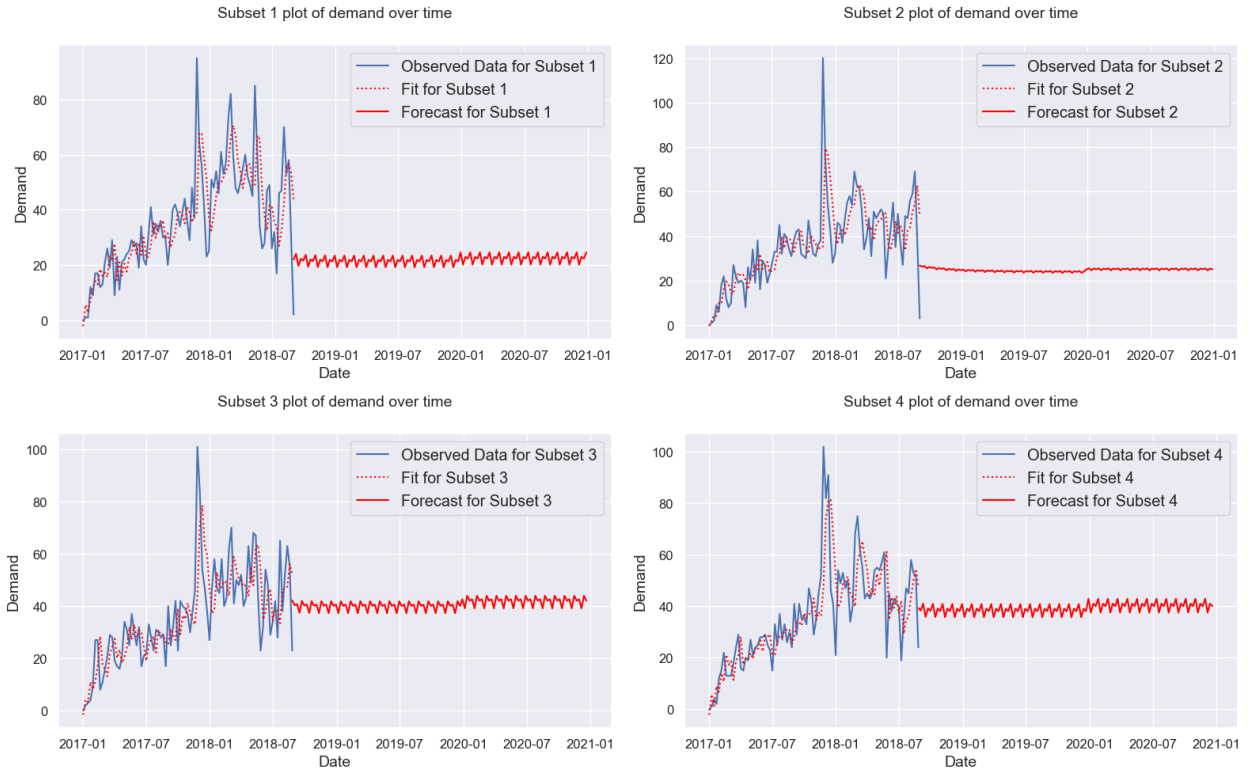


Figure 2.8: Damped Forecasts of demand for each subset.

Table 2.5: Expected total value of orders of entire region using damped forecasts.

Year	Expected total value of orders [R$]
2017	R$ 15,955,091.32
2018	R$ 23,589,540.68
2019	R$ 19,560,365.56
2020	R$ 20,523,247.32

Additionally, as mentioned, the 2016 data was removed due to the insufficient amount of available data for that year. This decision was made to ensure that the dataset used for the analysis is representative and robust, reducing the impact of data sparsity on predictions. However, it could be that there is an underlying trend and another reason as to why the demand was zero back then. Given that the dataset contains only a randomly picked 5% subset of the overall activity, it could also be that 2016 had much more data which randomly fell within the other 95% subset of the overall activity. Excluding the 2016 would mean that in reality even less than 5% of the overall activity is represented in the model, which could be deemed as too small of a subset to make reliable predictions regarding all activities in the region.

Extrapolating into the future imposes more risks on predictions for external events. An example here is the COVID-19 pandemic affecting the demand for electronic goods in 2020. Figure 2.8 displays an approximation, trying to account for this increase. However, in reality the demand for electronic goods is

obviously not a linear increasing line. This increase in demand did not happen from one day to another, but rather gradually over the whole year.

The last model limitation is its sensitivity to the α and β parameters. In a period of rapid change, such as the COVID-19 outbreak, more weight should be placed towards recent data and hence α should be greater. A way to increase the reliability of the model would be to have varying α and β parameters across the prediction period.

2.2.2 Identification of the most valuable customers

The most valuable customers to the electronic goods retailer could be interpreted in a number of ways. It could be the ones that have spent the most money, the ones placing the most orders, or even the ones with the greatest order weight. A big weight could mean that a customer has bought more than 1 item within a single order. Unique customers have been defined by unique locations, assuming that each individual customer has placed an order from the exact same location each time. In such a way, the total amount spent as well as the number of orders placed by each unique customer can be aggregated to filter out the most valuable customers. Given the limited data given, this assumption can be a good indication of the most valuable customers. However, there are limitations to this approach. For example, in reality the same customer could be placing an order from a different location each time.

Figure 2.9 plots the location of the 20 most valuable customers in terms of total expenditure, total purchased weight and number of orders. Table 2.6 ranks the 20 most valuable customers for each as well.

Table 2.6: 20 most valuable customers in terms of Expenditure, Product Weight and Number of orders.

Rank	Expenditure	Product weight	Number of orders
1	C_76	C_76	C_76
2	C_65	C_133	C_25
3	C_315	C_25	C_65
4	C_25	C_31	C_31
5	C_31	C_65	C_133
6	C_133	C_250	C_250
7	C_250	C_152	C_315
8	C_145	C_119	C_115
9	C_152	C_315	C_152
10	C_236	C_392	C_97
11	C_163	C_145	C_12
12	C_295	C_148	C_163
13	C_119	C_95	C_145
14	C_266	C_382	C_148
15	C_97	C_115	C_155
16	C_392	C_80	C_392
17	C_12	C_236	C_154
18	C_154	C_12	C_236
19	C_293	C_293	C_38
20	C_110	C_5	C_149

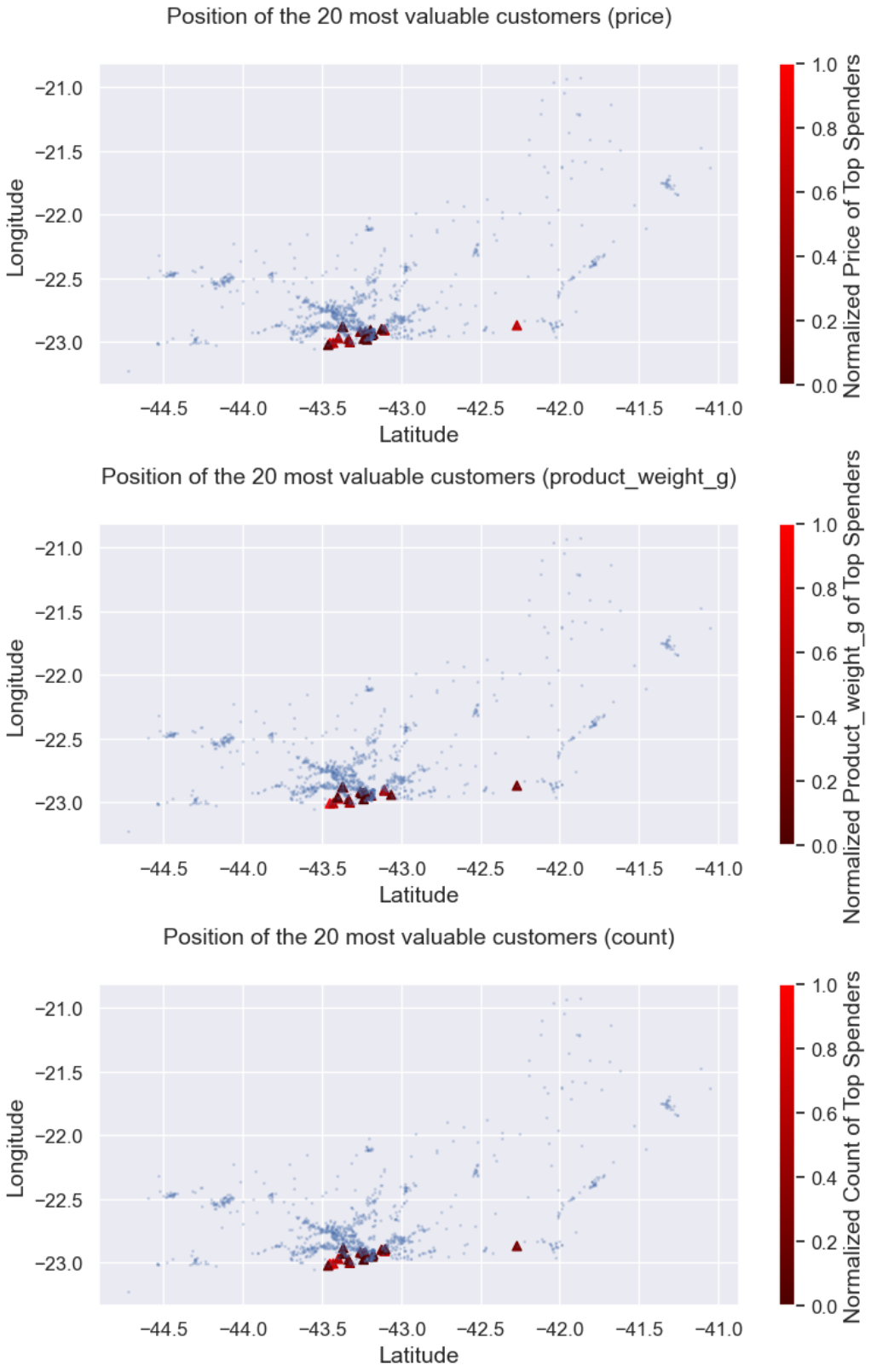


Figure 2.9: Locations of 20 most valuable customers in terms of expenditure, weight and number of orders.

It is observed that some of the customers fall within the top 20 lists for more than one of these parameters, especially when comparing the total expenditure with the order quantity metric and product weight. The relationships of total expenditure with order quantity, and product weight are further analysed using Figures 2.10 and 2.11 respectively. It can be seen that there is a linear correlation in both plots with a correlation coefficient of 0.95 and 0.91 respectively. This conclusion supports the statement of customers being in more than one top 20 list.

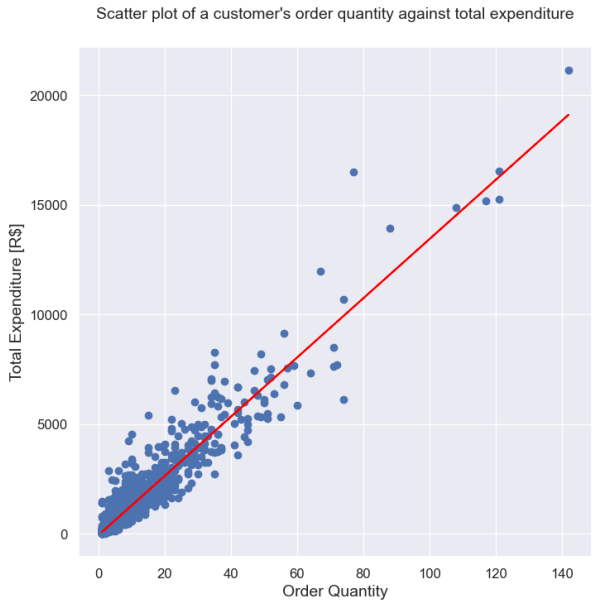


Figure 2.10: Scatter plot of expenditure over order quantity.

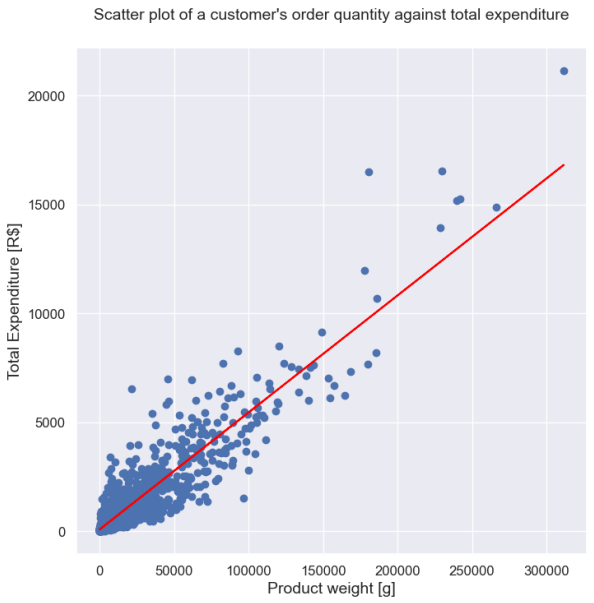


Figure 2.11: Scatter plot of expenditure over product weight.

Valuable customers can also be defined by the frequency of placed orders. A loyal customer would be someone who orders regularly from the electronic goods retailer. This does not necessarily mean that their total expenditure is high, as it could be less than someone who placed a single expensive order across the whole 2 years.

Another way to filter the most valuable customers can be combining two or more of these metrics, such as the order frequency and total spending. An approach is to calculate the weighted score for each customer combining order frequency, which is the inverse of the time between consecutive orders, and total spending. The formula used here as an example, assignes a weight of 0.7 to total spending and 0.3 to the inverse of order frequency, which is how often a customer bought an electronic good. Table 2.7 shows the 5 most valuable customers using this approach, with Customer 'C_76' taking the lead. It is also observed that customer 'C_76' was the top ranking customer for expenditure, total product weight and number of orders alone. It is therefore concluded that he is the most valuable customer in the dataset.

Table 2.7: Top 5 customers based on combined metrics of order frequency and total expenditure.

Customer ID	Time between orders	Total spending [R$]	Combined Score
C_76	4 days 00:28:14	13413.30	9389.31
C_315	7 days 01:37:53	12817.90	8972.53
C_65	4 days 17:16:04	9009.08	6306.36
C_133	5 days 06:20:56	7436.60	5205.62
C_250	6 days 05:50:09	6620.70	4634.49

Adding to the aforementioned, the retailer could define valuable customers as the ones with the closest distance to the facility from which the goods are dispatched for delivery. This would reduce the retailer's shipping costs, making it more efficient for them to deliver the electronic goods.

In conclusion, there are a number of ways to evaluate valuable customers and this simply depends on the online retailer's priorities.

2.3 Determining set of 10 facility locations using the k-means algorithm

This section of the report determines a preliminary set of 10 facility locations using the k-means algorithm. These facility locations are to be used as distribution warehouses that will support customer deliveries. The k-means algorithm is one of the most commonly used clustering algorithms, and is based on an iterative approach.

The objective of the k-means algorithm, is to identify groups with similar characteristics, given a collection of elements with a common set of properties. In this specific example, the objective it to identify delivery customers that could be served by the same facility. The algorithm determines the 10 distinct centroids of the clusters such that they provide the minimum distance between customers and the cluster centroid.

Elements of the group of nodes are represented as vectors $\vec{x}$. A cluster c would be centered around a centroid (mean) vector $\vec{\mu}$, that represents the cluster's centre of gravity, determined using Equation 2.4 where $|c|$ is the number of elements in the cluster.

$$\vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x} \tag{2.4}$$

The k-means algorithm requires an initial set of centroids to start the clustering process. These were initialized randomly, using the 'random_state=739' parameter in Python which seeds an underlying random number generator, making the randomness reproducible. The customer coordinates are then assigned to the nearest centroid and the centroids are recalculated, with this process being repeated until convergence, which is when the centroid locations no longer change.

Assumptions:

- Storage facilities have sufficient capacity for all goods imported and exported.
- Construction of storage facilities is physically viable at all cluster centroids identified.
- Neglecting considerations of congestion on goods to and from facilities at regional centroids.

Figure 2.12 plots the 10 optimal clusters with their corresponding cluster centroids, after implementation of the 'sklearn' package in Python. The customers belonging to the same cluster are plotted in the same colour to distinguish the different clusters. Table 2.8, summarises the 10 determined facility locations, specifying their coordinates, total demand and total weight that each cluster captures. Figure 2.13 plots the facilities with sizes proportional to the weight they aggregate. It is worth noting that the latitude and longitude coordinates in the provided data were swapped. Therefore, for plotting purposes they were swapped back correctly.

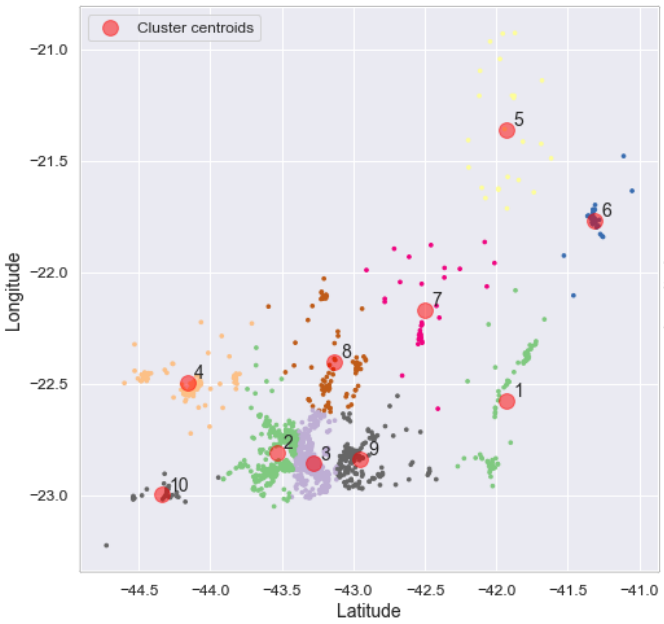


Figure 2.12: Plot of preliminary set of 10 facility locations using the k-means algorithm.

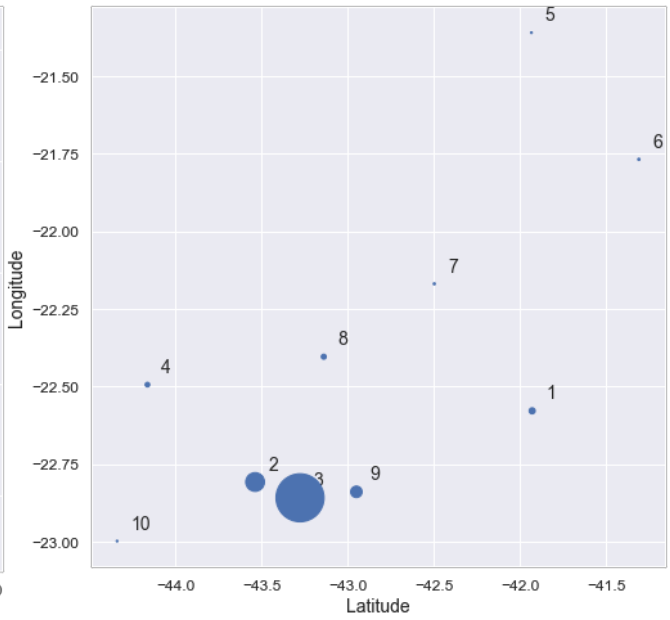


Figure 2.13: Plot of preliminary set of 10 facility locations with proportional aggregates weight

Table 2.8: Facility coordinates and aggregated total demand that eac h cluster captures.

Cluster	Latitude	Longitude	Total demand based on order quantity	Total demand based on product weight [g]
1	-41.93	-22.58	743	1,749,203
2	-43.54	-22.81	2377	5,357,338
3	-43.28	-22.86	6033	13,703,042
4	-44.16	-22.49	635	1,323,803
5	-41.93	-21.36	238	506,777
6	-41.31	-21.77	303	655,339
7	-42.50	-22.17	282	611,231
8	-43.14	-22.40	578	1,436,685
9	-42.93	-22.84	1430	3,282,898
10	-44.34	-23.00	126	509,918

The calculated inertia provides a measure of how well the kmeans has performed, equal to 35.67 here.

2.4 Inherent Limitations of the k-means algorithm

While the k-means algorithm is a popular clustering algorithm, it has several inherent limitations which could impact its performance to the underlying model of clusters.

Firstly, the final clustering is highly sensitive and dependent on the initial centroid selection. The initial cluster centroids were randomly generated as explained above, and this random initialization can result in convergence to suboptimal solutions, and different solutions every time if a seed is not set.

The process also requires an input of a predefined number of clusters, which may not always be the optimal one. Usually, finding the ideal number of clusters is part of the problem, and there is a need to undertake a trade-off between distances and groups. More clusters will result to lower travel distances to end customers. However, they will incur increased warehouse management and coordination costs. When more information is available, including costs, the elbow method could be used to identify the optimal number of clusters (Angeloudis, 2023a). An incorrect choice of the number of clusters may lead to formation of meaningless clusters or the merging of distinct clusters. The elbow method was applied here, as seen in Figure 2.14. After plotting the inertia values for a range of k values, the value that lies in the inflection point of the curve is chosen, in this case 5. Therefore using 5 clusters instead of 10 would be more optimal.

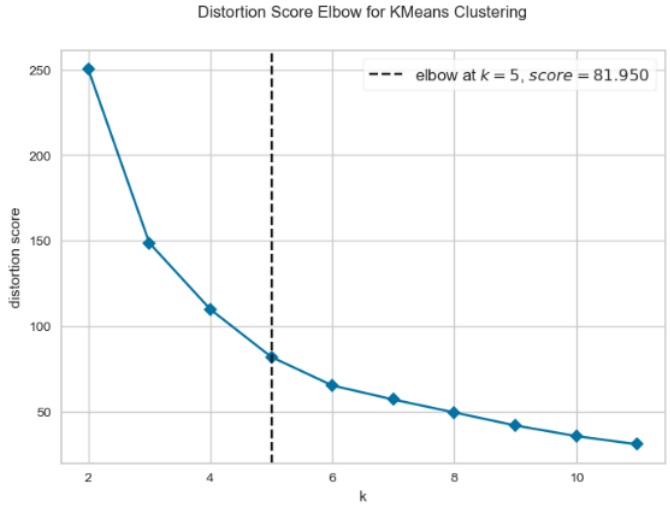


Figure 2.14: Plot of the Elbow Method.

Additionally, if a warehouse facility is not able to meet the demand of the region it serves, some regions will not be supplied with sufficient amounts of cargo. This would also mean that if a facility is shut down, the whole model is susceptible to failure. It therefore does not propose resilience and robustness which are critical for designing modern transport networks.

K-means clustering is sensitive to outliers, as it minimizes the sum of squared distances. Outliers can disproportionally influence the position of cluster centroids. It also assumes equally sized circular clusters, while it struggles with clusters of different shapes. The Euclidean distance formula is used to find distances, which calculates the distance between two points as $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, where the x coordinate corresponds to the latitude and the y coordinate to the longitude. However, the latitude and longitude values are not in kms and the Euclidean distance formula does not take into account the curvature of the Earth's surface, which causes inaccuracies.

Lastly, it does not take into account that some cluster locations may not be physically viable for the construction of facilities. This is proven in Figure 2.15, which shows the cluster locations on a map of Rio De Janeiro. Facility 1's coordinates obtained from k-means clustering are defined in the ocean rather than on land (Figure 2.16). Facility 8 has coordinates which lie exactly on a main road (Figure 2.17), while Facility 9 lies on residential houses (Figure 2.18). Therefore, facilities would not be able to be built at such locations, and a procedure which takes into account such limitations would be better.



Figure 2.15: Facility locations plotted on map of Brazil.

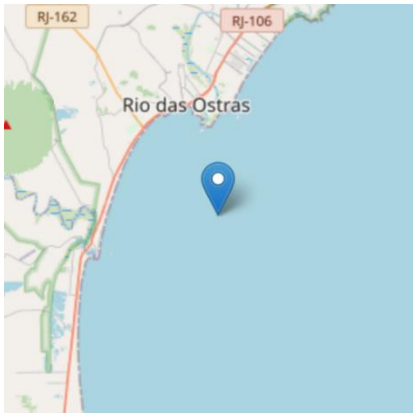


Figure 2.16: Facility 1 location.

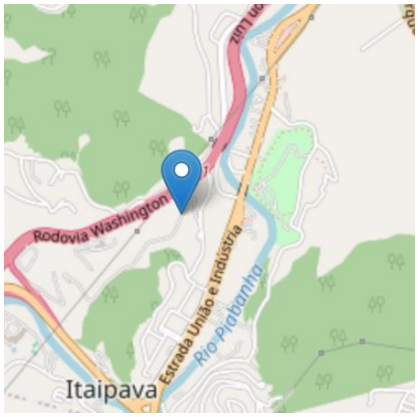


Figure 2.17: Facility 8 location.

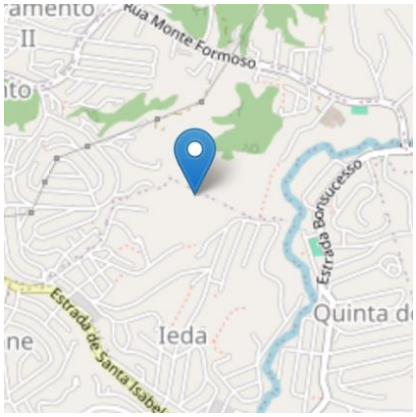


Figure 2.18: Facility 9 location.

2.5 Facility location problem

A version of the facility location problem is implemented in Python, to choose 1 facility out of the 10 seen in Figure 2.19 that would allow minimization of the total distance between customers and their assigned facility. Figure 2.19 shows in green the facility chosen, Facility 3, after implementing the facility location problem formulated in Section 2.6 of the report. The objective value of the problem, meaning the total transportation cost is 62,394.08 km.



Figure 2.19: Plot showing customers, facilities, and the chosen facility which minimises distance.

2.6 Mathematical formulation of the facility location problem

A facility location model can be used to build upon preliminary decision carried out using the k-means clustering, and incorporate operational considerations. The objective here, is to find the facility location that minimizes transportation cost to all destination points, meaning choosing one facility that would minimize the total distance in km between customers and their assigned facility. Given a matrix of distances among them, there is a need to pick the single best location for a facility that will serve the customers.

Table 2.9 shows the symbolic mathematical formulation for the facility location problem. However, a few assumptions are firstly listed below.

- There is sufficient land available at the locations of the cluster centroids, to allow for the construction of a facility.
- The transport network used in the model is assumed to be non-congestible.

- All customers and depots are identical in terms of material demands by customers, holding capacities at each facility, facility operation costs, and customer service level.
- The distance between customers and facilities can be calculated as the straight line distance between two points. The Haversine formula is therefore used to calculate the minimum distance between customers and the cluster centroids in this geospatial application. The Haversine formula calculates the distance between two points on a sphere, such as the Earth, using their latitudes and longitudes along the surface. This method is much better than the Euclidean distance formula which assumes a flat cartesian plane, hence more accurate results are obtained. It is based on spherical trigonometry which is appropriate here as the distance between lines of longitude varies with latitude on a sphere. Equation 2.5 shows the Haversine formula where r is the radius of the Earth, ϕ_1 and ϕ_2 are the latitudes of point 1 and point 2, and λ_1 and λ_2 are the longitudes of point 1 and point 2 (GeeksforGeeks, n.d.).

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_1 - \phi_2}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (2.5)$$

Table 2.9: Mathematical formulation of the facility location problem.

Sets:	
D	Set of customers.
S	Set of candidate facility locations.
Decision variables:	
x_{ij}	When customer i is served by facility j.
y_j	When facility j is open.
Parameter:	
d_{ij}	Distance from customer i and facility j.
Objective function:	
$\sum_{i \in D}\sum_{j \in S} x_{ij} d_{ij}$	Minimize the product of distances with the decision variables that indicate whether a facility and customer are connected.
Constraints:	
$\sum_{j \in S} x_{ij} = 1 \quad \forall i \in D$	Each customer will be served by 1 facility.
$\sum_{j \in S} y_i = 1$	Only 1 facility will be established as open.
$x_{ij} \leq y_j \quad \forall i \in D, j \in S$	Only established facilities can serve customers.
$x_{ij}, y_j \in \{0,1\} \quad \forall i \in D, j \in S$	x and y are Boolean (integers with values of 0 or 1).

The first step in order to solve the facility location problem involves finding the distance matrix between all of the cluster centroids found in Task 2.3, which are the candidates for the facility, and the customers. The distance matrix is calculated using the Haversine formula as displayed in the Jupyter notebook. However, this can be considered as a limitation of the problem since in reality the distance depends on the actual road network layout and not just the length of the straight line between two points.

The two sets of the model, D and S are defined as the set of customers and set of candidate facility locations respectively. Decision variables are values to be determined if the problem is solved, while they are linked to a quantifiable decision. The first decision variable, x_{ij}, is when customer i is served by facility

j, while the second one, y_j, is whether facility j is open or closed. Parameters are known values used to describe the problem premise, defined here as the distance between customer i and facility j. The aim of the task is to minimize the distance between customers and a facility location, and hence a few constraints are also defined. Constraints are equations that restrict values of decision variables. The first constraint is that each customer can only be served by one facility, therefore each customer i is assigned to exactly one facility j. Additionally, only one facility will be established and therefore the sum of the open facilities must be equal to one. This ensures that exactly one facility is open. The third constraint ensures that only established facilities can serve customers. It forces facility j to be activated after some customer i is assigned to j. Facility j is only allowed to serve a customer if it is open. Both x_{ij} and y_j are boolean variables, meaning integers with a value of either zero or one. Making them Booleans, allows for checks in Python, of whether facility j is open or closed, denoting *true* and *false*. The integer of 0 means closed and an integer of 1 means open. Lastly, the objective function is a measure against the optimization goal. There is a need to minimize the product of distances with the decision variables that indicate whether a facility and a customer are connected.

The facility location problem formulated in Table 2.9 is solved in python using PuLP, to find the facility location which optimizes the solution to the developed objective function. PuLP is an open-source linear programming package in Python which allows for the solving of linear programming optimization problems. As a result of solving the problem, facility S3 is opened, with an objective value of 62394.08km.

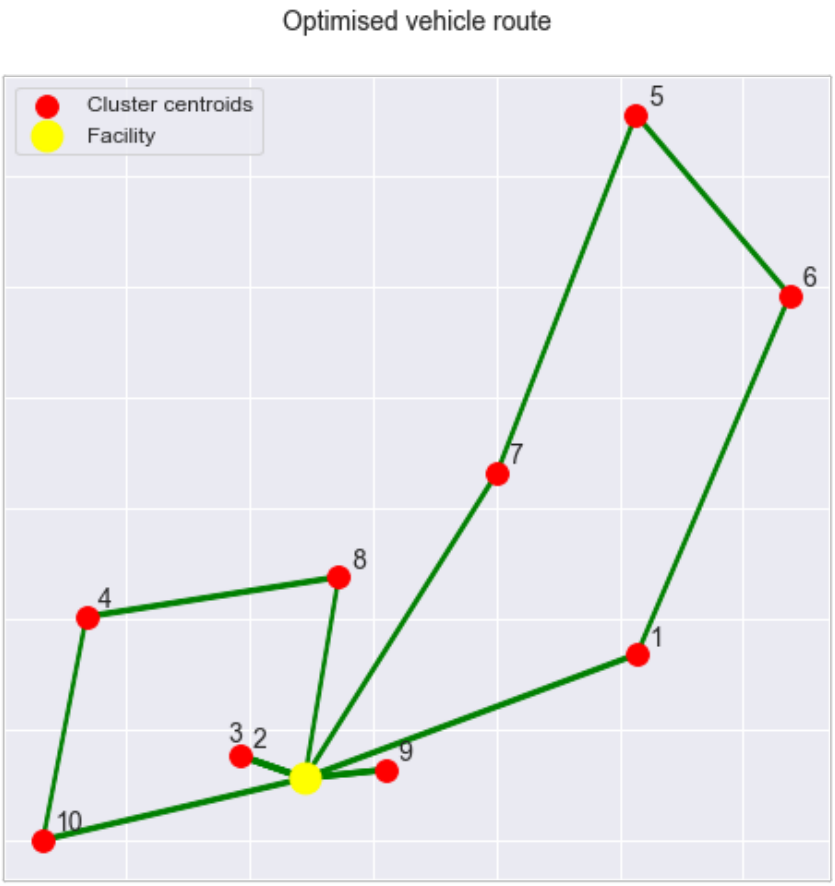
This result is not surprising considering that most of the customers are located in that part of Rio De Janeiro. This would mean that the distance between customers and the established facility is minimal. However, as can be observed in Figure 2.19, some customers seem to be outliers compared to the location of the majority and this increases the objective value.

2.7 Vehicle routing problem

The design of the parcel last-mile distribution network requested by the online electronic goods retailer serving customers in Rio De Janeiro, Brazil is finalized in this section of the report. Since the goods needed to be delivered have a weight, the delivery van would need to return back to the depot selected from the facility location problem, in order to pick up the appliances to be delivered in the next orders. The combination of customers to be visited at each trip needs to be determined, given a range of customer orders to be fulfilled by the vehicle of a given size.

A version of the vehicle routing problem, specifically the capacitated vehicle routing problem, is implemented here so that starting from the the facility selected in Task 2.5, vehicles visit all other cluster centroids obtained in Task 2.3. Hence, a total of 9 nodes should be visited by the delivery van. It is assumed that each vehicle can carry at most 4 tonnes of goods.

Figure 2.20 shows the route obtained after solving the mathematical formulation developed in Section 2.8 of the report. The routes taken are also displayed below. The number next to the nodes is the weight of cargo carried by the vehicle in grams after the visit to the specific node. In route 3 for example, this number decreases as the vehicle travels, since it delivers the required cargo.



Route 1:
S1, D2: 2,678,669 g
D2, S1: 0 g

Route 2:
S1, D3: 2,678,669 g
D3, S1: 0 g

Route 3:
S1, D7: 3,522,550 g
D7, D5: 2,911,319 g
D5, D6: 2,404,542 g
D6, D1: 1,749,203 g
D1, S1: 0 g

Route 4:
S1, D9: 3,282,898 g
D9, S1: 0 g

Route 5:
S1, D10: 3,270,406 g
D10, D4: 2,760,488 g
D4, D8: 1,436,685 g
D8, S1: 0 g

Figure 2.20: Optimised route of vehicle.

The objective value of the problem (total transportation cost) is 1026.72km.

However the aggregated weight demand included data covering nearly 2 years. Therefore, the way this problem was solved was assuming one vehicle every 2 years for only 5% of the overall activity in the region which does not make business sense. Section 2.8.1 of the report further evaluates this and solves the problem considering weekly demand for 100% of the region activity.

2.8 Mathematical formulation of the vehicle routing problem

There are a number of models and methods which can be used to solve this problem. The Sweep method uses a heuristic approach. It is a simple, geometric approach that can produce valid vehicle routing problem solutions very quickly. Since it is a quite crude heuristic approach however, these solutions are far from perfect, even though they can serve as a good starting point. The structure of tours depends on the choice of seed node, and if the depot does not coincide with the geometrical centroid, the tour lengths will vary significantly. Additionally, once a capacity constraint is violated, the tour is terminated, which might lead to capacity waste. The multiple travelling salesman problem does not yield an ideal solution either. Tours are uneven, but there is no way to take into account the capacity of the vehicle or the possibility that the nodes have different levels of demand. The savings method is another iterative approach, where at each iteration the largest known saving is considered, and added to a tour whose capacity is not violated. The algorithm terminates when all potential savings have been evaluated (Angeloudis, 2023b).

After considering the advantages and disadvantages of each method, the capacitated vehicle routing problem (CVRP) was deemed as the most appropriate. Even though it takes a lot of time to solve in larger problem instances, it is appropriate here where only 9 nodes should be visited. While the savings method is not the one explained in the formulation of the vehicle routing problem, it was also coded in Python to check that it yields the same result as the CVRP.

The objective here is that given a set of facilities with vertices and distinct demand levels, the sequence of vertices (tour) and number of tours that minimizes the total cost should be identified, reflecting the varying range of capacities. The tours are bounded by the payload capacity of the vehicle, set as 4 tonnes. The objective function seeks to minimize the total cost of electronic goods distribution. As before, assumptions need to be defined before the problem is formulated. These are listed below.

- Each vehicle can carry at most 4 tonnes of goods.
- There is only 1 vehicle.
- Total demand for each region is known and remains fixed. However in reality large fluctuations are expected.
- Capacity is not considered for the facilities. Warehouses do not have storage restrictions or throughput constraints.
- All tours start and end in the same depot vertex, Facility 3.
- All vertices are visited once (aside from the depot), except if their weight demand exceeds the vehicle capacity.

The first step in formulating the problem involves preparation of the customer demands, meaning the total weight that each cluster captures. These were previously defined in Table 2.8 above. However, it can be observed that Facility 2 in the table has a total weight greater than 4 tonnes. Since the vehicle capacity is less than the total demand there, a solution to the problem would not be reached, unless the vehicle travels to the specific node twice. Hence, whenever a customer has a demand greater than 4 tonnes, it is split into two separate notes. The total demand remains the same, but each node does not have more than 4 tonnes of demand on its own. The customer coordinates for that node are also duplicated, such that the nodes are on the exact same location. A distance matrix including the distances calculated using the Haversine formula between all of the coordinates is also created. As before, this is not accurate since in reality the actual road network layout would be used by the vehicle, and hence the distance between

points would not simply be the straight line distance. Nevertheless, this distance gives us an appropriate approximation.

The sets of the model are defined as S and D, where S is the set of the established facility, and D is the set of the customers, which are the other 9 cluster centroids. N is the set of vertices and includes both S and D, hence both the facility and the customers. The first decision variable, x_{ij}, is the link between i and j. It is a Boolean variable, hence if a link exists it is equal to 1 and if not it is equal to 0. The second decision variable, f_{ij}, is the amount of cargo transported from i to j. The parameters of the problem are the distances between vertices, the vehicle capacity (payload), which is given as 4 tonnes, and the weight demand of each of the customers. The optimization goal here is to minimize the distance of the vehicle routes, while satisfying the vehicle capacity. The objective function is therefore to minimize the summation of the product of the distances between links.

There are several constraints defined in order to reach the optimal solution. The first constraint is vehicle equilibrium and the second one is cargo equilibrium. This means that the cargo in minus the cargo out must be equal to the demand at that node. Additionally, the materials transported to any customer must be equal to their demand. The third constraint of the problem, is that the cargo carried by the vehicle must not at any point exceed the vehicle payload, 4 tonnes. The vehicle is unable to carry a weight greater than that. The problem is formulated such that the vehicle starts with exactly the amount of weight it will distribute per route, ensuring it is less than 4 tonnes. It then follows the established route, and delivers the demand required at each of the customers (nodes). At the end of each route, it returns back to the depot facility with a hold of zero weight. The fourth constraint, is therefore that the vehicle returns to the supply depot with no cargo. The fifth constraint ensures that no vehicle waits at any node. This prevents the vehicle from looping around a single customer, ensuring that no infinite loops develop. The sixth constraint ensures that the vehicle leaves the depot by setting the link between the supply facility and the customers equal to or greater than one. Constraint seven makes sure that all nodes are visited, since there cannot be a node which is simply left behind. The vehicle must deliver the appropriate cargo to all nodes. Another constraint involves sub-tour elimination. This improves the efficiency of finding an optimal solution by identifying and eliminating suboptimal tours within a larger tour. It helps in reducing the solution space, making it easier to find the optimal solution. Lastly, x is defined as an integer with a value of 0 or 1 since it is a Boolean, and f must have a positive value, since a vehicle cannot operate with negative cargo.

Table 2.10 shows the symbolic mathematical formulation for the vehicle routing problem. Facility refers to the facility selected in Task 2.5, while customers refer to the other 9 cluster centroids obtained in Task 2.3. The CVRP formulated is again solved in Python using PuLP. Figure 2.20 above displays the routes the vehicle takes.

Table 2.10: Mathematical formulation of the vehicle routing problem:

Sets: $S, D \subset N$	
N	Set of vertices.
S	Set of suppliers-facilities.
D	Set of customers.
Decision variables:	
x_{ij}	If vehicles travel from i to j.
f_{ij}	How much cargo is carried from i to j.
Parameters:	
d_{ij}	Distance between vertices i and facility j.
p	Vehicle payload = 4,000,000 grams.
Q_i	Demand by customer i.
Objective function:	
$\sum_{i \in N} \sum_{j \in N; i \neq j} x_{ij} d_{ij}$	Minimise the distance of vehicle routes, hence the product of distances with the decision variables that indicate whether links are connected.
Constraints:	
$\sum_{j \in N} x_{ji} - \sum_{j \in N} x_{ij} = 0 \quad \forall i \in D$	Vehicle equilibrium.
$\sum_{j \in N} f_{ji} - \sum_{j \in N} f_{ij} = Q_i \quad \forall i \in D$	Flow equilibrium: cargo in – cargo out=demand.
$f_{ij} \leq p \times x_{ij} \quad \forall i, j \in N$	Cargo flow must not exceed vehicle payload. Vehicle capacity limits tour size.
$f_{ij} = 0 \quad \forall i \in N, j \in S$	Return to the supply depot with no cargo.
$x_{ii} = 0 \quad \forall i \in N$	Ensure no vehicle waits at any node, meaning it does not do self-loops.
$\sum_{j \in D} x_{ij} \geq 1 \quad \forall i \in S$	Ensure that at least one vehicle leaves the depot.
$\sum_{j \in N} x_{ij} = 1 \quad \forall i \in D$	All nodes must be visited.
$x_{ij} + x_{ji} \leq 1 \quad \forall i, j \in D$	Sub-tour elimination.
$x_{ij}, \in \{0,1\} \quad \forall i \in N, j \in N$	x is a Boolean (integer with a value of 0 or 1).
$f_{ij} \geq 0 \quad \forall i \in N, j \in N$	Cargo cannot be negative.

2.8.1 Reflection on overall demand and resolving for weekly demand to make business sense

The overall weight demand used to solve the vehicle routing problem was obtained through aggregating the product weights from the provided dataset which only contains a 5% subset of the overall activities in the region. Also, since the dataset covers nearly 2 years, the solution obtained suggests having one vehicle travelling around 10 facilities over this period, per 5% of the overall activity, which makes no business sense. If accounting for 100% of the activities in the region, this would mean having 20 vehicles travelling around the same route to cover the whole demand in the region, which is not a desirable solution considering that delivery time is very important to customers nowadays.

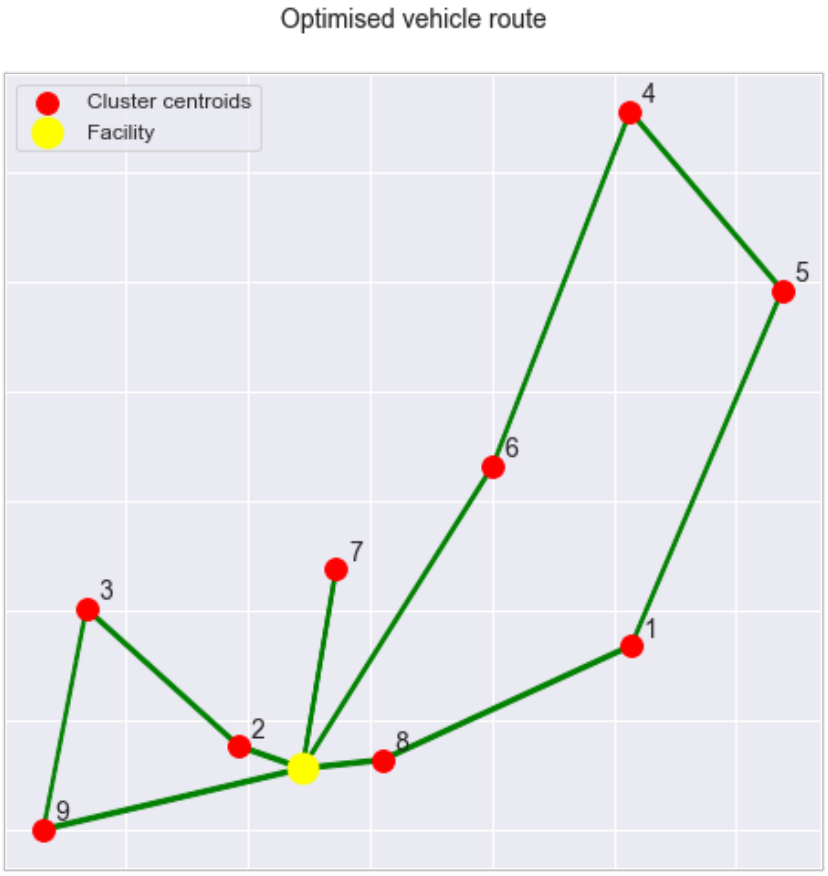
The most appropriate way to go about this is to adjust the demand to the projected 100% of the cargo, and assume weekly transportation times which are more realistic. Therefore, the vehicle routing problem is solved again, for the week deemed to be worst-case scenario, hence the one with the highest total load. This was found to be the week commencing on the 20/11/2017 which has a total aggregated weight of 963,936g for the 5% subset, and hence 19,278,720g for the whole region. The customers who placed an order during that week were identified and assigned to the same cluster they were assigned previously, as seen in Figure 2.21. New cluster centroids were not found since the facilities cannot be moved once they are built. The aggregated weight corresponding to each facility was recalculated based on the filtered customers in Table 2.11 and it is observed that none of the cluster centroids have a total weight greater than 4 tonnes, other than the main facility. The CVRP algorithm is therefore implemented again to find the new vehicle route with the same formulation as before. Figure 2.22 shows the optimized vehicle route for the week considered as the worst case scenario which has a total transportation cost of 866.22 km.



Figure 2.21: Filtered customers with assigned clusters.

Table 2.11: Weight demands per cluster considering weekly solution.

Cluster	Aggregated weight [g]
1	882,860
2	2,813,860
3	11,222,360
4	720,820
5	366,200
6	83,000
7	697,680
8	846,360
9	1,641,580
10	4,000



Route 1:

S1, D2: 3,538,680 g
D2, D3: 724,820 g
D3, D9: 4,000 g
D9, S1: 0 g

Route 2:
S1, D6: 3,671,320 g
D6, D4: 2,973,640 g
D4, D5: 2,607,440 g
D5, D1: 2,524,440 g
D1, D8: 1,641,580 g
D8, S1: 0 g

Route 3:
S1, D7: 846,360 g
D7, S1: 0 g

Figure 2.22: Optimised vehicle route for week commening 20/11/2017.

Since most of the weight demand is close to the main facility, it makes sense for the vehicle to have more routes closer to that point. This is because it cannot exceed the capacity of 4 tonnes. The way this problem was formulated was such that the vehicle leaves the facility with cargo of exactly equal to the weight it will distribute along its route. This is more efficient than starting with a cargo equal to 4 tonnes and returning with left over cargo.

2.8.2 Information needed to properly evaluate merits of plan

Even though the solution reached is good given the known parameters, more information would be needed to properly evaluate the merits of such a plan. Firstly, it still represents a single echelon supply chain. This means that the supply facility needs to be supplied with goods from elsewhere in order to distribute it to the customers, unless it is a factory and the electronic goods are produced there. However, even if this was the case materials would need to be transported to the factory to allow for manufactury of the goods. As already mentioned, the distances between nodes does not yield an accurate result since the vehicle would not be travelling in a straight line in reality. Modelling the actual road network in Rio De Janeiro would allow for accurate travel distances to be calculated. Knowing the investment cost per vehicle and the involved transportation costs would allow for more parameters to be included in the problem, putting an actual price to the yielded solution. Additionally, in reality there might be specific time slots of when to visit each node. This would place more constraints on the problem and could change the route taken. The time of when all the routes need to take place is also not stated. It could be over a day or over multiple days, depending on the time taken to travel to the nodes. It was also assumed that

the road network is non congestible, however in real life there will be traffic on the road network as well as traffic lights, further delaying the delivery of goods. The adopted approach only considers one vehicle. It might be cheaper to have multiple vehicles, or even vehicles with multiple capacities. It could also be, that deliveries are planned in a way as to minimize energy consumption. Similarly, trade-off analysis could be performed to obtain the optimal main facility numbers. Level of Service guarantees are nominally provided as part of inventory management. Increasing the number of facility numbers has both advantages and disadvantages which could be further analysed with more elaborate models. Having more than one main facility would improve response times and customer satisfaction, but would increase facility, staffing and inventory costs at the same time. Finding the level of service and comparing different facility options would therefore aid in obtaining the most optimal number (Angeloudis, 2023b). Lastly, the total demand for each region was assumed as known and fixed. However, in reality there will be large fluctuations and will vary in time. In conclusion, the problem is much more complex to how it has been solved and requires a lot more parameters and constraints to be defined so that its merits can be properly evaluated.

3. Conclusions

This report has discussed the limitations of the Holt Winter's Method and proposed improvements for the prediction of demand of online electronic goods in Rio De Janeiro, Brazil. Using an appropriate forecasting method, the expected total value of orders per year from the entire region were identified for years 2017, 2018, 2019 and 2020, and methods to identify valuable customers were discussed. A parcel last-mile distribution network was later designed, after using the k-means algorithm to identify 10 cluster centroids, used in the facility location problem in order to minimise the total distance between customers and their assigned facility. Lastly, the vehicle routing problem was implemented to propose the best approach for cargo delivery in the region to the electronic goods retailer. It was first solved for the data provided in the given dataset, however after discussion of whether this makes business sense it was deemed as insufficient. The problem was therefore solved again, assuming weekly delivery times, and designing the distribution system based on the worst case scenario. The objective function formulated focused on minimising the distance travelled within the transportation system. In conclusion, the solutions obtained were generated based on many assumptions which are discussed throughout the report, but would not be true in a modern transportation system. The imperative for advancing contemporary transit networks necessitates the development and implementation of more efficient and sophisticated models.

4. References

Angeloudis, P. (2023a) *Network Algorithms*. [Lecture] Imperial College London, 17th October.

Angeloudis, P. (2023b) *Supply Chain Networks*. [Lecture] Imperial College London, 23rd November.

Brownlee, J. (2020) *Time Series Forecasting With Prophet in Python*. https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/ [Accessed 6th December 2023].

Dash, S. (n.d.) *Gradient Boosting – A Concise Introduction from Scratch*. https://www.machinelearningplus.com/machine-learning/gradient-boosting/?utm_content=cmp-true [Accessed 6th December 2023].

GeeksforGeeks. (n.d.) *Haversine formula to find the distance between two points on a sphere*. https://www.geeksforgeeks.org/haversine-formula-to-find-distance-between-two-points-on-a-sphere/ [Accessed 5th December 2023].

Hyndman, R.J. & Athanasopoulos, G. (2018) *Forecasting: principles and practice*. 2nd edition. Melbourne, OTexts. https://otexts.com/fpp2/.

Kenton, W. (2021) *Sector Analysis: How it Works and Why It's Important*. https://www.investopedia.com/terms/s/sector-analysis.asp [Accessed 4th December 2023].

Nag, A. (2021) *Bayesian Optimization: A step by step approach*. https://towardsdatascience.com/bayesian-optimization-a-step-by-step-approach-a1cb678dd2ec [Accessed 6th December 2023].

Rebound Electronics. (n.d.) *The Demand for Consumer Electronics During the COVID-19 Outbreak*. https://reboundeu.com/insights/blog/the-demand-for-consumer-electronics-during-the-covid-19-outbreak/ [Accessed 2nd December 2023].

Statista. (n.d.) *Consumer Electronics, Brazil*. https://www.statista.com/outlook/cmo/consumer-electronics/brazil [Accessed 2nd December 2023].