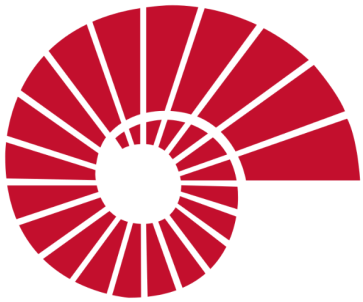

KOÇ UNIVERSITY

Industrial Engineering Department

INDR 491 Industrial Engineering Design Project

Fall 2022

Segmentation of Truckers for Borusan Logistics

1. Brief system description

'Borusan Lojistik' was founded in 1973 to serve the group companies within Borusan Holding'. In 2000, the company exceeded its boundaries of influence to present its experience and knowledge to companies outside the group. In 2012, they overtook Balnak: one of the ten great players in the market to improve their hold on it. Borusan Logistics includes a large variety of services. They mainly provide logistic support and services for the sellers. They offer two kinds of services. One is third-party logistic services, and the other is port management services.

Third-party Logistics services have special transportation services including warehousing and production, supply chain and vehicle logistics. The company develops the most appropriate solutions in line with the customer's unique requests and wants regarding their inputs and offers custom solutions suited to their needs in the operations. Third-party logistics include International Transportation Services, which Borusan Logistics handles in container, land, air, and rail transportation.

Through port management services, Borusan Logistics provides quality services through investments in strategically located Borusan Green Eco Port with expert and experienced staff. The port contains containers, general and project cargo, RO-RO, PCC, port, and terminal services with 5,000,000 tons of general cargo, 400,000 TEU of container, and 350,000 vehicles capacity. It can carry 20 lines of containers with its current 560m- long linear dock with a water depth of 14.5m.

The company aims to use modern management techniques effectively, give customers a competitive edge, and add value to its processes through continuous improvements. For example, one of their services, B-Yol, provides for sellers who want FBA service through German and England warehouses. B-Yol receives the products directly and delivers them to Germany and England warehouses quickly with the assurance of Borusan Logistics. They also have a business model provider that produces alternative delivery models in line with the increasing e-commerce volume named Bukoli. Their most popular service is known as ETA. This platform finds the best vehicle available for the transportation of goods within a certain distance in a country. Without returning to terminals, owners of trailers and trucks can instantly check all available transportation tasks compatible with their vehicle on the ETA Mobile application. With more than 100.000 certified truck and trailer drivers, the platform offers the best transportation choices. Borusan Logistics offers services in Turkey and Europe and international transportation services in the Middle East and Central Asia with partners in continental Europe, Iraq, Scandinavia, and several more countries.

2. System Analysis

The problem with the project is that the current segmentation system of the company's truck drivers is not good enough to force the truck drivers to maximize their performance. Therefore, the company wants to develop a new segmentation system where the grouping of the drivers into predetermined performance metric-based sub-groups will genuinely encourage and, in a way, force them to increase their performance and efficiency in terms of spending. This statement can be classified as our and the company's objective with this project. The predetermined performance metrics are net income per KM, fuel spending per KM, net income per ton, total revenue, recency, and frequency. With the implementation of this segmentation process, the company wants to create a competitive environment and increase the drivers' performance. This segmentation model also needs to be a very dynamic system to add to this objective. In other words, the system should force the low-class drivers to work harder, amplify their scores, and reach the upper classes, and the system being dynamic system could immensely help this. Higher class results in the truckers getting more bonuses and unique prizes. The direct scope of the project is the truck drivers who work for Borusan Logistics, and keep in mind that not all of these truck drivers work full-time for the company. Anybody with the right truck driving qualifications can sign up for the database for Borusan select a journey from the available pool for all the listed journeys and make the journey. So in a way, the workers can be thought of as independent freelance workers. However, some full-time workers are fully committed to working in the company. As stated, even though these people are the primary direct scope, indirectly, the scope is quite broad since the performance of these freelance employees affects the whole of the company and its reputation. On top of this, as the performance worsens, the customers' businesses will also be indirectly affected. So the whole company, sector, and customers are all tied to the scope.

During the fourth week, after signing all of the disclosure agreements, we finally obtained the dataset of all the journeys that the drivers have taken since the company's data collection program started. This database contained much information in terms of both the entries and the metrics being collected. Here are the initial data headings that were provided to us:

```
Index(['Id', 'SeferNo', 'MusteriSeferNo', 'SeferTarihi', 'PersonId',
       'SoforAdi', 'AracPlakasi', 'AtananTreylerPlakasi', 'AtananAracTipi',
       'AtananZeminTipi', 'AracGrubuMetni', 'CikisNoktasiIl', 'CikisIlce',
       'CikisLongitude', 'CikisLatitude', 'VarisNoktasiIl', 'VarisIlce',
       'VarisLongitude', 'VarisLatitude', 'UgrakSayisi', 'BosKM', 'FiiliKM',
       'PlanKM', 'YukCinsi', 'Tonaj', 'SeferDurumu', 'Alan', 'Bölge',
       'KarMerkezi', 'AracAtamaKanali', 'AracAtamaYapan', 'AracAtamaTarihi',
       'TedarikciNo', 'TedarikciAdi', 'OdeyenAdi', 'OdeyenKodu', 'FiyatTipi',
       'YakitTutari_Avans', 'YakitTutari_Bakiye', 'YakitKomisyonGeliri',
       'HedefYakitTutari', 'GerceklesenYakitTutari', 'HedefKomisyonBedeli',
       'GerceklesenKomisyonBedeli', 'EkHizmetMaliyeti', 'HedefMaliyet',
       'ToplamSatinAlmaMaliyeti'],
      dtype='object')
```

The data provided to us had several areas for improvement. Most of these metrics were irrelevant for our use since the data cannot be used for driver segmentation models. Also, many data entries had issues like negative, null, and invalid values; therefore, we had a solid phase of data parsing/preprocessing. In the next section, we will provide a more detailed explanation. Regardless, after analysing the data and understanding every metric, we have concluded that this data will be good enough for us to come up with a great segmentation model since the essential information that we have decided to use for our models, such as destination latitude and longitude, total fuel spending and more were already provided to us. Nevertheless, we will also derive more metrics from these provided metrics later.

3. Literature review and sector analysis

One of the primary research we have done was on the possible segmentation models we could use. We will pitch into more detail regarding these model alternatives in the next section. Besides this, research on models and parameter selection are also important factors to consider. We had to determine and use the metrics that would accurately measure the performance in many different metrics, such as spending, revenue obtained, and more. In terms of the research process, the first thing we did was to look for similar models of segmentation that grouped up truck drivers who work for logistics companies around the world in terms of their performance metrics. However, because companies keep such models confidential, no specific information can be found online. Therefore, as time progressed, we broadened our research, researched segmentation models for different occupations, and asked the company's data analyst team how we could apply such models to ours. They told us that we could mimic such models. We were advised to compare at least two statistical and machine learning models to test the performance of the models. Besides this, we have also researched what could be a good indicator of solid job performance for vehicle drivers. This mainly yielded results related to fuel usage, amount of

accidents, and more. In our weekly meetings, we discussed whether these topics benefit our model.

As mentioned, due to the privacy and protection of such data, unfortunately, we could not do a logistics sector analysis. Hence our truly one connection to the industry was the people from Borusan whom we interacted with during the weekly meetings. After our research, deriving the metrics from the existing segmentation models is helpful for the parameter selection step of the project. First, we had to complete the data parsing step. We have deleted all of the entries that are empty or null cells that belong to the columns of the metrics that we will use later on. The exact process was done on the entries with the invalid type entries. We have also computed a new metric column called distance using a Google Maps API, which took the latitude and longitude of the base and target locations (this data was present in our given dataset) as inputs and computed the Google Maps road distance in between the 2 locations. We have also encountered another logical issue in the data regarding this metric.

For example, one of the driver's several journeys consisted of distances much less than the amount he travelled. For instance, one case was when his origin location was about 50 KM away from his target location, but he travelled for 120 KM during the ride. However, since this was an unexplainable computational error that did not distinguish the case from the errorless cases (unlike a negative or empty value which we can detect), we could not find a way to catch them all in the whole dataset and, therefore, once we asked about this issue to the data analysts of Borusan Logistics. There could be more such cases in the database of the Borusan employees; they told us to ignore them. Also, when we looked at the data, we noticed that as the years went by, the prices constantly increased. After brainstorming why this trend was present, we have concluded that this could be due to the inflation of the Turkish Lira. By using a library containing all of the TL to Euro conversions for the past years, we could convert the monetary values to Euro. Although the Euro also gets affected by inflation over time, such an effect is not even close to being as severe as it is on the Turkish Lira and other worldwide currencies, and therefore even though it is not perfect, it could be considered as a good reference point. After all these steps of research, data cleansing, and data modification, we had enough research in our belts to implement our alternative models.

4. System Design

The current model, the company, used was simplistic, unfair, and inefficient in classifying the drivers. The system would not classify the drivers correctly according to the current metrics of the company's model, causing unfair classifications and not getting better benefits from the company, which could potentially cause demotivation to the drivers using the company's service. Our new methodology brought a more fair and justified way of classifying the drivers with more accurate performance measurement metrics. This is more motivating and better for the drivers since drivers have a higher chance of being in the class they belong to. Figure 1 explains the

motivation increase after the models that will be implemented. This figure shows the opportunities of the motivation sources for different segments of truckers. Implementing these models prevents the system from attributing privileges to drivers that do not deserve them. We made a model that could assign new drivers to the correct class and also be able to re-class the existing drivers according to the new parameters we put into the system.

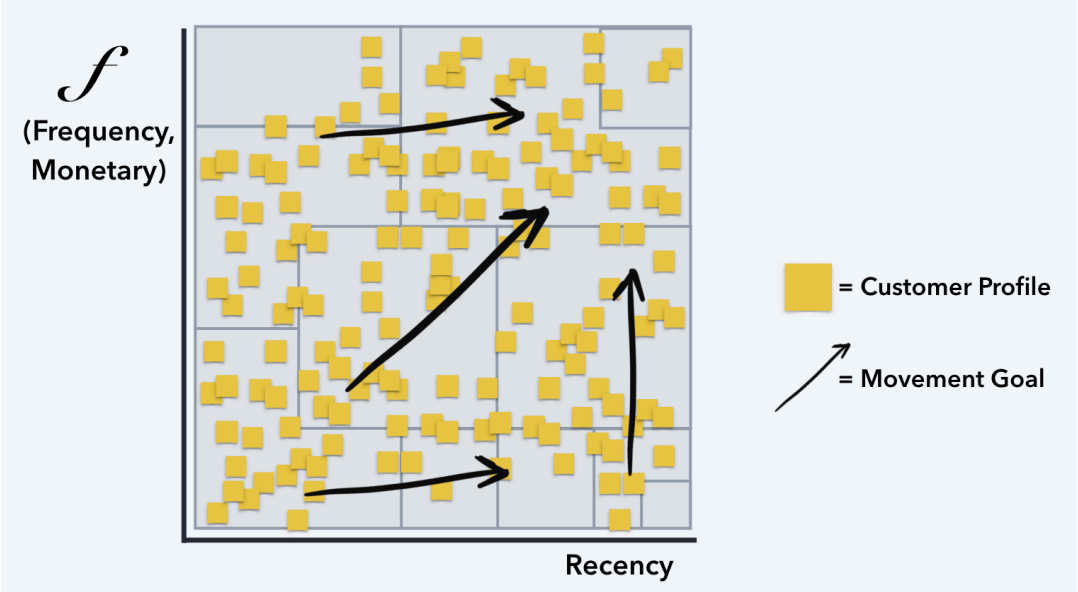

Figure 1: RFM Grid

We can consider several possible performance metrics/parameters for measuring the drivers' performance. Some standard truck performance metrics such as speed, anomalies in the driving speed can indicate delays that the driver possibly causes, and engine runtime, where this metric can record the amount of time the truck's engine spends running, which in turn can provide insights on the maintenance requirements, driver behaviour, truck utilization percentage and overall patterns of use of the driver, acceleration and braking, where rapid acceleration and braking can indicate how often the truck driver puts himself or herself in a situation where the safety of the truck is in danger. Also, constant rapid acceleration and braking are negative factors for fuel usage since these two actions require a lot of fuel and energy. There are some load management metrics, such as transit time, where this metric indicates the projected amount of time that it should take the driver to reach the desired destination, which can indicate a turn

towards bad performance if these times are not met for the driver. However, external factors such as road construction must also be considered for this metric, and therefore such periods should be updated constantly. For this metric, depending on how late the truck was to the destination, the points could be deducted from the driver accordingly. We can also have a binary version of this metric called on-time delivery, where there is a fixed amount of point deduction for any late truck and a fixed amount of bonus points for any on-time truck. Out of these possible metrics, we try to find similar ones that we already have in our dataset. The models for implementation are described below:

4.1) RFM Model

We have two models that we have derived as a result of the literature review for implementing the dataset of the company. We first used the RFM statistical model (Fig. 2). For this purpose, we found six metrics. For recency, we calculated the most recent trips for each driver. We gave a declining weight over the recent trips and, for this purpose, gave the most significant weight to the most recent trip. For frequency, we calculated the frequency of the trips, which is weighted by its recency, so that recent trips have a higher effect on the frequency metric. Lastly, for monetary, we used four different metrics. These are the incomes and costs. For income purposes, net income per KM, net income per Ton, and total revenue ranking. For cost purposes, we calculated the fuel that is spent per KM. For this metric to be fair, we extracted the ton they carry and recalculated the value. So drivers who carry heavier products would spend more fuel, but this effect is offset. For monetary values to be unaffected by inflationary pressures, we converted TL values to EUR, less affected by inflation and value losses.

```
ahcdf['Marginal_Cost'] = (ahcdf['Monetary']/ahcdf['FiiliKMsum'].apply(Decimal))
ahcdf['Fuel_Efficiency'] = (ahcdf['HedefYakitTutarisum']/ahcdf['GerceklesenYakitTutari'])
```

The three parameters, recency, frequency, and monetary, are the classification parameters for this model. Although the weights of these three metrics can be adjusted according to the objectives of the company, we gave the following weights to these parameters: 25% for the recency, 50% for the frequency, and 25% for the monetary. These weights can be chosen according to the nature of the business and can be changed. By this, we provide flexibility for the company so that they can change the parameters and other coefficients as they wish and therefore change the scores of drivers. To complete the model, using Python, we found the scores of the drivers and sorted them from highest to lowest. Then we created eight segments (classes) named Bronze 2 (being the worst), Bronze 1, Silver 2, Silver 1, Gold 2, Gold 1, Platinum 2, and

Platinum 1 (being the best). After dividing the final ranking according to percentiles predetermined by the company, we would need to build a minimum requirement for each class. The purpose of this minimum requirement will be to send a driver to a lower class if the driver is not at the same performance level as the drivers of that cluster. To measure the fitting level of the driver to its current cluster, we took the mean and standard deviation of each class and checked each driver's score distance to the mean of the current cluster. If the distance to the mean is more than one standard deviation, the driver is not fitting to that class. The standard deviation distance to the mean is a hyperparameter so that the company can change this parameter according to their needs and the nature of the business. According to this model, the distribution of the truckers is like in Figure 3.

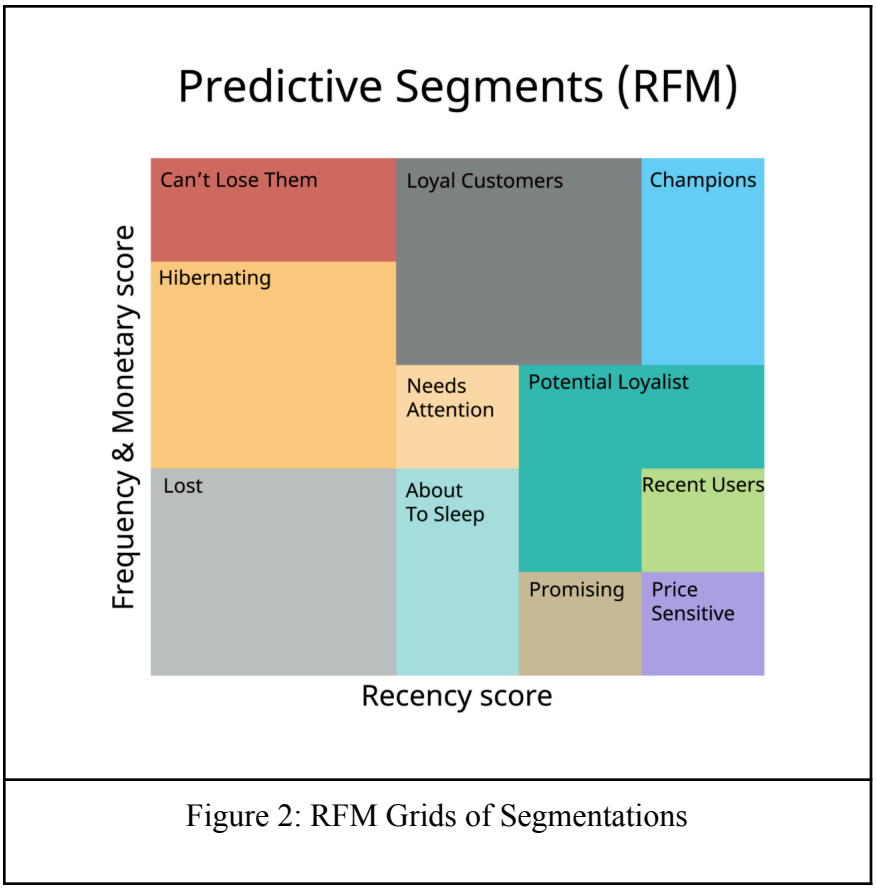

Figure 2: RFM Grids of Segmentations

4.2) K-Means Algorithm

In the upcoming models, we used machine learning algorithms. One model that we have used is the K-means algorithm. The K-means model creates a K amount of centroids which

assigns the closest node to itself. Each node is assigned to one centroid (whichever it is the closest to), and the centroid moves to the centre of mass of all the nodes assigned to it. It carries on iterating this algorithm until it reaches a point where it does not change (figure 4). We created 8 clusters (one for each class), and the resulting classification was not satisfactory for the company since it lacked differentiation of classes(the volume of the classes was nearly equal). We need to find a more complex algorithm to divide our classes and consider more factors, such as variation differences between classes and minimum requirements for each class.

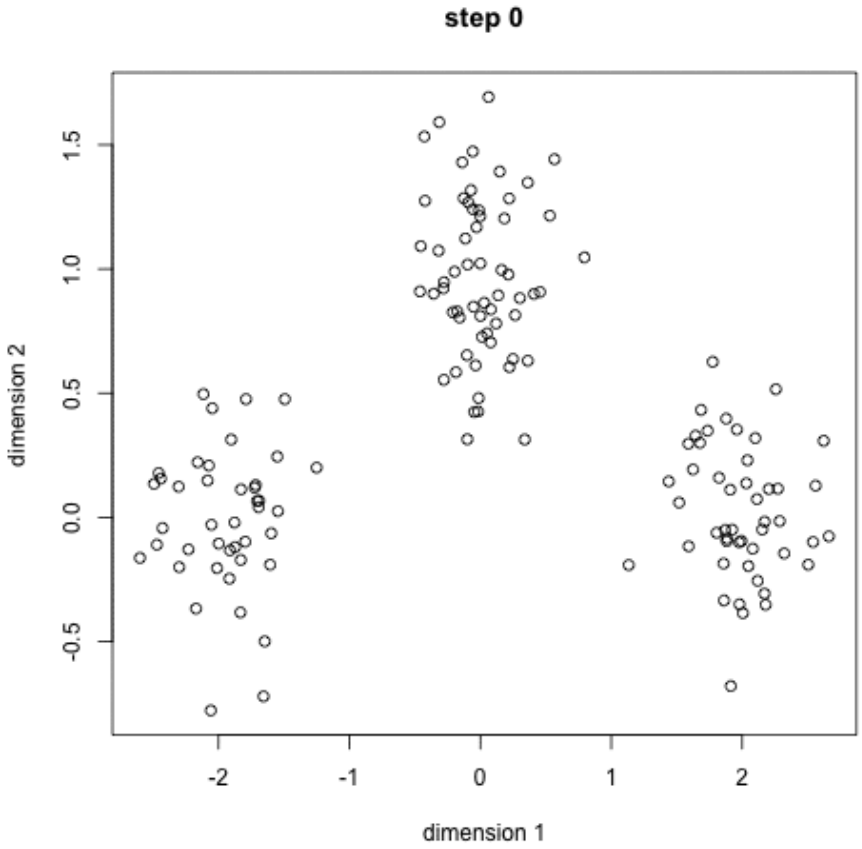

Figure 4: K-Means Algorithm

4.3) Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

We have decided to implement the second algorithm called Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM). Here, the idea is that each point has an assigned probability of joining the clusters. Therefore if we have 8 clusters, each point has eight probability percentages. The initial iteration is formed by randomly picking n number of nodes as the starting point of the groups, where n is the number of groups we want at the end. For each iteration, we are trying to maximize the cumulative chosen probabilities for each node. We continue like this until the iterations diverge into a steady state. The k-means algorithm only had a mean. Therefore whereas K-means could have only circular 2D groups, this algorithm can have elliptical groups, which is a more flexible and, therefore, better shape for grouping purposes, as seen in Figure 5.

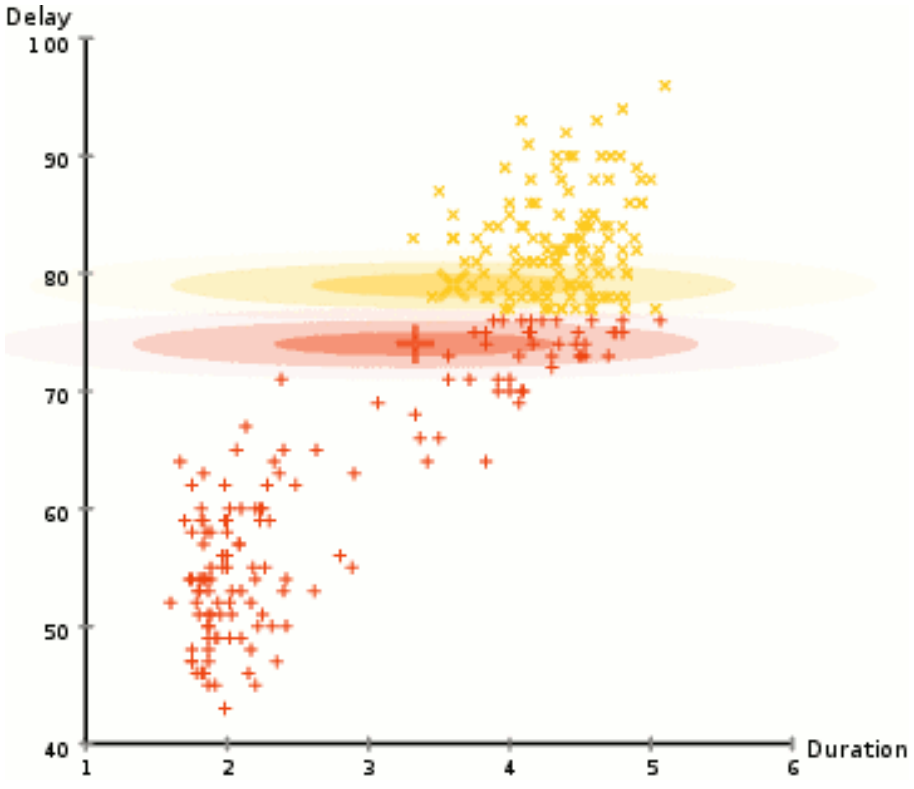

Figure 5: EM Using GMM

4.4) Agglomerative Hierarchical Clustering

The third algorithm that we have decided to implement is Agglomerative Hierarchical Clustering. In the beginning, every node is considered a group in itself, then whichever two groups have the minimum distance will be merged into a single group. This goes on and on until the required number of groups remains. Once several nodes are in the cluster, their mean coordinate is considered the center for the distance comparison. One advantage here is that there is no predetermined number of clusters here. Different implementations of this algorithm may include a factor for decreasing the variance. With the scikit library in python, we could implement such an algorithm. With the inclusion of variance reduction in our algorithm, we would get clusters with less variance within the cluster data points, which makes it more complex than the GMM model. The visualization of the Agglomerative Hierarchical Clustering algorithm is seen in Figure 6.

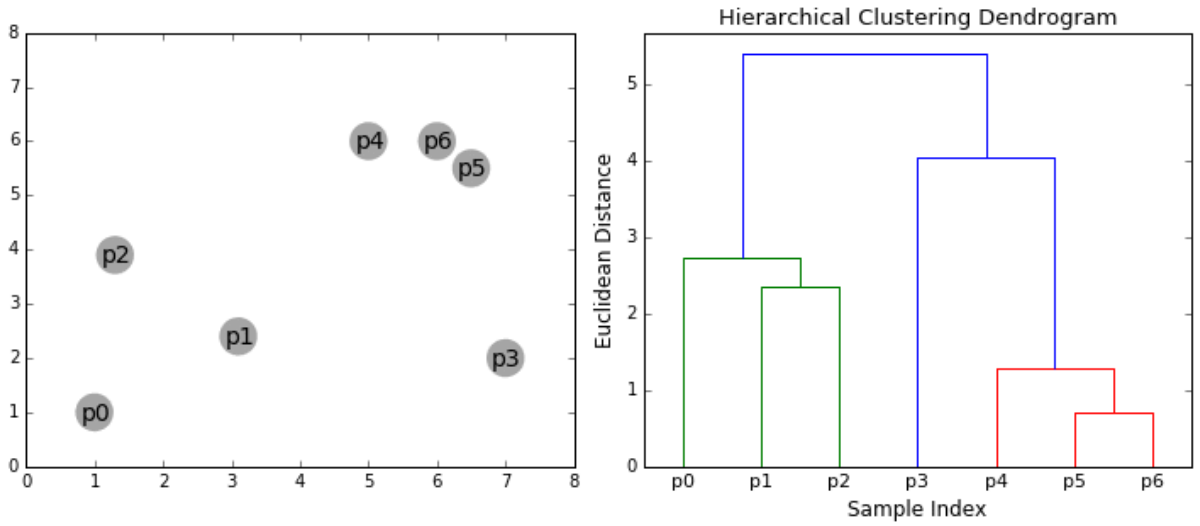

Figure 6: Agglomerative Hierarchical Clustering with Two-Parameter Setup

Even though the machine learning algorithms seem more dynamic and better, the statistical model provides more dynamism and flexibility, which we aimed to have initially. We found that by adding an extra driver or deleting a driver from the dataset, the RFM model

adapted better when we rerun our code. Because of this and the company's prioritizations, we chose the statistical model.

5. Implementation

Borusan Logistic is a sector leader logistic network company that needed higher performance of the truckers can be beneficial for the satisfaction of the customers. Using the segmentation, we have made the company enable their truckers to work more efficiently as truckers working with a higher performance leads to customer satisfaction to be higher due to sustainable service.

All in all, regarding our aim to choose the suitable model based on the company's needs, requirements, and data, we have tested various models to find the best implementation choice. First, we implemented a KNN machine learning algorithm as a test reference point, but this algorithm needed to be more complex, and the results could have been better for the company's needs.

Second, we have analyzed Agglomerative Hierarchical Clustering, one type of hierarchical Clustering where the population is divided into similar clusters having non-alike data points by the algorithm, which clusters the data points according to how similar they are until one cluster remains left. One single large cluster is formed, starting from a single one. As mentioned, we also take into account variance reduction in this method.

Agglomerative Hierarchical Clustering uses a bottom-up approach to form clusters. We have also examined the gaussian mixture where we utilize a function indicated by Gaussians with k where $k \in \{1,\ldots, K\}$ and K identify the dataset's clusters. Each k has the parameters μ (mean) covariance Σ (covariance) and π (mixing probability, i.e., the size of the function). We have taken a look at the PD cut, which separates scalar data points that are scored into different bins where the range within every bin is of equal size. Finally, we have discussed the statistical model, which interacts with the relationship between random and non-random variables. Using the model, we have divided our data set into commonly used percentages we have researched in the literature and got approval from the data analyst team to use these predetermined percentiles to divide the rankings into 8 clusters. However, we have realized that it was not a fair and distinct data separation system such that the percentage differences of each class were shallow and insignificant. To overcome this vagueness, within each class, we have calculated the difference between the mean of that class and the various pre-calculated standard deviations, allowing the different percentages between each class to be more apparent and the segmentation to be much more equitable as with this new model because it implies that the trucker that goes up from his class to an upper class earned his way up there.

Out of all the models we have analyzed, the company and we have decided that the statistical model is the appropriate model by the nature of the business and should be used in the sector's future for the segmentation of truckers. The reason for aiming dynamism in our selected model is that when a new trucker is added to or removed from the system, the system can run itself to adjust itself to rearrange the classes. Moreover, the dynamism is suitable for the addition or removal of a trucker and works with predetermined periods due to the trucker's extra earned scores. After the statistical model's implementation, we have found that eight classes in total should be formed, with platinum 1 being the best and bronze 2 being the worst-ranked class. We have also realized that updates are available for the minimum class requirements independently and set their personalized model, which will benefit the company's future action taken to segment their business. Our bar graph results and pie chart results for the segmentations are shown in the Appendix (Figures 1 and 2).

6. References

Aqsazafar. (2021, December 31). *Hierarchical clustering in python, step by step complete guide [2022]***. MLTut. Retrieved January 22, 2023, from https://www.mltut.com/**hierarchical-clustering-in-python-step-by-step-complete-guide/

***Gaussian mixture model*. GeeksforGeeks. (2022, July 31). Retrieved from https://www.geeksforgeeks.org/gaussian-mixture-model/**

Maklin, C. (2022, May 10). *Gaussian mixture models clustering algorithm explained*. Medium. Retrieved January 22, 2023, from https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

Aqsazafar. (2021, December 31). *Hierarchical clustering in python, step by step complete guide [2022]*. MLTut. Retrieved January 22, 2023, from https://www.mltut.com/hierarchical-clustering-in-python-step-by-step-complete-guide/

Seif, G. (2022, February 11). *The 5 clustering algorithms data scientists need to know*. Medium. Retrieved from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Soner. (2021, June 16). *All you need to know about pandas cut and Qcut functions*. Medium. Retrieved January 22, 2023, from https://towardsdatascience.com/all-you-need-to-know-about-pandas-cut-and-qcut-functions-4a0c1001c38b

7. Appendix

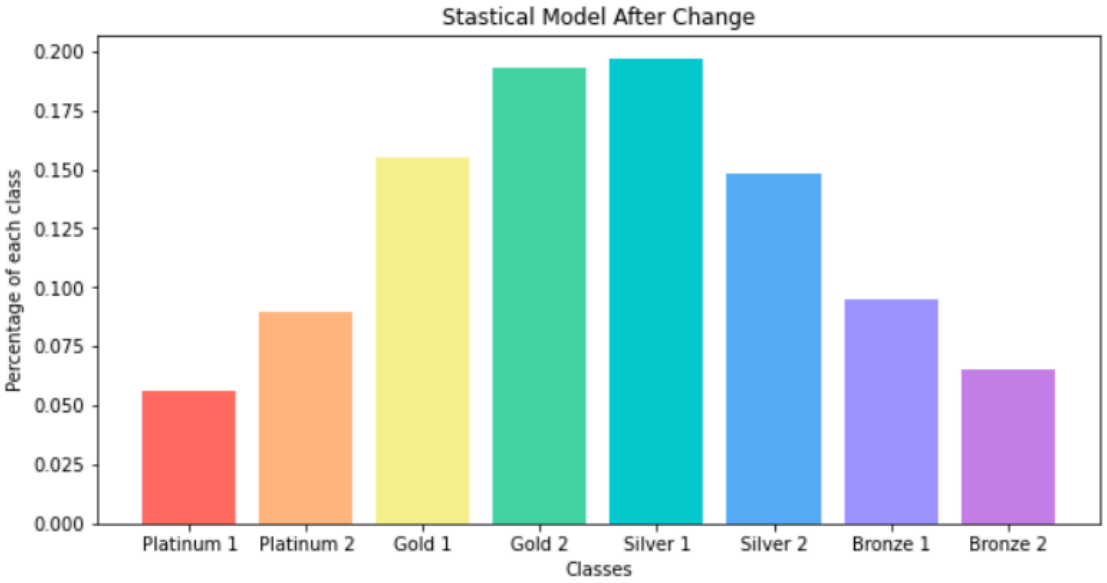

Figure 1: bar chart of the spread of ranking percentages.

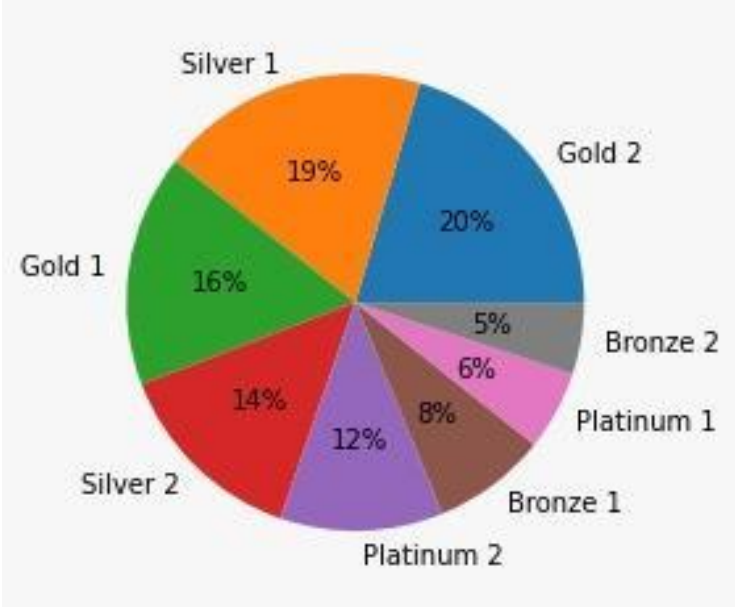

Figure 2: Pie chart of the ranking spread.