WHOLESALE CUSTOMERS DATA ANALYSIS

KATALLYST DATA SCIENCE INTERNSHIP

Adarsh Kumar (1805695)

ABSTRACT

Wholesale and distribution are important components for any country's economy. The movement of goods is a crucial part of the supply chain – and analytics can play a key role in this – it's no good manufacturing the right product if it can't get to the right retailer (or other business customer) at the right time. However, changes in retailer demand and technology mean the wholesale and distribution industries are changing.

Like many industries, new technology and global economic change have had an impact on wholesale and distribution businesses. An important task for many managers is to improve profit margins. Low sales figures, high inventory costs, and not enough hours in the day are the most common issues faced by managers in wholesale distribution. Why not consider industry specific data analytics to help improve business performance?

In this project, we investigate clients of a wholesale distributor. The dataset includes the annual spending in monetary units (m.u.) on diverse product categories. Our goal was to use various clustering techniques to segment customers. Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Thus, there is no outcome to be predicted, and the algorithm just tries to find patterns in the data.

Keywords: Clustering, Data Analytics, Machine learning.

INTRODUCTION

In this project, we will analyse a dataset containing data on various customers' annual spending amounts (reported in *monetary units*) of diverse product categories for internal structure. One goal of this project is to best describe the variation in the different types of customers that a wholesale distributor interacts with. Doing so would equip the distributor with insight into how to best structure their delivery service to meet the needs of each customer.
The dataset for this project can be found on the UCI Machine Learning Repository. For the purposes of this project, the features 'Channel' and 'Region' is also included in the analysis — with focus on the six product categories recorded for customers.

BACKGROUND

2.1 K-Means Clustering Model

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. A cluster refers to a collection of data points aggregated together because of certain similarities. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

K-Means Clustering Algorithm

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centres.

1. *Randomly select 'c' cluster centres.*
2. *Calculate the distance between each data point and cluster centres.*
3. *Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.*
4. *Recalculate the new cluster centre using:*

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

   where, 'c_i' represents the number of data points in i^{th} cluster.
5. *Recalculate the distance between each data point and new obtained cluster centres.*
6. *If no data point was reassigned then stop, otherwise repeat from step 3.*

2.2 Hierarchical Clustering Model

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) Identify the two clusters that are closest together, and (2) Merge the two most similar clusters. This iterative process continues until all the clusters are merged together.

Hierarchical Clustering Algorithm

Let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ be the set of data points.

1. *Begin with the disjoint clustering having level L (0) = 0 and sequence number m = 0.*
2. *Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r), (s)] = min d[(i), (j)] where the minimum is over all pairs of clusters in the current clustering.*
3. *Increment the sequence number: m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r), (s)].*
4. *Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r, s) and old cluster(k) is defined in this way: d[(k), (r, s)] = min (d[(k), (r)], d[(k), (s)]).*
5. *If all the data points are in one cluster then stop, else repeat from step 2.*

METHODOLOGY

The models compared in this study (K-Means Clustering, Hierarchical Clustering) have been used for numerous problems in different clustering task and have been chosen based on their popularity in industry. In addition, the data set is originated from a larger database referred on:
Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon.

3.1 The Data

The dataset consists of the annual spending in monetary units on diverse product categories including fresh products, milk products, groceries products, frozen products, etc. The data set contains a sample of 440 instances with 8 features/variables.

The description of the data is shown in table 1.

Sr. No.	Data Feature	Description	Feature Type
1.	Fresh	Annual spending on fresh products	Numeric
2..	Milk	Annual spending on milk products	Numeric
3.	Grocery	Annual spending on grocery products	Numeric
4.	Frozen	Annual spending on frozen products	Numeric
5.	Detergents_Paper	Annual spending on detergents & paper products	Numeric
6.	Delicatessen	Annual spending on delicatessen products	Numeric
7.	Channel	Hotel/Restaurant/Cafe/Nominal	Categorical
8.	Region	Regions- Lisnon, Oporto, Nominal	Categorical

Table 1. Data Description

3.2 Data Processing and Engineering

Further exploration of the dataset showed the need to create new features from the existing ones. This process is termed feature engineering; The new features created are:

Channel_Retail: Groups the feature *Channel* into two groups 0 and 1

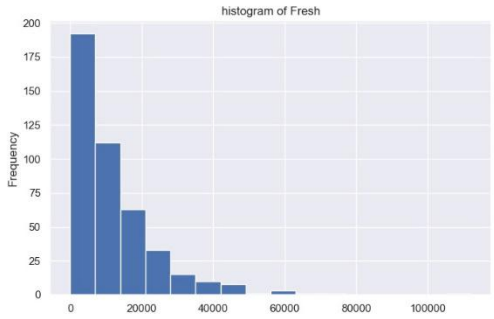
Region_Oporto: Groups the feature *Region* into two classes.

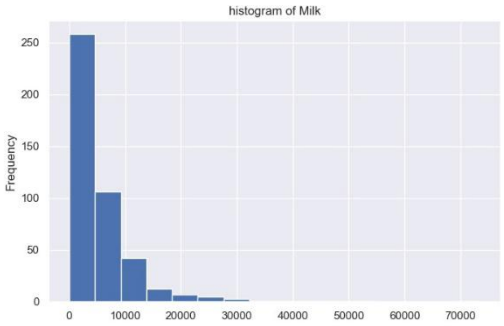
Region_other: Groups the *Region* into two classes

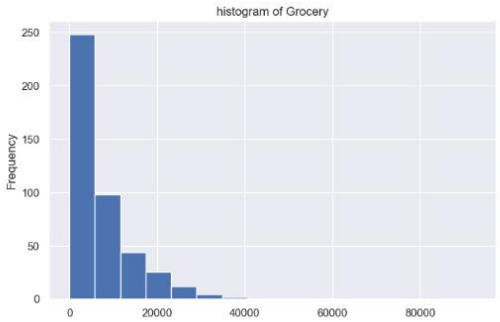
Next, detecting outliers in the data is extremely important in the data preprocessing step of any analysis. The presence of outliers can often skew results which take into consideration these data points. There are many "rules of thumb" for what constitutes an outlier in a dataset. Here, we have used Tukey's Method for identifying outliers: An outlier step is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal.

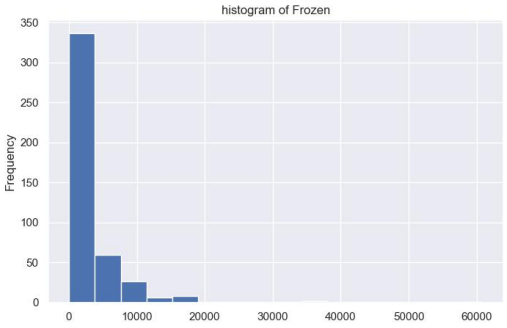
3.3 Data Visualization

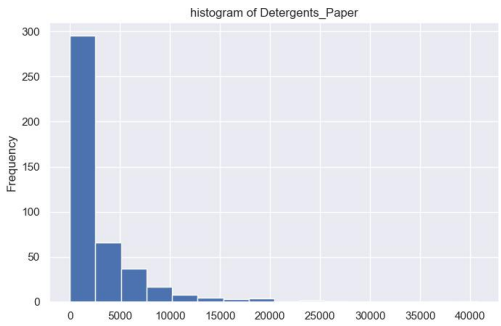
Creating histograms for each of the products available to get the insights of the requirement of these products in various regions and channel.

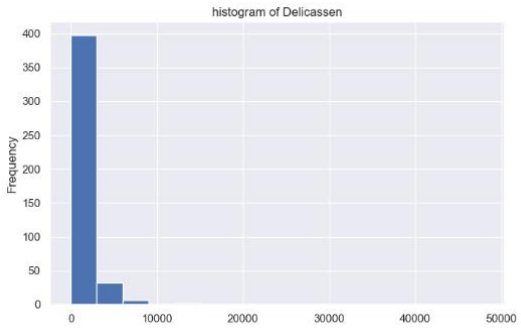




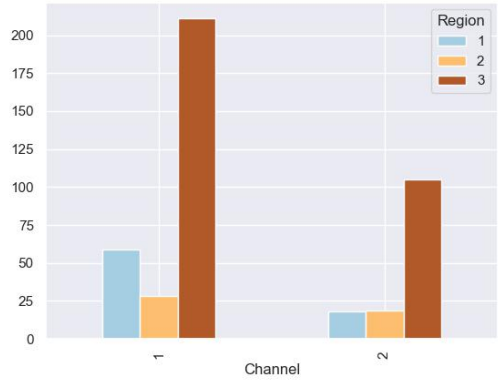






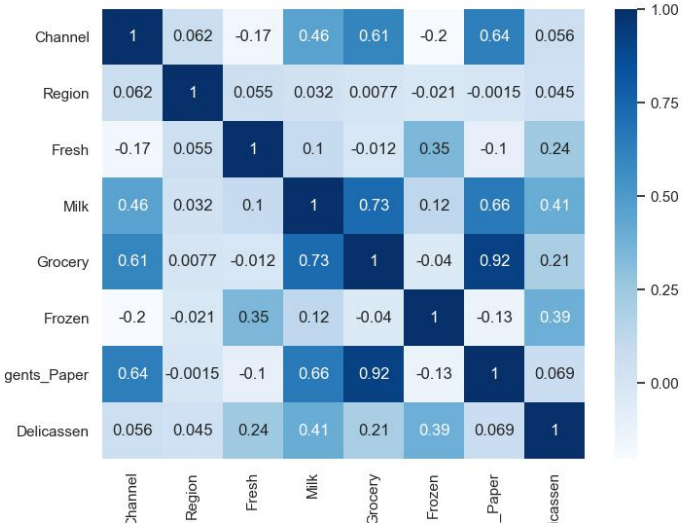


Plotting the frequency of each channel region wise we got to know that in Channel 1 and Channel 2 the products are delivered in region 3 are more as compared to respective regions.



Looking at the heatmap, there are a few pairs of features that exhibit some degree of correlation. They include:

- Milk and Groceries
- Milk and Detergents_Paper
- Grocery and Detergents_Paper



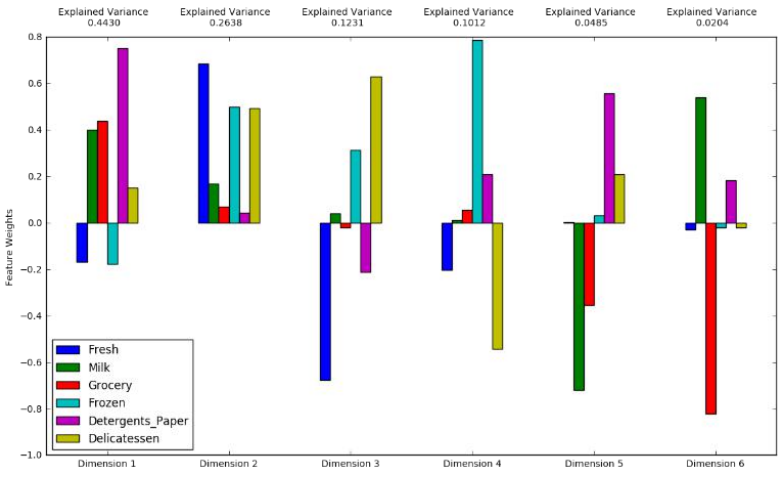
As we tried to predict the 'Milk' feature earlier, this confirms the suspicion that Milk isn't correlated to most of the features in the dataset, although it shows a mild correlation with 'Groceries' and 'Detergents_Paper'.

The distribution of all the features appears to be similar. It is strongly right skewed, in that most of the data points fall in the first few intervals. Judging by the summary statistics, especially the mean and maximum value points, of the features that we calculated earlier, we can expect that there are some outliers in each of the distributions. This conforms with the fact that there's a significant different between the mean and the median of the feature distributions.

3.3 Feature Transformation

In this section we will use principal component analysis (PCA) to draw conclusions about the underlying structure of the wholesale customer data. Since using PCA on a dataset calculates the dimensions which best maximize variance, we will find which compound combinations of features best describe customers.

Now that the data has been scaled to a more normal distribution and has had any necessary outliers removed, we can go ahead and apply PCA to the good data to discover which dimensions about the data best maximize the variance of features involved. In addition to finding these dimensions, PCA will also report the explained variance ratio of each dimension — how much variance within the data is explained by that dimension alone. Note that a component (dimension) from PCA can be considered a new "feature" of the space, however it is a composition of the original features present in the data.



The first and second features, in total, explain approx. 70.8% of the variance in our data.

The first four features, in total, explain approx. 93.11% of the variance.

In terms of customer spending,

- **Dimension 1** has a high positive weight for Milk, Grocery, and Detergents_Paper features. This might represent Hotels, where these items are usually needed for the guests.
- **Dimension 2** has a high positive weight for Fresh, Frozen, and Delicatessen. This dimension might represent 'restaurants', where these items are used for ingredients in cooking dishes.
- **Dimension 3** has a high positive weight for Deli and Frozen features, and a low positive weight for Milk, but has negative weights for everything else. This dimension might represent Delis.

- **Dimension 4** has positive weights for Frozen, Detergents_Paper and Groceries, while being negative for Fresh and Deli. It's a bit tricky to pin this segment down, but I do believe that there are shops that sell frozen goods exclusively.

When using PCA, one of the main goals is to reduce the dimensionality of the data — in effect, reducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of this, the *cumulative explained variance ratio* is extremely important for knowing how many dimensions are necessary for the problem. Additionally, if a significant amount of variance is explained by only two or three dimensions, the reduced data can be visualized afterwards.

3.6 Performance Metric

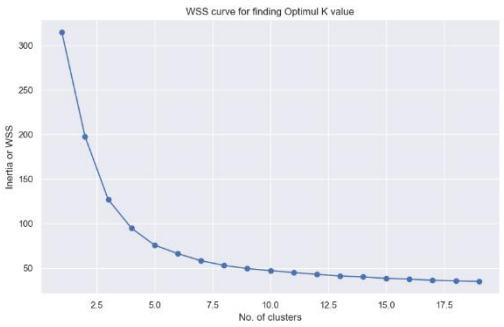
Our goal was to use various clustering techniques to segment customers. Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Thus, there is no outcome to be predicted, and the algorithm just tries to find patterns in the data.

RESULTS

In this section the results of the two models will be presented. The results were obtained by applying the two models: K-Means Clustering and Hierarchical Clustering.

4.1 K-Means Clustering:

The number of clusters is highly dependent on the type of scaling method used. When we were using Minmax Scaling, the number of clusters formed was found to be around 6 *(fig 4.1)* and it has been shown below in figure *(fig 4.2)*.



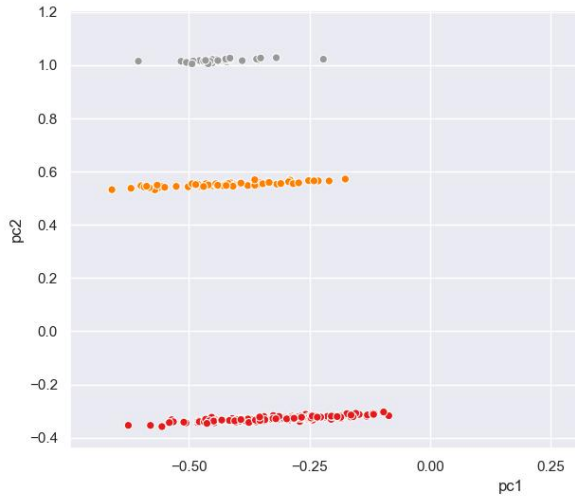


Fig 4.2: Clustering for Minmax Scaling

When we were using Normalizer Scaling, the number of clusters formed was found to be around 14*(fig 4.3)* and it has been shown below in figure *(fig 4.4)*.

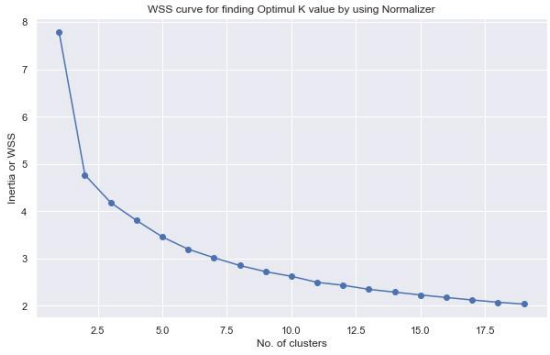
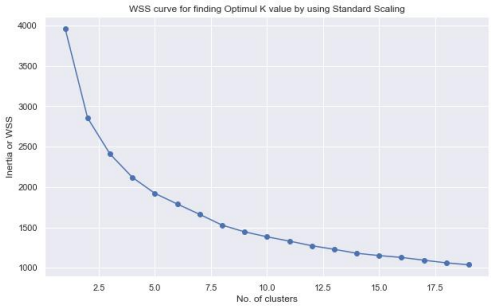


Fig 4.4: Clustering for Normalizer

When we were using Standard Scaling, the number of clusters formed was found to be around 15*(fig 4.5)* and it has been shown below in figure *(fig 4.6)*.



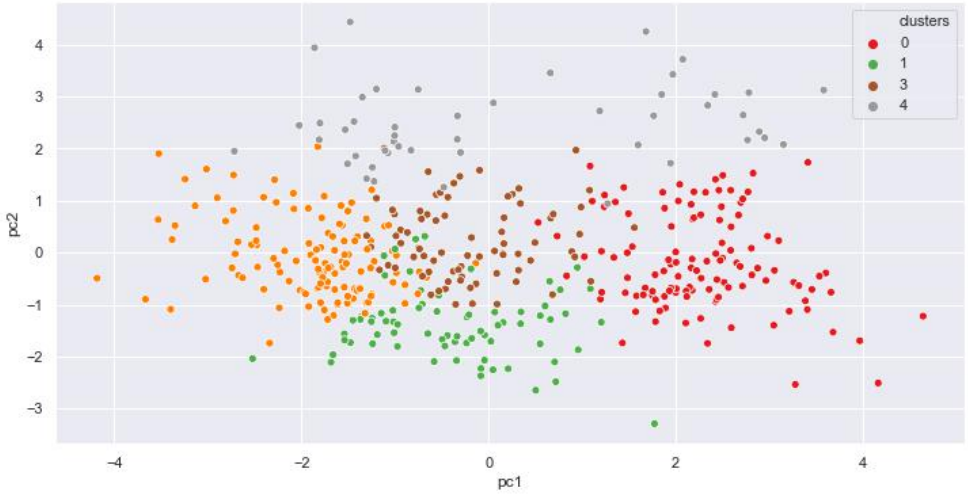
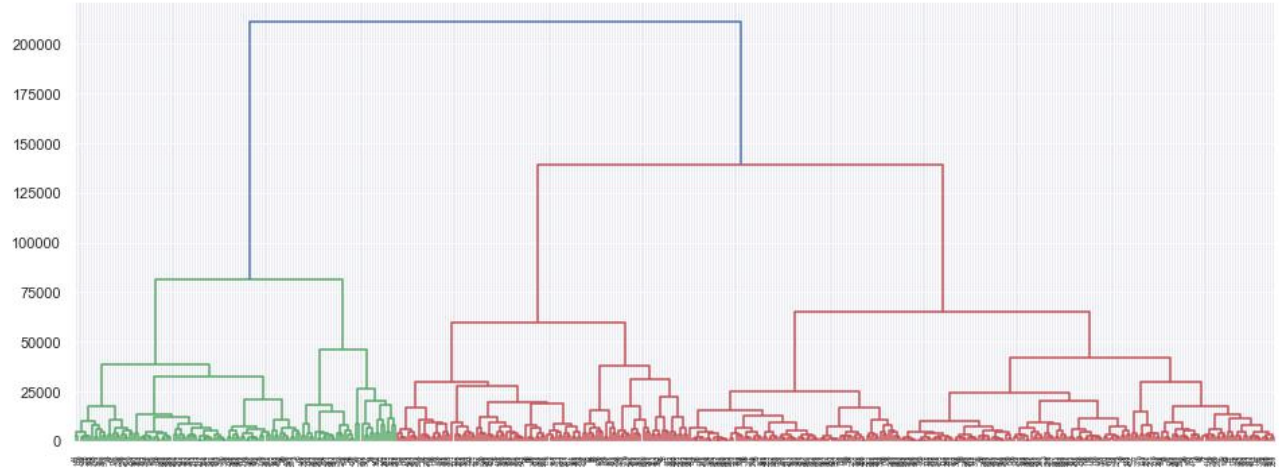


Fig 4.6: Clustering for StandardScaler

Since by using MinMaxScaler we can get the optimal value of K and the clusters formed by them can be easily identified.

4.1 Hierarchical Clustering:

A *dendrogram* is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from *hierarchical clustering.* The main use of a dendrogram is to work out the best way to allocate objects to
clusters.



From the above dendrogram we can assume that there will be around 6 clusters and as we move up along the y-axis the number of clusters decreases. In general, it is a mistake to use dendrograms as a tool for determining the number of clusters in data. Where there is an obviously "correct" number of clusters this will often be evident in a dendrogram. However, dendrograms often suggest a correct number of clusters when there is no real evidence to support the conclusion.

DISCUSSION AND FINDINGS

Companies will often run A/B tests when making small changes to their products or services to determine whether making that change will affect its customers positively or negatively. The wholesale distributor is considering changing its delivery service from currently 5 days a week to 3 days a week. However, the distributor will only make this change in delivery service for customers that react positively.

How can the wholesale distributor use the customer segments to determine which customers, if any, would react positively to the change in delivery service?

Making the change to the delivery service means that products will be delivered fewer times in a week. The wholesale distributor can identify the clusters to conduct the A/B test on, but the test should be done on one cluster at a time because the two clusters represent different types of customers, so their delivery needs might be different, and their reaction to change will, thus, be different. In other words, the control and experiment groups should be from the same cluster, at a time.

Additional structure is derived from originally unlabelled data when using clustering techniques. Since each customer has a *customer segment* it best identifies with (depending on the clustering algorithm applied), we can consider *'customer segment'* as an engineered feature for the data. Assume the wholesale distributor recently acquired ten new customers and each provided estimates for anticipated annual spending of each product category. Knowing these estimates, the wholesale distributor wants to classify each new customer to a *customer segment* to determine the most appropriate delivery service.

How can the wholesale distributor label the new customers using only their estimated product spending and the customer segment data?

To label the new customers, the distributor will first need to build and train a supervised learner on the data that we labelled through clustering. The data to fit will be the estimated spends, and the target variable will be the customer segment They can then use the classifier to predict segments for new incoming data.

CONCLUSIONS

We have described the variation in the different types of customers that a wholesale distributor interacts with and equipped the distributor with insight into how to best structure their delivery service to meet the needs of each. This process was very complex because there are lots of factors that should be taken into consideration. In order to implement achievable goals and successfully implement them, wholesale distributor chains always want to get insights of their customers' need. In this study, we used two machine learning algorithms (K-Means Clustering and Hierarchical Clustering). We observed that getting more data would generally increase the predictive power of our models.