II. Data Understanding

Collect initial data

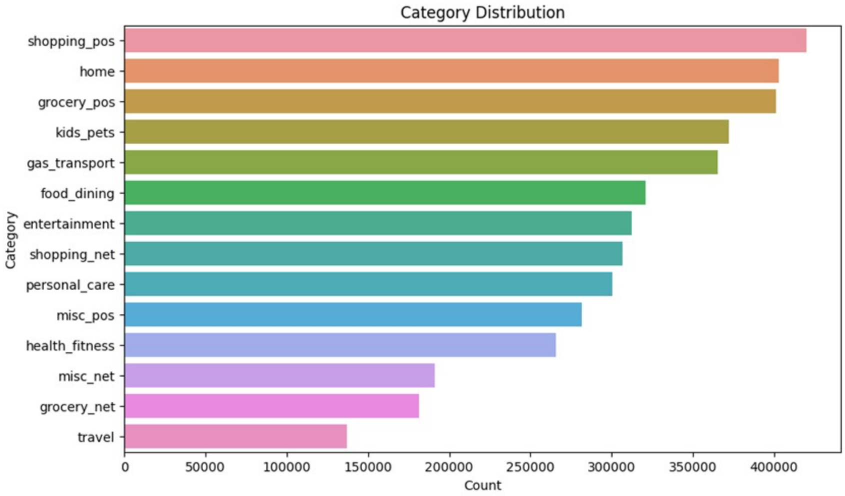
The customer dataset has 1000 rows and 15 columns with information about social security number, credit card number, customers' first and last names, gender, street address, city, state, latitude and longitude of customers' locations, city population, job, dob and account number.

There are 132 transaction csv files containing information about 4,260,904 customers' transactions such as credit card number, account number, transaction number, unix time, category, spending amount, whether the transaction is fraud, merchant name, merchant latitude and longitude. The transaction dataset is created by merging all 132 transaction files.

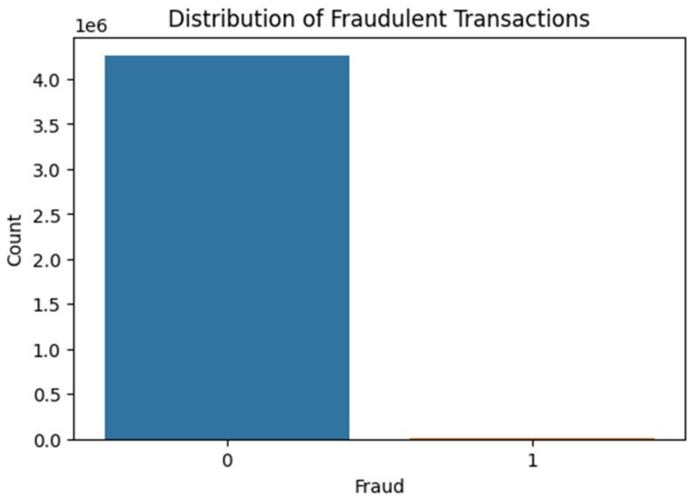
Describe data

The final dataset used for data analysis is the merged dataset combining the customer dataset and transactions dataset, comprising 4,260,904 rows and 23 columns. Among these columns, there are 9 categorical variables ('first', 'last', 'gender', 'street', 'city', 'state', 'job', 'category', 'merchant') that require conversion into numerical variables for compatibility with the linear regression model. Alternatively, if these variables are deemed irrelevant, they can be removed from the dataset. Notably, this dataset is free of missing values.In order to capture temporal patterns or trends in the data, the 'unix_time' variable should be transformed into datetime format. Moreover, it is advisable to conduct further investigation on features like 'dob', 'amt', 'lat', 'long', 'merch_lat', and 'merch_long' to extract more meaningful information pertaining to customers' spending behaviours

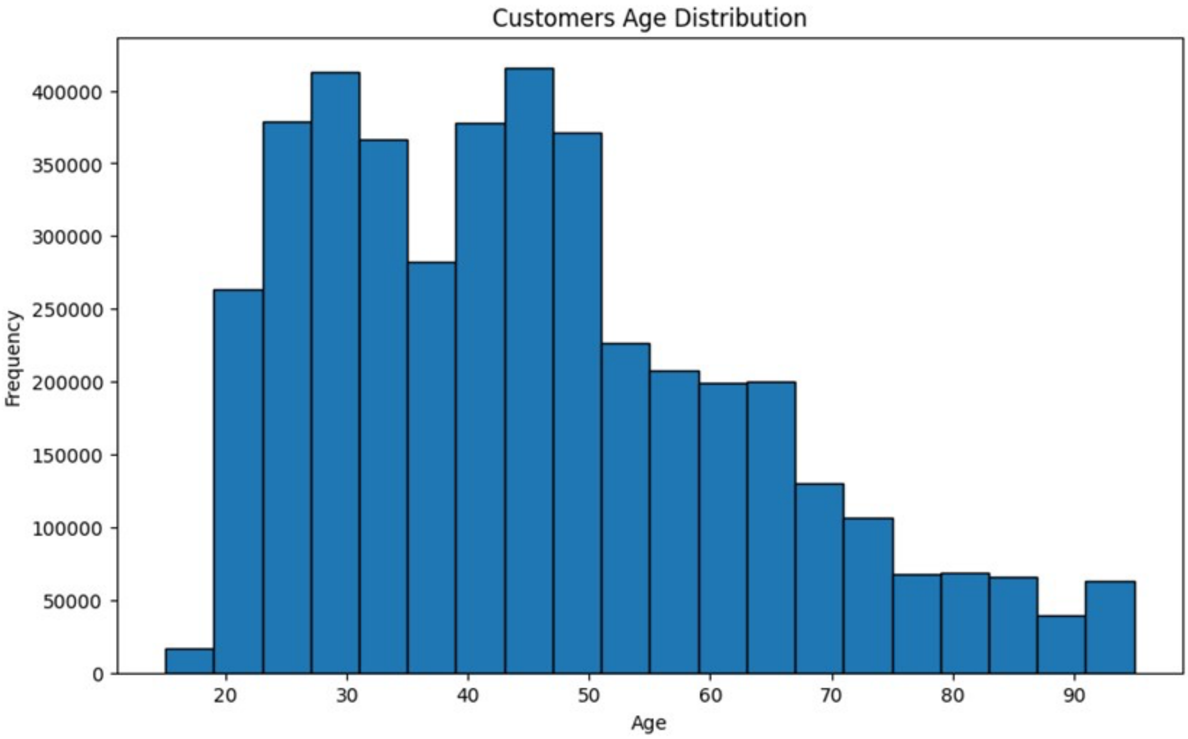
Explore data



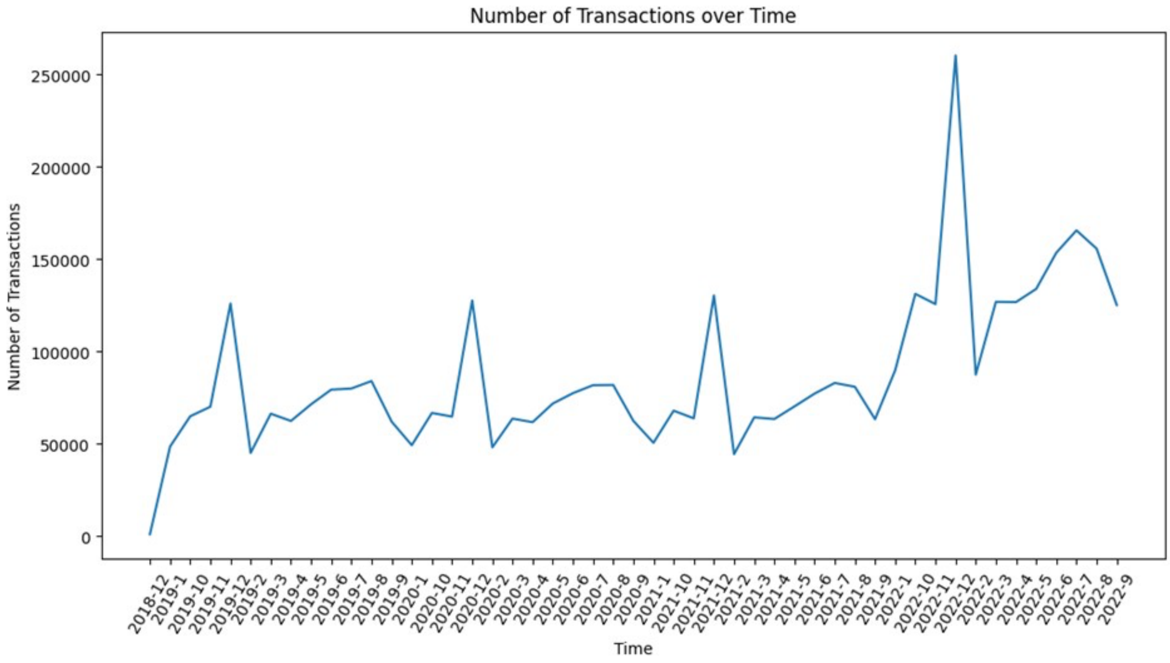
The given bar plot provides the information about the most common categories where the transactions were done. We can see that shopping ranking the highest followed by home shopping. The least amount of transactions were from travel, indicating the majority of the customers in the dataset spend less on travelling.



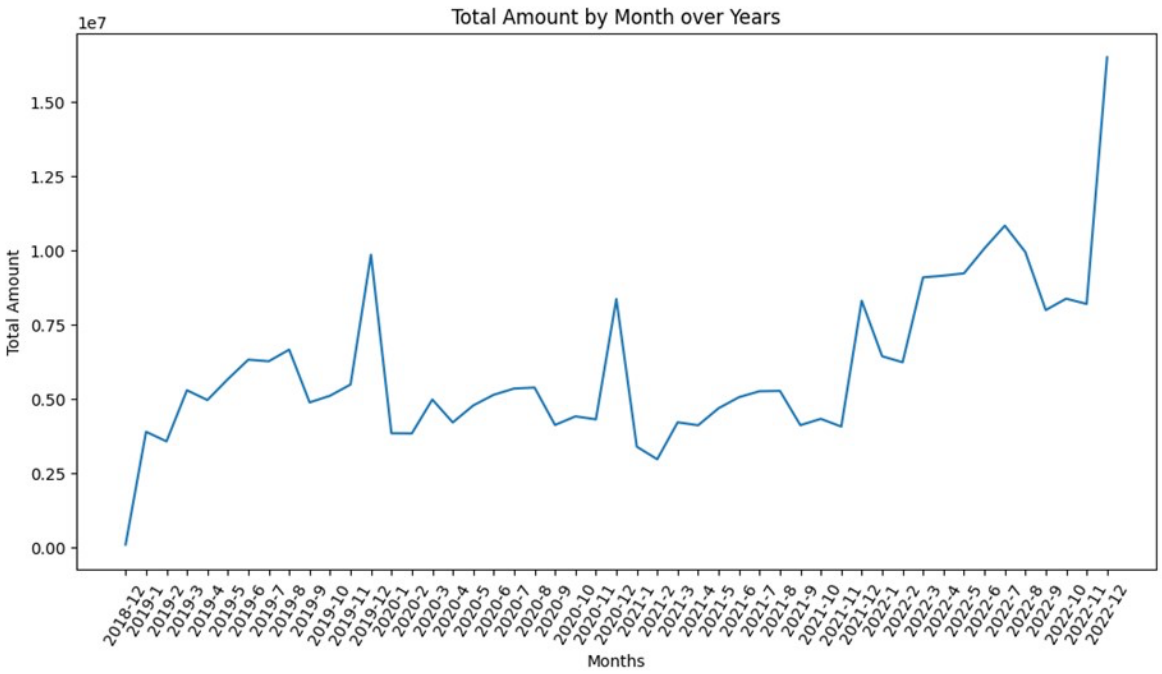
The amount of fraudulent transactions is only 5034 and 4,255,870 non fraudulent transactions out of a total of 4,260,904 values. This means that the data is heavily imbalanced in regards to fraud.



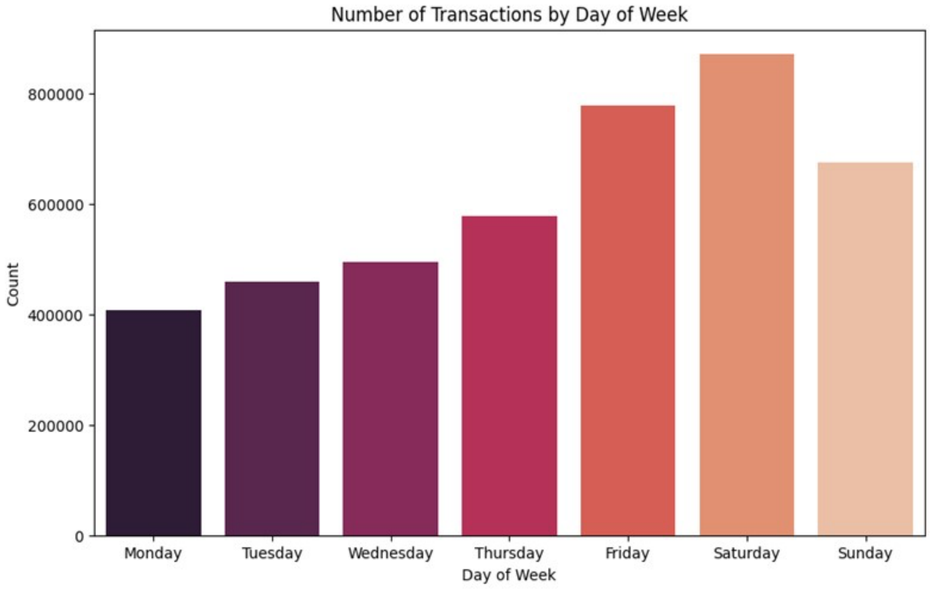
The majority of customers fall within the age range of 25 to 50 years, which corresponds to the phase in life where individuals typically establish a consistent income before approaching retirement age.



The given graph describes the total number of transactions over the years. A clear trend is observed, where the number of transactions rose during the end of the year, indicating that people spend more during the festive season. Post pandemic, it can also be interpreted that as people started using online banking more, their transactions spiked and the fall is still greater than the highest transactions in the previous years.



A clear seasonal pattern emerges in the data, showing a notable rise in the number of transactions and total spending amount during the holiday and festive season, particularly from November to December. Subsequently, there is a significant decline in both metrics. From February to August, there is a gradual increase in the number of transactions and total spending amount before experiencing a sharp surge once again in December. This cyclical pattern persists across multiple years, with the trend intensifying notably in 2022.



There is a noticeable trend indicating increased spending as the weekend approaches, particularly on Friday and Saturday. Conversely, spending tends to decline on Sundays as individuals prioritise staying at home to prepare for the upcoming workweek.

III. Data Preparation

Some data cleaning steps with feature engineering techniques to simplify the dataset and improve model performance are presented as below:

1. Extract temporal features from the "unix_time" feature, including year, month, day, and day of the week. These features capture temporal patterns and trends in the data. Remove the "unix_time" feature after extraction.

2. Calculating age from "dob" helps analyse spending patterns and predict its impact. Age interacts with other features to create informative variables, enabling personalised predictions for different age groups. Tailored financial advice and budgeting recommendations can be provided based on each group's needs. Remove the "dob" feature after calculating age.

3. Categorical feature encoding: Features like "gender," and "category" are categorical variables. We can encode them using techniques such as one-hot encoding to represent them numerically. This allows the model to understand the categorical information. Remove “gender” and “category” features after encoding.

4. Remove fraudulent transactions as they are not representative of normal spending behaviour. This eliminates noise and bias in the model, allowing it to focus on learning patterns from legitimate transactions. This improves the accuracy and reliability of predictions for customers' total spending amount.

5. Merchant-related features: Derive the distance between the customer's and merchant's locations ("merch_lat", "merch_long") to reflect proximity and accessibility. Including this distance feature captures the geographic influence on spending patterns. Calculate the average distance per customer over years and months.

6. Transaction aggregation: Aggregate transactions based on customer identifiers (e.g., "ssn," "cc_num") to summarise spending behaviour. Calculate statistics like total spent, average per transaction, maximum, minimum, and number of transactions by months over years. These features provide a summary of the customer's spending behaviour over time.

7. Aggregate demographic features ('lat', 'long', 'city_pop') by customer, then by year and month. These features capture geographic and population information that can impact spending patterns. 'Lat' and 'long' represent geographical data, while 'city_pop' indicates the population size of the customer's city. Population size can serve as an indicator of urbanisation and economic activity, which also influence spending patterns

8. Derive previous month amount and target month amount from "total_amt". Including the previous month amount helps the model learn the impact of past spending on current spending, capturing trends and temporal patterns. The target month amount represents the total spending for the next month and serves as the dependent variable for training the regression model

9. Remove redundant or irrelevant features: 'Unnamed: 0', 'ssn', 'cc_num', 'acct_num', 'trans_num', 'amt', 'total_amt', 'merch_lat', 'merch_long', 'merchant', 'is_fraud', 'first', 'last', 'street', 'city', 'state', 'zip', 'job', 'datetime', 'year_month', 'day_of_week', 'date'. Eliminating these features simplifies the dataset and enhances model performance by eliminating unnecessary or correlated features.

10. Extracting the target variable into a variable called y and removing it from the data frame allows us to create a separate dataset containing only the features we want to use to predict the target variable. This is important because our model should not have access to the target variable during training.

11. Saving all features from the dataset into variables called X to ensure that we have all the necessary features for training and testing our model.

12. Split the features and target variable of the dataset into training and testing sets with test size = 0.2. We then split the features and target variable of the training dataset into training and validation sets with test_size = 0.2. This choice ensures sufficient data for effective training on the training set.

Business Case 3

Helping the Marketing Team to send customised marketing emails to groups of customers presenting similar spending behaviour. (Varun Singh Chhetri)

I. Business Understanding:

This project aims to help the marketing team send customised marketing emails to a group of customers based on similar spending behaviours. By leveraging the clustering techniques, the marketing team can identify distinct customer segments with similar spending amounts, transactional patterns and spending behaviours. This can enable the marketing team to tailor their marketing messages according to the customer's needs and provide offers and deals to improve the effectiveness of the campaigns.
The given model can help the marketing team to identify what are the common spending traits and spending limits of the customer in each cluster. It can also help them understand which customer spends most amount of money in a particular category. This can mean, they can send the customer the products that fall into the category within the spending limit. This can make the customers more likely to buy the given product. Other than personalised customer offers, the team can also understand which age group spends the most amount of money and during what time of the year, people are more likely to spend.

II. Data Understanding:

Please refer to business case 1 for the data understanding

III. Data Preparation:

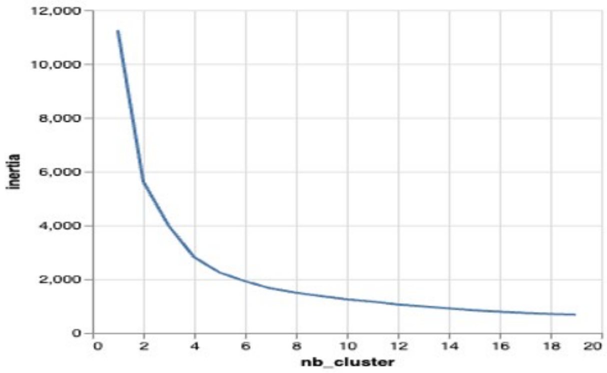
Initial data preparation was done as a team. Please refer to Business case 1 and note some changes I have made below.

1. Same as business case 1.
2. Same as business case 1.
3. Only the gender column was categorically encoded using one hot encoding. The category column was further mapped into four major categories – 'utility', 'miscellaneous', 'entertainment', and 'shopping'. These features were separately added to the new data frame, before training the model.
4. Same as business case 1.

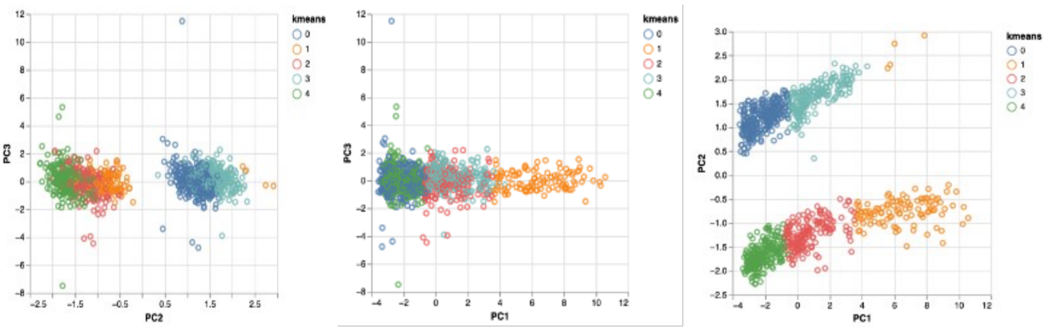
5. Same as business case 1.
6. Did not perform this step.
7. As I only needed the distance from customer to merchant, which was already calculated, I removed all the other columns containing the location information.
8. Did not perform this step.
9. A slightly different approach, where I only used the account number as a unique customer identifier, ‘age’, ‘job’ and ‘gender’. All the other columns were dropped, and the new features created using the feature engineering techniques were merged into a new data frame.
10. The steps 10-12 (as for business case 1) did not apply to the business task I am solving. Since in unsupervised learning, there is no need for data splitting. Instead, all the required features were merged into a new data frame, ‘X’, and further modelling and analysis were performed using the new table.

IV. Modelling:

During the modelling phase, several clustering models were considered for the analysis. There were 2 models that I considered for the data - K-means clustering and Hierarchical clustering.
The initial approach was the same for both models until feature engineering. To help the model better understand, the transactions dataset was used to find the customers spending for each year and each category. For all the separate data in these features, separate columns were created, and all the figures were calculated per account number. Columns ‘acct_num’ and ‘Job’ was dropped and will be used during the model analysis. Since K-means clustering is distance based, the distance is calculated using the Euclidean distance to ensure the larger features do not dominate. Hence Standard Scaling was performed. Using the elbow method, the number of clusters was calculated, and the graph was plotted as given below:

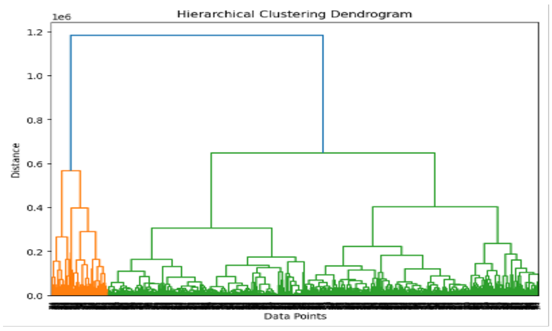


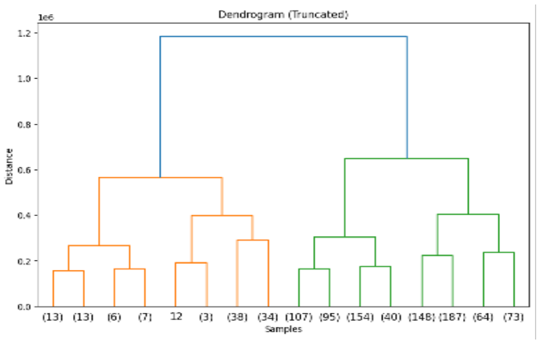
From the graph, the nb_cluster = 5 was determined. PCA was performed to understand the variance of the given features and visualise the correlation between the given features.



According to the plotted bar graphs, the variance relation with the features is signified and can be used to understand the specific feature's importance.

Most parameters were kept the same in the hierarchical clustering model. This distance metric was equal to Euclidean, and the linkage_used was complete. The number of clusters was kept the same so the feature evaluation would be easy during the data analysis. The given dendrogram and truncated dendrogram are plotted:





V. Evaluation:

Evaluating the performance of the clustering model can be a challenging task, as we know. That clustering is an unsupervised learning method. This means we do not have predefined labels or ground truth to compare the clusters. Since this is an exploratory technique, the evaluation metrics do not provide definitive performance measures. Hence, the model's performance analysis was performed by combining supporting visuals and using the silhouette score.

Silhouette score measures how well-defined the clusters are within the clustering algorithm. The score ranges from -1 to +1, where a score closer to +1 indicates the data points are well clustered and assigned to the correct clusters, while vice-versa for the score of -1.

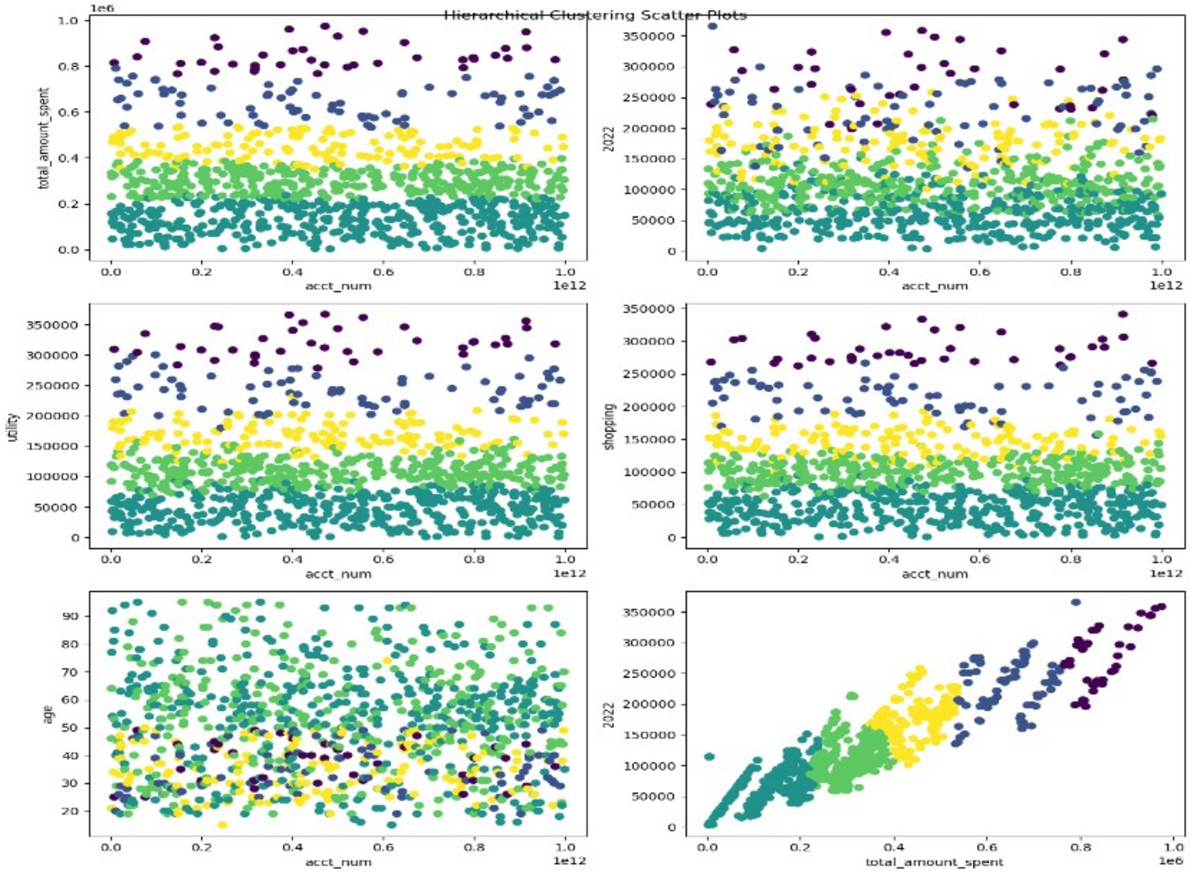
- K-means clustering – The silhouette score of 0.072936515614 was obtained, suggesting that the algorithm had a low level of separation and overlap. This indicates that the clustering results may not be very reliable or optimal.
- Hierarchical Clustering – The silhouette score of 0.44330879608 was obtained, suggesting that the model has a moderate level of separation and overlap. The data points are relatively well-clustered and assigned to appropriate clusters.

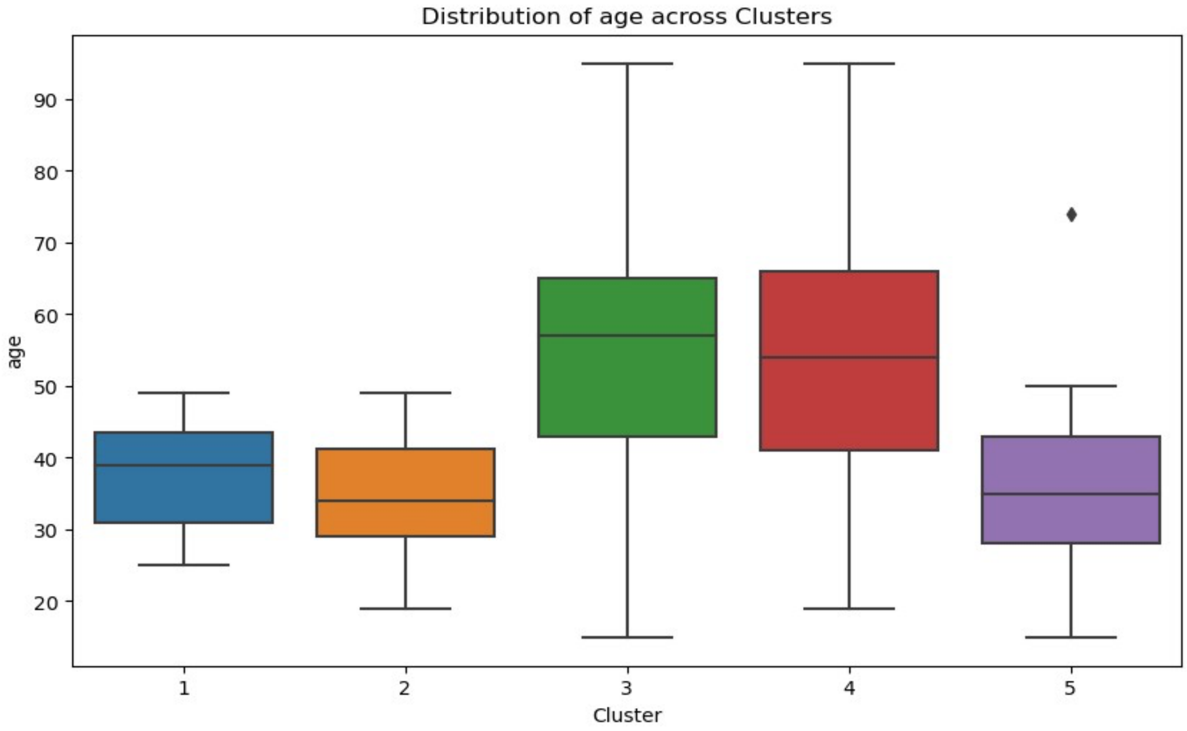
The comparison of scores clearly indicates that the hierarchical clustering model performs way better than the k-means, and it is recommended to use the latter one.

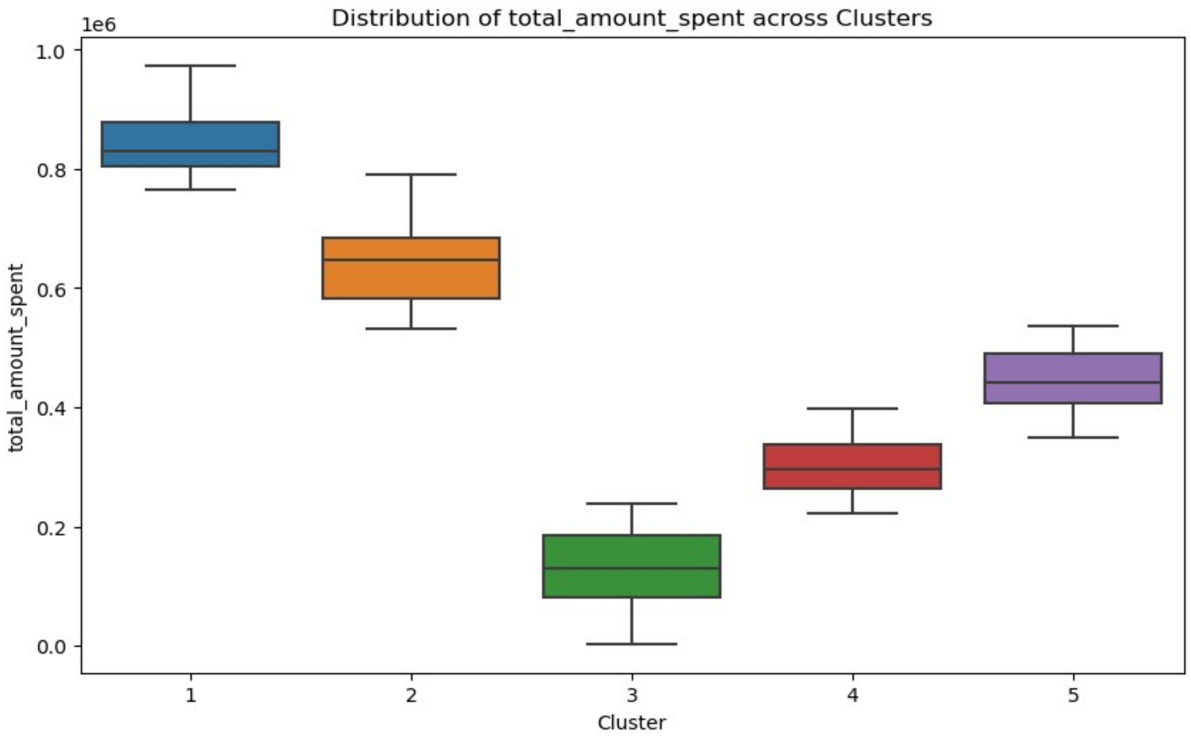
The given graphs were plotted to check if the hierarchical model was clearly able to distinguish the points in the clusters. The plotted graph can also be used to understand the patterns in clusters.

VI. Deployment:

The model scores and the visualisations help find some conclusions that might be useful to the marketing team.


Hierarchical Clustering Scatter Plots
total_amount_spent
1e6
acct_num
1e12
2022
utility
shopping
age
total_amount_spent
1e6


Distribution of age across Clusters
age
Cluster
1
2
3
4
5
20
30
40
50
60
70
80
90



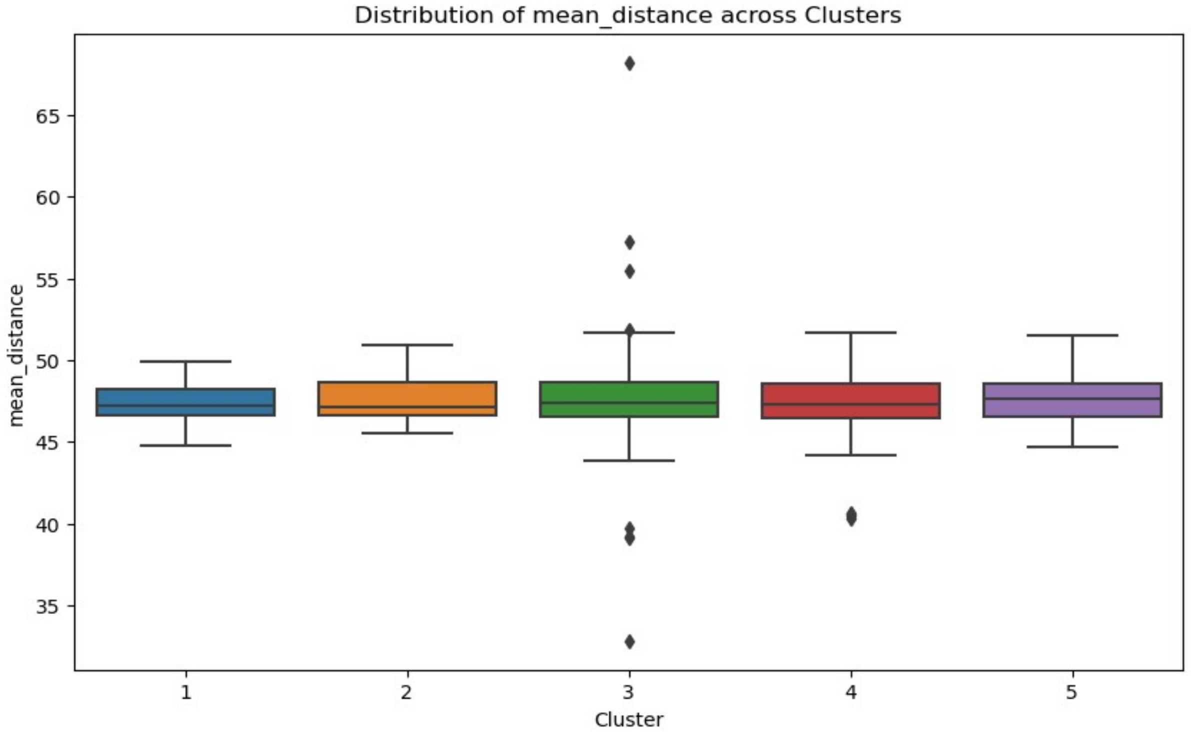


Fig 1: The distribution of age among the cluster will help the team identify any other person with the similar features that can be grouped to the specific cluster and a temporary score can be calculated.
Fig 2: The graph describes that all the account numbers in the cluster 1 have the highest spending score, and the team can target these customers, followed by Cluster 2. Cluster 3 has the lowest spending score, suggesting the team do not focus too much on them.

Fig 3: The given graph suggests that the average distance between the merchant and customer is around 45-50 miles, implying that distance might not be playing an important role because of factors like online shopping and free home deliveries.

The given model can provide sufficient insights to the marketing team, but it is also important to note that the model scores are moderately average, indicating further scope for improvement. These can be done by doing hyperparameter tuning or using some other clustering models like Mean Shift or DBSCAN as these models use linkage threshold while performing clustering and might have a positive or negative impact on the model. Although performance scores cannot provide the accuracy of the model and the best way to confirm it would be by implementing it on the customers. Their feedback can clarify if the model is helpful to them or not.

Data Privacy and Ethics:

Same as business case 1. The only column that contained sensitive information was the ‘acct_num’, which could reveal some information about the customer, but since the customers needed to be segmented, a unique identifier was needed that could map them to the cluster.