Report 3

Name: **Musadig Aliyev**

Project Title: **Implementations of Product Bundling and Pricing Strategies using Machine Learning Techniques in the Banking Industry**

Description of the Data:

In the Bundling model, I extracted raw transactional data for each service (merchant category) that I'm interested in for a period of 3 months. Initially, the data had 7 columns as below:

customer_no: A unique identifier for each customer.

trn_id: A unique identifier for each transaction.

mcc: A unique identifier for each merchant.

trn_dt: The date when the transaction occurred.

mcc_description: The description of merchant category code.

acc_amt_lcy: The monetary value of the transaction.

card_product_name: The name of card which transaction have been made.

After performing the initial preprocessing, I created a 'Merchant Category' column from the 'Merchant Category Code' using the international convention. For example, the code 5411 corresponds to the category 'Restaurants'

Description of the Measurement Strategies:

It is very important to understand to which category a set of variables belongs. The type of the variables determines the applicability of various statistical and data mining algorithms.

In our dataset, the 'merchant_category' and 'card_product_name' columns are categorical variables, representing the different categories of services/merchants and card products, respectively. These variables are in a nominal scale, meaning they have distinct categories without any inherent order or hierarchy.

On the other hand, the 'acc_amy_lcy' column is a numeric variable representing the amount in local currency. This variable is in a ratio scale, as it possesses a clear quantitative value and a meaningful zero point.

In bundling part of my research, I will build model using market basket analysis technique. Measuring the correctness or accuracy of market basket analysis typically involves evaluating the performance of association rules generated by the analysis. Here are a few commonly used measures to assess the correctness of market basket analysis:

Support: Support measures the frequency of occurrence of a particular itemset (combination of items) in the dataset. It indicates how frequently a specific itemset appears in the transactions. Higher support values indicate a higher occurrence of the itemset, implying greater importance.

Confidence: Confidence measures the reliability or strength of the association between items in the form of an "if-then" rule. It calculates the conditional probability of finding an item (consequent) given the presence of another item (antecedent). Higher confidence values indicate a stronger association between the items.

Lift: Lift measures the degree of dependency between the antecedent and the consequent in an association rule. It compares the observed support of the rule with the expected support if the items were independent. A lift value greater than 1 indicates a positive association, while a value less than 1 suggests a negative association.

Statistical Analysis:

Firstly, I have selected sample randomly from population to do statistical analysis. Random sampling helps to minimize sampling bias and increase the likelihood of obtaining results that are applicable to the entire population. However, it's important to note that random sampling does not guarantee perfect representation, as some variability is expected due to chance. So, I decided to test whether my random sample represents a population using the t-test for a continuous variable. The initial hypothesis for this test is as follows:

h0 = "There is no difference between the population mean and the sample mean."

h1 = "There is a significant difference between the population mean and the sample mean."

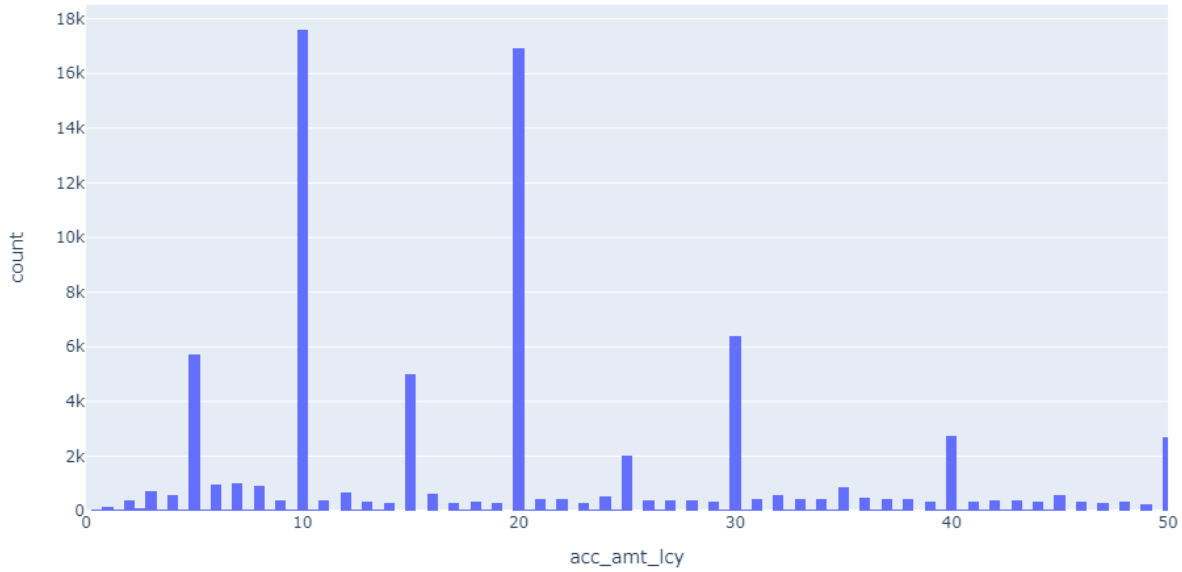
```
Population Mean: 21.94099747880097
Sample Mean: 22.122473836993994
Population Standard Deviation: 159.66986736108342
Sample Standard Deviation: 162.79757948129014
T-statistic: 1.1147356615070463
P-value: 0.2649639822404157
```

The t-statistic value of 1.1147 indicates the difference between the sample mean and the population mean, taking into account the variability within the sample. Since the t-statistic is close to zero and positive, it sugg ests that the sample mean is slightly higher than the population mean, but not significantly different.

The p-value associated with the t-test is 0.2649. This p-value represents the probability of observing the sample mean if the null hypothesis is true (i.e., the sample is drawn from the population). In this case, the p-value is relatively high (greater than 0.05 significance level), indicating that the observed difference between the sample mean and the population mean is not statistically significant. Therefore, we fail to reject the null hypothesis, which suggests that the sample is likely representative of the population.

After selecting a sample, I have decided to perform univariate analysis on the transaction amount column specifically for the petrol station merchant category. In case of numeric variables, we need to understand the central tendency and spread of the variable. These can be measured using various statistical metrics and visualization methods.

Histogram:



After visualizing data, I used moments to describe various characteristics of a probability distribution or a set of data.

First moment: The first moment is the mean, also known as the expected value. It represents the average value of a distribution or dataset.

Second central moment: The second central moment is the variance. It measures the dispersion or spread of a distribution or dataset. A higher variance indicates a wider spread of data points around the mean.

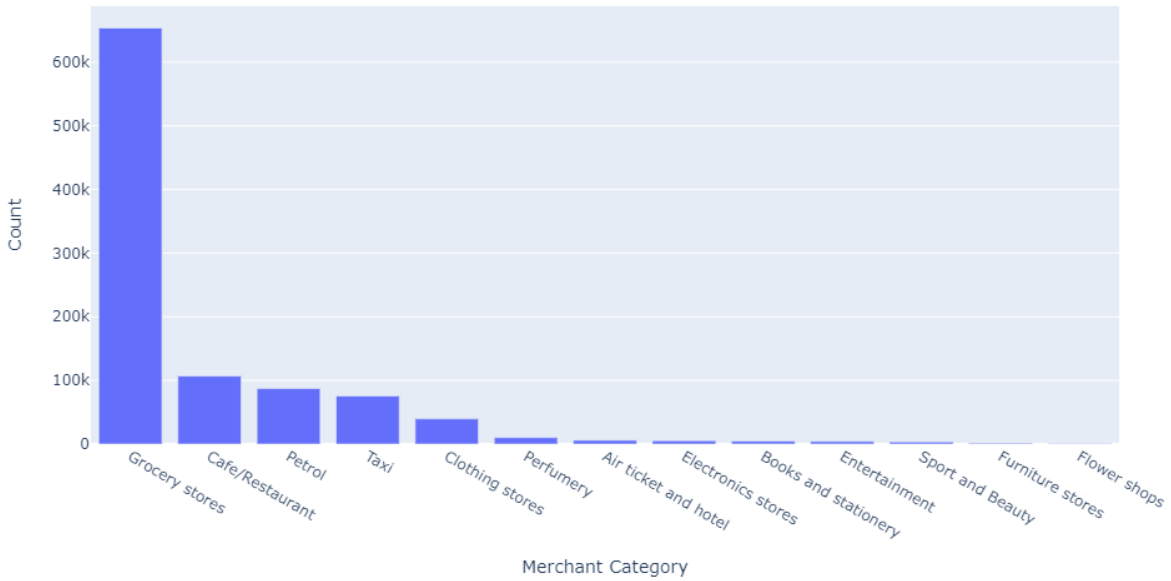
Third central moment: The third central moment is the skewness. It quantifies the asymmetry of a distribution or dataset.

Fourth central moment: The fourth central moment is the kurtosis. It describes the shape of a distribution or dataset by measuring the heaviness of the tails and the presence of outliers.

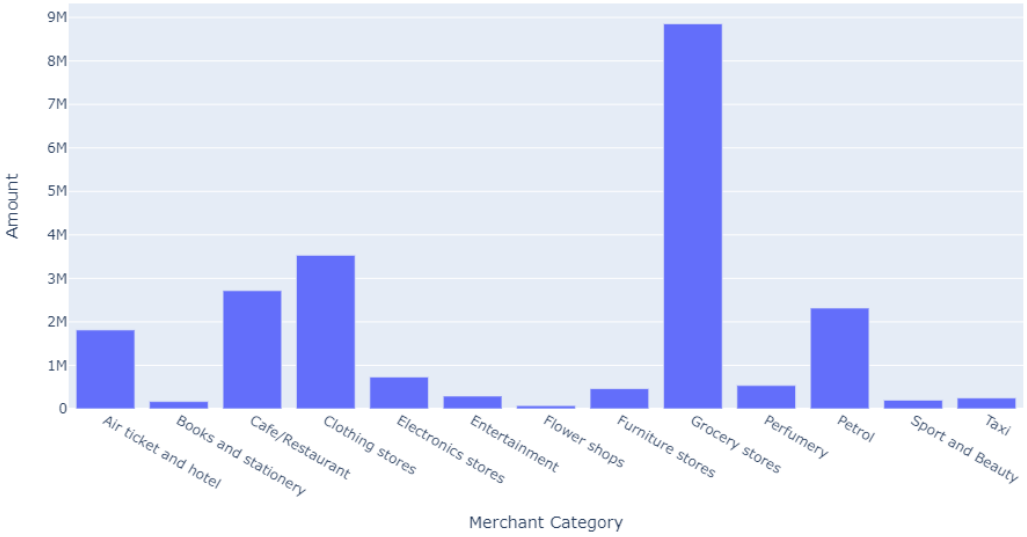
```
Mean: 26.76968894244116
Variance: 2267.748840113119
Skewness: 28.11798809811283
Kurtosis: 1612.2083578016152
```

Visualization:

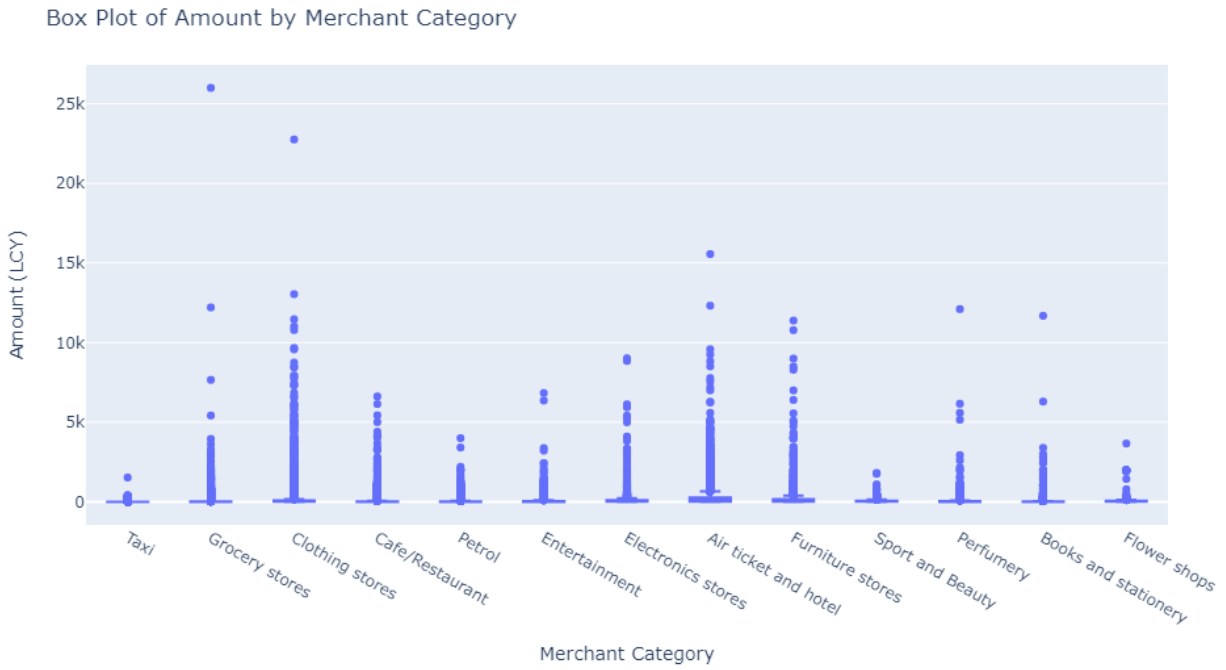
Bar chart:



Among the categories, grocery stores appeared to be the most popular, with a total usage count of 650,000 by customers. Café/Restaurants, petrol stations and taxis emerged as the next popular categories, each garnering a usage range of 75,000 to 100,000 transactions by customers. Flower shops, furniture stores and sport were less frequently used categories.



Although the hotel category is less frequently used by customers, its transaction amount is high. Conversely, while taxis rank as the fourth most frequently used category, their transaction amounts are relatively low.



We might observe noticeable outliers in the categories of hotels, books, perfumery, grocery, and clothing stores. On the other hand, the categories of restaurants, taxis, and sports do not exhibit many outliers in this graph.