**BANA 200: Foundations of Business Analytics**

**Starbucks Corporation Analysis**

**Assignment 1**

**Submitted by**

**Hari Sellappan (38562796)**

**INTRODUCTION:**

The study is about the ratings of customers at Starbucks Corporation conducted in Orange County. This study would be useful to conduct some insightful analysis to help senior management understand more about how to improve customer engagement and profitability. In this study, as a first step, the dataset is prepared for analysis by cleaning it. The dataset contains 10000 rows and 27 columns that include zip code-wise ratings, income, and profits at each customer level. As part of the exploratory data analysis, missing values are identified and rows corresponding to the missing values are eliminated. Later, to avoid input errors, the rating columns from X1 to X22 for those values less than 0 are replaced with 0 and rows with values greater than 5 are replaced with 5 to have the categories ordered from 1 to 5. Variables X1 – X22 are all measured on a 5-point scale (1 = terrible, 2 = poor, 3 = average, 4 = good, 5 = excellent). The same method is replicated for the column satis100 where less than 0 and greater than 100 are replaced with 0 and 100 respectively and for column recommend, values less than 0 and greater than 10 are replaced with 0 and 10 respectively. The summary of the original dataset is given below:

Text

Description automatically generated

**Q1: How many missing values are there for each variable and for the entire dataset?**

**Solution:**

The below table shows the column-wise missing values. Column X16 has the highest number of missing values whereas column X1 has the lowest number of missing values. The columns Zip code and Income have zero missing values. The overall missing values across all variables stand out to be 4926.

A picture containing diagram

Description automatically generated



**Q2: How many non-missing rows are there? Does it seem like we are throwing away a lot of data by removing all the rows with missing data?**

**Solution:**

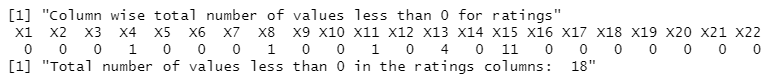


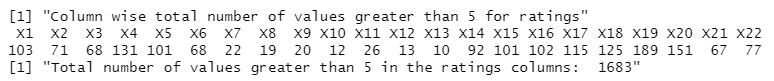
The dataset with any and all rows of missing values is removed. After removing the rows corresponding to missing values the shape of the dataset changed to 6121 rows and 27 columns. Nearly 39% of the data are removed. It is understood that the real-time dataset would have missing values. This may be due to many reasons like data collection problems, data entry errors or some customers would not wish to answer some questions, etc. Irrespective of the reasons, the missing values should be handled carefully as it biases the outcome. We lose a lot of information when 39% of data is removed whereas, for each column, it is not more than 3% of values missing. So, it is desirable to treat the missing values instead of removing them. In such case, for the categorical columns in the dataset like X1-X22, satis100, recommend and zip code, we can replace the missing values with **Mode** and for the numerical columns like profit, the income of the customer, we can replace the missing values with **Mean**. There are few other better methods to treat missing values like **KNN** rather than the naïve approach of filling the missing values with Mean and Median.

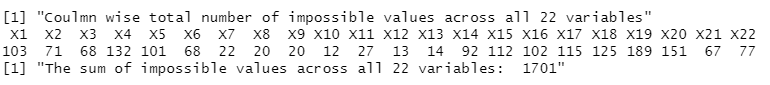
**Q3: How many impossible values are there for each variable? Also, report the total number of impossible values across all 22 variables (the sum).**

**Solution:**

This analysis was performed on the dataset with no missing values. The variables from X1 to X22 contain values outside the range of 1 to 5 which are defined as impossible values. The frequency of impossible values for each variable is calculated here and the sum of all impossible values is also reported here. It seems that column X15 has the highest number of values less than 0 and column X19 has the highest number of values more than 5. There are about 1701 impossible values in the variables from X1 to X22.







**Q4: Management has asked that for variables X1-X22, you replace the impossible values with better numbers and do a frequency count and report the total numbers of 1s, 2s, 3s, 4s, and 5s across all 22 variables AFTER replacement.**

**Solution:**

This replacement was performed on the dataset with no missing values. To correct the input errors, for any values less than 1 (< 1), replaced those values with 1 and for any values greater than 5(>5), replaced those values with 5. Later, to understand the frequency count of each category 1 to 5, frequency count is done across all 22 variables.

A picture containing calendar

Description automatically generated

There are a lot of customers who have given ratings of 4 and 5 which shows that most of the customers are happy with the service and there

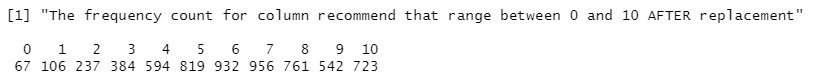


**Q5: The range of “satis100” should lie between 0 and 100, and the range of “recommend” should be between 0 to 10.**

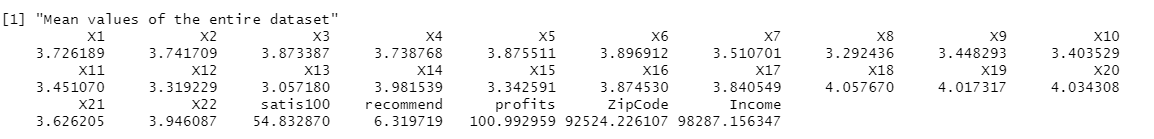
**Solution:**

This replacement was performed on the dataset with no missing values. To correct the input errors, for “satis100”, replaced any values that are less than 0 with 0, and replaced any values greater than 100 with 100. For column “recommend”, replaced any values that are less than 0 with 0, and replaced any values that are greater than 10 with 10.

For the column “recommend”, below are the counts of the number of unique values after replacing the impossible values. The rating of 7 has the highest number of counts which signifies that most customers likely recommend others.



The average values of all the columns are given below:



**SUMMARY:**

The summary of the dataset with no missing values is given below. By looking at the median of the rating columns in the dataset, it is understood that 50% of the customers have given ratings 3 or more for variables X1 to X22 and to decide whether they recommend others or not, about 50% of the customers have given a rating of 6 and above which shows that they likely recommend other people.

Table

Description automatically generated with medium confidence