Abhigna Ramamurthy

NUID: 002982276

ALY6040: DATA MINING APPLICATIONS

Module 1: Technical Practice

February 27, 2022

Table of Contents

[Introduction 2](#_Toc96895041)

[Data Clean up 2](#_Toc96895042)

[Data Analysis and Inference 3](#_Toc96895043)

[Outlier Analysis 4](#_Toc96895044)

[Correlation Analysis 5](#_Toc96895045)

[Future Analysis 6](#_Toc96895046)

[Conclusion 6](#_Toc96895047)

[References 7](#_Toc96895048)

# Introduction

The report below showcases exploratory data analysis of a Kickstarter campaign of an ice cream maker device company. There is a mixture of numerical and categorical data with 10,000 rows/11 columns of donation data. The goal is to clean and analyze the data to extract some insights through visualization and finally to suggest how machine learning can be employed.

# Data Clean up

The Kickstarter campaign data is put into an excel file with 10,000 rows and 11 columns. The initial structure of the dataset along with data type of columns, unique number of values and missing values can be seen in the below table 1.

Table

Description automatically generated

Table 1: Initial structure summary of the dataset.

There are missing data in Deposit Amount (0.4%), Ice Cream Products Consumed Per Week (1%) and Household Income (48%). As only 0.4% of Deposit Amount is missing and has a right skewed histogram, the 40 rows can be deleted (9960 x 11 is the new shape of the dataset) as Deposit Amount can be identified as main numerical contributor to the analysis and the deletion will ensure we are not disturbing the distribution of data. The missing 1% of data from Ice cream products consumed per week, with slightly normal distribution, can be replaced with the mean of the whole column as only 104 out of 9960 values are missing (Zach, 2021). Household Income has 48% of data missing. Considering that it’s not numerical data, there are two options, either to complete remove that column or to replace the mode of the column for the missing values. Weights both options, imputing the missing values with the mode makes more sense than completely loosing the column data.

# Data Analysis and Inference

In this section of the report, exploration of the available trying to find hidden patterns can be seen. Primary goal is to analyze every column and come up with visualizations to find patterns and form business questions for further exploration. The initial analysis of numerical values can be started with the below summary table 2. It shows the statistical analysis of the three numerical columns of the data.

Table

Description automatically generated

Table 2: Summary of numerical data.

We can make some interesting observation from this such as the median Deposit Amount collected is ***$100***. The population donated on an average have ***5*** ice cream products and 5 desserts per week.

Chart, histogram

Description automatically generated

Figure 1: Histogram of Donation date.

Donation drive was set up from July 1, 2019, to July 9, 2019. We can see from the histogram that the 7th and the first two days most of the donations were secured. On 7th July 2019, the highest donation of $10,000 was made. It is also important to note that on 3rd and 4th July there were no donations made. Total amount raised in these nine days is equal to $1,405,057.57.

Chart, bar chart

Description automatically generated

Figure 2: Number of donations per day grouped by gender.

An important categorical variable is “Gender”. During the nine days campaign we can see that majority of the money was donated by male population. We can also see an exception to this on day 1, i.e., 1st of July where female population almost doubled the donations made by male population.

Table

Description automatically generatedTable 3: Shows the preference of the Keurig device.

We can see that Silver has the highest demand but is in close competition with red and blue colors. This analysis would help the company concentrate on these colors initially. We can also see from the grouped analysis on gender that female population tend to like red or similar bright color device and male population like the silver color.

Chart, bar chart

Description automatically generatedFigure 3: Flavors of ice cream preference grouped by gender.

Another interesting survey data collected is on ice cream flavor. We can see that swirl has won the hearts of many people followed by vanilla. It is also interesting to note that a very close number to swirl votes said they had no preference. A lot of male population voted for no preference whereas majority of the female population preferred swirl.

The theory of economics suggests that if people are used to a particular device, they would be interested in upgrading them. We can see that in action here. The total number of donors who owned a Keurig before donated more than double that of people who did not. This shows that the main market focus needs to be on users who owned Keurig device before. It is also interesting to note that among the 9,960 donors, 3,235 were new donors. This is a good key performance indicator of the campaign.

## Outlier Analysis

Outliers are special observation data points which call for further analysis. Some can be influential data points which may steer the analysis into producing wrong outputs. The analysis is done on all the numerical values of the dataset below.

Graphical user interface, application

Description automatically generated with medium confidence

Figure 4: Box plot of Deposit Amount grouped by the ice cream flavor preferences.

We can see that $10,000 donated on 7th July shows up as an outlier. This could potentially be an influential variable; however, veracity of this donation can be crosschecked with the campaign coordinators.

Chart, box and whisker chart

Description automatically generated

Figure 5: Like figure 4 from above, ice cream products consumed per week is box plotted against the difference flavor here. We can see no outlier data point. The median consumption of ice cream products is five per week. This gives a basis on which financial forecasting can be done during further analysis.

Graphical user interface, application, timeline, Teams

Description automatically generated

Figure 6: Box plot of how many desserts consumed per week.

We can see that there is an outlier with 100 desserts per week as a data point. This could be an abnormality and due to its effects on the analysis, this data point can be removed.

## Correlation Analysis

Correlation analysis using correlation matrix and scatter plot matrix helps in further analysis of statistical correlation between the numerical values of the dataset. This helps in feature selection, regression, or other data modelling techniques.

Chart, treemap chart

Description automatically generated

Figure 7: Correlation matrix for the numerical values of Kickstarter dataset.

The correlation matrix values range from -1 to 1. Negative values showing a negative correlation and positive values indicate positive correlation. The diagonal is 1 as correlation between the same two values is the strongest. The matrix can be seen to be a symmetric matrix. The matrix suggests that none of the numerical data provide shows any strong negative or positive correlation as all the values are very close to 0. Hence, we can conclude that further information is required to do modelling on this data.

# Future Analysis

If we have further information on what the goal of the campaign was, we can evaluate if the campaign was a successful one or not. Recommendations on how to improve the aspects of the campaign can also be analyzed. As we can see from the Figure 1, 3rd and 4th of July, there were no donations made at all. Hence analysis on how we can utilize those days better can be recommended. Performance parameters of the campaign such as total donation amounts, incoming deposited amounts analytics like in Figure 1 and the duration of the campaign can give insights into how the campaign can be improved. Through the correlation analysis in Figure 7, we can conclude that the data as such is not fit for data modelling. Further data collection can help in campaign modelling such as forecasting deposit amounts, target consumers through previous campaign analytics and better consumer conversion patterns can be found using machine learning algorithms.

# Conclusion

Through this report exploratory data analysis using Python language was successfully done. Data cleaning is one of the first steps into any data analytics projects and understanding what method to use in different situations can be understood through this report. Visualizations using some powerful python libraries like bamboolib, plotly, and mathplotlib helped in visually analyzing the data. Insights such as gender groups, previous Keurig owners, favorite device colors, favorite ice cream flavors can be used for further analysis as mentioned in the above “Further Analysis” section.

# References

Getting Started - bamboolib. (n.d.). Welcome - we're joining forces with Databricks - bamboolib. <https://docs.bamboolib.8080labs.com/documentation/getting-started>

Grouped barplots — seaborn 0.11.2 documentation. (n.d.). seaborn: statistical data visualization — seaborn 0.11.2 documentation. <https://seaborn.pydata.org/examples/grouped_barplot.html>

Plotly Python Graphing Library. (n.d.). Plotly: The front end for ML and data science models. <https://plotly.com/python/>

Python Tutorial. (n.d.). Biggest Online Tutorials Library. <https://www.tutorialspoint.com/python/index.htm>

Zach. (2021). <https://www.statology.org/when-to-use-mean-vs-median/#:~:text=It's%20best%20to%20use%20the%20mean%20when%20the%20distribution%20of,when%20there%20are%20clear%20outliers.>