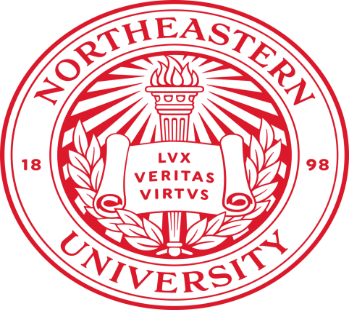
**Module 1 Technique Practice**

**Course:** [**ALY6040.**[**21452.202225**](https://northeastern.instructure.com/courses/97871)](https://northeastern.instructure.com/courses/92884) **– Data Mining**

Date: January 22nd,2022



**Submitted to Prof. Justin Grosz**

**By**

Aytaj Khankishiyeva

**Introduction**

This project is worked on the Kick Starter Dataset. The purpose of this project to observe variables, their data types, values and make exploratory data analysis. Using different functions, I am planning to look through summary statistics of variables about Kick Starter Dataset, create some graphs, charts to understand variables and gain meaningful insight from them. I will use quantitative method to explain and summarize dataset by numerically. Additionally, using visualization method to describe data with graphs, charts, plots and so on.

**Data Cleaning**

From the part of Data cleaning process, I checked to identify whether where any null values exist. I observed null values are on hand in Kick Starter dataset for the variable Deposit Amount with 40 amounts, Ice Cream Products consumed per week with 104 amounts and Household Income with 4801 amount that is almost half of whole variable. It is not good idea to remove missing values without observation. I must handle null values either removing null records or imputing with mean or median values. Firstly, I started to work on Deposit Amount’s null values, which are replaced with median values as the Deposit Amount variable has outliers and slightly skewed based on the observations from boxplot(figure2). Median value is close for substituting compared to mean, thereby median is better choice. Because median is more powerful to outliers rather than mean and 1-2 utmost values can affect the mean a lot, however, doesn’t change median a lot. Secondly, I replaced Ice Cream Products consumed per a week with mean, because observation from boxplot, I didn’t detect outliers for this attribute. Therefore, I replaced missing values with mean of this variable. Thirdly, Household Income attribute with 48% missing value isn’t significant to use with analysis. Because almost half of variable is unknown. From my perspective, better way is to delete entire column. Furthermore, I checked data to see whether there exists any special characters or duplicate values, but I couldn’t find any values.

The second step is checking data types of attributes. There is a variable named Donate ID, that is in numerical format can be converted to string. I didn’t need any data type conversion on this because this field doesn’t have any significance and will not be considered for analysis. There are categorical variables like Gender, Donate Date, Preferred Color of Device, Favorite Flavor of Ice Cream, Donated to Kick Starter Before and Do you own a Keurig that can be converted to integer format through concept of ‘OneHotencoder’, however it will be planned for next stages of analysis.

As a part of this process, it is required to check outliers from looking through boxplot. From the boxplot and histogram for Deposit Amount variable, I have identified outliers for this variable. Outliers for Deposit Amount are indicated between 151 and 400 and 10000 (maximum point). Outliers also are observed for How many desserts do you eat a week in the boxplot and histogram it occurs above upper broads of the variable.

**Exploratory Data Analysis**

For exploratory data analysis, I did obverse the number of observations I am trying to deal with and what kind variables and their data types. Then I focused on understanding the distribution of each attribute which one is numerical, which one is categorical data type, and their frequency level. The Kick Starter data set consists of 10000 observations/records and 11 attributes. The attributes of the dataset are Donate ID, Donate Date, Gender, Deposit Amount, Preferred Color of Device, Ice Cream Products Consumed Per Week, Favorite Flavor of Ice Cream, Donated to Kick Starter Before, Household Income, Do you own a Keurig, and How many desserts do you eat a week. There are 4 numerical attributes and 7 object type attributes.

Using the Statistical summary, I saw that the average of deposit amount is observed as about 141 and the minimum amount of deposit is shown as 100 and the maximum amount for deposit is 10000 with 9960 counts. It shows that there are 40 missing values for Deposit Amount. The average consumption of ice cream weekly is about 5 (4.96), minimum observation is 0, maximum level is 10 with 9896 counts. It is clearly seen that there are missing values for Ice cream product consumption per week. Maximum amount for Desserts you eat per a week is 100, and mean of this amount is above 5.

Now I am interested in some questions to answer using the exploratory data analysis. There are included how many desserts by each of gender are eating per a week, how many donation to kick starter before by each of gender, which color of Keurig device owned mostly or which not, which ice cream flavor consumed per a week mostly, favorite flavor of ice cream for each gender and their counts.

For giving answers the mentioned above business questions I created some visualizations which are found in Appendix. Graph is created to see how many desserts eaten by each of gender per a week. From the graph, I recognize that the most male eat 4 desserts per a week, the most of female eat 5 desserts per a week. About 120 male and 100 female don’t eat desserts weekly. The next graph which is about donated to kick starter before by each of gender, provides information that male is donated to kick starter before highly compared to female. Although that male is more interested to donation, however there is almost same level for male with female for don’t donated to kick starter before. Then I identified by helping graph which shows that count of owned Keurig in preferred color of device. From the observations, I saw that silver color is the highest demanded for Keurig owner. Black and no preference are the same level for Keurig owners. The most undesirable color for this device is red. The graphs which provide ice cream consumption per week in favorite flavor. The observations show that people who consume 8 ice creams per week, generally don’t have any preference, the people who consume 10 ice creams per week prefer mostly vanilla and as second choice is chocolate. Finally, favorite flavor of ice cream by each gender, female prefer swirl while male don’t have any preference for ice cream. Female prefer at least specialty flavor of ice cream while male prefer at least chocolate flavor of ice cream.

I didn’t make correlation to check relationship between variables. Because I have only 4 numerical variable and it is clearly seen that it is almost impossible to have relationship between them.

**Conclusion and Further Steps**

So I completed exploratory analysis, it is clearly seen that there is no target variable to create model and identify which variable and how they affect each other and specially to target variable. If dataset will have any adjustment, my next step will be data preprocessing and modelling. Before modeling the data, I must convert the categorical variables to numerical variables. Then dataset can be divided into testing and training dataset before proceeding with modelling.

**References**

Araujo, I. (2022, January 4). Bamboolib: One of the Most Useful Python Libraries You Have Ever Seen. Medium. Retrieved on January 22nd,2021, from

<https://towardsdatascience.com/bamboolib-one-of-the-most-useful-python-libraries-you-have-ever-seen-6ce331685bb7>

pandas.DataFrame.boxplot — pandas 1.4.0 documentation. (n.d.). Pandas. Retrieved on January 22nd,2021, from <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html>

Z. (2021f, September 28). How to Calculate Correlation Between Categorical Variables. Statology. Retrieved on January 22nd,2021, from <https://www.statology.org/correlation-between-categorical-variables/>

T. (2021e, September 21). *Data Cleaning in Data Mining: A Critical Step*. Trifacta. Retrieved on January 22nd,2021, from <https://www.trifacta.com/data-cleaning-in-data-mining/#:%7E:text=Data>

\

**Appendix**

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedShape, square

Description automatically generatedChart, histogram

Description automatically generatedA picture containing graphical user interface

Description automatically generatedChart

Description automatically generatedA picture containing timeline

Description automatically generated