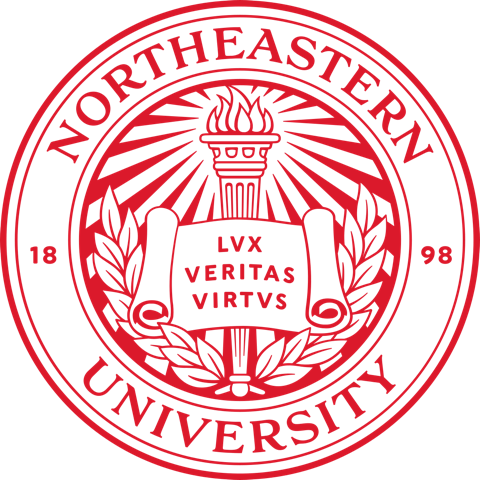
**Course:** ALY 6040

**Instructor’s Name:** Prof. Dr. Justin Grosz



**Submitted By:**

Sunil Raj Thota

**Date:** 10/01/2021

**Introduction**

The Kickstarter dataset from shows an insight into many donations with varying backgrounds and degrees of interest. Donation programs are either effective if they raise the required quantity through fundraising, or because they are unsuccessful when there is insufficient participation. In this, I have looked at the associations in the information to see if there were any markers of success of the project that I could use. I have gone through various characteristics of the information and displayed the findings after a primary data cleansing step.

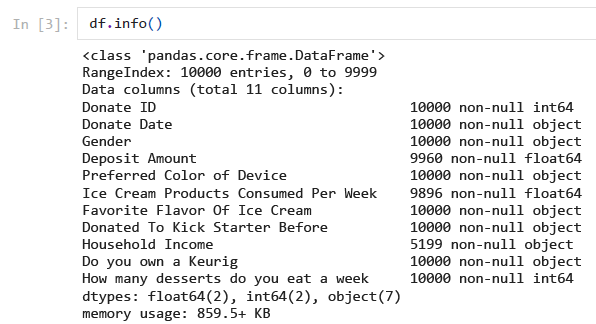
**An Overview of the Dataset**

**What did you do with the data in the context of exploration?**

I did not have a clear thought on this initially about the dataset. I went ahead and done some research online to know what to accomplish in this task. I have a goal in my mind when I decided to perform the analysis in the project. Let’s use Python to gain fresh insights into the dataset and explore further. I have structured the project into 2 parts they are data cleansing and exploratory data analysis. Performing data cleansing is essential to know the data more and make it useful for deriving insights. And also making use of proper tools and techniques to clean the data. This step is very important in any data science and analytics project that we are going to deal with. Mainly deals with the missing data, finding anomalies, and proper data types.

**How many entries are in the dataset?**

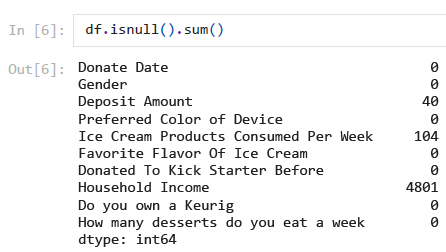
In this dataset, there are 10,000 records with 11 attributes with a mix of int, float, string, and date data types.



These are the fields that are present in this dataset:

* 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
* How many desserts do you eat a week

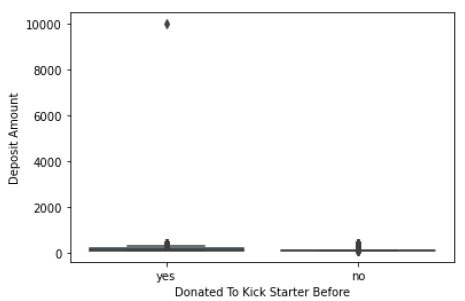
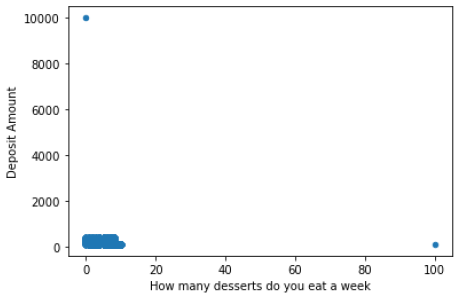
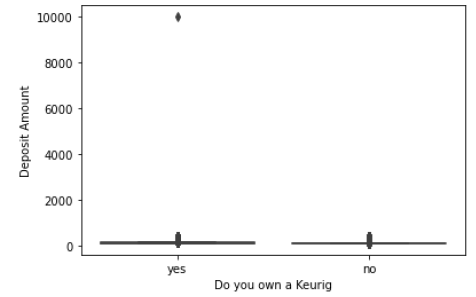
**Was there missing data? Duplications? How clean was the data?**



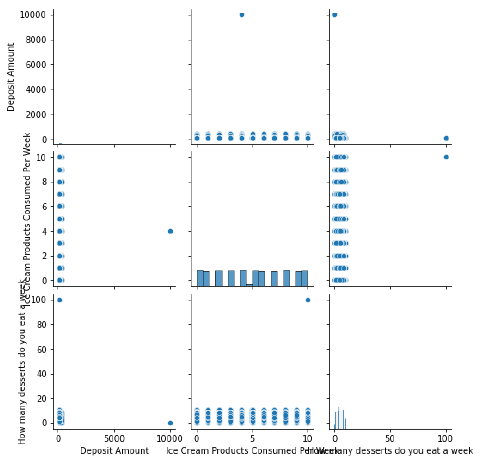
Yes, there is missing information in the 3 columns mainly “Deposit Amount”, “Ice Cream Products Consumed Per Week”, “Household Income”. Each of these columns has 40, 104, and 4801 missing values respectively. Other than this the rest of the dataset looks clean. As the Household Income column has 48% of the missing values. So, I am dropping this column. I also checked with the duplications if any but from the analysis, there is no duplicate information.

**Were there outliers or suspicious data?**

Throughout this assignment, I have gone through the data analysis steps and observed very few data points as outliers. There are techniques to detect and remove the outliers like Boxplots, Scatter Plots, IQR Scores, and Z-Scores. As there are not many outliers in the dataset, we will be not removing any of the data points and include them for further data analysis.

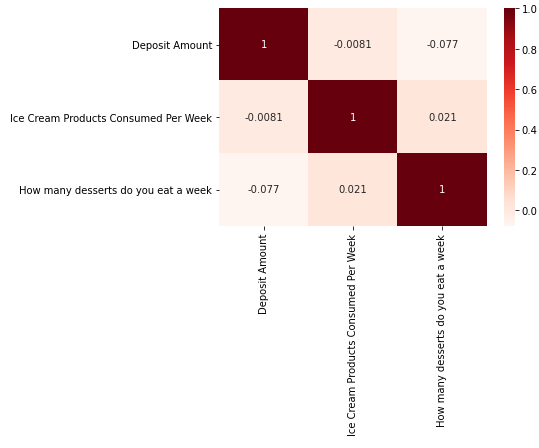
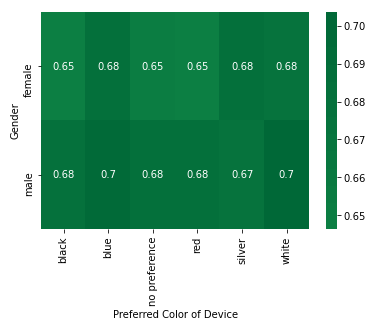
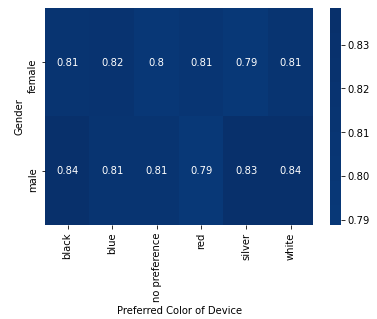
  

**What did you find? What intrigued you about the data? Why does that matter?**

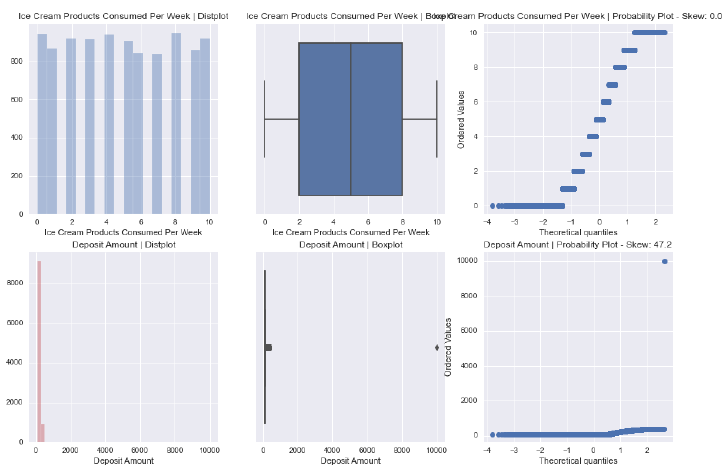


I have found a couple of interesting analyses from the dataset. 48% of the Household Income column has no values so I have removed that column. This gave me a broader approach to dive deep into the EDA. I have imputed the Deposit Amount Column with the mode value and the missing values are filled with it i.e., 100. Similarly, I have imputed the Ice Cream Product Consumed Per Week Column with the mean values of it and rounded it off to 2 decimals. After performing these steps, I have checked for the null values and found out to be nothing. I have included only numeric data types to plot a histogram and see the frequencies of the columns. Also, done some descriptive analysis to observe the count, mean, std, min, 25%, 50%, 75%, and max values for these 3 columns.

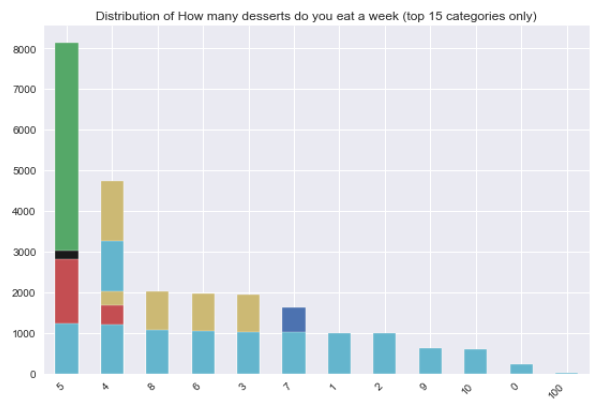
Also, I have observed this data how is distributed over the Scatter plots and found out to be fewer outliers. Used a Pair plot to visualize the data to interpret the connection between the continuous or categorical variables. It allows me to visualize bivariate distributions in the dataset.

Then, I have used the correlation plot to indicate the data points that lie no linear correlation between two variables and to indicate a positive linear correlation between the 2 variables. Identified the number of values with yes and no in the Donated to Kick Starter Before column and found out that 67.56% of them have donated to the Kick Starter. Also, I have identified that 81.36% own a Keurig. Used Seaborn’s heatmap to observe the preferred color of the device by gender. And, in this column, there are 6 categories in which the No Preference category has females with 0.68 whether they have donated before or not.



Also, I have utilized a library called “AutoViz’ which is mainly performed Automatic Visualizations of any dataset instantly. For this, we need to initiate the library and feed with respective parameters to automatically start visualizing the entire dataset. It generates quick Visualizations that are ready to observe and make strategic decisions with insights provided.



**What would your proposed next steps be? How do you plan to approach the cleansing of the data?**

I am looking forward to implementing and check other data columns and then perform some exploratory data analysis. After this is done, I will be working on the data cleaning techniques like Scatter, Box, and IQR Scores plots to identify the outliers if any. Also, drop/ impute the necessary columns with appropriate statistics. Once this is done, I will merge the learnings of Data Cleansing, EDA, and Data Modeling and utilize some ML Models to interpret the data insights. Mostly, I will be using classification models like Logistic Regression, Decision Trees, Gradient Boosting, etc.

Also, we should be capable of classifying our data integrity standards and also our industry requirements to be able to link data quality problems to industries' need.

**References:**

[1] Pandas documentation. pandas documentation - pandas 1.3.3 documentation. (n.d.). Retrieved October 1, 2021, from https://pandas.pydata.org/docs/