**Project 1**: Data Preprocessing

**CSC 177-01**: Data Warehousing and Data Mining

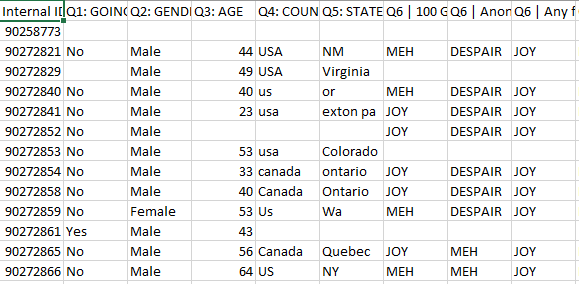
**Professor**: Jagan Chidella

**Group**: Jason Phillips, Mohammad Ameri, Ryon Faroughi, Youser Alalusi, Yusran Sadman

**Introduction to the Data:**

The data that was selected for this preprocessing assignment was a survey study of the types of emotions people feel when they are asked about various candies. In this survey, participants needed to choose from three different emotions when asked about different candies; these emotions ranged from “DESPAIR”, “MEH”, and “JOY”. Other questions that were presented to the participants of the survey include age, country of location, gender, and a few others. We chose this dataset because we thought it would be interesting to be able to find out which candies or chocolates were most popular amongst the people, and whether there were any relationships between the attributes.

**Issues with the Data:**

****

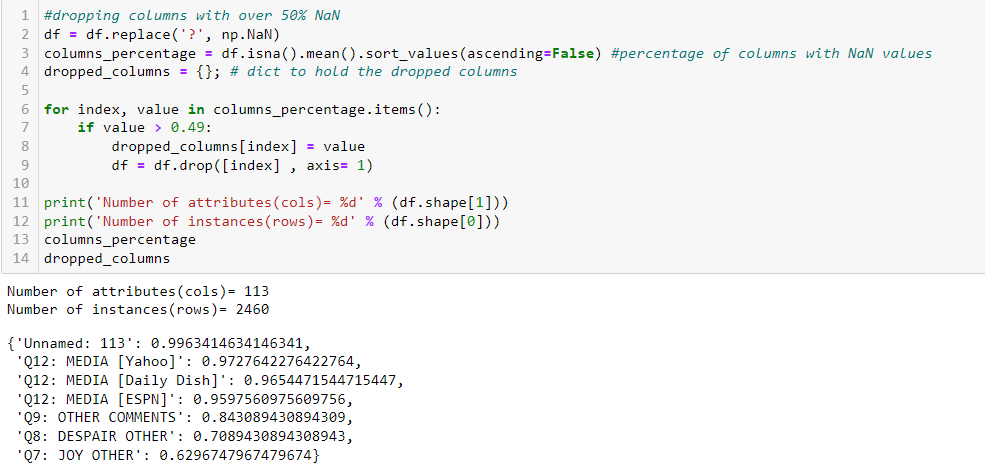
There were many issues with the initial dataset, which was ideal for this project. A few of the issues that we needed to fix include: missing values, unnecessary columns, off topic data, non-serious answers, outliers, and inconsistent formatting. Since this dataset was already “unclean” there was no need to modify the CSV file in order to preprocess it.

In order to address these issues in order to make the file usable and ready for data analysis, we followed many of the techniques that are displayed in the Tutorial 4 file on canvas. Some of the techniques we used were dropping columns, replacing missing values, dropping rows, and removing outliers. The main goal was to take this data and make it clean and presentable to the point where it can be properly analyzed.

**Cleaning the Data:**

There were many techniques used to clean the data, but we are only listing the major points here to demonstrate the order in which we did things, and why we deemed it necessary for this dataset.

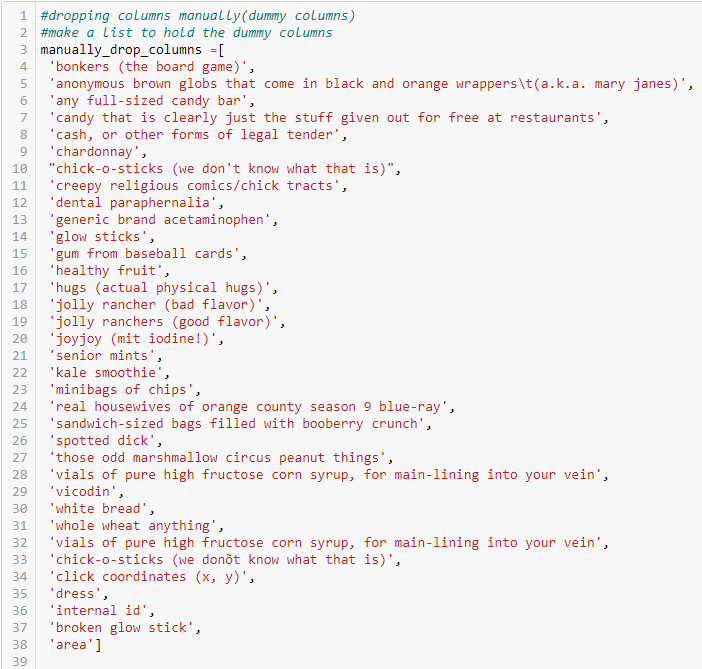
**Dropping columns with over 50% NaN**

After reviewing the data, we noticed there were many columns that had very little to no data in them. We decided to drop the columns that had over 50% NaN values.This was a necessary first step because we didn’t want to deal with attributes that didn’t provide much to the research.

**Renaming a Few Columns**

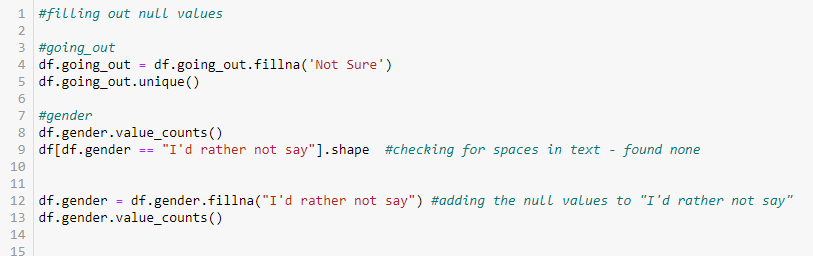
The next step we took was renaming a few of the columns, and removing the “Q” prefixes that were present in every column. This was done for better readability and would make calling certain columns later a lot easier.

**Dropping Dummy Columns**

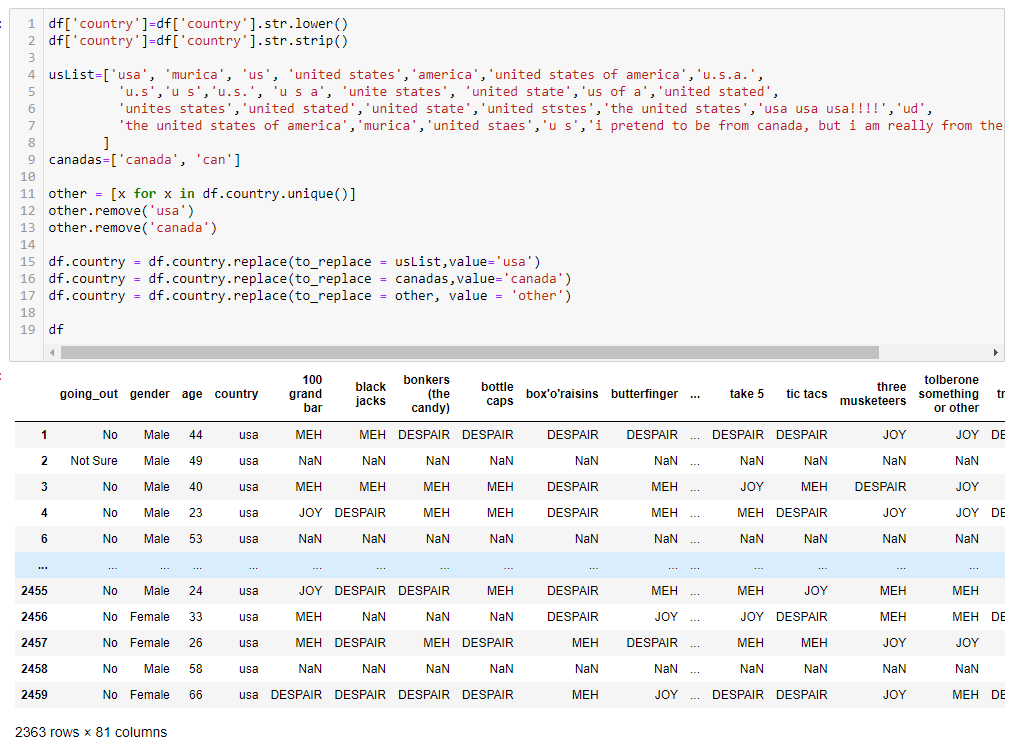
The next step we took was very crucial to be able to analyze the data properly. The dataset that we started with initially had a lot of columns that were not inherently related to the rest of the survey. For example, there was a question referring to “bonkers (the board game)” and not “bonkers” , the candy. We manually dropped the columns that would not help us when we analyze the different emotions people feel when asked about various candies.

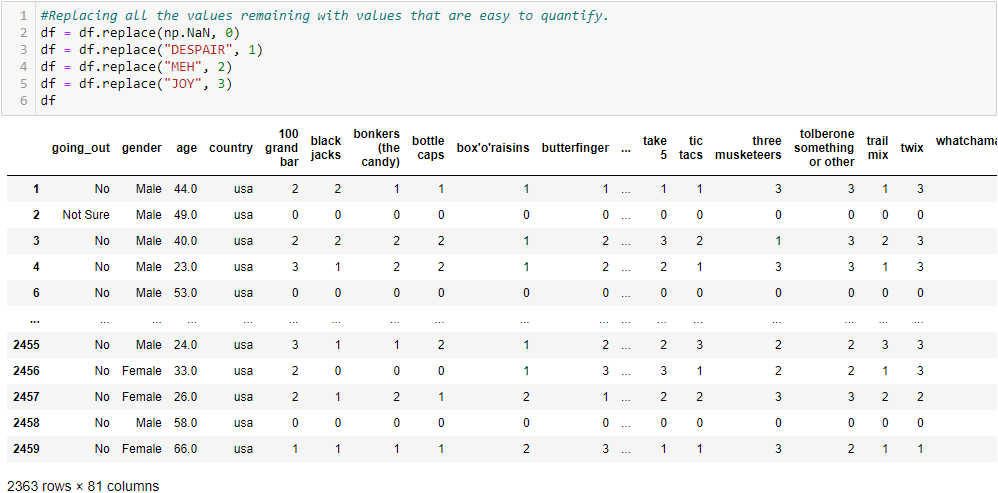
**Replacing Values**

After applying a few more tweaks to the data, the next action we took was to replace certain values in the dataset. Some participants would answer some questions and not others, and we did not want to completely disregard them in this study. Instead of removing them from the data frame, we decided to replace empty values with other answers such as “I’d rather not say”, which was a choice for some of the questions the participants were asked.

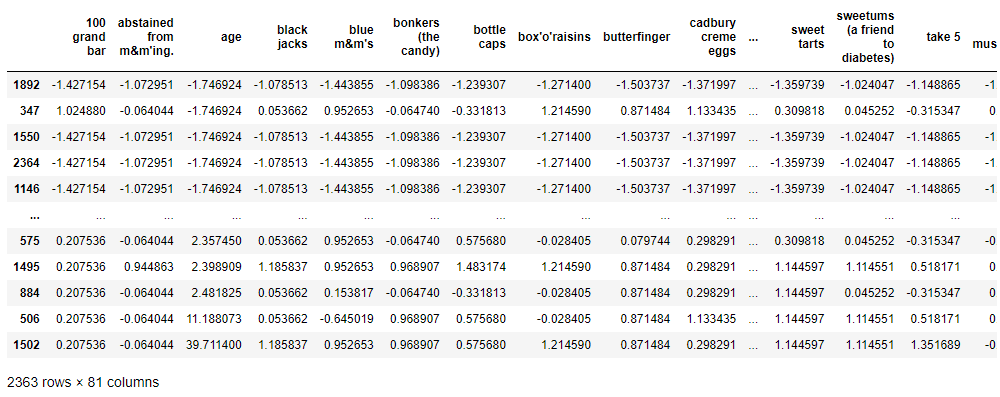
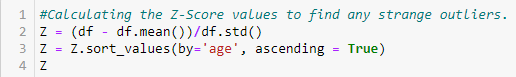
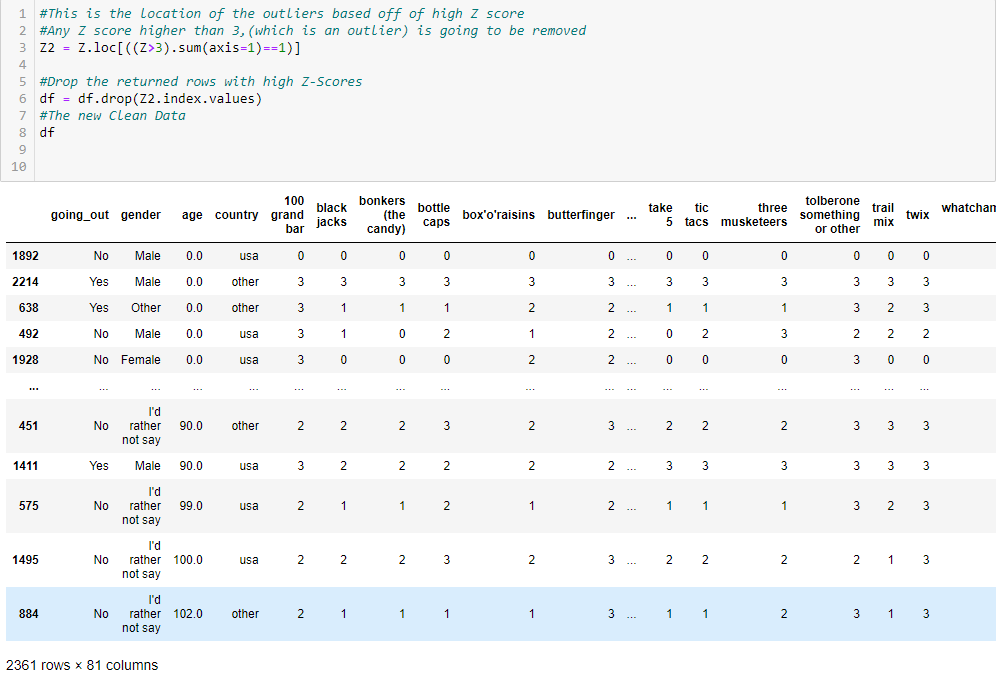
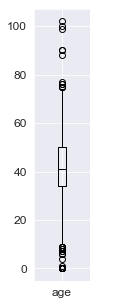
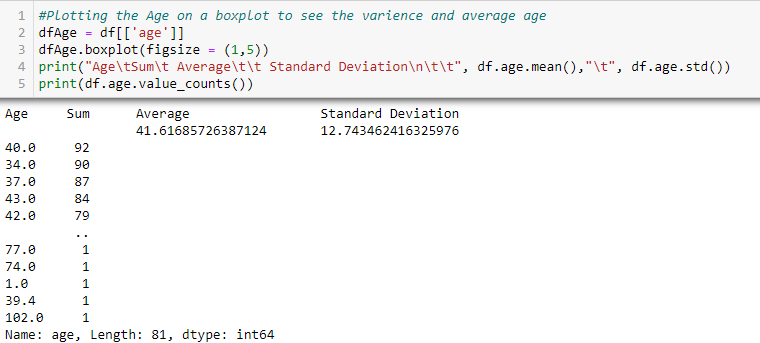


Other values that needed to be replaced included the inconsistently formatted answers. The biggest occurrence of this in the data set was with the country question. Most participants would answer “usa”, but many others would say “America” or something else that meant the same country. We took a list of every unique answer for that attribute and made the answers consistent.



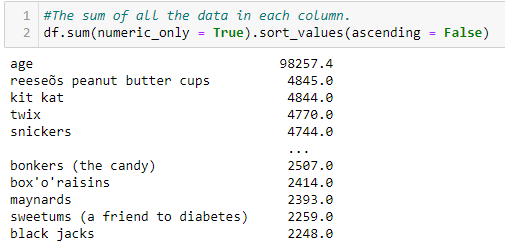
The final values we decided to replace were the answers for the candies. We thought it would be easier to analyze the data if the answers were given numeric values as opposed to a string value. So for every NaN, “DESPAIR”, “MEH”, and “JOY” value that occurred in the data set, they were replaced with 0,1,2, and 3 respectively. 

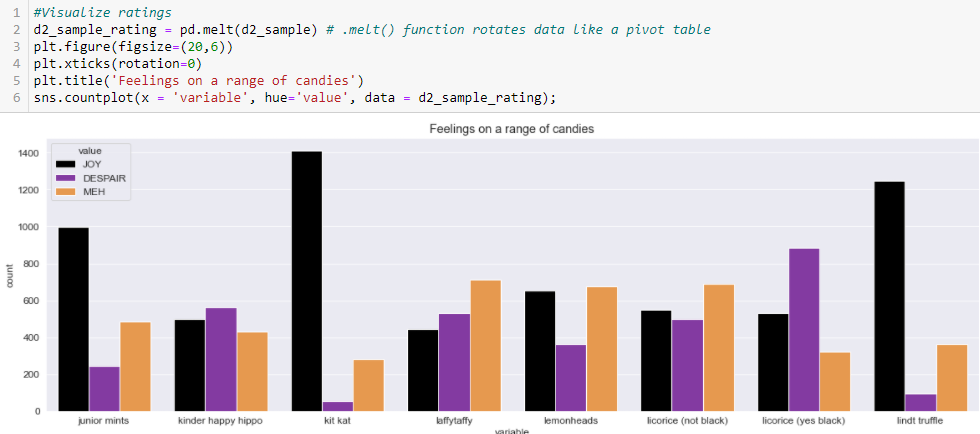
**Calculating the Z Score and Removing Outliers**

Now that data is formatted properly, we were ready to take the Z-Score to find any remaining outliers in the data.After calculating the Z-score for every row and column, we needed to drop any outliers that were present. We organized this dataframe to display age from lowest to highest, and we can see two outliers on the bottom. After selecting all the rows with a Z-score higher than 3, we then remove those rows from the dataframe.After running this code, we can see that only two rows were removed. These outliers were participants joking about their age and saying absurdly high numbers. We then boxplot the new data frame and analyze the results for the ages of participants.

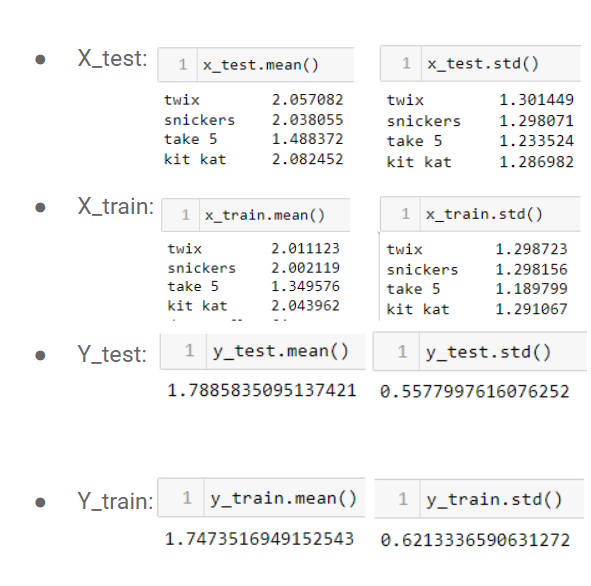
Here we can see that the average age of the participants were around 41 years old with a standard deviation of about 13 years.

**Displaying the Data:**

Now that the data is clean, we are able to take the data and analyze it in any way we see valuable. We first displayed the sum of the values of the data frame to see which candy was the most popular (reesees).

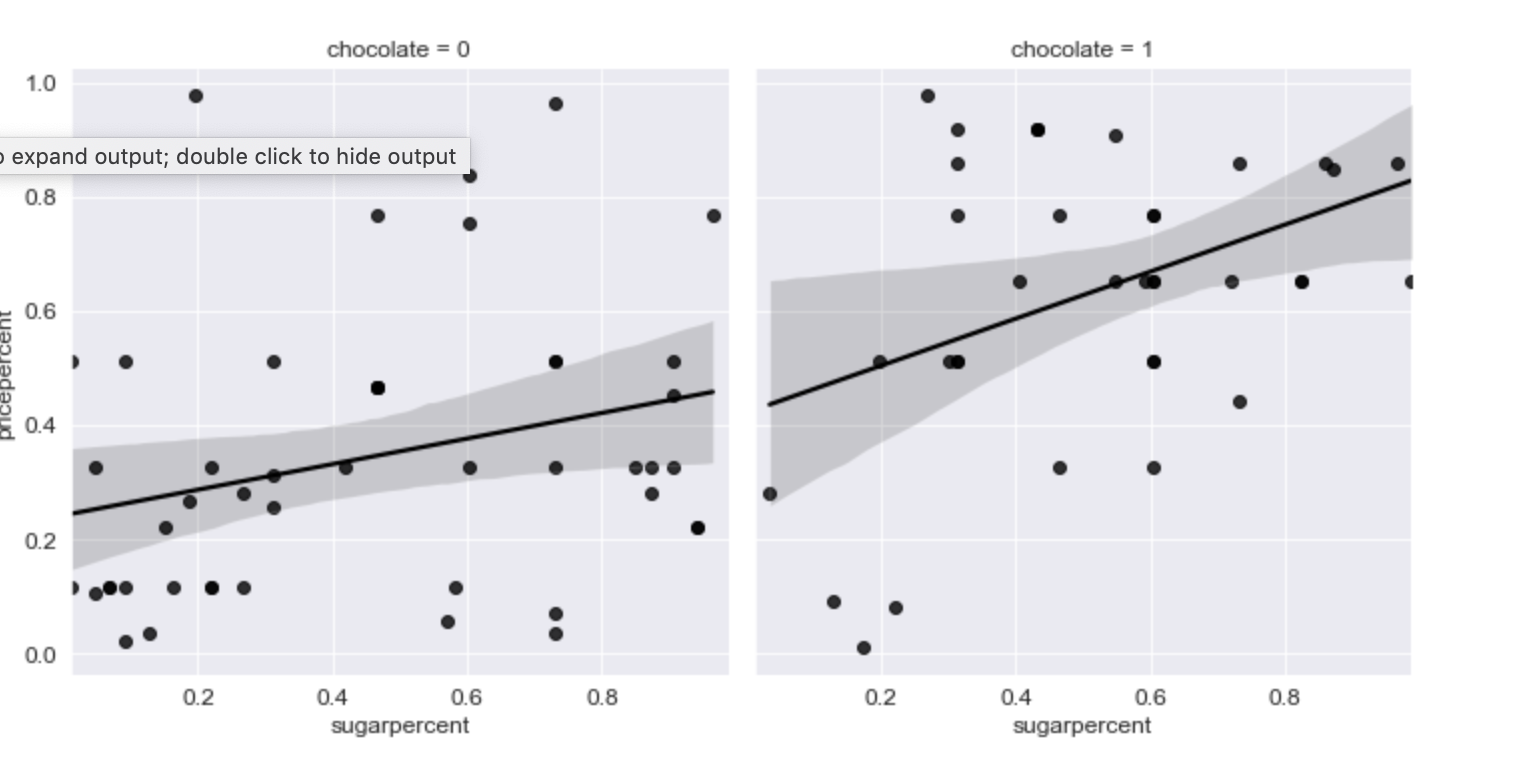
Then we decided to display information via graphs to be able to interpret the information a bit easier. The graphs that we used were for displaying information about different aspects of the cleaned data such as age, feelings of various candies, and gender to get a better understanding of our data.

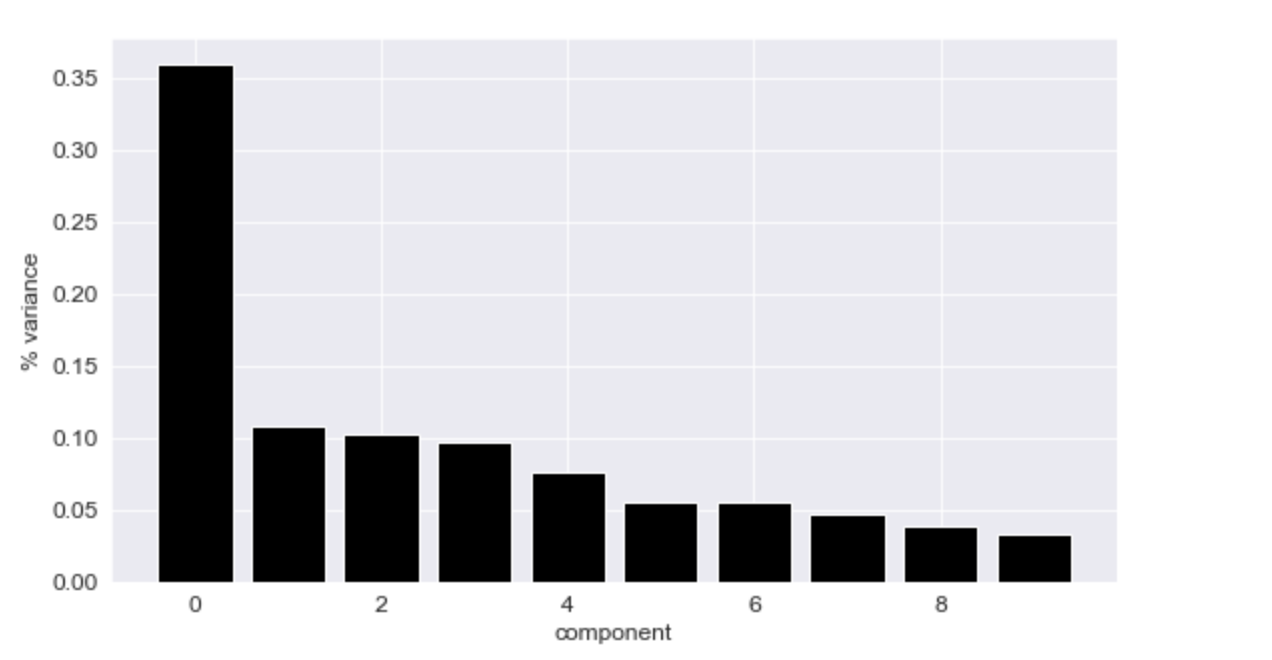
**Splitting the Data:**

When splitting the data, we chose the 80/20 split that was outlined in the Tutorial 4 file. We chose to encode the column “country” into a new column called “Country Joy”. This new column is going to be used for our Y test and training sets. For the purpose of the exercise, we chose four random candies: twix, snickers, take 5, and kit kats for the X training and testing sets. After splitting the data set we found the training/test mean and standard deviation of each candy shown in the screenshot above. In the x axis the train and test mean and standard deviation come close to each other. However in y axis the train and test mean are close to each other but the standard deviation came out to be different from each other. 

**PCA performed:**

Through the Pre\_Processing of our data, the number of rows and columns were reduced so that it became cleaner. We got rid of NaN data, unnecessary fields, duplicate data, outliers and sorted the data in such a way where the format was easily readable. Due to the removal of junk data the result is much quicker and more accurate.





**Conclusion:**

In Pre-Processing dataset it is crucial to clean up the data set at the beginning so the data is easily readable. We cleaned up the data set in a certain way so we can use the data in the next couple of projects. We split the data into training data and test data from which we calculated the mean and standard deviation. The data is sorted in a certain way, so that there is no longer any missing data or outliers that will affect the data set. It took us a long time to clean up and separate the data but now it is ready to proceed in the process of data mining.