Final Project - Black Friday Sales

Group Members: Mudit Kumar, Ahmed Khan, Vani Jaiswal

**Overview:**

Black Friday is one of the biggest sales events in America. Retail stores should be able to plan in advance what products sell the best. Knowing what items are needed will help boost sales, while minimizing cost of stocking unnecessary products. Trying to predict the amount of purchase based on other factors using various statistical methods could prove very useful.

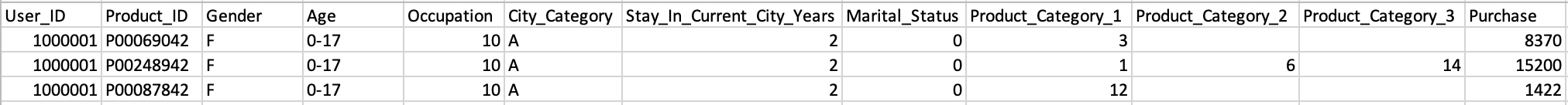
**Problem Selection:**

The problem that we decided to solve is to possibly predict the sales of products on black friday and understand customer purchases made upon different products. Our solution is to use a regression model where we try to predict the sales of products with the help of other data variables.

**Data Collection:**

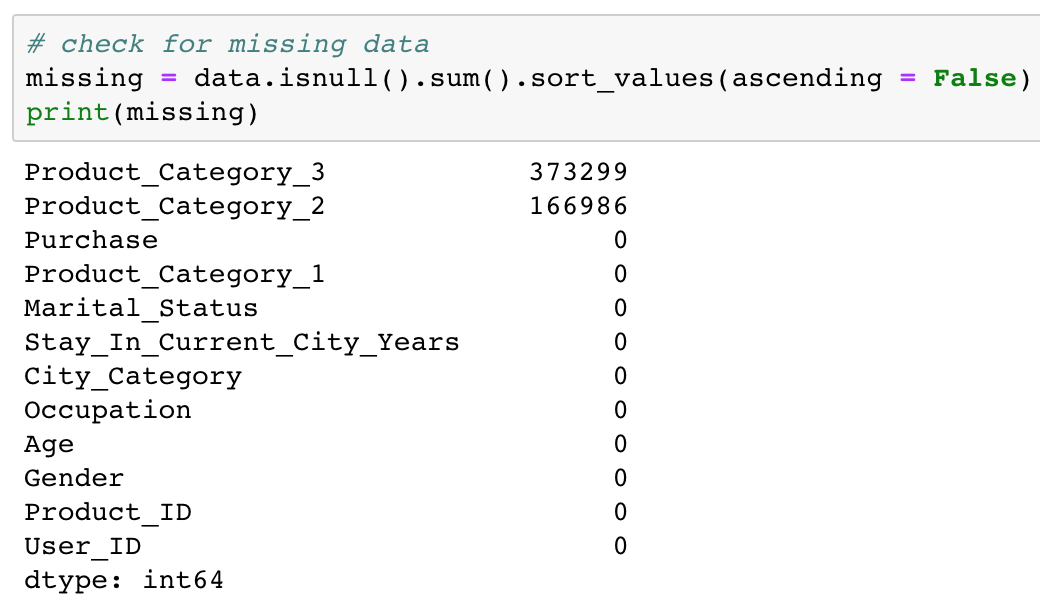
For this problem, we chose the black-friday dataset available on the Kaggle page. The data is a sample of transactions made in a store during Black Friday with 537,577 observations containing 12 features. The data can be found at [this](https://www.kaggle.com/mehdidag/black-friday) URL.

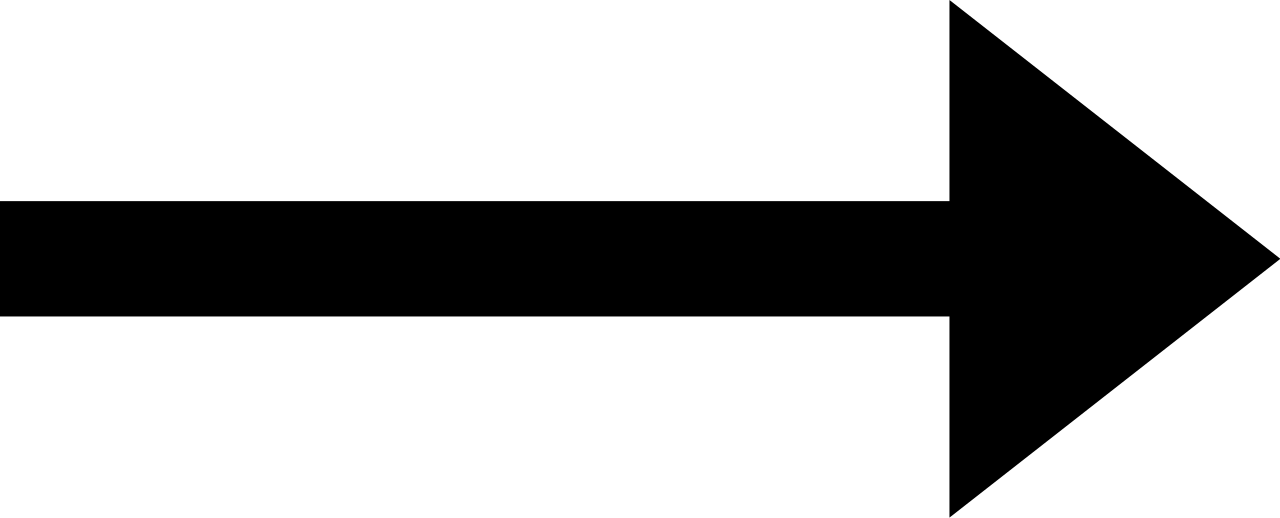
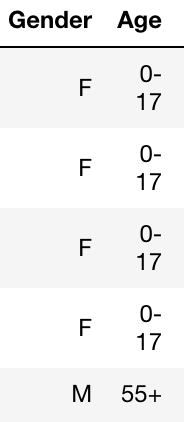
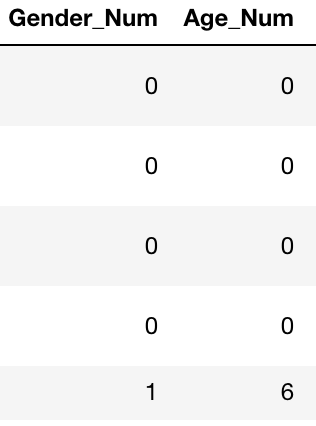
In order to better understand the data set being used, here are the first 3 rows:

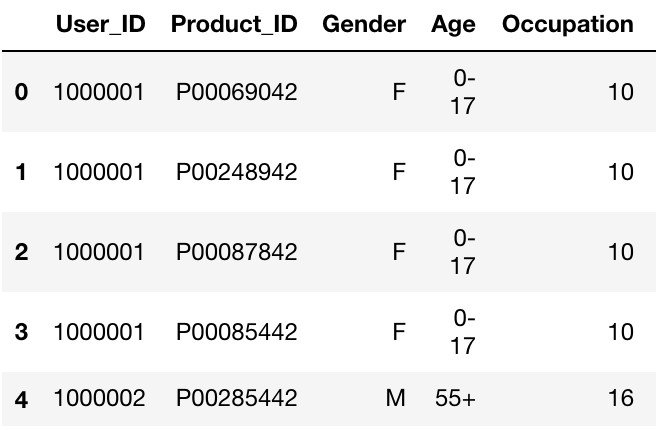
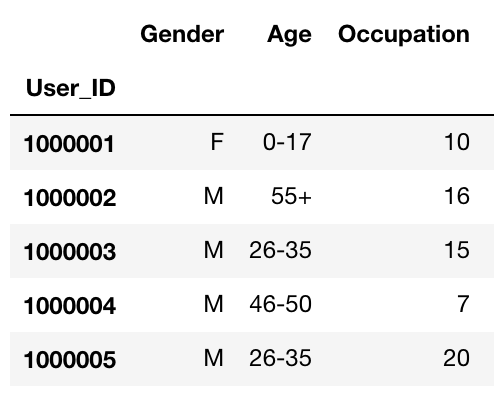


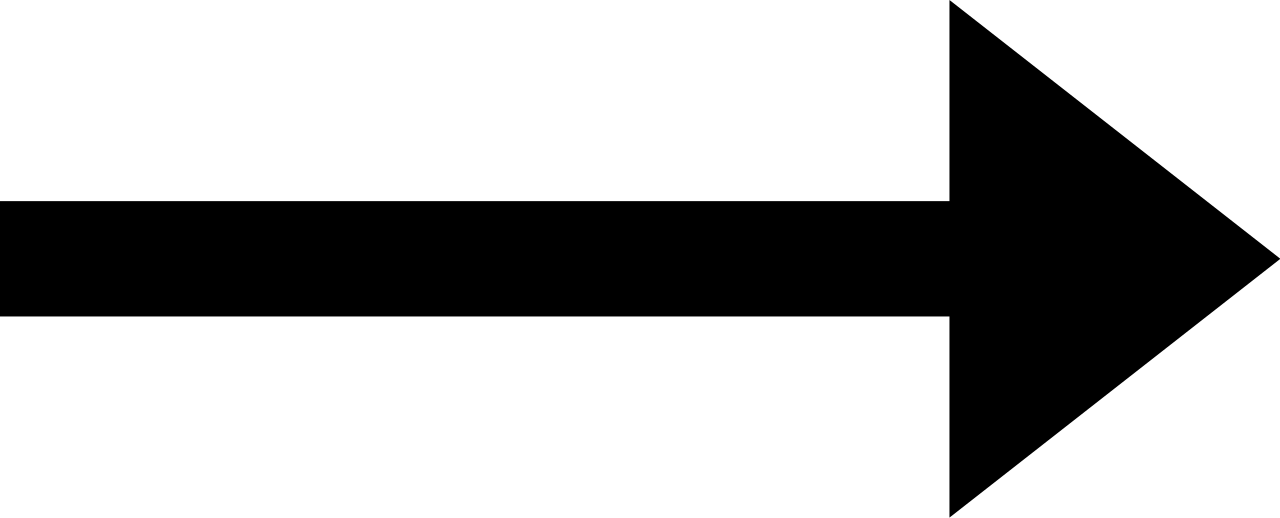
**Data Preparation:**

We initially cleaned the dataset by checking for missing values. What we found is that only two attributes contained missing data being ‘Product\_Category\_2’ and Product\_Category\_3’. Seeing that these attributes are categorical data that describe the product, we believe it is safe to replace these with 0 since it being empty implies that those categories don't describe the product.



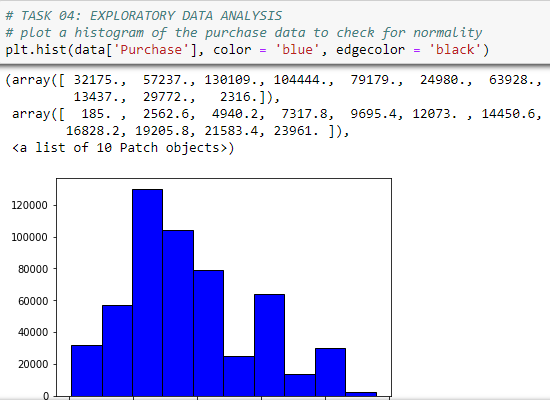
After exploring the dataset further, we noticed that a lot of the data was not in numerical form. Converting the data to it’s numerical representation is important in order to satisfy the requirements for simple and multiple regression. We proceeded by creating new columns of each necessary attribute to hold the numerical representation. Example:

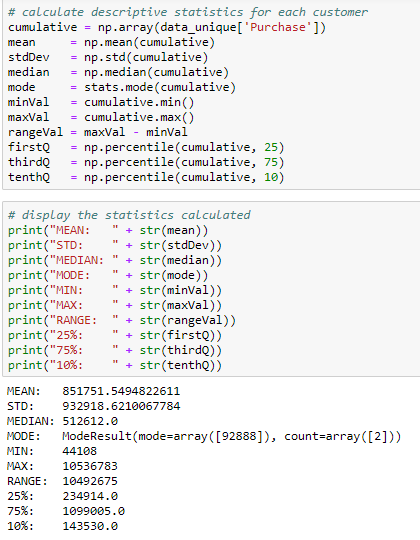
Finally, we noticed that the same customer was occupying multiple rows, which makes sense since each customer can have multiple transactions. However, this can misrepresent the data when analyzing the number of customers in the dataset. Therefore, we created a copy of the data by having each ID in the new table be unique, while accumulating the total purchase amount in dollars. In order to do this we had to drop some columns for the new table to make logical sense, but rest assured the original data set was not altered.



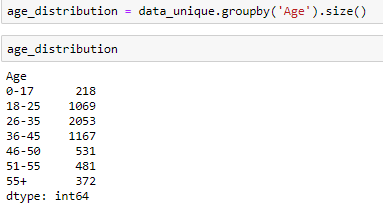
**Exploratory Data Analysis:**

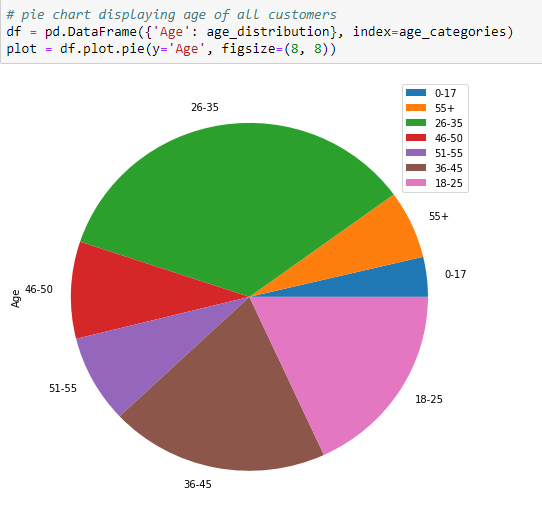
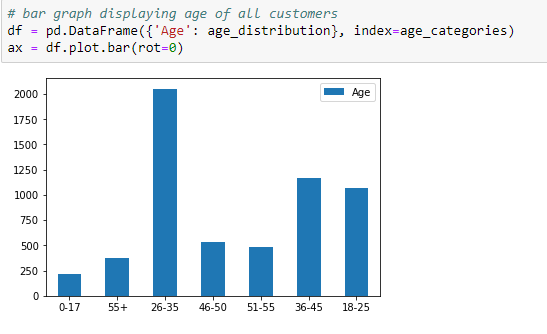
To explore the dataset, we performed various statistics on multiple variables and plotted their data. First, we plotted a histogram of the purchase data to check for normality and then calculated descriptive statistics for each customer. We found the histogram to be slightly skewed to the right.



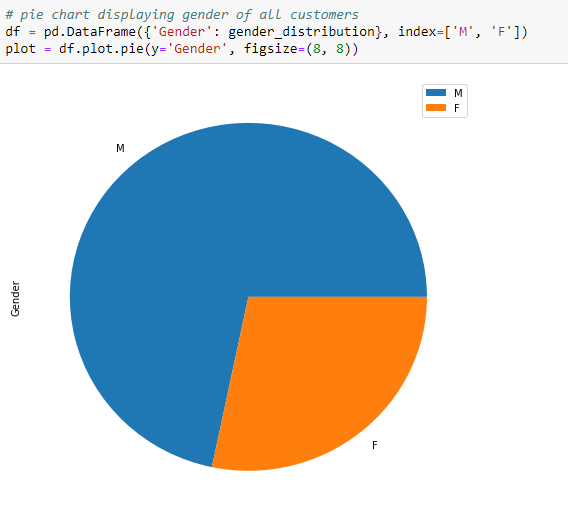


The maximum total of any one customer purchased was $10,536,783 where as the minimum total of any one customer purchased was $44,108. The mean is within one standard deviation of the median, which seems to imply normality amongst customers. Next, we figured out the categories for age which were 0-17, 26-35, 36 - 45, 46-50, 51-55, and 55+. The age distribution of the data samples are show below. We have also plotted the data using a pie chart and bar chart.

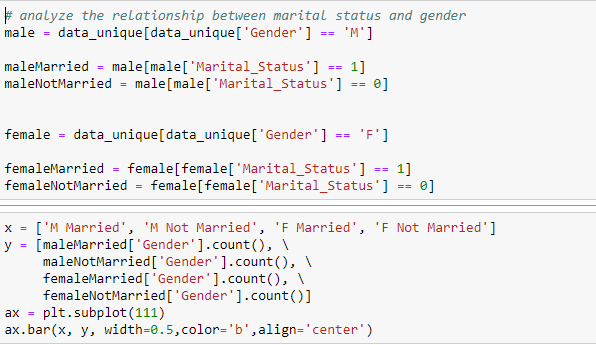


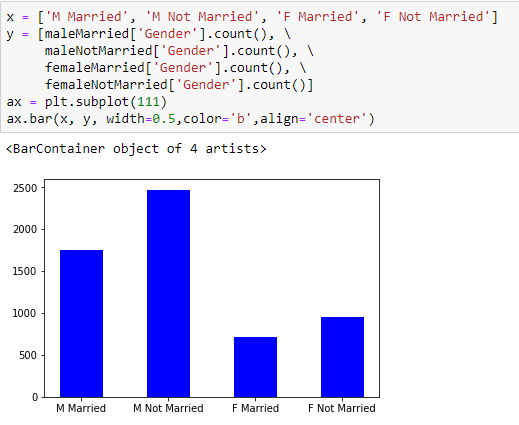


According to the data, the highest demographic age of customers at this particular retail store is the 26-35 years range, with 2,053. The lowest demographic age of customers at this particular retail store is the 0-17 years range, with 218.

Next, we looked at the gender distribution. 

The data shows that the majority of the retail stores customers are men. Then we analyzed the relationship between marital status and gender. We created variables that had value of 0 to be an unmarried males/females and variables that had a value of 1 were not married males/females.



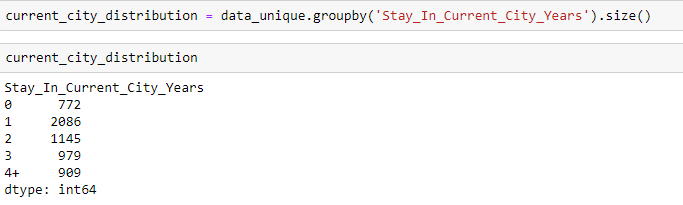


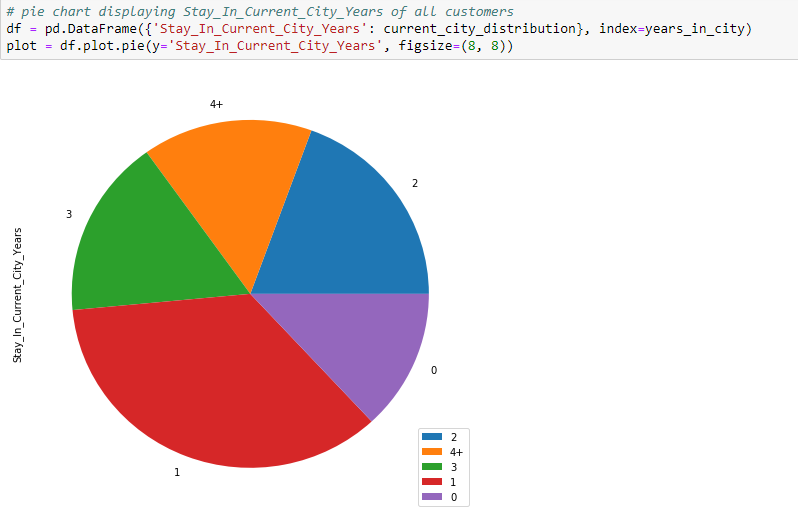
From this data, we concluded that non-married customers of each gender frequently visit the retail store more than married customers of each gender.

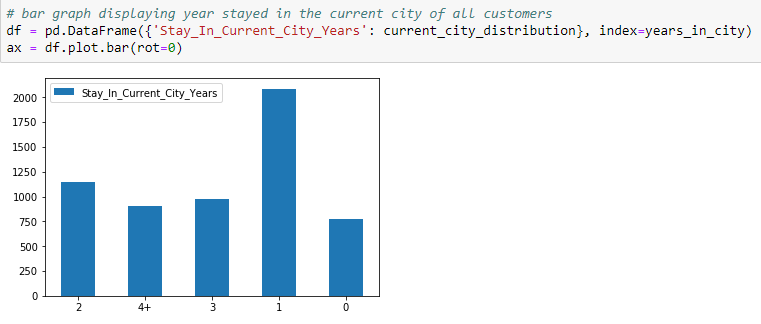
Afterwards, we grouped each product category sales compared to the gender of the customer.

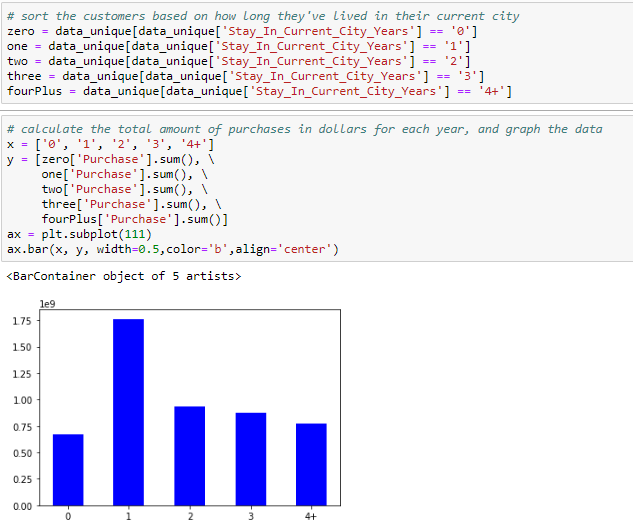


We can conclude from these respective categories that men buy more of each product than women do. The next variable we analyzed is the ‘Stay\_In\_Current\_City\_Years’. We plotted its distribution below.







From this data, we can conclude the highest number of customers have lived in their current city for 1 year, with 2086 where as the lowest number of customers have lived in their current city for less than a year, with 772. We then sorted the customers based on how long they've lived in their current city and calculated the total amount of purchases in dollars for each year, and graph the data. 

The data shows that residents who have lived in his/her current city for only a year spent the most amount of money. After living in the city for a year, residents have spent less and less each year. Residents who have lived in his/her city for less than a year spent the least amount of money.

We then sorted the customers based on what city category they live in and calculated the total amount of residents who live in each city category.



The data shows the most amount of customers live in city category C where as the least amount of customers live in city category A. Lastly, we calculated the total amount of sales of each product category.



The data suggests the product category that sells the most amount of units is product category 2 and the product category that sells the least amount of units is product category 3.

**Data Modelling:**

We did regression and classification on the dataset.For regression,we did linear and multiple regression with different parameters.We considered 'Gender\_Num','Age\_Num','Occupation','City\_Category\_Num', 'Stay\_In\_Years\_Num','Marital\_Status','Product\_Category\_1', 'Product\_Category\_2', 'Product\_Category\_3' for training and data['Purchase'] for testing.The highest R value is 0.12634023338141998.

For Classification we used 3 different model,Logistic regression,Decision tree and naive bayes and obtained results for different product on the basis of age and obtained same results for all the classifiers i.e 27 %,31 % and 67%.