The Black Friday Sales Project

Ibe Chukwudi Joshua, Ly Vi, Olichwier Shawn

December 9, 2018

Introduction:

The dataset that consists of variables that were recorded during the black friday sales event. This statistical methods to compare Black Friday sales. We will use methods such as Bootstrapping, permutation and ANOVA testing. With the use of these statistical tests to acertain various pararmeters; the parameter come in the form of but not only, the difference of means, tur value of average purchases made by various sub sets of the data set. Such subsets and their respective pararmeters could be the purchases based on the gender of the purchases. With that breakdown we can gain the sample mean of males to female, and then get the true difference between male and female shoppers. The purchases of different age groups are also observable from the dataset. Since marital status is also a variable in the Black friday observations of how marriage affects purchases on black friday would be looked into as well

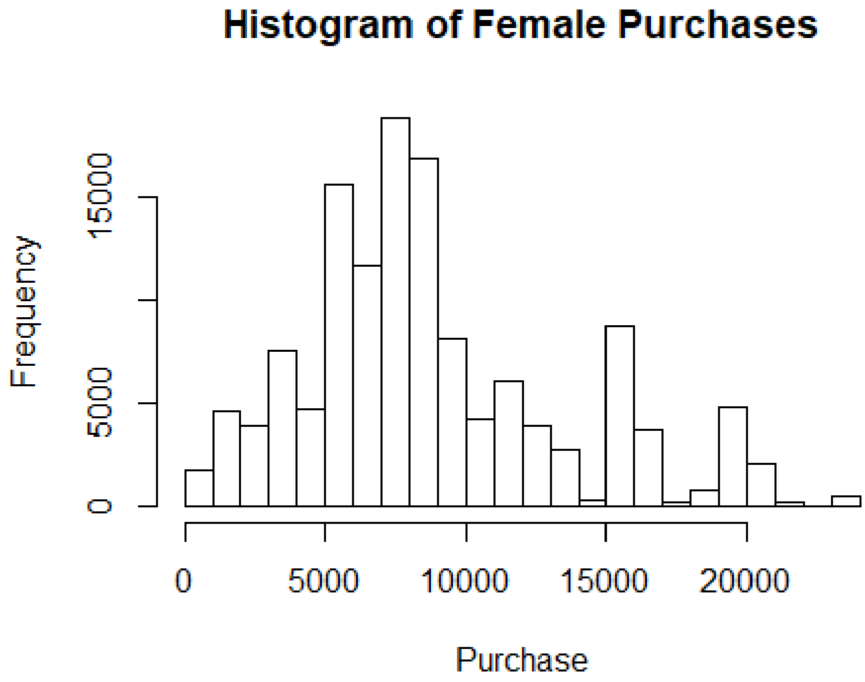
The data was retrieved from the website Kaggle and consists of half a million different observations of purchases. The data set has many variables like age, gender, occupation, and marital status, to name a few. We used only the few relevant variables for our research but a lot more information could be garnered if need be.

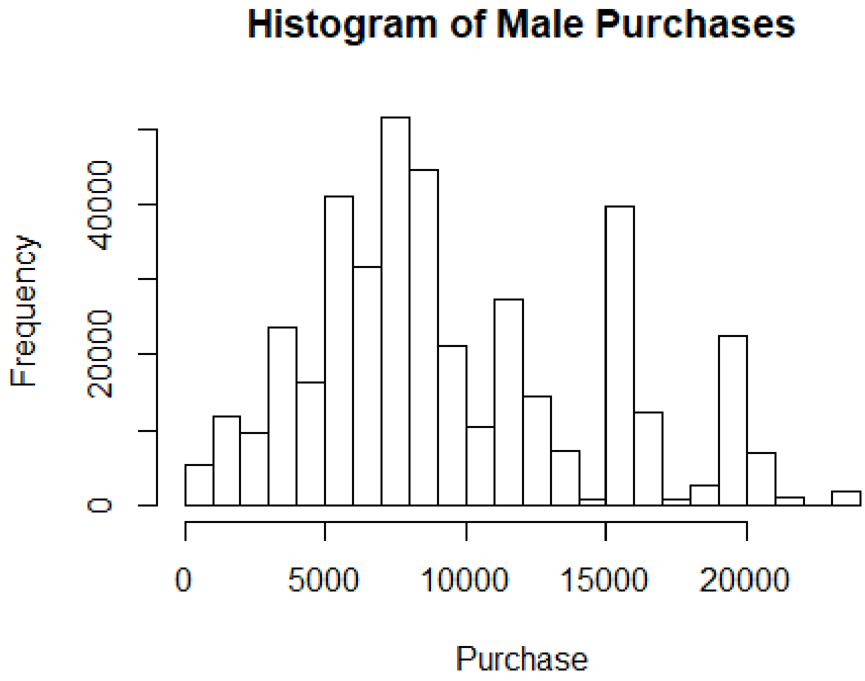
The use of statistics for our research is crucial. Are the values that we find in this data significant? Or are these values normal observations for Black Friday sales? Our statistical methods give us a basis for answering these questions. Sales data is increasingly important in our American life. Living in a society expects maximizationn of profits and resources, the constant ebb and flow of money is crucial to our economic growth. Studying datasets such as these can show strengths and weaknesses in our economic life. Are there demographics in which we can grow sales? Are there groups of individuals who are bedrocks for sales every year? Who is doing the most buying and contributing most to the growth of sales on Black Friday and throughout the holidays? These questions and others can be answered via careful consideration and our statistical inquiry.

The first step in any data analysis is to inspect your data. Looking at the entire dataset, while interesting, may not be very insightful. I will pick out relevant rows for my analysis.

```
##   Gender Purchase
## 1      F     8370
## 2      F    15200
## 3      F     1422
## 4      F     1057
## 5      M     7969
## 6      M    15227
```

This shows the gender of the person who made the purchase and the amount of their purchase. Perfect. Looking at the histogram of the our data can give us some insight as well.





The observation is that the data is bimodal, and asymmetric. We have some individuals spending a lot of money which is skewing our data. Usually we would run a qqplot on our data to be sure it is normal or not, but in our case, we can easily see it is not normal.

Let's state the test statistic of our bootstrap testing. The test statistic is the difference between the means of male and female shoppers. We set up the hypothesis as;

$$H_0: \mu_{men} = \mu_{women}$$

$$H_A: \mu_{men} \neq \mu_{women}$$

Below is our observed mean for both and our test statistic value.

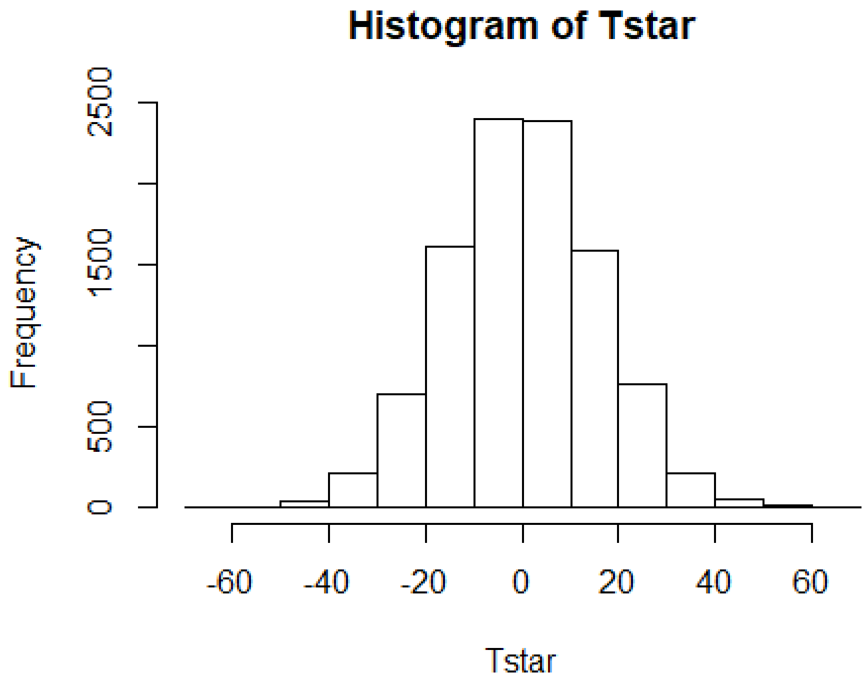
```
## [1] "This is the observed mean for the female shoppers."

## [1] 8809.761

## [1] "and the observed mean for our male shoppers."

## [1] 9504.772

## [1] "The observed difference of these means is"

## [1] 695.0104
```

Now the natural question that we should be asking is if our observed difference is statistically significant? Our observed value says that males spend more on average by about $700.00. Could this amount happen by random chance or do they spend a significant amount more than their female counterparts?

The only way to find out for sure is to run some tests. The first one that our research will test is bootstrap testing. This testing permutes the data with random organization to come up with many 'different' data sets. These will represent various other possible outcomes that the data could have taken. If our data is unique, then our value of the observed mean will stand out and be statistically relevant.

Let us begin our analysis. The great thing about bootstrap testing is that our data does not need to be normal!

After bootstrapping, the distribution of our bootstrap difference of means look like this.

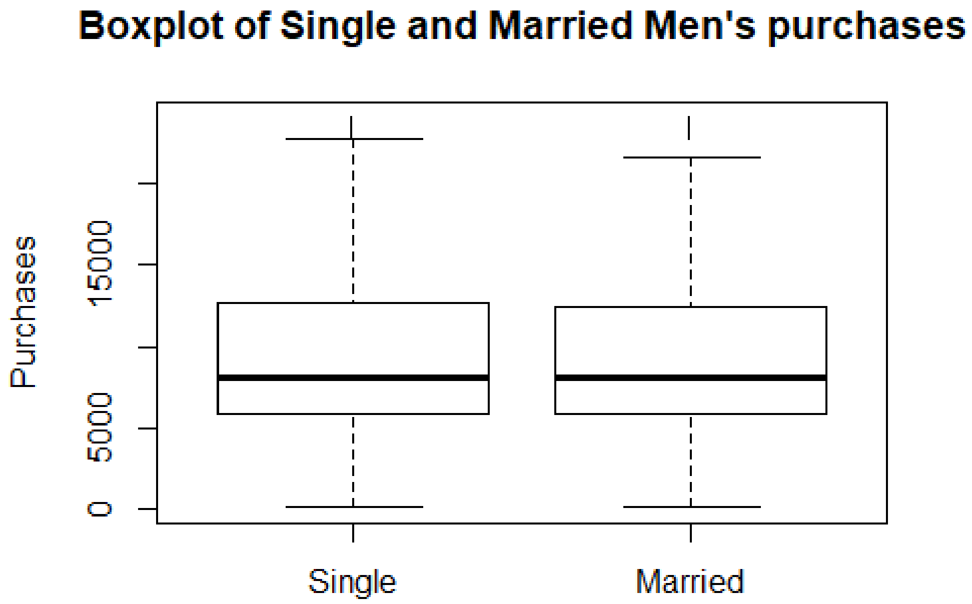


```
## [1] "Our p value is"

## [1] 9.999e-05
```

Our observed difference means was roughly $700.00. We can see here after bootstrapping the data that our value is nowhere close to being on this histogram. Our bootstrap distribution pvalue is 9.99^e-05, which is very close to zero. Even if we were to use a 99% confidence interval, we would reject the null hypothesis, which is that our observed values are due to random chance. Therefore, we can conclude from our bootstrap that there is significant evidence that men spend more than women on Black Friday.

In finding out that men on average spend more money on black friday sales than women it is worth trying to find out what type of men spend more or less, or is it just men in general.

Fortunately for us the data contains the marital status of thew men.



We get a

$$\mu_{single}$$

equal to 9518.54 dollars and

$$\mu_{married}$$

equal to 9484.62 dollars. This discrepency of $(9518.54-9484.62) = 33.92 dollars might be due to sample differences or sample collection we assume that single men spend on average the same. We are saying that

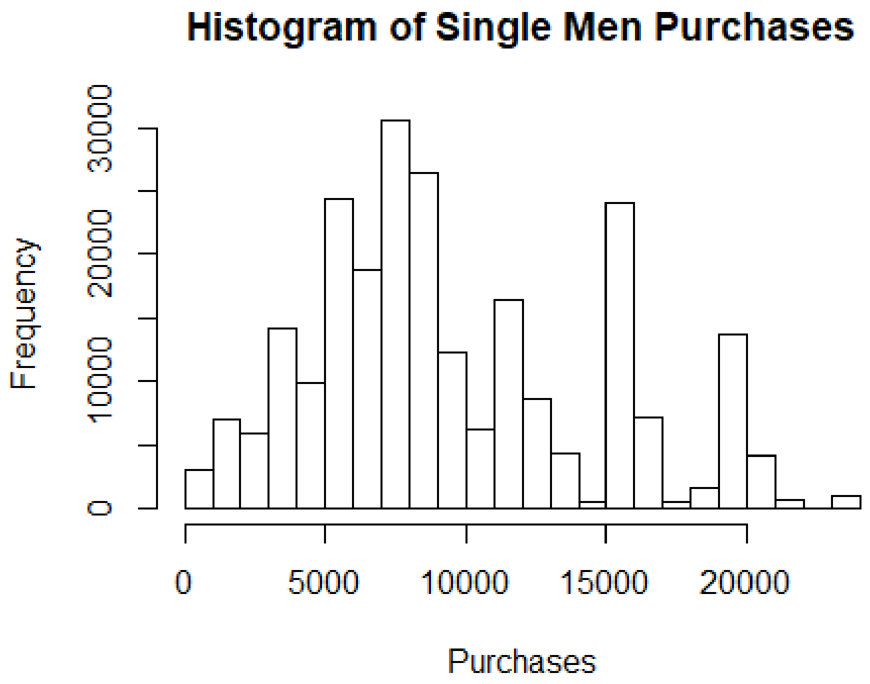
$$H_A: \mu_{single} - \mu_{married} = 0$$

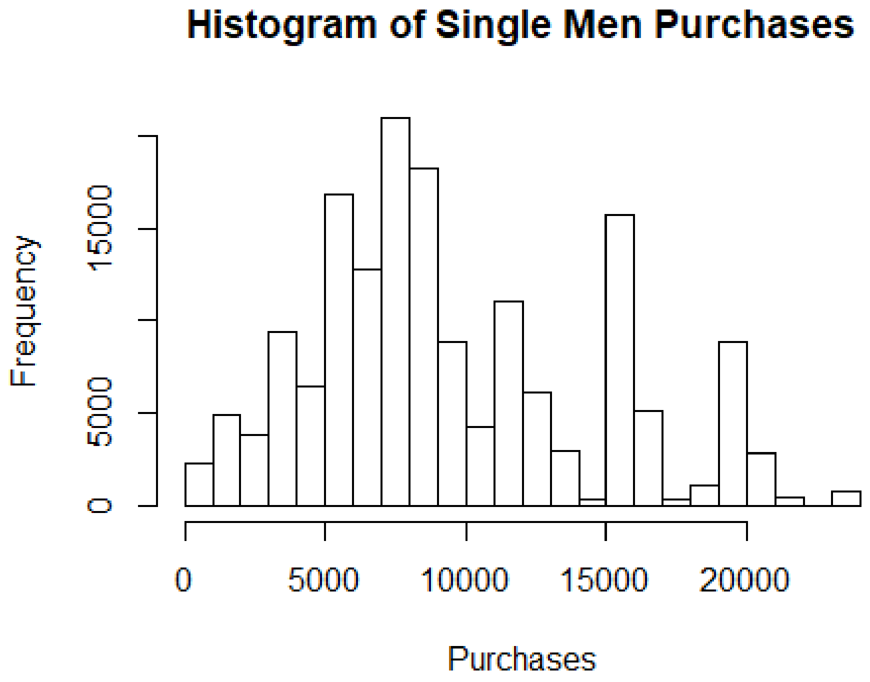
$$H_A: \mu_{single} > \mu_{married} = 0$$

Our test statistic is

$$\theta = \mu_{single} - \mu_{married}$$

We can observe what the distributions look like to know what would be an appropriate test to find out about our hypothesis.

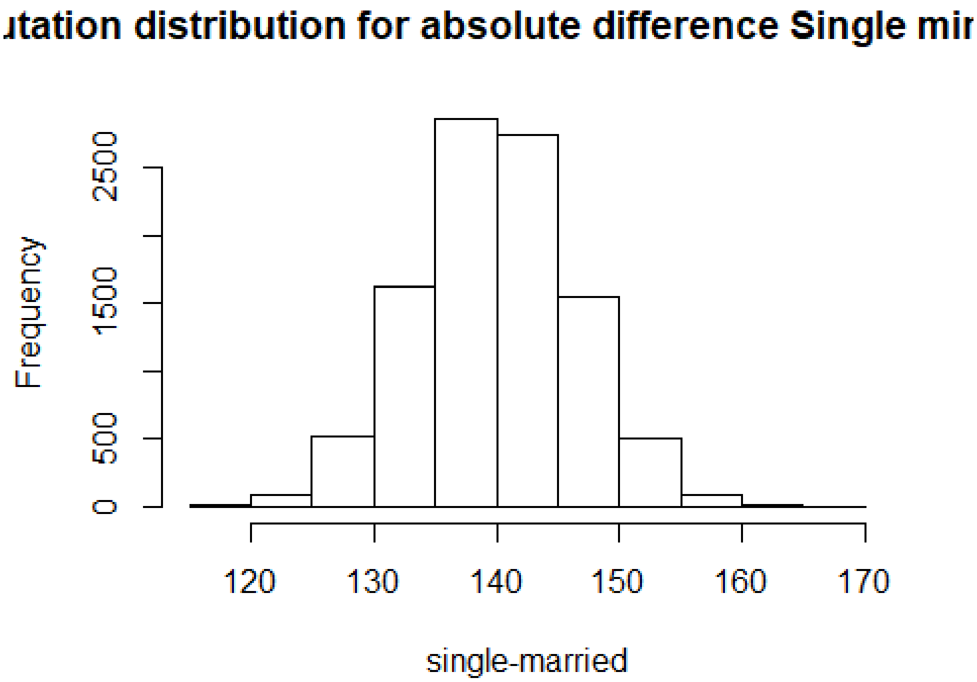




Both disttributions are bimodal, and asymmetric which means we can not use a t test. A permutation test might be more effective in finding out out test parameter of

$$\theta$$

After the permutation test we get a ditribution and results as such



```
## [1] "The p-value is"

## [1] 1e-04
```

The p-value optained was equal to 0.0001 which means that the mean difference will be less than observed amount of 33.92 is a 1 in 10000. Even with a test of 0.01 this would still mean that we have to reject that the mean purchases of single men to marrried men is greater.

To find out if different age groups affecting the Black Friday sales, we will first look at the two specific columns “Age” and “Purchase” from the data set.

```
##     Age Purchase
## 1  0-17     8370
## 2  0-17    15200
## 3  0-17     1422
## 4  0-17     1057
## 5   55+     7969
## 6 26-35    15227
```

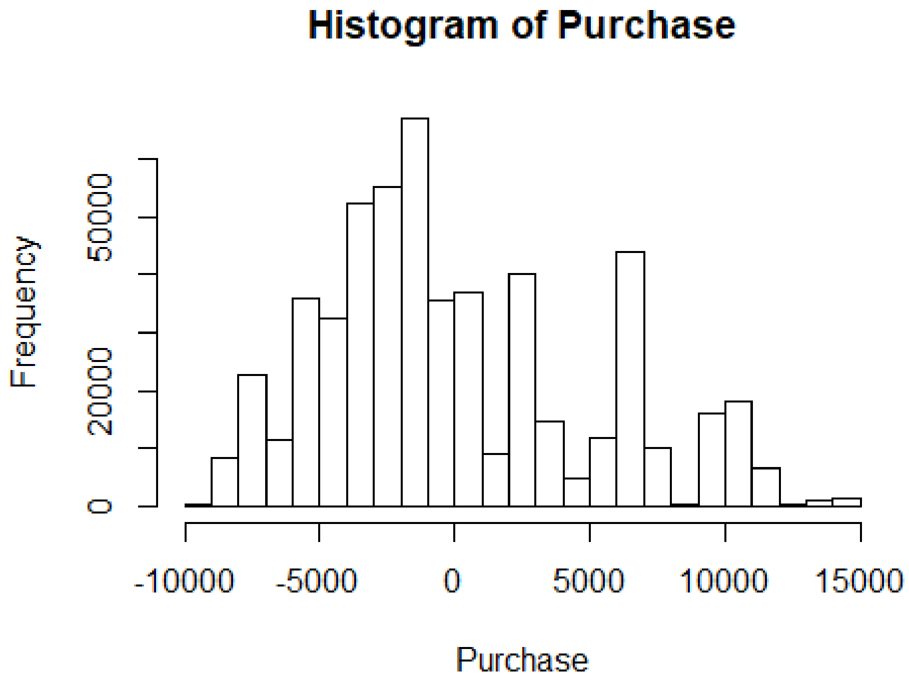
As we can see, we have many different age groups in the “Age” column. Those age groups and their total sale are listed below:

```
##     Age   Purchase
## 1  0-17  132659006
## 2 18-25  901669280
## 3 26-35 1999749106
## 4 36-45 1010649565
## 5 46-50  413418223
## 6 51-55  361908356
## 7   55+  197614842
```

Because there are more than 2 age group, we would perform the Anova test. Our NULL hypothesis (H0) would be the means from different age group are all equal. The althernative hypothesis would be all the means are different from each other.

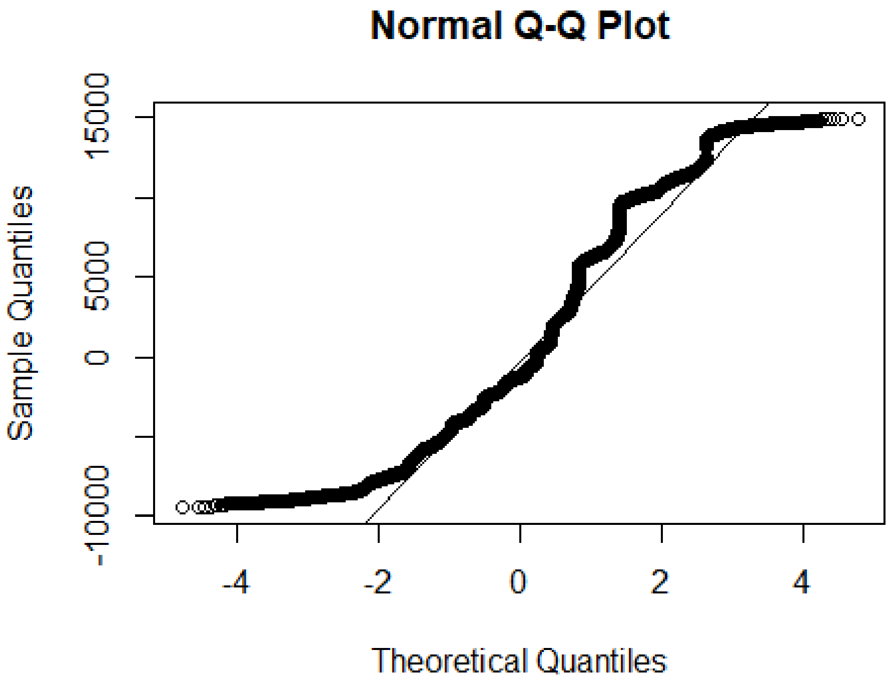
```
## Analysis of Variance Table
##
## Response: Purchase
##                Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age             6 6.4706e+09 1078430849  43.487 < 2.2e-16 ***
## Residuals 537570 1.3331e+13   24798822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is very small, we would reject the NULL hypothesis. However, we are not very confident with this results because we didn't check if the data is normally distributed when we first do the Anova test. Therefore, we now create a exploratory plot to ensure that our result is accurate. First, we created a histogram of the Purchase.



The histogram is a

bimodal right skwed, but we can't conclude any thing yet. We then perform the qqnorm to



check:

The plot doesn't follow the normal line at all. Therefore, we can conclude that the data provided is not normally distributed. That means our p-value from the previous Anova test is not accurate. At this stage, we have to perform permutation Anova test to get an accurate p-value because Permutation test doesn't require normal distrubition on origianal data. We resample from the data 500,000 times (the number of purchases in Black Friday dataset) without replacement. Then we repeat this process of resampling 10,000 times. Our NULL hypothesis (H0) would be the means from different age group are all equal. The althernative hypothesis would be all the means are different from each other.

```
## [1] 1e-04
```

We set our observed t-statistic equal to the F-value of the original Anova tets. We calculate the new p-value by finding the probility of the sum of all the resampled F-values that is larger or equal to the observed F-value with 10,000 (the number of repetitions). Eventually, we got our p-value equal to The p-value is 1e-04 which is really small so we have to reject th null hypothesis. Therefore, we can confidently conclude that the age difference does effect the retailsales significantly.

In order to find out which age groups have the moesPairwise t-test.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Purchase and Age
##
```

```
##       0-17    18-25   26-35   36-45   46-50   51-55  
## 18-25 1.0e-05 -       -       -       -       -      
## 26-35 5.2e-11 0.00026 -       -       -       -      
## 36-45 < 2e-16 6.4e-13 2.7e-05 -       -       -      
## 46-50 2.5e-07 0.24329 0.32753 0.00026 -       -      
## 51-55 < 2e-16 < 2e-16 < 2e-16 2.9e-12 < 2e-16 -      
## 55+   9.3e-15 9.9e-08 0.00052 0.32753 0.00031 0.00052
## 
## P value adjustment method: holm
```

There is not big significant among 18-25, 36-45 and 46-50. These three groups have the most positive impacts on the purchase. These seems to be true since these age groups has good income source and tend to spend a lot on good deal. In the opposite,the 0-17 doesn't have good income and the 51-over 55 is skeptical on buying on Black Friday.

Conclusion

In the findings we have seen that on average men spend more during the black friday sales even than women, single men in particular. And we also see that the age groups between 18-25, 36-45 and 46-50 have the most impact on purchases in terms of age. Results of such findings can be useful in a myraid ways. All areas of blackfriday and wholesale can be better prepared with such findings. Production companies can use this information to allocate resources toward greater focus on the customser that purchase the most which in this case are single men between the ages of 18-45. Advertisements can also be used to target that demographic for pre black friday adverts. This and more just a few ways to use the information gathered though the anaylsis of the black friday datasetr.