Northeastern University
College of Professional Studies

Final Project Report

ALY6010 - Probability Theory and Introductory Statistics

Instructor: Prof. Amin Karimpour

Date: 11th December 2021

By,
Jeseeka Shah,
Sanjana Chaudhari &
Supreeth Murugesh

Introduction:

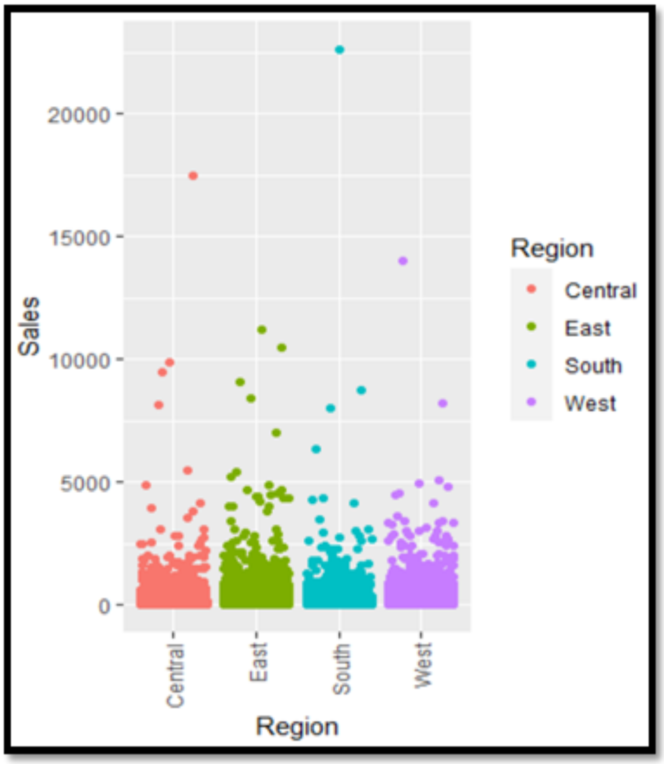
The superstore chain of stores which had 512,900 orders listed in the dataset at the start has reduced to 9994 orders as we are only considering the US market orders. The dataset contains the orders from the year 2011 to 2014. We have considered to keep 24 variables after carefully looking into the data. Additionally, we have created a new calculated variable total sale was obtained by product of quantity of sale and price of the item ordered.

Exploratory Data Analysis:

In the EDA performed our main aim was to answer few question that occurred while cleaning and selecting the data. The main questions that come to our mind was with respect to the profit earned by the supermarket in the US market: -

- Which category had the highest orders placed?
- What's the preferred shipping mode of Customers?
- What type of profit trends for different category are observed?
- What is the profit's earned per segment?
- How are the profits with respect to sales per order?

Let's investigate the questions one by one using visualization: -



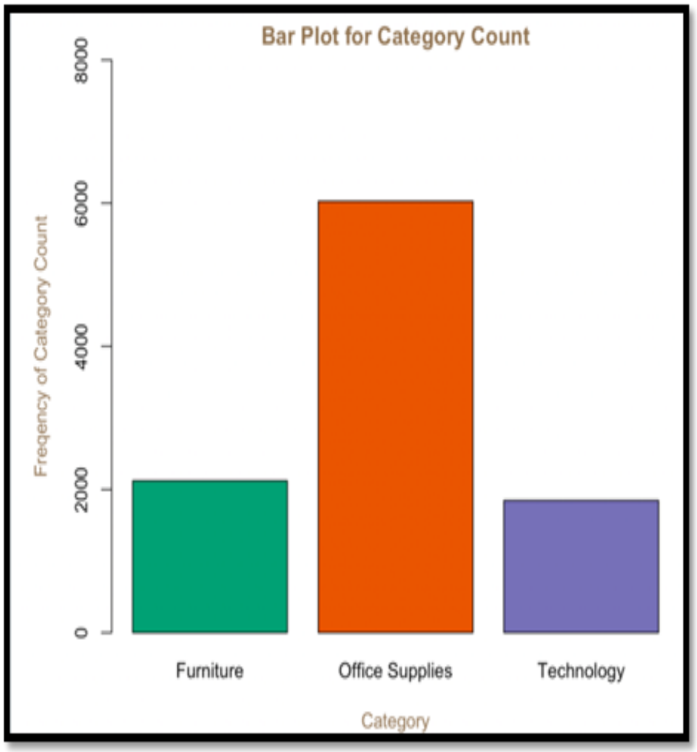
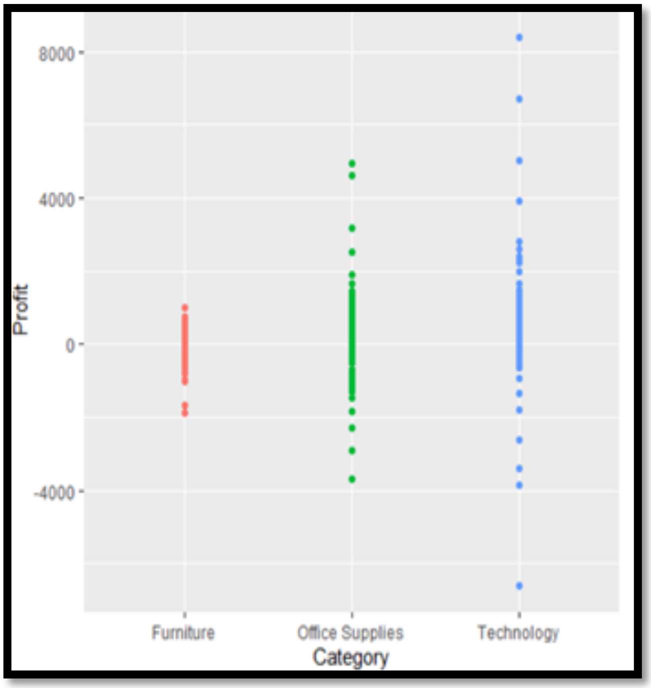


Figure 1: Sales achieved per region

Figure 2: Count of orders per Category

The figure 1 depicts sales per region achieved in the US market. The southern region of the US has the most sales followed up by central, west, and then east as per the max sales shown in the graph. And the figure 2 shows Superstore order based on categories has maximum count in office supplies which goes close to 6000 and the lowest count was in the category of technology which was less than 2000.



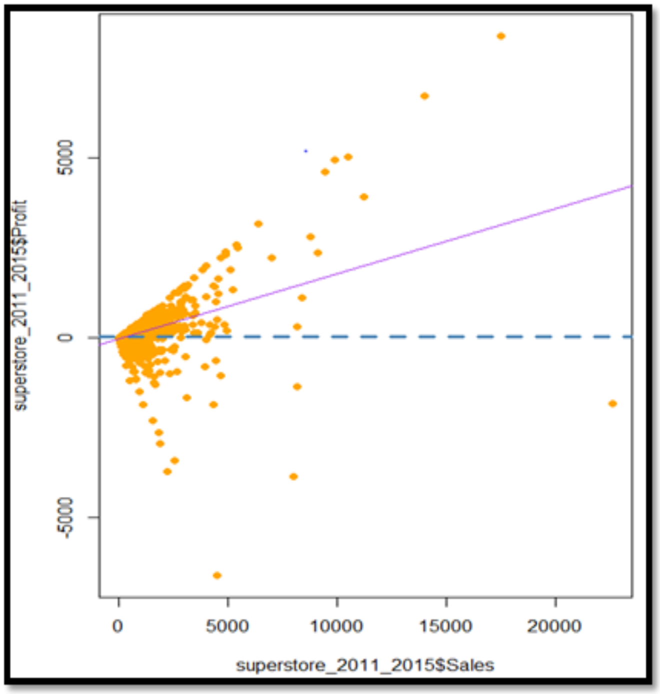


Figure 3: Profit per category

Figure 4: Profit per Sales

The Figure 3 shows the profit per categories. The category technology has the least and most profit achieved followed by office supplies and then furniture. And figure 4 shows the sales and profit relationship. The line in blues shows the mean values and the line in purple shows the linear positive relationship between profit and sales. The graph shows that sales are more than 20000 have incurred a loss. Whereas the sales < 20000 has incurred a profit roundabout to 10000.

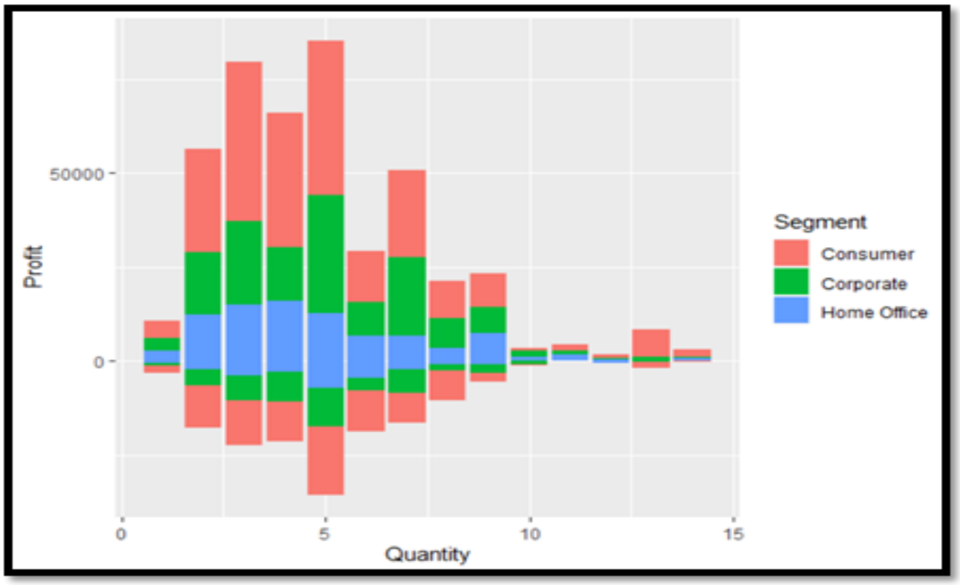


Figure 5: Profit earned per quantity based on segments

The Figure 5 depicts quantity and profit as per the segments. The maximum profit was achieved when the number of quantities of a product ordered was five. The most profit was earned by the consumer segment when order of quantity was 5 and less compared to orders more than five in quantity. The other segments such as corporate and home office followed similar trend.

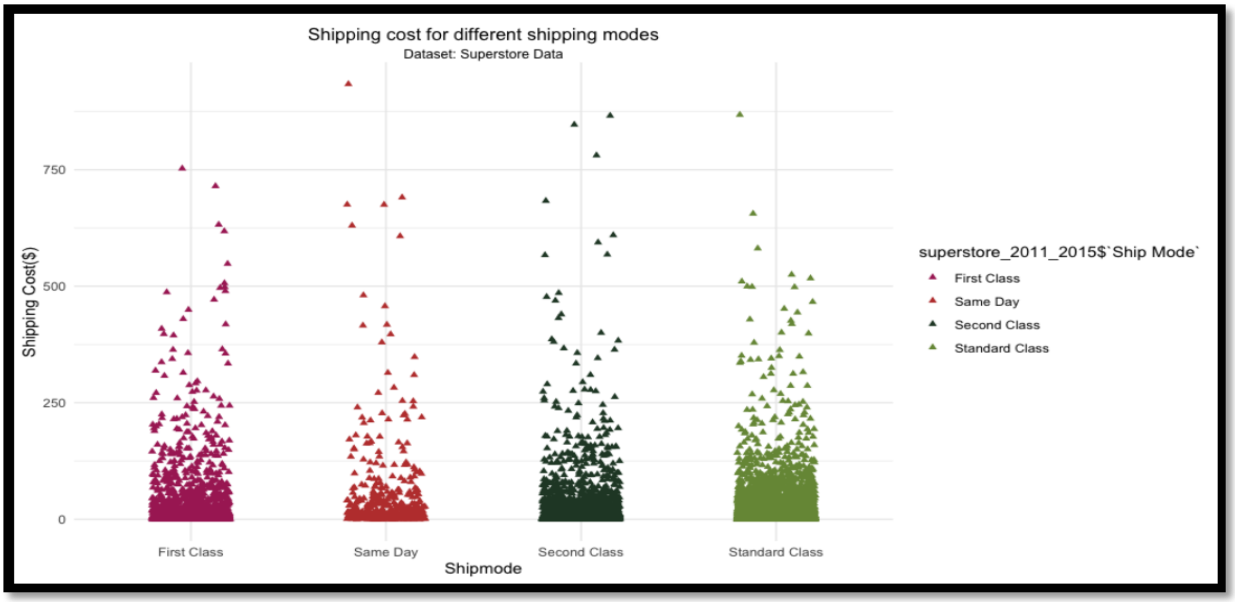


Figure 6: Cost for different shipping modes

The Figure 6 is a jitter plot of cost for different shipping modes. There are four types of shipping modes available for the orders. The Standard Class shipping mode chosen by 5968 which is around 60% of the orders. Shipping cost ranged from $0 to $800 for the standard class. In which,75% of the cost for Standard class ranged between $0 to $125. Comparing other shipping modes there were very few, 543 out of 9994 orders were marked to be delivered on the same day and the highest cost compared to other modes as well, was $825 for being delivery on the same day. Roundabout 35% of the orders were marked as First class and Second-class shipping modes.

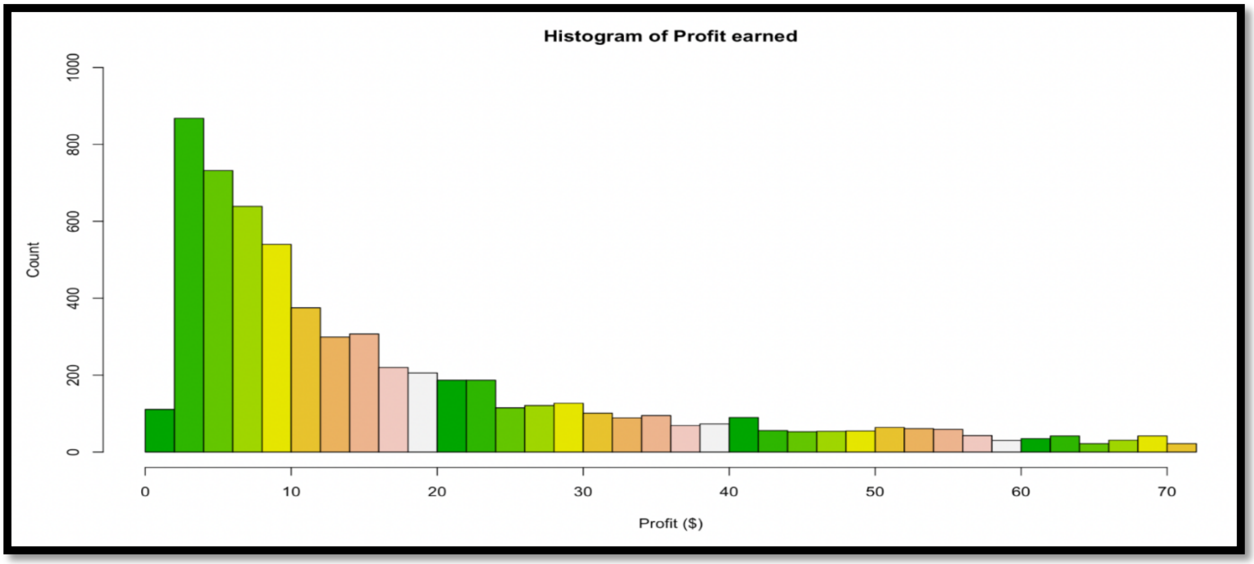


Figure 7: The profit earned trend

The orders made to the company had profits ranging from -$6599 which is points at loss and the maximum profit was $8399. The figure 7 shows histogram which is plotted after filtering the data by omitting the outliers, this are the orders with profits less than $1 and more than $70 which are outliers. Looking at the histogram, we can see that it is left skewed as maximum of the profit obtained is between $2 to $30. The number of ordered ranged around 2600 with profit ranging about $0 to $10.

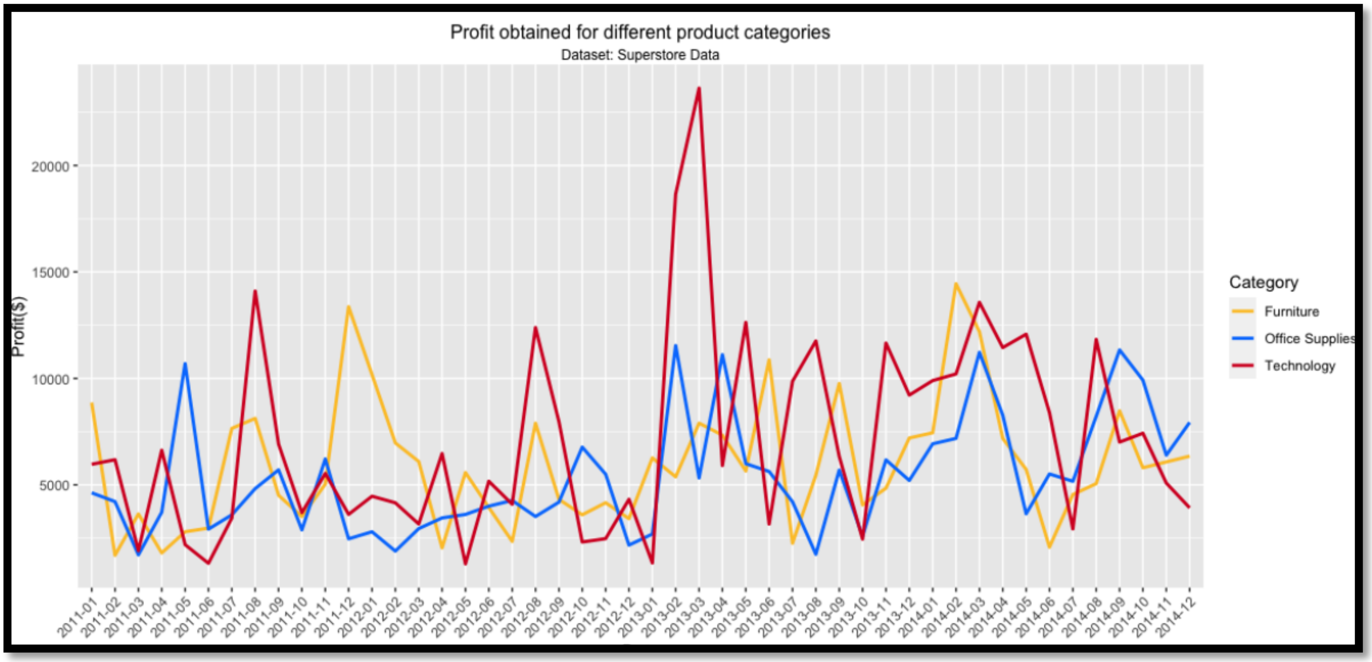


Figure 8: Profit obtained for the different product categories

The Figure 8 shows the time series analysis of profits gained for various category for a span of three years. The moderate looking technology category has always the earned the maximum profits for three years. In first quarter of 2013, Technology had made roof top reaching profit which is more than $20,000. However, the company made the least profit in the fourth quarter in time of 2012 where all 3 categories having no more than $5,000. Company should reconsider about discounts and offers during year end for Christmas or holiday season time.

Statistical analysis:

a. Hypothesis Testing:

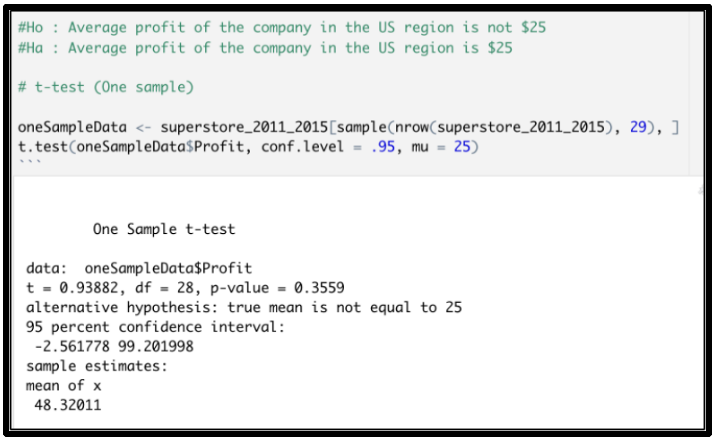
In this part of the report let us see some of the hypothesis that we constructed and the procedure we followed to obtain the results. Primarily our majority of the concentration is on the profit margin earned by the company, we are digging a step deeper into categories and the region of the US market where company saw some profits.

Looking at the histogram of the profits above we want to ask the question about the average profit earned by the company in the US market. So, our null and alternative hypothesis is as follows.

H_0: Average profit of the company in the US region is $25

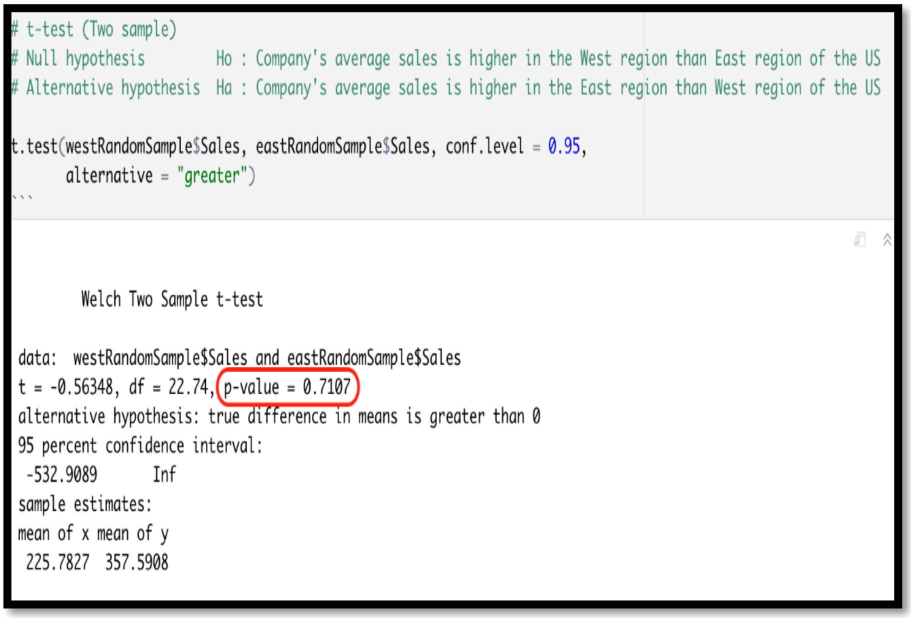
H_a: Average profit of the company in the US region is not $25

Based on the hypothesis constructed one sample t-test was performed by considering that company's profit to be 25 dollars in the US market where confidence interval is 95 percent with α = 0.05. We took a random sample of 29 from the dataset of 9,993 rows. We obtained a p-value of 0.6879 which is greater the 0.05. Hence we successfully reject the null hypothesis and accept the alternative hypothesis that Average profit of the company in the US region is not $25.



```
#Ho : Average profit of the company in the US region is not $25
#Ha : Average profit of the company in the US region is $25

# t-test (One sample)

oneSampleData <- superstore_2011_2015[sample(nrow(superstore_2011_2015), 29), ]
t.test(oneSampleData$Profit, conf.level = .95, mu = 25)
```
```

```

	One Sample t-test

data: oneSampleData$Profit
t = 0.93882, df = 28, p-value = 0.3559
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 -2.561778 99.201998
sample estimates:
mean of x
 48.32011
```

After discovering the insight about the average profit earned by the company, let us now hypothesize about the average sales in the West and east region of the US. As we can see below the null and alternate hypothesis constructed for the average sales done by the company.

$H_0$ : Company's average sales is higher in the West region than East region of the US
$H_A$ : Company's average sales is higher in the East region than West region of the US

We created two subsets, one containing east region data and other one containing west region data. Continuing with the subsets we created, we then took a random sample of 23 from both east and west region and performed two sample t-test.



```
t-test (Two sample)
Null hypothesis   Ho : Company's average sales is higher in the West region than East region of the US
Alternative hypothesis Ha : Company's average sales is higher in the East region than West region of the US

t.test(westRandomSample$Sales, eastRandomSample$Sales, conf.level = 0.95,
       alternative = "greater")
```
```

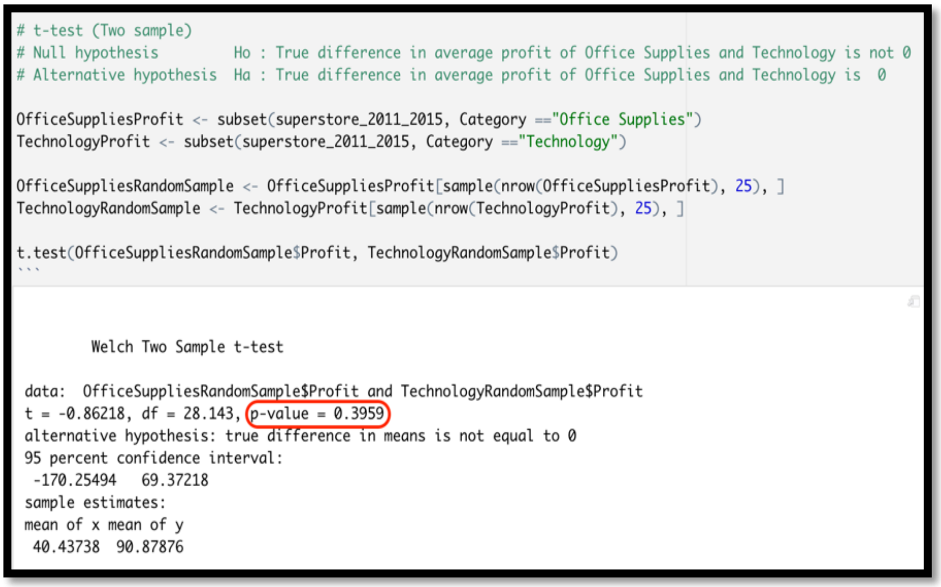
```

	Welch Two Sample t-test

data:  westRandomSample$Sales and eastRandomSample$Sales
t = -0.56348, df = 22.74, p-value = 0.7107
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -532.9089       Inf
sample estimates:
mean of x mean of y 
 225.7827  357.5908 
```

As shown in the figure above we obtained a p-value of 0.7107 which is greater the 0.05. However, to double check the result the average sales in the east and west region are $225 and $357 respectively.

Hence we successfully failed to reject the null hypothesis that Company’s average sales is higher in the East region than West region of the US.

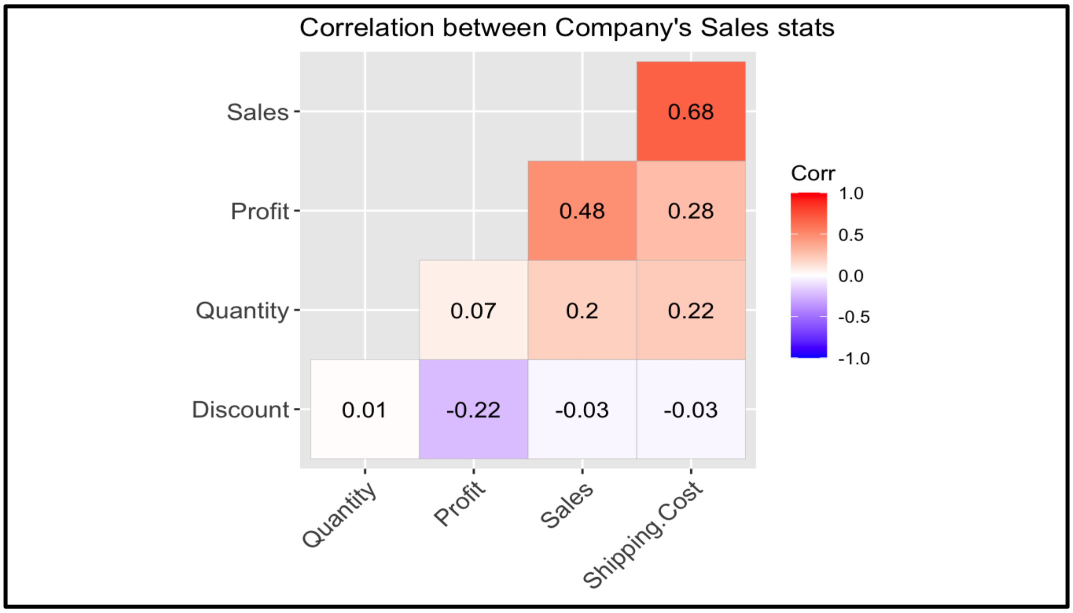
H_0 : Average profit earned of Office Supplies and Technology are same.
H_A : Average profit earned of Office Supplies and Technology are not same.

In the below image the first half shows that we have subset data based on Office supplies and Technology category. To perform t-test we have taken a random sample of 25 from the subset. Two sample t-test was performed to see if the Average profit earned in Office supplies and Technology category are same.



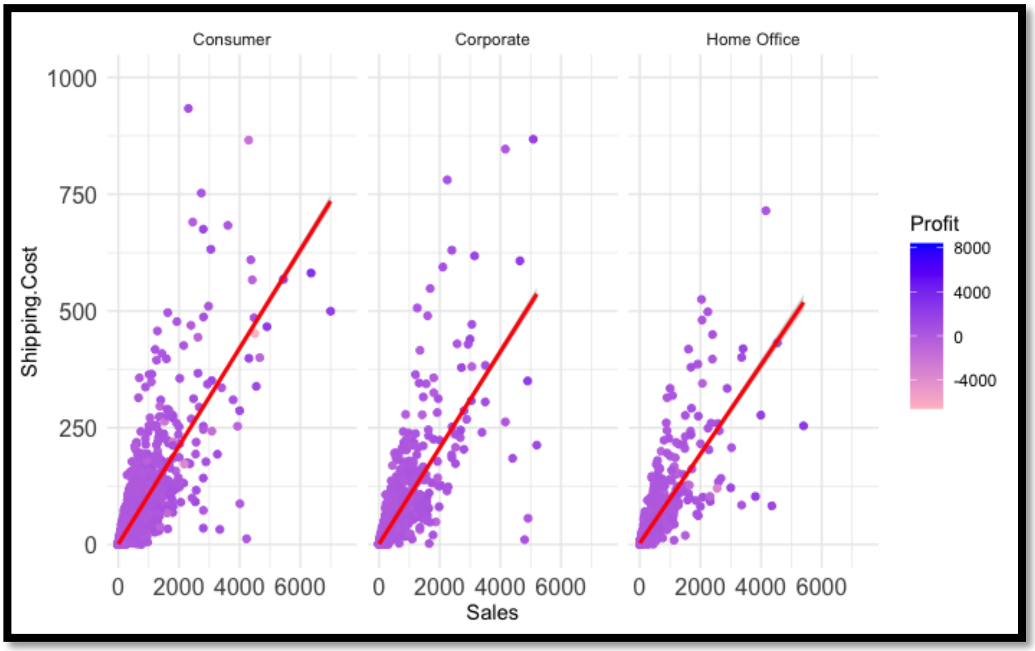
```
# t-test (Two sample)
# Null hypothesis        Ho : True difference in average profit of Office Supplies and Technology is not 0
# Alternative hypothesis  Ha : True difference in average profit of Office Supplies and Technology is  0

OfficeSuppliesProfit <- subset(superstore_2011_2015, Category =="Office Supplies")
TechnologyProfit <- subset(superstore_2011_2015, Category =="Technology")

OfficeSuppliesRandomSample <- OfficeSuppliesProfit[sample(nrow(OfficeSuppliesProfit), 25), ]
TechnologyRandomSample <- TechnologyProfit[sample(nrow(TechnologyProfit), 25), ]

t.test(OfficeSuppliesRandomSample$Profit, TechnologyRandomSample$Profit)
```
```


	Welch Two Sample t-test

data:  OfficeSuppliesRandomSample$Profit and TechnologyRandomSample$Profit
t = -0.86218, df = 28.143, p-value = 0.3959
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -170.25494   69.37218
sample estimates:
mean of x mean of y 
 40.43738  90.87876 
```

We obtained a p-value of 0.3959 which is greater the 0.05. However, to double check the result the average profit in the Office supplies and Technology category are $40.3 and $90.87 respectively. Hence we successfully reject the null hypothesis and accept the alternative hypothesis that Company's Average profit earned of Office Supplies and Technology are not same.

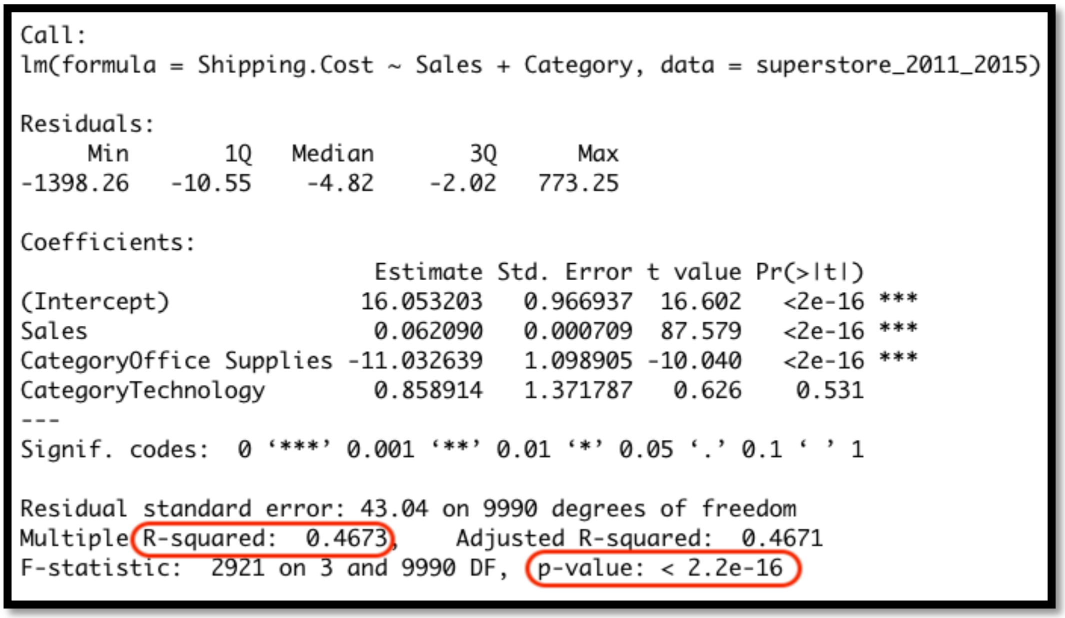


The above figure shows us the correlation stats between the numerical variables of the sales stats of the company in the US region. The values range between -1 to +1 where representing negative and positive relation between the variables. As we can see, Sales has a positive correlation with Shipping cost of the order and merely discount has any with Quantity of the product ordered. It is evident that Profit has a negative correlation with discount offered, more the discount offered less is the profit earned.

Let now visualize and analyze a regression model introducing a categorical variable with Sales and shipping cost associated with the order.



Regression model of Sales vs Shipping cost for different Customer categories



```
Call:
lm(formula = Shipping.Cost ~ Sales + Category, data = superstore_2011_2015)

Residuals:
     Min       1Q   Median       3Q      Max 
-1398.26   -10.55    -4.82    -2.02   773.25 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              16.053203   0.966937  16.602   <2e-16 ***
Sales                     0.062090   0.000709  87.579   <2e-16 ***
CategoryOffice Supplies -11.032639   1.098905 -10.040   <2e-16 ***
CategoryTechnology        0.858914   1.371787   0.626    0.531    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.04 on 9990 degrees of freedom
Multiple R-squared:  0.4673,	Adjusted R-squared:  0.4671 
F-statistic:  2921 on 3 and 9990 DF,  p-value: < 2.2e-16
```

Linear model of Shipping cost and sales made

As we can from the image above the R^2 value is **0.4673**, that means the linear model fits **46.73%** of the datapoint to the regression line. As the sales increases the shipping cost of the order increases. This justifies the positive correlation that we observed from the correlation chart. We can also see the increase in profit of the order in a gradient scale as shown in the legend scale. We have analyzed the above variation for different customer categories so and hence we have three separate graphs with 3 regression lines for Consumer, Corporate and Office supplies respectively.

Summary:

In this report we have attempted to answers few questions that would be a great insight for the business team and to mention marketing and sales team about the company's performance from 2011 to 2014. Some of the key points are:

- Sales achieved per region such as East, West, Central, South in the US market.
- Total order counts received for different categories.
- Profit earned per quantity of items sold based on different item segments.
- Most preferred shipping modes by the customers and it's dependency on the total sales made.
- Timeline of profit trend of the company.
- Hypothesis of company's profit and sales in different segments and categories.

References:

1. Coder, R. (2021, May 24). Boxplot in R. R CODER. https://r-coder.com/boxplot-r/
2. Jittered points — geom_jitter. (n.d.). Jitter Plot.https://ggplot2.tidyverse.org/reference/geom_jitter.html
3. W. (2021, January 21). How to set Colors for Bars in Bar Plot in R? TutorialKart. https://www.tutorialkart.com/r-tutorial/r-set-colors-for-bars-in-barplot/
4. Johnson, D. (2021, October 7). Scatter Plot in R using ggplot2 (with Example). Guru99. https://www.guru99.com/r-scatter-plot-ggplot2.html
5. Wetherill, C. (n.d.). How to Perform T-tests in R | DataScience+. Datascience. https://datascienceplus.com/t-tests/
6. The Open Educator - 9. Two Sample T-Test Unequal Variance. (n.d.). Open Educator. https://www.theopeneducator.com/doe/hypothesis-Testing-Inferential-Statistics- Analysis-of-Variance-ANOVA/Two-Sample-T-Test-Unequal-Variance
7. Selecting Random Samples in R: Sample() Function. (2021, November 27). ProgrammingR. www.programmingr.com/examples/neat-tricks/sample-r-function/

Appendix:

An attempt to complete this report helped us gaining a strong foundation on the basic descriptive analytics of data, hypothesis testing and linear model generation and it's interpretations. Using R we became familiar with libraries such as dplyr, ggplot, cor, color brewer for analysis and visualization.

The code used to complete this assignment is submitted in the Rmd R-Markdown format and output is in the html format as mentioned below.

1. Final-Project.html – R code output in HTML format
2. Final-Project.Rmd – R Code in R markdown format