BIRMINGHAM CITY
University

STATISTIC ANALYSIS OF ONLINE SHOPPING PORTAL DATA

Darshan Tamboli (STUDENT ID: 21171488)

CMP7205 Applied Statistics

12/12/21

CONTENTS

EXECUTIVE SUMMARY

This report evaluates the data of an online shopping portal data. The report's dataset is taken from the year 2020. The methods used to determine a relationship between the variables are linear regression, Pearson's Correlation, and the t-test. To support the observations, graphs such as the box plot, bar chart, Q-Q plot, and histogram have been applied.

The count of orders from different states and regions for different categories is presented using the bar chart. The frequency distribution of profit percentage is visualized with the help of a histogram.

While exploring the associations between variables, the impact of offered discount on sales is tested using the hypothesis testing, the Pearson's correlation value of 0.53 indicated that there is a significant positive association between sales and profit. Last, the regression model may be utilised to estimate the profit with the help of other variables.

1. INTRODUCTION

1.1 PROBLEM DOMAIN

Since its commercialization around 1990, the Internet's expansion has dominated any other technological communication development in human history. It is now commonly acknowledged that no communications medium or technological technology has ever expanded at the rate that the Internet has.

Due to the growing number of online users and items available on the Internet. Online purchasing has turned into a popular pastime that permits shoppers to experience the best way of life today.

It is a form of digital trade utilized for both B2B (business-to-business) as well as B2C (business-to-consumer) transactions, which are the procedures customers undergo to buy services and products through the internet. It has appeared in each aspect of life Beginning around 1990, associating individuals to enterprises on a successive and everyday basis.

On account of, increase in online shopping along with the desire of businesses to catch a huge percentage of the current as well as future

Internet market, hence, it is essential to interpret buyer’s characteristics on perspectives towards buying over the web (P. Georgiades, 2000).

This report attempts to explore the factors which can help to grow business rapidly. Significantly, knowing customers' online purchasing conduct is fairly identical to understanding the shopping conduct of everybody (Mokhtarian, 2005).

The discoveries of this analysis can make a contribution withinside the following manner: The retailer should therefore be able to priorities as well as can target the internet market to a greater degree

1.2 Statistical Questions

There are a few statistical questions to dig into the dataset which has details of United States e-commerce records from 2020 to get knowledge from it.

The subsequent questions will be addressed in this analysis report.

Q1. Do sales change when the discount is offered?

Q2. What is the relationship between sales and profit?

Q3. From which state retailer receives the most orders?

Q4. Which shipping mode customer prefers most?

Q5. Which category of product is most in demand?

Q6. From which region retailer receives the most orders?

2 METHODOLOGY

This section outlines the different types of techniques used to analyse the datasets along with graphs used to present the data for realisation. The reasons for selecting these techniques, in particular, are explained in the report's subsequent sections.

Each data set must be pre-processed before it can be analysed. This means that the data must be framed in a specific way for the conclusion to be drawn.

2.1 EDA (EXPLORATORY DATA ANALYSIS)

The first step Exploratory Data Analysis is used to study and evaluate dataset, as well as to describe their major properties, many times data visualisation is used. It assists in determining how to effectively transform data sources to obtain the answers you want, making it simpler to identify connections.

Histograms are used to visually show the frequency distribution of data and are used to visualise data skewness. Boxplots are used to visualise outliers in data by visually representing summary statistics. The frequency counts of entries for distinct categorical or nominal variables are presented by a ggplot (Hadley Wickham, n.d.).

EDA will help to find out the preferred shipping mode, most demanded category of products as well as the states and region from where the retailer receives the maximum orders. (i.e. Q3, Q4, Q5, Q6)

2.2 HYPOTHESIS TESTING

In statistics, The technique by which an analyst verifies an assumption about a population parameter is known as hypothesis testing (Paiva, 2010). First, preliminary assumptions are formed regarding the parameters or distribution. This claim is known as the null hypothesis and is symbolised by the symbol Ho. Then, an alternative hypothesis (symbolised as Ha) is defined, which is the opposite of what is claimed in the null hypothesis.

Hypothesis testing will help to test the change in sales when a discount is being offered (i.e., Q1)

2.3 PEARSON'S CORRELATION ANALYSIS

The Pearson correlation coefficient, often referred as Pearson's r, in a statistic. It measures linear correlation among sets of two different variables (Soetewey, 2020). Coefficient values range from -1 to +1. A value of 0 means that there is no association between the two variables.

If a value is higher than 0 then it denotes positive association; when one variable's value increases, the value of another variable increase as well. If a value is lower than 0 then it denotes negative association; when one variable's value increases, the value of another variable falls.

The Pearson's Correlation Coefficient, which quantifies the relationship between Discount, Quantity, Sales, and Profit, will be displayed in a correlation matrix. (i.e., Q2)

2.4 Regression Analysis

It is a collection of statistical processes for evaluating the connections between a dependent variable and one or a set of independent variables. It is commonly used for forecasting and prediction (Lantz, 2015)

The purpose of regression Is to predict the value of profit based on the other value of independent variables like sale, category, sub-category, and discount

3 DATASET

This section outlines the type of information obtained and its source, as well as an overview of the data set.

United States e-commerce records (Ammaraahmad, 2020)**.**

This is a public dataset of orders placed on an e-commerce website in the United States. The dataset contains information on over 3000 orders placed through an online portal in the year 2020 and contains 15+ columns. Its features enable to see an order from different dimensions: from order details, when a customer ordered, shipment mode for order, ordered from which region, city and state and postal code, Categories and subcategories of products, Quantity of products in order, how much discount applied and how much profit gained from the sale.

Table 1: Sample of E-commerce Dataset

Order Date	Order ID	Ship Mode	Customer	Segment	City	State	Region	Product ID	Category	Sub-Category	Sales	Quantity	Discount	Profit	ProfitorLo	Prof Perce	IsDiscounted
01-01-2020	CA-2017-1	Standard Class	GA-14725	Consumer	Lorain	Ohio	East	FUR-FU-1(	Furniture	Furnishings	48.896	4	0.2	8.5568	Profit	17.5	TRUE
01-01-2020	CA-2017-1	Standard Class	SC-20725	Consumer	Los Angeles	California	West	FUR-FU-1(	Furniture	Furnishings	474.43	11	0	199.2606	Profit	42	FALSE
01-01-2020	CA-2017-1	First Class	DP-13390	Home Offi	Franklin	Wisconsin	Central	OFF-BI-10	Office Supplies	Binders	3.6	2	0	1.728	Profit	48	FALSE
01-01-2020	CA-2017-1	Standard Class	JM-15250	Consumer	Huntsville	Texas	Central	OFF-ST-10	Office Supplies	Storage	454.56	5	0.2	-107.958	Loss	-23.75	TRUE

This report is meant to test the hypothesis to find out a change in customer's spending on the offered discount, discover the correlation between different attributes (discount, quantity, sales, and profit), forecast sales using linear regression.

3.1 DATA PRE-PROCESSING

i. Country column has been removed because all record has a common country so it's not going to help for any kind of statistic.
ii. Added column with calculated profit percentage to get an insight of profit in the form of a percentage.
iii. "Product name" and "Row ID" columns have less relevance so removed
iv. City and State column are present, so removed Postal Code
v. Added column to indicate Boolean status for Discounted/Not Discounted for hypothesis testing.
vi. Removed outliers where profit percentage is -100+

4 RESULTS AND DISCUSSION

4.1 EDA (EXPLORATORY DATA ANALYSIS)

A statistical summary of the deviation in profit percentage is presented in Table 1. The Median is 28 and the mean is 19.08, which states that this distribution is not a normal distribution and is skewed toward the left as the mean is towards the left of the median.

Table 2: Statical Summary of Profit Percentage

Profit Percentage	
Measure	**Result (Deviation in %)**
Minimum	-97.5
1st Quartile	8.75
Median	28
Mean	19.08
3rd Quartile	36.25
Maximum	50
Skewness	-1.742857

Figure 1 represents the Histogram, which shows the frequency distribution of profit percentage per order. The bars in this histogram are skewed to the left, Therefore, it is called a skewed left histogram. It shows the profit percentage earned per order. As per analysis most of the orders give profit but only a few have shown loss.

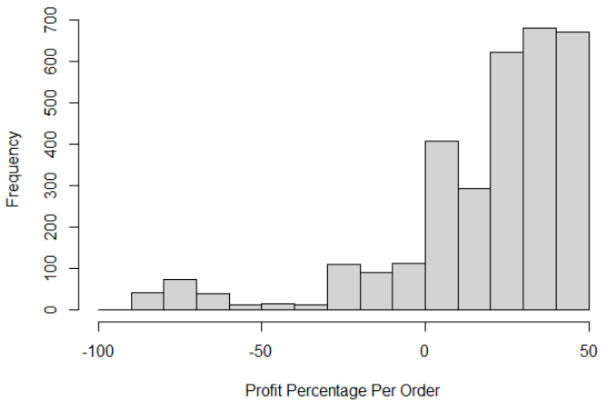

Fig 1: FREQUENCY DISTRIBUTION OF PROFIT PERCENTAGE

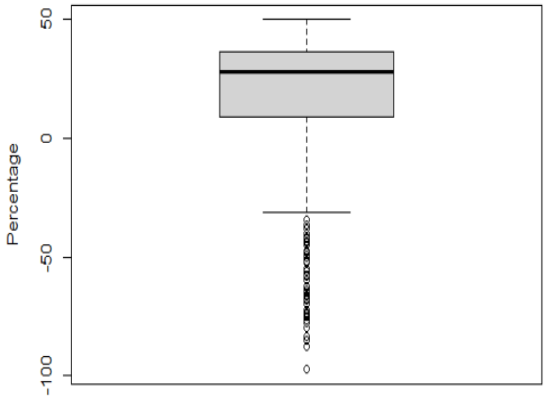

Fig 2: BOX PLOT OF PROFIT PERCENTAGE

Figure 2 represents a box plot of a profit percentage. A boxplot's box begins in the first quartile (25 percent) and finishes with the third quartile (75%). As a result, the box shows 50% of the center data, and the lines in it represent the median. The farthest data segments are drawn on either side of the box without counting outliers in the boxplot. Outliers in the boxplot are represented by circles. The above boxplot shows a few outliers in the data due to purchase pricing for the retailer. But if we remove these outliers then it can impact the data. Loss is expected in a few orders in the real scenarios so keeping these outliers.

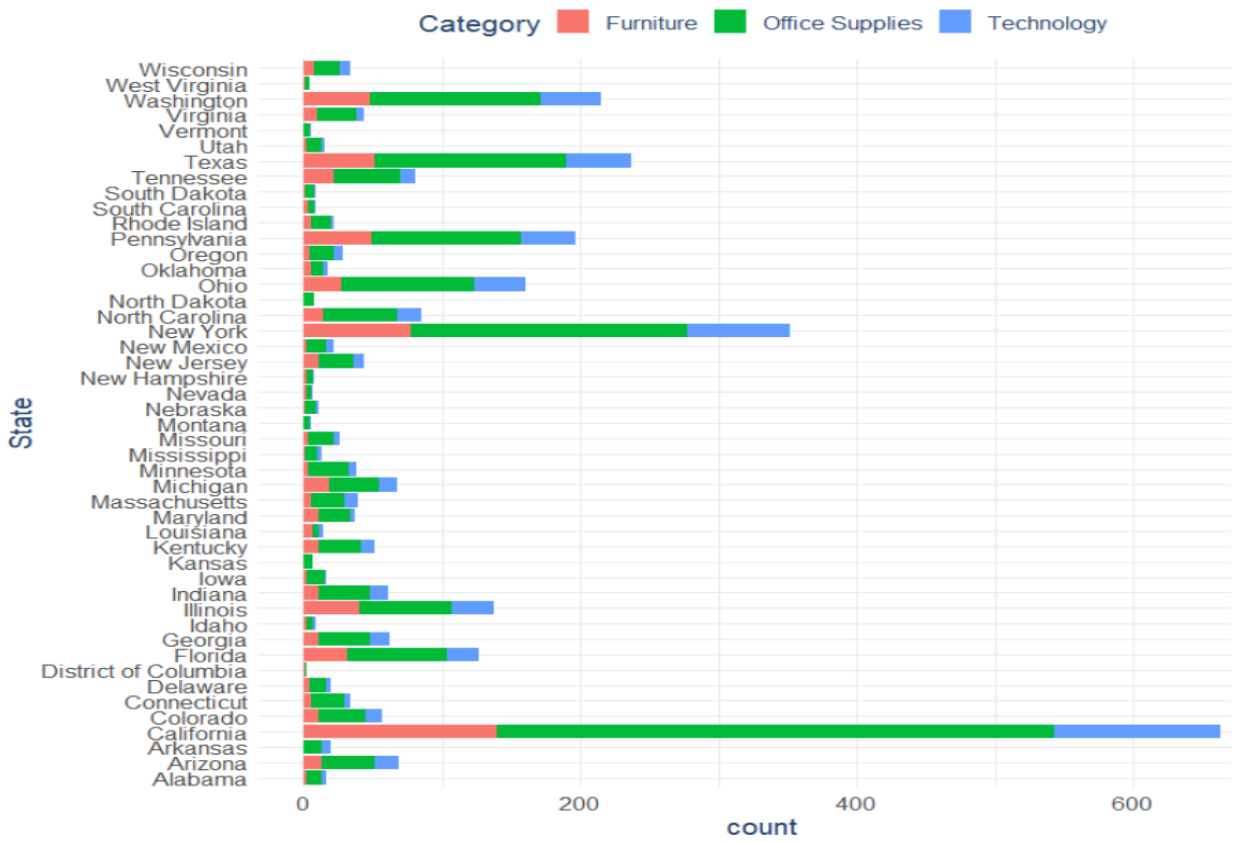

Fig 3: Bar chart of order count from different states and product categories

The above Bar chart shows (Fig 3), the maximum number of orders gets from California, New York, Texas, Washington, and Pennsylvania and the maximum orders are for the Office supplies category. After analyzing this bar chart businesses can focus on states like the District of Columbia, West Virginia, Vermont, Montana, Nevada, Kansas, etc., and furniture categories to boost a business.

Fig 4: Bar chart of order count from different Region and Shipment Mode

The above Bar chart (Fig 4) shows, the maximum number of orders gets from the West and East and maximum customers prefer Standard and Second-class shipment mode and get minimum orders from the south region and least customers choose the same day delivery.

4.2 Hypothesis Testing

Null Hypothesis (Ho): There is no difference in sales when a discount is offered.
Alternative Hypothesis (Ha): There is a difference in sales when a discount is offered.

It is a two-sided test because no directionality is implied and It's a two-sample test because a discount is applied or not are statistically independent samples

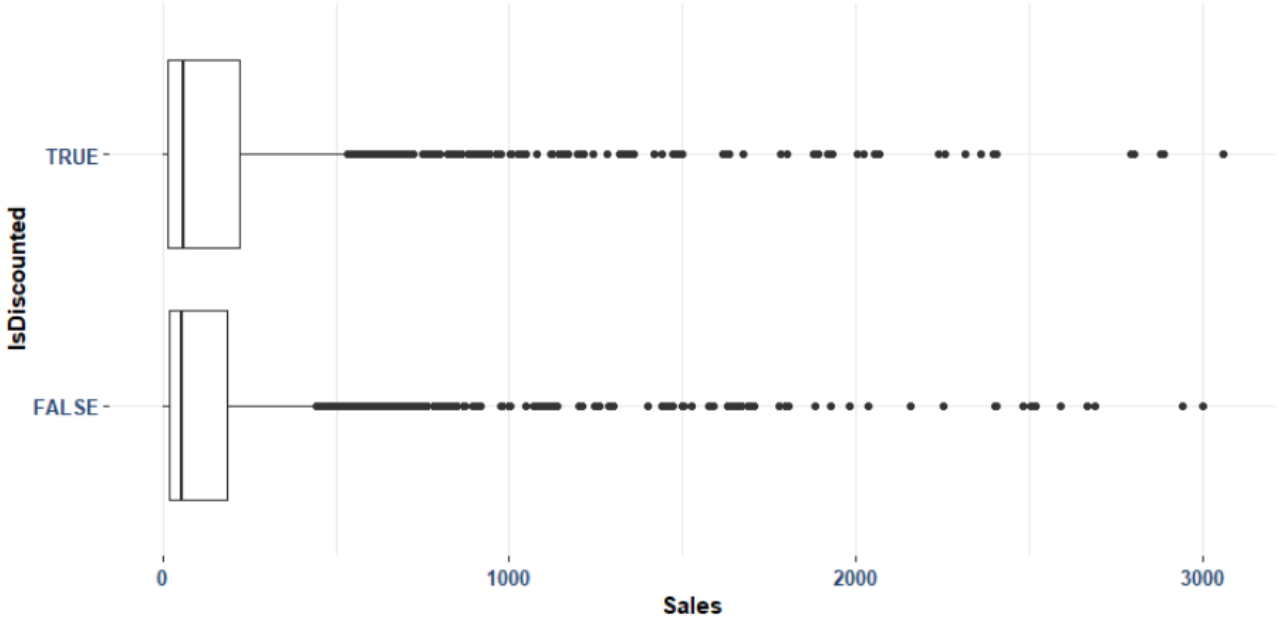

Fig 5: Boxplots of Sales with or without discount

Figure 5 represents the box plots of sales with or without discount. As per the figure, we can see the median in both cases is quite similar so there are high chances of retaining the null hypothesis. This can be confirmed with the hypothesis testing as shown in figure 6.

```
        Two Sample t-test

data:  Sales by IsDiscounted
t = -0.66761, df = 3170, p-value = 0.5044
alternative hypothesis: true difference in means between group FALSE and group
 TRUE is not equal to 0
95 percent confidence interval:
 -34.66315  17.05380
sample estimates:
mean in group FALSE  mean in group TRUE
           193.2486            202.0533
```

Fig 6: Hypothesis test to find a difference in mean sales between discount offered or no

As calculated in Figure 6, a p-value is 0.5044 > 0.05, therefore the probability that the null hypothesis is true.

4.3 Pearson's Correlation Analysis

Fig 7: Correlation Matrix of Discount, Quantity, Sales, Profit

Figure 7 represents the correlation coefficients between Discount, Quantity, Sales, and Profit. The non-significant correlations are highlighted through the use of a large cross on the correlation coefficients value (The Holm adjustment approach is used by default at a significance threshold of 5%).

The negative association seen between Discount and the price is - 0.29, as shown in the graph, whereas, the positive association seen between Sales and the Profit is 0.53. This suggests that there is a strong association between discounts and the price, as well as the sales and profit.

4.4 Linear Regression Analysis

The model to analyze is a linear regression between variables.

```
Call:
lm(formula = Profit ~ Sales + Category + Discount, data = USECOMDATA)

Residuals:
     Min       1Q   Median       3Q      Max
-1486.06   -23.51     3.33    19.30   904.87

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -14.12622    4.11025  -3.437 0.000596 ***
Sales                     0.16862    0.00454  37.141  < 2e-16 ***
CategoryOffice Supplies  42.39639    4.21363  10.062  < 2e-16 ***
CategoryTechnology       43.39282    5.10870   8.494  < 2e-16 ***
Discount               -177.34130    9.57723 -18.517  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.19 on 3167 degrees of freedom
Multiple R-squared:  0.3741,    Adjusted R-squared:  0.3733
F-statistic: 473.2 on 4 and 3167 DF,  p-value: < 2.2e-16
```

Fig 8: Linear Regression model of Discount, Category, Sales, Profit

In figure 8, the p-value of the model is 2.2e-16 which is a scientific sign of 0.00000000000000022, indicating that it is extremely near to zero and is less than 0.05. But the Multiple R2 value for the regression is 0.3741, which is a not good fit for the model. Let's try with more independent variables

```
Call:
lm(formula = Profit ~ Sales + Region + Quantity + Category +
    Sub.Category + Discount, data = USECOMDATA)

Residuals:
     Min       1Q   Median       3Q      Max
-1522.33   -21.42    -0.60    21.47   723.98

Coefficients: (2 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.189e+02  9.839e+00 -12.085  < 2e-16 ***
Sales                    1.774e-01  4.925e-03  36.013  < 2e-16 ***
RegionEast              -9.306e+00  4.333e+00  -2.148 0.031807 *
RegionSouth             -1.264e+01  4.978e+00  -2.539 0.011176 *
RegionWest              -9.544e+00  4.232e+00  -2.255 0.024184 *
Quantity                -1.041e+00  7.278e-01  -1.430 0.152756
CategoryOffice Supplies  1.214e+02  1.425e+01   8.519  < 2e-16 ***
CategoryTechnology       1.632e+02  1.012e+01  16.119  < 2e-16 ***
Sub.CategoryAppliances   4.356e+01  1.332e+01   3.271 0.001083 **
Sub.CategoryArt          2.690e+01  1.228e+01   2.191 0.028511 *
Sub.CategoryBinders      7.683e+01  1.217e+01   6.314 3.10e-10 ***
Sub.CategoryBookcases    1.061e+02  1.317e+01   8.059 1.08e-15 ***
Sub.CategoryChairs       1.210e+02  1.037e+01  11.670  < 2e-16 ***
Sub.CategoryCopiers      2.262e+02  2.135e+01  10.594  < 2e-16 ***
Sub.CategoryEnvelopes    3.775e+01  1.501e+01   2.516 0.011927 *
Sub.CategoryFasteners    2.395e+01  1.539e+01   1.557 0.119654
Sub.CategoryFurnishings  1.553e+02  1.002e+01  15.501  < 2e-16 ***
Sub.CategoryLabels       3.129e+01  1.371e+01   2.283 0.022518 *
Sub.CategoryMachines    -3.388e+01  1.766e+01  -1.919 0.055125 .
Sub.CategoryPaper        4.004e+01  1.186e+01   3.377 0.000742 ***
Sub.CategoryPhones      -1.885e+01  7.187e+00  -2.623 0.008747 **
Sub.CategoryStorage      6.268e+00  1.224e+01   0.512 0.608594
Sub.CategorySupplies            NA         NA      NA       NA
Sub.CategoryTables              NA         NA      NA       NA
Discount                -1.948e+02  1.025e+01 -18.998  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.33 on 3149 degrees of freedom
Multiple R-squared:  0.4678,    Adjusted R-squared:  0.4641
F-statistic: 125.8 on 22 and 3149 DF,  p-value: < 2.2e-16
```

Fig 9: Linear Regression model of Discount, Category, Sub Category, Sales, Profit, Region, Quantity

In figure 9, the p-value of the model is less than 0.05. The Multiple R^2 value for the regression is 0.4678, the model is performing better as compared to the previous model.

To improve model performance, we can add a non-linear relationship and convert a numeric variable to a binary indicator, and put it all together (Lantz, 2015).

```
Call:
lm(formula = Profit ~ Sales + Sales2 + Region + Sub.Category +
    Discount + Discount3 * Category, data = USECOMDATA)

Residuals:
     Min       1Q   Median       3Q      Max
-1516.91   -21.05    -4.23    17.97   664.95

Coefficients: (2 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.259e+01  6.182e+00   8.508  < 2e-16 ***
Sales                          9.839e-02  1.085e-02   9.068  < 2e-16 ***
Sales2                         3.910e-05  5.073e-06   7.708 1.70e-14 ***
RegionEast                    -1.012e+01  4.357e+00  -2.324 0.020198 *
RegionSouth                   -2.005e+01  4.979e+00  -4.028 5.77e-05 ***
RegionWest                    -1.572e+01  4.243e+00  -3.705 0.000215 ***
Sub.CategoryAppliances         2.869e+00  8.615e+00   0.333 0.739092
Sub.CategoryArt               -2.389e+01  7.107e+00  -3.361 0.000786 ***
Sub.CategoryBinders            1.935e+01  7.018e+00   2.757 0.005870 **
Sub.CategoryBookcases         -3.727e+01  1.137e+01  -3.279 0.001052 **
Sub.CategoryChairs            -2.235e+01  8.226e+00  -2.716 0.006636 **
Sub.CategoryCopiers            2.266e+02  2.082e+01  10.881  < 2e-16 ***
Sub.CategoryEnvelopes         -1.171e+01  1.103e+01  -1.062 0.288470
Sub.CategoryFasteners         -2.791e+01  1.153e+01  -2.420 0.015583 *
Sub.CategoryFurnishings       -8.055e+00  6.938e+00  -1.161 0.245760
Sub.CategoryLabels            -1.913e+01  9.273e+00  -2.063 0.039185 *
Sub.CategoryMachines           2.240e+00  1.784e+01   0.126 0.900097
Sub.CategoryPaper             -8.788e+00  6.371e+00  -1.380 0.167836
Sub.CategoryPhones            -4.717e+00  7.209e+00  -0.654 0.513025
Sub.CategoryStorage           -3.401e+01  6.944e+00  -4.898 1.02e-06 ***
Sub.CategorySupplies          -4.624e+01  1.200e+01  -3.852 0.000119 ***
Sub.CategoryTables            -1.259e+02  1.087e+01 -11.585  < 2e-16 ***
Discount                      -1.634e+02  1.558e+01 -10.492  < 2e-16 ***
Discount3                     -5.823e+01  9.748e+00  -5.973 2.59e-09 ***
CategoryOffice Supplies               NA         NA      NA       NA
CategoryTechnology                    NA         NA      NA       NA
Discount3:CategoryOffice Supplies  6.611e+01  1.305e+01   5.065 4.32e-07 ***
Discount3:CategoryTechnology      -4.034e+01  1.536e+01  -2.626 0.008690 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.31 on 3146 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4894
F-statistic: 122.6 on 25 and 3146 DF,  p-value: < 2.2e-16
```

Fig 10: Regression model of Discount, Category, Sub Category, Sales, Profit with non-linear relationship and binary indicator.

In figure 10, the p-value of the model is less than 0.05. The Multiple R^2 value for the regression is 0.4935, the model is performing fairly better as compared to the previous model, which tells us, these are the predictor for the Profit.

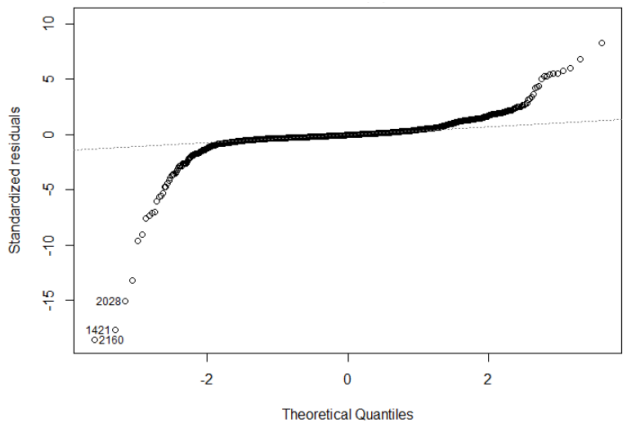

Fig 11: Normal Q-Q Plot for residuals of the regression model

Figure 11 shows that most of the residual points are close to the Q-Q (quartile-quartile) line.

5 CONCLUSION

The outcomes of this report answered all of the statistical questions highlighted at the start. The business is moving in a good direction as its earning profit in maximum orders. The very low number of orders getting from cities like District of Columbia, West Virginia, Vermont, Montana, Nevada, Kansas, etc., So the business can push some strategies like promotions and tie-ups to get a greater number of orders from these cities.

It is observed that fewer people placing orders with same-day delivery as well as least orders in furniture category. To tackle this organization can put promotions over social media and gain more orders for the furniture category as well as for same-day delivery. The currently offered discounts are not having a huge impact on sales figures. The total amount of sales is the most important factor to get more profit in the business. So, the company needs to create some strategies to boost total sales using a prediction model with discounts in different categories as well as subcategories.

This data report might be expanded in the future by other studies that evaluate consumer behaviour with regard to current market trends and offered items on the portal.

6 REFERENCES

1. P. Georgiades, J. DuPreez, B. Dowsland, and A. Simintiras, "Attitudes towards on-line purchase behavior: Comparing academics, students and other," presented at the European Business Management School (2000).
   Available at:
   https://www.researchgate.net/publication/242160162_Attitudes_toward_On-line_Purchase_Behavior_Comparing_Academics_Students_and_Others (Accessed: December 2021)

2. P. Cao and P.L. Mokhtarian, "The Internet and actual adoption of online purchasing: A brief review of recent literature," presented at the University of California (2005)
   Available at: http://docplayer.net/14928232-The-intended-and-actual-adoption-of-online-purchasing-a-brief-review-of-recent-literature.html (Accessed: December 2021)

3. United States E-Commerce records (2020)
   Available at: https://www.kaggle.com/ammaraahmad/us-ecommerce-record-2020
   (Accessed: December 2021)

4. Brett Lantz, 'Machine learning with R – Second edition', PACKT Publishing (2015)
   Available at:
   https://ebookcentral.proquest.com/lib/bcu/reader.action?docID=2122139&query=
   (Accessed: December 2021)

5. Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, K. Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, 'Bar charts' (No date)
   Available at: https://ggplot2.tidyverse.org/reference/geom_bar.html#arguments
   (Accessed: December 2021)

6. Antonio Paiva, 'Hypothesis Testing' (2010)
   Available at:
   https://www.sci.utah.edu/~arpaiva/classes/UT_ece3530/hypothesis_testing.pdf
   (Accessed: December 2021)

7. Antoine Soetewey, 'Correlation coefficient and correlation test in R (2020)
   Available at: https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/
   (Accessed: December 2021)

Appendix: R Code

- **EDA**

```
> USECOMDATA <- read.csv("USE-commerce2020.csv")
> summary(USECOMDATA$Prof.Percentage)
> skewness(USECOMDATA$Prof.Percentage)
> hist(USECOMDATA$Prof.Percentage, main="Distribution", xlab="Profit Percentage")
> boxplot(USECOMDATA$Prof.Percentage, main="Boxplot of Prof Percentage",
ylab="Percentage")
> ggplot(USECOMDATA, aes(y = State))
+ geom_bar(aes(fill = Category), position = position_stack(reverse = TRUE))
+ theme(legend.position = "top")
+ theme(text = element_text(size = 15, color = "#284571"))
> ggplot(USECOMDATA, aes(y = Region)) + geom_bar(aes(fill = Ship.Mode), position =
position_stack(reverse=TRUE))+theme(legend.position="top")+theme(text=element_text(s
ize = 15, color = "#284571"))
```

- **Hypothesis Analysis**

```
>ggplot(USECOMDATA, aes(x=Sales, y=IsDiscounted)) + geom_boxplot()+
theme(axis.text.x=element_text(face="bold",color="#284571",size=10,angle=0),
axis.text.y=element_text(face="bold",color="#284571",size=10))
> t.test(Sales ~ IsDiscounted, data = USECOMDATA, var.equal = TRUE, paired = FALSE)
```

- **Pearson's correlation Analysis**

```
> library(ggstatsplot)
> ggcorrmat( data = USECOMDATA[, c("Discount", "Quantity", "Sales", "Profit")], type
="parametric",#parametric for Pearson
  colors = c("darkred", "white", "steelblue") # change default colors) +
theme(axis.text.x=element_text(face="bold",color="#284571", size=10,angle=0),axis.text.y
= element_text(face="bold", color="#284571",size=10))
```

- **Regression Analysis**

```
> model <- lm(Profit ~ Sales + Category  + Discount, data=USECOMDATA)
> summary(model)
> model <- lm(Profit ~ Sales  + Region + Quantity + Category + Sub.Category + Discount,
data=USECOMDATA)
> summary(model)
> USECOMDATA$Sales2 <- USECOMDATA$Sales^2
> USECOMDATA$Discount3 <- ifelse(USECOMDATA$Discount >= 0.3, 1 , 0)
> model2 <- lm(Profit ~ Sales + Sales2 + Region  + Sub.Category + Discount +
Discount3*Category, data=USECOMDATA)
> summary(model2)
> plot(model2)
```