Impact of Sales Discounts, Shipping Modes and Customer Variables on Sales Performance in North America within 2009 to 2012.

For the Bachelor of Science Honours Degree in Financial Mathematics and Industrial Statistics

By

D.M.B.D.Dissanayake(SC/2021/12461)

Supervisor :

Ms. K.C.N. Shanthidevi

Department of Mathematics

University of Ruhuna

Matara

Declaration

I, D.M.B.D.Dissanayake, declare that the presented project report titled, "Impact of Sales Discounts, Shipping Modes and Customer Variables on Sales Performance in North America within 2009 to 2012" is uniquely prepared by me based on the group project carried out under the supervision of Ms. K.C.N. Shanthidevi, Department of Mathematics, Faculty of Science, University of Ruhuna, as a partial fulfillment of the requirements of the level II , Case Study course unit, MIS 2231 of the Bachelor of Science Honours Degree in Financial Mathematics and Industrial Statistics in Department of Mathematics,Faculty of Science, University of Ruhuna, Sri Lanka.

It has not been submitted to any other institution or study program by me for any other purpose.

Signature: . . . . . .
Date: . . . . . .

Supervisor's Recommendation

I certify that this study was carried out by D.M.B.Dissanayake under my supervision.

Signature:. . . . . .
Date:. . . . . .
Department of Mathematics
Faculty of Science
University of Ruhuna.

Acknowledgement

First, we would like to express my sincere gratitude to our supervisor Ms. K.C.N. Shanthidevi madam who helped us to learn a lot about this case study. Her ideas and comments aided in the completion of this study. Without her support and guidance, this study would not have been possible.

As well as, we would like to thank our course coordinator Dr. A.W.L. Pubudu Thian sir who gave us many more valuable instructions to fulfill our case study successfully.

And also, we extend our thanks to all the instructors who helped us to make this research a success. Also, we express our sincere appreciation to all our batch mates for sharing their knowledge with us.

Finally, we are thankful to everyone who played a part; big or small, to make this research a success.

Table of Contents

Declaration ii
Supervisor's Recommendation ii
Acknowledgement iii
List of Figures v
List of Tables vii
Abstract ix

1 Introduction **1**
1.1 Background of the study 1
1.2 Problem statement 2
1.3 Objective of the study 2
1.3.1 Research objectives 3
1.4 Research questions 4
1.5 Hypothesis 4
1.5.1 Hypothesis 1 4
1.5.2 Hypothesis 2 5
1.6 Significance of the study 5

2 Literature Review **6**
2.1 Introduction 6
2.1.1 Sales 7
2.1.2 Unit Price 7
2.1.3 Sales Promotion 7
2.1.4 Shopping Cost 7
2.1.5 Shopping Duration 8
2.1.6 Customer Segment 8
2.1.7 Product Categories 8

3 Materials and Methods **9**
3.1 Research approach 9
3.2 Conceptual model 10
3.3 Research design 10

3.3.1 Multiple linear regression 10
3.3.2 Selecting the best regression equation 11

4 Data **12**
4.1 About the data set 12
4.2 Metadata 12
4.3 Data Dictionary 13
4.4 Data set preparation 15
4.4.1 Check null values 15
4.4.2 Check NA values 15
4.4.3 Remove the variables that do not provide additional information to predict sales growth. 15
4.4.4 Create new variable 15
4.4.5 Assign values for categorical variable 16
4.4.6 Check Outliers 17

5 Results **19**
5.1 Exploratory data Analysis 19
5.2 Quantitative Analysis 24
5.2.1 Correlation Analysis 24
5.2.2 Assumptions for multiple linear regression 24
5.2.3 Estimate model parameters 27
5.2.4 Assess model fit 29

6 Discussion and conclusion **34**
6.1 Discussion 34
6.2 Conclusion 36

7 Appendix **37**

Bibliography **38**

List of Figures

3.1 *Conceptual model* 10

4.1 *Label for categories in data set* 16
4.2 *Histogram of sale* 17
4.3 *Box plot of sales* 17
4.4 *Histogram of log(sales)* 18
4.5 *Box plot of log(sales)* 18

5.1 *Descriptive Statistics Table* 19
5.2 *Histogram of log(sales)* 20
5.3 *Box plot of log(sales)* 20
5.4 *Histogram of Discount* 20
5.5 *Scatter plot of log(Sales) VS Discount(Sales promotion)* 20
5.6 *Histogram of Unit Price* 21
5.7 *Scatter plot of log(Sales) vs Unit Price* 21
5.8 *Histogram of Shipping Cost* 21
5.9 *Scatter plot of log(Sales) vs Shipping Cost* 21
5.10 *Histogram of Product category* 22
5.11 *Scatter plot of log(Sales) vs Product Category* 22
5.12 *Histogram of Shipping Mode* 22
5.13 *Scatter plot of log(Sales) vs Shipping Mode* 22
5.14 *Histogram of Customer Segment* 23
5.15 *Scatter plot of log(Sales) vs Customer Segment* 23
5.16 *Histogram of Shipping Duration* 23
5.17 *Scatter plot of log(Sales) vs Shipping Duration* 23
5.18 *Correlation Table* 24
5.19 *Normal probability plot with outlier* 24
5.20 *Normal probability plot without outlier* 24
5.21 *Correlogram* 25
5.22 *Standardized Residuals vs Fitted* 26
5.23 *Full model summary* 27

5.24 *Backward Elimination Steps* 28
5.25 *Backward Elimination Summary* 29
5.26 *ANOVA table of full model* 30
5.27 *ANOVA table of reduced model* 30

List of Tables

4.1 *Data Dictionary Table* 14
4.2 *Dependent and Independent variables* 14

Abstract

This study investigates the impact of unit price, sales promotion, shipping costs, shipping mode, customer segment, product category and shipping duration on online shopping company sales growth. Here, we will use a multiple linear regression model and the regression technique to investigate this link. This study aims to investigate the impact of these factors on the growth in sales of online retailers. Thus, this will provide insightful information about which elements are most crucial to generating a high volume of sales for the online retailer.

Key words: Customer Segment, Sales Discount, Sales growth, Shipping cost, Shipping Duration, Shipping Mode, Unit price

Chapter 1

Introduction

Overview

The structure of this chapter is as follows. Firstly concentrates on providing an overview of online purchasing and important procedures used during the model-development process. The topic of study and aims will be clarified in uture investigations.

1.1 Background of the study

Online shopping is a form of e-commerce that allows consumers to purchase goods or services directly from a seller over the Internet using a web browser or mobile application. Customers can find a product of interest by visiting the retailer's website directly or by searching through alternative suppliers using a shop search engine, which shows the availability and price of the same product from different e-tailers.

Online markets are now widely used worldwide due to the use of the Internet and new electronic devices. It has increased due to the recent Covid-19 pandemic. Nowadays, everyone including adults, youth and students, uses online shopping without any age limit.

online sales revenue continued to grow significantly, people started researching this area. The purpose of this study is to study the effect of unit price, sales promotion, shipping cost, shipping duration, shipping mode,customer segment, and product cate-

gory on the sales growth of online shopping companies in North America between 2009 and 2012 through the data collected by the Kaggle website.

1.2 Problem statement

As online shopping is becoming popular day by day we should aware of what are the factors that can improve sales capacity as customer is not contacting the seller directly. Our aim is to study about what factors are the most important to increase sales and what are the changes that should be done in this sector to gain a better revenue. So, we are going to identify the relationship between Unit price, Sales promotion, Shipping cost, Shipping mode, Customer segment, Product category and Shipping duration with sales.

At the end of the study, we will have a clear idea about what are the factors that need to be improved to have better revenue while what are the things that need to be changed to get better results.

1.3 Objective of the study

The main objective of this study is to examine the Impact of Unit price, Sales promotion, Shipping cost, Shipping duration, Shipping mode, Customer segment, Product category on Sales growth in Online Shopping Companies in North America. In addition to that, studying the individual contributions of these variables to sales growth, identifying whether they have positive or negative impact on sales growth.

we are planning to achieve our goal using the multiple regression model and then the step wise method to select the best model.

1.3.1 Research objectives

- To analyze the relationship between unit pricing and sales growth in an online shopping company.
- To assess the impact of sales promotions on consumer purchasing decisions sales performance.
- To examine the effect of shipping costs on consumer behavior and its implications for sales growth in an online shopping company.
- To investigate how shipping duration influences sales growth in an online shopping.
- To evaluate the impact of different customer segments on the sales growth of an online shopping.
- To study the influence of product categories on the sales growth of an online shopping .
- To explore the relationship between shipping modes and the sales growth of an online shopping company.

The mentioned regression model will be taking using the R software and the Mini tab software respectively. Since we are getting the best fitted line equation, we will not be able to get the exact correct values for the revenue, but we can get the most approximate answer accordingly.
At the end of the study, the outcome is going to be a great asset to the online shopping companies as they can use our results to overcome their shortcomings and can be used to make decisions in the future.

1.4 Research questions

To obtain meaningful research findings, the following research questions have been developed for this study:

- How do the unit price base on the sales growth of online shopping company?
- How does the impact of sales promotions offer on consumer purchasing decisions and resultant sales?
- What is the effect of shipping costs on consumer behavior and its implications for sales performance?
- How do the shipping duration base on the sales growth of online shopping company?
- How does the impact of customer segment on the sales growth of online shopping company?
- How does the impact of product category on the sales growth of online shopping company?
- Is a relationship between ship mode and the sales growth of the company?

1.5 Hypothesis

1.5.1 Hypothesis 1

- Null hypothesis (H0):

  There is no linear relationship between dependent variable and independent variables of the online shopping company.

- Alternative hypothesis (Ha):

  There is a linear relationship between dependent variable and independent variables of the online shopping company.

1.5.2 Hypothesis 2

- Null hypothesis (H0):

  Reduced model is suitable.

- Alternative hypothesis (Ha):

  Full model is needed.

1.6 Significance of the study

This study holds significant importance for several reasons. As this aims to examine the relationship between Unit price, Sales promotion, Shipping cost, Shipping duration, Shipping mode, Customer segment, Product category on Sales growth. This will help to identify the key factors that contribute for development of online shopping in the region and for businessmen this will help to increase the revenue of their business and to have new ideas about the business.

Also, the findings of this study are important to identify the effective strategies that use in online shopping to have more profit and the effective ways to survive in the online market while facing the competition among other people.

Chapter 2

Literature Review

Overview

In this chapter, we mainly focuses on a survey of scholarly sources (such as books, journal articles, and theses) related to our topic and key features.

2.1 Introduction

As the technology has advanced a lot when compared to past few decades, this thing also changed and got personalized it in our daily life. They are also attracted to the online shopping of commodities as they no more have a time following behaviour in perspective due their busy schedule. There have been specifically a significant increase in e-commerce throughout the world with expanding use of internet and smart devices.

Online shopping features has revolutionized retail providing consumers with accesibility, convenience and a huge selection of goods at low prices. There lot of number of people shopping online increasing worldwide, e-commerce businesses are always searching for ways to increase sales performance and remain competitive in their respective markets. The pricing strategy that businesses use, which takes into account elements like unit price, discounts, and delivery costs, is one important component that affects online sales.

2.1.1 Sales

The article Khemvaraporn [2006] by Preedee Khemvaraporn, published in 2006 by Assumption University, wanted to create an online platform that enhance tha sales of motorcycle store. This site helped to get easy access to information about promotions,discounts and speatial offers.

2.1.2 Unit Price

Zeithaml [1988] study shows that consumers can hold those mental constructs about price, quality, and value that can result in the sales performance being affected not only by price but mainly by the fact that customers see the value they get. Sometimes, a high unit price can improve the sales performance as long as customers believe it is of better quality and offers them more value for the money.

2.1.3 Sales Promotion

The Sujata et al. [2016]investigates what impact advertising costs and sales promotion expenses have on sales performance in the Indian telecom industry. According to the main findings, there is a significant predictive relationship between the money spent on advertising and sales promotion and sales performance which means that adequate marketing communication strategies can improve sales volume.

2.1.4 Shopping Cost

Bell et al. [1998] has investigated the how fixed and variable cost impact on consumer buying behavior. They provided insights into how retailers can optimize these factors to improve their sales performance.

2.1.5 Shopping Duration

Wang et al. [2021] explored how sales performance is affected by peak moments of interaction with consumers during live streaming e-commerce (LSE) on Taobao Live. It is important for the duration in which the most interactions occur to be just right. Using video data, this research measures intensity and duration as well as live interaction variables through moving window and computer vision techniques .This study gives important information about the time and change connected with live interactions in e-commerce platforms and how vital they are when it comes to making more sales.

2.1.6 Customer Segment

One of the most importants pointed out by Kotler and Keller in Kotler et al. [2016] was the significant segmentation of the customer in order to drive sales performance. The competence of the marketing team to tailor their strategies in a way that satisfies the needs and wants of the different groups by using the division of the market into smaller segments is one of the ways in which the business can achieve improved sales through a well-structured sales performance model are totally correct.

2.1.7 Product Categories

Wan et al. [2012] studied the impact of product variety on operations and sales performance. For my research, I only get impact of product variety on sales performance. Their research, product variety was found to have both direct and indirect effects on sales. Overall, combining the direct and indirect impacts of product variety on sales, product variety initially raises sales at a diminishing rate when product variety is at a low level.

Chapter 3

Materials and Methods

Overview

This chapter gives a brief introduction about, How we model the research and what techniques we are going to use to get success our research.

3.1 Research approach

The overall research approach for this study is quantitative research approach.

Quantitative research approach

Quantitative research is the process of collecting and analyzing numerical data. It can be used to find patterns and averages, make predictions, test causal relationships, and generalize results to wider populations. The multiple linear regression model is one of the tools of the quantitative approach. It is used to determine a mathematical relationship among several variables. In our case, dependent variable is continuous and we have several independent variables. Therefore, we selected multiple linear regression as our model.

3.2 Conceptual model

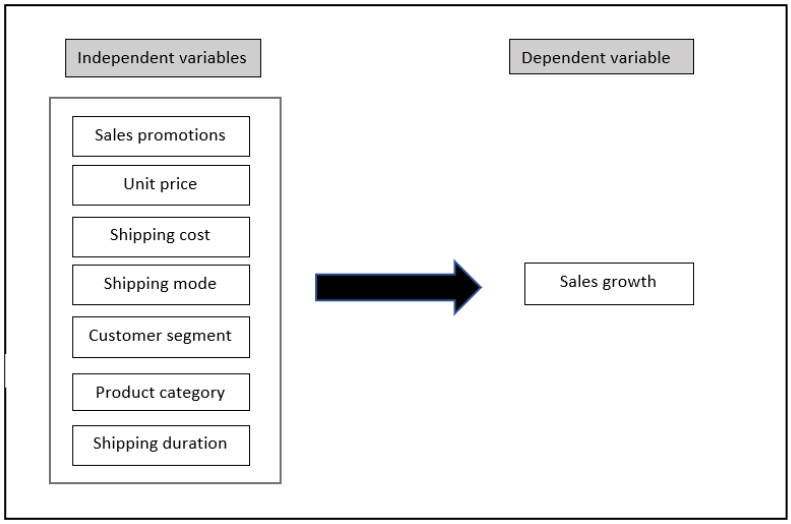

Figure 3.1: *Conceptual model*

3.3 Research design

The primary goal of the study is to use multiple linear regression to assess how Unit price, Sales promotion, Shipping cost, Shipping duration, Shipping mode, Customer segment, Product category impact on Sales growth in Online Shopping. The step wise approach was adopted due to the fact that it allows for the recognition of the most relevant independent variables. After that, using the most relevant independent variables, the best-fitted regression model is constructed for making the conclusions.

3.3.1 Multiple linear regression

Multiple linear regression is a statistical technique used to model the relationship between multiple independent variables and a dependent variable. A population model for a multiple linear regression model is written as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + +\beta_n X_n \tag{3.1}$$

Where,

Y = Dependent variable

X_1, $X_{1,...}$, X_n = Independent variables

β_0 = Intercept of Y

β_1, $\beta_{2,....}$, β_n = Slope of line of X_i

Matrix Notation of the Model

- We can interpret this in matrix notation.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

3.3.2 Selecting the best regression equation

Then, the best regression equation should be selected related to the equation of this linear combination.This selection process aims to reduce the set of predictor variables to find the most important predictor variable or variables while maintaining a good explanation of the data. For this process, there are mainly three methods.

1)All possible regression
This method is also called as stepwise selection method. This is a combination of forward selection method and backward elimination.

2)Backward elimination
This method starts with the equation that includes the all predictor variables and step by step remove the least significant predictor that means the predictor variable with the highest p-value until the subtraction of predictors does not significantly change the model.

3)Forward selection method
This method starts with an empty equation and step by step add the most significant predictor that means the predictor variable with the lowest p-value until the addition of predictors does not significantly improve the model.

Chapter 4

Data

Overview

In this chapter, we concern about the data set what we selected. This will gives brief introduction about data set and special characteristic related to data set.

4.1 About the data set

The dataset is about a broad view into what affects for online shopping behavior. It constructed with immense care to ensure it effectively examines an array of factors that influence customers' purchasing intentions in the increasingly significant realm of digital commerce.

The R code used to read this data is in appendix. 7

We have 8399 observations and 22 variables. The response variable 'Sales growth' is continuous, and the predictors are mixed with numerical and categorical variables.

4.2 Metadata

The source of these data is Kaggle.

(`https://www.kaggle.com/datasets/thedevastator/online-shopping-consumer-behavior`

The data collection took place at 9.50 am on 9th April 2024

4.3 Data Dictionary

Variable Descriptions

1. SALES GROWTH: Revenue generated from goods or services sold of online shopping company.

2. DISCOUNT: Price reduction of a product or service offered to customers.

3. SHIP MODE: Method of transportation used to deliver orders to customers.

   EX: Delivery Truck (1), Express Air (2), Regular Air (3)

4. UNIT PRICE: Cost of a single item or unit of a product of the company.

5. SHIPPING COSTS: Charges incurred by customers for delivering their orders.

6. CUSTOMER SEGMENT: Categorized groups of customers with what products they want to buy.

   EX : Consumer, Corporate, Home Office, Small Business,

7. PRODUCT CATEGORY: Classification of items based on attributes or functions on the online platform.

   EX : Furniture, Office Supplies, Technology

8. SHIPPING DURATION: Difference between ship date and order date 4.4.4

Variable	Type	Missing Data Indicators
SALES GROWTH	Numeric , Continuous	NA
DISCOUNT	Numerical	NA
SHIP MODE	Categorical	NA
UNIT PRICE	Numeric	NA
SHIPPING COSTS	Numeric	NA
CUSTOMER SEGMENT	Categorical	NA
PRODUCT CATEGORY	Categorical	NA
SHIPPING DURATION	Numerical	NA

Table 4.1: *Data Dictionary Table*

Dependent Variable	Independent Variable
SALES GROWTH	DISCOUNT
	SHIP MODE
	UNIT PRICE
	SHIPPING COSTS
	CUSTOMER SEGMENT
	PRODUCT CATEGORY
	SHIPPING DURATION

Table 4.2: *Dependent and Independent variables*

Here, from above variables most of them are continuous . The dependent variable SALES GROWTH, is a measure which represents the total value of all sales within a one purchase .

4.4 Data set preparation

4.4.1 Check null values

We check whether this data set has null value by using below R code,

- is.null(data)

We found that this data set hasn't any null values.

4.4.2 Check NA values

Then we check is this data set has NA values.

- is.na(data)

Here also found that this data set hasn't any NA values.

4.4.3 Remove the variables that do not provide additional information to predict sales growth.

In the data set, some variables not affect to the company sales growth. So we remove the columns called 'ROWID', 'ORDERPRORITY' , 'ORDERQUANTITY' , 'PROFIT', 'CUSTOMERNAME' , 'REGION', 'PRODUCTSUBCATEGORY', 'PRODUCTNAME', 'PRODUCTCONTAINER', 'PRODUCTMARGIN', 'DATASET'.

4.4.4 Create new variable

In the data set there are two variable called SHIPDATE and ORDERDATE. So we create new variable using below R equation and label it as SHIPPING DURATION.After that we plan to check whether SHIPPING DURATION will affect to the sales in a online shipping company.

SHIPPING DURATION <- SHIPDATE – ORDERDATE

4.4.5 Assign values for categorical variable

In the data set there some categorical variable.For the calculation we assign values for the categories.Below table shows the assigned values for categories.4.1

Categorical Variable	Name	Label
SHIPMODE	Delivery Truck	1
	Express Air	2
	Regular Air	3
CUSTOMERSEGMENT	Consumer	1
	Corporate	2
	Home Office	3
	Small Business	4
Product Category	Furniture	1
	Office Supplies	2
	Technology	3

Figure 4.1: *Label for categories in data set*

4.4.6 Check Outliers

We intend to follow some steps for handling outliers. First, we hope to plot boxplot to identify outliers.

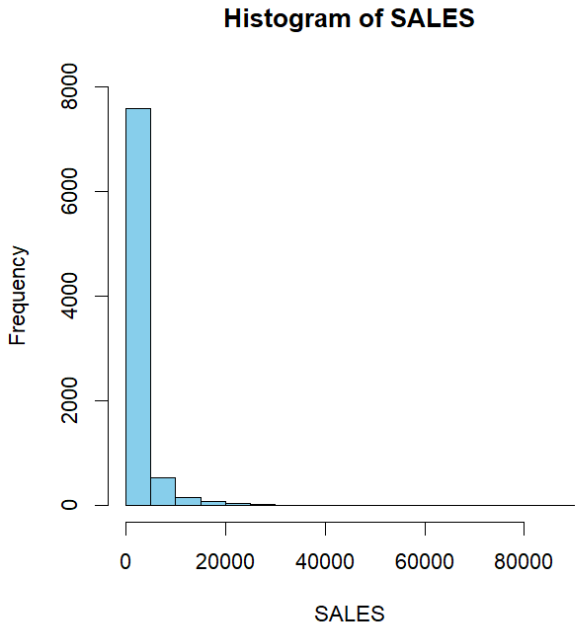

Figure 4.2: *Histogram of sale*

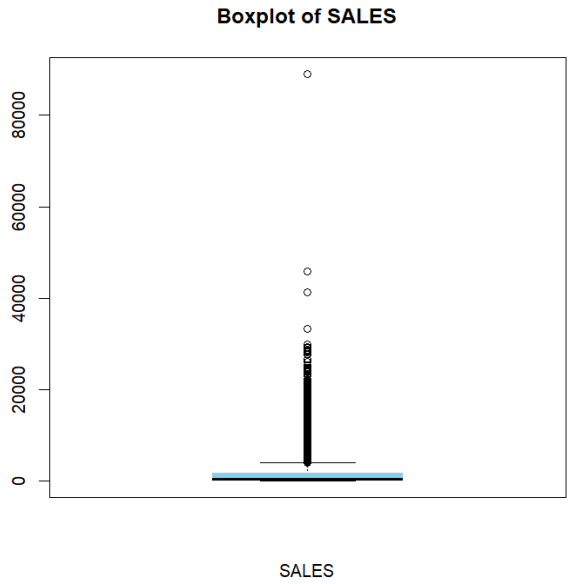

Figure 4.3: *Box plot of sales*

According to the figure 4.3 sales has the outliers. Therefore, we use a transformation for Sales to avoid the outliers. We use logarithm value of sales as the transformation for Sales.

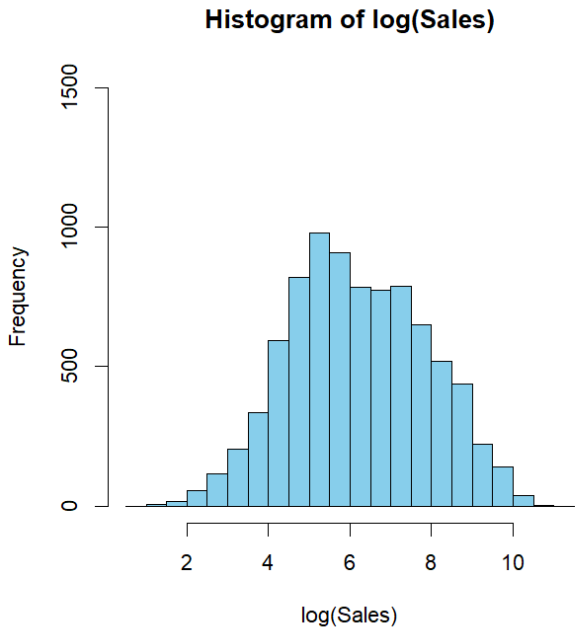

Figure 4.4: *Histogram of log(sales)*

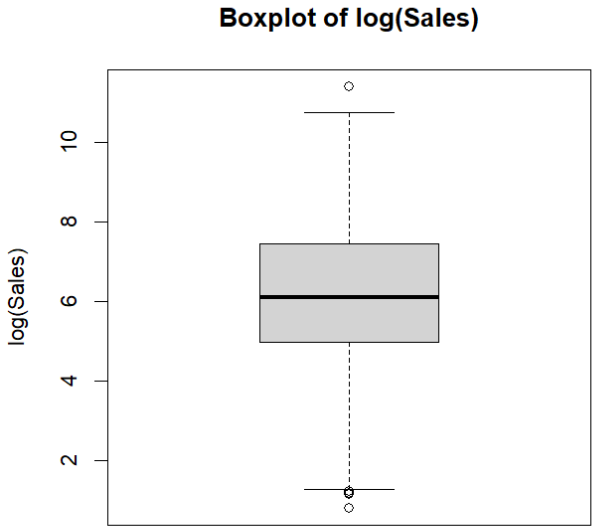

Figure 4.5: *Box plot of log(sales)*

This outliers can be ignored when compared to the previous one.So let's consider there is no any outliers of log sales.

Therefore, we decited to predict result using these data.

Chapter 5

Results

Overview

In this chapter we mainly focused on Exploratory data Analysis and it include tabular summarization (mean,median,mode,variance) and some graphical summarization about the data we selected.

5.1 Exploratory data Analysis

The below table 5.1 illustrates the summary statistics of each variable. According to this table, the total count of observations is 8399 and there are no missing values which is indicated by N*.

```
     SALES           UNIT.PRICE          DISCOUNT        SHIPPING.COSTS    PRODUCT.CATEGORY
 Min.   : 0.8065   Min.   :   0.99   Min.   :0.00000   Min.   :  0.49   Min.   :1.000
 1st Qu.: 4.9642   1st Qu.:   6.48   1st Qu.:0.02000   1st Qu.:  3.30   1st Qu.:2.000
 Median : 6.1080   Median :  20.99   Median :0.05000   Median :  6.07   Median :2.000
 Mean   : 6.1982   Mean   :  89.35   Mean   :0.04967   Mean   : 12.84   Mean   :2.041
 3rd Qu.: 7.4439   3rd Qu.:  85.99   3rd Qu.:0.08000   3rd Qu.: 13.99   3rd Qu.:2.000
 Max.   :11.3971   Max.   :6783.02   Max.   :0.25000   Max.   :164.73   Max.   :3.000
 CUSTOMER.SEGMMENT   SHIP.MODE     Shipping.Duration
 Min.   :1.000     Min.   :1.00   Min.   : 0.000
 1st Qu.:2.000     1st Qu.:2.00   1st Qu.: 1.000
 Median :2.000     Median :3.00   Median : 2.000
 Mean   :2.437     Mean   :2.61   Mean   : 2.033
 3rd Qu.:3.000     3rd Qu.:3.00   3rd Qu.: 2.000
 Max.   :4.000     Max.   :3.00   Max.   :92.000
>
```

Figure 5.1: *Descriptive Statistics Table*

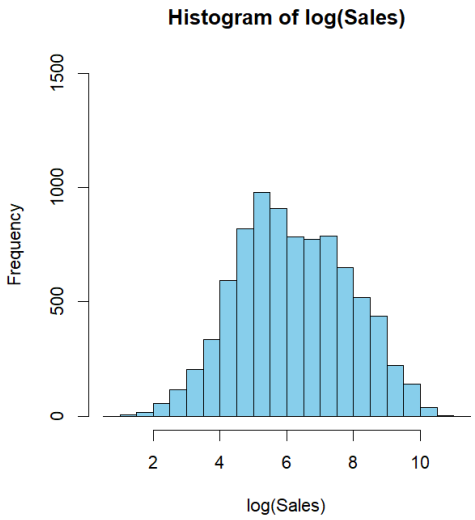

Figure 5.2: *Histogram of log(sales)*

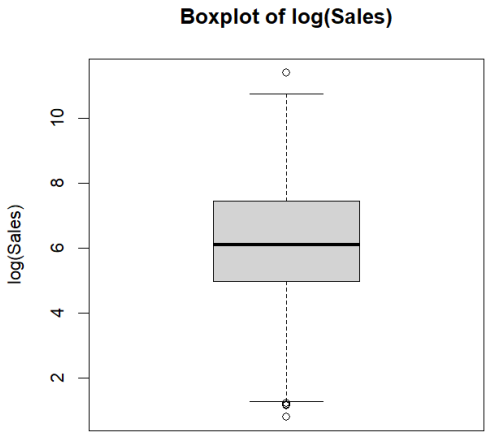

Figure 5.3: *Box plot of log(sales)*

Above histogram illustrates that the SALES 5.2, has the symetric distribution and also according to the box plot 5.3, there are few no.of outliers in SALES.

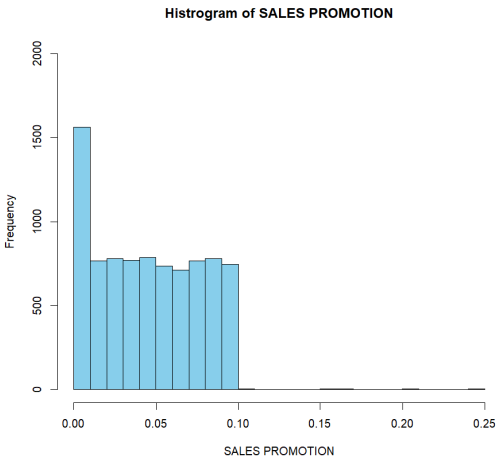

Figure 5.4: *Histogram of Discount*

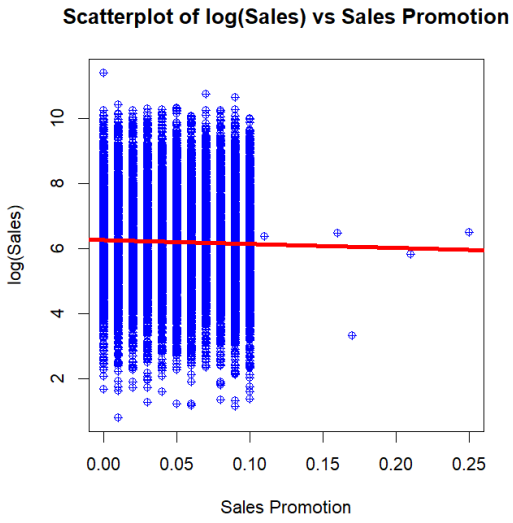

Figure 5.5: *Scatter plot of log(Sales) VS Discount(Sales promotion)*

According to the figure 5.4 and figure 5.5, this data set is symmetric because the majority of data points spread around the mean.
The scatter plot of sales promotion vs sales shows that they have no relationship.

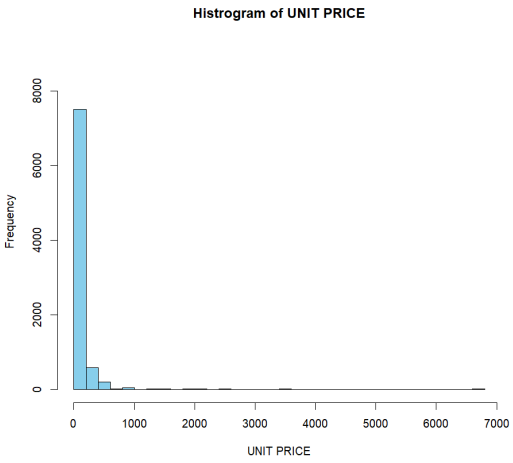

Figure 5.6: *Histogram of Unit Price*

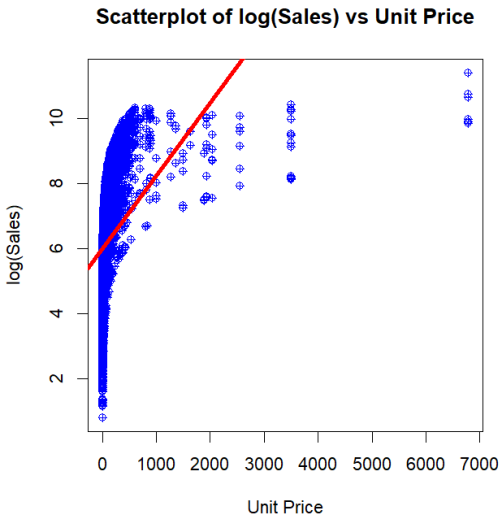

Figure 5.7: *Scatter plot of log(Sales) vs Unit Price*

According to the figure 5.6 and figure 5.7, this data set is right skewed because the majority of data points are in the left to the mean.
The scatter plot of Unit Price vs sales shows that they have strongly positive relationship.

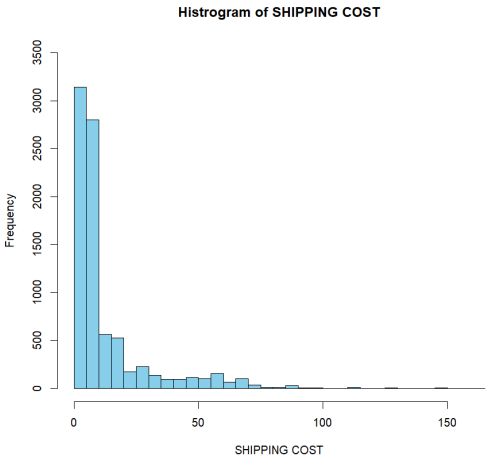

Figure 5.8: *Histogram of Shipping Cost*

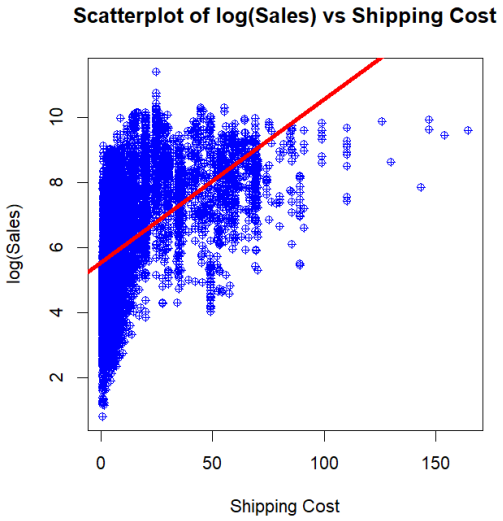

Figure 5.9: *Scatter plot of log(Sales) vs Shipping Cost*

According to the figure 5.8 and figure 5.9, this data set is right skewed because the majority of data points are in the left to the mean.
The scatter plot of Shipping Cost vs Sales shows that they have positive relationship.

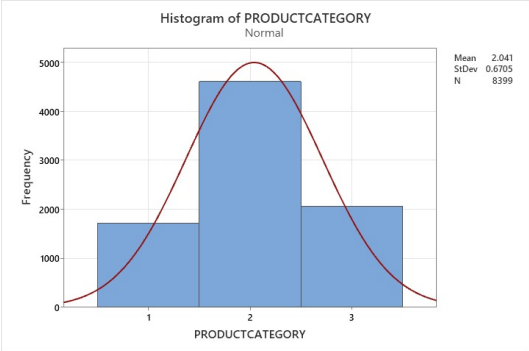

Figure 5.10: *Histogram of Product category*

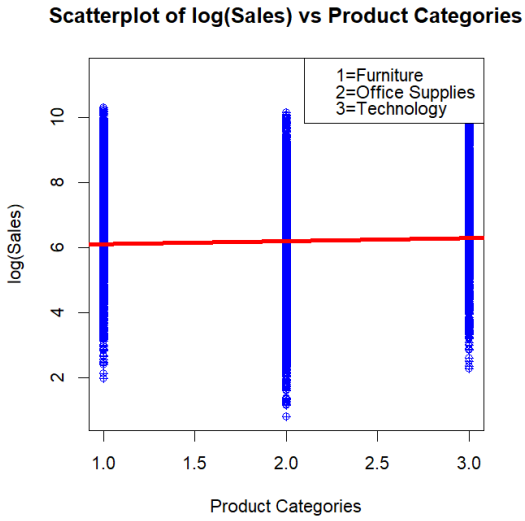

Figure 5.11: *Scatter plot of log(Sales) vs Product Category*

According to the figure 5.10 and figure 5.11, this data set is symmetric because the majority of data points are spread around the mean.The scatter plot shows, there is no any relationship between these 2 variables.

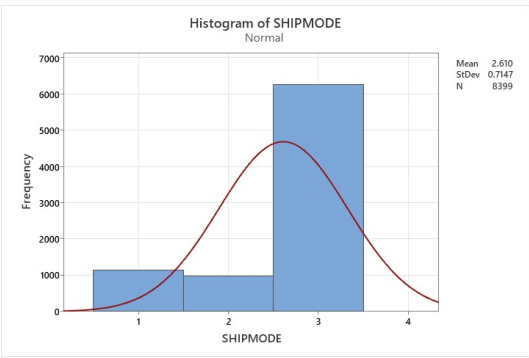

Figure 5.12: *Histogram of Shipping Mode*

Figure 5.13: *Scatter plot of log(Sales) vs Shipping Mode*

According to the figure 5.12 and figure 5.13, this data set is left skewed because the majority of data points are right to the mean.
The scatter plot of Shipping Mode vs Sales shows that they have strong negative relationship.

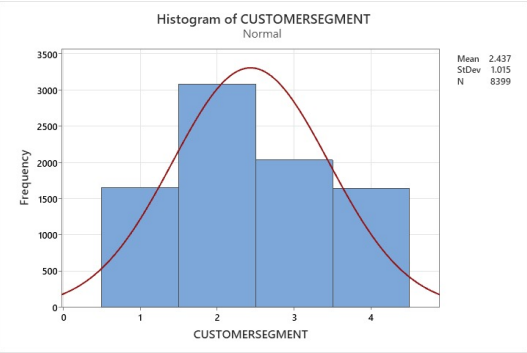

Figure 5.14: *Histogram of Customer Segment*

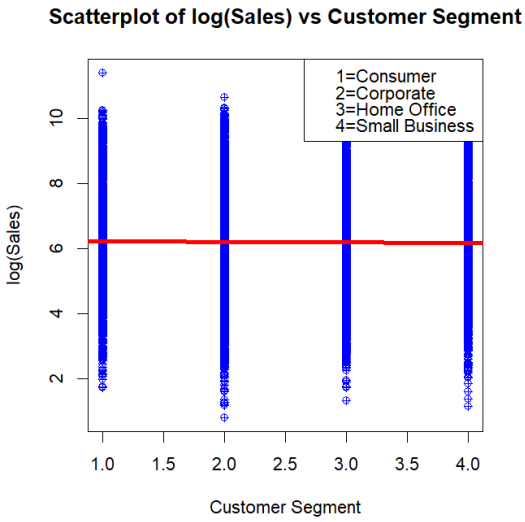

Figure 5.15: *Scatter plot of log(Sales) vs Customer Segment*

According to the figure 5.14 and figure 5.15, this data set is Symmetric.
The scatter plot of Shipping Mode vs Sales shows that they have no relationship.

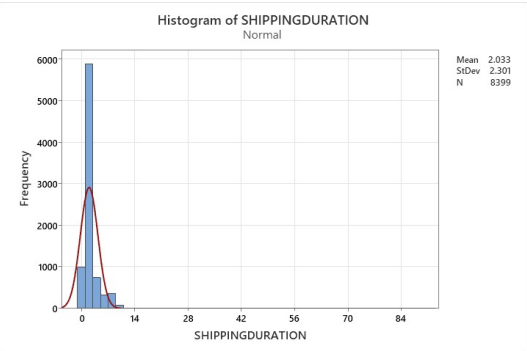

Figure 5.16: *Histogram of Shipping Duration*

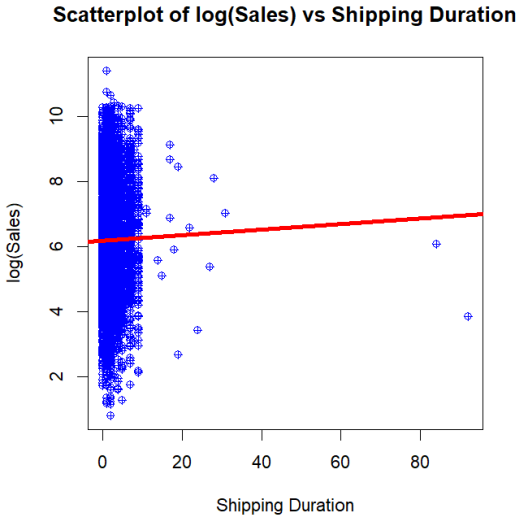

Figure 5.17: *Scatter plot of log(Sales) vs Shipping Duration*

According to the figure 5.16 and figure 5.17, this data set is Symmetric.
The scatter plot of Shipping Mode vs Sales shows that they have week positive relationship.

5.2 Quantitative Analysis

In previous explotary analysis we found out that there are no missing values and null values in this dataset.In this part we are going to estimate model parameters and assess model fit.

5.2.1 Correlation Analysis

	LOGSALES	UNIT.PRICE	DISCOUNT	SHIPPING.COSTS	PRODUCT.CATEGORY	CUSTOMER.SEGMMENT	SHIP.MODE	SHIPPING.DURATION
LOGSALES	1.000000000	0.3880355734	-0.022256795	5.129586e-01	0.033537235	-0.0057648377	-0.416828168	1.170554e-02
UNIT.PRICE	0.388035573	1.0000000000	0.001332397	2.399594e-01	0.082330660	-0.0202654994	-0.211038783	-6.766746e-04
DISCOUNT	-0.022256795	0.0013323969	1.000000000	-1.955711e-03	-0.006964634	-0.0063606055	-0.002859526	-2.809981e-03
SHIPPING.COSTS	0.512958591	0.2399593750	-0.001955711	1.000000e+00	-0.402540443	0.0005012832	-0.673094854	-3.895245e-05
PRODUCT.CATEGORY	0.033537235	0.0823306605	-0.006964634	-4.025404e-01	1.000000000	-0.0010291684	0.313085616	9.852340e-03
CUSTOMER.SEGMMENT	-0.005764838	-0.0202654994	-0.006360605	5.012832e-04	-0.001029168	1.0000000000	0.001783574	-2.080807e-03
SHIP.MODE	-0.416828168	-0.2110387826	-0.002859526	-6.730949e-01	0.313085616	0.0017835740	1.000000000	1.215416e-03
SHIPPING.DURATION	0.011705538	-0.0006766746	-0.002809981	-3.895245e-05	0.009852340	-0.0020808074	0.001215416	1.000000e+00

Figure 5.18: *Correlation Table*

The figure 5.18 illustrates the correlation table of the variables. According to this figure, the highest correlation is between unit price and log(sales) which in scaled as 0.388. The lowest correlation is between ship mode and shipping cost growth which is scaled as -0.673.According to this correlation matrix we can see that there is no relationship between two independent variables.

5.2.2 Assumptions for multiple linear regression

Before start the model fitting part we check whether our data set satisfies multiple linear regression assumptions. Those assumptions are:

1.The residual values are normally distributed.

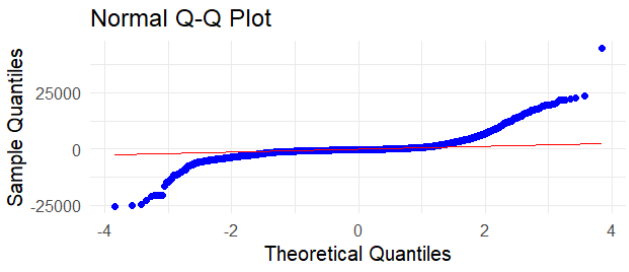

Figure 5.19: *Normal probability plot with outlier*

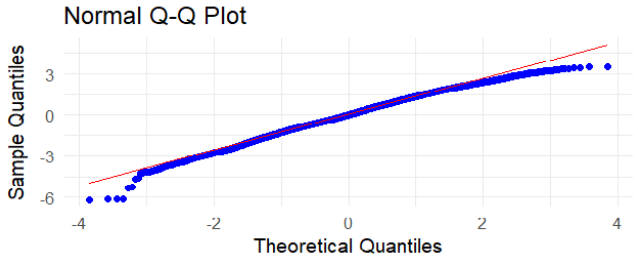

Figure 5.20: *Normal probability plot without outlier*

After observing figure 5.19 and figure 5.20 we can clearly see figure 5.20 is normally distributed than the figure 5.19.Therefore, residuals are normally distributed.

2.A linear relationship between the dependent and the independent variables.

In exploratory data analysis part, we discuss there is a linear relationship between the dependent and the independent variables by using scatter plots.

3.Multicollinearity - To check, the independent variables are not highly correlated with each other.

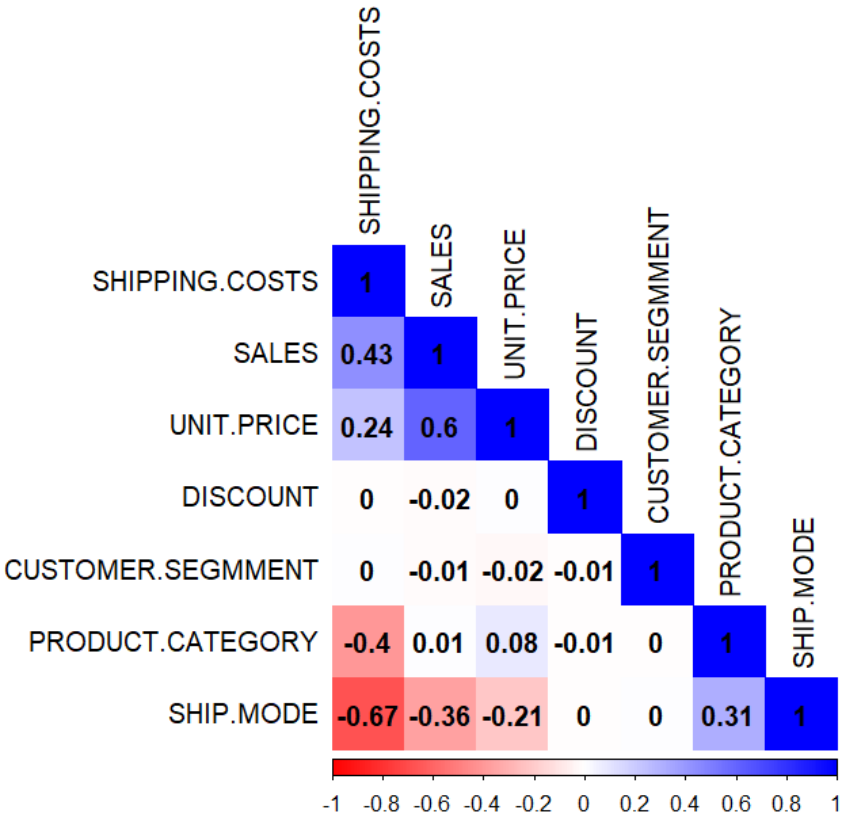

Figure 5.21: *Correlogram*

This called correlation heat map. The figure 5.21 shows, that correlation between each and every two independent variable is less than 0.8.

4.Homoscedasticity - To check, the variance of the residuals errors is similar across the value of each independent variable.

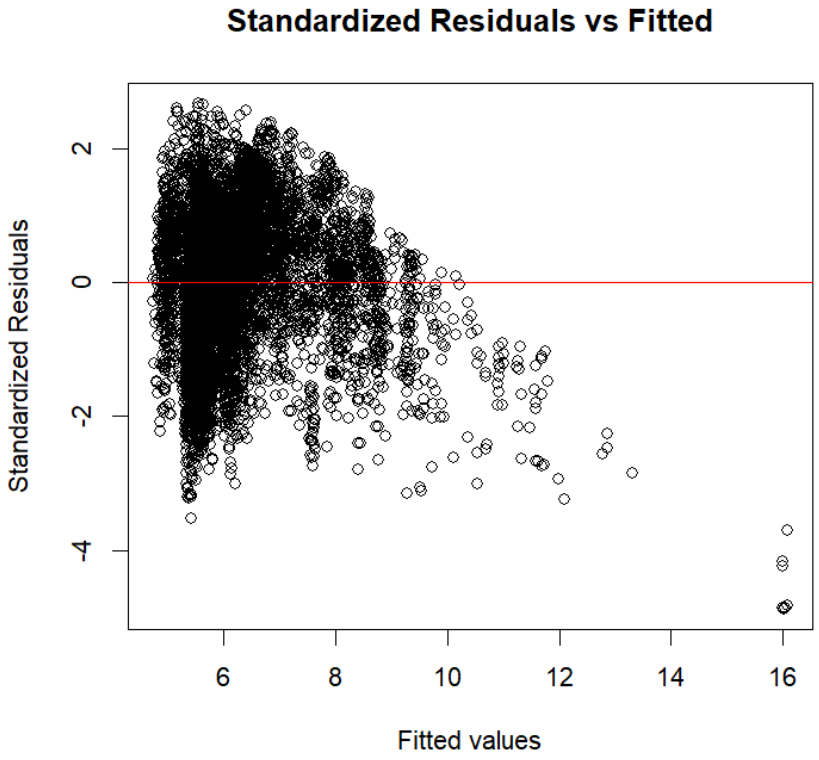

Figure 5.22: *Standardized Residuals vs Fitted*

The residuals have constant variance at every level of the predictor.

We prove these four assumptions for our data set.Therefore, we can start the model fitting part.

5.2.3 Estimate model parameters

Full model

Here we write the regression equation for all the variables.

y = logarithm value of Sales = log(SALES)

```
Call:
lm(formula = y ~ UNITPRICE + DISCOUNT + ship_mode_numeric + product_category_numeric +
    SHIPPINGCOSTS + customer_segment_numeric + SHIPPINGDURATION,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1673 -0.8347  0.0055  0.9294  3.5106

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.082e+00  1.042e-01  48.756   <2e-16 ***
UNITPRICE                 1.319e-03  5.198e-05  25.380   <2e-16 ***
DISCOUNT                 -1.071e+00  4.494e-01  -2.384   0.0172 *
ship_mode_numeric        -3.035e-01  2.720e-02 -11.159   <2e-16 ***
product_category_numeric  6.111e-01  2.386e-02  25.612   <2e-16 ***
SHIPPINGCOSTS             4.567e-02  1.186e-03  38.527   <2e-16 ***
customer_segment_numeric -1.665e-03  1.409e-02  -0.118   0.9059
SHIPPINGDURATION          6.986e-03  6.213e-03   1.124   0.2609
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.31 on 8391 degrees of freedom
Multiple R-squared:  0.3919,    Adjusted R-squared:  0.3914
F-statistic: 772.5 on 7 and 8391 DF,  p-value: < 2.2e-16
```

Figure 5.23: *Full model summary*

Full model:

$$log(SALES) = 5.082 + 0.001319(UNITPRICE) - 1.071(DISCOUNT) - 0.3035(SHIPMODE) + 0.6111(PRODUCTCATEGORY) + 0.04567(SHIPPINGCOST) - 0.001665(CUSTOMERSEGMENT) + 0.00698(SHIPPINGDURATION)$$

Backward elimination

Backward elimination is a stepwise variable selection method used in regression analysis to identify the most relevant independent variables that should be included in a regression model. Backward eliminationis frequently used in multiple linear regression. Main steps are,

• In this procedure the complete regression equation is determined containing all the variables .

• Then variables are checked one at a time and the least significant is dropped from the model at each stage.

• The procedure is terminated when all of the variables remaining in the equation provide a significant contribution to the prediction of the dependent variable Y.

```
Start:  AIC=4548.6
y ~ UNITPRICE + DISCOUNT + ship_mode_numeric + product_category_numeric +
    SHIPPINGCOSTS + customer_segment_numeric + SHIPPINGDURATION

                           Df Sum of Sq   RSS    AIC
- customer_segment_numeric  1      0.02 14408 4546.6
- SHIPPINGDURATION          1      2.17 14410 4547.9
<none>                                 14408 4548.6
- DISCOUNT                  1      9.75 14418 4552.3
- ship_mode_numeric         1    213.80 14622 4670.3
- UNITPRICE                 1   1105.99 15514 5167.8
- product_category_numeric  1   1126.38 15534 5178.8
- SHIPPINGCOSTS             1   2548.71 16957 5914.6

Step:  AIC=4546.62
y ~ UNITPRICE + DISCOUNT + ship_mode_numeric + product_category_numeric +
    SHIPPINGCOSTS + SHIPPINGDURATION

                           Df Sum of Sq   RSS    AIC
- SHIPPINGDURATION          1      2.17 14410 4545.9
<none>                                 14408 4546.6
- DISCOUNT                  1      9.75 14418 4550.3
- ship_mode_numeric         1    213.81 14622 4668.3
- UNITPRICE                 1   1106.70 15515 5166.2
- product_category_numeric  1   1126.36 15534 5176.8
- SHIPPINGCOSTS             1   2548.71 16957 5912.7

Step:  AIC=4545.88
y ~ UNITPRICE + DISCOUNT + ship_mode_numeric + product_category_numeric +
    SHIPPINGCOSTS

                           Df Sum of Sq   RSS    AIC
<none>                                 14410 4545.9
- DISCOUNT                  1      9.77 14420 4549.6
- ship_mode_numeric         1    213.78 14624 4667.6
- UNITPRICE                 1   1106.43 15516 5165.2
- product_category_numeric  1   1127.58 15538 5176.7
- SHIPPINGCOSTS             1   2549.41 16959 5912.1
```

Figure 5.24: *Backward Elimination Steps*

```
Call:
lm(formula = y ~ UNITPRICE + DISCOUNT + ship_mode_numeric + product_category_numeric +
    SHIPPINGCOSTS, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1652 -0.8384  0.0078  0.9269  3.5097

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.092e+00  9.769e-02  52.118   <2e-16 ***
UNITPRICE                1.319e-03  5.197e-05  25.386   <2e-16 ***
DISCOUNT                -1.072e+00  4.493e-01  -2.386   0.0171 *
ship_mode_numeric       -3.035e-01  2.720e-02 -11.159   <2e-16 ***
product_category_numeric  6.114e-01  2.386e-02  25.627   <2e-16 ***
SHIPPINGCOSTS             4.568e-02  1.185e-03  38.534   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.31 on 8393 degrees of freedom
Multiple R-squared:  0.3918,    Adjusted R-squared:  0.3914
F-statistic:  1081 on 5 and 8393 DF,  p-value: < 2.2e-16
```

Figure 5.25: *Backward Elimination Summary*

For this study, the best regression equation:

$$\textit{log(SALES)= 5.092 + 0.001319(UNITPRICE) - 1.072(DISCOUNT) - 0.3035(SHIPMODE) + 0.6114(PRODUCTCATEGORY) + 0.04568(SHIPPINGCOST)}$$

R squared for full model = 0.3919

R squared for reduced model = 0.3918

5.2.4 Assess model fit

Here are the regression equations of two models

Full model:

$$\text{log(SALES)= 5.082 + 0.001319(UNITPRICE) - 1.071(DISCOUNT) - 0.3035(SHIPMODE) + 0.6111(PRODUCTCATEGORY) + 0.04567(SHIPPINGCOST) - 0.001665(CUSTOMERSEGMENT) + 0.00698(SHIPPINGDURATION)}$$

Reduced model:

$$\text{log(SALES)= 5.092 + 0.001319(UNITPRICE) - 1.072(DISCOUNT) - 0.3035(SHIPMODE) + 0.6114(PRODUCTCATEGORY) + 0.04568(SHIPPINGCOST)}$$

Analysis of Variance Table for full model and reduced model(ANOVA)

```
Analysis of Variance Table

Response: y
                          Df  Sum Sq Mean Sq   F value    Pr(>F)
UNITPRICE                  1  3567.5  3567.5 2077.6657 < 2.2e-16 ***
DISCOUNT                   1    12.3    12.3    7.1566  0.007483 **
ship_mode_numeric          1  2782.8  2782.8 1620.6976 < 2.2e-16 ***
product_category_numeric   1   370.8   370.8  215.9711 < 2.2e-16 ***
SHIPPINGCOSTS              1  2549.4  2549.4 1484.7525 < 2.2e-16 ***
customer_segment_numeric   1     0.0     0.0    0.0145  0.904022
SHIPPINGDURATION           1     2.2     2.2    1.2643  0.260877
Residuals               8391 14407.8     1.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.26: *ANOVA table of full model*

```
Analysis of Variance Table

Response: y
                          Df  Sum Sq Mean Sq   F value    Pr(>F)
UNITPRICE                  1  3567.5  3567.5 2077.8442 < 2.2e-16 ***
DISCOUNT                   1    12.3    12.3    7.1572  0.007481 **
ship_mode_numeric          1  2782.8  2782.8 1620.8369 < 2.2e-16 ***
product_category_numeric   1   370.8   370.8  215.9897 < 2.2e-16 ***
SHIPPINGCOSTS              1  2549.4  2549.4 1484.8801 < 2.2e-16 ***
Residuals               8393 14410.0     1.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.27: *ANOVA table of reduced model*

Hypothesis 1

According to the full model,

• Null hypothesis (H0): There is no linear relationship between dependent variable and independent variables of the online shopping company.

• Alternative hypothesis (Ha): There is a linear relationship between dependent variable and independent variables of the online shopping company.

1) H0:$\beta(1) = 0$ *against* $\beta(1) \neq 0$

The p-value of unit price is 0.000 and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected.This suggests that there is a linear relationship between log(Sales) and unit price.

$\beta(1) = 0.001319$

2) H0:$\beta(2) = 0$ *against* $\beta(2) \neq 0$

The p-value of discount is 0.007 and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected.This suggests that there is a linear relationship between log(Sales) and discount.

$\beta(2) = -1.071$

3) H0:$\beta(3) = 0$ *against* $\beta(3) \neq 0$

The p-value of ship mode is 0.000 and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected.This suggests that there is a linear relationship between log(Sales) and ship mode.

$\beta(3) = -0.3035$

4) H0:$\beta(4) = 0$ *against* $\beta(4) \neq 0$

The p-value of product category is 0.000 and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected.This suggests that there is a linear relationship between log(Sales) and product category.

$\beta(4) = -0.6111$

5) H0:$\beta(5) = 0$ *against* $\beta(5) \neq 0$

The p-value of shipping cost is 0.000 and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected.This suggests that there is a linear relationship between log(Sales) and shipping cost.

$\beta(5) = 0.04567$

6) H0:$\beta(6) = 0$ *against* $\beta(6) \neq 0$

The p-value of customer segment is 0.904 and it is greater than 0.05. That means the p-value is not in the rejection region. So, the null hypothesis is accepted.This suggests that there is not a linear relationship between log(Sales) and customer segment.

$\beta(6) = 0$

7) H0:$\beta(7) = 0$ *against* $\beta(7) \neq 0$

The p-value of shipping duration is 0.2609 and it is greater than 0.05. That means the p-value is not in the rejection region. So, the null hypothesis is accepted.This suggests that there is not a linear relationship between log(Sales) and shipping duration.

$\beta(7) = 0$

Hypothesis 2

This hypothesis is named as partial F test. In this test, the reduced model is compared with the full model.

- Null hypothesis (H0): Reduced model is suitable
- Alternative hypothesis (Ha): Full model is needed

$$F_{(\text{partial})} = \frac{SS_{\text{Reg}}(\text{FULL}) - SS_{\text{Reg}}(\text{REDUCED})}{MS_{\text{Error}}(\text{FULL})} \tag{5.1}$$

According to this equation,

$$F_{(\text{partial})} = [(3567.5+12.3+2782.8+370.8+2549.4+2.2)-(3567.5+12.3+2782.8+370.8+2549.4)]/1.7 \tag{5.2}$$

$$F_{(\text{partial})} = [9285 - 9282.8]/1.7 \tag{5.3}$$

$$F_{(\text{partial})} = 1.2941 \tag{5.4}$$

Test statistic = 1.2941

Now, this value is compared with test statistic with k + 1 - p and n - (k + 1) degrees of freedom at the 0.05 significance level.

n= 8399 , k = 7 , p = 6

$$F_{(\text{table},0.05,2,8391)} = 2.996802 \tag{5.5}$$

F table value = 2.996802

That means, the test statistic < F table value which shows that it is not in the rejection region. Therefore, the null hypothesis is not rejected. That means the reduced model is suitable for this study.

Reduced model:

log(SALES)= 5.092 + 0.001319(UNITPRICE) - 1.072(DISCOUNT) - 0.3035(SHIPMODE) + 0.6114(PRODUCTCATEGORY) + 0.04568(SHIPPINGCOST)

Chapter 6

Discussion and conclusion

6.1 Discussion

The objective of conducting this study is to identify Impact of Sales Discounts, Shipping Modes and Customer Variables on Sales Performance in North America within 2009 to 2012. The dataset includes 8399 observations without any missing values and null values.

When we plot the box plot for dependent variable,it has outliers and not normally distributed. After that,we check the outliers in our data set using IQR method and create a without outlier data set. And plot the box plot. It also has the significance number of outliers. So, we transform the dependent variable to log values. Log values of sales has a few no.of outliers.Then we select this as our data set.

The distribution of log(Sales),discount,product category,customer segment and shipping duration show symmetric while ship mode and shipping cost appear Left and right skewness respectively.The scatter plots illustrate the graphical interpretation of the relationships between the independent variables and the log(Sales). Before start the model fitting part we check whether our data set satisfies multiple linear regression assumptions.The data set with log(Sales) satisfied all 4 assumptions. To get the best fitted model, we apply R Studio. For this study, the regression equation of full model is,

$$\log(\text{SALES}) = 5.082 + 0.001319(\text{UNITPRICE}) - 1.071(\text{DISCOUNT}) - 0.3035(\text{SHIPMODE}) + 0.6111(\text{PRODUCTCATEGORY}) + 0.04567(\text{SHIPPINGCOST}) - 0.001665(\text{CUSTOMERSEGMENT}) + 0.00698(\text{SHIPPINGDURATION})$$

Full model R square value is 39.19 percent which does not give any conception about data set. Therefore we obtain the best regression equation,we use backward elimination method which is used and the obtained reduced model equation is,

$$\log(\text{SALES}) = 5.092 + 0.001319(\text{UNITPRICE}) - 1.072(\text{DISCOUNT}) - 0.3035(\text{SHIPMODE}) + 0.6114(\text{PRODUCTCATEGORY}) + 0.04568(\text{SHIPPINGCOST})$$

Reduced model R square value is 39.18 percent which also does not give any conception about data set. In regression analysis, hypothesis testing is used to determine linear relationship between independent variables and dependent variable in hypothesis 1. And hypothesis 2, we check what is the good model for our study.According to the results obtained by the hypothesis 2, called as partial F test, the reduced model is more suitable than the full model of this study

6.2 Conclusion

According to the findings, I decided that our model does not significantly correspond to the data and is not appropriate for creating the type of estimations we had hoped for based on the results. Since we have chosen a real-world problem with real-world data collection, we are unable to forecast a perfect solution as theoretical results.

As a conclusion, our R-squared value of 39.18 percent indicates that this data collection fails to create a well-fitted model with a high R-squared value. Therefore, this dataset is not suitable for predicting future decisions or providing reliable insights for sales growth strategies.

Future research should focus on selecting more appropriate datasets and considering additional variables to better capture the complexities of sales growth in online shopping contexts.

Chapter 7

Appendix

To access the data set:

- https://docs.google.com/spreadsheets/d/1ySDSWUgycgAXE4aIMm6fZ2UZqUF2d0PmQgf4st
  edit?usp=sharing

To read the data set using R software:

- data <- read.csv("online shopping.csv")

To read and analyzing data set using R software:

- https://docs.google.com/document/d/1eXnu_Vh_zH-mohHK0cVDQpFlKsSOJmW2hqMr8Uw2pO
  edit

Bibliography

D. R. Bell, T.-H. Ho, and C. S. Tang. Determining where to shop: Fixed and variable costs of shopping. *Journal of marketing Research*, 35(3):352–369, 1998.

P. Khemvaraporn. Sales promotion website for motorcycle store. 2006.

P. Kotler, K. L. Keller, M. Brady, M. Goodman, and T. Hansen. *Marketing Management 3rd edn PDF eBook*. Pearson Higher Ed, 2016.

J. Sujata, P. Sandeep, and C. Abhijit. Impact of advertising and sales promotion expenses on the sales performance of indian telecommunication companies. *Indian Journal of Science and Technology*, 9(46), 2016.

X. Wan, P. T. Evers, and M. E. Dresner. Too much of a good thing: The impact of product variety on operations and sales performance. *Journal of Operations Management*, 30(4):316–324, 2012.

Y. Wang, J. Liu, and Y. Fang. Live streaming e-commerce: The impact of the intensity, duration, and phases of peak interaction on sales performance. In *14th China Summer Workshop on Information Management (CISWIM)*, volume 6, 2021.

V. A. Zeithaml. Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *Journal of marketing*, 52(3):2–22, 1988.