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

B.L.A.W.Lakshika

Supervisor :

Ms. K.C.N. Shanthidevi

(B.Sc., M.Sc.(Japan))

Department of Mathematics

University of Ruhuna

Matara

2024

Declaration

I, B.L.A.W.Lakshika, 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. Shanthidev, 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 by me to any other institution or academic program for any other purpose.

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

Supervisor's Recommendation

I certify that this study was carried out by B.L.A.W. Lakshika under my supervision.

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

Department of Mathematics
Faculty of Science
University of Ruhuna

Acknowledgement

First of all, my sincere gratitude goes to our supervisor Ms. K.C.N. Shanthidevi for her unending guidance and support for us to success this research.

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

And also, I extend my thanks to all the supervisors in the supervisory board for their great support towards me.

Moreover, I extend my thanks to all the instructors who helped me to make this research success.

Also, I express my sincere appreciation to all our batch mates for sharing their knowledge with me.

Finally, I'm thankful to everyone who played a part, big or small, to make this research a success.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

NA	Not Available

Abstract

This study investigates impact of unit price, sales promotion, shipping costs, shipping mode, customer segment, product category and shipping duration on online shopping companies' sales growth. Here, we are going to analyze this relationship using regression approach in multiple linear regression model. The objective of this study is to examine how these factors affects to the sales growth of the online shopping company. So, this will contribute the valuable insights to the what features are most important to get many sales for the online shopping company.

Key words: Sales growth, Sales Promotions, Shipping cost, Unit price, shipping Mode, Customer Segment, Product Category, Shipping Duration, North America

Chapter 1

Introduction

Overview

This chapter is organized as follows. First focuses on giving an understanding of the online shopping and key steps used in the process of developing the model. In later research problem and objectives of this research will be defined.

1.1 Background of the study

Over the last few years, e-commerce has become an indispensable part of global retail. Like many other industries, buying and selling goods has undergone a substantial transformation following the advent of the internet.Users has rapidly increased due to the recent Covid-19 pandemic. Nowadays, It has no age limit to use.

Online shopping is a form of e-commerce which allows consumers to buy goods or services from a seller over the Internet using a web browser or a mobile app whitout visiting the live in sellers place. When a consumer search a product they are interested in, they can use a shopping search engine to look up other vendors product or visit directly to the retailer's website. And also using that, they can know if product is available for sale and Various prices at which vendors sell. So,consumers all over the world now profit from the perks of online transactions.

As the revenues from online sales continued to grow significantly people started to do research about this field. This study aims to study about 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 within 2009 to 2012 through the data collected by Kaggle website.

1.2 Problem statement

online shopping is the hot topic that hovering these days. So It's better to be aware, what are the factors that can improve sales capacity as customer is not contacting the seller directly. In this project we focus on 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. Consequently, in order to obtain important insights into boosting long-term economic stability and tendency, it is necessary to investigate and evaluate the connection between these characteristics and sales growth.

1.3 Objective of the study

In this study, we mainly focus on to get the clear idea about what factors that affect to the sales growth and examine that what factors need to be improved to have better sales growth while what are the things that need to be changed to get better results.

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. At the end of the study, we are going to have the ability to predict the sales growth when the other relative factor's details are given.

1.3.1 Research objectives

we are planning to achieve our goal using the multiple regression model and then the step wise method to select the best model. It will be discussed elaborately in the data section. The primary objective of this study are :

- 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.

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):

  Reduce model is suitable.

- Alternative hypothesis (Ha):

  Full model is needed.

1.6 Significance of the study

The most significant of this research is to analyze factors like price, promotions, shipping details, and customer demographics, it will pinpoint the key ingredients for online success in this region.

From this study ,Businesses can leverage these insights to skyrocket their revenue and explore fresh strategies. Plus, the study reveals effective tactics to thrive in the competitive online marketplace.

In the other hand we can study online shopping behaviors of people in a developed continent.

Chapter 2

Literature Review

Overview

In this chapter, we delve into relevant academic literature, such as books, articles, and dissertations, to establish the existing knowledge base on our topic and its key characteristics.

2.1 Introduction

In the rapidly developing world, people want to do everything easily and in less time without any effort. So people use buy or sell things in a digital marketplace.Online shopping is the one of e-commerce system among them.

Unlike retail service there are some factors that affect to the sustainability and future development of the online shopping system. So researchers did there research to find out what are them and how they will affect to the online shopping system. In here we selected some factors from those researches for our study.

2.1.1 Sales

Researchers highlighted the significant relationship and direct impact between online shopping and sales growth.Alzoubi et al. [2022] Research shows that specialized sales strategies, such email marketing campaigns and product suggestions, can increase sales growth and conversion rates. Using both online and physical channels, a multi-channel sales strategy can increase sales volume and improve customer engagement.But Ranganathan and Grandon [2002] highlighted that Online merchants face a number of problems in improving online sales. There is little understanding of the factors affecting online sales.Verhoef et al. [2017] have demonstrated that the utilization of data analytics and artificial intelligence in sales forecasting and optimization has the potential to enhance sales performance and profitability.

2.1.2 Unit Price

Product unit prices have a considerable effect on the sales of online stores. Roth et al. [2017] Studies show that the availability of unit pricing has a significant impact on a number of shop price image aspects.And it tells that there is a positive influence of unit price presence and unit price prominence on the consumers' intention to shop at a online store through the store price image. According to research by Hamby et al. [2018], customers are more willing to buy goods with lower unit pricing, particularly when comparing comparable products offered by several online sellers. Likewise, a Johnson et al. [2020] study found that giving discounts on large purchases or establishing competitive unit prices are examples of strategic pricing tactics that can enhance the number of online sales.

2.1.3 Sales Promotion

Sales promotion is a type of promotion, and promotion is one of the primary components of the "marketing mix." SETIAWAN [2021]. This is the short term increase in the value of the product or service motivate the consumer Fitri [2018].This goal is to have an immediate and direct impact on the purchasing behavior of customers Arsta and Respati [2021]. There are two types of sales promotions. They are consumer-oriented and trade-oriented promotions. Samples, coupons, premiums, contests and sweepstakes, refunds and rebates, bonus packs, price reductions, loyalty programs, and event marketing are examples of consumer oriented sales promotion activities. Dealer contests and incentives, trade allowances, point-of-purchase displays, sales training programs, trade fairs, cooperative advertising, and other activities are examples for trade-oriented promotions Belch and Belch [2018])

2.1.4 Shipping Cost

According to Chen and Ngwe [2018], Shipping fees are an important aspect of online retail for both consumers and sellers. And also they find that contingent free shipping shifts demand to more popular products, and that the effects of category-level price changes on profits depends on the active shipping policy. High shipping prices have been shown in studies by Manapul et al. [2022] to discourage customers from finishing their orders, which can result in abandoned carts and lower conversion rates.

2.1.5 Shipping Duration

The no of days that going to receive the goods ordered is an important factor that impacts online shops' sales, even though it isn't indicated in the terms used. Ma [2016] research suggests that faster shipping options could improve customer satisfaction and increase conversion rates. Wang et al. [2019] Studies examine how shipping speed and clarity of delivery information affect customer satisfaction with online retailers. Quick delivery services can improve customers' shopping experiences and increase revenue; they are often appreciated by customers.

2.1.6 Customer Segment

The importance of customer segmentation in optimizing sales performance through customized discounts and promotions was highlighted in a 2019 study by Wang et al. [2019]. Through the examination of client factors like preferences, purchasing patterns, and demographics, companies can develop marketing strategies that are specifically tailored to appeal to particular customer categories. For example, offering Eco-friendly product discounts to client segments that care about the environment might improve sales performance in sustainability-related product categories.

2.1.7 Product Categories

Ravula [2023] conducted an analysis which demonstrated the influence of shipping options on sales performance in several product categories. They discovered that consumer preferences for delivery alternatives were influenced by variables such as product size, value, and perishability. Offering free delivery on expensive electronics, for instance, could increase sales, while offering expedited shipping for presents or other things that must be delivered quickly could increase customer satisfaction and encourage return business.

Chapter 3

Materials and Methods

Overview

This chapter gives a brief introduction about, How we model the research and what techniques or we are going to use to for this research.

3.1 Research approach

The overall approach used to conduct the study is quantitative research approach. The use of quantitative methods allows researchers to collect and analyze numerical data related.

3.1.1 Quantitative research approach

Quantitative research is a systematic approach to investigating events by collecting and analyzing numerical data. It focuses on measuring variables and testing hypotheses to understand relationships or explain Causality. 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 to solve our model and will used to detect the relationship patterns.

3.2 Conceptual model

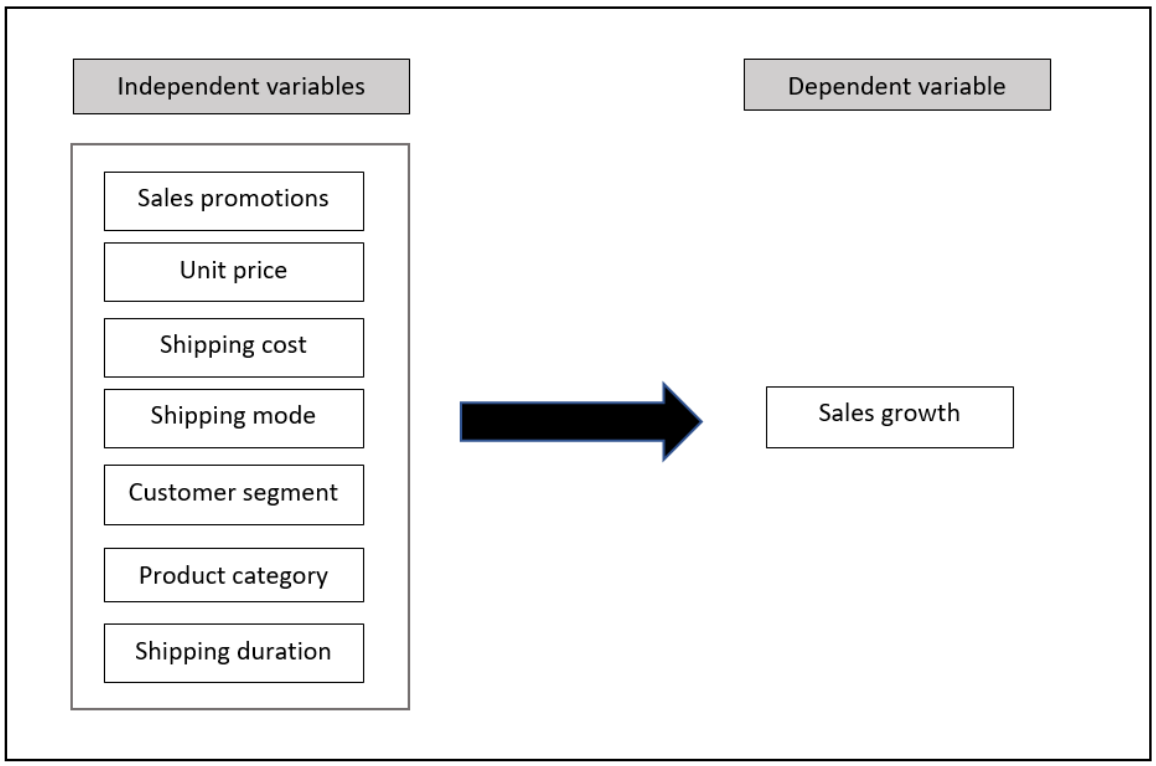

Figure 3.1: *Conceptual model*

3.3 Research design

This is a quantitative case study which adopts multiple linear regression. In this study, we get the Sales growth as a linear combination of multiple independent variables, including Sales promotion, Unit price, Shipping cost, Customer segment, Product category and Shipping duration.

As a quantitative research approach, multiple linear regression involves the use of numerical data, making it suitable for analyzing large dataset.

3.3.1 Ordinary Least Square Method

This is a process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve. During the process of finding the relation between Dependent and Independent variable, the trend of outcomes are estimated quantitatively.

Fitted line model :

$$\hat{y_i} = \beta_0 + \beta_1 x_i : i = 1, 2, ..., n \tag{3.1}$$

$y_i = ObservedValue$
$\hat{y}_i = FittedValue$
$\beta_0 = Intercept$
$\beta_1 = Slope$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \bar{y} + \hat{\beta}_0\ \bar{x}$$

$\hat{\beta}_0$=Estimated Intercept
$\hat{\beta}_1$ =Estimated Slope

3.3.2 Multiple linear regression

Multiple linear regression is a statistical technique that allows researchers to examine the simultaneous effects of independent variables on the dependent variable, while controlling for potential confounding factors. A population model for a multiple linear regression model is written as,

$$\mathbf{y} = \beta_0 + \beta_1 X_1 + \beta_2 X2 + ... + \beta_k X_k + \epsilon$$

is called a multiple linear regression model with k regressors.
Where,

y = Dependent variable

β_0 = Intercept of y (value of y when all other parameters are set to 0)

$X_1, X_2, X_3,$ = Independent variable

ϵ = Error term

$\beta_1, \beta_2, \beta_3,$ = Slop of line Xi

And also we can write it in matrix notation,

.

Matrix Notation of the Model

- In matrix notation, the model is given by:

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

where

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

Then, the most suitable 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, we can use three methods.

- Forward selection method.

  It begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model. Once the model no longer improves with adding more variables, the process stops.

- Backward selection method.

  The backward elimination process begins by fitting a multiple linear regression model with all the independent variables. The variable with the highest p-value is removed from the model, and a new model fits. This process is repeated until all variables in the model have a p-value below some threshold, typically 0.05.

- All possible regression method. .

  This algorithm fits all regressions involving one regressor, two regressors, three regressors, and so on. The selection criterion is recorded for each regression. Once the procedure finishes, the champion for each subset size is determined.

Becuase forward selection method or all possible regression may increase the risk of over fitting We used the backward elimination method because it aligns with the objective

of studying the individual contributions of variables to Sales Growth while simplifying the model and also this method ensures than all remaining predictor variables in the best regression equation are statistically significant and draw meaningful conclusions on this study.

Chapter 4

Data

Overview

This chapter will provide a overview of the data set and any unique characteristics associated with it.

4.1 About the data set

The dataset examines factors that influence consumers' online buying habits. 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. 6

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

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, most of the variables are continuous. The dependent variable Sales Growth, is a measure which represents the total value of online sales growth within the North America. Therefore, it is typically not based on the categorical variables.

4.4 Data set preparation

Data cleaning is a crucial step in the data analysis process. It involves identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to ensure that the data is accurate, reliable, and ready for analysis.

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','PROVINCE'

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

Check null values

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

- is.null(data)

As shown in figure 4.1 There are also no any null values present in the data set too. Therefore, there is no need to use any missing data handling techniques.

```
> print(paste("Total null data:", null_data))
[1] "Total null data: 0"
```

Figure 4.1: Null value Statistics

Check NA values

Then we check is this data set has NA values.

- is.na(data)

```
> # Check for missing values in specific columns
> missing_per_column <- data.frame(colSums(is.na(OSE)))
> print(missing_per_column)
                 colSums.is.na.OSE..
SALES                              0
DISCOUNT                           0
SHIPMODE                           0
UNITPRICE                          0
SHIPPINGCOSTS                      0
CUSTOMERSEGMENT                    0
PRODUCTCATEGORY                    0
SHIPPINGDURATION                   0
```

Figure 4.2: Missing Value Table

Here also found that this data set hasn't any NA values.We don't want to use any missing value replacement techniques.

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.4.3

Outliers

According to the box plots obtained from the R software, there are outliers in Sales growth, Unit Price, Shipping Cost,Shipping Duration. figure 4.4 and figure 4.5 shows that there are uncountable outliers in dependent variable (Sales Growth) as well as independent variables.
We use different method to remove these outliers such as IQR method, Get logarithms,Get square roots and Get cubic root of the variables but they didn't satisfy the

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 4.3: *Label for categories in data set*

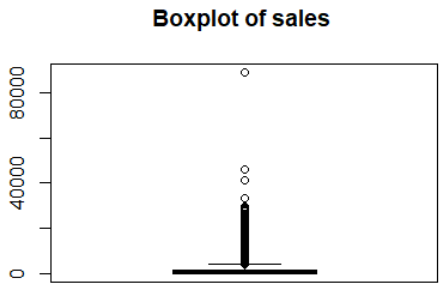

Figure 4.4: *Boxplot of Sales Growth(Dependent Variable)*

Figure 4.5: *Boxplot of Independent Variables*

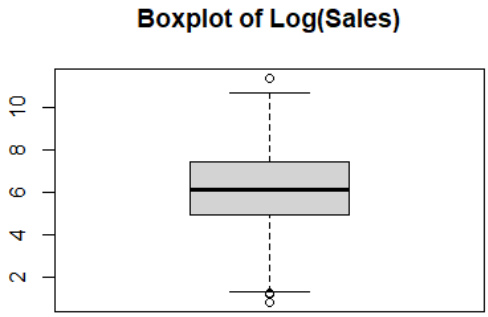

Figure 4.6: Boxplot log(Sales Growth)

assumption of regression. Therefore, we use a transformation for Sales to avoid more outliers. We use log Sales as the transformation for Sales. According to the boxplot of log(Sales), there are some outliers in log(sales). However, we can't remove these as this is a time series data set.
According to the figure 4.6 here three outliers in the data set. comparing to the large data set it is an uncountable outliers.

```
> Before <- summary(model)$r.squared
> Before
[1] 0.4675886
```

Figure 4.7: *Model r squared summary (before)*

```
> After <- summary(model2)$r.squared
> After
[1] 0.3918914
```

Figure 4.8: *Model r squared summary (after)*

According to the figure 4.7 and figure 4.8, We can see that R squared values with outliers, is greater than without outliers. It doesn't mean that these two models are useless just because that they are less than 50% .Because this is a real world data set, We can see independent variables in this model aren't effectively capturing the factors that influence the dependent variable. There could be inherent randomness in the data that this model isn't accounting for.

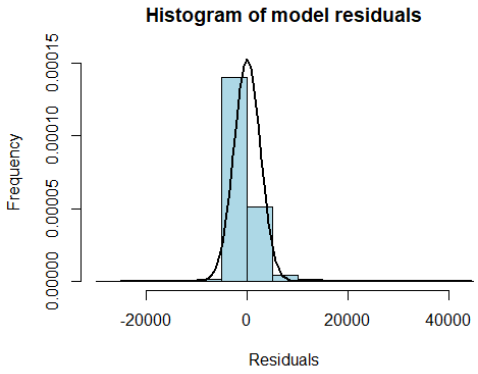

Figure 4.9: *Histogram - residual (For Orginal data set)*

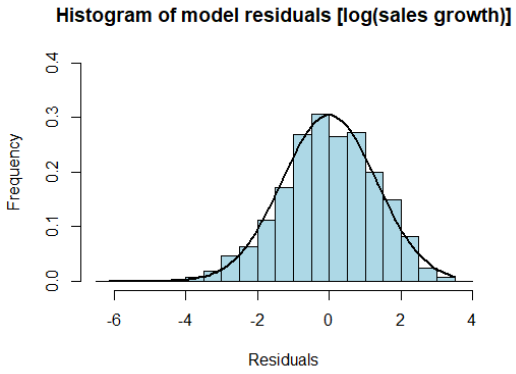

Figure 4.10: *Histogram - residual (log(Sales Growth) as Dipendent)*

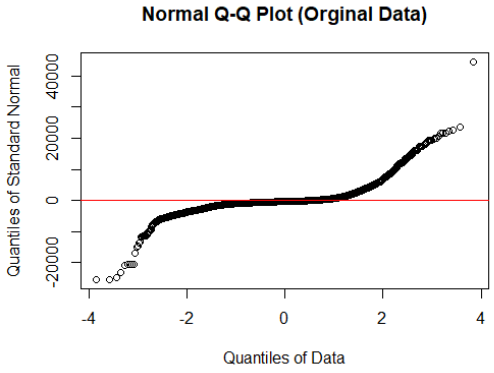

Figure 4.11: *Q-Q plot for orginal data set*

Figure 4.12: *Q-Q plot (After Get log(Sales Growth)*

In above both histogram plots, the data points are shown as a Symmetrically distributed.It means the residuals follow a average normal distribution in both scenarios.But After observing figure 4.11 and figure 4.12, we can clearly see figure 4.12 is normally distributed than the 4.11. but residuals are still not normally distributed. This cause be non-linearity in the relationship between the independent and dependent variable.

After observing above graphs, we choose to get logarithm because model2 (dependent variable <- log(salesgrowth)) is more suitable than model 1 (dependent variable <- salesgrowth).
Therefore, the first assumption of the regression is satisfied by our new model.

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

```
> library(psych)
> OSEnew <- OSE[, c("UNITPRICE","SHIPPINGCOSTS", "DISCOUNT","SHIPPINGDURATION","SALES","SALES_LOG")]
> descriptive_stat<-describe(OSEnew)
> descriptive_stat
                 vars    n    mean      sd median trimmed    mad  min      max    range  skew kurtosis    se
UNITPRICE           1 8399   89.35  290.35  20.99   44.77  25.22 0.99  6783.02  6782.03 14.12   270.94  3.17
SHIPPINGCOSTS       2 8399   12.84   17.26   6.07    8.78   5.35 0.49   164.73   164.24  2.55     7.74  0.19
DISCOUNT            3 8399    0.05    0.03   0.05    0.05   0.04 0.00     0.25     0.25  0.07    -0.96  0.00
SHIPPINGDURATION    4 8399    2.03    2.30   2.00    1.71   1.48 0.00    92.00    92.00 14.24   478.73  0.03
SALES               5 8399 1775.88 3585.05 449.42  944.26 566.28 2.24 89061.05 89058.81  5.39    60.88 39.12
SALES_LOG           6 8399    6.20    1.68   6.11    6.19   1.82 0.81    11.40    10.59  0.05    -0.52  0.02
```

Figure 5.1: *Descriptive Statistics Table*

Below figure 5.1 shows various statistical measures for different variables. The mean values of Discount, shipping duration,log(sales) are close to the median, which indicates that the data is fairly symmetrical. The standard deviation of dependent variable log(sales) is relatively small when considered with other variables. It indicates that there are a wide range of values for log(sales).

```
> ##to get quantiles
> # Select only numeric columns from the cleaned_data
> numeric_data <- OSEnew[sapply(OSEnew, is.numeric)]
> # Apply the quantile function to the numeric data
> quantiles <- sapply(numeric_data, quantile)
> quantiles
     UNITPRICE SHIPPINGCOSTS DISCOUNT SHIPPINGDURATION     SALES  SALES_LOG
0%        0.99          0.49     0.00                0     2.240  0.8064759
25%       6.48          3.30     0.02                1   143.195  4.9642071
50%      20.99          6.07     0.05                2   449.420  6.1079579
75%      85.99         13.99     0.08                2  1709.320  7.4438506
100%   6783.02        164.73     0.25               92 89061.050 11.3970774
```

Figure 5.2: Quantiles Statistic

To display the distributions of numerical data, we plotted histograms and scatter plots. This will aid in determining whether they follow any pattern (normal, skewed, etc.) for histograms and get the relationship between log(sales) and each independent variables.

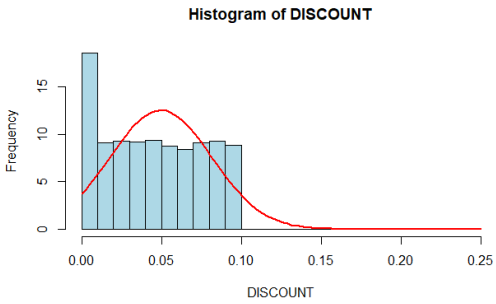

Figure 5.3: *Histogram of Discount*

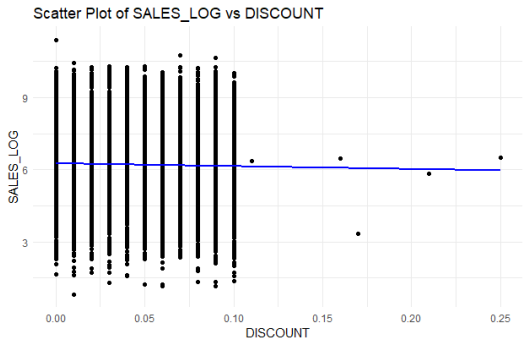

Figure 5.4: *Scatter plot of log(sales) VS Discount*

According to the figure 5.3 and figure 5.4, this data set is symmetric because the majority of data points spread around the mean.

The scatter plot of log(sales) vs. Discount shows that they have no relationship.

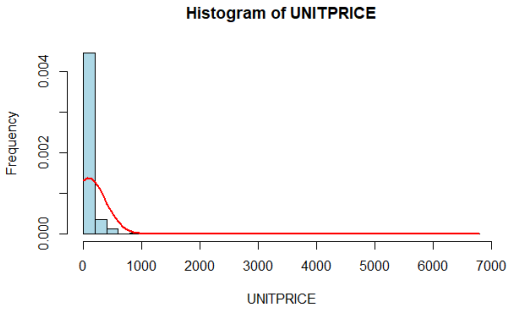

Figure 5.5: *Histogram of Unit Price*

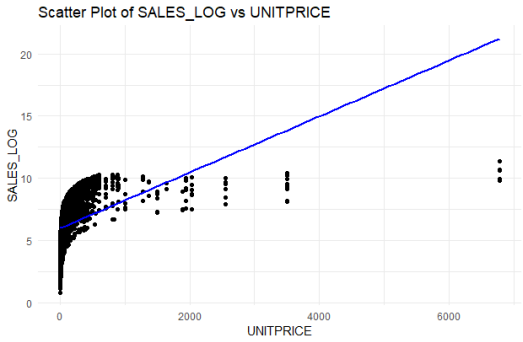

Figure 5.6: *Scatter plot of log(sales) VS Unit Price*

According to the figure 5.5 and figure 5.6, this data set is skewed because the majority of data points are in the right if the mean.

The scatter plot of log(sales) vs. Unit Price shows that they have strongly positive relationship.. It means, if the Unit Price is increasing then the log(sales) will go up

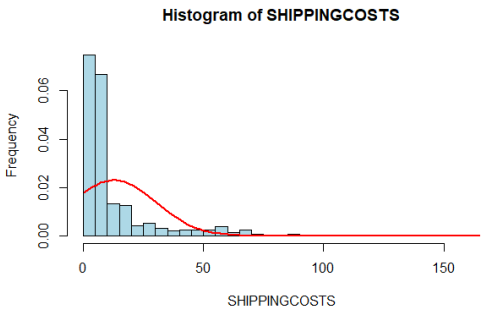

Figure 5.7: *Histogram of Shipping Cost*

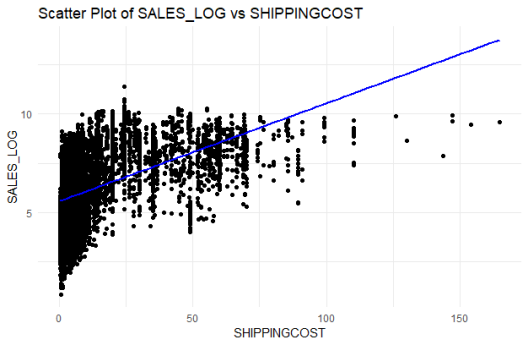

Figure 5.8: *Scatter plot of log(sales) VS Shipping Cost*

According to the figure 5.7 and figure 5.8, this data set is right skewed because the majority of data points are in the left to the mean.

The scatter plot of log(sales) VS Shipping Cost shows that they have positive relationship.

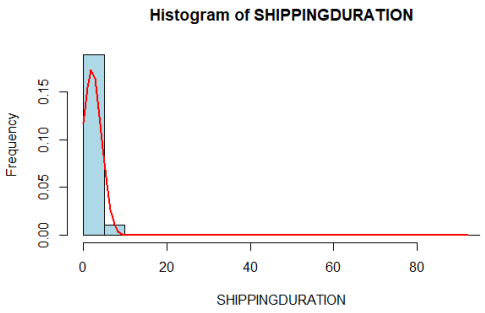

Figure 5.9: *Histogram of Shipping Duration*

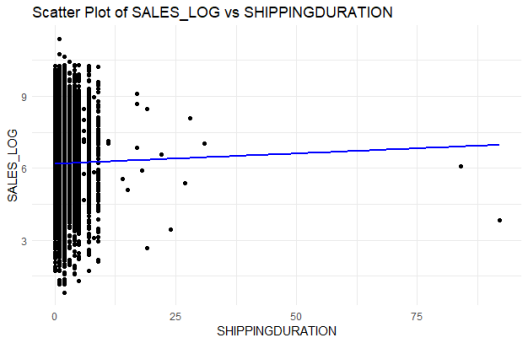

Figure 5.10: *Scatter plot of log(sales) VS Shipping Duration*

According to the figure 5.9 and figure 5.10, this data set is symmetric because the majority of data points are spread around the mean.

The scatter plot of log(sales) VS Shipping Duration shows that they have week positive relationship..

5.2 Quantitative Analysis

As mentioned earlier, there are no missing values and null values in this dataset. Here, we used model that have log(sales) as dependent variable for this analysis.Please refer the appendix for the associated R code.

5.2.1 Assumptions for multiple linear regression

Before fit the model there are assumption to check whether dataset is suitable for to fit the regression model.

1. First we have to check whether residual values are normally distributed.For that we draw histogram and Q-Q plot. According to figure 4.10 and 4.12 shows that this assumption satisfy for this data set.

2. Secondly we check linearity, there must be a linear relationship between the dependent and the independent variable. As shown in figure 5.4,figure 5.8,figure 5.6,figure 5.10 assumption also satisfied by our data set.

3. After that we check no multicollinearity, It means that the independent variables are not highly correlated with each other (r value should be less than 0,8).According to the correlation matrix 5.12 this is also satisfied.

4. Finally we check homoscedasticity. It assume that the variance of the residual errors is similar across the value of each independent variable.One way of checking that is through a plot of the predicted valued against the standardize residual values to see if the points are equally distributed across all the values of the independent variables. according to the below figure 5.11 least equally spread around the 0 in in the plot. So, this assumption also satisfy by our data set.

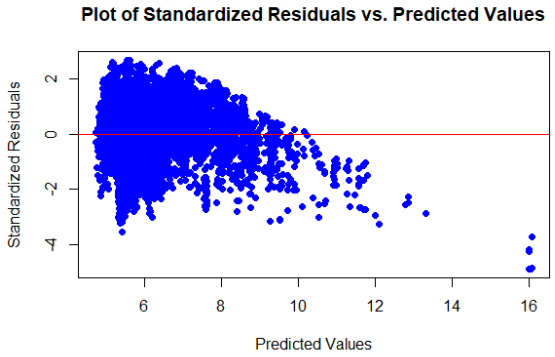

Figure 5.11: *Plot of Standardized Residuals vs. Predicted Values*

Now we can model the regression model for further analysis.

5.2.2 Correlation Analysis

```
> #multicollinearity
> # Compute correlation matrix
> correlation_matrix <- cor(OSE[, c("UNITPRICE","SHIPPINGCOSTS", "DISCOUNT","SHIPPINGDURATION","SALES_LOG")])
> print(correlation_matrix)
                  UNITPRICE SHIPPINGCOSTS     DISCOUNT SHIPPINGDURATION   SALES_LOG
UNITPRICE         1.0000000000  2.399594e-01  0.001332397    -6.766746e-04  0.38803557
SHIPPINGCOSTS     0.2399593750  1.000000e+00 -0.001955711    -3.895245e-05  0.51295859
DISCOUNT          0.0013323969 -1.955711e-03  1.000000000    -2.809981e-03 -0.02225680
SHIPPINGDURATION -0.0006766746 -3.895245e-05 -0.002809981     1.000000e+00  0.01170554
SALES_LOG         0.3880355734  5.129586e-01 -0.022256795     1.170554e-02  1.00000000
```

Figure 5.12: *Correlation matrix*

1. Research question 1: Relationship between unit price and sales growth of online shopping company.

```
> # Calculate correlation coefficient and p-value
> correlation_result <- cor.test(OSEnew$UNITPRICE,OSEnew$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): 0.3880356
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 6.15833e-300
```

Figure 5.13: *correlation coefficient and p-value of unit price and log(sales).*

The correlation coefficient of 0.3880 between unit price and log(sales) implies that the relationship between these two variables is weakly positive. This suggests that when the unit price increasing, the log(sales) also increase. Also, this relationship is marginally significant at the 6.15833e-300 level($p < 0.01$), suggests that it is unlikely to have occurred by random chance.

2. Research question 2: Relationship between Shipping cost and sales growth of online shopping company.

```
> # Calculate correlation coefficient and p-value
> correlation_result <- cor.test(OSEnew$SHIPPINGCOSTS,OSEnew$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): 0.5129586
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0
```

Figure 5.14: *correlation coefficient and p-value of Shipping cost and log(sales).*

According to the correlation of 0.5129586 between shipping cost and log(sales), there is highly positive relationship. But it is not statistically significant at 0.05 level as its significance level is 0. This means that there suggests a much stronger correlation.

3. Research question 3: Relationship between Discount and sales growth of online shopping company

```
> # Calculate correlation coefficient and p-value
> correlation_result <- cor.test(OSEnew$DISCOUNT,OSEnew$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): -0.0222568
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0.04138056
```

Figure 5.15: *correlation coefficient and p-value of Discount and log(sales).*

The correlation between discount and log(sales) is -0.0222568 . It implies that the relationship between these variables is weakly negative. p-value of 0.0413 is statistically significant at the 0.05 level. This means it's unlikely (less than 5% chance) to observe a correlation as negative (or more negative) if there truly were no relationship between discount and log(sales).

4. Research question 4: Relationship between Shipping Duration and sales growth of online shopping company.

```
> # Calculate correlation coefficient and p-value
> correlation_result <- cor.test(OSEnew$SHIPPINGDURATION,OSEnew$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): 0.01170554
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0.2834311
```

Figure 5.16: *correlation coefficient and p-value of Shipping duration and log(sales).*

The correlation between log(sales) and shipping duration very low and positive because it is 0.01170554. Despite the weak correlation, the p-value is statistically significant at the 0.05 level. This means it's unlikely ($p<0.05$) to observe a correlation as positive (or more positive) if there were truly no relationship between the variables.

5. Research question 5: Relationship between customer segment and log(sales) of online shopping company.

```
> correlation_result <- cor.test(OSE$customer_segment_numerical,OSE$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): -0.005764838
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0.5973255
```

Figure 5.17: *correlation coefficient and p-value of Customer segment and log(sales).*

The correlation between customer segment and log(sales) is -0.005764838.It convey that there is a significant negative correlation with customer segment and log(sales).And also, the significance level is 0.5973255 ($p > 0.05$)which implies this is not statistically significant at the level 0.05, indicating that there is little evidence of a meaningful linear relationship between these variables.

6. Research question 6: Relationship between product category and log(sales) of online shopping company.

```
> correlation_result <- cor.test(OSE$product_category_numerical,OSE$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): 0.03353724
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0.002112359
```

Figure 5.18: *correlation coefficient and p-value of product category and sales growth.*

The correlation between product category and log(sales) is 0.03353724. there is lowly positive relationship. But it is not statistically significant at 0.05 level as its significance level is 0.002112359. This means it's unlikely ($p<0.05$) to observe a correlation as positive (or more positive) if there were truly no relationship between the variables.

7. Research question 7: Relationship between ship mode and log(sales) of online shopping company.

```
> correlation_result <- cor.test(OSE$ship_mode_numerical,OSE$SALES_LOG, method = "pearson")
> # Print the correlation coefficient (r)
> cat("Correlation Coefficient (r):", correlation_result$estimate, "\n")
Correlation Coefficient (r): -0.4168282
> # Print the p-value
> cat("p-value:", correlation_result$p.value, "\n")
p-value: 0
```

Figure 5.19: *correlation coefficient and p-value of Shipping mode and sales growth.*

There is a weekly negative relationship between shipping mode and log(sales).But it is not statistically significant at 0.05 level as its significance level is 0. This means that there suggests a much stronger correlation.

5.2.3 Estimate model parameters

In this section, relationship between the dependent and independent variables are mathematical represented. Then, which independent variables are to be included is chosen to assess the model's fit. For this process, there are three methods namely forward selection method, backward elimination, and all possible regression method.So, Backward elimination method is used to fit this model.

```
> full_model <- lm(SALES_LOG ~DISCOUNT+ship_mode_numerical+UNITPRICE+SHIPPINGCOSTS+customer_segment_numerical+
product_category_numerical+SHIPPINGDURATION,data = OSE)
> summary(full_model)

Call:
lm(formula = SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE +
    SHIPPINGCOSTS + customer_segment_numerical + product_category_numerical +
    SHIPPINGDURATION, data = OSE)

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 ***
DISCOUNT                  -1.071e+00  4.494e-01  -2.384   0.0172 *
ship_mode_numerical       -3.035e-01  2.720e-02 -11.159   <2e-16 ***
UNITPRICE                  1.319e-03  5.198e-05  25.380   <2e-16 ***
SHIPPINGCOSTS              4.567e-02  1.186e-03  38.527   <2e-16 ***
customer_segment_numerical -1.665e-03  1.409e-02  -0.118   0.9059
product_category_numerical  6.111e-01  2.386e-02  25.612   <2e-16 ***
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 5.20: *Summary of full model*

In a backward elimination is a feature selection technique used in building regression models. It starts with a model that includes all possible explanatory variables and iteratively removes the least significant variable until a stopping point is reached.

Full model,

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

In here we first get the full model and remove the greatest p value (This value should be >p) from summary and built a new model and use nested anova for full model and reduce model. From that we check whether p-value associated with the F-statistic is less than 0.05 ($p<0.05$) .If it is, then we reject null hypothesis(Full model is suitable). We repeatedly do this until we get the all the summary p values less than 0.05.

We get hypothesis as follows,

- H_0: Reduce model is suitable.
- H_1: Full model is needed.

In the first step we remove shipping duration that has p = 0.2609 (>p) (figure 5.20).Then we get the model without shipping duration and get it as a reduce model1. And check whether nested anova tables' F Pr(>F) is greater than 0.05. from figure 5.21

```
> # Reduce 1
> reduce_model1 <- lm(SALES_LOG ~DISCOUNT+ship_mode_numerical+UNITPRICE+SHIPPINGCOSTS+product_category
_numerical+SHIPPINGDURATION,data = OSE)
> anova(full_model,reduce_model1)
Analysis of Variance Table

Model 1: SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE + SHIPPINGCOSTS +
    customer_segment_numerical + product_category_numerical +
    SHIPPINGDURATION
Model 2: SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE + SHIPPINGCOSTS +
    product_category_numerical + SHIPPINGDURATION
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1   8391 14408
2   8392 14408 -1 -0.023976 0.014 0.9059
```

Figure 5.21: *Nested ANOVA for full model and reduce model*

We can see that nested anova F Pr(>F)= 0.9059. So we accept null hypothesis.Now we can get reduce model 1 as new model with R squared = 0.3919.

```
> summary(reduce_model1)

Call:
lm(formula = SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE +
    SHIPPINGCOSTS + product_category_numerical + SHIPPINGDURATION,
    data = OSE)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1658 -0.8352  0.0041  0.9277  3.5097

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.078e+00  9.843e-02  51.589   <2e-16 ***
DISCOUNT                  -1.071e+00  4.493e-01  -2.383   0.0172 *
ship_mode_numerical       -3.035e-01  2.720e-02 -11.160   <2e-16 ***
UNITPRICE                  1.319e-03  5.197e-05  25.389   <2e-16 ***
SHIPPINGCOSTS              4.567e-02  1.185e-03  38.529   <2e-16 ***
product_category_numerical 6.111e-01  2.386e-02  25.614   <2e-16 ***
SHIPPINGDURATION           6.988e-03  6.213e-03   1.125   0.2607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.31 on 8392 degrees of freedom
Multiple R-squared:  0.3919,    Adjusted R-squared:  0.3915
F-statistic: 901.4 on 6 and 8392 DF,  p-value: < 2.2e-16
```

Figure 5.22: *Summary for reduce model 1*

Again we check the p values from the summary of reduce model1. figure 5.22 shows that Shipping duration's p value=0.2607.So we remove that variable from reduce model1 equation and build a new equation as reduce model2 and do nested anova test for reduce model1 and reduce model2. Then we get the table as figure 5.23 shown,

```
> anova(reduce_model1,reduce_model2)
Analysis of Variance Table

Model 1: SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE + SHIPPINGCOSTS +
    product_category_numerical + SHIPPINGDURATION
Model 2: SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE + SHIPPINGCOSTS +
    product_category_numerical
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1   8392 14408
2   8393 14410 -1   -2.1718 1.265 0.2607
```

Figure 5.23: *Nested ANOVA for reduce model1 and reduce model2*

This time also check hypothesis.In here also we can see that F Pr(>F)=0.2607 value is greater than 0.05. So we accept the null hypothesis and get reduce model2 is suitable. Then assign reduce model 2 as new model.

We get summary for reduce model 2 5.24

```
> summary(reduce_model2)

Call:
lm(formula = SALES_LOG ~ DISCOUNT + ship_mode_numerical + UNITPRICE +
    SHIPPINGCOSTS + product_category_numerical, data = OSE)

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 ***
DISCOUNT                  -1.072e+00  4.493e-01  -2.386   0.0171 *
ship_mode_numerical       -3.035e-01  2.720e-02 -11.159   <2e-16 ***
UNITPRICE                  1.319e-03  5.197e-05  25.386   <2e-16 ***
SHIPPINGCOSTS              4.568e-02  1.185e-03  38.534   <2e-16 ***
product_category_numerical 6.114e-01  2.386e-02  25.627   <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 5.24: *Summary of reduce model 2*

Again we check p value form figure 5.24. In this stage the model is statistically significant as the p value of remaining variable,Discount,Ship mode numerical,unit price,Shipping cost ,Product category numerical is less than 0.05 which suggests that they are strong predictors for explaining sales.

For this study, the best regression equation is:

$$SALE\_LOG = 5.092 + 0.001319(UNITPRICE) - 1.072(DISCOUNT) -0.3035(SHIPMODE) +0.6114(PRODUCTCATEGORY) +0.04568(SHIPPINGCOST)$$

In this model, the R-squared value in is 39.18%. This indicates that approximately 31.18% of the variance in the sales growth can be explained by the independent variables in the model. The adjusted R-squared value of this model is 39.14%. The difference between the R-squared value and adjusted R-squared value is very small (about 0.04). That says that those additional variables might not be substantially improving the model's explanatory power after considering the increased complexity due to more variables.

5.2.4 Assess model fit

Here are the regression equations of two models.

Full model:

$$Sales\_log = 5.082 + 0.001319(Unit\ price) - 1.071(Discount) - 0.3035(shipping\ mode) + 0.6111(Product\ category) + 0.04567(shipping\ Cost) - 0.001665(Customer\ segment) + 0.00698(shipping\ duration)$$

Reduced model:

$$Sales\_log = 5.082 + 0.001319(Unit\ price) - 1.071(Discount) - 0.3035(shipping\ mode) + 0.6111(Product\ category) + 0.04567(shipping\ Cost)$$

```
> anova(full_model)
Analysis of Variance Table

Response: SALES_LOG
                           Df  Sum Sq Mean Sq   F value    Pr(>F)
DISCOUNT                    1    11.7    11.7    6.8353  0.008953 **
ship_mode_numerical         1  4117.8  4117.8 2398.1860 < 2.2e-16 ***
UNITPRICE                   1  2233.0  2233.0 1300.4986 < 2.2e-16 ***
SHIPPINGCOSTS               1  1792.7  1792.7 1044.0290 < 2.2e-16 ***
customer_segment_numerical  1     0.0     0.0    0.0014  0.970029
product_category_numerical  1  1127.6  1127.6  656.7078 < 2.2e-16 ***
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 5.25: *ANOVA table of full model*

```
> anova(reduce_model2)
Analysis of Variance Table

Response: SALES_LOG
                           Df  Sum Sq Mean Sq   F value  Pr(>F)
DISCOUNT                    1    11.7    11.7    6.8359 0.00895 **
ship_mode_numerical         1  4117.8  4117.8 2398.3921 < 2e-16 ***
UNITPRICE                   1  2233.0  2233.0 1300.6103 < 2e-16 ***
SHIPPINGCOSTS               1  1792.7  1792.7 1044.1187 < 2e-16 ***
product_category_numerical  1  1127.6  1127.6  656.7511 < 2e-16 ***
Residuals                8393 14410.0     1.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.26: *ANOVA table of reduced model*

1. Hypothesis 1

   The full model, figure 5.25) considered for checking this hypothesis.

   - $H_0 : \beta_1 = 0 \quad and \quad H_1 : \beta_1 \neq 0$

     The p-value of Unit price is 2.2×10^{-16}, which is less than 0.05. This implies that the p-value is in the rejection region. So, the null hypothesis is rejected. Therefore, $\beta_1 \neq 0$. This suggests that there is a linear relationship between Unit price and log(Sales).According to the reduced model, $\beta_1 = 0.001319$..

   - $H_0 : \beta_2 = 0 \quad and \quad H_1 : \beta_2 \neq 0$

     The p-value of Discount is 2.2×10^{-16} and it is less than 0.05. That means the p-value is in the rejection region. So, the null hypothesis is rejected. Therefore, $\beta_2 \neq 0$. This suggests that there is a linear relationship between Discount and log(Sales).According to the reduced model, $\beta_2 = -1.071$.

   - $H_0 : \beta_3 = 0 \quad and \quad H_1 : \beta_3 \neq 0$

     The p-value of Shipping mode is 2.2×10^{-16}, which is less than 0.05. This implies that the p-value is in the rejection region.So, the null hypothesis is rejected. Consequently, $\beta_3 \neq 0$. This suggests that there is a linear relationship between Shipping mode rate and log(Sales). According to the reduced model,$\beta_3 = -0.3035$.

   - $H_0 : \beta_4 = 0 \quad and \quad H_1 : \beta_4 \neq 0$

     The p-value of Product category is 2.2×10^{-16}, which is less than 0.05. This implies that the p-value is in the rejection region.So, the null hypothesis is rejected. Consequently, $\beta_4 \neq 0$. This suggests that there is a linear relationship between Product category rate and log(Sales). According to the reduced model,$\beta_4 = +0.6111$

   - $H_0 : \beta_5 = 0 \quad and \quad H_1 : \beta_5 \neq 0$

     The p-value of Shipping Cost is 2.2×10^{-16}, which is less than 0.05. This

implies that the p-value is in the rejection region.So, the null hypothesis is rejected. Consequently, $\beta_5 \neq 0$. This suggests that there is a linear relationship between Product category rate and log(Sales). According to the reduced model,$\beta_5 = +0.04567$

- $H_0 : \beta_6 = 0 \quad and \quad H_1 : \beta_6 \neq 0$

  The p-value of Customer Segment is 0.970029, which is greater than 0.05. This implies that the p-value is not in the rejection region.So, there is not much evidence to reject the null hypothesis.Consequently, $\beta_6 = 0$. This suggests that there is no linear relationship between Customer Segment and log(Sales). According to the reduced model,$\beta_6 = 0$

- $H_0 : \beta_7 = 0 \quad and \quad H_1 : \beta_7 \neq 0$

  The p-value of Shipping Duration is 0.260877, which is greater than 0.05. This implies that the p-value is not in the rejection region. So, there is not much evidence to reject the null hypothesis. Consequently, $\beta_7 = 0$. This suggests that there is no linear relationship between Shipping Duration and log(Sales).According to the reduced model,$\beta_6 = 0$

2. Hypothesis 2

- H_0 : Reduce model is suitabel.
- H_1 : Full model is needed.

  In here we use F partial test,

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

$$F_{\text{partial}} = \frac{SS_{\text{Reg}}(\text{FULL}) - SS_{\text{Reg}}(\text{REDUCED})}{MS_{\text{Error}}(\text{FULL})} \tag{5.2}$$

5.3 Discussion and Conclusions

5.3.1 Discussion

In this study, Impact of Sales Discounts, Shipping Modes and Customer Variables on Sales Performance in North America within 2009 to 2012. period is observed. So, the dataset includes 8399 observations without any missing values and null values.

The role of product categories highlights the importance of a targeted approach in marketing and inventory management. Shipping modes and customer segments, while not significant in this study, should not be entirely disregarded as they might have indirect effects or become significant in different contexts or with different datasets.

Because we have outliers in the dependent variable ,we use logarithm to minimize the outliers . After that we get log(sales) as our dependent variable.

Here, the distribution of Discount, Shipping duration and log(sales) appear symmetric while the Unit price and Shipping cost shows skewness. The scatter plots illustrate the graphical interpretation of the relationships between the independent variables and the log(sales).

The regression analysis is a powerful statistical technique used to examine the relationships between the variables and make predictions. For this study, the regression equation of full model is

$$Sales\_log = 5.082 + 0.001319(Unit\ price) - 1.071(Discount) - 0.3035(shipping\ mode) + 0.6111(Product\ category) + 0.04567(shipping\ Cost) - 0.001665(Customer\ segment) + 0.00698(shipping\ duration)$$

But, in this model, the contribution from each variable to the dependent variable log(sales) is very low from the variables customer segment and Shipping category.That means these variables are not statistically significant at the level 0.05. Therefore, to obtain the best regression equation, backward elimination is used and the obtained reduced.

model equation is,

$\text{Sales}_log = 5.082 + 0.001319(Unitprice) - 1.071(Discount) - 0.3035(shippingmode) + 0.6111(Productcategory) + 0.04567(shippingCost)$

In the backward elimination model, this equation is obtained by removing the least significant predictor variable in each step if there are variables with p value greater than 0.05. That means the variable which has the highest p value greater than 0.05 is removed and continued until the model is significant with the remaining variables.

To ascertain if independent factors significantly affect the dependent variable, regression analysis and hypothesis testing are utilized.
Based on the outcomes of the hypothesis . According to the results obtained by hypothesis 1,

$\beta_1 \neq 0$, $\beta_2 \neq 0$,$\beta_3 \neq 0$, $\beta_4 \neq 0$, and $\beta_5 \neq 0$,$\beta_6 = 0$,$\beta_7 = 0$

This is further proved by the reduced model equation.

And also, We use reduced model and partial F test to find suitable model.

5.3.2 Conclusions

This study provides valuable insights into the factors affecting sales performance in North America from 2009 to 2012.

Because the data set is a time-sires data set collected in the real world, there are many outliers in that data set. there may be a errors in the predictions made by the model.

According to the findings, our model's R-squared value of 39.18% indicates a statistically non-significant fit (p-value > 0.05) for the data. This suggests that the model, in its current form, is not suitable for making accurate predictions for sales growth.

Future research should explore datasets that better capture the complexities of online sales growth, such as including variables related to customer demographics, product categories, and marketing channels. Additionally, considering a wider range of

variables and potentially using more sophisticated models might improve the model's ability to predict future sales trends.

Chapter 6

Appendix

To access the data set:

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

To read and analyzing data set using R software:

- https://docs.google.com/document/d/1CnXS8DxGLrsP76aBnwBlYjDiGvGJ3Zn4B3GEEX9pF4
  edit?usp=sharing

Bibliography

H. Alzoubi, M. Alshurideh, B. A. Kurdi, K. Alhyasat, and T. Ghazal. The effect of e-payment and online shopping on sales growth: Evidence from banking industry. *International Journal of Data and Network Science*, 6(4):1369–1380, 2022.

I. Arsta and N. Respati. The effect of sales promotion on purchase decisions mediated by brand image (study on e-commerce tokopedia in bali). *American Journal of Humanities and Social Sciences Research (AJHSSR)*, 5(12):205–215, 2021.

G. E. Belch and M. A. Belch. *Advertising and promotion: An integrated marketing communications perspective*. mcgraw-hill, 2018.

C. Chen and D. Ngwe. *Shipping fees and product assortment in online retail*. Harvard Business School Boston MA, 2018.

F. R. Fitri. The influence of web quality and sales promotion toward impulse buying behavior with openness personality as moderating variable. *Jurnal Akuntansi, Manajemen dan Ekonomi*, 20(1):48–55, 2018.

S. Hamby, E. Taylor, A. Smith, K. Mitchell, and L. Jones. Privacy at the margins| technology in rural appalachia: cultural strategies of resistance and navigation. *International Journal of Communication*, 12:21, 2018.

G. A. Johnson, S. K. Shriver, and S. Du. Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1):33–51, 2020.

S. Ma. Fast or free shipping options in online and omni-channel retail? the mediating role of uncertainty on satisfaction and purchase intentions. *The International Journal of Logistics Management*, 28, 11 2016. doi: 10.1108/IJLM-05-2016-0130.

K. Manapul, J. Isidro, M. Hernandez, and R. Fernandez. Influence of shipping fees in customer purchase decisions for online retailers residing within the philippines. *Journal of Business and Management Studies*, 4:203–212, 03 2022. doi: 10.32996/jbms.2022.4.1.23.

C. Ranganathan and E. Grandon. An exploratory examination of factors affecting online sales. *Journal of Computer Information Systems*, 42(3):87–93, 2002. doi: 10.1080/08874417.2002.11647507. URL `https://www.tandfonline.com/doi/abs/10.1080/08874417.2002.11647507`.

P. Ravula. Impact of delivery performance on online review ratings: the role of temporal distance of ratings. *Journal of Marketing Analytics*, 11(2):149–159, 2023.

S. Roth, L. Himbert, and S. Zielke. Does unit pricing influence store price image dimensions and shopping intentions for retail stores? *European Journal of Marketing*, 51(7/8):1396–1413, 2017.

K. P. SETIAWAN. Pengaruh fashion involvement, sales promotion, dan hedonic shopping motivation terhadap e-impulse buying fashion product pada marketplace shopee. 2021.

P. C. Verhoef, A. T. Stephen, P. Kannan, X. Luo, V. Abhishek, M. Andrews, Y. Bart, H. Datta, N. Fong, D. L. Hoffman, et al. Consumer connectivity in a complex, technology-enabled, and mobile-oriented world with smart products. *Journal of Interactive Marketing*, 40(1):1–8, 2017.

Z. Wang, Y. Zuo, T. Li, C. Chen, and K. Yada. Analysis of customer segmentation based on broad learning system. pages 75–80, 12 2019. doi: 10.1109/SPAC49953.2019.237870.