Data Modelling

Factors Affecting Sales

University of London

18/09/2023

ABSTRACT

The research is aimed at exploring and measuring the factors impacting on the number and volume of sales from orders made for imports and export trading. Notably, the research involves the development of a series of a model for examining the impact of the various selected variables on the sales variable. To accomplish the research, the initial strep is the development of the discussions around the models and then the analysis of the dataset with 180519 valid observations and 53 variables. The main variables used for modelling in this research include sales, discount, product price, total amount, and sales for each individual customer. The outcome from the exploratory analysis demonstrates that all the predictor variables have significant positive impacts on the sales amounts. In the result, the product price has the highest influence on the value of sales, while the order discount has the least influence on the sales. In addition, this research demonstrates that the product price significantly effects the customer's decision on product order placement for cargo delivery. Since customers are mostly influenced by the prices, this research recommends necessary reforms on the management policies and executive decisions to maximize customer's satisfaction and decision to increase the sales level. It also provides support for new innovations and further research works in new related areas.

Keywords: Sales, Modelling, Product Price, Discount, Ordering

Table of Contents

1. INTRODUCTION

1.1. Background

This study involves the investigation of factors affecting sales in a mixed market globally. Sales is the response variable (dependent variable) for this research. The dependent variables item discount, product price, total amount, and Sales for each individual customer. In the past research works, there have been studies on the factors affecting product prices. For example, Ahmed et al (2022) mentions factors such as the production costs, cost of distribution, market situations, operation fees, and cost of marketing and sales promotion. Guizzardi et al (2021) provides a discussion about dynamic product pricing using the Big Data Technology, to drive sales and prices from an informed ground, based on data analytics.

1.2. Previous Research

There have been various research works focusing on the factors affecting sales or sales performance in various perspectives. For example, Talukder & Jan (2017) and Xu, Tang & Zhou (2020) studied factors affecting Sales People's Performance, from the human resource perspective. In this case, the author presents factors such as salaries and remuneration, job satisfaction index, intrinsic inspiration, extrinsic inspiration and organizational effort (Talukder & Jan, 2017). Other studies such as Hutt (2015) and Weinberg (2015) did investigations on the factors affecting the sales of various companies, and the independent variables in each case were competition, consumer loyalty, pricing, and promotion. All these studies have motivated a different approach to investigating the factors influencing the sales. There are two choices of parameters of sales from which to select, the number of sales orders and the value of sales made. For the purpose of standardization, the number of sales does not reflect the best performance in sales. It is possible for a customer to make a few or one order with high sales amount (Johnson, 2021). Similarly, it is possible for a client to make several orders whose total sale amount is significantly low. Therefore, the amounts of sales speak better about sales performance of the cross-border trades than the number of orders (Priester, Robbert & Roth, 2020). This study is special, unique and original in the sense that it is entirely about international trade

(imports and exports businesses). Other factors not included in this study include the shipping cost, Number of shipping days (real), Scheduled number of shipping days (scheduled), Benefit realised from each order and Sales for each individual customer. These will probably be considered in the future versions of this research.

1.3. Research Problem

On the factors affecting sales volume, it has been found that if high shipping cost discourage customers from making orders and prompts them to search for better deals in other outlets. On the other hand, low costs of shipping and free offers have been found to be great incentive for promoting customers' decisions to order for products (Al-Turjman, 2017). There have been deficiency of research works on the factors affecting sales especially in the field of logistics. For example, research by Cai, Liu & Liu (2022) presents investigation of sales efficiency and effect. However, Djoni, Oktaviani & Kirbrandoko (2016) attempts to explore the factors affecting sales performance in Indonesia. Additionally, Fabra, & Reguant (2020) discusses factors that affect sales prices in the European Union region. This study seeks to bridge the gap in literature by conducting an empirical process of analysis to measure the effect of the selected factors on the sales (Le-Hoang, 2020). At the same time, the recommendation from this study serves as essential baseline for informing the production level, the pricing and the distribution of products in the market. It will guide the management of various businesses on the control of the independent variables to maximize the number of sales orders and sales values.

1.4. Research Objective

For this study, the objectives are:

1) To investigate the effect of Order discount on the value of Sales
2) To investigate the effect of Price of the ordered items on the Sales
3) To find out the effect of Order total cost on the value of Sales
4) To establish the effect of Product price on the value of Sales
5) To study the effect of Sales for each individual customer on Sales

The investigations involve the creation of model with the sales as the dependent variable and the five selected factors as the independent variables.

1.5. Research Questions

From the study objectives, this research seeks to answer the research questions below:

1) How does the Order discount affect the value of Sales?
2) How does the Price of the ordered items affect the value of Sales?
3) How does the Order total cost affect the value of Sales?
4) How does the Product price affect the value of Sales?
5) How does the Sales for each individual customer affect the value of Sales?

1.6. Hypothesis

With the five independent variables, the hypotheses of the study are set as shown below:

Order discount

Alternative Hypothesis $H1_0$: Order discount has positive impact on the value of Sales

Null Hypothesis $H1_a$: Order discount Does not have positive impact on the value of Sales

Price of the ordered items

Alternative Hypothesis $H1_0$: Price of the ordered items has positive impact on the value of Sales

Null Hypothesis $H1_a$: Price of the ordered items does not have positive impact on the value of Sales

Order total cost

Alternative Hypothesis $H1_0$: Order total cost has positive impact on the value of Sales

Null Hypothesis $H1_a$: Order total cost does not have positive impact on the value of Sales

Product price

Alternative Hypothesis $H1_0$: Product price has positive impact on the value of Sales

Null Hypothesis $H1_a$: Product price does not have positive impact on the value of Sales

Sales for each individual customer

Alternative Hypothesis $H1_0$: Sales for each individual customer has positive impact on the value of Sales

Null Hypothesis $H1_a$: Sales for each individual customer does not have positive impact on the value of Sales

This research aims at using multiple variable ordinary least square regression to perform the hypothesis tests with the constructed model.

1.7. Structure of the Paper

This paper is divided into four sections. The first section is the introduction, which provides the background of the research, and the algorithm or the statistical method to be applied. It also describes the dataset to be used. The second section is the methodology section which describes the details of the methods to be applied in the analysis. It also describes the models used in the regression. The third section is the results section, which presents all the results generated in the analysis. It contains the figures and tables used for presenting the outcomes of the analysis from data modelling, benchmarking and related investigations. The last section is the discussion of the result, with the decisions on whether the hypotheses are rejected or if there was failure to reject the hypotheses. It also provides comment on the benefits and limitations of the dataset, the methods and modelling. Finally, it provides comment on what future direction to be taken in the modelling.

2. METHODS AND MATERIALS

2.1. Choice of the Method

Since this study involves a causal-effect relationship, this study required a quantitative research method for analysing the influence of the independent factors on the sales performance. The preferred method of data analysis was the OLS regression (specifically, multiple variable linear regression). This is accomplished using R Studio application. The data is prepared by cleaning and storing in a text format (csv), which is easy to load in the R environment for data analysis. At the same time, the R script used for the analysis is placed in the same directory as that of the data for easy loading of the data, and for easy execution and replication by an experienced and motivated R programmer. Prior to the actual regression test for hypothesis testing, this study applies data validation tests such as the test for normality, the test for multicollinearity and homoskedasticity test (Yu et al, 2018). Another vital test is the descriptive statistics for displaying the properties of each variable. There was no need for performing autocorrelation test in this paper since the assumption of autocorrelation is rarely a failure cross-sectional form of data (Zang et al, 2020). Another form of essential test carried out in this study is the ANOVA test, the analysis of variance in the dependent and independent variables.

2.2. The Data

The data for this study was obtained from a secondary online data sources, with the result being the extraction of the dataset “DataCoSupplyChainDataset.csv”. The secondary data was obtained from various online websites and media sources that were essential for this study. There was no primary source or method of data collection available for collecting data sample of such a large size. The dataset consists of both numerical data variables and various categorical variables. The dataset has 180519 valid observations, which are sufficiently large and suitable for use in the quantitative research data analysis. For the quantitative method application in this study, only a few relevant numerical variables were selected for the actual data analysis and hypothesis testing.

Having collected the data samples, the cleaning of the data was done by removal of all rows with missing values, to remain with a suitable sample for valid data analysis. After cleaning of the data, the second step was performing the exploratory data analysis and the validation tests.

2.3. Data Modelling and Analysis

On the data modelling, only one regression model is created and used for performing the tests at various instances. The first step is the allocation of the selected variables to the model parameters. The table below presents a summary of the variables and model parameters to which they are assigned before the modelling.

Table 1: Model Variable Definition

Variables	**Model Parameter**
Sales	Y_i
Order discount	X_1
Price of the ordered items	X_2
Order total cost	X_3
Product price	X_4
Pales per customer	X_5
intercept	β_0
Regression coefficient estimate where i = 1 to 5	β_i
Error Term	ε_{ij}

The second step is the construction of the models or modelling and analysis. The model is

$$Y_i = \beta_0+\beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 +\beta_5 * X_5 + \varepsilon_{ij} \quad \text{Eq.1}$$

The third step is the application of the model equation in the multiple variable linear regression. The parameters of interest in the analysis are the coefficient of regression and the p – values estimated for each of the five independent variables. The coefficients of regression determine the effect of the independent variables on the sales, which is the only dependent variable in the model. The p – values are used as the measures of statistical significance.

2.4. Significance of the Tests

Since the regression test is done at the 95% confidence level, the alpha value for the test is 0.05 or 5%. Therefore, any coefficient with a p – value of between 0 to 0.05 signifies that the coefficient is statistically significant and that there is justification to reject the null hypothesis. On the other hand, having a p – value greater than 0.05 implies that the coefficient is not statistically significant and does not justify the decision to reject the null hypothesis.

3. RESULTS

3.1. Descriptive Statistics

The first section of the results is the descriptive statistics. Figure 1 below shows the descriptive statistics of the five selected model variables.

	data.Sales	data.Order.Item.Discount	data.Order.Item.Product.Price	data.Order.Item.Total	data.Sales.Per.Customer
nbr.val	180519.00	180519.00	180519.00	180519.00	180519.00
nbr.null	0.00	10028.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00	0.00
min	9.99	0.00	9.99	7.49	7.49
max	1999.99	500.00	1999.99	1939.99	1939.99
range	1990.00	500.00	1990.00	1932.50	1932.50
sum	36784735.01	3730378.40	25495158.68	33054402.38	33054402.38
median	199.92	14.00	59.99	163.99	163.99
mean	203.77	20.66	141.23	183.11	183.11
SE.mean	0.31	0.05	0.33	0.28	0.28
CI.mean.0.95	0.61	0.10	0.64	0.55	0.55
var	17496.17	475.28	19525.17	14410.48	14410.48
std.dev	132.27	21.80	139.73	120.04	120.04
coef.var	0.65	1.05	0.99	0.66	0.66

data.Sales	data.Order.Item.Discount	data.Order.Item.Product.Price	data.Order.Item.Total
2.884225	3.039770	3.190993	2.888422

data.Product.Price	data.Sales.Per.Customer
3.190993	2.888422

Figure 1: Descriptive Statistics

The descriptive statistics presents all variables with large variances and standard deviations. Their measures of skewness are between 2 and 4, meaning that the data for each variable is skewed to the right. To reduce the impact of this variance the data is scaled before modelling. The three assumptions made on the data prior to the inferential test are the test for normality, test for heteroskedasticity and the test for multicollinearity. The results for each test of the assumptions are shown below.

3.2. The Assumptions of OLS Regression

3.2.1. Test for Normality

In determining if the data meets the normality assumption, the Kolmogorov-Smirnov or the KS test was done. The result of the KS test is shown below. The criterion for determining the normality is the P-value. If the P value is higher than the alpha value of 0.05, then the data is assumed to be normal.

```
> ks.test(narrowed_data$data.Sales, narrowed_data$data.Product.Price)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  narrowed_data$data.Sales and narrowed_data$data.Product.Price
D = 0.23466, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(narrowed_data$data.Sales, narrowed_data$data.Order.Item.Discount)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  narrowed_data$data.Sales and narrowed_data$data.Order.Item.Discount
D = 0.13279, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(narrowed_data$data.Sales, narrowed_data$data.Order.Item.Product.Price)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  narrowed_data$data.Sales and narrowed_data$data.Order.Item.Product.Price
D = 0.23466, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(narrowed_data$data.Sales, narrowed_data$data.Order.Item.Total)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  narrowed_data$data.Sales and narrowed_data$data.Order.Item.Total
D = 0.08778, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(narrowed_data$data.Sales, narrowed_data$data.Sales.Per.Customer)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  narrowed_data$data.Sales and narrowed_data$data.Sales.Per.Customer
D = 0.08778, p-value < 2.2e-16
```

Figure 2: Test for Normality

All the variables show p-values of $2.26e^{-16}$, which is below the alpha value. The implication of the p-values is that the data in all the five variables are normally distributed. Additional test for normality is seen in the residuals of the histograms. If the residual histogram assumes a bell-shaped curve, then it has a normal distribution. Figure 3 below shows the residual histogram.

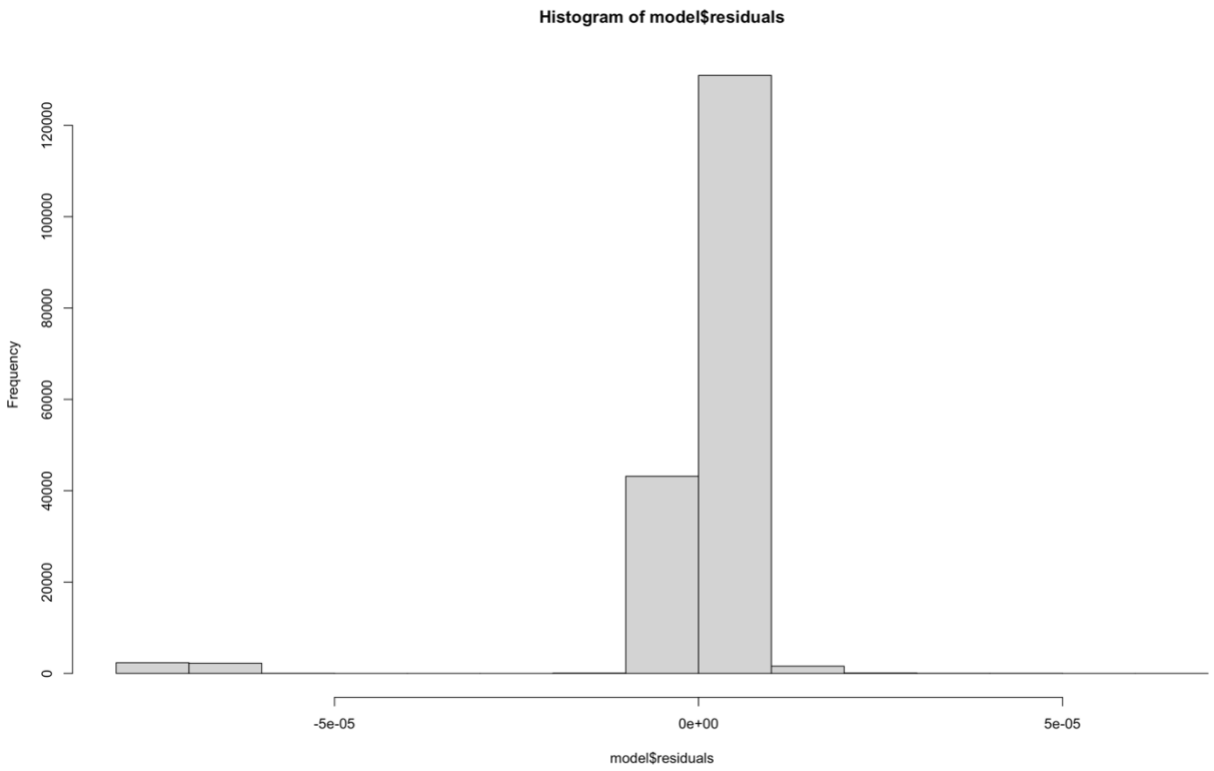

Figure 3: Residual Histogram

It is evident from the residual histogram of the model that the normality test meets the assumption of normality of the data. The other related test is the observation of the normal Q-Q curve as seen in figure 4 below.

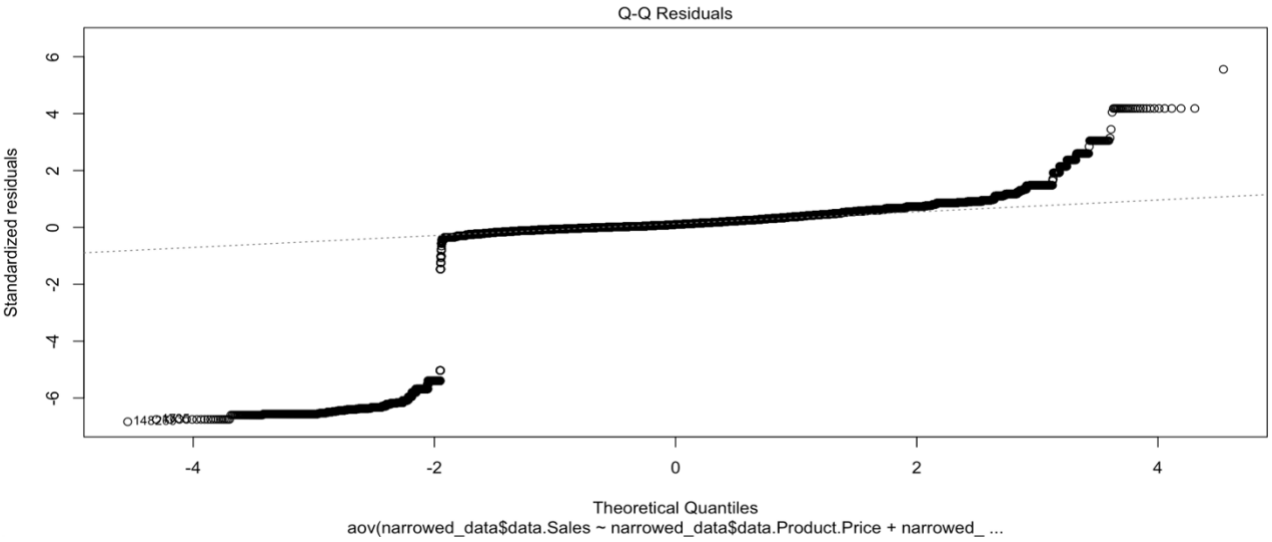

Figure 4: Normal Q-Q Curve

Since most of the data points are clustered around the normal line, the data is said to be in normal distribution, or, to have met the assumption on normality. There are two alternative approaches to use.

First, we use the normality test on KS graph demonstrating the residual scatters around the normal line. We also check if the p-value is higher than the alpha of 0.05 or observing the bell-shaped histogram, the data passes all the tests for normality and is normally distributed for all variables. The ultimate implication of the normality is that the residual model, which influences the value of sales has normal distribution.

3.2.2. Test for Heteroskedasticity

To test the data for heteroskedasticity, the two methods used were Breusch-Pagan test, and the NCV Test. While carrying out the two tests, the parameter of interest is the p-value. If the p-value of any test is below the alpha of 0.05, then the data is proven to meet the heteroskedasticity assumption. The tests are shown in figure 5 below.

```
> ## NCV Test
> car::ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 18251.17, Df = 1, p = < 2.22e-16
> ## Breusch-Pagan Test
> lmtest::bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 11993, df = 3, p-value < 2.2e-16
```

Figure 5: Breusch-Pagan and NCV Test

The two tests each has a p-value below the alpha of 0.05, so, it there is justification to conclude that the residual has a constant variance and there is no heteroscedasticity in the data. This conclusion confirms the graphical conclusion. Still on the assumption of heteroskedasticity, the scatter plot below shows no symptom of heteroskedasticity while it fits the prediction of the Y (dependent variables) and the residual data.

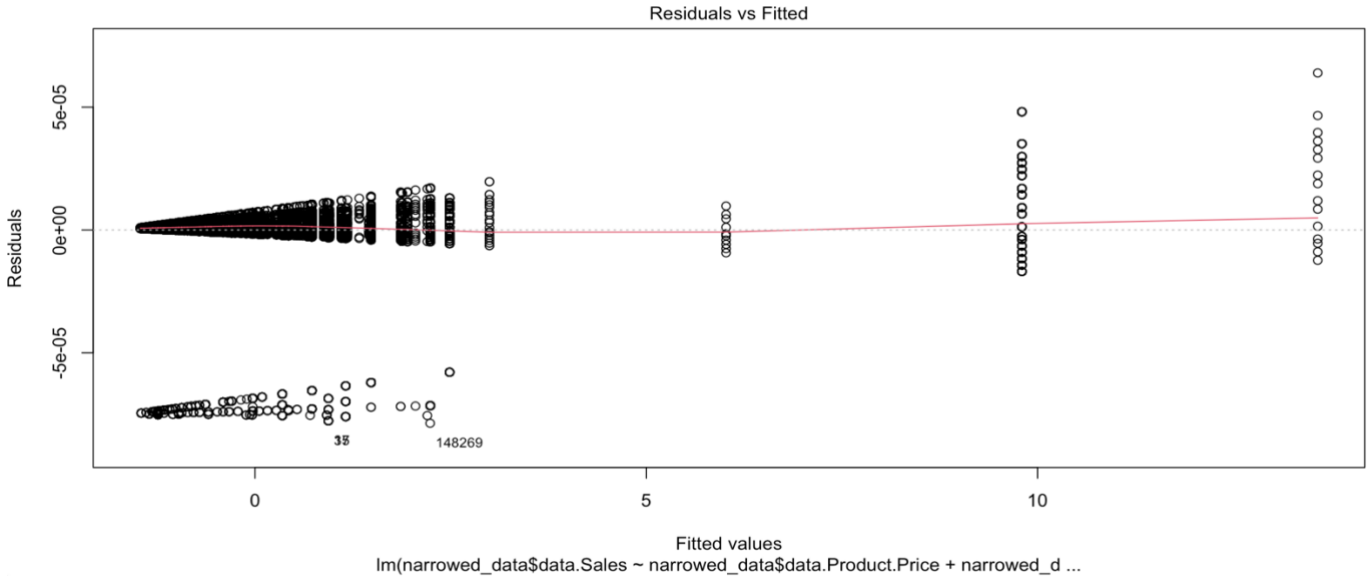

Figure 6: Fitting Residual Prediction

As observed in the plot, the data points or the scattered dots are evenly distributed above and below the axis 0 and does not create a definite pattern. Consequently, from the plotted residual versus fitted prediction, it is concluded there is no indicator of heteroskedasticity, and there is a constant variation of all the residual components. The same pattern of distribution is confirmed with the plots of the model below, including the normal Q-Q plot, scale location and residual versus leverage.

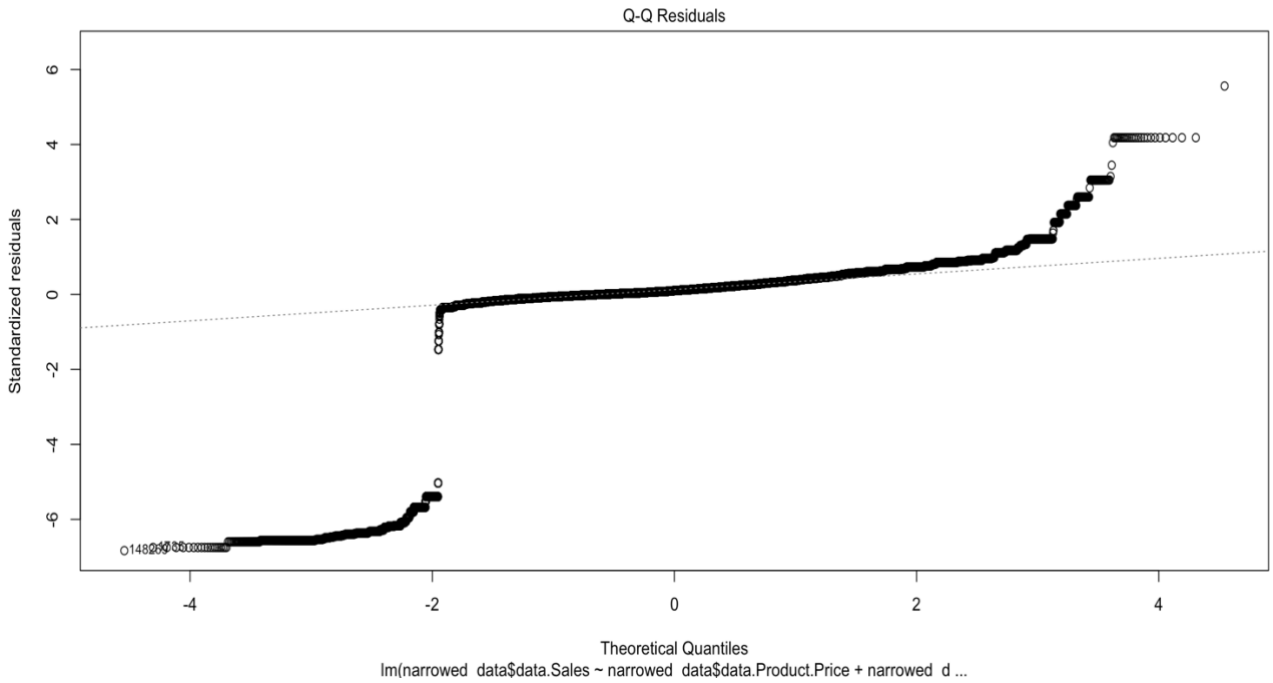

Figure 7: Normal Q-Q Plot of the Residual

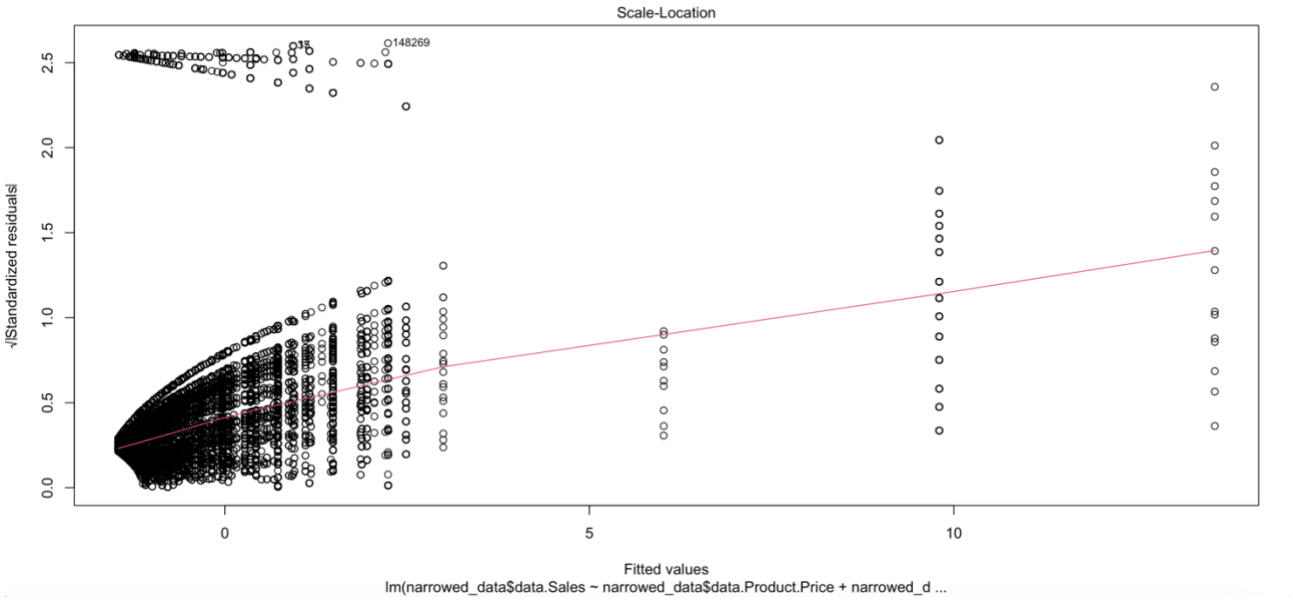

Figure 8: Scale Location (Standardized Residual)

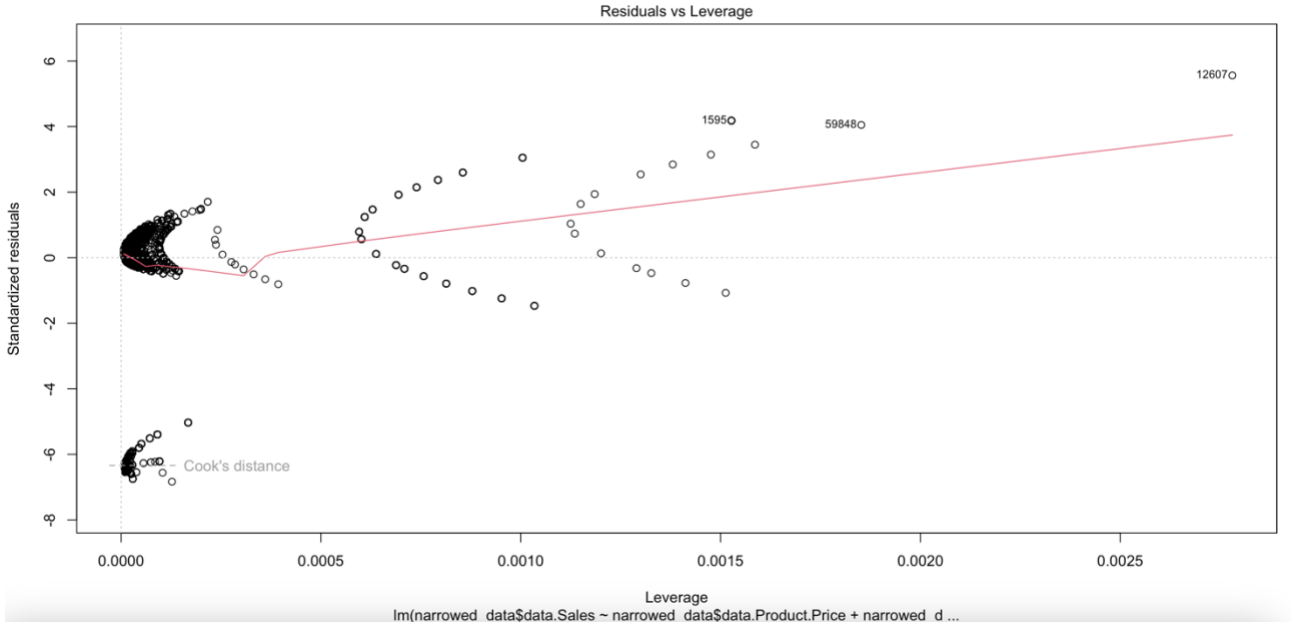

Figure 9: Standardized Residual Vs. Leverage

As the three plots confirm, the assumptions of no homoskedasticity are acceptable, and so, it can be stated that the model used to predict sales does not have heteroskedasticity in it.

3.2.3. Testing for Multicollinearity

In the regression analysis, multicollinearity test is done to check if two or more of the independent or predictor variables have high correlation with one another, so that they fail to offer unique and independent details on the model of regression. The criterion used for checking the presence of multicollinearity is the VIF score. If the VIF score falls in the range of 1 to 5, then it is concluded that the predictor variables in the model have moderate correlation among them.

```
narrowed_data$data.Product.Price narrowed_data$data.Order.Item.Discount
                        2.659704                                1.376533
narrowed_data$data.Order.Item.Total
                        2.696851
```

Figure 10: VIF Test for Multicollinearity

In the VIF test, Product Price has a VIF value of 2.66, Order discount has a VIF of 1.38, Price of the ordered items has a VIF value of 2.70. All the variables in the model have VIF scores in the range of between 1 and 5. It can be concluded that there is moderate correlation among all the four independent variables. In that case, the overall conclusion is that there is no multicollinearity among the predictor variables and that the assumption of multicollinearity fails.

3.3. Correlation Test

The correlation test was conducted as part of the inferential analysis to determine the correlation and the interaction among all the variables. The coefficients of correlations are presented in the correlation matrix in figure 11 below.

```
> round(cor(narrowed_data),2)
                          data.Sales data.Order.Item.Discount data.Order.Item.Product.Price
data.Sales                      1.00                    0.62                          0.79
data.Order.Item.Discount        0.62                    1.00                          0.49
data.Order.Item.Product.Price   0.79                    0.49                          1.00
data.Order.Item.Total           0.99                    0.50                          0.78
data.Product.Price              0.79                    0.49                          1.00
data.Sales.Per.Customer         0.99                    0.50                          0.78
                          data.Order.Item.Total data.Product.Price data.Sales.Per.Customer
data.Sales                                0.99                0.79                    0.99
data.Order.Item.Discount                  0.50                0.49                    0.50
data.Order.Item.Product.Price             0.78                1.00                    0.78
data.Order.Item.Total                     1.00                0.78                    1.00
data.Product.Price                        0.78                1.00                    0.78
data.Sales.Per.Customer                   1.00                0.78                    1.00
>
```

Figure 11: Correlation matrix

Taking sales as the dependent variable, the main interest in the correlation analysis is the correlation between sales and each of the independent variables. In that regard, the coefficient of correlation

between sales and order discount is 0.62. This is a positive coefficient, indicating a positive correlation between sales and the order discount. The coefficient of correlation between sales and price of the ordered items is 0.79. This positive coefficient indicates the positive correlation between sales and price of the ordered items. The coefficient of correlation between sales and order total cost is 0.99, a positive coefficient pointing to the positive corelation between sales and order total cost. The coefficient of correlation between sales and product price is 0.79, a positive coefficient and an indicator that sales have a positive correlation with the sales and order total cost. Finally, the coefficient of correlation between sales and Sales for each individual customer is 0.99, also a positive correlation coefficient. It shows that the sales data has positive correlation with the Sales for each individual customer.

3.4. Regression

The regression analysis was conducted with the regression model to determine the effect of each predictor variable on the dependent variable (Sales). In addition, the regression analysis also tests the significance of the coefficients in the decision on whether to reject or fail to reject the null hypothesis. The regression analysis result with the full model is shown in figure 12 below.

```
Call:
lm(formula = narrowed_data$data.Sales ~ narrowed_data$data.Product.Price +
    narrowed_data$data.Order.Item.Discount + narrowed_data$data.Order.Item.Product.Price +
    narrowed_data$data.Order.Item.Total + narrowed_data$data.Sales.Per.Customer,
    data = narrowed_data)

Residuals:
       Min         1Q     Median         3Q        Max
-7.871e-05 -1.260e-07  1.192e-06  3.112e-06  6.395e-05

Coefficients: (2 not defined because of singularities)
                                          Estimate Std. Error   t value Pr(>|t|)
(Intercept)                             -5.101e-15  2.711e-08 0.000e+00 1.000000
narrowed_data$data.Product.Price         1.519e-07  4.422e-08 3.435e+00 0.000593 ***
narrowed_data$data.Order.Item.Discount   1.648e-01  3.181e-08 5.181e+06  < 2e-16 ***
narrowed_data$data.Order.Item.Product.Price       NA         NA        NA       NA
narrowed_data$data.Order.Item.Total      9.075e-01  4.452e-08 2.038e+07  < 2e-16 ***
narrowed_data$data.Sales.Per.Customer             NA         NA        NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.152e-05 on 180515 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.535e+14 on 3 and 180515 DF,  p-value: < 2.2e-16
```

Figure 12: Regression Analysis Result

From the results of the regression analysis, the product price has a coefficient of $1.43e^{-07}$. This is a positive coefficient, and an indicator that the product price has a positive influence on the Sales. The regression of the sales on the product price is shown in figure 13 below.

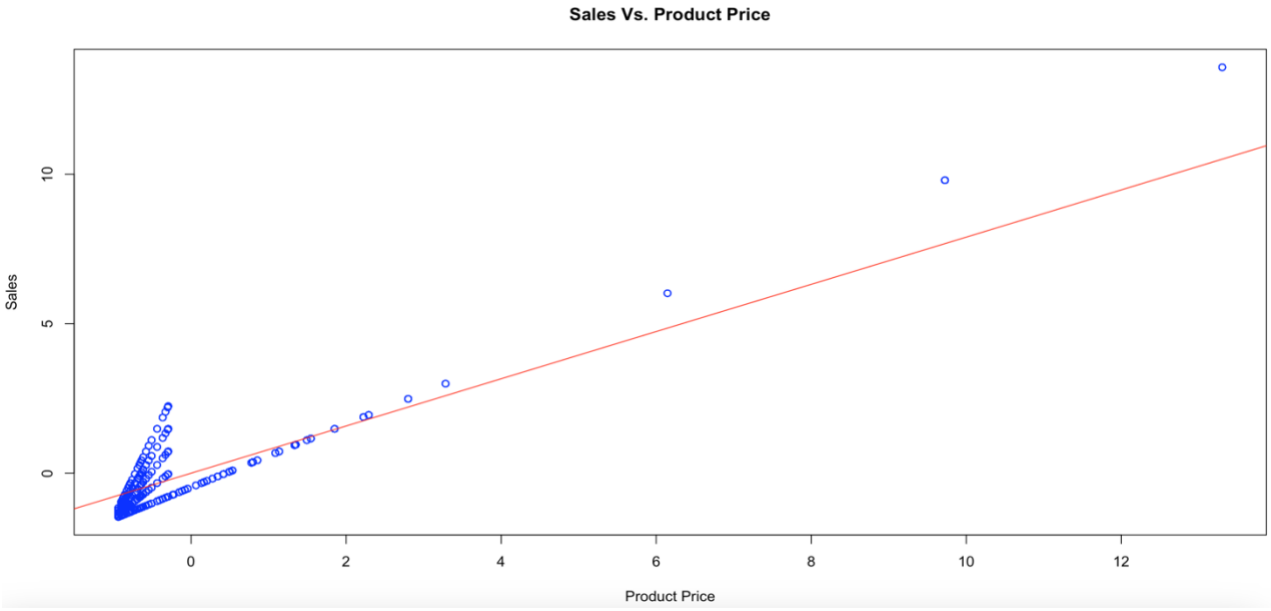

Figure 13: regression of Sales on Product Price

The product price further has a p- value of 0.000593 (< 0.05 at the 95% significance level). The p – value shows that the coefficient of the product price is statistically significant and that the null hypothesis (that the product price does not have a positive impact on the sales) should be rejected.

The order total cost has a coefficient of 1, a positive coefficient, and an indicator that the order total cost has a positive influence on the Sales. The regression of the sales on the order total cost is shown in figure 14 below.

Figure 14: Regression of Sales on Order discount

The Order total cost further has a p- value of 2.00^{e-16} (< 0.05 at the 95% significance level). The p – value shows that the coefficient of the Order total cost is statistically significant and that the null hypothesis (that the Order total cost does not have a positive impact on the sales) should be rejected. The discount has a coefficient of 1, a positive coefficient, and an indicator that the order discount has a positive influence on the Sales. The regression of the sales on the order discount is shown in figure 15 below.

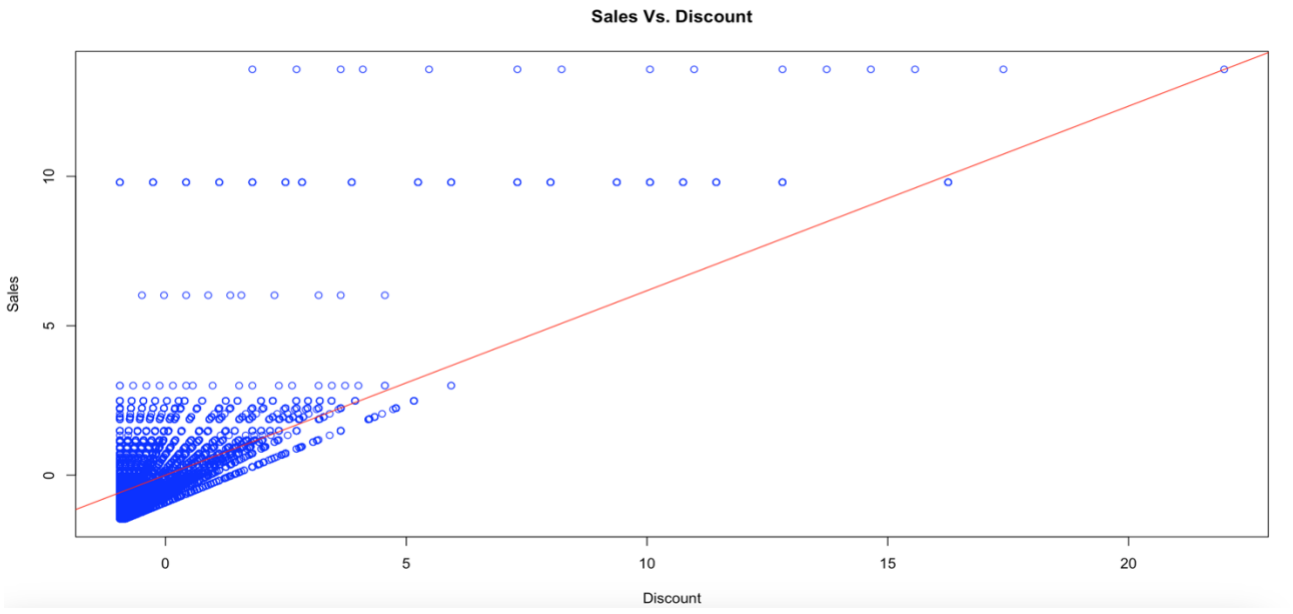

Figure 15: Regression of Sales on Discount

The Order discount further has a p- value of 2.00^{e-16} (< 0.05 at the 95% significance level). The p – value shows that the coefficient of the Order discount is statistically significant and that the null hypothesis (that the Order discount does not have a positive impact on the sales) should be rejected. Price of the ordered items and Sales for each individual customer, each have positive coefficients, but their values are negligible since they are too small and are therefore shown as NA. These coefficients show that the two variables, each has a positive influence on the sales, but with negligible strength. The p-values in both cases are also presented as NA, since they are too low and below 0.05. The tests are therefore statistically significant and the null hypotheses on the Price of the ordered items and Sales for each individual customer not having positive impact on the sales, are rejected. The ANOVA test in the regression process shows the result below.

```
> summary.aov(model)
                                      Df Sum Sq Mean Sq   F value Pr(>F)
narrowed_data$data.Product.Price       1 112647  112647 8.490e+14 <2e-16 ***
narrowed_data$data.Order.Item.Discount      1  12740   12740 9.602e+13 <2e-16 ***
narrowed_data$data.Order.Item.Total         1  55132   55132 4.155e+14 <2e-16 ***
Residuals                          180515      0       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: ANOVA

The ANOVA test shows that all the p values are less than 0.05, hence, the three null hypotheses are rejected.

The second regression analysis was done as an improvement of the regression model, retaining only the three predictor variables with strong effects on the sales (product price, order discount and order total cost). The result of the new egression is shown below.

```
Call:
lm(formula = narrowed_data$data.Sales ~ narrowed_data$data.Product.Price +
    narrowed_data$data.Order.Item.Discount + narrowed_data$data.Order.Item.Total,
    data = narrowed_data)

Residuals:
       Min         1Q     Median         3Q        Max
-7.871e-05 -1.260e-07  1.192e-06  3.112e-06  6.395e-05

Coefficients:
                                      Estimate Std. Error   t value Pr(>|t|)
(Intercept)                         -5.101e-15  2.711e-08 0.000e+00 1.000000
narrowed_data$data.Product.Price     1.519e-07  4.422e-08 3.435e+00 0.000593 ***
narrowed_data$data.Order.Item.Discount  1.648e-01  3.181e-08 5.181e+06  < 2e-16 ***
narrowed_data$data.Order.Item.Total     9.075e-01  4.452e-08 2.038e+07  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.152e-05 on 180515 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.535e+14 on 3 and 180515 DF,  p-value: < 2.2e-16

>
```

Figure 17: Regression with Improved Model

The ANOVA test for the improved regression model is shown below.

```
> summary.aov(model)
                                      Df Sum Sq Mean Sq   F value Pr(>F)
narrowed_data$data.Product.Price       1 112647  112647 8.490e+14 <2e-16 ***
narrowed_data$data.Order.Item.Discount      1  12740   12740 9.602e+13 <2e-16 ***
narrowed_data$data.Order.Item.Total         1  55132   55132 4.155e+14 <2e-16 ***
Residuals                           180515      0       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 18: ANOVA with the Improved Regression Model

In the improved regression model, all the three predictor variables have positive coefficients and therefore, positive impacts on the sales. The p-values in the regression and ANOVA results are all less than the alpha value of 0.05, leading to the rejection of all the null hypotheses.

4. DISCUSSION AND CONCLUSION

4.1. Summary

In the correlation analysis table and the regression results, all the predictor variables have positive correlations, and positive influences on the value of sales made in the business transactions. Since all the null hypotheses were rejected, the conclusion in this study is that the Product Price,
Order discount, Price of the ordered items, Order total cost and Sales for each individual customer, have positive impacts on the sales.

4.2. Prediction

The model can be used as an interface to predict sales with 95% confidence once given the parameters of product price, item order discount and item order total. Below is a user interface code that can be applied to a company's intranet to predict sales and help in business decisions. This is in line with current world views of using AI and machine learning to help business leaders.

4.3. Recommendation

The results of this study lead to the recommendation that companies need to focus on promoting the growth of sales by maximizing the four factors. However, the results of the improved model of regression shows that only three variables Product Price, Order discount and Order total cost have strong effects on the sales and needs to be prioritized. Naturally, it is perceived that consumers prefer lo prices for products, and therefore, high prices should lead to low sales. However, this study reveals that high prices cause higher sales. Perhaps the reason for this behaviour is that majority customers of associate higher product prices with higher product quality.

4.4. Future Studies

The current study was entirely quantitative and concentrated on the quantitative data in determining the factors affecting sales. Additionally, there are several variables that were excluded from the regression models, yet they have potential significant impacts on the sales. These include the shipping cost, Number of shipping days (real), Scheduled number of shipping days (scheduled), Benefit

realised from each order and Sales for each individual customer. It will be essential to include these in the future studies. At the same time, the future studies should include comparative analysis of sales, to compare the sales by different categories such as payment Type, Delivery Status, Category Name, customer state, Department Name, Market, Order City, and Order Country. Other important categories include the Order Region, Order State and Order Status.

BIBLIOGRAPHY

Ahmed, S., Chowdhury, B., Khalil, I., Haque, R., Senathirajah, A. R. B. (2022). Analysing The Factors Affecting Sales Performance Amongst Malaysian Smes: A Structural Path Modelling Approach. International Journal of eBusiness and eGovernment Studies, 14 (3), 560-577. doi:10.34111/ijebeg.202214127.

Al-Turjman, F. (2017). "Price-based data delivery framework for dynamic and pervasive IoT". Pervasive and Mobile Computing. 42: 299–316. doi:10.1016/j.pmcj.2017.05.001.

Cai, A., Liu, M., Liu, . (2022). "A methodology for evaluating salespeople performance considering efficiency and effect: A case study of a liquor company in China". Frontiers in Psychology. 13: 923198. doi:10.3389/fpsyg.2022.923198.

Djoni, D., Oktaviani, R. & Kirbrandoko, K. (2016). Factors that Affect the Sales Performance of PT SKP (A Case Study of Sales Force of Moorlife Indonesia in Jabodetabek). Indonesian Journal of Business and Entrepreneurship. 2. 122-129. https://www.researchgate.net/publication/307085954_Factors_that_Affect_the_Sales_Performance_of_PT_SKP_A_Case_Study_of_Sales_Force_of_Moorlife_Indonesia_in_Jabodetabek.

Fabra, N., & Reguant, M. (2020). A model of search with price discrimination. European Economic Review, 129, 103571. https://doi.org/10.1016/j.euroecorev.2020.103571.

Guizzardi, A., Pons, F. M. E., Angelini, G. & Ranieri, E. (2021). "Big data from dynamic pricing: A smart approach to tourism demand forecasting". International Journal of Forecasting. 37 (3): 1049–1060. doi:10.1016/j.ijforecast.2020.11.006.

Hutt, M. (2015). Cross-Functional Working Relationships in Marketing. Journal of the Academy of Marketing Science, 23(4), 351-357.

Johnson, E. R. (2021). "Dynamic Pricing Strategies for Perishable Products in Retail". Retail Analytics. 1 (1): 45–63. doi:10.1007/978-3-030-61367-6_3.

Le-Hoang, P. V. (2020). Factors affect customer satisfaction: the case of cargo delivery services. Independent Journal of Management & Production. 11. 1342.

https://www.researchgate.net/publication/343374983_Factors_affect_customer_satisfaction_the_case_of_cargo_delivery_services.

Priester, A., Robbert, T. & Roth, S. (2020). "A special price just for you: effects of personalized dynamic pricing on consumer fairness perceptions". Journal of Revenue and Pricing Management. 19 (2): 99–112. doi:10.1057/s41272-019-00224-3.

Talukder, K. & Jan, M. (2017). Factors Influencing Sales People's Performance: A Study Of Mobile Service Providers In Bangladesh. Academy of Marketing Studies Journal. 21. 1-20. https://www.researchgate.net/publication/329933654_FACTORS_INFLUENCING_SALES_PEOPLE'S_PERFORMANCE_A_STUDY_OF_MOBILE_SERVICE_PROVIDERS_IN_BANGLADESH.

Weinberg, C. B. (2015). An Optimal Commission Plan for Salesmen's Control Over Price. Management Science, 21(8), 937-943.

Xu, M., Tang, W. & Zhou, C. (2020). "Price discrimination based on purchase behaviour and service cost in competitive channels". Soft Computing. 24 (4): 2567–2588. doi:10.1007/s00500-019-03760-7.

Yu, T., Ruyter, K. D., Patterson, P., Chen, C. (2018), "The Formation of a Cross-Selling Initiative Climate and Its Interplay with Service Climate" (PDF), European Journal of Marketing, 52 (7/8): 1457–1484, doi:10.1108/EJM-08-2016-0487.

Zang, Z., Liu, D., Zheng, Y., Chena, C. (2020), "How do the combinations of sales control systems influence sales performance? The mediating roles of distinct customer-oriented behaviours", Industrial Marketing Management, 84 (1): 287–297, doi:10.1016/j.indmarman.2019.07.015.