Text

Description automatically generated with medium confidence

Logo

Description automatically generated

Final Project

By,

Tanisha S

Sharon Appoline Rosary

Vishaka Mohan Mohana

ALY6010- Probability Theory and Introductory Statistics

Instructor: Amin Karimpour

Date: December 12, 2021

Introduction

***Data Analysis in Sales Industry***

Today, many sales companies and industries use data analysis to analyse sales data to make better business decisions. The data gives an idea about the sales in the market and the product’s performance in the market. Data Analysis will help the company understand the pattern, and we can decide which market or product to focus on to increase the sales or profits of the company. Sales analysis is the process of analysing your data in order to assess your sales team's performance in relation to its objectives. It offers information on top-performing and underperforming products/services, selling issues and market opportunities, sales forecasts, and revenue-generating sales activities.

Regular sales data analysis gives a better grasp of what consumers are buying and helps figure out why they're acting the way they are. Patterns in lead conversions and drop-offs can also be discovered. All of these factors help improve sales process.

***Regression Analysis:***

Regression analysis is a set of statistical processes used in statistical modelling to estimate the relationships between a dependent variable (often referred to as the 'outcome' or 'response' variable) and one or more independent variables (often referred to as ‘predictors’ ,‘covariates’ , ‘explanatory variables’ , or ‘features’). Linear regression is the most frequent type of regression analysis, in which one finds the line (or a more sophisticated linear combination) that best fits the data according to a set of mathematical criteria.

Initial EDA

***Description of the Dataset***

The dataset contains numerical and categorical variables and contains 24031 rows. The dataset holds sales, profits, shipping cost, quantity entities across the segments such as Consumer, Corporate, and Home Office. Segments have multiple categories furniture, office supplies, Technology. Categories is again fragmented to subcategories which holds product categories where there will be difference in shipping charges and sales across market places and regions. Profits details is given for each product.

From the initial analysis we have mainly focused on employing inferential statistics and hypothesis testing based on p-values using t-tests. The key take-aways from the analysis are:

* We have strong evidence to support the claim that the average sales recorded in 2014 is greater than the average sales incurred in 2013 i.e., 20.16666 USD.
* When compared markets with maximum sales for same average profit, we failed to reject the null hypothesis since there wasn’t sufficient evidence to support the alternative hypothesis, i.e., the average profit of Asia Pacific and Europe aren’t same. Thus, the data support the assumption that both Asia Pacific and Europe have same average profit.

Thus, this project gave us an overall insight on how to perform hypothesis testing, gather evidence, frame questions, and decide whether to reject or fail to reject null hypothesis to answer those questions.

***Problem Statement:***

Performing regression analysisto understand whether the sales influence profit, Does discount have an influence on sales and Do profit and discount have an influence on sales.

***Library used:***

To perform the tasks, we have used packages as mentioned below:

*R code:*

A picture containing graphical user interface

Description automatically generated

***Import Dataset:***

Using read\_csv() function, We have imported the dataset and assigned to final\_project which holds approximately 24000 row data. Below is the screenshot for reference.

*R code:*



***Data Cleaning:***

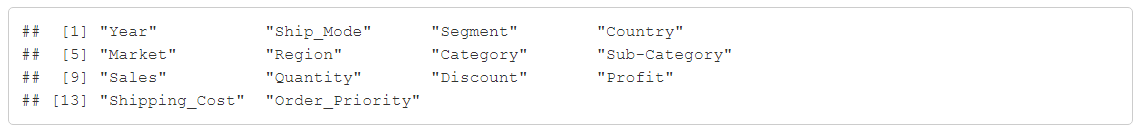
In this step, we have dropped unwanted variables from the dataset such as Row\_ID, order\_ID, Order\_Date, Ship\_Date, Customer\_ID, Customer Name, City, State, Postal code, product\_ID and Product Name.

*R code:*

Graphical user interface, application, Word

Description automatically generated

Below is the attributes used for regression analysis using names() function in R code.



And also, We have altered the necessary records using mgsub() function to gain a better understanding of the data, we have updated “EU” to “Europe”, “APAC” to “Asia Pacific” and “EMEA” to “Emirates” created a data frame for the updated dataset.

*R code:*

Chart, scatter chart

Description automatically generated

Below the dataset after data cleaning and altering necessary records.

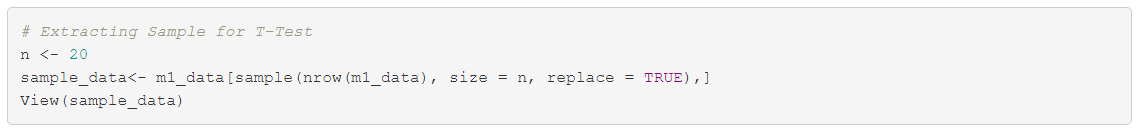
Graphical user interface, application

Description automatically generated

Analysis

***Random sample from dataset:***

In this step, we have selected random data (sample size=20) from the dataset which contains 20 rows of the dataset for hypothesis testing. Below is the R code for sample extraction with sample size, n=20:

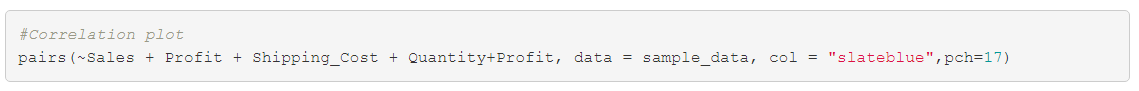


Below is the sample data which contains 20 rows from the dataset.

Table, calendar

Description automatically generated with medium confidence

A correlation matrix graph can be visually presented for highlighting the relationship / correlation between variables.



Calendar

Description automatically generated

In the above plot:

• The distribution of each variable is shown on the diagonal.

• On the bottom of the diagonal, the bivariate scatter plots with a fitted line are

displayed.

• As the correlation coefficient increases, the observations group closer

together in a linear shape.

• There is a positive correlation between Profit and shipping cost and sales and profit. And we also notice that there is positive relationship between all the attributes.

***Q1 : Does sales influence profit?***

In this task we are validating correlation between sale and profit by plotting a scatter plot using ggplot() function. By regression model understanding whether sale impacts the profit.

R code:

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Based on the criteria listed, the value of r in this case (r= 0.68) indicates that there is a positive, linear relationship of moderate strength between sales and profit hence we can conclude that sale influence profit from the scatterplot.

The hypothesis can be given as:

Null Hypothesis: H0: β1=β2

Alternate Hypothesis: H1: βj ≠0 for at least one j≠0

Below is the R code for creating regression model using lm() and summary() for Profit and Sales. Here Sales is independent variable or predictor while Profit is dependent variable or response variable.



*Output:*

Text

Description automatically generated

From the above model,

* The final equation will be **profit=0.20617\*sales-15.78206**.
* The above equation tells us that for every 1000 sales there will be an increase in profit by 0.20617.
* The regression coefficient is 0.20617, which is positive and indicates a positive relationship.
* The regression results are statistically significant at 99% confidence interval.
* The p-value is 5.93e-06 <<< significance level hence, we have enough evidence to reject the null hypothesis. Hence, Profit has significant influence on Sales.

***Q2: Does discount have an influence on sales?***

In this task we are validating correlation between discount and sales by plotting a scatter plot using ggplot() function. By regression model understanding whether discount impacts sales.

*R code:*

Graphical user interface, text, application

Description automatically generated

*Output:*

Chart

Description automatically generated

Based on the above plot we can see that the data doesn’t resemble any kind of pattern which tells us that there is no relationship between sales and discount.

The hypothesis can be given as:

Null Hypothesis: H0: β1=β2

Alternate Hypothesis: H1: βj ≠0 for at least one j≠0

Below is the R code for creating regression model using lm() and summary() for Discount and Sales. Here Discount is independent variable or predictor while Sales is dependent variable or response variable.



*Output:*

Text

Description automatically generated

From the above regression model,

* The p-value is 0.29>>significance level, hence there is not enough evidence to reject the null hypothesis. This indicates that there is no relationship between discount and sales.

***Q3: Do profit and discount have an influence on sales?***

In this task we are validating correlation between sales and profit, discount by plotting a scatter plot using ggplo0074() function. By regression model understanding whether profit and discount impacts the sales.

*R code:*

Graphical user interface, text, application

Description automatically generated

*Output:*

Chart, line chart, scatter chart

Description automatically generated

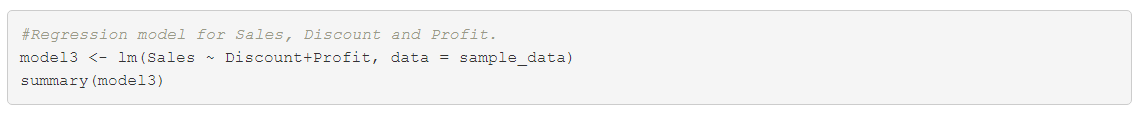
Based on the criteria listed, the value of r in this case (r= 0.74) indicates that there is a positive, linear relationship of moderate strength between sales and profit+discount hence we can conclude that profit and discount together influence sales from the scatterplot.

The hypothesis can be given as:

Null Hypothesis: H0: β1=β2

Alternate Hypothesis: H1: βj ≠0 for at least one j≠0

Below is the R code for creating regression model using lm () and summary () for Discount+Profits and Sales. Here, Discount and Profit are independent variables or predictors while Sales is dependent variable or response variable.



*Output:*

Text

Description automatically generated with low confidence

From the above model,

* The final equation will be **sales=571.33\*Discount+3.91\*Profit+35.19.**
* The coefficients are positive which indicates a positive relationship between Sales and Profits+Discount.
* The p-value is 1.0421e-05 << significance level and hence we have enough evidence to reject the null hypothesis. Thus, Discount and Profit together have significant influence on Sales. We can also conclude that the regression results are statistically significant at 95% confidence interval.

Conclusion

From the above analysis we can conclude the following things:

* Profit and Sales have a positive relationship as the data is showing an uphill pattern and thus, if there is an increase in sales profits will also increase. The p-value<significance level which also tells us that there is a positive relationship between profit and sales.
* Profit and Discount together has a positive influence on Sales. The p-value<< significance level hence it concludes that these is a positive relationship among the variables. Even from the scatterplot we can conclude the same as the data has an uphill pattern.
* Discount alone has no relationship with sales, we can conclude this by looking at the scatterplot which shows no pattern among the variables. Also, the p-value > significance level which indicates that there is no relationship between them.

Bibliography

1. Bluman Elementary Statistics: A Step by Step Approach 10th edition, McGraw Hill
2. R. Kabacoff,  R in Action 2nd edition, Manning
3. Miller, S Reading a Regression Table: A Guide for Students. (n.d.). Steven v. Miller. [*http://svmiller.com/blog/2014/08/reading-a-regression-table-a-guide-for-students/*](http://svmiller.com/blog/2014/08/reading-a-regression-table-a-guide-for-students/)
4. Linear Regression Example in R using lm() Function – Learn by Marketing. (n.d.). [*https://www.learnbymarketing.com/tutorials/linear-regression-in-r/\*](https://www.learnbymarketing.com/tutorials/linear-regression-in-r/\)
5. finnstats. (2021, May 13). Correlation Analysis Different Types of Plots in R
6. [*https://www.r-bloggers.com/2021/05/correlation-analysis-different-types-of-plots-in-r/*](https://www.r-bloggers.com/2021/05/correlation-analysis-different-types-of-plots-in-r/)