Text

Description automatically generated with medium confidence

Logo

Description automatically generated

Final Project Milestone 2

By,

Tanisha Srivastava

Sharon Appoline Rosary

Vishaka Mohan Mohana

ALY6010- Probability Theory and Introductory Statistics

Instructor: Amin Karimpour

Date: November 27, 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 these factors help improve sales process.

***Hypothesis Testing***

In statistics, the purpose of statistical inference is to draw conclusions about a population based on data obtained from a sample of that population (Davis R, 2006). The strength of evidence from the sample can be evaluated by hypothesis testing. In simple words, hypothesis tests are used to validate whether a hypothesis about a population parameter is true or not. A random sample from the population is obtained and a hypothesis test on the sample data can be performed for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from which the sample was drawn (Zach H, 2021).

*T-test and its types:*

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. The t-test is one of the tests used for the purpose of hypothesis testing in statistics. There are several different types of t-test that can be performed depending on the data and type of analysis required.

There are three main types of t-test:

* If the groups come from two different populations (e.g., two different variants of car, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test).
* A Paired sample t-test compares means from the same group at different times (e.g. measuring before and after an experimental treatment).
* If there is one group being compared against a standard value (e.g., comparing the mean height of male in the U.S to a height of 75 inches), perform a one-sample t-test.

***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.

***Problem Statement:***

The agenda of this project is to compare sales during 2013 & 2014, compare profits of markets with maximum sales and compare shipping cost of standard class and second-class shipping mode by performing hypothesis testing using inferential statistics.

***Library used:***

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

*R code:*

Graphical user interface, text, application, email

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

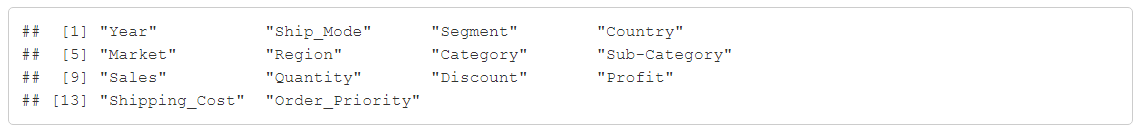
* In this step, we have dropped unwanted variables from the dataset such as Row\_ID, order\_ID, Order\_Date, Ship\_Date, Customer\_ID, Customer Name, City, State, Postal code, product\_ID and Product Name.

*R code:*

Graphical user interface, application, Word

Description automatically generated

*Output:*

- 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 is the dataset after dropping unwanted variables from the dataset.

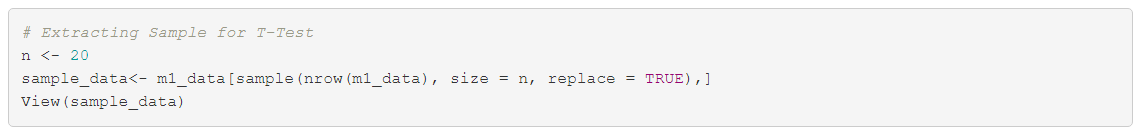
Graphical user interface, application

Description automatically generated

AnalysisSection

*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

*Normal Distribution check:*

In most cases, visual examination, as detailed in the preceding section, is unreliable. A significance test may be used to determine if data exhibit a significant departure from normalcy by comparing the sample distribution to a normal distribution. When it comes to data analysis, it's common to focus on the features of a specific set of statistics, observations, or measurements. We might be interested in knowing the center and dispersion of this core value. We could wish to look at outliers (extreme numbers) or investigate the distribution or trend of the data values. There are several graphs to explore the distribution.

*Normality test for Sales distribution:*

*R code:*

Graphical user interface, application

Description automatically generated



*Output:*

Chart, line chart

Description automatically generatedChart

Description automatically generated

***Observation:*** The first figure indicates normal distribution for sales. We can notice that bell shaped distribution constituting normal distribution. In the figure 2, This is a plot of the inverse of the standard normal cumulative versus the sales observations. The underlying distribution of the data is normal, the points fall along the regression line(y=x). Stragglers at either end of the normal probability plot are outliers. Curvature at both ends of the plot indicates long or short distribution tails. Hence, we can conclude that the sample of sales is normally distributed.

*Normality test for Profit distribution:*

*R code:*

Graphical user interface, text

Description automatically generated



*Output:*

Chart, line chart

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

***Observation:*** The first figure indicates normal distribution for Profit. We can notice that bell shaped distribution is present constituting normal distribution. In figure 2, This is a plot of the inverse of the standard normal cumulative versus the profit observations. The underlying distribution of the data is normal, the points fall along the regression line(y=x). Hence, we can conclude that the sample of profit is normally distributed.

*Normality test for Shipping Cost distribution:*

*R Code:*

Text

Description automatically generated



*Output:*

Chart, line chart

Description automatically generatedChart

Description automatically generated

***Observation:*** The first figure indicates normal distribution for shipping cost. We can notice that bell shaped distribution is present therefore we can conclude that the sample is normal distribution. In the figure 2, This is a plot of the inverse of the standard normal cumulative versus the shipping cost observations. The underlying distribution of the data is normal, the points fall along the regression line(y=x). Hence, we can conclude that the sample of shipping cost is normally distributed.

1. ***The company claims that the average sale in 2014 has improved as compared the average sale is 2013. Is it true?***

Average sales for 2013: The below R code is used to get average sale for the year 2013 using dplyr(), knitr() functions.





The below R code filters records for the year 2014 using dplyr() function.



In this task we are performing hypothesis testing using t.test() function to check average sale during 2014 is greater than average sale during 2013 with 95% confidence level and 5% alpha.

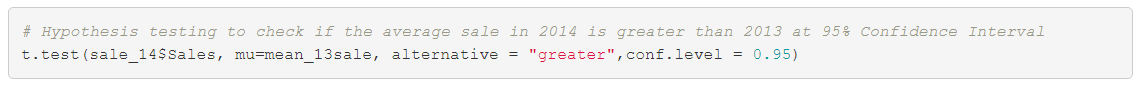
The null and alternative hypothesis can be given as,

*Null Hypothesis* - H0: µ1 ≤ µ2

*Alternate Hypothesis* - H1: µ1 > µ2

where µ1 and µ2 are mean sales recorded during 2014 and 2013.

*R Code:*



*Output:*

Text, letter

Description automatically generated

we can infer the following:

* t is the t-test statistic value (t = 14.595),
* df is the degrees of freedom (df= 8360),
* p-value is the significance level of the t-test (p-value = 2.2^{-16}).
* conf.int is the confidence interval of the mean at 95% (conf.int = [292.2287, ]);
* sample estimates are the mean value of the sample (mean = 301.3826).

From the above result, we can conclude that statistically significant p-value 2.2^{-16} is less than 0.05 (2.2^{-16} ≤ 0.05). It provides significant evidence against the null hypothesis since the null hypothesis has a less than 5% chance of being right (and the results are random). As a result, the null hypothesis is rejected, and the alternative hypothesis is accepted. At 95% confidence level, mean of sales during 2014 is significantly greater than mean of sales during 2013 with a p-value 2.2^{-16}.

1. ***Is the average profit same for markets with maximum sales?***

In order to get an insight on sales per market, we use describeBy() function to get descriptive statistics of sales across markets.

*R Code:*



*Output:*

Table

Description automatically generated

In next step, we are plotting bar plot to visually analyse markets with maximum sales.

*R Code:*

Graphical user interface, text, application

Description automatically generated

*Output:*

Chart, bar chart

Description automatically generated

***Observation:*** From the above graph we can notice that Asia Pacific has maximum sales compared to Emirates and Europe. Asia Pacific has 20% more sales compared to Europe market. Emirates Market had the least sales.

Below is the R code for filtering profits across marketplace for analysis:

Text

Description automatically generated with medium confidence

*Output:*

A picture containing table

Description automatically generatedText

Description automatically generated

The above two tables depict the profits across Asia Pacific and Europe markets. A Density-plot to check normality of profit distribution for Asia Pacific and Europe is shown below.

Chart, line chart, histogram

Description automatically generated

***Observation:*** In the above density graph, the purple line shows the distribution of profits in Asia pacific and blue line shows the distribution of profits in Europe. We can also see that there exist negative profits. The cumulative density is more between the range -25 to 25.

In the below task, we are performing hypothesis testing using t.test() function to compare mean profits for Asia pacific and Europe markets with 95% confidence interval and 5% alpha.

The null and alternative hypothesis can be given as,

*Null Hypothesis*: H0: µ1=µ2

*Alternate Hypothesis*: H1: µ1 ≠µ2

where µ1 and µ2 denotes mean profit of Asia Pacific and Europe.

*R code:*



*Output:*

Text, letter

Description automatically generated

we can infer the following:

* t is the t-test statistic value (t = -0.27733),
* df is the degrees of freedom (df= 14.904),
* p-value is the significance level of the t-test (p-value =0.7853).
* conf.int is the confidence interval of the mean at 95% (conf.int = [-58.00896, 44.65796]);
* sample estimates are the mean values of the sample (mean of Asia Pacific= 14.2680, mean of Europe = 7.5925).

***Observation:*** From the above result, it does not provide enough evidence to reject the null hypothesis since statistically significant p-value 0.7853 is greater than 0.05 (0.7853>= 0.05). As a result, we fail to reject the null hypothesis, and the alternative hypothesis is rejected. Hence at 95% confidence level, we can support the assumption that both Asia Pacific and Europe have same mean profit.

1. ***Is the mean shipping cost for second class lesser than mean shipping cost for standard class mode?***

First step is to validate distribution across ship mode using Density-plot. Below is the R code where we are validating shipping cost distribution for each ship mode.

*R Code:*

Graphical user interface, text, application

Description automatically generated

*Output:*

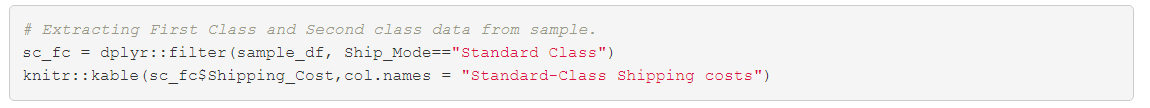
Chart, histogram

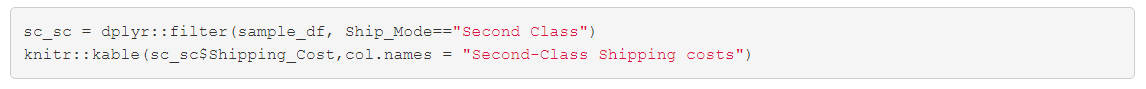
Description automatically generated

***Observation:*** In the above density-plot graph we can notice that bell shaped distribution is present for all the shipment mode. we can conclude that the sample is normal distributed and that the density is deep for standard class than second class.

Next step is to filter Standard-class and second-class ship mode data from the sample. Using dplyr (), filter() and knitr() functions we can shipping cost for each ship mode.

*R code:*





*Output:*

A picture containing table

Description automatically generated

Shape

Description automatically generated

Next step is to check if shipping costs for second-class are lesser than standard-class.

In the below task, we are performing hypothesis testing using t.test() function to check shipping costs for second-class are lesser than standard-class with 95% confidence interval and 5% alpha.

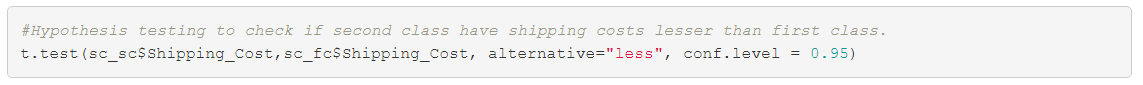
The null and alternative hypothesis can be given as,

*Null Hypothesis*: H0: µ1 ≥ µ2

*Alternate Hypothesis*: H1: µ1 < µ2

where µ1 and µ2 denotes mean shipping cost of second-class and standard-class mode.

*R code:*



Text

Description automatically generated

we can infer the following:

* t is the t-test statistic value (t =0.45906),
* df is the degrees of freedom (df= 1.2441),
* p-value is the significance level of the t-test (p-value =0.6433).
* conf.int is the confidence interval of the mean at 95% (conf.int = [ , 111.3645]);
* sample estimates are the mean values of the sample (mean of first class= 33.9100, mean of second class =23.75533).

***Observation:*** From the above result, it does not provide enough evidence to reject the null hypothesis since statistically significant p-value 0.6433 is greater than 0.05 (0.6433 >= 0.05). As a result, we fail to reject the null hypothesis, and the alternative hypothesis is rejected. Hence at 95% confidence level, the shipping costs for second class of the population is significantly greater than the shipping costs for standard class with p-value 0.6433.

Summary

The report mainly focused on framing questions about the data, employing inferential statistics and hypothesis testing to find answers for the framed questions. The hypothesis testing was performed 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.
* When compared shipping costs for both second class and standard class mode, we failed to reject the claim that second class contains shipping cost lesser than standard class. Thus, we can support the assumption that the shipping costs for second class are more than that of standard class.

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.

Bibliography

[1] Zach, H. (2021, April 29). Introduction to Hypothesis Testing. Statology. <https://www.statology.org/hypothesis-testing/>[1] Zach, H. (2021, April 29). Introduction to Hypothesis Testing. Statology. <https://www.statology.org/hypothesis-testing/>

[2] Davis, R. B., & Mukamal, K. J. (2006). Hypothesis Testing. *Circulation*, *114*(10), 1078–1082. <https://doi.org/10.1161/circulationaha.105.586461>

[3] Bevans, R. (2020, December 14). *An introduction to t-tests*. Scribbr. [https://www.scribbr.com/statistics/t-test/Links to an external site.](https://www.scribbr.com/statistics/t-test/)

Appendix

We are attaching an additional file which contains the source R code and the data set used for analysis in this report. We have attached the code file R Script\_milestone2.R and data set final\_project.csv.