**100594631**

The first dataset has passed through prior preparation and cleaning to suit the assessment.

**BRIEF EXPLORATION OF THE DATASET**

In the ‘Segment’ column, which contains the customer targets ‘Consumer and Corporate’, it can be seen that **Consumer** have more orders, **25051** while **Corporate** customers have **23538** orders. Percentage-wise(relative frequency), Consumer have 52% while Corporate have 48% as seen in the table.

Looking at the Quantity column, which contains the number of products each customer orders for, it can be seen from the frequency distribution that it is highly skewed to the left, meaning most customers tend to order for low quantity of goods from 1 to 4, instead of going for higher quantities like 7 and above.

As seen from the graphs (histogram and boxplot), it can be deduced that the **Quantity** of ordered goods is highly right-skwed, meaning there are most customers makING orders of between 1 to 4, then abrubtly reduced towards 10, with the average total orders lies close to 2.5 and many outliers as indicated in the boxplot.

**The effect of Discout on Sales**

it is generally perceived that granting discount on goods bougth always have positive inpact on the sales made by customers, below are the tables of sales made with and without discount in different reqions concerned:

As seen from the tables, only Southeast Asia and EMEA have great impact of the discount on their total cost as they both have tangible reduced cost displacing that of two other countries each.

The correlation coefficient of Discount and Total Sales is **-0.06087718.**  This is somewhat against the general believe as it shows that the discount has a negative and near-average average effect on sales, which means that the more discount given to customers, the less they make orders for goods.

The scatterplot below says it all, there is no regular pattern between Sales and Discount.

**Verdict:** And as seen, it is advisable for the store to discontinue the granting of discounts to customers as it provides a negative effect.

**STATISTICAL TESTS**

**ONE-SAMPLE T-TEST**

**Problem Statement**

Prior to the introduction of granting discounts to customers, the business owner recorded an average sale of **1,300**, which was expected to increase after the discount was introduced, hence the need to carry out a statistical test to determine whether such hypothesis is true or otherwise, using 0.05 significance level and assuming the TotalSales is normally distributed:

**Hypothesis**

Null Hypothesis (): Total Sale mean is 1,300 (TotalSale = 1,300)

Alternate hypothesis (): Total Sale is not 1,300 (TotalSale != 1,300)

**p-value**: The p-value is observed to be **3.765**, smaller than 5% significance threshold; as a result, the NULL hypothesis is considered to be rejected.

**95 percent confidence range:** The range for my test's 95 percent confidence level is 1176.238 to 1237.999.

**Mean of X:** The mean of Totalsale in the given sample data is **1207.119.**

**Inference:** After carrying out the T-Test, I got the P-Value to be much lower than 0.05. As a result, the NULL hypothesis which puts the mean sales of the Super store within the period covered at **13,000** is **REJECTED**. Hence, I come to a conclusion that granting Discounts to customers has no effect on sales as the 95% confidence interval puts the hypothesized mean between 1176.238 and 1237.999. i.e discount does not boost sales as expected.

So generally, the Discount has no statistically significant on the total sales.

**TWO-SAMPLE T-TEST**

**Problem Statement**

Considering the population, the business owner assumed that the mean value in the Segment column, which contains the Consumer and the Corporate are equal. A paired T-test will be carried out to confirm if this claim is true or otherwise.

**Hypothesis**

Null Hypothesis () : Consumer mean is same as Corporate mean (difference = 0)

Alternate hypothesis () : Consumer mean is **not** same as Corporate mean (difference != 0)

**p-value**: The p-value is observed to be **0.5675**, which is above the 5% significance level, therefore I refused to reject the NULL hypothesis.

**95 percent confidence range:** The range of our test's results from -79.81753 to 43.76482 is the 95 percent confidence interval. The null hypothesis cannot be ruled out for postulated means between -79.81753 and 43.76482 at the 5% level of significance.

**Inference**

After carrying out the T-Test, I got the P-Value to be much higher than 0.05.

Consequently, I am unable to disprove the NULL hypothesis that the mean sales of Consumers, is the same as the mean sales of Corporate as there is no enough prove to ascertain this claim, hence we cannot come to a conclusion that the difference between the two variables is not zero

So generally, it is statistically significant that mean value possessed by both variables are equal.

**CHI-SQUARE TEST**

**Problem statement**

Still on the Superstore dataset, the store owner was confused about the choice of shipping mode based on each region, he wishes to know if there is any pattern or significant relationship between Ship mode and Region where his customers are based.

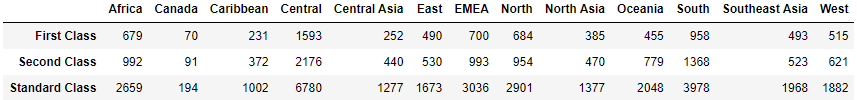
Hence, I test on 5% significance level whether or not there is a relationship between Discount and Sales.

**Hypothesis**

Null Hypothesis (): Ship mode is independent of Region

Alternate hypothesis (): There is a relationship between Ship mode and Region

**The ShipMode and Region contingency table**



**P-value:** The Chi-squared value gotten is 111.84, while the p-value is **2.688**, what is extremely less than the specified 0.05 significance threshold, consequently, I refused to accept the NULL hypothesis that both variables, Ship mode and Region are not dependent on each other.

**Inference**

I can assert that there is substantial proof that the association between Ship mode and Region is statistically significant as the p-value gotten is less than the 5% significance level.

**KRUSKAL WALIS TEST**

**Problem Statement**

Without assuming the data to have normal distribution, a comparison test is to be carried out on the Order Priority which contains four variables, **Critical, High, Low and Medium** as against the Total price of goods bought to determine if they have identical distribution, this will help the store determine if there is any category that requires more attention.

**Hypothesis**

Null Hypothesis () : same median ranks(identical distribution)

Alternate hypothesis () : at least one of the median rank is different(unidentical distribution)

**p-value**: The p-value is observed to be **0.2**, which is above the 5% significance level, therefore I refused to reject the NULL hypothesis.

**Inference**

After carrying out the Kruskal Wallis Test, I can conclude that because the p-value is below the significance threshold of 0.05, there are no discernible differences between the fore groups.

Therefore, I am unable to disprove the NULL hypothesis that the difference in TotalSale in the four groups is Zero

So generally, the Null Hypothesis () is statistically significant.

NOTE: there is no need for further **Pairwise Comparison** since the result shows that there is no significant difference between the four categories.

**ANALYSIS OF VARIANCE(ANOVA)**

**Problem statement**

A popular Pastry Chef in the city center, who has been making delicious Meat pie over the years approached me, he basically produces **three** set with different recipes, due to lack of knowledge, he couldn’t separate the sales of all set of meat pies, so he wishes to know if there is actually a difference between the popularity of the piece of snacks in order for him to know which one to focus one.

In this regard, I would deploy an Analysis of Variance test in order to determine if indeed there is at least two difference between the mean sales of the three kind of meat pies he makes.

**NOTE: problem statement and sample dataset are randomly made up by me**

**Hypothesis**

Null Hypothesis () : MeatPie\_A = Meatpie\_B = Meatpie\_C (difference = 0)

Alternate hypothesis () : Consumer mean is **not** same as Corporate mean (difference != 0)

**p-value**: The p-value is observed to be **0.591**, which is above the 5% significance level, therefore I refused to reject the NULL hypothesis.

**f-value:** The greater the value of the f-value, the likelihood of rejecting the null hypothesis increases. The f-value in this instance is 0.529, which is low.

**Inference**

After carrying out the ANOVA, I got the P-Value to be much higher than 0.05, and f-Value also small.

Therefore, I am unable to disprove the NULL hypothesis that the mean sales of the meatpie are equal, or that one of the varieties is the most popular as there is no enough prove to ascertain this claim.

In other words, the choice of recipe used in preparing each type of meatpie has no effect on the sales, so the pastry Chef can continue to do the business not minding which one is most popular as we cannot come to a conclusion that the difference between the three varieties is not zero

So generally, the Null Hypothesis () is statistically significant.

**WILCOXON SIGNED RANK TEST**

**Problem Statement**

The plots, QQ and Kernel Density gave justification to the fact that the distiribution is heavily skewed to the left, i.e it is not normally distibuted.

To further check for the normality of the distribution, a **Shapiroko-Wilk test** was conducted using a 95% confidence interval. After the test, a P-value of **2.2**, which is extremely smaller than the 5% significance level.

This brought me to the conclusion that the distribution is **NOT normally distributed.**

Therefore I can go ahead to check the median of the distribution if its 18000 has its meant to be believed.

**Hypothesis**

Null Hypothesis (): The median(mean) is 18000

Alternate hypothesis (): The median(mean) is not 18000

**p-value**: The p-value is observed to be **0.0009131**, which is less than 5 percent level of significance, the NULL hypothesis is consequently deemed to be rejected.

**Inference:** After carrying out the Wilcoxon signed rank test, I got the P-Value to be lower than 0.05. As a result, the NULL hypothesis which puts the mean value at **13,000** is **REJECTED**.

So generally, the Discount has no statistically significant on the total sales.

**LINEAR REGRESSION**

**Problem statement**

Using the **Sales** as the target variable, a linear regression test is be carried out to check if the independent variables have significant predicting effect.

**Hypothesis**

Null Hypothesis (): Target variables are not impacted by the predictors.

Alternate hypothesis (): Target variables are impacted by the predictors.

**Linear model Equation:** **Sales = 39.00 + 0.37Segment – 0.53Ship.Mode + 0.59Region \_ 0.38Order.Priority + 66.67Quantity – 186.96Discount**

The above equation implies that when all independent variables equal zero, the target(Sales) variable would equal 39.00, which is the intercept of the model.

**p-value**: The p-value is observed to be **2.2**, which is less than 5 percent level of significance, the NULL hypothesis is consequently deemed to be rejected.

**Inference:** After carrying out the Significance test for linear regression, I got the P-Value to be lower than 0.05. As a result, the NULL hypothesis which says the predictor variables have no influence on the target variablesis **REJECTED**.

So generally, the independent variables have statistical significance on the dependent variable.

**CONCLUSION**

Prior the conduction the statistical tests, a brief exploration of the major dataset was done, this helps in uncovering hidden insights and patterns that helped in approaching the tests using the best methods.

The tests were properly carried out giving the problem statement, the hypothesis, break down of the result and lastly stating the inference and discoveries. This test as exposed me to a lot of theories and approaches in the field of statistics and R-programming language inclusive.

**REFERENCE**

Chowdhury, V. (2022) *Superstore dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final (Accessed: January 4, 2023).

*Elementary statistics with R* (no date) *Elementary Statistics with R | R Tutorial*. Available at: https://www.r-tutor.com/elementary-statistics (Accessed: January 4, 2023).

Bevans, R. (2022) *Choosing the right statistical test: Types & examples*, *Scribbr*. Available at: https://www.scribbr.com/statistics/statistical-tests/ (Accessed: January 23, 2023).

*Learn R Programming* (2021) *GeeksforGeeks*. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/learn-r-programming/ (Accessed: January 23, 2023).