**REPORT ON HYPOTHESIS TESTING FOR A CAR SHARING SERVICE**

PROBLEM STATEMENT

Electric car sharing is a system of taking and returning cars according to the slots available. The cars are in specified postal code areas where the public can access them.

As a data scientist, I have been approached by the company to analyse data about car sharing activities. An investigation has to be carried out concerning a claim regarding blue cars in the service. This will surely require a hypothesis test. The following is the outline of this process.

WHAT IS THE CLAIM?

The null hypothesis will be derived from a claim about the blue cars. There is a need to first conduct a descriptive analysis of the data in order to come up with a claim.

IMPORTANCE OF HYPOTHESIS TESTING

It provides statistical evidence for a claim. A claim that has been statistically approved can be incorporated within the marketing or business structure in order to maximise profits. It also minimises risk.

DATA DESCRIPTION

Data collection: data was carried out by the company itself. I was not involved in the collection. It is assumed that the data was continuously collected while business was ongoing ie, information was logged in every time cars were taken or returned.

Data sourcing: data was obtained from an open source.

The link to the dataset and its description of columns can be found here: <http://bit.ly/DSCoreAutolibDataset> and [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary)

Data understanding: There are 1605 rows and 13 columns, they are;

Postal code 0

date 0

n\_daily\_data\_points 0

dayOfWeek 0

day\_type 0

BlueCars\_taken\_sum 0

BlueCars\_returned\_sum 0

Utilib\_taken\_sum 0

Utilib\_returned\_sum 0

Utilib\_14\_taken\_sum 0

Utilib\_14\_returned\_sum 0

Slots\_freed\_sum 0

Slots\_taken\_sum

Some columns needed to be dropped as they are not relevant to the analysis.We retained postal code,date,day, day type,blue cars returns, daily points and blue cars taken.

DATA ANALYSIS

Complete analysis located on google colab notebook savd in the github repository that can be accessed through this link

was carried out in three steps which are cleaning,exploratory data analysis and hypothesis testing.

After importing libraries and loading dataset, data cleaning carried out by:

-renaming columns

-dropping columns

-checking for duplicates

-checking for outliers

-checking for null values

Exploratory data analysis involved univariate and bivariate analysis.

Univariate analysis

Value count - monday(2374), calculating means, mode, median, skewness, kurtosis, quartiles and range of specified variables.

Visualisation through bar charts and pie chart

Bivariate analysis was carried out through plotting scatter plots, line graphs, heat map and combined bar chart. Scatter plots were challenged using pearson correlation coefficient. It revealed a strong correlation between blue cars taken and blue cars returned.

It is important to note that the variable of interest was NUMBER OF BLUE CARS TAKEN

In univariate analysis, the mean was found to be 125.92695.This was our claim

**HYPOTHESIS TESTING PROCEDURE**

1.State the null and alternate hypothesis

2.Get sample data

3.Identify test statistic

4.Set alpha

5.Obtain critical z value

6 Analyse data

7. Get p value

8. Analyse results

9.Conclusion

1.State the null hypothesis statements.

As mentioned, the mean of the blue cars returned was found to be 125.9269505750699

Ho : mean of blue cars taken = 125.92695(claim)

H1 : mean of blue cars taken != 125.92695

2.Get sample data

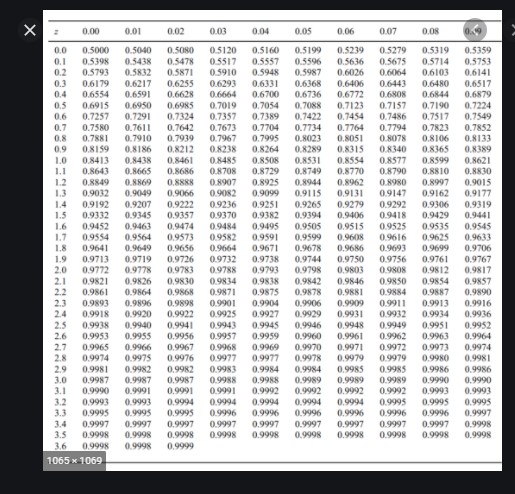
Simple random sampling was incorporated and an unbiased sample data of 40 rows was obtained.

3.Identify test statistic.

Z Test will be best to use because the sample size is 40 rows. The recommended number is any number more than 30.

4.determine alpha level

alpha = 0.05, for a two tail test, then we divide this by two. alpha is now 0.025

5.Obtain critical z value

From the z score, 0.025 corresponded with a critical value of + or - 1.96 as shown

6.Analyse data

These were the results of the analysis carried out with python

Sample mean = 106.2

Population mean(hypothesised) = 125.92695057506994

Standard deviation of population = 185.42657876727893

7.Obtain p-value

Using N=40, the p-value was calculated using the formula

Z Test = (x̄ – μ) / (σ / √n)

The results is -0.6728495507117408

8.Analyse results

* critical value is -1.96
* observed value -0.672849

The observed p-value is within range of the critical value.

TEST SENSITIVITY

How does sample size affect the results of the test?

Changing the sample size does not alter the results, this is demonstrated in the google colab notebook.

CONCLUSION

There is now statistical evidence to guide my decision, I FAIL TO REJECT THE NULL HYPOTHESIS.

Report compiled by,

Velvin samkul.