**HYPOTHESIS TESTING FOR AN AUTOLIB DATASET:**

**PROBLEM STATEMENT**

Hypothesis testing is an act whereby an assumption is tested regarding a population parameter. The test tells the statistician or the data analyst or the data scientist whether or not his primary hypothesis is true.

This dataset was an autolib dataset that contained details about the operation of cars within Paris. It showed a compilation of dates when the blue cars were picked from and returned to the particular addresses. The claim being investigated was whether or not the average number or blue cars taken was different from the average number of blue cars returned during the weekdays.

The null hypothesis is the hypothesis the statistician believes to be true. The alternative hypothesis to be untrue, making it effectively the opposite of a null hypothesis.

The null hypothesis is that the number of blue vehicles taken in a weekday are equal to the number of blue vehicles returned.(Null Hypothesis -> Ho: u1 = u2)

The alternative hypothesis is that the number of blue vehicles taken in a weekday are not equal to the number of blue vehicles returned.

Hypothesis testing helps in data analysis and measure the validity and reliability of the research. It provides a basis or evidence to prove the validity of the research.

(Alternative Hypothesis -> Ha: u1 != u2 )

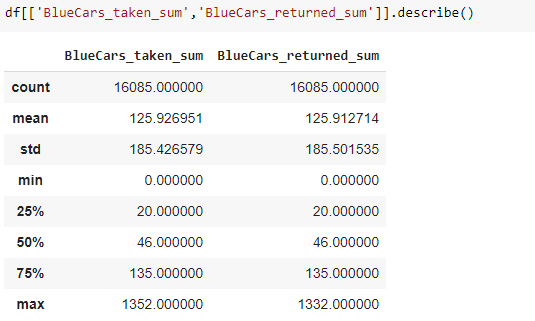
**DATA DESCRIPTION**

The dataset that was used for this investigation was an open dataset about cars in Paris. It contains columns such as postal code,day of week,bluecars\_taken and bluecars\_returned.The dates ranged between January and July of 2018. With also had the number of daily data points that were available for aggregation on the particular days of aggregation within the specified time periods

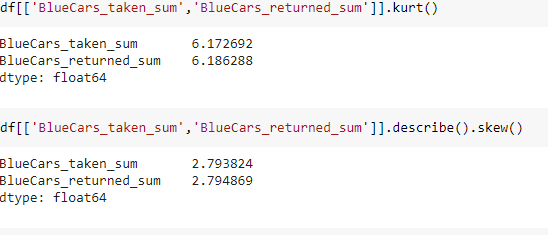
.The days of the week were the usual monday to friday with the specifications and special assignments of days. Weekday or weekend the dataset had the specific days within the time period. The blue cars that were taken and returned, the utilib data and the slots set of data were also contained in the dataset.

Our main focus being blue cars,investigating if the cars taken are the cars returned.we conducted a descriptive statistics(measures of central tendency and measures of spread on the blue cars,the results are as follows.)

**MEASURES OF CENTRAL TENDENCY**



**MEASURES OF DISPERSION**

****

**HYPOTHESIS TESTING PROCEDURES.**

Inour analysis we used:

scipy.stats.shapiro(x)[source]¶

Perform the Shapiro-Wilk test for normality.

The Shapiro-Wilk test tests the null hypothesis that the data was drawn from a normal distribution.

Parameters x : array\_like

Array of sample data.

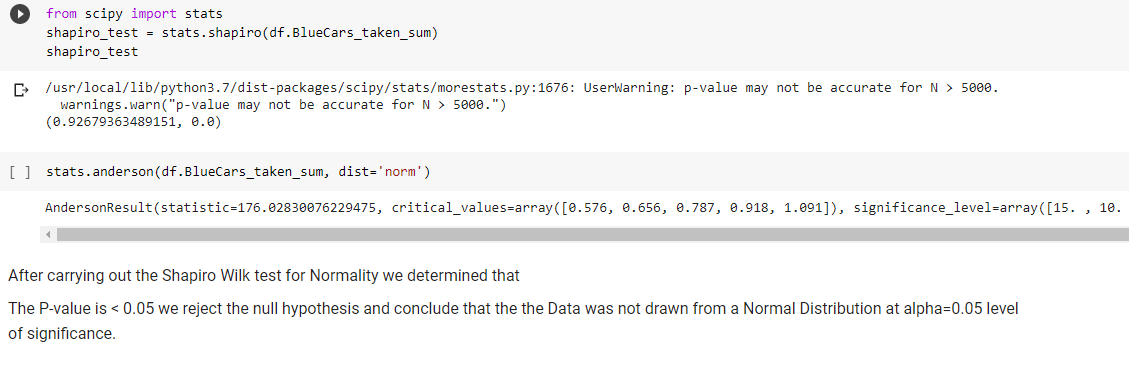
Returns :statistics: float

The test statistic.

p-value float

The p-value for the hypothesis test;

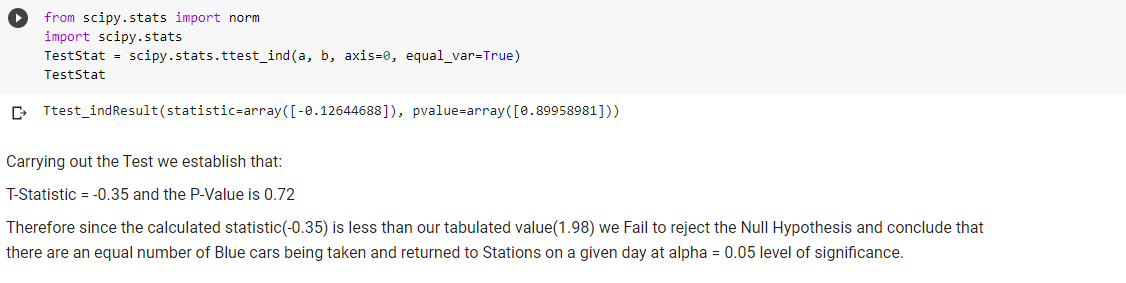
The analysis is shown below:



After carrying out the Shapiro Wilk test for Normality we determined that

The P-value is < 0.05 we reject the null hypothesis and conclude that the Data was not drawn from a Normal Distribution at alpha=0.05 level of significance.

**T TESTS**

****Carrying out the Test we establish that:

T-Statistic = -0.35 and the P-Value is 0.72

Therefore since the calculated statistic(-0.35) is less than our tabulated value(1.98) we Fail to reject the Null Hypothesis and conclude that there are an equal number of Blue cars being taken and returned to Stations on a given day at alpha = 0.05 level of significance.

**HYPOTHESIS TESTING RESULTS**

After carrying out the Shapiro Wilk test for Normality we determined that

The P-value is < 0.05 ,T-Statistic = -0.35 and the P-Value is 0.72 ,thus the calculated statistic(-0.35) is less than our tabulated value(1.98) we Fail to reject the Null Hypothesis and conclude that there are an equal number of Blue cars being taken and returned to Stations on a given day at alpha = 0.05 level of significance

**DISCUSSION OF TEST SENSITIVITY.**

Test Sensitivity in a statistical test is the measure of performance of a binary classification test(True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).The probability of a null hypothesis being true.In our case it 92%.

**SUMMARY AND CONCLUSION**

In our analysis,we performed data cleaning where we removed missing values,we dropped duplicates and identified unique values.We also performed exploratory data analysis(univariate and bivariate) and plotted graphs for better visualization.

We performed shapiro wilk test to test the null hypothesis of blue cars on a certain day that happened to be a weekday,if the cars borrowed are the cars returned.From our analysis we Fail to reject the Null Hypothesis and conclude that there are an equal number of Blue cars being taken and returned to Stations on a given day at alpha = 0.05 level of significance.Since the evidence did not support the claim(alternative hypothesis)