**IP WEEK 4**

**AUTOLIB ELECTRIC CAR-SHARING SERVICE HYPOTHESIS TESTING**

**Problem Statement**

Autolib is an electric car-sharing service which began in Paris and operates in several French cities. It is a concept that has now gone global. This model allows drivers to rent and return electric cars through this simple four-step process: subscribe, rent, drive and pass it on. Autolib provides an accessible answer to automobile ownership. In this study, we want to conduct an hypothesis test on a sample of the dataset to ascertain whether the null hypothesis will be accepted or rejected. Below is out null and alternative hypotheses:

**Null Hypothesis**: On average, the number of bluecars taken on weekends are greater than or equal to the number of bluecars returned on weekends.

**Alternative Hypothesis**: On average,the number of bluecars taken on the weekend are less than that of the bluecars returned on the weekend.

This hypothesis is important because it gives an insight into whether on average the number of bluecars taken on the weekend is greater or equal to the average number of bluecars returned on the weekend.

**Data Description**

The source of this data has not been provided, but it is our assumption that the data is sourced from an open access database and is also credible. The dataset contains the following columns that were used in our analysis:

Postal code: It shows the postal code of the specific area (Paris).

Date: It is the period when this data was captured (January - June 2018).

Day of the week and Day type: They show the day of the week and whether the day is a weekday or weekend.

BlueCars Taken Sum: It shows the total number of bluecars that have been taken during that day in a given area.

BlueCars Returned Sum:It shows the total number of bluecars that have been returned during that day in a given area.

Other columns which were also important but not used in this study are: Utilib Taken Sum, Utilib Returned Sum, Utilib-14 Taken Sum, Utilib-14 Returned Sum, Slots Freed Sum, and Slots Taken Sum.

As per our analysis of the data, the original dataset contained 16085 rows,and 13 columns. On average, bluecars taken was 125.926951 and bluecars returned was 125.912714 during the period when this data was captured. The analysis also shows that the day of the week which recorded maximum values is the 6th day, which is a weekend. This implies that most people hire and return cars during the weekend. A detailed descriptive statistics is captured on the Colab Notebook. One assertion in this experiment is that no record contains null values. Another assertion is that there are no columns that contain any outliers.

**Hypothesis Testing Procedure**

Before starting to test the hypothesis, I will start by defining the question and the metric of success, understanding the context, recording the experimental design and evaluating data relevance. I will then import all the libraries that I will need in my analysis and hypothesis testing, read and check the data. I will then perform data cleaning before performing EDA (Univariate and Bivariate Analysis). After conducting EDA, I will begin the hypothesis testing, which I will start by performing the normality test. This helps us see whether the data is normally distributed. After the normality test, I will choose my sample using Stratified Random Sampling. This enabled me to organize my data into two stratas (Bluecars) and (Day Type-Weekend). I used the groupby function in python to pick the sample.

The null and alternative hypothesis are interesting because of the trend depicted by the data. As per the data, the number of blue cars taken on weekends tend to be greater than or equal to those returned. In practice, we know that many activities usually occur during the weekends, for instance people tend to travel a lot to spend time with other family members or for leisure activities, and so on. This is the logic behind my null and alternative hypothesis.

I will use a T-test to test the validity of the hypothesis. This is because I will be working on the average number of bluecars taken and bluecars returned. This statistic requires my data to be normally distributed. In a case where the data might not be normal, I will fix that through log10 transformation. The alpha level that I will use is 0.05.

**Hypothesis Testing Results**

First, my data was not normally distributed. It was skewed to the right, but once I transformed it using log10, it changed to form a somewhat normal distribution. The results of the ttest were as follows:

P\_value - 0.9293323888923636

Statistical Value - 0.0886962666222155

Given that the p\_value was greater than our alpha(0.05), the null hypothesis was accepted.

Point of Estimation bluecars taken- 8.219551013281048

Point of Estimation bluecars returned- 8.046937105003977

Confidence Interval for bluecars taken - (46.5011118823617, 52.496843127863244)

Confidence Interval for bluecars returned - (46.521567327476916, 52.476387682748026)

**Discussion of Test Sensitivity**

When we increase the sample size, the confidence of our estimates increases. This is because we have more information, hence giving better precision and lowering the level of uncertainty.

**Summary and Conclusions**

In summary, we started by reading our data and tidying it up. We then performed EDA and afterwards the hypothesis test. After conducting the hypothesis test, we failed to reject the null hypothesis because the p\_value was greater than alpha. So that implies that it is actually true that On average, the number of bluecars taken on weekends is greater than or equal to the number of bluecars returned on weekends.