**Hypothesis Testing Report**

Autolib Electric-Car-Sharing Service

Samwel Omondi - DS14B

# **Background of Study**

Autolib' was an electric car sharing service which was inaugurated in Paris, France, in December 2011. It closed on 31 July 2018. It was operated by the Bolloré industry and complemented the city's bike sharing system, Velib', which was set up in 2007. The Autolib' service maintained a fleet of all-electric Bluecars for public use on a paid subscription basis, employing a citywide network of parking and charging stations. As of 3 July 2016, 3,980 Bluecars had been registered for the service and had more than 126,900 registered subscribers; Autolib' furthermore offered 1,084 electric car stations in Paris agglomeration with 5,935 charging points

Electric car sharing services greatly contribute to the preservation of the environment and facilitate mobility in the bustling city of Paris. Electric 'Bluecars' are silent and have zero emissions locally, and their affordability and popularity continues to prevent more and more people from buying an own car

# **Problem Statement**

Working as a data Scientist for the Autolib electric car-sharing service company to investigate a claim about the blue cars from the provided Autolib dataset. The dataset to use for this project can be found here in ([link](http://bit.ly/DSCoreAutolibDataset%22)).

In this study, we seek to underscore the claim that at any given period the number of blue cars taken from stattion 75015 has no difference with number of blue cars taken fro stattion 75017.

*Null hypothesis:* The number of blue cars taken from station 75015 on equal to the number of blue cars taken from station 75017.

*Alternative hypothesis:* There is a difference between the number of blue cars taken from station 75015 and number of blue cars taken from station 75017.

This hypothesis will help us understand the level of difference in terms of bluecars taken in different areas of interest.

# **Data Description**

The c and a glossary of the same is highlighted below:

| **Column name** | **explanation** |
| --- | --- |
| **Postal code** | postal code of the area (in Paris) |
| **date** | date of the row aggregation |
| **n\_daily\_data\_points** | number of daily data poinst that were available for aggregation, that day |
| **dayOfWeek** | identifier of weekday (0: Monday -> 6: Sunday) |
| **day\_type** | weekday or weekend |
| **BlueCars\_taken\_sum** | Number of bluecars taken that date in that area |
| **BlueCars\_returned\_sum** | Number of bluecars returned that date in that area |
| **Utilib\_taken\_sum** | Number of Utilib taken that date in that area |
| **Utilib\_returned\_sum** | Number of Utilib returned that date in that area |
| **Utilib\_14\_taken\_sum** | Number of Utilib 1.4 taken that date in that area |
| **Utilib\_14\_returned\_sum** | Number of Utilib 1.4 returned that date in that area |
| **Slots\_freed\_sum** | Number of recharging slots released that date in that area |
| **Slots\_taken\_sum** | Number of rechargign slots taken that date in that area |

A detailed understanding of this dataset is included in our analysis (colab [analysis link](https://colab.research.google.com/drive/1q_EuWtfbo8Hyy5Z36IGMM9NSi51AD4q-?usp=sharing)).

Provide information about the data necessary to understand the rest of the report including a precise statement of the random variable.

Provide a description of the source of your data and the data collection procedures, the descriptive statistics, and some assertions about the model that is consistent with the data.

# **Hypothesis Testing Procedure**

To work on this project, we performed the following analysis with Python;

## Defined the hypothesis

claims that at any given period the number of blue cars taken from stattion 75015 has no difference with number of blue cars taken fro stattion 75017

*Null hypothesis:* The number of blue cars taken from station 75015 on equal to the number of blue cars taken from station 75017.

*Alternative hypothesis:* There is a difference between the number of blue cars taken from station 75015 and number of blue cars taken from station 75017.

ma = number of blue cars taken from station 75015

mu = number of blue cars taken from station 75017

Ho: ma = mu

H1: ma ≠ mu (Claim)

## Loaded our dataset in our notebook

The dataset to use for this project can be found here in ([link](http://bit.ly/DSCoreAutolibDataset%22)).

Here, we checked and ensured that we had a valid and reliable dataset.

## Found and dealt with outliers, anomalies, and missing data within the dataset.

Dropped the outliers, no missing data, no duplicates and no anomalies found.

## Sampling

A stratified random sampling technique was used in selecting a sample for this study. We grouped the dataset by postal code, in our case we are looking at activities in 75015 & 75017

## Plot appropriate univariate and bivariate summaries recording our observations.

Conducted univariate (Measures of Central Tendency & Dispersion) and bivariate analysis on our sample data. Showed visuals to give us a better understanding of the dataset we were dealing with.

One interesting observation that we drew from this analysis was that, the highest area with blue cars taken on a weekday is 75017, whereas 75015 leads in the number of blue cars taken on the weekend.

## Hypothesis testing

The most suitable hypothesis test that we shall use in this study is a t-test. We conducted an independent Welch's t-test using scipy.stats, with 0.05 as the significance level.

## Test Results Interpretation

From the test results, we rejected the null hypothesis because the p-value was significantly less than the alpha value.

# **Hypothesis Testing Results**

Based on test conducted, we have found the following results Statistic=8.7356 and pvalue=0.00000000000002. Maintaining a 95% confidence interval which yields a 5% margin of error, the p-value is less than the alpha value of 0.05 hence we reject the null hypothesis.

We can therefore conclude that the is a significant difference between the number of blue cars taken from station 75015 and number of blue cars taken from station 75017.

# **Discussion of Test Sensitivity**

In this study we conducted a t-test and our sample size is calculated using Yamane (1967) formula and maintaining a 95% confidence interval which yields a 5% margin of error. This is because 95% confidence level and a 0.05 (5%) margin of error is most commonly used.

The null hypothesis in in this study was rejected because the p-vaue of 0.00000000000002 is significantly less than alpha value of 0.05.

# **Summary and Conclusions**

In summary, we have analysised the claim and can say that there is a significant difference between the number of blue cars taken from station 75015 and number of blue cars taken from station 75017.

Further analysis and tests need to be done in order to have an all round conclusion regarding the next areas of expansion for Autolib Electric-Car-Sharing service.