**HYPOTHESIS TESTING REPORT**

**INTRODUCTION**

In this project I will seek to understand electric car usage by solving for another research question.We will work 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.To do this, we need to identify some areas and periods of interest via sampling stating the reason to the choice of method,Then perform hypothesis testing with regards to the claim that we will have made.

**1)Problem statement:**

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 that period.

**Null hypothesis**: the average number of blue cars taken is not different that of the blue cars returned

**Alternative hypothesis**: the average number of blue cars taken is different that of the blue cars returned**(CLAIM)**

**2)Data description**

The dataset I used for this investigation was an open dataset about cars in Paris. It contains variables like the postal code of the area which was Paris, the dates of data collection. The dates ranged between January and July of 2018. Which 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.

**3)Hypothesis testing procedure**

The dataset is large with 16, 085 \* 13 entries. Therefore, I picked a sample from it using a stratified sampling method. I used this method of sampling because with it, the sample would be unbiased and have better accuracy in results in comparison to other sampling methods. Stratified sampling required me to group the data into different sets of strata and then randomly pick a sample from the strata. However, considering I used python programming to generate my sample, it was quite easy and efficient. Since the dataset only had the count and dates of picked and taken bluecars, I used those to determine my hypothesis.

The logic behind my null and alternative hypothesis is that I wanted to know whether the average number of cars that were picked in a day was similar to the average number of cars that were returned on that very day.

From my stratified sampling, I got a sample of 1154\* 4 entries. This meant that n > 30. As such, I used the z-score to determine the p-value. I did perform normality tests on the data. The alpha level of significance that I used was 0.05.

**4)Hypothesis testing results**

From the hypothesis test, I found that there was not sufficient evidence to prove that the average means of the blue cars taken and the bluecars returned are not equal. The z-score was -0.10617667206769237 and as a result, the null hypothesis was not rejected.

The p -value was 0.45772108965010394 which was greater than the significance level.

**5)Summary and conclusions**

The project was comprehensive and demanding. I performed exploratory data analysis with hypothesis testing as its implementation. Conclusively, I failed to reject the null hypothesis because there was not enough evidence for me to reject the null hypothesis.