**Hypothesis Testing Report for Electric Car sharing service**

**Problem Statement**

In Paris, cars are on the road for only about 5% of the time and remain parked the 95% of the remaining time.To address this problem, the autolib company was formed.

Autolib is a car sharing service company that provides easy and convenient access to electric cars, operating in several French cities. There are three types of cars to choose from, the Bluecar(which is the primary car) or the utility cars Utilib and Utilib 14. The service is attained through subscription after which the subscriber receives a Radio Frequency Identification (RFID) badge. The scheme allows a driver to rent a car at a station and return it to a station in a different location at a nominal charge. By use of the RFID, an invoice is charged to a subscriber's account on their usage.

In the city of Paris, a driver is estimated to save about 7,000 Euros per year by using the Autolib service in the stead of buying a car.

I am 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 since their demand is quite high compared to the rest.

The null and alternative hypotheses are as stated below;

Null hypothesis: The average number of blue cars taken from 75015 is not different from the number of blue cars taken from 75016

Alternative Hypothesis: The average number of blue cars taken from 75015 is different from the number of blue cars taken from 75016.

We chose this hypothesis because we wanted to test whether the number of blue cars taken from these two postal codes were different since they had the highest number of blue cars taken overall.

**Data Description**

Our main interest is to test a claim about the blue cars. This is our random variable. The dataset and glossary to use for this project can be found here [<http://bit.ly/DSCoreAutolibDataset>] and here [[Link]](http://bit.ly/DSCoreAutolibDatasetGlossary) respectively.

The provided dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging).

Below is a description of the columns found in the dataset.

Postal code - postal code of the area (in Paris)

Date - date of the row aggregation

n\_daily\_data\_points - number of daily data points 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 blue cars taken that date in that area

BlueCars\_returned\_sum - Number of blue cars 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 recharging slots taken that date in that area

**Hypothesis Testing Procedure**

Seeing as the null and alternative hypothesis have already been stated, we are going to perform a two tailed test and calculate the test statistic( z statistic).

We will use a confidence level of 95% and a corresponding significance level (alpha) of 5% and calculate the p value. This means that there is a 5% chance that we will accept the alternative hypothesis when our null hypothesis is actually true. If the p value is less than the significance level then we will reject the null. If it’s greater than the significance level then we will fail to reject the null because there won’t be enough evidence to prove that the alternative hypothesis is true.

We will use the z statistic because of my sample size(50) and also cause the population std deviation is known. If we was using a sample size less than 30 we would have used the t test.

**Hypothesis Testing Results**

We failed to reject the null hypothesis.

The z statistic and p values were 0.7386 and 0.7699 respectively. We chose a 5% level of significance but since this is a two tailed test, we shall compare the p-value against 0.025. Since the p value is higher than our level of alpha, we will fail to reject the null hypothesis. This means that the p value does not fall in the rejection region in this case.

**Test Sensitivity**

Our sample size directly impacts our p-value. Large sample sizes produce small p-values even when differences between groups are not meaningful. On the other hand, a sample size that is too small can result in a failure to identify a difference when one truly exists. Therefore, using a larger sample size would give us a small p value and reducing the sample size would produce a large p value.

**Summary and Conclusions**

Below is an outline of the project process

* Clearly defined the null and alternative hypothesis
* Looked for outliers and other anomalies such as missing data and duplicates. Our dataset was clean. We did not drop the outliers because they were too many and would have affected our results.
* Performed exploratory data analysis on the dataset. We did both univariate and bivariate analysis. We did pandas profiling for the univariate analysis and plotted heat maps and pair plots for the bivariate analysis bit.
* Performed hypothesis testing
* Drew a conclusion that stated; Since the p value(0.7699) is more than the level of significance(0.025) , we fail to reject the null hypothesis. There isn't enough evidence to show that the number of blue cars taken from postal code 75015 is different from the ones taken from 75016.