**A DATA REPORT ON AUTOLIB ELECTRIC CAR-SHARING SERVICE HYPOTHESIS TESTING.**

**1 Problem Statement**

 We have been tasked 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 a provided Autolib dataset. In an effort 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.

The data that we will be investigating features electric car usage among different cars one being the bluecars, the utilib and the utilib 14 cars all from the Autolib electric car-sharing company. The dataset has 13 columns and 16085 rows. Among the 13 columns we have the column date that features the only continuous random variable within the dataset. Our main focus will be on the discrete random variables of the electric cars within the company.

Since I set out to investigate whether the number of blue cars taken in area X is different from the number of blue cars from area Y therefore, my null hypothesis became that the number of blue cars in postal code 75015 is equal to the number of blue cars in postal code 75017 and the alternative hypothesis became that the number of blue cars in postal code 75015 are not equal to the number of blue cars in postal code 75017.

H0: Number of blue cars in postal code 75015 = Number of blue cars in postal code 75017.

H1: Number of blue cars in postal code 75015 ≠ Number of blue cars in postal code 75017

This hypothesis was important because we could appropriately understand whether there exists a difference or a statistical significance when different areas are considered within the dataset.

**2 Data Description**

This dataset had 16085 rows and 13 columns. 12 of columns had discrete random variables and only the date column contained a continuous random variable. The Autolib dataset contained zero missing values and this made the data cleaning process less hefty. The dataset also contained no duplicate values hence the anomalies were less if none existent. The only anomaly spotted was the naming of the postal code column that was renamed to include an underscore.

The source of the data was the Autolib electric car-sharing company where a link to the dataset was given and I used the pandas programming library to import and load the dataset from the source into my programming environment. There existed correlation between some of the column as some of the column describe the same car.

**3 Hypothesis Testing Procedure**

To test the for the hypothesis stated above in the problem statement, I first started by defining the Null and the alternative hypothesis which were:

H0: Number of blue cars in postal code 75015 = Number of blue cars in postal code 75017.

H1: Number of blue cars in postal code 75015 ≠ Number of blue cars in postal code 75017

I then ventured into stating the level of significance or the level of significance which I decided was going to be 0.05 or 5 %. Later I selected the appropriate statistical test as my input was a binary classification. Afterwards I endeavored to get the p- value to be compared to the level of significance to access whether we can reject or fail to reject the null hypothesis.

I wanted to test the hypothesis that would provide information whether there existed a difference or statistical significance between the as per different areas defined by the postal code chosen at random. That is why I structured my null and alternative hypothesis to do or investigate just that. By looking at the structure of the alternative hypothesis it is important to note that the test will be a two tailed test.

I used the Sharpiro-Wilk test to check for normality of the data given while also accessing the p- value to mitigate and weigh in on whether to reject or fail to reject the null hypothesis. Since the assumption was asserted that the data was indeed not normally distributed the choice of test statistic became fairly routine. I used the Mann-Whitney U test as the non-parametric method of checking for the p-value to be compared to the significance level. I used 0.05 as the level of significance for my test.

**4 Hypothesis Testing Results**

On carrying on with the test I found out that the Mann-Whitney u test gave a statistic of 4734.5 and a p- value was 5.313439023634663e-21. This p-value was less than the 0.05 significance level hence I had to reject the null hypothesis and ascertain that indeed there was a difference when different areas were concerned in terms of the variability of the dataset. Therefore, the number of cars in postal code 75015 were not equal to the number of cars in postal code 75017. Since the level of significance was 0.05 then the test had a confidence level of 0.95 that is a confidence level of 95%.

**5 Discussion of Test Sensitivity**

The confidence level of the test was at 0.95 that is at 95% because the significance level was 0.05. So we had the 95% ability to test whether we had a statistical significance or not as the alpha is the probability of making the type one error, that is wrongly rejecting the null hypothesis when it is true. However, our beta was 0.10 that is the probability of making the type two error that is failing to reject the null hypothesis when it was false. So the power of the test was at 90% or 0.90 and this means the test was fairly accurate and sensitive inn it ability to accurately test for a statistical difference or significance within the parameters defined.

**6 Summary and Conclusions**

As confirmed above our test was fairly accurate hence with the confidence level of 95% and the power of the test at 90%, we can therefore ascertain that since we rejected the null hypothesis that stated that there was no difference of the number of blue cars between two different areas defined by their individual postal code. Hence it is necessary and logical to infer that the hypothesis test had statistical significance and there indeed exist a difference and a level of variability in the number of cars where different areas or location is concerned. This is necessary as the Autolib company can now strategize and evaluate which areas have low numbers so as relevant improvements can be made within those areas.