**AUTOLIB BLUECAR ELECTRIC CAR SHARING HYPOTHESIS TESTING PROJECT REPORT**

1. **Problem Statement**

The Autolib car-sharing company in Paris serves people in multiple areas across France and is keen to understand the blue car usage at the different postal codes in Paris.

We have been tasked to analyze and formulate a hypothesis for two different postal codes and compare the average of cars taken from the stations.

1. **Hypothesis**

The null hypothesis (H0) is:

#hypothesis testing,

#the claim is that the mean average for the sample of cars taken over the weekend at postalcode 75015(strata1) is equal to that in postal code 75017(strata2)

#H0 : µ 1 = µ 2

#Ha: µ 1 ≠ µ 2

# Significance Level is α = 0.05

H0: Mean postal code 75015 == Mean postal code 75017

H1: Mean postal code 75015=/ Mean postal code 75017

The hypothesis test will help establish whether the means are significantly different.

1. **Data Description**

The dataset provided is described as follows

|  | **Column name** | **explanation** |
| --- | --- | --- |
| **0** | Postal code | postal code of the area (in Paris) |
| **1** | date | date of the row aggregation |
| **2** | n\_daily\_data\_points | number of daily data poinst that were availabl... |
| **3** | dayOfWeek | identifier of weekday (0: Monday -> 6: Sunday) |
| **4** | day\_type | weekday or weekend |
| **5** | BlueCars\_taken\_sum | Number of bluecars taken that date in that area |
| **6** | BlueCars\_returned\_sum | Number of bluecars returned that date in that ... |
| **7** | Utilib\_taken\_sum | Number of Utilib taken that date in that area |
| **8** | Utilib\_returned\_sum | Number of Utilib returned that date in that area |
| **9** | Utilib\_14\_taken\_sum | Number of Utilib 1.4 taken that date in that area |
| **10** | Utilib\_14\_returned\_sum | Number of Utilib 1.4 returned that date in tha... |
| **11** | Slots\_freed\_sum | Number of recharging slots released that date ... |
| **12** | Slots\_taken\_sum | Number of rechargign slots taken that date in ... |

From the data set that was provided http://bit.ly/DSCoreAutolibDataset, we identified postal codes 75015, and 75017 as the two areas of interest we will focus on and specifically over the weekend.

We performed cleaning on the dataset, checking for missing values and outliers.

1. **Sampling**

From the initial dataset as stated above, we first used stratified sampling to extract a sample of 50% for our analysis and hypothesis testing

strata = df.groupby('Postal code', group\_keys=False).apply(lambda grouped\_subset : grouped\_subset.sample(frac=0.5))

strata

We then focus on the two postal codes indicated above during weekends and acquire two samples that we will use to test our hypothesis.

1. **Hypothesis Testing Procedure**

After creating our two samples, we perform t-test and calculated the p-value so as to draw our conclusions:

import scipy.stats as stats

stats.ttest\_ind(df\_strata1["BlueCars\_taken\_sum"], df\_strata2["BlueCars\_taken\_sum"], equal\_var=True)

Ttest\_indResult(statistic=5.922218283021847, pvalue=5.13908298095769e-07)

From this we, our p-value is too low compaired to our significance level and we therefore reject the Null hypothesis.

Our pvalue=5.13908298095769e-07 Our level of significance being 0.05.

This means our p value is way too low than the significance level hence we will reject the null hypothesis.we conclude that the average bluecars taken from the different postalcodes are not equal over the weekends

1. **Conclusion**

From this we, our p-value indicates no significance statistically and therefore we reject the null hypothesis. The mean average of blue cars taken from each station is different during the weekend.