Autolib Car Sharing Service

# Problem Statement

As a Data Scientist for the Autolib electric car-sharing service company, I have been tasked to investigate a claim that on average, only all clients who take blue cars over the weekend, will return them during the weekend. The rest will return them during weekdays. Thus, we would like to check if the mean number of blue cars taken on Saturday/Sunday is equal to the mean number of blue cars returned on Saturday/Sunday, from the provided Autolib dataset.

In an effort to do this, we need to identify some areas of interest via sampling stating the reason for the choice of method, then perform hypothesis testing with regards to the claim that we will have made.

Our hypotheses are:

H₀ :μ(Blue Cars taken on weekend) = μ(Blue Cars returned on weekend)

H₁ :μ(Blue Cars taken on weekend) ≠ μ(Blue Cars returned on weekend)

# Data Description

Our dataset contains 16085 entries and 13 columns which are as described below:

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

We cleaned the data using the integrity rules i.e Accuracy, Completeness, Consistency, Uniformity & Validity.

1. Dropped irrelevant columns
2. Checking for outliers and anomalies. However, we didn’t drop them since they were a very significant number and removing them would skew our data in a different direction.
3. Checking for missing data
4. Check for duplicates

We used cluster sampling for our data for ease of processing and due to time constraints. We made an assumption that the sample will be representative of the general population and that our sample has the same statistical distribution as the population. We created clusters by postal code. Each postal code has approximately 156 entries, thus we used random sampling to select our cluster from the list(postal code 75015). We filteredr our data frame to only contain data for the weekend values as our hypothesis covers only weekend data.

We then did the exploratory data analysis(Univariate and bivariate).

# Hypothesis Testing Procedure & Results

We used the following steps for hypothesis testing:

* Step 1: Formulate the null hypothesis and the alternative hypothesis:

Our null & alternative hypothesis are:

H₀ :μ(Blue Cars taken on weekend) = μ(Blue Cars returned on weekend)

H₁ :μ(Blue Cars taken on weekend) ≠ μ(Blue Cars returned on weekend)

* Step 2: Identify a test statistic and significance level that can be used to assess the truth of the null hypothesis.

For our sample, we used a significance level(alpha) of 5%. We also decided to use a 2 paired sample z=test as when we split the data into 2 as per our hypothesis, we have 2 samples, the dataset is normally distributed (as evidenced by the QQ plot we did)our sample contains more than 30 values and our samples are independent of each other. We have satisfied the assumptions required for our statistic.

* Step 3: Computing the test-statistic and P-value. The smaller the P-value, the stronger the evidence against the null hypothesis.

We computed our z-statistic as -0.262 and p-value as 0.7935

* Step 4: Analyze the results and either accept or reject the null hypothesis.

Our p-value is 0.7935 which is greater than our significance level, giving us no evidence to reject our null hypothesis. Thus, we shall accept our null hypothesis.

* Step 5: Interpreting the Results

Our results mean that there is a 79.35% chance that the mean number of blue cars taken over the weekend is equal to the mean number of blue cars returned during the weekend. Thus, we can confidently say that clients who take blue cars over the weekend, will return them during the weekend.

Our point estimates are 0.00081247,165.0044733, and the confidence interval is 95%.

# Summary & Conclusion

We can conclude that there is a 79.35% chance that the mean number of blue cars taken over the weekend is equal to the mean number of blue cars returned during the weekend. Thus, we can confidently say that clients who take blue cars over the weekend, will return them during the weekend.