**Autolib Electric Car-sharing Service Company**

* [**Github Repository**](https://github.com/Brian-Onyango/moringa-school-course/blob/master/Moringa-Core/Statistics%20and%20Visualization%20Module/Week%204/Moringa_Data_Science_Core_W4_Independent_Project_2022_03_Brian_Onyango.ipynb)

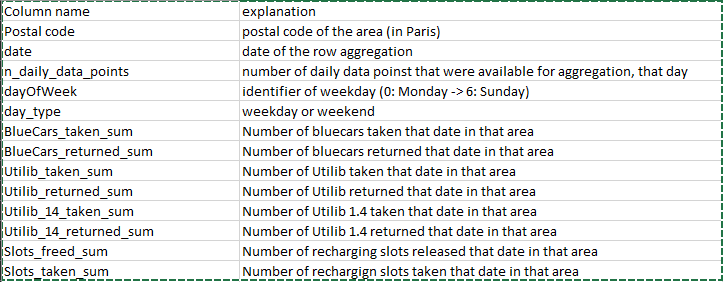
1. **Introduction**

Autolib' was an electric car-sharing service that was inaugurated in Paris, France, in December 2011. There is a claim that the number of blue cars taken on Saturday is equal to the number of cars taken on Sunday. We will work as a Data Scientist for the Autolib electric car-sharing service company to investigate this claim about the blue cars from the provided Autolib dataset.

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

1. **Problem Statement**

* To investigate a claim about the blue cars from the provided Autolib dataset.
  + Claim: The number of Blue cars taken on Sunday (day 6) was different than on Saturday (day 5)
* Understanding the dataset:
  + 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).
  + [Dataset](http://bit.ly/DSCoreAutolibDataset)
  + [Dataset Glossary](http://bit.ly/DSCoreAutolibDatasetGlossary)



* From the dataset, our interest will be in two columns, namely:
  + dayOfWeek
  + Bluecars\_taken\_sum
* Hypothesis:
  + Null Hypothesis, H0: The number of blue cars taken on Saturday is equal to the number of blue cars taken on Sunday.
  + Alternative Hypothesis, H1:The number of blue cars taken on Saturday is not equal to the number of blue cars taken on Sunday.
* Through testing this hypothesis, we will be able to compare the BlueCars usage on Saturday and Sunday. This can primarily enable us to determine which day during the weekend is the busiest for the Blue cars.

1. **Data Description**

* Describing the data we will use in this study:
  + dayOfWeek - This shows the number assigned to each day of the week. We will be using 5 - Saturday and 6- Sunday for our study.
  + Bluecars\_taken\_sum - This shows the sum of cars taken per day for every postal code.
* This dataset was provided by the Autolib electric car-sharing service company and the following procedures were carried out to ensure it’s viable for this study:
  + Data Preparation and cleaning
    - Checking validity -Dropping irrelevant columns to our study(only two relevant columns were left)
    - Checking completeness - Checking for the presence of missing values and dealing with them
    - Checking consistency -Checking for duplicates and removing them
    - Checking uniformity - Renaming of columns to a uniform naming style
    - Checking outliers and anomalies - Checking and removing outliers and anomalies.
  + Descriptive Analysis
    - Univariate analysis
      * Bar Plots - Checking relationship between day and number of Cars taken.
      * The measure of central tendency - Calculating the Mean, Median, and Mode
      * Measures of distribution - Calculating the Range, IQR, variance, std, skew, and kurtosis.
    - Bivariate analysis
      * Scatter Plots - Checking if there is any correlation between day of week and number of blue cars taken.
      * Correlation - Calculating if there is any correlation between day of week and number of blue cars taken.
* Sampling Method:
  + From the data set, we used the clustering method to group the data into two categories i.e Sunday dataset and the Saturday dataset.
  + The Sunday dataset will be our Population dataset and has the following features:

count 2274.000000

mean 146.071240

std 206.289498

min 0.000000

25% 27.000000

50% 59.000000

75% 152.000000

max 1164.000000

* + After clustering, we used random sampling to pick 100 records as our sample from the Saturday dataset. This will be our sample dataset and has the following features:

count 2000.000000

mean 155.257000

std 229.109182

min 0.000000

25% 23.000000

50% 58.000000

75% 161.250000

max 1352.000000

1. **Hypothesis Testing Procedure**

For the hypothesis testing, the following procedure is followed:

* + **Step 1: Specify the Null Hypothesis**
    - Null Hypothesis, H0: The number of blue cars taken on Saturday is equal to the number of blue cars taken on Sunday.
  + **Step 2: Specify the Alternative Hypothesis**
    - Alternative Hypothesis, H1:The number of blue cars taken on Saturday is not equal to the number of blue cars taken on Sunday.
  + **Step 3: Set the Significance Level (a)**
    - The significance level (denoted by the Greek letter alpha— a) is set at 0.05. This means that there is a 5% chance that we will accept our alternative hypothesis when our null hypothesis is actually true. The smaller the significance level, the greater the burden of proof needed to reject the null hypothesis, or in other words, to support the alternative hypothesis.
  + **Step 4: Calculate the Test Statistic and Corresponding P-Value**
    - For this study, we will use T-statitics with the following assumptions:
      * The scale of measurement applied to the data set follows a continuous or ordinal scale, such as the scores for an IQ test.
      * A simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
      * The data, when plotted, results in a normal distribution, bell-shaped distribution curve. When a normal distribution is assumed, one can specify a level of probability (alpha level, level of significance, p) as a criterion for acceptance.
      * Reasonably large sample size is used. Larger sample size means the distribution of results should approach a normal bell-shaped curve.
      * The homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.
    - For our dataset:
      * A simple random sample was used.
      * A large sample size (2000 records) was used.
      * T-statistics = 1.79303783765275
      * P-value = 0.07311800223205009
  + **Step 5: Drawing a Conclusion**
    - P-value <= significance level (a) => Reject our null hypothesis in favor of our alternative hypothesis. Our result is statistically significant.
    - P-value > significance level (a) => Fail to reject our null hypothesis. Our result is not statistically significant.
    - Since our P-value > significance level (a), we fail to reject our null hypothesis and conclude that:
      * There is evidence that the number of blue cars taken on Saturday is equal to the number of blue cars taken on Sunday.

1. **Hypothesis Testing Results**

* T-statistics = 1.79303783765275
* P-value = 0.07311800223205009
* Significance Level = 0.05
  + P-value <= significance level (a) => Reject our null hypothesis in favor of our alternative hypothesis. Our result is statistically significant.
  + P-value > significance level (a) => Fail to reject our null hypothesis. Our result is not statistically significant.
* Since our P-value > significance level (a), we fail to reject our null hypothesis and conclude that:
  + There is evidence that the number of blue cars taken on Saturday is equal to the number of blue cars taken on Sunday.

1. **Discussion of Test Sensitivity**

Since our null hypothesis is not rejected, with a significance level of 0.05, we can confidently say that there is a 95% chance that this finding is true. However, with a huge dataset, this confidence level might decrease giving life to alternative hypotheses.

1. **Summary and Conclusions**

This project went through a process of data collection, data preparation, data cleaning. Descriptive analysis and finally hypothesis testing. All these processes were necessary to enable us to come up with an accurate finding. We can now confidently and statistically justify the claim that the number of blue cars taken on Saturday equals the number of blue cars taken on Sunday.