## **DATA REPORT**

* The project management link is (Trello): [Trello](https://trello.com/b/0gFJUJ9G/ip-week-4)
* The GitHub link is: [GitHub](https://github.com/allanstar-byte/IP_WEEK_4.git)
* Google colaboratory link:[Colab](https://colab.research.google.com/drive/1NCCKw7yzYsvwnhdoDihCEjq7Gtf7LUBt?usp=sharing)

## **Business Understanding**

### **Business Overview**

The electric car-sharing service company needs to understand electric car usage over time so as to maximize their profits.

### **Business Objective**

The main objective of this report is to identify the picking and returning hours of different stations that should be prioritized

### **Business Success Criteria**

To compile a list of stations, hours and postal codes for the management to make paramount decisions.

### **Assessing the Situation**

1. **Resource Inventory**
   1. Dataset:

-Autolib [[Link](http://bit.ly/autolib_dataset)]

b.Software( Github, Google Collaboratory, Trello)

1. **Assumptions**
   1. The data provided is correct and up to date
2. **Constraints**
   1. There are no constraints

### **Data Mining Goals**

Our data mining goals for this project are as follows:

* To calculate picking and returning hours
* To calculate popular station
* To calculate postal code of the most popular stations

**Data Mining Success Criteria**

Our success criteria will be measured by the following criteria;

* We target the stations with the most blue car counter at different hours

## **Data Understanding**

### **Data Understanding Overview**

For this project, we are using the availed dataset by the company. That is;

* Autolib dataset - This dataset gives the information for the month of April.

### **Data Description**

We have one dataset available for this project. A detailed description of the datasets is provided as follows:

* **Autolib -** Data was extracted from opendataparis.com, where the Autolib availability information was

available in real-time. The accessed database was the following:

Name: Stations Autolib: Disponibilité en temps réel

Producer : Autolib

Date : see below

License : Open Database License

### **Verifying Data Quality**

The data set has some missing values majorly on the Location and displated\_comment columns.

## **Data Preparation**

These are the steps followed in preparing the data

#### **Loading Data**

Loaded the datasets from the CSV to the Google Colaboratory.

#### **Cleaning Data**

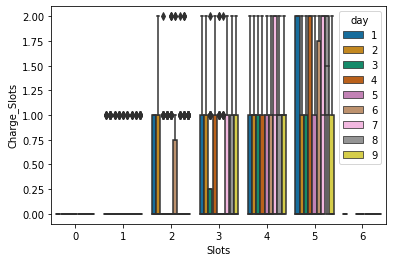
While doing data cleaning I noticed that the dataset had some large amount of missing values which i decided to fill with zero values.

The dataset had no duplicated values

There was a syntax error on the column naming due to the white spaces in between the names.

1. **Outlier detection and removal**

The charges for the slots have some outliers which are denoted by the dots which were dropped.



## **Analysis**

During the analysis, i was able to single out the following insights:;

1. The most popular picking hours is the 3 hours
2. The top 5 most popular stations at picking are:

* Alfortville/Charles de Gaulle/16
* Alfortville/Etienne Dolet/174
* Alfortville/Jean-Baptiste Preux/17
* Alfortville/Joseph Franceschi/1
* Alfortville/Malleret/16

1. The top 5 most popular stations overall are:

* **Ãpinay-sur-Seine/Foch/53**
* **Ãpinay-sur-Seine/Fitzelin/64**
* **Ãpinay-sur-Seine/Avenir/1**
* **Yerres/Raymond PoincarÃ©/200**
* **Yerres/Pierre Brossolette/92**

The above analysis was done using Python. The full analysis can be found in the following notebook.[[Link](https://colab.research.google.com/drive/1NCCKw7yzYsvwnhdoDihCEjq7Gtf7LUBt?usp=sharing)]

## **Recommendations**

From our analysis, we would recommend that the management prioritize the above-listed stations and hours. Our main reason behind this recommendation would be that the most popular stations will attract more blue cars thus promising a higher return on investment.

1. **Evaluation**

On a scale of 1 to 10 I would evaluate the project to 5 out of 10.This is because the project is done half way.There is remainder of modelling, hyper-parameter tuning and model deployment.