# PROBLEM STATEMENT.

The dataset was for an autolib dataset for an electric car-sharing service company located in Paris.The dataset contains different dates for which the electric cars were taken and returned in the company. As a Data Scientist for the Autolib electric car-sharing service company i have been tasked to investigate a claim about the blue cars from the provided Autolib dataset.

## Null hypothesis:

*There is a difference in the average number of number of blue cars taken on saturday to that of the blue cars taken on sunday*

## Alternative hypothesis:

*There is a difference in the average number of number of blue cars taken on saturday to that of the blue cars taken on sunday ( CLAIM)*

The importance of my hypothesis is to help me make a decision which day between Saturday or Sunday the cars are readily available to the customers and enable me make a conclusion if there is a difference or no difference in the rate at which cars are taken from the company between saturday and sunday(weekends)

# Data Description.

The dataset I used in the project was an open dataset for Autolib electric car sharing service company located in Paris . The data was collected in a period of 1/1/2018 and 6/19/2018. It contained the daily data points which is the number of daily data points that were available for aggregation that day.It had days of the week from Monday to Sunday of which they were assigned 0-Monday to 6-Sunday. Each day was given its day type, that is it is either a weekday or a weekend. It had postal codes for the area in Paris and number of electric cars,that is-the blue car,Utilib and Utilib 14, taken or returned in the respective areas represented by the postal codes. I was also provided with the slots freed and slots taken that is- the number of recharging slots released that date in that area and number of recharging slots taken that date in that area respectively.

The data used was already collected and thus ready to be analyzed and used for hypothesis testing . However ,if I was to collect the data by myself I would send teams of people to collect data for the Autolib Electric car sharing company in all areas of Paris where the company is located by giving them postal codes that they will use to know the location of the Autolib company in different areas. This is to ensure that the data collected is accurate and without bias.

# Hypothesis testing procedure.

The dataset was large as it contained 16085 rows and 9 columns . Although, from the dataset i was interested with those days that were weekend(sunday,saturday) only which blue car is taken . i located weekend and the total columns to rows were 4541 \* 10. I then picked a sample here using the stratified sampling method. The reason i used the stratified method is because the method ensures each subgroup within the population receives proper representation within the sample. During my analysis and hypothesis testing, something interesting i came across was that the descriptive statistic for blue car taken and blue car returned was almost similar. This was a good sign as it creates a pattern to determine the future trends of business operations relating to blue cars in the Autolib electric car sharing company.

From the stratified sample ,I got a sample of 454 rows and 5 columns .I used the sample to calculate the z score and p value in hypothesis testing. I considered 5%(0.05) as my alpha level of significance.

I also performed a normality test from a sample dataset for a blue car taken to see if it follows a normal distribution or not. I used the Shapiro-Wilk Test and Quantile\_Quantile(Q-Q) Plot. From the normality test I conclude that the distribution of the blue car taken are not normally distributed over the population. This is so because the sample of bluecar taken does not look like gaussian from Shapiro-Wilk Test and also points in Quantile\_Quantile(Q-Q) Plot are not aligned along the red line. The reason why the bluecar is not normally distributed is due to the presence of outliers . The reason why the outliers were not not dropped is because they are an actual representation of the numbers of bluecar taken.

# Hypothesis testing results

From the hypothesis test, we found that there was not sufficient evidence to prove that there is a difference in the average number of blue cars taken on saturday to that of the blue cars taken on sunday. The z score calculated was -0.1415(rounded to 4decimal places) and p values calculated from the z score was 0.4437.

In the interpretation of the p -value, it was greater than the alpha level of significance which was 0.05. Hence i failed to reject the null(accepted the null).The z critical 1.959963984540054 and confidence intervals were between (139.78922 and 173.9024)

# Discussion of test sensitivity.

The true positive rate (also known as the sensitivity of a test) is the percentage of the population that will get a positive result. The total population can be identified by the test to be 100 percent accurate.Although , any population test that is 100 percent sensitive is extremely rare. A test will be conducted.In our case we conducted a hypothesis test for the blue car taken on weekends. In my case the sensitivity was 100-((5)/2) = 97.5

**NOTE**: *Was a two tail test*.

# Summary and conclusion.

In the course of my project , I performed exploratory data analysis using univariate and bivariate analysis and implemented my solution with hypothesis testing. I failed to reject the null hypothesis because there was not enough proof that there is a difference in the average number of blue cars taken on saturday to that of the blue cars taken on sunday ( CLAIM).