# **PROBLEM STATEMENT**

## **DATA DESCRIPTION AND DEFINITION OF THE RANDOM VARIABLE**

* I have been presented with the autolib dataset of car usage and been tasked with investigating a claim on the electric blue cars.
* The random variable I will be investigating will be the discrete variables of the blue cars taken.
* The question I am trying to answer is whether more blue cars were taken in month 5 as opposed to month 1.

## **STATEMENT OF HYPOTHESIS**

* The question **;**Is the number of blue cars taken greater in month 5 as opposed to month 1?:
  + Null Hypothesis**;**
    - There is no difference between the number of blue cars taken in month 1 weekdays as opposed to month 5 weekdays.
  + Alternative Hypothesis**;**
    - There is a greater number of cars taken in month 5 weekdays as opposed to month 1 weekdays.(claim)
* The above claim is interesting as it can give us insight into whether blue car usage is increasing month over month and thus serve as a basis for a further expansion of the fleet if the null hypothesis is to be rejected.

## **DATA DESCRIPTION**

* The data to be used in the hypothesis testing will be the sum of the blue cars taken during month 1 and month 5.
* The random variable being tested is a discrete variable in the form of the sum of blue cars taken during a specific day throughout the month in question.
* The descriptive statistics are as follows:

## **HYPOTHESIS TESTING**

* The first step will be to clean the dataset and remove any outliers
* Step two will involve selecting the variables of interest in the study namely month 1 and month 5.
* Step three will be to select a sample from the 2 dataframes that I will use to run the tests and compare the results to see whether I should reject the null hypothesis or not.

### Reasons For The Hypothesis

* I selected the hypothesis above to see whether bluecar usage increased between the first and fifth month which could act as a strong argument for increasing the fleet size of the blue cars as it would indicate increased demand for the blue cars.

### Description of test statistic

* The test statistic I will employ in my analysis will be a t-test
* The reasons for using a t-test are the small sample size that I have selected to work with (30 variables in each sample).
* The second reason is in wanting to compare the means between the 2 sample sizes I have selected for the analysis.
* The test will be a two tailed t-test.
* The alpha level to be used is α 0.05

## **HYPOTHESIS TESTING RESULTS**

* The p-value that I got after running the t-test is 0.065.
* The value that I got from the t-test is -1.87.
* Thus going back to the criterion of rejecting or accepting the null hypothesis the test seems to indicate that I reject the alternative hypothesis as the p value is greater than the set alpha level of 0.05.

## **SUMMARY AND CONCLUSION**

* The summary of the procedures to conduct the hypothesis testing were:
  + Importing the data-set and doing the necessary data cleaning.
  + Doing EDA.
  + Preparing a sample from the data-frames of interest
  + Conducting a t-test
  + Finding the p-value of the result got from the t-test
* From the results obtained it is clear that I should reject the alternative hypothesis.