**PROBLEM STATEMENT:**

As a data scientist at Autolib Paris, I have been handed the task of utilizing the Autolib Dataset to determine the conclusive values of the BlueCars variable in two distinct postal codes (these are 75001 and 95880) and their activity during the weekday.

To do this, I have formulated two hypotheses that will be tested and form the foundation for solving the above problem. The main hypothesis, and the alternative hypothesis, which opposes this:

**Null Hypothesis:**

There are less BlueCars taken in postal code 75001 than postal code 95880 during the weekday in total.

**Alternative Hypothesis:**

The total number of BlueCars taken in postal code 95880 is less than those taken in postal code 75001.

Primarily this is important in a business perspective as it would require a rethink or creation of a distribution system that can be actively used to maximize consumer supply if a situation where the unavailability of the product is identified.

**DATA UNDERSTANDING:**

The main focus is on the following keywords (these are the columns of interest found in the Autolib Dataset):

* **Postal code** - Postal code of the area of interest (in Paris).
* **day\_type** - Weekday or weekend.(In our case it will be the Weekend).
* **BlueCars\_taken\_sum** - Number of bluecars taken that date in that area.

The random variable of interest here is the BlueCars\_taken\_sum which essentially is the number of BlueCars in use throughout the day, and will be used to assess the BlueCar usage in the target post codes.

Data Source:

The data was extracted from opendataparis.com where the client Autolib has made it available in real time.The section of use for this case is an excel datatable named Autolib\_dataset.Inorder to properly cleanse this data, high performance tools have been made available:

* Python (hosted on a Jupyter notebook) - This is the programming language and toolbox of choice for handling of the coding requirements for our case.It will also house the final program.
* Pandas Library - A high performance database manipulation tool for handling large datasets.It is a part of the vast python library
* Matplotlib - A dataset visulazation tool for numerical data presentation and in our case will be used to assist in normality plotting..

From the data we will be selecting random samples from the target population (Postal Codes) which will be used to draw conclusions about the population.

The key statistic that is evident in all the data is the relatively high skewness of the data.

**HYPOTHESIS TESTING:**

The procedure to be followed will be as follows:

* Filtering the data set to ensure that only the two postal zones are utilised for sampling.
* Carry out stratified random data sampling from the filtered dataset above, the number of samples will be 60 from both postal codes.
* Carry out normality testing this will aid in the determination of the test to use for statistical tests according to their parameters.
* Create confidence value (alpha),which will be used to test our hypothesis.
* Compare test statistic result to confidence values.
* Conclude from the above.

Apart from the business wise reason stated above, the other driving force for selecting the null and alternative hypothesis was to ensure that the two highest likelihood( more than or less than) are taken into account .These two are closer to real world situations and would prove more applicable to test for.

We will use a two sample z-test being that our sample has more than 30 values, is gaussian, from both qqplot, and Shapiro test we got the following values when testing for normality(alpha=0.05):

* For 75001 =Statistics= 0.980, p = 0.445
* For 95880 =Statistics= 0.978, p = 0.352

We will use alpha value of 0.02, giving us a 98% confidence level.

**HYPOTHESIS TESTING RESULTS:**

Given the above p-values, which indicate that the samples are normal and representations of the mean,a direct analysis of the describe() function gives us a great pointer that we should reject the null hypothesis and keep the alternative hypothesis.

Given our very low z-score, it is apparent that the data is a very close representation of the mean of the original dataset and as such, it is valid to keep the alternative hypothesis.

**CONCLUSION:**

To conclude, the processes were as follows:

* Use a large sample size to ensure representativeness of the results, increase power of the test and increase accuracy.
* Due to statistical requirements and those caused by the above statement, a two sample z test was used.