**HYPOTHESIS TESTING REPORT**

**CONTEXT:** I will work as a Data Scientist for the Autolib electric car-sharing service company to investigate a claim about the blue cars from the provided Autolib dataset.

The claim is that the average number of Bluecars taken on weekdays is different from the average number of Bluecars taken on weekends.

**PROBLEM STATEMENT**

The data that we will be describing is on the Autolib network (car-sharing and recharging) and the random variables under investigation are Bluecar taken, day type and day name (a variable we will create).

The null and alternative hypothesis that we are testing are as follows;

**Null hypothesis:** There is a difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

**Alternative hypothesis:** There is no difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

This hypothesis is important/interesting because we investigate if the average number of Bluecars taken changes when people have more free time over the weekend (assuming more people are not working over the weekend).

**DATA DESCRIPTION**

The data description of the dataset we are using is as follows

1. Postal code - postal code of the area (in Paris)
2. Date - date of the row aggregation
3. N\_daily\_data\_points - number of daily data points that were available
4. dayOfWeek - identifier of weekday (0: Monday -> 6: Sunday)
5. Day\_type - weekday or weekend
6. BlueCars\_taken\_sum - Number of bluecars taken that date in that area
7. BlueCars\_returned\_sum - Number of bluecars returned that date in that area
8. Utilib\_taken\_sum - Number of Utilib taken that date in that area
9. Utilib\_returned\_sum - Number of Utilib returned that date in that area
10. Utilib\_14\_taken\_sum - Number of Utilib 1.4 taken that date in that area
11. Utilib\_14\_returned\_sum - Number of Utilib 1.4 returned that date in that area
12. Slots\_freed\_sum- Number of recharging slots released that date in that area
13. Slots\_taken\_sum - Number of recharging slots taken that date in that area

The source of the data is from the Autolib website and the method of data collection is the Archival research; where we had access to manuscripts, documents or records from libraries, depositories and the internet to get our data.

Please find the descriptive statistics in the python colab notebook, link is below.

**HYPOTHESIS TESTING PROCEDURE**

How I plan to test our hypothesis is by first sampling our data through stratified sampling where we will use the variable day name as the sampling frame and a sample size of 70% of the population, which comes to 7000.

With an appropriate sample, we move on to choosing the test statistic that we will use for the analysis.

The logic behind the null (ie there is a difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends) and alternative hypothesis (there is no difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends) is that through this we are able to come up with insights that investigate the claim that there is a difference between the Bluecars taken on weekdays and weekends.

The test statistic we are going to be using is the z test. The reason for choosing the z test is because we are testing whether two population means are different when the variances are known and also we have a population that is non normal with a sample size that is large. The alpha level we will use is 0.05.

**HYPOTHESIS TESTING RESULT**

The value of the test statistics ie zscore is -10.362477294999469

The p value is 0.0000000 and the result of our hypothesis test is to reject the null hypothesis that states there is a difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

**DISCUSSION OF TEST SENSITIVITY**

From our test we have come to a conclusion to reject the null hypothesis, meaning we choose our alternative hypothesis that states that there is no difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

The power of the test (1-ß) for the given alpha level is the probability of rejecting the null hypothesis when the null hypothesis is false giving us a type 2 error

(1-ß) = 1 - 0.05 = 0.95

The effect of changing the sample size is that as the sample size gets larger, the z value increases therefore we are more likely to reject the null hypothesis

**SUMMARY AND CONCLUSION**

The process of the project is as follows

1. Defining the Question
2. Reading the Data
3. Checking the Data
4. External Data Source Validation
5. Tidying the Dataset
6. Exploratory Analysis
7. Statistical Sampling
8. Statistical distribution.
9. Hypothesis Conclusion.

Our hypothesis was;

**Null hypothesis:** There is a difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

**Alternative hypothesis:** There is no difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

Where we performed the z test to come up with the result of rejecting the null hypothesis, meaning that there is sufficient evidence to conclude that there is no difference between the average number of Bluecars taken on weekdays and the average number of Bluecars taken on weekends.

We were also able to see this in the time series analysis.

For the complete hypothesis test analysis please refer to the colab notebook;

<https://colab.research.google.com/drive/19DGCrEObISyZDIehe45ZQ4JGeZwjOnGM?usp=sharing>