**PROBLEM STATEMENT**

This is an analysis was performed on the Autolib Electric Car Share data set to investigate claims regarding the number of blue cars taken as provided by the Autolib Dataset. The datasets used are as follows:

Main Autolib Dataset - <http://bit.ly/DSCoreAutolibDataset>

Description Dataset - <http://bit.ly/DSCoreAutolibDatasetGlossary>

Two main claims will be examined with the first exploring whether or not there was a significant difference in the number of Blue cars taken on Monday and those on Friday.

**Null Hypothesis:** There is no difference between the number of blue cars taken on Monday and number of blue cars taken on Friday.

**Alternative Hypothesis:** There is a difference between the number of blue cars taken on Monday and number of blue cars taken on Friday.

The second claim examines differences between the number of blue cars taken over the weekend in two selected postal codes, based on similar daily points. The postal codes examined were (75001 and 92000) as they both shared fairly similar daily data points with a slight (0.5 difference).

**Null Hypothesis:** There is no difference between the number of blue cars taken on the weekend in postal area 75001 and postal area 92000.

**Null Hypothesis:** There is a difference between the number of blue cars taken on the weekend in postal area 75001 and postal area 92000

Exploring the hypothesis from this angle allows us to monitor performance of the blue car over the unique time periods which are weekdays and weekends. Examining weekday patterns could reveal surges over the week while narrowing down to postal codes over the weekend allows a more detailed analysis. We can compare how different locations perform over the weekend, perhaps revealing which ones are the most lucrative.

**DATA DESCRIPTION**

The data set contained 16,085 rows and 13 columns which have the following descriptions:

● Postal Code - postal codes for different areas in Paris

● The Date of action.

● N\_daily\_data\_points - number of daily data points available for aggregation that day.

● Day of the week : 0 for Monday, 1 for Tuesday, 2 for Wednesday, 3 for Thursday, 4 for Friday, 5 for Saturday and 6 for Sunday.

● The day type - weekday or weekend

● BlueCars\_taken\_sum - number of blue cars taken that date in that area.

● BlueCars\_returned\_sum - number the number of blue cars returned that date in that area.

● Utilib\_taken\_sum - number of Utilib cars taken that date in that area.

● Utilib\_returned\_sum - number of Utilib cars returned that date in that area.

● Utilib\_14\_taken\_sum - number of Utilib 14 cars taken that date in that area.

● Utilib\_14\_returned\_sum - number of Utilib 14 cars returned that date in that area.

● Slots\_freed\_sum - number of recharging slots released that date in that area.

● Slots\_taken\_sum - number of recharging slots taken that date in that area.

The data source is obtained from the Autolib Company and provided by Moringa LMS, which prompted secondary data collection methods. The models that seemed consistent with the data were normal, with the variable of interest being the sum of blue cars returned relative to certain parameters. The variable was examined relative to the day of the week i.e Monday and Friday. It was also examined relative to the mentioned postal codes over the weekend.

**HYPOTHESIS TESTING PROCEDURE**

To test the hypothesis, the data was split into different horizons namely weekdays and weekends, further being divided to data frames containing information from Monday and Friday. This resulted into the Mondays dataset with 237 rows and 13 columns and Fridays dataset with 227 rows and 13 columns. It was also further broken down to a Postal dataset containing weekend values from which the selected area codes were obtained with 414 rows and 13 columns. This division made it possible to calculate the required t and z statistics as parameters such as mean, standard deviations and standard errors could be computed providing a foundation for the critical and p-values to be calculated.

The first hypothesis tested was H₀ :μ(Monday) = μ(Friday)

H₁ :μ(Monday) ≠ μ(Friday

The second hypothesis tested was : H₀ :μ(92000) = μ(75001)

H₁ :μ(92000) ≠ μ(75001)

The first statement is interesting since it allows us to monitor whether there are variations of the blue cars taken between the start and the end of the week. The second provides more details on weekend performances in certain areas. Through examining measures of central tendencies, we could determine which days of the week and which postal areas on the weekend are the busiest. This makes strategic resource allocation more efficient. To test on variations of car usage between Mondays and Fridays, a z-test was used since the length of the dataset was larger than 30. To examine differences between the selected postal codes over the weekend, a t-test was used since the dataset contained 8 items. The data was assumed to be normally distributed due to factors such as independence and data clustering. The alpha level for both tests was set at 0.05 which resulted in a 95 % confidence interval.

For the first test, the null hypothesis is rejected if the p-value is less than alpha or is not rejected if the p-value is greater than alpha. For the second test, the null hypothesis is rejected if the test statistic is found to be greater than the critical value, and fails to be rejected if the converse results are observed.

**HYPOTHESIS TESTING RESULTS**

We rejected both null hypotheses which stated:

**Null Hypothesis:** There is no difference between the number of blue cars taken on Monday and number of blue cars taken on Friday.

**Alternative Hypothesis:** There is a difference between the number of blue cars taken on Monday and number of blue cars taken on Friday.

**Null Hypothesis:** There is no difference between the number of blue cars taken on the weekend in postal area 75001 and postal area 92000.

**Null Hypothesis:** There is a difference between the number of blue cars taken on the weekend in postal area 75001 and postal area 92000.

For the first test H₀ :μ(Monday) = μ(Friday)

H₁ :μ(Monday) ≠ μ(Friday

The test statistic was: -31.21237810698284 and the p value 7.237721174341743e-214 which led to rejection of the null hypothesis. The population point estimate for the average number of blue cars taken on Monday was 110.83890338161203 while the population point estimate for the number of blue cars taken on Friday was 127.16744610280581.The sample point estimate for the number of blue cars taken on Monday was 101.42999535302694 while the sample point estimate for the number of blue cars taken on Friday was 122.5309086000665. The 95% confidence interval for this parameter on Mondays was (90.71037793987082, 131.60891439374333) while that of Fridays was (103.07326336791935, 151.04034297289965.

For the second test H₀ :μ(92000) = μ(75001)

H₁ :μ(92000) ≠ μ(75001)

The test statistic was: 0.16819101386167865 and the critical value was -3.1824463052842638 which led to rejection of the null hypothesis. The point estimate for the mean of blue cars taken for the dataset was 128.495 with a 95% confidence interval of (123.06139032759609, 128.7925108225438).

**DISCUSSION OF TEST SENSITIVITY**

The power of the test was 95% while the possibility of committing a type one error was 5% which was fairly low given the data set. Increasing the sample size would decrease the probability of a type two error which leads to failure to reject the null hypothesis when it is false.

**SUMMARY AND CONCLUSION**

The project split the dataset into various samples that were easier to analyse and revealed the following: the number of blue cars taken varies among the various days of the week, with activity being different between weekdays and weekends. Furthermore, blue car usage over the weekend also varies among the various postal codes. The observations could be attributed to different activities over the week such as Monday being the start of the business week and Friday being the end. Different area codes also support different activities at social, professional and recreational capacities which explains varied usage over a similar timeframe which is the weekend. The results presented were found to be valid and can be applied to decision making.