**INDEPENDENT PROJECT 9**

INVESTIGATION OF USAGE OF BLUE CARS IN PARIS FOR THE AUTOLIB COMPANY

Working as a data scientist for the Autolib electric car-sharing service company, we are tasked to investigate a claim about the usage of blue-car(a category of electric cars) in Paris.

1. **Problem Statement**

In this project, the procedure being done is hypothesis testing to confirm a claim that the mean number of blue cars borrowed is greater than 120. In addition it is to be determined if there are more vehicles being borrowed on weekends tha

Hypotheses

Null hypothesis : The mean of blue cars borrowed is not greater than 120

Alternative hypothesis : The mean of blue cars borrowed is greater than 120

H0 : μ <= 120

Ha : μ > 120

It is important to investigate this hypothesis since by doing so, it will be determined if the number of borrowers is rising hence an expectation of higher profits or otherwise.

1. **Data description**

In this dataset there are 13 variables as follows

* Postal code
* Date (date of the row aggregation)
* Data points(number of daily data points available for aggregation on that day)
* Day of the week(identifier of weekday (0: Monday -> 6: Sunday))
* Day type(weekday or weekend)
* Sum of Blue cars taken(Number of bluecars taken that date in that area)
* Sum of Blue cars returned(Number of bluecars returned that date in that area)
* Utilib cars taken(Number of Utilib taken that date in that area)
* Utilib cars returned(Number of Utilib returned that date in that area)
* Utilib 14 cars taken(Number of Utilib14 taken that date in that area)
* Utilib 14 cars returned(Number of Utilib14 returned that date in that area)
* Sum of freed slots(Number of recharging slots released that date in that area)
* Sum of taken slots(Number of recharging slots taken that date in that area)

The variable being tested here is the mean of the blue cars taken in the dataset.

This project will need only day of the week, day type and sum of blue cars taken.

The data source is a dataset from the Autolib company records.([link](http://bit.ly/DSCoreAutolibDataset)) The data was collected through daily tallying of numbers of cars being borrowed and returned.

Descriptive statistics

* The mean of the sum of taken blue cars(**μ**) is 129.926951
* Upper quartile = 135.0
* Median = 40.0
* Lower quartile = 20.0

1. **Hypothesis testing**

The hypothesis testing is being done by

* Formulating and understanding the null and alternative hypotheses
* Deciding on a significance level
* Calculating the test statistic and the corresponding p-value
* Drawing a conclusion based on the result of the test

The hypotheses for this project are important since by knowing the estimates of the mean, one is able to guage the trends over some time, if this is done repeatedly, and the resulting cumulative information may give insights that steer major decisions, such as opening a new station, or increasing the number of cars in a certain station, leading to maximizing profits.

In this test, a z test will be used since we have all the parameters for it (that is **μ, x̄, σ,**and **n**. Moreover, the sample size is well over 30

**4. Hypothesis Testing result.**

The result showed that the null hypothesis is not to be rejected since the p-value is greater than α

* P-value = 0.1 while α = 0.05
* The z test statistic was -1.2385868646565052
* The interval around the mean is [120.23114474169826, 155.34265525830176]

**5. Summary and conclusion**

This project involved analysing car sharing data for the Autolib company. Hypotheses were made surrounding the claim. The Python language was used to test the hypotheses.

Besides, univariate and bivariate analyses were done to record summaries about the properties of the dataset. In addition are point estimate and interval calculations.

The hypotheses testing led to the conclusion that there is no sufficient evidence to show that the mean of the population is greater than 120.

In conclusion, The decision made through this testing is to not reject the null hypothesis which states that the population mean is less than or equal to 120

Sensitivity of the testing.

The probability of making a type I error is 0.05 since we used this as our significance value.

The power of the test is 1 - β which is 0.9 since β = 0.1(in most cases β is estimated to be 0.1 and may not change significantly as α)