ELECTRIC CAR SHARING PROJECT

1. BUSINESS UNDERSTANDING

Link to Google Colab Notebook: [Python Notebook]
Link to GitHub Repository: [GitHub Repo]

1.1. UNDERSTANDING THE PROBLEM

The role assumed here is that of a Data Scientist working for an electric car-sharing service company. The task is to process data from the different stations from which the cars are dispatched in order to understand usage of the electric cars over time. This will be achieved by answering the research question: Is the mean of the number of blue_cars taken from postcodes starting with '75' similar to that of the other postcodes. The '75' postcodes represent the most well known tourist destinations of Paris, France e.g. The Louvre (Museum where the famous MonaLisa painting resides) with postcode 75001, and, The Eiffel Tower (an iconic tower famously sold twice by the famous con man Victor Lustig, twice, to unsuspecting investors in the early 1900s, despite it being a landmark owned by the government.) with postcode 75007.

1.2. PROBLEM STATEMENT

The Problem statement is to determine if the mean of the number of blue_cars taken from postcodes starting with '75' is at least similar to that of all the Paris postcodes. To investigate this, our hypothesis will be:

1. The Null Hypothesis is that the mean of blue_cars taken in postcodes starting with '75' during the weekdays is greater than or equal to that of all the Paris postcodes during the weekdays.

2. The Alternate Hypothesis is that the mean of blue_cars taken in postcodes starting with ‘75’ during the weekdays is less than that of all the Paris postcodes during the weekdays.

Paris is one of the most visited areas in the world by tourists, and due to this, the economic activity in Paris is vibrant, with industries related to tourism and hospitality thriving in these regions. The ‘75’ postcodes also represent areas in and around the city centre, which additionally serves as the economic hub of France. It would therefore be prudent to investigate if the use of blue_cars in these regions is similar, considering the traffic they receive due to the economic activities in these areas.

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

The data was collected from Autolib Dataset [Link] and the description of the dataset was found on [Link]. The City of Paris, together with the Île-de-France region, was the first major European city, which successfully deployed a public electric car-sharing program. The program, named Autolib, was initiated by the mayor of Paris, Bertrand Delaone who was searching for another traffic option to complement the city’s much acclaimed bike sharing service, Velib. Autolib is operated by the Bolloré Group enterprise, which won the contract to develop the service and to supply the area with electric cars and stations [Link].

2.2. DATA DESCRIPTION

Column	**Description**
Postal code	postal code of the area (in Paris)
date	date of the row aggregation
n_daily_data_points	number of daily data points that were

	available
dayOfWeek	identifier of weekday (0: Monday -> 6: Sunday)
day_type	weekday or weekend
BlueCars_taken_sum	Number of bluecars taken that date in that area
BlueCars_returned_sum	Number of bluecars returned that date in that area
Utilib_taken_sum	Number of Utilib taken that date in that area
Utilib_returned_sum	Number of Utilib returned that date in that area
Utilib_14_taken_sum	Number of Utilib 1.4 taken that date in that area
Utilib_14_returned_sum	Number of Utilib 1.4 returned that date in that area
Slots_freed_sum	Number of recharging slots released that date
Slots_taken_sum	Number of recharging slots taken that date in

The column BlueCars_taken_sum holds our target variable from which we will get the mean of the postal codes. The other columns will be features from which correlation of the data will be analysed.

2.3. SAMPLING STRATEGY

2.3.1. TARGET POPULATION

The target population is Paris Metropolitan in France. The dataset contains data from January 1st 2018 up until 19th June 2018.

2.3.2. SAMPLING METHOD

We will use a Probabilistic sampling method to ensure randomness as this will yield an unbiased result. The Probabilistic sampling method chosen is Stratified Random Sampling and it will be appropriate, considering the data and the goals of this project.

2.3.3. SAMPLE FRAME AND SIZE

Considering the postcodes are in the city area of Paris, the city will be extracted from the larger Paris Metropolitan area via the postcodes, since the postcodes of Paris City start with '75'. This will ensure better accuracy of results, since we will be gathering random samples from areas that are appropriate to the city of Paris, from where most economic activities take place, rather than the residential areas.

2.4. DESCRIBING THE QUESTION

2.4.1. SPECIFYING THE QUESTION

Determine if the mean of the number of blue_cars taken from postcodes starting with '75' is at least similar to that of all the Paris postcodes. To investigate this, our hypothesis will be:

1. The Null Hypothesis is that the mean of blue_cars taken in postcodes starting with '75' during the weekdays is greater than or equal to that of all the Paris postcodes during the weekdays.

2. The Alternate Hypothesis is that the mean of blue_cars taken in postcodes starting with '75' during the weekdays is less than that of all the Paris postcodes during the weekdays.

2.4.2. DEFINING THE METRIC FOR SUCCESS

It will be deemed successful if the Null hypothesis fails to be rejected, i.e. if it is true.

2.4.3. EXPERIMENTAL DESIGN

1. Loading Datasets and Preparing the Data.
2. Data Cleaning to deal with Anomalies and Outliers.
3. Exploratory Data Analysis (Univariate and Bivariate Analysis).
4. Hypothesis Testing to Implement the Solution.
5. Conclusions and Recommendation.

2.5. HYPOTHESIS TESTING PROCEDURE

The procedure to be followed will be as below:

- Specifying the Null and Alternate hypothesis (as we have already done).
- Setting the significance level (alpha) at 5 % (0.05) as this is a standard agreeable by proxy. This means that the confidence interval will be 95% (0.95).
- Calculating the z-score since the sample will include more than 100 rows of data.
- Calculating the corresponding p value to ensure that any type I error is mitigated hence giving the result more confidence.
- Drawing the conclusions from the relationship between the p value and the level of significance (alpha).

For the year 2018, the most visited country in the world was France with 89 million visitors [Link], with Paris being the most visited city [Link]. Eiffel Tower and the Louvre Museum being the two most known destinations of tourism in Paris [Link]. Furthermore, apart from the local population that uses the cars for their errands and co-operates hiring them instead of buying their own fleet, tourists also use them en masse due to the hype around electric cars [Link]. Due to the reasons above, the vibrant nature of their locations should offer an interest when the averages of customer engagement with BlueCars are considered, when compared to the rest of the Paris Metropolitan.

3. DATA PREPARATION

3.1. SELECTING DATA

We'll use all of the columns relevant to the Blue Car as they are all relevant to the study. Columns related to Utilib and Utilib_14 will be dropped.

3.2. DATA CLEANING

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

The first thing done was to rename the columns to make them uniform and readable. The columns were then checked to see if they were of the appropriate types / dtypes. After this, missing values in the datasets were checked for and were found to be none. The data was also found to be consistent there being no duplicated data.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

4.1.1. UNIVARIATE DATA ANALYSIS

a) Numerical Data

There were a lot of outliers in the BlueCar taken (2215), BlueCar returned (2213), Slots freed (3235) and Slots taken (3234) columns. These are too many to remove as this will affect the accuracy of the data analysis, and the result could be inconclusive and/or incorrect. The outliers suggest that the data could possibly be data that does not have a normal distribution.

b) Categorical Data

The category of interest is the week category 'day_type' that classifies the days into weekday or weekend. The weekday accounts for almost 12,000 rows of data whereas the weekend accounts for about 4,000 rows. Most of the activity seems to happen during the weekday, therefore, we'll be using samples from the weekday for the hypothesis testing. Furthermore, since we'll be using the z-score in our analysis, the larger the data, the more accurate the results will be.

c) Summary Statistics

Statistic	*Bluecar taken*	*Bluecar returned*	*Slots freed*	*Slots taken*
mode	12	13	0	0
range	1352	1332	360	359

Standard deviation	185.427	185.502	52.120	52.146
Variance	34383.016	34410.819	2716.522	2719.208
1st Quartile (Q1)	20	20	0	0
Median (Q2)	46	46	0	0
3rd Quartile (Q3)	135	135	5	5
Skewness	2.406	2.412	2.597	2.597
Kurtosis	6.173	6.186	6.455	6.443
mean	125.927	125.913	22.630	22.630

d) Univariate Analysis Recommendation

The data is heavily skewed to the right i.e. leptokurtic, as was suspected due to the large number of outliers. This suggests that our initial decision to keep them is justified as this is not a normally distributed dataset. Furthermore, the Bluecar taken and returned columns seem to have similar statistical bearing, hence using either as the target variable in place of the other would be justified. We have decided to use the Bluecar taken column as our target variable, however, future analysis can be done with the Bluecar returned column and the results compared.

4.1.2. BIVARIATE DATA ANALYSIS

a) Numeric

Strong positive correlation was found among BlueCar taken, BlueCar returned, Slots freed and Slots taken variables. They were linear correlations with pearson's coefficients greater than 0.94. The pairplots and heatmap further cemented this observation.

b) Categorical

Contrary to what we believed in the univariate analysis, the weekend seems to have more activity. This could be due to the weekday having more days (5) than the weekend (2), however, more Blue cars were taken during the weekends.

c) Bivariate Analysis Recommendation

From the above, we can see that the weekends are the busiest, and this could be due to weekend activities that people have time to engage in, requiring them to use the service. Even though this contradicts the univariate analysis, we will still use the weekday to conduct our hypothesis testing, since the weekdays have more days, hence more rows of data to work with. The more the data, the better our model will be.

4.2. HYPOTHESIS TESTING

a) Specifying the Null and Alternate hypothesis.

- The Null Hypothesis is that the mean of blue_cars taken in postcodes starting with '75' during the weekdays is greater than or equal to that of all the Paris postcodes during the weekdays.
- The Alternate Hypothesis is that the mean of blue_cars taken in postcodes starting with '75' during the weekdays is less than that of all the Paris postcodes during the weekdays.

b) Setting the significance level (alpha) at 5 % (0.05) as this is a standard agreeable by proxy. This means that the confidence interval will be 95% (0.95).

c) Sampling technique will be Stratified Random Sampling because the sample chosen contained 22 post codes, each with either 111 or 112 rows of data. The technique will select 30 samples randomly from each postcode.
d) Calculating the z-score since the sample will include more than 100 rows of data.
e) Drawing the conclusions from the relationship between the p value and the level of significance (alpha).

4.2.1. HYPOTHESIS TESTING RESULTS

Previously, we had found the population mean to be 125.927 with a population deviation of 185.427. The sample mean upon conducting hypothesis testing was 354.38 with a sample deviation of 212. These values tell us that the mean of the City Centre postcodes i.e. those that start with '75' are greater than the population mean, this is in line with the null hypothesis. To confirm this, the Z-score was calculated and was found to be 1.23. The critical Z-score value for an alpha of 0.05 is 1.645 (Since it is a one-tailed test). If our calculated Z-score was above 0.96, we would reject our Null hypothesis, however since it is below, we cannot reject it, since the 95% Confidence Interval has not been crossed. The p calculated was 0.109 which is greater than the alpha (0.05) level. It is therefore not statistically significant, indicating strong evidence for the null hypothesis. This cements the decision to fail to reject the Null Hypothesis.

5. CONCLUSION

The Test sensitivity was done by changing the sample size to 90 samples per postcode. There was no significant difference, proving the rigidity of the method. Considering the mean for the '75' code postcodes was higher than the mean of all the postcodes, the Z-score (1.23) being less than the Z-critical (1.625), and the p value (0.109) being greater than the alpha (0.05), the Null Hypothesis failed to be rejected. This means that indeed, these postcodes that start with '75' i.e. those related to regions within the City Centre and it's immediate environs, have a higher mean of bluecars being taken daily.

6. RECOMMENDATION

The test suggests that the City Centre area of Paris experiences more customer engagement with the Bluecars. This suggests that the economic activities related to the City have an effect on the rate at which Bluecars are used. For future research, data regarding why the customer used a bluecar, e.g. Leisure, Work, Tourism etc, should be gathered to enable analysis that will show which activities boosted the use of the Bluecars most.