DATA REPORT FOR SHARED ELECTRIC CAR SERVICE

1.0 Introduction

According to wikipedia **Autolib'** was an electric car sharing service which was inaugurated in Paris, France, in December 2011. It closed on 31 July 2018. It was operated by the Bollore' industrial group, and complemented the city's bike sharing system,Velib', which was set up in 2007. The Autolib' service maintained a fleet of all-electric Ballore' Bluecars for public use on a paid subscription basis, employing a citywide network of parking and charging stations. As of 3 July 3, 2016, 3,980 Bluecars had been registered for the service and had more than 126,900 registered subscribers; Autolib' furthermore offered 1,084 electric car stations in Paris agglomeration with 5,935 charging points. Since the beginning of its operations in Paris, Autolib' expanded its business to the cities of Lyon and Bordeaux. Bolloré also signed deals to begin operating offshoots of Autolib' in London and Indianapolis in 2015, Turin in 2016 and Singapore in 2017.

2.0 Problem Statement

Autolib electric car-sharing service company to investigate a claim about the blue cars from the provided Autolib dataset.

In an effort to do this, we need to identify some areas and periods of interest via sampling stating the reason to the choice of method, then perform hypothesis testing with regards to the claim that we will have made. An example of claim to test would be "Is the number of Bluecars taken in area X different than in area Y? Is it greater in area X than in area Z? etc". The selected periods of interest will either be weekdays or weekends but not a mix of both. You can also consider postal codes 75015 vs 75017 to some of the areas of interest.

3.0 Formulating the Hypothesis

The hypothesis will therefore be:

1. The Null hypothesis
   The number of blue cars taken from postcodes that start with '75' are higher than those taken from all the other postcodes in paris.
2. The Alternate Hypothesis
   The number of blue cars taken from postcodes that start with '75' are similar to those taken from all the other postcodes in paris.

It is important to find out if there's any difference from the operation of the autolib cars depending on the location. This will be done by establishing the number of cars taken from different locations basing on the postcodes in Paris.

The significance level will be 0.05(95%)

4.0 Data Description

Dalberg data insights obtained real time data. The main aim was to see if the number of the blue cars taken from postcodes starting with '75' are more than those taken from all other postcodes in Paris. The data used was downloaded from this link. It was collected in real time and captured various data points for the autolib company. Autolib is an electric car services in Paris that provides cars on hire. The details of the data was also provided as indicated below:

Description of Data

Column name	Explanation
Postal code	postal code of the area (in Paris)
date	date of the row aggregation
n_daily_data_points	number of daily data poinst that were available for aggregation, that day
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 in that area
Slots_taken_sum	Number of rechargign slots taken that date in that area

5.0 Data Preparation

After understanding the data provided in details, the next thing was preparing the data for analysis by cleaning the data. This was done by first previewing the data, checking the shape of the data, data types, outliers, missing values, duplicates, removing the

white spaces, renaming the columns, ensuring the data types are in the right format and finally saving the new dataframe for further analysis.

Below are the steps followed in preparing the data for analysis:

a. Importing the libraries
b. Loading the data
c. Identifying the dataframes' data type
d. Verifying Data Quality
e. Loading the cleaned dataframe
f. Data Analysis and Findings
g. Giving Recommendation

The softwares and libraries used were (Github, VsCode, Python, Jupyter, Numpy, Pandas, Matplotlib, seaborn, Scipy, Stats, Shapiro etc)

6.0 Analysis and Hypothesis Testing

I performed univariate analysis, by establishing the measures of central tendency, then thereafter performed bivariate and multivariate analysis. It was evidence that there were so many outliers in the dataset that before dropping them, the data points of various variables were spread widely from the mean leading to a leptokurtic distribution.

It was evident that the autolib blue cars were on a high demand with the highest average mean of 125.96 followed by utilib 14 with a mean of 8.6 and lastly utilib with an average mean of 3.7. The data points for the blue cars were spread out over a wide range with a standard deviation of 185. This was also because it had lots of data captured as compared to the other cars which had a low measure of variation from mean. The autolib had a standard deviation of 5 while autolib 14 had a standard deviation of 12 with a mean of 3 and 8 respectively.

This huge variation led to the dropping of the outliers which made the distribution of the curve change from having a heavy tail (leptokurtic) to a platykurtic for blue cars and mesokurtic for autolib and autolib 14. The data points for the blue cars are clustered around the mean with a standard deviation of 28.25 and a mean of 35.5. This is a good indication that the data is more reliable than when it is widely spread out around the mean. For the utilib and utilib 14, the data points are closely spread out within the mean as the standard deviation is the same as the mean.

The dataset was large as it contained 16085 rows and 13 columns . Despite noticing that the weekend seemed to have high numbers of cars in operation, as indicated in the bivariate analysis, I preferred to get the sample size from the weekday data so that it

can capture activities for the others cars for proper comparison and also giving all variables in the data a chance to be sample and obtain a good analysis. It was evident that most of the activity occurs during the week day thus this forms part of the sample size for analysis which will also be used for hypothesis testing. Since we'll be using the z-score, the larger the data, the more accurate the results will be. I therefore used a cluster sampling method. The entire paris city had data totaling to 3431 rows while the data for weekdays only had a total of 2463 rows. A clear indication that most variables will have a fair chance of being sampled.

The descriptive statistic for blue car taken and blue car returned was almost similar. This was a good sign as it creates a pattern to determine the future trends of business operations relating to blue cars in the Autolib electric car sharing company.

From the sample data ,I got a sample of 30 at a random state of 10 and obtained 660 rows and 13 columns. Having that small data as a sample size, it was effective to calculate the z-score and the p-value to determine whether to reject or accept the null hypothesis. Rejecting the null hypothesis means that the alternate hypothesis is true. I used the sample to calculate the z score and p value in hypothesis testing. I considered 5%(0.05) as my alpha level of significance.

The z-score tells us that the sample mean is 1.23 standard deviations away from the population mean. This is within the 1.645 critical value (since it is a one-tailed test), which is the area where 95% confidence level lies. Therefore, the null hypothesis couldn’t be rejected. The p-value was 0.108 and a z-score was 1.23. Which indicates that there's a 10.8% likelihood of the blue cars taken from postcodes that start with ‘75’ to be similar to those taken from all the other postcodes in Paris.

Thus making an event of the alternate hypothesis have a limited chance (10.8%) of happening. The alternate hypothesis is rejected and the null hypothesis accepted. The p-value confirmed that the rate of picking blue cars from the postcodes that start with ‘75’ is greater than those picked from the other postcodes in Paris.

I also performed a normality test from a sample dataset for a blue car taken to see if it follows a normal distribution or not. I used the Shapiro-Wilk Test and Quantile_Quantile(Q-Q) Plot. From the normality test I conclude that the distribution of the blue car taken are not normally distributed over the population. This is so because the sample of bluecar taken does not look like gaussian from Shapiro-Wilk Test and also points in Quantile_Quantile(Q-Q) Plot are not aligned along the red line. The distribution for the number of daily points, slots taken and slots freed sum were the only ones that the distribution looked gaussian with a p-value of 1.0 and a statistics of 1.0.

7.0 Conclusion

After performing the Test sensitivity, it was evident that by changing the sample size to 90 samples per postcode, there was no significant difference thus proving the rigidity of the method. It was also noted that those postcodes that begun with ‘75’ had a higher mean for blue car collection than 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 could however not be rejected. It is therefore confirmed that these postcodes that start with ‘75’ have a higher average mean of collection of bluecars daily.

8.0 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.

Github Link

The link to the github repository that contains the data preparation and analysis is found (**here**)