AUTOLIB CAR SHARING HYPOTHESIS TEST REPORT.

INTRODUCTION: BACKGROUND INFORMATION.

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 Bolloré industry and complemented the city's bike sharing system, Velib', which was set up in 2007. The Autolib' service maintained a fleet of all-electric Bolloré Blue Cars or the utility cars Utilib and Utilib 14. for public use on a paid subscription basis, employing a citywide network of parking and charging stations.

Introduced by the former Socialist mayor Bertrand Delaone, it was a proposal for the implementation of an individual transport option for people who didn't find themselves in need of a vehicle regularly, but may want to use a vehicle once in a while.

The so-called Bluecar, designed in collaboration of Bolloré with the Italian factory Pininfarina, is a tiny bubble shaped four-seater car with three doors. The car has a lithium-metal-polymer battery, can travel up to 250 km and can reach speeds of 130 km/h . Charging the car takes four hours. There are around 3,000 Bluecars and 870 Autolib stations spread throughout the region. Autolib is a widespread and a high density network with more than 4,400 parking spaces [8

1) Problem statement:

This dataset was an autolib dataset that contained details about the operation of electric car sharing service cars within Paris. It showed a compilation of dates when the blue cars, utilib and utilib 1.4 cars were picked from and returned to the respective postal codes.

The goal of this project is to demonstrate the concept of hypothesis testing by investigating a claim about the blue cars using the autolib dataset. More specifically, I will be comparing the blue cars taken and dropped off during the weekend between January and June when the data was collected.

I came up with the null and alternative hypothesis for this research as follow;

Null Hypothesis - The means of blue cars taken and cars returned during the weekends are equal.

Ho : μ1 = μ2 (where μ1 is the mean for Blue cars taken and μ2 is the mean for blue cars returned.)

Alternate Hypothesis - The average number of blue cars picked up during the weekend is different from that of the blue cars returned.

$$Ha : \mu 1 \neq \mu$$

The interest of this hypothesis is crucial for the understanding of the data and its distribution. It will also help to successfully extrapolate data from the sample to the larger population and determine whether the data from the sample is statistically significant

2) Data description

The so-called Bluecar, designed in collaboration of Bolloré with the Italian factory Pininfarina, is a tiny bubble shaped four-seater car with three doors. The car has a lithium-metal-polymer battery, can travel up to 250 km and can reach speeds of 130 km/h. Charging the car takes four hours. There are around 3,000 Bluecars and 870 Autolib stations spread throughout the region. Autolib is a widespread and a high density network with more than 4,400 parking spaces.

The dataset I used for this investigation was an open dataset about cars in Paris. It contains variables like the postal code of the area which was Paris, the dates of data collection. It also contained the number of free charging slots and slots taken respective to each postal code.The dates ranged between January and June of 2018, which also had the number of daily data points that were available for aggregation on the particular days of aggregation within the specified time periods. The blue cars that were taken and returned, the utilib data were also contained in the dataset. The problem under investigation was on the averages which would make the null and alternative hypotheses.

. Given the above null and alternative hypotheses, the main variable of interest is the number of blue cars picked up and returned. This specific attribute is available in the dataset for various postal codes during each day of the week beginning on 1/1/2018 to 6/19/2018, . The full dataset as well as dataset description can be found here and here respectively. Below is a summary of the attributes used in this analysis

Column Name	Column Description	Data Type	No. of Records
Postal code	This attribute identifies the postal code where the observation was recorded	Integer	16085
Date	Date when the observation was recorded	Date	16085
Day of Week	The day of the week when the observation was recorded	String	16085
Day Type	Whether the day was a weekday or a weekend	String	16085
Blue Cars Taken	Total number of blue cars picked up on a particular day for a particular postal code	Integer	16085
Blue Cars Returned	Total number of blue cars returned on a particular day for a particular postal code	Integer	16085

It was a set of data that was already collected. However, if i were to collect such comprehensive data, i would use my data response team to go out in the field, collect the data and perform the analysis from which conclusions would later on be made.

Summary of Data Cleaning Sampling Technique

Some of the preliminary steps performed before the hypothesis testing procedure include data cleaning. Our data did not have any missing values and I retained the outliers because they contained important information for the test univariate and bivariate data analysis as well as discussion of the sampling approach. Specific details and results of these steps can be found on the python notebook.

3) **Hypothesis testing procedure**

Sampling Technique

For the hypothesis test, we started by selecting a 70% sample from the 16,085 entries. We obtained a sample of 3,179 entries from the 4541 entries(weekend entries).

The hypothesis is tested on a sample of the population. Since I wanted to compare samples from different postal codes, I used stratified random sampling. Pick up and drop off stations are unique to postal codes. A pickup station can only exist in one postal code or city. Therefore, a car pickup or drop off recorded in a station belongs to a unique city. Using only one stratum ensures that each record of usage has an equal chance of being selected during sampling. Each stratum has no overlapping sample therefore no bias in the sampling technique. The sample size for each stratum is proportionate to the target population size of that particular stratum

The logic behind my null and alternative hypothesis is that, since it was not going to be easy for me to manually group the data and the sample in them, I decided to work with the mean. It was interesting for me to know whether the average number of cars that were picked in a day was similar to the average number of cars that were returned on that very day. This would determine the future trends of business operations relating to blue cars in the autolib electric car sharing company.

From my stratified sampling to conduct hypothesis testing, I used a two-sample z-test and calculated the p-value in order to either reject or accept the null hypothesis. Below are the reasons why I chose to use the z-test as the appropriate test statistic:

1. The sample size is greater than 30.
2. Data points are independent from each other.

3. The sample data has been randomly selected from a population, so each item has an equal chance of being selected.
4. The data did not follow a normal distribution

The level of significance used in the hypothesis testing is 0.05 or 5%. Therefore, if the p-value calculated from the test statistic is less than 0.05 then we reject the null hypothesis. If the value is greater than or equal to 0.05, we accept the null hypothesis.

4) **Hypothesis testing results**

I used *ztest* from *statsmodels.stats* libraries in python to conduct the z test and calculate both the z statistic and the p-value and obtained the following results:

This is our p-value 0.8481097406849015

This is value of the z test -0.19153081478438222

We fail to reject the Null Hypothesis

From the hypothesis test, we found that there was not sufficient evidence to prove that the average means of the blue cars taken and the bluecars returned are not equal. The z-score was -0.19153081478438222 and our p value was

0.8481097406849015 which was greater than our significance level and as a result, the null hypothesis was not rejected. The z-critical value was 1.959963984540054 with the confidence interval being :

Confidence interval:

(145.4251659011278, 160.7220502045658)

5) **Discussion of test sensitivity**

Sensitivity in a statistical test is the measure of performance of a binary classification test. It measures the proportion of the actual positive i.e. the probability of a null hypothesis being true. In this case the sensitivity was 91%.

6) Summary and conclusions

The project was comprehensive and demanding. I performed exploratory data analysis with hypothesis testing as its implementation. Conclusively, I failed to reject the null hypothesis because there was not enough evidence for me to reject the null hypothesis. Some of the findings include:

a) There were no missing values in the data.
b) I retained the outliers since they contained important information
c) I used a stratified sample for the hypothesis test'
d) The data did not follow a normal distribution
e) We failed to reject the null hypothesis

We have successfully defined the null and alternate hypothesis, executed the sampling technique and carried out hypothesis testing which led to the failing to reject the null hypothesis. We concluded that there was not sufficient evidence to prove that the average means of the blue cars taken and the bluecars returned are not equal.

Link to github repository containing worked out python notebook