Autolib electric car-sharing service company

A report on hypothesis testing of the Autolib dataset

Introduction

The data to carry out the study was obtained from The Paris Autolib electric car-sharing sharing scheme. Bolloré is the group that ran Autolib and is named after Vincent Bolloré. An autolib is an electric car sharing service with a signature grey fleet of vehicles rolled out in Paris and the surrounding Ile-de-France region in 2011.
It eventually saw 4,000, battery-powered cars stationed at over 1,100 self-service docking stations across the city and the surrounding suburbs.It worked in the same way as the bike sharing scheme Velib' where users sign up for a subscription and pay a small fee each time they use the car depending on how long they use it for.

Problem Statement

In an effort to do this, we need to identify some areas and periods of interest via sampling stating the reason for the choice of method, then perform hypothesis testing with regards to the claim that we will have made. An example of a claim to test would be "Is the number of Blue Cars 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 are either 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.

The dataset used was named Autolib Dataset.The provided dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging). The variables under investigation include:

- Postal code
- Date
- Number of daily data points. (n_daily_data_points)
- Day of week. (dayOfWeek)
- Type of day (day_type)
- Sum of taken blue cars (BlueCars_taken_sum)
- Sum of returned blue cars (BlueCars_returned_sum)
- Sum of taken Utilib cars (Utilib_taken_sum)
- Sum of returned Utilib cars (Utilib_returned_sum)
- Sum of taken Utilib 14 cars (Utilib_14_taken_sum)
- Sum of returned Utilib 14 cars (Utilib_14_returned_sum)
- Sum of freed slots (Slots_freed_sum)
- Sum of taken slots (Slots_taken_sum)

Hypotheses

First hypothesis

H_o: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekday **p = 92723**

H_1: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekday **p ≠ 92723** (claim)

This hypothesis is stated in order to find out if the blue cars used in different regions; 75015 and 75017 are equal or the same. This is interesting and can help in determining which areas have more demand for blue cars than others during the weekday.

Second hypothesis

H_o: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekdays; **population </= 92723.**

H_1:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekdays; **population > 92723** (claim)

This hypothesis is stated so as to find out whether the difference in blue cars in the regions with postal code 75015 is less than or equal to 75017 or if the difference between postal code 75015 is greater than 75017 during the weekday

Third hypothesis

H_o: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekend **p = 47707**

H_1: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekend **p ≠ 47707** (claim)

This hypothesis is stated in order to find out if the blue cars used in different regions; 75015 and 75017 are equal or the same. This is interesting and can help in determining which areas have more demand for blue cars than others during the weekend.

Fourth hypothesis

H_o: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekends; **population </= 47707.**

H_1:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekends; **population > 47707** (claim)

This hypothesis is stated so as to find out whether the difference in blue cars in the regions with postal code 75015 is less than or equal to 75017 or if the difference between postal code 75015 is greater than 75017 during the weekends.

Data Description

In the dataset, data provided is obtained from three types of electric cars in the city of Paris,France. The different electric cars include:

- Blue cars
- Utilib
- Utilib_14

The information also includes cars taken and returned to different charging stations in different regions of different area codes. The instance at which this occurs is recorded daily for six months in the year 2018. The respective variables include:

1. Postal code -postal code of the area (in Paris)
2. Date - date of the row aggregation
3. Number of daily data points. (n_daily_data_points) - number of daily data points that were available for aggregation, that day
4. Day of week. (dayOfWeek) - identifier of weekday (0: Monday -> 6: Sunday)
5. Type of day (day_type) - weekday or weekend
6. Sum of taken blue cars (BlueCars_taken_sum) - Number of blue cars taken that date in that area
7. Sum of returned blue cars (BlueCars_returned_sum) - Number of blue cars returned that date in that area
8. Sum of taken Utilib cars (Utilib_taken_sum) - Number of Utilib taken that date in that area
9. Sum of returned Utilib cars (Utilib_returned_sum) - Number of Utilib returned that date in that area
10. Sum of taken Utilib 14 cars (Utilib_14_taken_sum) - Number of Utilib 1.4 taken that date in that area
11. Sum of returned Utilib 14 cars (Utilib_14_returned_sum) - Number of Utilib 1.4 returned that date in that area
12. Sum of freed slots (Slots_freed_sum) - Number of recharging slots released that date in that area
13. Sum of taken slots (Slots_taken_sum) - Number of recharging slots taken that date in that area

The data source was obtained through a secondary method of data collection that is from a web page. Data was acquired from Moringa School LMS which I believe was originally obtained from this link

below:https://opendata.paris.fr/explore/?refine.theme=Equipements,+Services,+Social&disjunctive.theme&disjunctive.publisher&disjunctive.keyword&disjunctive.modified&disjunctive.features&sort=modified&refine.modified=2017&refine.modified=2019 The page a reliable site with very accurate information making our data valid

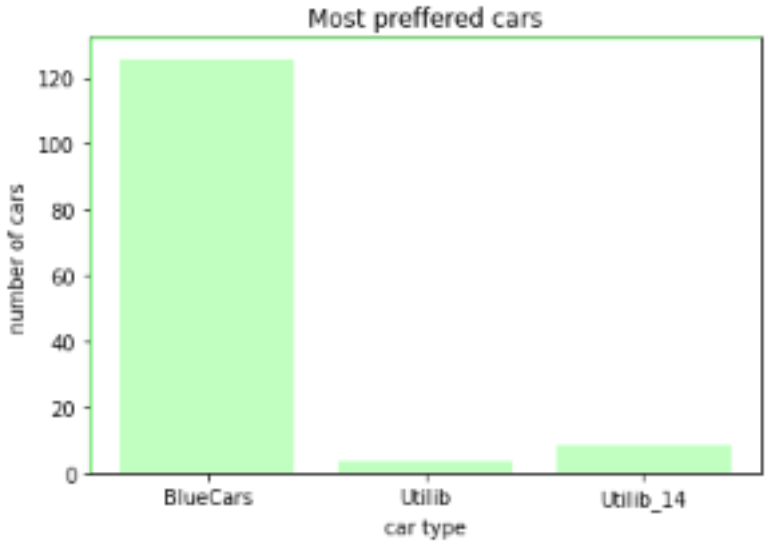
The data obtained had only three categorical variables:Postal code, Day of week. (dayOfWeek), Type of day (day_type) while others were numerical variables.

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

The process followed in the descriptive statistics

- Data Cleaning: There were no missing values in the dataset. There were outliers that were removed and the anomalies were dealt with.
- Univariate Analysis :Each variable was plotted inorder to view its observation characteristics.
- Bivariate Analysis: The variables were plotted each other in order to answer some of the bivariate questions below.

Which electric car was most used during weekdays?



Hypothesis Testing Procedure

Step 1: The first step is usually for the statistician to state the two hypotheses so that only one can be right. I will formulate the null hypothesis H_0 and the alternative hypothesis H_a.
Step 2: The next step is to formulate an analysis plan, which outlines how the data will be evaluated. Then identify a test statistic that can be used to assess the truth of the null hypothesis.
Step 3: The third step is to carry out the plan and physically analyze the sample data. This involves computing the P-value. The smaller the P-value, the stronger the evidence against the null hypothesis.
Step 4: The fourth and final step is to analyze the results and either accept or reject the null hypothesis. They compare the P-value to an acceptable significance value α (sometimes called an alpha value). If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

Why the hypotheses for the study?:
First hypothesis
$\mathbf{H_o}$: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekday **p = 92723**
$\mathbf{H_1}$: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekday **p ≠ 92723** (claim)

This hypothesis is stated in order to find out if the blue cars used in different regions; 75015 and 75017 are equal or the same. This is interesting and can help in determining which areas have more demand for blue cars than others during the weekday.

Second hypothesis
$\mathbf{H_o}$: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekdays; **population </= 92723.**
$\mathbf{H_1}$:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekdays; **population > 92723** (claim)

This hypothesis is stated so as to find out whether the difference in blue cars in the regions with postal code 75015 is less than or equal to 75017 or if the difference between postal code 75015 is greater than 75017 during the weekday

Third hypothesis

H_o: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekend **p = 47707**

H_1: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekend **p ≠ 47707** (claim)

This hypothesis is stated in order to find out if the blue cars used in different regions; 75015 and 75017 are equal or the same. This is interesting and can help in determining which areas have more demand for blue cars than others during the weekend.

Fourth hypothesis

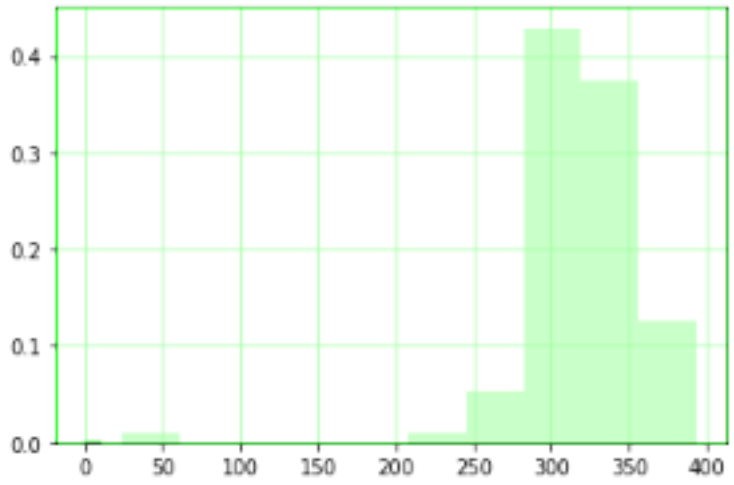
H_o: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekends; **population </= 47707.**

H_1:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekends; **population > 47707** (claim)

Test statistic

For the hypothesis I will use the z statistic, this is based on the assumption that normality of the data is normal distribution. A t test cannot be carried out since the sample sizes are greater than 30 .The figure below shows that the data is somewhat a normal distribution.Z-tests can also be helpful when we want to test a hypothesis. Generally, they are most useful when the standard deviation is known.

A graph of the distribution of blue cars in the regions 75015 and 75017 during the weekday



The alpha level for the study will be 5 %. A lower significance level will increase the chance of committing type II error. A type II error produces a false negative, also known as an error of omission.

Hypothesis Testing Results

First hypothesis

$\mathbf{H_0}$: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekday **p = 92723**

$\mathbf{H_1}$: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekday **p ≠ 92723** (claim)

Result

This is a two tailed test.There is a strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and accept the alternative hypothesis. The z score is -16.19 and p-value is 5.7757e-59. This makes the p-value less than the critical value of 1.96 and -1.96. Therefore there is enough evidence against the null hypothesis. The point estimate used was the means of both the population and sample and the standard deviation of the sample. The confidence interval was 95% for this hypothesis

Second hypothesis

$\mathbf{H_0}$: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekdays; **population </= 92723.**

$\mathbf{H_1}$:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekdays; **population > 92723** (claim)

This is a right-tailed test. There is a strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and accept the alternative hypothesis. The z score is -17.0162 and p-value is 3.1108e - 65. This makes the p-value less than the critical value of 1.645 . Therefore there is enough evidence against the null hypothesis. The point estimate used was the means of both the population and sample and the standard deviation of the sample. The confidence interval was 95% for this hypothesis

Third hypothesis

$\mathbf{H_0}$: The population of blue cars is the same in different regions with the postal codes 75015 and 75017 during the weekend **p = 47707**

H_1: The population of blue cars is different in different regions with postal codes 75015 and 75017 during the weekend **p ≠ 47707** (claim)

Result

This is a two tailed test.There is a strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and accept the alternative hypothesis. The z score is -9.633 and p-value is 6.2217e-65. This makes the p-value less than the critical value of 1.96 and -1.96. Therefore there is enough evidence against the null hypothesis. The point estimate used was the means of both the population and sample and the standard deviation of the sample. The confidence interval was 95% for this hypothesis

Fourth hypothesis

H_0: The population of blue cars in postal code 75015 is less than or equal to postal code 75017 on weekends; **population </= 47707.**

H_1:The population of blue cars in postal code 75015 is greater than postal code 75017 on weekends; **population > 47707** (claim)

This is a right-tailed test. There is a strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and accept the alternative hypothesis. The z score is -9.633 and p-value is 6.2217e-65. This makes the p-value less than the critical value of 1.645. Therefore there is enough evidence against the null hypothesis. The point estimate used was the means of both the population and sample and the standard deviation of the sample. The confidence interval was 95% for this hypothesis

Discussion of test sensitivity

Sensitivity measures how often a test correctly generates a positive result for the population who have the condition that's being tested for (also known as the “true positive” rate). Power of the test on a significant level of 1% was conducted. The findings all rejected the null hypothesis for the alternative. This shows that the test has no type 2 statistical error. There was also no difference in the findings when the sample size was reduced or added.

Summary and Conclusions

The hypothesis test for the project was formulated in line with the clients requirements. The null hypothesis and alternative hypothesis were created on the number of bluecars in the regions with postal code 75015 and 75017. The hypotheses all looked to test the claim of number blue cars on weekdays and weekends.

The first process was to create the null hypothesis followed by the alternative. The next step is to formulate an analysis plan, which outlined how the data would be evaluated. Then identify a test statistic that can be used to assess the truth of the null hypothesis. The third step was to carry out the plan and physically analyze the sample data. This involved computing the P-value. The smaller the P-value, the stronger the evidence against the null hypothesis.The fourth and final step was to analyze the results and either accept or reject the null hypothesis. Here, we reject the null hypothesis. The P-value was compared to an acceptable significance value α (sometimes called an alpha value of 0.05). If $p \leq \alpha$, that the observed effect was statistically significant, the null hypothesis was ruled out, and the alternative hypothesis is valid.

In all the hypotheses tests carried out, the null hypothesis was rejected. This shows that the test carried out had a significant observation against the null hypothesis. The sensitivity of the test carried shows no difference in the output when measured against the power of the sample and also on reducing the sample size. This shows that there was no Type II error purported in the study.