Autolib Electric Car Sharing Service.

Hypothesis Report

1. Overview

Autolib was one of the electric car-sharing services founded in Paris, France in 2011. Autolib electric cars were referred to as bluecars and new models introduced Utilib and Utilib 1.4. The company was a project by the mayor of France as a traffic option for the city. Through the services, the people of Paris who subscribe to this service are able to obtain a car from any of the Autolib pickup stations and can leave them at any other pick-up stations in the city. The company also provides car charging services at various charging stations.

1.1 Research Question

The main question that this analysis sort to investigate was whether the use of blue cars was changing over time(either increasing or decreasing) or if it remained the same across different time periods.

1.2 Objective

This research seeks:

- To analyze the adoption of Blue cars over time.
- Analyze whether there was a significant change in the usage of blue cars over time.

1.3 Success Criteria

For the analysis to be successful, it should provide sufficient insights as to whether there is any statistical significance in the difference in the mean number of cars used between two time periods.

1.4 Assessing the situation

1.4.1 Resources

I. Data analyst experts

II. Data sets:

Autolib data [here]

Data description[here]

III. Sofware(Python, Github, Google colab, Google suite)

1.4.2 Assumptions

I. The data provided is correct

1.4.3 Constraints

The data may be biased

2. Data Understanding

2.1 Data Mining Goals

The data mining goals for this analysis are:

- Perform exploratory analysis and analyze how Autolib's blue car performed across different regions and time zones.
- Perform hypothesis analysis and assess if there is a difference in means between two different time periods.

2.2 Data Description

The main dataset used in this study was collected from various Autolib stations and car logs of the electric vehicles. All the data needed for the study was provided in one data file.
The original data had 16,086 rows in 13 columns.
The description of each column is shown in the table below.

NAME OF COLUMNS	DESCRIPTION
Postal code	Postal code of the area (in Paris)
Date	date of the row aggregation
n_daily_data_points	The number of daily data points that were available
DayOfWeek	identifier of weekday (0: Monday -> 6: Sunday)
day_type	weekday or weekend
BlueCars_taken_sum	The number of bluecars taken that date in that area.
BlueCars_returned_sum	The 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	The number of Utilib 1.4 taken that date in that area.
Utilib_14_returned_sum	The number of Utilib 1.4 returned that date in that area.
Slots_freed_sum	The number of recharging slots released that date in that area.
Slots_taken_sum	The number of recharging slots taken that date in that area.

2.3 Data Preparation

2.3.1 Loading and reviewing the data

The data was analyzed using python software. For this, the data was first loaded into the software and the top 5 and bottom five rows were previewed. We then checked the shape of

the data frame which was 16,085 rows by 13 columns. Data types and data information was also reviewed at this stage.

2.3.2 Data cleaning

To prepare the data for analysis, the following steps were taken to clean the data.

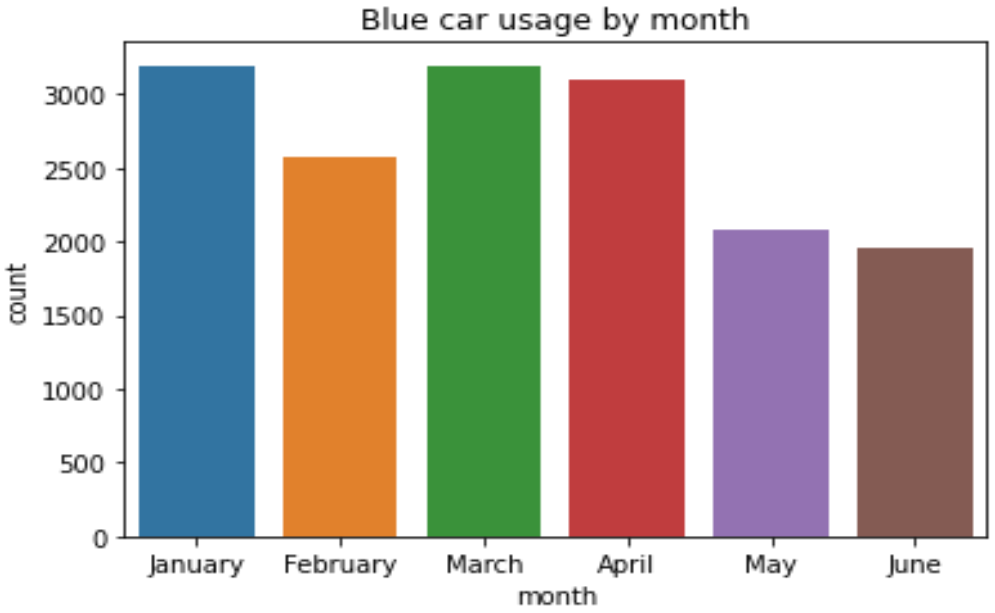
- Check for missing values: There were no missing values in this dataset.
- Replace nan entries with zero: assuming that nan meant no cars or charging slots were returned or freed.
- Check for duplicates: The data set did not have any duplicates.

3. Hypothesis Testing

3.1 Problem Statement

For Autolib Company to grow and continue to run and provide services, the company products need to be adopted more, that is, the market penetration of autolib should be increasing over time.

The analysis sought to investigate if the usage of these cars was increased, decreased, or remained the same. To do this. The analysis looked at two months, January and June. the bivariate analysis showed that more blue cars were taken in January than in June, where Jan had one of the highest records and June had the least.



From this, we created a null hypothesis with the assumption that there is no difference in the meant of the blue cars take in January and in June. The claim was that there is a difference in the means of blue cars taken in the month of January and in the month of June. therefore the null hypothesis and alternative hypothesis are as follows:

Ho: There is no difference in the means of Blue cars take in the month of January and the month of June.

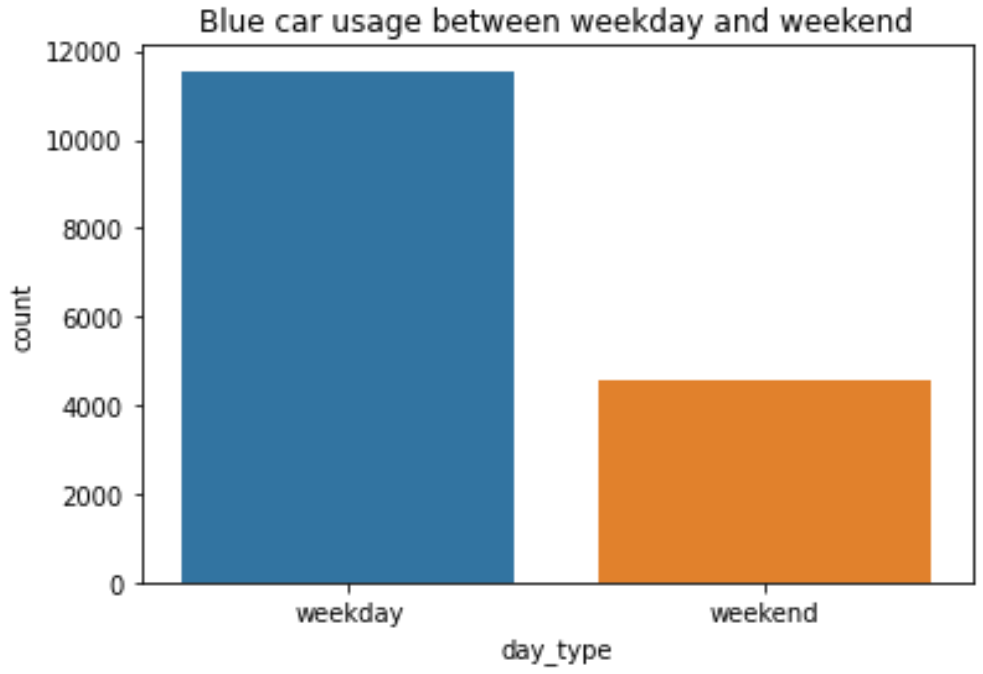
Ha: There is a difference in the means of Blue cars take in the month of January and the month of June.

This was to investigate whether the decline seen in the visualization above was statistically significant. Was it an indicator of declining interest in blue cars?

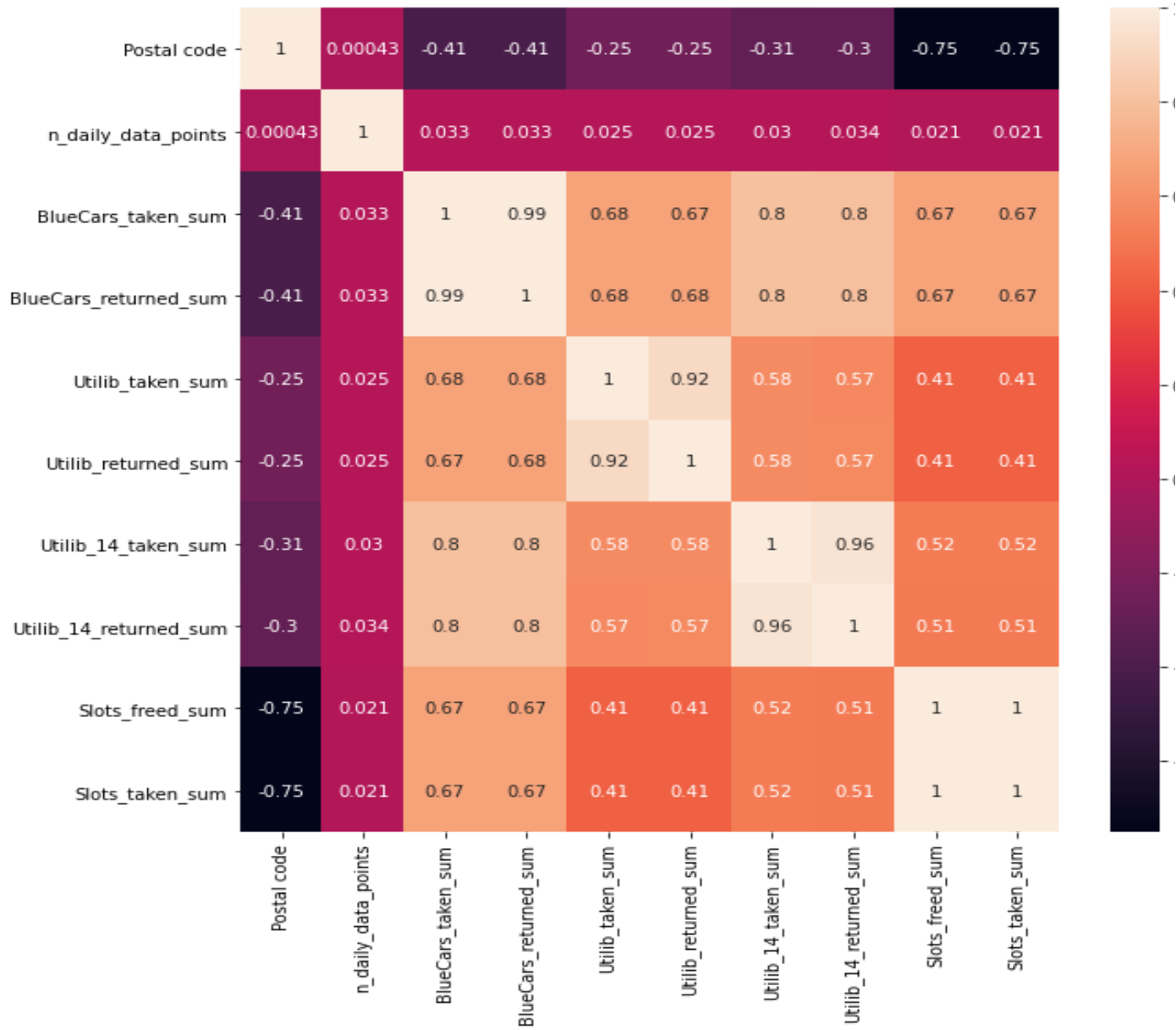
3.1 Data Description

The data set provided information on Autolib's Blue cars taken and returned in the year 2018 between the months of January and June. The data was leptokurtic in terms of kurtosis.

Bivariate analysis showed that more care was taken during the weekdays than on the weekend as shown below.



The correlation between Blue cars take and blue cars returned was a perfect positive correlation. The same was observed for the Utilib, Utiib 14, and charging slot services also provided by the company. All of Autolibs services were positively correlated to each other as seen in the heatmap.



3.3 Hypothesis Testing Procedure

Since the data set was very large, the first this we did before begin the hypothesis testing was obtaining sample data. To ensure that our sample was a proper representation of the entire dataset, we began by grouping the data into postal codes, and then random sampling was used to select random entries from each postal code. Once that was done, we created sample data one for all the entries for the month of January and the other for the month of June. These were the two samples used to test out the hypothesis.

The hypothesis followed the following procedure:

Step 1: Defining the null and alternative hypothesis.

Ho: There is no difference in the means of Blue cars take in the month of January and the month of June.

$H0 : \mu1 = \mu2$

Ha: There is a difference in the means of Blue cars take in the month of January and the month of June.

$Ha: \mu1 \neq \mu2$

This is a two-tailed test as the claim implies that the means of the two samples are not equal.

Since the analysis would be comparing the means of two samples, a two-sample t-test was used to determine whether to reject or not to reject the null hypothesis. Some of the conditions that our sample had to meet in order for us to use the t-test are:

- Data collected follows a continuous scale
- Normally distributed data
- Variables are independent of each other
- Variance is the same

The sample data was found to meet the above conditions and thus we proceeded with the two-sample t-test. The alpha used is 0.05 to allow for a 5% margin of error and a confidence level of 95%.

Step 2: Computing the T-statistics, P-value, and t-critical.

At the 95% confidence level, the tabulated t critical value was |1..833|. This meant that if the calculated t-statistic is greater than the critical t value, then the point falls in the rejection region and we reject the null hypothesis.

3.4 Hypothesis Testing Results

The results of our computations were:

T-statistic : 0.48995

P- value : 0.312083

We do not reject the null hypothesis.

The test statistic was 0.48995 which falls under the rejection do not reject region of the curve.

The p-value was 0.312083 which is greater than the alpha, hence we reject the null hypothesis. The p-value also implies g that the probability of getting the observed result when the null hypothesis is true is 31.208%

The point estimate shows that the average number of blue cars taken for the month of January was approximately |2.78| points different than the number of blue cars taken in the month of June.

3.6 Summary and Conclusions

From this study, the hypothesis testing showed that the usage of blue cars was not significantly different in the two time periods analyzed. This proved helpful as it showed that the decline was seen in the graphical visualization of the monthly analysis, the means in the two-period had no statistically significant differences.

We conducted a sensitivity analysis at the end of our analysis by adjusting the sample size and the p-value was still greater than the alpha. This shows that the test done was not sensitive and held true in an even bigger sample hence proving the results to be valid.

The data used for this study can be found here for the main data set, and here for the data description. The analysis explained in this report was done n a python notebook which can be found here.