Autolib Electric Car Sharing Hypothesis Testing

Business Understanding

Business Overview

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. The program started in 2011 with an initial fleet of 250 eco-friendly electric Blue Cars. The program as at 2015 had more than 3000 cars operating on the streets of Paris and within the whole region. There are around 860 Autolib stations where users can subscribe, pick up or drop off the cars. As well, there are 4400 parking spaces and charging points reserved exclusively for Bluecars. Electric car sharing services greatly contribute to the preservation of the environment and facilitate mobility in the bustling city of Paris. Electric Blue Cars are silent and have zero emissions locally, and their affordability and popularity continues to prevent more and more people from buying their own car.
We, a data science team, appointed by Autolib will understand a dataset given by Autolib and investigate a claim about the blue cars.

Business Objective

The objective of our claim is to intensively explore the given dataset and come up with a research question around the blue cars. From this research question, we come up with a claim about ble cars and find statistical evidence to prove it. Our research question, we are trying to check for equality in the mean of blue cars taken for two postal code areas: 75015 and 75017.

Business Success Criteria

Our success criteria will be determined by the ability to explore the data intensively and come up with a viable conclusion and recommendations out of the results. We are to come up with hypotheses and correctly interpret our results. In our case we are to statistically find evidence to reject the null hypothesis, which is that the mean of blue cars taken in postal code 75015 is statistically equal to the mean of blue cars taken in postal code 75017.

Assessing the situation

- **Assumptions**
  The data provided is accurate.
- **Resources**
  1. **Datasets**

We have been provided with two datasets:

- Autolib Daily Events Dataset [http://bit.ly/DSCoreAutolibDataset]
- Column Explanation [http://bit.ly/DSCoreAutolibDatasetGlossary]

2. **Softwares needed**

   Github
   Google collab
   Google suite

- **Implementation plan**

Phase	**Time-Frame**
Formulation of Research Question	30 minutes
Business Understanding	30 minutes
Data Understanding	30 minutes
Data Preparation and Cleaning	2 hours
Data Analysis	3 hours
Hypotheses formulation and computation	2 hours
Conclusion and recommendation	1 hour
Report writing	2 hours

Data Understanding

Data Mining Goals

The data mining goal for this project was to investigate the blue cars usage in Paris and perform hypothesis tests on a chosen parameter of the dataset.

Data Success Criteria

Successful data understanding would mean that we clean our data, do our analysis then now go ahead and do hypotheses testing.

Data Description

We will be working on two datasets.

1. Glossary dataset (http://bit.ly/DSCoreAutolibDatasetGlossary)

   This first dataset is a glossary, describing the columns used in the other main dataset containing records. It has 2 columns(the column name and the explanation column) and 13 rows.
   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 for aggregation, that day
   dayOfWeek - identifier of weekday (0: Monday -> 6: Sunday)
   day_type - weekday or weekend
   BlueCars_taken_sum - Number of blue cars taken that date in that area
   BlueCars_returned_sum - Number of blue cars 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 recharging slots taken that date in that area

2. Autolib Daily Events Dataset (http://bit.ly/DSCoreAutolibDataset)
   This dataset contains the daily records and has a shape of 13 columns by 16085 rows. The columns ae the above stated columns in the glossary dataset

The random variable that we are investigating is the blue cars taken which is our response variable and the postal code variable. The primary variable is the postal code variable.

Problem Statement

Our research question is to determine the statistical difference between the mean number of blue cars taken on weekdays for 75015 and mean of blue cars taken on weekdays for 75017. Therefore our hypotheses are as below:
Null hypothesis (H0) : Mean of blue cars taken in postal code 75015 is statistically equal to the mean of blue cars taken in postal code 75017
Alternative Hypothesis (H1) : Mean of blue cars taken in postal code 75015 is statistically not equal to the mean of blue cars taken in postal code 75017

We chose this analysis because we are interested in knowing the competitiveness of postal code areas that are close by so as to know whether resources are distributed equally and fairly.

Data Preparation

Loading data

We imported the needed libraries and loaded our datasets on google collab and created dataframes for them.

Exploring the data

We checked the shape of our datasets. Our main data frame contains 13 columns and 16085 rows. We checked the datatypes of each column and a description of the statistical summary of the numeral columns.

Data cleaning

1. Checked for duplicate values and found no duplicate values.
2. Renaming column names by removing any spaces that were in our columns and later changed all the column names to lowercase so as to make the data uniform.
3. Checked for missing values and found that our data frame had no missing values.
4. Checking for outliers - We found quite a number of outliers and decided not to drop them because dropping such a large part of our data would seriously affect the validity of our results. Besides, the outliers wouldn't affect our analysis.

Data Analysis

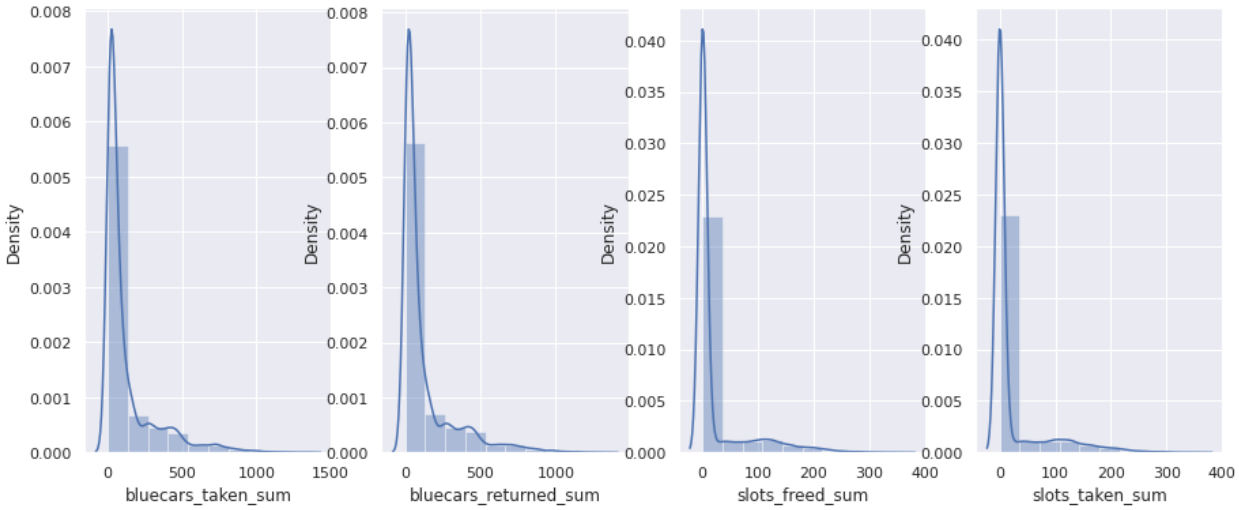
We now went ahead and did our analysis that involved of:

- Univariate Analysis
- Bivariate analysis
- Hypothesis Testing
- Parameter Point estimate
- Confidence Interval Construction

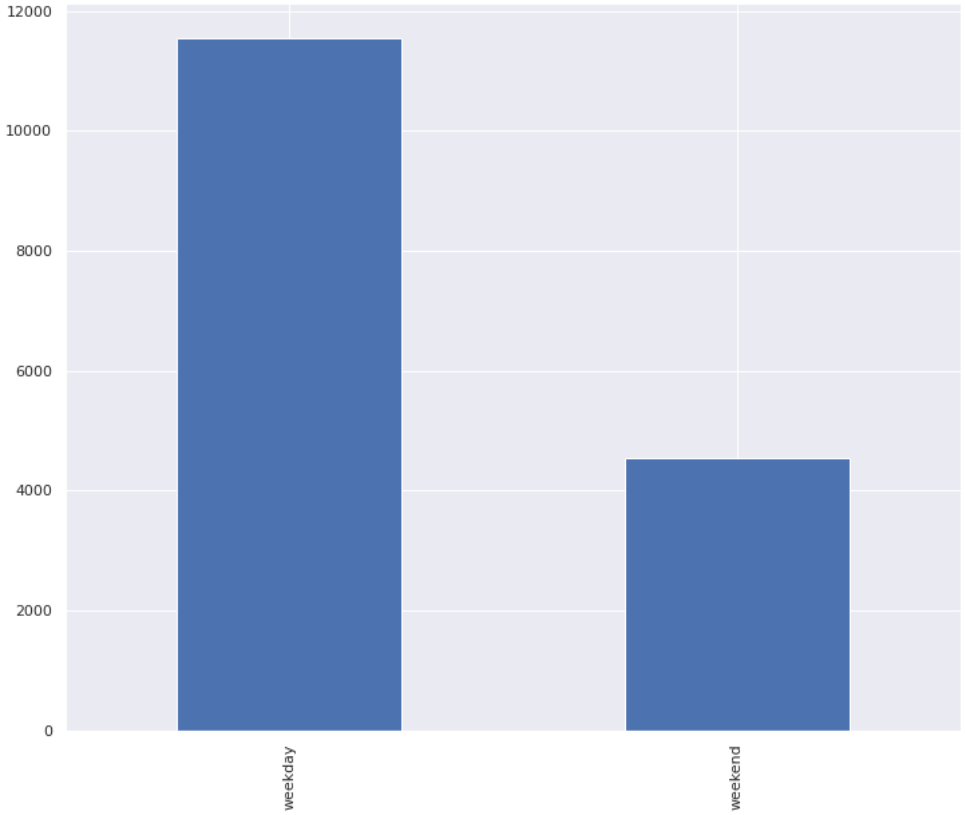
Univariate analysis

We defined our numerical columns and a descriptive statistics of skewness, kurtosis, standard deviation, mean, mode, median and variance of the data. We then did a range definition of the columns we are interested in i.e the blue cars.

We discovered that our data is leptokurtic i.e has an excess positive kurtosis.Also our data has a high positive skewness in all the variables.This was proven by plotting distplots as shown below:



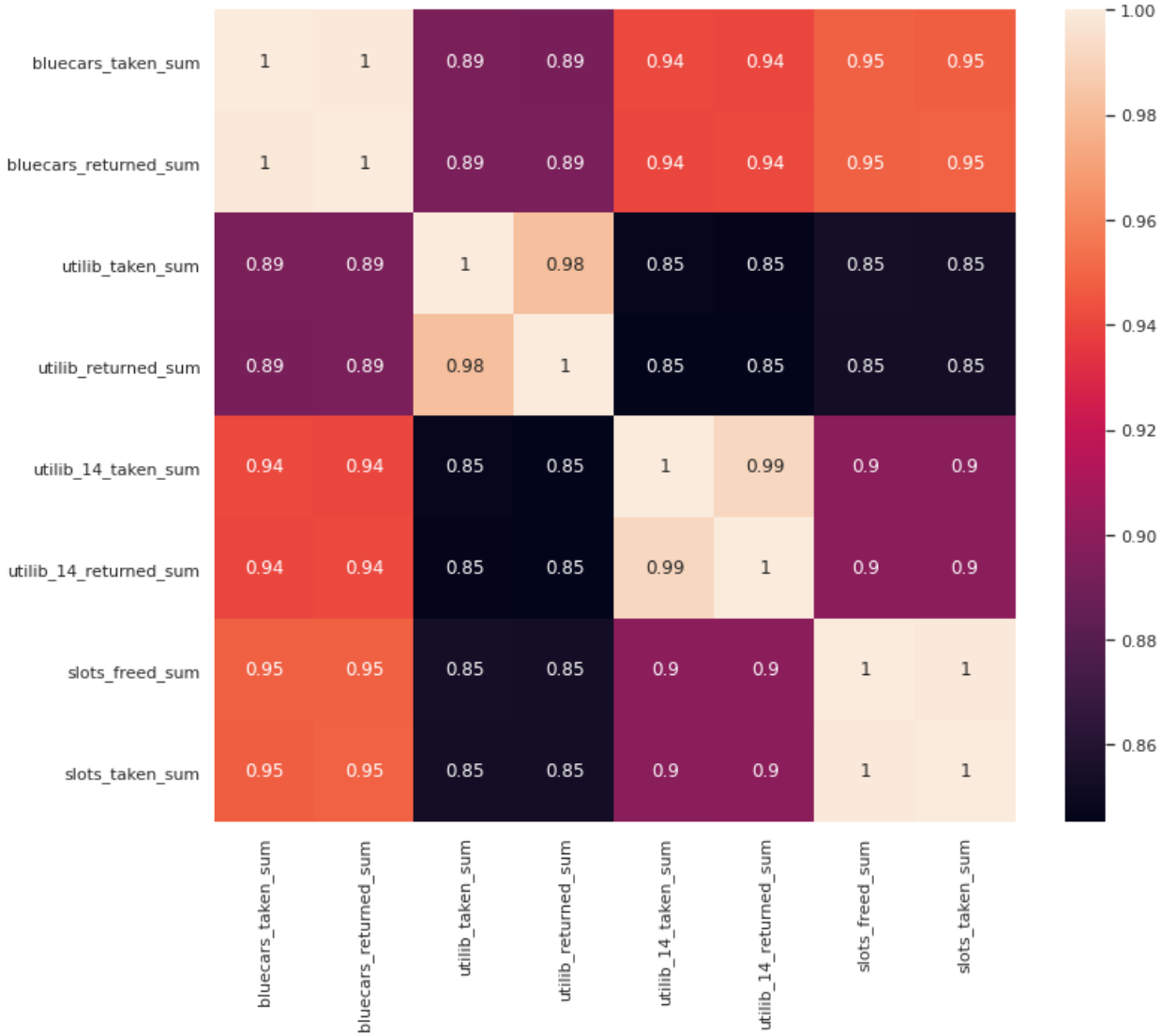
We then plotted histograms for our columns of interest. Figure is too large hence can be found in the notebook attached.
We then grouped by weekdays and weekends and plotted for side to side comparison and noted that weekdays are more busy hence decided to focus our study on weekdays records. The bar plot is as below:



We then plotted boxplots to see our outlier distribution although we did not remove them because they were too many and they seemed viale for the analysis.

Bivariate analysis

We plotted scatter plots and a heat map to confirm the positive correlation between interested columns. The heatmap is as attached:



Hypotheses Testing

Our claim was aimed towards understanding the relationship between two neighbouring postal codes and studying the distribution of resources between two postal codes. We are looking to improve services in Autolib through this by coming up with viable recommendations. We first created a sub data frame of weekdays because we saw most of the action is on weekdays. We then grouped by the two postal codes before finally getting a sample of 0.2 fraction from each

dataframe. We used 0.2 because the population dataframe was not that large, it had 112 rows. Our final sample therefore had 22 rows.
We will then perform our test by the standard procedure:

- Specifying the Null Hypothesis

Mean of blue cars taken in postal code 75015 is statistically equal to the mean of blue cars taken in postal code 75017

- Specifying the Alternative Hypothesis

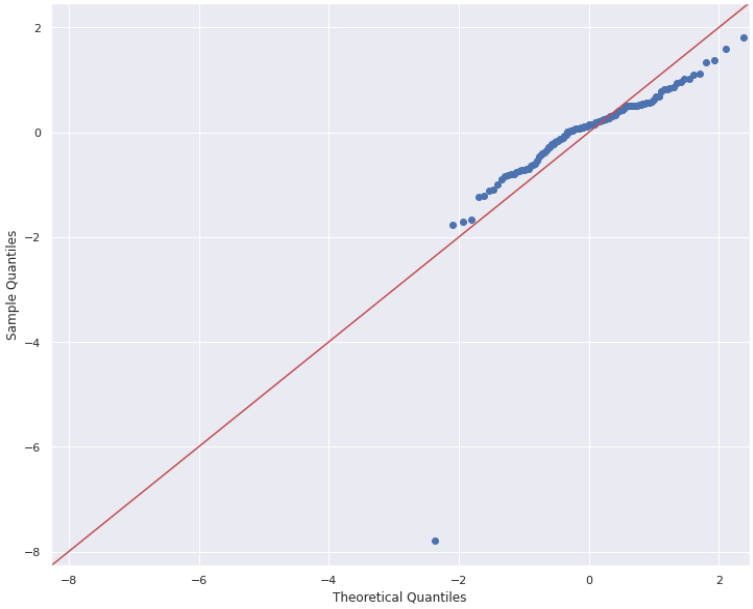
Mean of blue cars taken in postal code 75015 is statistically not equal to the mean of blue cars taken in postal code 75017

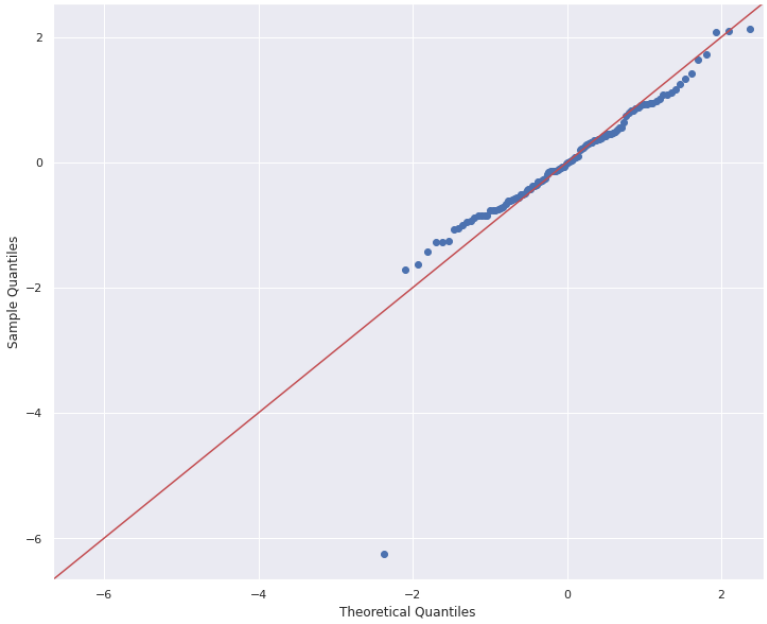
- Setting the Significance Level (a)

We set our significance level to 0.05 the standard level for most tests.

- Calculating the Test Statistic and Corresponding P-Value

We are going to use a t test because our final sample is less than 30 and the data is normally distributed as we confirmed with qq plots below:





We found a p value of 1.3111418291241822e-07 which is less than our set alpha 0.05

- Drawing a Conclusion

After grouping the data by weekdays and postal codes 75015 and 75017, you gauge the difference between the two samples based on blue cars taken. Based on your calculations, the difference between the two groups is statistically significant with a p-value of 1.3111418291241822e-07, well below the defined alpha of 0.05. You conclude that your results support the alternative hypothesis that the Mean of blue cars taken in postal code 75015 is statistically not equal to the mean of blue cars taken in postal code 75017. We thus rejected our null hypotheses.

Point Estimate

Point estimation is the process of finding an approximate value of some parameter of a population from a random sample of the population. The population parameter can be the mean (average) of the population.
The average number of blue cars taken from postal code 75015 is 151.54 points higher than it is for the average number of blue cars taken from postal code 75017.
This was attained by finding the difference between the sample mean of the first postal code 75015 and the sample mean of postal code 75017. This brought us to (847.8636363636364 - 696.3181818181819) hence our answer 151.54.

Confidence Level

Our sample mean of blue cars taken for postal code 75015 is 847.8636363636364
From the confidence level results above, we can be 95% certain that the population mean data is between 804.6888754766074 and 851.0789816662497 for postal code 75015 since it is within this interval.

Our sample mean of blue cars taken for postal code 75017 is 696.3181818181819
From the confidence level results above, we can be 95% certain that the population mean data is between 680.8482501527757 and 712.2946069900814 for postal code 75017 since it is within this interval.

Test Sensitivity

Test sensitivity is the probability of measurement in hypotheses tested that can be measure in four ways:
• true positive (TP): truly alternative features that are called significant.
• false positive (FP): truly null features that are called significant (also called type I error)
• true negative (TN): truly null features that are called insignificant.

• false negative (FN): truly alternative features that are called insignificant (also called type II error).
It is not possible to completely eliminate the probability of a type I error in hypothesis testing. However, there are opportunities to minimize the risks of obtaining results that contain a type I error.

One of the most common approaches to minimizing the probability of getting a false positive error is to minimize the significance level of a hypothesis test. Since the significance level is chosen by a researcher, the level can be changed. For example, the significance level can be minimized to 1% (0.01). This indicates that there is a 1% probability of incorrectly rejecting the null hypothesis.

However, lowering the significance level may lead to a situation wherein the results of the hypothesis test may not capture the true parameter or the true difference of the test
Similar to the type I error, it is not possible to completely eliminate the type II error from a hypothesis test. The only available option is to minimize the probability of committing this type of statistical error. Since a type II error is closely related to the power of a statistical test, the probability of the occurrence of the error can be minimized by increasing the power of the test i.e increasing sample size or increasing significance level.

Conclusion

After performing our test our results rejected the null hypothesis. There is sufficient evidence to conclude that the mean number of blue cars taken from postal code 75015 is statistically significantly different from the mean number of blue cars taken in postal code 75017. The two postal codes seem to be in the same area as they both start with 75'.

Recommendation

We can therefore see that there is likely a competitive nature of blue cars even in neighbouring areas.
We can also recommend treating each postal area individually as evidence has it that customer engagement in even close postal codes is different.