Auto Lib Data Report

Moringa Core Week 4 Project

Summary

We are working as a Data Scientist for the 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 for 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 can be 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.

Problem Statement

As a data scientist assigned to this project it's my job to figure out how many blueCar vehicles are being used in Paris on a daily basis, using the Autolib dataset.

Ho Null Hypothesis : The number of blueCar vehicles used per day are greater than 200 (P > 200)

Ha Alternate Hypothesis : The number of blueCar vehicles used per day are less than 200 (P < 200)

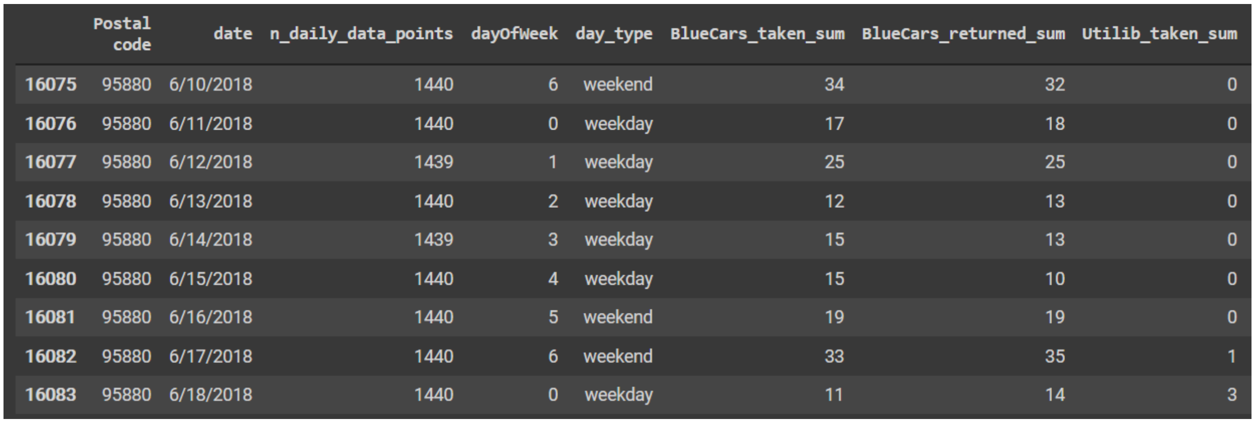
Data Description

The data used in this project is from Autolib dataset provided by the client, along with the data file, there is a data description attached named columns_descripton:

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 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 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 recharging slots taken that date in that area

Here is how some of the data looks imported into a notebook:



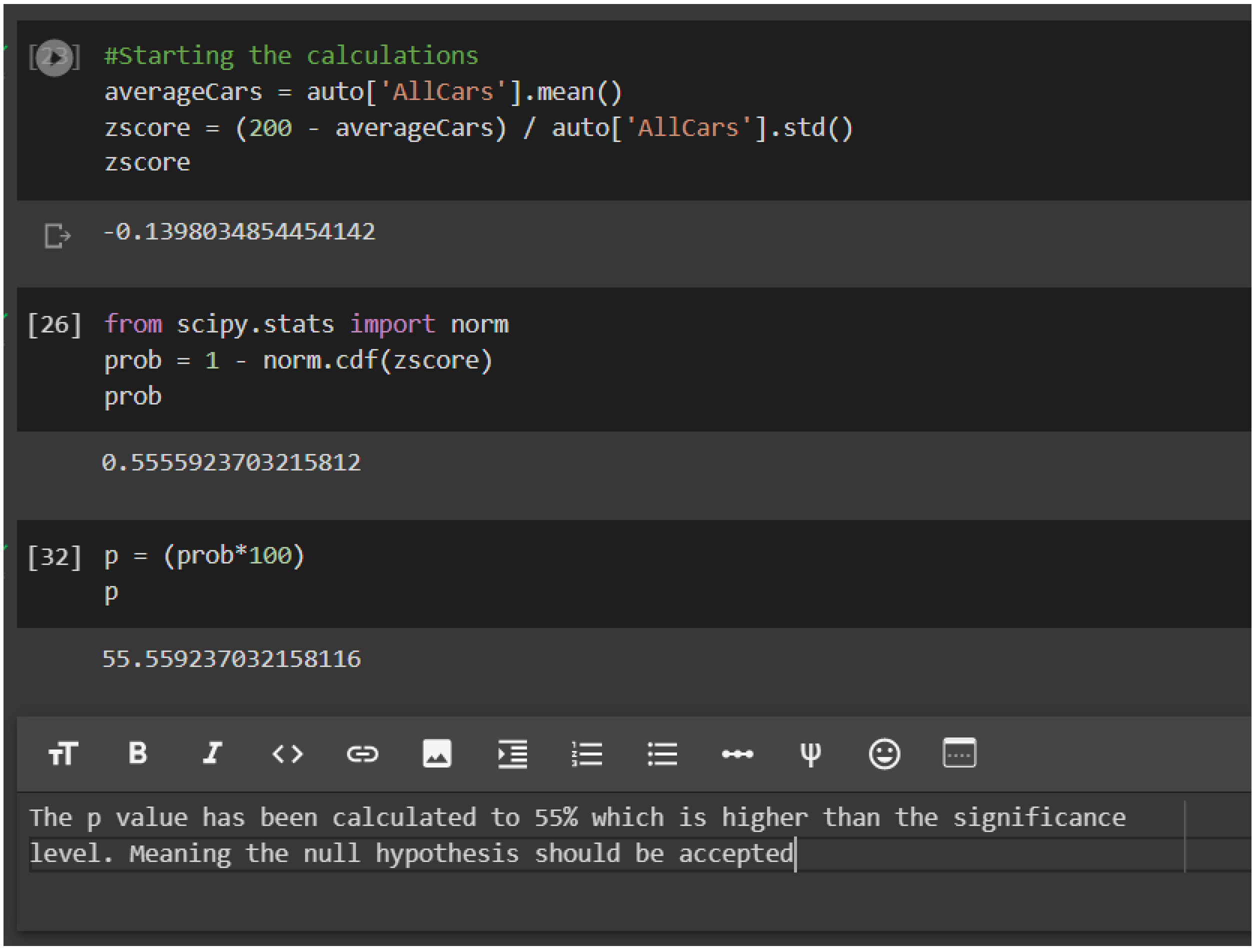
	Postal code	date	n_daily_data_points	dayOfWeek	day_type	BlueCars_taken_sum	BlueCars_returned_sum	Utilib_taken_sum
16075	95880	6/10/2018	1440	6	weekend	34	32	0
16076	95880	6/11/2018	1440	0	weekday	17	18	0
16077	95880	6/12/2018	1439	1	weekday	25	25	0
16078	95880	6/13/2018	1440	2	weekday	12	13	0
16079	95880	6/14/2018	1439	3	weekday	15	13	0
16080	95880	6/15/2018	1440	4	weekday	15	10	0
16081	95880	6/16/2018	1440	5	weekend	19	19	0
16082	95880	6/17/2018	1440	6	weekend	33	35	1
16083	95880	6/18/2018	1440	0	weekday	11	14	3

Hypothesis Testing Procedure

The testing procedure to check if my hypothesis is correct is as follows:

1. Import and view the data
2. Perform any cleaning required i.e removing duplicates, null values
3. Perform univariate analysis
4. Perform bivariate analysis
5. Perform the hypothesis test
    a. This is a one sided left tailed test, **α** = 0.05
6. Compute the p value
7. Publish the results

Hypothesis Testing Result



```
[23] #Starting the calculations
     averageCars = auto['AllCars'].mean()
     zscore = (200 - averageCars) / auto['AllCars'].std()
     zscore

     -0.1398034854454142

[26] from scipy.stats import norm
     prob = 1 - norm.cdf(zscore)
     prob

     0.5555923703215812

[32] p = (prob*100)
     p

     55.559237032158116
```

The p value has been calculated to 55% which is higher than the significance level. Meaning the null hypothesis should be accepted

The z score of the test came out to -0.1398 and the probability of the number of blue cars used adding up to more than 200 per day is at 55% which is higher than the significance level of 5%. This means that the null hypothesis is true

Discussion of Test Sensitivity

The test sensitivity is a measure of the ability of a screening test to detect a true positive, which avoids the type 1 and 2 errors possible in probability testing

In the case of this test the sample data from 2018 might not be indicative of all the days recorded and the number of blue vehicles used per day will not be above 200. This could be due to a variety of reasons i.e. human error in recording, it was a period of high activity of the blue cars.

Summary & Conclusion

The tests prove that on average the number of blue cars used in Paris in 2018 were more than 200 cars per day. The probability states that if you select a random day from the recorded days the chances there were more than 200 vehicles is at 55%, which means that majority of the days are >200