# **Autolib Car Sharing Data Report**

**Name:** Kennedy Njoroge

## Table of Contents

[**Autolib Car Sharing Data Report**](#_f2wksdcjyr9d) **1**

[**Revision History**](#_x1brzpuimvyb) **1**

[**Glossary**](#_ezr78er7j87h) **3**

[**Problem Statement**](#_i9i90vv0zgb0) **3**

[**Business Understanding**](#_b5zrv8hdzuiu) **4**

[**Data Understanding**](#_4g4p47kgklcn) **8**

[**Data Preparation**](#_i4aehfj1enia) **12**

[**Data Analysis**](#_299s5j8jccgy) **12**

[**Evaluation**](#_hag80y7dtz6) **12**

[**Recommendations**](#_ftmzwedf55yc) **12**

[**References**](#_uuay7rn2ljs6) **13**

1. **Project Links**

|  |  |  |
| --- | --- | --- |
| **Item** | **Description** | **URL** |
| Github Project URL |  | <https://github.com/jupitershot/Autolib-Electric-Car-Sharing> |

## Problem Statement

### **Problem Statement**

Autolib data set has been provided showing electric cars hired and returned per day and per postal codes for blue cars, utilib cars and utilib 1.4.

Main requirement is to identify the cars hired average numbers between different postal codes so as to analyze business performance and make region specific corrective measures where need be.

In France,postal codes are used to identify the location of a place. First two digits represent the district, and the last three are for subdivisions (town or neighborhood). The dataset provided has below distinct postal codes

[75001, 75002, 75003, 75004, 75005, 75006, 75007, 75008, 75009,75010, 75011, 75012, 75013, 75014, 75015, 75016, 75017, 75018,75019, 75020, 75112, 75116, 78000, 78140, 78150, 91330, 91370,91400, 92000, 92100, 92110, 92120, 92130, 92140, 92150, 92160,92170, 92190, 92200, 92210, 92220, 92230, 92240, 92250, 92260,92270, 92290, 92300, 92310, 92320, 92330, 92340, 92350, 92360,92370, 92380, 92390, 92400, 92410, 92420, 92500, 92600, 92700,92800, 93100, 93110, 93130, 93150, 93170, 93200, 93230, 93260,93300, 93310, 93350, 93360, 93370, 93390, 93400, 93440, 93500,93600, 93700, 93800, 94000, 94100, 94110, 94120, 94130, 94140, 94150, 94160, 94220, 94230, 94300, 94340, 94410, 94450, 94500,94700, 94800, 95100, 95870, 95880]

As per <https://en.wikipedia.org/wiki/Departments_of_France>, the postal codes above below to below administrative divisions also known as departments in france

|  |  |  |
| --- | --- | --- |
| # | Postal Code | Administrative Division |
| 1 | 75XXX | Paris |
| 2 | 78XXX | Yvelines |
| 3 | 91XXX | Essonne |
| 4 | 92XXX | Hauts-de-Seine |
| 5 | 93XXX | Seine-Saint-Denis |
| 6 | 94XXX | Val-de-Marne |
| 7 | 95XXX | Val-d'Oise |

### **Hypothesis**

Research question to answer is : Are the cars returned in Paris same as Seine Saint Denis?

**Null Hypothesis** : Average number of cars returned in Paris over weekends is same as those returned in Seine Saint Denis.

**Alternative Hypothesis**: Average number of cars returned in Paris over weekends is not the same as those returned in Seine Saint Denis.

Ho : μ of Paris = μ of Seine Saint Denis

Ha : μ of Paris ≠ μ of Seine Saint Denis

**Why Choose Paris verus Seine Saint Denis?**

75XXX (Paris) is the central business district which is financial and tourism center of paris hence the demographics and expendable income of residents and visitors is higher.

93XXX (Seine Saint Denis) is located in the suburbs of Paris. 21% of the population is composed of emigrants where over 50% of the emigrants are from Africa. In 2018, the poverty rate of Seine Saint Denis was twice the national average at 28%, the unemployment rate was 3 percentage above the national average and 4 percentage points above the national average at 12.7%. Source : <https://en.wikipedia.org/wiki/Seine-Saint-Denis>

From above, the two regions have contrasting demographics and have differing economic growth hence the interest to compare performance.

### **Autolib Background**

Autolib’ was an electric car sharing company owned by Bollore industrial group that operated in France in multiple cities including Paris between 2011 and 31st July 2018.

Autolib owned cars which were which could be leased by public and returned after use. Since they were electric cars, there was charging stations located at different locations where customers could pick and drop back the cars after use.

The model of electric cars used by Autolib are:

* Blue car - A passenger electric car.

More background details about this car are available on this [link](https://en.wikipedia.org/wiki/Bollor%C3%A9_Bluecar)

* Utilib car - Is a red blue car. Apart from passenger model, it has a delivery variant has two seats to create more space for cargo. It’s red in colour.
* Utilib 1.4



Figure 1: An Autolib blue car at a charging station. Source: France 24



Figure 2: An Autolib utilib car cargo variant with no backseats. Source: [https://insideevs.com](https://insideevs.com/)

**Project Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Time** | **Resources** | **Risks** |
| Business Understanding | 1 hour | Data Analyst/Scientist |  |
| Data understanding  and preparation | 3 hours | Data Analyst/Scientist | Insufficient data set |
| Data Preparation | 1 hour | Data Analyst/Scientist |  |
| Exploratory Data Analysis | 3 hours | Data Analyst/Scientist |  |
| Sampling | 1 hour | Data Analyst/Scientist |  |
| Hypothesis testing | 1 hour | Data Analyst/Scientist |  |
| Total | 10 hours | Data Analyst/Scientist |  |

**Success Criteria**

The research will be successful when below are accomplished.

* Data is successfully explored
* Null and Hypothesis tests are well documented
* Sampling done is representative of the sample
* Hypothesis testing is done

**Resources**

* Datasets
  + Dataset extracted from http://bit.ly/DSCoreAutolibDataset

Data was originally extracted from opendataparis.com by Dalberg Data Insights.

* + Data description provided here http://bit.ly/DSCoreAutolibDatasetGlossary
* Tools used
  + Jupyter Notebook
  + Github
  + Moringa school canvas access for data and problem statement access

**Assumptions**

* Data provided is sufficient to fulfill the research objectives

**Constraints**

* Data is summarised per day hence analysis by hour of the day not possible.

**Cost**

Analysis time.

## Data Understanding

**Data Structure and Exploration**

Below is the definition of the provided Autolib Data Definition. This is determined after data exploration on Jupyter notebook and data description file review.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Column** | **Number of Records** | **Data Type** | **Sample values** | **Meaning of Column** |
| 1 | Postal code | 16085 | Integer | 75001, 75002, 75003,75004, 75005, 75006 | postal code of the area (in Paris) |
| 2 | date | 16085 | String | 2018-03-29, 2018-03-09, 2018-05-08, 2018-01-23, 2018-05-21 | Number of cars available at the station. Redundant column as value is same as Blue Car counter. |
| 5 | n\_daily\_data\_points | 16085 | Integer | 1438, 1440, 1439, 1439, 1439 | Number of daily data points that were available for aggregation, that day |
| 6 | dayOfWeek | 16085 | Integer | 3, 3, 0, 2, 2, 2, 4 | identifier of weekday (0: Monday -> 6: Sunday) |
| 7 | day\_type | 16085 | String | ['weekend,' 'weekday'] | weekday or weekend |
| 8 | BlueCars\_taken\_sum | 16085 | Integer | [0 1 2] | Number of bluecars taken that date in that area |
| 9 | BlueCars\_returned\_sum | 16085 | Integer |  | Number of bluecars returned that date in that area |
| 10 | Utilib\_taken\_sum | 16085 | Integer |  | Number of Utilib taken that date in that area |
| 11 | Utilib\_returned\_sum | 16085 | Integer |  | Number of Utilib returned that date in that area |
| 12 | Utilib\_14\_taken\_sum | 16085 | Integer |  | ID of the station |
| 13 | Utilib\_14\_returned\_sum | 16085 | Integer |  | Number of Utilib 1.4 taken that date in that area |
| 14 | Slots\_freed\_sum | 16085 | Integer |  | Number of recharging slots released that date in that area |
| 15 | Slots\_taken\_sum | 16085 | Integer |  | Number of recharging slots taken that date in that area |

**Correlation**

* Correlation analysis done using correlation function.

**Distribution**

* Univariate distribution checked using histograms
* Outliers - from initial observations, the numeric data is not heavily dispersed.

Box plots used to check presence of outliers.

**Observations**

* **Correlation:** There is a direct correlation between ‘Cars’ and ‘Blue Cars Counter’ column is noted with a value of 1. ‘Cars’ column has similar values as ‘blue car counter’ column hence the correlation.
* 1105 distinct stations noted. One station has no address.
* **Data Collection period:** Data collected across from 1st to 9th April 2019.
* **Data quality issues:** There are two columns with missing values i.e ‘displayed comments’ - 4,899 and ‘Scheduled at’ - 4,965.
* Multiple columns having status of stations i.e status, subscription status and rental status depending on context.
* Kind and station type columns define the attributes of a station.
* No outliers noted.

**Observation Recommendations**

* Make cars column to have totals of blue car, utilib and utilib 1.4 cars to remove current redundancy with blue car.
* Derive timestamp from year, month, day, hour, minute column.
* Data set does not have number of cars hired or returned. This will have to be derived.
* ‘Displayed comments’ and ‘Scheduled at’ columns have significant null values. Scheduled values is validly null since the scheduled dates are only provided for stations with future status. This data set is relevant when exploring future expansion hence will be maintained.

Displayed comments column is irrelevant when answering research questions hence is recommended as a candidate for dropping.

* Scheduled at is a string with timestamp values. Field data type should be converted to timestamp.

## Data Preparation

1. Data Columns Selection/Reduction

* By visual data exploration, the columns that are not of interest to answering of the hypotheses are identified and dropped. The selected columns are:

['Postal code' ,'date', 'dayOfWeek', 'day\_type', 'BlueCars\_taken\_sum', 'BlueCars\_returned\_sum']

1. Data Cleaning

* No null values exist hence no cleaning action.

1. Constructing data set/Feature Engineering

* No new column created.

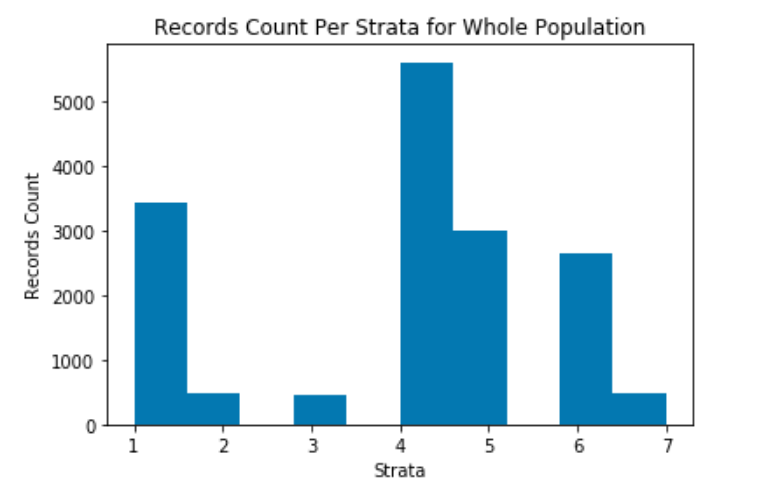
1. Outliers dropping

‘BlueCars\_taken\_sum’ and 'BlueCars\_returned\_sum' have outliers. The outliers are observed to be valid data. Therebeing, they were not in the first iteration. However, post z score calculation a z statistic of 13 was obtained which is due to data not being normally distributed due to outliers. Therebeing, the outliers were dropped using IQR approach.

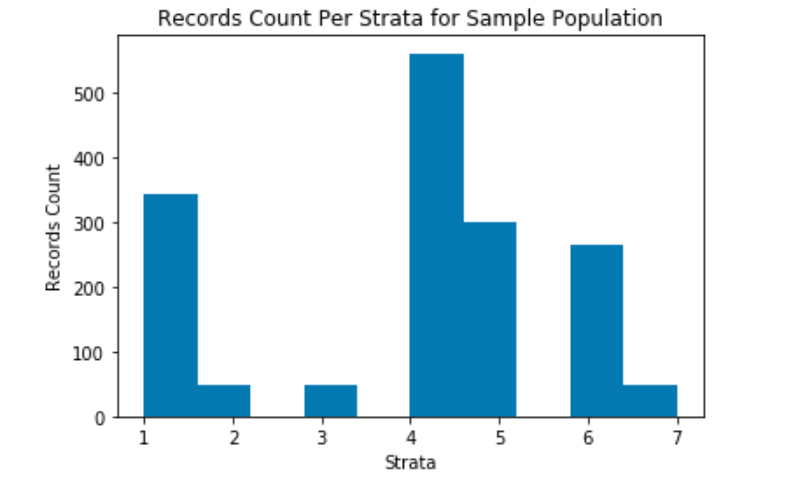
## Sampling

10% Sample size was taken for records in region 1 and 5.

**Distribution Before Sampling**



**Distribution after sampling**



## Hypothesis Testing Procedure

* Sampling dataset was obtained
* Hypothesis test was laid down i.e
* Significant level of 0.05 % was adopted
* Z test was adopted as the records count is greater than 30.
* The z-score and p-value was calculated using the mean and standard deviation of the two samples.
* The p\_value was compared against the significance level

## Hypothesis Testing Results

Outcome: The Null that mean of paris is same during weekends as mean of Seine-Saint-Denis is rejected with a P-Value of 4.

The average daily returned cars in Paris is 465 against that of Seine-Saint-Denisof 41.

Notably, this is after dropping of outliers. The outliers were dropped as data was not normally distributed which causes failure in usage of z test. It's proposed a further test mechanism is explored which includes the outliers. Dropping of may result to Type 1 error.

## Summary and Conclusions

* Different operational and marketing strategies should be adopted across the two administrative regions of Paris and Seine-Saint-Denis as their performance is significantly different.
* Further tests for weekday performance should be evaluated
* A different test approach which caters for non-normal distribution should be adopted.

## References

* Autolib Company Details, <https://en.wikipedia.org/wiki/Autolib%27>
* Blue Car at a charging station <https://www.france24.com/en/20141208-paris-electric-car-infrastructure-go-nationwide>
* Utilib car, <https://insideevs.com/news/324551/bollore-expands-autolib-by-adding-utilib-variant/>
* Seine Saint Denis, <https://en.wikipedia.org/wiki/Seine-Saint-Denis>
* Departments of France, <https://en.wikipedia.org/wiki/Departments_of_France>