# **CRISP - DM ELECTRIC CAR SHARING PROJECT**

The following assessment will follow the CRISP-DM method to help solve the problem. It will display the Business Understanding, Data Understanding, Data Preparation, and Evaluation stages. The Recommendation section will be our conclusion on the results, process and thoughts process, highlighting the insights gained from conducting the project.

The code can be found on:[Electric Car Sharing Service](https://github.com/AlexTwenji777/Electric-Car-Sharing-IP)

## **BUSINESS UNDERSTANDING**

### 1.1. UNDERSTANDING THE PROBLEM

The role assumed here is that of a Data Scientist working for an electric car-sharing service company. The task is to process data from the different stations from which the cars are dispatched in order to understand usage of the electric cars over time. This will be achieved by answering the research question: *‘Identify the most popular hour of the day for picking up a shared electric car (Bluecar) in the city of Paris over the month of April 2018.’*

To achieve this, the Data Report needs to have an objective account, with insights majorly coming from the dataset. External information can only be used for supporting information.

### 1.2. DATA MINING GOAL

The goal of this project can be found in the research question, which is, to Identify the most popular hour of the day for picking up a shared electric car (Bluecar) in the city of Paris over the month of April 2018.

To further understand the data, the following questions need to be answered as well:

* The most popular day with the most rides.
* The least popular day with the least rides.
* The most popular hours (working or home hours).
* The least popular hours (working or home hours).
* The most popular hours (working or home hours) for returning cars.
* The most popular station.
* The most popular stations determined by the most popular hours to visit them.
* The most popular hour.
* The least popular hour.
* Most Popular Post Code for picking up Blue Cars.
* Does the above Post Code belong to the most popular station overall?

### 1.3 ANALYSIS

#### **1.3.1 Data**

Data was extracted from opendataparis.com, where the Autolib availability information was

available in real-time. The accessed database was the following:

Name: Stations Autolib: Disponibilité en temps réel

Producer : Autolib

License : Open Database License

Any further publication should include this mention and respect the terms of the Open Database

License.

The sample holds data from April 1 to April 9, 2018, which was obtained from Dalberg Data Insights. The data collected had different stages, however, the time period we are considering had data collected every minute from October 30, 16:59 PM to July 31, 2018, 23:59 PM. Some gaps exist in the data largely due to the downloading process failing. This however, should not prove to be a big problem since missing files were rare, except for a major gap from November 11 to November 14, 2017 (of which is not in the scope of our data anyway).

#### **1.3.2 Risks**

Financial Risks are negligible since the project will be carried out on open source software services such as Google Colab, GitHub, Jupyter Notebook and Atom. Time is the only constraint since this needs to be done in one day.

#### **1.3.3 Limitations**

One of the limiting factors is that the insights should come from the data. This is not a major limitation as external sources can be used to support the analysis performed from the data. Another limitation is that the data only encompasses data for 9 days in just 1 month. Due to this being a school research project, this should not be a problem as the purpose is to understand how to work with large data and datasets. Finally, the dataset only considers one month, but this will not affect the process of understanding how to manipulate large datasets for future projects.

#### **1.3.4 Assumptions**

The major assumption concerns the data only being 9-days worth of data. We will assume that this data is representative of a longer time period. There is no way to mitigate this as it is for the purpose of a School Project with a goal of understanding how to work with large Datasets on Python using Pandas and other related Libraries.

## **DATA UNDERSTANDING**

### 2.1 DATA COLLECTION

The dataset used can be found on the following link: [Autolib\_dataset(2)](http://bit.ly/autolib_dataset).

Documentation explaining the above dataset can be found on the following link: [Autolib\_description](https://drive.google.com/file/d/13DXF2CFWQLeYxxHFekng8HJnH_jtbfpN/view)

### 2.2 DATA DESCRIPTION

The columns found in the dataset have been described, in order of their appearance, below:

Table1: Table of Column Keys

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Address | Station’s physical address. |
| Cars | Number of cars available at the station - similar to Bluecars counter column. |
| Bluecar counter | Number of Bluecars available at the station. |
| Utilib counter | Number of Utilibs available at the station. |
| Utilib 1.4 counter | Number of Utilib 1.4 available at the station. |
| Charge Slots | Number of Charging slots available at the station. Can only be greater than 0 if charging status is ‘operational’. |
| Charging Status | If the station is operational for recharging. Slots and cars can be available in all scenarios except stations designated as future. |
| City | Station’s City. |
| Displayed comment | Comments about extra facilities and infrastructure in and around the station. |
| ID | Station’s ID. |
| Kind | ‘CENTER’ = no resources, ‘PARKING’ = no charge slots, but can have Bluecars & Utilib, ‘STATION’ & ‘SPACE’ = can have all resources. |
| Geo point | Station’s GPS coordinates. |
| Postal code | Station’s Postcode. |
| Public name | Station’s Name. |
| Rental Status | If Station is available for renting cars. Resources only available when ‘operational’. ‘Broken’ can have Slots but no other resource. (Bluecars, Utilib or charging Slots) |
| Scheduled at | Planned opening date. Only those with ‘future’ status do NOT have null values. |
| Slots | Number of parking slots available in the station. |
| Station type | No resources available for ‘sub\_center’ - only 1 location. read*[1]* |
| Status | ‘Closed’ can have resources. ‘Scheduled’ has no resources, which is actually the status if there is a Scheduled at date. |

|  |  |
| --- | --- |
| Subscription status | If possible to get an Autolib subscription at the station. ‘Future’ has no resources but others can have resources. |
| Year | Year in numeric |
| Month | Month in numeric |
| Day | Days in numeric |
| Hour | Hours in numeric |
| minute | Minutes in numeric |

[1] Was this a selling point for Autolib subscriptions.

Furthermore, the known value types of the columns are in Table 2 below.

Table 2: Table of column types and known values.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Values** |
| Address | String |  |
| Cars | Number | 0 - 7 |
| Bluecar counter | Number | 0 - 7 |
| Utilib counter | Number | 0 - 4 |
| Utilib 1.4 counter | Number | 0 - 5 |
| Charge Slots | Number | 0 - 3 |
| Charging Status | String | ‘Nonexistent’, ‘operational’, ‘broken’, ‘future’, some with typing errors. |
| City | String |  |
| Displayed comment | String |  |
| ID | String |  |
| Kind | String | ‘STATION’, ‘SPACE’, ‘PARKING’, ‘CENTER’. |
| Geo point | String |  |
| Postal code | Number |  |
| Public name | String |  |
| Rental Status | String | ‘Nonexistent’, ‘operational’, ‘broken’, ‘future’, some are empty. |
| Scheduled at | String | datetime. |
| Slots | Number | 0 - 7 |
| Station type | String | ‘Station’, ‘full\_station’, ‘sub\_center’. |
| Status | String | ‘Ok’, ‘closed’, ‘scheduled’. |

|  |  |  |
| --- | --- | --- |
| Subscription status | String | ‘Nonexistent’, ‘operational’, ‘broken’, ‘future’. |

## **3. DATA PREPARATION**

### 3.1 SELECTING DATA

For this analysis, the Utilib counter and Uilib 1.4 counter columns will be ignored. We will only work with the Bluecar counter column, in reference to the car brand as this is our primary service company to deal with.

The Cars column can be dropped because it’s values are similar to those in the Bluecar counter column. The Displayed comment column can also be removed as it serves no purpose in the kind of analysis being conducted here. Geo point column can also be dropped. Kind column can be dropped in the initial analysis as we only want to see the overall peak hour of using the car service. It can later be reintroduced to rank the ‘Kind’ with the most engagement with customers. For the same reason, Station type can be dropped.

Postal Code, Public name and ID seem to be similar in that Postal Code seems to be the numeric address. Investigate this with pandas then we can use one of the other. Subscription status can be dropped for the initial analysis, then later reintroduced to determine the most popular location for subscription.

### 3.2 DATA CLEANING

1. Cars and Blue Car Counter hold the same information.
2. Each ID corresponds to a specific Public Name. They both correspond to a specific Address except 1 outlier. Postal code has much less unique values than the 3.
3. We’ll drop ID and Public Name regardless, as we want to know the Addresses and Postal Codes.

## **4. DATA ANALYSIS**

Order of precedence -

1. Charging Status If station is operational for recharging. ‘Nonexistent’, ‘operational’, ‘broken’, ‘future’. Slots and cars can be available in all scenarios except stations designated as future. It was investigated and those marked as ‘future’ were not in the data.
2. Charge Slots. 0 - 3. Number of Charging slots available at the station. Can only be greater than 0 if charging status is ‘operational’. There were no outliers.
3. Status. ‘Ok’, ‘closed’, ‘scheduled’. ‘Closed’ can have resources. ‘Scheduled’ has no resources, which is actually the status if there is a Scheduled at date. There were no outliers.
4. Rental Status. ‘Nonexistent’, ‘operational’, ‘broken’, ‘future’, some are empty. If Station is available for renting cars. Resources only available when ‘operational’. ‘Broken’ can have Slots but no other resource. (Bluecars, Utilib or charging Slots). There were no outliers.
5. Scheduled at. Datetime. Planned opening date. Only those with ‘future’ status do NOT have null values. There were no outliers.
6. Station type. ‘Station’, ‘full\_station’, ‘sub\_center’. No resources available for ‘sub\_center’ - only 1 location. Read*[1]*. There were no outliers
7. Slots. 0 - 7. Number of parking slots available in the station.
8. Bluecar counter. 0 - 7. Number of Bluecars available at the station.
9. City - we need Paris only.
10. ID - will help in linking a station in different time frames to determine if cars were returned. Postal Code, Public name and ID seem to be similar in that PC seems to be the numeric address. Investigate this with pandas then we can use one of the 3.

### 4.1 PER DAY ANALYSIS

The following questions should be answered by the model:

1. The most popular day with the most rides including day of the week.
2. The least popular day with the least rides including day of the week.

### 4.2 HOURS ANALYSIS

The following questions should be answered by the model:

1. The most popular hours (working or home hours)
2. The least popular hours (working or home hours)
3. The most popular hour.
4. The least popular hour.
5. The most popular hours (working or home hours) for returning cars.
6. The least popular hours (working or home hours) for returning cars.

### 4.3 STATION ANALYSIS

The following questions should be answered by the model:

1. The most popular station.
2. The most popular stations determined the most popular hours to visit them.

### 4.4 POSTCODE ANALYSIS

The following questions should be answered by the model:

1. Most Popular Post Code for picking up Blue Cars.
2. Does the above Post Code belong to the most popular station overall?

## **5. EVALUATION**

### 5.1 PER DAY ANALYSIS

1. The most popular day with the most rides was 4th April 2018, which was a Wednesday. The other days in the top 3 are 5th and 6th April 2018, which were Thursday and Friday respectively.
2. The least popular day with the least rides was 1st April 2018, which was a Sunday. The second least popular day was 9th April, which was a Monday. The third least popular was 8th April, which was also a Sunday.

From the above we can see a pattern emerging where Sunday seems to be the least popular day of picking up cars.

### 5.2 HOURS ANALYSIS

1. The most popular working hours, which in Paris is from 8am to 6pm, Monday to Friday, were: 12pm (with 64 rides), 3pm (with 57 rides) and 4pm (with 51 rides). As for home hours during the weekday, the most popular hours were: 3am (with 68 rides), 9pm (with 52 rides) and 1am (with 51 rides) .
2. The least popular working hours were: 2pm (with 35 rides), 1pm (with 41 rides) and 9am (with 44 rides). As for home hours during the weekday, the least popular hours were 11pm (with 31 rides), 8pm(with 37 rides) and 10pm (with 41 rides).

From the above, the most popular hour for working days is 3am which is considered as home time. The least popular hour is 11pm which is also a home time. Overall, working hours accounted for 576 rides. Home hours in general experienced the most traffic, with 658 rides. The weekdays had more rides than the weekend.

Considering the weekends only;

1. The most popular day hours (between 8am and 6pm) were: 12pm (with 14 rides), 10pm and 8pm (both with 13 rides each). As for night hours, the most popular hours were: 3am (16 rides), 10pm and 6am (both with 14 rides each).
2. The least popular day hours were: 5pm and 6pm (both with 5 rides each). 9am, 11am and 1pm all had 7 rides each, which was the next lowest value. As for night hours, midnight, 1 am and 4 am had the lowest rides. Each of them had 5 rides.

The weekend also has its highest rides at 3am with 16 rides, but the least rides are spread out between different time frames as indicated in the point 2 above. The night rides were also more during the weekend (118), as compared to the day rides (98). In total, with both the weekday and weekend combined, the total day rides were 576 and the total night rides were 658. The least popular hours during the whole week were: 11pm (31 rides), 2pm (35 rides) and 8pm (37 rides). The most popular hours were: 3am (84 rides), 12pm (78 rides) and 3pm (66 rides).

As for returns:

1. The most popular working hours on the weekday were: 10am (with 44 returns), 3pm (with 43 returns) and 2pm (with 40 returns). As for home hours during the weekday, the most popular hours were: 10pm (with 46 returns), 11pm and 5am (with 44 returns).
2. The least popular working hours were: 4pm (with 14 returns), 11am (with 22 rides) and 8am (with 28 rides). As for home hours during the weekday, the least popular hours were 7pm (with 24 rides), 8pm(with 25 rides) and 4am (with 41 rides).

From the above, the most popular hour for working days is 10pm which is considered as home time. The least popular hour is 4pm which is a working hour time. Overall, working hours accounted for 366 returns. Home hours in general experienced the most traffic, with 461 returns.

Considering the weekends only;

1. The most popular day hours (between 8am and 6pm) were: 9am (with 19 returns), 8am (with 18 returns) and 11am (with 17 returns). As for night hours, the most popular hours were: 10pm (31 returns), 11pm (23 returns) and 9pm (both 19 returns).
2. The least popular day hours were: 10am (4 returns), 6pm (7 returns) and 5pm (8 returns). As for night hours, midnight, 1 am and 7 am had the lowest returns. Each of them had 3.

The weekend also has its highest returns at 10pm with 31 returns, but the least rides are spread out between different time frames as indicated in the point 2 above. The night returns were also more during the weekend (157), as compared to the day returns (131). In total, with both the weekday and weekend combined, the total day returns were 497 and the total night returns were 618. The least popular hours during the whole week were: 4pm (25 returns), 7am (27 returns) and midnight (35 returns). The most popular hours were: 10pm (77 returns), 11pm (67 returns) and 9am (56 returns).

### 5.3 STATION ANALYSIS

The address of the most popular station for picking up cars was ‘8 Avenue de la Porte de Montrouge’ with 15 pick up. ‘8 rue Goubet’ has 11 and ‘19 rue de Chateaubriand’ has 10. Returns top addresses are ‘ 8 Avenue de la Porte de Montrouge’ with 17, ‘19 rue BrÉmontier’ with 11 and ‘220 rue Marcadet’ with 10. For all activity considered, i.e. both returns and pick up, the most popular stations were ‘8 Avenue de la Porte de Montrouge’ with 32 total pickups and returns, ‘19 rue de Chateaubriand’ and ‘19 rue BrÉmontier’ both with 20. The least popular stations were quite a number, which all had no activity whatsoever.

During the week, working hours,the most popular stations for pick ups within a specific hour are ‘1 Rue du Colonel Pierre Avia’ with 6 pickups at 8am, ‘41 boulevard de Rochechouart’ with 6 pickups at 3pm and ‘37 quai de Grenelle’ with 6 pickups at 3pm as well. For returns, the results are ‘41 avenue Bosquet’ with 7 returns at 6pm, ‘220 rue Marcadet’ with 6 returns at 11pm and ‘2 avenue de la Porte de Charenton’ with 6 returns at 3pm.

For home hours, the top addresses were ‘356 rue des PyrAcnaces’ with 6 pickups at 1am, ‘8 rue Goubet’ with 6 pickups at 2am and ‘42 bis avenue Georges Mandel’ with 6 pickups at 1am. The returns were ‘210 rue de Courcelles’ with 6 returns at 11pm, ‘8 Avenue de la Porte de Montrouge’ with 6 returns at 11pm and ‘1 Rue du Colonel Pierre Avia’ with 6 returns at 1 am.

As for the weekend, the addresses are ‘19 rue Tronchet’ with 6 pickups at 12pm, ‘24 rue de la Paix’ with 4 pickups at 10am and ‘5 Place du Commerce’ with 4 pickups at 8am. Returns were ‘2 rue GacOral Renault’ with 7 returns at 8am, ‘150 rue Marcadet’ with 6 returns at 9am and ‘28 rue Championnet’ with 5 returns at 12pm.

During the night hours, the top stations were ‘210 rue de Courcelles’ with 6 returns at 6pm, ‘112 Avenue de Suffren’ with 5 pickups at 6pm and ‘35 rue Vauvenargues’ with 4 pickups at midnight. Returns were ‘210 rue de Courcelles’ with 6 returns at 6am, ‘112 Avenue de Suffren’ with 5 returns at 6 am and ‘35 rue Vauvenargues’ with 4 returns at midnight.

As can be noted from the above analysis, the top stations in general did not have a single hour in which they had dominant returns when compared to the other stations.

### 5.4 POSTCODE ANALYSIS

Most Popular Post Code for picking up and returning Blue Cars was 75015, which did not correspond to the top address ‘8 Avenue de la Porte de Montrouge’. The address’ Postcode was 75014, which was the 5th highest rank.

## **6. RECOMMENDATION**

Since no data was provided about the pick up and return times, calculations had to be done to determine them by subtracting consecutive available cars within the same address, arranged by time of day and eventually days. This was a lengthy process and to avoid mistakes in future, it is recommended that the data is collected including these two important variables. From the analysis, the most popular days seem to be the weekdays, especially Wednesday to Friday, while the least popular are the weekends. This could be caused by heightened economic activity during the weekdays as most work is carried out then. The most popular time for both activities (pick ups and returns) is home time rather than working time. This could be a result of people leaving their workplace, hence able to engage with the service mostly after or before work. Additionally, in between the day, at around lunchtime there seems to be large activity as people could be free to engage with the service. As was observed, the top stations in general did not have a single hour in which they had dominant activity when compared to the other stations. This is to be expected as the busier the station, the more traffic will be spread out within different hours. Finally, the most popular address was not in the most popular post code. This is because there were more busy stations in the 75015 postcode, even though they were not the top station. This could suggest that postcode 75015 is in a busier environment, where there’s more customer engagement with the product when compared to other postcodes.