***USAGE RATES AND CHARACTERISTICS FOR AUTOLIB (AN ELECTRIC CAR-SHARING SERVICE COMPANY)***

1. **BUSINESS UNDERSTANDING**

For this research project, the client is an electric car-sharing service company. As a Data Scientist, I have been tasked with responding to various questions that the company has regarding their electric car usage over time. By analyzing and processing data from their stations I need to help the company understand usage rates and characteristics over a given period of time.

The main research question that my analysis needs to respond to is identifying the most popular hour of the day for picking up a shared electric car(Bluecar) in Paris over the month of April 2018.

Some of the other research questions that this project seeks to respond to include:

* Most popular hour for returning cars.
* Most popular station.
* Overall.
* During the most popular picking hour.
* Most popular postal code for picking up Blue cars and relation of this postal code to most popular station.
* Overall.
* During the most popular picking hour.
* How the results change when Utilib and Utilib 1.4 are considered instead of Blue cars.

This project will be successfully completed when the answers to the queries above are provided.

1. **DATA UNDERSTANDING**

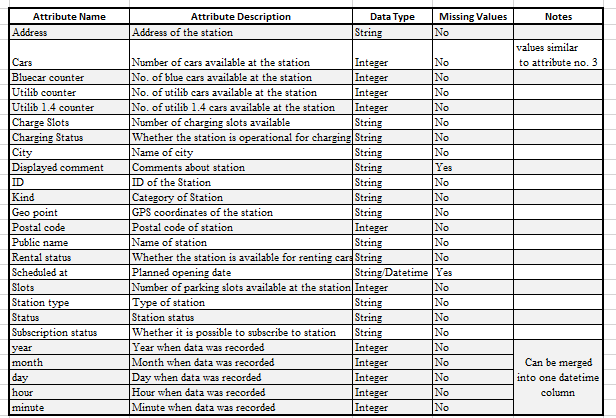
*Data Collection*

The dataset necessary for this research project was provided by the client (Autolib), which eliminated any challenges associated with data collection and acquisition.

*Data Exploration*

The main dataset required for this project was available in a csv file labelled Autolib\_Dataset. A word document containing the description of the dataset was also provided.

The Autolib\_Dataset table contains twenty-five attributes. Below is a summary table describing the characteristics of the different attributes available in the dataset.



1. **DATA PREPARATION**

*Data Cleaning and Transformation*

I used Pandas and Numpy libraries in Python to perform the following data cleaning steps:

* I checked the validity of the dataset and determined that some attributes were irrelevant for the research project. The attributes that were dropped include Address, Subscription status, Displayed comment, Geo point and Scheduled at.
* I also standardized the naming format of the attributes by using lowercase on all the column names.
* I create a new attribute called ‘time’. This attribute combines five attributes i.e year, month, day, hour and minute attributes into a ‘time’ attribute with the format YYYY-MM-DD HH:MM:SS. I also converted the data type of this attribute from string to datetime.
* I verified that there were no missing values in the remaining attributes.
* I checked and confirmed that there were no duplicates within the remaining records and attributes.
* In the dataset description provided by the client, the attribute labelled ‘cars’ and ‘bluecar counter’ should have the same values. I checked the accuracy of this statement within the dataset and confirmed that records in the two columns matched.
* I also confirmed the accuracy of the ‘hour’ attribute to make sure it contained values for 0 to 24 only since those are the available hours in a day. I performed this step for the ‘minute’ attribute as well to make sure values were between 0 and 60.
* Finally, I saved the cleaned dataset as a csv file within my colab notebook for analysis.

1. **ANALYSIS**

I used SQL to analyze the data and respond to the research questions posed by the client.

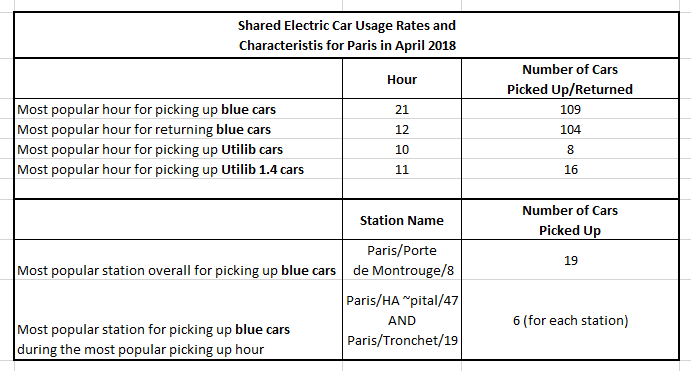
I exported the cleaned dataset as a csv file within the colab notebook and loaded the saved csv file into sqlite database. I then wrote various queries to answer the research questions.

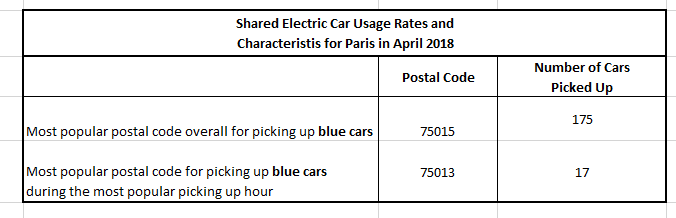
In order to identify whether a car was picked up or returned, I analyzed the change in the number of cars for each station at subsequent times. For instance, for a station named ‘Paris/Suffren/2’ which is in Paris, I ordered the newly created ‘time’ attribute in ascending order meaning from the earliest date and time to the latest date and time. Then I proceeded to check how the number of blue cars changed from the earliest time all the way to the latest time for that station.

If on the 4th of April at 7:03 am the number of blue cars was 6 at that station and at 9:45am the number changed to 4 cars, then it means that 2 blue cars must have been picked up at the 9am hour. On the other hand, if at 11:30am the number of blue cars was recorded as 5 then it must be the case that 1 car had been returned at the 11am hour. No change in the number of cars for subsequent times implies that a car was neither picked up nor returned.

I used CTEs and joins in SQL to write queries that show the changes in the number of cars for each station over time. I also used WHERE and ORDER BY clauses to filter data so as to display only stations in Paris and also sort the time and date in ascending order. Finally, I wrote calculations to get the difference in number of cars over time then used the aggregate function SUM as well as the GROUP BY clause to aggregate the changes and group the total change by hour. A clearer description of this process can be found in the colab notebook via the provided GitHub link.

Below is a summary of the results for the various questions addressed in the analysis.





From the result set in the colab notebook, we also notice that overall, the most popular station is not in the most popular postal code in regards to the rates of picking up blue cars.

However, when looking at the most popular hour only, the most popular station is in the most popular postal code for picking up blue cars.

We also observe that the most popular hour for picking up the shared eclectic cars varies in Paris for blue cars and Utilib cars. This is also the case when you compare blue cars and Utilib 1.4 cars as shown in the summary table above.

1. **RECOMMENDATION**

From the data analysis results described above, we can see that in Paris, blue car usage is much higher compared to Utilib and Utilib 1.4 cars. In the span of 9 days, 109 blue cars were picked at various stations in the most popular hour unlike 16 and 8 Utilib 1.4 and Utilib cars respectively. This information is vital since Autolib can now ensure that they prioritize making blue cars largely available in all their stations especially during the 21st hour. The analysis has also allowed us to identify the most popular stations and postal codes within the city of Paris where blue cars were picked up more frequently.

The above insights obtained in this project will allow Autolib to make informed business decisions within the company. In case there is a need for expansion, the company can optimally select areas to focus on to ensure continued high rates of usage in those popular geographical areas. The company can also ensure that the most popular station and postal codes stay operational and are performing optimally. For instance, Autolib can ensure that all available charging stations in those popular stations are working more often than not. Finally, the company can further investigate why the less popular stations are performing poorly in terms of usage and take necessary steps to boost usage in those stations and areas.

1. **EVALUATION**

The dataset available for this research project provided useful information regarding the usage rates and characteristics of shared electric cars in the city of Paris. These rates were for the first nine days of April 2018.

The project was conducted in various steps as outlined in the CRISP DM methodology. This framework ensured that the project was executed in a well organized manner and that reliable results were obtained at the end.

The project can be replicated further to identify the usage patterns for other time periods as well as other cities where Autolib is operational. Furthermore, the results obtained in this project can be displayed in an interactive dashboard which the client can access.