**A DATA REPORT ON PROCESSING STATIONS DATA FOR AN ELECTRIC CAR-SHARING SERVICE COMPANY TO UNDERSTAND ELECTRIC CAR USAGE OVER TIME.**

Compiled by: David Muna Mwangi

Data Scientist, Ajira

(See Section 4. Analysis for link to Python Notebook)

1. **Business Understanding**

Breaking down the whole problem into smaller problems that are easily rectified. We start by answering the research question which is;

* Identifying 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.

Having answered this question, we will now be able to answer additional research questions since the main question is the backbone of the whole research undertaking.

During the data collection period, there were occasions where downloading the data collected failed and this created discrepancies in the data collected.

The main need for understanding usage over time is so that the company can identify peak hours and popular stations for picking an electric car. With this information in place, the company can identify areas that need an upgrade as well as identify reasons as to why certain stations are not so popular.

Popularity of a station might arise because of factors such as location of the station and staff customer service. Some of the leading questions that will help in analysing the datasets include and are not limited to;

* What is the most popular hour for returning cars?
* What station is the most popular?
  + Overall?
  + At the most popular picking hour?
* What postal code is the most popular for picking up Blue cars? Does the most popular station belong to that postal code?
  + Overall?
  + At the most popular picking hour?
* What is the correlation between variables in the dataset?

1. **Data Understanding**

The following are the resources/datasets provided by the Jones Electric Car Sharing Company . To understand the data, we first have to describe what it looks like and the shape/form it has taken so that we can identify how we want to format it for easier analysis. The table below describes the shape and gives links to preview the datasets as they are before any preparation takes place. The shapes of the datasets were obtained using autolib.shape. See section 1.3 on accessing information about the dataset in our python notebook. The link will be provided in the sections below. For this research undertaking, we only have one dataset and its description.

**NB:** The third row in the table below describes the new excel.csv file we exported after cleaning the raw dataset. This will be the file that we will use for analysis.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Shape/Size/Type** | **Preview** |
| Autolib\_dataset (2).csv | 5000 rows x 25 columns | [Link](http://bit.ly/autolib_dataset) |
| Autolib\_DDI\_DB\_description\_MoringaSchool\_w4.docx | Word Document | [Link](https://drive.google.com/a/moringaschool.com/file/d/13DXF2CFWQLeYxxHFekng8HJnH_jtbfpN/view?usp=sharing) |
| autolib\_cleaned.csv | 5000 rows x 20 columns | File generated in Notebook |

From the dataset, the columns were not standardised. We changed all column names to lowercase, standardised the data types of the respective variables in the dataset as shown below. We can also see that the columns for Month, Year, Scheduled at, Displayed comment and Cars were dropped. This is because the column ‘Cars’ was identical to ‘Bluecar counter’. We also removed any spaces between column names and replaced the spaces with an underscore for easier reference. Month and Year columns were unnecessary while the remaining two columns had a lot of missing values so we dropped them.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 address 5000 non-null object

1 bluecar 5000 non-null int64

2 utilib 5000 non-null float64

3 utilib\_14 5000 non-null float64

4 charge\_slots 5000 non-null float64

5 charging\_status 5000 non-null object

6 city 5000 non-null object

7 id 5000 non-null object

8 kind 5000 non-null object

9 geo\_point 5000 non-null object

10 postal\_code 5000 non-null int64

11 public\_name 5000 non-null object

12 rental\_status 5000 non-null object

13 slots 5000 non-null int64

14 station\_type 5000 non-null object

15 status 5000 non-null object

16 subscription\_status 5000 non-null object

17 day 5000 non-null int64

18 hour 5000 non-null int64

19 minute 5000 non-null int64

dtypes: float64(3), int64(6), object(11)

1. **Data Preparation**

After describing the data, we cleaned, explored, and verified its integrity/quality for analysis. We started by;

* Dropping irrelevant columns (Month, Year, Scheduled at, Displayed comment and Cars)
* Renaming columns, fixing syntax errors and putting column names them in lower case for standardisation
* Viewing and removing Outliers in our dataframe and creating a new dataframe
* Checking our dataframe for null values
* Backward filling the null values since the values at the top are zero, we want to fill with whatever values which were below
* Removing duplicates from our dataset
* Exporting cleaned dataframe to a csv file and previewing it

The main tool we used for data cleaning and analysis was Python Pandas using a Google Collaboratory Notebook.

1. **Analysis**

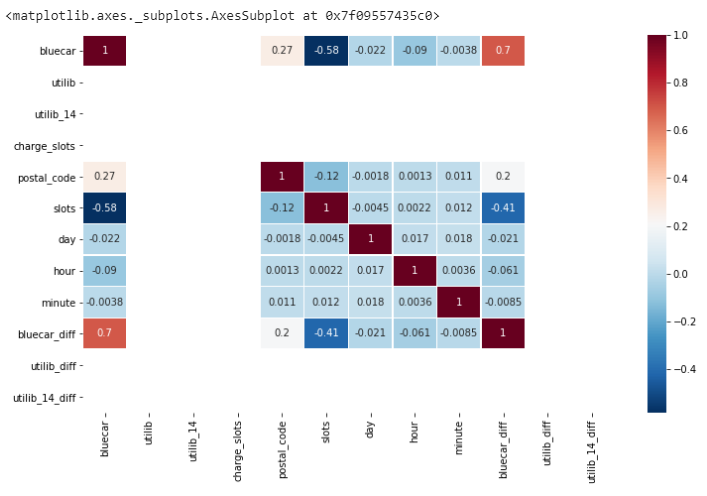
We analysed the dataset using Pandas by answering the research questions in the business understanding part to derive information from the data. Here is the link to the copy of a google collaboratory notebook that contains the data preparation and analysis. [Analysis Notebook -Github](https://github.com/davidmuna/electric-car-sharing-service)

1. **Recommendation**

The results obtained from the analysis are in line with my expectations. Paris is a city filled with romance and adventure and it is plausible to conclude that the majority of the residents therein are constantly up and about after 2100 hours to engage in nightlife activities filled with romance.This is a reason why 2100 hrs has been identified as the most popular hour in Paris and overall. See section 1.5 under “What station is the most popular?” . It is therefore beneficial for the company to identify reasons why Paris is most popular and use that information to identify where their next upcoming stations will be set up.

Additionally, other less popular stations, their managers and staff members can learn a few pointers from the managers and staff members of the most popular station overall and in the popular hour which are Paris/Porte de Montrouge/8 and Paris/Tronchet/19 respectively.

The figure below shows that the number of “slots” is correlated to the number of blue cars available at a given station which translates to whether or not a given station is popular/convenient. Slots show a -0.58 correlation coefficient with bluecars column. The interpretation is that 1 shows maximum correlation while zero shows no correlation. Therefore, the company can work on increasing the number of slots available to increase usage of a station.



1. **Evaluation**

In this stage, we took a look at the generality of the analysis and the accuracy and/or degree to which it met the expectations of the success criteria for the electric car sharing company.. From the results obtained from the analysis as seen from the link above, we can conclude that the model is accurate since all outliers and null values were dealt with accordingly.

Measures should be put in place to ensure that acquisition of data is clean and crisp to prevent instances of missing data, incorrectly filled data or incorrectly formatted data. This in the end reduces the time taken to clean the data and verify its integrity thereby increasing the accuracy of analysis and improving models for business success criteria in the electric car sharing company.