**DATA REPORT ON AUTOLIB:**

**GITHUB LINK:** [**https://github.com/LKabaya/Moringa\_Prep\_Week4\_IP.git**](https://github.com/LKabaya/Moringa_Prep_Week4_IP.git)

1. **Business Understanding**

Autolib is a car rental company that lets you rent an electric car from various stations in cities in France. A car can be picked, dropped and charged at any of the Autolib stations. The data provided shows statistics on the rental car usage for a period of 9 days in April. There are peak hours when most customers pick or return cars and there are also stations that are generally more popular.

We seek to draw insights from the data so as gain a better understanding of the business and identify trends in the customers usage of the service. Determining the busiest hours and stations will help in better resource allocation and in improving service delivery to its customers.

We seek to answer the following questions:

1. **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.**
2. **Identify the most popular hour for returning cars**
3. **Identify the most popular station overall and at the most popular picking hour**
4. **Identify the most popular postal code for picking up Bluecars- Does the most popular station belong to that postal code overall and at the most popular picking hour?**
5. **Consider how the above parameters would change if we evaluated Utilib and Utilib 1.4 instead of Bluecars**
6. **Data Understanding**

The dataset contains data collected for a period of 9 days in the month of April 2018.

The dataset has different statistics on different fields, whose description is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Type | Values | Comments |
| Address | String |  | address of the station |
| Cars | Number | [0-7] | Number of cars available at the station - redundant with Bluecar counter, always the same value |
| Bluecar counter | Number | [0-7] | Number of Bluecars available at the station |
| Utilib counter | Number | (0-4] | Number of Utilibs available at the station |
| Utilib 1.4 counter | Number | [0-5] | Number of Utilib 1.4  available at the station |
| Charge Slots | Number | [0-3] | Number of Charging slots available at the station |
| Charging Status | String | {"nonexistent","operational","broken","future", some typos} | Whether the station is operational for recharging. Mainly "nonexistent", "operational" or "broken": charge slots can only be greater than 0 when "operational"; slots and vehicles can be available in all situations (except future stations that have 0 resources) |
| City | String |  | City |
| Displayed comment | String |  | Some comments like "station within parking, access through …" |
| ID | String |  | ID of the station |
| Kind | String | {"STATION","SPACE,"PARKING",CENTER"} | "CENTER" have no resources at all; "PARKING" do not have charge slots, but can have bluecars and utilib; "STATION" and "SPACE" can have all resources |
| Geo point | String |  | GPS coordinates of the station |
| Postal code | Number |  | Postal code of the station |
| Public name | String |  | Name of the station |
| Rental status | String | {"nonexistent","operational","broken","future", a few empty} | Whether the station is available for renting vehicles. Resources are only available when "operational", except for "broken" which can have Slots, but none of the other resources (Bluecars, utilib or charging slots). |
| Scheduled at | String | datetime | Planned opening date: non empty values only for stations that have "future" in one of the statuses. |
| Slots | Number | [0-7] | Number of parking slots available at the station? |
| Station type | String | {"station","full\_station","subs\_center"} | No resources available for "subs\_center" - which is just one location. Was that actually a selling point for Autolib subscriptions? |
| Status | String | {"ok","closed","scheduled"} | No resources available for "scheduled", which is the status if there is a "scheduled at" date. Yet there can be resources associated with "closed" stations |
| Subscription status | String | {"nonexistent","operational","broken","future"} | Whether it is possible to subscribe to the autolib service in that station? No resources available when "future", but other values can have resources |

Based on our understanding of the data and the business objectives we seek to achieve, the key data fields are: ***Cars, Postal Code, City, Kind, ID and Hour***

Upon running descriptive statistics, we unearth the following insights about the data types of our dataset:

Variables types

|  |  |
| --- | --- |
| **Numeric** | 10 |
| **Categorical** | 13 |
| **Boolean** | 0 |
| **Date** | 0 |
| **Text (Unique)** | 0 |
| **Rejected** | 3 |
| **Unsupported** | 0 |

The columns ‘DISPLAYED COMMENT’ and ‘ SCHEDULED AT’ have more than 90% missing values and should therefore be dropped.

The dataset has 142890 duplicate rows which should be dealt with at the data cleaning stage.

1. **Data Preparation:**

Upon review of our data description and the descriptive statistics, we prepare our data by ensuring that it is Clean, Consistent, Uniform, Valid and Complete.

*Completeness:*

Checked for missing values

*Validity:*

Dropped the following columns which were not necessary in answering our data mining questions: *Unnamed:0, Bluecar counter, Public name, Displayed comment, Scheduled at, Subscription status , Station type and Geo point*

*Accuracy*:

Remove outliers that lie more than 1.5 times IQR below Q1 or above Q3

*Consistency:*

Checked for duplicates and dropped duplicated rows.

*Uniformity:*

Data-type conversion to datetime format to help in manipulating the data

1. **Analysis**

**What was 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?**

Picking up means that there will be a negative change in cars at the station. The hour with the highest rate of negative change in cars

Most popular hour is hour 19, with total pickups of 5534 cars in the dataset period

**What is the most popular hour for returning cars?**

Returns result in increase of cars at the station therefore Determine hour with the highest rate of positive change in car

Hour 20 has the overall highest rate of returns at 8565

**What station is the most popular?**

**Overall?**

This is the station with most activity whether positive or negative

Station with the most activity cumulatively over the dataset period is **paris-portedauteuil-parking with a frequency of 395**

**At the most popular picking hour**

First determine most popular picking hour overall(for all stations) which is Hour 19 with a frequency of [ ]

The most popular station at hour 19 is the station with highest rate of change at hour 19 which is **boulognebillancourt-jeanjaures-245 with frequency of 32**

**What postal code is the most popular for picking up Bluecars?**

The most popular Postal code is 75015 with a frequency of 7366

**Does the most popular station belong to that postal code?**

**Overall?**

**At the most popular picking hour?**

1. **Recommendation**

Autolib should invest more resources in the postal code and stations which have the most activity. Being Postal Code 75015, Station **boulognebillancourt-jeanjaures-245** and **paris-portedauteuil-parking**

Since the busiest hours are hours 19 and 20. Autolib should have more staff at the stations at that time and ensure that there are enough cars available at those hours.

Autolib should ensure that it has adequate cars, charging ports and parking resources at the stations which has the highest activity.

1. **Evaluation**