# **DATA REPORT: DDI\_AUTOLIB PROJECT**

### **Business Understanding :-**

Humans are the most mobile creatures on planet earth. At the heart of their movements is cars. Cars manufacturing has experienced great disruption over the years especially with the onset of electric vehicles. That is why numerous companies have very much invented not only on production but also car-sharing apps.

**The key business objective of this report is to draw insights from the dataset provided and understand electric car-usage over time.**

**Research questions:**

1. 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?
2. What is the most popular hour for returning cars?
3. What was the overall popular station?
4. What was the popular picking hour?
5. What postal code is the most popular for picking up Bluecars? Does the most popular station belong to that postal code? Both overall and was at the most popular picking hour?
6. Do the results change if you consider Utilib and Utilib 1.4 instead of Bluecars?

#### **Data Understanding:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **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** |

This phase includes a detailed documentation of data description, exploration of the data - done through running descriptive statistics on the datasets and trying to find out of any correlations and regression across variables and any relationships before we dive into data preparations.

**Data Description:-**

1. **Our majorly comprises of two data types strings and numbers.**
2. **We have 21 columns in our dataset.**
3. **The key resources from our dataset are : BLUECAR COUNTER, UTILIB COUNTER, UTILIB 1.4 COUNTER, CHARGE SLOTS AND PARKING SLOTS.**

#### **Data Preparation:-**

The data preparation phase included three tasks. These are

* **Selecting data the dataset - Autolib\_IDD**
* **CREATING AN SQL DATABASE ENVIRONMENT TO LOAD HUGE CHUNKS OF DATA IN MILLIONS -** using **import os, import csv, import sqlite3, and from sqlalchemy import create\_iengine**
* **Cleaning data** - I made the following changes :
  1. **DROPPED** the following columns because if their irrelevance in answering the questions('**Unnamed', 'Displayed Comment', 'ID', 'Geo Point', 'Scheduled at ''Address', 'Cars**')
  2. **CHECKING FOR ACCURACY BY DEALING WITH OUTLIERS**
  3. **DEALT WITH MISSING VALUES -** the column **Scheduled at** and **Comments** columns had too many missing values therefore it got dropped
  4. **DEALT WITH DUPLICATES -**
* **Constructing data - Derived attributes:** New fields (columns) constructed was the **DATE** variables - which combined the YEAR, MONTH, DAY, HOUR, MINUTE columns into one variable inorder to get time Cars were returned.
* **Formatting data -** Our final cleaned dataset has relevant columns with no duplicates, null values or irrelevant columns for our later analysis.

#### **Analysis:-**

The data analysis phase used the answer the business questions we had.

1. First I ensured the City was **‘PARIS’ by dff[dff['city'] == 'Paris']**
2. Next used the **Min() function** grouped by Bluecar\_counter to get the least popular car counter on the blue cars
3. The most popular Postal Code per hour for returning blue cars dff.groupby('postal\_code',dff.date.dt.hour,['bluecar\_counter'].count().sort\_values(ascending = False) ~ **78005 most popular with 14419 units of cars**
4. **Most Popular Station - for picking blue cars**

**dff.groupby('public\_name','cars'].sum().sort\_values(ascending = False).head(500)**

1. **Most Popular hour for picking hour for blue cars : 0 = 129372 units**

**dff.groupby('public\_name',dff.date.dt.hour.['bluecar\_counter'],['bluecar\_counter'].count().sort\_values(ascending = False)**

**6. Most Popular Station at the most popular picking hour for Utilib\_counter**

**dff.groupby('public\_name',dff.date.dt.hour.['utilib\_counter'],['utilib\_counter'].count().sort\_values(ascending = False)**

#### 

#### **Recommendation:-**

This analysis fostered the following recommendations:

1. Most Clients returned the cars in the wee hours of the night which was midnight and therefore the need to consider hiring night shift personnel for seamless handling of clients when they return the vehicles.
2. The station with the highest traffic could be turned into a high priority station with special infrastructure.
3. Utilib compared to Blue cars counter had lower traffic, but its conspicuous client behaviour still makes it a car type of interest as clients somehow prefer it.
4. Paris is a high profile city with irregular client preference when it comes to cars selections. Therefore more research can be done and invested in on the various loopholes on behavioral tendencies of our clients.

#### 

#### **Evaluation:-**

This analysis clearly brings out favourite hours of returning vehicles and vehicle preferences for the clients based on diverse factors such as traffic, need of cars, tourism in Paris as a famous romantic city etc etc.

Therefore marketing and brand strategy for our car sharing app could escalate to strategic places where most tourists enter the city, the need for sharing a car and client age groups and all.

**The Github repository for my project is at:**

[**GITHUB WEEK 4 PROJECT REPOSITORY**](https://github.com/kweyuchesa/MORINGA_WEEK_4_IP)