The report to accompany my Research with Autolib Dataset.

Business Understanding.

The Dataset comes from an electric car sharing service company, The dataset objective is to identify the most popular hour of the day for picking up a shared electric car in the city of paris over the month of April 2018. To determine the most popular hour of returning cars,the most popular hour station overall and at the most popular picking hour.

Also I am to determine the most popular postal code for picking the blue cars,and whether the most popular station belongs to that postal code.

Data understanding.

The data set consists of 5000 records and 24 columns that is Address,Cars,Bluecar Counter,

Utilib Counter,Utilib 1.4 counter,charge slots,charging status,city,Displayed comment, ID, Kind,Geo Point,Postal Code,Public Name,Rental Status,Scheduled at, Slots,Station Type,Status,Subscription Status,year,Month,day,hour and minute of which comprises of different data types.

To display this info

df.info()

RangeIndex: 5000 entries, 0 to 4999

Data columns (total 25 columns):

# Column Non-Null Count Dtype

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

0 Address 5000 non-null object

1 Cars 5000 non-null int64

2 Bluecar counter 5000 non-null int64

3 Utilib counter 5000 non-null int64

4 Utilib 1.4 counter 5000 non-null int64

5 Charge Slots 5000 non-null int64

6 Charging Status 5000 non-null object

7 City 5000 non-null object

8 Displayed comment 111 non-null object

9 ID 5000 non-null object

10 Kind 5000 non-null object

11 Geo point 5000 non-null object

12 Postal code 5000 non-null int64

13 Public name 5000 non-null object

14 Rental status 5000 non-null object

15 Scheduled at 47 non-null object

16 Slots 5000 non-null int64

17 Station type 5000 non-null object

18 Status 5000 non-null object

19 Subscription status 5000 non-null object

20 year 5000 non-null int64

21 month 5000 non-null int64

22 day 5000 non-null int64

23 hour 5000 non-null int64

24 minute 5000 non-null int64

To find the shape of the dataset

df.shape

(5000, 25)

The relevant columns that will help solve the business at hand are

# Column Non-Null Count Dtype

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

0 Address 5000 non-null object

1 Cars 5000 non-null int64

2 Bluecar counter 5000 non-null int64

3 Utilib 1.4 counter 5000 non-null int64

4 Charge Slots 5000 non-null int64

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 year 5000 non-null int64

18 month 5000 non-null int64

19 day 5000 non-null int64

20 hour 5000 non-null int64

21 minute 5000 non-null int64

Data Preparation

In this stage of CRISP-DM Methodology covers all the activities to construct the final dataset from the initial raw data.

The relevant columns that will help solve the business at hand are

# Column Non-Null Count Dtype

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

0 Address 5000 non-null object

1 Cars 5000 non-null int64

2 Bluecar counter 5000 non-null int64

3 Utilib 1.4 counter 5000 non-null int64

4 Charge Slots 5000 non-null int64

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 year 5000 non-null int64

18 month 5000 non-null int64

19 day 5000 non-null int64

20 hour 5000 non-null int64

21 minute 5000 non-null int64

To determine the most popular hour of picking up the car in the city of Paris,we compare the most popular hour against the city of Paris.

df[df['City'] =='Paris']['hour'].value\_counts()

This determines that Hour 21 is the most popular in the city of Paris as it has the highest count.

The determine the most popular hour for returning cars is

df['hour'].value\_counts()

The hour 21 has the most counts for returning the cars

Determining the popular hour the blue car is picked in the city of Paris.

df[(df['Bluecar counter'])&(df['hour'].value\_counts()) & (df['City']=='Paris')]['Bluecar counter'].count()

The hour 7 is the popular hour

Analysis.

Github link: https://github.com/ArthackA/moringaIP4.git