Project
OMAC

Data : GoBike Sharing System

Date: Oct 14th 2021

This project is to explore "Ford Gobike Sharing System"

- starting with posing questions about the data
- **First: Wrangling Data**

. . . Gathering Data
. . . Assessing our data
. . . Cleaning the data

- **2nd : Visualization and Communicate Data Findings**
- answering the questions

This data downloaded from Kaggle website (https://www.kaggle.com/franckjay/fordgobike- data), collected by Ford GoBike System Sharing, for the period from end of June 2017 until first of July 2018 (one year). There are other data files cover all the period of complete year of 2018 and 2019, but I believe our downloaded data with (1338864) rows will do for our current study. Ford GoBike runs business of bike sharing; they have huge numbers of heavy duty, durable fleet of bikes, specially designed and manufactured for this purpose.
Distributed in the most important areas in California forming a network of stations in a chosen venues for public use.
The data consist of (1338864) rows, (16) columns. . First group of data: start station related information such as start time, id, name, location defined by its latitude & longitude. Same information types for the end station, besides bike id, user_type, age of members, gender of members and bike share for a trip.

Posing questions

1. When are the most numbers of trips taken in terms of:
    - period of day (morning, afternoon, evening and night)
    - weekdays (monday to sunday)
    - month of year (january till december)
2. What is the average trip duration?
3. What is the average trip duration in term of:
    - period of day (morning, afternoon, evening and night)
    - weekdays (monday to sunday)
    - month of year (january till december)
4. Are the above depend on user type or not?
5. what other insights we may find from our data exploration & explanation?

First: Wrangling Data

Gathering Data:

Duly downloaded from the above-mentioned website, the data consist of multiple files (7 csv format files). So after reading the data in pandas, it concatenated together in pandas into one csv format file.

Assessing Data:

- the collected data covers the the period:
  (from '2017-06-28 09:47:36.3470', till '2018-06-30 23:58:48.2930')
- the trip duration is the very important variable in our project:
  - mean trip duration time = about 16 minutes
  - minimum trip duration time= about 1 minute
  - maximum trip duration time= about 1440 minutes= one day

  other information about trip duration time:

  - trip duration time median = 9.5 minutes
  - trip duration time first quantile = about 6 minutes
  - trip duration time third quantile = about 15 minutes

  - mean of member birth year = 1981
  - minimum of member birth year = 1932
  - maximum of member birth year = 2000
  - median of member birth year = 1984
  - first quantile of member birth year = 1976
  - third quantile of member birth year = 1989

- the data contains some outliers: such as some persons (members) has birth year before 1900. this means that we have some persons who are participating in this activity have more than 120 years age.
- user types consist of two categories: customer and subscriber.
- start_station_id, start_station_name, end_station_id and end_station_name contains cells with null values
- total numbers of stations = 308

Quality Issues:

- reindex the dataset (as our dataset consists of multiple data sets, need for reindexing after concatenating them together).
- drop null rows in the dataset
- modifying data type for some variables (columns):
- start_time (from string (object) to datetime)
- end_time (from string (object) to datetime)
- start_station_id (from "float" to "int" then to "string" type)
- end_station_id (from "float" to "int" then to "string" type)¶
- bike_id (from "int" to "string" type)
- member_birth_year (from "float" to "int" type)
- drop incorrect data values in member_birth_year

Tidiness Issues

- Creating columns from other columns with multiple variables are stored in one column.

Cleaning:

- creating a copy dataframe during the cleaning process
- do the necessary to correct the quality and tidiness issues above-mentioned. using the standard steps in cleaning processes (**Define** the problem or dirty/untidy data, **Code** the proper coding to correct the issues then finally **Test** our action to see if the issue has been corrected)

Data structure: our original data consist of the following variables

Number of Columns= 16 columns
Number of Rows = 1338864 rows

- duration_sec: duration of each trip in second
- start_time: trip start time
- end_time: trip end time
- start_station_id: start station ID number
- start_station_name: start station name
- start_station_latitude
- start_station_longitude
- end_station_id: end station ID number
- end_station_name: end station ID number
- end_station_latitude
- end_station_longitude
- bike_id: Bike ID number
- user_type: divided into two categories: Customer & Subscriber
- member_birth_year:
- member_gender: divided in three categories: Male, Female and Other
- bike_share_for_all_trip: Yes or No

after the wrangling and cleaning process 5 other columns were added: all the following added columns were derived from start_time column

- year: year at which any trip start (2017 or 2018)
- month: month of year the trip start (of course we have 12 months Jan, Feb ...till Dec)
- day: day of week the trip start (7 days, start from Mon, Tue .... till Sun)
- hour: the hour of day that any trip start (start from 00, 01, 02 ... till 23)
- period of day: it consist of 4 categories (Morning, Afternoon, Evening and Night

the total size of our data now after complete wrangling and cleaning will be:

Number of Columns= 21 columns
Number of Rows = 1210333 rows

Exploration of Data

- the most numbers of trips are taken in June month then May, the least numbers of trips are taken in July.
- most member birth year lying between 1982 and 1992 (member ages of about 26 to 36 year: base year is 2018).
- most numbers of trips taken are in May and June and the least number of trips are taken in July.
- most numbers of trips are taken in Tuesdays, Wednesdays and Thursdays, the least Number of trips are taken in Sundays
- most number of trips are taken in morning and afternoon times the least number of trips are taken at night.
- users type are distributed between subscribers 89.2% and customers 10.8%.
- member gender consist of male: 74.9%, female: 23.7% and other: 1.4%.
- bike share for all trip: yes: 9.1% and No: 90.9%

Now it is time to answer our questions, and summarize the main findings:

1. **When are the most numbers of trips taken in terms of:**
   - period of day (morning, afternoon, evening and night)
   - weekdays (monday to sunday)
   - month of year (january till december)

starting with the first term (period of day): most trips are taken at the afternoon and morning times, the percentage of trips that are taken at afternoon time is about 43%, we may give the following percentage for each period of day:

- afternoon time: about: 43%
- morning time: about : 42%
- evening time: about : 14%
- night time: about : 1%

in the other hand the most trips are taken at weekdays are in Tuesday, Wednesday and Thursday (about 17.5% plus / minus, the peak value is on Tuesdays) then it descends gradually till it reaches to the undermost values (about 8.5% and 7.5%) on Saturdays and Sundays respectively (or on weekends)

in term of month of years: the maximum numbers of trips are taken in June month with about 15.2% then May with about 14% of the total trips. the least number of trips are taken in July with only about 3%.

2. **What is the average trip duration:**
   the average trip duration takes about 16 minutes. Though our minimum trip duration is only about one minute, the maximum trip period is about 1440 minutes this equals 24 hours or one entire day.

3. **What is the average trip duration in term of:**
   - period of day (morning, afternoon, evening and night)
   - weekdays (monday to sunday)
   - month of year (january till december)

the average trip duration regardless of any constraints of time is 16 minutes
in term of period of day, the longest average trip duration are on mornings and afternoons, the average trip durations are:

- morning time : about 13.5 minutes
- afternoon time: about 13.0 minutes
- evening time : about 11.5 minutes
- night time : about 10.0 minutes

in case of weekdays, the longest average trip duration are on Sundays and Saturdays (or on weekends)
the distribution of average trip duration on the weekdays has its maximum value of about 17 minutes on Sundays then about 16 minutes on Saturdays (the maximum average is in weekends) then the value line goes down to the least value of about 11 minutes on Mondays

in case of months of year, the longest trip duration is in July with about 18.25 minutes then November and September with about 15 minutes, the least average trip duration is in June and December with about 11 minutes.

4. **Are the above depend on user type or not**

- the answer is yes,
- the average trip duration in minutes for each user type are as follows:
- period of day
    - in case of subscribers: the average trip duration at morning and afternoon times will be about 11.5 minutes, this value will come down at evening and night to about 10 minutes.
    - in case of customers: the average trip duration will be at its peak in the morning time of about 35 minutes, then it goes down in afternoon time to about 27 minutes then to 23 minutes at evening time and finally to about 13 minutes at night period.

- weekdays:
    - the same scenario is noted on weekdays, as the average trip duration for the subscribers are more balanced scale (almost constant or not changing much from day to day) and further less than the customers average trip duration. the subscriber average trip duration in its highest level in Sundays for about 13 minutes then Saturdays in 12 minutes then it goes to its minimum at Wednesdays about 10 minutes.
    - in the contrary the customer average trip duration have a maximum value of about 33 minutes on Sundays, 32 minutes on Saturdays, 31 minutes in Tuesdays and it goes down to its minimum in Mondays with average trip duration equals about 18 minutes.

- going to month scale:
    - the same pattern will continue in case of subscribers with average trip duration of about 10 minutes for all months of year with a rise in this value to about 13 minutes in November.
    - in case of customers: our peak value of the average trip duration is in July with a value of about 80 minutes, then it goes down dramatically to about 38 minutes in April and it reaches its bottom

value in June and November with average trip duration equal to about 20 minutes.

5. **what other insights we may find from above.**
   we can conclude from above results that although the least number of trips are taken on Saturdays and Sundays (or on weekends), but the average trip duration has its maximum on weekends (Saturdays and Sundays). this means that there is a limit numbers of members will hire these bikes on weekends (so the number of trips are at its minimum) but they will keep the bike for a longer time and maybe for daylong (they may need these bikes in having fun and in their weekend various activities) and this will make the average trip duration has its utmost value on weekends.
   in the contrary the number of trips are taken on other days (working days) especially on Tuesdays, Wednesdays and Thursdays is much higher from these taken on Weekends, but the average trip duration on these days (working days) is much less or limit. Apparently members use the bikes to reach their work locations.
   in a bar plot (in univariate section) for proportion of bike hiring during the hours of day (along 24 hours): the maximum proportion (about 12%) is on 8 o'clock, morning and 5 o'clock afternoon. this means that the most bike sharing activities are used in going to work (8 o'clock in the mornings) and back home (5 o'clock in the afternoons) at rush hours. the second position (9.5%) goes to the period of (9 am) and (6 pm) each of them is time after the rush hour, going work time (8 am + 9 am) then back home (5 pm + 6 pm). accordingly the total proportion for these four hours (12% + 12% + 9.5% + 9.5%= 43%). and this a nother clue of using the bikes in reaching work locations and back home.

   Mohamed Makki