Divvy_Bike_Analysis

Wei_Cheng_Yu

Data Preparation:

First, I imported every library I needed such as *Numpy, Pandas, Matpltlib* and *Seaborn* for data manipulation and visualisation.

I downloaded the datasets Divvy_Trips_2017_Q1 and Divvy_Trips_2017_Q2 and read the data using the 'pd.read_csv()' function, and I have a quick look by using *'df.head()'*. After concatenating these two datasets using the 'pd.concat()' method, I had a dataset with a shape of (1551505, 12) .

I needed to cleanse and transform some data before using it. Thus, I use 'df.isnull().sum()' function to find out there are some null values in *gender* and *birthday,* I assumed gender and birthyear(ages) are important factors for analysis, so I prefer to delete all null values using 'df.dropna()' function instead of replacing them with another values. After this cleansing, I have the shape of (1234638, 12) dataset finally. Then, I checked if there were duplicate by using 'df.duplicated().sum()'*;* check the datatype of each categories by using 'df.dtypes' function. After these checks, I found out there is no duplicates or inappropriate datatype in the dataset.

My goal is to optimise the profit of Divvy Bike company, so I selected some critical factors such as *tripduration, from_station_name, to_station_name, usertype, gender* and *birthday* to analyse. I created a new data frame by doing 'df[['tripduration'. 'from_station_name', 'to_station_name', 'usertype', 'gender', 'birthday']]*'.*

Data Exploration:

1. I obtained the numbers of each user type using ‘value_counts()’ function, which showed that there are 1234155 people registered as subscriber, 479 as customer and 4 as dependent. The percentage of subscriber is 99.9%, which outweighs than the other two user .I used the ‘mean()’ to calculate the average usage time for each user type, and it shows that the average time used from subscriber is 699.7 seconds, from customer is 865 seconds, from dependent is 454 second. Besides, I visualised the average usage time for each user type using ‘plt.bar()’, ‘plt.xlabel’, ‘plt.ylabel’.

2. I used the ‘value_counts()’ method to get the number of male and female users. There are 935864 male users and 298784 users. For better understanding, I used pie chat to present the difference. Cause pie chart would show us the percentage of these two gender. I used the function ‘plt.pie(x, labels, autopct, explode)’ function to display the data clearly. After visualisation, I know that 75.8% users are male and 24.2% users are female.

3. I transformed birth year data into age data by using ‘age=abs(df_new[‘birthyear’] - 2017)’, then I found out there are incorrect data among the birthyear category. I used ‘min()’ and ‘max()’ to get the youngest is 1 years old and the oldest is 118 years old, there must be some errors in some rows. The Divvy bike age policy is limited age from 16 and above, and I assume that the people older than 80 are unable to ride bike. Thus, I used ‘df.drop()’ to delete the birthyear before 1937 and after 2001, resulting in a correct range of 16 to 79 years old. Then, I used ‘plt.hist(x, bins, edgecolor)’ to visualise the age data with *bins = range(16, 80, 10)*, which

means the range from 16 years old to 80 years old with 10 years old intervals, setting ‘edgecolor = ‘black’’ for better observation.

The result turned out the majority of age is between 26 to 36 years old, after the age range is 36 to 46, 46 to 56, 16 to 26, 56 to 66 and 66 to 80 years old.

4. I want to know the relationship between gender and trip duration. First, I added the ‘age ’ column to the data frame using ‘df_new[‘age’] = age’. I used Seaborn scatterplot to see the relationship between gender and age. I used ‘sns.scatterplo t(data, x, y)’ to visualise the relationship. In the first graph, I showed the relationship without separating genders. The *Relationship between Age & Trip Duration* graph shows that people from all ages used the bike around 4000 secon ds. I used the function ‘df_new[‘tripduraion’].describe()’ and ‘df_new[‘tripduratio n’].mode()’ to have accurate distribution. The mean of trip duration is 699.6 seconds, the standard deviation is 1028.6 seconds, the min is 60 seconds, the first quartile is 350 seconds, the second quartile is 565 seconds, the third quartile is 898 seconds and the max is 8609.6 seconds. The mode of trip duration is 332 seconds. The other thing we can tell is that the aged 20~30 tend to spend more time on bike rides.

   On the second figure, I tried to show the relationship between age and two genders. I used ‘plt.subplots()’ function to enable me to draw two graphs in a figure. Then, I also used the scatterplot to demonstrate the relationship. From these two scatterplot, I can know the two fact. First, There are more senior male users than female users. Second, younger age between 20~40 male users had had spent more time than same age of female users.

I also used the function *'*describe()*'* and 'mode()' to acquire details. The mean of trip duration in male users is 672.9 seconds, the standard deviation is 989 seconds, the minimum is 60 seconds, the first quartile is 338 seconds, the second quartile is 541 seconds, the third quartile is 861 second and the mode is 332 seconds.

The mean of trip duration in female users is 783.6 seconds, the standard deviation is 1140 seconds, the minimum is 60 seconds, the first quartile is 399 seconds, the second quartile is 648 seconds, the third quartile is 1010 second and the mode is 342 seconds. The statistics shows that female users had had spent more time on bike than male users.

5. I wanted to know which stations were very busy. So, I used 'describe()' function to analyse the *'from_station_name'* and *'to_station_name'* columns. The results shows that there 577 different stations, with Clinton St & Washington Blvd being the busiest station, people took bike from Clinton St & Washington Blvd up to 21752 times; as for the terminal station, most people would leave their bikes at Clinton St & Washington Blvd. So, Clinton St & Washington Blvd is busiest station both in taking bikes or leaving bikes.

   I visualized the *'from_station_name'* and '*to_sstation_name* dataset and tried to find out the most important five stations. I used 'sns.barplot(data, x, y)' function to show my finding.

   From the bar graph, Clinton St & Washington Blvd, Clinton St & Madison St, Canal St & Adams St, Kingsbury St & Kinzie St are the top five important stations; besides, the number of taking bike is 21752, 18317, 17393, 13973, 13928 times for the stations mentioned above. There is same pattern for end stations, Clinton St &

Washington Blvd, Clinton St & Madison St, Canal St & Adams St, Kingsbury St & Kinzie St are still the most popular stations, the number of leaving bike is 20532, 20317, 16754, 13848, 13393 times.

Critical Thinking:

My assumption is that subscriber is the people who need to commute between home and office daily; customer is the people who need the bike ride occasionally, this group of people have multiple choices of transportation; dependent is the people who need bike ride for emergency.

1. Based on Observation 1 and 4, we can tell that most of user are subscriber, which says that the combination of user is really heathy. Subscriber membership ensures a long-term relationship with clients; However, there is room to improve profit. Subscriber used bike around 700 seconds and the mode for each gender is around 340 seconds, the pricing policy for subscriber is 45 minutes for free, the operator can decrease using time for free to increase the profit without having bad influence on the majority.

   Besides, the prices are $108/year, $3.3/trip and $15/day respectively, we can know that subscriber spend 1 dollar for average 78 seconds(700/(108/12)), and customer spend 1 dollar for average 58 seconds(865/15). I think the operator can increase the annual membership fee slightly to gain more profit.

2. Based on Observation 2, with 75.8% users being male and only 24.2% female, there is an opportunity to attract more female users by providing safety equipment or insurance.

3. Based on Observation 3, the users of age range between 16~26 are much less than age range from 26~36. The reason I think is because the majority of the 16~26 age range is student. They can't afford 108 dollars membership per year. So, maybe the operator can have discount for student.

4. Based on Observation 4, we can tell that the younger users (20~40) tend to occupy the bike for more time, besides there are many outliers who used the bike for more than 20000 seconds. The situation is much obvious when the users are male. In order to cut down the maintenance expense of bike, I suggest the operator can set different price policy when facing extreme use. For example, if you occupy the bike more than six hours, the fee will be rise from 0.17 to 0.34 per minutes.

5. Based on Observation 5, the busiest stations are Clinton St & Washington Blvd, Clinton St & Madison St, Canal St & Adams St, Kingsbury St & Kinzie St. I think the operator can place more bikes at these location to manage rush hours Besides, I suggest the operator can create an appointment system to make sure the users who are in rush to get the bikes on time. Users would need to pay for an appointment, the company can increase the profit by doing so, and the people can get their bikes even in very busy stations.