Data Science Challenge

Applicant name: Monica Ethayananth
Email: monica160694@gmail.com

Source Data

The NYC Taxi and Limousine Commission (TLC) publicly shares data on trips taken by Yellow taxis, Green taxis, and "For-Hire-Vehicles" (FHVs). Please focus on trips taken by Yellow taxis in June 2017.

You may find the Yellow taxi data dictionary and taxi zone lookup table helpful.

Exercise

Imagine that you decide to drive a taxi for 10 hours each week to earn a little extra money. Explain how you would approach maximizing your income as a taxi driver.

If you could enrich the dataset, what would you add? Is there anything in the dataset that you don't find especially useful?

Solution:

The main objective is to make some extra money by driving 10 hours per week.

Step 1:

Using the data provided, identify the peak times of the day there are greatest number of commuters using the yellow taxi cab.

Using the drop off and pick time ('tpep_pickup_datetime' and 'tpep_dropoff_datetime') we can identify the hour of the day.

The more the demand i.e. more people commuting using taxi the possibility to get more rides increases. Plus, we can add a peak hour extra price.

After identifying the hours. Will identify the distance of the trips at different hours. The longer distance will yield more pay. Since the taximeter will run longer. But we don't want to drive very far off since the chances of getting a ride on the return way may be harder(which will be a waste of time and resource).

The figure 1 is the plot between the hours and the mean and median distance at different time.

The jupyter notebook used to derive the below the graphs/ figures is attached along with this document.

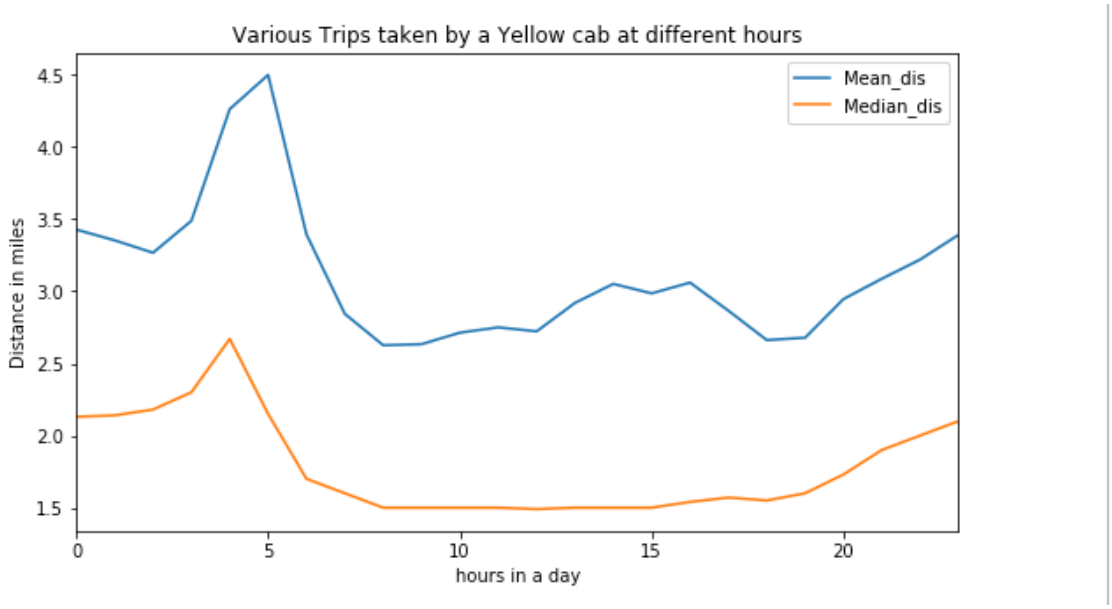


Figure 1

From the above graph we can infer that most of the trips happen at early morning hours and evening hours.

Step 2:

Trips that happen to and from the airports with a RatecodeID 2 and 3

From the lookup table we have identified those places as 2: JFK and 3: Newark.

Trips to the airport are quite profitable with an average total amount of $67 dollars.

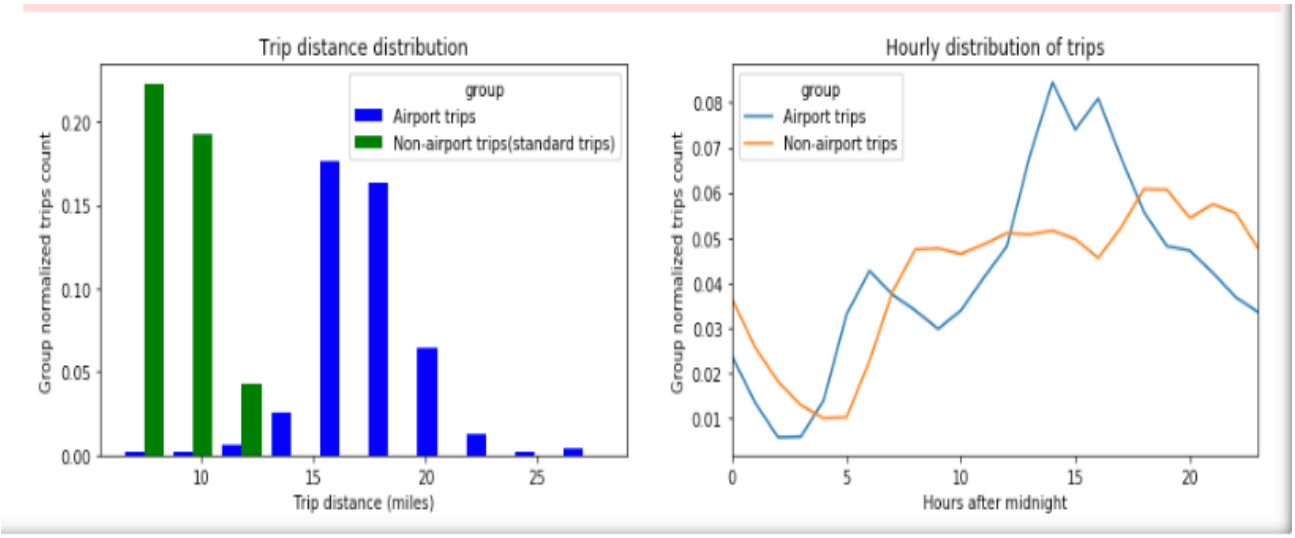
```
In [12]: #let find out the average trips from or to the airport
         #from the dictionary lookup table we know the  rate code id of value of 1 is standard 2 is JFK
         #lets find the number of from the airpoprts.
         airport_trips = data[(data.RatecodeID ==2) | (data.RatecodeID==3)]
         shape1 = airport_trips.shape[0]
         print("Shape of the airport trips: ",shape1)
         print("Average Fare of trips calculated by the taximater: $",airport_trips.fare_amount.mean())
         print("Average Total: $",airport_trips.total_amount.mean())

         Shape of the airport trips:  243347
         Average Fare of trips calculated by the taximater: $ 53.20434050964261
         Average Total: $ 67.40676881161467
```

Figure 2

Identify the number of trips to airport and the distance of the trips and other trips.

We do know that there are more standard trips than airport trips from this figure 3 below



We also know that standard trips during the day time have a average steady growth in availability unlike airport trips.

We also know that the standard trips are shorter and quicker than long trips to the airport.

Step 3:

Identify the location of pick up and drop of where most rides originate and ends

```
In [40]: data['PULocationID'].value_counts().head(10)

Out[40]: 237    379518
         161    358793
         236    343195
         162    331402
         186    328348
         170    315156
         230    309831
         234    307904
         48     301831
         142    275837
         Name: PULocationID, dtype: int64
```

```
In [39]: data['DOLocationID'].value_counts().head()

Out[39]: 161    364035
         236    350582
         237    332505
         170    310753
         230    298525
         Name: DOLocationID, dtype: int64
```

From the data provided in taxi+_zone_lookup table the locations are identified as follows

4	5	Staten Island	Arden Heights	Boro Zone

```
In [62]: a1.LocationID ==161) | (data1.LocationID == 236)|(data1.LocationID==237)|(data1.LocationID==170)
```

Out[62]:

	LocationID	Borough	Zone	service_zone
160	161	Manhattan	Midtown Center	Yellow Zone
161	162	Manhattan	Midtown East	Yellow Zone
169	170	Manhattan	Murray Hill	Yellow Zone
185	186	Manhattan	Penn Station/Madison Sq West	Yellow Zone
229	230	Manhattan	Times Sq/Theatre District	Yellow Zone
235	236	Manhattan	Upper East Side North	Yellow Zone
236	237	Manhattan	Upper East Side South	Yellow Zone

From the above data obtained we now know that most of the rides are located in the Manhattan region. Most of the rides end at Manhattan are located at Midtown Center and originate at upper east side south and north. All these rides have a small percentage of tips.

We also have additional information that on an average the short trips yield considerable tips.

Conclusion:

To make a major profit driving 10 hour a week can be achieved by driving at **early morning hours or late evening hours** and concentrate the rides to **the Manhattan midtown center and upper east side area.** These zones have more demand and hence the chance to obtain riders is high which reduces the taxi drivers wait time. And doing a bunch of short trips will yield a profit from tips. The morning hours marked by the high demand will yield us an **additional fare for peak hours.** Once in a while evening airport tips can be beneficial as well since the average fare for an airport trip amount is quite high.

Data Relevance and enrichment

The columns ‘VendorI’D and ‘Store_and_fwd_flag’ is not really needed to identify the best ways to make more money. These columns come in handy for other process like identify how often taxi driver lose connection and who provides the information this data. It is not helpful for the current scenario.

The information of number of driver active at a given time will be helpful in estimating supply of drivers at a given time offering rides.

The jupyter notebook called final.ipynb is attached along with this document