FULLSTORY

Data Science Challenge

Submitted By:

Swapnil Vijay

Swapnilvijay9@gmail.com

Mob: +1 706-461-6444

Problem Statement:

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?

Approach:

To start the problem, I would go by below stated points to make a conclusion:

1) Check the data size, number of rows/columns
2) Check the basic statistics of each variable; min/max/avg/count
3) Remove any unwanted/invalid data based on the step 2 or based on the data dictionary
4) Feature Engineering to help in further analysis
5) Data Exploration to identify the trends/hidden insights from the dataset
6) Make suggestions

Step 1:

- Dataset provided is of Yellow Taxi trip for 1 month - June 2017
- Number of Rows: 9656993 and Number of Columns: 17
- Available Columns: 'VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'passenger_count', 'trip_distance', 'RatecodeID', 'store_and_fwd_flag', 'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge', 'total_amount'

Step 2:

- See the Code for detail analysis
- Analysis from this step:
    - Vendor ID does not give important information related to the problem statement
    - Min and Max passenger counts are 0 and 9 respectively
    - Min trip distance is 0; we need to remove the records having trip distance = 0
    - Max Rate Code ID is 99 which is not in the data dictionary. We need to check the other invalid values present in the Rate
    - Code ID.
    - There are some negative fare amount; we need to remove the records having fare amount < 0

- There are some negative extra charges; we need to remove the records having invalid extra charges.
- There are some negative MTA_tax charges; we need to remove the records having invalid MTA_tax charges.
- There are some negative tip amount; we need to remove the records having invalid tip amount.
- There are some negative tolls amount; we need to remove the records having invalid tolls amount.
- There are some invalid values for improvement surcharges; we need to remove the records having invalid improvement
- surcharges.
- There are some negative values for total amount; we need to remove the records having invalid total amounts.

Step 3:

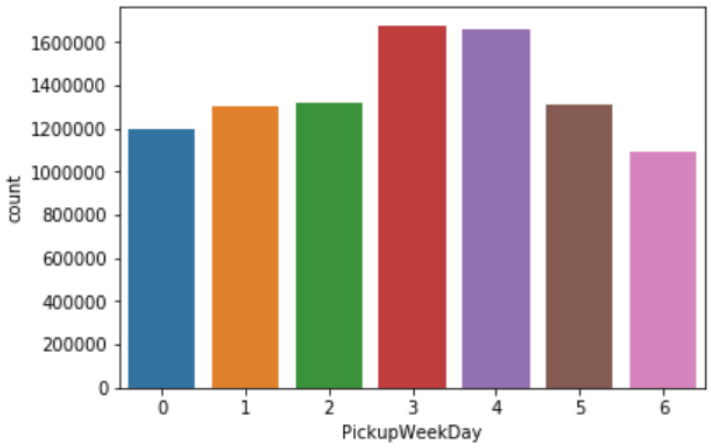
- Observations made from above step, dataset is cleaned by removing invalid dataset and outliers:
  - Passenger Count != 0
  - Trip Distance > 0
  - RateCode ID should be between [1, 6]
  - Fare Amount >= 0
  - Extra should be (0, .5, 1) based on the data dictionary
  - MTA Tax should be ($0.00, $0.5) based on the data dictionary
  - Tip Amount >= 0
  - Tolls Amount >= 0
  - Total Amount >= 0

Step 4:

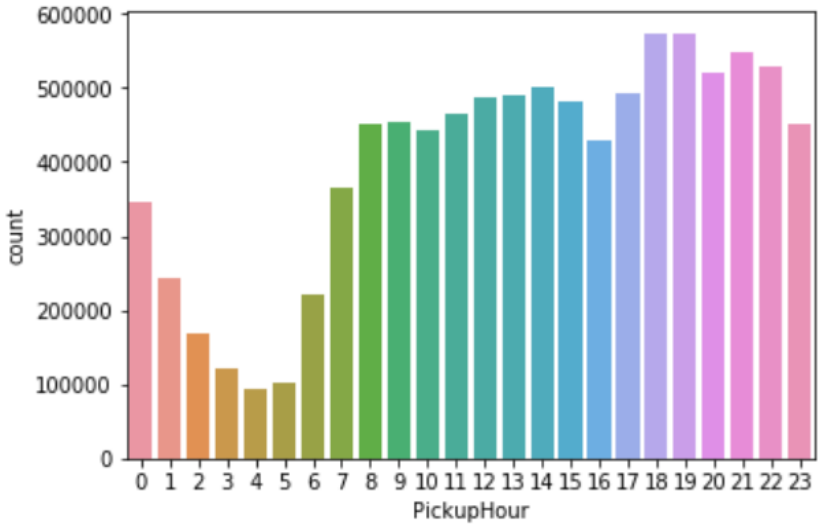
- PickUp Week Day and Drop Off Week Day are created to get the most busy days in a week
- Pickup Hour and Drop Off hour are created to get the frequency distribution of trips across the full 24 hours

Step 5:

- Number of Trips by Weekday; Thursday and Friday have the most numbers of trips out of 7 days



- Number of Trips by Pickup Hour; Hour 18, 19, 21, 22 and 20 have the most number of trips in the day



- Top 5 pickup locations based on number of trips: 237, 161, 236, 162 and 186

	PULocationID
PULocationID	
237	378064
161	356561
236	341940
162	329360
186	326718

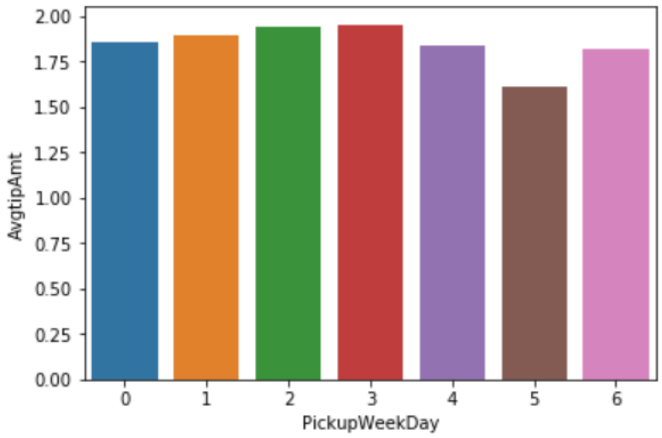
- Top 10 routes w.r.t the number of trips are (PickUpID_DrioOffID) - 264_264, 237_236, 236_237, 236_236, 237_237, 239_142, 239_238, 237_162,237_161, 142_239

Route	Route
264_264	107208
237_236	52658
236_237	44214
236_236	42845
237_237	40004
239_142	25251
239_238	24962
237_162	24869
237_161	24220
142_239	23133

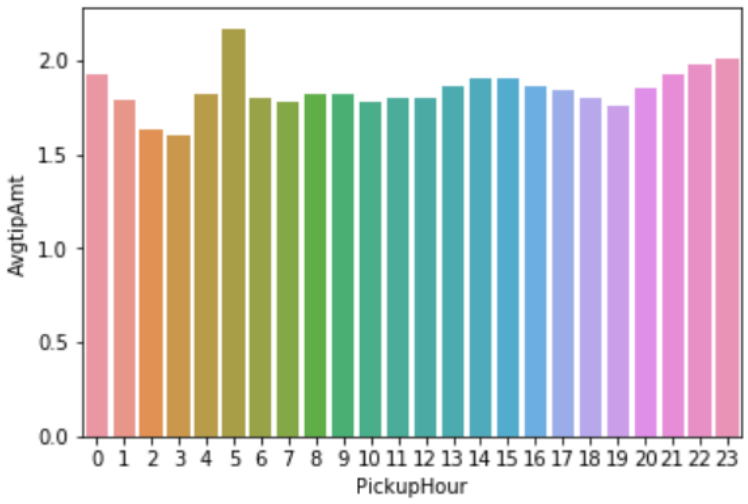
- Top 10 routes w.r.t the tip amount are 117_223, 23_113, 242_168, 201_234, 145_12, 187_23, 113_5, 43_86, 239_39, 67_1; But the number of trips on these routes are very minimal (1 or 2)

Route	Route	tip_amount
117_223	1	200.000000
23_113	1	120.000000
242_168	1	120.000000
201_234	1	92.000000
145_12	3	83.333333
187_23	1	83.000000
113_5	1	75.880000
43_86	1	75.240000
239_39	1	70.840000
67_1	1	70.000000

- Wednesday and Thursday have the highest Avg Tip Amount



- Hour 5, 21, 22, 23, and 0 have the highest Avg Tip Amount



Step 6:

- Coming to our problem statement, if a driver wants to earn maximum while driving 10 hours per week, we can suggest multiple options/suggestions:
    - Choose to drive on Wednesday, Thursday and Friday based on the number of Trips and Tip amount; If the driver drives on these days, the idle time for the driver will be less as there are more customers on these days as well as they give more tip based on the overall average.
    - Choose to drive in certain hours of the day eg. Between 10 PM – 1 AM in the morning based on the number of customers during that time of the days and the tip amount
    - Choose to drive on certain routes based on the number of customers, tip amount, relative distance etc.
        - Routes - 264_264, 237_236, 236_237, 236_236, 237_237, 239_142 are the busiest routes based on the number of trips available
        - Routes - 117_223, 23_113, 242_168, 201_234, 145_12 have the most tip amount but very few trips (can also be treated as outliers as there are very few numbers of trips between these routes)
        - Routes - 264_264, 237_236, 236_237, 236_236, 237_237, 239_142 have a decent tip (more than 10% of total tip amount) as well as number of trips (more than 25000 trips)

Conclusions:

- Can make more sophisticated suggestions based on time spent on the problem, new features, combination of different factors, analysis based on trip duration, total amount and tip amount, Taxi Vendor
- Can enrich the results by including more data from other months and year, demographics of the area, income level, busy hours, latitude and longitude of the PickUp and DropOff locations

- I think we can deduce some or other insights from the variables give in the dataset. Can create new features like date, time, weekday, weekend, trip duration and then discard the Pickup and Drop Off date time variables.