PROJECT 1 REPORT

TASK 1 AND 2

Analyze the data of the Green taxis for NY and NJ area and come with associations rules for each attributes

Brijesh Goharia

09.30.2019
Mining of large datasets

INTRODUCTION

I first reviewed the csv file on an excel sheet to get an idea on how the data is structured and brainstormed on different ideas and ways I could approach this project.

There were a bunch of ideas which I was very excited to implement and review the outcomes. Following are the insights that I wanted to find for inter-state, intra-state and entire dataset:

- Perform association analysis of multiple attributes with the pickup hour (i.e. of the day for the entire month. These multiple attributes include: Total count of rides, mean distance travelled, total revenue and average tips

- Analyze the relation between the RateCode ID and respective average fare amount

- Explore all the statistics related to the airport i.e. ride counts to airport, total passengers traveling to the airport, average passenger count to airport and comparison of JFK and Newark airport

- Find the total distance travelled in each day for all kind of trips

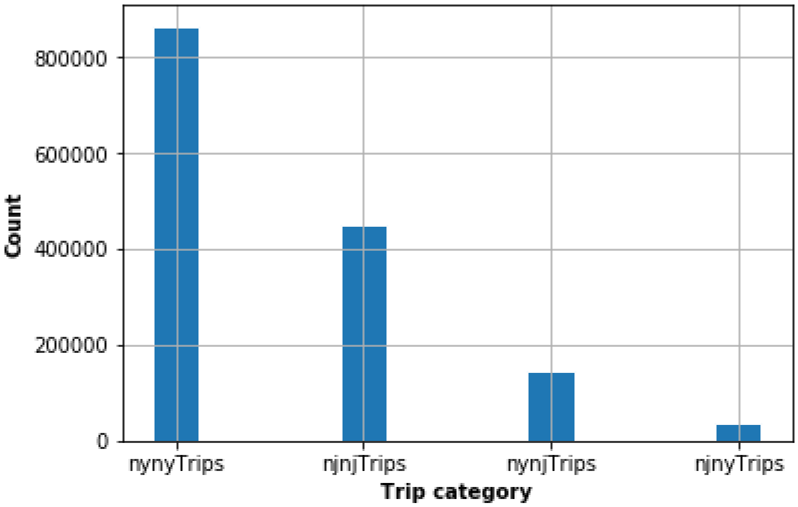
Model Implementation:

TASK 1: Analyzing NYC_GREEN_TAXI data using “MAP-REDUCE”

1.1 Creation of Intra and Inter Borough Pairs:-

First I started by grouping the data into four categories NJ-NJ, NY-NY, NY-NJ, NJ-NY using the map-reduction approach.

Inference: from fig (a) I conclude that maximum trips in the month of September 2015 is between inter-borough (NY-NY, NJ-NJ) than intra-borough (NY-NJ, NJ-NY)



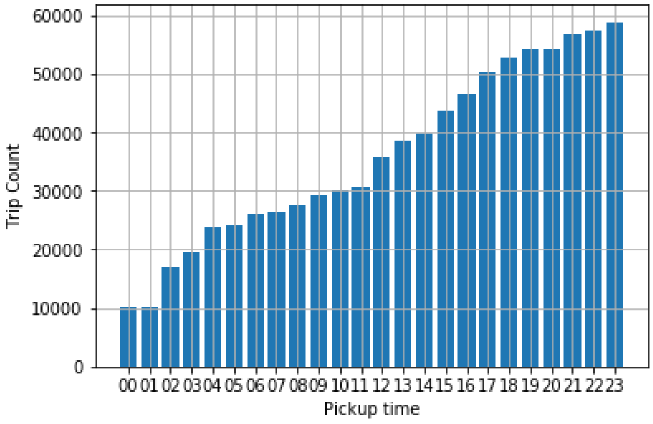
I then used these groups for the following analysis.

1.2 Perform association analysis of multiple attributes with the pickup hour:-

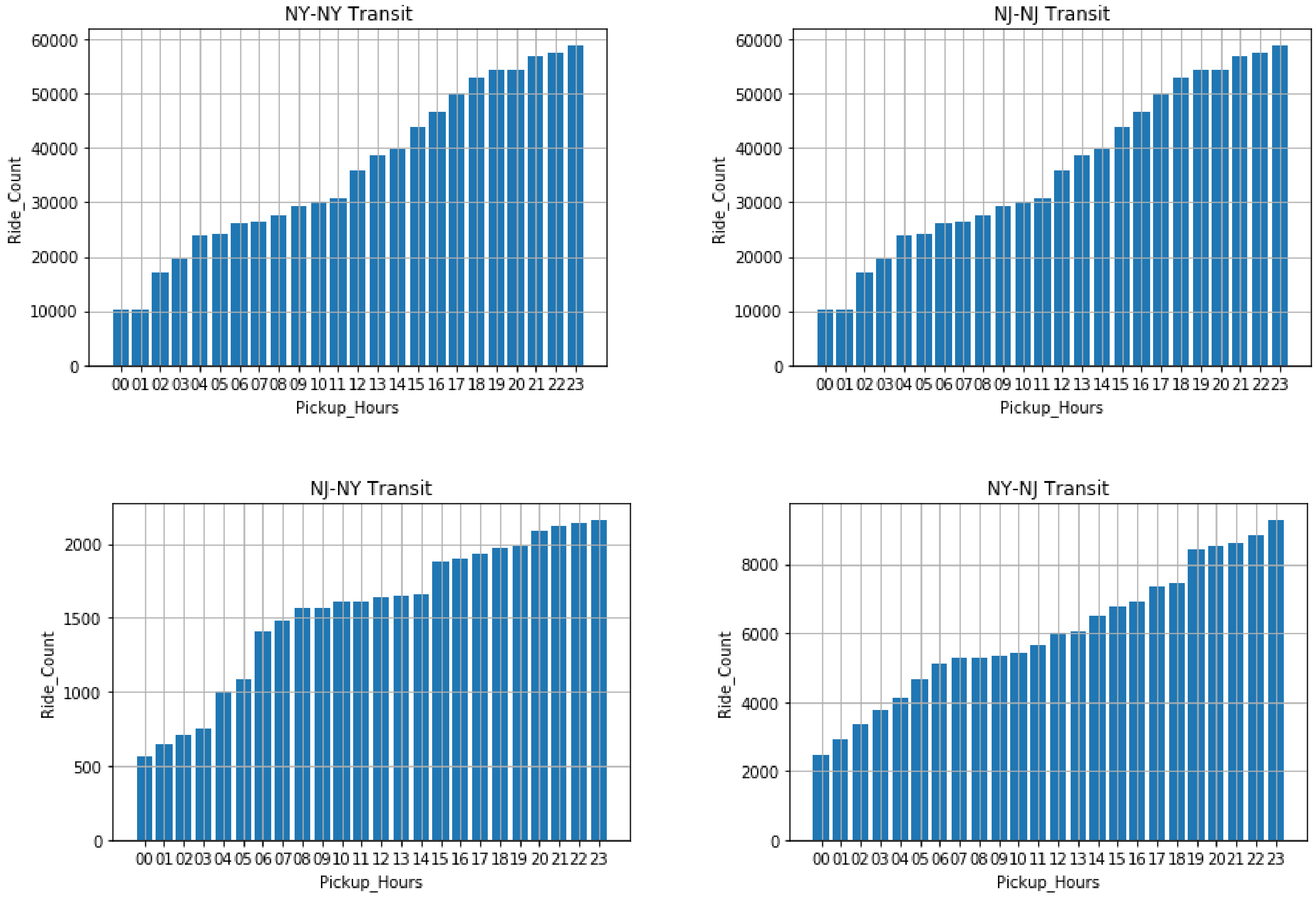
A) Pick-Up hours v/s Total Trips (fig. b-f)

In order to provide the taxi service about the demand for the cabs during the specific hours of the day I grouped the total rides according to pickup hours and deduced the prime hours of the day.

Inference: This insight can help drivers to choose their time of service accordingly

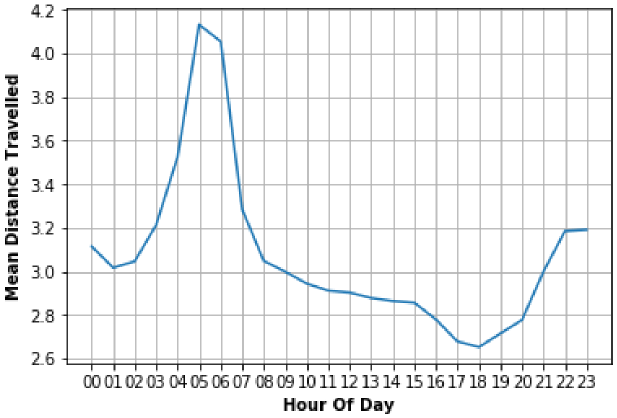


Further I went ahead and plotted the same graphs on the four categories as well to find maximum and minimum trips occuring inter and intra Borough during specific hours of the day.



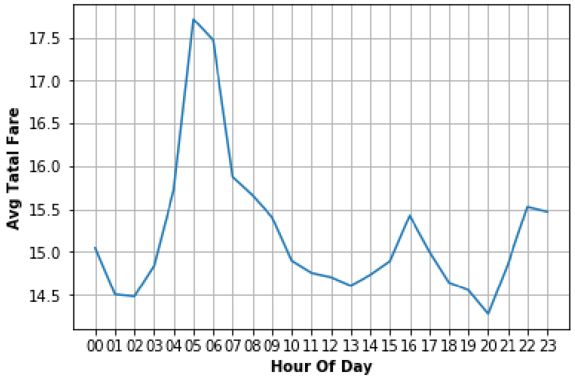
B) Pick-Up hours v/s Mean distance travelled:

Inference: This plot will help the drivers who wish to go for comparatively longer trips to choose the time (4:00 pm - 7:00 pm) of their service accordingly.



C) Pick-Up hours v/s Average Revenue:

Inference: The above plot gave us the further curiosity to relate mean distance traveled and pick-up hours with the Average revenue earned during those hours. The graph appropriately validates our assumption that Average Revenue is higher during 4:00 pm - 7:00 pm.



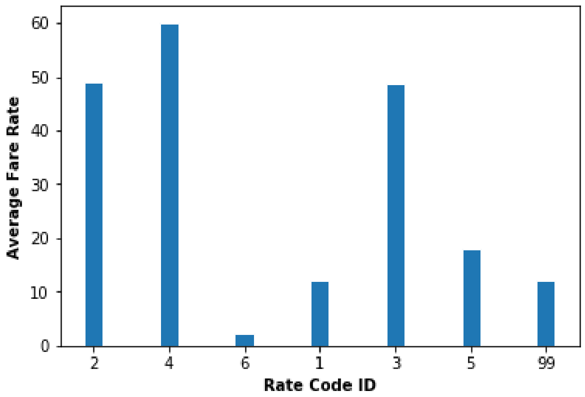
D) Pick-Up hours v/s Average Tips:

Inference: This plot helps the drivers and the cab service to analyse the average tips the drivers can expect while driving during specific time of the day. To add more granularity on the tip amount I performed it on all four groups and concluded as follows:

Groups	Pick-up Time	Average Tip
NY - NJ	07:00 - 11:00	$ 2.80 - $ 3.20
NY - NY	05:00 - 08:00	$ 1.20 - $ 2.00
NJ - NJ	07:00 - 10:00	$ 0.95 - $ 1.05
NJ - NY	04: 30 - 07:00	$ 2.50 - $ 3.50

1.3 Relation between the RateCode ID and their respective Average Fare:-

Inference: Average Fare Amount for RateCode ID 2(JFK), 3(Newark), 4(Nassau or Westchester) is way more than other RateCode ID consisting of Standard rate, Negotiated fare, Group Ride.



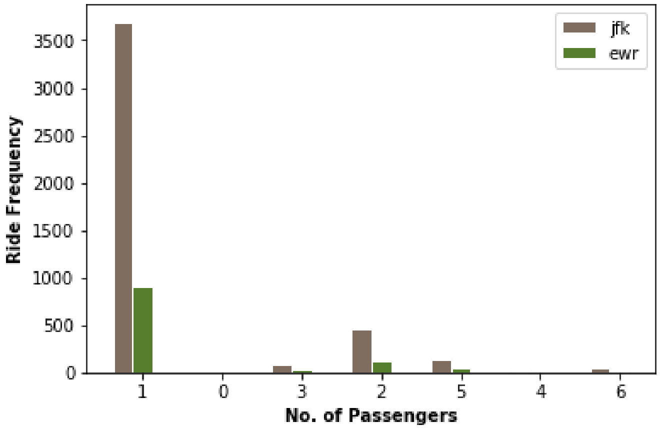
1.4 Airport Statistics:

Here I plotted a histogram for comparison of “Trip Counts” by “Number of Passengers” in the RateCodeId (2=JFK Airport, 3=EWR Airport)

Inference:

-Help green nyc taxi service to gain insights whether to opt for pooled cabs for the specific time of the day

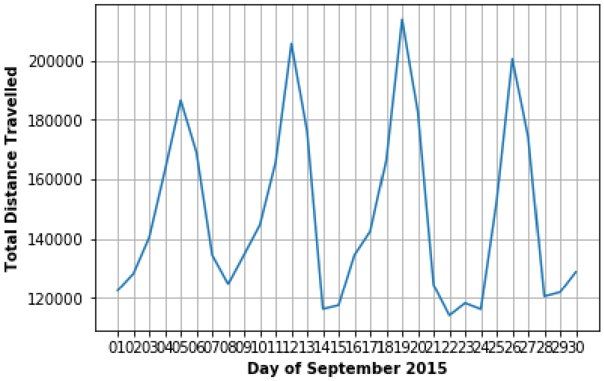
-From No. of Trips statistics as well as No. of Passenger Count Statistics, I can conclude that JFK Airport is more busy compared to EWR(Newark) Airport



1.5 Total distance travelled in each day for all kind of trips:

Here I plotted the days of the month to find the maximum distance travelled. This insight further helps the drivers in addition to our aforementioned hourly busy slots to choose the day of their choice to do the business as well.

inference : Total distance travelled is minimal on Mondays, Tuesdays of all weeks in the month of september (all trips). This can help the drivers to take a week-off with the minimum loss in the revenue.



1.6 Intra-Borough Relationship (NY-NY trips):

Further we also calculated the number of trips between UpTown and DownTown and vice versa as follows: Total Ny to Ny Trips: 865242.

Trips from Downtown to Uptown: 466550 ; Trips from Uptown to Downtown: 398692

Further Scope for Analysis: I can determine the number of passengers travelling in either direction during the specific hour of the day. This can help the drivers to position themselves within the hot-zones.

Task 2: Build Data-Frames from the given data set and apply similar inferences from Task 1.

1. Comparison between Data-Frames and Map Reduce -

- In simple Map & Reduce Method, each element in RDDs is initially a unicode. Hence, while performing different tasks based on attributes, assigning Data-Type to different attributes was time consuming as it required more steps to pre-process the data. On the other hand, while implementing Data Frame, once done with the data schema, it becomes comparatively easier to perform various tasks with less steps as the data needs minimal preprocessing and sql queries can be performed with ease
- Data Frames are well structured hence its easier to visualize the data and also structured data requires less time to process

Therefore, Data Frames is easier to use compared to map & Reduce Method

2. Reflect on the homeworks/labs I did So far.Which one would be a good candidate to use Data-Frames?

- FP-Plants
- FP-Crime
- FP-Ingredients