MAST30034: Applied Data Science

Assignment 1

Introduction

The aim of this project is to gain an initial insight into the New York City Taxi and Limousine Service Trip Record Data. The aim will be achieved through performing an initial analysis,

along with a visualization of the results.

In this project, by analyzing the datasets, we intent to understand the problem of” What will affect taxi drivers’ profit?”. We divide the problem into two aspects.

Aspect 1: What will increase the demand of taxi in New York?

Aspect 2: What will bring taxi drivers more tips in New York?

## Data and Attribute Selection

### Data period

The data we focus on is the Yellow Taxis data in September 2018.

As mentioned above, for aspect 1 we would like to discuss if the sports event and weather will increase the demand of taxi.

The US tennis Open is one of the Major Tennis Tournaments around the world which usually begins from the end of August to the beginning of September Annually.

In 2018, the tournament was from 27 Aug 2018 to 9 Sep 2018, we decided to specifically look at the last two days of the event (8th Sep and 9th Sep) when the Men's and Women’s Singles were held.

Further, we chose the yellow taxis to observe because they are the only vehicles permitted to respond to a street hail from a passenger in all five boroughs.

### Attributes Selected

For aspect 1 we would like to know if the sports event affects the demand of taxi in New York. Thus, we need to find out if there is an increasing of trips which have the drop-off location around the sports venue.

For aspect 2 we list several potential causes that may influence the tips amount: special date (include the sport events, holiday, weekend), trip distance, number of passengers and time.

The final attributes we used are listed below:

* tpep\_pickup\_datetime: To discovery the total number of trips for different time periods.
* DOLocationID: To discovery if there is an increasing around a specific area.
* Tip\_amount: To discovery the factors that influence the tip amount.
* Passenger\_count: To discovery if there are any relationship between number of passenger and tip amount.
* Trip\_distance: To discovery if there are any relationship between trip distance and tip amount
* Fare\_amount: To discovery if there are any relationship between fare amount and tip amount

## Pre-processing and Cleansing

### Missing data

First of all, we give an overview of our data set. We have 8040133 instances and 17 columns in our dataset, then we check if there are any missing values. Fortunately, we didn’t detect any missing values, but there might be any disguised missing values we need to beware of.

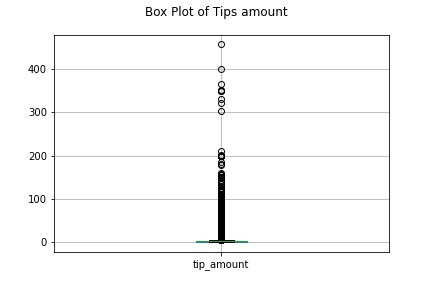
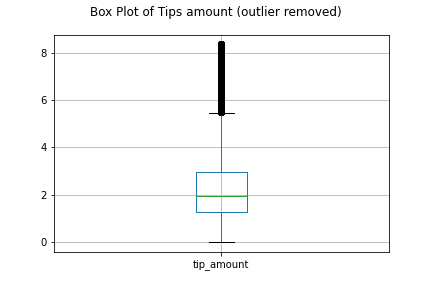
### data cleansing steps

Then we observed some noisy data which are invalid and unreasonable and we aim to clean them.

* Tpep\_pickup\_datetime & Tpep\_dropoff\_datetime: A few data were recorded on 2001/2002/2008/2019. We remove all the invalid date data.
* Passenger\_count: A few data show the passenger count is 0, the taxi driver might mis record the data. Therefore, we remove 0 passenger trip data.
* Trip\_distance: A few data show the trip distance is 0 but the total amount is not 0, that is not explainable, we removed all 0 trip distance data
* Payment\_type: We would like to analysis the tips; thus, we only need the data paid with credit card. Remove all the data that payment type is not 2.
* Fare\_amount: A few data have negative values on the fare amount, the fare amount should have 2.5 as the initial charge, thus, we remove all the data which has the fare amount less than 2.5.
* Total\_amount: A few data show negative total amount which are not reasonable, the total amount should be the sum of all the fees, therefore we remove the data which has the total amount less than 2.5 since the initial charge for fare amount is 2.5 and the total amount can not be less than the fare amount.

### Outliers

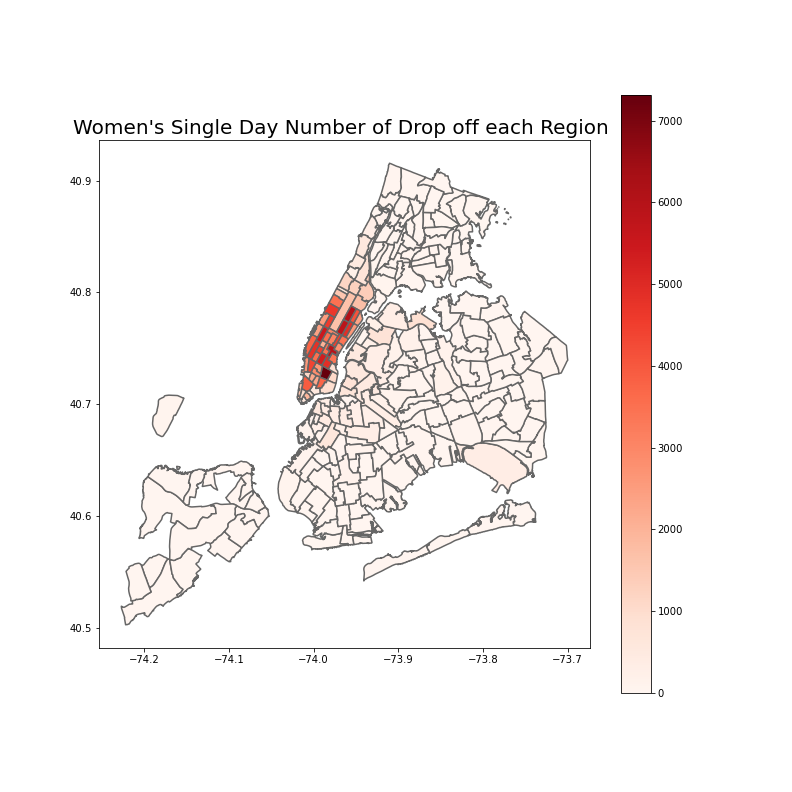
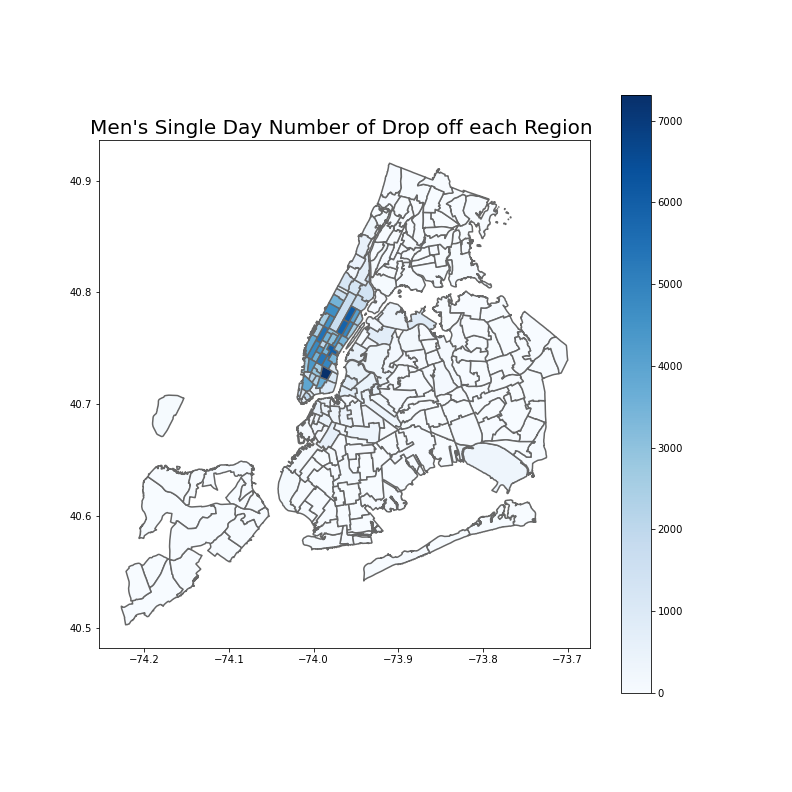
Through the above observations, we found that there might be some outliers in our data.

We visualized the tips amount by using box plot, we could see there are few passengers gave an excessive number of tips (over $300). We use the Interquartile Rule to detect any outliers.

### pre-processing and/or feature engineering steps

For further visualization we need to count the total number of trips by day and time, we did feature engineering and add two new columns which only include date and time(by hours) from “tpep\_pickup\_datetime”.

## Visualisation



We inspected both green cars and yellow cars data

In this aspect, we consider if the date, weather, time or trip distance may influence the tips amount.

Since we

Data and Attribute Selection

The sporting event we chose is the 2018 US Open (tennis) which usual

(4 marks)

Clearly states data period (1m)

Clearly states the three (or more) attributes to be analysed (1m)

Convincing justification for data period (1m)

Convincing justification for three (or more) attributes to be analysed (1m)

Pre-processing and Cleansing (4 marks)

Clearly states pre-processing and/or feature engineering steps (1m)

Clearly states data cleansing steps (1m)

Adequately investigates data for possible anomalies/outliers (1m)

Appropriate justification for pre-processing steps, as well as steps for handling missing data (1m)

Visualisation: Quality/Clarity (6 marks)

No marks possible without geospatial visualisation Geospatial visualisation is present (i.e heatmap, choropleth) (1m)

Appropriate granularity. Is it easily understandable to see what the visualisation is trying to show? Are there too many data points? (1m)

Geospatial visualisation clearly expresses a story, particularly if it raises “interesting” areas of further analysis or indicates an area that does not need further analysis (1m)

Appropriate choice of dimension, colour scheme, legend and formatting (1m)

Appropriate explanation of what the visualisation shows without being overly verbose (2m)

Analysis of result(s) (3 marks)

An appropriate summary statistic to describe the chosen attributes (1m)

Appropriate analysis of the relationship between the attributes (2m)

Quality and clarity of report (3 marks)

Quality and clarity of report (3 marks)

Quality writing, spell-checked, correct grammar, and comprehensible sentence structures (1m) Identifies potential stakeholders, motivation for the report and real-life use cases (1m) Provides recommendations for potential stakeholders based on analysis of findings (1m)

<https://www.wunderground.com/calendar/us/nj/newark/KEWR/date/2018-9>