Applied Data Science (MAST30034)

Project 1: Quantitative Analysis

HUATING JI
Student ID: 1078362

August 15, 2021

1 Introduction

Some useful Latex tips (refer to the code panel):

“ and ” should be used instead of ”. For example: “test” vs ”test”.

Remember, your report should be between 5 to 8 pages and between 1500 to 2000 words.

The project records nearly 10 years of taxi and limousine traffic in New York City. They cover all types of New York City taxis: yellow cars, green cars, and FHV Trip Records, which has just emerged in recent years. Faced with so much data, we need to choose information properly, process information carefully and analyze information carefully. One thing we should note is that since 2018, the way of recording the geographical location of vehicles has changed, from the original longitude and latitude to the area code of each region. When faced with this kind of information, we need to convert the previous longitude and latitude coordinates into area numbers. At the same time, we can also use some external data to analyze the information we already have and enhance the comprehensiveness of the data. For this report, it mainly processed and analyzed the data of yellow cars in the second half of 2018.

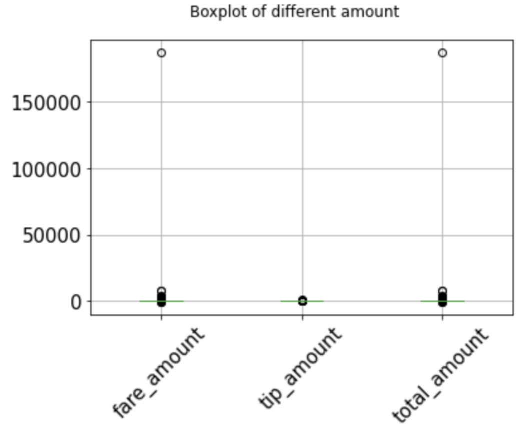
2 Preprocessing

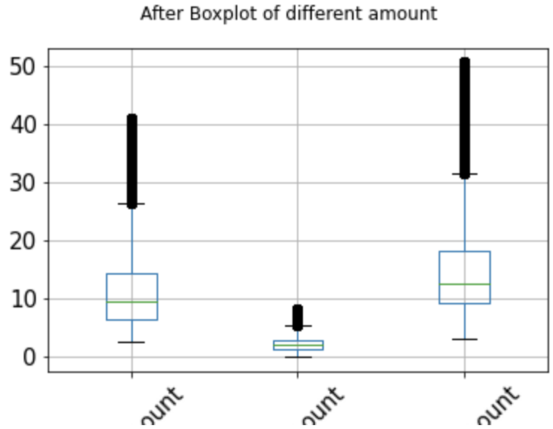
the general steps of preprocessing data

1. read the data dictionary

- choose the suitable data period
- remove the noisy data
- itemize remove the outliers.
- remove the useless attribute
- save the processed file.

In the aspect of pre-processing data, when we get the data, the first step is not to rush to process the data. After we have a general understanding of the model, we can open the Data Dictionary and start to check to understand the data type. So that we can process the data later. When preprocessing data, we can first eliminate some data with obvious errors. Then, using the image of boxplot, to remove outliers.



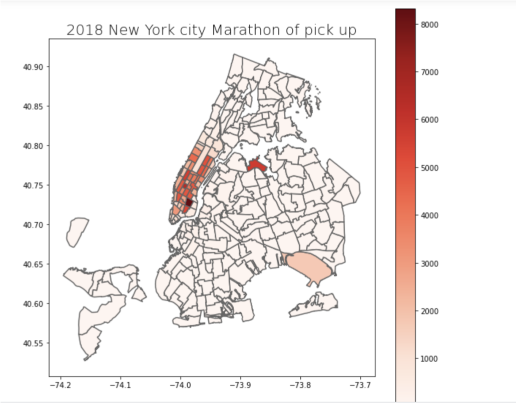


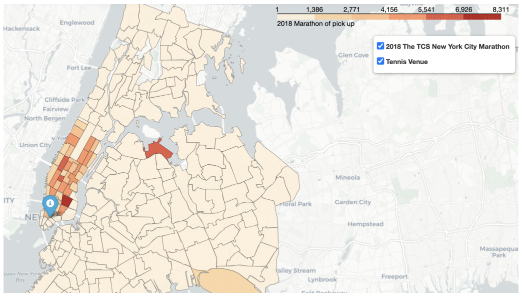
For example, there is a sharp contrast between the pre-processed boxplot and the processed boxplot. After processing outliers, the processed boxplot shows that the fare amount and total amount are mainly distributed around 10 dollars. And remove the useless attributes can save a lot of time. such as the payment type, we just need the payment of credit card, so we remove the other payment type. At the same time, we also can choose the way of feature engineering to combine the extra data. Some of them can train the better model.

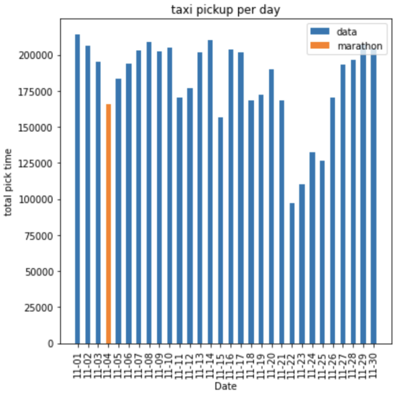
3 Preliminary Analysis

3.1 External Dataset 1:event

In addition to the information provided by the system, we can also find some external data to enrich our database. For example, the weather forecast can be used to study what kind of weather to choose the most people on the bus. Or if the taxi industry is affected by extreme cold in New York. In the same way, we can use New York's holidays and major events as an external data point to see if more people take taxis to get around during events and holidays. On November 4, 2018, for example, an annual marathon in New York City, we found from the image, from the overall population, fewer in taxi surrounding the marathon, cause the reason for this may be because of the particularity of the marathon, in order to protect the safety of athletes and spectators, the organizers may have to clean up the surrounding traffic. But compared to the number of normal weekend, number or similar, there may be people want tourist destination choice taking a taxi, but the number did not significantly increase, but compared to usual fell over the weekend, so we can be seen from the graph, although there is a marathon, someone may because can't insist on, choose to take a taxi on his way to watch attractions, But since it's a Sunday and most people stay home, major events can occasionally have a negative impact on the taxi industry.







3.2 External Dataset Visualisation2:weather

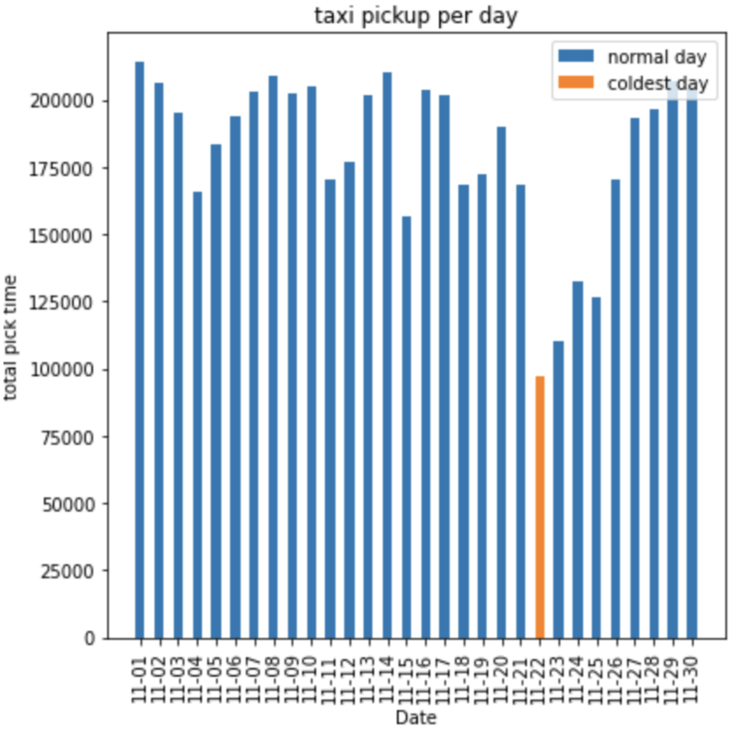
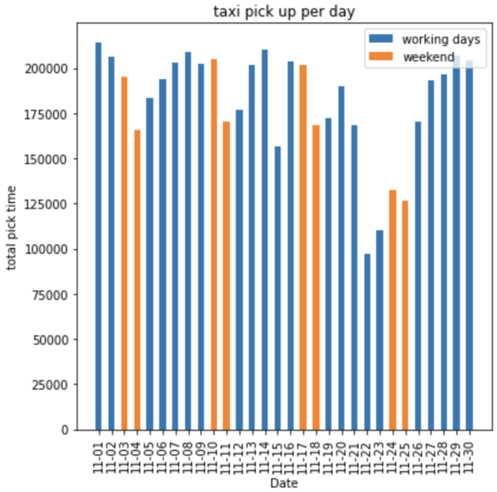


Figure 1: Some caption

Thanksgiving Day, November 22, 2018, is usually the busiest day of the month and a busy day for the taxi industry, but from the image, we can see that November 22 is the fewest day of the month for the taxi industry. It's a dip in the bar chart, which is very unusual. According to the weather forecast, on November 22, 2018, the temperature in Manhattan, New York dropped from 20 degrees to more than 10 degrees, making it the ”coldest Thanksgiving” since 1996. So fewer people travel that day, people stay at home, and drivers get fewer orders that day. Thus, we can preliminarily find that in extremely cold weather, the taxi industry will be affected by the weather. So that we can pay more attention to the data obtained from the outside and use it to analyze the internal data to find out when the taxi industry is best to make money.

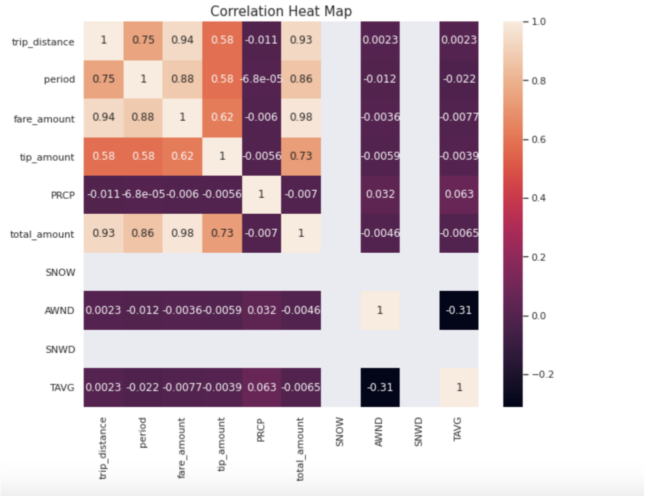
3.3 External Dataset 3:weekend



As can be seen from the bar chart, the blue line showed a clear upward trend, while the orange line showed a clear downward trend, especially on November 22, with a direct decrease of nearly 70,000 hours. This shows that on weekdays, people are busy going to work, so most of them choose to take a taxi, while on weekends, the number of people who choose to take a taxi drops sharply. It may be that on weekends, most families choose to drive private cars to go out for fun, so as to save money on taxi fares. But it can also be that the weekdays are just too hard, and most people choose to relax at home on the weekends, spend time with family, or have dinner with close friends. Therefore, most people do not choose to go out, which leads to a decrease in the number of people taking taxis. So to some extent, holidays will also affect the operation of taxis.

4 Statistical Modelling

4.1 Model



the Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables. The value of correlation can take any values from -1 to 1.Correlation between two variables can also be determined using scatter plot between these two variables. From the correlation heat map, we can find the correlation of total amount,tip amount and trip distance are close. so we choose Linear Regression Model in python. We convert the tip amount, trip distance, total distance to the DataFrame and split the train and test. we use 2019-11 as the test value.

4.2 Results

The results of the score is 1, this value is not reality. Because it is too close. I think that I do the wrong way.

4.3 Discussion

From my statistical model, i think i need you preprocessed the test value firstly, because we scales the train data. Although this statistical modelling is default, i think we can try another way in R. input the sample data in R to predict the tip amount, change the categoryl to the vector or use the continuous value to do the test. And then split value to the train and the test. Both of them are be scaled. Then take them to the glm model. Watch the model, if the model is not good, we can use AIC to continuous model it. Finally, to find the accuracy of model.

5 Recommendations

Through the understanding and research of the data recorded by taxi drivers, I would like to make some suggestions to the taxi drivers. As a taxi driver, you must be concerned about how to pick up more passengers, where are the hot spots and where can you make more money. First advice is to care about the weather every day, when the sudden bad weather, through the above research, we found that the passengers a lot less, is most people choose not to go out, but will have to go out forget to bring my umbrella emergency need a taxi, the hot spots is the supermarket, office building or shopping centre downstairs. Weather can affect the taxi industry, but it can also help the taxi industry. The second suggestion is that in the morning rush hour, the demand for taxis will increase, so you can leave early and go to the places with the most people in the morning rush hour, so that you can get the orders quickly, so that the income will increase. Thirdly, you can pay attention to the surrounding large concert, the general concert around the car is difficult to park. Many people will take taxis home after the concert. Concerts are the hot spot.These are my suggestions. I hope they will be helpful to you. Finally, I wish everyone happiness.

6 Conclusion

First, I used the driving records of yellow taxis in November 2019 as a sample to test how the income of taxi drivers would be affected. Then I used external data to find out that November 22, 2019 was the coldest Christmas Eve in recent decades, so I used the bar chart to observe the income of that month, and it was found that the lowest income day in November was November 22, so I preliminatively assumed that the driver's income was related to the weather. Then I found out that there was a major marathon in November. In order to explore the impact of the holding of a major marathon on the driver's income, I chose to compare the income of the day with that of the day with the bar chart, and found that the impact was small and could be ignored. So I combined the data of 2018, August, September and October to predict the relationship with weather, and finally found that it had a relatively large impact. However, after drawing the heatmap, I found that the driver's income was also related to tips and distance, so I chose linear Regression model to predict the model.

References

[1] “Taxi Fare.” Taxi Fare - TLC. Accessed September 7, 2019. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[2] Knuth: Computers and Typesetting, https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[3] Knuth: Computers and Typesetting, https://world.huanqiu.com/article/9CaKrnKf4Rs

[4] Knuth: Computers and Typesetting, https://www.nyrr.org/tcsnycmarathon/Race-Day/The-Start

[5]