Analysis of Pickups Data of NYC of Uber

Questions We Try to Answer

Statistics analysis of order data are widely used in the development of current enterprises. With the help of order data analysis, an enterprise can grasp the development trends of the industry, so that their own development can meet better with the market economy situation, which helps the achieving of higher economic profit. When it comes to the role of order data statistics, it is mainly reflected in the following perspectives:

1. Better grasps of the development trends of the industry;
2. More scientific application of resources;
3. Effective guidance provision for enterprise development.

Statistical analysis of order data is a key content that modern enterprises must evolve when participating in market competition. The analyze of order data of an enterprise helps the realization of its' economic benefits, as well as enhancing its' own competitive advantages in the process of participating in market competition. Order data statistics analysis from an objective perspective can provide effective guidance for the future growth and development of the company.

The emergence of Internet ride-hailing platforms has not only reconstructed the offline ride-hailing market, but also provided more profitable possibilities for other idle resources in the market. Since birth, Uber has firmly occupied the first position in the non-China travel market. While developing rapidly, it has to continuously optimize the allocation of resources for social automobile travel in order to lower its' cost as well as enhancing its' benefit.

Therefore, aiming at this goal, in this project, we are trying to find out the relationship between time and location with rides and hoping to draw conclusions as well as providing corresponding suggestions for future operating strategies.

We come up with the following initial questions to explore:

1. What is the relationship between time and rides?
2. What is the peak period of passenger car use?
3. What is the most popular spot of rides?

Data

We get our Uber pickups data in New York City from April to September of Year 2014 from Kaggle (https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city (https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city)).

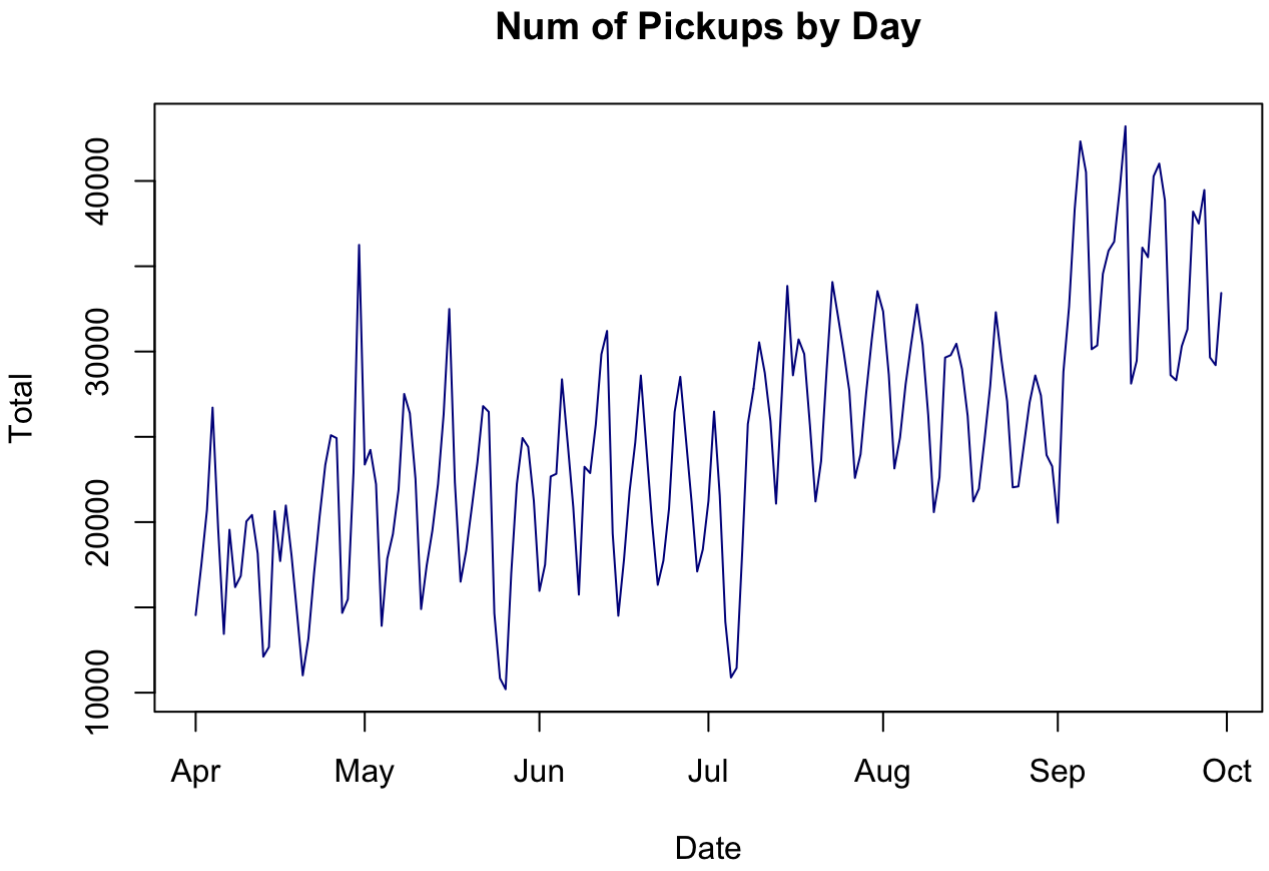
The dataset contains over 4.5 million Uber pickups order and gives information on the timing and location (that is date and time, the latitude and longitude) of each Uber pickup, as well as The TLC base company code affiliated with each Uber pickup. Before we carry out further analysis, we convert the date-time string into different format we need to use.

Methodology

We apply line-plot, pie-chart, bar-plot and heat map for descriptive analysis and time series analysis for quantitative analysis, findings and results are shown in the following part.

Findings

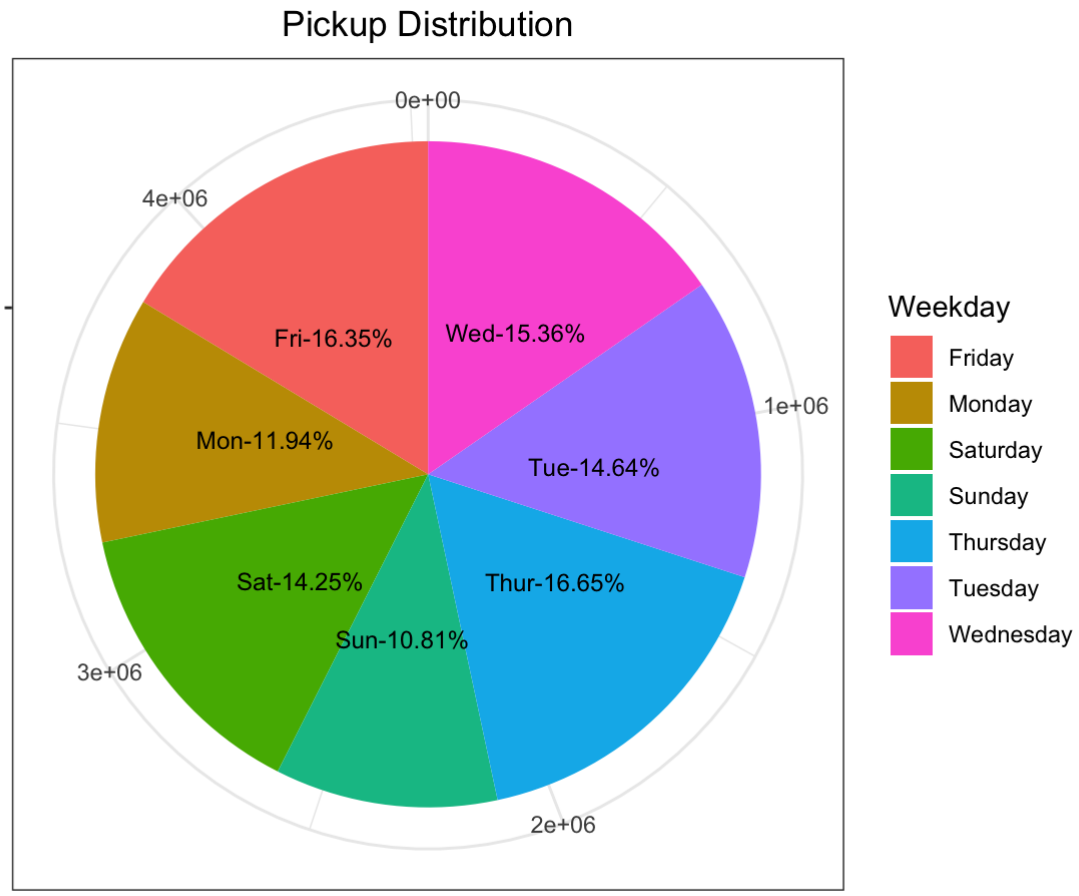
We first draw the pickup trends by hours.



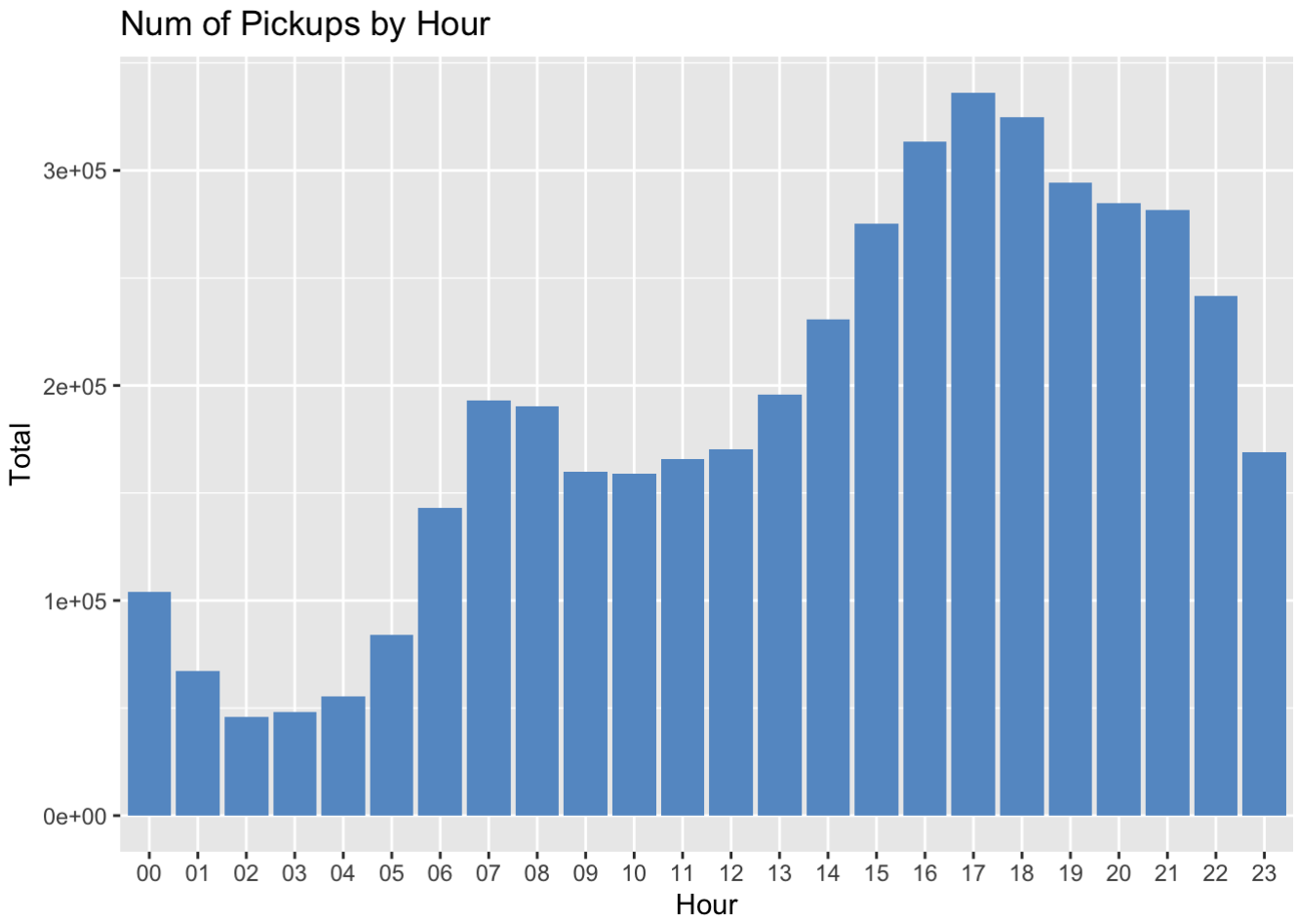
Viewing from the whole period, we can spot out that the daily number of pickups enjoy a steady rise even though fluctuation exists. In September, it reaches a peak. Also, the first figure clearly illustrates that the number of pickups exerts weekly trends, which we will discuss later in the quantitative analysis part.

We suggest that the rise comes from the increase of business range Uber provides, or the seasonal trend that ride-hailing market exerts. Unfortunately, we don't have enough historical data to prove the seasonal trend.

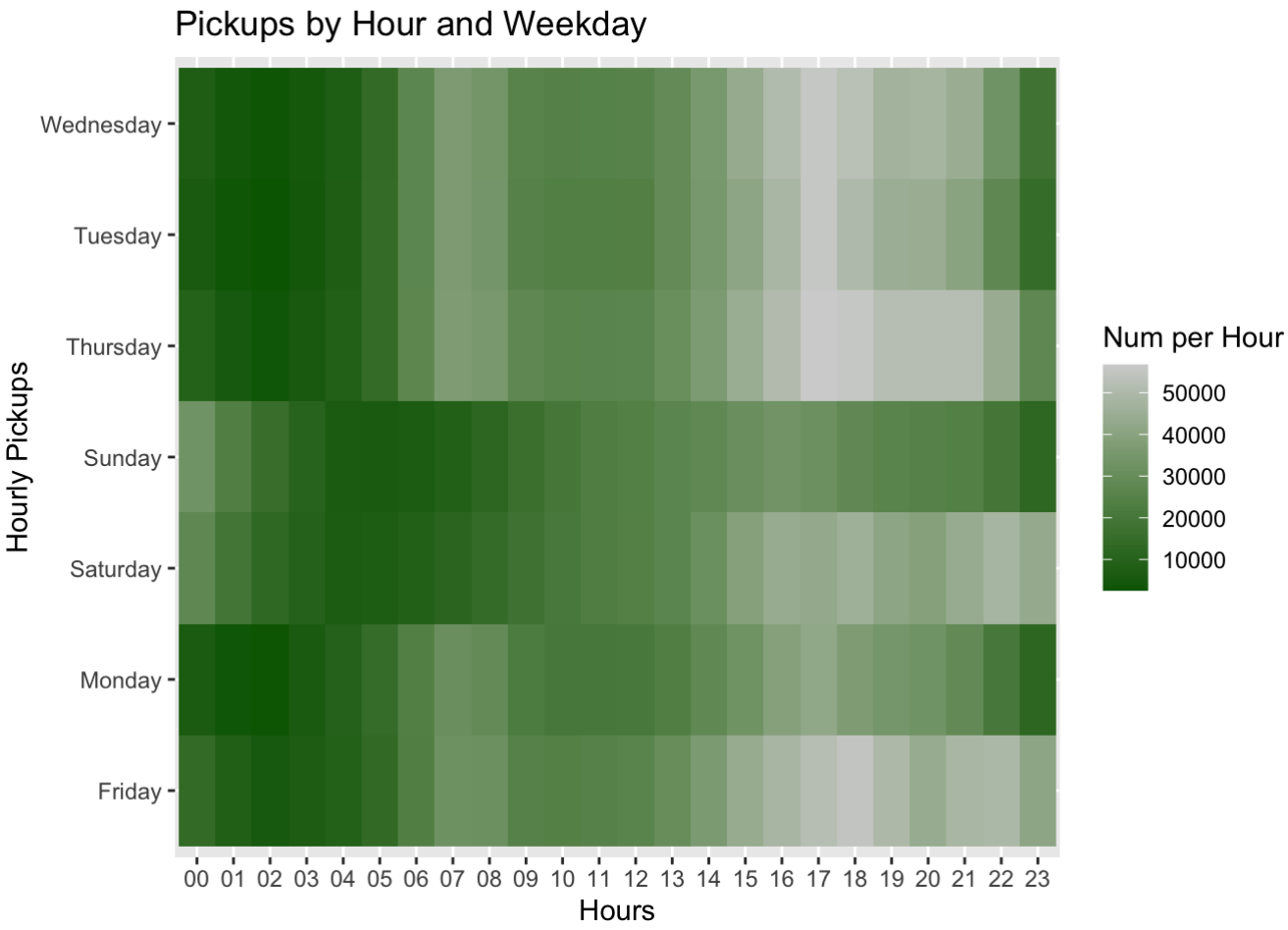
Moving to the results by aggregated by weekdays, we can tell that there do not exists or exists little significant differences between the number of pickups when grouping by weekdays, this contradicts with what the line-plot illustrates.

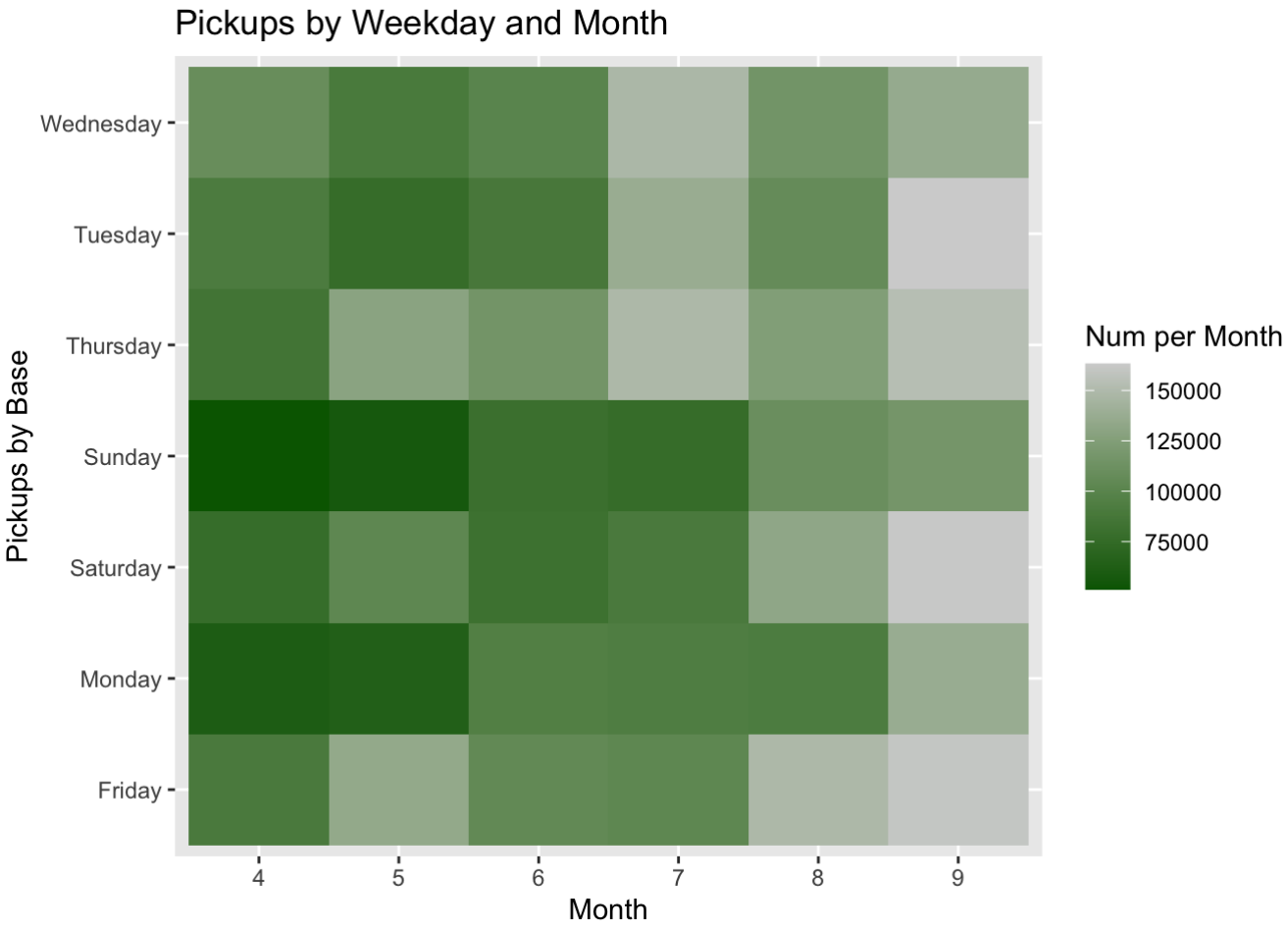


The bar-plot illustrates that obvious hourly trend exists, with peak occurs during evening hours and valleys are reached during late-night.



The heat map is an icon that displays the area of the a certain range that the visitor is keen on. We then draw the heat map of pickups of time and location to dig further.


Pickups by Hour and Weekday
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
Hourly Pickups
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hours
Num per Hour
50000
40000
30000
20000
10000


Pickups by Weekday and Month
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
Pickups by Base
4
5
6
7
8
9
Month
Num per Month
150000
125000
100000
75000

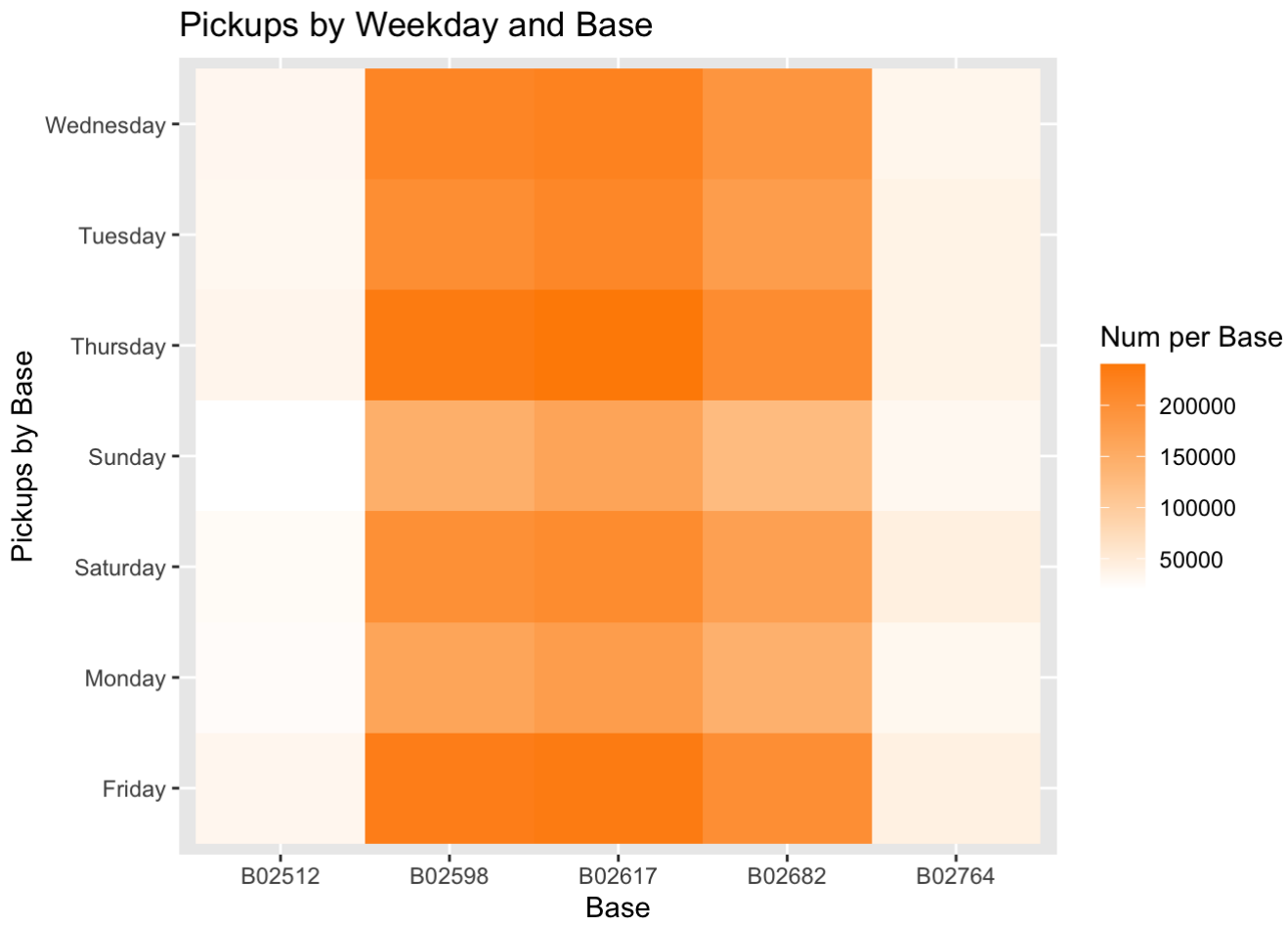
The two heat maps of pickups by both weekday as y-axis and hour, month as x-axis correspondingly are shown above.

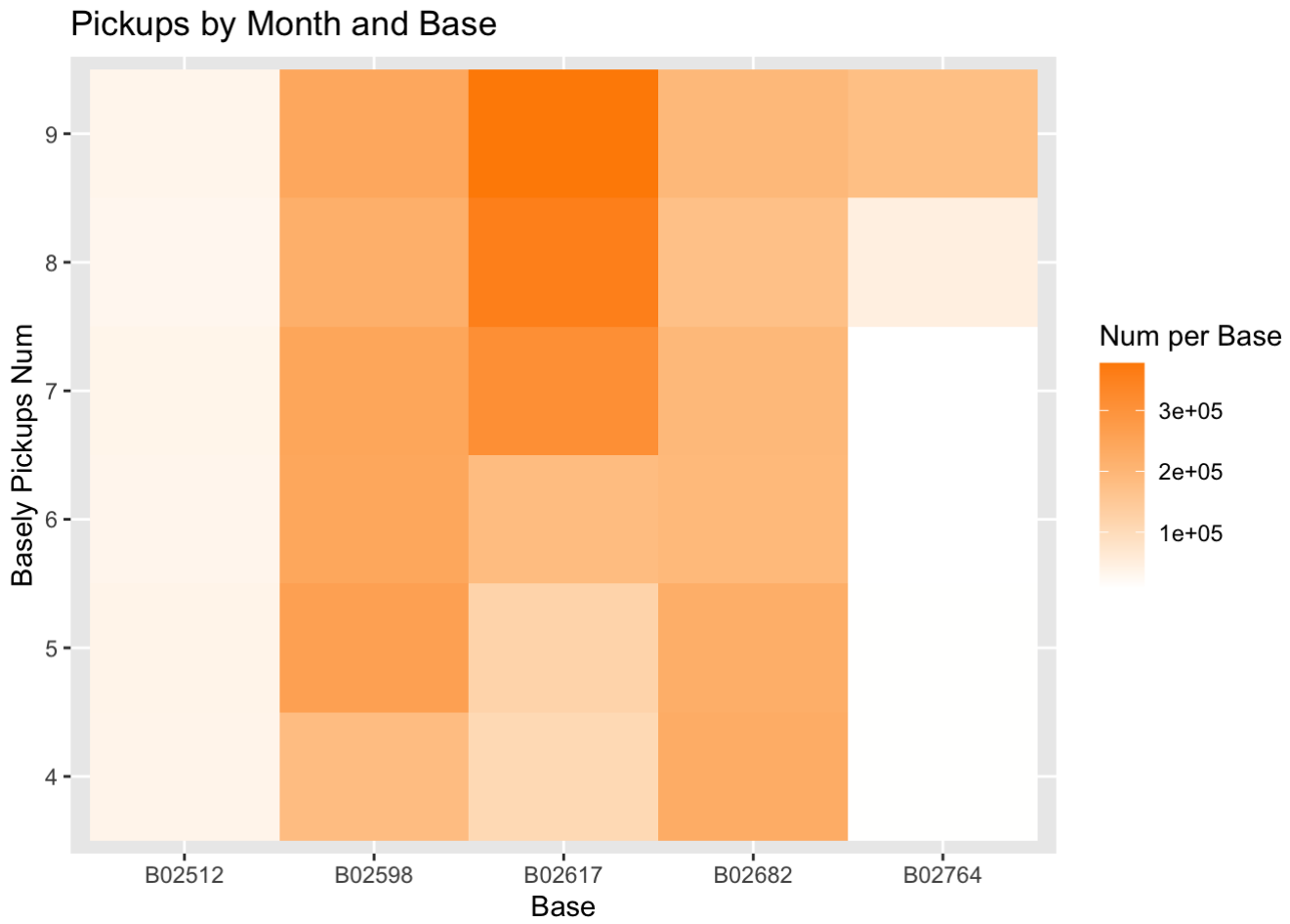
We can spot out that by month, gradual increase of the number of pickups occurs from April to September. Actually, the number of trips in September 2014 is almost as much as twice times the number of trips in April 2014.

By weekday, Thursday enjoys the highest volume of pickups while Sunday and Monday only receive relatively low pickup orders.

By hour, we can tell that the peak period for weekdays is rush hours, that is the time people logging from home to work or from work to home, which happens at 7am to 9am and 4pm to 10pm; the peak period for weekends occurs during Saturday night.

The heat maps of pickups by base and time are drawn below.





From these figures, we can tell that Base B02617 is the most popular spot that ride will occurs, while B02512 and B02764 behaves bad in comparison. Moreover, in Thursday and Friday, the number of pickups are higher than other times for Base B02598, B02617 and B02682.

Till now, we spot out the trend of number of pickups by time and base through plots and figures, in the last part, we tend to find out the equation of the trend by applying time series analysis.

We calculate the fluctuation of the number of pickups by weekdays since weekly trends exerts and get a differenced sequence. This sequence passes relative test and tends to be stationary (the basic demand of time series analysis). Our model fitting is carried out based on the sequence.

We get a model of ARIMA(1,0,0) for the weekly differenced daily pickups sequence, the model passes the residual test, that is to say, the residuals have no significant trend, appearing to be ‘random’, indicating information of the sequence has been fully extracted. It is described in the form as below.

$$y_t = 0.6024 * y_{t-1} + y_{t-7} - 0.6024 * y_{t-8} + \varepsilon_t$$

Thus, it is clear that the number of pickups exerts weekly trends.

To sum up, we give the brief answers for out initial questions.

1. What is the relationship between time and rides? The number of pickups follows a ARIMA(1,7,0) time series model in the form of $y_t = 0.6024 * y_{t-1} + y_{t-7} - 0.6024 * y_{t-8} + \varepsilon_t$, in a simple word, the number of pickups exerts weekly trends.

2. What is the peak period of passenger car use? The peak month of passenger car use is September; the peak period for weekdays is rush hours, that is the time people logging from home to work or from work to home, which happens at 7am to 9am and 4pm to 10pm; the peak period for weekends occurs during

Saturday night.

3. What is the most popular spot of rides? The most popular spot of rides is B02617, while B02512 and B02764 are the most unpopular bases.

Therefore, we can adjust our vehicle scheduling system according to the peak period as well as the popular spot of passenger car use, this can improve the efficiency of transport capacity, leading to greater profits.