Interim Report

Hukai Luo

13 November 2018

1 Introduction

The goal of this project is to predict the *duration of taxi rides in NYC* based on features like trip coordinates or pickup date and time.

Since there are lots of parameters when we are estimating the trip duration, we will first study and visualise the original data, engineer new features, and examine potential outliers. Finally, we will choose some important parameters to curve fit the data using the least square method.

First of all, let's load the given data

```
library('tibble')
library('data.table')
train <- as.tibble(fread("/Users/luohukai/Documents/GitHub/final-project-hul17011/train.csv"))
test <- as.tibble(fread("/Users/luohukai/Documents/GitHub/final-project-hul17011/test.csv"))
sample <- as.tibble(fread("/Users/luohukai/Documents/GitHub/final-project-hul17011/sample.csv"))
```

Then find the structure of the data

```
library('dplyr')
summary(train)
combine <- bind_rows(train %>% mutate(dset = "train"),
                     test %>% mutate(dset = "test",
                                     dropoff_datetime = NA,
                                     trip_duration = NA))
combine <- combine %>% mutate(dset = factor(dset))
```

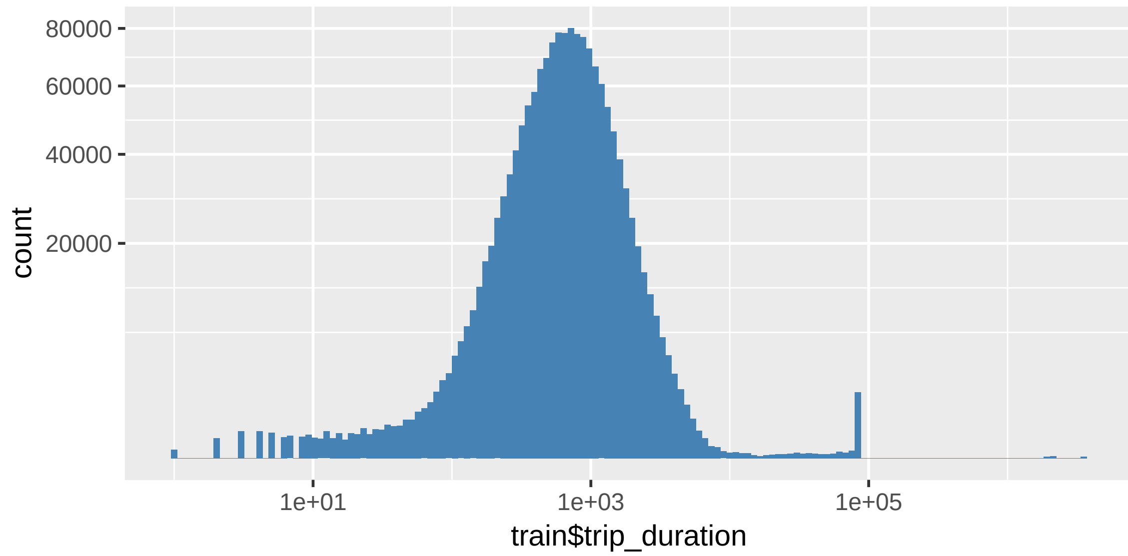
We find the data contains several factors: **vender_id** takes only 1 or 2 which represents two taxi companies; *pickup_datetime*; *dropoff_datetime*; *passenger_count*; *pickup_longitude*; *pickup_latitude*; *dropoff_longitude*; *dropoff_latitude*; *store_and_fwd_flag*; *trip_duration* which is measured in seconds.
In order to make the data easy to use, we will make some change to the data.

```
library('lubridate')
train <- train %>%
  mutate(pickup_datetime = ymd_hms(pickup_datetime),
         dropoff_datetime = ymd_hms(dropoff_datetime),
         vendor_id = factor(vendor_id),
         passenger_count = factor(passenger_count))
```

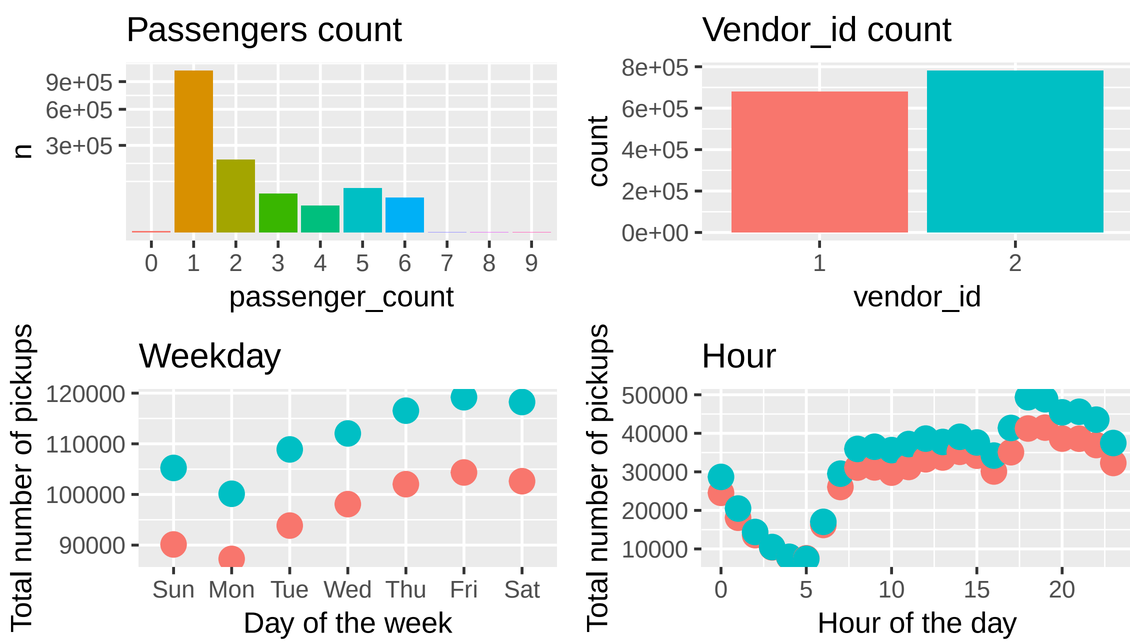
2 Plot by single parameter

Now in order for us to get a better understanding of the data, we will begin by having a look at the distributions of the individual data features. First of all, let's plot the target feature trip_duration.

```r
library('ggplot2')
p1 <- ggplot(train, aes(train$trip_duration)) +
  geom_histogram(fill = "steelblue", bins = 150) +
  scale_x_log10() +
  scale_y_sqrt()
p1
```



Comments: Most trips will ends in nearly 1000 seconds, but there will also be some exceptions. Then we can also plot the distribution of passenger_count,Vendor_id,day of the week, hour of the day



Comments: Most trips only have 1 passenger; Thursday,Friday and Satuaday are the most busy days; there is a strong dip during the early morning hours and another dip around 4pm.

3 Relations

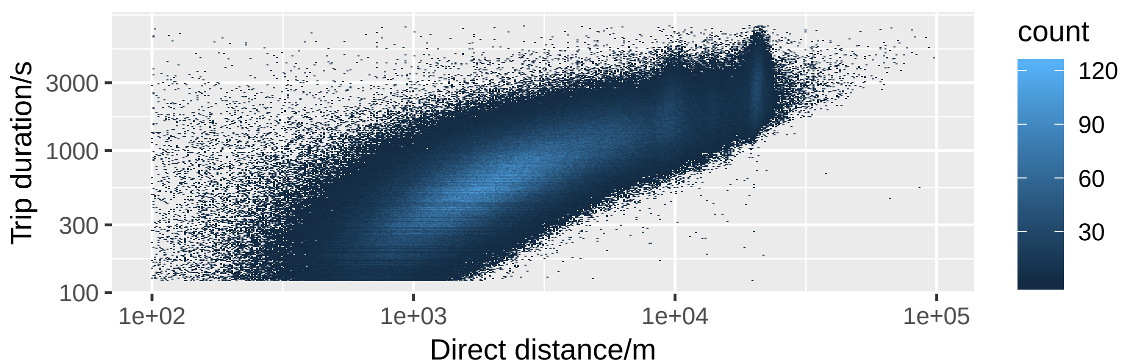
While the previous section looked primarily at the distributions of the individual features, here we will examine in more detail how those features are related to each other and to our target trip_duration. In this project, we will assume that the trip_duration is only related to **Trip distance**, **passenger numbers**, **vender_id**, **day of the week**, **hour of the day**.

3.1 Trip distance vs trip_duration

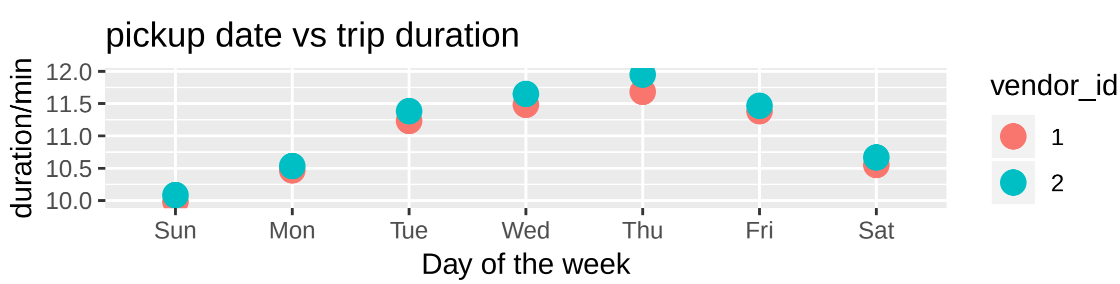
First, we need to calculate the exact trip distance by the pickup and dropoff location.

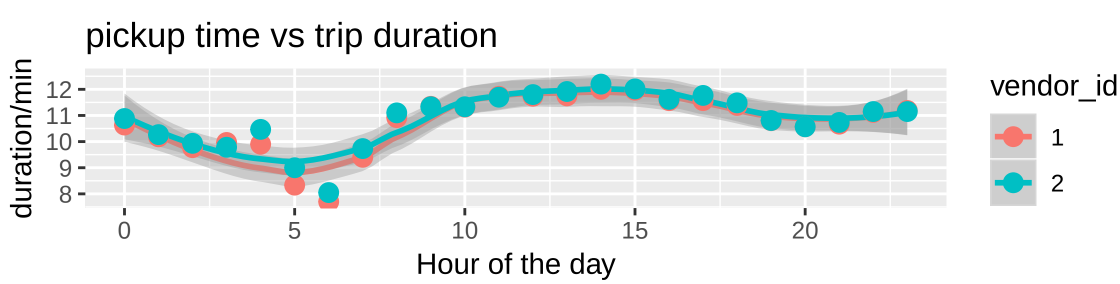
```
library('geosphere')
pick_coord <- train %>%
  select(pickup_longitude, pickup_latitude)
drop_coord <- train %>%
  select(dropoff_longitude, dropoff_latitude)
train$dist <- distCosine(pick_coord, drop_coord)
```

Then plot the Trip distance vs trip_duration distribution.

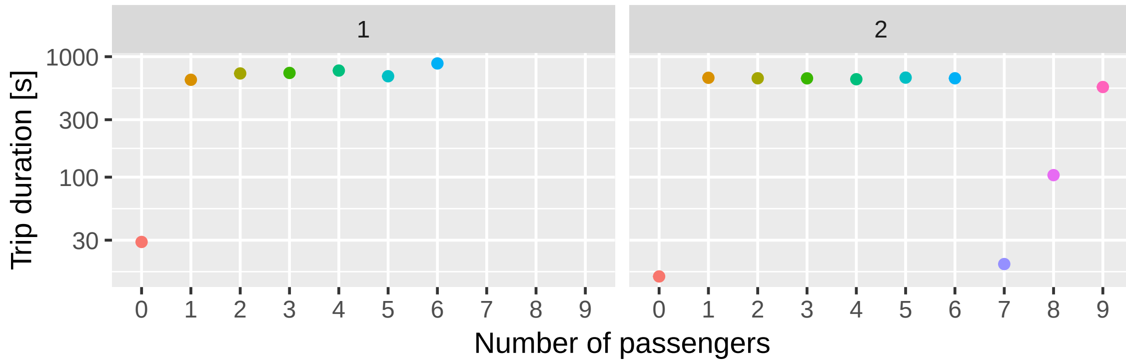


3.2 Pickup date/time vs trip_duration





3.3 Passenger number vs trip_duration



From this plot, we can find the trip_duration doesn't have a strong relationship with the passenger numbers, the difference in the picture may only reveal the impact of different distance. For those passenger numbers = 7, 8, 9, we don't think that they are reasonable, so we just won't consider them.

4 Prediction/least square method

In this project, we assume the trip_duration distribution function has three parameters: trip_distance d, pickup date w, pickup time t.