MA615 Final Project

Kaiwei Xiao

2022-12-17

```
Q1 = read.csv(
  "2022-Q1_LRTravelTimes.csv")
```

It's not reasonable but for some reason I just want to find if there is any relationship between the stop id and travel time.

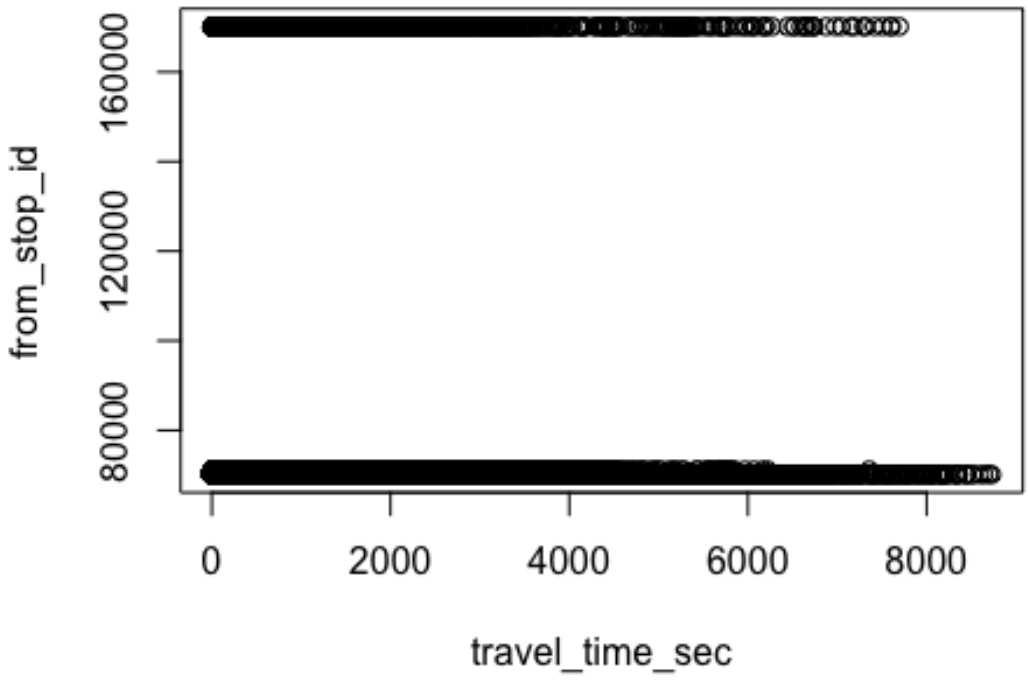
Below is a simple linear regression

```
m1 = lm(from_stop_id ~ travel_time_sec, data = Q1)
summary(m1)
```

```
##
## Call:
## lm(formula = from_stop_id ~ travel_time_sec, data = Q1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3849  -3471  -3078  -2494 103880
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.396e+04  6.891e+00   10732   <2e-16 ***
## travel_time_sec -1.001e+00  6.810e-03    -147   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16940 on 15879909 degrees of freedom
## Multiple R-squared:  0.001359,   Adjusted R-squared:  0.001359
## F-statistic: 2.161e+04 on 1 and 15879909 DF,  p-value: < 2.2e-16
```

The significant coefficient do shows some relationship. But how? Let's plot them.

```
plot(from_stop_id ~ travel_time_sec, data = Q1)
```



It seems like if we do a logistic regression there will be more sense there. First I map thoese stop id into 0 to 1 range. Then I do a logistic regression

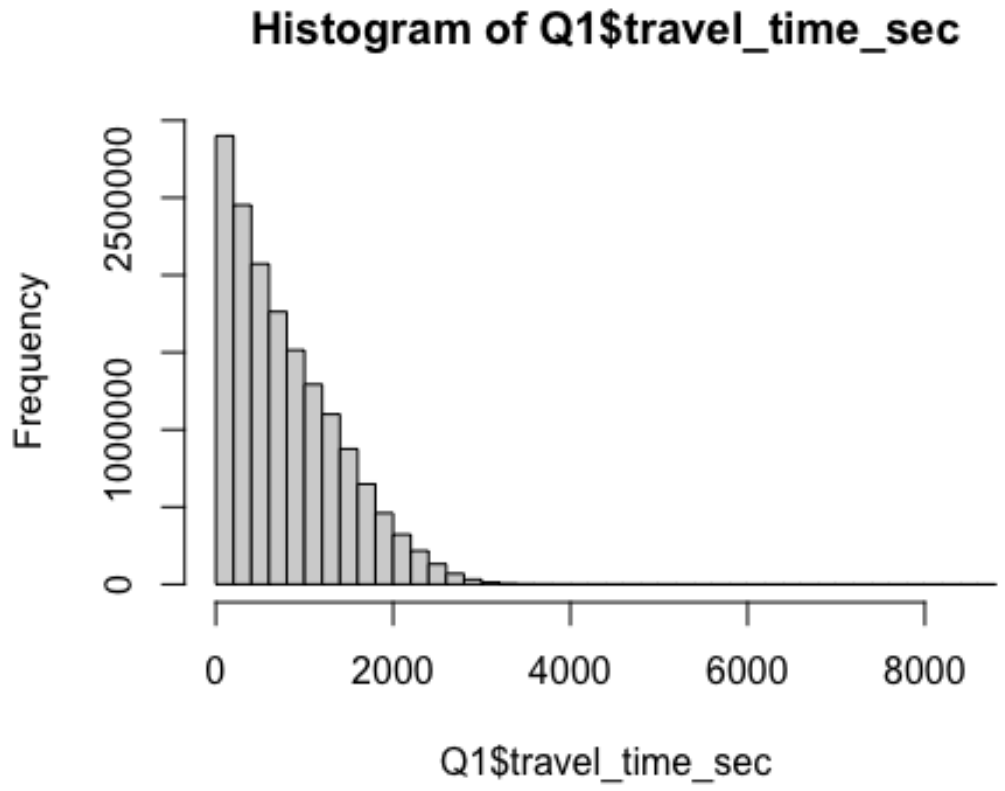
```
library(scales)
tstop = rescale(Q1$from_stop_id, to=c(0,1))
m2 = glm(tstop ~ travel_time_sec, data = Q1, family = "binomial")

## Warning in eval(family$initialize): non-integer #successes in a bino
mial glm!

summary(m2)

##
## Call:
## glm(formula = tstop ~ travel_time_sec, family = "binomial", data = Q
1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2850  -0.2554  -0.2365  -0.2099   3.4885
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.183e+00  2.283e-03 -1394.2   <2e-16 ***
```

```
## travel_time_sec -3.769e-04  2.612e-06  -144.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4129534  on 15879910  degrees of freedom
## Residual deviance: 4107054  on 15879909  degrees of freedom
## AIC: 4218846
##
## Number of Fisher Scoring iterations: 6
```

The summary give us evidence that the model fit well. Which means if you know the travel time, you can give a probability that which stop id you are start from.

Let's check the travel time histogram

```
hist(Q1$travel_time_sec, breaks = 50)
```



Most of the travel time are very short and close to 0.

##Shiny app

for the shiny app there are two version

The first version is driving map, which allow you to add waypoint on it. Simply type in the departure address and final destination (waypoints if needed), you can get the driving routine. Also use your map key and API key if possible.

#published link: https://jpz61m-jeremy-x0.shinyapps.io/Driving_Map/

The second version is transit map. It allows you travel through public transportation but no waypoints due to the API limitation.

#published link: https://jpz61m-jeremy-x0.shinyapps.io/Transit_Map/