Assignment 4: Linear Regression

Team Members: Vedant Naik, Rohan Joshi

Problem Statement:

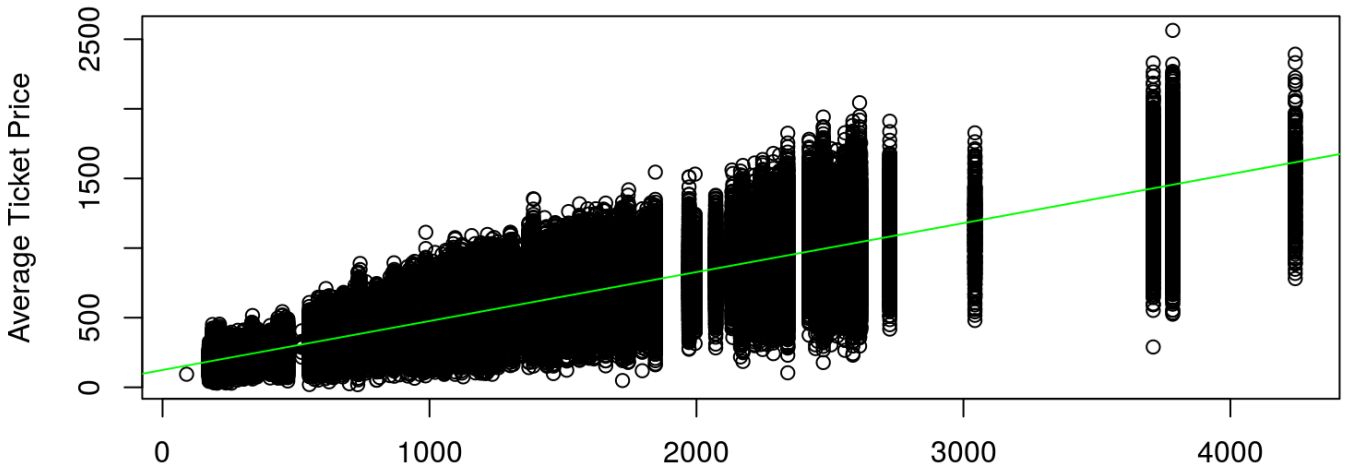
The price of a ticket depends, in part, on the amount of fuel consumed in the particular itinerary (a factor of distance and winds). Furthermore, prices have a tendency to increase over time. To understand airline pricing, computer a simple linear regression that models the cost of tickets for different airlines. Give a new ranking of airlines with respect to price.

The output of the Map-Reduce program is used as an input for this report.

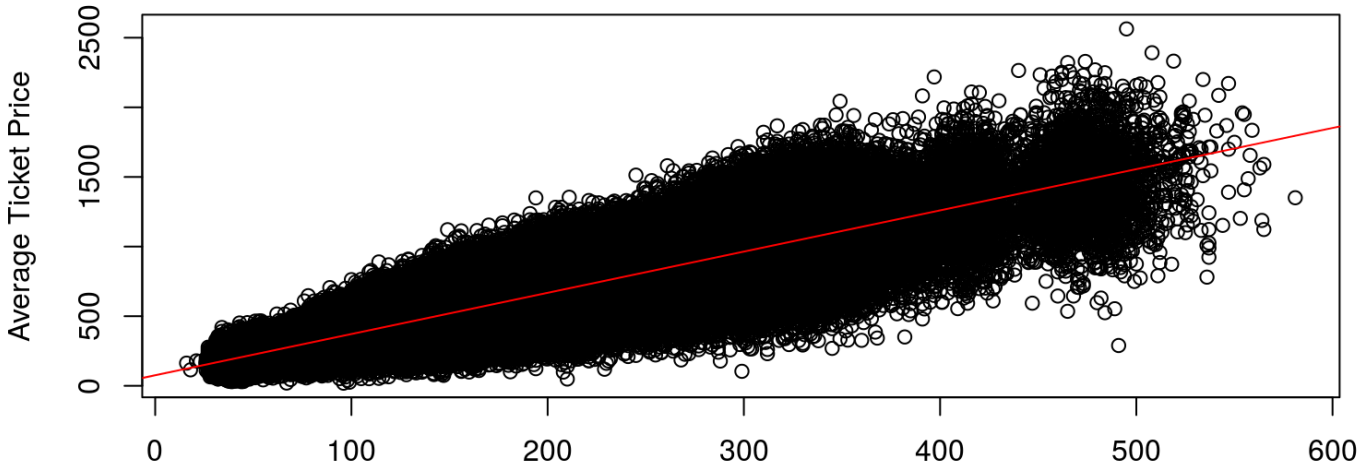
Airlines active in 2015

HA, EV, MQ, OO, US, B6, WN, UA, DL, NK, VX, AS, F9, AA (Of these, NK does not have any data in 2010-2014)

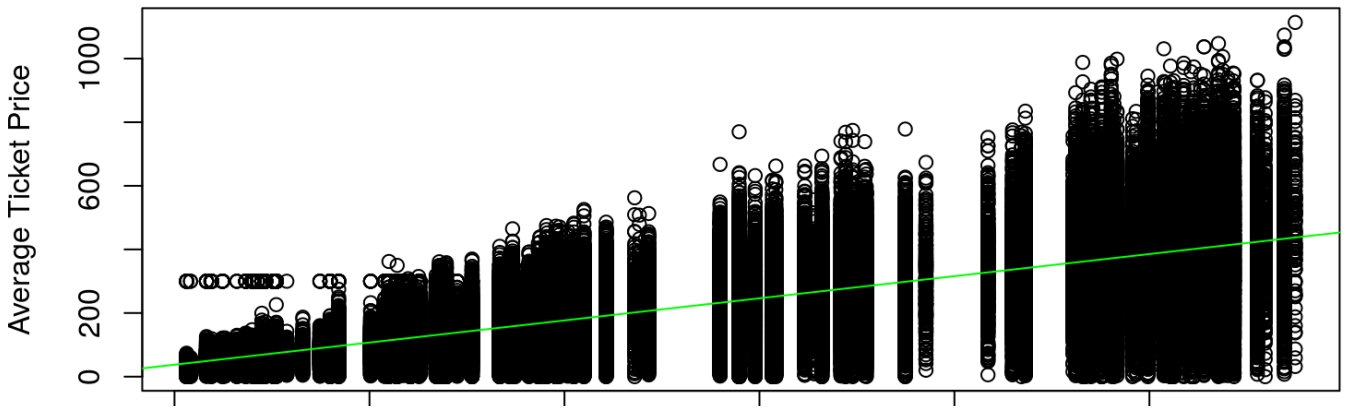
Graphs showing a linear fit of the variable to price for each airline



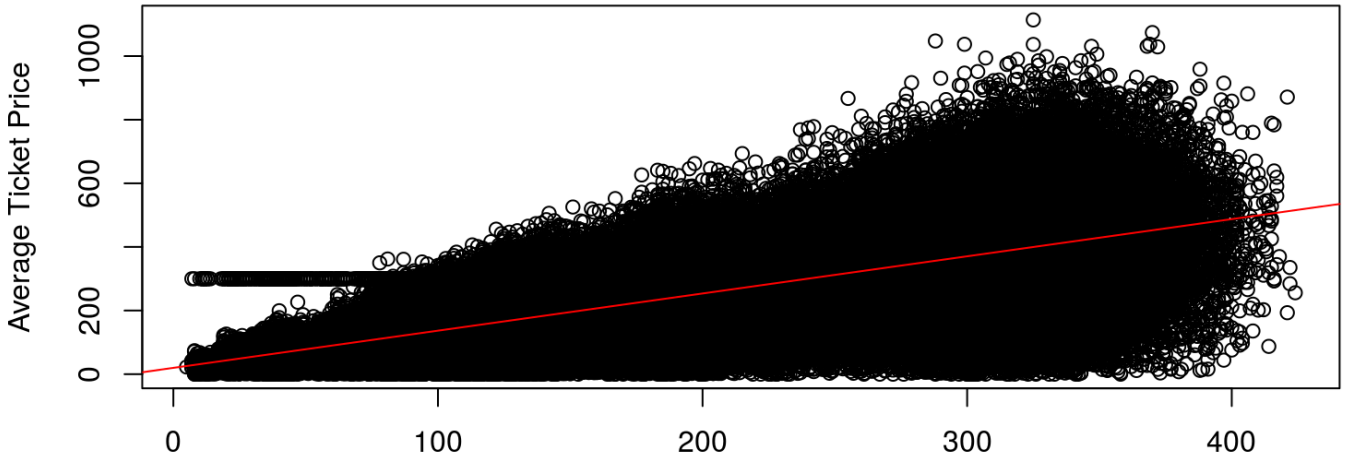
AA Distance Plot - Intercept : 124.086238284994 Slope: 0.351675432543031 MSE: 13577.55785696(



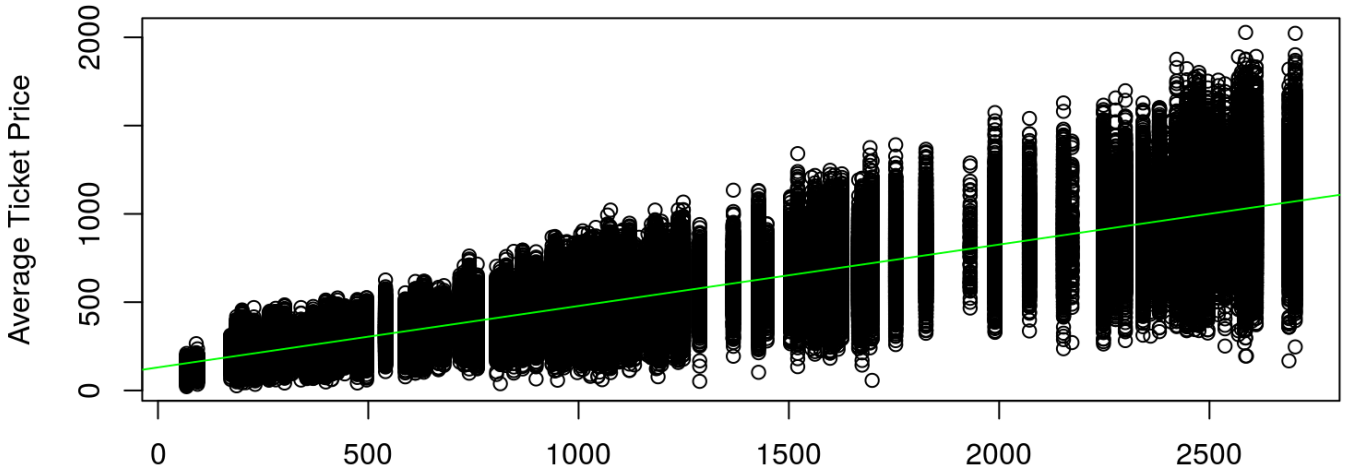
AA Time Plot - Intercept : 75.6213130714896 Slope: 2.96063458495978 MSE: 12861.5945295545



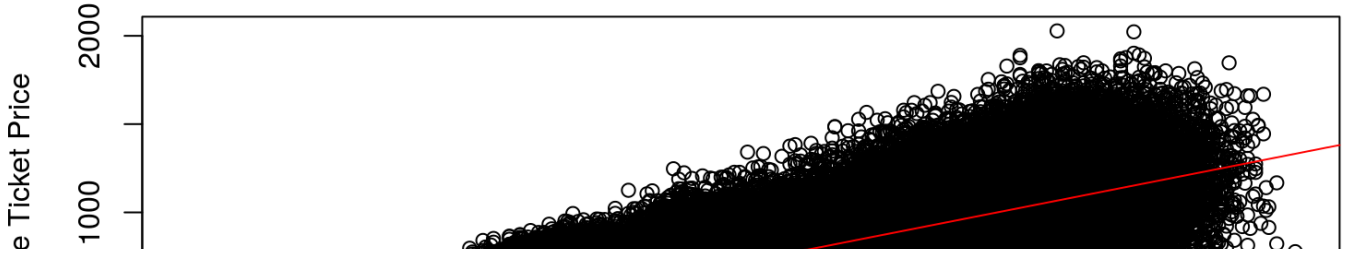
AS Distance Plot - Intercept : 37.4980949746815 Slope: 0.139147772036619 MSE: 8585.494659253

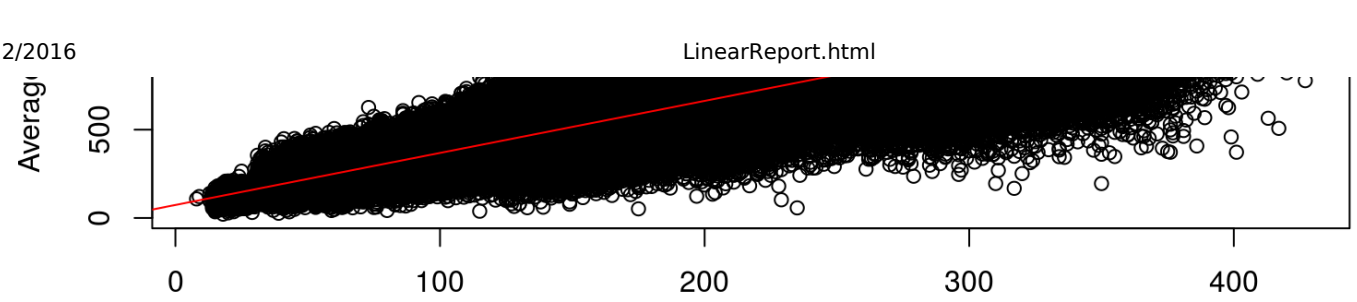


AS Time Plot - Intercept : 19.6417486866346 Slope: 1.16939132996612 MSE: 8530.9021207004

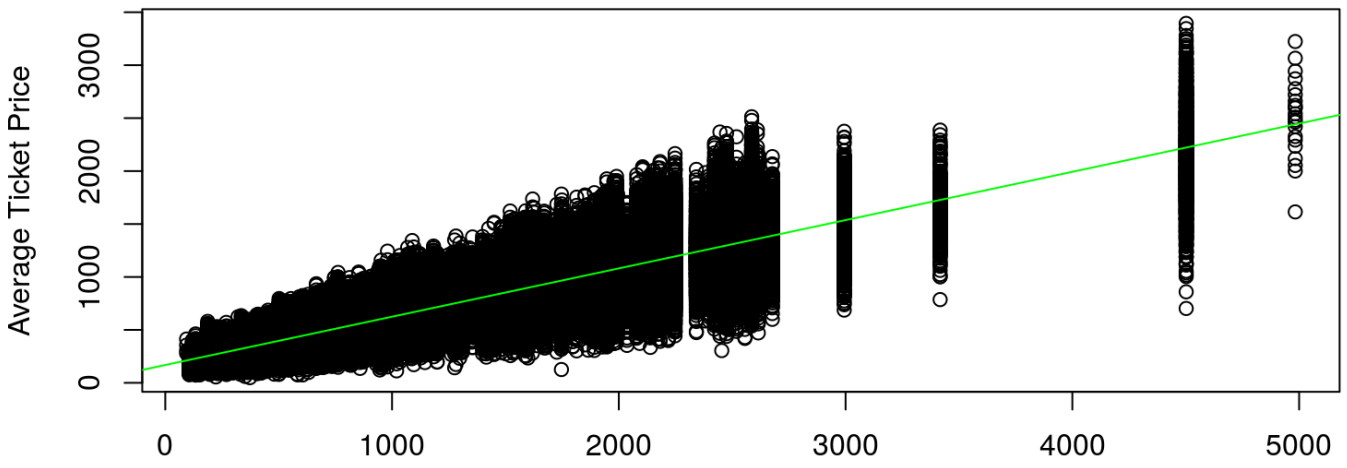


B6 Distance Plot - Intercept : 130.419447340971 Slope: 0.347895125752539 MSE: 14140.18720647

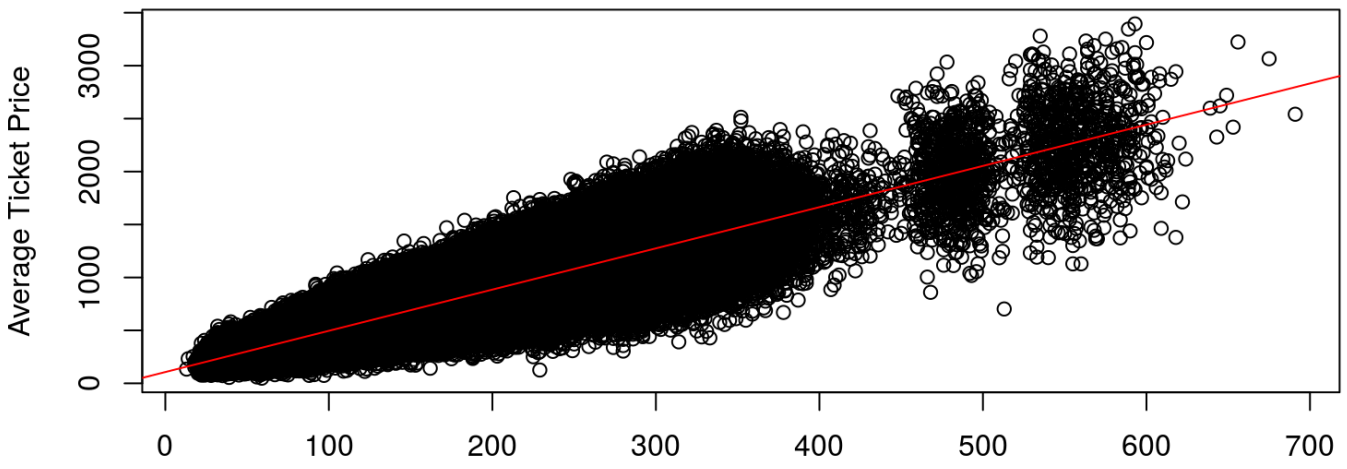




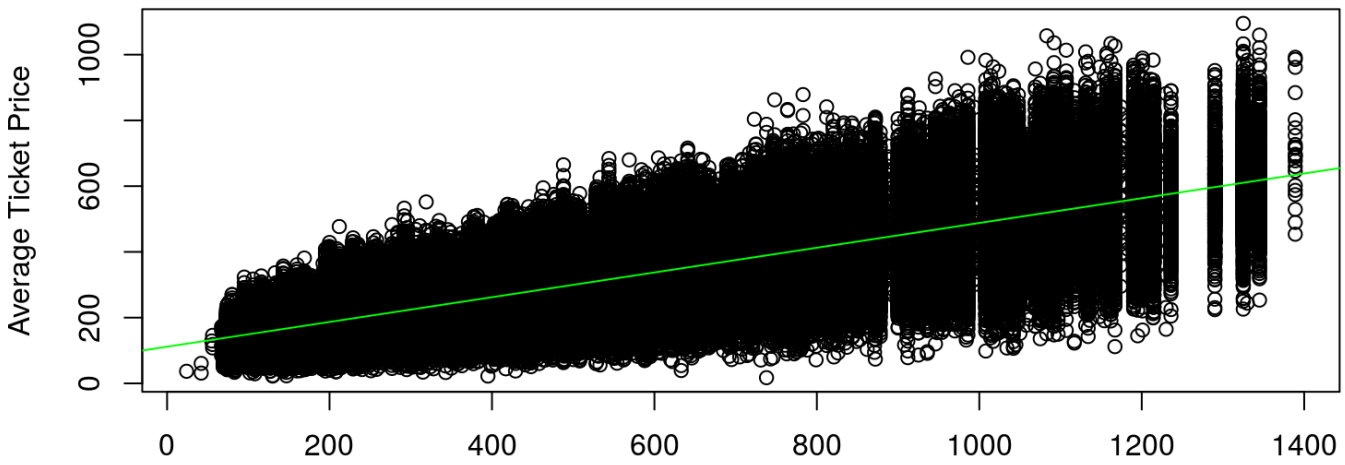
B6 Time Plot - Intercept : 73.3813505192907 Slope: 2.9470081179088 MSE: 13340.3881618974



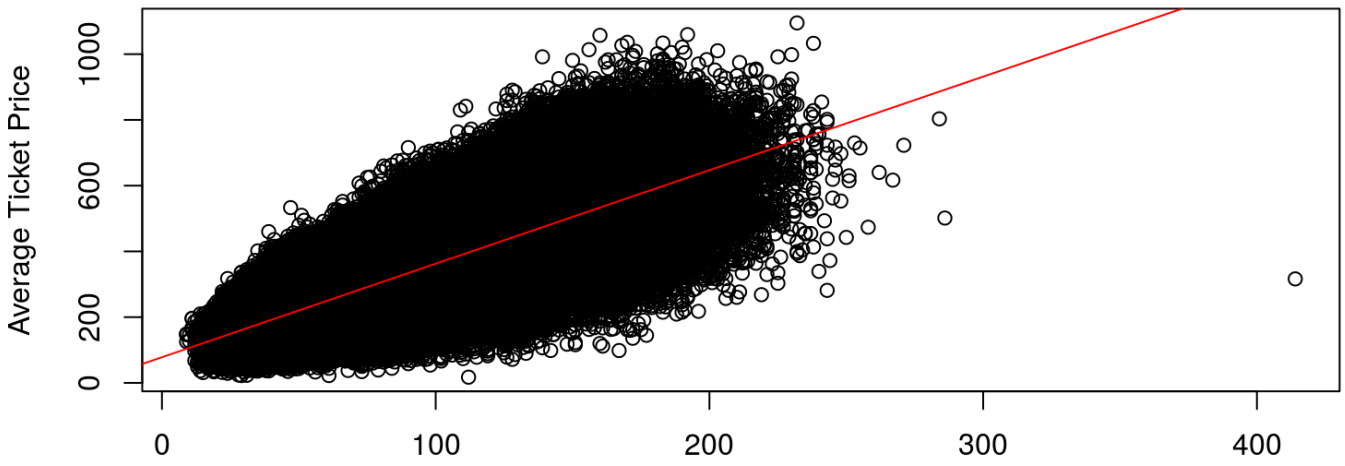
DL Distance Plot - Intercept : 168.470512931104 Slope: 0.456147256058837 MSE: 14733.19932261



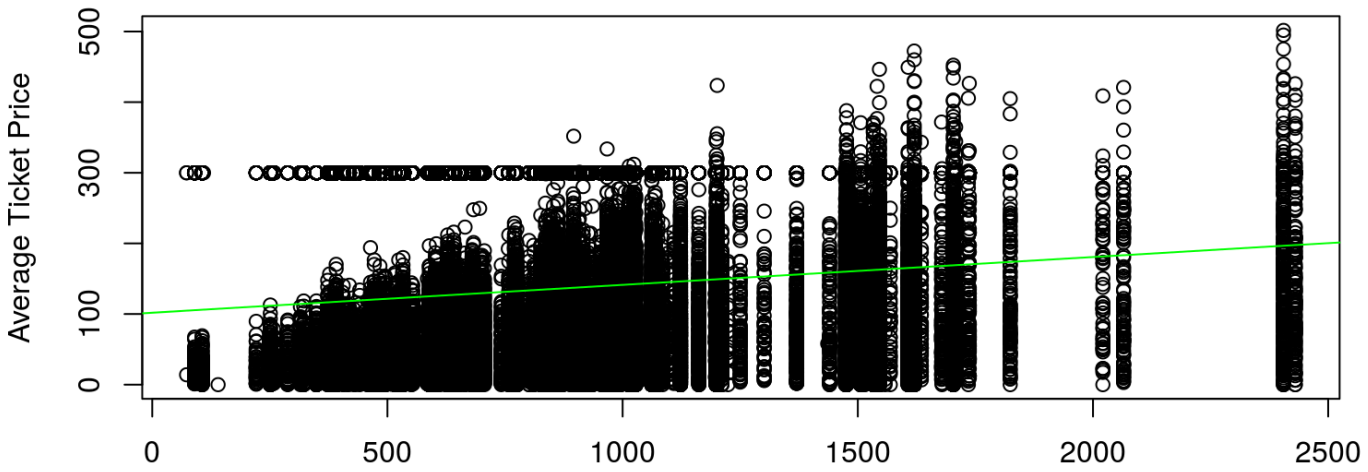
DL Time Plot - Intercept : 106.365396558424 Slope: 3.89311687591794 MSE: 13751.4586001922



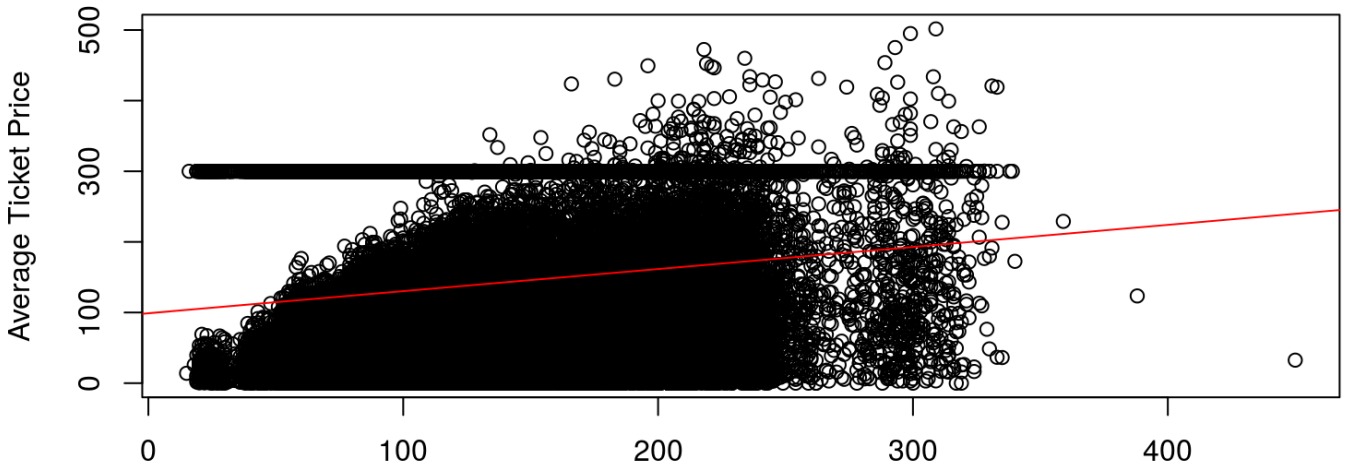
EV Distance Plot - Intercept : 111.526430394886 Slope: 0.37629290319073 MSE: 4451.2219213577



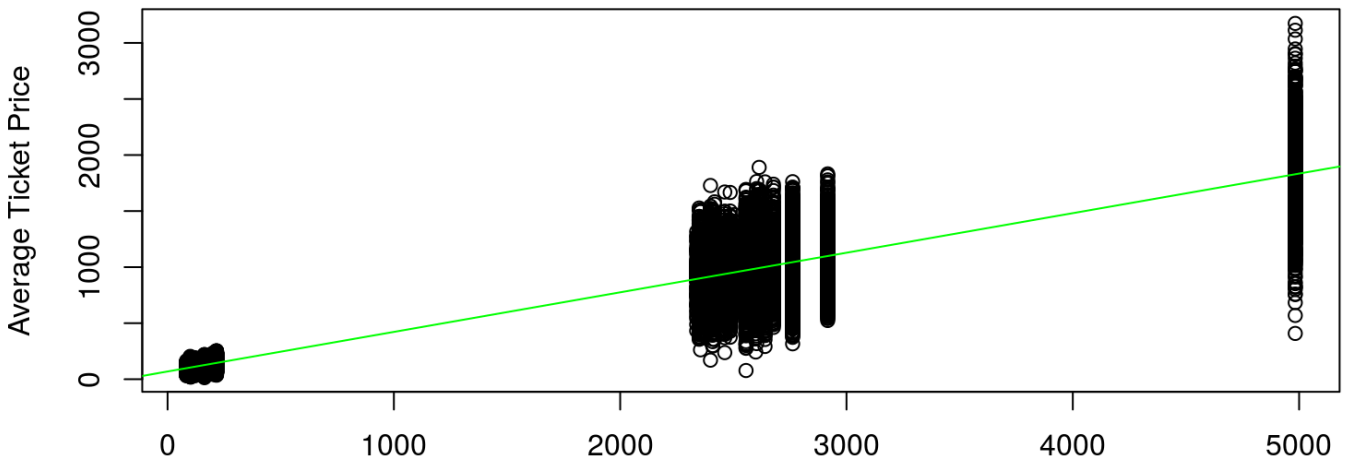
EV Time Plot - Intercept : 77.9429680630528 Slope: 2.84470153843118 MSE: 4576.24079109492



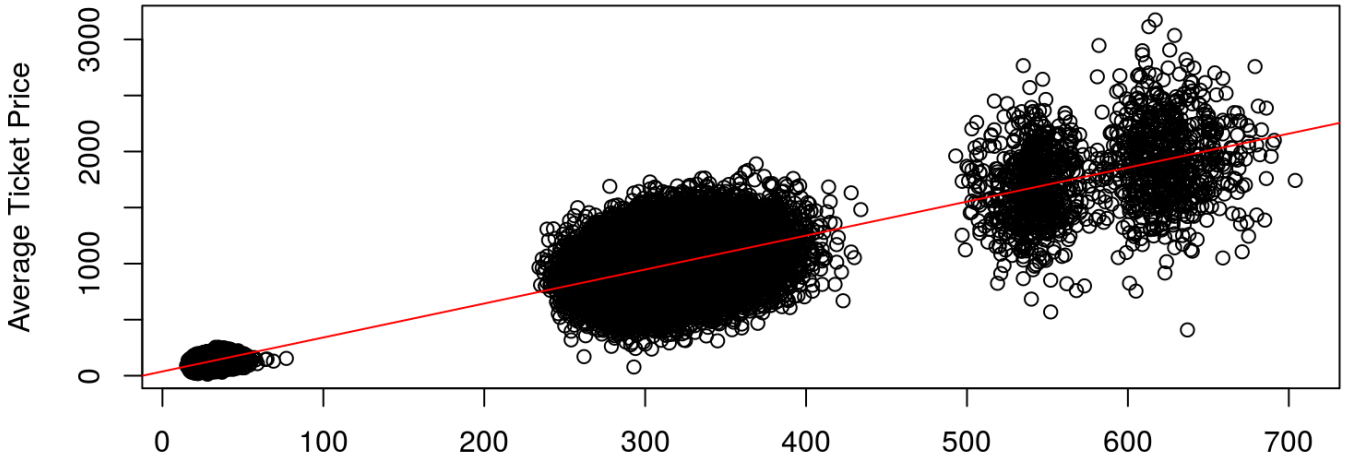
F9 Distance Plot - Intercept : 101.78401875101 Slope: 0.0394154447417115 MSE: 13603.771588555



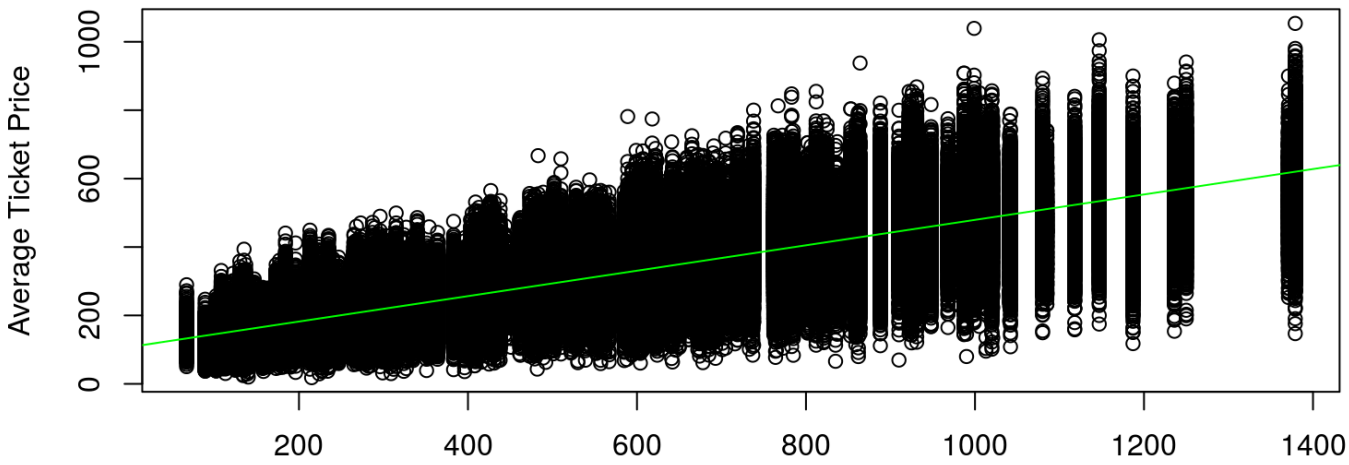
F9 Time Plot - Intercept : 98.9101025272196 Slope: 0.312860086852667 MSE: 13603.4615366985



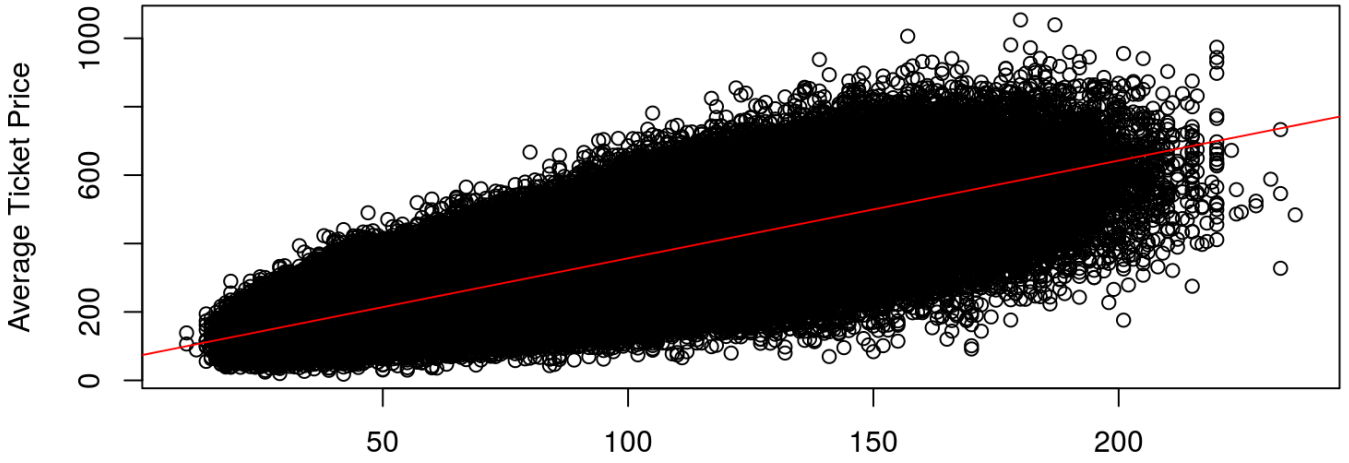
HA Distance Plot - Intercept : 68.7246721712862 Slope: 0.353121270252018 MSE: 8650.069487284



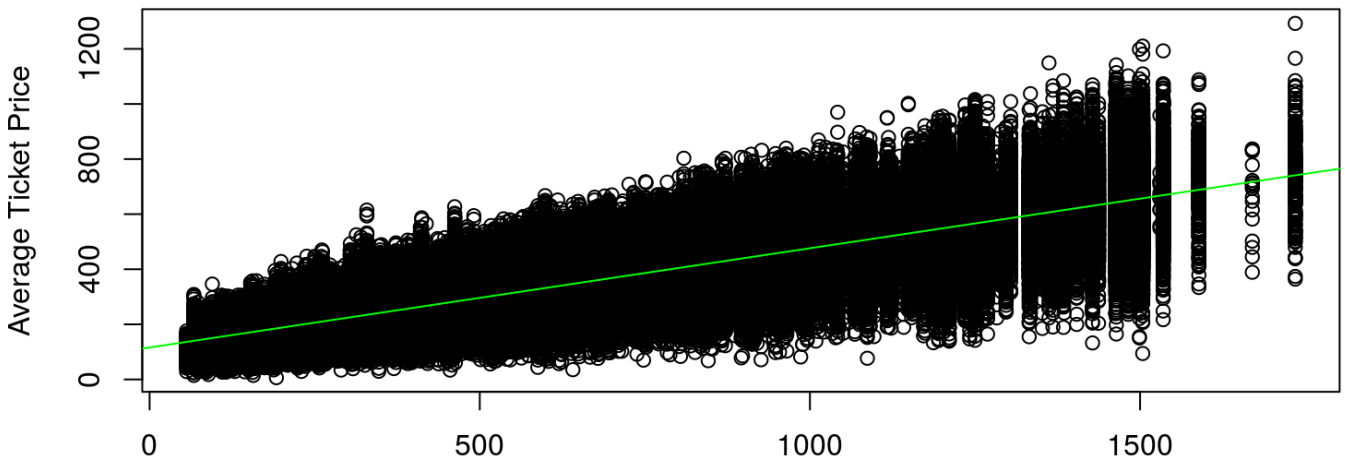
HA Time Plot - Intercept : 37.6362324011018 Slope: 3.03036379186358 MSE: 8545.44231029723



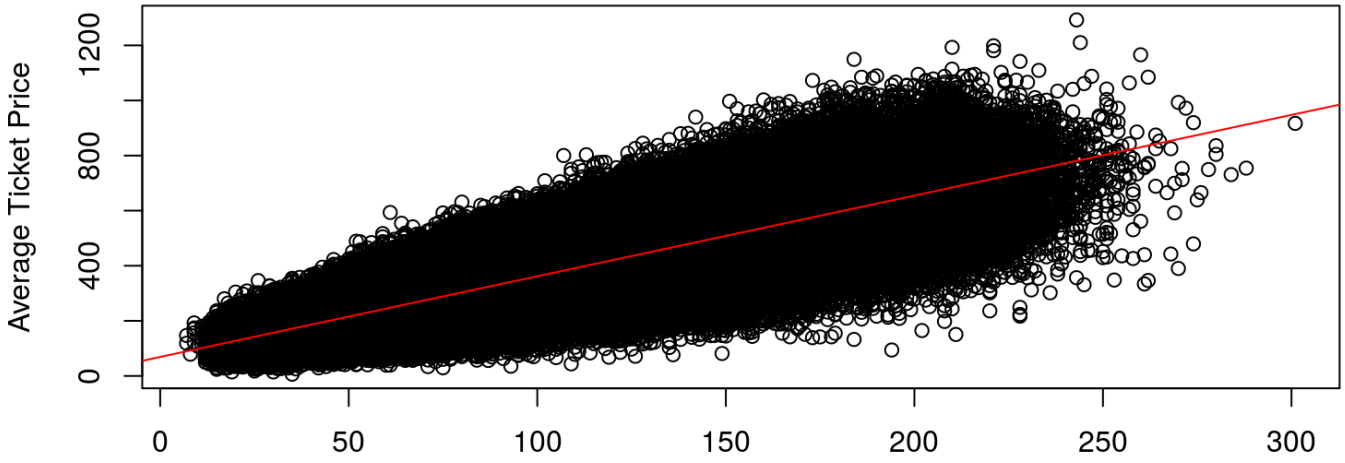
MQ Distance Plot - Intercept : 107.813350595992 Slope: 0.371448809852199 MSE: 4479.661485395



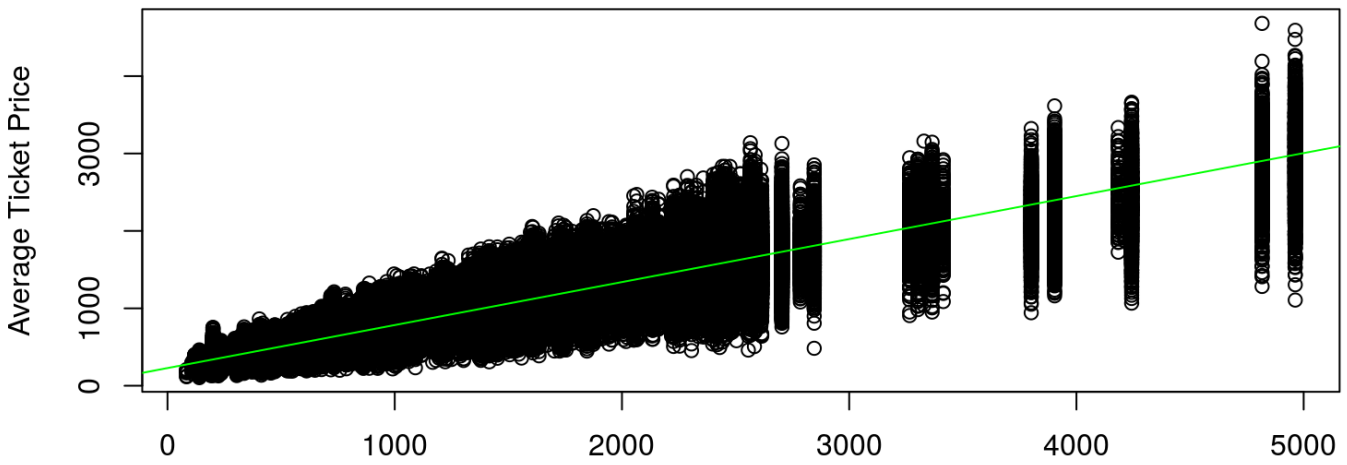
MQ Time Plot - Intercept : 71.1741707630537 Slope: 2.85578006936557 MSE: 4458.22986775931



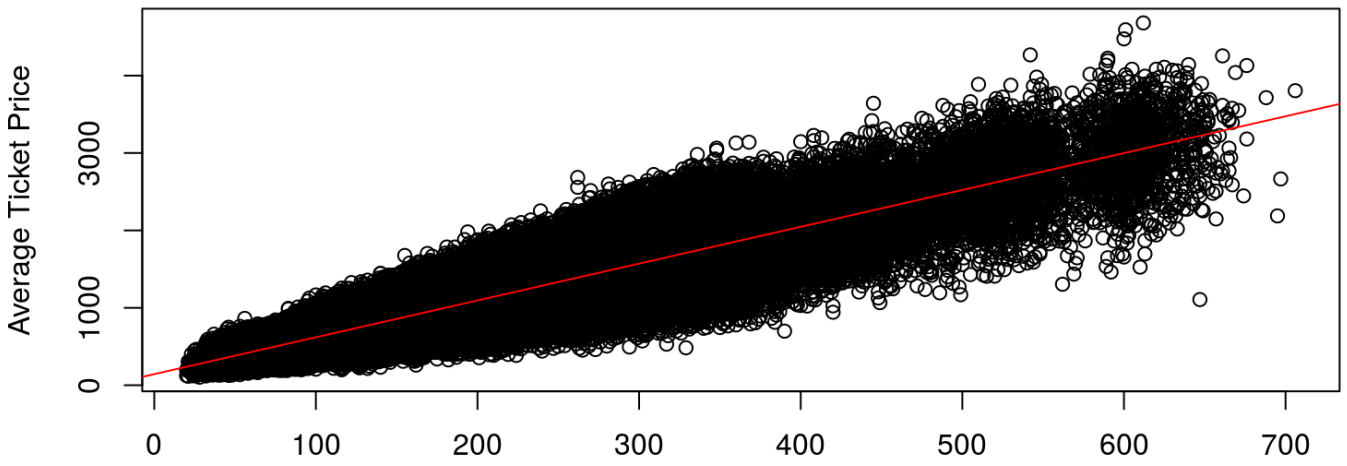
OO Distance Plot - Intercept : 116.049220266678 Slope: 0.360118667866591 MSE: 4294.574647569



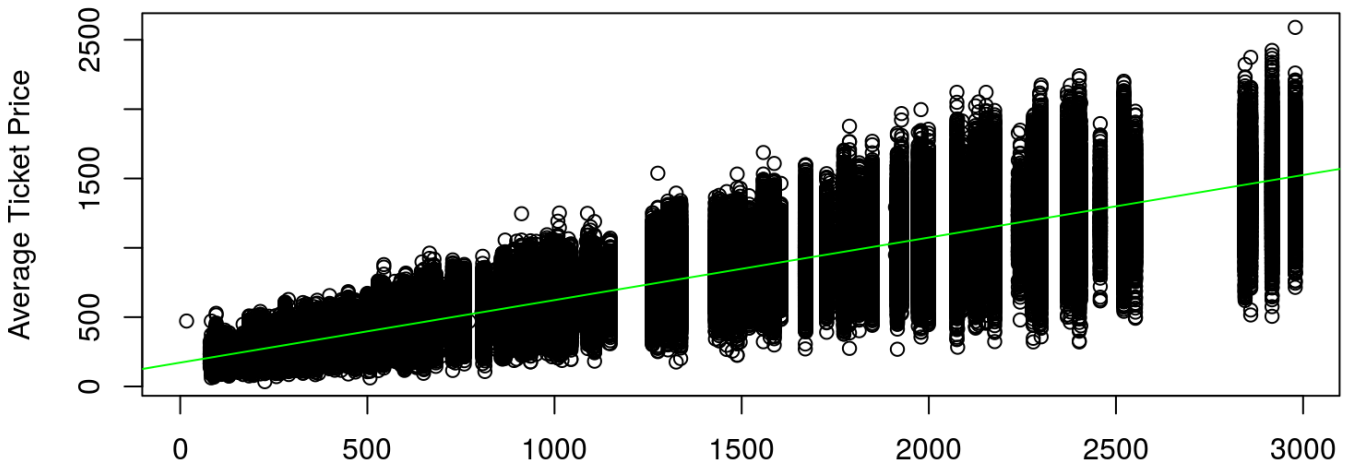
OO Time Plot - Intercept : 68.8546785630441 Slope: 2.92840258605386 MSE: 4347.19274570734



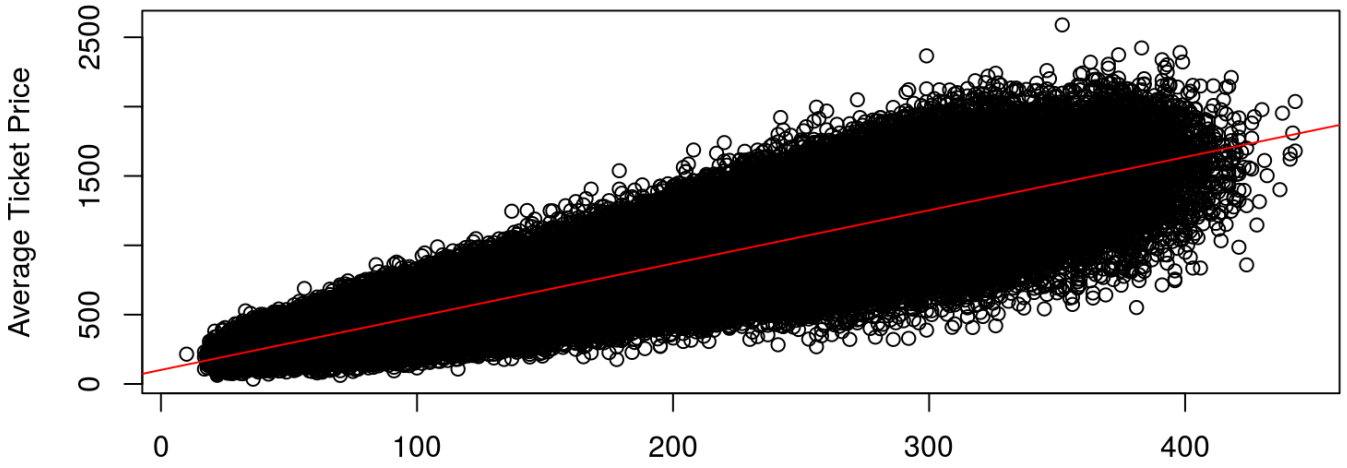
UA Distance Plot - Intercept : 227.987299123021 Slope: 0.555418930284021 MSE: 34763.34403527(



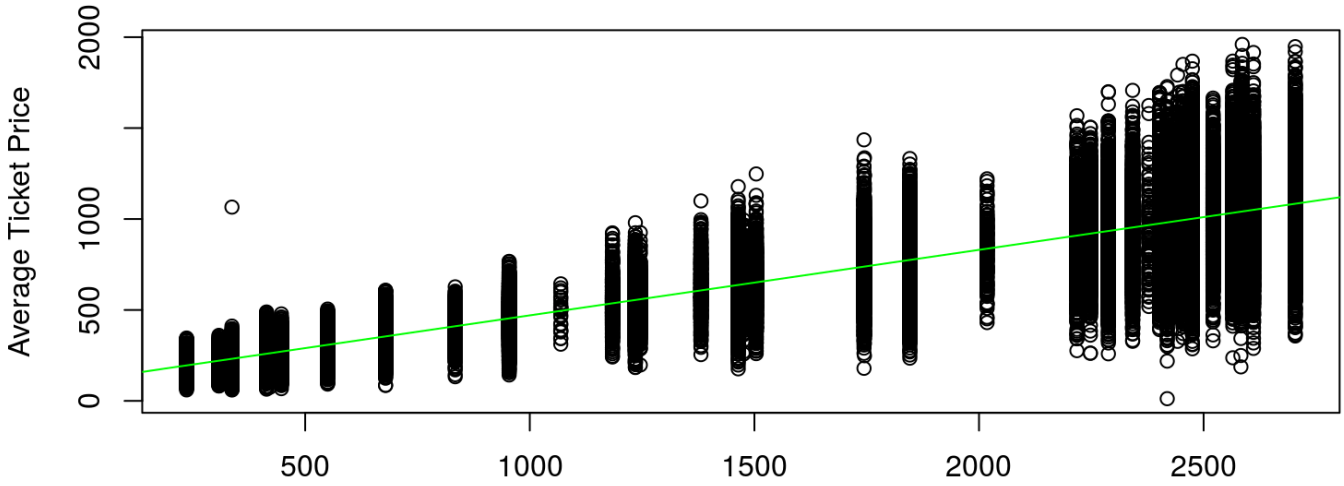
UA Time Plot - Intercept : 143.334713805998 Slope: 4.75801240098415 MSE: 31938.4678665511



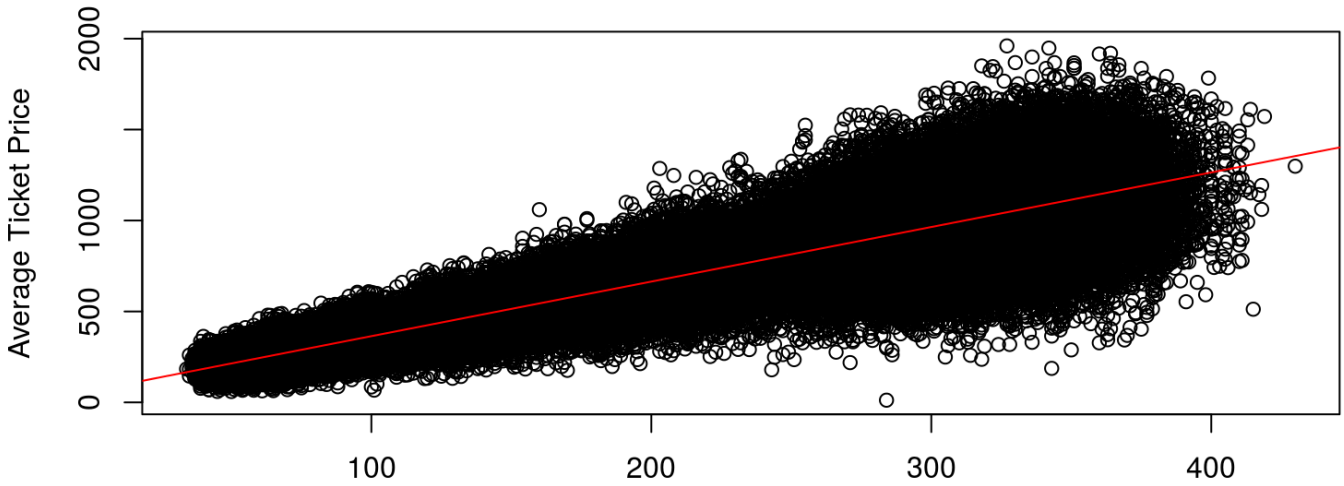
US Distance Plot - Intercept : 171.755791213112 Slope: 0.450953692436455 MSE: 15267.759472313



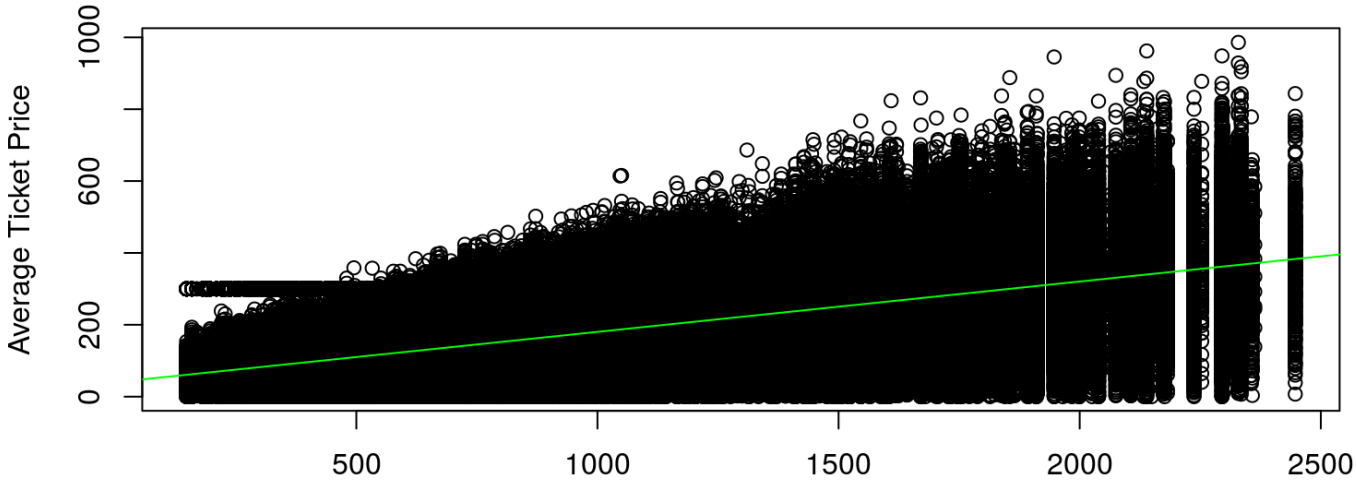
US Time Plot - Intercept : 100.921488173712 Slope: 3.83821929115149 MSE: 14408.6753384195



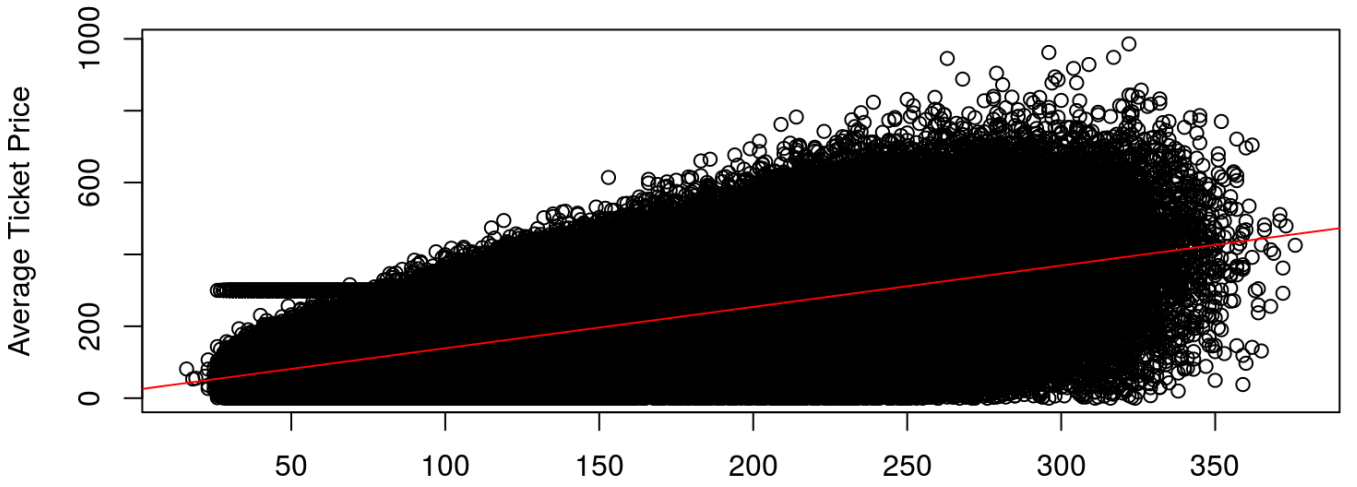
VX Distance Plot - Intercept : 109.857395412785 Slope: 0.360221069263937 MSE: 23789.92859698



VX Time Plot - Intercept : 63.8684288970711 Slope: 3.00074415445163 MSE: 21864.4935870901



WN Distance Plot - Intercept : 39.9666575532328 Slope: 0.140184207651658 MSE: 4043.581644638



WN Time Plot - Intercept : 23.7334954112909 Slope: 1.15062066674314 MSE: 3999.49982587404

Is distance traveled or flight time a better variable?

In order to decide which model is better for this dataset, we compare the MSE (Mean Square Error) values of each model, for each carrier. Of all the carriers seen so far 11 carriers have better fit for time. And 2 have a better fit for Distance.

MSE values can be calculated as follows:

```
# lrDistance <- lm(dfForCarrier$AvgPrice ~ dfForCarrier$Distance)
# summaryDistance <- summary(lrDistance)
# mseDistance <- mean(summaryDistance$residuals^2)

# lrTime <- lm(dfForCarrier$AvgPrice ~ dfForCarrier$Time)
# summaryTime <- summary(lrTime)
# mseTime <- mean(summaryTime$residuals^2)
```

We maintain a count for each variable, and increment it when the MSE for that is lower, as we iterate through all carriers

```
#  if(mseTime > mseDistance){
#    mseDistanceLowerCount <- mseDistanceLowerCount + 1
#  } else {
#    mseTimeLowerCount <- mseTimeLowerCount + 1
#  }
```

Thus we conclude the linear regression model for time has a better fit over the given data.

Cheapest Airline

We can use the time variable to find the cheapest airline. Using the LR Model for time, we can get an estimated ticket prices for each airline. (Explained in detail in Analysis.)

ANALYSIS

We used the MSE to dertermine the best fit. In most of the cases (for most of the carriers) in this dataset, time variable gives a better fit. To automate the process of identifying which of the two variables are better, we used a naive approach of counting the lower MSEs for each carrier. (Discussed earlier.)

Once the better variable was identified, we need to find which airline is the cheapest. For this, we used the mean value of Time, from the enitre dataset. Here, 111.055472. Using this value, we predict the price for each carrier using the Linear Regression model for time.

```
# timePredVal <- lrTime$coefficients[[2]]*timePredFor +  lrTime$coefficients[[1]]
```

The slope is obainted from lrTime $coefficients[[2]], and intercept from lrTime $coefficients[[1]]

We store the timePredVal for each carrier in a map. Later, using this map, we find the cheapest airline, and rankings for each airline.

Output:

The cheapest airline in this dataset: F9

The rakings are given:

```
for (x in strsplit(finalResultList, ",")) {
  print(paste(x[2],x[3]))
}
```

```
## [1] "F9 1"
## [1] "AS 2"
## [1] "WN 3"
## [1] "HA 4"
## [1] "MQ 5"
## [1] "EV 6"
## [1] "OO 7"
## [1] "VX 8"
## [1] "B6 9"
## [1] "AA 10"
## [1] "US 11"
## [1] "DL 12"
## [1] "UA 13"
```

Conclusion:

We are using MSE to determine the best fit. In this case, the time variable has a better linear regression model. Hence, we use this to predict prices of all carriers. We are predicting based on the mean value of time from the entire dataset(here, 111.05). Of all the predicted prices(for all carriers) F9 gives the cheapest predicted price. Hence, F9 is the cheapest.

References:

For calculating MSE using summary of model given by lm.
http://stats.stackexchange.com/questions/107643/how-to-get-the-value-of-mean-squared-error-in-a-linear-regression-in-r (http://stats.stackexchange.com/questions/107643/how-to-get-the-value-of-mean-squared-error-in-a-linear-regression-in-r)