A4: Linear Regression

Harshali Singh, Vishal Mehta

Implementation

In the Mapper, we have filtered out the records with year less than 2010 or greater than 2015 as we do not need to take them into consideration. Key: Carrier Code Value: Year, Avg Ticket Price, Distance Traveled, Actual Elapsed Flight Time All the pairs which pass the sanity Test are sent to Reducer.

In Reducer, we first check if the airline is active in 2015 (we check all the values for Year 2015). For the airline which was active in 2015, all the values for that carrier (for year 2010-1014) are reduced and written to output.

Linear Regression

The R script reads all the *.txt files containing inidividual carrier data.

We have computed 2 linear regression models for each carrier in R. 1. Price (Response variable), Distance Traveled(Explanatory variable) 2. Price (Response variable), Flight Time(Explanatory variable)

```
## Loading required package: methods
```

Graph Plot

There are 26 plots generated (2 for each carrier). The Plots can be seen attached at the end of report.

Conclusion

```
## [1] "Distance is a better variable"
## [1] "Adjusted R-Squared for Price-Distance: 0.694110649045351"
## [1] "Adjusted R-Squared for Price-Time: 0.690103550137566"
## [1] "New Ranking of Airlines:"

##    Carrier      Slope
## 6       F9 0.03898392
## 13      WN 0.13759992
## 2       AS 0.14233357
## 3       B6 0.34554326
## 1       AA 0.34952469
## 7       HA 0.35839949
## 9       OO 0.36011867
## 12      VX 0.36415036
## 8       MQ 0.36767280
## 5       EV 0.37524156
## 11      US 0.45000596
## 4       DL 0.46287182
## 10      UA 0.55801289
```

Better Variable

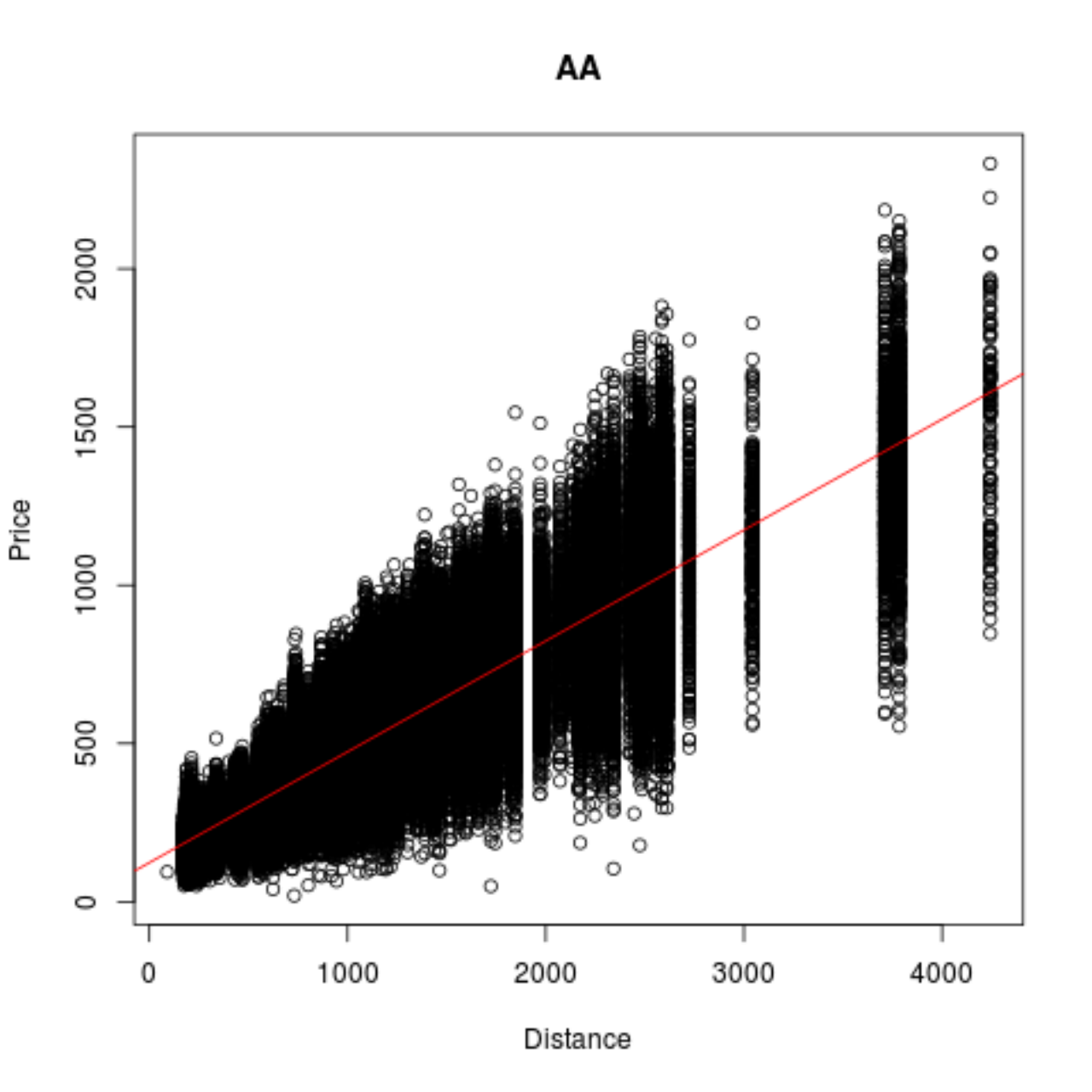
R-squared (also known as the Coefficient of Determination) is the “percent of variance explained” by the model.

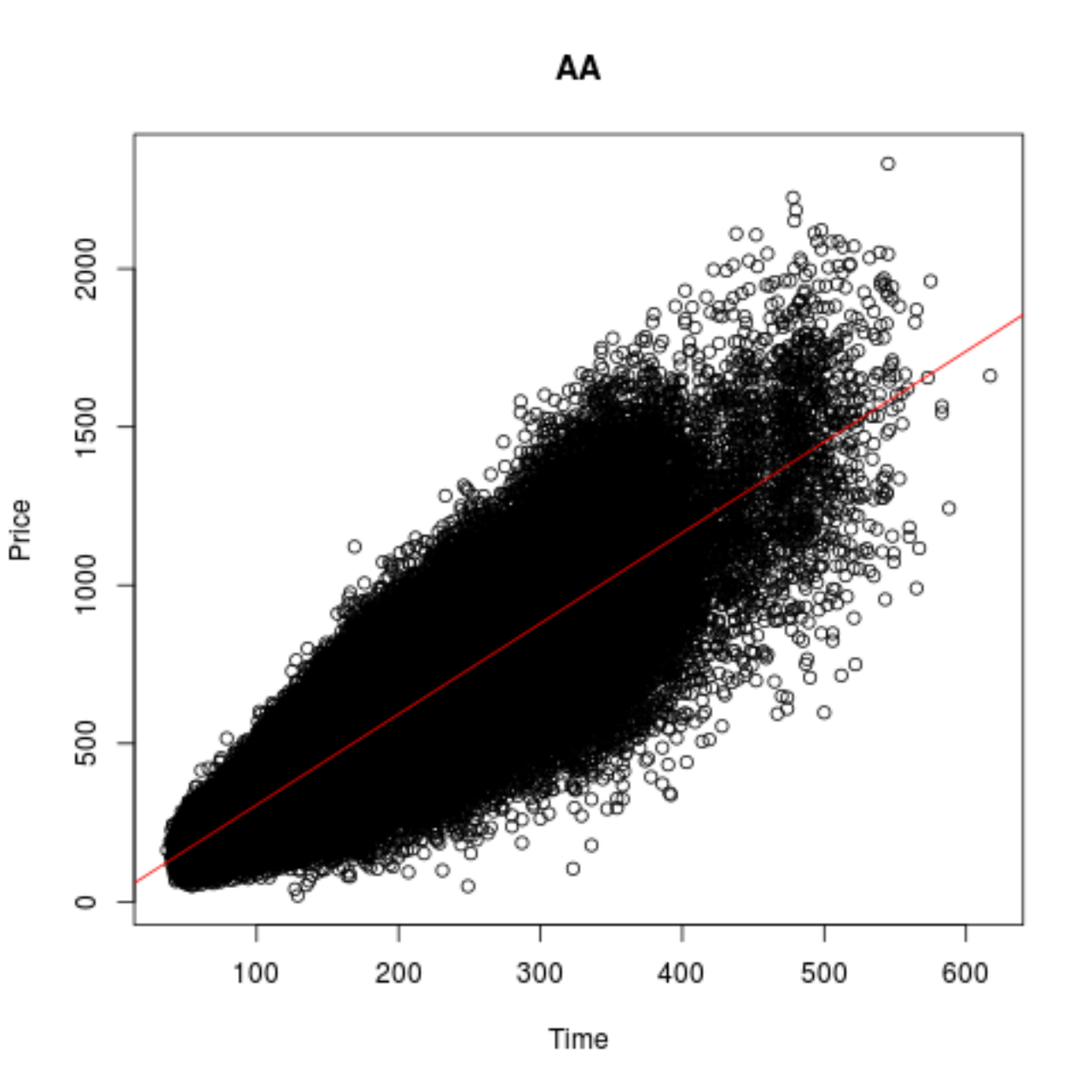
We see from the output that our Price-Distance model causes 69.4% of the variation in price using the Distance as the explanatory variable. The Price-Time model causes 69.0% of the variation in price using Time as the explanatory variable. Higher the adjusted R-squared, higher the percent that is the closest to the line of best fit. Hence, among the two (Distance or Time) Distance is a better explanatory variable.

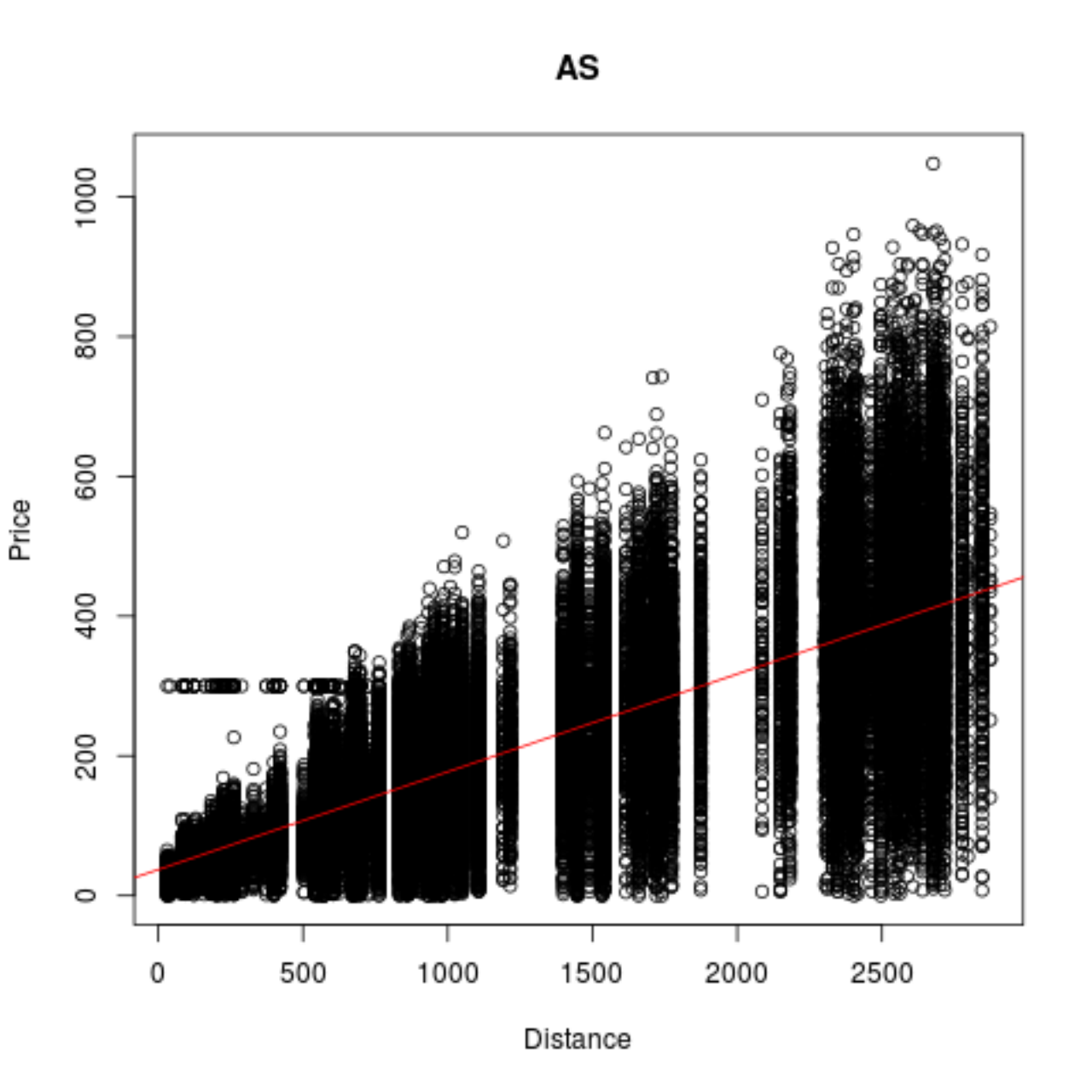
We have used Adjusted R-squared to decide the model’s explanatory power. Adjusted R-squared is an unbiased estimate of the fraction of variance explained, taking into account the sample size and number of variables.

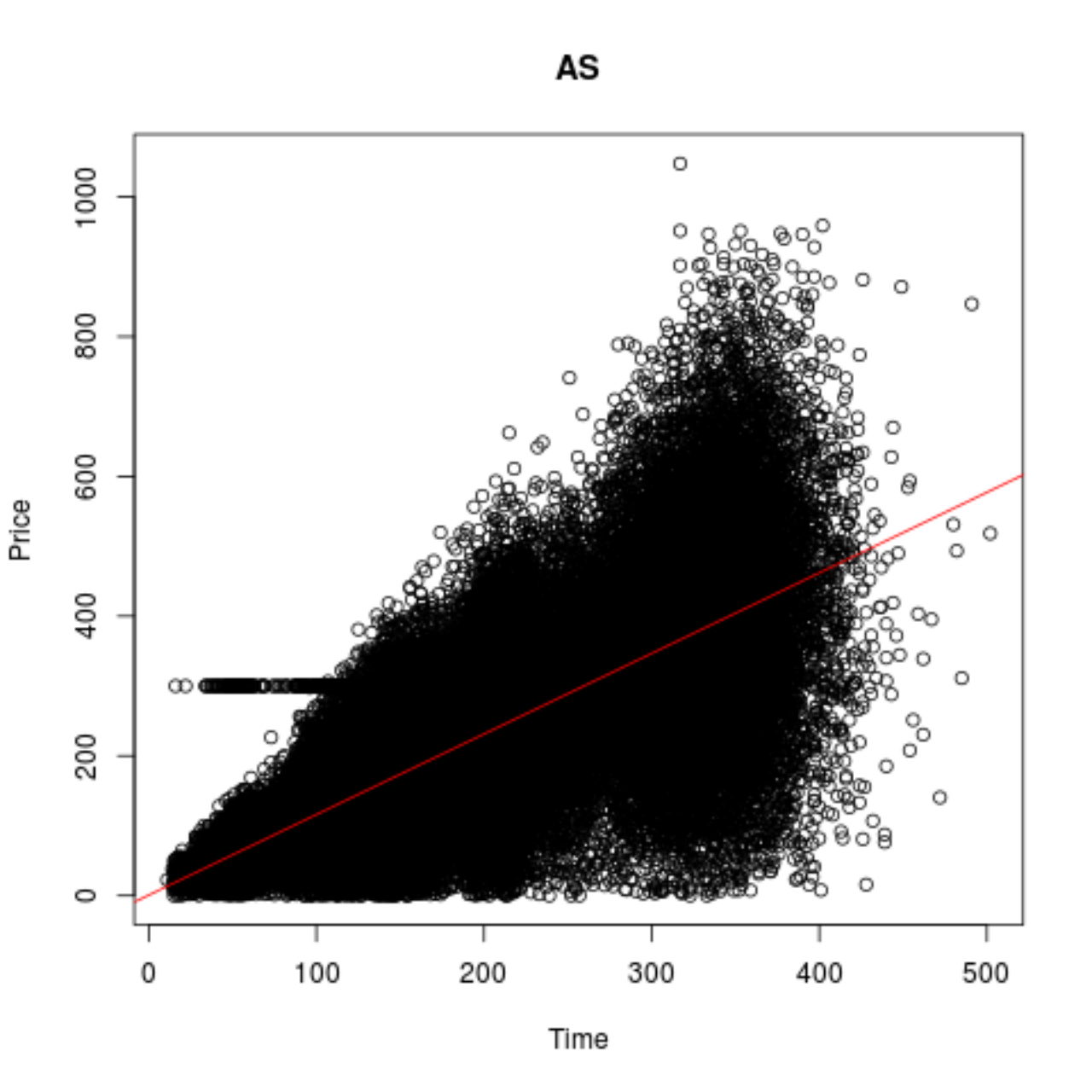
Cheapest Flight

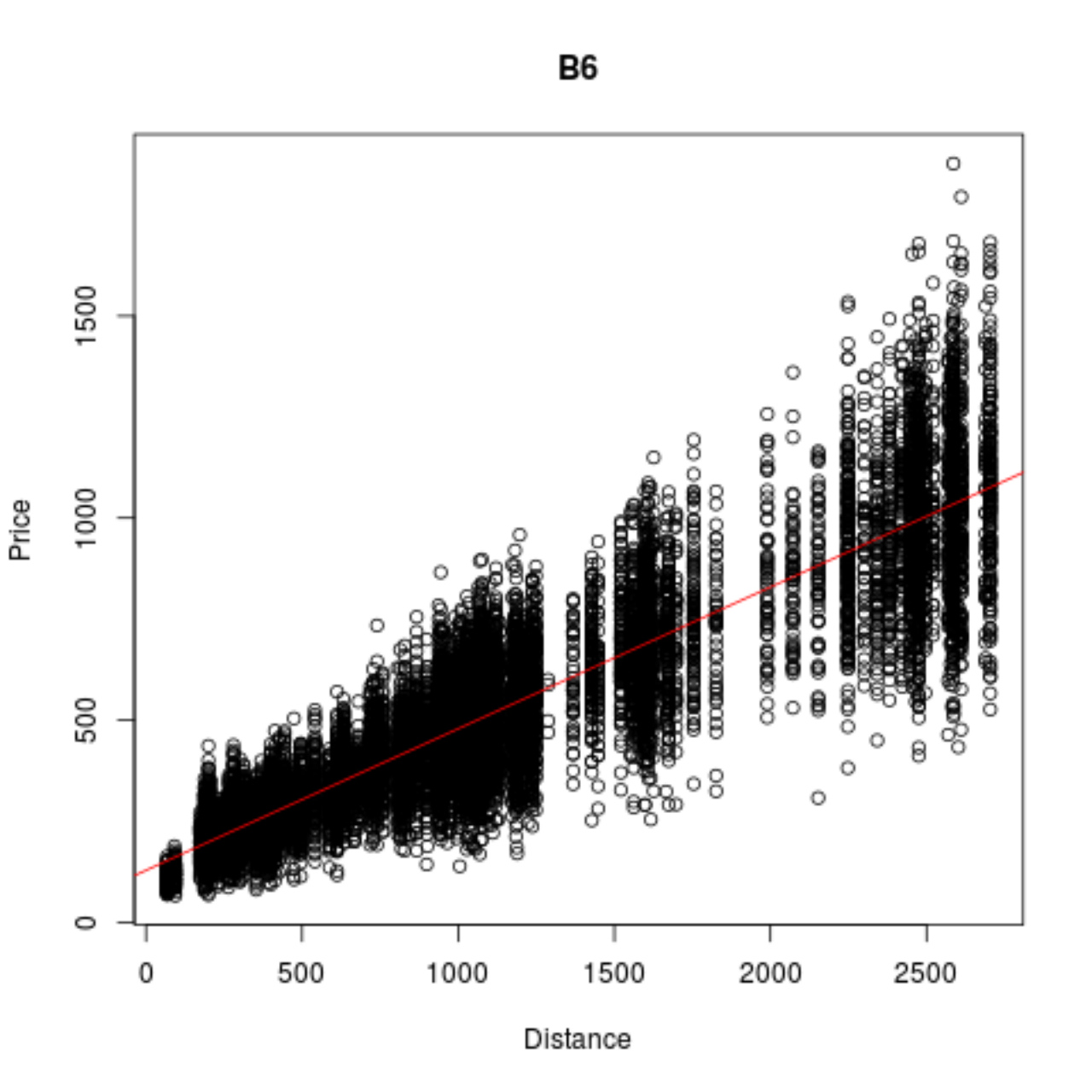
After deciding the better regression model, we have considered Slope and Intercept as the factor to rank the airlines. Flight with the minimum slope and intercept can be said as the cheapest flight which is Frontier Airline, F9.

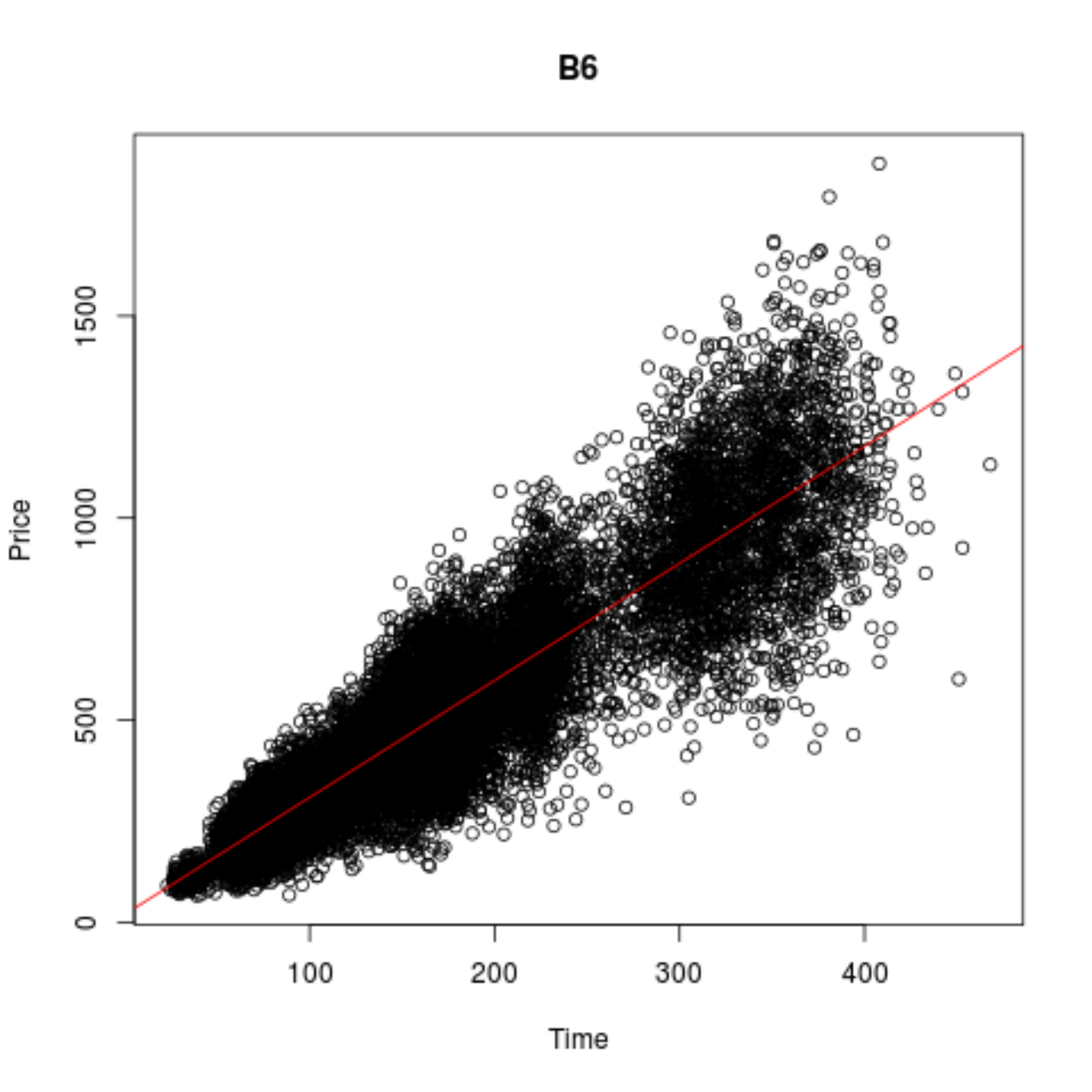

AA
Price
0
500
1000
1500
2000
0
1000
2000
3000
4000
Distance

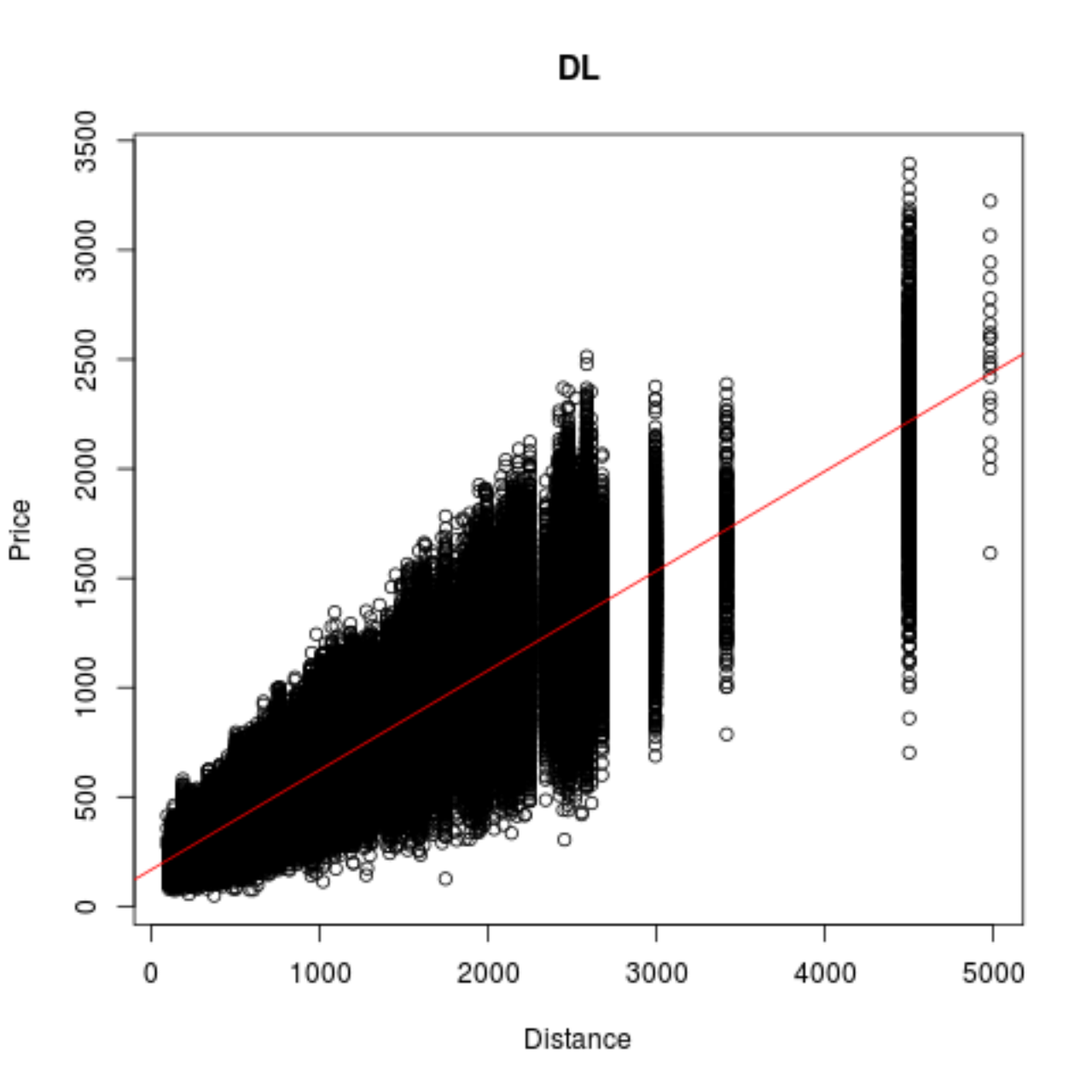

AA
Price
0
500
1000
1500
2000
100
200
300
400
500
600
Time

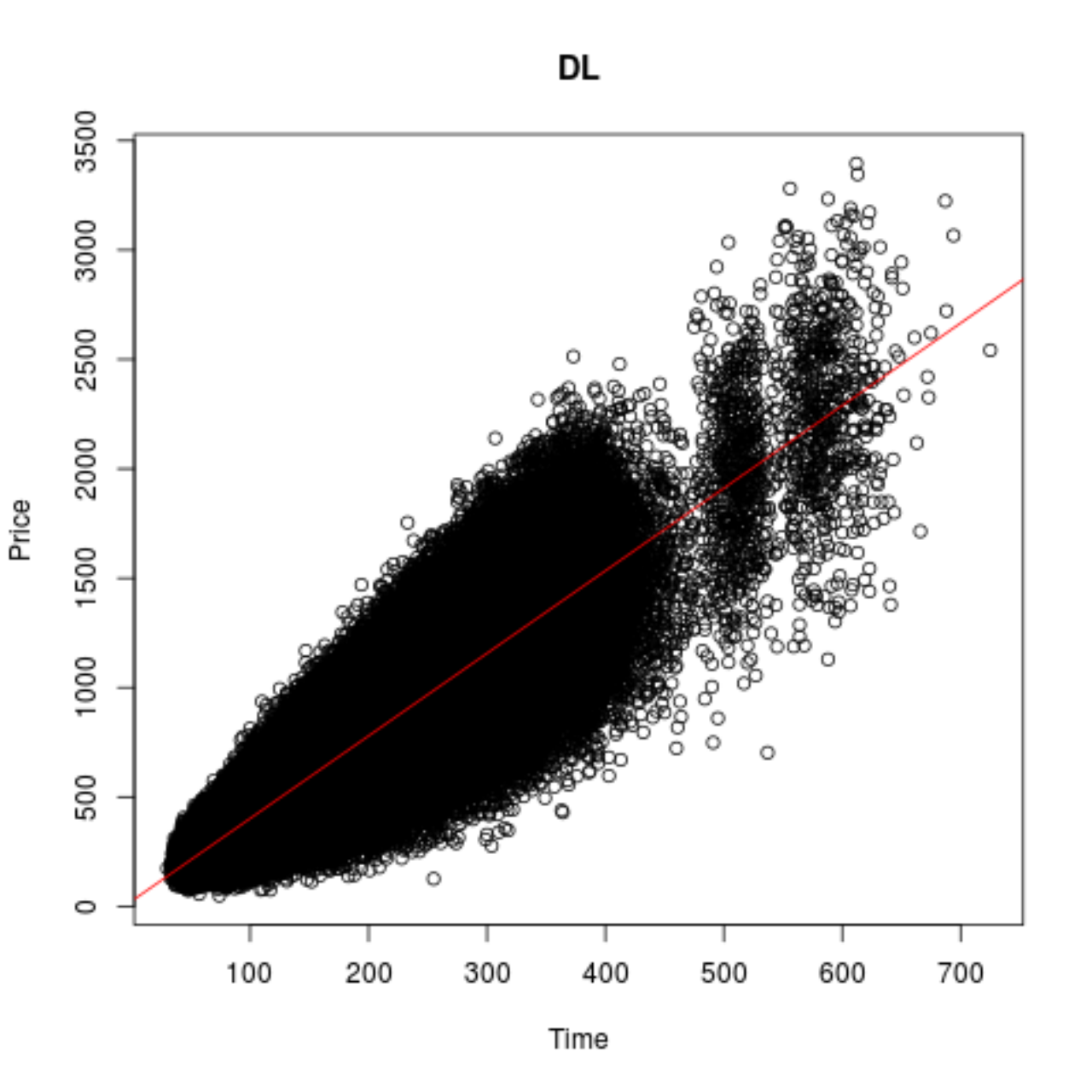

AS
Price
Distance
0
200
400
600
800
1000
0
500
1000
1500
2000
2500

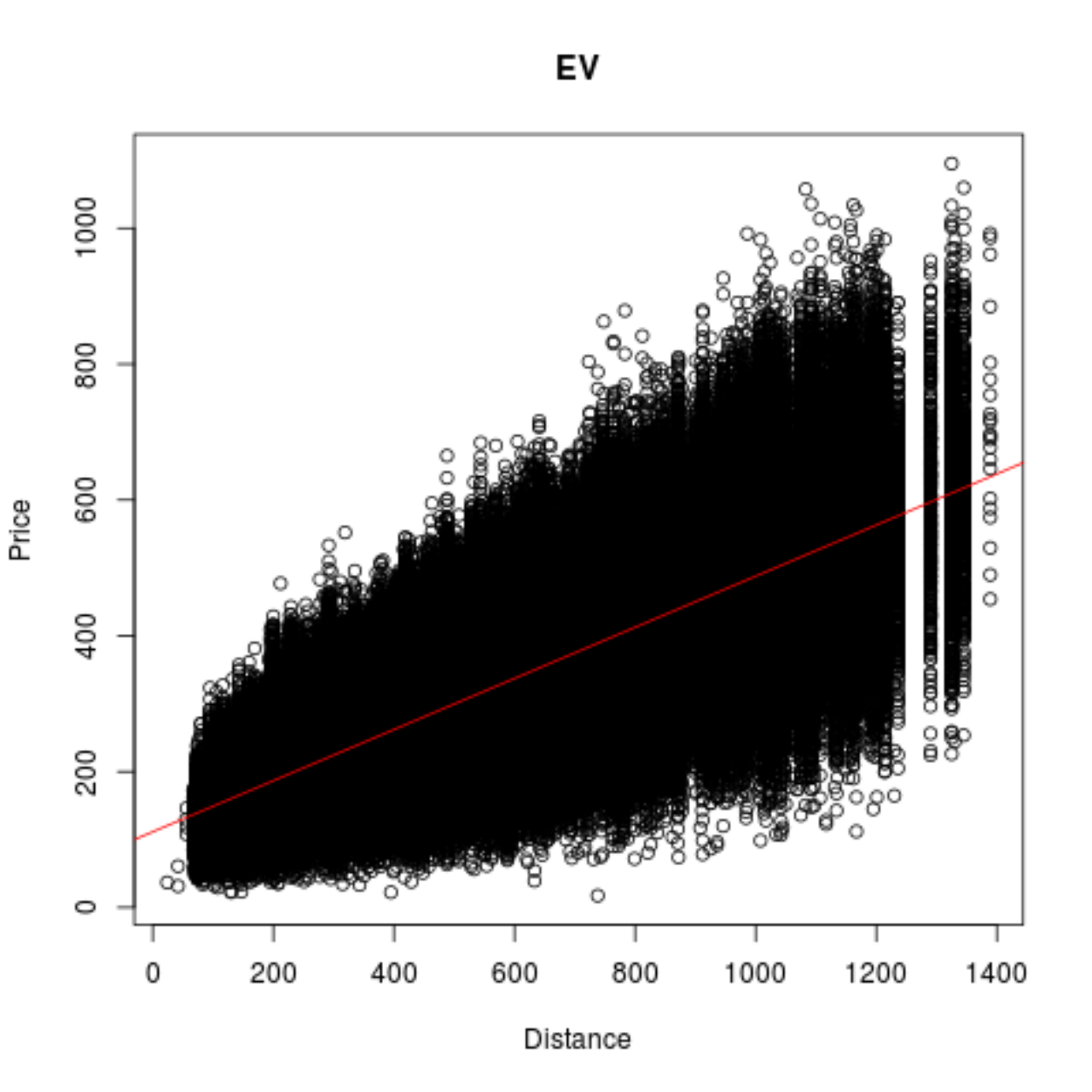

AS
Price
Time
0
200
400
600
800
1000
0
100
200
300
400
500

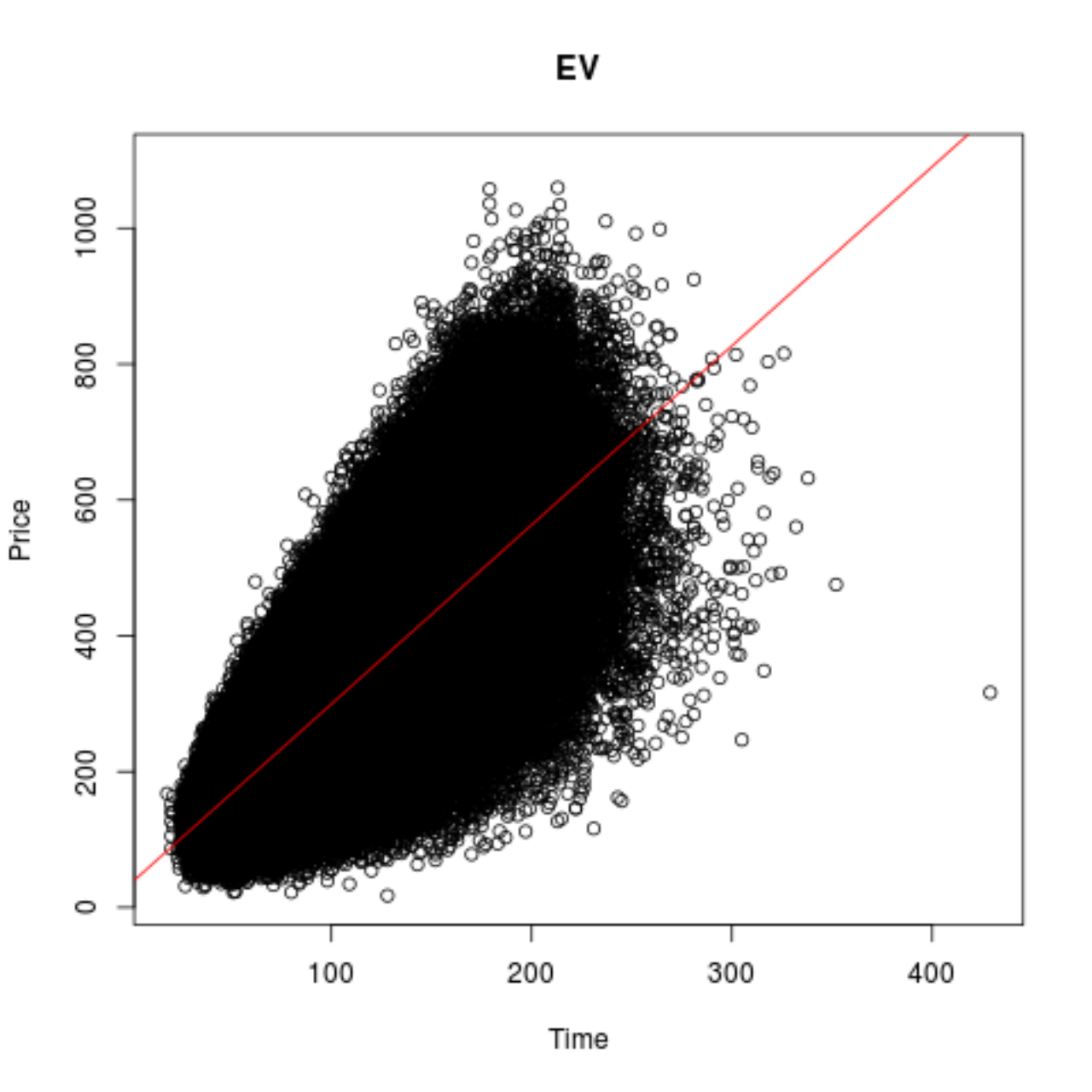

B6
Price
0
500
1000
1500
0
500
1000
1500
2000
2500
Distance

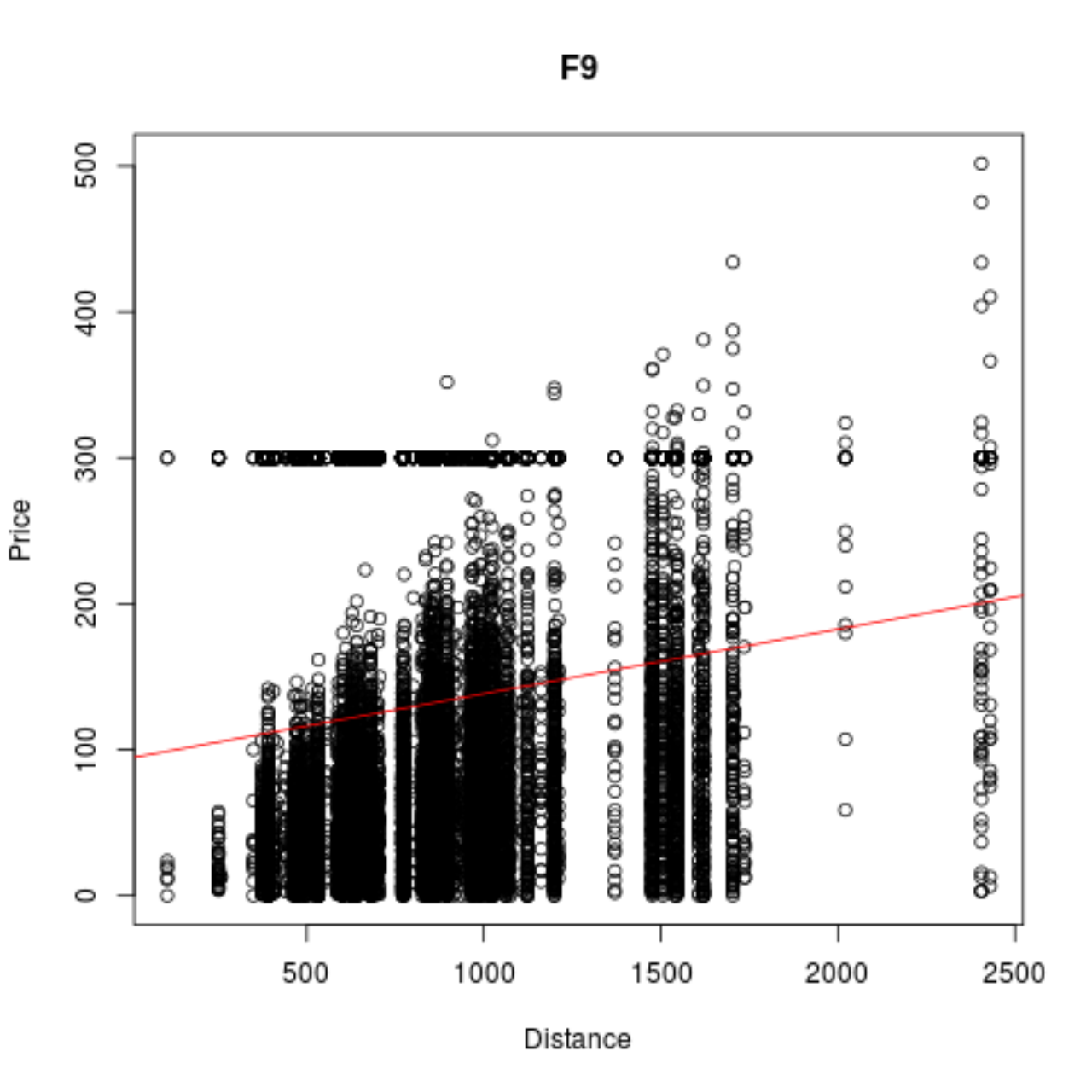

B6
Price
0
500
1000
1500
100
200
300
400
Time

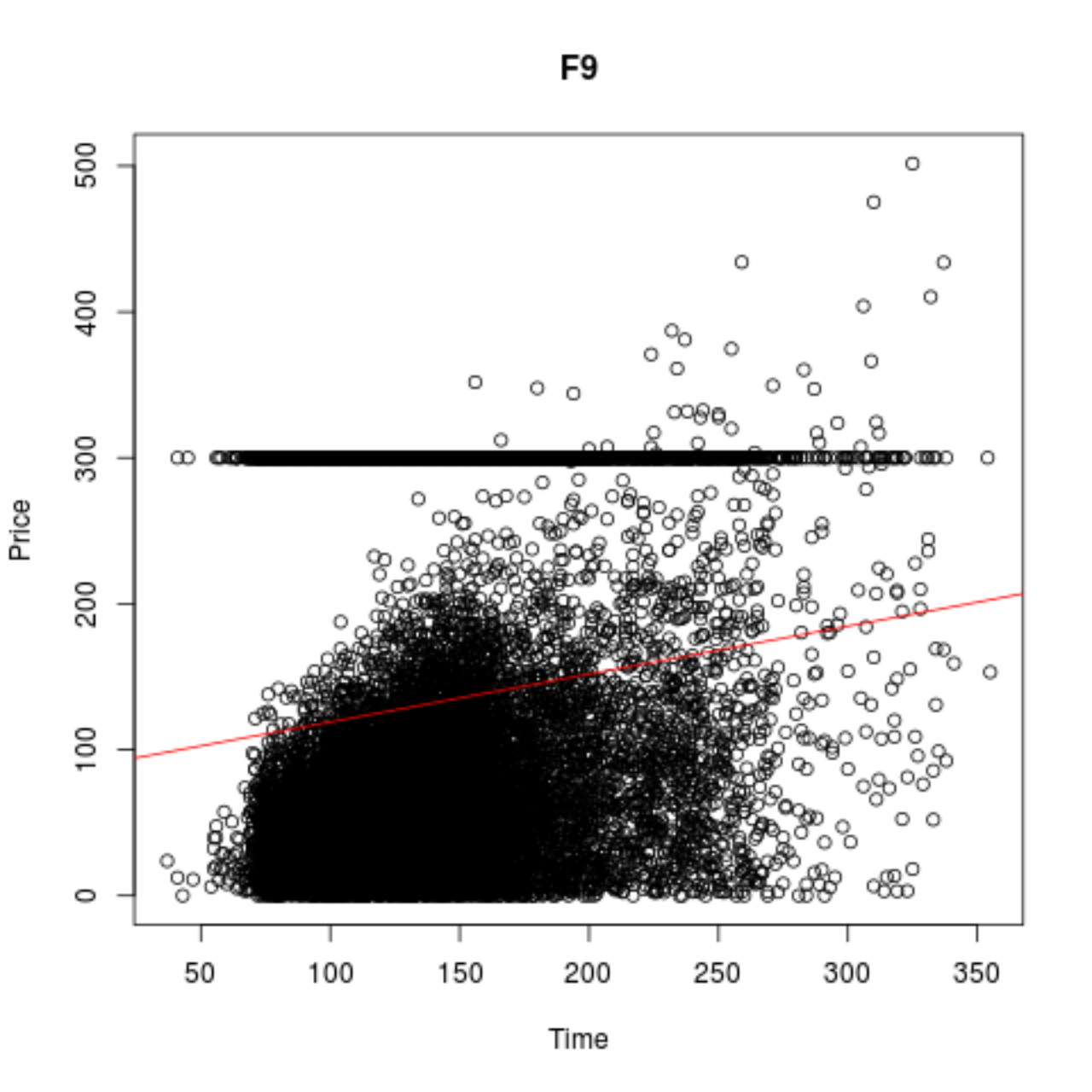

DL
Price
Distance
0
500
1000
1500
2000
2500
3000
3500
0
1000
2000
3000
4000
5000

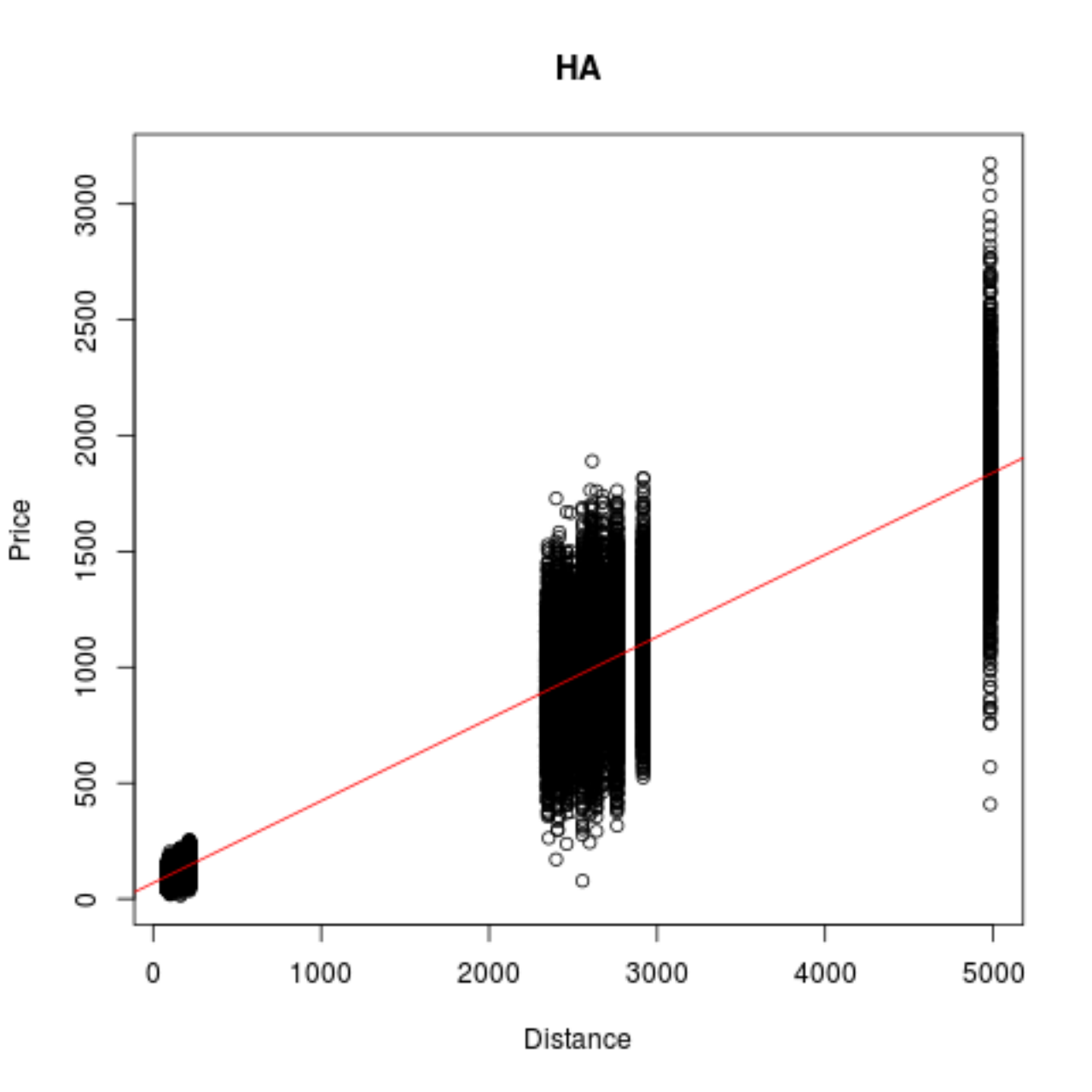

DL
Price
Time
0
500
1000
1500
2000
2500
3000
3500
100
200
300
400
500
600
700

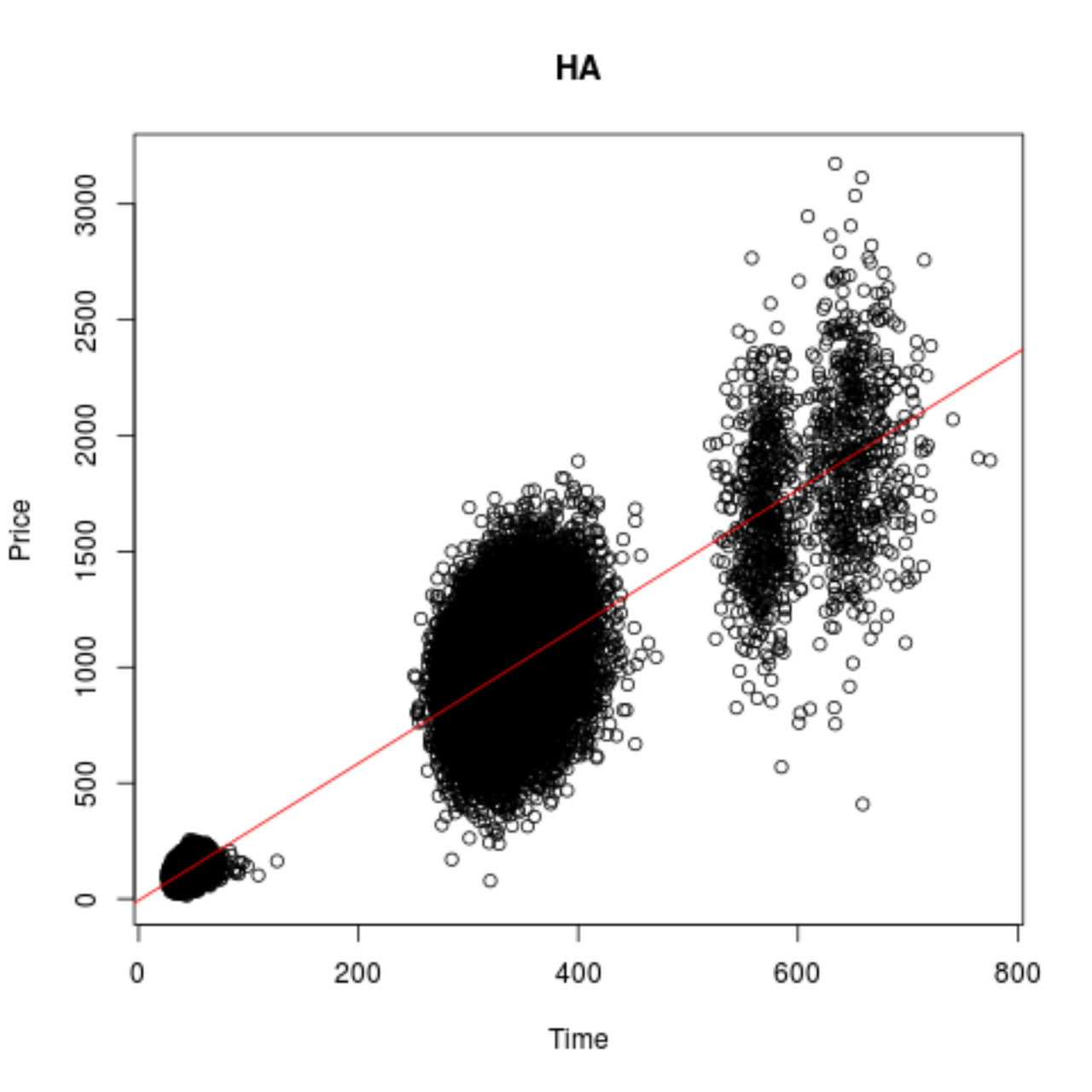

EV
Price
Distance
0
200
400
600
800
1000
1200
1400

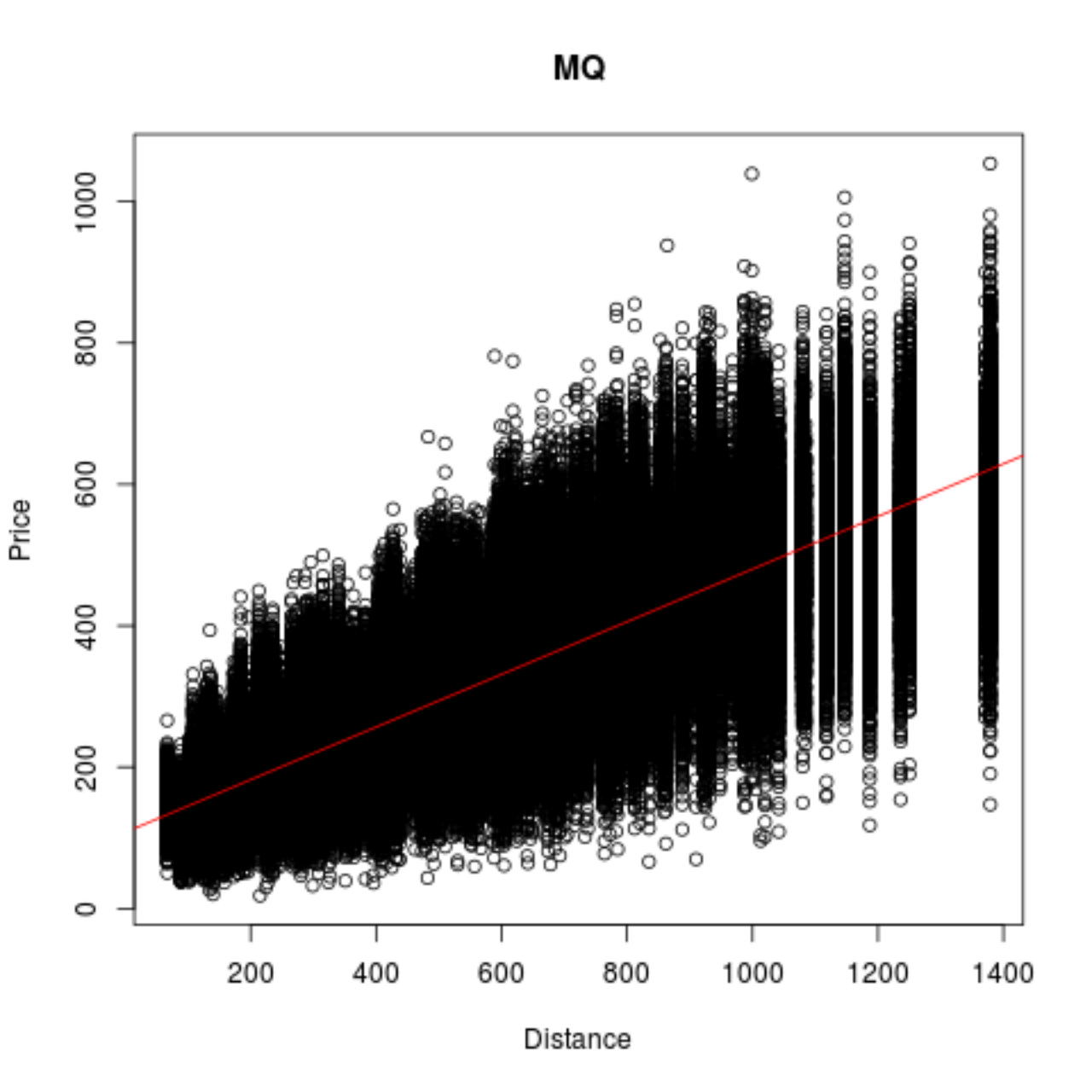

EV
Price
0
200
400
600
800
1000
100
200
300
400
Time

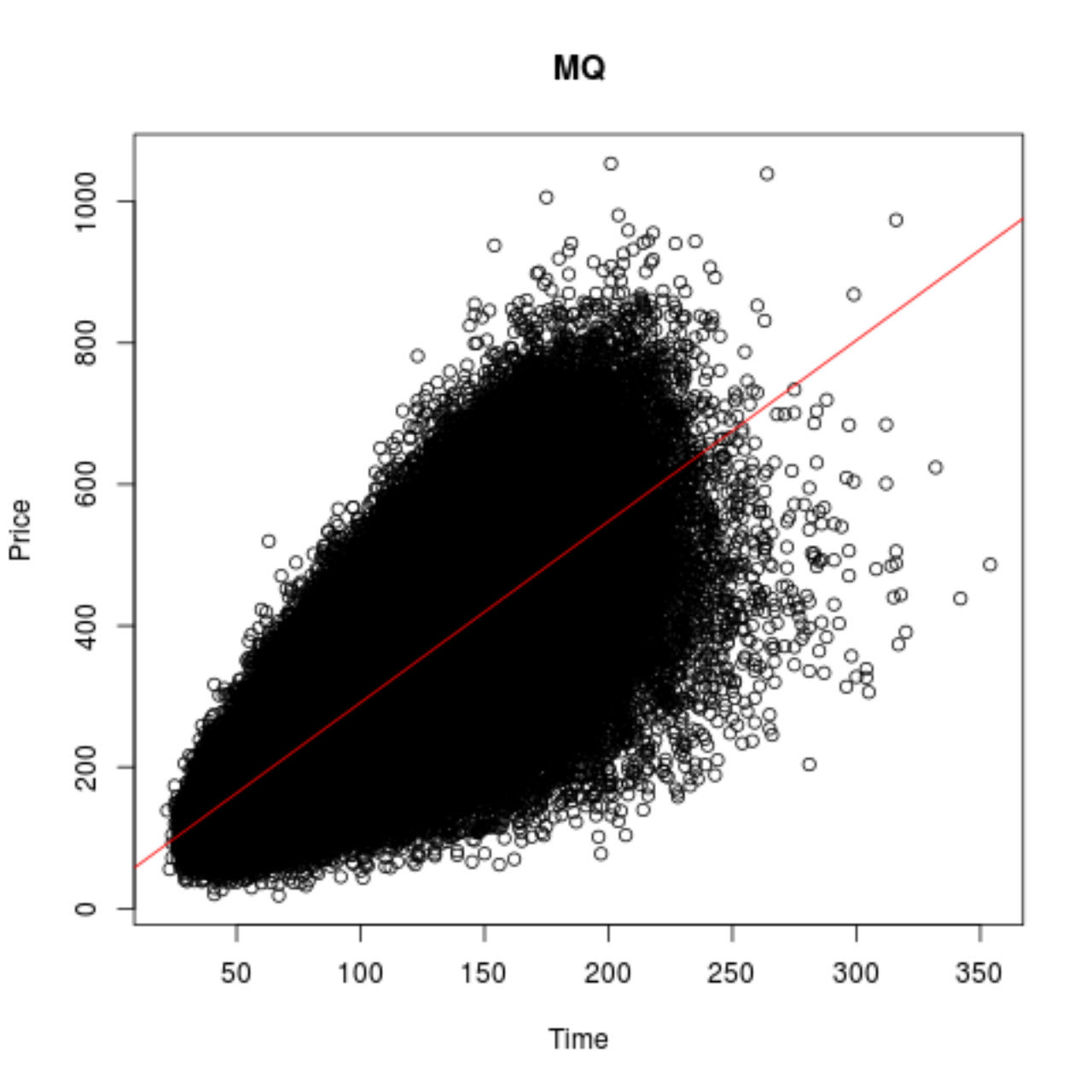

F9
Price
Distance
0
100
200
300
400
500
500
1000
1500
2000
2500

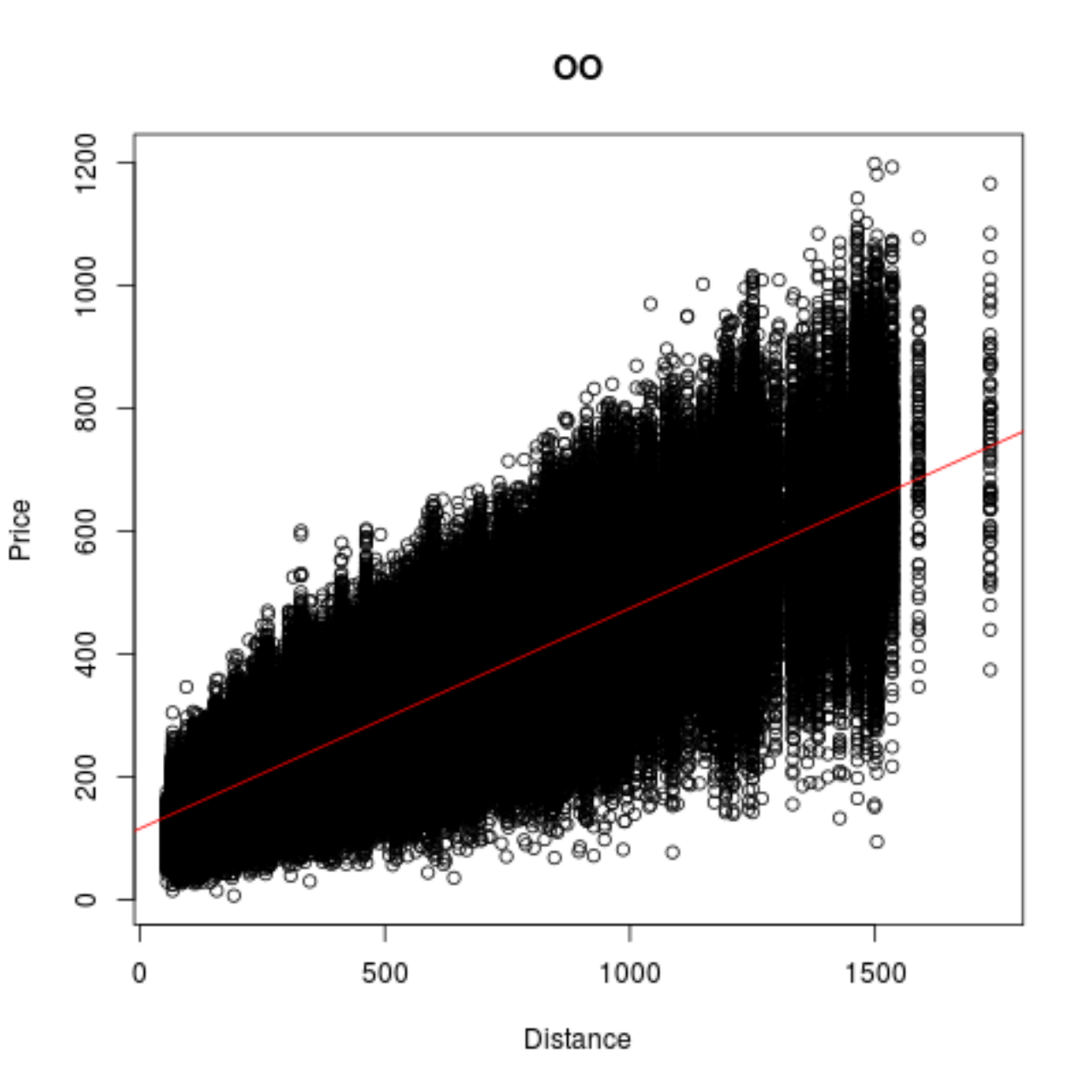

F9
Price
Time
0
100
200
300
400
500
50
100
150
200
250
300
350

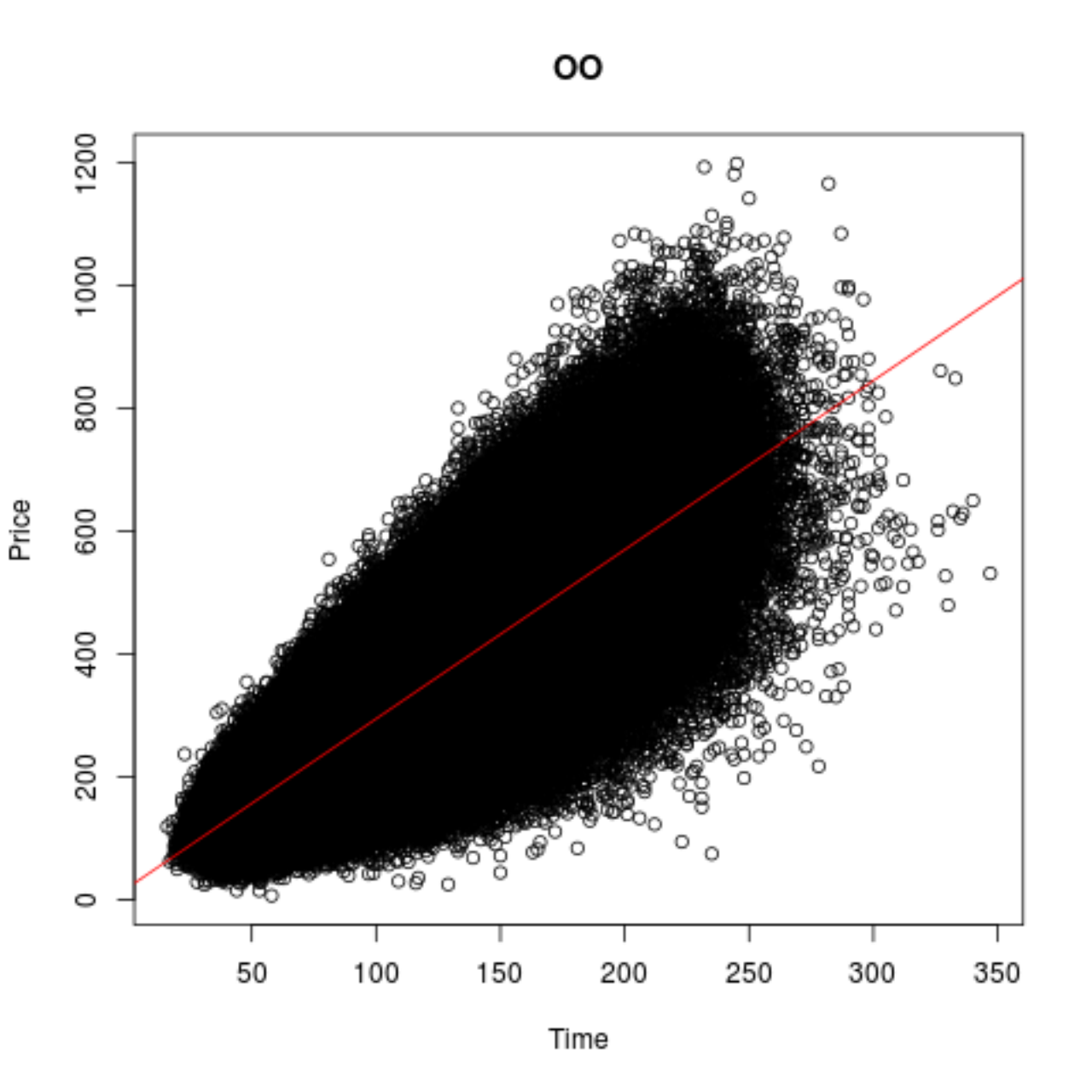

HA
Price
0
500
1000
1500
2000
2500
3000
0
1000
2000
3000
4000
5000
Distance

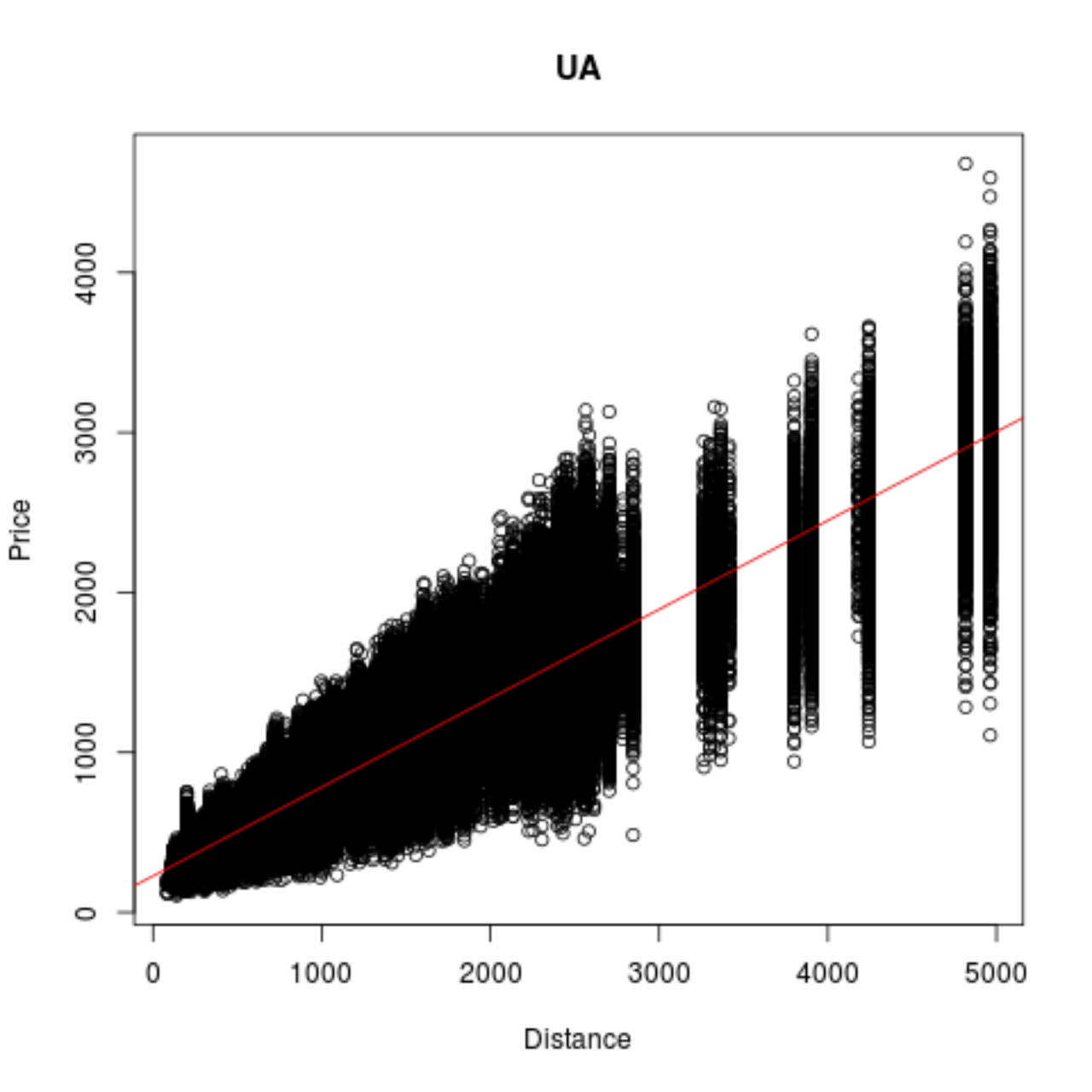

HA
Price
0
500
1000
1500
2000
2500
3000
0
200
400
600
800
Time

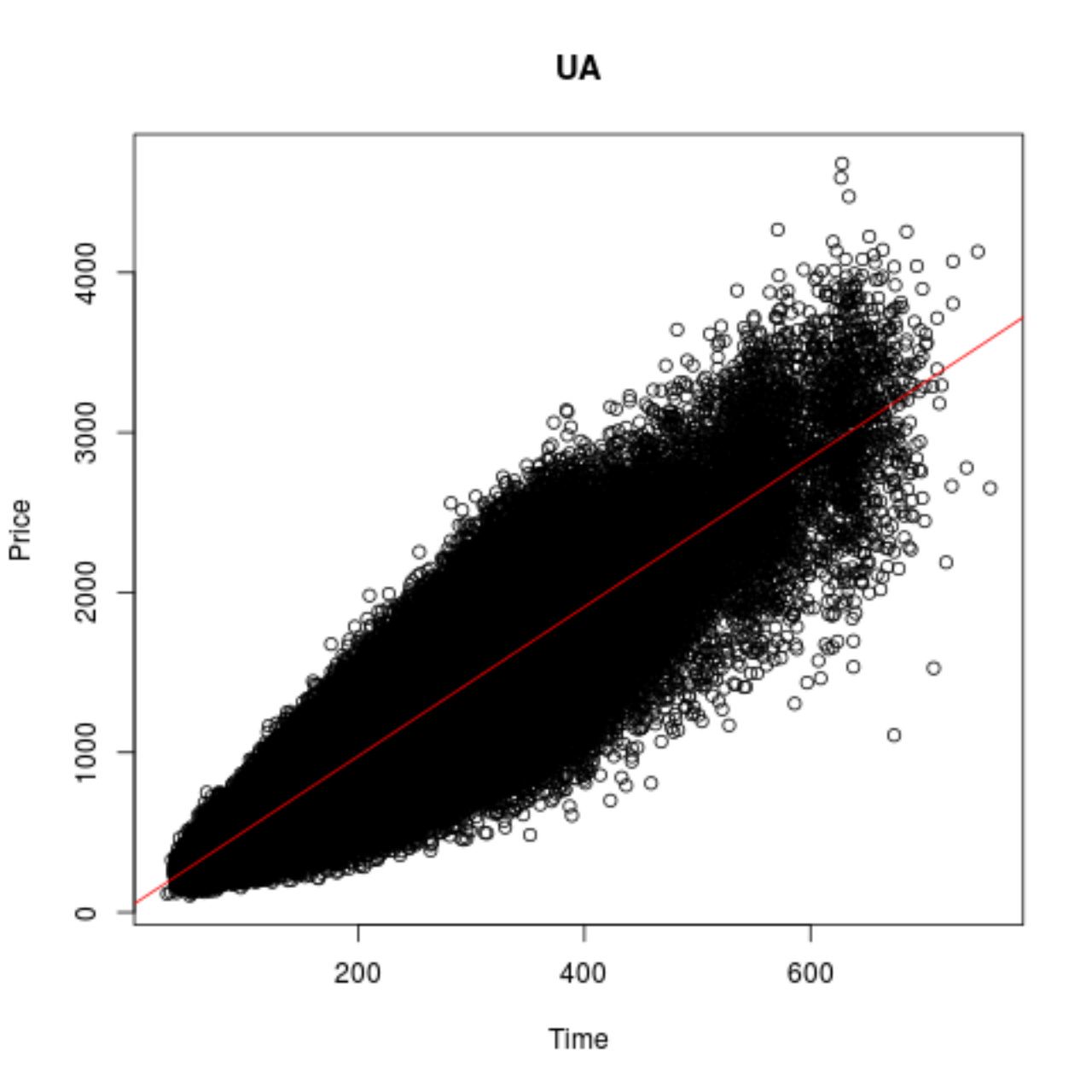

MQ
Price
Distance
0
200
400
600
800
1000
200
400
600
800
1000
1200
1400

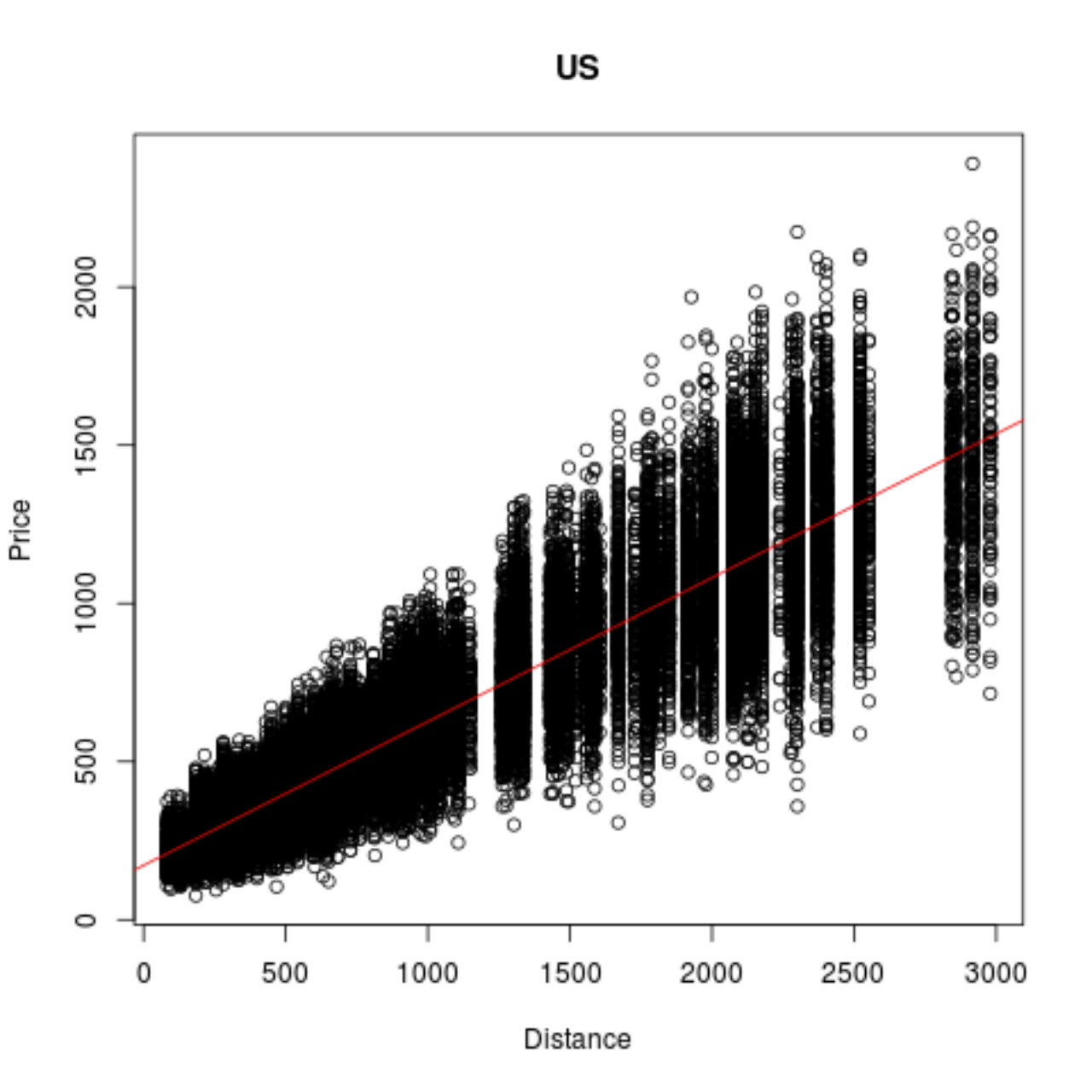

MQ
Price
Time
0
200
400
600
800
1000
50
100
150
200
250
300
350

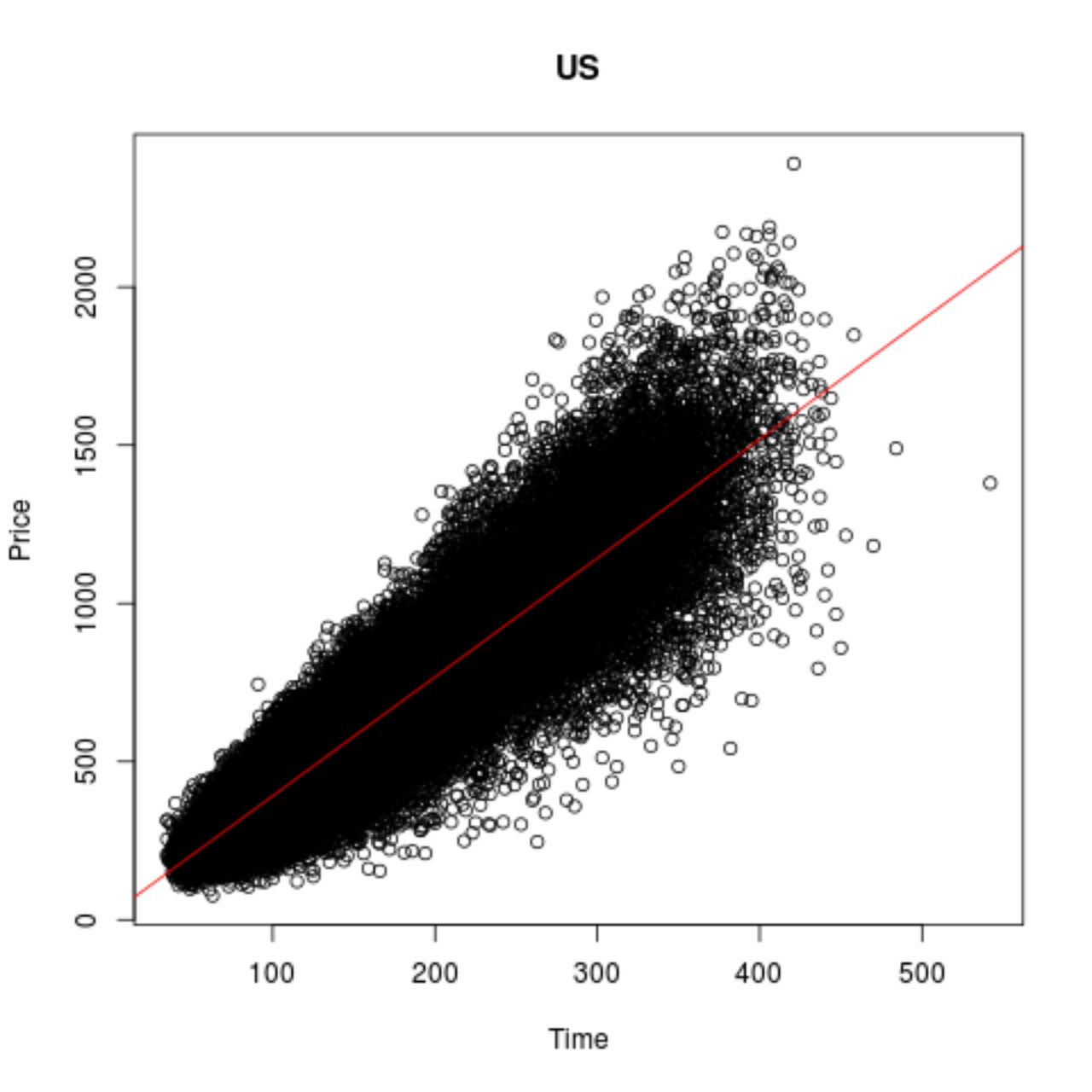

OO
Price
Distance
0
200
400
600
800
1000
1200
0
500
1000
1500

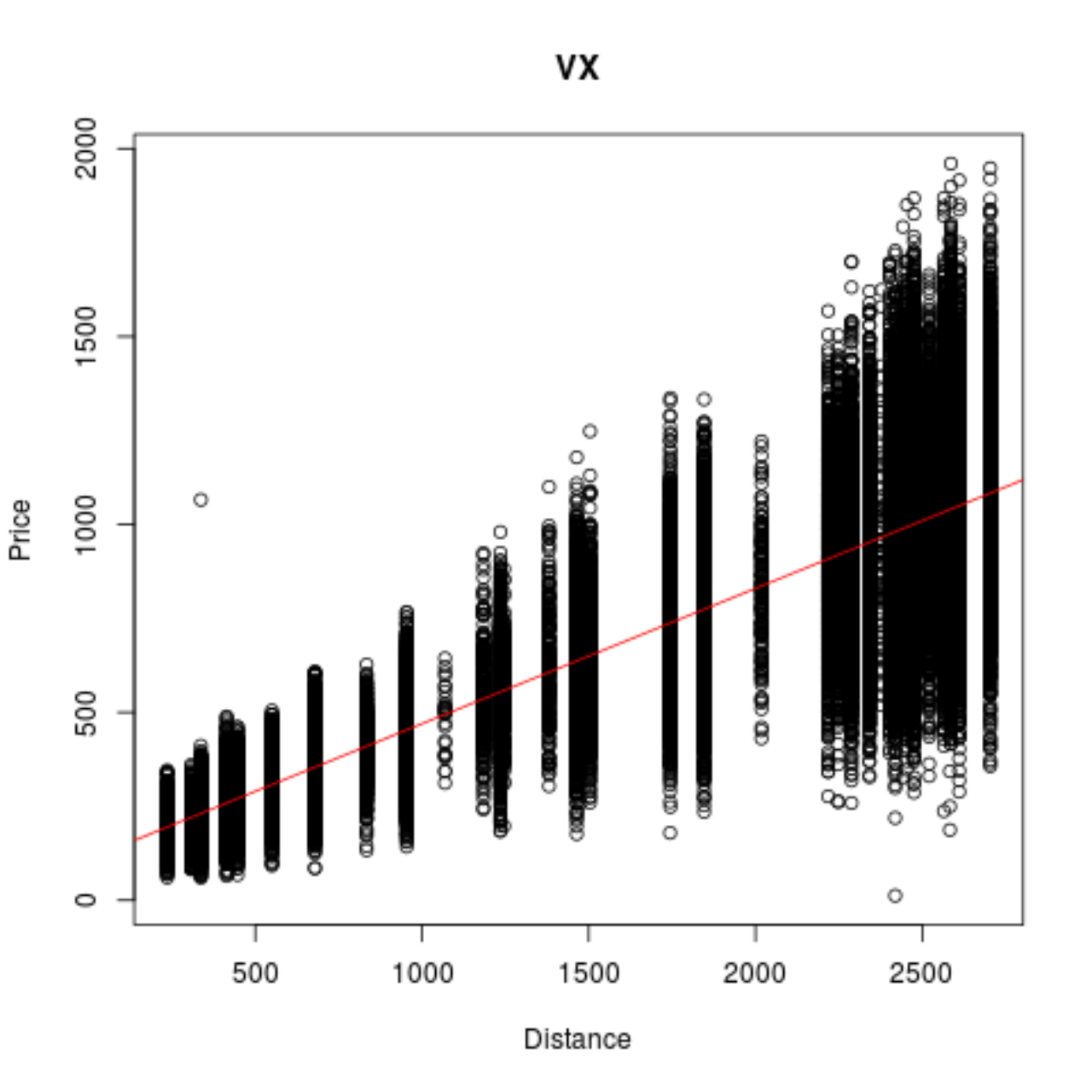

OO
Price
Time
0
200
400
600
800
1000
1200
50
100
150
200
250
300
350

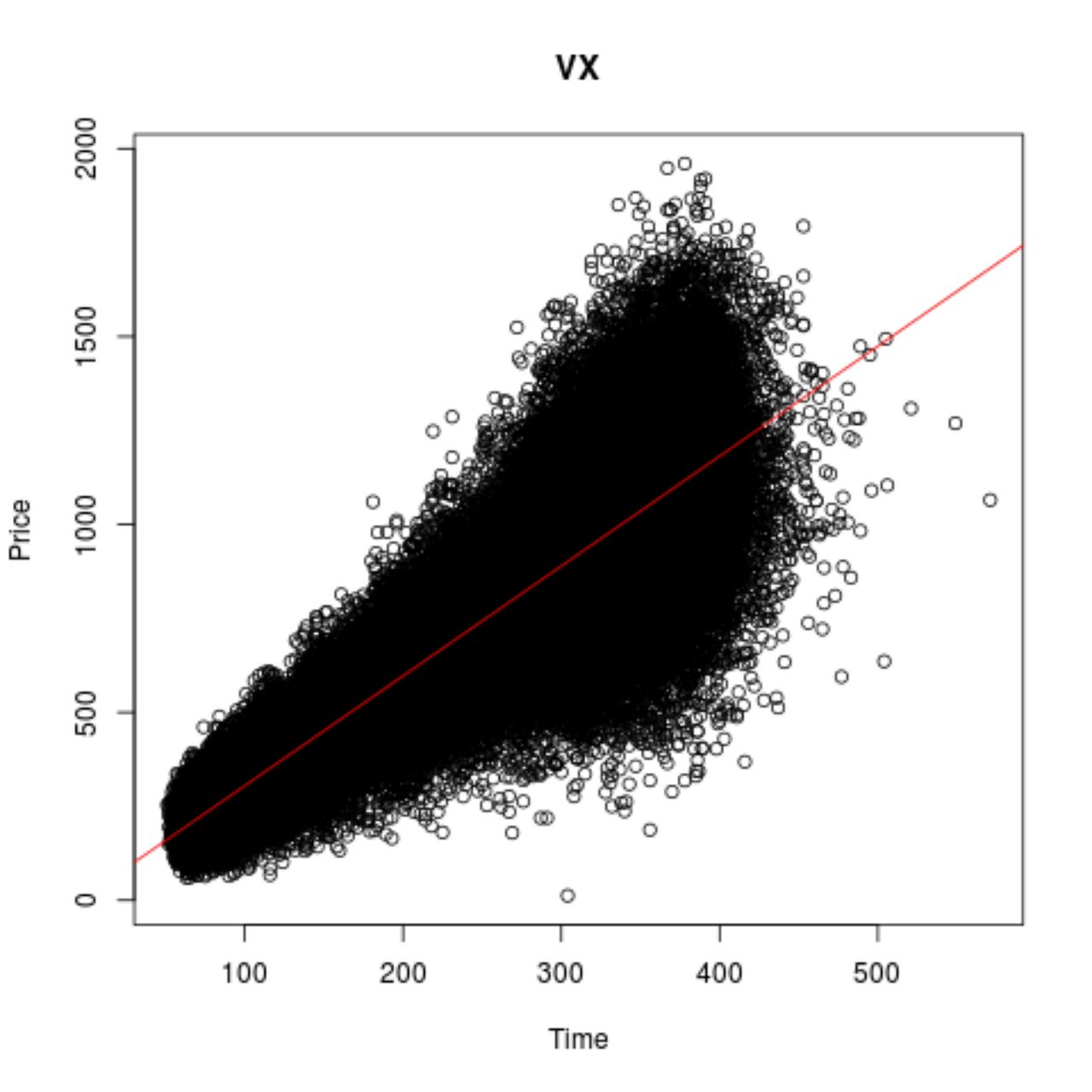

UA
Price
Distance
0
1000
2000
3000
4000
5000

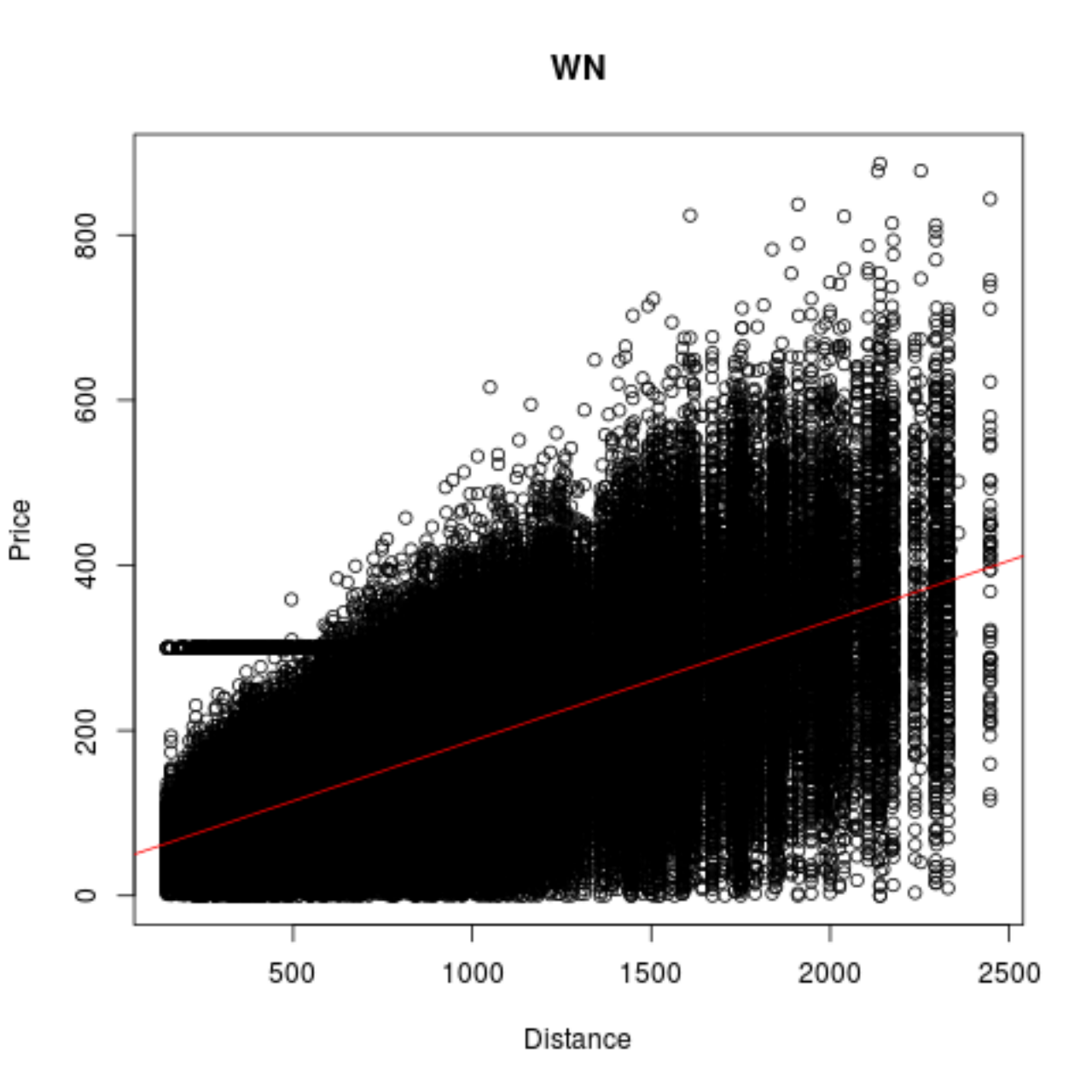

UA
Price
0
1000
2000
3000
4000
200
400
600
Time

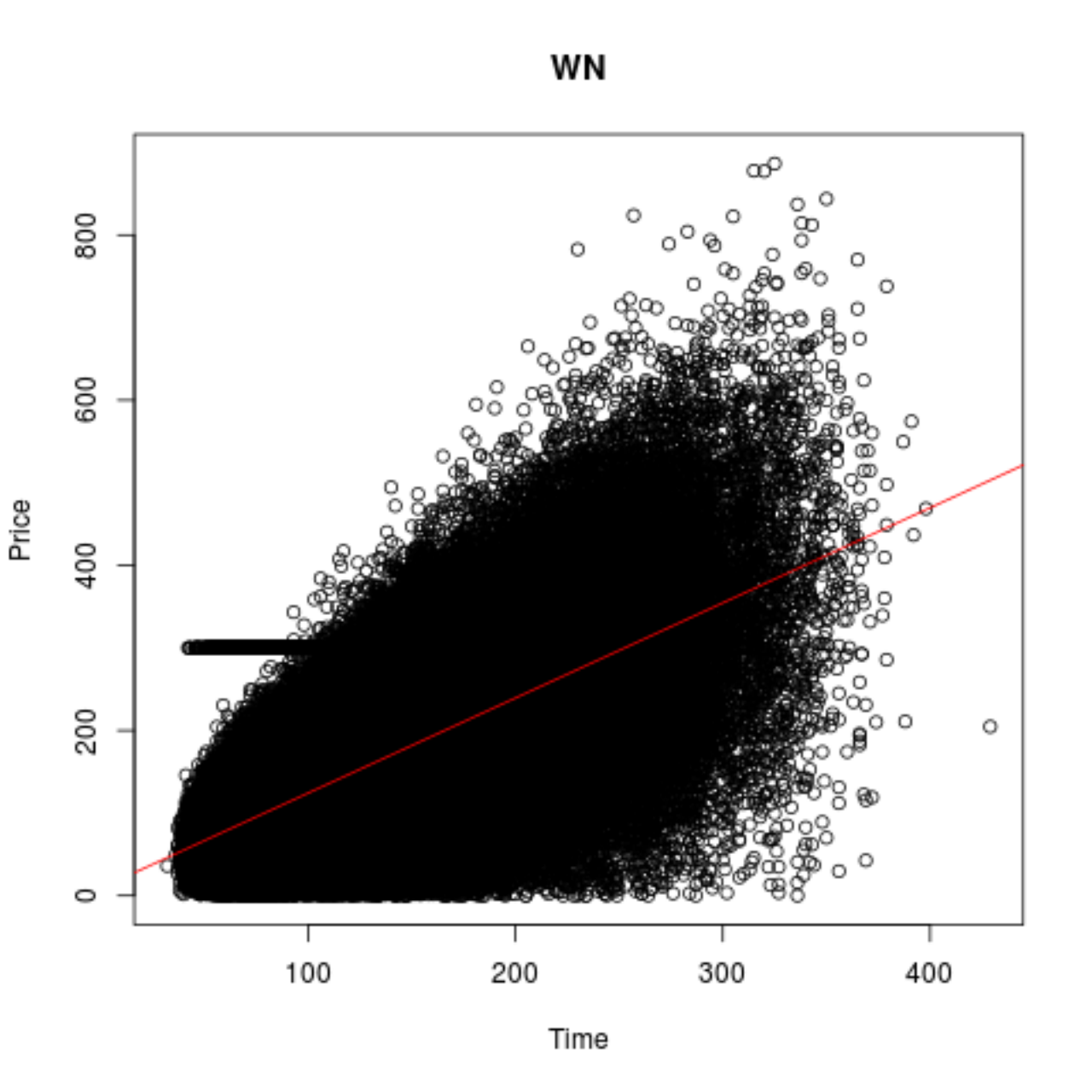

US
Price
Distance
0
500
1000
1500
2000
2500
3000


US
Price
0
500
1000
1500
2000
100
200
300
400
500
Time


VX
Price
Distance
0
500
1000
1500
2000
500
1000
1500
2000
2500


VX
Price
Time
0
500
1000
1500
2000
100
200
300
400
500


WN
Price
0
200
400
600
800
500
1000
1500
2000
2500
Distance


WN
Price
0
200
400
600
800
100
200
300
400
Time