# Automobile exploratory data analysis by Sanyam Chaudhary

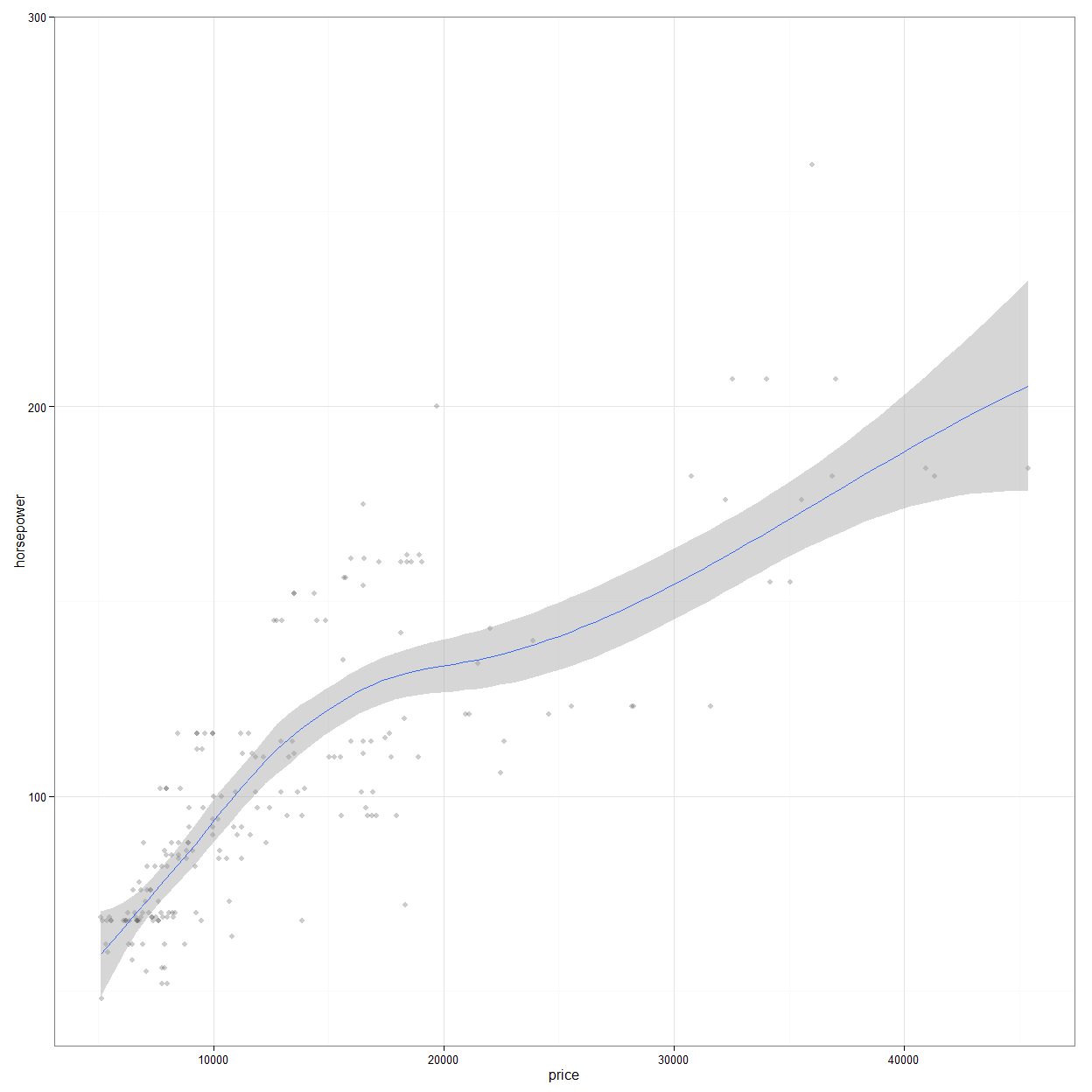
EDA for the given data set was done and below are the summary of findings. Remaining document outside of summary section is the entire report.

# Summary

Based on given data price of automobiles is dependent on Horsepower, CurbWeight, MPG , Aspiration ,Drive wheels and engine types. Plots of these are below. Even though the bodystyles also posed important but in the given data set the convertible, hardtop entires are small in frequency as seen in the final plot of Summary section ( count is 6, 8). Therefore we cannot conclusively say that convertible and hardtop contribute to price. Eventhough graph seems to suggest that.

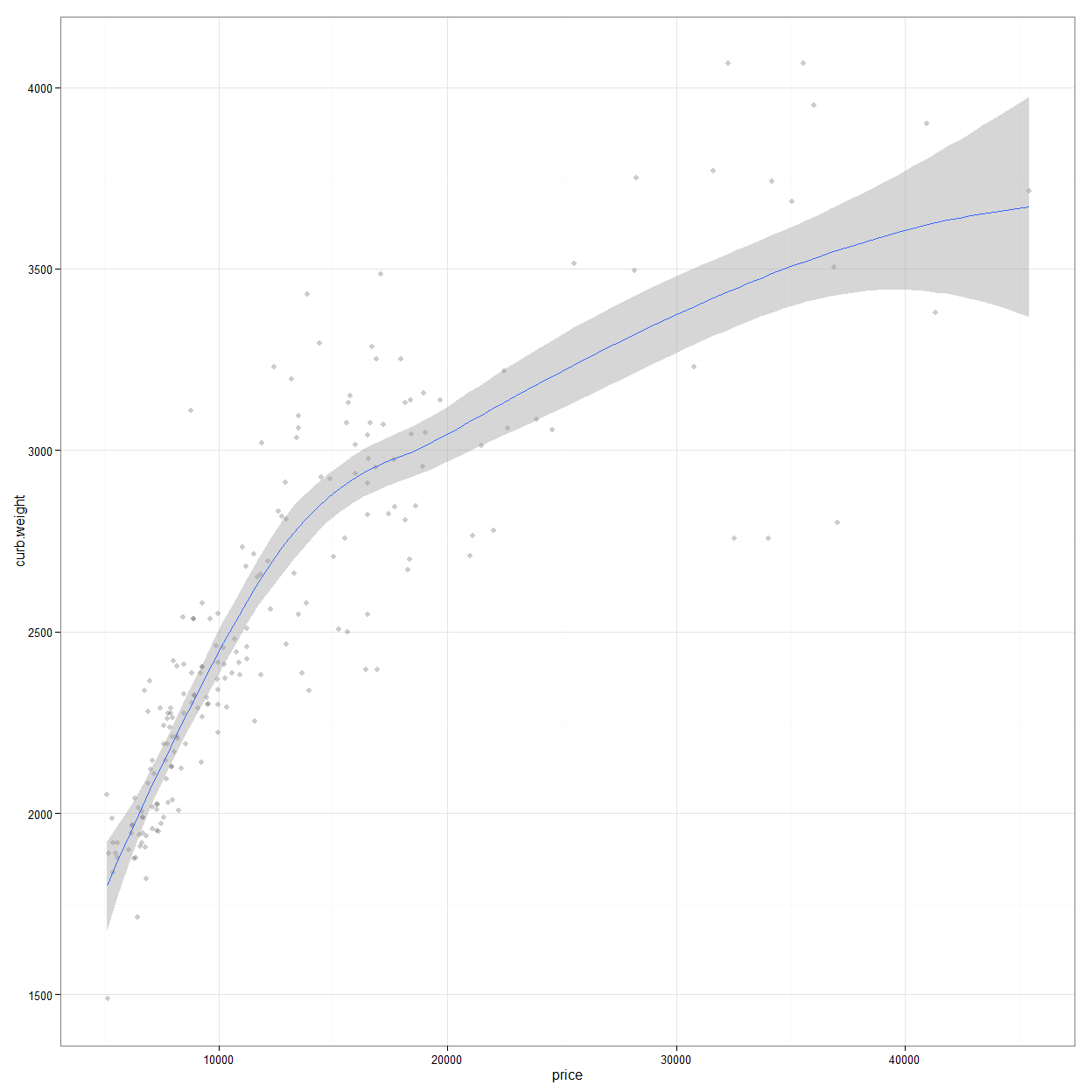
### Horsepower

As we can also see from correlation plot( below in the details section) that horsepower is +vly related to price



### Curb Weight

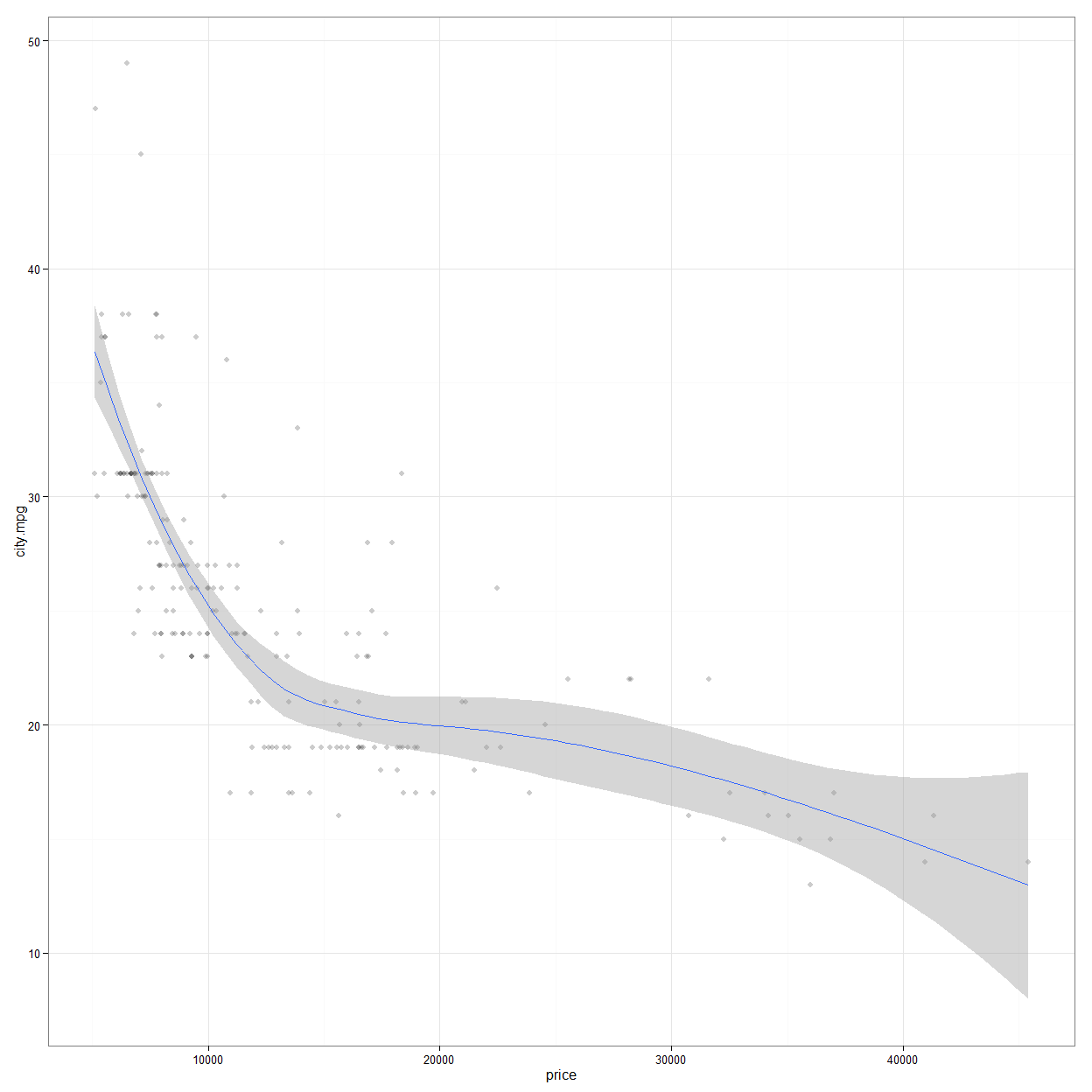
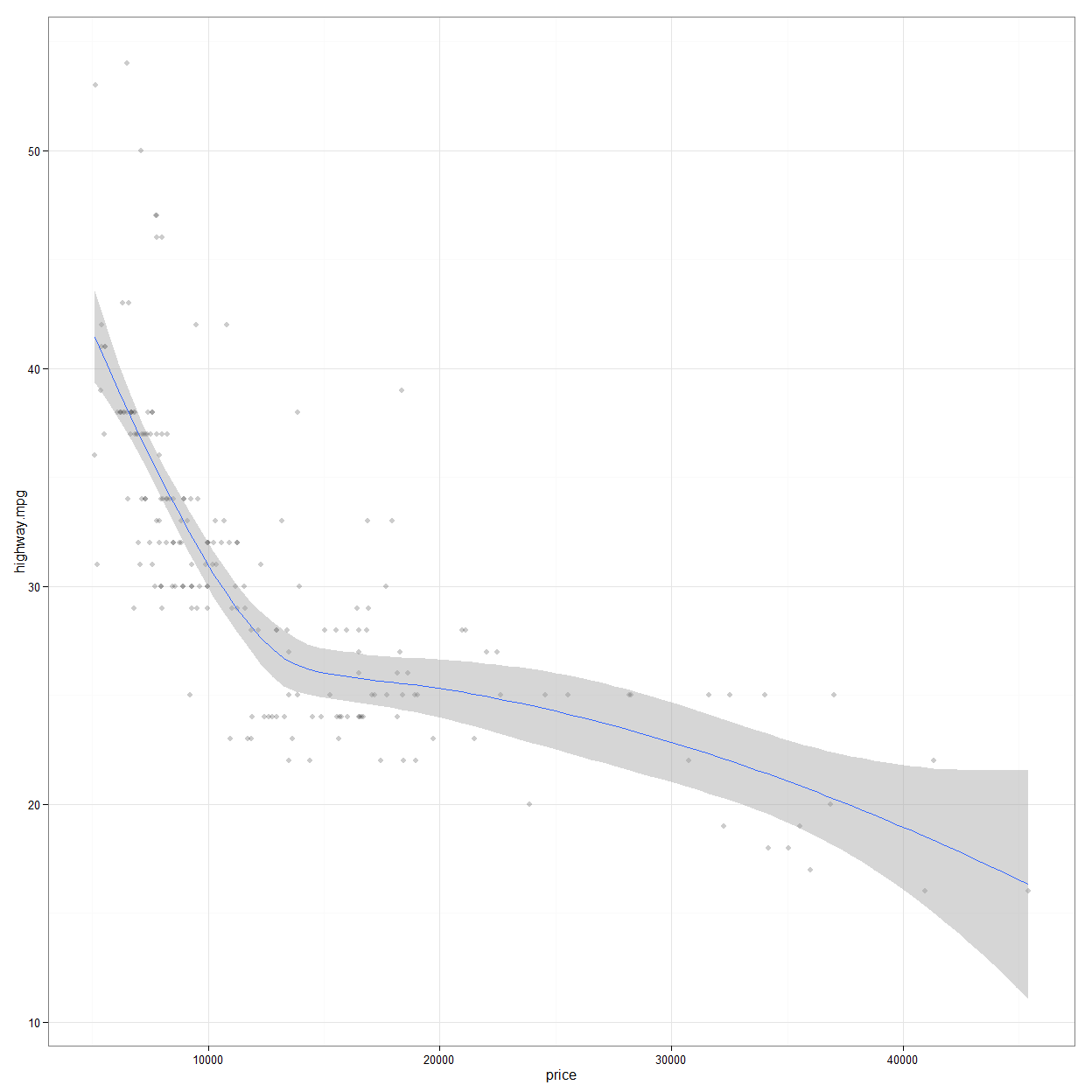
As we can also see from correlation plot below that curb weight is +vly related to price



Above plot suggests that costlier car gets have heavy weight. Intuitively this could be due to following ( just my opinions, not backed by data)

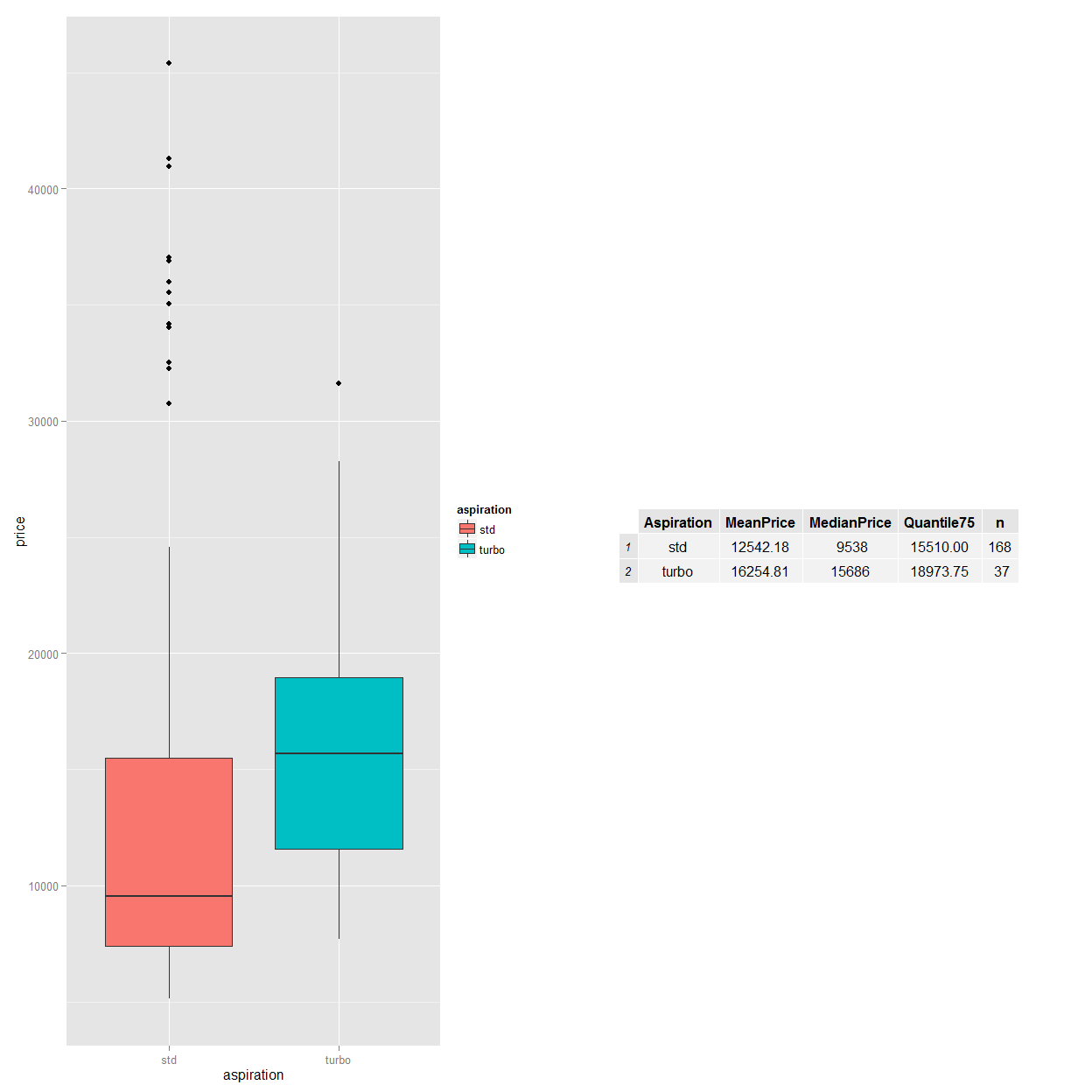
1. Fuel tank on costlier cars are big as we discovered from previous point that they require more fuels(not sure if this curb weight is with fuel weight in fuel tank or without it)
2. Costlier car runs faster hence the material used should make car stick to road at higher speed and different wind pressure.

### MPG

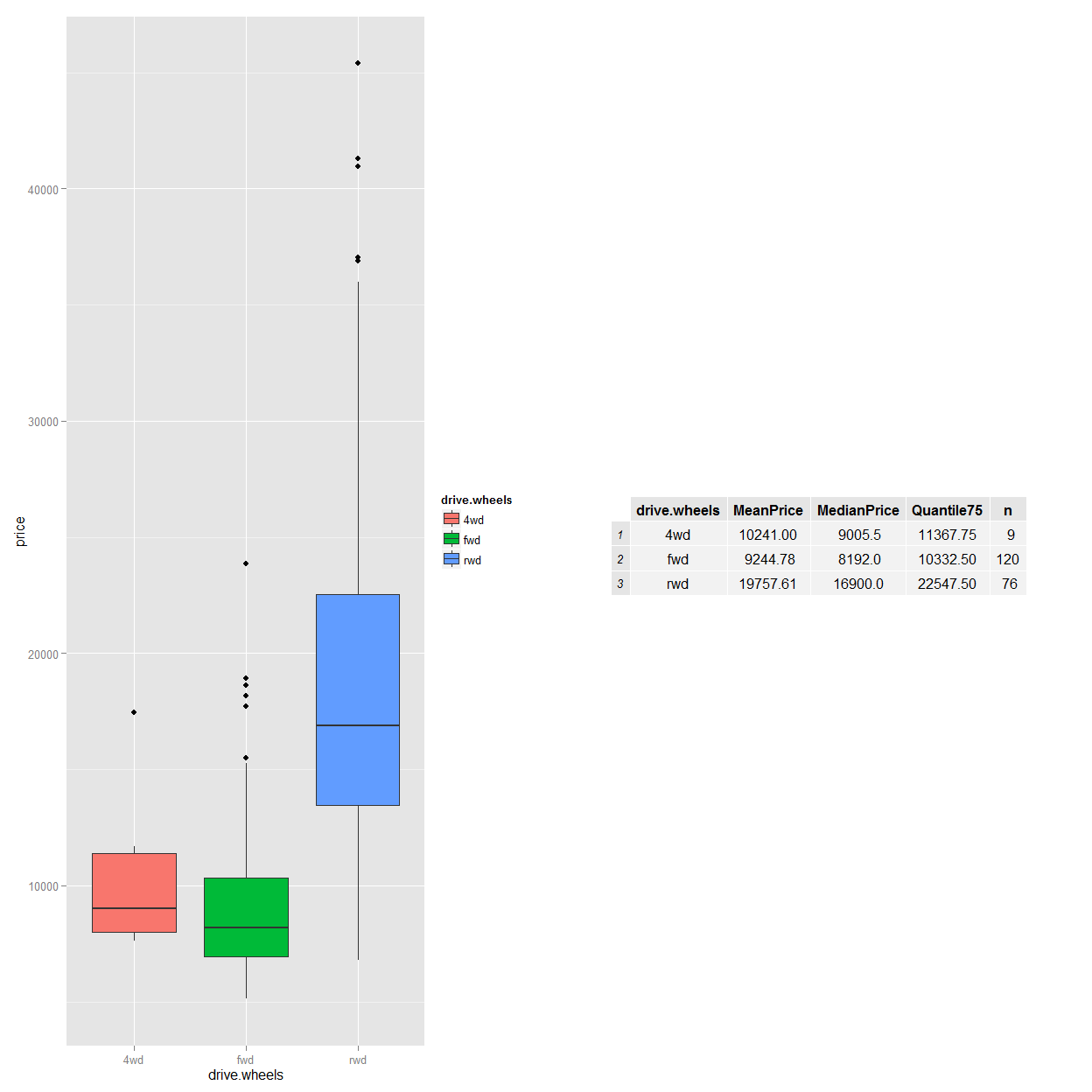
Based on the data, we can conclusively say that higher the price, lower is the milegae both for city and highway. Obviously the city and highway mpg are very highly correlated among themselves so we could have got away with plotting just one above.

### Aspiration vs price



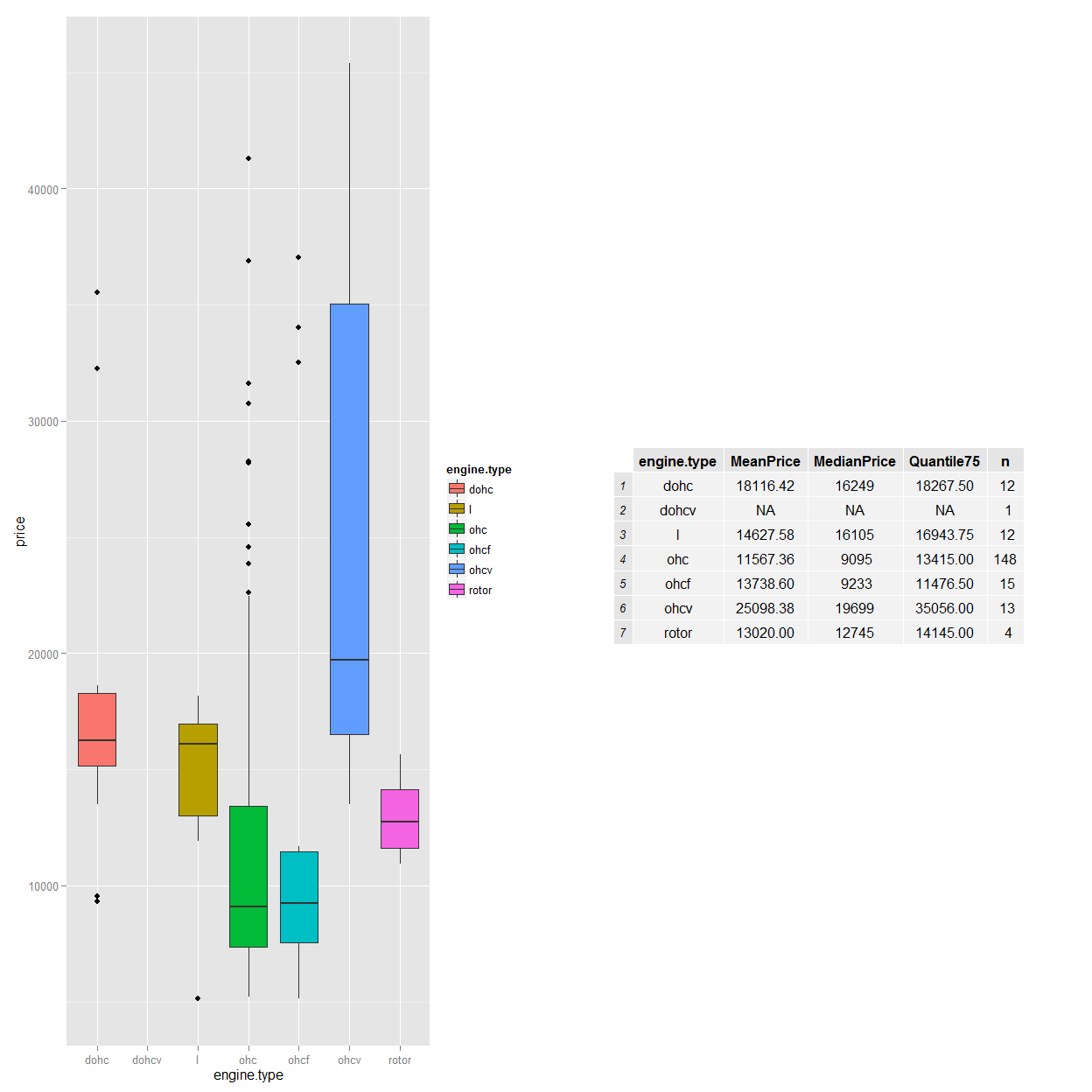
From the above we can see that turbo aspiration has higher price ( across Mean, Median and Quantile), hence it definitely contributes to price.

### drive.wheels vs price



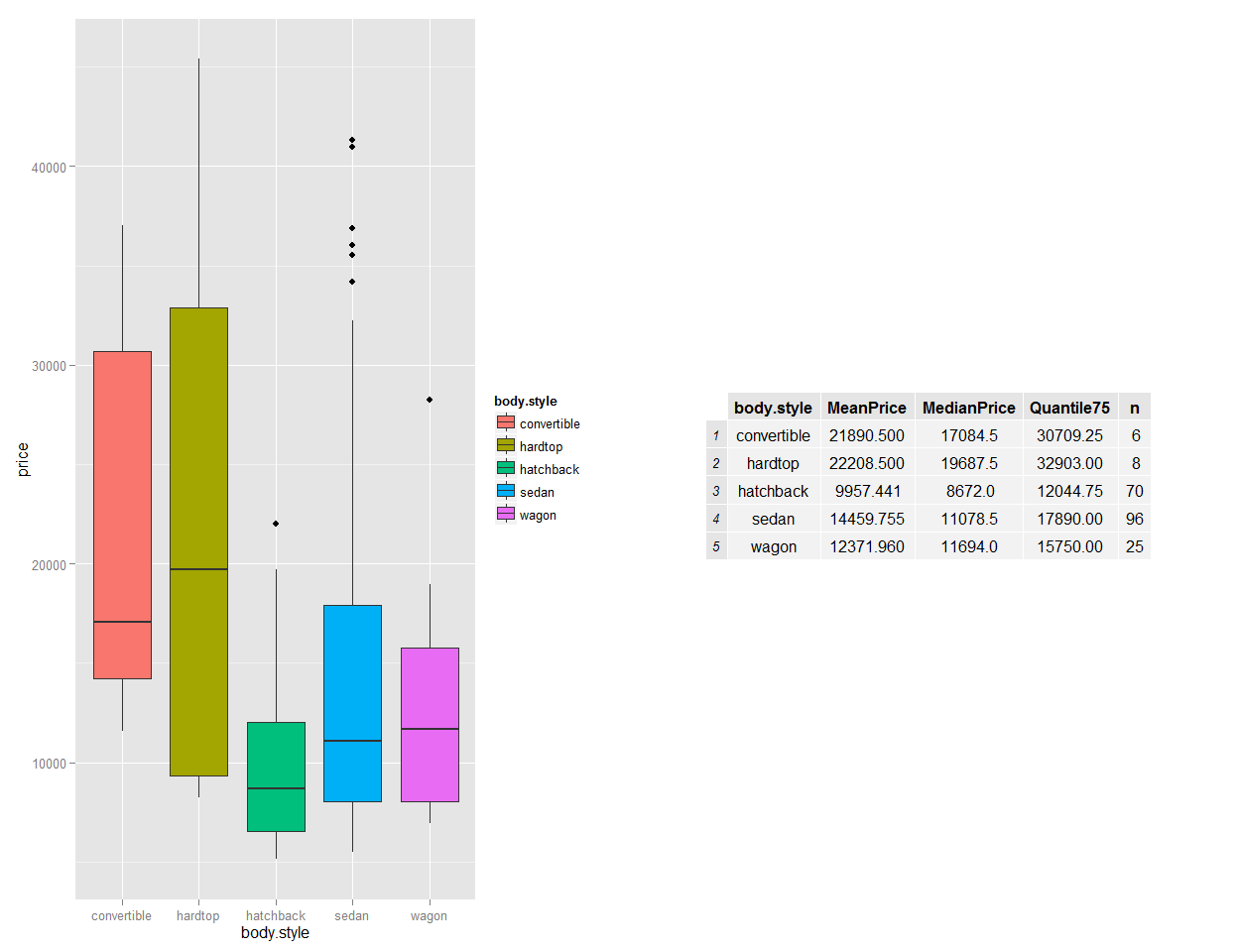
As we can see that rwd has high Mean, Median and Quantile price with comparable counts with fwd. We can conclusively say that rwd automobiles are costlier.

### Engine type vs price



ohcv engine type has third highest frequency in engine types among 7 engine types in data set and clearly higher than all of them.

### body.style vs price



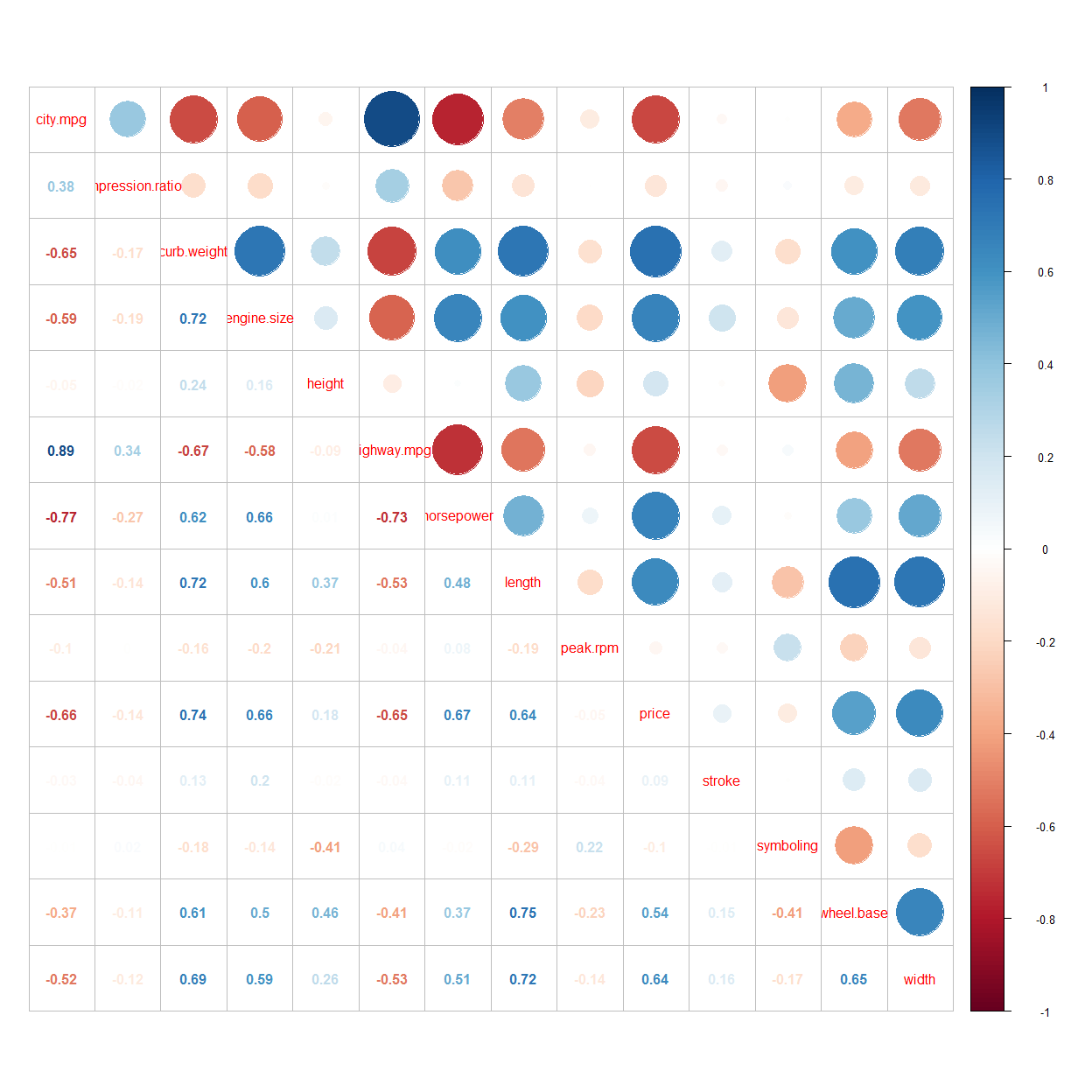
It appears both intuitively and from the box plot that convertibly and hardtop are very pricy however the number of entries in this group is significantly low than others. So from given data we are not going to pick body style as the determining factor.

### Detailed report

## [1] "symboling" "normalized.losses" "make"   
## [4] "fuel.type" "aspiration" "num.of.doors"   
## [7] "body.style" "drive.wheels" "engine.location"   
## [10] "wheel.base" "length" "width"   
## [13] "height" "curb.weight" "engine.type"   
## [16] "num.of.cylinders" "engine.size" "fuel.system"   
## [19] "bore" "stroke" "compression.ratio"  
## [22] "horsepower" "peak.rpm" "city.mpg"   
## [25] "highway.mpg" "price"

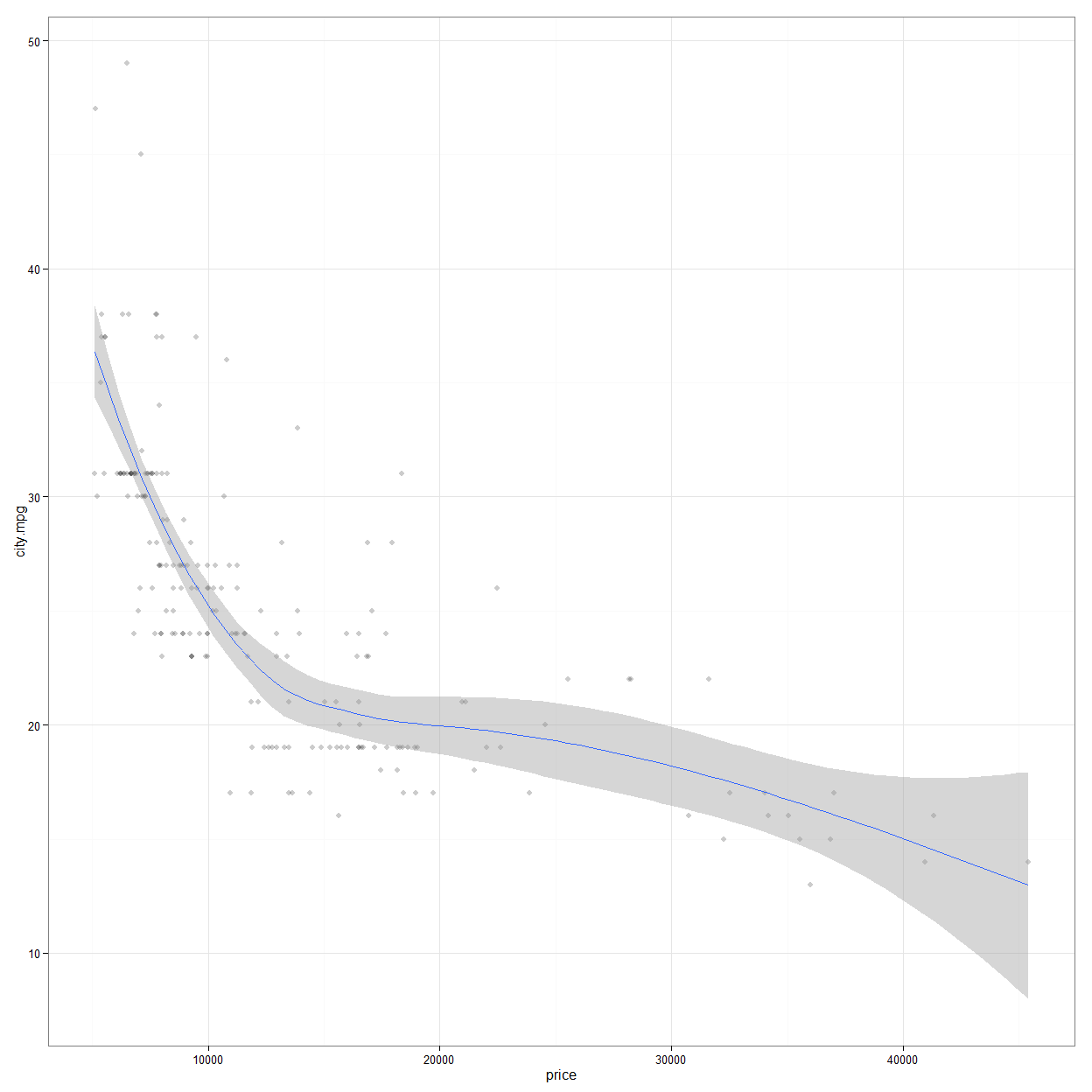
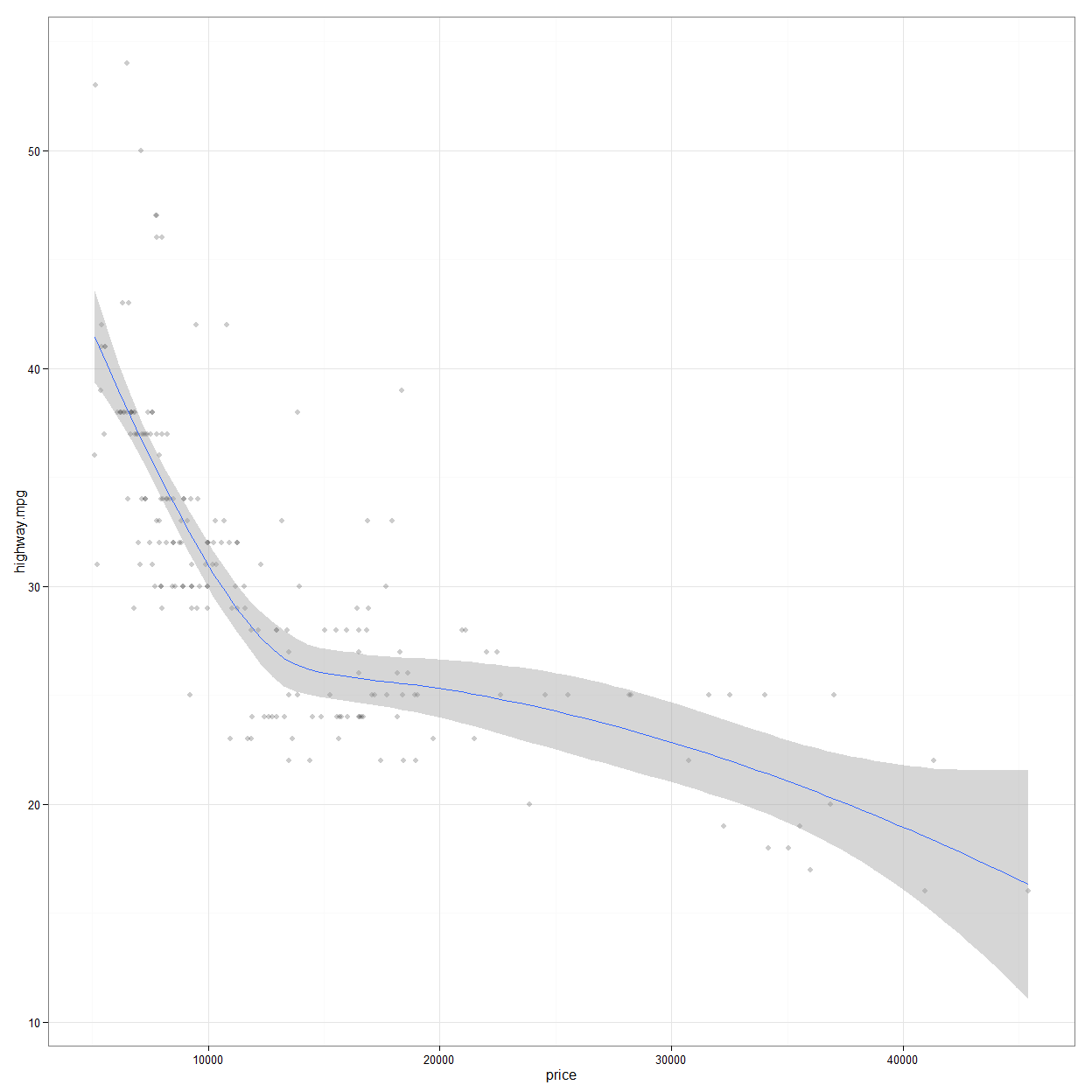
## symboling normalized.losses make fuel.type   
## Min. :-2.0000 ? : 41 toyota : 32 diesel: 20   
## 1st Qu.: 0.0000 161 : 11 nissan : 18 gas :185   
## Median : 1.0000 91 : 8 mazda : 17   
## Mean : 0.8341 150 : 7 honda : 13   
## 3rd Qu.: 2.0000 104 : 6 mitsubishi: 13   
## Max. : 3.0000 128 : 6 subaru : 12   
## (Other):126 (Other) :100   
## aspiration num.of.doors body.style drive.wheels engine.location  
## std :168 ? : 2 convertible: 6 4wd: 9 front:202   
## turbo: 37 four:114 hardtop : 8 fwd:120 rear : 3   
## two : 89 hatchback :70 rwd: 76   
## sedan :96   
## wagon :25   
##   
##   
## wheel.base length width height   
## Min. : 86.60 Min. :141.1 Min. :60.30 Min. :47.80   
## 1st Qu.: 94.50 1st Qu.:166.3 1st Qu.:64.10 1st Qu.:52.00   
## Median : 97.00 Median :173.2 Median :65.50 Median :54.10   
## Mean : 98.76 Mean :174.0 Mean :65.91 Mean :53.72   
## 3rd Qu.:102.40 3rd Qu.:183.1 3rd Qu.:66.90 3rd Qu.:55.50   
## Max. :120.90 Max. :208.1 Max. :72.30 Max. :59.80   
##   
## curb.weight engine.type num.of.cylinders engine.size fuel.system  
## Min. :1488 dohc : 12 eight : 5 Min. : 61.0 mpfi :94   
## 1st Qu.:2145 dohcv: 1 five : 11 1st Qu.: 97.0 2bbl :66   
## Median :2414 l : 12 four :159 Median :120.0 idi :20   
## Mean :2556 ohc :148 six : 24 Mean :126.9 1bbl :11   
## 3rd Qu.:2935 ohcf : 15 three : 1 3rd Qu.:141.0 spdi : 9   
## Max. :4066 ohcv : 13 twelve: 1 Max. :326.0 4bbl : 3   
## rotor: 4 two : 4 (Other): 2   
## bore stroke compression.ratio horsepower peak.rpm   
## 3.62 : 23 3.40 : 20 Min. : 7.00 68 : 19 5500 :37   
## 3.19 : 20 3.03 : 14 1st Qu.: 8.60 70 : 11 4800 :36   
## 3.15 : 15 3.15 : 14 Median : 9.00 69 : 10 5000 :27   
## 2.97 : 12 3.23 : 14 Mean :10.14 116 : 9 5200 :23   
## 3.03 : 12 3.39 : 13 3rd Qu.: 9.40 110 : 8 5400 :13   
## 3.46 : 9 2.64 : 11 Max. :23.00 95 : 7 6000 : 9   
## (Other):114 (Other):119 (Other):141 (Other):60   
## city.mpg highway.mpg price   
## Min. :13.00 Min. :16.00 ? : 4   
## 1st Qu.:19.00 1st Qu.:25.00 13499 : 2   
## Median :24.00 Median :30.00 16500 : 2   
## Mean :25.22 Mean :30.75 18150 : 2   
## 3rd Qu.:30.00 3rd Qu.:34.00 5572 : 2   
## Max. :49.00 Max. :54.00 6229 : 2   
## (Other):191

# Relationship of numerical variables with Price

I want to start of with correlation matrix to see whats the correlation among different variables and plot the ones which had any meaningful relationships. I am going to use corrplot package for this. Below are the numeric columns plotted , chose the method as "kendall" based on class discussion that its more robust. 

As we can see from the plot above that mpg has -ve relationship with the price of the vehicle. Lets take a look at these plots in more details.

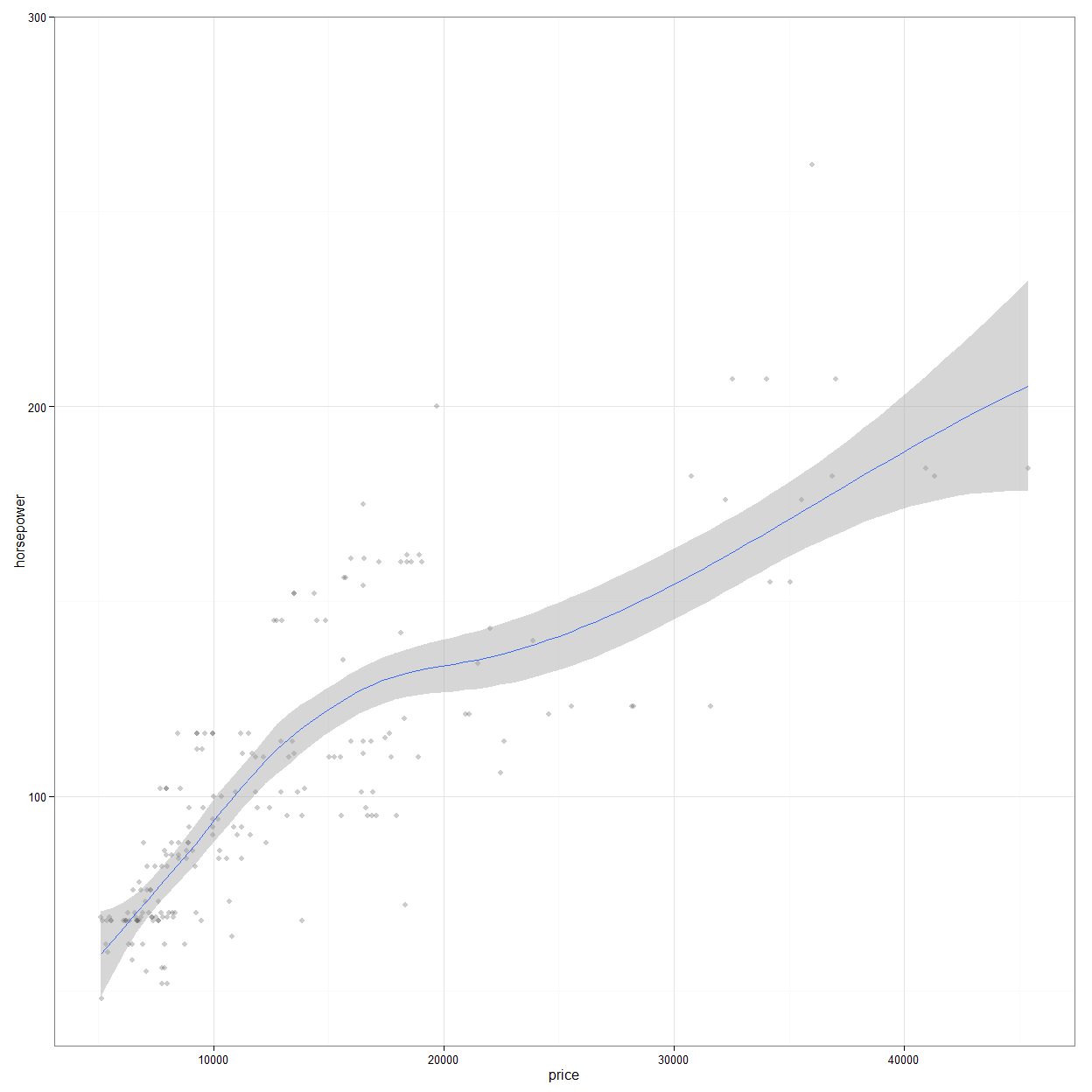
## Price vs MPG

Based on the data, we can conclusively say that higher the price, lower is the milegae both for city and highway. Obviously the city and highway mpg are very highly correlated among themselves so we could have got away with plotting just one above.

## Horsepower vs Price

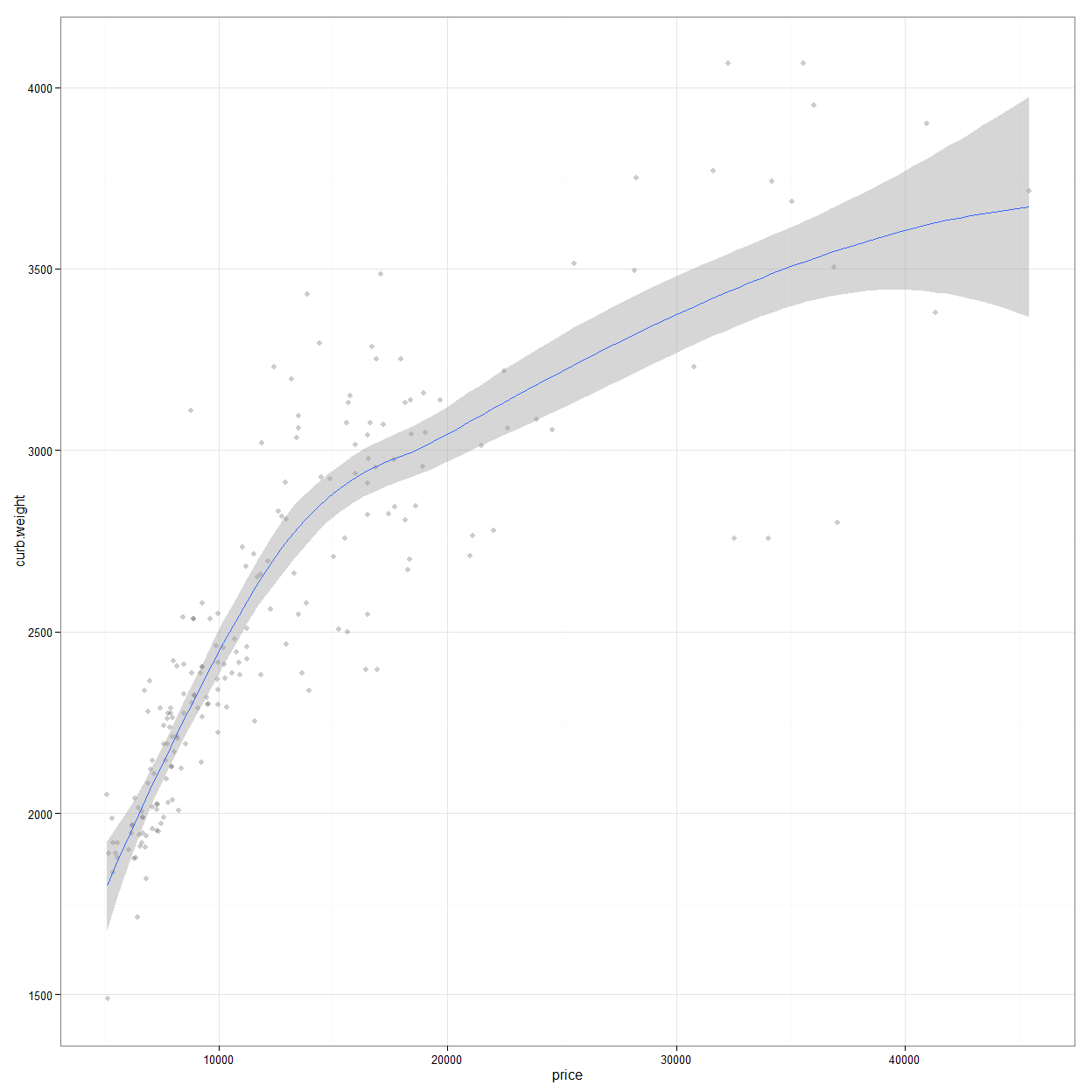
As we can also see from correlation plot that horsepower is +vly related to price



We can conclusively say that horsepower is +vely related to price and is a factor in determining the price

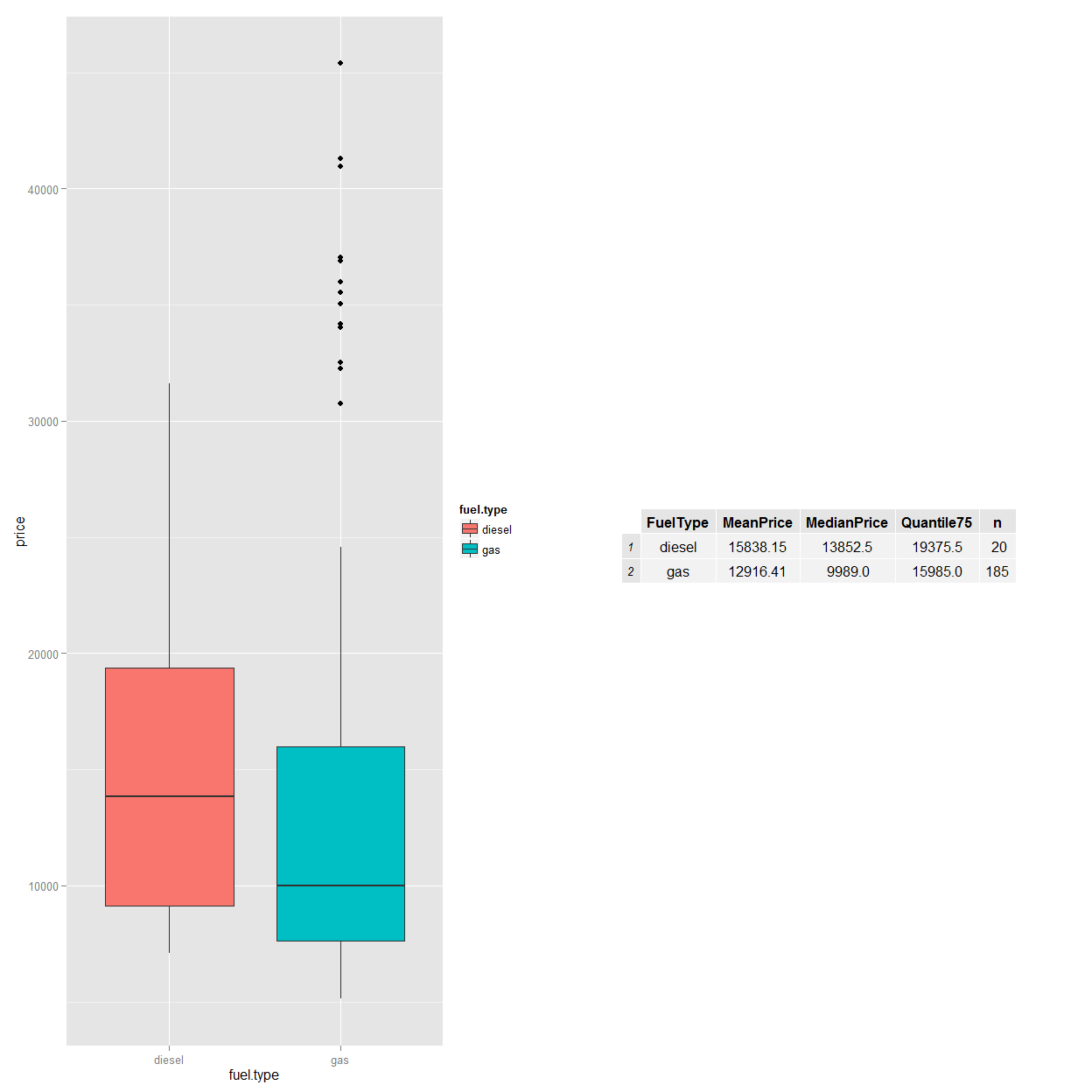
## Curb weight vs Price

As we can also see from correlation plot that curb weight is +vly related to price

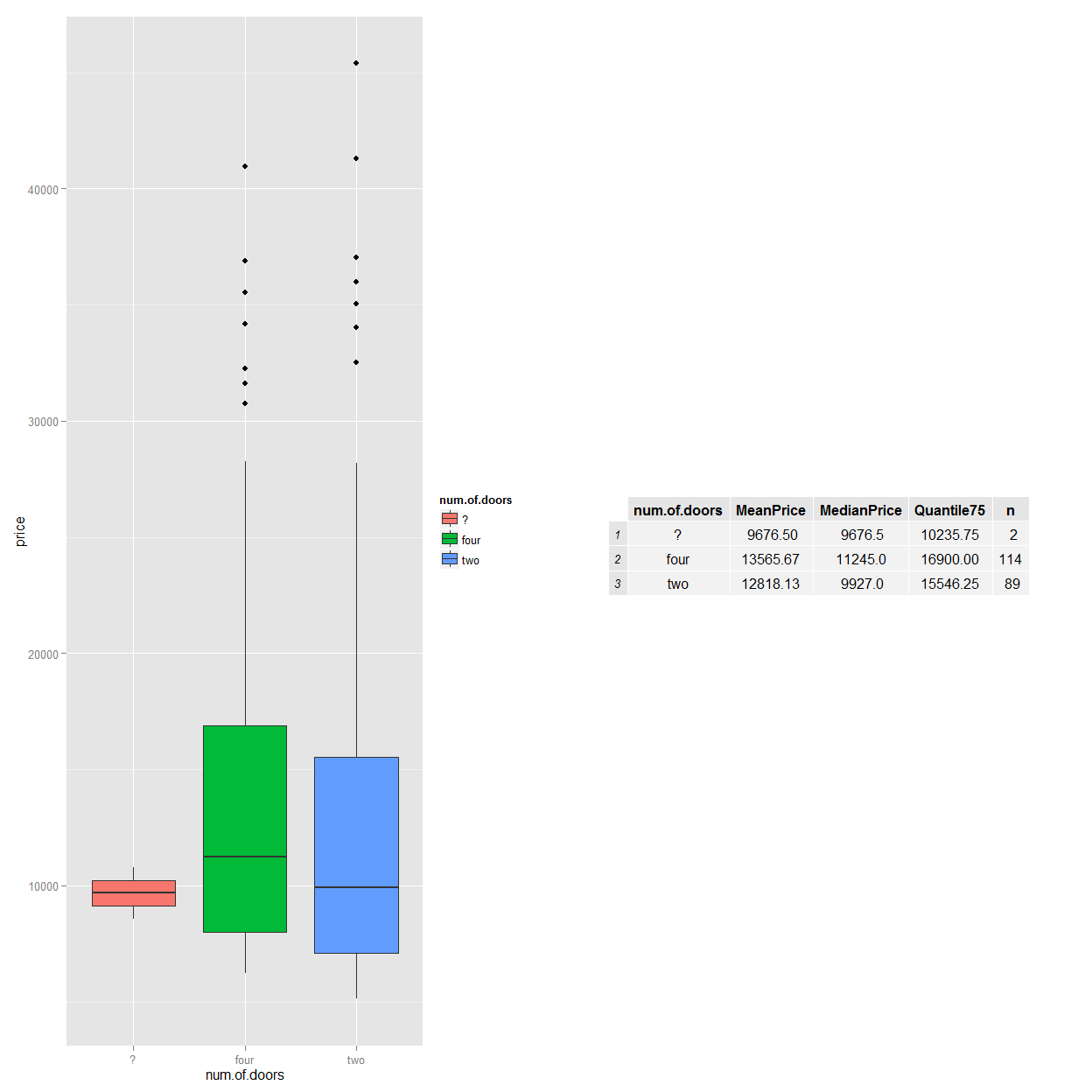


# Relationship of non numerical variables with Price

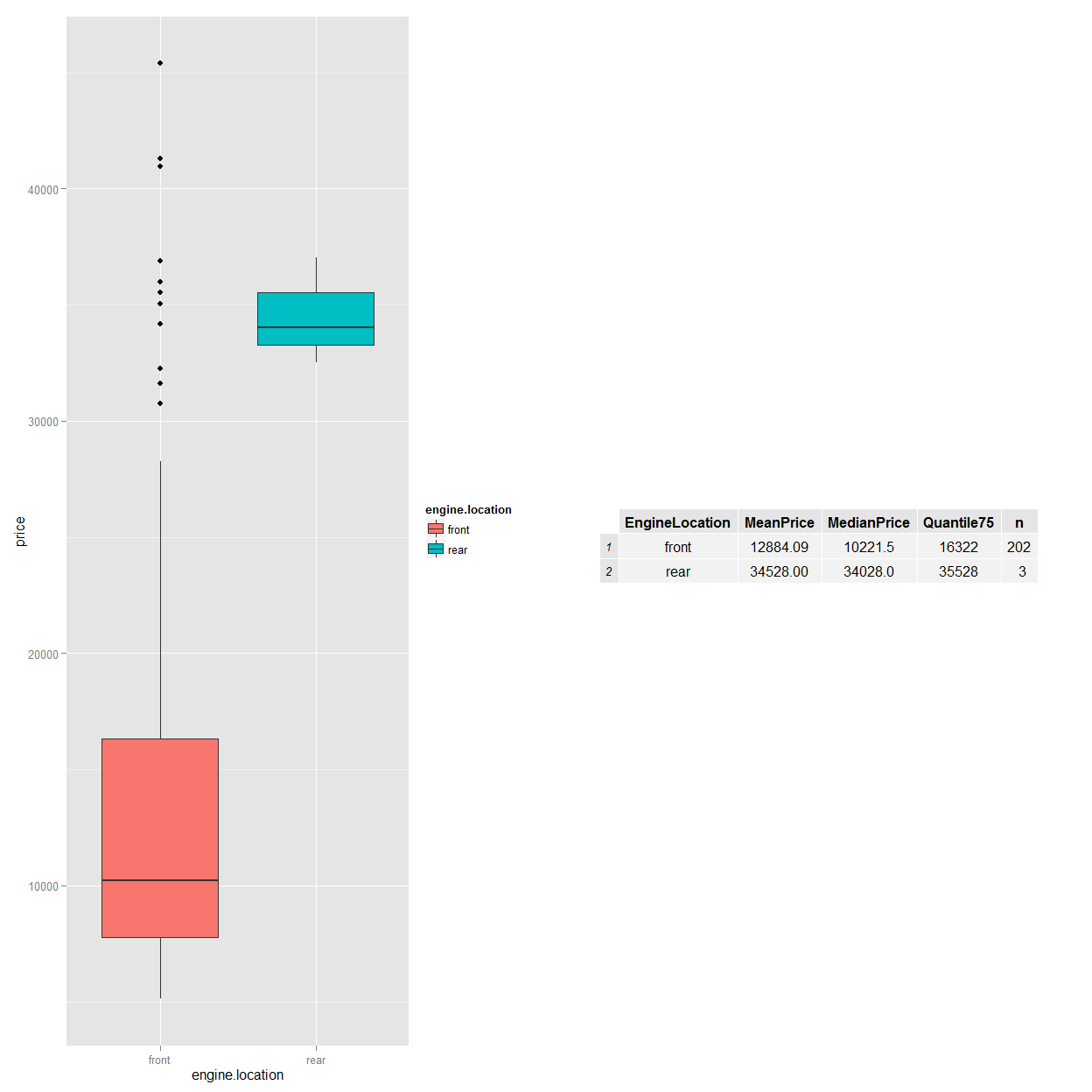
## Fuel type vs price

 Nothing conclusive could be said about Price and fuel type from this data.

### Door vs price



## Engine location vs price



## Fuel System vs price

