**Statistical Data Mining I**

**Homework 1**

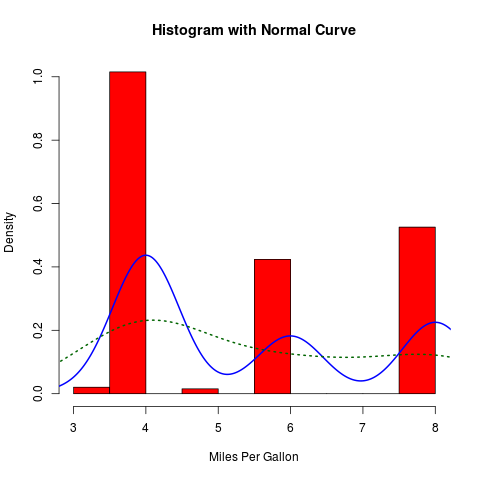
1) **Tasks**:

* Perform exploratory data analysis on data
* Pre-process the data
* Submit the cleaned dataset in .Rda format

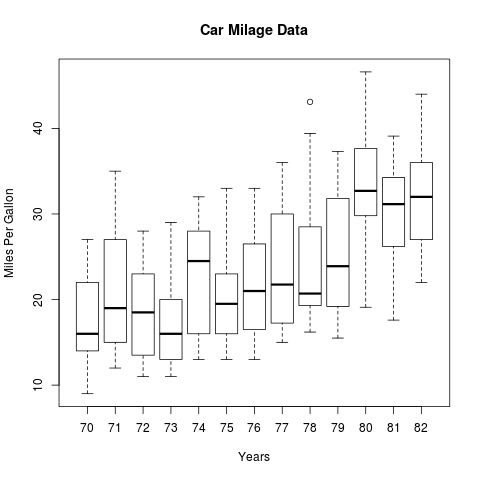
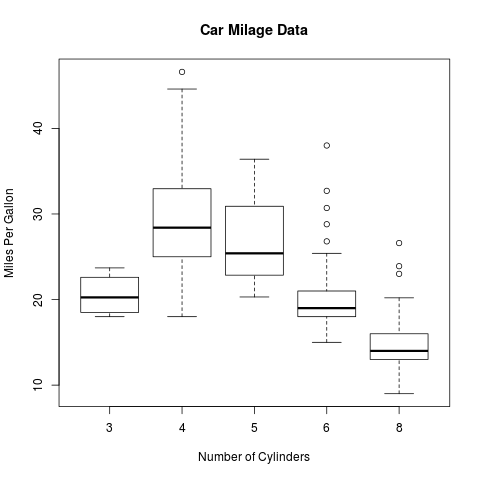
Dataset to be used: Auto from ISLR package

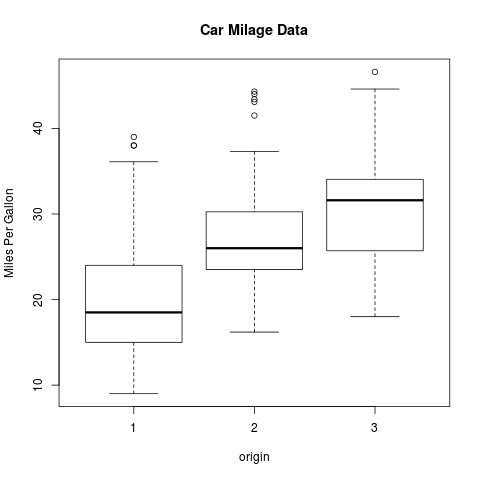
**Answers**:

* The Auto dataset of ISLR package has in total 9 variables out of which the ‘**names**’ variable would have the zero significance since it is just a kind of tagging attribute. So, I have removed the ‘names’ column in order to build the fitting predictive model for mpg.
* **Histograms**: The histogram is distributed normally.



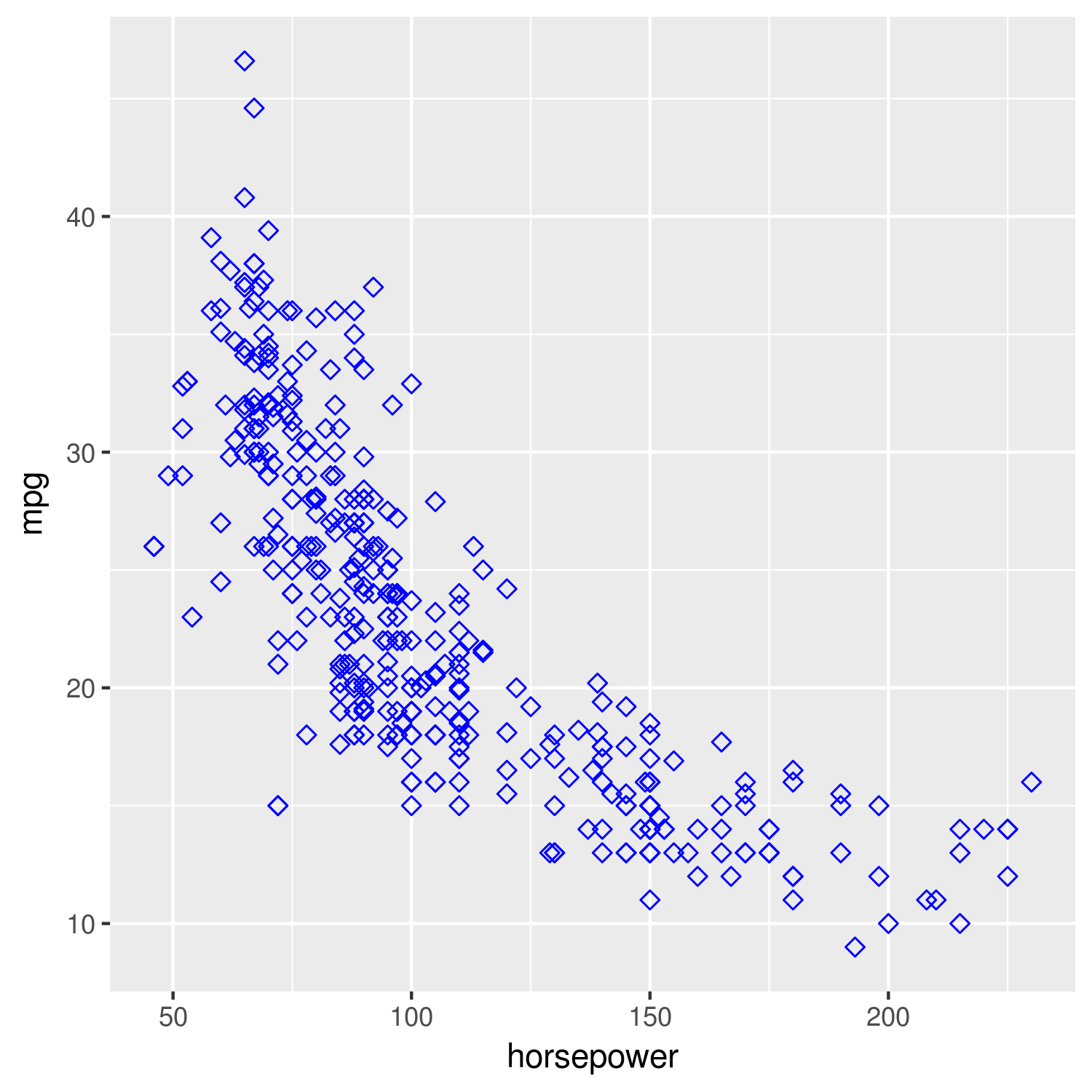
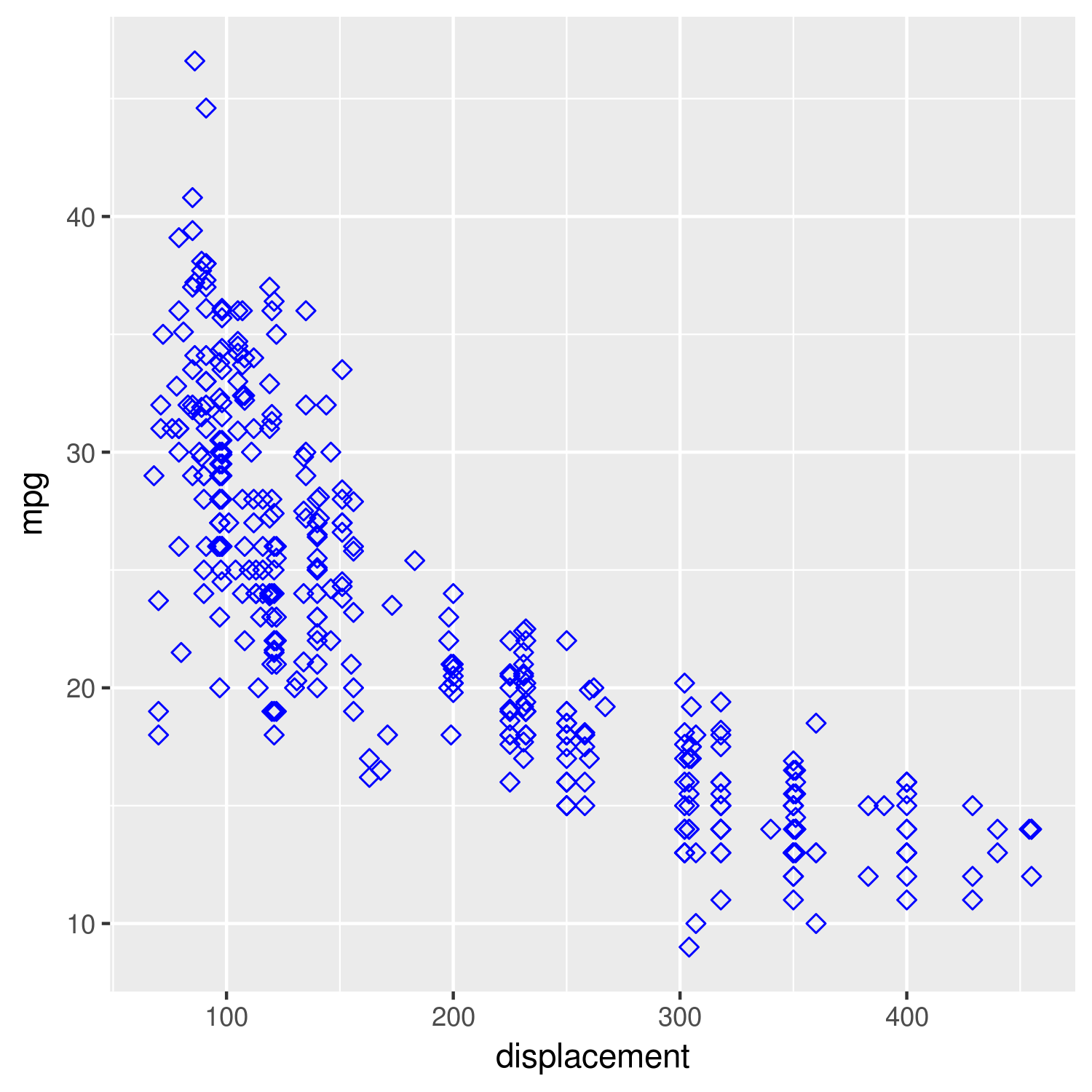
* The predictors Year, Origin and Cylinders are discrete variables so I have used Box-whiskers plot for analyzing them with respect to MPG (Miles per gallon). Below are the findings:

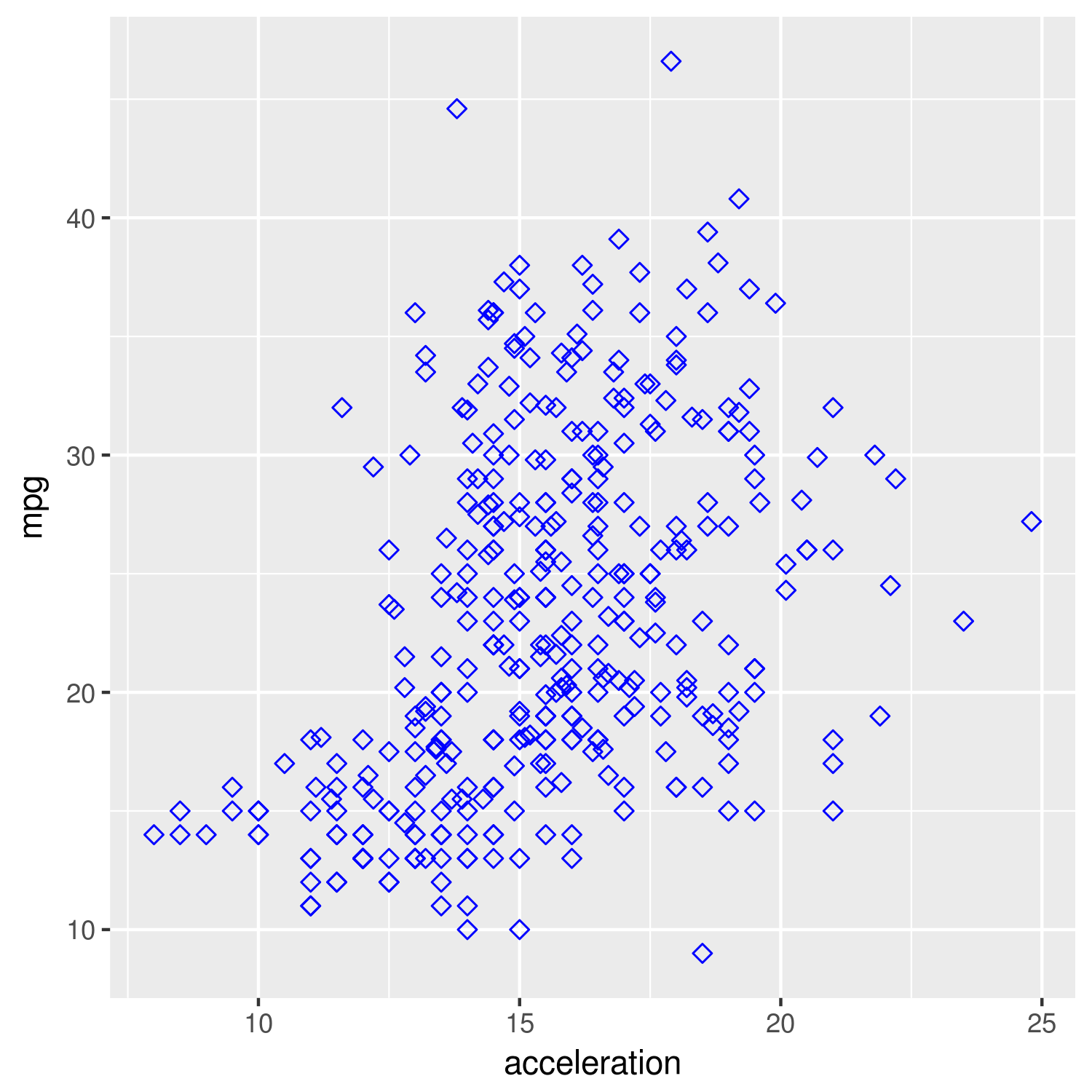


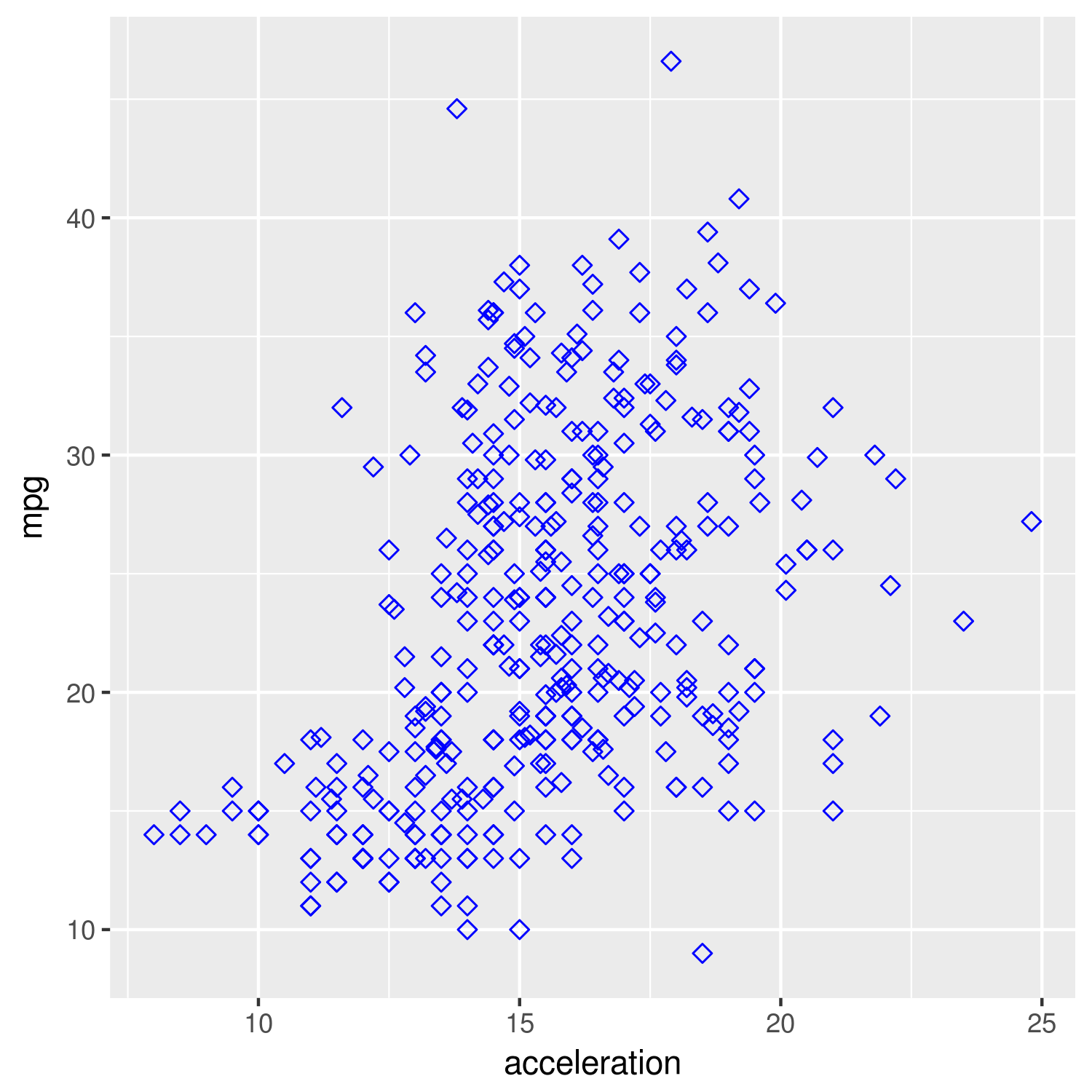


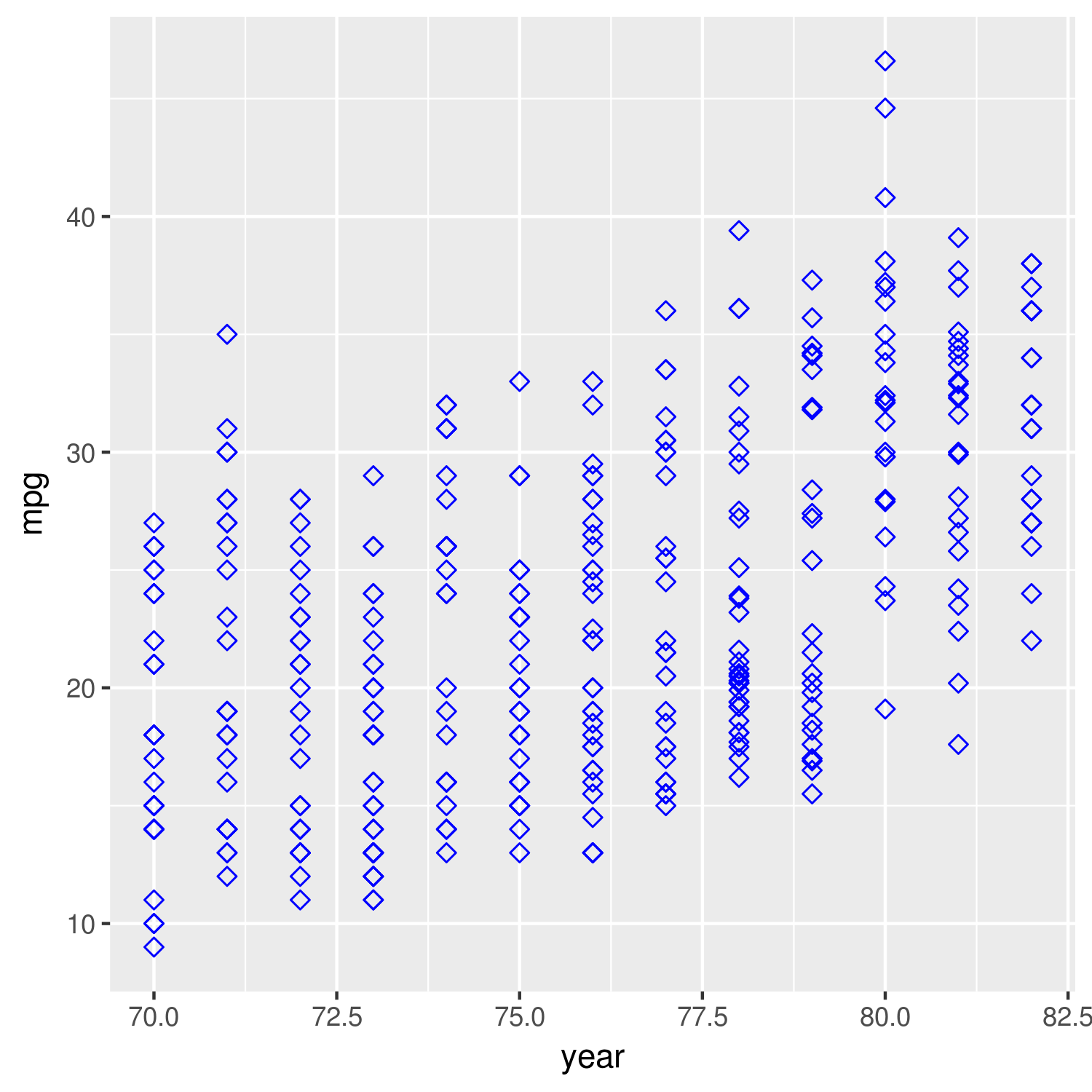
* As evident from the above plots there are some outliers which should be treated for cleaning the data. After removing some of the outliers and cleaning the data I have used scatter plots for plotting the predictors against MPG and have found some useful analysis.

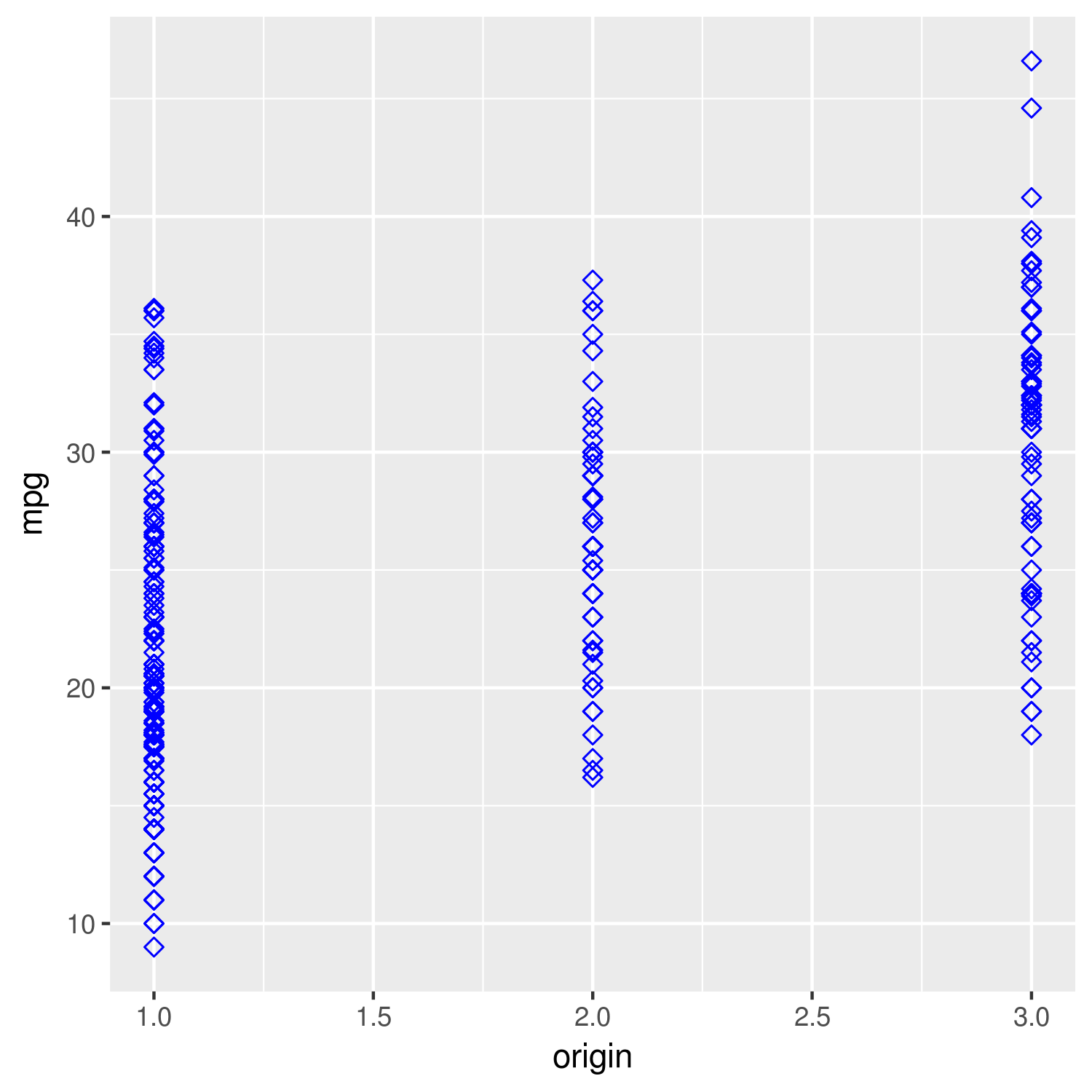
Also the predictors Cylinders, acceleration, weight, Displacement, Horsepower are the continuous entities so we are using scatter plot for analyzing them.

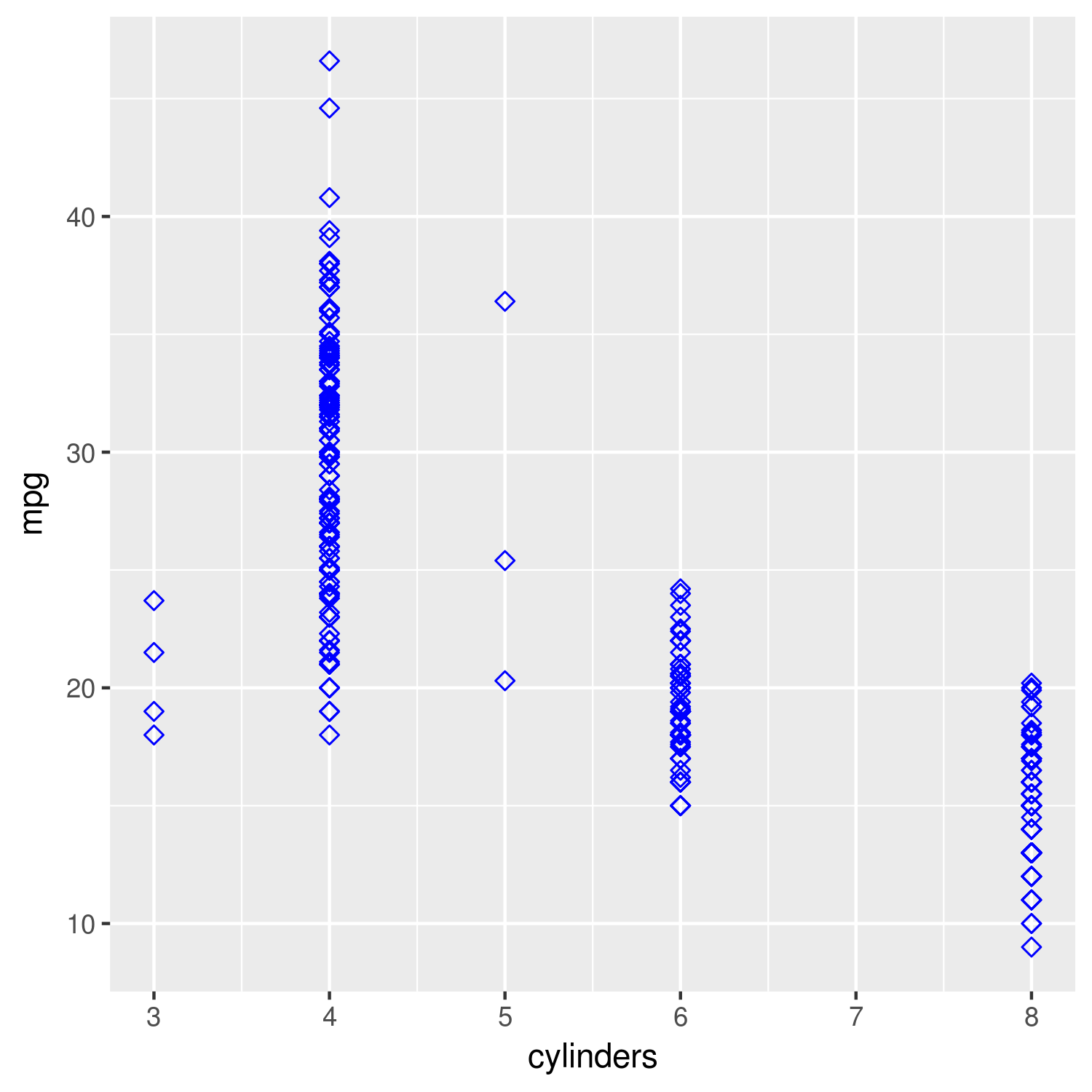












* + I’ve used histograms with density function, box plots and bivariate scatter plots to investigate the data. From histograms it’s evident that the data is distributed normally in most of the cases.
  + I’ve used box plots for the discrete predictors – Cylinders, Year and Origin against ‘mpg’ as the response variable. There are some outliers in each case which has been removed from the dataset.
  + After removing the outliers the plots looks well arranged.
  + I’ve used Bivariate Scatter plots for continuous predictors like Weight, Horsepower, Displacement etc. and can see the relationship of each of them with response variable ‘mpg’. After removing outliers the scatter plots are also looking well bind.
  + From the scatter plots we can see that MPG is strongly and inversely related with the Weight, Displacement and Horsepower.
  + From the scatter plot for Acceleration vs Mpg, the data is looking very scattered. Thus, the acceleration field might not be assisting while building the predictive model.
  + On the other hand, Year, Weight, Displacement, Horsepower have a strong relationship with MPG.
* The cleaned data is saved as a cleanData.Rda file.