Assignment9

John Bute

2024-10-02

Assignment 9

Q1: Which of the predictors are quantitative, and which are qualitative? Quantitative predictors: mpg, cylinders, displacement, horsepower, weight, acceleration, year Qualitative predictors: origin, name

Q2: find the range of each quantiative predictor (we also need to delete missing values)

```
data <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/Auto.csv")
data <- na.omit(data)
quantitative_columns <- c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration")
ranges <- sapply(data[quantitative_columns], function(x){
  limits <- range(x)
  names(limits) <- c("Lower Limit", "Higher Limit")
  return(limits)
})
ranges
##               mpg cylinders displacement horsepower weight acceleration
## Lower Limit   9.0         3           68         46   1613          8.0
## Higher Limit 46.6         8          455        230   5140         24.8
```

Q3: Find the standard deviation and mean of each quantitative predictor

```
f <- function(x) {
  c(mean = mean(x), sd = sd(x))
}
mean_deviation <- sapply(data[quantitative_columns], f)
mean_deviation
##            mpg cylinders displacement horsepower    weight acceleration
## mean 23.445918  5.471939      194.412  104.46939 2977.5842    15.541327
## sd    7.805007  1.705783      104.644   38.49116  849.4026     2.758864
```

Q4: now we remove observations 10-85, and recalculate the range, mean, and standard deviation

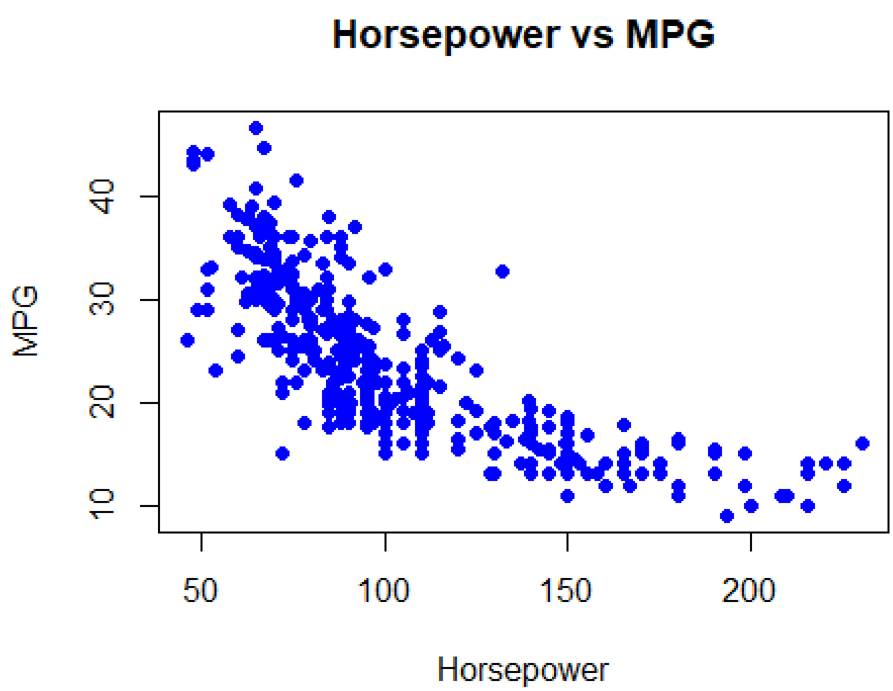
```
new_function <- function(x) {
  mean_value <- mean(x, na.rm = TRUE)
  sd_value <- sd(x, na.rm = TRUE)
  limits <- range(x, na.rm = TRUE)
  names(limits) <- c("Lower Limit", "Higher Limit")
  return(c(mean = mean_value, sd = sd_value, limits))
}
subset_data <- data[-c(10:85), ]
quantitative_columns <- c("mpg", "cylinders", "displacement", "horsepower",
                          "weight", "acceleration", "year")
mean_deviation <- sapply(subset_data[quantitative_columns], new_function)
mean_deviation
##                    mpg cylinders displacement horsepower    weight acceleration
## mean         24.404430  5.373418    187.24051  100.72152 2935.9715    15.726899
## sd            7.867283  1.654179     99.67837   35.70885  811.3002     2.693721
## Lower Limit  11.000000  3.000000     68.00000   46.00000 1649.0000     8.500000
## Higher Limit 46.600000  8.000000    455.00000  230.00000 4997.0000    24.800000
```

```
##                   year
## mean          77.145570
## sd             3.106217
## Lower Limit  70.000000
## Higher Limit 82.000000
```

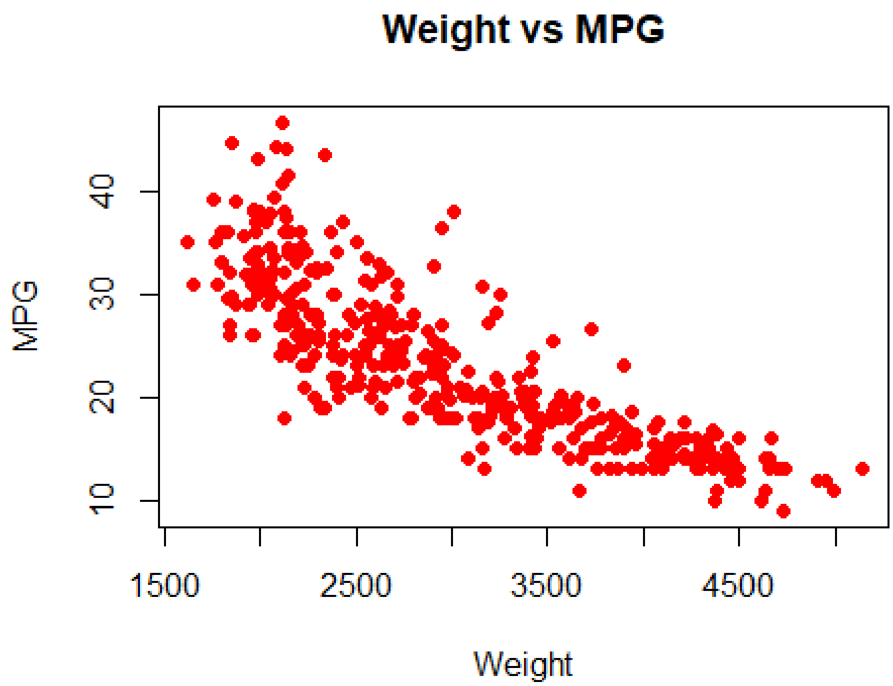
Removing 76 of the dataset's observations does not change the nature of the Auto dataset.

Q5: Using the full dataset, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings

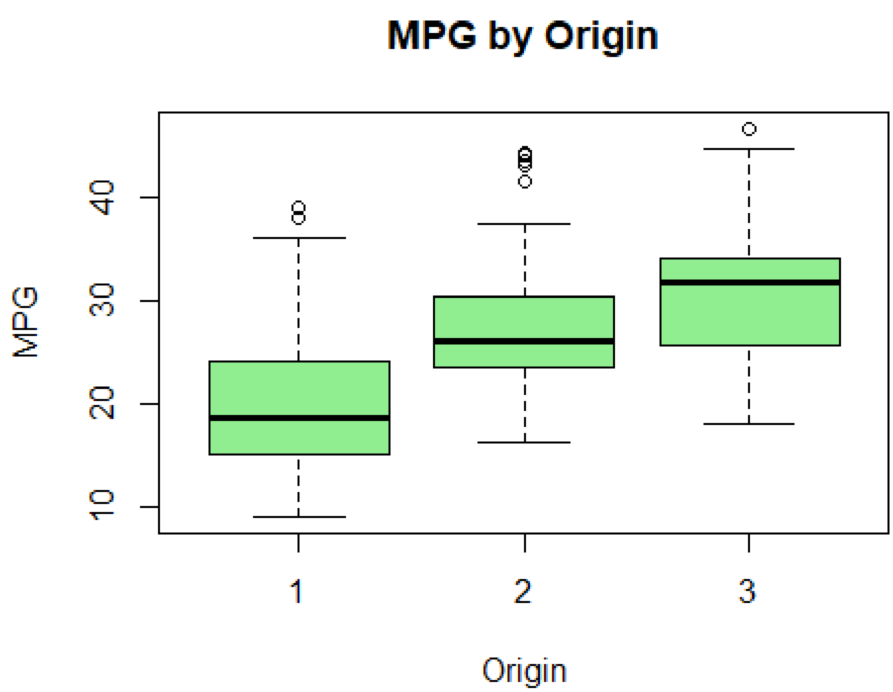
```r
plot(data$horsepower, data$mpg,
     main = "Horsepower vs MPG",
     xlab = "Horsepower", ylab = "MPG",
     pch = 19, col = "blue")
```



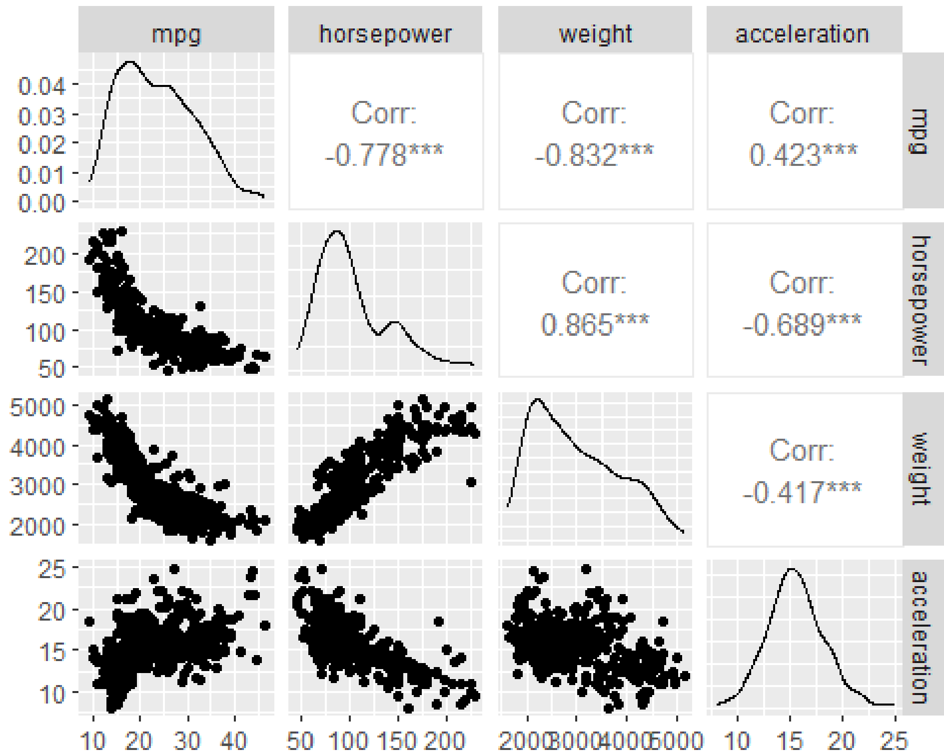

```r
plot(data$weight, data$mpg,
     main = "Weight vs MPG",
     xlab = "Weight", ylab = "MPG",
     pch = 19, col = "red")
```



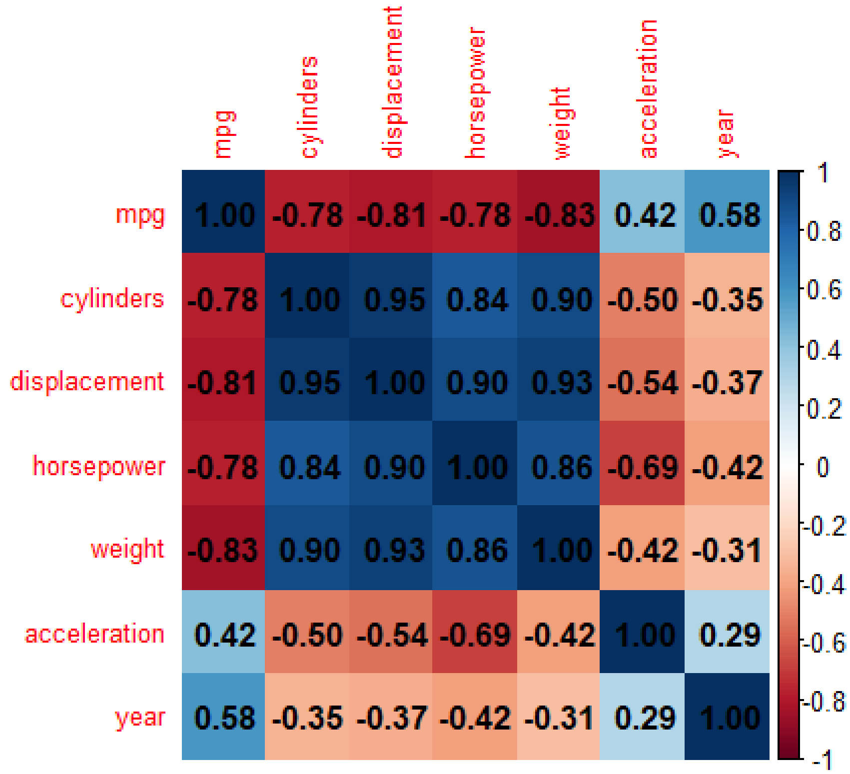

```r
boxplot(data$mpg ~ data$origin,
        main = "MPG by Origin",
        xlab = "Origin", ylab = "MPG",
        col = "lightgreen")
```



```r
install.packages("GGally")
## package 'GGally' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\johnb\AppData\Local\Temp\RtmpkBFPxt\downloaded_packages
library(GGally)
ggpairs(data[, c("mpg", "horsepower", "weight", "acceleration")])
```




```r
install.packages("corrplot")
## package 'corrplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\johnb\AppData\Local\Temp\RtmpkBFPxt\downloaded_packages
library(corrplot)
cor_matrix <- cor(data[, c("mpg", "cylinders", "displacement", "horsepower",
                           "weight", "acceleration", "year")], use = "complete.obs")
corrplot(cor_matrix, method = "color", tl.cex = 0.8, addCoef.col = "black")
```



Horsepower vs MPG:

Plot: A scatterplot showing how horsepower relates to mpg. Analysis: There is a clear negative correlation between horsepower and mpg. Vehicles with higher horsepower tend to have lower fuel efficiency, which makes sense since more powerful engines typically consume more fuel.

Weight vs MPG:

Plot: A scatterplot showing how weight relates to mpg. Analysis: There is a negative relationship between vehicle weight and fuel efficiency (mpg). Heavier cars generally have lower fuel efficiency, which makes sense because more weight means more energy is required to move the vehicle.

MPG by Origin:

Plot: A boxplot of mpg by origin (the region where the car was manufactured). Analysis: There is a noticeable difference in fuel efficiency between cars from different regions. Cars from origin 3 tend to have generally better mpg.

Pairwise Relationships:

Plot: A pairwise plot matrix (ggpairs) of several important variables like mpg, horsepower, weight, and acceleration. Analysis: This visualization allows us to see the pairwise relationships between mpg, horsepower, weight, and acceleration. For instance, it reinforces the negative correlation between mpg vs horsepower and weight. Furthermore, horsepower and weight are positively correlated.

Correlation Matrix:

Plot: A correlation matrix to summarize the relationships between the quantitative predictors. Analyis: The correlation matrix shows that mpg is negatively correlated with horsepower, weight, and cylinders, while it is positively correlated with year (newer cars tend to have better fuel efficiency).

Q6. Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Horsepower: There is a strong negative relationship between horsepower and mpg. As horsepower increases, mpg tends to decrease. This suggests that horsepower could be a strong predictor of fuel efficiency. Weight: Similarly, weight has a negative relationship with mpg. Heavier cars consume more fuel, making this another important predictor. Cylinders: The number of cylinders in an engine is likely to influence fuel consumption, as engines with more cylinders tend to be larger and less fuel-efficient. Year: There is a positive correlation between year and mpg, indicating that newer cars tend to be more fuel-efficient. Possibly, over time, technological advancements allowed us to better manage fuel. Acceleration: slightly positively correlated with mpg, as cars that accelerate easier are most of the time smaller, such as a mini-cooper. Smaller cars tend to have higher acceleration (as shown by the negative correlation between acceleration and weight), and therefore smaller weight = better mpg.

Given these relationships, all these variables are useful in predicting mpg, but the most useful ones are weight, horsepower, displacement, and cylinders as they display the strongest correlations with mpg.