Math 185 Homework 5

Lucien Chen

2024-05-25

Homework 5

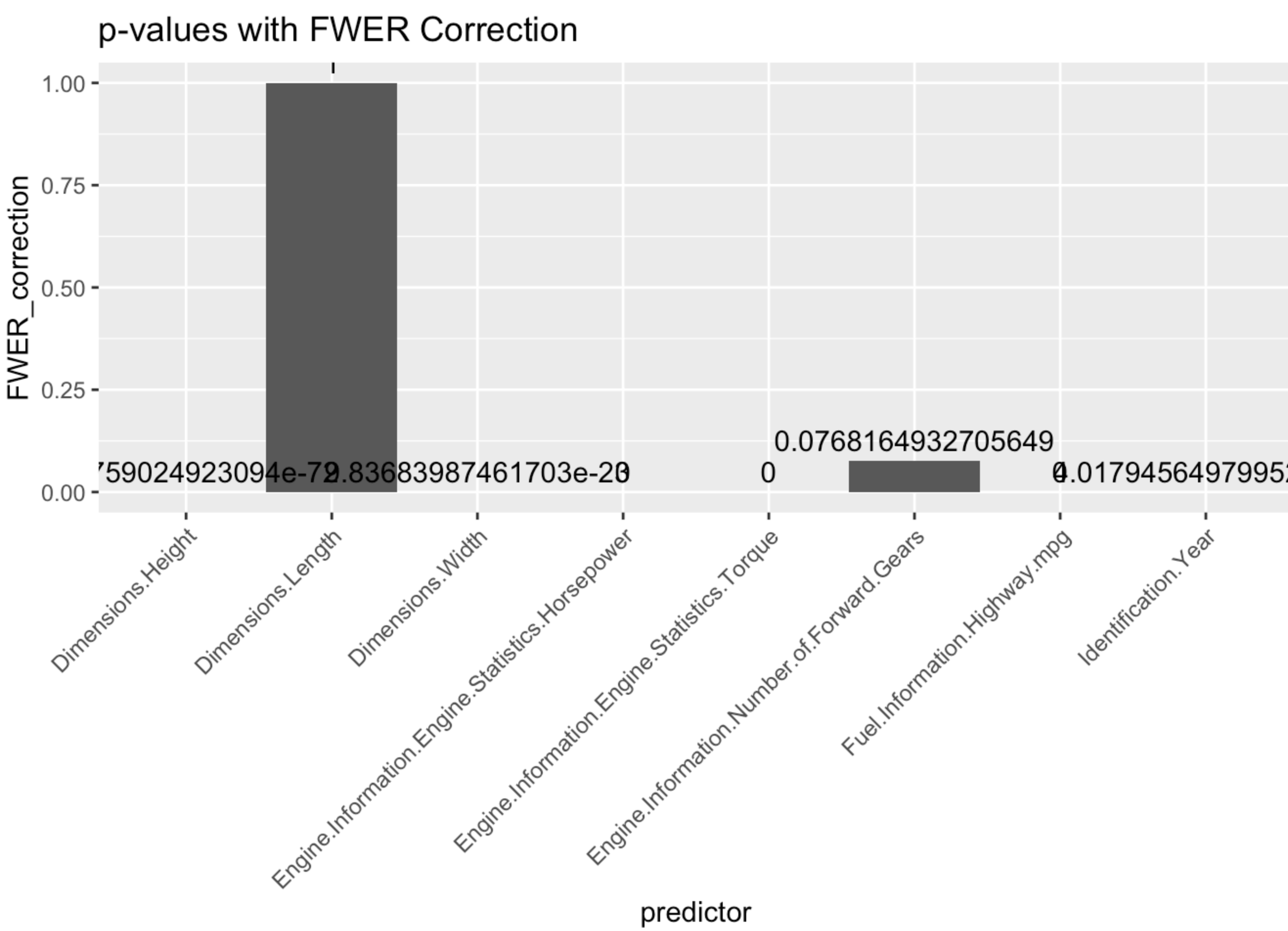
```
library(readr)
library(dplyr)
library(ggplot2)
df <- read.csv("cars.csv") %>% select(-c(Engine.Information.Engine.Type, Identification.ID, Identification.Model.
Year))
```

Question 1

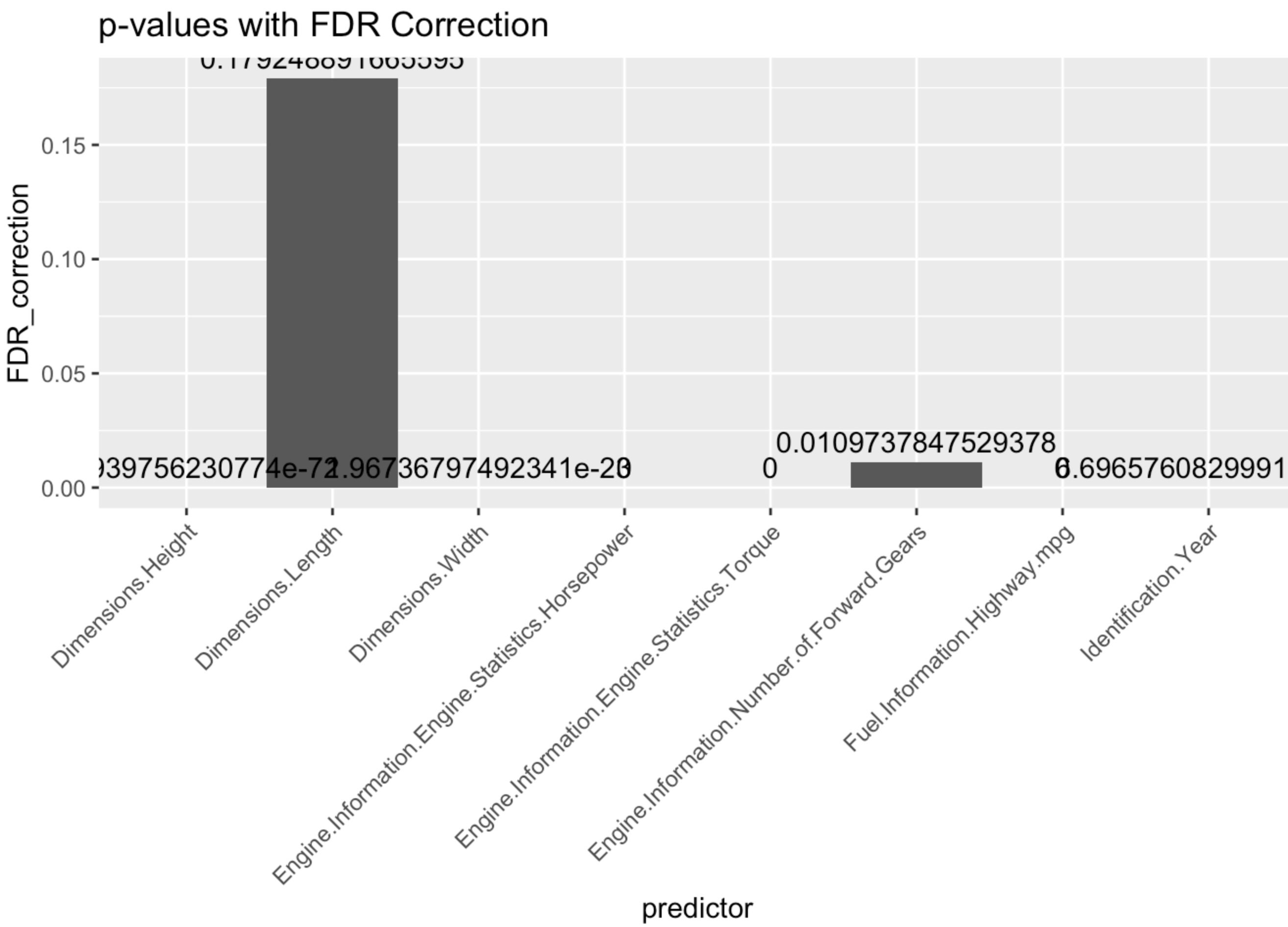
```
cols <- names(df %>% select(-Fuel.Information.City.mpg))
response <- df$Fuel.Information.City.mpg

p_vals <- list()

for (col in cols) {
  predictor <- df[[col]]
  if (is.numeric(predictor)) {
    test <- cor.test(response, predictor)
    p_vals[col] <- test$p.value
  } else if (is.factor(predictor)) {
    test <- aov(response ~ predictor)
    p_vals[col] <- test$p.value
  }
}

p_values <- data.frame(
  predictor = names(p_vals),
  p.value = unlist(p_vals)
)

p_values
```

```
##                                                                                    predictor      p.value
## Dimensions.Height                                                         Dimensions.Height 7.996988e-73
## Dimensions.Length                                                         Dimensions.Length 1.792489e-01
## Dimensions.Width                                                           Dimensions.Width 1.229605e-23
## Engine.Information.Number.of.Forward.Gears         Engine.Information.Number.of.Forward.Gears 9.602062e-03
## Fuel.Information.Highway.mpg                                   Fuel.Information.Highway.mpg 0.000000e+00
## Identification.Year                                                     Identification.Year 5.022432e-11
## Engine.Information.Engine.Statistics.Horsepower Engine.Information.Engine.Statistics.Horsepower 0.000000e+00
## Engine.Information.Engine.Statistics.Torque         Engine.Information.Engine.Statistics.Torque 0.000000e+00
```

```
p_values$FWER_correction <- p.adjust(p_values$p.value, method="bonferroni")
p_values$FDR_correction <- p.adjust(p_values$p.value, method="BH")
ggplot(p_values, aes(x=predictor, y=FWER_correction)) + geom_bar(stat="identity") + geom_text(aes(label = FWER_co
rrection), vjust = -0.5) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(title="p-values with F
WER Correction")
```



```
ggplot(p_values, aes(x=predictor, y=FDR_correction)) + geom_bar(stat="identity") + geom_text(aes(label = FDR_corr
ection), vjust = -0.5)+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(title="p-values with FDR
Correction")
```



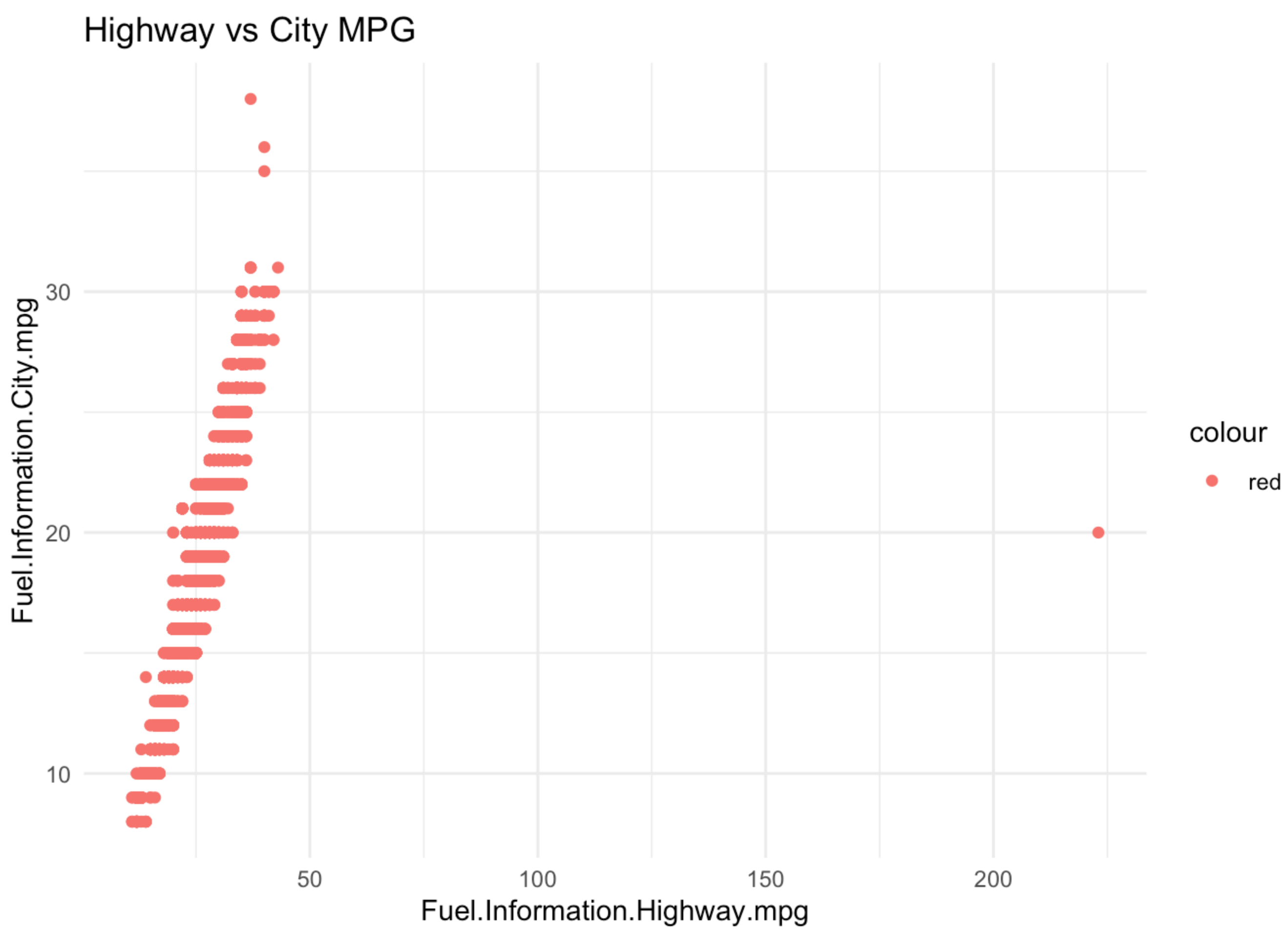
As we can see from the results of our corrections, the Bonferroni procedure is much more conservative than the Benjamini-Hochberg procedure. The p-values for the variables are much higher, i.e Dimensions.Length and Engine.Information.Number.of.Forward.Gears, and so at a given alpha level, say $\alpha = 0.05$, we would end up failing to reject number of forward gears as a significant predictor, whereas under the BH procedure, we would reject. In both cases, we would fail to reject length as a significant predictor although in the Bonferroni procedure, the p-value becomes much higher (close to 1).

Question 2

```
response <- df$Fuel.Information.City.mpg
predictor <- df$Fuel.Information.Highway.mpg

corr <- cor(response, predictor)
p_value <- cor.test(response, predictor)$p.value
print(c(corr, p_value))
```

```
## [1] 0.8656173 0.0000000
```

```
ggplot(df, aes(x=Fuel.Information.Highway.mpg, y=Fuel.Information.City.mpg, color="red")) + geom_point() + theme_
minimal() + labs(title="Highway vs City MPG")
```

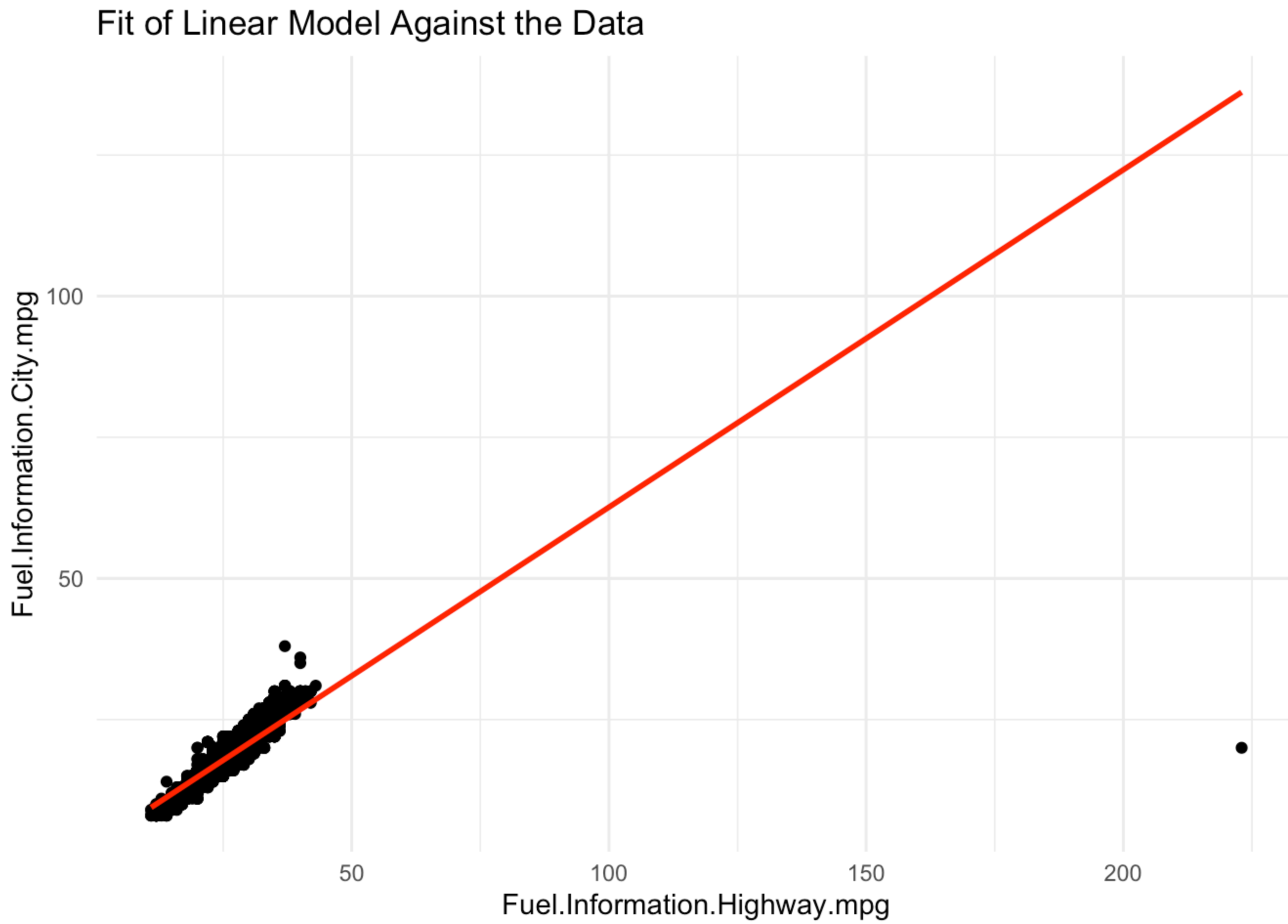


```r
lr <- lm(response ~ predictor)
summary(lr)
```

```
##
## Call:
## lm(formula = response ~ predictor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.127   -0.994   -0.201    0.787   13.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.857981   0.121239   23.57   <2e-16 ***
## predictor   0.597618   0.004853  123.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 5074 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7492
## F-statistic: 1.516e+04 on 1 and 5074 DF,  p-value: < 2.2e-16
```

```r
ggplot(df, aes(x=Fuel.Information.Highway.mpg, y=Fuel.Information.City.mpg)) + geom_point() + geom_smooth(method
="lm", se=F, color="red") + labs(title="Fit of Linear Model Against the Data") + theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

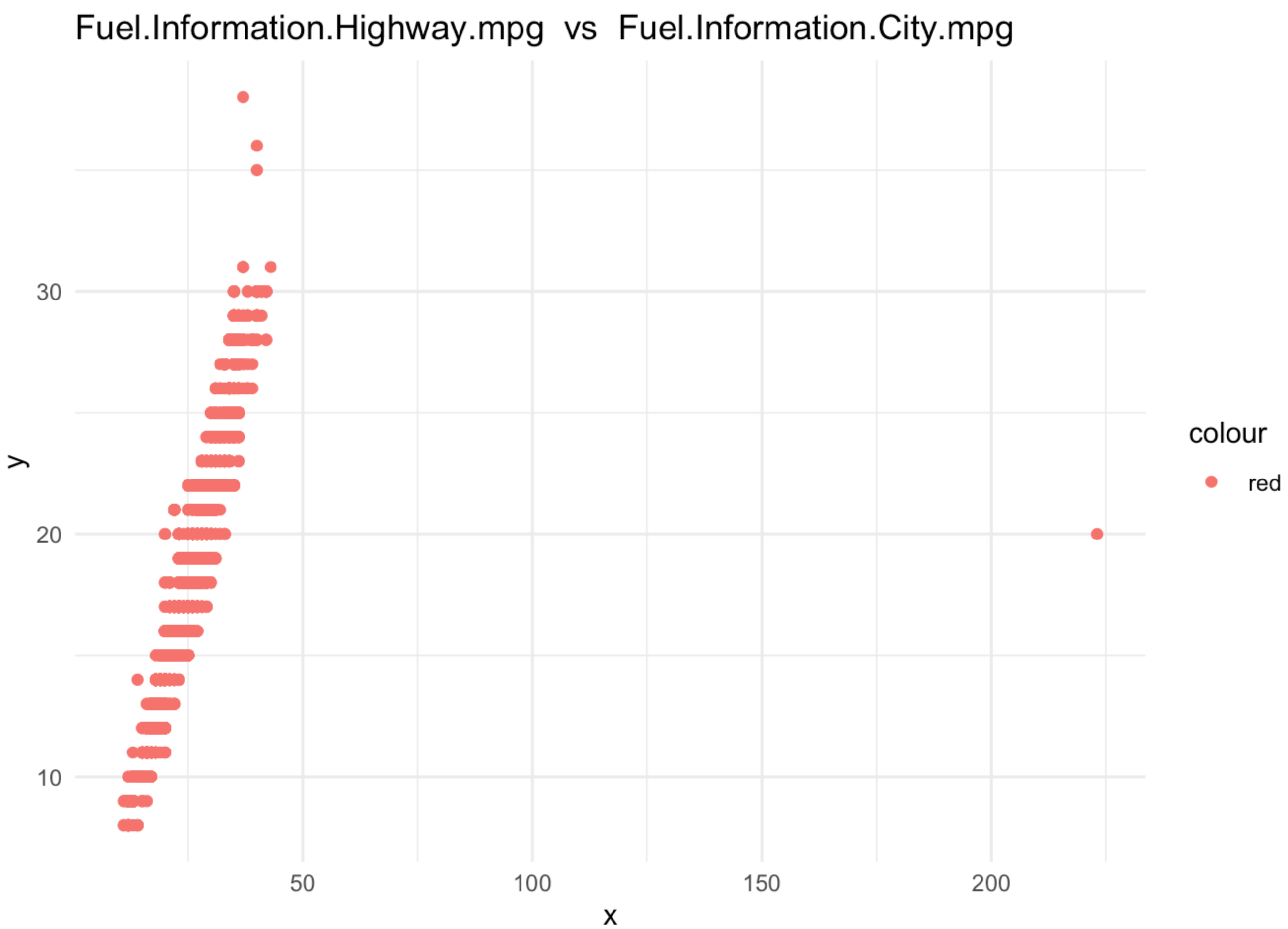


As we can see from the above code, the association is quite high, with a correlation of ~0.866 between the highway and city mpg. We can also see that our cor test has a p-value of 0 which indicates that there is significant association between the two variables. In the scatterplot, we can see that there is quite a good linear relationship between the variables with the exception of one outlier on the x-axis. Our linear regression also fits the data quite well, barring the outlier mentioned before (however this is likely to be an error).

Question 3

```
association_analysis <- function(data) {
  cols <- names(data)
  x <- data[[cols[1]]]
  y <- data[[cols[2]]]
  corr <- cor(x, y)
  p_value <- cor.test(x, y)$p.value
  print(c(corr, p_value))
  scatter <- ggplot(mapping=aes(x=x, y=y, color="red")) + geom_point() + theme_minimal() + labs(title=paste(cols
[1], " vs ", cols[2]))
  print(scatter)
  lr <- lm(y ~ x)
  print(summary(lr))
  lr_plot <- ggplot(mapping=aes(x=x, y=y)) + geom_point() +  geom_smooth(method="lm", se=F, color="red") + theme_
minimal() + labs(title=paste("Line of best fit for", cols[1], "and", cols[2]))
  print(lr_plot)
  return()
}

paired_data <- df %>% select(c(Fuel.Information.Highway.mpg, Fuel.Information.City.mpg))
association_analysis(paired_data)
```

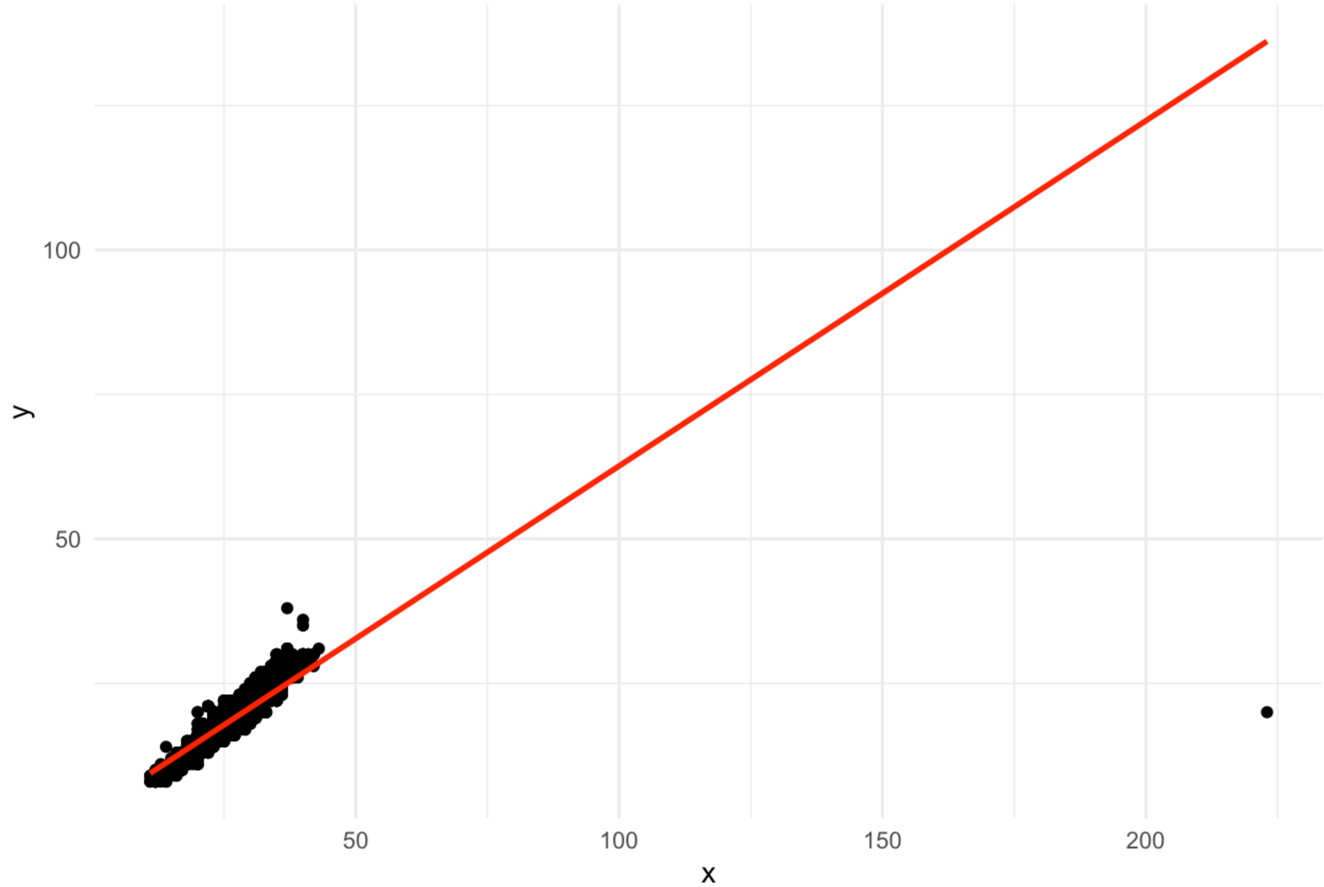
```
## [1] 0.8656173 0.0000000
```


Fuel.Information.Highway.mpg vs Fuel.Information.City.mpg
30
20
10
y
50
100
150
200
x
colour
red

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.127   -0.994   -0.201    0.787   13.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.857981   0.121239   23.57   <2e-16 ***
## x           0.597618   0.004853  123.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 5074 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7492
## F-statistic: 1.516e+04 on 1 and 5074 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Line of best fit for Fuel.Information.Highway.mpg and Fuel.Information.City.mpg




```
## NULL
```

As we can see, our function produces the same output as in Question 2.