Homework 7

Greg DePaul

2023-03-28

Problem 2 - Cars Exploratory Data Analysis.

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file and its corresponding .html file.

```
my_data <- read.csv(file = 'Cars.csv')
```

(a) Conduct a visual inspection of the data in "Cars.csv" and then read the data into R.

```
my_data$horsepower = as.numeric(my_data$horsepower)
```

(b) Are there missing values? If so, replace missing values by "NA".

```
## Warning: NAs introduced by coercion
```

```
#which(my_data$horsepower=='')
```

```
sapply(my_data,class)
```

(c) Check the variable types. Which variables do you think should be treated as quantitative and which should be treated as qualitative/categorical? Fix the problems that you have identified (if any).

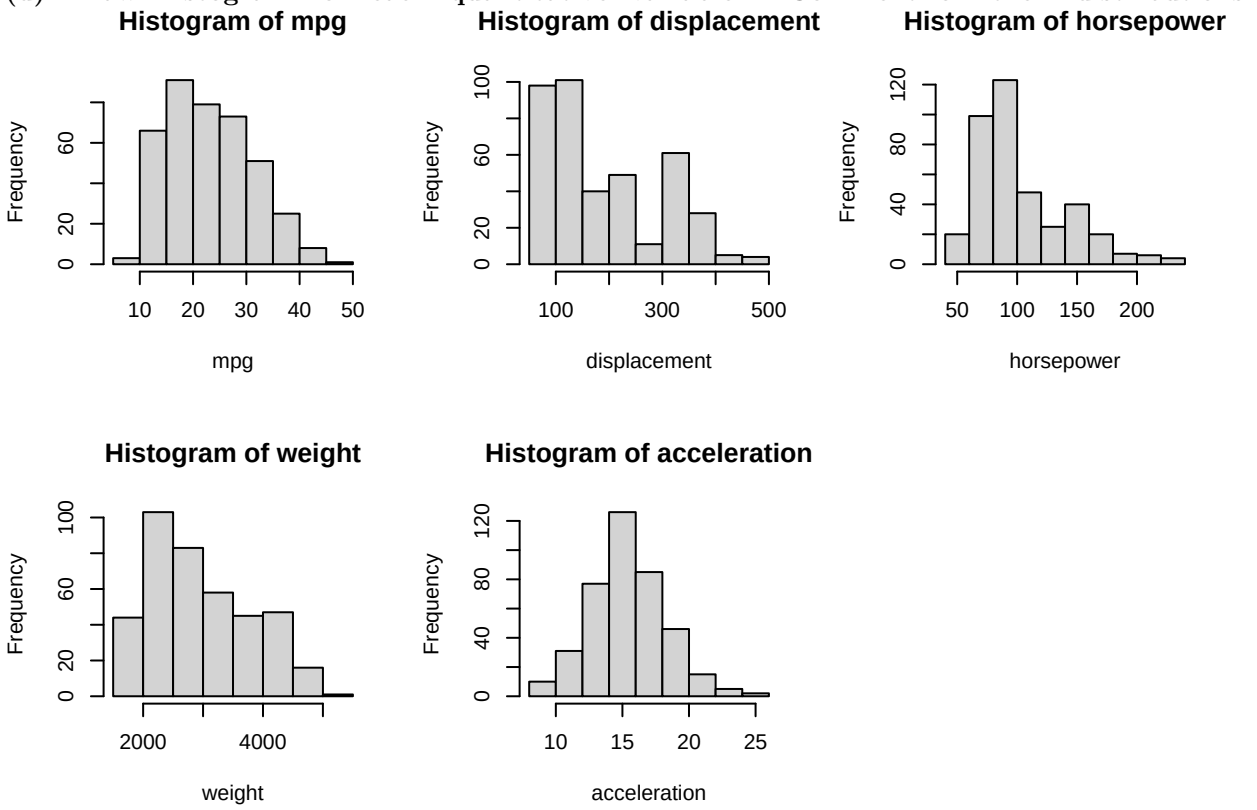
```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    "numeric"    "integer"    "numeric"    "numeric"    "integer"    "numeric"
## country.code
##    "integer"
```

Quantitative: mpg, displacement, horsepower, weight, acceleration

Categorical: cylinders, country.code

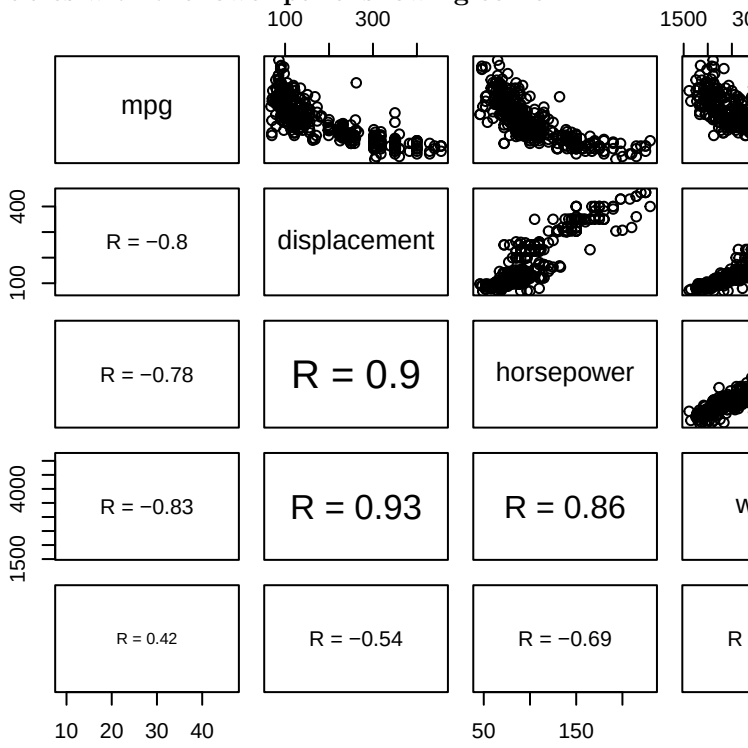
```
par(mfrow = c(2, 3))
hist(my_data$mpg, xlab='mpg', main='Histogram of mpg')
hist(my_data$displacement, xlab='displacement', main='Histogram of displacement')
hist(as.numeric(my_data$horsepower), xlab='horsepower', main='Histogram of horsepower')
hist(my_data$weight, xlab='weight', main='Histogram of weight')
hist(my_data$acceleration, xlab='acceleration', main='Histogram of acceleration')
```

(d) Draw histogram for each quantitative variable. Comment on their distributions.

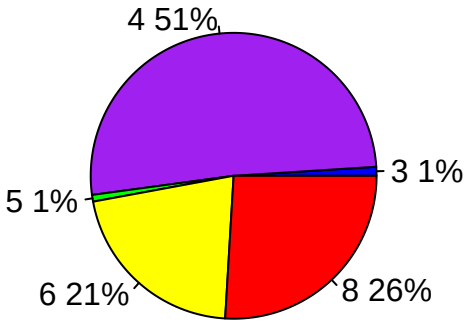



```
panel.cor <- function(x, y){
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use="complete.obs"), 2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

mpg <- my_data$mpg
displacement <- my_data$displacement
horsepower <- as.numeric(my_data$horsepower)
weight <- my_data$weight
acceleration <- my_data$acceleration
pairs(~ mpg + displacement + horsepower + weight + acceleration, lower.panel = panel.cor)
```

(e) Draw scatter plot matrix among quantitative variables with the lower panel showing correlation coefficients. Comment on their relationships.



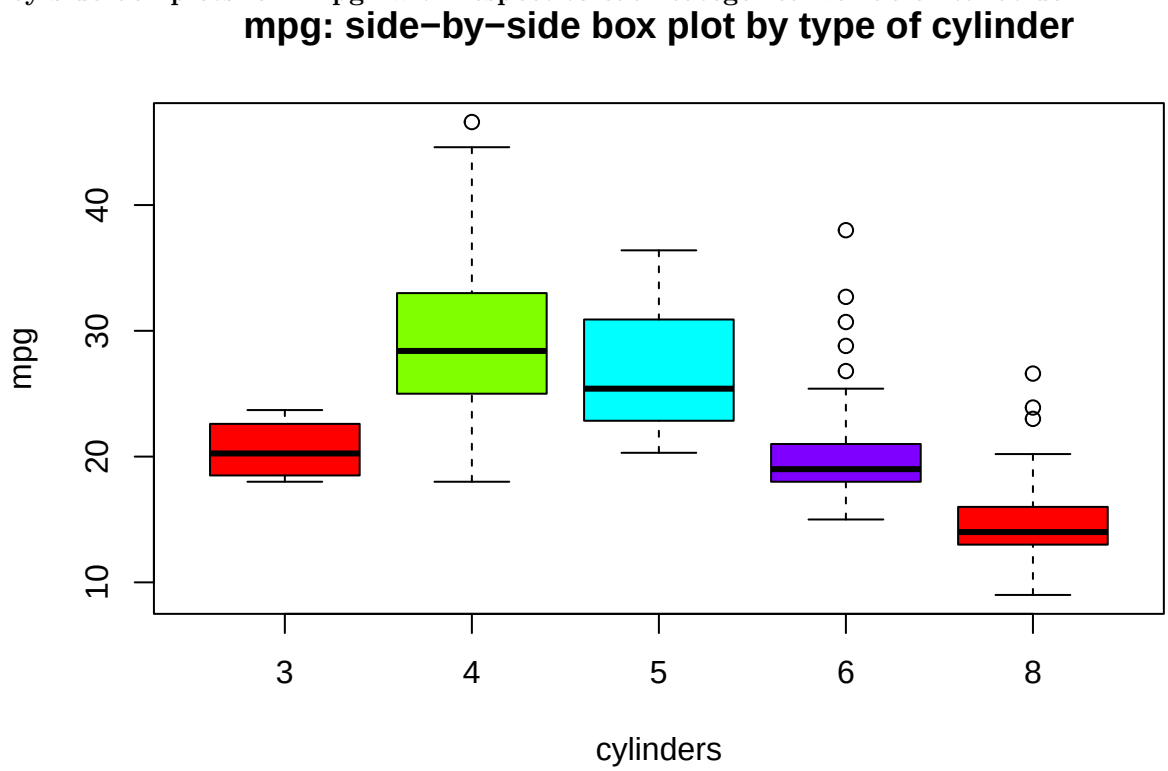

```
par(mfrow = c(1, 2))
n <- nrow(my_data)
pct <- round(100*table(my_data$cylinders)/n)
lbls <- names(pct)
vals <- as.numeric(pct)
lab <- paste(lbls,vals, sep=' ')
lab <- paste(lab,'%',sep='')
pie(table(my_data$cylinders),labels=lab,col=c('blue','purple','green','yellow','red'),main='Frame: car

n <- nrow(my_data)
pct <- round(100*table(my_data$country.code)/n)
lbls <- names(pct)
vals <- as.numeric(pct)
lab <- paste(lbls,vals, sep=' ')
lab <- paste(lab,'%',sep='')
pie(table(my_data$country.code),labels=lab,col=c('blue','purple','green','yellow'),main='Frame: country
```



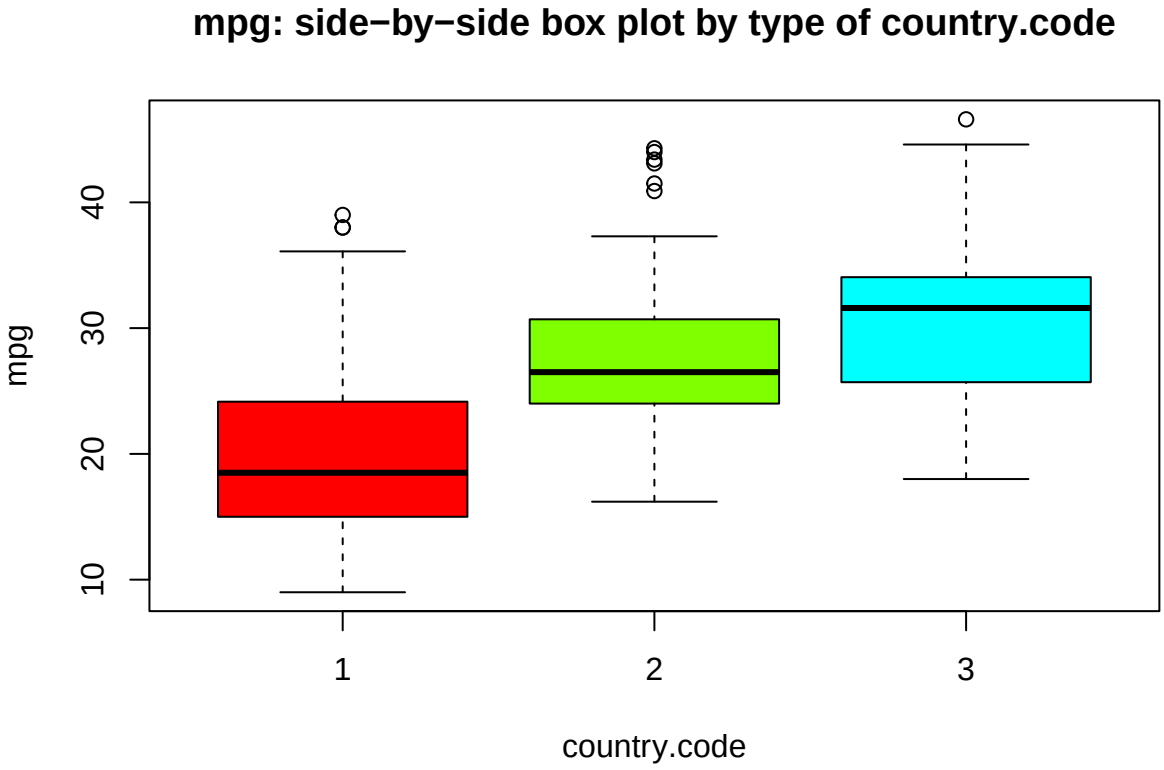
(f) Draw pie chart (with class percentage) for each categorical variable.

```
boxplot(my_data$mpg ~ my_data$cylinders,main='mpg: side-by-side box plot by type of cylinder',
xlab='cylinders',ylab='mpg',col=rainbow(4))
```

(g) Draw side-by-side box plots for "mpg" with respect to each categorical variable. What do you observe?




```
boxplot(my_data$mpg ~ my_data$country.code,main='mpg: side-by-side box plot by type of country.code',
xlab='country.code',ylab='mpg',col=rainbow(4))
```



Problem 3 - Cars Regression with Categorical Variables.

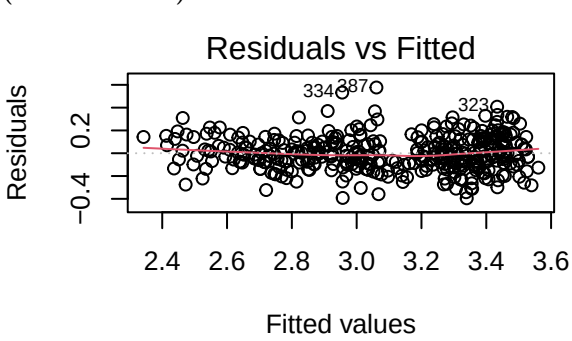
In this question, we consider models for "mpg" using "cylinders", "horsepower", and "weight" as predictors, where "cylinders" should be treated as a categorical variable.

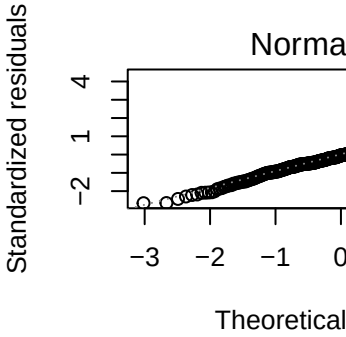
```
my_data$logmpg <- log(my_data$mpg)
```

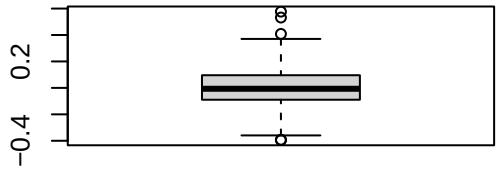
(a) Decide on whether you'd like to make any transformation of the "mpg".

```
my_data$cylinders.f  <- factor(my_data$cylinders)
fit_1 = lm(logmpg ~ cylinders.f + horsepower + weight, data=my_data)
par(mfrow = c(2, 2))
plot(fit_1,which=1) ##residuals vs. fitted values
plot(fit_1,which=2) ##residuals Q-Q plot
boxplot(fit_1$residuals) ## residuals boxplot
```

(b) Fit a first-order model with the (transformed) variables. Conduct model diagnostics. Does this model appear to be adequate?








```
fit_1$coefficients
```

(c) Derive the regression function for cars with 4 cylinders.

```
##   (Intercept)  cylinders.f4  cylinders.f5  cylinders.f6  cylinders.f8
##  3.7637094434  0.2664930665  0.3581547202  0.1287068811  0.1739937306
##    horsepower        weight
## -0.0026939945 -0.0001998106
```

If we set $cyl = 4$, then

$$log(mpg) \approx 3.7637094434 + 0.2664930665 - 0.0026939945 \cdot horsepower - 0.0001998106 \cdot weight$$

```
fit_2 = lm(logmpg ~ cylinders.f + horsepower + weight + cylinders.f:horsepower + cylinders.f:weight, da
print(fit_2$coefficients)
```

(d) Fit a model including interactions between "cylinders" and "horsepower", and "cylinders" and "weight". Derive the regression function for cars with 4 cylinders.

```
##             (Intercept)            cylinders.f4            cylinders.f5
##            -13.57939312             17.72135437             19.28318035
##            cylinders.f6            cylinders.f8              horsepower
##             17.31614208             17.30915646              0.84061135
##                  weight cylinders.f4:horsepower cylinders.f5:horsepower
##             -0.02786500             -0.84607037             -0.85790610
## cylinders.f6:horsepower cylinders.f8:horsepower     cylinders.f4:weight
##             -0.83989621             -0.84321244              0.02771060
##     cylinders.f5:weight     cylinders.f6:weight     cylinders.f8:weight
##              0.02754281              0.02760573              0.02771215
```

If we set $cyl = 4$, then

$log(mpg) \approx -13.57939312 + 17.72135437 + 0.84061135 \cdot horsepower - 0.02786500 \cdot weight - 0.84607037 \cdot horsepower + 0.02771060 \cdot$

```
summary(fit_1)
```

(e) Compare the two models using the function anova(). What do you find?

```
##
## Call:
## lm(formula = logmpg ~ cylinders.f + horsepower + weight, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39521 -0.08939 -0.00672  0.09552  0.57659
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.764e+00  9.133e-02  41.211  < 2e-16 ***
## cylinders.f4  2.665e-01  7.600e-02   3.507 0.000507 ***
## cylinders.f5  3.582e-01  1.161e-01   3.084 0.002186 **
## cylinders.f6  1.287e-01  7.867e-02   1.636 0.102650
## cylinders.f8  1.740e-01  8.427e-02   2.065 0.039623 *
## horsepower   -2.694e-03  4.495e-04  -5.994 4.72e-09 ***
## weight       -1.998e-04  2.319e-05  -8.615  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1495 on 385 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.8067
## F-statistic: 272.9 on 6 and 385 DF,  p-value: < 2.2e-16
```

```
anova(fit_1)
```

```
## Analysis of Variance Table
##
## Response: logmpg
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cylinders.f   4 32.049  8.0123 358.397 < 2.2e-16 ***
## horsepower    1  2.894  2.8942 129.460 < 2.2e-16 ***
## weight        1  1.659  1.6594  74.225 < 2.2e-16 ***
## Residuals   385  8.607  0.0224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit_2)
```

```
##
## Call:
## lm(formula = logmpg ~ cylinders.f + horsepower + weight + cylinders.f:horsepower +
##     cylinders.f:weight, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.37686 -0.08275 -0.00220  0.09000  0.62173
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -13.57939   35.12262  -0.387    0.699
## cylinders.f4             17.72135   35.12269   0.505    0.614
## cylinders.f5             19.28318   35.14373   0.549    0.584
## cylinders.f6             17.31614   35.12311   0.493    0.622
## cylinders.f8             17.30916   35.12288   0.493    0.622
## horsepower                0.84061    1.86548   0.451    0.653
## weight                   -0.02786    0.06256  -0.445    0.656
## cylinders.f4:horsepower  -0.84607    1.86548  -0.454    0.650
## cylinders.f5:horsepower  -0.85791    1.86549  -0.460    0.646
## cylinders.f6:horsepower  -0.83990    1.86548  -0.450    0.653
## cylinders.f8:horsepower  -0.84321    1.86548  -0.452    0.652
## cylinders.f4:weight       0.02771    0.06256   0.443    0.658
## cylinders.f5:weight       0.02754    0.06256   0.440    0.660
## cylinders.f6:weight       0.02761    0.06256   0.441    0.659
## cylinders.f8:weight       0.02771    0.06256   0.443    0.658
##
## Residual standard error: 0.1451 on 377 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8245, Adjusted R-squared:  0.818
## F-statistic: 126.5 on 14 and 377 DF,  p-value: < 2.2e-16
```

```
anova(fit_2)
```

```
## Analysis of Variance Table
##
## Response: logmpg
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## cylinders.f               4 32.049  8.0123 380.6930 < 2.2e-16 ***
## horsepower                1  2.894  2.8942 137.5138 < 2.2e-16 ***
## weight                    1  1.659  1.6594  78.8428 < 2.2e-16 ***
## cylinders.f:horsepower    4  0.587  0.1467   6.9691 2.021e-05 ***
## cylinders.f:weight        4  0.086  0.0214   1.0188    0.3974
## Residuals               377  7.935  0.0210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interactions model isn't very necessary. That's which the coefficients are virtually identical.

```
newX = data.frame(cylinders.f = factor(4), horsepower = 100, weight = 3000)
predict(fit_1, newX, interval='confidence', level=0.95)
```

(f) Construct a 95% prediction interval of "mpg" for a car with 4 cylinders, 100 horsepower and 3000 pounds under these two models. What do you observe?

```
##        fit      lwr      upr
## 1 3.161371 3.126228 3.196514
```

```
predict(fit_2, newX, interval='confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 3.132879 3.086912 3.178846
```

To two intervals are very similar, although the interaction model yields a very slightly small interval.