test.R

harry

2020-09-11

```
library("tidyverse")
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(imager)
```

```
## Loading required package: magrittr

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

##
## Attaching package: 'imager'

## The following object is masked from 'package:magrittr':
##
##     add

## The following object is masked from 'package:stringr':
##
##     boundary
## The following object is masked from ’package:tidyr’:
##
##     fill

## The following objects are masked from ’package:stats’:
##
##     convolve, spectrum

## The following object is masked from ’package:graphics’:
##
##     frame

## The following object is masked from ’package:base’:
##
##     save.image
```

```
’
1. This exercise relates to the College data set, which can be found in the file College.csv on the
course’s public webpage (https://scads.eecs.wsu.edu/index.php/datasets/). The dataset contains a
number of variables for 777 different universities and colleges in the US. The variables are
’
```

```
## [1] "\n1. This exercise relates to the College data set, which can be found in the file College.csv o
```

(a) Use the read.csv() function to read the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it as a list of lists or use the pandas dataframe to store your data. Call the loaded data college. Ensure that your column headers are not treated as a row of data.

```
college<-read.csv("college.csv", header=TRUE)
```

(b) Find the median cost of books for all schools in this dataset.’

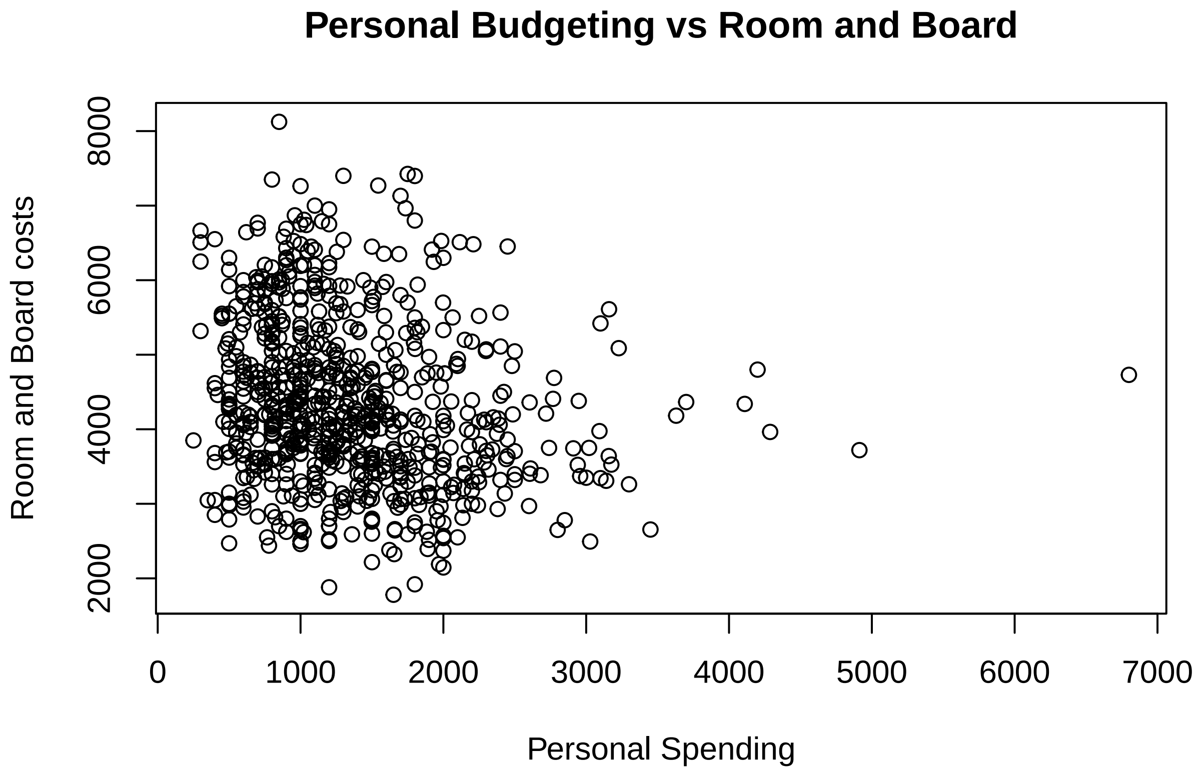
```
summary(college)
```

```
##       X               Private               Apps           Accept
##  Length:777         Length:777         Min.   :   81   Min.   :   72
##  Class :character   Class :character   1st Qu.:  776   1st Qu.:  604
##  Mode  :character   Mode  :character   Median : 1558   Median : 1110
##                                        Mean   : 3002   Mean   : 2019
##                                        3rd Qu.: 3624   3rd Qu.: 2424
##                                        Max.   :48094   Max.   :26330
##      Enroll       Top10perc       Top25perc      F.Undergrad
##  Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
##  1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
##  Median : 434   Median :23.00   Median : 54.0   Median : 1707
##  Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
##  3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
##  Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
##   P.Undergrad        Outstate       Room.Board       Books
##  Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
##  1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
##  Median :  353.0   Median : 9990   Median :4200   Median : 500.0
##  Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
##  3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
##  Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
##     Personal          PhD            Terminal       S.F.Ratio
##  Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
##  1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
##  Median :1200   Median : 75.00   Median : 82.0   Median :13.60
##  Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
##  3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
##  Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
##   perc.alumni        Expend        Grad.Rate
##  Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :21.00   Median : 8377   Median : 65.00
##  Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :64.00   Max.   :56233   Max.   :118.00
```

```
#median cost of books = 500
```

(c) Produce a scatterplot that shows a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title. '

```
plot(college$Personal,college$Room.Board,xlab='Personal Spending',
     ylab='Room and Board costs',main='Personal Budgeting vs Room and Board')
```



(d) Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools. Ensure whatever figures you produce have appropriate axis labels and a title. ’

```
private <- subset(college,Private == "Yes")
public <- subset(college,Private == "No")
```

```
private_data <- data.frame(
  type = c("private full time" ,"private part time"  ),
  value = c( private$F.Undergrad,private$P.Undergrad )
)

private_plot <- private_data %>%
  ggplot( aes(x=value, fill=type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = ’identity’,binwidth = 100) +
  scale_fill_manual(values = c("#96b3a2","#404080" )) +
  labs(fill="",title = "Distribution of full and part time undergrads in private schools.",
  x = "Student Population" , y = "Frequency")
```

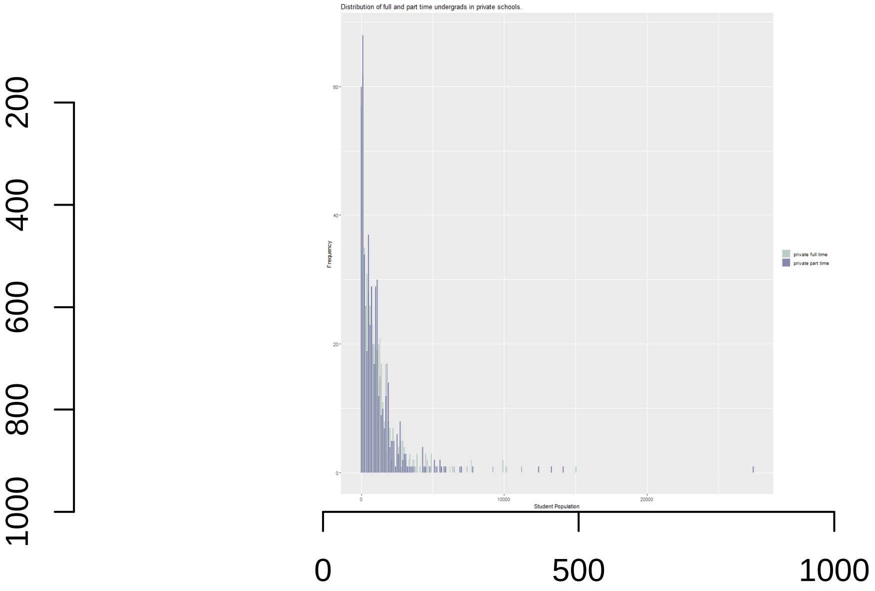
```
public_data <- data.frame(
  type = c("public full time" ,"public part time"  ),
  value = c( public$F.Undergrad,public$P.Undergrad )
)
```

```r
public_plot <- public_data %>%
  ggplot( aes(x=value, fill=type)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity',binwidth = 100) +
  scale_fill_manual(values = c("#96b3a2","#404080" )) +
  labs(fill="",title = "Distribution of full and part time undergrads in public schools.",
       x = "Student Population" , y = "Frequency")
```

```r
png(filename="private.png",width=1000, height=1000)
print(private_plot)
dev.off()
```

```
## pdf
##   2
```

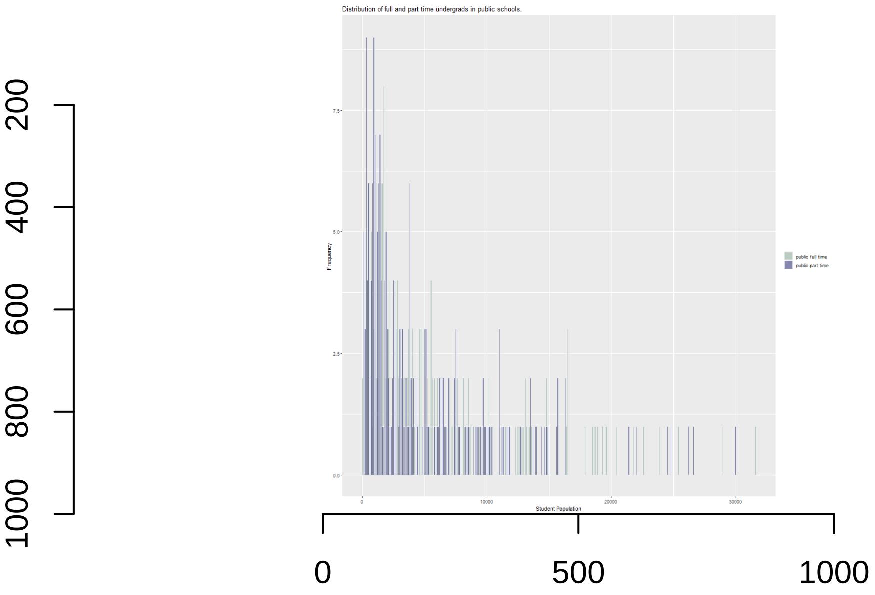
```r
im<-load.image("private.png")
plot(im)
```



```r
png(filename="public.png",width=1000, height=1000)
print(public_plot)
dev.off()
```

```
## pdf
##   2
```

```r
im<-load.image("public.png")
plot(im)
```

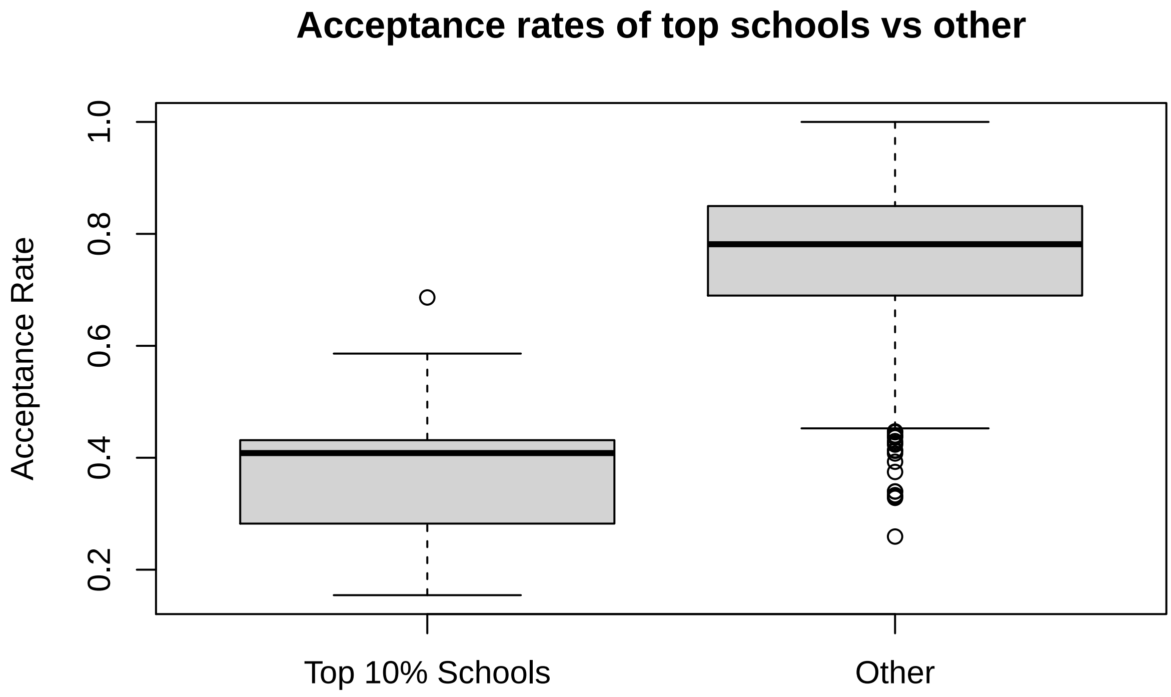


(e) Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories (Yes and No). Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 75%.'

```r
college <- transform(college, Top = ifelse(Top10perc >= 75, "Yes", "No"))
college <- transform(college, Accept.Rate = Accept / Apps)

top_schools <- college$Accept.Rate[which(college$Top=="Yes")]
not_schools <- college$Accept.Rate[which(college$Top=="No")]
```

Now produce side-by-side boxplots of the schools' acceptance rates (based on Accept and Apps) for each of the two Top categories. There should be two boxes on your figure, one for top schools and one for others. How many top universities are there?

```r
boxplot(top_schools,not_schools, main = "Acceptance rates of top schools vs other",
        ylab = "Acceptance Rate", names = c("Top 10% Schools", "Other"))
```

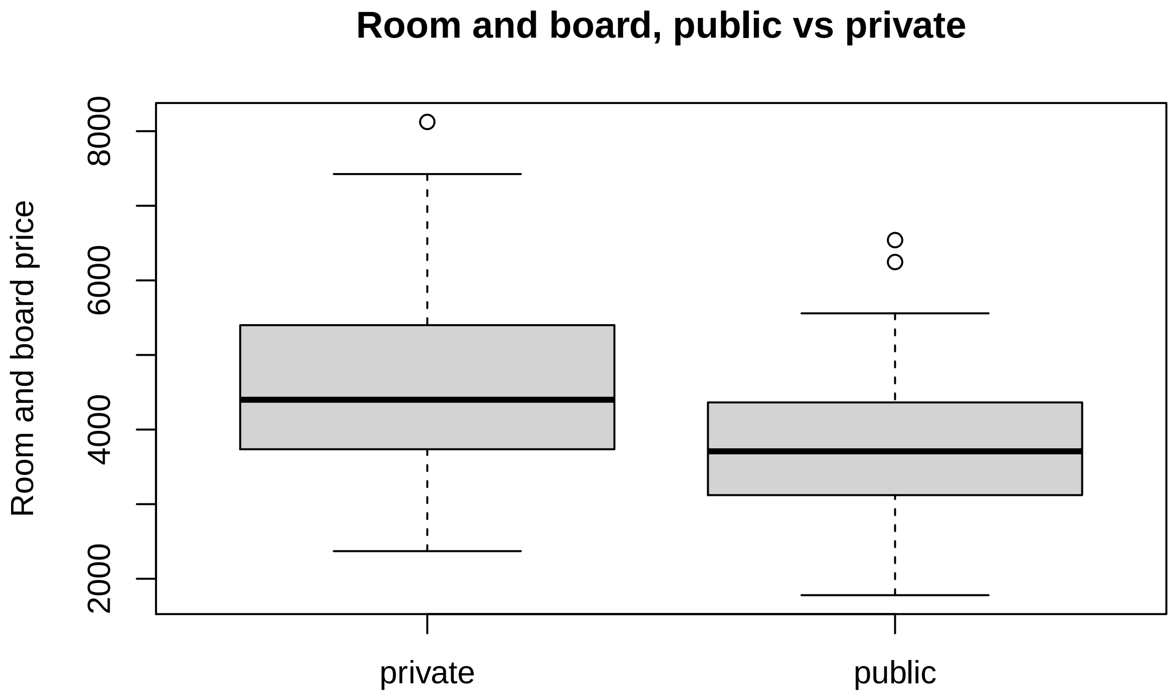


```r
top_colleges <- college[college$Top == "Yes", ]

nrow(top_colleges) #NROW = 26, Therefore there are 26 top universities
```

```
## [1] 26
```

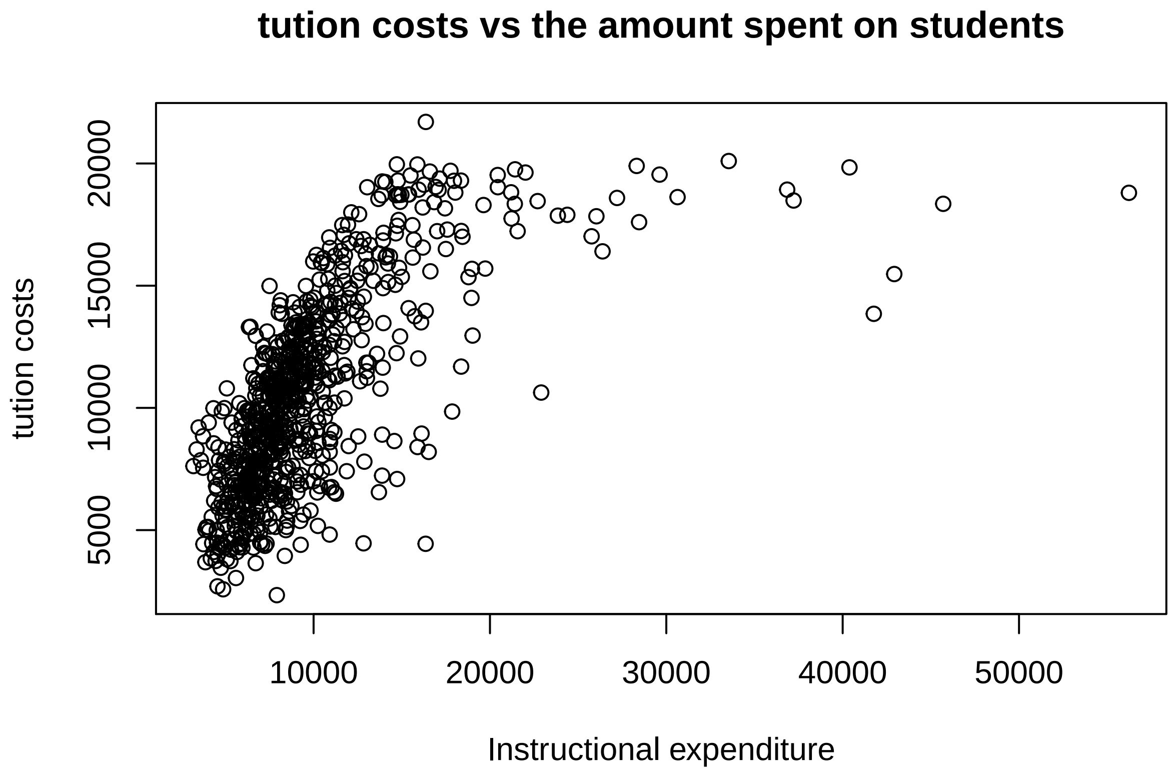
(f) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge. ”

```r
private_room <- private$Room.Board
public_room <- public$Room.Board

boxplot(private_room,public_room, main = "Room and board, public vs private",
        ylab = "Room and board price", names = c("private", "public"))
```



```
#In most to all cases, on average, Room and Board expenses are greater for private universities
#than they are for their public counterparts.
```

```
expenses = college$Expend
tuition = college$Outstate
plot(expenses, tuition,main = "tution costs vs the amount spent on students",
     xlab="Instructional expenditure", ylab="tution costs")
```



```
#Tuition costs appear to rise at a rate much faster than the expenditure per student increases.
```

2. This exercise involves the Auto.csv data set found on the course website. The features of the dataset are as follows: • mpg: miles per gallon • cylinders: number of cylinders • displacement: volume of air displaced by cylinders • horsepower: power of the car (rate of work) • weight: how much the car weighs in lb • acceleration: rate at which car accelerates • year: when the car was made • origin: where the car comes from (1=USA, 2=Germany, 3=Japan) • name: the make and model of the car Make sure that rows with missing values have been removed from the data. For part, show both the code you used and any relevant outputs.

(a) Specify which of the predictors are quantitative (measuring numeric properties such as size, or quantity), and which are qualitative (measuring non-numeric properties such as color, appearance, type etc.)? Keep in mind that a qualitative variable may be represented as quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings. Quantitative predictors:

```
# ’o    MPG
# ’o    Cylinders
# ’o    Displacement
# ’o    Horsepower
# ’o    Weight
# ’o    Acceleration
# ’o    Year
```

Qualitative predictors:

```
# ’o    Name
# ’o    Origin
```

(b) What is the range, mean and standard deviation of each quantitative predictor?

```
#read in data
autos<-na.omit(read.csv("auto.csv", header=TRUE))
autos$horsepower <- as.numeric(as.character(autos$horsepower))
```

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

```
#remove missing rows
autos <- autos[complete.cases(autos),]

summary(autos)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin          name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

```
max(autos$mpg)-min(autos$mpg)
```

```
## [1] 37.6
```

```
sd(autos$mpg)
```

```
## [1] 7.805007
```

```
max(autos$cylinders) - min((autos$cylinders))
```

```
## [1] 5
```

```
sd((autos$cylinders))
```

```
## [1] 1.705783
```

```r
max(autos$displacement) - min((autos$displacement))
```

```
## [1] 387
```

```r
sd((autos$displacement))
```

```
## [1] 104.644
```

```r
max(autos$horsepower) - min(autos$horsepower)
```

```
## [1] 184
```

```r
sd((autos$horsepower))
```

```
## [1] 38.49116
```

```r
max(autos$weight) - min(autos$weight)
```

```
## [1] 3527
```

```r
sd(autos$weight)
```

```
## [1] 849.4026
```

```r
max(autos$acceleration) - min((autos$acceleration))
```

```
## [1] 16.8
```

```r
sd(autos$acceleration)
```

```
## [1] 2.758864
```

```r
max(autos$year) - min((autos$year))
```

```
## [1] 12
```

```r
sd(autos$year)
```

```
## [1] 3.683737
```

Quantitative predictors: • MPG • Range:37.6 • Mean: 23.45 • Std: 7.805007

- Cylinders • Range:5 • Mean:5.472 • Std: 1.705783
- Displacement • Range:387 • Mean:194.4 • Std: 104.644
- Horsepower • Range:184 • Mean:104.5 • Std: 38.49116

- Weight • Range: 3527 • Mean:2978 • Std: 849.4026
- Acceleration • Range: 16.8 • Mean:15.54 • Std: 2.758864

o Year • Range:12 • Mean: 75.98 • Std: 3.683737 (c) Now remove the 40th through 80th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
#remove 40->80 inclusive
autosC <- autos[-c(39:79),]

#get the mean
summary(autosC)
```

```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1649  
##  1st Qu.:17.85   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2226  
##  Median :23.50   Median :4.000   Median :146.0   Median : 92.0   Median :2789  
##  Mean   :23.95   Mean   :5.413   Mean   :190.2   Mean   :102.8   Mean   :2943  
##  3rd Qu.:29.80   3rd Qu.:6.000   3rd Qu.:258.0   3rd Qu.:115.5   3rd Qu.:3522  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :4997  
##   acceleration        year          origin          name          
##  Min.   : 8.00   Min.   :70.0   Min.   :1.000   Length:351        
##  1st Qu.:14.00   1st Qu.:74.0   1st Qu.:1.000   Class :character  
##  Median :15.50   Median :77.0   Median :1.000   Mode  :character  
##  Mean   :15.59   Mean   :76.5   Mean   :1.595                     
##  3rd Qu.:17.00   3rd Qu.:79.0   3rd Qu.:2.000                     
##  Max.   :24.80   Max.   :82.0   Max.   :3.000                     
```

```
max(autosC$mpg)-min(autosC$mpg)
```

```
## [1] 37.6
```

```
sd(autosC$mpg)
```

```
## [1] 7.809443
```

```
max(autosC$cylinders) - min((autosC$cylinders))
```

```
## [1] 5
```

```
sd((autosC$cylinders))
```

```
## [1] 1.663988
```

```
max(autosC$displacement) - min((autosC$displacement))
```

```
## [1] 387
```

```r
sd((autosC$displacement))
```

```
## [1] 101.1749
```

```r
max(autosC$horsepower) - min(autosC$horsepower)
```

```
## [1] 184
```

```r
sd((autosC$horsepower))
```

```
## [1] 37.52519
```

```r
max(autosC$weight) - min(autosC$weight)
```

```
## [1] 3348
```

```r
sd(autosC$weight)
```

```
## [1] 812.3924
```

```r
max(autosC$acceleration) - min((autosC$acceleration))
```

```
## [1] 16.8
```

```r
sd(autosC$acceleration)
```

```
## [1] 2.722163
```

```r
max(autosC$year) - min((autosC$year))
```

```
## [1] 12
```

```r
sd(autosC$year)
```

```
## [1] 3.546323
```

Quantitative predictors: o MPG • Range: 37.6 • Mean: 23.95 • Std: 7.809443

o Cylinders • Range:5 • Mean:5.413 • Std: 1.663988

o Displacement • Range: 387 • Mean:190.2 • Std: 101.1749

o Horsepower • Range: 184 • Mean:102.8 • Std: 37.52519

o Weight • Range: 3348 • Mean:2943 • Std: 812.3924

o Acceleration • Range: 16.8 • Mean:15.59 • Std: 2.722163

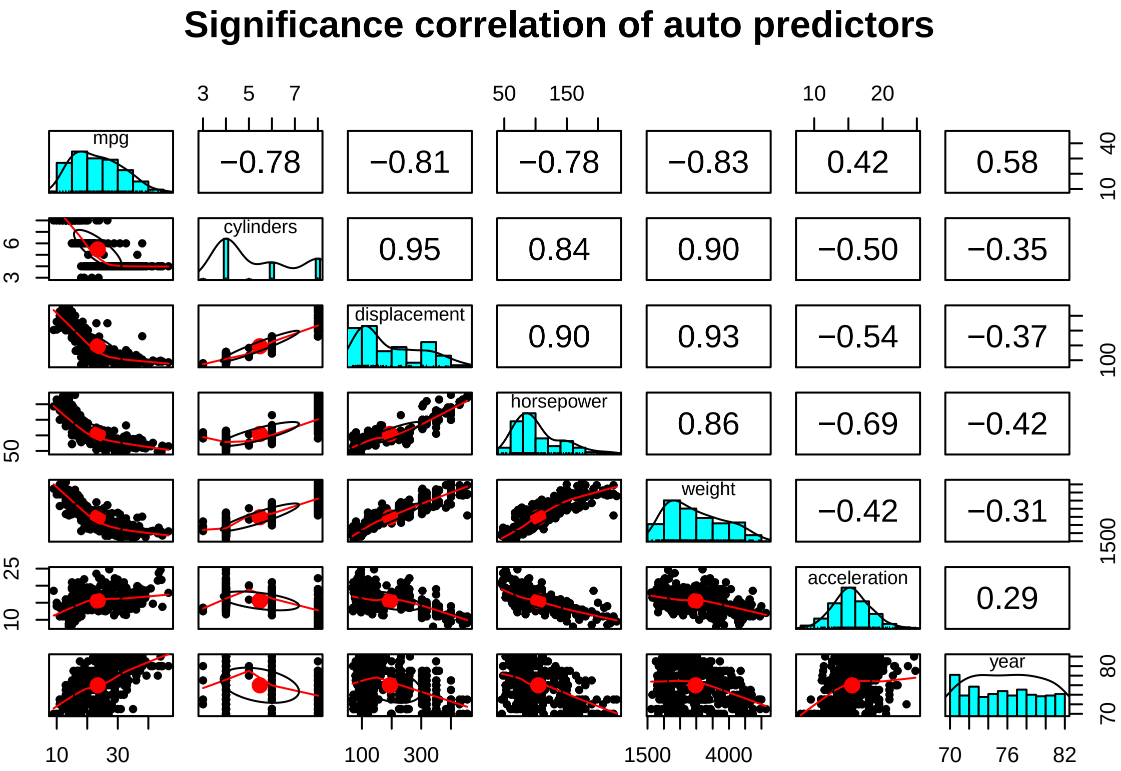
- Year • Range:12 • Mean: 76.5 • Std: 3.546323

(d) Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create a correlation matrix for the relevant variables.

```
#install.packages("psych")


library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
autos_numeric <- subset(autos, select=-c(name,origin))

pairs.panels(autos_numeric, method = "pearson", density = TRUE,
             ellipses = TRUE,main = "Significance correlation of auto predictors")
```



(e) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Which, if any, of the other variables might be useful in predicting mpg? Justify your answer based on the prior correlations.

Based on the correlation matrix, The most likely candidates for predicting mileage would be cylinders displacement and horsepower , being that they all strongly correlate negatively with mileage, The hypothesis would be whether cars with larger engines and power typically have worse mileage compared to smaller low powered engines.