Assignment Number or Description

Your Name

June, 1, 2018

Assignment Template

1.

1a. Load the `College` data set from the ISLR package. How many and what types of variables are in the dataset?

```
library(ISLR)
data(College)

# Dimension and data types
dim(College)
```

```
## [1] 777  18
```

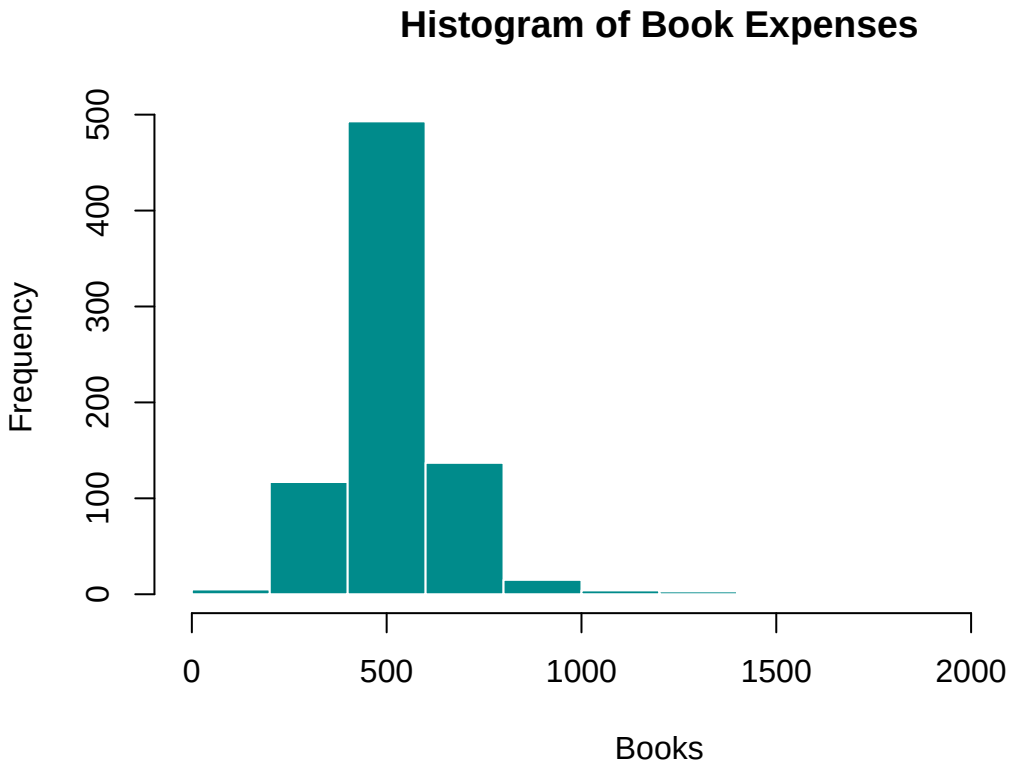
```
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

There are 18 variables. All of them are numeric except "Private" which is a factor (qualitative) variable.

1b. Plot a histogrami of one of the variables with filled in bars and no border

```
hist(College$Books, col = "darkcyan", border = "white", xlab = "Books", main = "Histogram of Book Expens
```



Question 2.

2a. Install the GGally packages using the command 'install.packages("GGally"). (This command only needs to be run once.

```
install.packages("GGally")
```

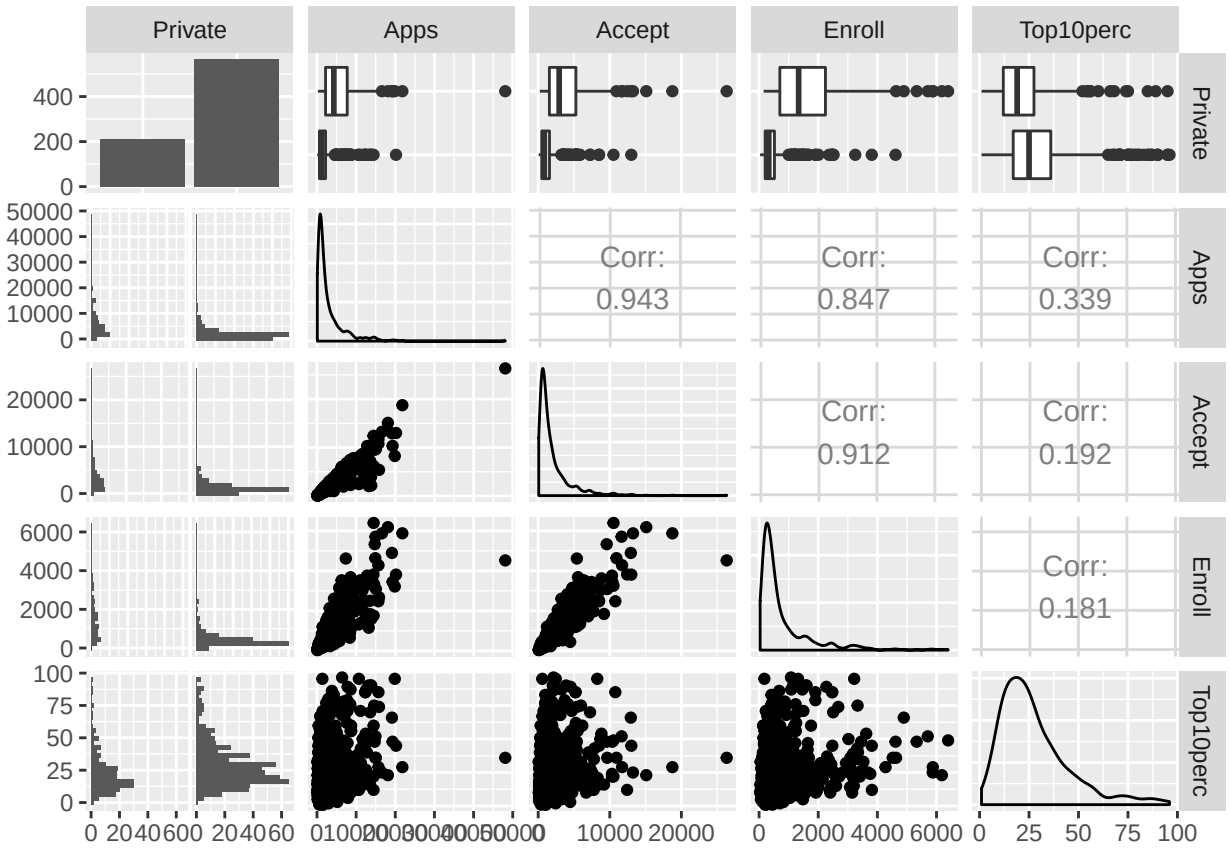
2b. Create a new dataframe with 5 of the variables from the College data set. Plot the correlation structure of this new dataset using the ggpairs function from the GGally package. Include "Private" as one of the variables in your dataset.

```
library(GGally)

# Subset data to create new dataset
new.data <- College[, c("Private", "Apps", "Accept", "Enroll", "Top10perc")]

# Correlation plot
ggpairs(new.data)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2c. Describe if any of the variables in your plotted subset of the data are correlated. Describe the relationship of the "Private" variable to one of the other variables in your subset.

Applications are positively correlated with the acceptance rate (.91) and the enrollment rate (.87). The number of applications are somewhat positively correlated with a college having a large percent of students in the top 10 percent of their class. The percent of students in the top 10 percent of their class is not highly correlated with acceptance or enrollment, however, with correlation below .20 for both.

The plots for the "Private" variable show the proportion of private schools vs public schools. The plots also show the distribution of each of the numeric variables for the two classes of schools. The boxplots show that there is a different median number for enrollment for Private vs. Public schools.