

# Assignment 2: R Basics and Exploratory Data Analysis

WenLi

9/4/2020

## 1. College Data Analysis

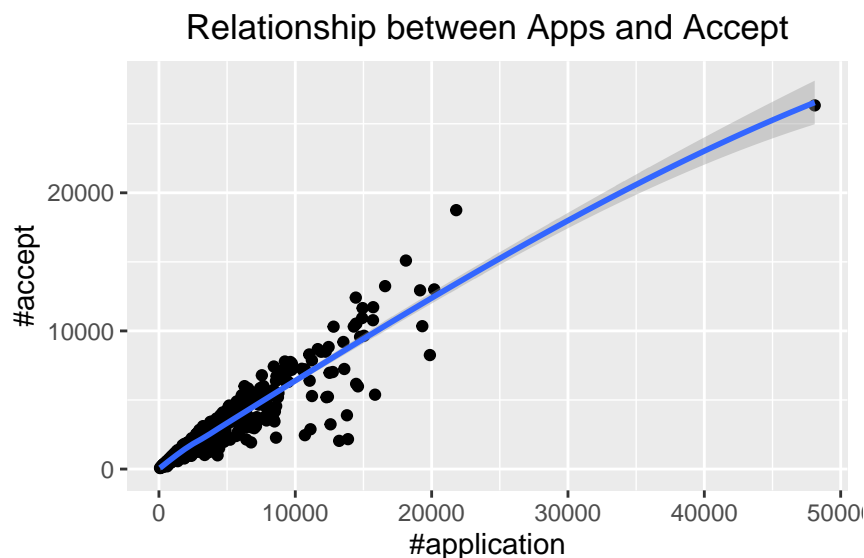
```
library(ggplot2)

##### a. Read the data into R.
College = read.csv("College.csv", header = TRUE)

##### b. Find the median cost of books.
median(College$ Books)

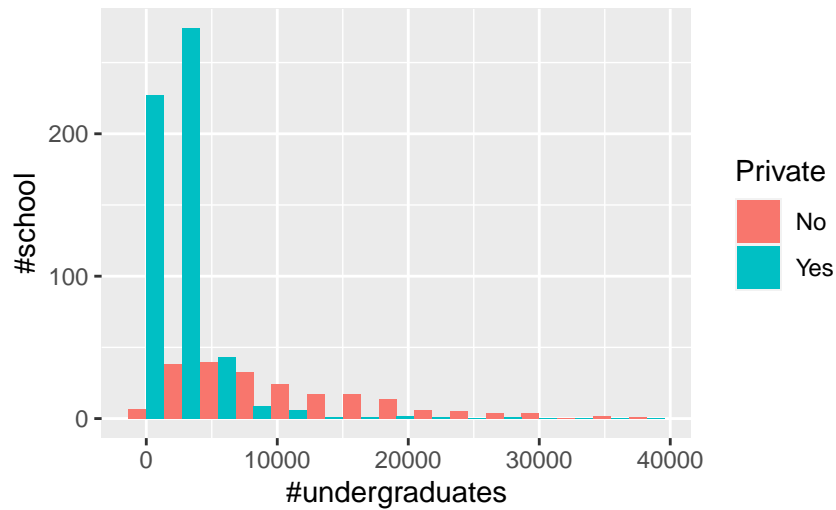
## [1] 500

##### c. Produce a scatterplot showing relationship between Apps and Accept.
ggplot(data=College, aes(x=Apps, y=Accept)) + geom_point() + geom_smooth() +
  labs(title="Relationship between Apps and Accept",
        x="#application", y="#accept") +
  theme(plot.title = element_text(hjust = 0.5))
```



```
##### d. Showing the overall enrollment numbers (P.Undergrad plus F.Undergrad)
ggplot(College, aes(x = P.Undergrad+F.Undergrad, fill = Private)) +
  geom_histogram(bins = 15, position = "dodge") +
  labs(title="Overall enrollment numbers of shcools",
        x="#undergraduates", y="#school") +
  theme(plot.title = element_text(hjust = 0.5))
```

Overall enrollment numbers of schools



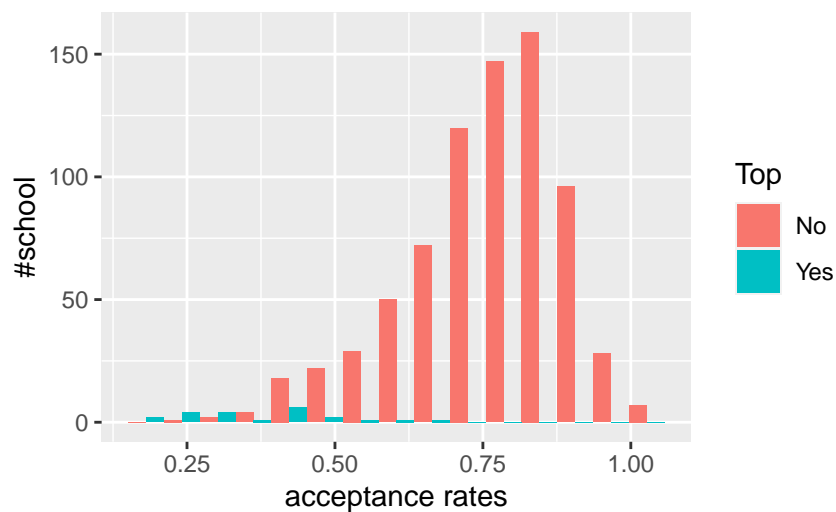
##### e. Schools' acceptance rates for each of the two Top categories

```
Top <- as.factor(ifelse(College$Top10perc > 75, "Yes", "No"))
summary(Top)
```

```
## No Yes
## 755 22
```

```
AccRate = College$Accept / College$Apps
ggplot(College, aes(x = AccRate, fill = Top)) +
  geom_histogram(bins = 15, position = "dodge") +
  labs(title="Acceptance rates for top schools and others",
       x="acceptance rates", y="#school") +
  theme(plot.title = element_text(hjust = 0.5))
```

Acceptance rates for top schools and others

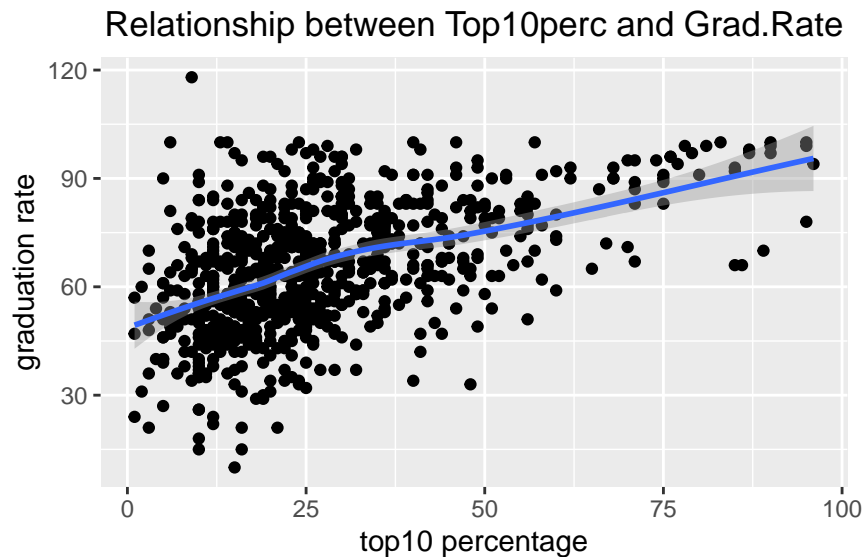


```
cat ("The number of top universities is ", sum(College$Top10perc > 75))
```

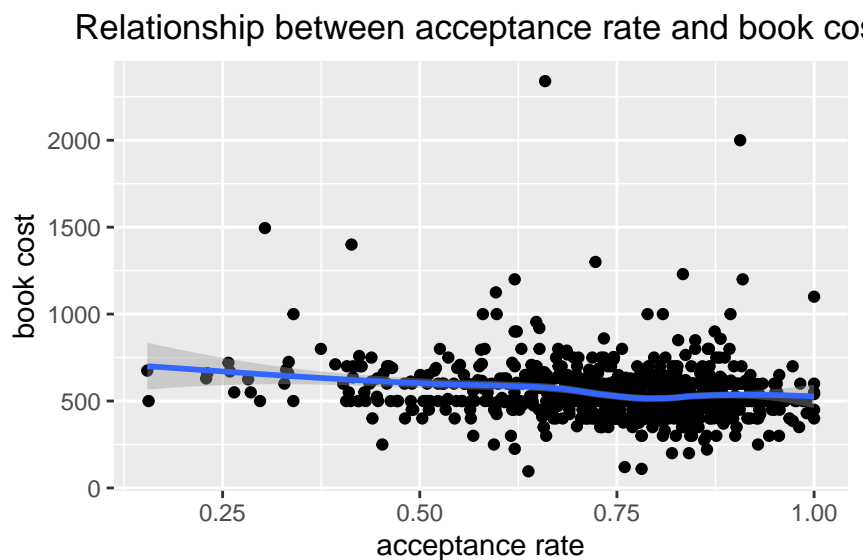
```
## The number of top universities is 22
```

```
##### f. Two new plots of any type,
```

```
# Colleges that have the most students from top 10% of high school class  
# usually have better graduation rate but do not obtain the highest graduation rate.  
ggplot(data=College, aes(x=Top10perc, y=Grad.Rate)) + geom_point() + geom_smooth() +  
  labs(title="Relationship between Top10perc and Grad.Rate",  
        x="top10 percentage", y="graduation rate") +  
  theme(plot.title = element_text(hjust = 0.5))
```



```
#Universities with various acceptance rates seem to have very close book cost.  
ggplot(data=College, aes(x=AccRate, y=Books)) + geom_point() + geom_smooth() +  
  labs(title="Relationship between acceptance rate and book cost",  
        x="acceptance rate", y="book cost") +  
  theme(plot.title = element_text(hjust = 0.5))
```



## 2. Auto Data Analysis

```
library(ggplot2)
library (PerformanceAnalytics)

##### Read the data into R.
Auto = read.csv("Auto.csv", header = TRUE, na.strings = "?")
Auto = na.omit(Auto)
dim(Auto)

## [1] 392  9

##### a. Specify which of the predictors are quantitative, and which are qualitative?
# Quantitative: mpg cylinders displacement horsepower weight acceleration year.
QuantitativePredictors = c("mpg", "cylinders", "displacement", "horsepower",
                           "weight", "acceleration", "year")
# Translate origin into factor
Auto$originFactor = factor(Auto$origin, labels = c("USA", "Germany", "Japan"))
table(Auto$originFactor, Auto$origin)

##
##           1    2    3
##   USA      245    0    0
##   Germany    0   68    0
##   Japan      0    0   79

# Qualitative: name origin originFactor
QualitativePredictors = c("name", "origin", "originFactor")

##### b. What is the range, mean and standard deviation of each quantitative predictor?
Quantitatives = which(names(Auto) %in% QuantitativePredictors)
sapply(Auto[, Quantitatives], range)

##           mpg cylinders displacement horsepower weight acceleration year
## [1,]   9.0           3           68           46      1613           8.0    70
## [2,]  46.6           8          455          230     5140          24.8    82

sapply(Auto[, Quantitatives], mean)

##           mpg      cylinders displacement      horsepower      weight acceleration
##  23.445918    5.471939   194.411990    104.469388   2977.584184    15.541327
##           year
##  75.979592

sapply(Auto[, Quantitatives], sd)

##           mpg      cylinders displacement      horsepower      weight acceleration
##   7.805007    1.705783   104.644004    38.491160    849.402560    2.758864
##           year
##   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?
sapply(Auto[-seq(40, 80), Quantitatives], range)

##           mpg cylinders displacement horsepower weight acceleration year
## [1,]   9.0           3           68           46      1649           8.0    70
```

```
## [2,] 46.6      8      455      230  4997      24.8  82
```

```
sapply(Auto[-seq(40, 80), Quantitatives], mean)
```

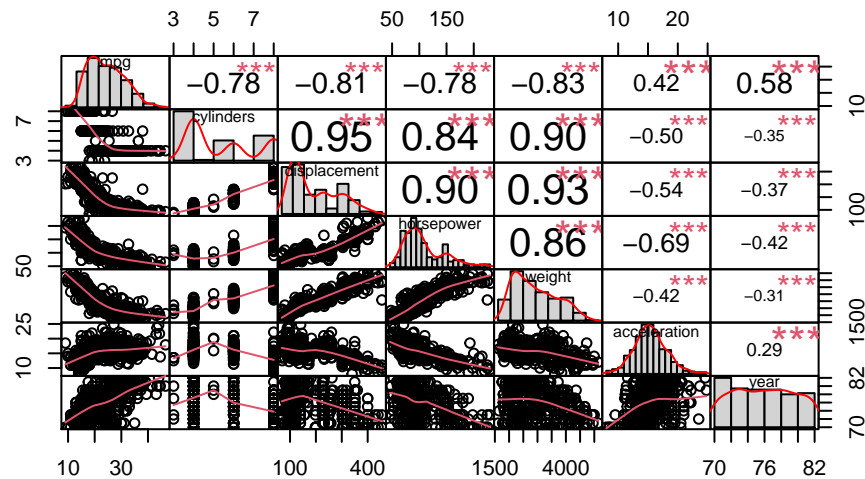
```
##      mpg      cylinders displacement  horsepower      weight acceleration
## 23.931054  5.424501  190.943020  103.019943 2948.934473  15.581766
##      year
## 76.492877
```

```
sapply(Auto[-seq(40, 80), Quantitatives], sd)
```

```
##      mpg      cylinders displacement  horsepower      weight acceleration
##  7.826817  1.667975  101.726508  37.711797  815.903085  2.730831
##      year
##  3.550345
```

##### 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.

```
chart.Correlation(Auto[, Quantitatives], histogram=TRUE, pch=19)
```



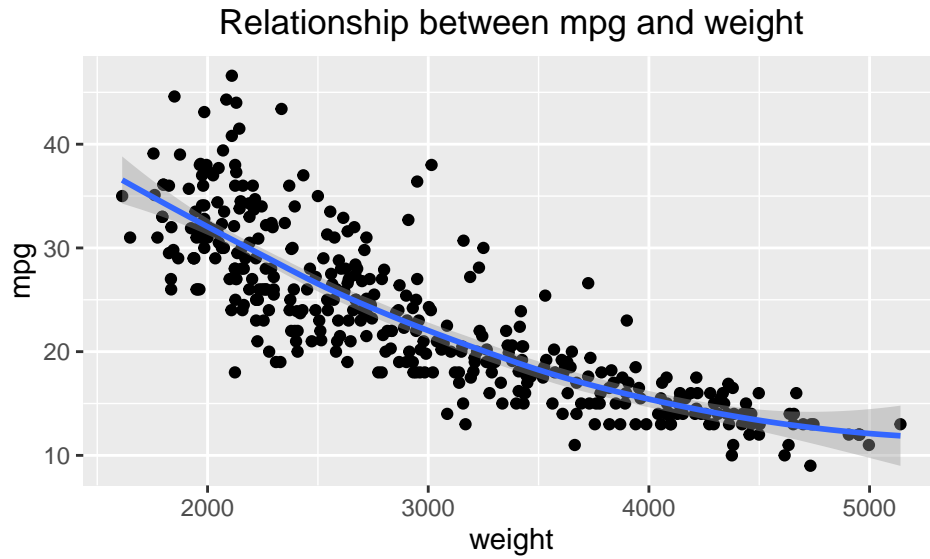
```
Matrix = cor(Auto[, Quantitatives])
round(Matrix, 2)
```

```
##      mpg cylinders displacement horsepower weight acceleration year
## mpg      1.00      -0.78      -0.81      -0.78      -0.83      0.42  0.58
## cylinders -0.78      1.00      0.95      0.84      0.90      -0.50 -0.35
## displacement -0.81      0.95      1.00      0.90      0.93      -0.54 -0.37
## horsepower  -0.78      0.84      0.90      1.00      0.86      -0.69 -0.42
## weight      -0.83      0.90      0.93      0.86      1.00      -0.42 -0.31
## acceleration 0.42     -0.50     -0.54     -0.69     -0.42      1.00  0.29
## year        0.58     -0.35     -0.37     -0.42     -0.31      0.29  1.00
```

##### 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.

#From the correlation graph, we can find that all of the quantitative variables show some

```
# correlation with mpg. for example, as weight goes up, mpg goes down.
ggplot(data=Auto, aes(x=weight, y=mpg)) + geom_point() + geom_smooth() +
  labs(title="Relationship between mpg and weight",
       x="weight", y="mpg") +
  theme(plot.title = element_text(hjust = 0.5))
```



```
# and better acceleration with higher mpg
ggplot(data=Auto, aes(x=acceleration, y=mpg)) + geom_point() + geom_smooth() +
  labs(title="Relationship between mpg and acceleration",
       x="acceleration", y="mpg") +
  theme(plot.title = element_text(hjust = 0.5))
```

