Week 2 R Code

The data for these courses can be found here: Important Github for Specialization Data Examples

```
library("ggplot2")
library("RCurl")
library("purrr")
library("tidyr")
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:RCurl':
##
##     complete
```

The following daataset contains measurements related to the impact of three advertising medias on sales of a product, P. The variables are:

- `Youtube`: the advertising budget allocated to YouTube. Measures in thousands of dollars
- `facebook`: the advertising budget allocated to Facebook. Measures in thousands of dollars
  `-newspaper`: the advertising budget allocated to a local newspaper. Measures in thousands of dollars
- `sales`: the values in the i^{th} row of the sales columns is a measurement of the sales (in thousands of units) for product P for company i.

```
url = getURL(paste0("https://raw.githubusercontent.com/bzaharatos/-Statistical",
                    "-Modeling-for-Data-Science-Applications/master/Modern%20R",
                    "egression%20Analy", "sis%20/Datasets/marketing.txt"))

marketing = read.csv(text=url, sep= "")
data(marketing)
```

```
## Warning in data(marketing): data set 'marketing' not found
```

```
head(marketing)
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

Exploratory Data Analysis

Before we model the data, let's first exlpore it. We'll check for any missing values. the look at univariate and bivariate summaries of the data

Missing Data and Univariate Explorations

Are there any missing values coded as `NA`? Or, are there any odd values for variables, e.g. `9999` or `0` possibly standing in for a missing value?

```
dim(marketing)
```

```
## [1] 200   4
```

```
cat("There are", sum(is.na(marketing)), "missing data values.\n")
```

```
## There are 0 missing data values.
```

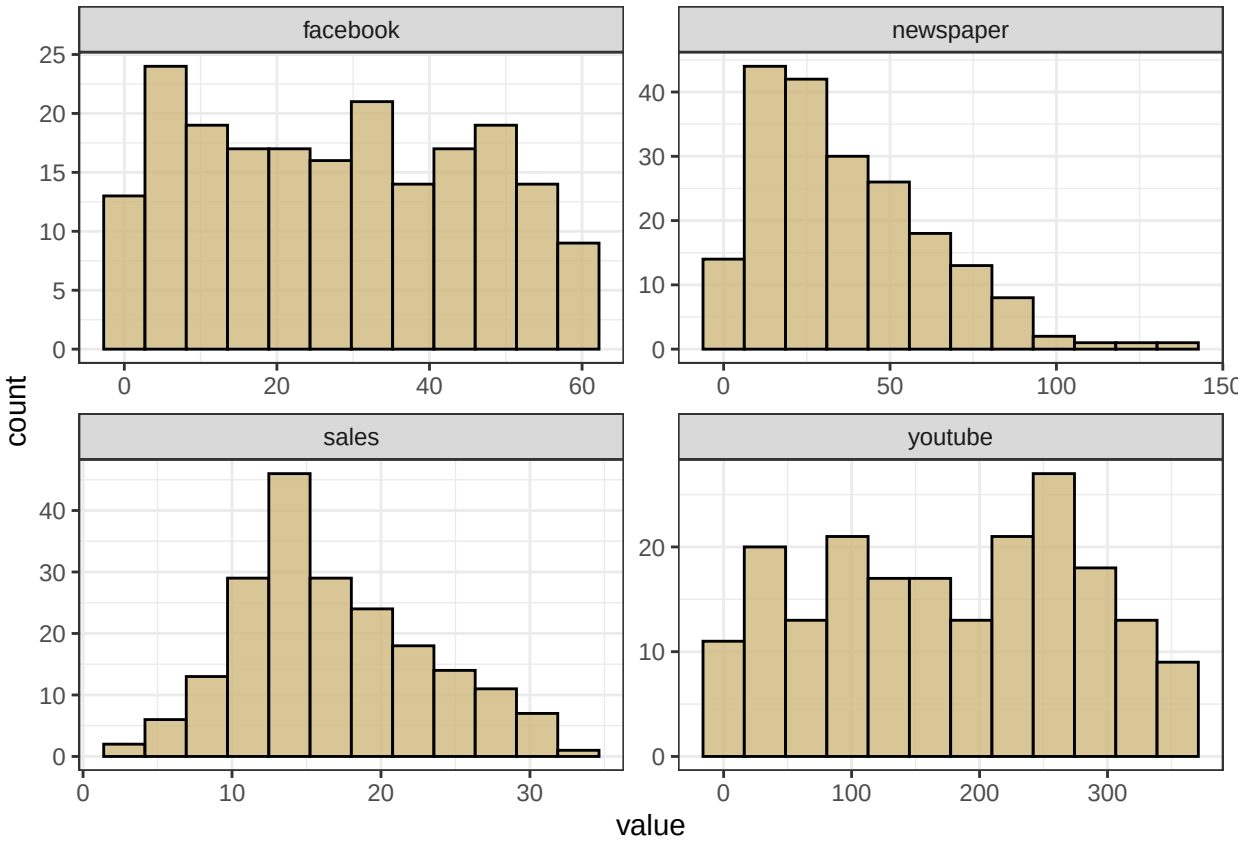
```
summary(marketing)
```

```
##     youtube          facebook       newspaper          sales      
##  Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92  
##  1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45  
##  Median :179.70   Median :27.48   Median : 30.90   Median :15.48  
##  Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83  
##  3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88  
##  Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40
```

As we can see, this dataset has 200 rows and 4 columns. There are 0 missing values. No alarming statistics shown in the summary.

The issue with this dataset is we can't split into test and training datasets.

```
marketing %>%
  keep(is.numeric) %>%
  gather %>%
  ggplot(aes(value)) +
    facet_wrap(~key, scales="free") +
    geom_histogram(bins=12, color="black", fill="#CFB87C", alpha=0.8) +
    theme_bw()
```



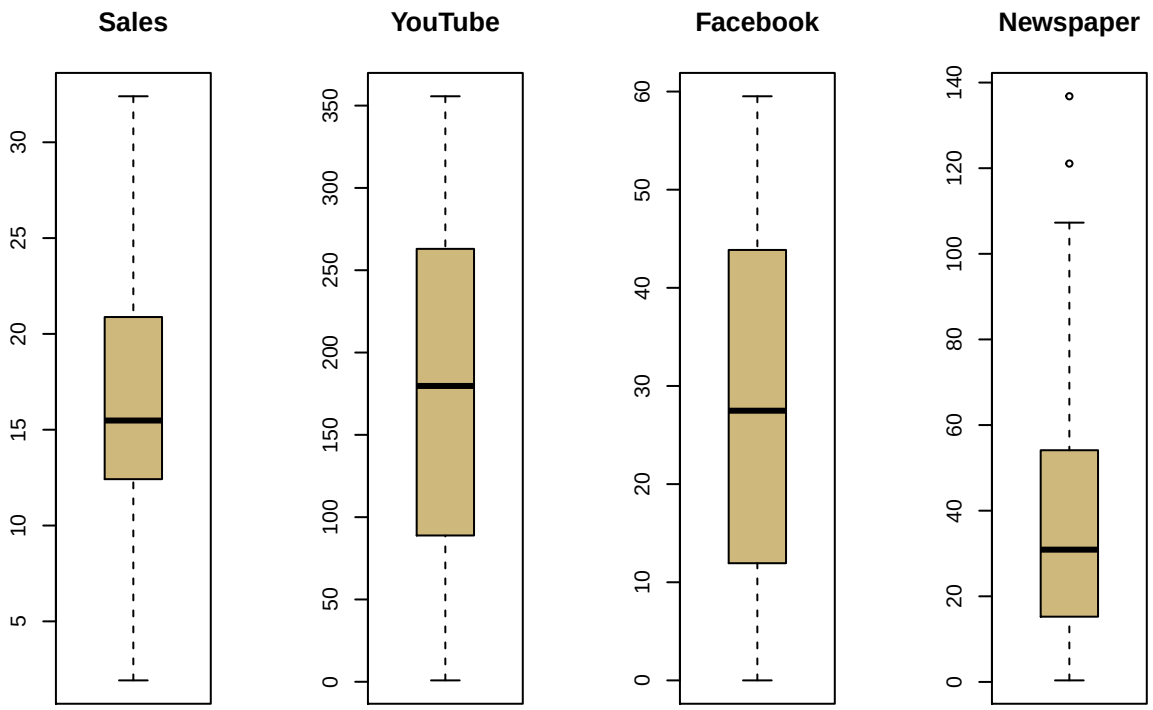
None of the predictors don't look normally distributed. However, our assumption doesn't require the predictors to be normally distributed. We also notice the newspaper variable potentially has some outliers.

`Sales` sort of has a bell shaped curve. The response variable should be normal, according to our assumption. However, that is a lower-tiered assumption on the list. Our Least squares regression doesn't require normality; however, our inferences do.

As we said earlier, `newspaper` potentiall has some outliers. Let's look at some boxplots to see in further detail.
R classifies potential outliers by the "IQR criterion". This means observations above $q_{0.75}$ **+ 1.5 X IQR** or below $q_{0.25}$ **- 1.5 X IQR** are classified as outliers where
- $q_{0.25}$ is the first quartile
- $q_{0.75}$ is the third quartile
- IQR is the interquartile range, defined as the difference between the third and first quartile.

A boxplot will flag the outliers:

```
par(mfrow = c(1,4))
boxplot(marketing$sales, main="Sales",col = "#CFB87C")
boxplot(marketing$youtube, main="YouTube",col = "#CFB87C")
boxplot(marketing$facebook, main="Facebook",col = "#CFB87C")
boxplot(marketing$newspaper, main="Newspaper",col = "#CFB87C")
```




```
cat("The outliers for the Newspaper variable are ",
    boxplot.stats(marketing$newspaper)$out[1], " and ",
    boxplot.stats(marketing$newspaper)$out[2])
```

```
## The outliers for the Newspaper variable are  136.8  and  121.08
```

These two outliers can affect the regression.

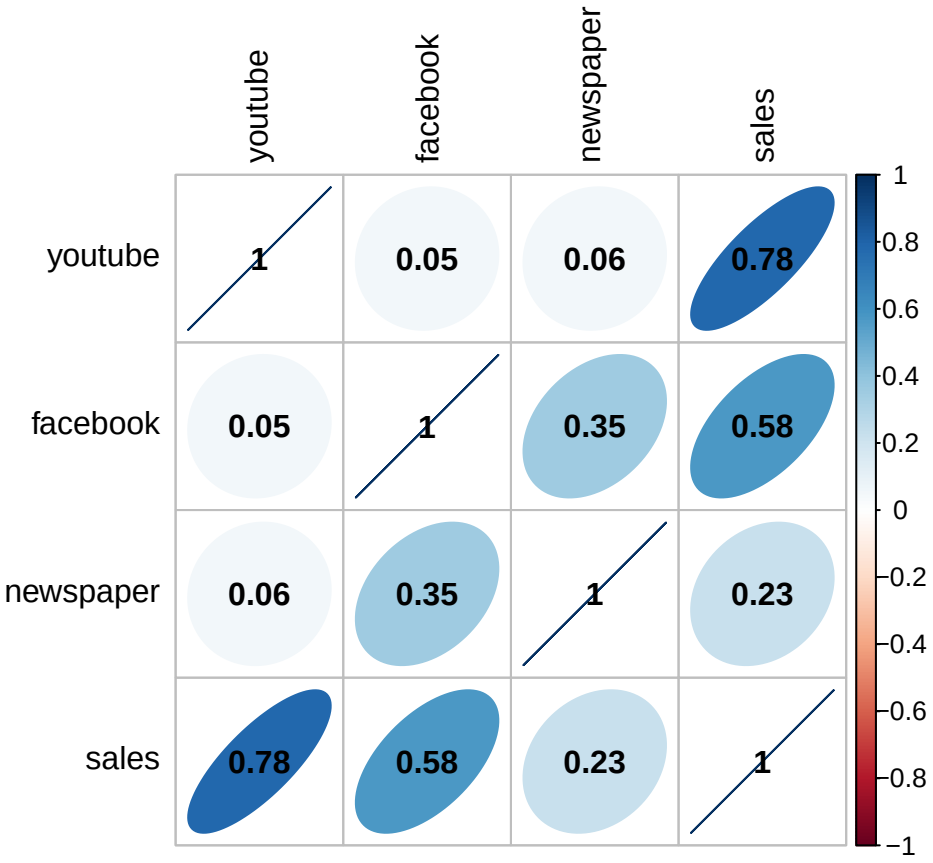
Bivariate Explorations

Let's now explore how the variables may or may not relate to each other. First, calculate the correlations between variables. Correlations can help us measure the strength of the liner relationship between variables. We'll do this with the `corrplot()` function.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(marketing), method="ellipse", addCoef.col="black", tl.col="black")
```



We can see sales high high correlation with YouTube, a bit with Facebook and almost none with newspaper. This shows the measure of the strength of the linear relationship.

```
#pairs(marketing, main="Marketing Data", pch=21, bg=c("#CFB87C"))
#Will need to double click to see actual relationship better
# Image saved as img in file. R markdown not allowing pair plot.
```

Looking at plot we can see the relationship between the different variables.

Linear Regression Modeling

Sums of squares and R^2 for simple linear regression

First, let's fit the entire dataset. In addition to running a summary of the model, (using `summary`), we'll also run an "analysis of variance", using the `anova()` function. The analysis of variance decomposes the total variability (TSS) into the explained variabiliy (ESS), and the residual/unexplained variability (RSS). It also produces an "F-test that we'll learn how to interpret in the next module.

```
lm_marketing = lm(sales ~ facebook, data = marketing)
summary(lm_marketing)
```

```
##
## Call:
## lm(formula = sales ~ facebook, data = marketing)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8766  -2.5589   0.9248   3.3330   9.8173
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.17397    0.67548  16.542   <2e-16 ***
## facebook     0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.13 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
anova(lm_marketing)
```

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## facebook    1 2590.1 2590.08  98.422 < 2.2e-16 ***
## Residuals 198 5210.6   26.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output we see that:
- ESS = 2590.1 - RSS = 5210.6 - TSS = 2590.1 + 5210.6 = 7800.7

** Rest of code in own notes**