CIND 123 - Data Analytics: Basic Methods

Assignment 3 (10%)

[Daria Yip]

[CIND 123 - DHD, 500721106]

Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

Sample Question and Solution

Use `seq()` to create the vector $(2, 4, 6, \ldots, 20)$.

```
#Insert your code here.
seq(2,20,by = 2)
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20
```

Question 1

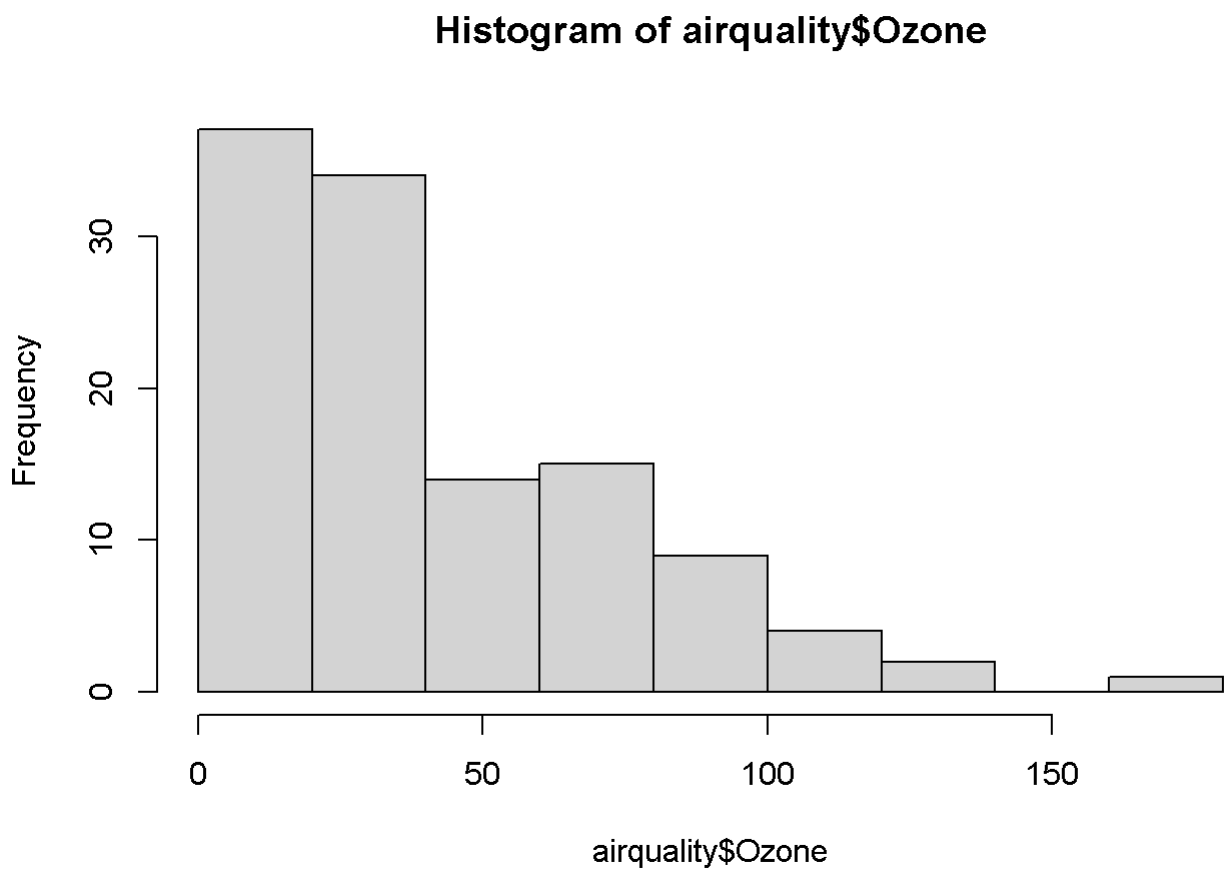
Use the following commands to install the `airquality` dataset and load the `datasets` package into your session.

```
#install.packages("datasets")
library(datasets)
data(airquality)
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

a. Use a histogram to assess the normality of the `Ozone` variable, then explain why it does not appear normally distributed.

```
hist(airquality$Ozone)
```

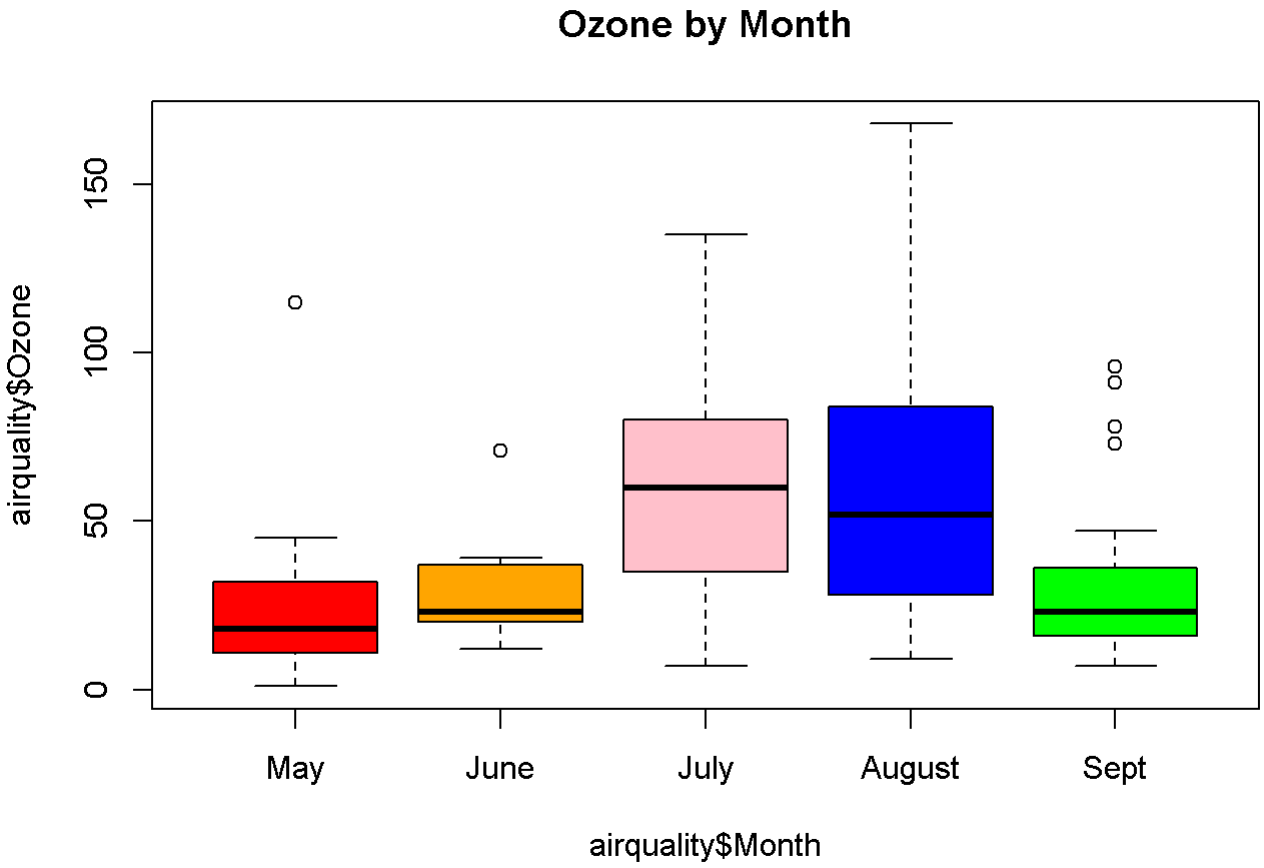



```
#It does not appear normally distributed as it is skewed to the right side. When we think of
#normally distributed, we picture a peak in the centre and the left and right side as
#symmetrical. However, it is clear that the peak is on the left hand side. Based off the
#summary of the data, this peak is due to the 37 NAs and the rest of the data is "imbalanced"
#in its representation of the frequency of certain Ozones.
summary(airquality$Ozone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   18.00   31.50   42.13   63.25  168.00      37
```

b. Create a set of boxplots that shows the distribution of `Ozone` in each month. Use different colors for each month.

```
boxplot(airquality$Ozone ~ airquality$Month, data = airquality, main = "Ozone by Month", names = c
("May", "June", "July", "August", "Sept"),
                                          col = c("red", "orange", "pink", "blue", "green"), n
otch = FALSE)
```



##Question 2

Use the following commands to install the `marketing` dataset and load the `datarium` package into your session.

```
#install.packages("datarium")
library(datarium)
data("marketing", package = "datarium")
str(marketing)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ youtube  : num  276.1 53.4 20.6 181.8 217 ...
##  $ facebook : num  45.4 47.2 55.1 49.6 13 ...
##  $ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

a. Find the covariance between the `Sales` and the advertising budget of `newspaper`. Comment on the output, in terms of the strength and direction of the relationship.

```
cov(marketing$newspaper, marketing$sales) #Covariance
```

```
## [1] 37.3556
```

```
cor(marketing$newspaper, marketing$sales) #Correlation
```

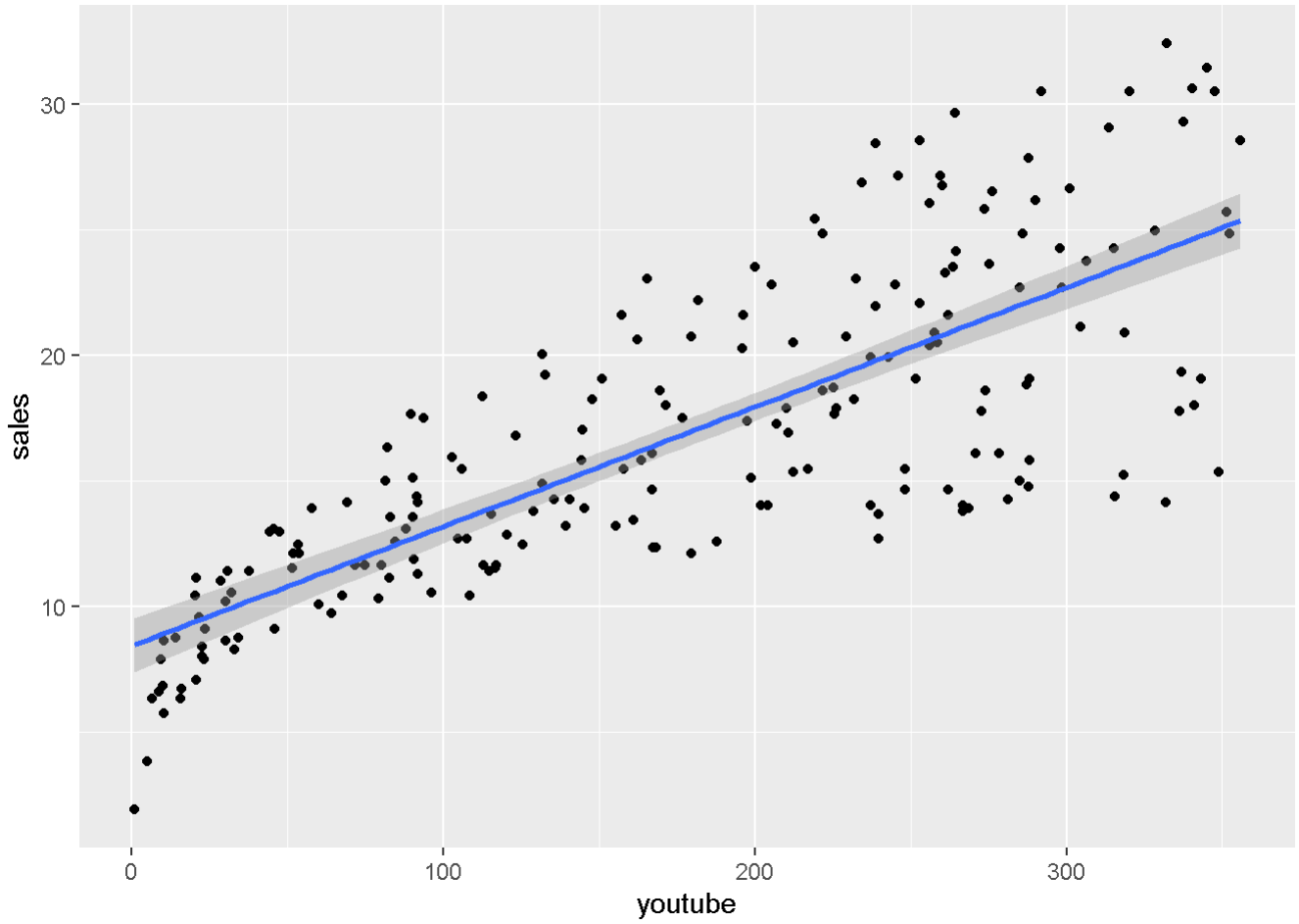
```
## [1] 0.228299
```

```
#The output we got was 37.3556. Covariance only indicates the direction for the relationship
#not the strength of the relationship. This is a positive relationship which means that
#the two variables 'sales' and 'newspaper' move in the same direction. If one tends to
#increase, the other will also tend to increase. If we want to know the strength, we can
#look at the correlation which is 0.228299. This means that the strength of the relationship
#is not that strong as it is closer to 0 than 1. It is a weak positive linear relationship.
```

b. Plot the `Sales` as a function of the `Youtube` variable using a scatterplot, then graph the least-square line on the same plot. Hint: You may use the `ggplot()` function from `ggplot2` package.

```
#install.packages('ggplot2')
library("ggplot2")
ggplot(marketing, aes(youtube, sales))+ geom_point(aes(shape = NULL)) +
  geom_smooth(method='lm') #scatterplot with regression line
```

```
## `geom_smooth()` using formula 'y ~ x'
```



c. Use the regression line to predict the `Sales` amount when `newspaper` budget is `$136.80K` . Comment on the difference between the output and the expected value.

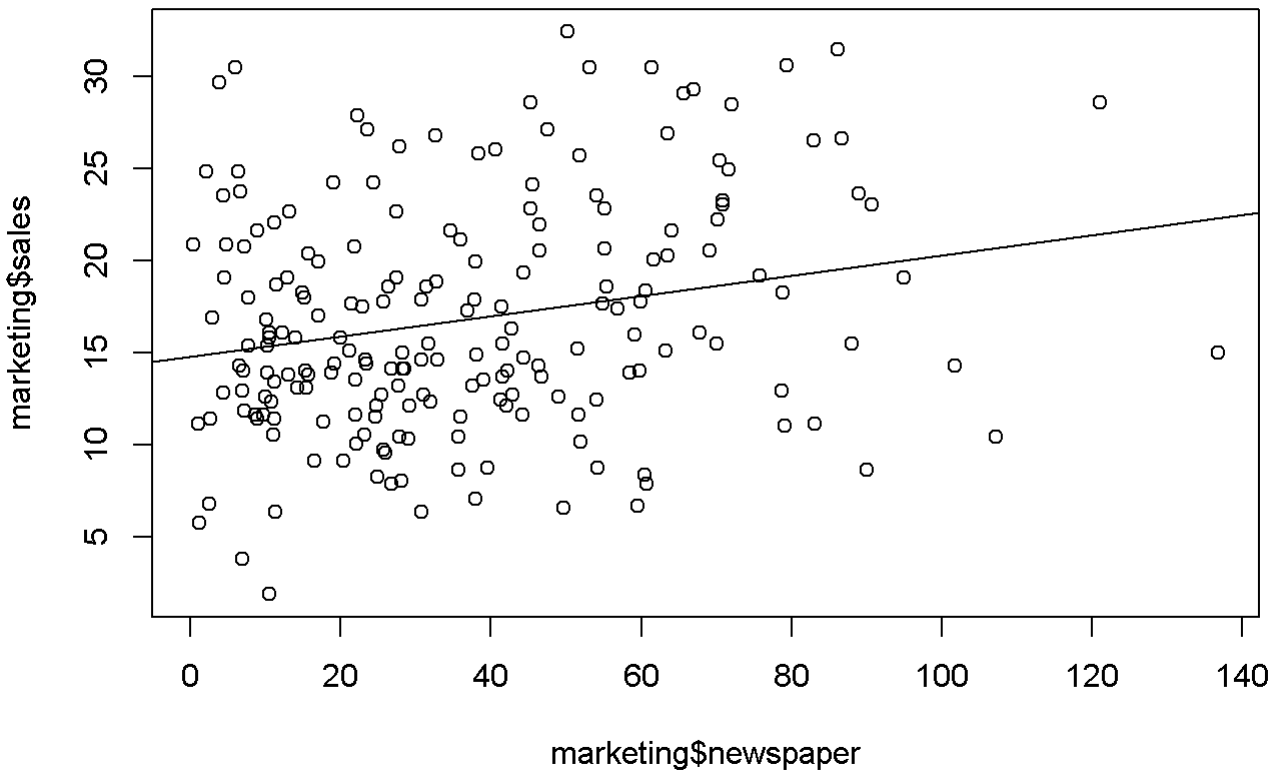
```
newmodel <- lm(sales~newspaper, data = marketing) #regression line
newspaperValue <- data.frame("newspaper" = 136.8) #dataframe creation
prediction <- predict(newmodel, newspaperValue) #predict function
prediction - marketing$sales[which(marketing$newspaper== '136.8')]
```

```
##        1
## 7.303704
```

```
abs(prediction - marketing$sales[which(marketing$newspaper== '136.8')]) #abs version
```

```
##        1
## 7.303704
```

```
plot(marketing$newspaper,marketing$sales) #graph of the relationship to see it visually
abline(newmodel) #regression line visualization
```




```
#The difference between the output and expected value is 7.303704. This means that the
#prediction was higher than the actual value of 15. If we look at the dataset in a graph,
#we can see that there is no clear trend. If we want to see the difference without the sign #then
 we can add the abs() function as shown above. Thus, we can conclude that the line of
#best fit may not be completely accurate and we should leave some room for error. By also
#looking at the graph, we can see that 136.80 is an outlier data point so that may also
#explain the difference between the expected and the actual. As the variance in the data is #larg
e, this line of best fit may not be completely accurate.
```

d. Use `newspaper` and `facebook` variables to build a linear regression model to predict `sales`. Display a summary of your model indicating Residuals, Coefficients, …, etc. What conclusion can you draw from this summary?

```
dmodel <- lm(sales~newspaper+facebook, data = marketing)
summary(dmodel)
```

```
##
## Call:
## lm(formula = sales ~ newspaper + facebook, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6347  -2.5739   0.8778   3.3188   9.5701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.026705   0.753206  14.640   <2e-16 ***
## newspaper    0.006644   0.014909   0.446    0.656
## facebook     0.199045   0.021870   9.101   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.14 on 197 degrees of freedom
## Multiple R-squared:  0.3327, Adjusted R-squared:  0.3259
## F-statistic: 49.11 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
#From this summary, we can see that the model predicts certain points that fall far away
#from the actual points (max is 9.5701 and min is -18.6347). The coefficients tell us the
#expected value of sales if newspaper and facebook were zero. The slope tells us that for
#every 1 increase in sales, the required newspaper/facebook goes up by 0.006644 or 0.199045
#respectively to reach that number. Generally, the standard errors are not to large
#relative to their coefficient counterparts except for newspaper which is bigger. This may
#signify a greater area for error in our estimates vs. actual. Since our t-values for
#facebook and (intercept) is generally large and is large relative to SE, this could
#indicate a relationship between facebook and sales. Newspaper and sales may be a smaller
#relationship, if any. We can also see the p-value for newspaper is over 5% which indicates
#that it has no significant relationship with sales. However, according to the p-value and
#the stars, the sales to facebook relationship is significant. The residual SE suggests that
#our average amount that our sales will deviate from the regression line is about 5.14.
#The Multiple R-Squared, as it is closer to 0 than 1, indicates a regression that may
#not be a good explanation of the variance in the sales variable. About 33.27% of variance
#found in sales can be explained by the other two parameters. Our adjusted R-Squared also
#mentions something similar. Finally, the F-Statistic distance from 1 is small considering
#the amount of data points. This suggests that it is hard to ascertain if there is a
#relationship between the predictor and response variables. Therefore, one can conclude that
#this regression model needs a bit of work before it can be used to accurately predict sales
#and we can consider that there may be a significant relationship between facebook and sales
#but not newspaper and sales.
```

e. Use the regression line to predict the `Sales` amount when `newspaper` budget is `$136.80K` and `facebook` is `$43.92K` .

```
newspaperandfb <- data.frame("newspaper" = 136.80, "facebook" = 43.92)
prediction2 <- predict(dmodel, newdata = newspaperandfb)
prediction2 #output is 20.67767
```

```
##        1
## 20.67767
```

```
#for comparison
marketing$sales[which(marketing$newspaper == '136.8' & marketing$facebook == 43.92)] #actual
```

```
## [1] 15
```

f. What is the difference between the output in (e) and the output in (c)

```
abs(prediction2 - prediction)
```

```
##        1
## 1.626038
```

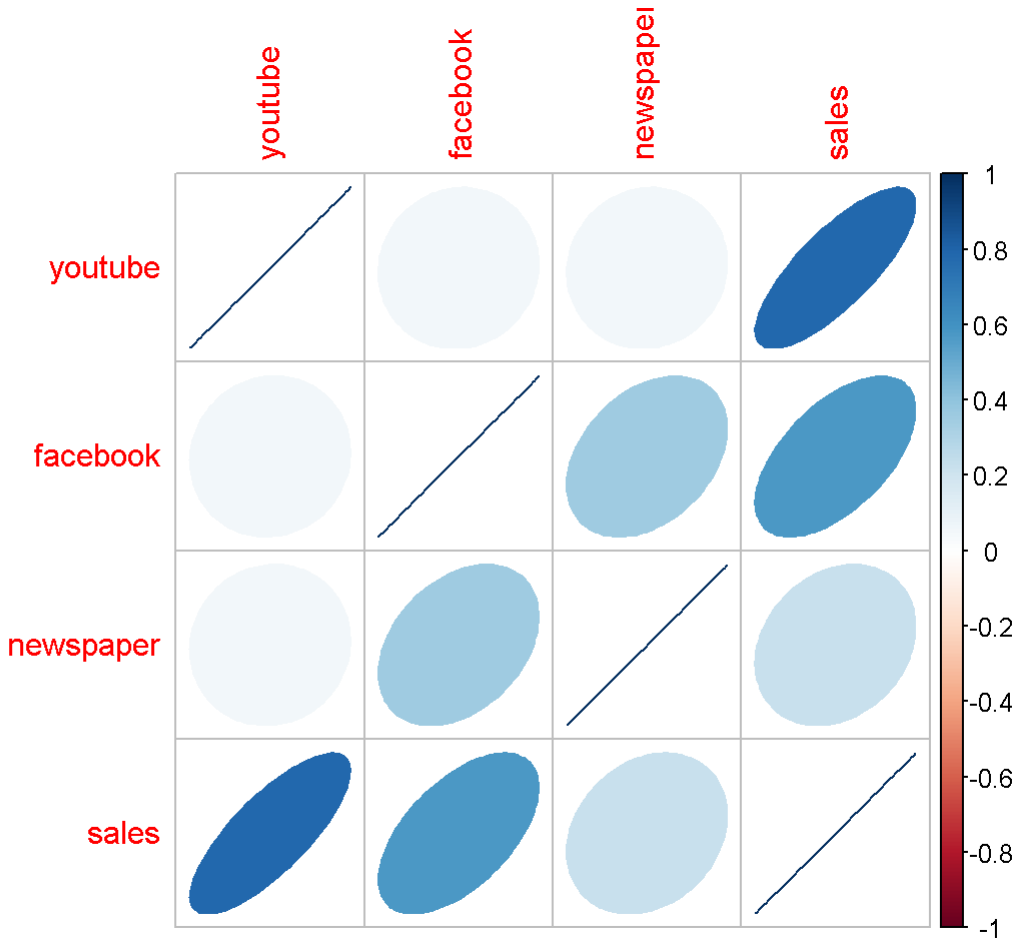
```
#The difference is 1.626038. (e) adds in facebook to the criteria of (c) and is a very small
#difference.
```

g. Display the correlation matrix of the variables: `youtube`, `facebook`, `newspaper` and `sales`. What conclusion can you draw?

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
table_cor_marketing <- cor(marketing)
corrplot(table_cor_marketing, method = "ellipse")
```




```
#The size of the circle is proportional to the correlation coefficient. The colour blue
#indicates a positive correlation and red signifies a negative correlation. Sales and
#youtube have the strongest correlation and positive correlation (as one goes up, the other
#also tends to go up). As correlation does not equal causation, we still need to investigate
#other components, but generally we can say that there is a strong linear relationship
#between sales and youtube. The other sigificant relationship to mention is sales and
#facebook, but it is not as strong as sales and youtube. The matrix displays only positive
#correlations and the insignificant relationships in grey.
```

h. In your opinion, which statistical test should be used to discuss the relationship between `youtube` and `sales` ? Hint: Review the differnce between Pearson and Spearman tests.

```
#In my opinion, the statistical test that should be used to discuss the relationship
#between youtube and sales should be the Pearson test as it is designated for numerical
#variables and Spearman is for categorical variables. As we progress throughout levels,
#one might consider that our dataset is small and generally going with Spearman is "safer"
#but in this case, we will use the rule of thumb and stick with Pearson as the statistical
#test of choice.
```

##Question 3

Install the `carData` dataset on your computer using the command `install.packages("carData")` . Then load the `CanPop: Canadian Population Data` into your session using the following command. The CanPop` has 16 rows and 2 columns and represent the decennial time-series of Canadian population between 1851 and 2001.

```
#install.packages("carData")
library("carData")
data("CanPop", package = "carData")
str(CanPop)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ year      : num  1851 1861 1871 1881 1891 ...
##  $ population: num  2.44 3.23 3.69 4.33 4.83 ...
```

a. Which of the two variables is the independent variable and which is the dependent variable? Explain your choice.

```
#The Year is the independent variable and the Population is the dependent variable. This
#is because we can think of independent as the one that we can change and dependent is the
#one that changes because we changed the independent. It tends to be that the population
#changes year to year rather than the year changes population to population.
```

b. Assuming that year and population are linearly related, give the equation and the graph of the least-squares regression line. Hint: use lm() function.

```
yearmodel <- lm(population ~ year, data = CanPop) #This gives us that the equation is with
#the intercept of -337.0986 and the slope of 0.1813. Therefore, based on the equation of
#y = a + bx, we have y = -337.0986 + 0.1813x OR population = -337.0986 + 0.1813(year).
plot(CanPop$year, CanPop$population, las = 1, type = "n")
abline(yearmodel)
```

c. Explain the meaning of the slope and y-intercept for the least-squares regression line in (b).

```
#From part b, the intercept iss -337.0986 and the slope is 0.1813. This means that at
#x = 0, the y value will be -337.0986. In this case, we can interpret it as in Year 0,
#the population was -337.0986. But of course, as that does not make sense, we can see the
#first issue of using a linear regression line to represent population. Next, the slope is a #posi
tive slope, representing that the population grows from year to year by 0.1813 units of measure. S
ince this is a least-squares regression line, it is the line of best fit based on #the data points
given and is not an actual representation of what happened but a general #"average" of what happen
ed and is predicted to happen based on past trends.
```

d. In year 2020, what would you predict the population's size to be. Does the value of the predicted size matches your expectations? Explain.

```
year2020 <- data.frame("year" = 2020)
predict(yearmodel, year2020)
```

```
##        1
## 29.19844
```

```
#The predicted size is 29.19844. Currently, there are approximately 8 billion people in the
#world. However, even if population is measured in millions, that would be much less than
#the amount of people that we have today. If it was measured in billions, that would be
#much more than what we have today. Generally, population is best measured with an
#exponential or logistical graph rather than a linear graph. Thus, the value of the
#predicted size does not match my expectations and is inaccurate.  .
```