Copy the Comments below into an R script. . Run your Script and copy your code solution as illustrate as directed below in item A-4. Upload you project in the M5 Project assignment in Module 6.

**#REFERENCES**

#Ref: https://www.statmethods.net/advgraphs/ggplot2.html

#Ref: https://www.datanovia.com/en/blog/ggplot-point-shapes-best-tips/

#Ref: http://environmentalcomputing.net/plotting-with-ggplot-colours-and-symbols/

#Ref: Andy Fields: Discovering Statistics Using R

#Ref: R Kabacoff: R in Action

**#DIRECTIONS**

#After each numbered item, Copy the comment and then your solution onto THIS Project Specification Sheet. #See Item A-4 as an example.

#A-1 LOAD DATA SET: <. Album Sales 2.dat > & display the first 6 records

**albumSalesData <- read.delim("G:/NEU/Coursework/2020 Q4 Fall/ALY 6010 PT & IS/Album Sales 2.dat")**

**head(albumSalesData, 6)**

**> head(albumSalesData, 6)**

**adverts sales airplay attract**

**1 10.256 330 43 10**

**2 985.685 120 28 7**

**3 1445.563 360 35 7**

**4 1188.193 270 33 7**

**5 574.513 220 44 5**

**6 568.954 170 19 5**

#A-2 EXPLORE the DATA SET: List the names of the 4 variables then display the dimensions of the dataset. Finally display the basics quartile statistics of the 4 variables,

**names(albumSalesData)**

**dim(albumSalesData)**

**summary(albumSalesData)**

**> names(albumSalesData)**

**[1] "adverts" "sales" "airplay" "attract"**

**> dim(albumSalesData)**

**[1] 200 4**

**> summary(albumSalesData)**

**adverts sales airplay attract**

**Min. : 9.104 Min. : 10.0 Min. : 0.00 Min. : 1.00**

**1st Qu.: 215.918 1st Qu.:137.5 1st Qu.:19.75 1st Qu.: 6.00**

**Median : 531.916 Median :200.0 Median :28.00 Median : 7.00**

**Mean : 614.412 Mean :193.2 Mean :27.50 Mean : 6.77**

**3rd Qu.: 911.226 3rd Qu.:250.0 3rd Qu.:36.00 3rd Qu.: 8.00**

**Max. :2271.860 Max. :360.0 Max. :63.00 Max. :10.00**

#A-3 CREATE the LINEAR REGRESSION MODEL of Sales vs Advertisements. Save the model as an R object named < albumSales.3ad >

**albumSales.3ad <- lm(sales~adverts, data = albumSalesData)**

**summary(albumSales.3ad)**

**> albumSales.3ad <- lm(sales~adverts, data = albumSalesData)**

**> summary(albumSales.3ad)**

**Call:**

**lm(formula = sales ~ adverts, data = albumSalesData)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-152.949 -43.796 -0.393 37.040 211.866**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.341e+02 7.537e+00 17.799 <2e-16 \*\*\***

**adverts 9.612e-02 9.632e-03 9.979 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 65.99 on 198 degrees of freedom**

**Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313**

**F-statistic: 99.59 on 1 and 198 DF, p-value: < 2.2e-16**

**#**A-4 What is the CORRELATION COEFFICIENT between Advertisements and Album sales. Save the value as an object named < r >

**r <- cor(albumSalesData$sales, albumSalesData$adverts)**

**r**

**> r**

**[1] 0.5784877**

#A-5 Display the 3 most basic DESCRIPTIVE STATISTICS for Sales and Advertisements

**descr(albumSalesData$sales)**

**descr(albumSalesData$adverts)**

**> descr(albumSalesData$sales)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**10.0 137.5 200.0 193.2 250.0 360.0**

**> descr(albumSalesData$adverts)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**9.104 215.918 531.916 614.412 911.226 2271.860**

#A-6 What is the COEFFICIENT OF DETERMINATION of Advertisements vs Sales

**coeffOfDetermination <- lm(sales ~ adverts, data=albumSalesData)**

**summary(coeffOfDetermination)$r.squared**

**> coeffOfDetermination <- lm(sales ~ adverts, data=albumSalesData)**

**> summary(coeffOfDetermination)$r.squared**

**[1] 0.3346481**

#A-7 Compute Total Sum of Squares from a mathematics basic formula

**anova(albumSales.3ad)**

**sum(anova(albumSales.3ad)[, 2])**

**> anova(albumSales.3ad)**

**Analysis of Variance Table**

**Response: sales**

**Df Sum Sq Mean Sq F value Pr(>F)**

**adverts 1 433688 433688 99.587 < 2.2e-16 \*\*\***

**Residuals 198 862264 4355**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> sum(anova(albumSales.3ad)[, 2])**

**[1] 1295952**



#A-8 Compute the Residual Sum of Squares from a mathematical formula

**sum(resid(albumSales.3ad)^2)**

**> sum(resid(albumSales.3ad)^2)**

**[1] 862264.2**



#A-9 Display the intercept and slope of the regression line of Sales vs Advertisements

**coef <- coef(coeffOfDetermination)**

**> coef <- coef(coeffOfDetermination)**

**(Intercept) sales**

**134.13993781 0.09612449**

#A10 Create the regression line equation (e.g., y = a + bx) and display the y coordinate at x = 500

**x <- 500**

**y <- coef[1] + coef[2]\*x**

**y**

**> x <- 500**

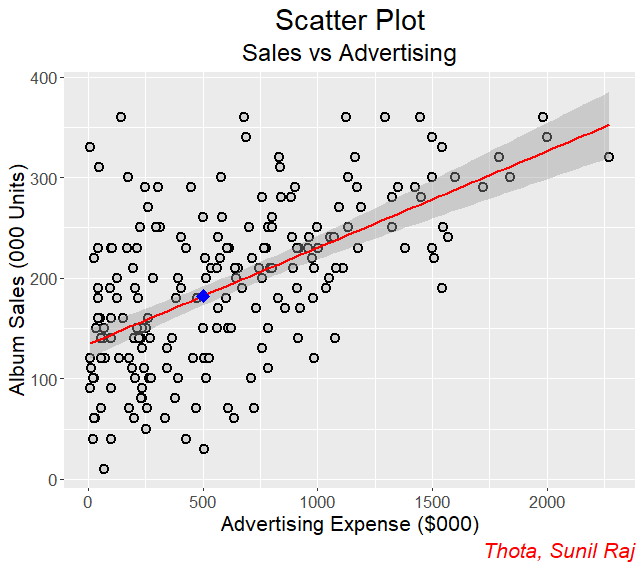
**> y <- coef[1] + coef[2]\*x**

**> y**

**(Intercept)**

**1682.507**

#A-11 Create a plot identical to that which is shown: REPLACE THIS IMAGE WITH YOUR PLOT ANNOTATED WITH **YOUR NAME** AND CHANGE THE REGRESSION LINE TO **RED** AND THE ORANGE TRIANGLE TO **BLUE** Replace the Professor's last name with your own.



**ggplot(albumSalesData,**

**aes(x = adverts, y = sales)) +**

**labs(**

**title = "Scatter Plot",**

**subtitle = "Sales vs Advertising",**

**x = "Advertising Expense ($000)",**

**y = "Album Sales (000 Units)",**

**caption = "Thota, Sunil Raj"**

**) +**

**theme(**

**plot.title = element\_text(hjust = 0.5, size = 22),**

**plot.subtitle = element\_text(hjust = 0.5, size = 18),**

**plot.caption = element\_text(color = "red", face = "italic", size = 16),**

**text = element\_text(size = 16)**

**) +**

**geom\_point(**

**shape = 21,**

**fill = "lightgray",**

**color = "black",**

**size = 2.74,**

**stroke = 1.1**

**) +**

**geom\_smooth(**

**method = lm,**

**color = "red",**

**fullrange = TRUE,**

**formula = y ~ x**

**) +**

**geom\_point(**

**aes(x = x, y = y),**

**size = 3,**

**fill = "blue",**

**shape = 23,**

**col = "blue",**

**stroke = 1**

**)**