## BAN 502

## Mandy Hawkins

## Module 2 Assignment 1: Correlation and Simple Linear Regression Assignment

# 

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts ----------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

## Task 1

#a. Describe this dataset. 6 columns: Ozone, Solar.R, Wind, Temp, Month, Day  
#b. How many variables and observations are there? 153 obs of 6 variables  
#c. Is there any missing data? Yes, 37 missing rows in Ozone and 7 missing rows in Solar.R  
#d. Which variable is likely to be the response (Y) variable? Ozone?  
  
# Read-in the airqualtiy data set (a default R dataset) as a data frame called "air".  
air= airquality   
#Examine the structure and summary of the dataset  
str(air) #all variables numeric

## '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 ...

summary(air) #37 missing rows in Ozone and 7 missing rows in Solar.R

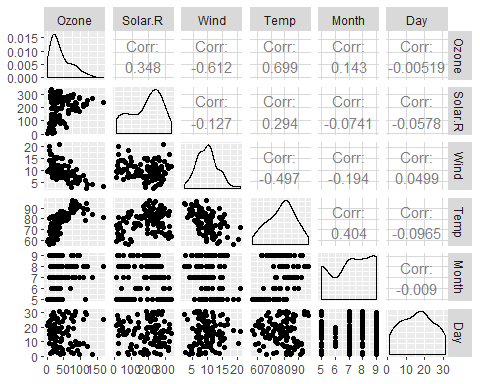
## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day   
## Min. :5.000 Min. : 1.0   
## 1st Qu.:6.000 1st Qu.: 8.0   
## Median :7.000 Median :16.0   
## Mean :6.993 Mean :15.8   
## 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :9.000 Max. :31.0   
##

## Task 2

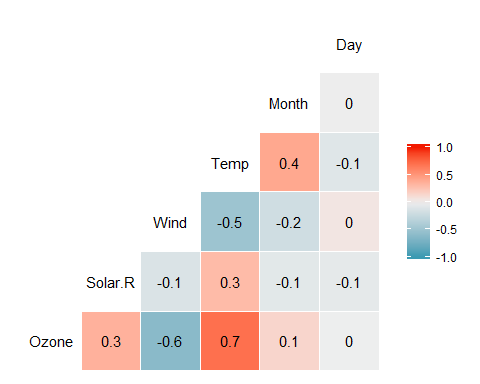
#Delete missing data  
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))  
  
#How many rows and columns remain in this new (air2) data frame? 111 columns and 6 rows

## Task 3

ggpairs(air2)#develop a visualization of and to calculate correlation for the combinations of variables in this dataset.



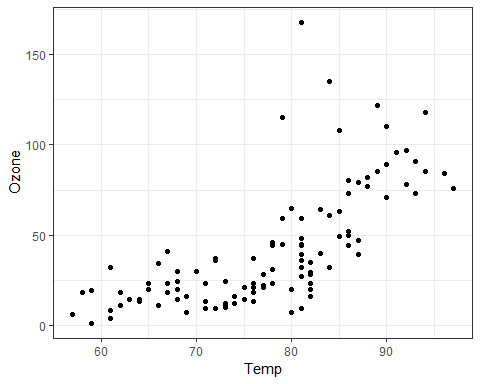
ggcorr(air2, label = TRUE) #"ggcorr" to develop a correlation matrix for the variables. "label = TRUE" in the ggcorr function to show the correlation values.



#a. Which variable is most strongly correlated with the "Ozone" variable? Temp  
#b. Which variable is least strongly correlated with the "Ozone" variable? Day

## Task 4

#Plot "Temp" (x axis) versus "Ozone" (y axis) using the "ggplot" function. Choose an appropriate chart type. Describe the relationship between "Temp" and "Ozone".   
ggplot(air2,aes(x=Temp, y=Ozone)) +  
 geom\_point() + theme\_bw()



#As temp increases, Ozone increases

## Task 5

#Create a linear regression model (called in model1) using "Temp" to predict "Ozone".   
#a. Discuss the quality of this model (mention the R square value and signi???cance of the predictor variable). The P value is significantly less than .05 so that is good meaning x is a predictor of y but R squared is .48 which is ok. Visually looking at the graph as the temp increases, ozone increases so yes it seems that x is a predictor of y.  
#b. Use the code "con???nt(model1)" to generate 95% con???dence intervals for the coe???cients. In what range does the slope coe???cient likely fall? 1.964787 and 2.913433  
  
model1=lm(Ozone~Temp, air2)#create linear regression model  
  
summary(model1)

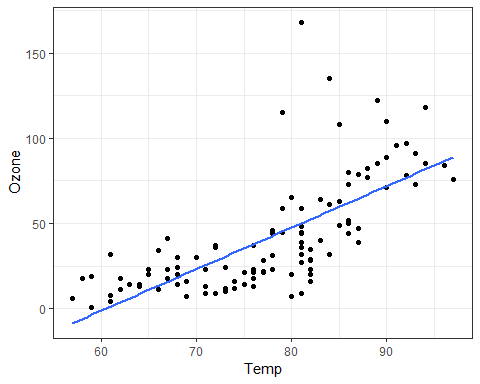
##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 10.192 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.92 on 109 degrees of freedom  
## Multiple R-squared: 0.488, Adjusted R-squared: 0.4833   
## F-statistic: 103.9 on 1 and 109 DF, p-value: < 2.2e-16

confint(model1,level=0.95) # to generate 95% con???dence intervals for the coe???cients.

## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

## Task 6

#Re-do Task 4 to include the regression line.  
ggplot(air2,aes(x=Temp, y=Ozone)) +  
 geom\_point() + geom\_smooth(method= "lm", se = FALSE) + theme\_bw()



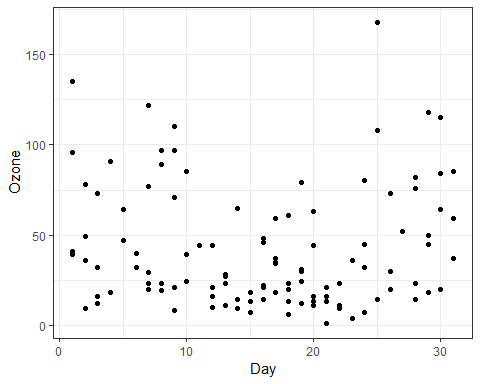
## Task 7

#Develop a prediction for "Ozone" when "Temp" is 80. Include the prediction interval for this prediction. -0.1510188 (lower) 95.11646 (upper)  
task7prediction = data.frame(Temp=c(80))  
predict(model1, newdata=task7prediction, interval = "predict")

## fit lwr upr  
## 1 47.48272 -0.1510188 95.11646

## Task 8

#Plot "Day" (x axis) versus "Ozone" (y axis) using the "ggplot" function. Choose an appropriate chart type. Describe the relationship between "Day" and "Ozone".  
ggplot(air2,aes(x=Day, y=Ozone)) +  
 geom\_point() + theme\_bw()



## Task 9

#Create a linear regression model (called in model2) using "Day" to predict "Ozone".   
#a. Discuss the quality of this model (mention the R square value and signi???cance of the predictor variable). The P value is greater than .05 so that is not good and the R squared is awful. Visually, it doesn't appear that Day is a predictor of ozone.  
#b. Use the code "con???nt(model2)" to generate 95% con???dence intervals for the coe???cients. In what range does the slope coe???cient likely fall? -0.745321 and 0.7056539  
  
model2=lm(Ozone~Day, air2)#create linear regression model  
  
summary(model2)

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.00 -24.23 -11.04 19.96 126.08   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.41536 6.64353 6.384 4.32e-09 \*\*\*  
## Day -0.01983 0.36604 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.43 on 109 degrees of freedom  
## Multiple R-squared: 2.693e-05, Adjusted R-squared: -0.009147   
## F-statistic: 0.002936 on 1 and 109 DF, p-value: 0.9569

confint(model2,level=0.95) # to generate 95% con???dence intervals for the coe???cients.

## 2.5 % 97.5 %  
## (Intercept) 29.248109 55.5826192  
## Day -0.745321 0.7056539

## Task 10

#Re-do Task 8 to include the regression line.  
ggplot(air2,aes(x=Day, y=Ozone)) +  
 geom\_point() + geom\_smooth(method= "lm", se = FALSE) + theme\_bw()

