Module 2 - Assignment 1

#install.packages("GGally")  
library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 2.0.1 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 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

library(knitr)

## Task 1

air = airquality

1. Describe the dataset. This dataset contains daily readings of the air quality in New York from May 1, 1973 to September 30, 1973. The values in the dataset are used to represent the variables of the ozone, solar.r, wind, temp, month and day.
2. There are 6 variables and 153 observations.
3. Yes, there is missing data in the Ozone and Solar.R columns.
4. The response (Y) variable is likely to be our “Ozone” attribute.

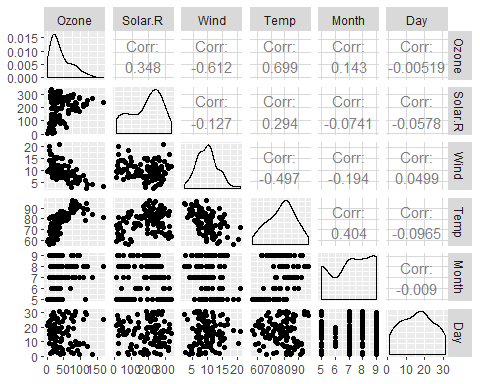
## Task 2

air2 = air %>% filter (!is.na(Ozone)) %>% filter (!is.na(Solar.R))

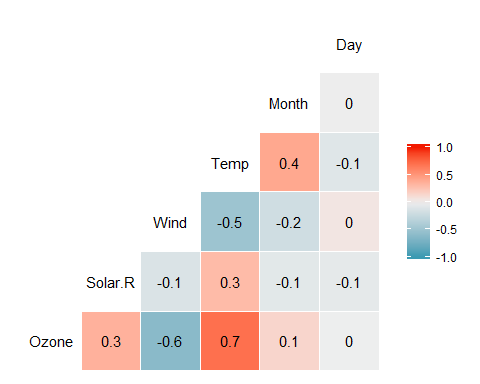
In the new air2 data frame there are 111 rows and 6 columns.

## Task 3

ggpairs(air2)



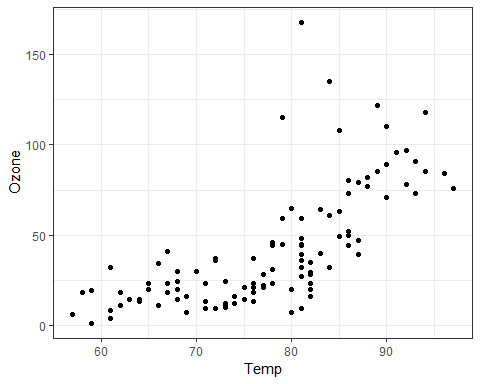
ggcorr(air2, label = TRUE)



1. Temperature is the most strongly correlated variable with the “Ozone” variable.
2. The Day variable is the least strongly correlated variable with the “Ozone” variable.

## Task 4

ggplot(air2, aes(x=Temp, y= Ozone))+geom\_point()+theme\_bw()



As the temperature increases we can see a corresponding increase of the Ozone variable. With a visulation test we can see that there is not a strong linear relationship.

## Task 5

model1 <- lm(Temp~Ozone,air2)  
summary(model1)

##   
## Call:  
## lm(formula = Temp ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.980 -4.775 1.825 4.228 12.425   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.37059 1.05151 65.97 <2e-16 \*\*\*  
## Ozone 0.20006 0.01963 10.19 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.851 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

1. Our p-value for this model is indeed less than 0.05 which tells us that our X variable is a significant predictor of our Y variable. Also, our R-squared value is 0.488 so we can assume that this a good quality model.

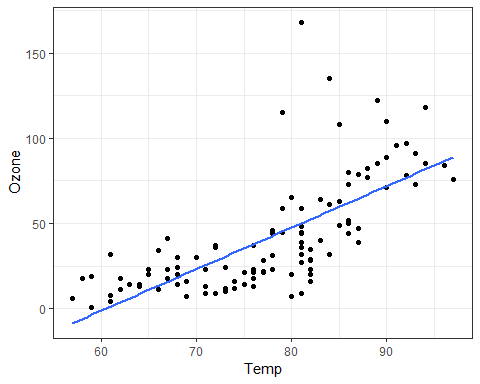
confint(model1, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 67.2865285 71.4546496  
## Ozone 0.1611525 0.2389608

The slope coefficient will fall in the range of 67.287-71.456.

## Task 6

ggplot(air2, aes(x=Temp, y= Ozone))+geom\_point()+theme\_bw()+geom\_smooth(method="lm", se=FALSE)



summary(model1)

##   
## Call:  
## lm(formula = Temp ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.980 -4.775 1.825 4.228 12.425   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.37059 1.05151 65.97 <2e-16 \*\*\*  
## Ozone 0.20006 0.01963 10.19 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.851 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

## Task 7

#manually  
69.37059 + 0.20006\*80

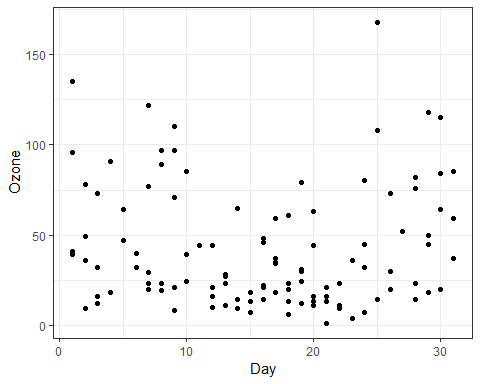
## [1] 85.37539

#Using predict function  
prediction <- data.frame(Ozone=c(80))  
predict(model1, newdata = prediction, interval = "predict")

## fit lwr upr  
## 1 85.37512 71.65702 99.09322

## Task 8

ggplot(air2, aes(x=Day, y= Ozone))+geom\_point()+theme\_bw()



From our graph we can see that the relationship between Day and Ozone is not linear. There is no significant relationship between the two variables.

## Task 9

model2 <- lm(Day~Ozone,air2)  
summary(model2)

##   
## Call:  
## lm(formula = Day ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.9502 -6.9834 0.0621 6.5152 15.1123   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.003116 1.342588 11.920 <2e-16 \*\*\*  
## Ozone -0.001358 0.025063 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.747 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

1. The quality of this model is very poor/ not good. The p-value is greater than 0.05 which tells us that the X is not a significant predictor of Y. Also, our R-squared value is very low which tells us that the variables are not significant and our model is not good.

confint(model2, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 13.34214999 18.66408196  
## Ozone -0.05103151 0.04831554

1. The slope coefficent will likely fall in the range of 13.34-18.66.

## Task 10

ggplot(air2, aes(x=Day, y=Ozone))+geom\_point()+theme\_bw()+geom\_smooth(method="lm", se=FALSE)

