# BAN503 - Module 2 Assignment 1

## Silas, Kimberly

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

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

## v ggplot2 3.1.0 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## v tidyr 0.8.3 v stringr 1.4.0   
## v readr 1.3.1 v forcats 0.4.0

## -- 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

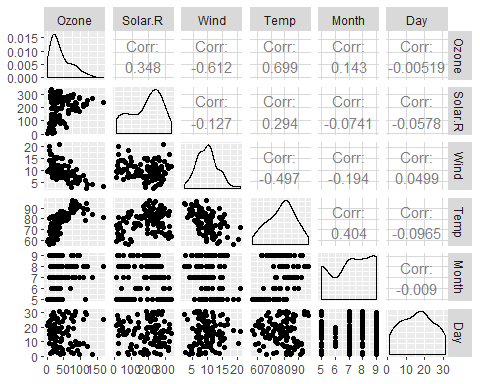
air = airquality

A.This dataset represents the airquality in New York in the 70’s that was collected by the NY state conservation and the National weather service. B.there are 6 variables and 153 observations C.there is missing data in the dataset D.the response variable will most likely be ozone

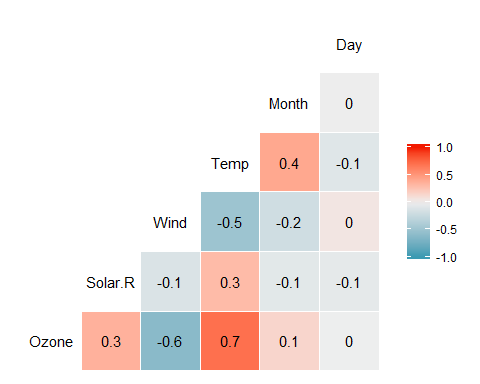
library(tidyverse)  
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))

there are 111 observations remaining in the 6 variables of the dataset

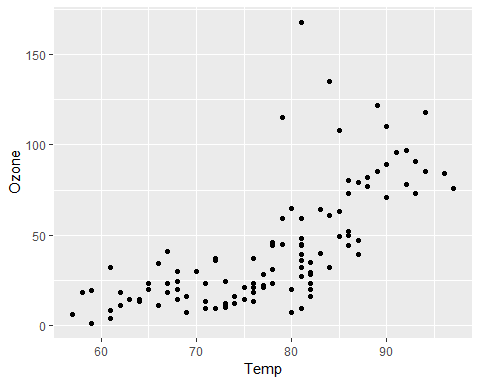
library(tidyverse)  
ggpairs(air2)



ggcorr(air2, label = TRUE)

 Temp has the strongest correlation to Ozone While day has no correlation and Month has the weekest correlation

library(tidyverse)  
ggplot(air2, aes(x=Temp,y=Ozone))+  
 geom\_point()

 As the temp increases the the ozone also increases in a positive correlation

library(tidyverse)  
model1 = lm(Ozone ~ Temp, air)   
summary(model1)

##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.729 -17.409 -0.587 11.306 118.271   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -146.9955 18.2872 -8.038 9.37e-13 \*\*\*  
## Temp 2.4287 0.2331 10.418 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.71 on 114 degrees of freedom  
## (37 observations deleted due to missingness)  
## Multiple R-squared: 0.4877, Adjusted R-squared: 0.4832   
## F-statistic: 108.5 on 1 and 114 DF, p-value: < 2.2e-16

A. The R squared is good at a .4877 and the pvalue is way below .05 so the model is a good model

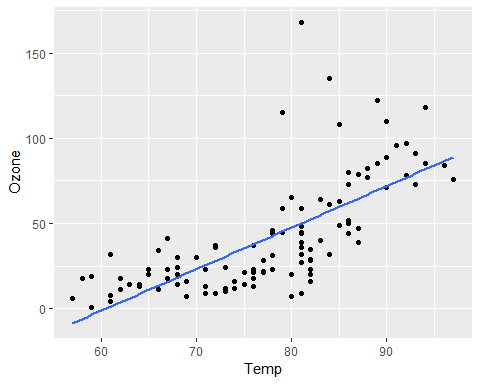
confint(model1, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) -183.222241 -110.768741  
## Temp 1.966871 2.890536

There are 3695 rows in this new data frame

The value of the slope is 2.8 and they intercept at -110.77

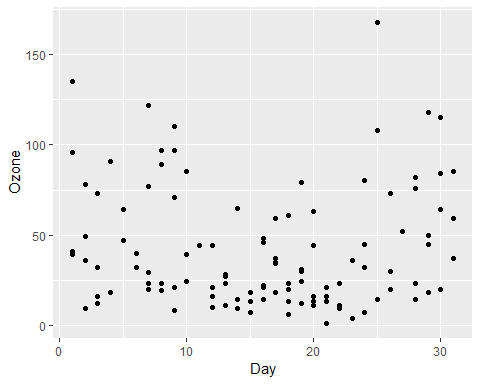
library(tidyverse)  
ggplot(air2, aes(x=Temp,y=Ozone))+  
 geom\_point()+  
 geom\_smooth(method="lm", se = FALSE)



testdata = data.frame(Temp = c(80))  
predict(model1, newdata = testdata, interval = "predict")

## fit lwr upr  
## 1 47.30077 0.110647 94.4909

library(tidyverse)  
ggplot(air2, aes(x=Day,y=Ozone))+  
 geom\_point()

 There appears to be no relationship between day and ozone

library(tidyverse)  
model2 = lm(Ozone ~ Day, air)   
summary(model2)

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.86 -24.29 -10.86 21.16 126.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.90387 6.28832 6.823 4.53e-10 \*\*\*  
## Day -0.04986 0.35306 -0.141 0.888   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.13 on 114 degrees of freedom  
## (37 observations deleted due to missingness)  
## Multiple R-squared: 0.0001749, Adjusted R-squared: -0.008595   
## F-statistic: 0.01994 on 1 and 114 DF, p-value: 0.8879

A. The R-squared is extremly low and the pvalue is high makeing the model not very good.

confint(model2, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 30.446753 55.3609902  
## Day -0.749275 0.6495535

The value of the slope is 0.64 and they intercept at 55.36

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

