## Correlation and Simple Linear Regression Assignment

#install.packages("GGally")

library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.3.0  
## 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

air = airquality

ncol(air) #Number of columns

## [1] 6

nrow(air) #Number of rows

## [1] 153

str(air)

## '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 ...

airblank <- mapply(anyNA, airquality)  
airblank

## Ozone Solar.R Wind Temp Month Day   
## TRUE TRUE FALSE FALSE FALSE FALSE

This dataset has 153 observations and 6 various variables. We can see that with the variables the date (month and day) has been capture. With that the temperature and wind speed have been captured. The other variables surround Ozone and Solar factors. There is missing data in the Ozone column and in the Solar.R column.

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

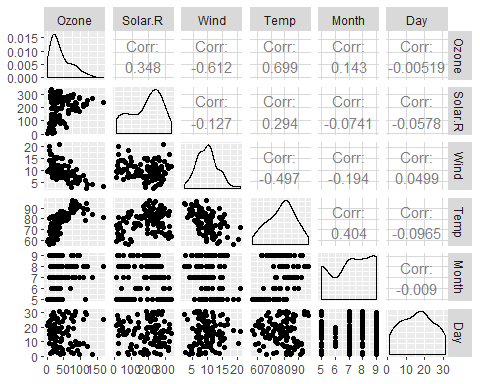
## [1] 6

nrow(air2)

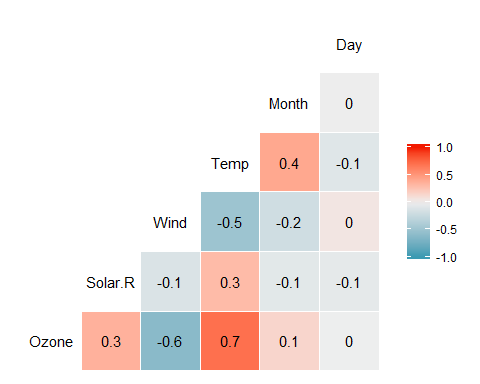
## [1] 111

There are now 111 observations with 6 variables in the new data frame.

ggpairs(air2)

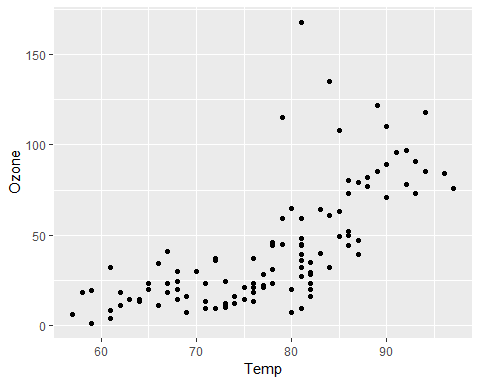


ggcorr(air2, label = TRUE)



The variable that is most strongly correlated with the “Ozone” variable is the “Temp”" variable. The variable that is least strongly correlated with the “Ozone” variable is the “Wind”" variable.

air2[1:6] <- lapply(air2[1:6], as.numeric)  
ggplot(air2, aes(x = Temp, y = Ozone)) + geom\_point()



In the scatterplot we can see that as the Temperature rises so does the Ozone variable.

model1 = lm(Ozone ~ Temp, air2)  
summary(model1)

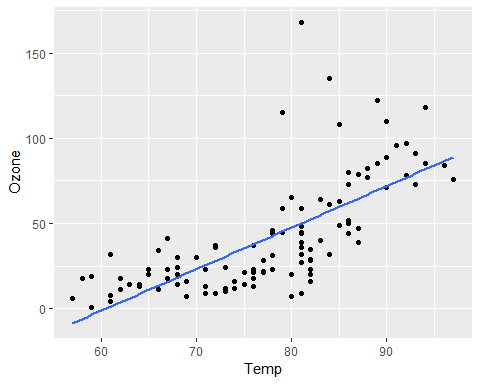
##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 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)

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

This is indeed a good model. We see that the p-value (< 2e-16) is significantly lower than .05 and a R square value this is relatively high at .488. We could see hight for the R square value but it will do the good. Through this we can see that Temp and Ozone are highly correlated.  
This slope coefficient falls between 1.96 and 2.91.

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



In the scatterplot, with the regression line, we can see that as the Temperature rises so does the Ozone variable.

-147.646 + 2.439\*80

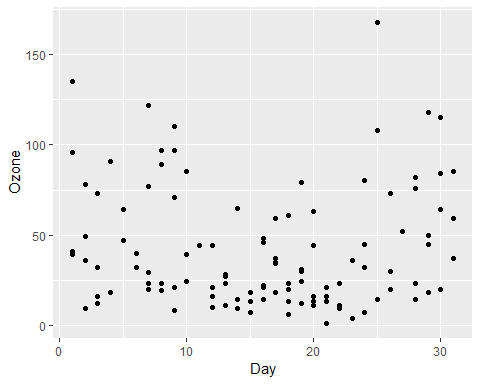
## [1] 47.474

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

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

The predicted Ozone value when the temp is 80 is 47.48

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



There seems to be no relationship or correlation between the day of the month and the Ozone variable.

model2 = lm(Ozone ~ Day, air2)  
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)

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

Seeing that the R square value is extremly low and the p-value is nearly 1 then I would say that this model is not a good model. The slope coefficient will likely fall between -.75 and .71.

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

