## Module 2 - Assignment 1

### Simple Linear Regression and Correlation

# 

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

library(ggcorrplot)

air = airquality  
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 ...

summary(air)

## 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   
##

a.) This data set reprsents the daily air quality in New York from May to September in 1973. b.) There are 153 observations and 6 variables. c.) Yes, there is missing data in this data set. d.) The response variable is likely to be Ozone.

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

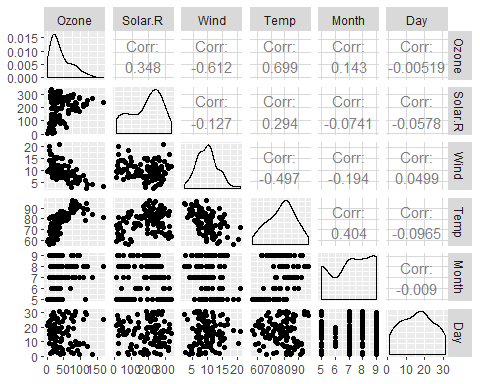
## [1] 111

ncol(air2)

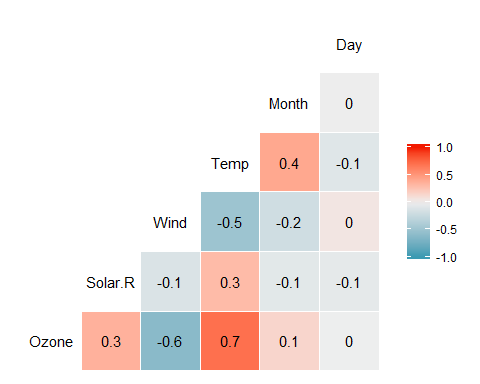
## [1] 6

There are 111 rows and 6 columns.

ggpairs(air2)

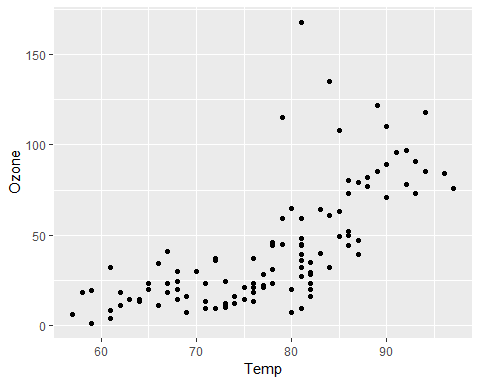


ggcorr(air2, label=TRUE)



a.) Temperature has the strongest correlation with the “Ozone” variable.  
b.) The variable “Day” has the weakest correlation with the “Ozone” variable.

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



The relationship between Temperature and Ozone is positive and fairly linear.

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

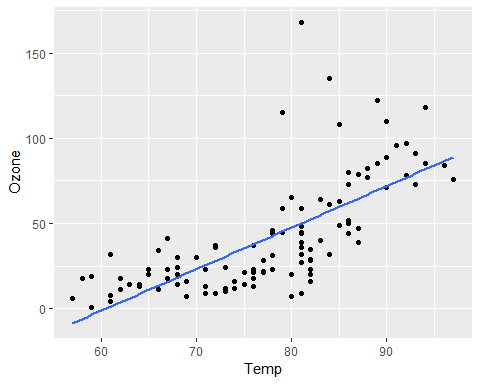
confint(model1)

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

a.) The R squared value for this model is pretty good at 0.488. The p-value being less than 0.05 determines that there is a significant relationship betwen the variables.

b.)The slope coefficient falls between 1.964787 and 2.913433.

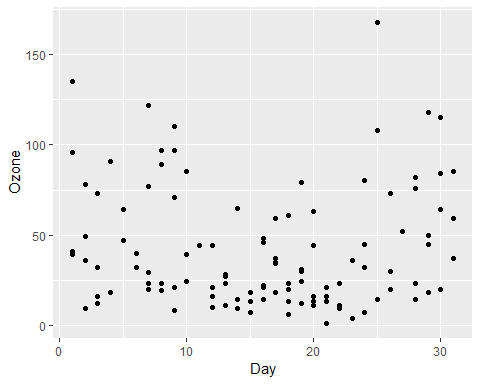
ggplot(air2, aes(x=Temp , y=Ozone))+  
 geom\_point()+  
 geom\_smooth(method="lm", se= FALSE)



Prediction = data.frame(Temp = 80)  
predict(model1, newdata=Prediction, intereval = "predict")

## 1   
## 47.48272

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



Based of the scatterplot, there is minimal relationship between “Day” and “Ozone”.

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

a.)The R squared value in model2 is very low meaning that there is high variance in the dataset around the regression line. The p value is higher that 0.05 meaning that there is very little relationship betweent these two variables.

b.)The slope coefficient falls between -0.745321 and 0.7056539.

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

