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

### Hackett, Evan

library(tidyverse) #Loads in Tidyverse

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- 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) #Loads in GGally

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

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

library(ggplot2)

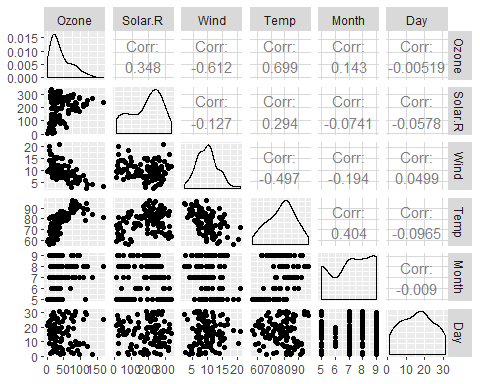
air = airquality #Loading in Dataset

This dataset is found within R and contains 6 variables with 153 observations. There is missing data within the dataset. The variable most likely to be the Y variable is the Ozone variable.

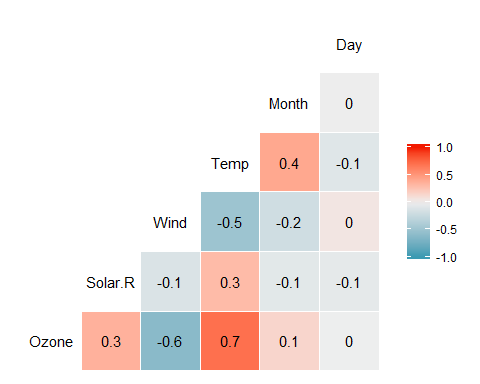
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))

After removing any rows with missing data we are left with the same six variables but there are only 111 rows of observations now.

ggpairs(air2)

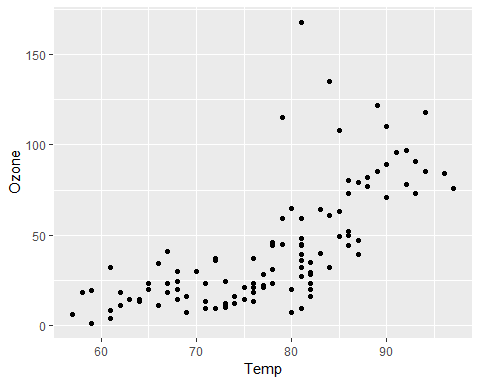


ggcorr(air2, label = TRUE)



The variable that is most strongly correlated with the Ozone variable is the temperature variable.  
The variable that is least strongly correlated with Ozone layer are the day and month variables, this is to be expected with dates, so if we exclude those it becomes Solar.R.

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



It appears that there is a positive correlation between the two variables - as temperature increases so does the Ozone variable.

model1 = lm(Ozone ~ Temp, air2) #Creates Linear Model  
summary(model1) #Analyze Linear Model

##   
## 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

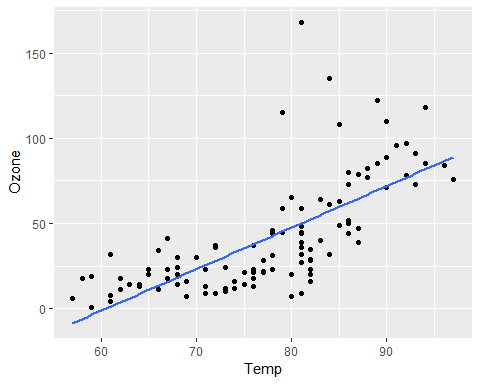
confint(model1) #Produces Confidence Interval

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

Overall the quality of this linear model is not too terrible. The preidctor variable is statistically significant as the P-value is less than .05 and the R squared value of .488 is not that bad, it can explain roughly half the data points.

The range of the slope coeffeicient typically falls between 1.964787 and 2.913433.

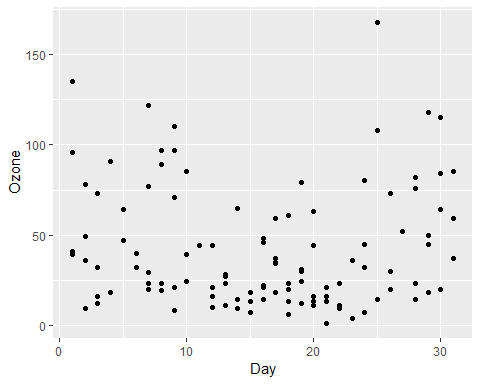
ggplot(air2,aes(x=Temp,y=Ozone)) +geom\_point() + geom\_smooth(method="lm", se = FALSE)



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

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

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



model2 = lm(Ozone ~ Day, air2) #Creates Linear Model  
summary(model2) #Analyze Linear Model

##   
## 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) #Produces Confidence Interval

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

The second model we created to see if day was a good indicator variable for Ozone is a terrible model. The model has an R squared value that is near zero and the indicator variable is not statistically significant.

R is telling us that we should be 95% confident that the slope coefficient will fall between -.74521 and .7056539

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

