# Simple Linear Regression and Correlation

## Ariel Romero

library(tidyverse)

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

## Registered S3 method overwritten by 'rvest':  
## method from  
## read\_xml.response xml2

## -- Attaching packages --------

## v ggplot2 3.1.1 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 -----------------  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

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

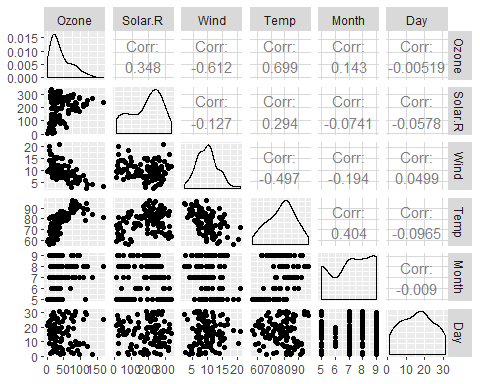
air = airquality  
  
head(air)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

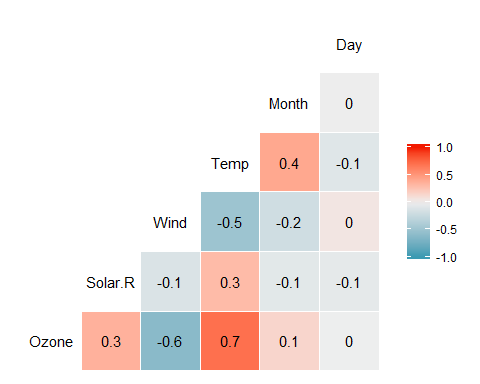
1. The “air” dataset has data that measures the air quality in New York. The data captured ranges from May 1973 to September 1973.
2. There are 6 variables and 153 observations in the dataset.
3. Yes, there is missing data. Primarily in the Ozone and Solar.R fields.
4. The variable to most likely be the response variable is “Ozone”

In the “air2” column, there are 6 columns and 111 rows.

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

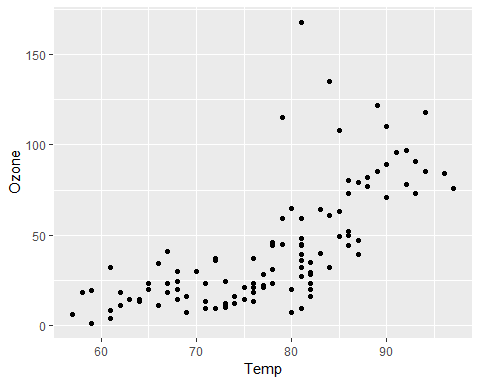


ggcorr(air2, label = TRUE)



1. Temperature is the variable that is most strongly correlated with the Ozone variable.
2. Day is the the variable that is the least strongly correlated with the Ozone variable.

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



Ozone and Temp have a positive correlation.

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

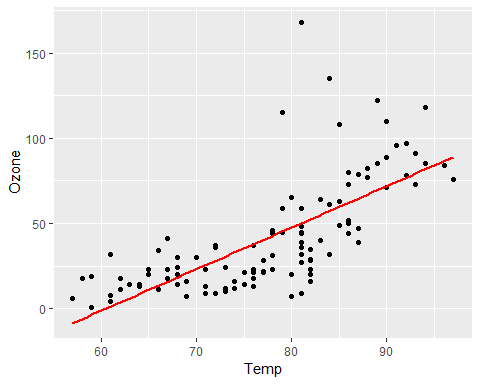
1. Based on the R squared value,0.488, it seems that a good chunk of the data falls within the regression model, thus making it reasonable predictor. The p value associated with this value is less than .05, which means the the regression slope is significant.

confint(model1)

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

The range in which the slop coefficient likely falls is between 1.964787 and 2.913433

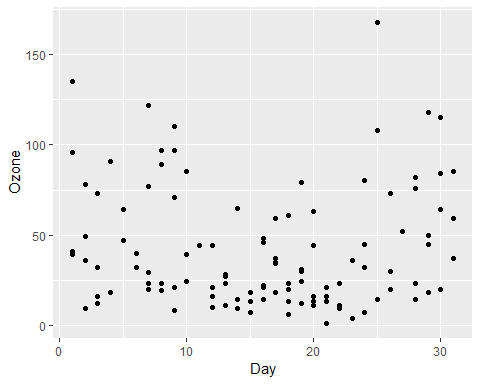
ggplot(data=air2, aes(x=Temp, y=Ozone)) + geom\_point() + geom\_smooth(method="lm",se = FALSE, colour="red")



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

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

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



It appears that there is no correlation 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

1. In this linear regression model does appear to be a good predictor for how Day affects Ozone. The R squared value is practically 0, which means that almost none of the data falls on within the model. The p value is also greater than .05 which makes the slope none significant.
2. The range that the slop coefficient likely falls in is between -0.745321 and 0.7056539

confint(model2)

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

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

