## Simple Linear Regression and Correlation

### Adam Samstag

library(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)

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

##   
## Attaching package: 'GGally'

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

#### Task 1

air = airquality

1. This data set has the records for 153 days of air quality variables such as temperature, wind speend and solar radiation.
2. There are 153 onservations and 6 variables.
3. Yes there are several pieces of missing data.
4. Ozone

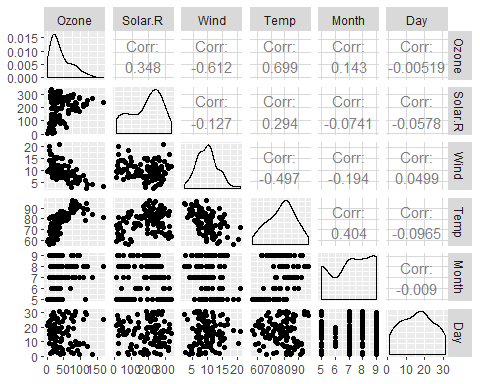
#### Task 2

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

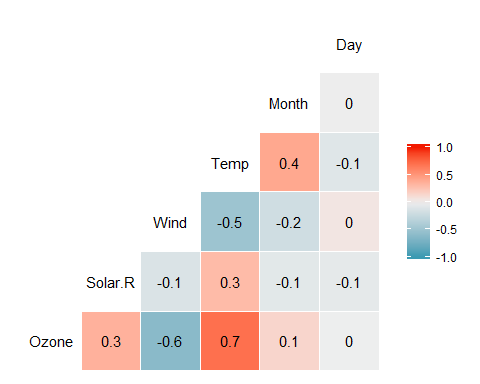
There are now 111 rows and 6 columns.

#### Task 3

ggpairs(air2)

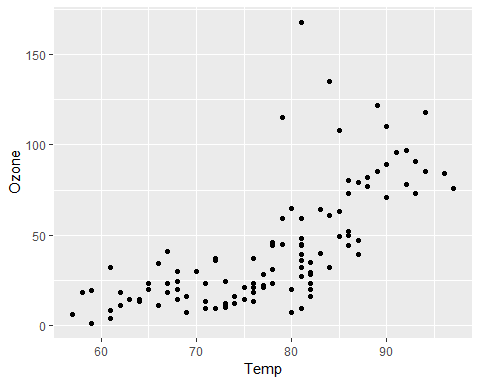


ggcorr(air2, label = TRUE)

 a. It appears that Temp is the most correlated with Ozone b. Day is the least correlated variable with Ozone

#### Task 4

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



Based on the scatter plot you can see that if the Temp is less than 85 degees the Ozone level is lower than 75. If the temp is over 85 degrees the ozone level is over 75.

#### Task 5

model1 =lm(Ozone ~ Temp, data = 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

The p value of this model is <0.05 making it a good predictor for Ozone, however the R-sqaured value is not very good, it isn’t bad but it is also not high enough to be a good variable.

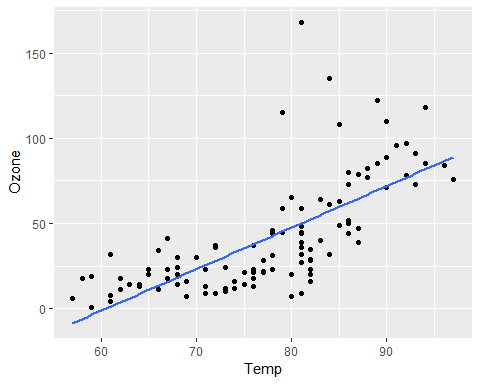
confint(model1, level = 0.95)

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

Between -184.81 and -110.47

#### Task 6

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



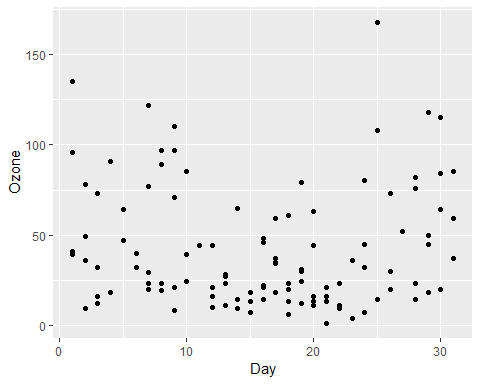
#### Task 7

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

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

#### Task 8

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

 There is no correlation between Day and Ozone.

#### Task 9

model2 =lm(Ozone ~ Day, data = 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

This is not a good moodel, the p-vaule is >0.05 and the R-sqaured value is small meaning that Day is not a good predictor for Ozone.

confint(model2)

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

Between 29.24 and 55.58

#### Task 10

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

