## Correlation and Simple Linear Regression Assignment

library("tidyverse")

## -- 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")

##   
## Attaching package: 'GGally'

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

### Task 1

air= airquality

#### Describe the data set

1. There are 6 variables and 153 observations in the airquality dataset
2. Yes there is missing data from the dataset. The majority of the missing data comes from the Ozone variable.
3. The variable that is likely to be the response (Y) variable is the Ozone variable.

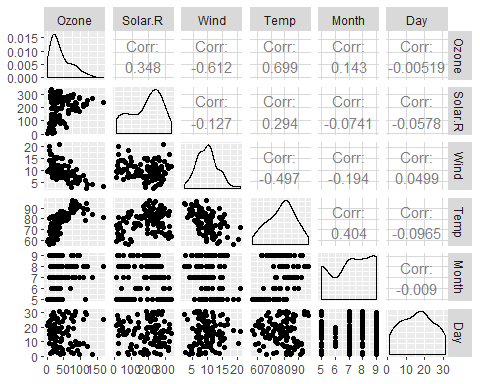
### Task 2

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

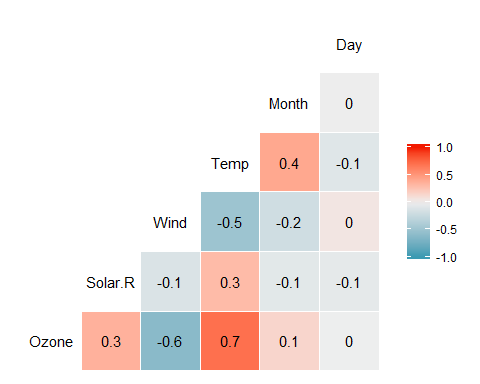
In the new data frame (air2) we have 111 observations and 6 variables. This means 42 variables were removed for having no data.

### Task 3

ggpairs(air2)



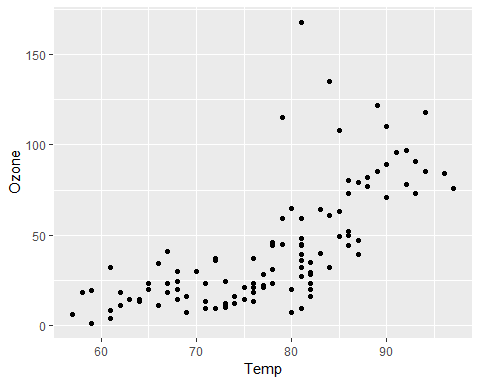
ggcorr(air2, label = TRUE)



a. The Temperature is most strongly correlated with the “Ozone”variable. b. The Day is the least correlated with the “Ozone” Variable.

### Task 4

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

 As the ozone increases the temperature increases as well. There is a positive linear correlation between temperature and ozone.

### Task 5

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

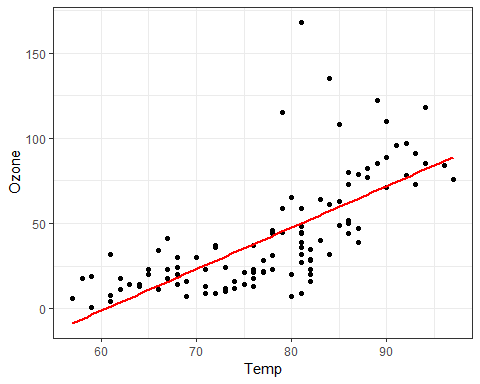
confint(model1)

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

1. R Squared is 23.09 on 109 degrees from freedom. The predictor variable is a significant becuase the p value is less than 0.05.
2. For 95% confidence intervals the slope confident likely falls in the middle range of the dataset.

### Task 6

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

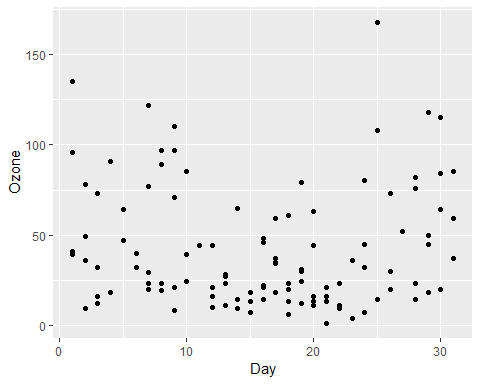


### Task 7 & Task 8

Testdata= data.frame(Temp = c(80))

Predict(air2, newdata= testdata, interval = “predict”)

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



At the moment their appears to be no relationship in the day and the Ozone. The Ozone is majority lower on all days of the month.

### Task 9

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

1. The model is not significant. The P-value is greater than 0.05. The p-value is 0.9569. The residual standard error is 33.43, the multiple R-squared is 2.693e-05, and the adjusted R-squared is -0.09147.
2. The slope coefficent likely falls within the data from 29.2481 to 55.5286.

### Task 10

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

