# Module 2: Assignment 1

## Correlation and Simple Linear Regression Assignment

### Rachel Urban

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)

## 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  
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   
##

1. The dataset contains data about the daily air quality measurements in New York from the months of May to September in 1973.  
   b.There are 6 variables and 153 observations.  
   c.Yes, there is missing data in the Ozone and Solar.R columns.  
   d.The Ozone variable is most likely to be the response variable.

### Task 2

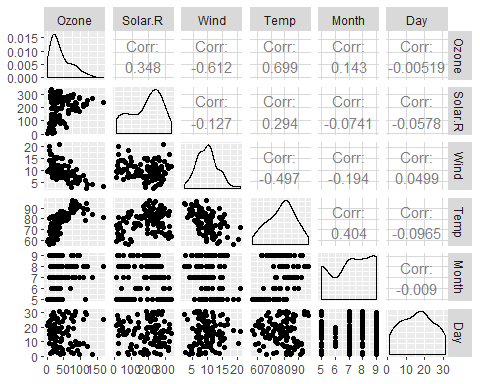
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))  
summary(air2)

## Ozone Solar.R Wind Temp   
## Min. : 1.0 Min. : 7.0 Min. : 2.30 Min. :57.00   
## 1st Qu.: 18.0 1st Qu.:113.5 1st Qu.: 7.40 1st Qu.:71.00   
## Median : 31.0 Median :207.0 Median : 9.70 Median :79.00   
## Mean : 42.1 Mean :184.8 Mean : 9.94 Mean :77.79   
## 3rd Qu.: 62.0 3rd Qu.:255.5 3rd Qu.:11.50 3rd Qu.:84.50   
## Max. :168.0 Max. :334.0 Max. :20.70 Max. :97.00   
## Month Day   
## Min. :5.000 Min. : 1.00   
## 1st Qu.:6.000 1st Qu.: 9.00   
## Median :7.000 Median :16.00   
## Mean :7.216 Mean :15.95   
## 3rd Qu.:9.000 3rd Qu.:22.50   
## Max. :9.000 Max. :31.00

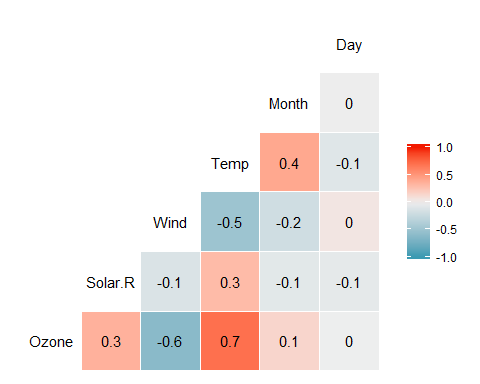
There are 111 rows and 6 columns in the new dataframe air2.

### Task 3

ggpairs(air2)

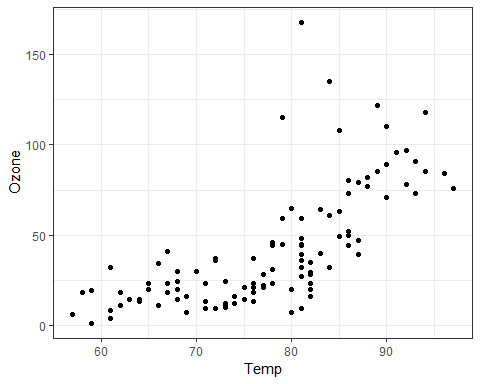


ggcorr(air2, label = TRUE)

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

### Task 4

ggplot(air2,aes(Temp,Ozone))+   
 geom\_point()+  
 theme\_bw()

 The relationship between Temp and Ozone can be described as that the temperature increases, so does the 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

1. The R-squared value is okay, could be better. The significance of the predictor variable is very good as the p-value is <0.05. I would say the quality of the model is pretty good overall.

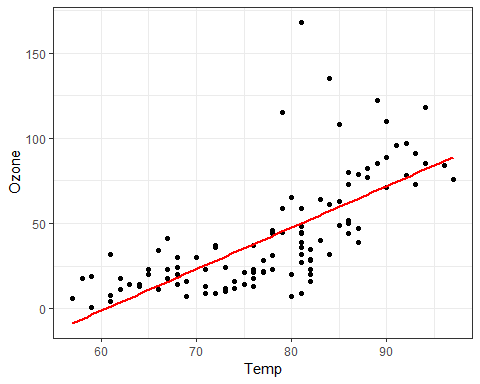
confint(model1)

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

The slope coefficient likely falls between 1.96 and 2.91.

### Task 6

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



### Task 7

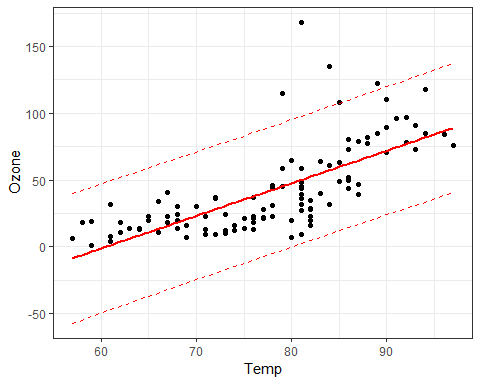
tempdata = data.frame(Temp = 80)  
predict(model1,newdata=tempdata,interval = "predict")

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

temp\_var <- predict(model1, interval="prediction")

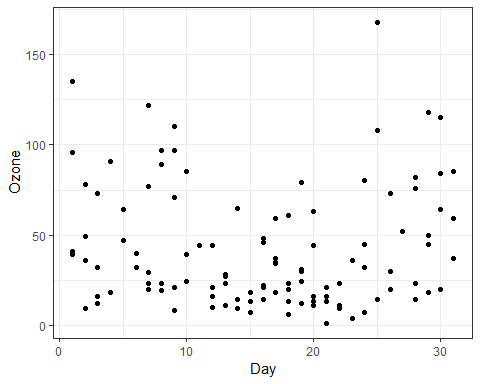
## Warning in predict.lm(model1, interval = "prediction"): predictions on current data refer to \_future\_ responses

pred\_interval = cbind(air2, temp\_var)  
  
ggplot(pred\_interval,aes(Temp,Ozone)) +   
 geom\_point() +   
 geom\_smooth(method="lm",se=FALSE,color="red") +   
 geom\_line(aes(y=lwr),color="red",linetype="dashed")+  
 geom\_line(aes(y=upr),color="red",linetype="dashed")+  
 theme\_bw()



### Task 8

ggplot(air2,aes(Day,Ozone))+  
 geom\_point()+  
 theme\_bw()

 The relationship between the variables Day and Ozone is that as the days go on the ozone seems to increase slightly.

### 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

1. The quality of the model appears low based on the very low R-sqaured value and the visual results. The significance of the predictor value is showing pretty good as the p-value is less than 0.05.

confint(model2)

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

The slope coefficient likely falls between -0.75 and 0.71.

### Task 10

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

