# Alex Lugo

## Simple Linear Regression and Correlation

Load libraries

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

#install.packages("GGally")  
library(GGally)

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

##   
## Attaching package: 'GGally'

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

#install.packages("car")  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

#install.packages("lmtest")  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

**Task 1** The dataset contains 153 observations and 6 variables.There are 37 missing observations in the Ozone variable and 7 missing observations in the Solar.R variable. Ozone is likely the response variable.

air = airquality  
str(air)

## 'data.frame': 153 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

summary(air)

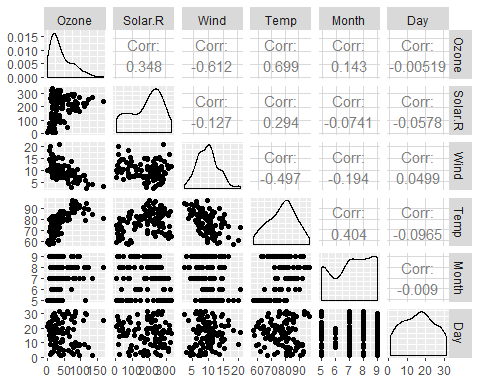
## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd 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   
##

**Task 2** The below deletion of rows with missing data has reduced the total amount of rows to 111. There are still 6 columns.

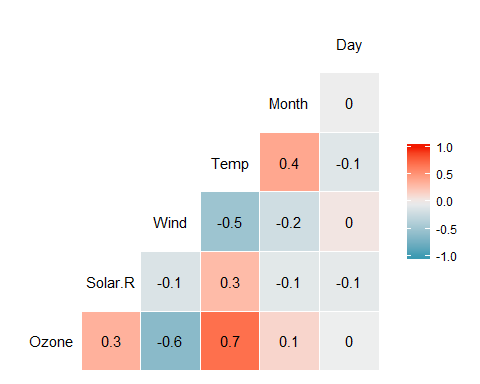
air2 = air %>% drop\_na()

**Task 3** Temp is most strongly correlated to Ozone. The least correlated variable to Ozone is Day.

ggpairs(air2)

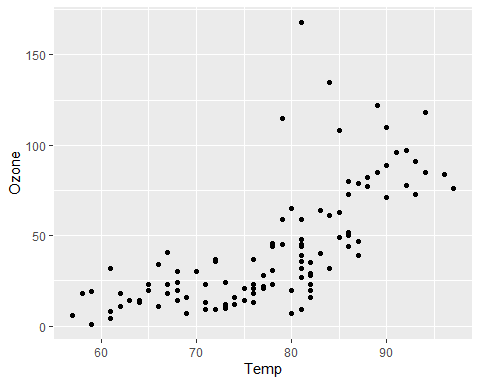


ggcorr(air2, label = TRUE)



**Task 4** The below plot shows that Ozone has a positive correlation to Temp.

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



**Task 5** a. The quality of the model is good. The R squared value is not quiute as high one might want but when we take a look at the p-value, it is showing evidence that Temp is a significant predictor of Ozone.

1. The range for the slope coefficient is between 1.96 and 2.91.

model1 = lm(Ozone ~ Temp, air2)  
summary(model1)

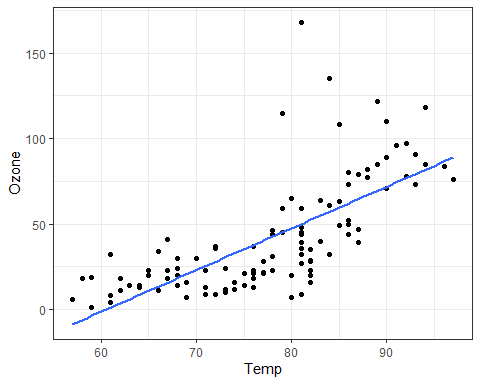
##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 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, level = 0.95)

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

**Task 6**

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



**Task 7** Ozone of 47.48 is the prediction when Temp is 80.

predict(model1, data.frame(Temp = 80))

## 1   
## 47.48272

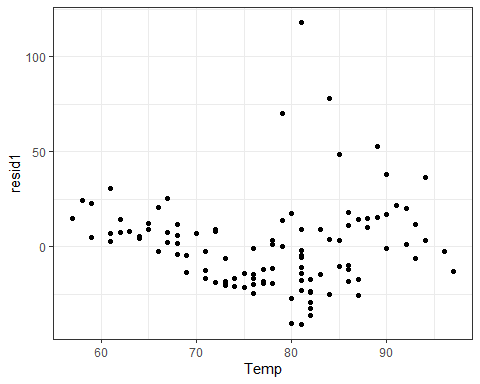
**Task 8** 1. Predictor and response variable have linear relationship: Referring to Task 6, we can see that the variables do have a linear relationship.

1. Model residuals are independent: The dwtest function below is telling us that the residuals are independent because the p-value is greater than 0.05.
2. Model residuals exhibit constant variance: Below plot of residuals shows that there doesn’t appear to be change in the residuals.
3. Model residuals are normally-distributed: The below histogram appears to show a somewhat noramlly-distributed chart of the residuals.

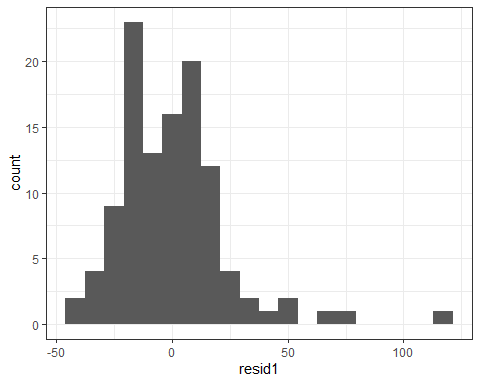
dwtest(model1)

##   
## Durbin-Watson test  
##   
## data: model1  
## DW = 1.8644, p-value = 0.2123  
## alternative hypothesis: true autocorrelation is greater than 0

air2 = air2 %>% mutate(resid1 = model1$residuals)  
ggplot(air2 ,aes(x= Temp,y = resid1)) + geom\_point() + theme\_bw()



ggplot(air2 ,aes(x= resid1)) + geom\_histogram(bins = 20) + theme\_bw()



**Task 9** The model in Task 5 might be used to predict ozone levels for a day given the temperature. I would be somewhat cautious with the model because there are other factors that are also similarly correlated to ozone. I would say the model would give a decent indication of what range ozone could be at but other factors could swing the actual ozone higher or lower than predicted by the model.