## Simple Linear Regression and Correlation - Module 2 Assignment 1

### Nicholas Barrett, 02/02/19

### Task 1 - Read-in the airquality dataset as a df called “air”.

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

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   
##

1. All of the variables (Ozone, Solar.R, Wind, Temperature, Month and Day) in the new “air” dataset are represented as numeric.
2. The “air” dataset contains 153 observations of 6 variables
3. The Ozone and Solar R variables as missing data (37 in “Ozone” and 7 in “Solar.R”)
4. Ozone levels is the most likely response variable (Y)

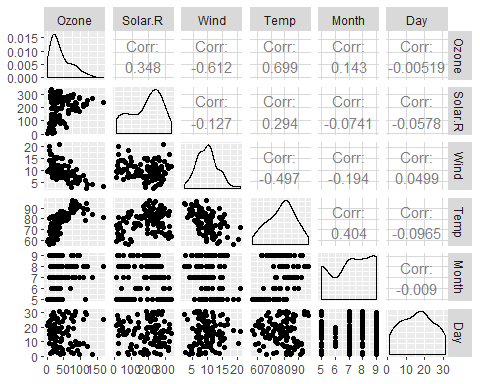
### Task 2 - Missing data in the two variables: “Ozone” and “Solar.R” Three options - Delete the rows with missing data, delete the columns with missing data or impute values to replace missing values. Delete rows with missing data.

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

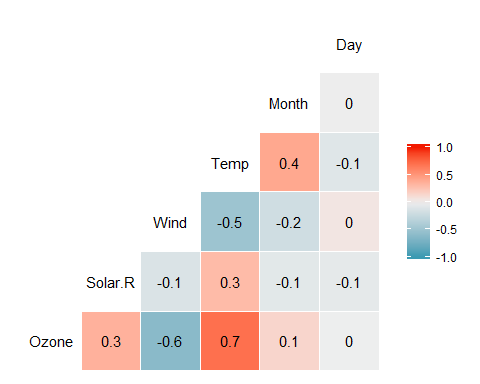
The new “air2” dataset has 111 rows and 6 columns.

### Task 3 - Use the code provided to develop a visualization of and to calculate correlation for the combinations of variables in this dataset. Use “ggcorr” to develop a correlation matrix for the variables.

ggpairs(air2)



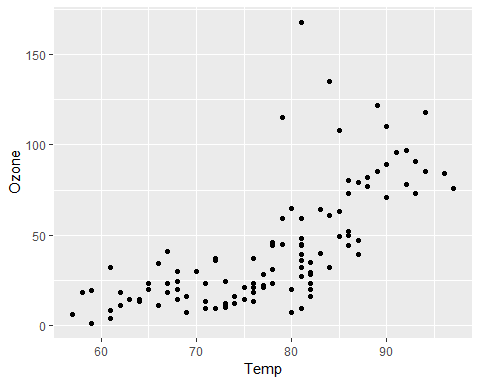
ggcorr(air2,label = TRUE)



1. The variable that most strongly correlates with the “Ozone” variable is “Temp”.
2. The variable that least strongly correlates with the “Ozone” variable is “Wind”.

### Task 4 - Plot “Temp” vs. “Ozone” using “ggplot” using appropriate chart typle. Describe the relationship between the two variables.

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



Based on the plot above, “Temp” appears to have an increasing linear relationship with “Ozone” with some outliers. As temparatures rise, ozone levels appear to rise.

### Task 5 - Create a linear regression model (model1) using “Temp” to predict “Ozone”.

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

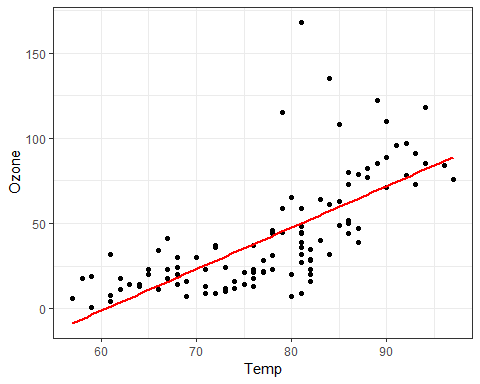
1. The above “model1” model for the “air2” appears to be a decent predicative model for determining Ozone from the Temp variables. The Temp, predictor, variable is significant with a p-value less than 0.05. The R-squared value is pretty good at 0.4833.
2. Based on the 95% confidence interval below,the slope co-efficent falls between 1.964787 and 2.913433

confint(model1)

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

### Task 6 - Re-do Task 4

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



### Task 7 - Develop a prediction and prediction interval for “Ozone” when “Temp” is 80.

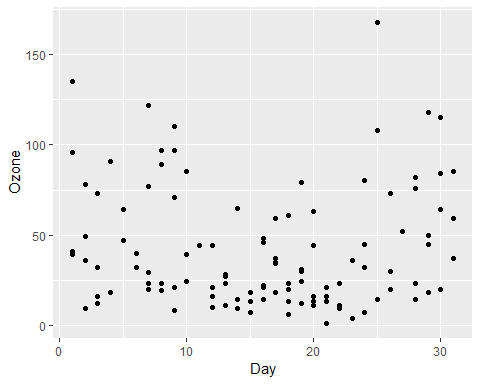
testdata = data.frame(Temp = 80)  
predict(model1, newdata = testdata, interval = "predict")

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

The prediction for Ozone when Temp is 80 is 47.48272 with a predicaiton interval between -0.1510188 lower and 95.11646 upper.

### Task 8 - Plot “Day” vs. “Ozone” using ggplot. Describe the relationship between the two variables.

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



Considering the plot above, there does not appears to be an apparent or significant linear relationship between “Day” and “Ozone”.

### Task 9 - Create a linear regression model called “model2” using “Day” to predict “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. The above “model2” model for the “air2” appears to be a very poor predicative model for determining Ozone from the Day variables. The Day, predictor, variable is insignificant with a p-value more than 0.05 (0.957). The R-squared value is also bad at -0.009147.
2. Based on the 95% confidence interval below,the slope co-efficent falls between -0.745321 and 0.7056539

confint(model2)

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

### Task 10 - Re-do Task 8 to inlcude the regression line.

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

