### Module 2: Assignment 1

### Correlation and Simple Linear Regression Assignment

### BAN502

### Mayberry, Brandon

Libraries

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  
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   
##

glimpse(air)

## Observations: 153  
## Variables: 6  
## $ Ozone <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, ...  
## $ Solar.R <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256,...  
## $ Wind <dbl> 7.4, 8.0, 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, 74, 69, 66, 68...  
## $ Month <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...  
## $ Day <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...

1. This dataset is the daily air quality in New York from May to September 1973.
2. There are 6 variables and 153 observations
3. There is missing data as ozone and Solar.R with ozone having 37 NA’s and Solar.R with 7 NA’s respectively.
4. I would think that ozone is the response variable.

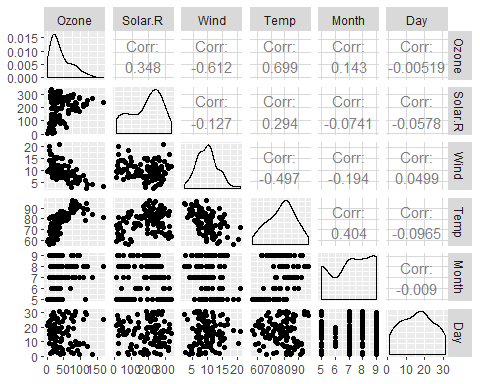
Task 2

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

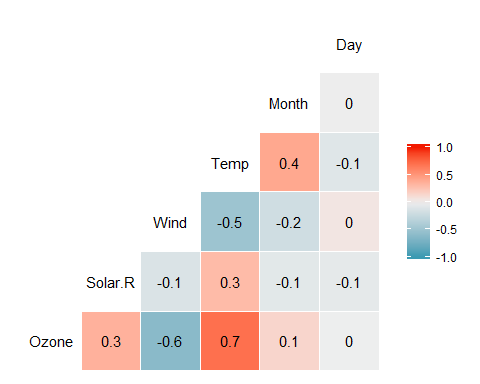
111 rows and 6 columns remain

Task 3

ggpairs(air2)

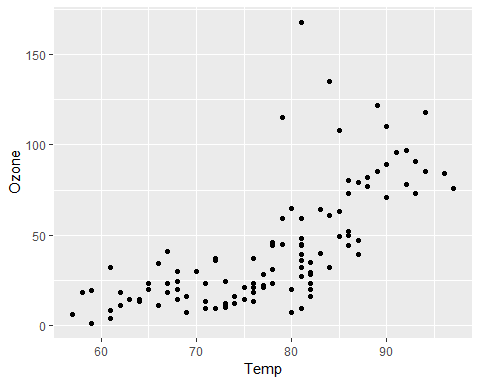


ggcorr(air2, label=TRUE)

 a. The Temperature variable is most strongly correlated to the Ozone variable. b. The Day variable is least strongly correlated to the Ozone variable.

Task 4

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

 As the temperature increases, we see an inrease in the Ozone level.

Task 5

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

##   
## Call:  
## lm(formula = Temp ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.980 -4.775 1.825 4.228 12.425   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.37059 1.05151 65.97 <2e-16 \*\*\*  
## Ozone 0.20006 0.01963 10.19 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.851 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. This model seems to be a fairly decent model. R square is a respectable .4833 and the P value is significantly less than .05, meaning the there is a high significance of the predictor variable.

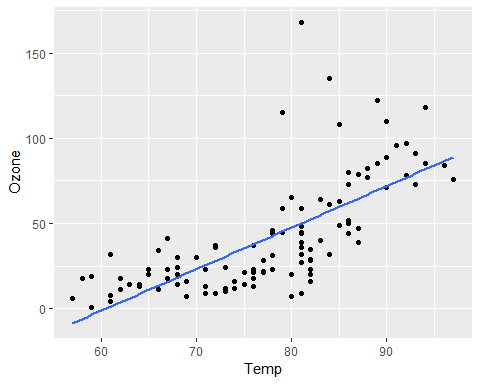
confint(model1)

## 2.5 % 97.5 %  
## (Intercept) 67.2865285 71.4546496  
## Ozone 0.1611525 0.2389608

1. The slope coefficient will likely fall between 0.1611525 and 0.2389608

Task 6

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



Task 7

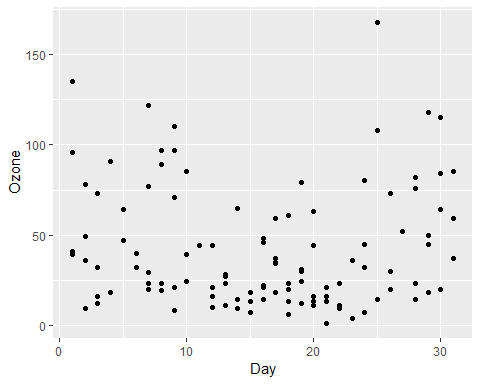
testdata = data.frame(Ozone = c(80))  
predict(model1, newdata = testdata, interval = "predict")

## fit lwr upr  
## 1 85.37512 71.65702 99.09322

The prediction interval is 71.65702 to 99.09322

Task 8

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

 There is no relationship between Ozone and Day. They do not correlate at all.

Task 9

model2 = lm(Day ~ Ozone, air2)  
summary(model2)

##   
## Call:  
## lm(formula = Day ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.9502 -6.9834 0.0621 6.5152 15.1123   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.003116 1.342588 11.920 <2e-16 \*\*\*  
## Ozone -0.001358 0.025063 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.747 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 this model is very poor. The R square is very close to zero with -.009147 and the p value is close to 1 at .9569 meaning there is little to no significance in the predictor variable.

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 13.34214999 18.66408196  
## Ozone -0.05103151 0.04831554

1. The slope coefficient will likely fall between -.05103151 and .04831554

Task 10

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

