# Module 2 - Correlation and Simple Linear Regression

## Sox, Morgan

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

## ── Attaching packages ────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(GGally)

##   
## Attaching package: 'GGally'

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

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   
##

is.na(air)

## Ozone Solar.R Wind Temp Month Day  
## [1,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [5,] TRUE TRUE FALSE FALSE FALSE FALSE  
## [6,] FALSE TRUE FALSE FALSE FALSE FALSE  
## [7,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [8,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [9,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [10,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [11,] FALSE TRUE FALSE FALSE FALSE FALSE  
## [12,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [13,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [14,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [15,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [16,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [17,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [18,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [19,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [20,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [21,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [22,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [23,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [24,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [25,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [26,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [27,] TRUE TRUE FALSE FALSE FALSE FALSE  
## [28,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [29,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [30,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [31,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [32,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [33,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [34,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [35,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [36,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [37,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [38,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [39,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [40,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [41,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [42,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [43,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [44,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [45,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [46,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [47,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [48,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [49,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [50,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [51,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [52,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [53,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [54,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [55,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [56,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [57,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [58,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [59,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [60,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [61,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [62,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [63,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [64,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [65,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [66,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [67,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [68,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [69,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [70,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [71,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [72,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [73,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [74,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [75,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [76,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [77,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [78,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [79,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [80,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [81,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [82,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [83,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [84,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [85,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [86,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [87,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [88,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [89,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [90,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [91,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [92,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [93,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [94,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [95,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [96,] FALSE TRUE FALSE FALSE FALSE FALSE  
## [97,] FALSE TRUE FALSE FALSE FALSE FALSE  
## [98,] FALSE TRUE FALSE FALSE FALSE FALSE  
## [99,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [100,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [101,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [102,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [103,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [104,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [105,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [106,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [107,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [108,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [109,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [110,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [111,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [112,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [113,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [114,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [115,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [116,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [117,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [118,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [119,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [120,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [121,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [122,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [123,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [124,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [125,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [126,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [127,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [128,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [129,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [130,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [131,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [132,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [133,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [134,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [135,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [136,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [137,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [138,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [139,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [140,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [141,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [142,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [143,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [144,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [145,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [146,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [147,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [148,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [149,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [150,] TRUE FALSE FALSE FALSE FALSE FALSE  
## [151,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [152,] FALSE FALSE FALSE FALSE FALSE FALSE  
## [153,] FALSE FALSE FALSE FALSE FALSE FALSE

sum(is.na(air))

## [1] 44

Task 1:

There are 153 obeservations in the dataset and 6 variables. We see there is a variable for Ozone, Solar, Wind, Temperature, Month, and Day. the majority of variables are integers except wind which is a numerical variable. We see several missing values in the Ozone and Solar.R variables. There is a total 44 missing observations. Ozone is likely the Y variable based on the other variables in the dataset.

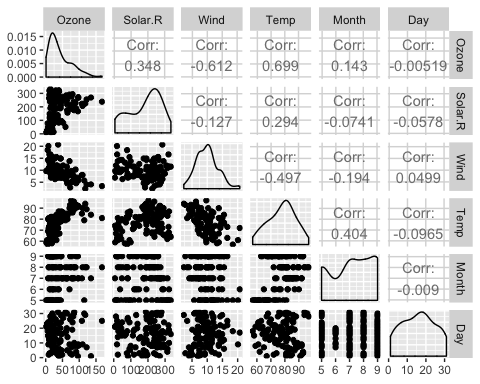
air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))  
summary(air2)

## 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

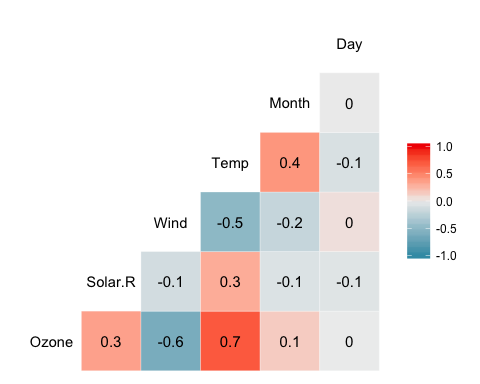
Task 2:

There are now 111 Observations in the new air2 dataset.

ggpairs(air2)



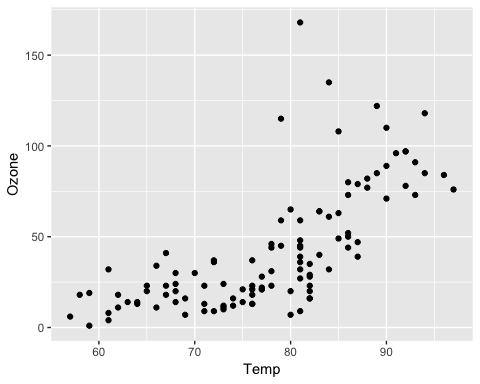
ggcorr(air2, label = TRUE)



Task 3:

1. Temperature is the most strongly correlated with Ozone.
2. Day is the least strongly corelated to Ozone.

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



Task 4:

Based on the graph we see that as temerature increases the Ozone measure also increases. It seems like a slight trend but not a direct one to one relationship.

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

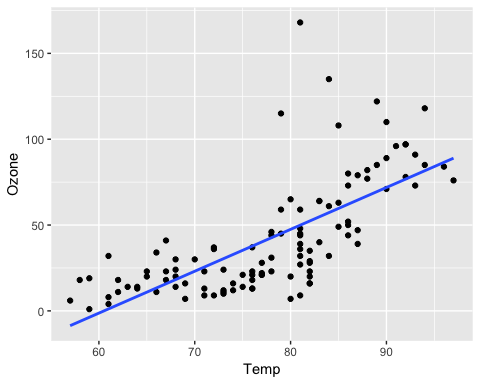
We see that as we add one additional degree of temperature we see a 2.4391 increase in the Ozone measure. The p-value associated with the Temp variable is very small meaning Temperature is a signifigant predictor of Ozone. Also noticed is that the Multiple R-squared value is 0.488 and is fairly a moderately strong fit. These two findings combined help make a case that our model is fairly good.

confint(model1)

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

The slope coefficient falls between 1.964787 and 2.913433.

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

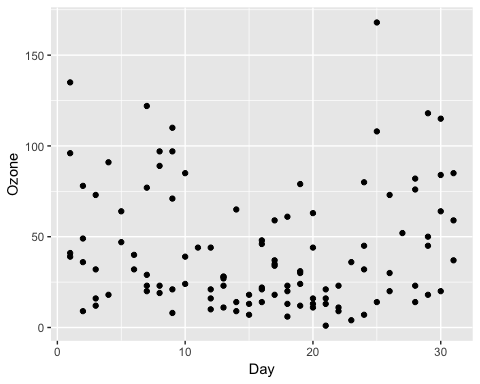
 Task 7:

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

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

Task 8:

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

 Based on the graph above we do not see any trends in the data points as the number of days gets higher or lower. This will be an interesting series to graph an explore the linear regression.

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

Task 9: We see that as we add one additional unit of Day the Ozone measure decreases by 0.01983 units. We also see that the Multiple R-squared value is very close to zero and the p-value is .9569 meaning it is not signifigant and therefore not a good model. The slope will likely fall between -0.745321 and 0.7056539.

confint(model2)

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

Task 10:

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

