# Module 2 - Assignment 1

## Roberts, Damon

### Simple Linear Regression

### Preparing RStudio

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.1 v dplyr 0.8.1  
## v tidyr 0.8.3 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()

library(GGally)

##   
## Attaching package: 'GGally'

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

### Task 1

air = airquality  
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,...

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   
##

###### This dataset contains air quality measurements taken daily in New York from May to September in 1973. It includes 153 observations with 6 variables: Ozone, Solar.R, Wind, Temp, Month, and Day. There are 37 missing observations of the Ozone variable, and 7 missing Solar.R observations. Because this dataset studies air quality, Ozone is the most likely response (Y) variable.

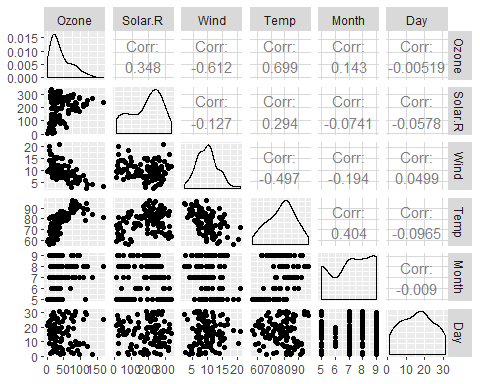
### Task 2

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

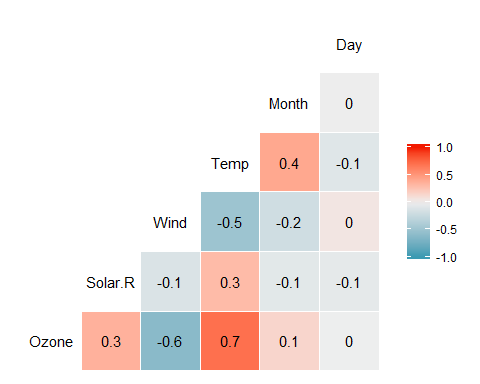
###### After filtering for missing values, the new dataset *air2* includes 111 observations with 6 variables.

### Task 3

ggpairs(air2)



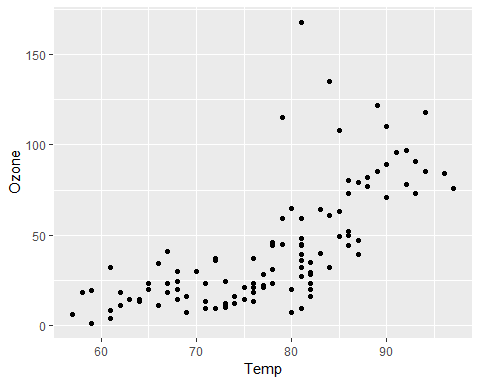
ggcorr(air2, label = TRUE)



###### Temp and Wind both appear to be strongly correlated with Ozone values, with correlations of 0.699 and -0.612, respectively. The least-correlated variable appears to be day, with a correlation of -0.005.

### Task 4

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



###### Based on this plot, it appears Ozone increases gradually as Temp increases. However, the relationship does not sppear to be linear. While the relationship between Temp and Ozone may be linear in lower Temp values, there seems to be an increase in the correlation at Temp values higher than 80.

### 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

confint(model1)

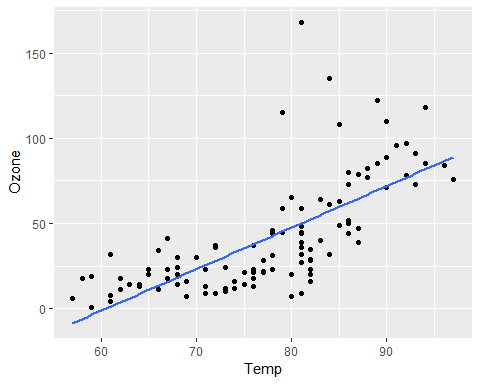
## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

###### From this model, we calculated an R-squared value of 0.488, which suggests the model is of good quality. We also see a p-value of nearly 0, which is less than 0.05 and a significant value.

###### We’re also able to construct a 95% confidence interval to estimate where the slope of the linear regression line may fall. At 95% confidence, we can estimate the slope will be between 1.96 and 2.91.

### Task 6

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



### Task 7

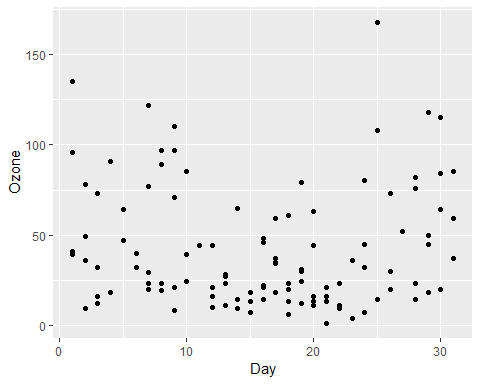
testdata = data.frame(Temp = 80)  
predict(model1, newdata = testdata, interval = "predict")

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

###### Based on this regression model, we could expect to see an Ozone of 47.48 with a Temp value of 80. This falls within the prediction inverval of [-0.15, 95.12].

### Task 8

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



###### Based on this plot, there doesn’t appear to be strong correlation, if any at all, between Day and Ozone. The majority of Ozone values remain below 50 regardless of day. Before day 10 and after day 25, there tend to be more Ozone observations above 50, but these do not indicate a linear relationship between Day and ozone.

### 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

confint(model2)

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

###### After creating a linear regression model to predict Ozone based on Day, we can see there’s little evidence for a relationship between the two. The R-squared value is very small at 2.69e-05, indicating virtually no correlation between the variables. In addition, a p-value of 0.9569 is much larger than our preferred 0.05. At 95% confidence, we estimate the slope coeffecient will likely fall between -0.75 and 0.71.

### Task 10

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

