Assignment\_1

Jordan Whitaker

5/22/2020

library(tidyverse)

library(GGally)

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 ...

Task 1: There are 153 rows and 6 columns (Ozone, Solar.R, Wind, Temp, Month, Day). Ozone and Solar.R have NA Values

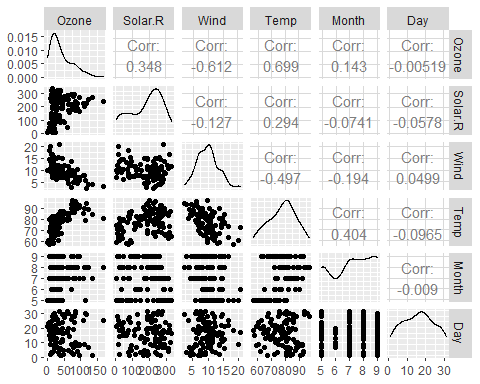
air2 <- air %>% drop\_na()  
str(air2)

## 'data.frame': 111 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 23 19 8 16 11 14 ...  
## $ Solar.R: int 190 118 149 313 299 99 19 256 290 274 ...  
## $ Wind : num 7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...  
## $ Temp : int 67 72 74 62 65 59 61 69 66 68 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 7 8 9 12 13 14 ...

Task 2: After removing the NA values, there are 111 rows with 6 variables left.

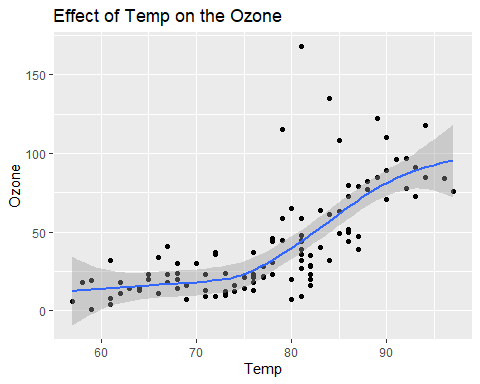
ggpairs(air2, label=TRUE)

## Warning in warn\_if\_args\_exist(list(...)): Extra arguments: "label" are being  
## ignored. If these are meant to be aesthetics, submit them using the 'mapping'  
## variable within ggpairs with ggplot2::aes or ggplot2::aes\_string.

 Task 3: The variable most strongly correlated with Ozone is the Temp variable. While the weakest correlated variable is Day

ggplot(data = air2, aes(x=Temp, y=Ozone)) +   
 geom\_point() +   
 geom\_smooth() +  
 labs(title='Effect of Temp on the Ozone')

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Task 4: The above chart shows a strong relationship between Ozone and Temp

model1 = lm(Ozone ~ Temp, air2)  
model1

##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Coefficients:  
## (Intercept) Temp   
## -147.646 2.439

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

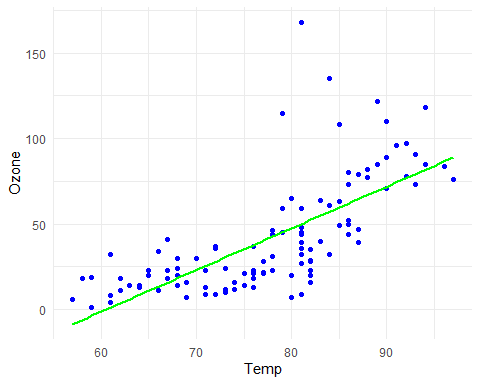
confint(model1)

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

The model shows that the Temp variable is a significant indicator of ozone with a <2e-16 P-Value and the R-Squared is good at .488. Overall, this model is representative of a relationship between temperature and ozone. The intercept likely lies within the range of -184.81 and -110.47 based on the confint function.

ggplot(data = air2, aes(x=Temp, y=Ozone)) +   
 geom\_point(color="blue") +   
 geom\_smooth(method="lm", se=FALSE, color="green") +  
 theme\_minimal()

## `geom\_smooth()` using formula 'y ~ x'



labs(title='Effect of Temp on the Ozone')

## $title  
## [1] "Effect of Temp on the Ozone"  
##   
## attr(,"class")  
## [1] "labels"

Task 6: ^^^

Task 7: When Temp is 80, based on the above model, I would predict that the Ozone will be ~50.

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

residuals(model1)

Task 8: The model looks to fit most of the assumptions of a linear regression model. The relationship is clearly linear with a postiive slope. The variance of residiuals looks to be somewhat the same from a given value of X. Observations are independent and I would assume that for a fixed value of X, Y would be normally distrubuted.

Task 9: This model shows a clear relationship between temperature and ozone. However, when providing advice or promoting this model, it should be noted that although the highest correlated variable was used as a predictor, a single predictor may not be as powerful. Also, this is a relatively small dataset with only 111 observations.