# Module 2 Assignment 1

## Ban 502

### Jennifer Brill

library("tidyverse")

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

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.8  
## ✔ 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  
require("datasets")  
data("airquality")  
str(airquality)

## '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 ...

coll<- mapply(anyNA,airquality)  
coll

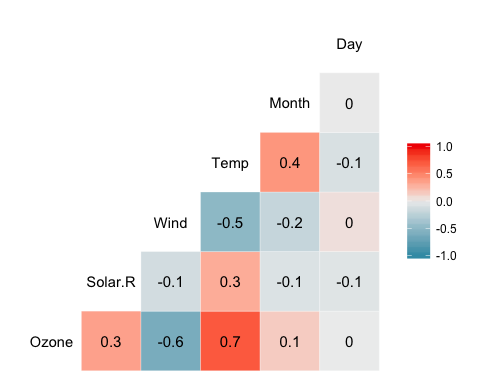
## Ozone Solar.R Wind Temp Month Day   
## TRUE TRUE FALSE FALSE FALSE FALSE

The dataset is fairly small. There are 6 variables and 153 observations. There is missing data in the dataset. The ozone variable is most likely to be the response (Y) variable

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

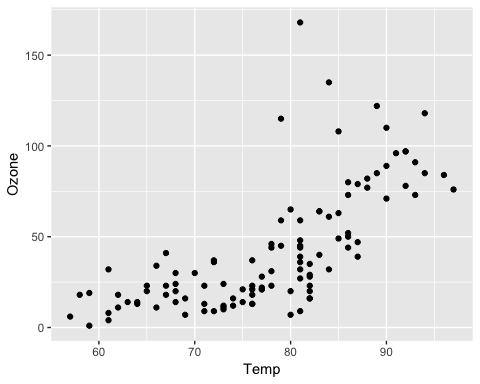
There are 111 rows and 6 columns in the new air2 data frame.

ggcorr(air2, label = TRUE)



Temperature is most strongly correlated with the Ozone variable and Day is least strongly correlated with the Ozone variable.

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



Temperature has a slight positive correlation with 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

Discuss quality of model (mention R square value and significance of predictor variable).

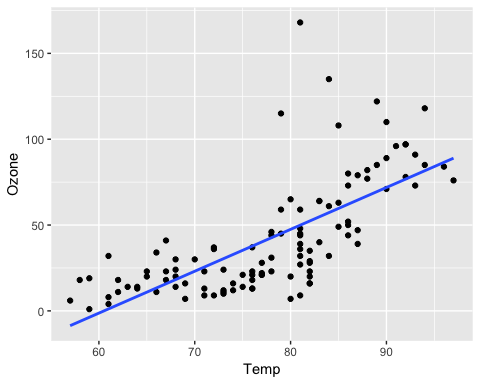
The p value is incredibly small meaning temperature is a significant predictor of ozone. The r-square valued is significant and shows that the two variables are positively correlated.

confint(model1)

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

The slope Coefficient is likely to fall between 1.96 and 2.91.

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

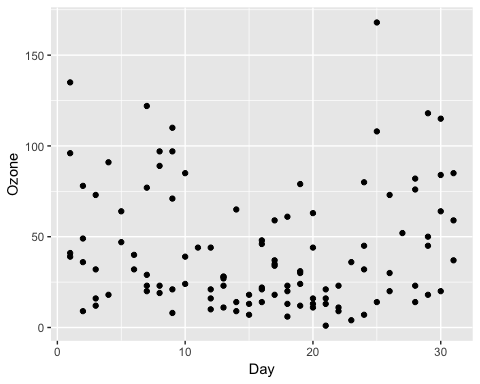


testdata = data.frame(Temp = c(80))  
predict(model1, newdata = testdata, interval = "predict")

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

Ozone is estimated to be 47.48 when temperature is 80. The prediction intervatl is between -0.15 and 95.12

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



Day and ozone do not appear to have a strong correlation.

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

The p value is above 0.05 which means it is not significant and the r seuwared value is incredibly small which means there is no real correlation between Day and Ozone.

confint(model2)

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

The slope coefficient is likely to fall between -0.75 and 0.71

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

