library(tidyverse)  
library(GGally)  
library(car)  
library(lmtest)

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   
##

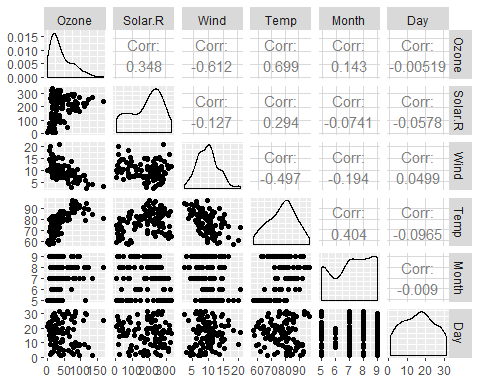
This data set has 153 observations (rows) of 6 variables (columns). There is missing data. The response variable is likely Ozone.

air2 = air %>% drop\_na()  
summary(air2)

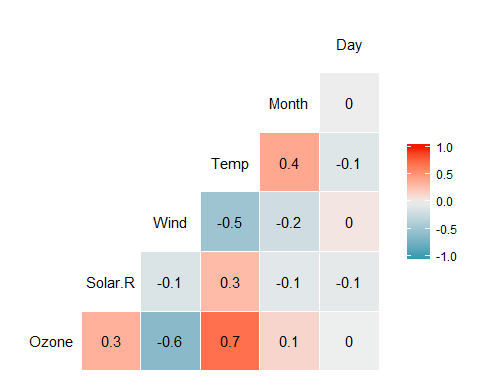
## Ozone Solar.R Wind Temp   
## Min. : 1.0 Min. : 7.0 Min. : 2.30 Min. :57.00   
## 1st Qu.: 18.0 1st Qu.:113.5 1st Qu.: 7.40 1st Qu.:71.00   
## Median : 31.0 Median :207.0 Median : 9.70 Median :79.00   
## Mean : 42.1 Mean :184.8 Mean : 9.94 Mean :77.79   
## 3rd Qu.: 62.0 3rd Qu.:255.5 3rd Qu.:11.50 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

After we removed rows with missing data, there are now 111 observations (rows) of 6 variables (columns).

ggpairs(air2)

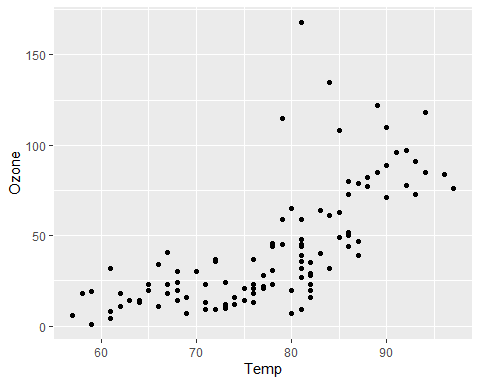


ggcorr(air2, label=TRUE)



The “Temp” variable is the most strongly correlated with the “Ozone” variable, and the “Day” variable is the least correlated with the “Ozone” variable.

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



There is a positive correlation between the “Temperature” and “Ozone” variables. As the temperature value rises, so does the ozone value.

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

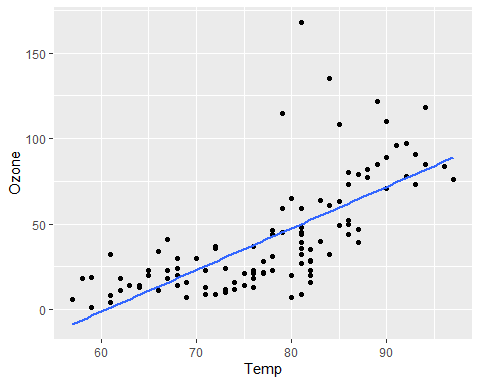
A. This model is pretty useful for determining the correlation between Temperature and Ozone. It gives us a value for the y-intercept and slope at about -147.65 and 2.44 respectively. The p-value was determined to be <2e-16 which is well below the desired 0.05, so we can determine that Temperature is indeed a significant predictor of Ozone. Additionally, the adjusted R-squared value is 0.4833 which, while not terribly close to 1, is still a fine and useful metric in this model.

confint(model1)

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

B. Confint shows us a slope value of about 2.91 for the upper 97.5% confidence interval. Since the regression model gives us a slope of 2.43, we can assume that the slope coefficient likely falls in the upper 50th percentile range in terms of confidence, which is…just okay.

ggplot(model1, aes(Temp, 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

Prediction of Ozone value = 47.48272 when temperature = 80.

**Assumption 1:** The predictor and response variable have a linear relationship. Appropriate modeling for this assumption already completed. As mentioned above, it seems resonable to assume that these two variables have a linear relationship based on the p-value and r-square value.

**Assumption 2:** Model errors (residuals) are independent. Durbin-Watson Test.

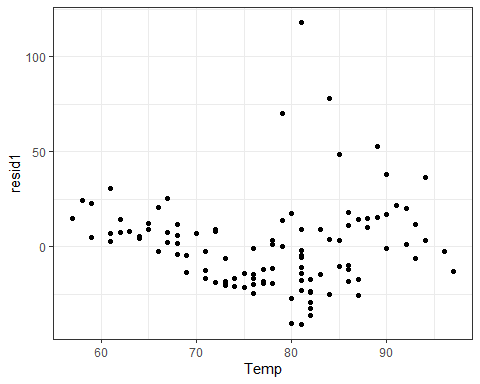
dwtest(model1)

##   
## Durbin-Watson test  
##   
## data: model1  
## DW = 1.8644, p-value = 0.2123  
## alternative hypothesis: true autocorrelation is greater than 0

We fail to reject the null hypothesis with a p-value greater than 0.05. Value is 0.2123. Therefore, the variables are likely independent.

**Assumption 3:** Model residuals exhibit constant variance

air2 = air2 %>%   
 mutate(resid1 = model1$residuals)  
  
ggplot(air2,aes(Temp, resid1)) + geom\_point() + theme\_bw()

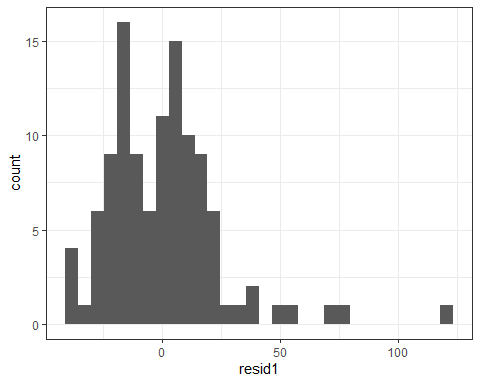


The variance of the residuals looks constant, with the exception of a few outliers.

**Assumption 4:** Model residuals are Normally-distributed

ggplot(air2,aes(x=resid1)) + geom\_histogram() + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The residuals histogram shows a (relatively) normal distribution, albeit skewed. I think it is safe to assume this assumption is valid, although we may benefit from removing a few outliers.

The model contstructed in Task 5 would be used to determine the correlation between the Temp and Ozone variables. When recommending the model, I would suggest to not only make decisions based off the p-value. It is important to also take into consideration the R-square value. Since this value is only 0.4833, which is not all that close to 1 (but still useful), it may be worth it to use other models when investigating the correlation between these two variables.