Mod2Assign1

# load libraries  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 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)

## Warning: package 'GGally' was built under R version 3.6.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

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

library(car)

## Warning: package 'car' was built under R version 3.6.2

## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

library(lmtest)

## Warning: package 'lmtest' was built under R version 3.6.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

**Task 1**

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 ...

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   
##

There are 153 observations and 6 columns. There is missing data for Ozone and Solar.R. I’m guessing that Ozone will be the variable we are concerned about, or perhaps Solar.R.

**Task 2**

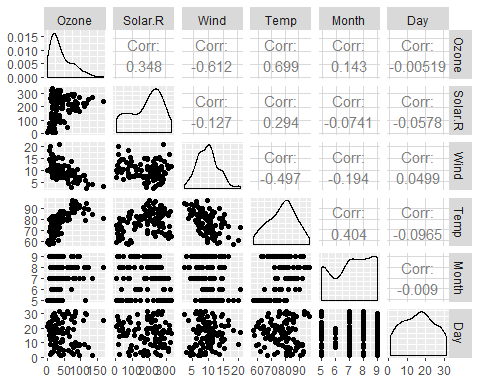
air2 = air %>% drop\_na() #delete any row with an NA value  
str(air)

## '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 ...

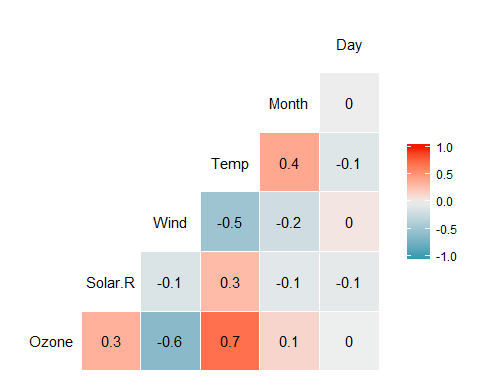
After deleting the rows with NA in them, we have 111 observations, still with 6 columns.

**Task 3**

ggpairs(air2)



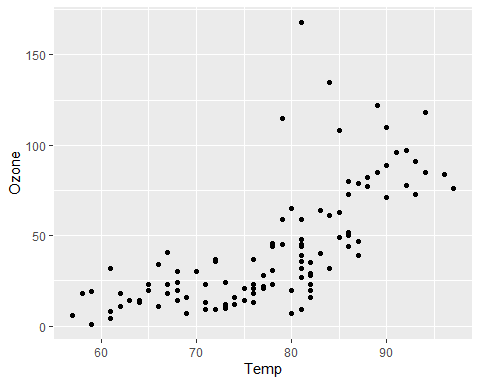
ggcorr(air2, label = TRUE)



Which variable is most strongly correlated with the “Ozone” variable? Temperature has a correlation of 0.699 Which variable is least strongly correlated with the “Ozone” variable? Day has a correlation of -0.005

**Task 4**

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



From the graph we can see that the ozone does go up as the Temperature goes up.

**Task 5**

model1 = lm(Ozone ~ Temp, air2) #create linear regression model  
summary(model1) #examine the model

##   
## 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

1. The p-values are both much less than 0.05, so there does seem to be a correlation. R-squared is 0.488, indicating that there is some strong positive correlation.

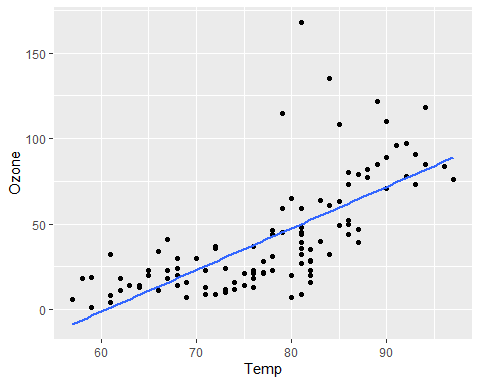
confint(model1, level = 0.95)

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

1. For a 95% confidence interval, the slope coefficient is likely to fall between 1.96 and 2.91

**Task 6**

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



**Task 7**

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

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

ANSWER: The predicted value is 47.48, which looks right from looking at the regression line on the graph.

**Task 8**

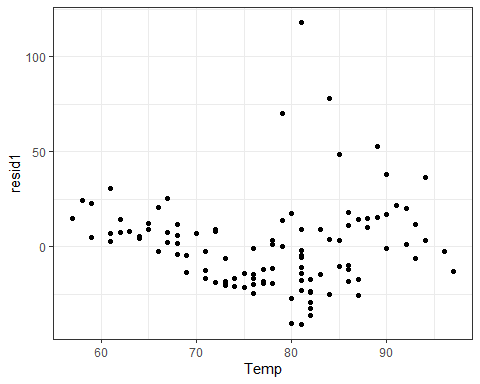
1. The predictor and response variable have a linear relationship ANSWER: Plotting Temp vs. Ozone shows that there does appear to be a linear relationship.

dwtest(model1)

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

1. Model errors (residuals) are independent (recall that a residual is the difference between a predicted value and the actual value) ANSWER: The p-value is not less than 0.05, which shows that the residuals are independent.

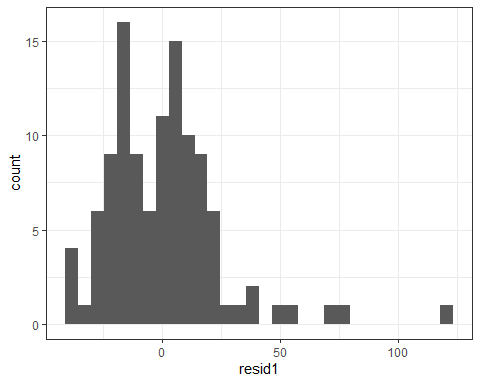
air2 = air2 %>% mutate(resid1 = model1$residuals) #add the model residuals to our data frame  
ggplot(air2,aes(x=Temp,y=resid1)) + geom\_point() + theme\_bw()



1. Model residuals exhibit constant variance ANSWER: From plotting the residuals from the model, we can see that they appear to be clustered around 0, and that there is no particular pattern to the variance. They do not spead out as the values of Temp get larger.

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

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



1. Model residuals are Normally-distributed ANSWER: Plotting the residuals as a normal curve shows that the residuals have something resembling a normal relationship.

**Task 9**

Since there is a correlation between temperature and ozone, the general public can be alerted ahead of time if the ozone levels are predicted to be high. The higher the temperature, the higher the ozone level.