Load libraries

library(tidyverse)

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

## 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: Read-in the airqualtiy data set (a default R dataset) as a data frame called “air”. Details concerning this dataset can be found here: <http://rpubs.com/Nitika/linearRegression_Airquality>. Describe this dataset. How many variables and observations are there? Is there any missing data? Which variable is likely to be the response (Y) variable?

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   
##

*The air quality dataset contains 6 variables and 153 observations. There are records with missing data as we can see in the ozone and Solar.R variables. It is likely that the response variable will be the ozone variable.*

### Task 2: In Task 1 you would have discovered that there is missing data in two of the variables: Ozone and Solar.R. We have three approaches that we can typically select from to deal with missing data:

### 1. Delete the rows with missing data

### 2. Delete the columns with missing data

### 3. Impute (i.e., estimate or guess) values to replace the missing values.

### Here we’ll choose deletion of rows with any missing data. Use the Tidyverse drop\_na function to remove any row with missing data. Save your new data frame (with missing data removed) as a data frame named “air2”.

### How many rows and columns remain in this new (air2) data frame?

air2<-drop\_na(air)  
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

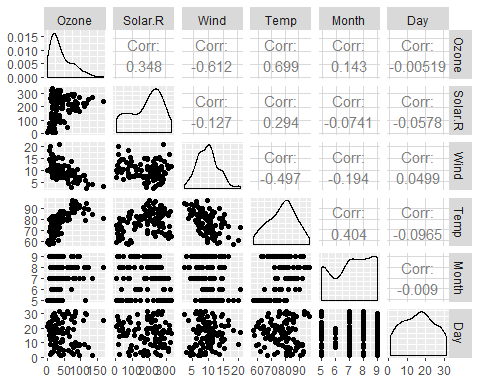
*After dropping the missing data, we have 6 variables and 111 observations.*

### Task 3: Use the ggpairs function to develop a visualization of and to calculate correlation for the combinations of variables in this dataset. Then use the “ggcorr” function to develop a correlation matrix for the variables. Hint: Use “label = TRUE” in the ggcorr function to show the correlation values.

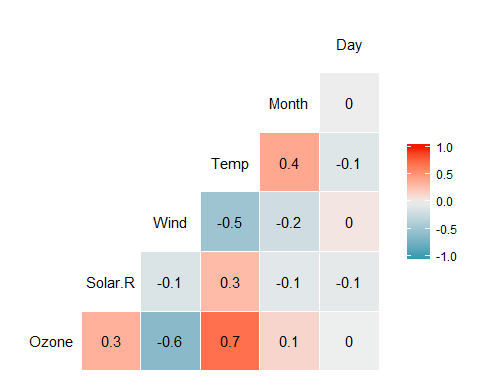
### Which variable is most strongly correlated with the “Ozone” variable?

### Which variable is least strongly correlated with the “Ozone” variable?

ggpairs(air2)



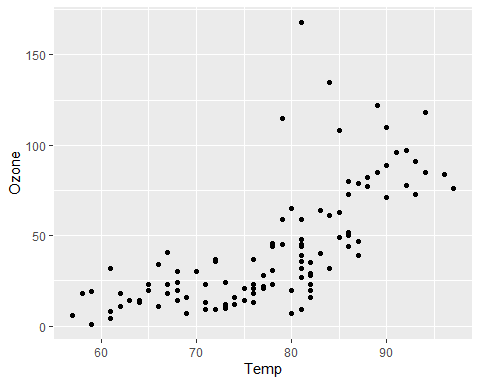
ggcorr(air2, label=TRUE)



*The variable most strongly correlated with Ozone is Temp with a correlation coefficient of 0.7, while the variable least strongly correlated with Ozone is Day with a correlation coefficient of 80.*

### Task 4: Plot “Temp” (x axis) versus “Ozone” (y axis) using the “ggplot” function. Choose an appropriate chart type. Describe the relationship between “Temp” and “Ozone”.

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



*Based on the scatterplot above, we see that the relationship between temperature and ozone is positive and linear.*

### Task 5: Create a linear regression model (called model1) using “Temp” to predict “Ozone”.

### a. Discuss the quality of this model (mention the R square value and significance of the predictor variable).

### b. Use the code “confint(model1)” to generate 95% confidence intervals for the coefficients. In what range does the slope coefficient likely fall?

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)

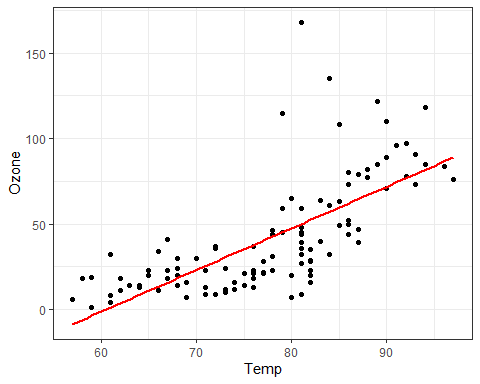
## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

*Based on the linear model above, we see that temperature appears to be a strong predictor of ozone. We have a highly significant predictor variable coefficient, and the R^2 value is 0.488 which is not bad. Additionally, the sign of the predictor coefficient is positive, which is to be expect given the context of the data.*

*Based on the confidence interval, we see that the range in which the slope coefficient is likely to fall is between [1.96, 2.91].*

### Task 6: Re-do Task 4 to include the regression line. Hint: Add “geom\_smooth(method=”lm“, se = FALSE)”.

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



*Based on the visual, it appears that the regression line falls between most points, and overall appears to be fairly accurate.*

### Task 7: Develop a prediction for “Ozone” when “Temp” is 80.

temp80 = data.frame(Temp = c(80))  
predict(model1, newdata = temp80, interval = "predict")

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

*Based on our model, we would predict that when Temp=80, Ozone=47.48.*

### Task 8 Perform appropriate model diagnostics to verify whether or not the model appears to meet the four linear regression model assumptions. Provide a brief comment on each assumptions validity for this model.

1. The predictor and response variable have a linear relationship:  
   *This assumption was verified through the plotting in which it was apparent that the predictor and response variables do have a linear relationship.*
2. Model errors (residuals) are independent (recall that a residual is the difference between a predicted value and the actual value)

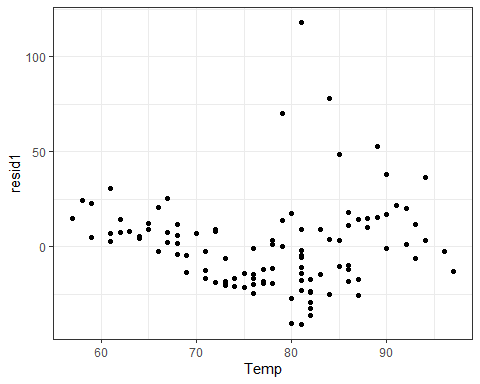
dwtest(model1)

##   
## 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 at 0.21. This suggests that the residuals are likely independent.*

1. Model residuals exhibit constant

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

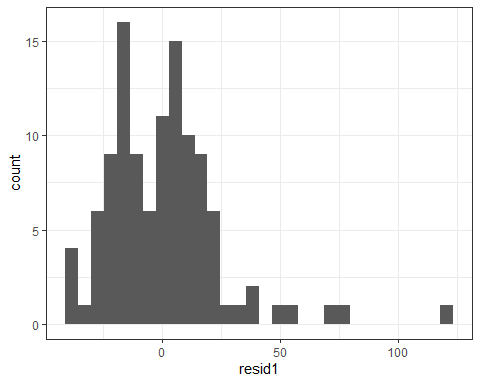


*Based on the plot of the residuals, there does not appear to be change in the variance of residuals.*

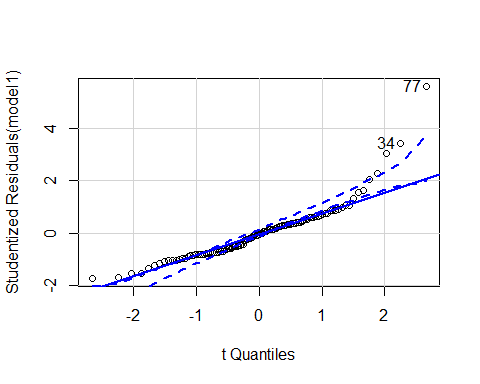
1. Model residuals are Normally-distributed

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

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



qqPlot(model1)



## [1] 34 77

*Based on the histogram and the Normal Probability Plot, it does appear that the model residuals are normally-distributed.*

### Task 9 How might the model that you constructed in Task 5 be used? Are there any cautions or concerns that you would have when recommending the model for use?

*The model from task 5 may be used to predict ozone levels at certain temperatures. Cautions that I would provide would be to restrict the values to non-negative Ozone levels, as well as impractical temperature levels.*