Intro to Data Science HW 7

```
# Enter your name here: Victoria Haley
```

Copyright Jeffrey Stanton, Jeffrey Saltz, and Jasmina Tacheva

Attribution statement: (choose only one and delete the rest)

```
# 1. I did this homework by myself, with help from the book and the professor.
```

The chapter on **linear models** ("Lining Up Our Models") introduces **linear predictive modeling** using the tool known as **multiple regression**. The term "multiple regression" has an odd history, dating back to an early scientific observation of a phenomenon called **"regression to the mean."** These days, multiple regression is just an interesting name for using **linear modeling** to assess the **connection between one or more predictor variables and an outcome variable**.

In this exercise, you will **predict Ozone air levels from three predictors**.

A. We will be using the **airquality** data set available in R. Copy it into a dataframe called **air** and use the appropriate functions to **summarize the data**.

```
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  
## 
```

B. In the analysis that follows, **Ozone** will be considered as the **outcome variable**, and **Solar.R**, **Wind**, and **Temp** as the **predictors**. Add a comment to briefly explain the outcome and predictor variables in the dataframe using **?airquality**.

```
#Ozone is the y-variable (dependent variable)
#Solar.R, Wind, and Temp are the x-variables (independent variables)
```

```
#The changes in Solar.R, Wind, and Temp will impact Ozone
```

C. Inspect the outcome and predictor variables – are there any missing values? Show the code you used to check for that.

```
anyNA(air$Ozone)
```

```
## [1] TRUE
```

```
#There are missing values in Ozone
anyNA(air$Solar.R)
```

```
## [1] TRUE
```

```
#There are missing values in Solar.R
anyNA(air$Wind)
```

```
## [1] FALSE
```

```
#There are no missing values in Wind
anyNA(air$Temp)
```

```
## [1] FALSE
```

```
#There are no missing values in Temp
```

D. Use the **na_interpolation()** function from the **imputeTS package** (remember this was used in a previous HW) to fill in the missing values in each of the 4 columns. Make sure there are no more missing values using the commands from Step C.

```
library(imputeTS)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
air$Ozone <- na_interpolation(air$Ozone)
anyNA(air$Ozone)
```

```
## [1] FALSE
```

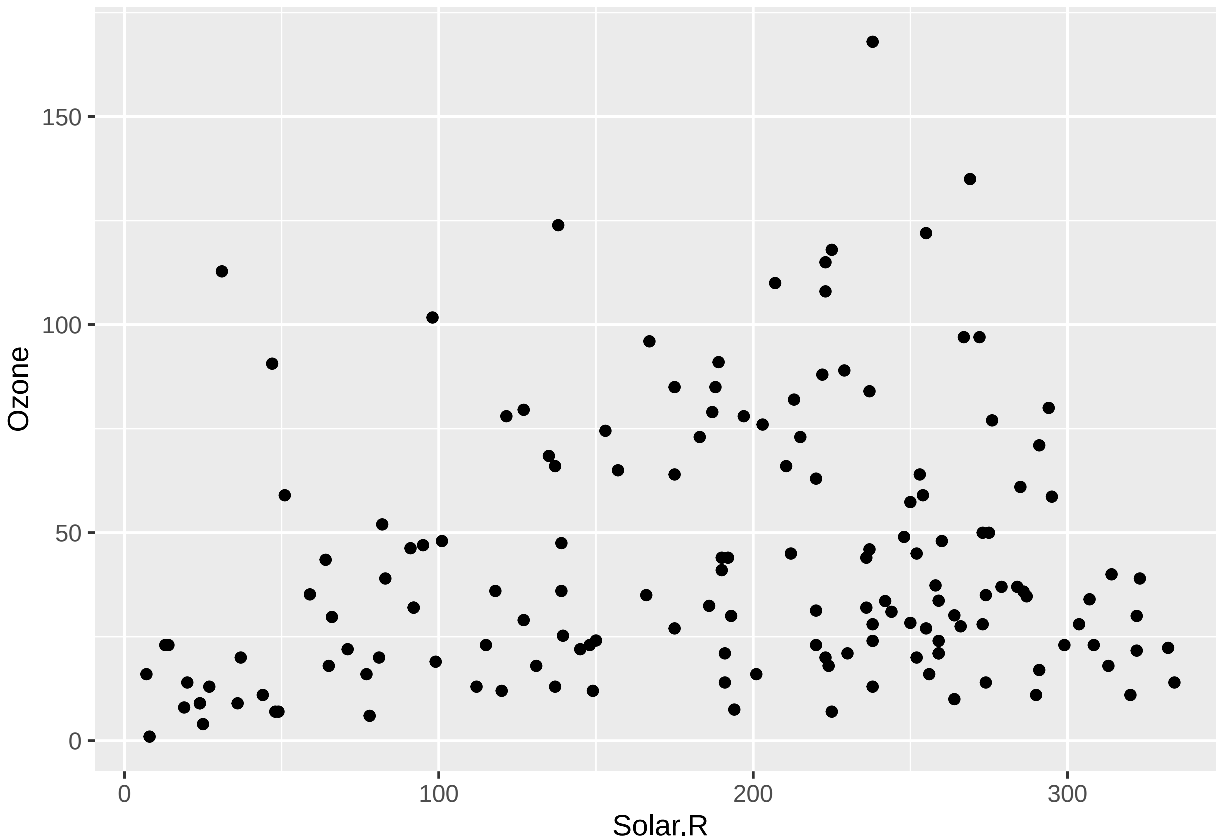
```
air$Solar.R <- na_interpolation(air$Solar.R)
anyNA(air$Solar.R)
```

```
## [1] FALSE
```

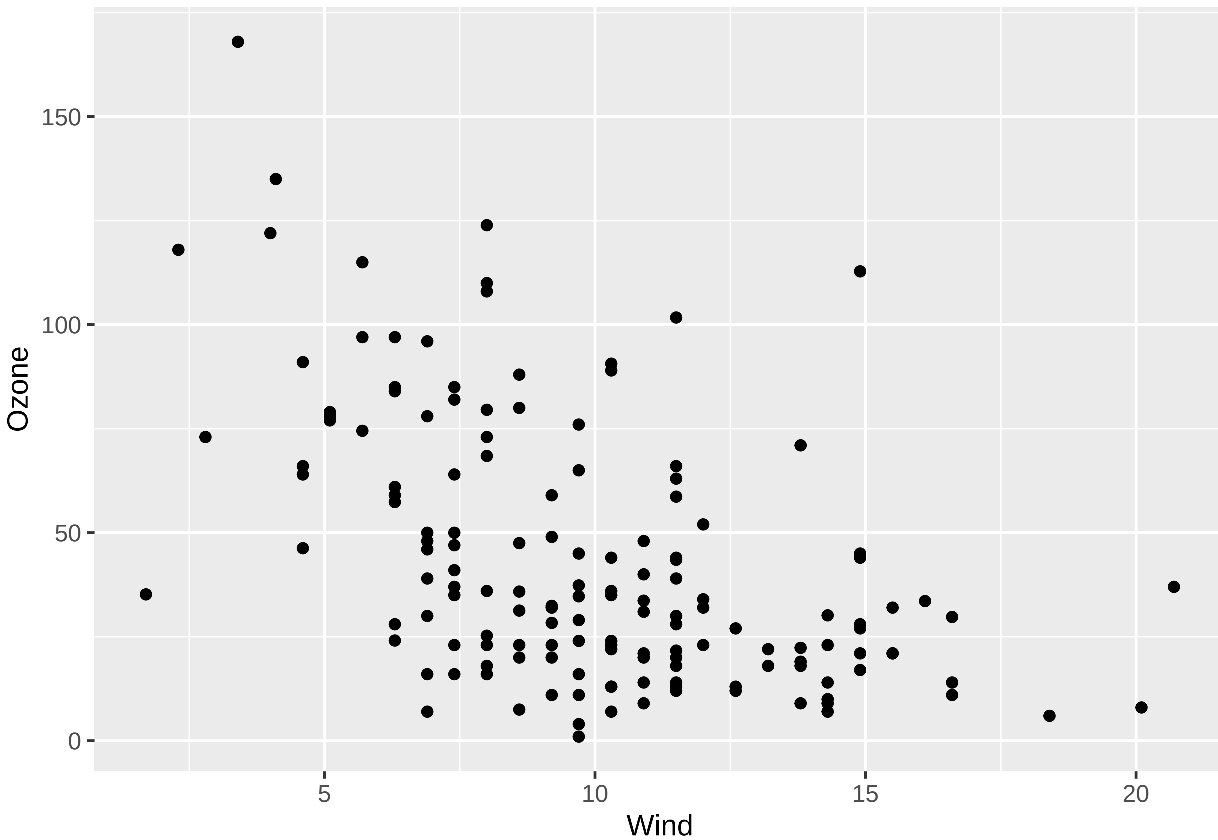
```
#Now there are no missing values in Ozone or Solar.R
```

E. Create **3 bivariate scatterplots (X-Y) plots** (using ggplot), for each of the predictors with the outcome. **Hint:** In each case, put **Ozone on the Y-axis**, and a **predictor on the X-axis**. Add a comment to each, describing the plot and explaining whether there appears to be a **linear relationship** between the outcome variable and the respective predictor.

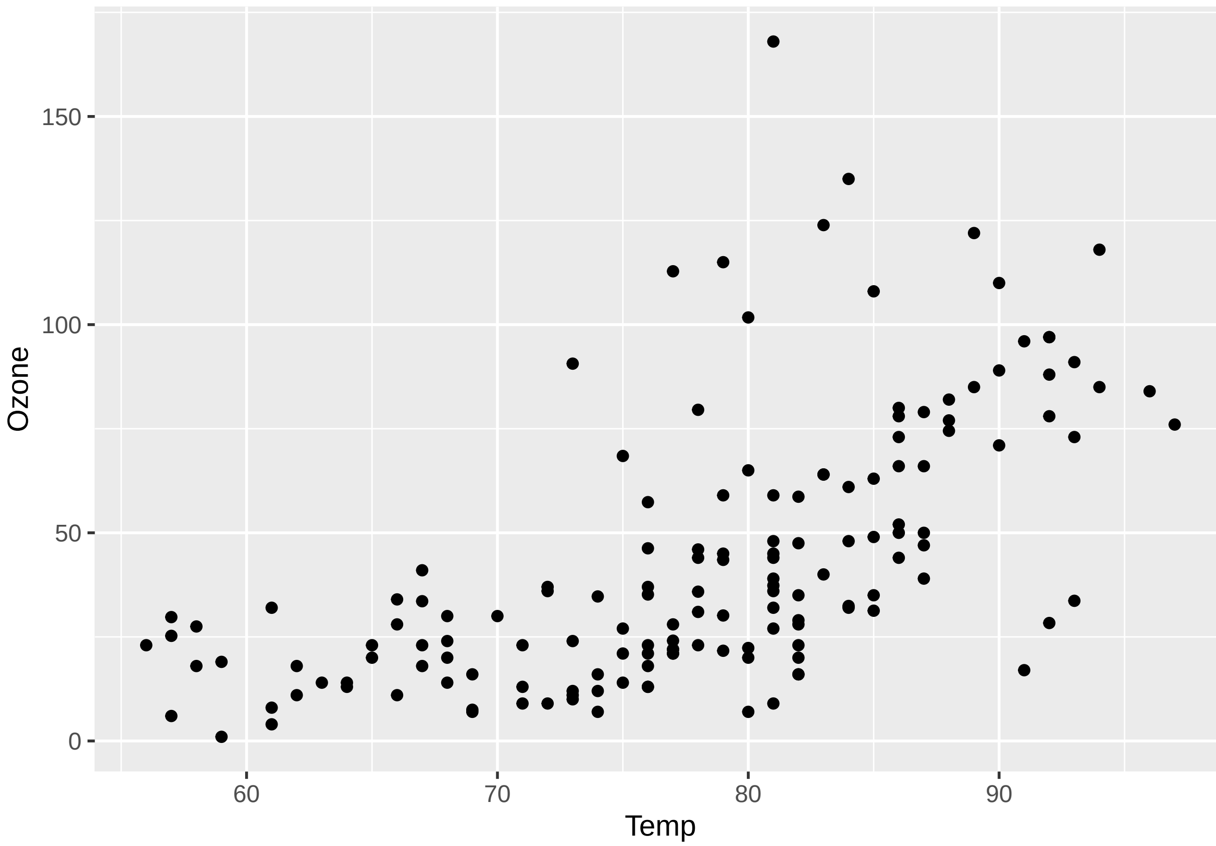
```
library(ggplot2)
ggplot(air, aes(x=Solar.R, y=Ozone)) + geom_point()
```




```
#scatterplot 1: Solar.R by Ozone. It doesn't look like the change in Solar.R influences Ozone, so there
ggplot(air, aes(x=Wind, y=Ozone)) + geom_point()
```




```r
#scatterplot 2: Wind by Ozone. It looks like an increase in Wind does influence a decrease in Ozone, so
ggplot(air, aes(x=Temp, y=Ozone)) + geom_point()
```




```
#scatterplot 3: Temp by Ozone. It does appear as though an increase in Temp causes a slight decrease in
```

F. Next, create a **simple regression model** predicting **Ozone based on Wind**, using the **lm( )** command. In a comment, report the **coefficient** (aka **slope** or **beta weight**) of **Wind** in the regression output and, **if it is statistically significant**, **interpret it** with respect to **Ozone**. Report the **adjusted R-squared** of the model and try to explain what it means.

```
OzonexWind <- lm(formula = Ozone ~ Wind, data=air)
summary(OzonexWind)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.332 -18.332  -4.155  14.163  94.594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.0205     6.6991  13.288  < 2e-16 ***
## Wind         -4.5925     0.6345  -7.238 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.56 on 151 degrees of freedom
## Multiple R-squared:  0.2576, Adjusted R-squared:  0.2527
## F-statistic: 52.39 on 1 and 151 DF,  p-value: 2.148e-11
```

```
#The coefficient of Wind is -4.59.
#The p-value of the coefficient of wind is 2.15e-11, which is statistically significant, indicates that
#The adjusted R-squared is 0.25, which means that the change in wind accounts for 25% of the variation
```

G. Create a **multiple regression model** predicting **Ozone** based on **Solar.R**, **Wind**, and **Temp**. **Make sure to include all three predictors in one model – NOT three different models each with one predictor.**

```
OzoneModel <- lm(formula=Ozone ~ Solar.R + Wind + Temp, data=air)
```

H. Report the **adjusted R-Squared** in a comment – how does it compare to the adjusted R-squared from Step F? Is this better or worse? Which of the predictors are **statistically significant** in the model? In a comment, report the coefficient of each predictor that is statistically significant. Do not report the coefficients for predictors that are not significant.

```
summary(OzoneModel)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.651 -15.622  -4.981  12.422 101.411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.16596   21.90933  -2.381   0.0185 *
## Solar.R       0.01654    0.02272   0.728   0.4678
## Wind         -2.69669    0.63085  -4.275 3.40e-05 ***
## Temp          1.53072    0.24115   6.348 2.49e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.26 on 149 degrees of freedom
## Multiple R-squared:  0.4321, Adjusted R-squared:  0.4207
## F-statistic: 37.79 on 3 and 149 DF,  p-value: < 2.2e-16
```

```
#The adjusted R-squared for the multiple regression is 0.42, or 42%. This is better than the adjusted R
#The predictors that are statistically significant are Wind and Temp. The coefficient of Wind is -2.69,
```

I. Create a one-row data frame like this:

```
predDF <- data.frame(Solar.R=290, Wind=13, Temp=61)
```

and use it with the **predict( )** function to predict the **expected value of Ozone**:

```
test=data.frame(predDF)
predict(OzoneModel, test, type="response")
```

```
##       1
## 10.9464
```

```
#The expected value of ozone is 10.95
```

J. Create an additional **multiple regression model**, with **Temp** as the **outcome variable**, and the other **3 variables** as the **predictors**.

Review the quality of the model by commenting on its **adjusted R-Squared**.

```r
TempModel <- lm(formula=Temp ~ Ozone + Solar.R + Wind, data=air)
summary(TempModel)
```

```
##
## Call:
## lm(formula = Temp ~ Ozone + Solar.R + Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.831  -4.802   1.174   4.880  18.004
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.693222   2.796787  26.707  < 2e-16 ***
## Ozone        0.139055   0.021907   6.348 2.49e-09 ***
## Solar.R      0.015751   0.006737   2.338  0.02072 *
## Wind        -0.580176   0.195774  -2.963  0.00354 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.313 on 149 degrees of freedom
## Multiple R-squared:  0.4148, Adjusted R-squared:  0.403
## F-statistic: 35.21 on 3 and 149 DF,  p-value: < 2.2e-16
```

```r
#The adjusted R-squared of this model is 40%. This means that the changes in the predictors account for
```