In Class Exploration 2

20 Pts

Read the prompt for the problem on Page 443 in the text entitled *Investigating Predicting Wine Prices*. The data can be downloaded from:

```
Grapes <- read.csv("https://raw.githubusercontent.com/nick3703/MA376/master/Wine.csv")
```

1. Identify the observational/experimental units
2. What is the response variable? Create a visualization that allows you to explore the response. Describe the distribution of the response variable.
3. Take a look at the data. What do you notice about the price index of the 1961 vintage? In context of the study does this make sense? Why or why not?
4. Write out, using proper statistical notation, both the statistical as well as the fitted (predicted) model for the single means model. Note, to find the fitted model you will need to find the mean and the standard error of the residuals.
5. It is common belief that wines tend to get better with age, and so maybe older wines tend to be more expensive than younger wines. Create a scatterplot comparing price to age. Describe the association. Are there any potential outliers?
6. Let's now consider the univariate regression using age to explain Price Index from Age. Write out the statistical model and give the fitted model.
7. What linear regression assumption is violated? Why might taking the Log of your response variable be appropriate here?

To fix this we will fit the model:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma) \\
i &= \text{Wine} \\
y_i &= \text{Log of Price Index for Wine } i \\
x_i &= \text{Age of Wine } i
\end{aligned}
$$

The R Code to fit the model is given below:

```
log_model <- lm(LogPriceIndex~Age,data=Grapes)
summary(log_model)
```

```
##
## Call:
## lm(formula = LogPriceIndex ~ Age, data = Grapes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8545 -0.4788 -0.0718  0.4562  1.2457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.02520    0.24723  -8.192 1.52e-08 ***
## Age          0.03543    0.01366   2.593   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5745 on 25 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.1804
## F-statistic: 6.725 on 1 and 25 DF,  p-value: 0.01567
```

8. Assuming the validity conditions are met, interpret the P-value for Age in context of the problem.
9. Explain, in context of the problem, what it would mean for there to be an interaction between age and summer temperature.
10. Now, fit a model with log(price index) as the response, and age, summer temperature, harvest rain, and winter rain as predictors. Report the prediction equation.
11. Is this model **better** than the model only using Age as a predictor. How are you deciding?