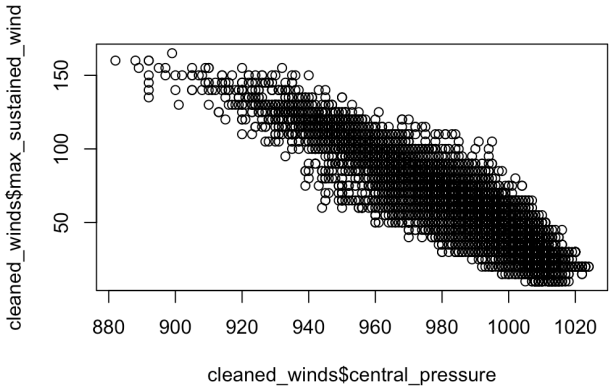
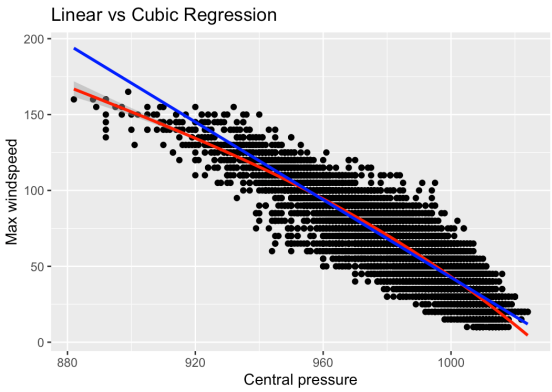
Glenn Turner
AIT580-M3-L2

Part 2 Steps:
Using your work from M3 Lab Part 1, produce and interpret a scatterplot and regression line & equation for wind vs pressure and respond to the following:

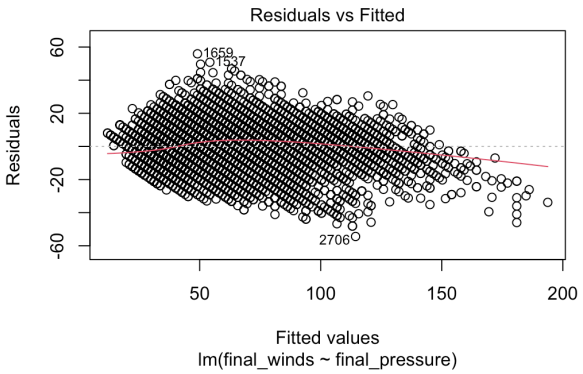
1. Does the data show any changes over time for the number or intensity of hurricanes?
    - Support your answer with a data summary.
2. Explain how you accounted for missing data in the dataset

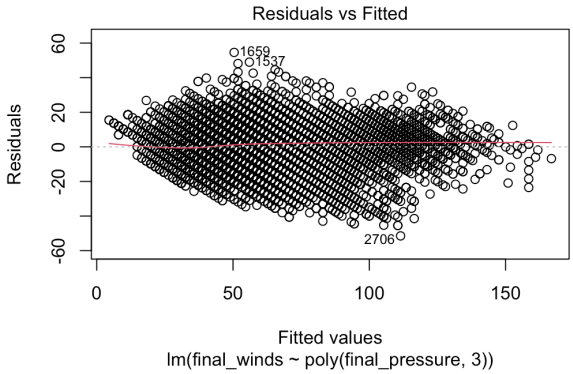
I first cleaned the data by deleting the -99 wind data points using subset. Then I plotted the data on a scatterplot.



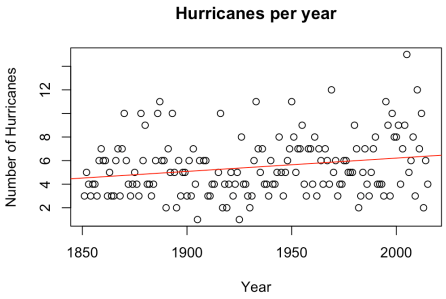


Then, I wanted to plot both a linear and a polynomial regression line using ggplot2. I found that using a cubic model gave me the best results without overfitting the data. In, particular, you'll see that the residuals plot is much better with the cubic model than the linear model.





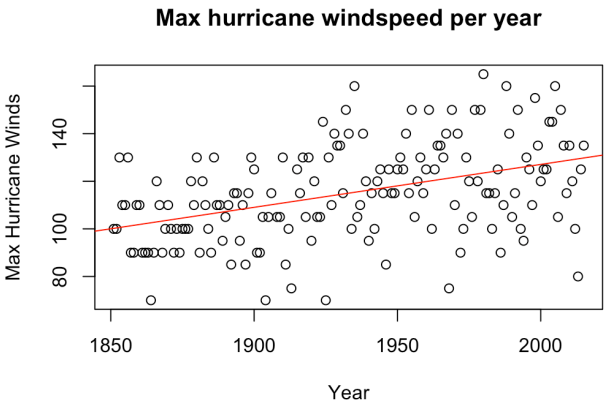
The data does show evidence that the number of hurricanes per year is increasing:




```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -16.402594   7.659411  -2.141  0.03374 *
HU_per_year$year   0.011309   0.003961   2.855  0.00487 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.42 on 161 degrees of freedom
Multiple R-squared:  0.0482,    Adjusted R-squared:  0.04228
F-statistic: 8.152 on 1 and 161 DF,  p-value: 0.004867
```

In fact, in the summary of the regression you can see that our p-value is .004, so we can reject the null hypothesis that the number of hurricanes per year is constant. While the relationship isn't very consistent, it is still statistically significant. The case for stronger hurricanes is a bit more clear cut. By graphing the maximum wind speed recorded in a hurricane each year, we can see a significant increase in wind speed over the dataset. In other words, the strongest hurricane of the year is getting even stronger over time. This trend is both more consistent and more significant than the number of hurricanes per year.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -233.63652   57.89120  -4.036 8.39e-05 ***
year           0.18037    0.02994   6.025 1.11e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.29 on 161 degrees of freedom
Multiple R-squared:  0.184,     Adjusted R-squared:  0.1789
F-statistic:  36.3 on 1 and 161 DF,  p-value: 1.111e-08
```

In accounting for missing data, I employed a couple of methods. For the wind speed data points of -99, those are clearly a measurement error. I deleted those few rows from the data. The missing pressure data was also interesting. For the purposes of this analysis, I could not impute the pressure data using wind data. This is because the goal was to make a regression and discuss the relationship. Therefore, using the wind data to impute would only reinforce whatever relationship I had already found. Regardless, I was able to use a full 18,000 data points to create my models, which is more than enough. If I needed a historical record of pressure data, I would definitely impute that data using the wind speed given their strong correlation, but for the purposes of this analysis, that was not necessary or helpful.