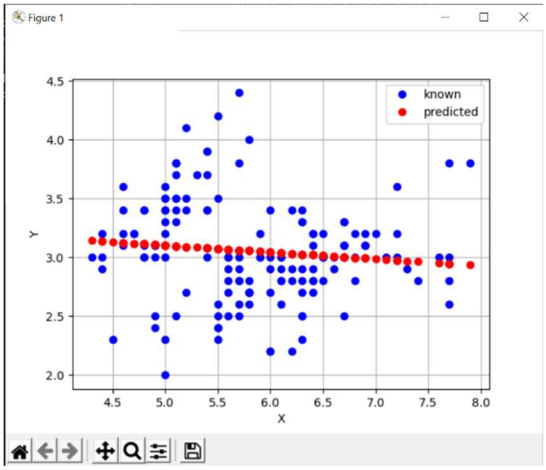
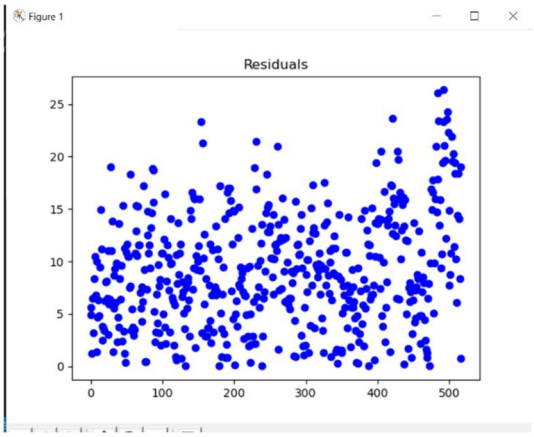
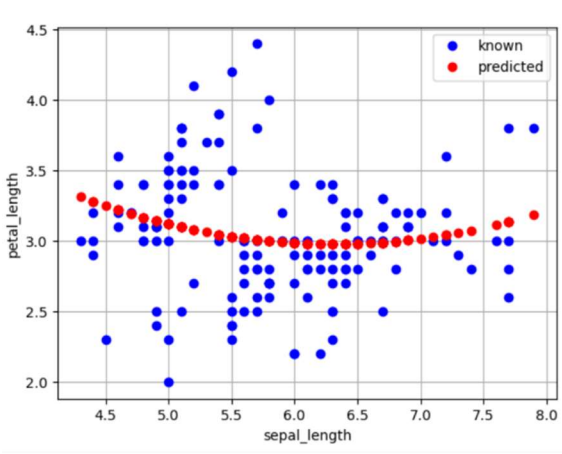
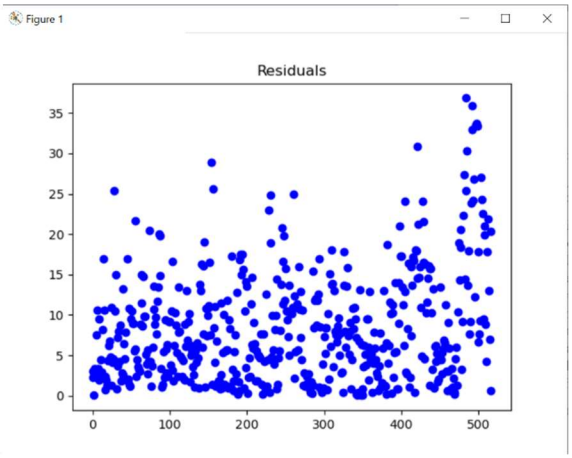
Part I: Iris





Simple Linear Regression:
Y = -0.05726823x + 3.38863738
Mean residual: 2.789333333333335
RSS: 1281.2624764359466
RSq: -11.540700573588813





Quadratic Linear Regression:
Y = $0.08326623x^2$ -1.05179644+ 6.30019539
Mean residual: 2.789333333333806
RSS: 1281.9038088642162
RSq: -11.546977786961543

Part II

Introduction

For my initial dataset, I decided to use UCI Machine Learning's forest fires dataset, which is a regression dataset to predict the area burned by forest fires. I took an interest in this dataset because I have relatives who live in areas affected by wildfires. Where they lived, they told me that ash was falling from the sky, that it wasn't safe to go outside, that they knew lots of people who needed to evacuate because the fires spread so quickly. This dataset uses forest fires from Portugal and data including the location of the fire, the temperature during the fires, and the rain during that time. The documentation of this data noted that it is a multiple linear regression task and that the area burned by fires relies on a lot of features but I am going to see how much of this dataset I can determine based on area and one other variable. Overall, I found that temperature is not sufficient to predict the area burned by fires and that in order to get a decent model of the fires, I likely need to take more variables into account.

Preprocessing and Minor Issues

Preprocessing the data included changing the categorical variables e.g. day and converting them to integers but otherwise all of the columns were okay and the data was pretty easy to use.. This dataset has 13 columns so to test which were the most relevant columns for my analysis, I used a covariance matrix that I normalized. There was an issue with my normalization code where it graphed my outliers as -1000 so I decided to use numpy's built-in functions.

Another bug I ran into during my analysis was that when I tried to graph a quadratic line fitted to my plot, I ended up with a squiggly line that went through every point as if it were a single point. I kept my line plot but plotted it as a red circle instead of as a line. I did not have time to fix this bug but I think it's an error with how numpy is calculating my points. It worked on my smaller test array that I had used to develop my line so I think the line is correct but I'm not sure what's going on with the line not being plotted as a line.

Part I: Plotting the Covariance Matrix

The first thing I will try to do is to plot the covariance matrix to see if any of the features are promising for linear regression. Unfortunately, none of the features really had a significant effect alone on area. I recall that in the iris covariance dataset, all of the iris features had a high correlation with themselves, which makes sense because there should be a one-to-one correlation between iris's data and themselves.

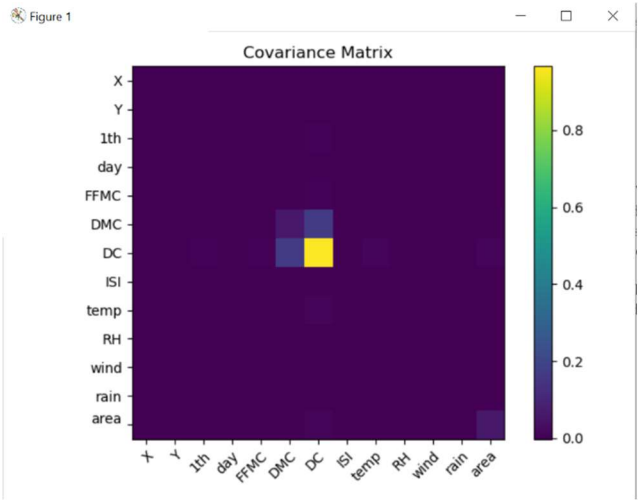


Figure 1: Covariance matrix of the forest fires dataset.

What I thought was most interesting was that even the features that should have one-to-one correlation, such as plotting area by area, don't have any covariance. Area appears to have 0 covariance for all its features. I think this could be because there isn't a linear trend for the dataset. Covariance can only predict linear trends so it is possible there could be a strong polynomial model for these

features and you wouldn't be able to see that based solely on covariance.

Based on my initial perceptions of this dataset, I assumed that there would likely be a correlation based on temperature and area but it actually seems to match the covariance matrix. Based on the graph, it looks like there is no correlation. There are a lot of outliers in the area. Most of the fires appear to be small or relatively small fires. They only burned around 0 to 100 acres and it looks like based on those we could recreate an effective trend line. However, it looks like there are at least two high leverage outliers. This is a point that lies so far outside the trend of our data that it tilts the data towards that outlier even though most of the data follows a different trend. If it were just these two points, then we could possibly ignore them or weight them less in our analysis. But it looks like there are a lot of moderately large fires, particularly around 15 to 30 degrees, that burned more acres than the average fire along this path and don't lie in the general trend. It looks like this is not a good model for these points.

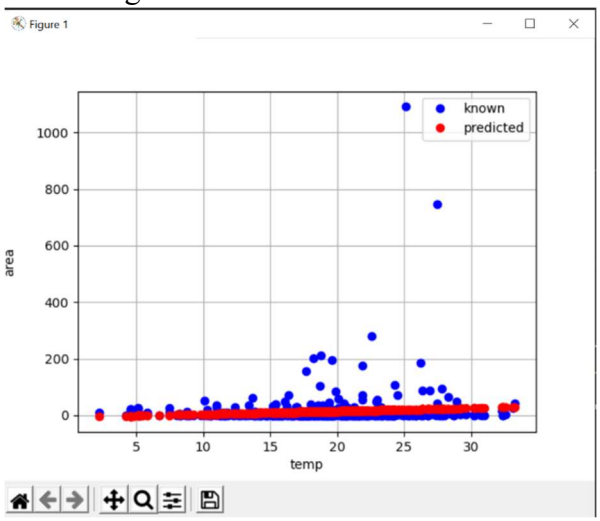


Figure 2: Simple linear model: Y=1.07284525x - 7.41602487

This is confirmed by my plotting of the residuals in this dataset. For clarity, I am plotting the absolute value of my residuals. If this is a good model, most of the residuals will hover around 0 with random differences from the means. There should be no trends and no correlation.

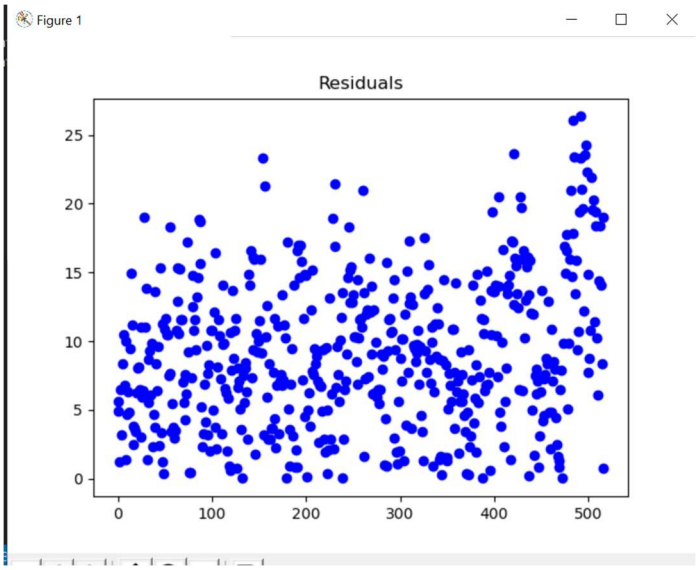


Figure 3: Residual plot of the simple linear model.

Residual Statistics	
Mean residual	-8.17988394584138
RSS	58142.597862653296
RSq:	-20.047543726914185

From my statistics, the average fire burns 8 fewer acres than the mean, there is a huge amount of variance in the dataset and RS, and line explains only 80% of our data. All of these statistics point to a bad model, which is further exemplified by the plot above.

[[0.05498481]
 [-0.89906319]
 [8.36273492]]

My next task was to plot my regression plot using polynomial regression, in this case a quadratic.

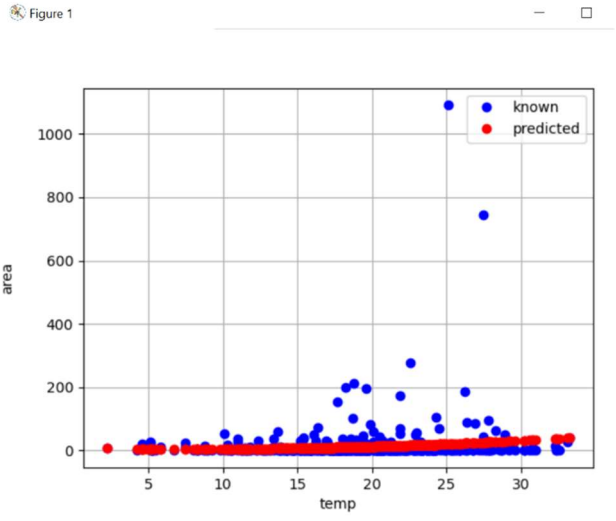


Figure 4: $Y=0.05498481x^2-0.89906319x+8.36273492$

Looking at this prediction, it doesn't look much improved from the simple linear regression model. We still have the same outliers that interfere with the simple analysis of the dataset. The curve of this line actually looks worse for the fires that burned between 0 and 100 acres and the outliers don't really match the trend of this dataset. Looking at the outliers, I am not sure that there necessarily is an equation that would be accurate for this point.

This is confirmed by the plot of my residuals and their means:

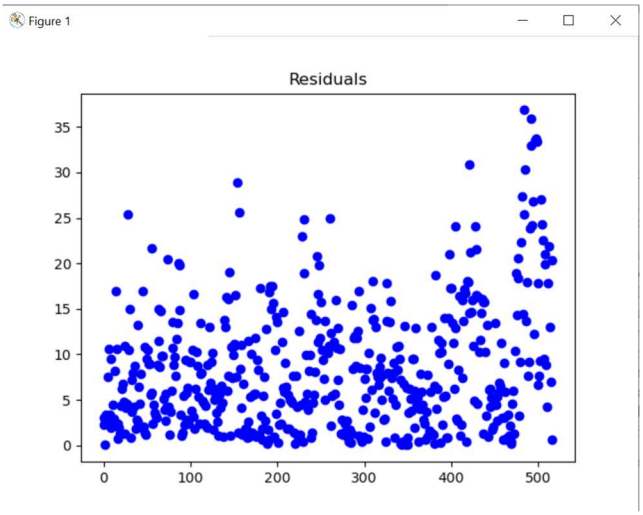


Figure 5: Residuals of the quadratic line.

Residual Statistics	
Mean residual	-8.179883945841857
RSS	61610.21869419466
RSq	-21.302817893586838

The mean residual is about the same but the RSS is much higher and R^2 is slightly higher, meaning that the data is a slightly worse fit. The quadratic model is not better than the linear model at predicting the trends, in fact it is slightly worse but neither one of them allows us to predict anything about the forest fires dataset.

Conclusion

I learned that some datasets are not well suited for single linear regression. Neither one of these models were an example of overfitting but I got the idea of what happens when you have a trend that doesn't really fit into a standard dataset. Further work includes writing multiple linear regression code and trying to figure out how to make an effective multiple linear regression analysis to predict the forest fires.