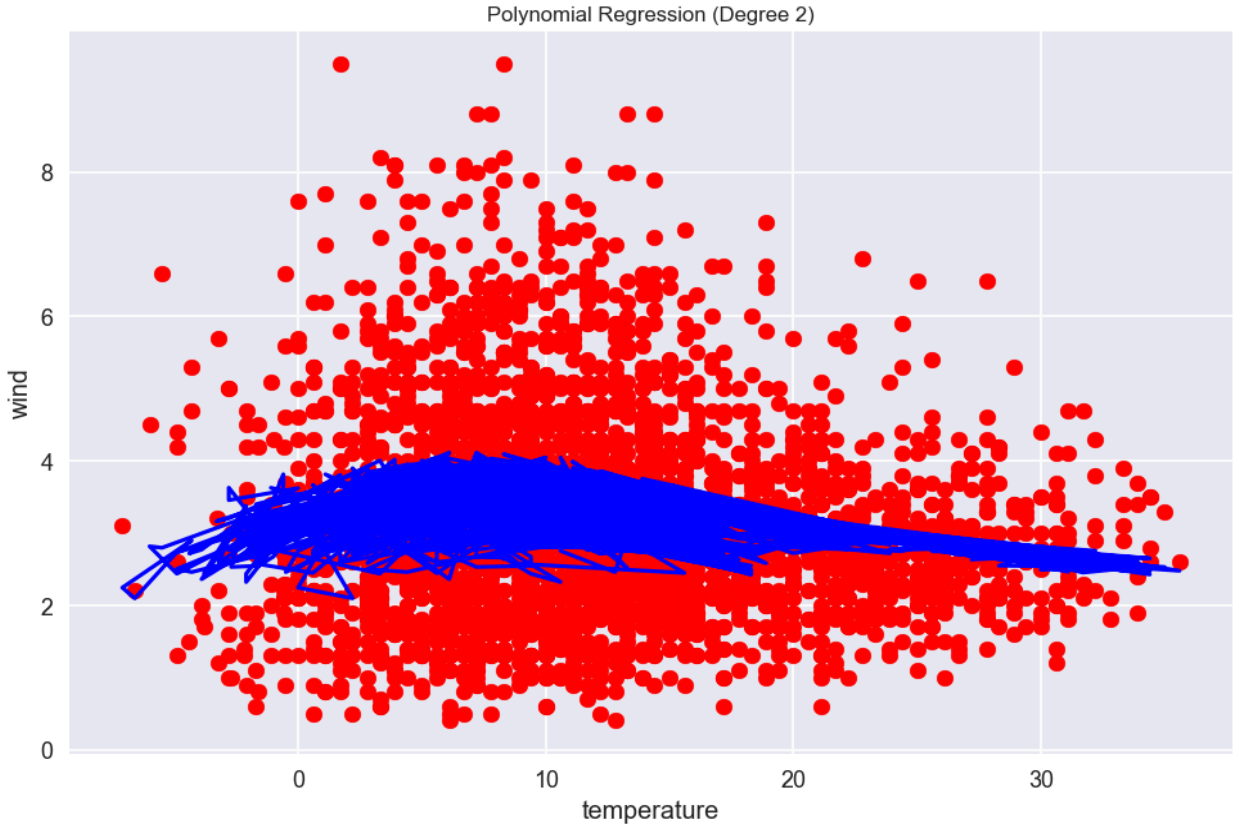
The data I am using is seattle_weather.csv which can be found here("https://www.kaggle.com/datasets/ananthr1/weather-prediction").

For data preparation I convert the weather type to an integer value for multiple linear regression. I also ignore the date value when using the dataset.

For model building I decided to use the maximum and minimum temperatures along with weather for my linear regression model to predict wind. Those three variables seemed the most relevant to affect wind. I then used polynomial regression with maximum and minimum temperatures for wind prediction, and found that there is a non linear relationship between temperature and wind. After that I found the best subset for the model, and discovered that using all of the variables(excluding date and wind) are best to predict the wind.

And for the logical regression I decided to use precipitation, maximum temperature, minimum temperature, and wind to predict weather. I did not use date because it does not seem to be as useful in predicting weather.

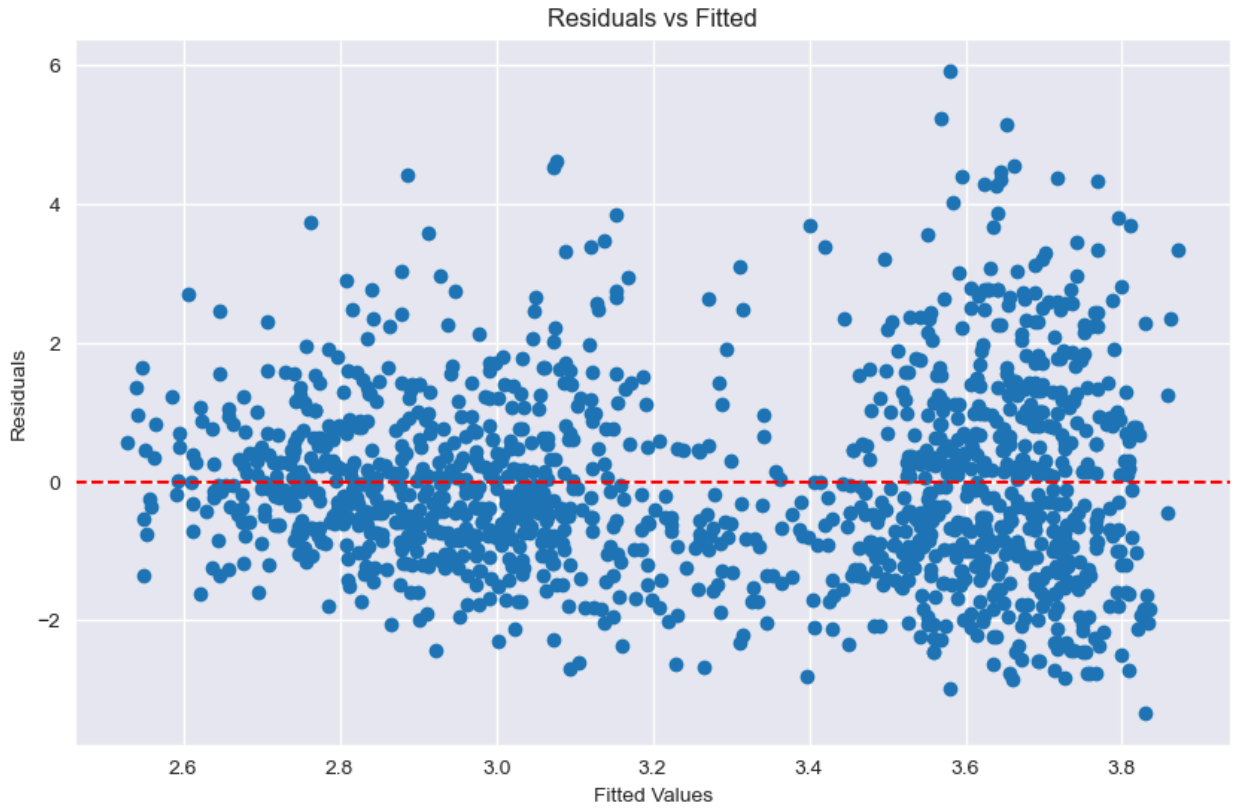
For my linear regression model I got a r2 score of around 0.09, and a Mean Square Error score of around 1.76. These results show very little dependence on the independent variable from the dependent variable, and a large error margin. I also made a polynomial regression model and made a graph from it which showed that there was a non-linear relationship between maximum/minimum temperature and wind. See below.



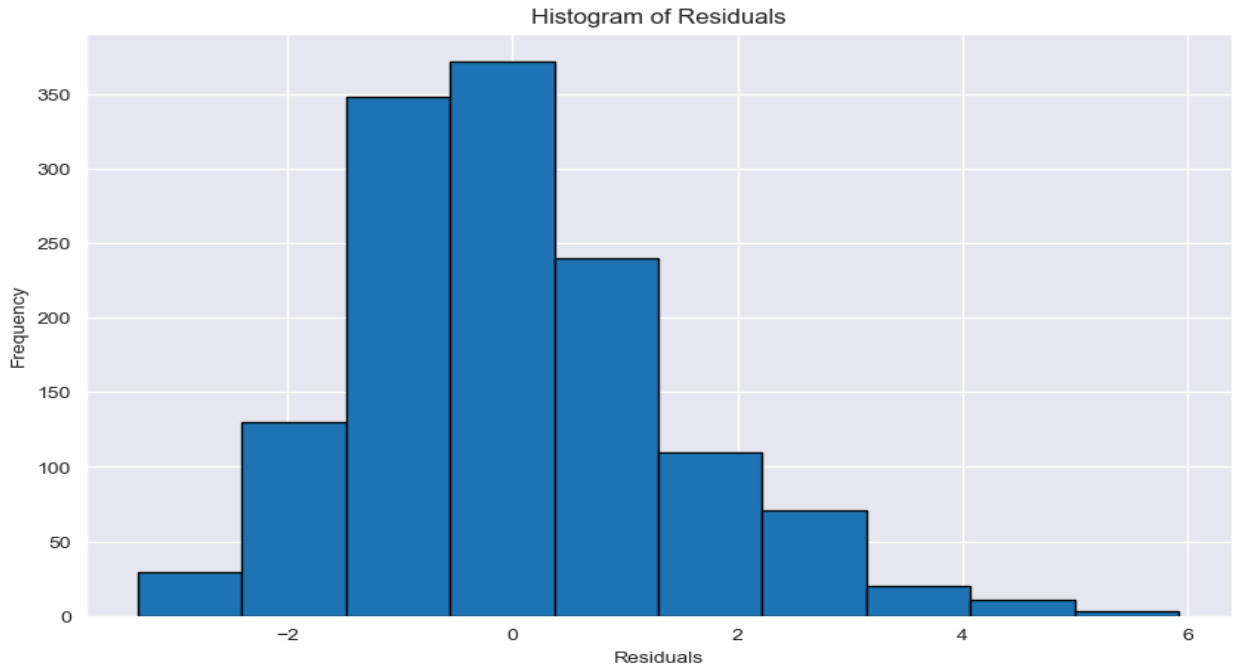
The blue lines represent the polynomial regression line, and you can see there is a bump around 8 for the temperature. The VIF(Variance Inflation Factor) scores are as follows.

	Variable	VIF
0	const	19.366013
1	precipitation	1.303752
2	temp_max	5.701006
3	temp_min	4.941618
4	wind	1.137178
5	weather	1.399480

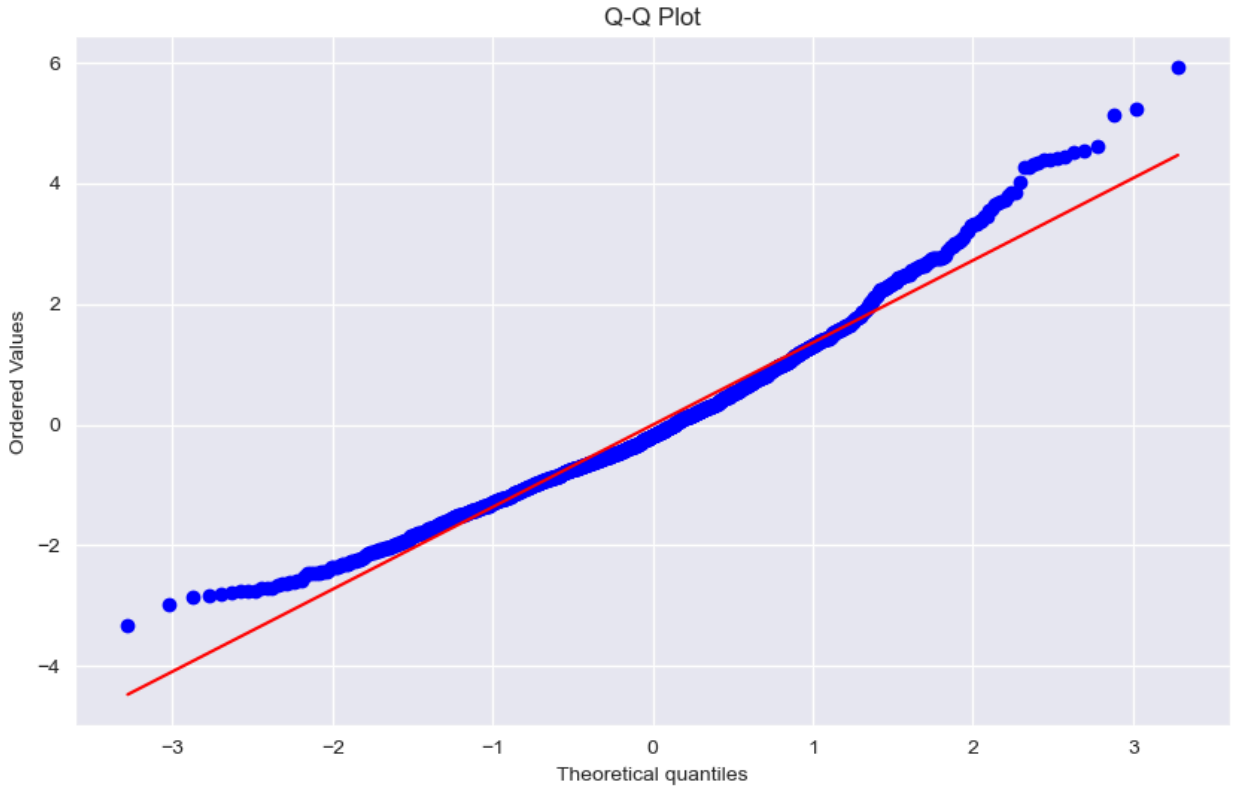
Here We can see that the VIF score of const(constant) is rather high. Which means there likely is a small variance for the other variables. After calculating the VIF scores I analyzed the residual values I made a residual vs fitted plot. see below.



The plot shows that the residuals are fairly scattered which is good, and other than noticing there might be slightly more values above 0 than below there is nothing to note. The next graph I made was a histogram of residuals which can be seen in the next page.



The histogram shows that 0 is the most common residual which is good, but the histogram is skewed, which shows that although the values do have some linear relationships they are not linear. The final graph I made is Q-Q plot



As you can see the plot has some resemblance to a linearity but it is not linear. Finally the coefficients of the different values are very low which

means if the data was linear the increases based off of the values would also be low

For the logistic regression; its accuracy is around 0.83, which is 83% which is fairly high. The coefficient and odds ratio are below.

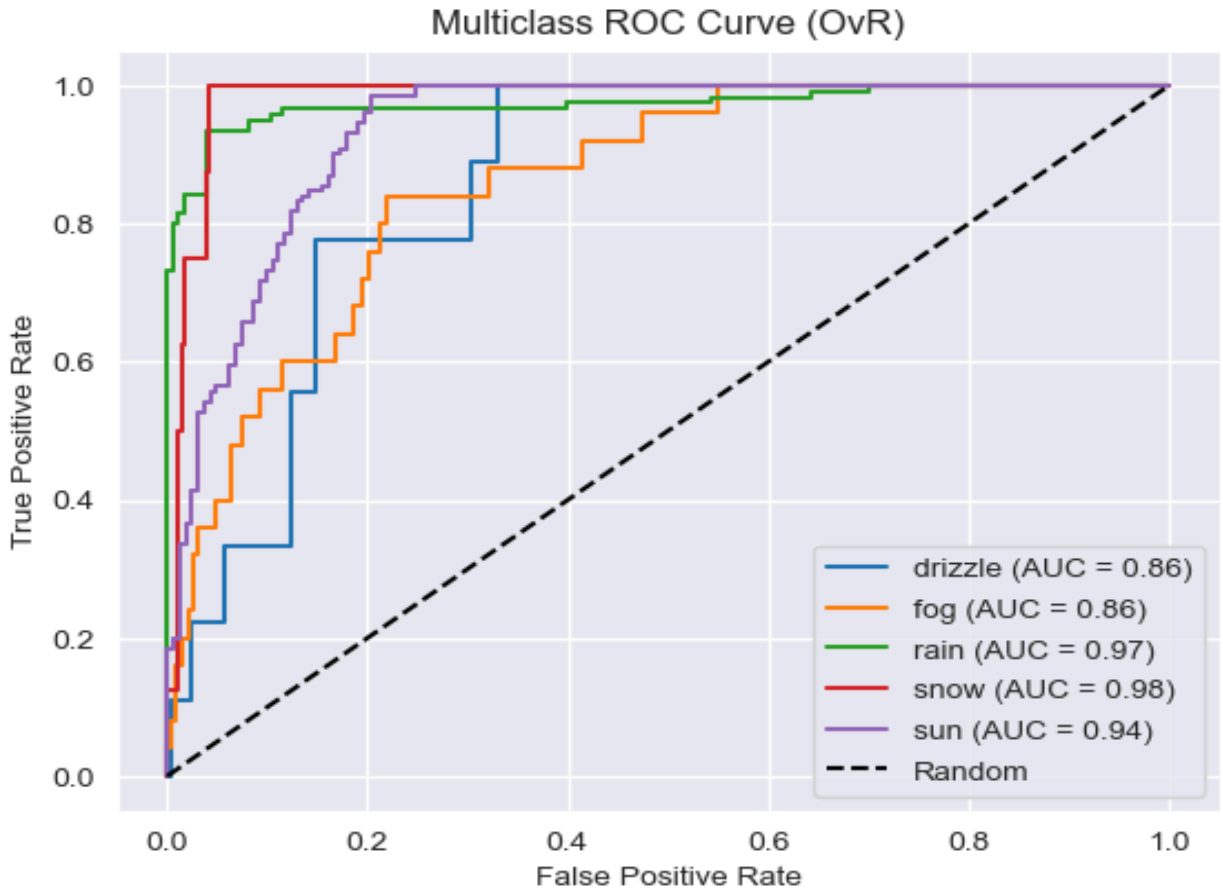
coefficients ratio

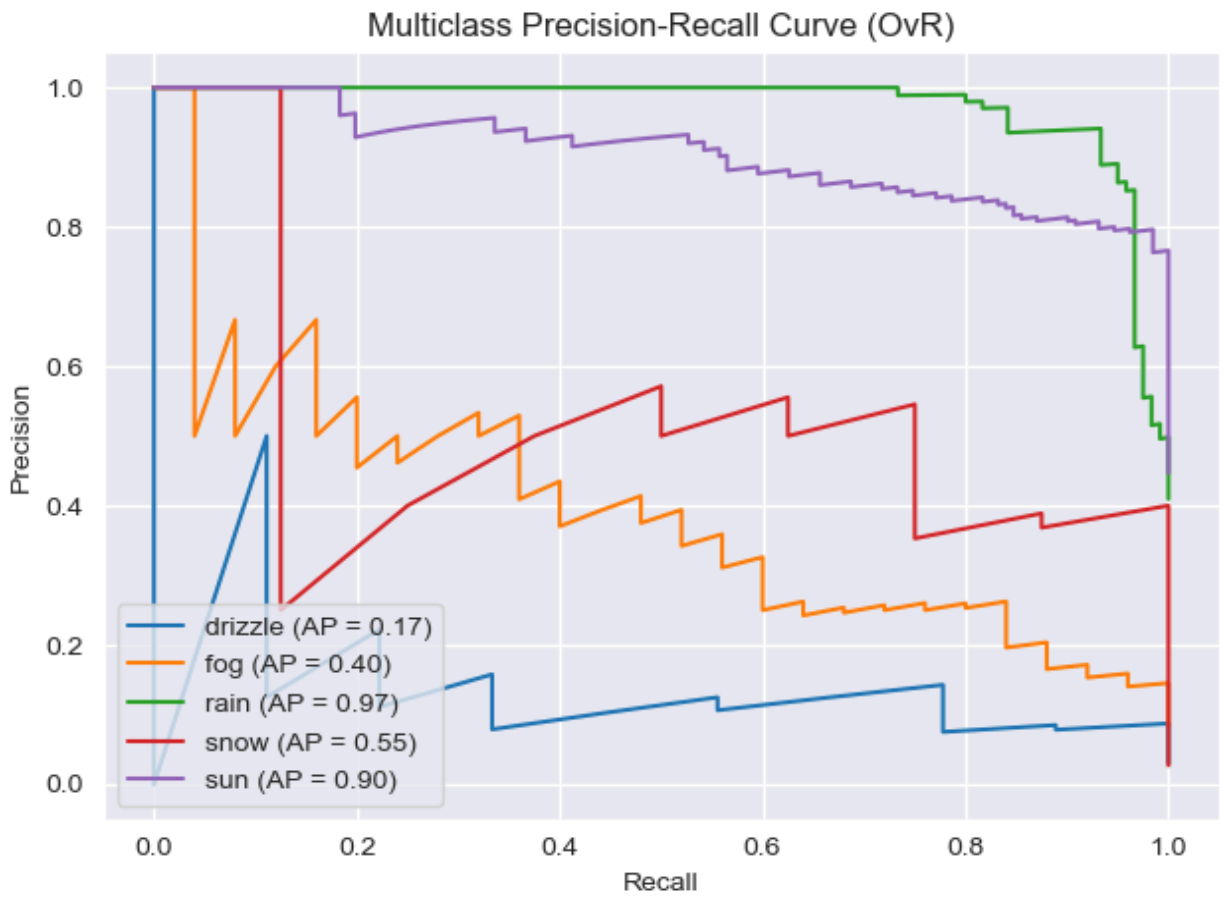
	precipitation	temp_max	temp_min	wind
drizzle	-1.437623	0.102792	-0.017748	-0.328235
rain	-2.145363	0.050149	0.078630	-0.183944
sun	4.066312	-0.046129	0.198024	0.056449
snow	4.142122	-0.279923	-0.208975	0.358095
fog	-4.625448	0.173111	-0.049931	0.097635

odds ratio

	precipitation	temp_max	temp_min	wind
drizzle	0.237492	1.108261	0.982408	0.720194
rain	0.117025	1.051428	1.081804	0.831982
sun	58.341398	0.954919	1.218992	1.058073
snow	62.936230	0.755842	0.811415	1.430601
fog	0.009799	1.188998	0.951295	1.102561

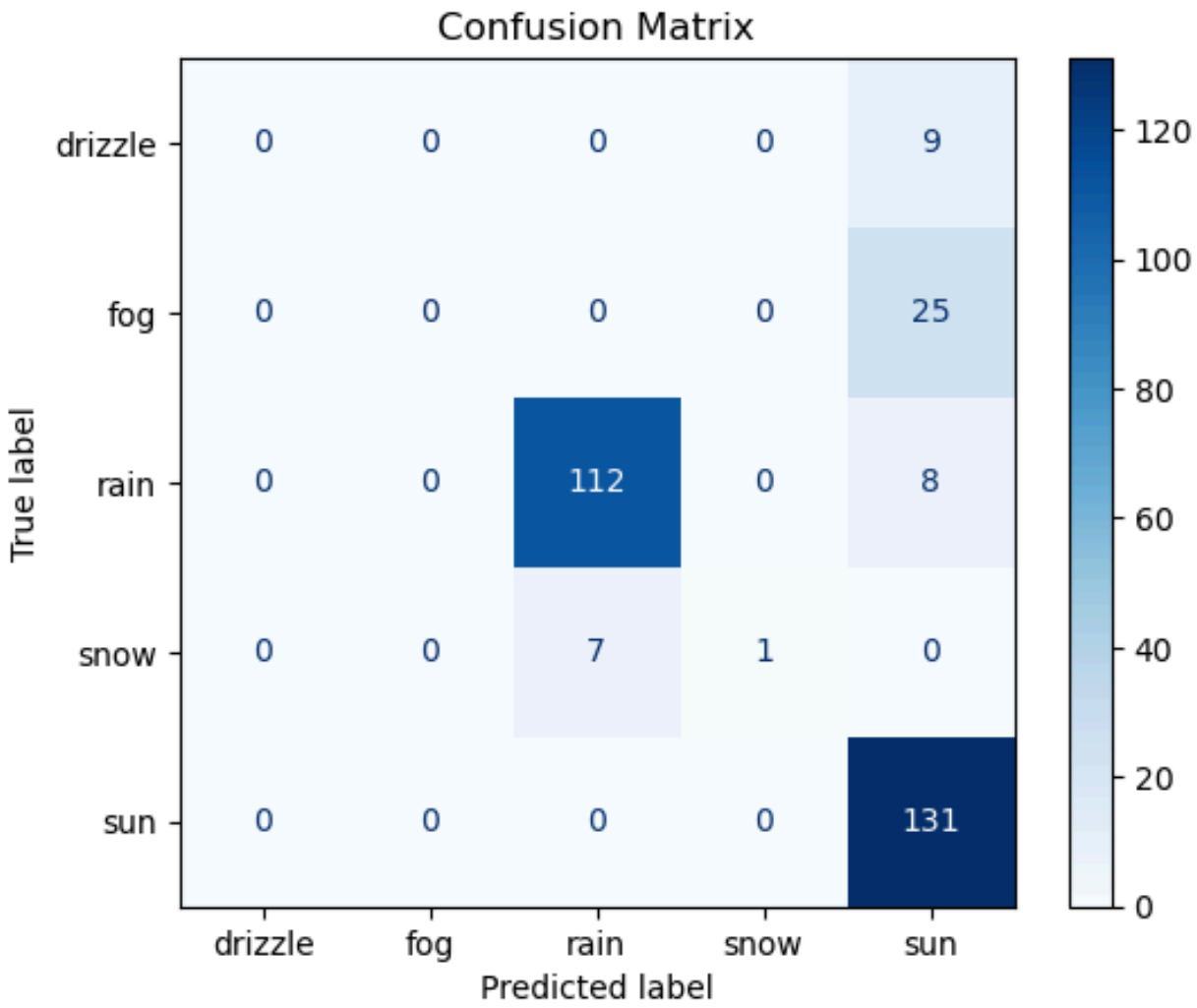
Most of the values are small with the notable exceptions of precipitation and sun/snow, which has a very strong correlation. The correlation between snow/sun and precipitation is very intriguing, unfortunately I do not know how to investigate it further. Next I graphed the ROC and Precision-Recall curve





The high AUC(Area Under Curve) represents a good model, but the second graph tells a different story. The second graph shows that the model is great at predicting rain and sun, but bad at the others, and especially bad at detecting drizzle. This could be due to environmental factors. For example there may be very little snow, fog, or drizzle in the state where the weather was recorded, meaning the model would be much better trained for predicting if it is going to rain or if it is going to be sunny.

Finally I made a confusion matrix. See the next page.



The confusion matrix shows that it detected rain when it rained a lot, and sun when it was sunny a lot, but did poorly at detecting fog, drizzle, and snow.

This project was very challenging to find a good dataset for and figure where the different types of regression models are used. I figured out that for linear regression numbers that have a large range are better for the linear regression model, while smaller ranges are better for the logical regression model. After figuring these things out, I slowly but surely figured out the rest.