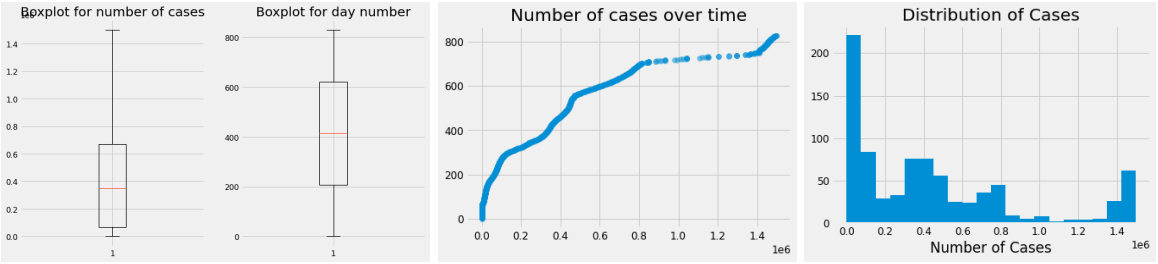
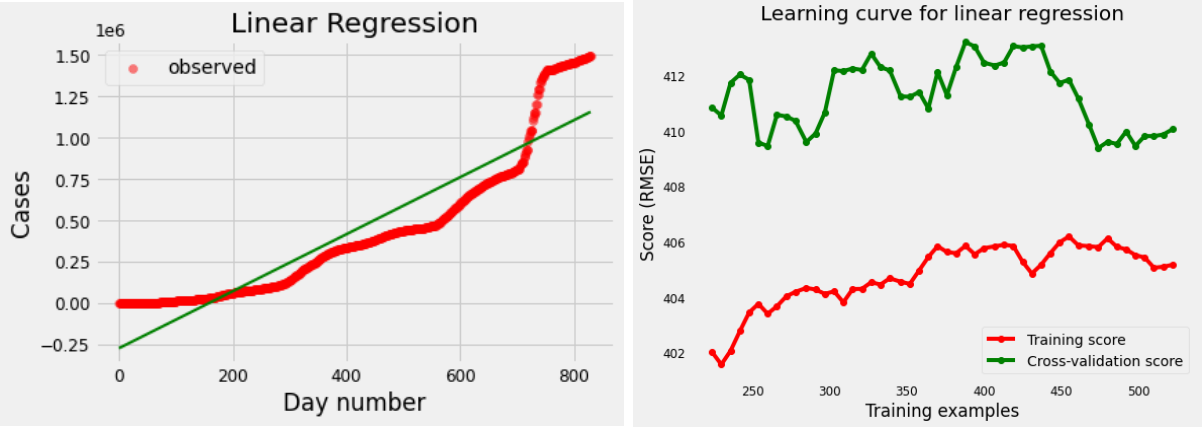
Mia Rodgers: https://github.com/miamrodgers/4310-ML/blob/main/ME5_simpleRegression_COVID19/simpleLinearRegression.ipynb
Alex Larsen: https://github.com/alarsen123/ML-HW/blob/main/ME5_simpleRegression_COVID19/ME5_simpleRegression_COVID19/simpleLinearRegression.ipynb
ME5

In this assignment, we explored COVID-19 data made available by the John Hopkins University Center for Systems Science and Engineering. Specifically, we were looking at data for number of cases per day in Washington with 830 observations. We plotted some graphs in the exploratory data analysis step to better understand our data.



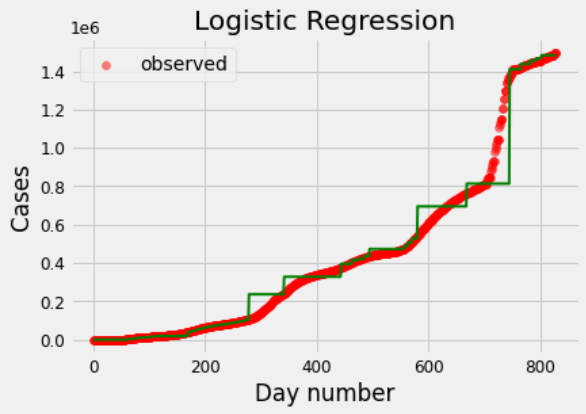
First, we trained a linear regression model, which had a training score of 0.854 and a testing score of 0.849.

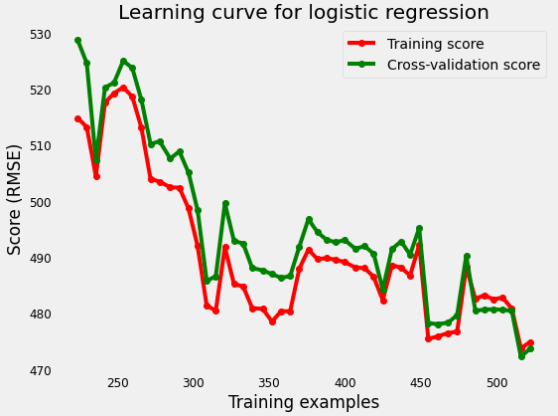
We then applied 10-fold cross-validation and got an average RMSE of 170929.528 and an average R2 of 0.848.



This makes sense since we can see a linear relationship, but the actual data seems to be more quadratic which is why the accuracy is good, but not great. The RMSE for the cross-validation is much better, which makes sense.

Then we tried applying logistic regression. The accuracy for the training set was 0.126 and for the test set, it was 0.056. This obviously is not very good. However, the visualization of the model looks somewhat decent. We again applied 10-fold cross-validation, but for this model, we kept getting warnings and ended up scaling the data as suggested. So, for the cross-validation, we got an average RMSE of 218743.59 and an average R2 of 0.736





Overall, it seems that linear regression is better for this problem.