Error Identification and removal

Values more than three standard deviations away from the mean were assumed to be errors or anomalies and removed from the dataset prior to analysis.

Irrelevant Data removal

The temperature and CO2 datasets were trimmed to the sates covered by the arctic Ice dataset. All results from the year 2023 were removed from analysis, because the arctic ice dataset used a different data type in 2023. Given most of the model is trained on Goddard data, it would perform poorly predicting NTSC results.

Blank values

Because the linear regression model averages monthly values into annual values and the data is seasonal, any missing monthly values could skew the mean (for example missing winter data would increase mean temp). As there are multiple features from different datasets in our regression model, it is uncommon for an error to occur across multiple features. This means data from the other features of the same time can be used to help calculate the interpolated value and also preserves the relationship between the different features within the model.