National Oceanic and Atmospheric Administration Weather Data Analysis

Abstract

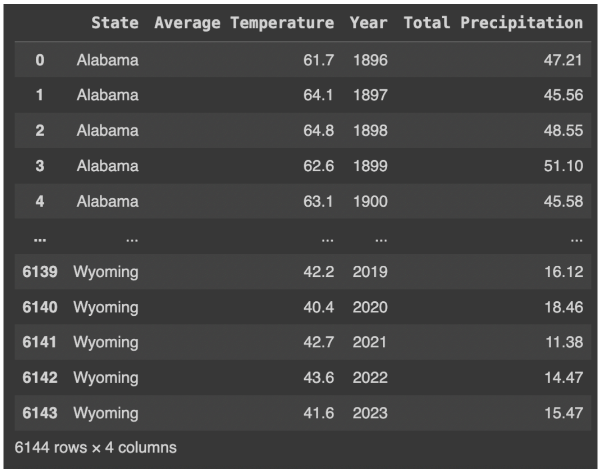
A few observations from the data analysis and machine learning model characterize the results of the United States weather analysis project. The first key finding is that states in the Southeast have seen the greatest increase in temperature from (1896 to 2023), but have seen a much lower rate of temperature increase compared to their Northern counterparts. To that end, the research also shows that no states have seen a decrease in average temperature over this period. Another important observation is that the Western United States has become dryer over this period than states east of Colorado, which have seen increases in total precipitation over this time. Considering both factors, the weather changes in the Northeast can be characterized as becoming warmer and wetter while the Southwest has become both dryer and warmer. The machine learning model focused on predicting average temperature and precipitation trends using polynomial regression. The model showed that temperature predictions were more accurate and consistent than those for precipitation, especially for coastal states.

Introduction

The National Oceanic and Atmospheric Administration (NOAA) is a United States Government agency that collects and analyzes weather data nationwide. The open-source data collected by NOAA is collected with a high accuracy standard so that it can be used in various scientific endeavors. This project leverages temperature and precipitation data for each state over the last one hundred and twenty-six years to identify trends in climate data and predict future climate conditions for specific locations. This project narrows the scope of the weather data by solely focusing on temperature and precipitation values at the State and National levels. This enables the capability to build purely numeric graphs and visualize trends geographically. This project aims to explore several key questions, including: how mean temperatures and precipitation totals are evolving and at what rate, determining patterns in specific regional climate changes, and identifying which states exhibit the most predictable and unpredictable weather patterns. The predictions developed through these models will be useful in many facets including the task of identifying where more public and climate-related infrastructure will be needed in the future and which places will see an influx of immigration from people looking for more optimal climate conditions.

Data Description

The particular data this project utilizes is from a subset of the NOAA website that enables a user to access climate data for a specific region within a specified time period. This means that to compare multiple locations this project concatenates tables from different locations. A typical chart generated by the website contains a date, a measured value, a ranking of that value compared to other values, and the anomaly or the measurement of the difference of the specific value from the mean. All the tables are scraped from a specific download link that generates the table as a page object based on the page inputs. The data this project gathered includes averages and totals from each calendar year to maintain a consistent time period across all state data. However, it is important to point out that the data is only for the contiguous United States because there was no consistent data on Hawaii and Alaska. After the data is scraped, it is stripped only to contain relevant information in the form of measurements and years. Each data frame is then mapped to the state it corresponds to and added to the full dataset. A small sample of the processed data is included below:



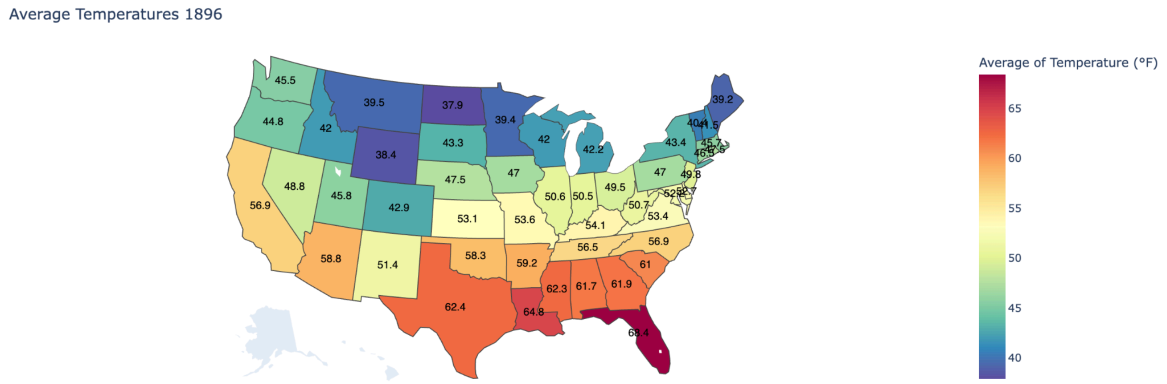
	State	Average Temperature	Year	Total Precipitation
0	Alabama	61.7	1896	47.21
1	Alabama	64.1	1897	45.56
2	Alabama	64.8	1898	48.55
3	Alabama	62.6	1899	51.10
4	Alabama	63.1	1900	45.58
...	...	...	...	...
6139	Wyoming	42.2	2019	16.12
6140	Wyoming	40.4	2020	18.46
6141	Wyoming	42.7	2021	11.38
6142	Wyoming	43.6	2022	14.47
6143	Wyoming	41.6	2023	15.47

6144 rows × 4 columns

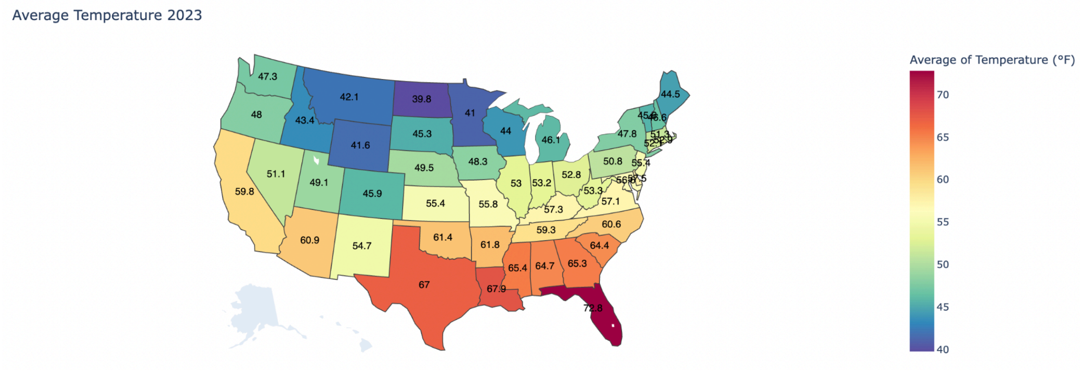
With the scraping, cleaning, and formatting of the data completed, the collected information is now capable of being used to show trends and changes in weather. For this, a range of tools and graphs help visualize how temperature and precipitation have changed over time. Some libraries that are used for these model predictions and visualizations of the collected data include Plotly, GeoPandas, and Sci-kit-learn. To utilize the data most effectively the data first needs to be quantified on its own to identify interesting trends and patterns, then used to develop predictions based on the findings.

Methods

In order to identify the trend characteristics of the data it made sense to examine geographic tendencies around the temperature and precipitation values. This meant grouping the data for the beginning and end years and trying to see any apparent patterns. The first graph generated shows changes in temperature over time on the United States map.

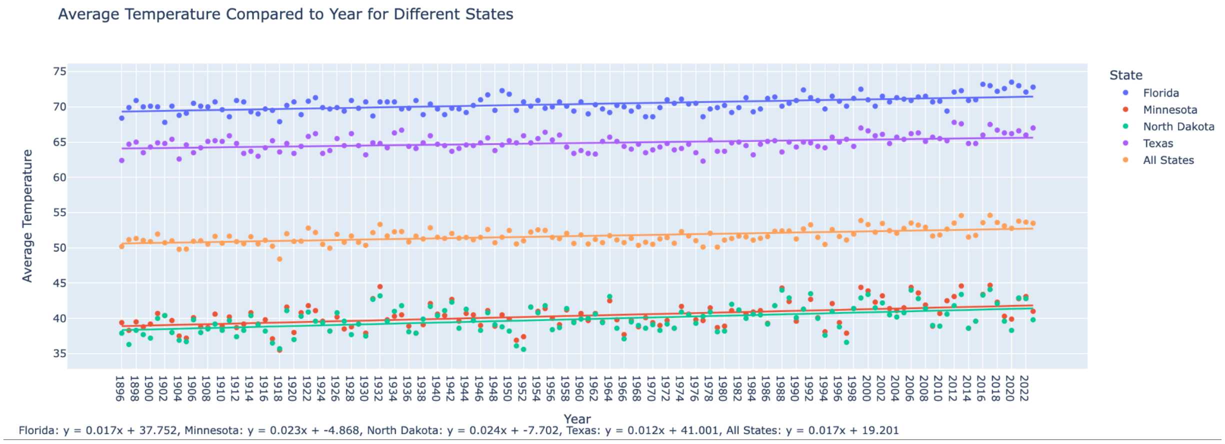


Average Temperatures 1896

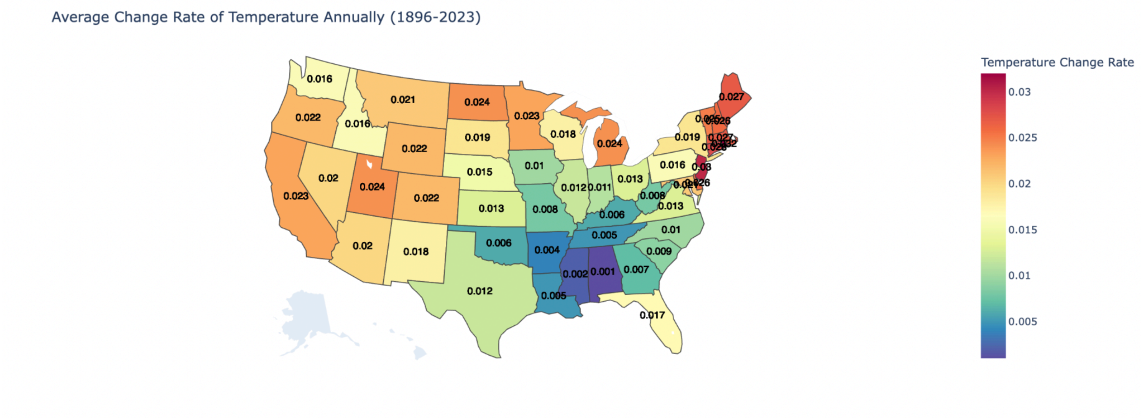


Average Temperature 2023

From these graphs, it is easy to see that mean temperatures today are around 3-5 degrees warmer than in 1896. There is also a much less noticeable change in average temperature for states in the upper midwest compared to other regions of the United States. With this in mind, it makes sense to compare temperature data from states in the upper midwest to states in the south to see if there was a noticeable change. This is the resulting graph:

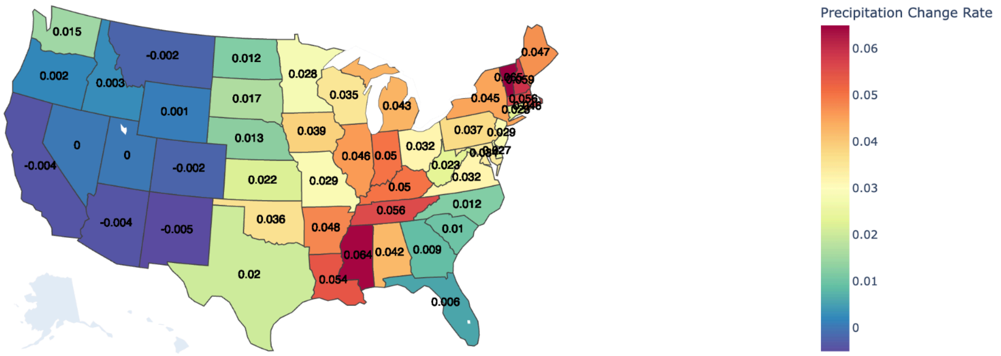


This plot shows a somewhat surprising result by revealing that the states in the Midwest had a much higher rate of temperature increase than the states in the South with mean temperature increases of 2.4% and 2.3% annually compared to the much lower national average of 1.7%. As a result of these findings, it is relevant to visualize where exactly the rate of temperature increase is larger in order to identify which states are likely facing the largest impacts of changing temperatures using the same least squares regression techniques as above applied to all states.

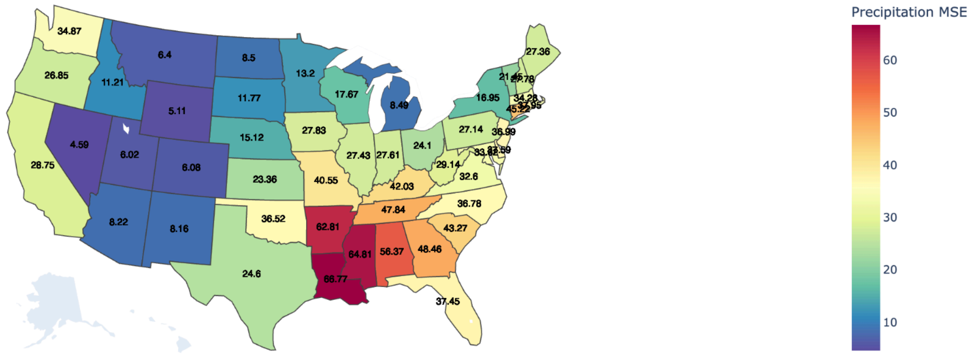


This graph shows that the northern areas of the United States, especially the Northeast, have had the largest rate of changing temperatures across the United States. This is also interesting because over 17% of the population lives in the upper Northeast meaning that these higher change rates affect many Americans. It is also interesting to see that Southern States like Mississippi and Alabama have seen the lowest rates of temperature increase in the last 128 years.

When it comes to analyzing changes in precipitation patterns it made sense to build off the previous techniques by first visualizing precipitation changes across the United States before leveraging these findings to identify areas with unfavorable weather changes. In order to do this it is applicable to utilize a least squares model to see which areas have experienced the largest and smallest changes in precipitation rate over the duration of the data collection.

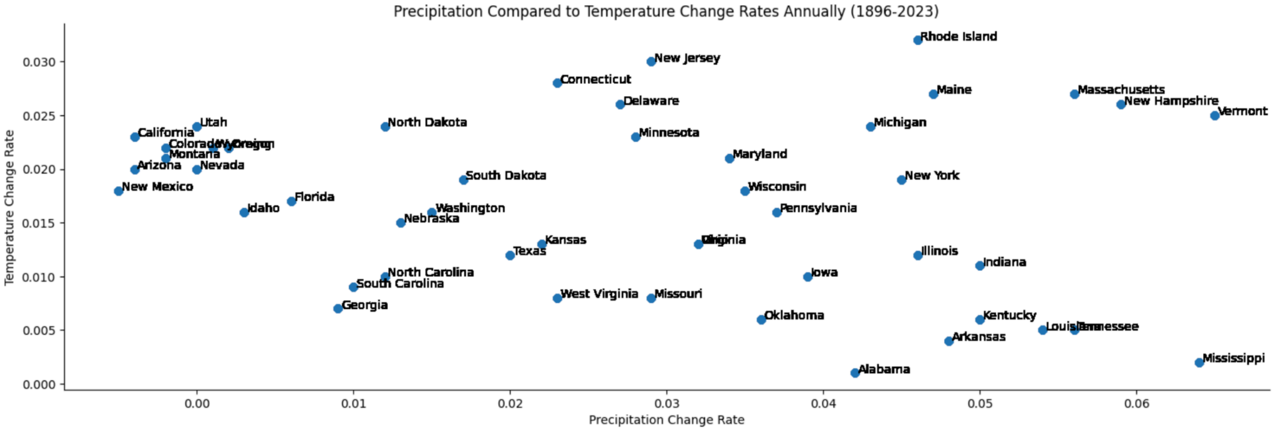


Average Change Rate in Precipitation Annually (1896-2023)



Mean Square Error of Precipitation Predictions by State (1896-2023)

This analysis shows that the Western half of the United States has typically become dryer whereas the eastern half has seen increases in precipitation. Looking at the mean square error plot shows that the linear model is typically better at predicting variance in precipitation for inland states as opposed to coastal states due to the larger mean square error values seen around the coasts compared to the interior. This also provides context as to which states have faced detrimental climate conditions in relation to both factors. Putting it all together visualizing both factors on a plot shows which areas have seen the largest change rates in relation to both categories.

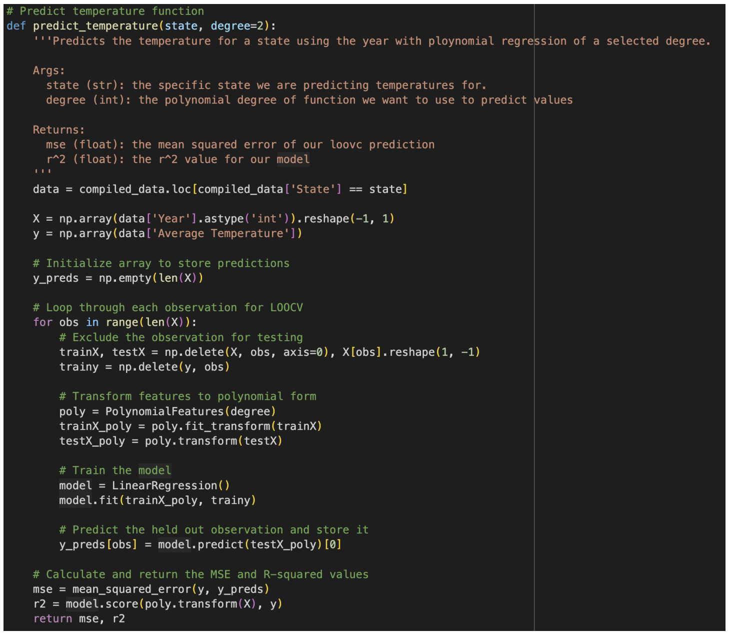


Graphing the precipitation and temperature change trends shows that a state like California has faced considerable climate adversity by experiencing warmer and dryer trends opposed to Georgia which has seen more gradual changes in both regards. It also shows that states in the upper Northeast have tended to become both significantly warmer and wetter over this period. Building off these findings, it becomes relevant to use similar techniques to pinpoint and predict future climate conditions for specific locations using a machine learning model.

Machine Learning Model

In an endeavor to understand and predict climatic trends, particularly average temperature and precipitation, it made sense to use polynomial regression models. This choice was made based on the non-linear nature of climate data. Lower-degree quadratic and linear models had visible correlation trends in their residual plots meaning they were underfit for making predictions with the data.

Here is a code snippet used for one of the models:

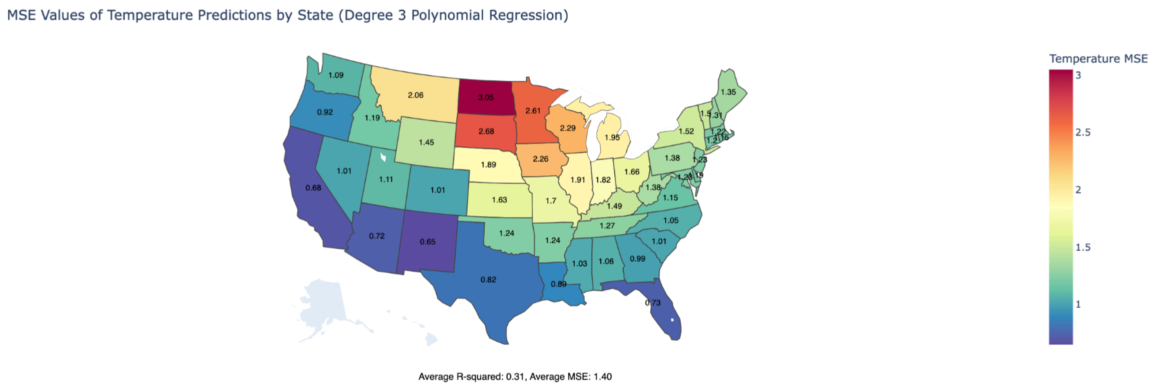


```python
# Predict temperature function
def predict_temperature(state, degree=2):
    '''Predicts the temperature for a state using the year with ploynomial regression of a selected degree.

    Args:
      state (str): the specific state we are predicting temperatures for.
      degree (int): the polynomial degree of function we want to use to predict values

    Returns:
      mse (float): the mean squared error of our loovc prediction
      r^2 (float): the r^2 value for our model
    '''
    data = compiled_data.loc[compiled_data['State'] == state]

    X = np.array(data['Year'].astype('int')).reshape(-1, 1)
    y = np.array(data['Average Temperature'])

    # Initialize array to store predictions
    y_preds = np.empty(len(X))

    # Loop through each observation for LOOCV
    for obs in range(len(X)):
        # Exclude the observation for testing
        trainX, testX = np.delete(X, obs, axis=0), X[obs].reshape(1, -1)
        trainy = np.delete(y, obs)

        # Transform features to polynomial form
        poly = PolynomialFeatures(degree)
        trainX_poly = poly.fit_transform(trainX)
        testX_poly = poly.transform(testX)

        # Train the model
        model = LinearRegression()
        model.fit(trainX_poly, trainy)

        # Predict the held out observation and store it
        y_preds[obs] = model.predict(testX_poly)[0]

    # Calculate and return the MSE and R-squared values
    mse = mean_squared_error(y, y_preds)
    r2 = model.score(poly.transform(X), y)
    return mse, r2
```

The code takes an input of a specified state and a polynomial degree with which to fit the model. It then defines the inputs (X: Year) and the expected outputs (y: Average Temperature) and utilizes LOO-CV (Leave-one-out Cross-validation) to iteratively leave out an observation, fit the model to the rest of the data, and then predict the final value and add it to a list of predictions. It does this by fitting the model with the sklearn.preprocessing.PolynomialFeatures function which transforms the features (X and y) into a polynomial form, before using Linear Regression for fitting. To finish off the model then calculates the MSE between the predicted y values and actual y values, along with the R^2 value of the model predictions using the polynomial transformed original X values and y values.

Predicting Average Temperature

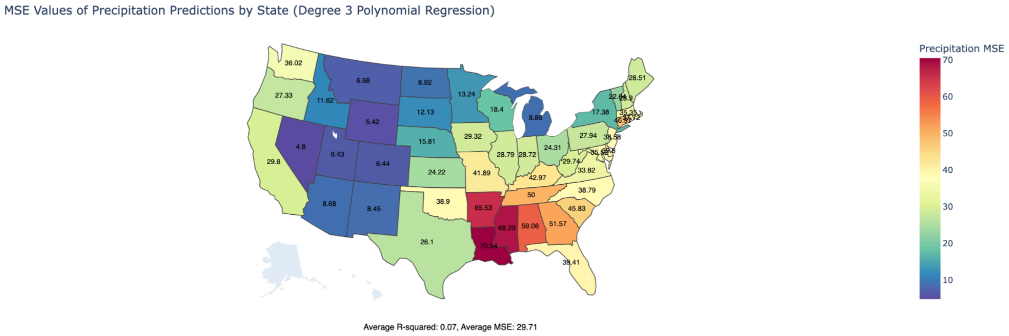
Initially, the project sought to predict the Average Temperature using just the year as a variable. Employing the above polynomial regression model with a degree of three, our model was most effective in its predictions. The model yielded an average R^2 value of 0.31 across different states, indicating a moderate level of predictability of temperature based on the year. This was complemented by a relatively low mean squared error (MSE) of 1.4, suggesting a reasonable fit of the model to the data. Below is a graph by state of the MSE for the average temperature predicted using the year.



An analysis of the MSE by state revealed an interesting pattern: coastal states generally showed lower MSE and higher R^2 values compared to inland states. This could be indicative of a stronger and more consistent temperature trend in coastal areas, possibly linked to factors such as oceanic temperature influences, which are less variable than land-based climate factors.

Predicting Average Precipitation

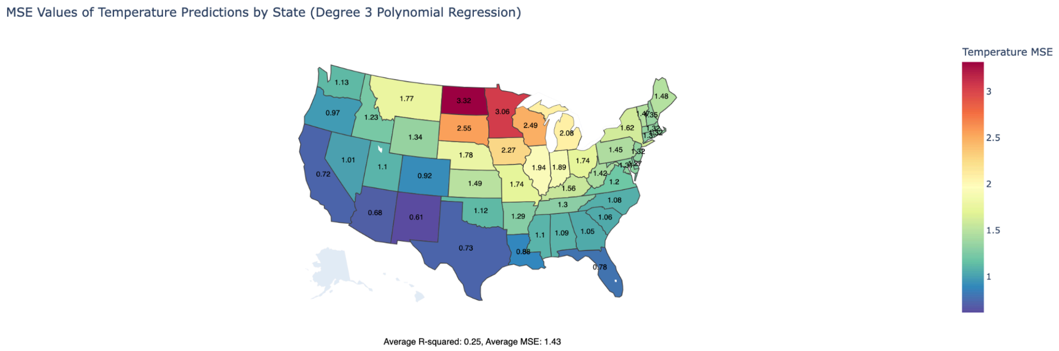
The next step was to apply a similar polynomial regression approach to predict average precipitation based on the year. This updated model is very similar to the above model but with the X variable changes to precipitation. This model, however, demonstrated limited efficacy, with an average R^2 value across states at a low 0.07, accompanied by a high MSE of 29.71. This suggests that precipitation is considerably more variable and less predictable based solely on the year. Below is a graph by state of the MSE for the average precipitation predicted using the year.



The regional differences in MSE, particularly lower values in the Western United States (excluding coastal regions), might reflect unique patterns in these areas, possibly influenced by factors like ocean currents, terrain, and historical weather patterns.

Multivariable Polynomial Regression

This model was then expanded to predict average temperature using both the average precipitation and the year as predictors. Contrary to the expectations, this multivariable approach resulted in a decrease in the model's predictive power (average R^2 across states dropped to 0.25, and MSE increased slightly to 1.43). This outcome suggests that precipitation, within the context of the dataset, does not significantly contribute to predicting temperature and may introduce additional variability or complexity into the model. Below is a graph by state of the updated MSE values with the multivariable polynomial regression model.



Discussion

Through the process of analysis of the NOAA weather data, a few identifiable apparent weather trends are happening across the United States. The first significant finding is that all the states in the study have experienced increasing temperatures over the measured period (1896-2023). On the lowest end, Iowa only saw a mean temperature increase of 1.3 degrees Fahrenheit while Taxes saw a much more significant 4.6-degree increase in mean temperature. This did not prove to be a great indicator of temperature mean change rates as midwestern states tended to have higher change rates than states with greater temperature mean increases. For context one of the states that saw the greatest increase in temperature was Texas and one of the states that saw the smallest change in average temperature was North Dakota, but their temperature change rates were 0.012 and 0. 024 respectively. This analysis went on to show that the upper Northeast is the region with the highest typical temperature mean increases in the United States. The next observation the project made is that the western United States is the only part of the country seeing a general trend of decrease in precipitation rates. States like California and Arizona have seen the largest decreases in precipitation at -0.004 and -0.006 per year respectively. East coast states saw the complete opposite of this pattern with Vermont seeing a mean precipitation rate increase of 0.065 per year. Comparing states with both these measurements yielded the general trend that the Southeast is experiencing the smallest change rate among both categories, the western United States is becoming warmer and dryer, the East Coast is becoming warmer and wetter the fastest, and the Midwest sits in the middle for both categories. The machine learning models in the study have highlighted important aspects of climate change in the United States. The successful prediction of temperature trends, particularly in coastal states, emphasizes more predictable patterns of warming in these regions, which is crucial for climate change strategies. Conversely, the challenges faced in accurately predicting precipitation highlight the complex and variable nature of this climate factor. Additionally, the multivariable regression analysis, which showed limited improvement when combining temperature and precipitation data, suggests a more complex relationship between these climate factors.