Making the Rain

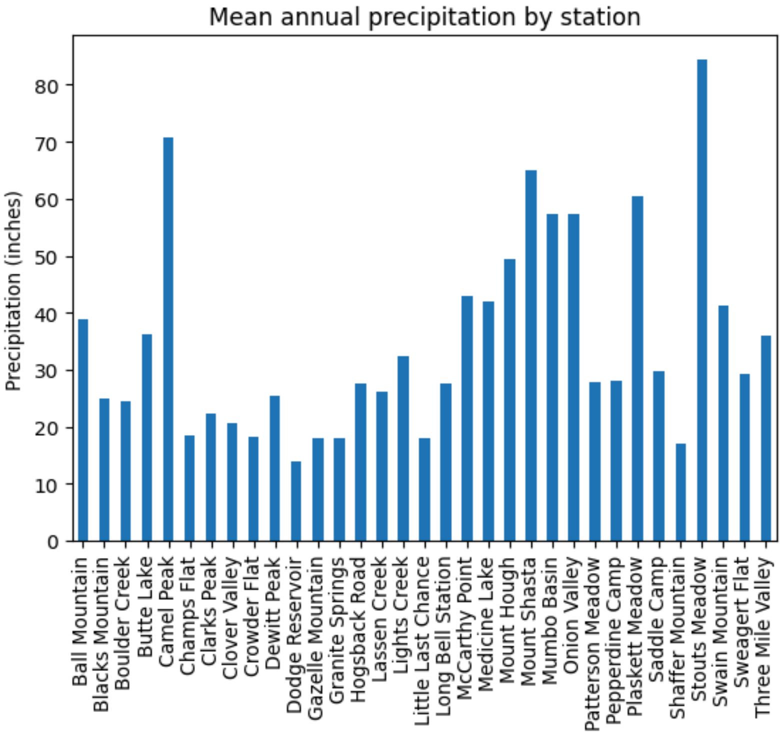
Bryan Zanoli, Daniel Arias, Kyle Stefun, Victor Gomez

Introduction

Growing up in California, some of us have witnessed the dramatic shifts between wet winters and dry spells in Northern California, which affect everything from agriculture to wildfire risk. This project will analyze historical annual precipitation data dating back to 1944 and attempt to gauge the effectiveness (or lack thereof) of this data to predict patterns of precipitation.

Selection of data

The data set includes all historical annual precipitation data for the California Department of Water Resources's 33 Northern California rain stations by year. Each annual entry includes data related to the Station: name and station location by county as well as direct latitude and longitudinal coordinates.

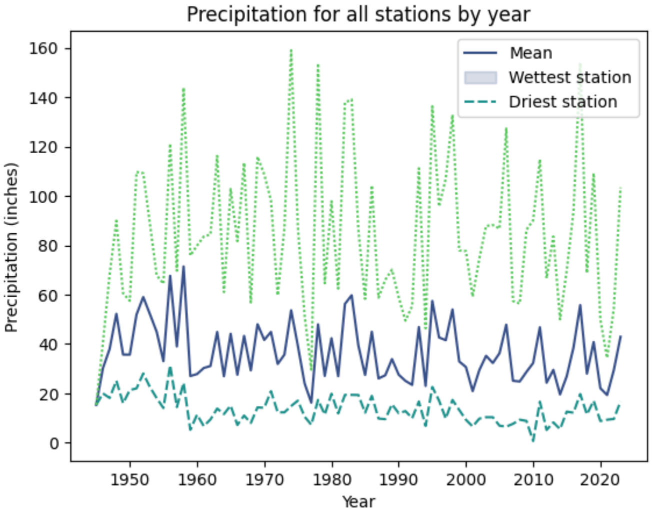


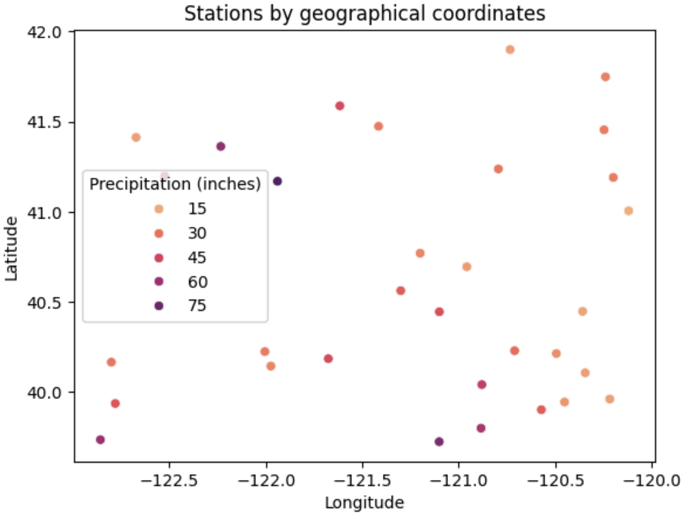
It was determined that this data would not be sufficient for an accurate prediction, or even a strong analysis of what factors contribute to an accurate prediction, so additional information was sought. The Open-Meteo weather API was used to pull additional data based on the coordinates of the weather stations, including elevation, temperature, and wind speed, and merged with the existing data set.

The data was then further curated to add necessary features that averaged out the monthly totals from the weather station data, to match the structure of the annual rain station data.

Methods

In the course of exploring the data, two distinct ways of looking at the data were found. The data could be viewed as a sequence of changing values over time, or as a series of geographical points.





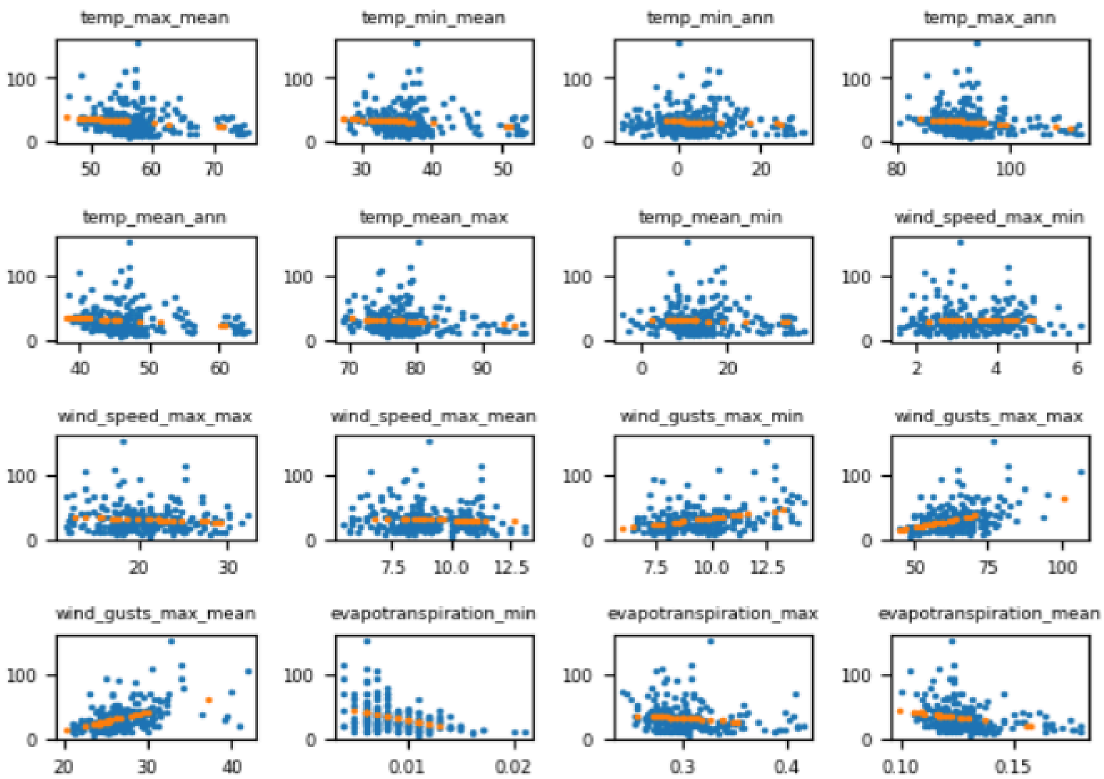
Based on these, two different approaches were examined for making predictions based on the data set.

Method - Linear Regression

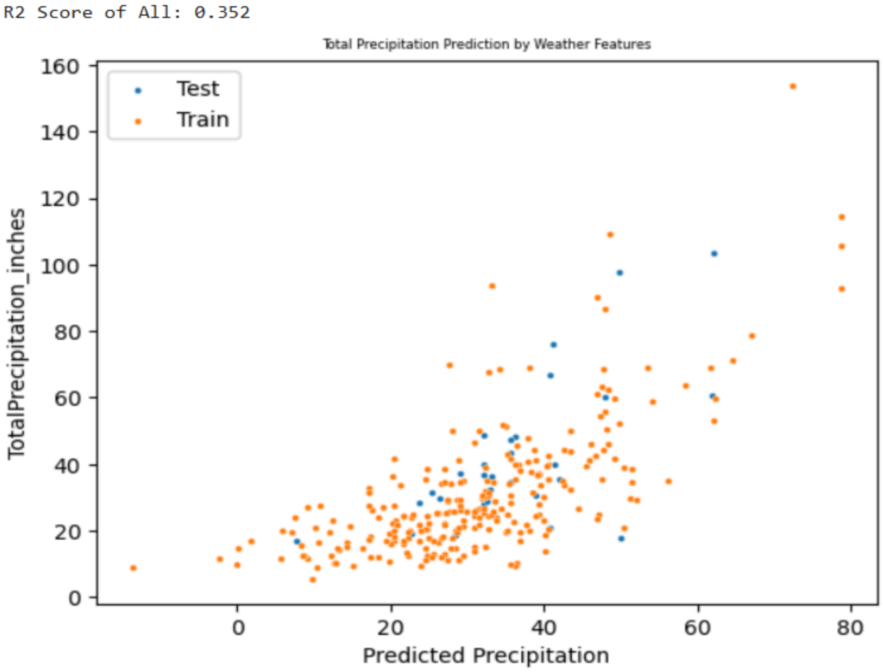
Linear regression analysis first began with feature selection to determine which, if any, features or combination of features would provide the best predictive ability.

To provide categorical prediction associations, the features were broken down into geographic vs weather based feature sets and evaluated separately. It’s worth noting that while beyond the exercise of this analysis, it could be beneficial to evaluate the correlative predictive capabilities of geographic features combined with their weather feature counterparts.

The initial analysis of the weather features showed the below results. Although the chart is poor visual quality, the results are obvious–no feature shows a strong correlation with total precipitation.

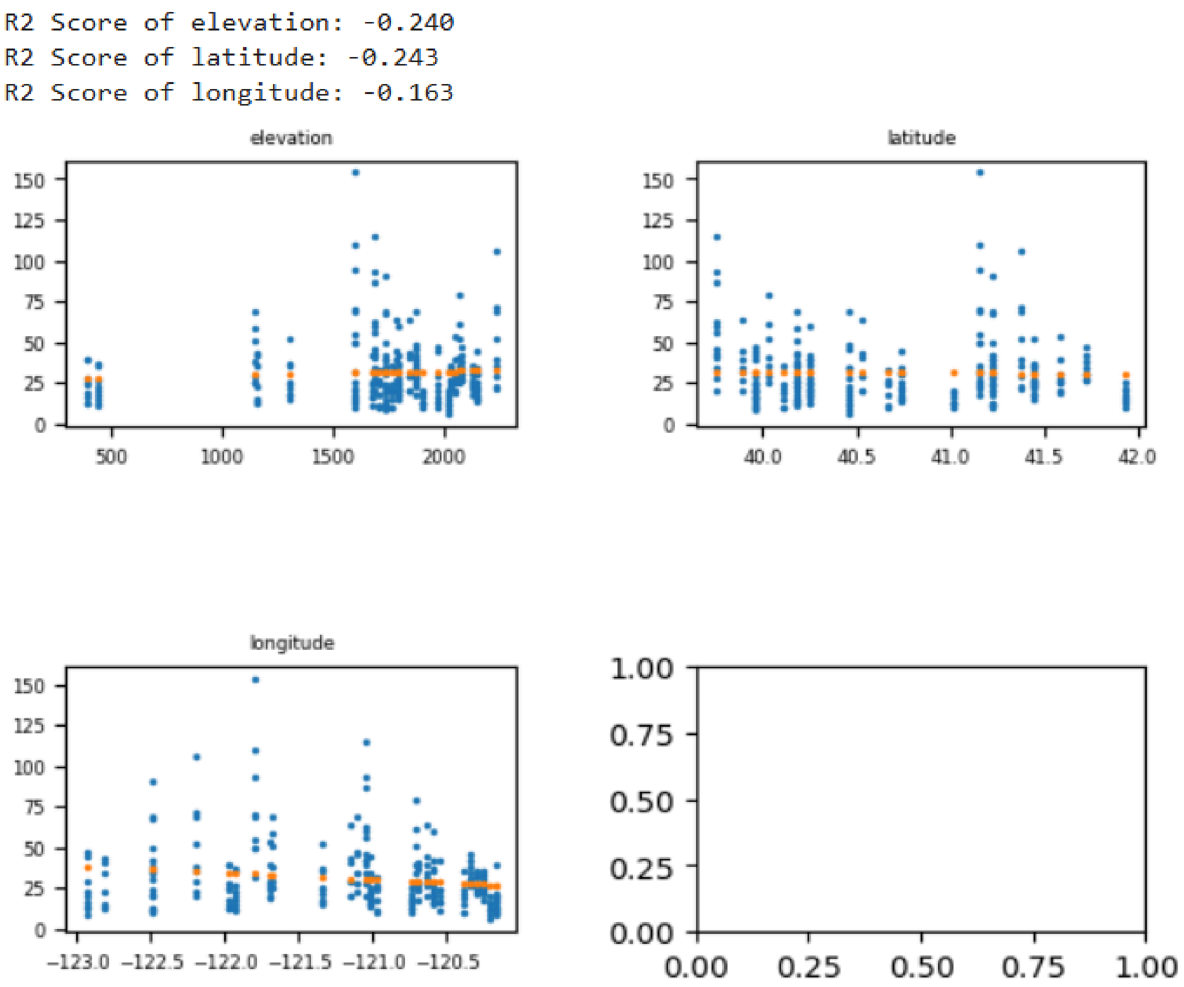


As sometimes can occur with linear regression, a combination of features may result in a stronger association with the target. Shown below, the combined predictive power of all features results in a relatively low R2 score of .35. Analyzing the chart further, it becomes clear that the model does not do well in predicting precipitation, with predicted amounts peaking at 80 instead of the actual peak near 140, and even including negative values in its predictions.



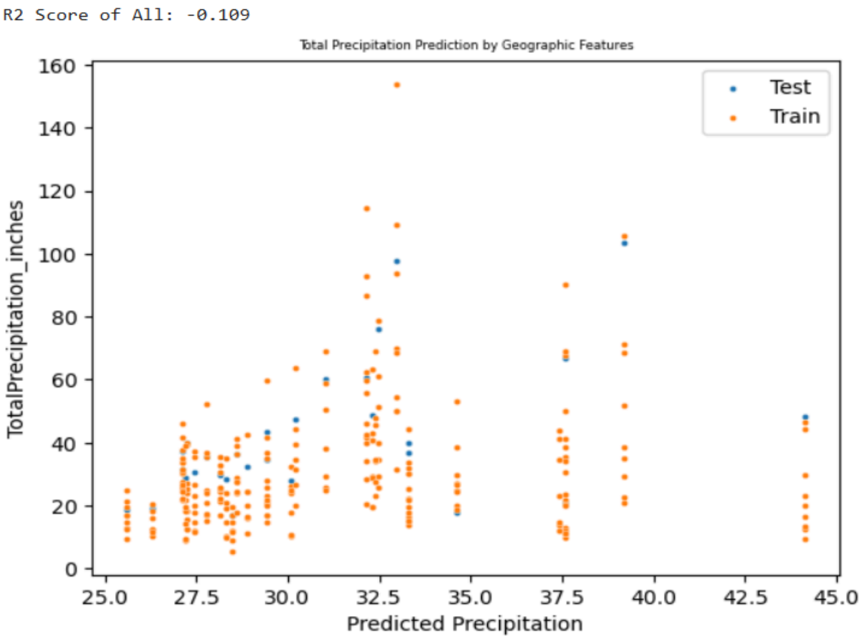
```
R2 Score of All: 0.352
```

Repeating this analysis with geographical features, the results are in fact worse.



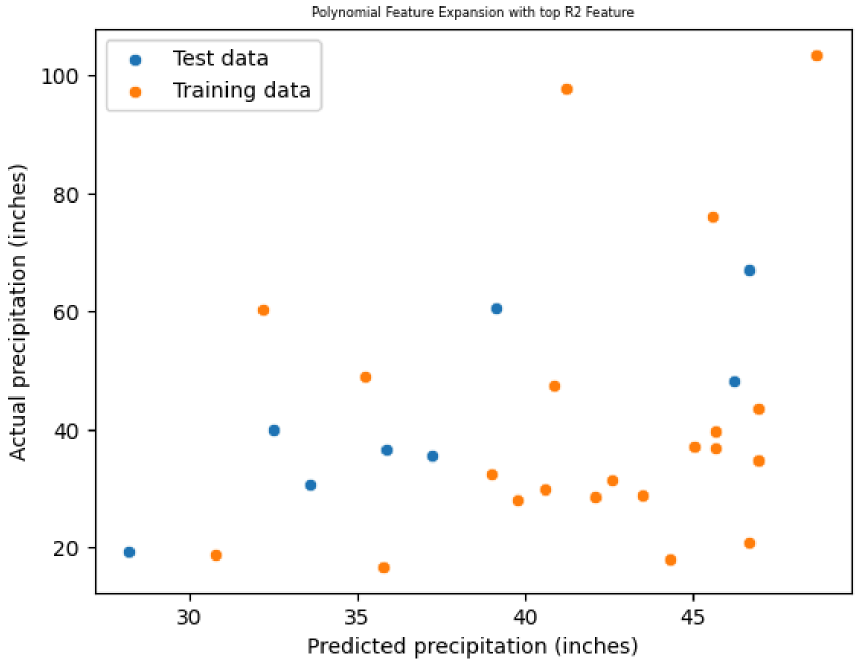
```
R2 Score of elevation: -0.240
R2 Score of latitude: -0.243
R2 Score of longitude: -0.163
```

The combined geographical predictive power of the graphical features indicates poor performance. In fact the R2 of the new model is negative, indicating a completely underfit model.



```
R2 Score of All: -0.109
```

Additional testing was carried out to gauge the value of polynomial features. The most promising feature to be identified demonstrated an R-squared value of .408, but plotting based on this finding still proved unimpressive:

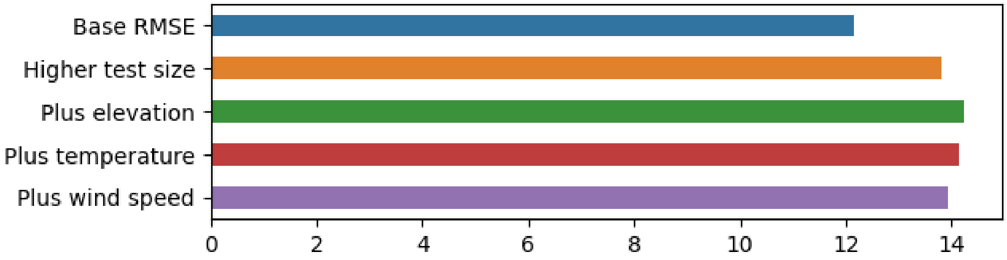


Method - k Nearest Neighbors

Linear regression proved to be a poor fit for the problem at hand. The other approach taken to attempt to predict precipitation was geographical: that is, given the data from the weather stations and their geographical coordinates, could we use that to predict precipitation in between those specific points?

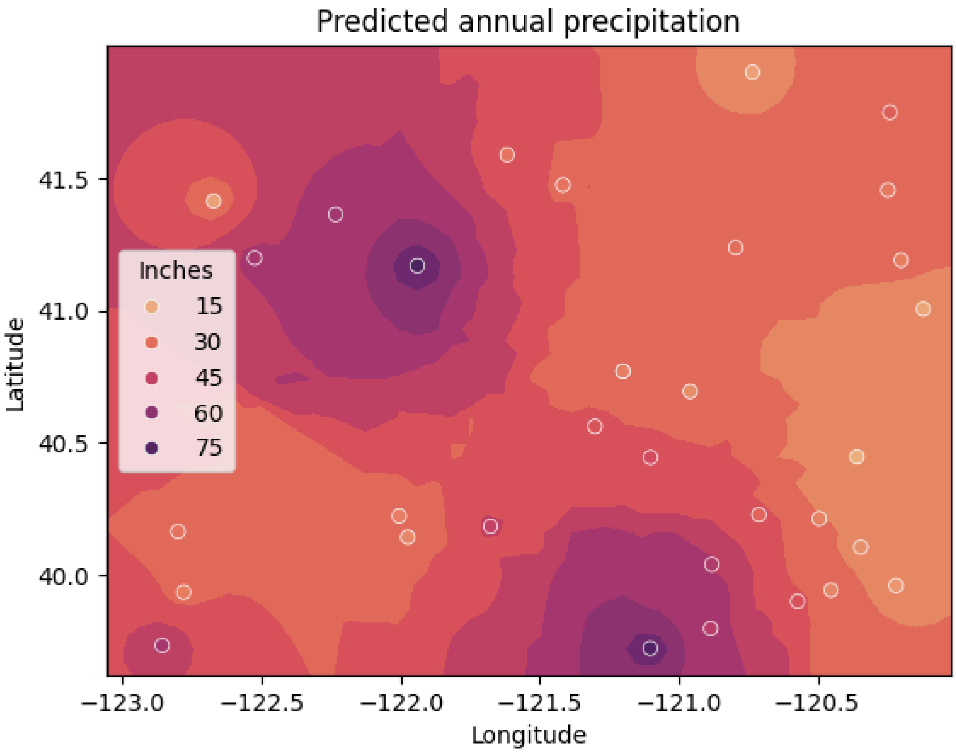
To this end, kNN regression was used to make predictions based on the stations closest to the prediction location. To test the predictive potential of the model, train/test splitting was conducted. It was found that splits with smaller proportions of training data had significantly higher average errors, so the train/test split was set to favor a large amount of training data. With only 33 stations, the randomness of this split resulted in a sizable variation of errors, so the split was repeated a large number of times and the average error was taken.

Additional experiments were run to gauge the value of adding more predictors, such as elevation and temperature:



Experiments with additional predictors did not seem to be effective; the best result was found using the stations' latitude and longitude as the sole predictors. With a data set this small, perhaps there was too much noise in these predictors for the model to establish a meaningful correlation.

The predictive power of this model looks to be modest at best, with an RMSE of 12 annual inches of precipitation. Errors notwithstanding, here is what a precipitation map of the region would look like, using this model to predict annual rainfall amounts:



While the accuracy of the model is questionable, visualizing it like this does show that the prediction does make sense within the limits of the data we're working with. With more data, it's possible that this model could be used to make reasonably accurate predictions based on geographical data.

Discussion

While the linear regression approach had some minor predictive power, it proved to be too inaccurate to be useful in this context. More data might be beneficial, but it is also possible that this is not the type of problem that the model is well suited to solving.

It makes some intuitive sense that the kNN model might be useful when approaching the problem from a geographical perspective: we understand that locations that are physically close by will likely experience similar weather patterns. The contour map generated by the model shows that it can be used to make predictions that make sense, based on the data. However, it is still limited by the lack of data itself.

Given that neither the linear regression or kNN regression model benefited from additional predictors, the best course would seem to be to collect data from additional locations. It's possible that this would not only benefit the model by filling in empty spaces in the map, in the case of the kNN model, but also create an opportunity to use additional predictors that would be of benefit, rather than simply adding noise due to the small sample size.

A sensible next step for the project would be to use the weather API to collect information from different locations in between the weather stations provided in the original data. This would provide more data to train the model as well as more data for testing. The kNN model showed noticeable improvement from increasing the proportion of training data in the train/test split; adding additional training data to the core data set is very likely to improve its predictive accuracy.

Summary

At least one of the models used (kNN) showed some potential predictive power. However, the original data set proved to be insufficient for consistently accurate predictions, and even augmenting it with additional information such as elevation, temperature and wind speed provided no benefit to either model. The approach most likely to improve the models' predictive power is to expand the data set to include new geographical locations.

References

Annual Precipitation Data for Northern California 1944-Current
https://data.ca.gov/dataset/annual-precipitation-data-for-northern-california-1944-current
California Department of Water Resources

Open-Meteo
https://open-meteo.com/
Open source weather API

GitHub repository

https://github.com/csumbbryan/cst383-Team2-Project-Rain