Predicting droughts in North America

Group 8
Spencer Ramsey
Tessa Cayton
Travis Harry

Reason

Rising temperatures and global warming has caused many environmental challenges. Droughts can have lasting effects on food supply and the economy. If we could predict droughts based on current weather/climate patterns it could help areas work to prepare for water shortages before they occur.

Main Hopes

The goal of our analysis is to see if we can predict a drought and its severity based off of weather/climate patterns and landscape/soil conditions.

Additional questions we are interested in answering are which factors have a stronger correlation with drought severity and how far out can we accurately predict droughts into the future.

Data and Storage

Data Sources

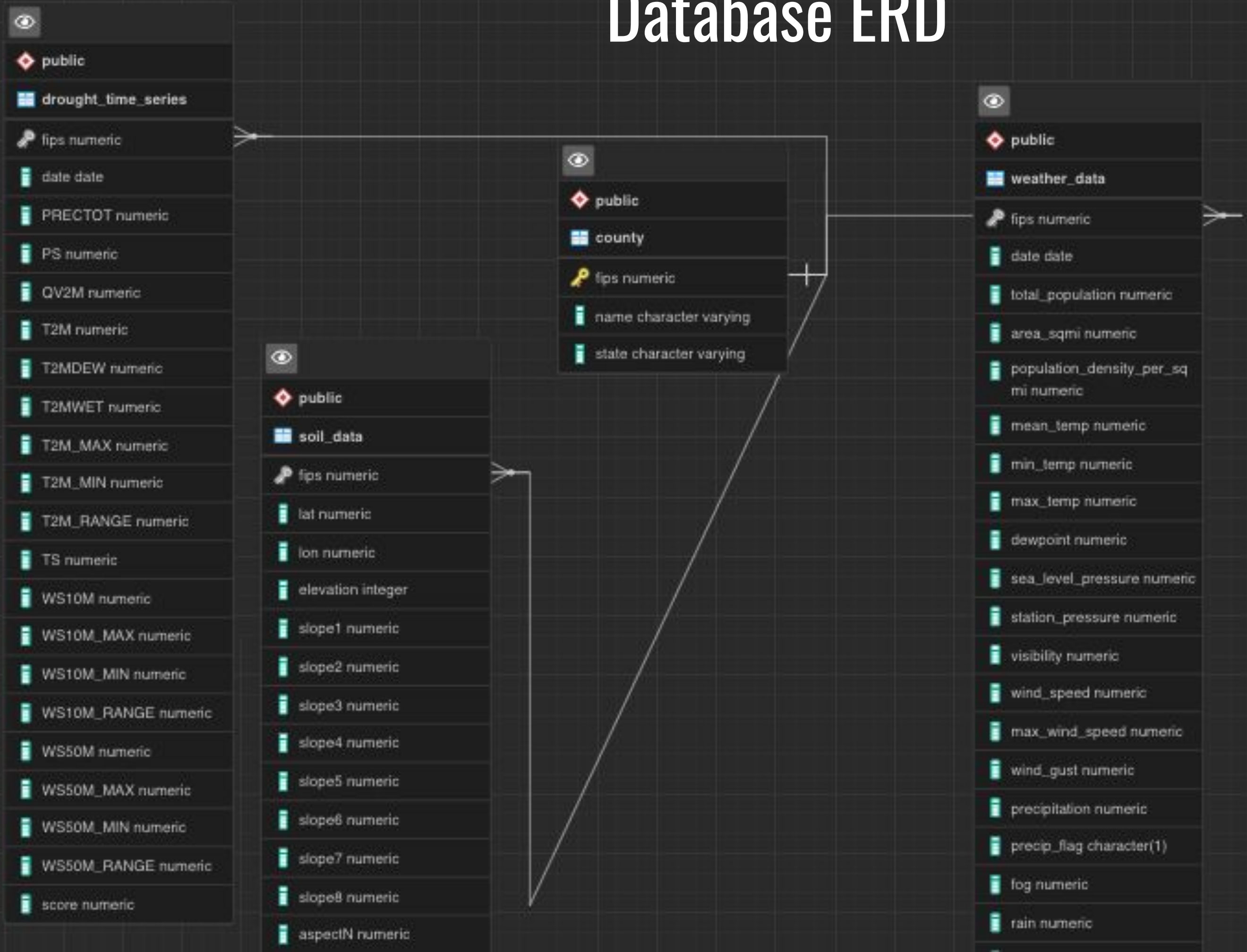
All data sets were found on Kaggle. FIPs (county) information was pulled from the USDA.

- https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data
- https://www.kaggle.com/cdminix/us-drought-meteorological-data
- https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697

Database

- The Postgres DB has been set up with 4 tables of static data. The data was pulled from Kaggle and a few other sources (see above for specifics) and cleaned using python scripts before being imported into the database.
- In order to streamline and reduce the data, we dropped all the county and state information from each individual table and only left the county fips code as a unique identifier.
- The DB is connected with the machine learning model by hosting it in the cloud through AWS/RDS.

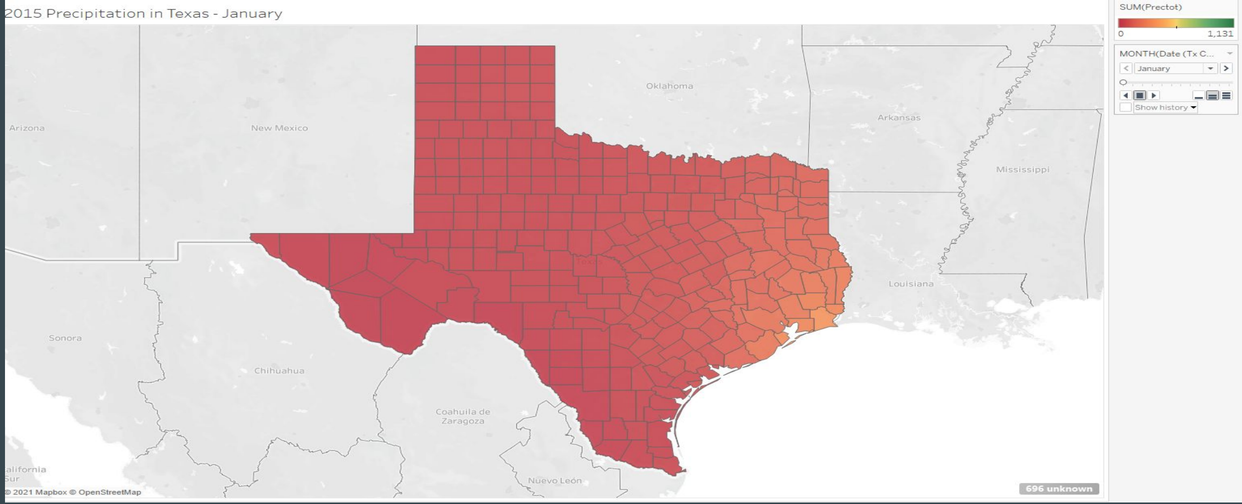
Database ERD

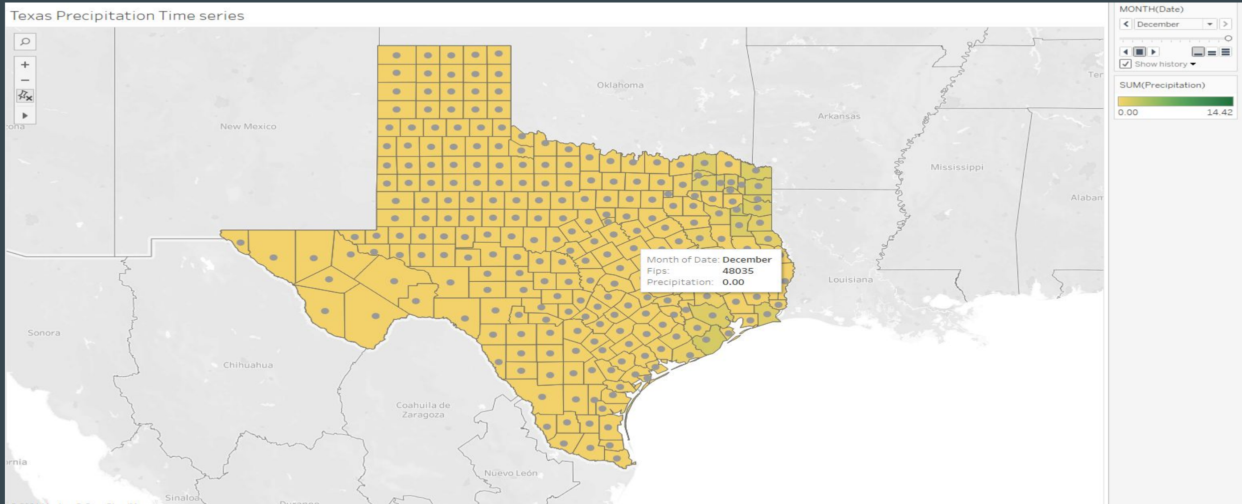


Visualizations

Tableau

[Tableau Dashboard]



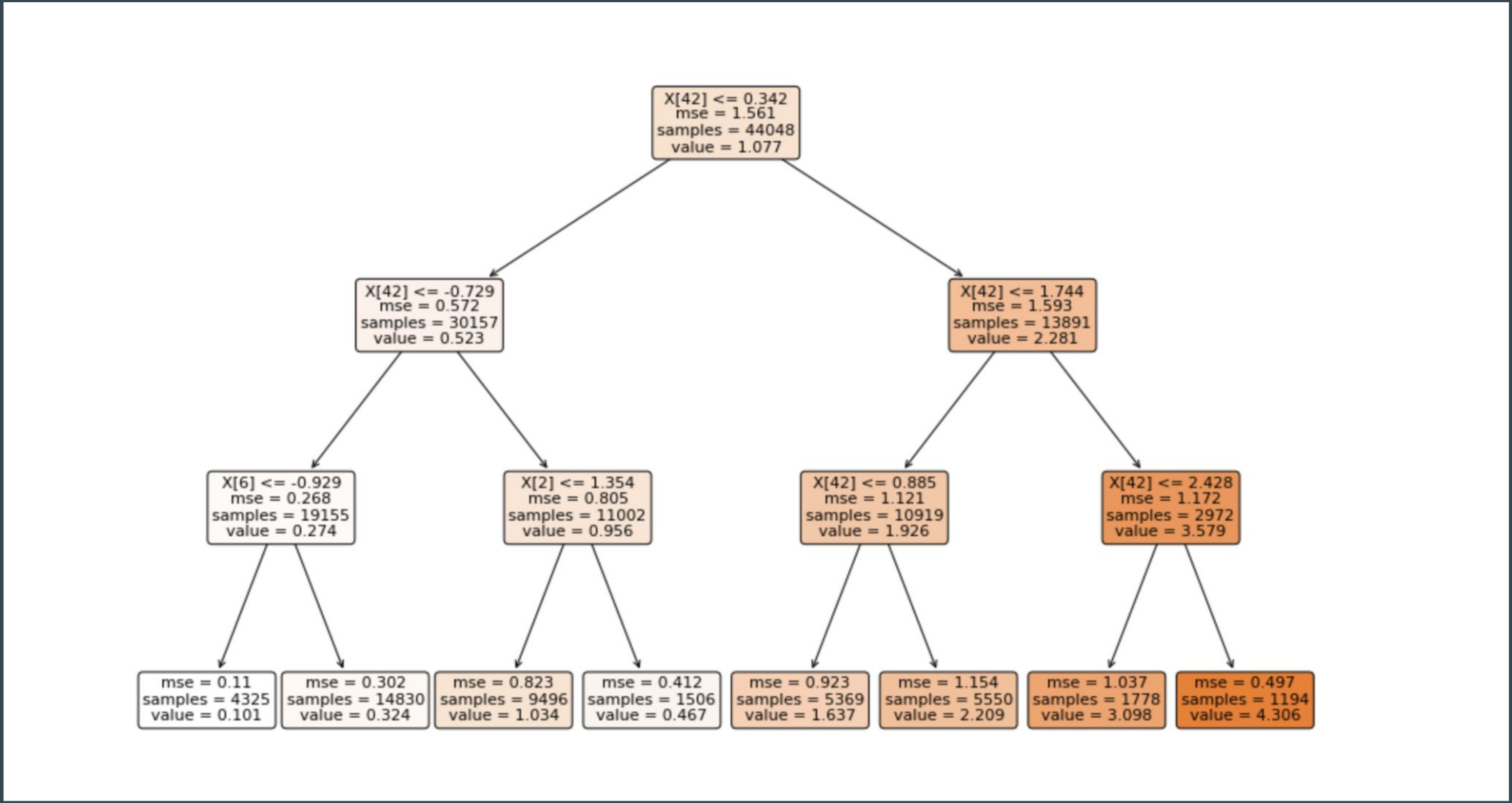


Machine Learning

Data Exploring

- In order to prepare the data for the Random Forest Regressor, we first merged our two datasets based on the county FIPS code.
- We limited the results only to Texas since we wanted to achieve higher accuracies by limiting the location to a single state.
- Lastly we dropped some data which was redundant or others which were identifiers like latitude and longitude.

Random Forest Diagram



Current Exploring results

- After Segment 1 we received the following results: Mean Absolute Error (MAE): 0.5480466754695936 Mean Squared Error (MSE): 0.5444936385385971 Root Mean Squared Error (RMSE): 0.7378981220592699
- The MSE increased, which indicates that the soil data actually made it more difficult for the Random Forest Model to predict outcomes correctly. We may be better served by leaving out the soil data.

Model Accuracy

One Week Ahead
Mean Absolute Error (MAE): 0.11395177801839734
Mean Squared Error (MSE): 0.03870444769778442
Root Mean Squared Error (RMSE): 0.19673445986350338

Two Weeks Ahead
MAE: 0.15552836606280845
MSE: 0.0651923262596554
RMSE: 0.25532787991062666

One Month Ahead
MAE: 0.21093419709423947
MSE: 0.1054075453231857
RMSE: 0.3246652819800505