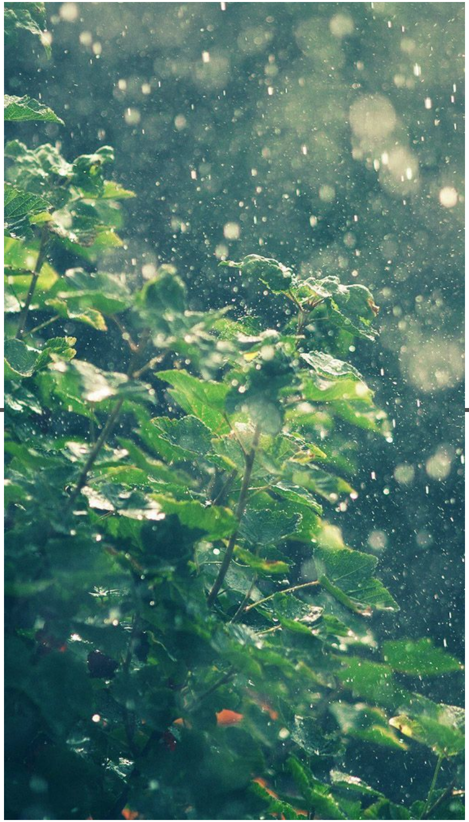


PREDICTIVE CLIMATE

Data Estimating Future Weather Patterns

Group Members: Enmanuel Baez, Keivon Chamanara, Nishita Cheekatamarla

DS 4002 – 11/20/23

PROJECT OVERVIEW

1. Project Details
2. Data Acquisition + Explanation
3. Analysis Plan + Justification
4. Tricky Analysis Decision
5. Bias + Uncertainty Validation
6. Results + Conclusion
7. Next steps

PROJECT DETAILS

MOTIVATION:
Since 1901, average precipitation and temperature has increased by 0.20 inches and 0.17°F per decade, in the US. (EPA 2023)

Resulting in soil erosion, rising sea levels, and extinctions. Could scientists of the past have predicted this? Today, can we?

RESEARCH QUESTION:
Can multilinear regression models accurately predict average temperature, based on historical US climate data?

NULL HYPOTHESIS:
Precipitation, PDSI, minimum temperature, and maximum temperature are NOT accurate predictors of average temperature.

MODELING APPROACH
Use multilinear regression model on US climate data to predict average temperature and find p-values from model's residuals.

GOAL
To investigate how accurately a multilinear regression model can predict average temperature based on past US climate data and either reject or fail to reject the null hypothesis.

Data Acquisition & Explanation

Final Dataset

Data Dictionary

Column	Description
Date	Date of data entry, providing year and month
Avg Temp	Average temperature in °F
Min Temp	Minimum temperature in °F
Max Temp	Maximum temperature in °F
Precipitation	Average amount of precipitation in inches over the duration of one month
PDSI	Average Palmer Drought Severity Index (PDSI) value over the duration of one month

Acquisition

1. Data gathered from National Centers for Environmental Information.
2. Downloaded 5 separate files, each containing one a weather parameter.
3. We merged those 5 files into one using Python.

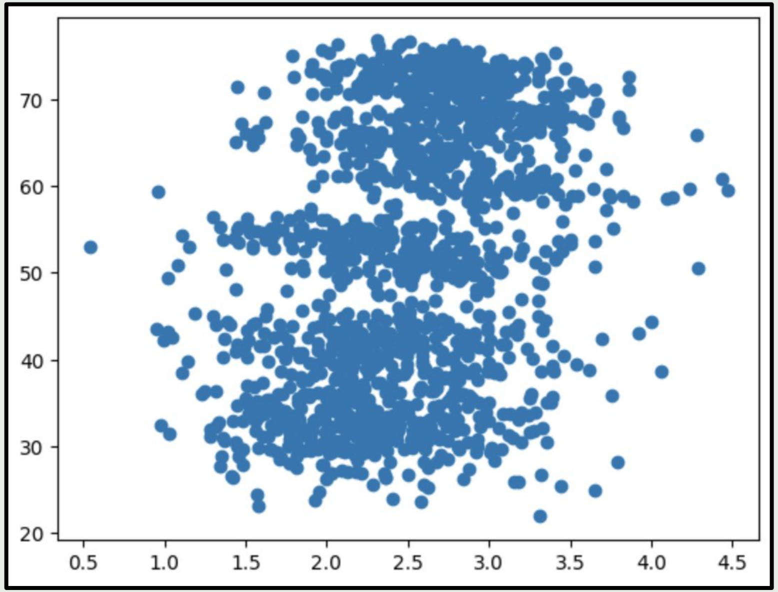
Cleaning

1. Checked the dataset for NA values.
2. Removed duplicate variables resulting from merging datasets.
3. Created dataset shown to the left.

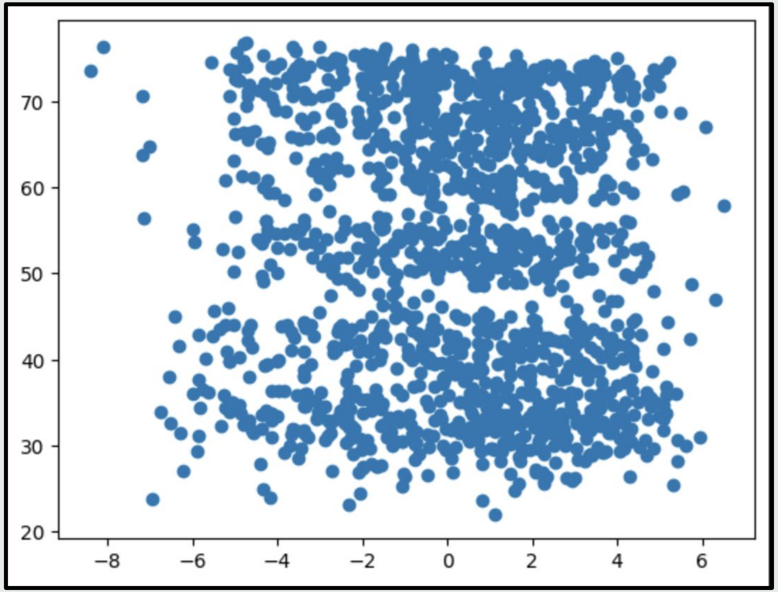
Analysis Plan + Justification

EXPLORATORY DATA ANALYSIS

Average Temperature vs Precipitation

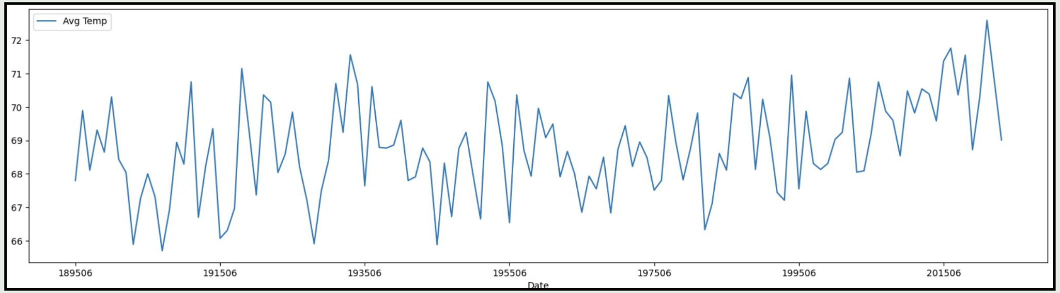


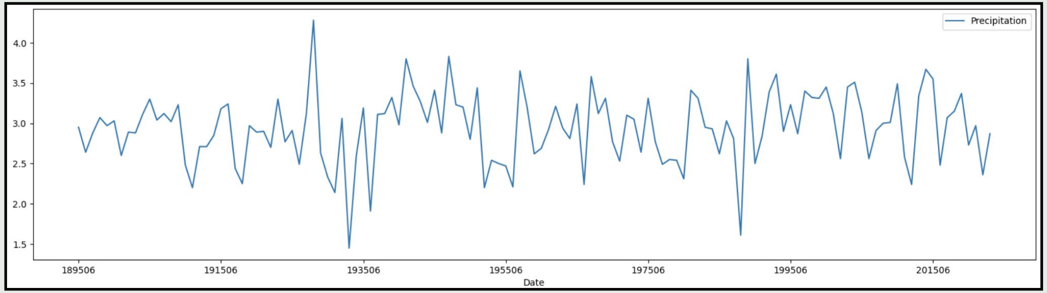
Average Temperature vs PDSI



EXPLORATORY DATA ANALYSIS (continued)

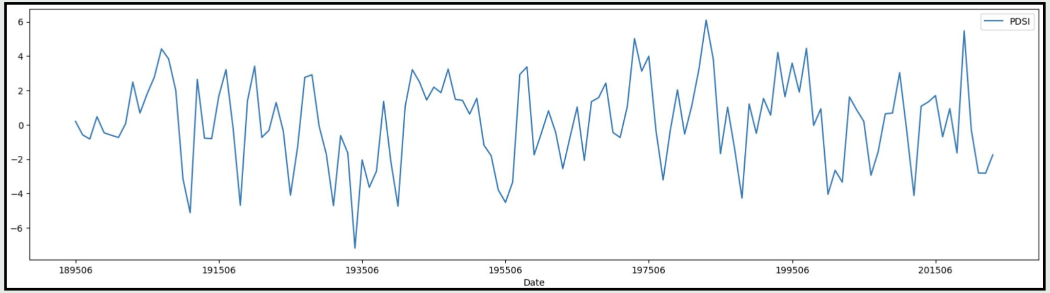
Average Temperature in June over time →





Average Precipitation in June over time ←

Average PDSI in June over time →



Analysis Plan

1. After doing our EDA, we noticed patterns in the variables over time.

2. We used multilinear regression to create a predictive model of average temperature.

3. We found p-values for each of the independent variables.

TRICKY ANALYSIS DECISION

- **Accounting for seasonality in average temperature, minimum temperature, and maximum temperature?**
  - Planned to use simple linear regression.
  - Can't use simple linear regression model with highly variable data due to seasonal changes.

- **What kind of model to use: ARIMA or Linear Regression?**
  - Using the statsmodel package, we performed seasonal decomposition.
  - Removed seasonality from data, subtracting its values from our values.
  - Then allowed for multilinear regression.

BIAS & UNCERTAINTY VALIDATION

Spatial Considerations:

a. This data is from all over the United States. However, weather and climate conditions vary a lot across the country.

Multilinear Regression:

b. The data was modified to remove the seasonality in order to use multilinear regression. This may have skewed the results.

RESULTS & CONCLUSIONS

Precipitation and PDSI

- The p-value for precipitation was 0.611 and 0.872 for PDSI, meaning that we cannot reject our null hypothesis.
- These variables most likely do not play a role in determining average temperature.

Minimum and Maximum Temperature

- The p-values for minimum and maximum temperatures were 0, meaning that we can clearly reject our null hypothesis.
- This makes sense, as minimum and maximum temperature play a role in calculating the average temperature.

NEXT STEPS

Flip our analysis

a. Instead of trying to see if certain variables affect average temperature, we would like to see how other variables can be predicted.

Use different models:

b. Ignore seasonality and use ARIMA.

Predict future values

c. Using our model, we would like to see if we could use it to predict future average temperatures, as it was very difficult using the statsmodels package.

Thank You

REFERENCES

SOURCE #1:

[1] NOAA National Centers for Environmental Information, "Climate at a Glance: Time Series," [Online]. Available: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/national/time-series. [Accessed: March 25, 2024]

GITHUB REPOSITORY LINK: https://github.com/keivonc/ds4002project2