Project 6 - Linear Regression
Logan Liddiard, Eric Johnson

Git Repository: https://github.com/Logsterx7/cs5830_project6
Project Slides:https://docs.google.com/presentation/d/1QhaoZb9bbhD9V6bMLQTERE81rNhA9GEm59GPfJIWsKo/edit?usp=sharing

Introduction

The domain of this project is a hydrologic dataset that measures water intake and outtake of a river at different measurement stations once a month. The goal of this project is to use linear regression to predict the baseflow of the river, and in our case, the predicted baseflow with year increments, using fields such as precipitation, irrigation pumping, the year and evapotranspiration. This is important as it can allow for farmers, ecologists, and even cities and towns to prepare for possible droughts or overflows in a year, and recognize patterns. Overall, we discovered that by using the fields listed above, we could get a fairly accurate model that accounted for roughly 60% of the river baseflow variation.

Dataset

The dataset is focused on providing information of a river that branches between Nebraska, Colorado and Kansas. There are measurement stations periodically spread across the river and branches that take measurements once a month. The dataset contains the location of the gauging station, the river segment id, the evapotranspiration, precipitation, irrigation pumping, and observed baseflow. However this data set alone doesn't tell us much. It even has inconsistent data in there as well. If you look at the observed values before the year 1950 you will see that there is a dramatic difference. There is belief that this data is miscalculated. Due to this we decided to cut off any data before 1950 and only focus on the 50 years after that. We then needed to group each year up and take the mean of all values for that year so that we could further analyze the data.

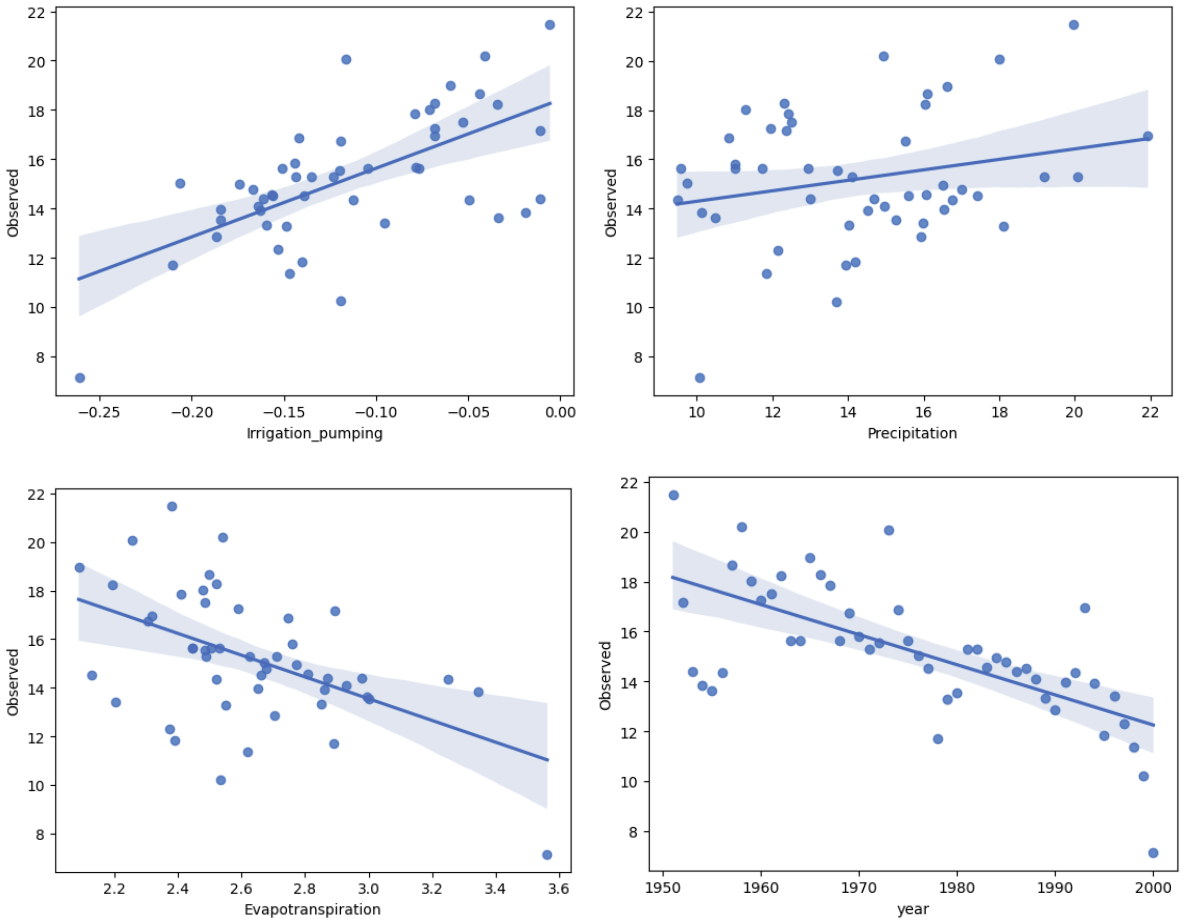
Analysis Technique

The analysis technique that we used was linear regression. This method allowed us to try and fit a line to multiple independent variables in order to try and predict the observed baseflow of a river. This was a suitable technique as there were many independent numerical measurements, and a numerical target variable that we wanted to predict. The only problem is we weren't sure what values to feed our linear regression model since different combinations can yield different results. Ultimately we decided to look at a heatmap of our data to see if we could find any patterns for our values that may help us increase our r-squared value. Sure enough the results of

this allowed us to get stronger predictions. It shot up from our initial 11 percent average to a much stronger value.

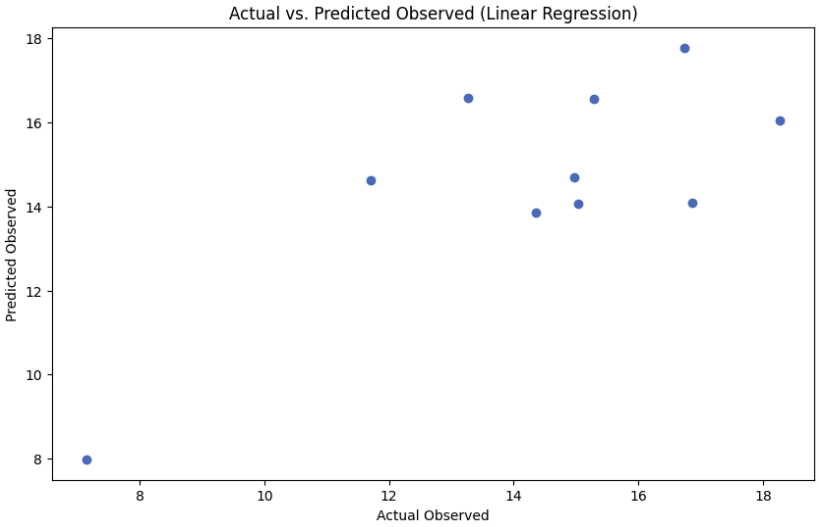
Results

Initially, we could not identify any significant positive or negative trends between any of the fields and the observed baseflow. By grouping the data by year, and then averaging the attributes throughout each grouped year, we were able to generalize the data and recognize patterns and correlation between the fields and the target, as seen by the graphs below:



Overall, the average coefficients for the linear regression lines for precipitation, irrigation pumping, year and evapotranspiration resulted in being 0.27, -1.26, -0.14, and -3.26 respectively.

By using these averaged and grouped fields, we were able to yield an average R-squared score between many tests of 0.60, which would look similar to the graph below:



Technical

While the data was clean in the sense that there weren't many missing values, if any, it was very hard to understand. Since the data did not have easily visible correlation, we decided to start looking at averages over the years, so we grouped the data by year, and took the average of the values found within that year.

We used Linear Regression as we had multiple, continuous independent variables and a target continuous variable. Linear Regression allowed us to plot data and determine if there was a relationship between the variables, and create a model for data prediction based on key factors.

Initially, we tried looking at relationships between the fields and the target based on overall data, specific river segments, years, etc. We even tried to clean the data up to the point where we would only look at seasons and segments. But doing this yielded no results. After no patterns could be found, we decided to generalize patterns by looking at strictly year groups, as well as averages. By creating a correlation matrix, we were able to determine there were patterns by using field averages over year intervals, and we created our model using that discovery.