Samuel McMillan and Jensen Judkins
A02261822 and A02310135

Project 6: Linear Regression Analysis

Introduction

For our project, we were given a dataset focusing on the base flow of a river. Using this we were tasked with creating a model using appropriate features and creating a linear regression model to predict the base flow of the river. Looking at the different features and breaking apart the different features were able to make some predictions. It is imperative that baseflow is kept track of for this helps people know what water is where for them to use for farming and other necessary city functions. We hope to be able to figure out the predicted values for base flow to solve these issues.

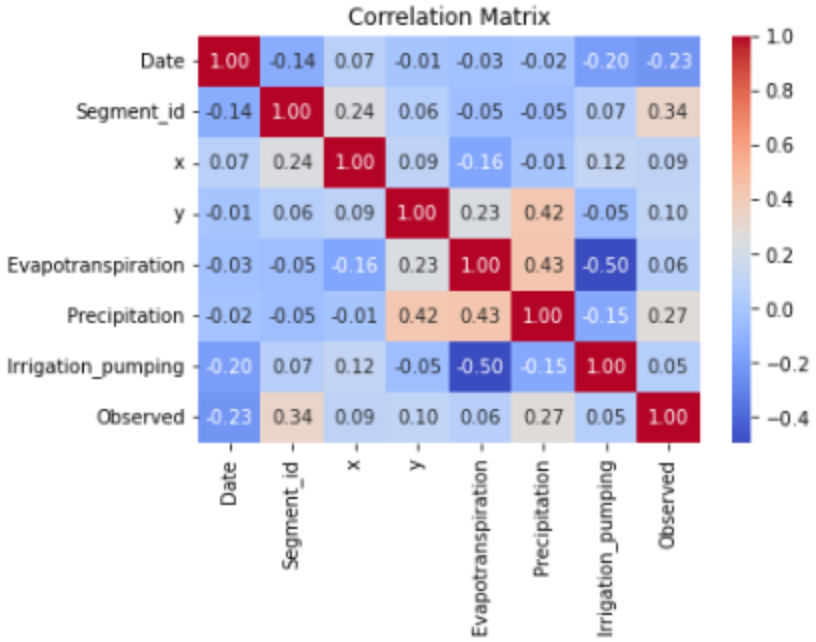
Link to Presentation
Link to GitHub Repo

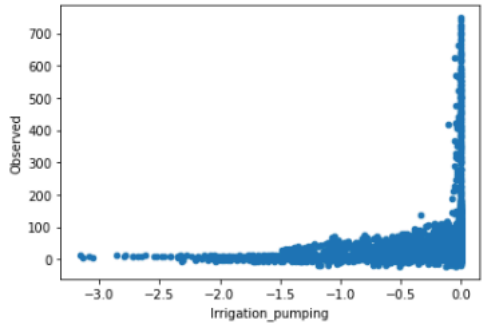
Dataset

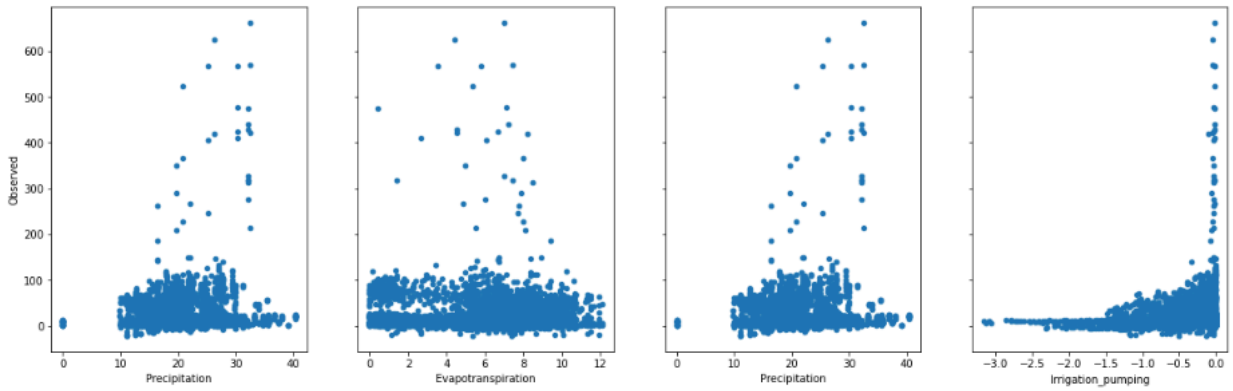
The dataset we were given contains several key points. This feature list contains 'Segment_id' points which indicate the area in which the data was collected, 'date' points which we were able to separate into months, years, weeks and day of the month, 'Precipitation' the precipitation amount of an area adjacent to the river segment in the given month, 'Irrigation pumping' the amount of groundwater pumped out for irrigation in an area adjacent to the river segment in the given month,evapotranspiration is the amount of an area adjacent to the river segment in the given month, and x and y coordinates which indicate the spatial location of the gaging station at which observations are obtained.

Analysis Techniques

First for our analysis we needed to decide which features we were going to use in the regression model. This was very difficult but we were able to find a few very interesting points. The first indication that we found was that two x and y coordinates had an enormous amount of variation in the baseflow recordings (fig 1 and 2). We decided that for the purpose of our model and in order to create a more reliable prediction, we should remove these points. We removed seven of the forty-three x coordinates that resulted in the highest base flows from the dataset. We also reviewed and found that precipitation levels made the model extremely inaccurate (fig 3) and we decided that it would be fair to remove those as well given that no human intervention would be able to change the weather anytime soon. This still resulted in over 7800 data points and we felt that this was an adequate amount of data points in which to base our model. We also removed all data points with no irrigation water that was pulled from it. This simplified the model even more.


Correlation Matrix
Date 1.00 -0.14 0.07 -0.01 -0.03 -0.02 -0.20 -0.23
Segment_id -0.14 1.00 0.24 0.06 -0.05 -0.05 0.07 0.34
x 0.07 0.24 1.00 0.09 -0.16 -0.01 0.12 0.09
y -0.01 0.06 0.09 1.00 0.23 0.42 -0.05 0.10
Evapotranspiration -0.03 -0.05 -0.16 0.23 1.00 0.43 -0.50 0.06
Precipitation -0.02 -0.05 -0.01 0.42 0.43 1.00 -0.15 0.27
Irrigation_pumping -0.20 0.07 0.12 -0.05 -0.50 -0.15 1.00 0.05
Observed -0.23 0.34 0.09 0.10 0.06 0.27 0.05 1.00
Date
Segment_id
x
y
Evapotranspiration
Precipitation
Irrigation_pumping
Observed
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Observed
Irrigation_pumping


Observed
Precipitation
Evapotranspiration
Precipitation
Irrigation_pumping

These top graphs just show looking at the dataset to understand it better. Bottom of the two graphs shows before and after removing all data points with a irrigation_pumping value of 0 because all irrigation_pumping values of zero could be seen as an outlier for this dataset.

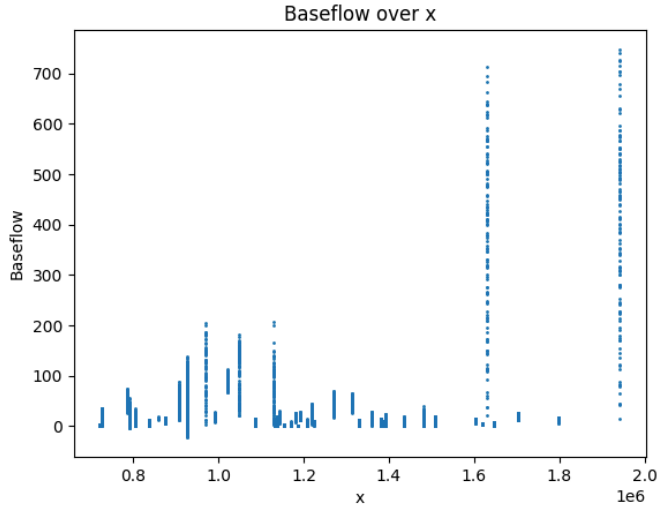


Fig 1

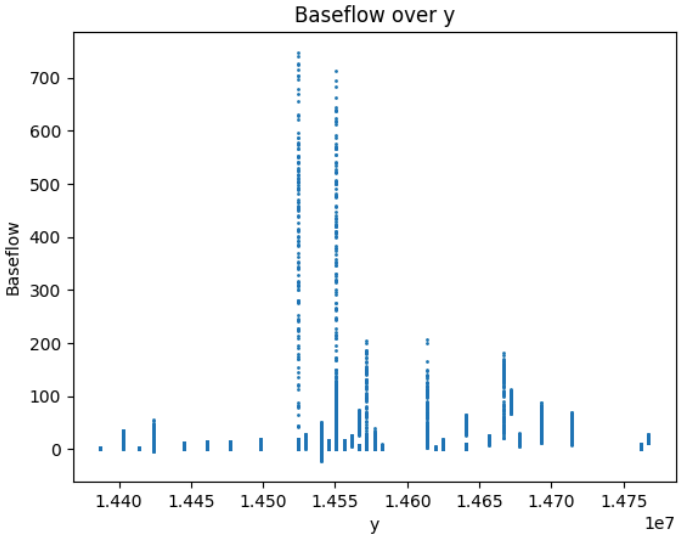


Fig 2

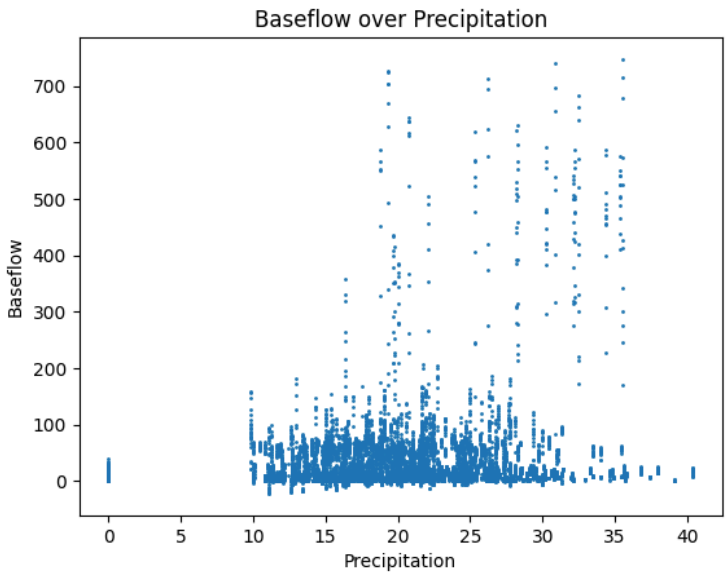
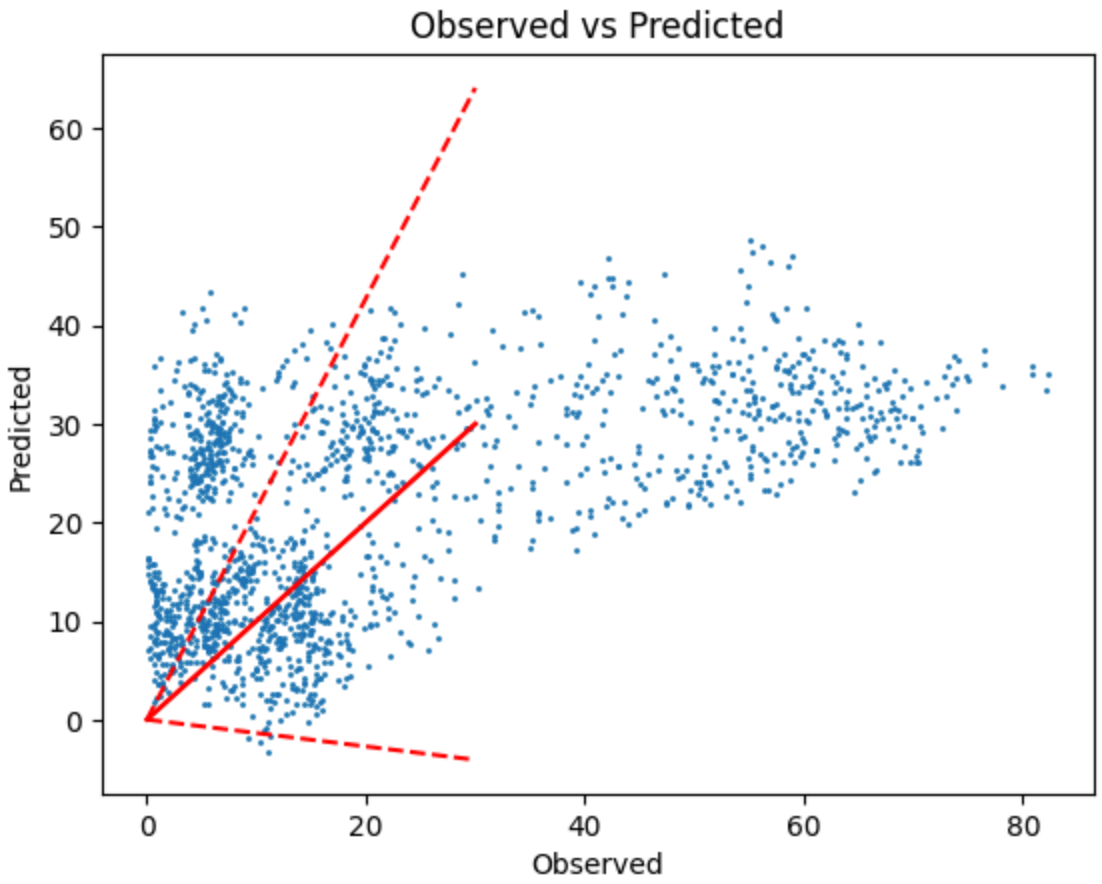


Fig 3

Following the filtering of our dataset we decided to use a multi feature linear regression model which had four key features: 'week', 'year', 'Segment_id', 'Evapotranspiration'. We used these features as the weeks would be able to account for seasonal changes more accurately than months, years could correlate to the changes in human based irrigation pumping, segment id could account for areas of low average baseflow, and evapotranspiration would be able to account for the monthly erosion/weathering of the area in which the sample was taken.

Results

From our analysis we were able to come up with a model that had a .294 R^2 score. THis is quite low and we acknowledge that but it was the highest score that we could come up with with the time allotted. Some reasons that this score could be so low is the vast number of 0 baseflow recordings and the high variance within the dataset. With certain days ranging from 0 to over 700 in baseflow, creating an accurate model was very difficult. Below is a graph pertaining to our model along with a 20% test dataset. We see the linear line correlating to our findings as well as two dotted lines that contain a 95% accuracy rate with all of the data. What we would like to see is that those dotted lines were more parallel to one another as well as being closer together to the hard line the model actually came up with. Along with that since the chart is in 'Observed vs Predicted' we would have liked to see that the values created a better straight linear line from 0,0 to 80,60 which would have meant that the model was more accurate in its predictions with regards to the linear regression line. Following the chart there is the output of our analysis and we cannot reject the null hypothesis for our remaining featureset as they do not contain a significant p value.



Results fig

Dep. Variable:	Observed	R-squared:	0.294
Model:	OLS	Adj. R-squared:	0.294
Method:	Least Squares	F-statistic:	818.5
Date:	Fri, 22 Mar 2024	Prob (F-statistic):	0.00
Time:	09:34:51	Log-Likelihood:	-33466.
No. Observations:	7862	AIC:	6.694e+04
Df Residuals:	7857	BIC:	6.698e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	356.9012	22.336	15.979	0.000	313.118	400.685

week	-0.0571	0.013	-4.481	0.000	-0.082	-0.032
year	-0.1808	0.011	-15.970	0.000	-0.203	-0.159
Segment_id	0.2282	0.004	53.497	0.000	0.220	0.237
Evapotranspiration	-0.6695	0.061	-10.954	0.000	-0.789	-0.550

Output of sm.OLS(y,X).fit()

Technical

In order to figure out what data should be used when, we decided to take out outliers initially. In addition to this, we removed all irrigation values that were equal to 0 since irrigation does play a significant impact in farming. We decided to do individual r squared value finders for each of the attributes. The r^2 values were so low that we didn't use those in this report but if you want to see what was done you can look at our code. Another mistake that was made was looking at the correlation between attributes. Since all the attributes were not super significantly correlated to the final outcome, we had to forgo analyzing correlation coefficients and move to r^2 values. However, it was overall, a great and fun project!