Empirical Analysis and Visualizations of Quantitative Data from Student Literacy and State Assessment

Project Increment 1

David L Downing
Thoa Nguyen
5310 Methods in Empirical Analysis
Dr Sayed Shah
Fall 2022
University of North Texas
https://github.com/dldowning/2022-5310/

Project Description:

1. Idea description

This is a data set that was collected in high school classrooms. There are approximately 400 observations with a dozen features. First, we will do some data cleaning to eliminate null value or duplicate data. Then, we will perform EDA to refer to the critical process of performing initial investigations on data. It will help us to discover patterns, to spot anomalies, to test hypotheses and check our assumptions with those summary statistics and graphical representations. Then, we will perform t-tests and ANOVA to look at statistically significant variations between groups. We will test assumptions of normalization and variance on the populations. Also, we will perform logistic regression and a decision tree model to try to predict one of the test scores given the other independent variables. The focus will be on exploratory data analysis, statistical tests, quantitative analysis, and visualizations.

2. Goals and Objectives:

We will perform some visualizations such as Grouped Bar plot, Pie Chart, Histogram and Box plot to make data more straightforward and use these results to decide which methods should fit our predictions.

Our goal is to be comprehensive with our visualizations of the features that are available. We want to deeply explore the data and analyze what is available, so we can continue the procedure.

The objective will be to end with an understanding of the data that we have but also to communicate the results of our analysis through visualizations. We expect to finish with a predictor model that will be trained from our data set to predict the dependent variable of TOSLSTOT which is the total science literacy as determined by a well-documented assessment tool.

3. Motivation

There is a large enough data set with N ~= 400 to use for analysis of students in the North Texas region. This dataset has not been overmined so we would like to explore it to find what conclusions can be reached for our analysis. This will give us an opportunity to practice the skills developed in class and to extend our learning into a field that has a growing need for data science and machine learning.

4. Significance

Before making decisions with information, we want to ensure that the data based decisions are not done in haste. We want to make sure there is no bias, there is statistical significance, the predictions done are made with assumptions that are checked, and the metrics match the needs of the decisions we are making.

Being able to make data based decisions in an education environment is a powerful tool to add to the school district's ability to meet the needs of their learners. Knowing which features to use, what their analyses look like, and which are good predictor variables would make it easier to identify which students need which interventions.

5. Objectives

We want to be able to describe each of the feature variables in detail. We want a correlation matrix made between them with descriptive visualizations. We want to be able to compare that there are significant differences between populations. We want to be able to compare the linear regression of features to a decision tree prediction of their state assessment score based on other features.

y= TOSLSTOT

x=[]

The list of features that will be selected as independent variables for the model will be determined through empirical analysis and testing.

6. Features

We'll take some of the info from df.describe and put it here

Project Increment 1

```
Data columns (total 32 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Gender      1260 non-null   object 
 1   Teacher     1290 non-null   object 
 2   Period      1279 non-null   float64
 3   Student ID  443 non-null    object 
 4   Gender.1    441 non-null    object 
 5   Ethnicity   441 non-null    object 
 6   Economic    440 non-null    float64
 7   LEP         440 non-null    float64
 8   SCICOUR     439 non-null    object 
 9   CTE         438 non-null    float64
 10  TOT         438 non-null    object 
 11  BIO         425 non-null    float64
 12  ELA         423 non-null    float64
 13  Alg         418 non-null    float64
 14  GPA         440 non-null    float64
 15  TOSLS1      418 non-null    float64
 16  TOSLS2      418 non-null    float64
 17  TOSLS3      418 non-null    float64
 18  TOSLS4      418 non-null    float64
 19  TOSLS5      418 non-null    float64
 20  TOSLS6      418 non-null    float64
 21  TOSLS7      418 non-null    float64
 22  TOSLS8      418 non-null    float64
 23  TOSLS9      418 non-null    float64
 24  TOSLSTOT    418 non-null    float64
 25  BRAINS1     450 non-null    float64
 26  BRAINS2     450 non-null    float64
 27  BRAINS3     450 non-null    float64
 28  BRAINS4     450 non-null    float64
 29  BRAINS5     450 non-null    float64
 30  BRAINSTOT   450 non-null    float64
 31  SORT        1280 non-null   float64
dtypes: float64(25), object(7)
memory usage: 323.9+ KB
```

```
"""Data Dictionary - we made this to help us with our interpretation
this dictionary created by us from reading the research papers

Gender          M or F
Teacher         categorical grouping of students
Period         numeric category for grouping students by location
Student ID      identifier for student
Gender.1        1=M, 0=F
Ethnicity       0=American Indian, 1=Asian, 2=Black, 3=Hispanic 4=Two or More, 5=White ** Do piechart
Economic       1=in economic need, 0=not in economic need (defined by free and reduced lunch program)
LEP            0 = limited english proficiency, 1=proficient in english
SCICOUR         number of  course credits earned in science classes
CTE            number of course credits earned in career/tech/engineering classes
TOT             ?
BIO            grade on biology state assessment
ELA            grade on english state assessment
Alg            grade on algebra state assessment
GPA            grade point average in high school
TOSLS1         these are measurements of scientific literacy
TOSLS2         float64
TOSLS3         float64
TOSLS4         float64
TOSLS5         float64
TOSLS6         float64
TOSLS7         float64
TOSLS8         float64
TOSLS9         float64
TOSLSTOT       I think this is total of TOSLS 1 through 9?
BRAINS1        float64
BRAINS2        float64
BRAINS3        float64
BRAINS4        float64
BRAINS5        float64
BRAINSTOT      "Behavior, related attitudes, and intentions towards science" survey info
SORT           float64
"""
```

Related Work (Background)

For Increment 1, our main approach will be Exploratory Data Analysis to understand the data and figures. Also, by making some visualization, we have the general picture of the correlations and the relationships among data. We have done a first pass at using a random forest regressor to predict the outcome variable. The research that was done before used a multiple linear regression so we are looking to improve upon their results.

Dataset

This dataset was taken from a high school. Some of the data is census data, some is test data, some is records data, and some is survey data. We obtained it from a journal search of published dissertations through the library. A copy of the raw data is available in our github.

Details design of Features

We have some categorical features and some continuous features. We did some cleaning to get the features loaded into our model and run the regression. We plan to further do some one hot encoding on some of the features such as the ethnicity feature. We also intend to check if applying a minmaxscaler or some normalization will boost our RMSE scores.

Analysis

The main analysis idea for this report will focus on how to perform data visualization and their results. So far, we have done some early visualizations and reported some descriptive statistics. Some interesting stuff in this correlation heatmap. Highest correlation is ELA and BIO which is the English and Science tests. you might think Math and Science would be higher. Unsurprisingly, the 3 standardized tests (BIO, ELA, ALG) and GPA are much more correlated than anything else. It's sad socially that LEP and Economics are somewhat correlated, but nice that LEP and GPA are not correlated.

Implementation

In the Increment 1 report, we will perform the Exploratory Data Analysis (EDA). EDA is the process of visualizing and analyzing data to extract insights from data.
The process will be involved:

1. Understanding the data

- Checking for general information of the dataset

```
[ ] df.describe()
```

	Period	Economic	LEP	CTE	BIO	ELA	
count	1279.000000	440.000000	440.000000	438.000000	425.000000	423.000000	418.0
mean	4.517592	0.411364	0.115909	0.760868	81.402353	80.529551	79.9
std	2.063422	4.252440	1.166238	1.042858	13.445400	10.726388	15.1
min	1.000000	0.000000	0.000000	0.000000	38.000000	25.000000	9.0
25%	3.000000	0.000000	0.000000	0.000000	74.000000	74.000000	70.2
50%	4.000000	0.000000	0.000000	0.500000	84.000000	82.000000	83.0
75%	6.000000	0.000000	0.000000	1.000000	92.000000	88.000000	93.0
max	8.000000	89.000000	24.000000	5.000000	100.000000	100.000000	100.0

8 rows × 25 columns

- Checking for data types

```
[ ] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1295 entries, 0 to 1294
Data columns (total 32 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Gender      1260 non-null   object 
 1   Teacher     1290 non-null   object 
 2   Period      1279 non-null   float64
 3   Student ID  443 non-null    object 
 4   Gender.1    441 non-null    object 
 5   Ethnicity   441 non-null    object 
 6   Economic    440 non-null    float64
 7   LEP         440 non-null    float64
 8   SCICOUR     439 non-null    object 
 9   CTE         438 non-null    float64
 10  TOT         438 non-null    object 
 11  BIO         425 non-null    float64
 12  ELA         423 non-null    float64
 13  Alg         418 non-null    float64
 14  GPA         440 non-null    float64
 15  TOSLS1      418 non-null    float64
 16  TOSLS2      418 non-null    float64
 17  TOSLS3      418 non-null    float64
 18  TOSLS4      418 non-null    float64
 19  TOSLS5      418 non-null    float64
 20  TOSLS6      418 non-null    float64
 21  TOSLS7      418 non-null    float64
 22  TOSLS8      418 non-null    float64
 23  TOSLS9      418 non-null    float64
 24  TOSLSTOT    418 non-null    float64
 25  BRAINS1     450 non-null    float64
 26  BRAINS2     450 non-null    float64
 27  BRAINS3     450 non-null    float64
 28  BRAINS4     450 non-null    float64
 29  BRAINS5     450 non-null    float64
 30  BRAINSTOT   450 non-null    float64
 31  SORT        1280 non-null   float64
dtypes: float64(25), object(7)
memory usage: 323.9+ KB
```

- There are 1295 rows and 32 columns before cleaning data

```
r,c = df.shape
r,c
```

```
(1295, 32)
```

2. Data Cleaning

- Checking for null values and remove null values

```
[ ] df.isnull().sum()
```

```
Gender          35
Teacher          5
Period          16
Student ID     852
Gender.1       854
Ethnicity      854
Economic       855
LEP            855
SCICOUR        856
CTE            857
TOT            857
BIO            870
ELA            872
Alg            877
GPA            855
TOSLS1         877
TOSLS2         877
TOSLS3         877
TOSLS4         877
TOSLS5         877
TOSLS6         877
TOSLS7         877
TOSLS8         877
TOSLS9         877
TOSLSTOT       877
BRAINS1        845
BRAINS2        845
BRAINS3        845
BRAINS4        845
BRAINS5        845
BRAINSTOT      845
SORT            15
dtype: int64
```

```
[ ] df = df.dropna() # drop null values
    #consider using mean as well
```

- After removing the null, we have 360 rows and 32 columns

```
df.isnull()
```

	Gender	Teacher	Period	Student ID	Gender.1	Ethnicity	Economic	LEP	S
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...
434	False	False	False	False	False	False	False	False	
435	False	False	False	False	False	False	False	False	
436	False	False	False	False	False	False	False	False	
437	False	False	False	False	False	False	False	False	
438	False	False	False	False	False	False	False	False	

360 rows × 32 columns

```
[ ] r,c = df.shape
    r,c
```

```
(360, 32)
```

- Checking for duplicate data. If yes, remove them. There is no duplicate data, so **we still maintain 360 rows and 32 columns after cleaning data.**

```
[ ] df.drop_duplicates()
```

	Gender	Teacher	Period	Student ID	Gender.1	Ethnicity	Economic	LEP	SC
0	M	Lewis	4.0	43954	1	0	0.0	0.0	
1	F	Howell	6.0	47436	0	0	0.0	0.0	
2	m	Lewis	4.0	59755	0	0	0.0	0.0	
3	M	Marshall	5.0	35449	1	1	0.0	0.0	
4	M	Lewis	3.0	43956	1	1	0.0	0.0	
...	...	...	...	...	...	...	...	...	
434	M	Lehmann	2.0	57499	1	3	1.0	1.0	
435	F	Brennan	2.0	65507	0	3	1.0	1.0	
436	M	Howell	4.0	36145	1	5	1.0	1.0	
437	M	Howell	4.0	39634	1	5	1.0	1.0	
438	F	Nyholm	5.0	40738	0	5	1.0	1.0	

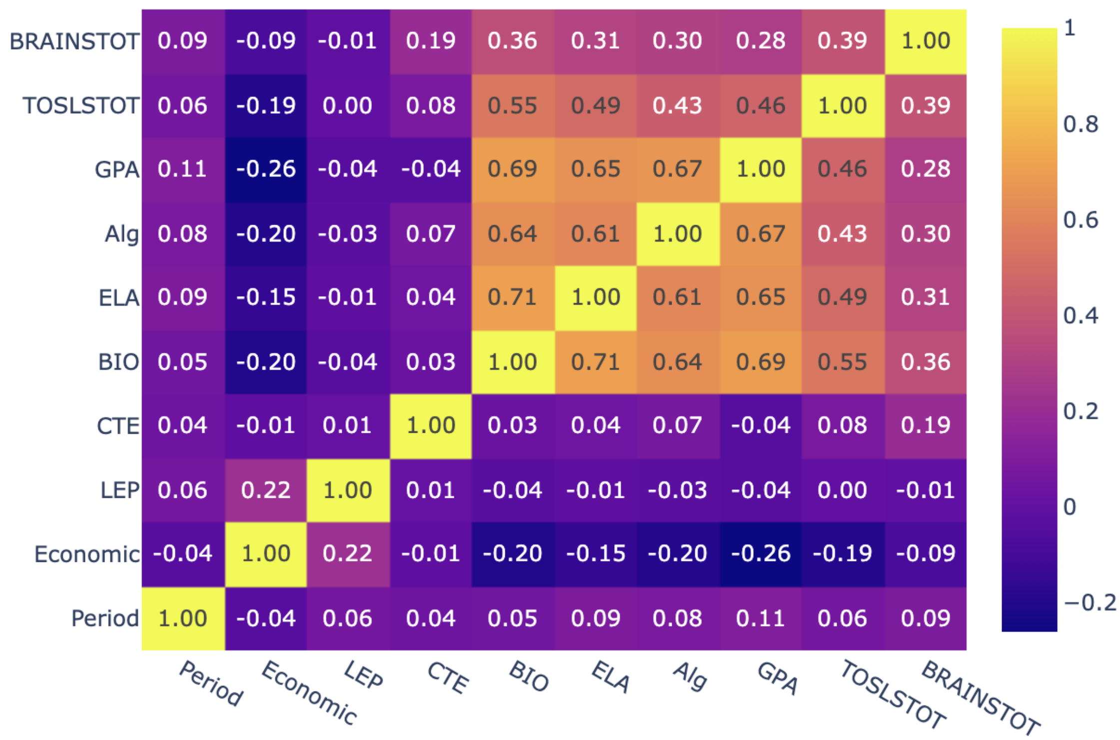
360 rows × 32 columns

3. Analyze the relationship between variables

In this step, we will perform the visualizations, such as heat map, histogram, scatter plot, bar chart, pie chart, … By analyzing these visualizations, we would be able to decide which modeling is best fit for the data
To visualize, we use plotly Python library.
In increment 2, we will improve the titles and axis labels for our visualizations. These are the most informative visualizations. We have more in our ipynb.

- Heat map

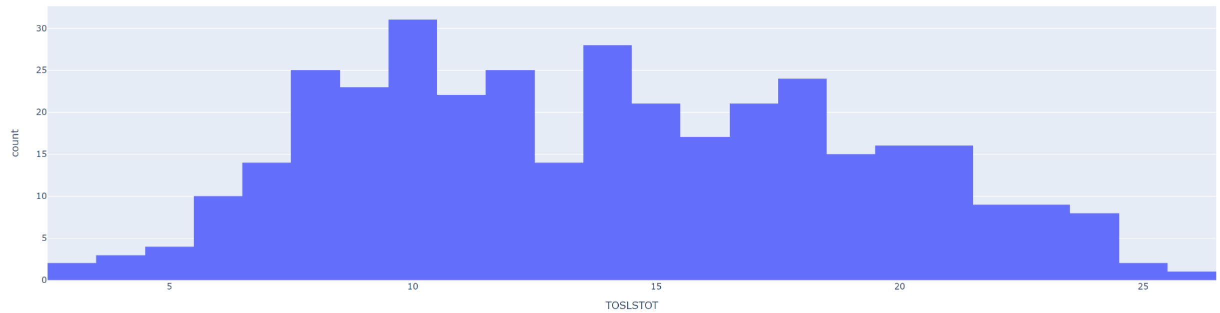
Project Increment 1



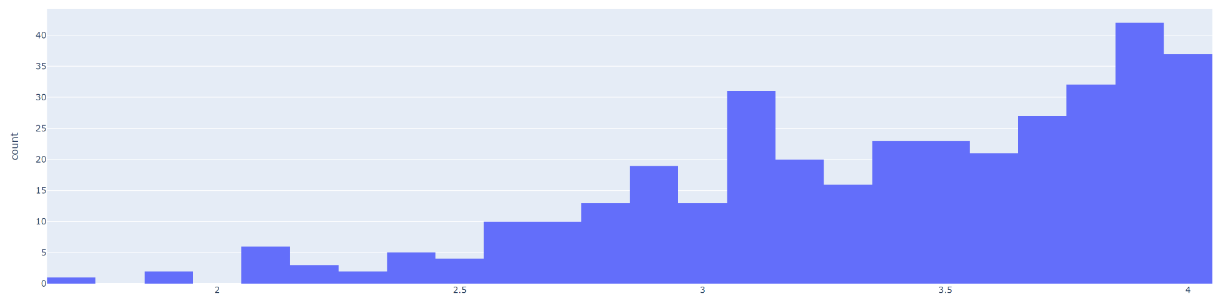
Some interesting stuff in this correlation heatmap. The highest correlation is ELA and BIO which is the English and Science tests. you might think Math and Science would be higher. What is unsurprising is the 3 standardized tests (BIO, ELA, ALG) and GPA are much more correlated than anything else. Its sad socially that LEP and Economic are somewhat correlated,, but nice that LEP and GPA are not correlated

- Histogram

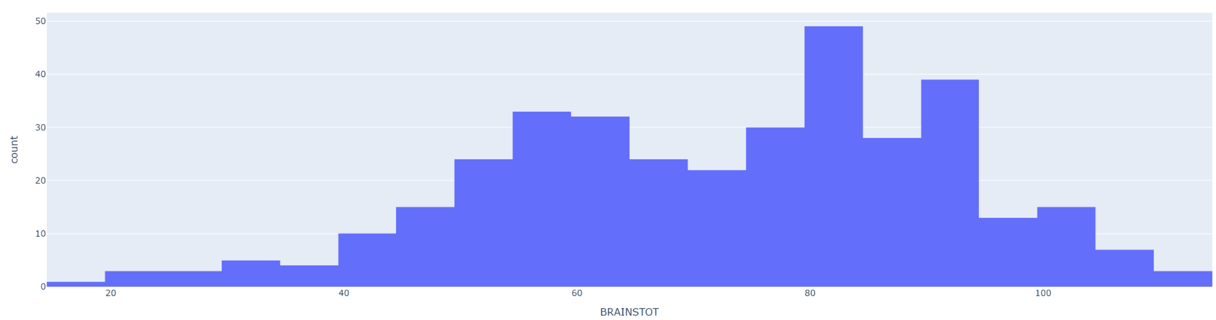
```
fig = px.histogram(df, x="TOSLSTOT")
fig.show()
#this looks fairly normal, might play with bin size
```




```
fig = px.histogram(df, x="GPA")
fig.show()
#not a normal distribution
```




```
fig = px.histogram(df, x="BRAINSTOT")
fig.show()
#interesting similar to TOSLSTOT
```

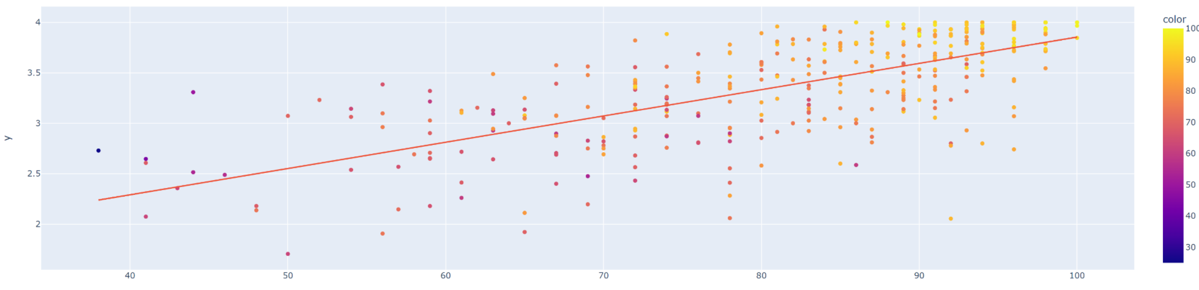


We see some normality in these continuous variables, but there is some skew as well. For increment 2, we will try different binning parameters.

- Scatter plot

Project Increment 1

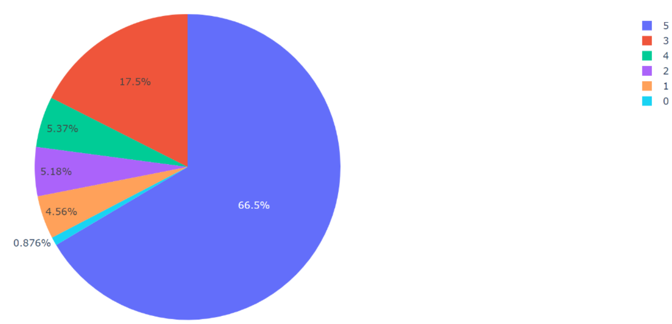
```
fig = px.scatter(x=df['BIO'], y=df['GPA'], color=df['ELA'], trendline="ols")
fig.show()
#can see the general pattern between BIO GPA, indicitive that a linear regression would do... okay. that's what the research paper did. we can do better.
```



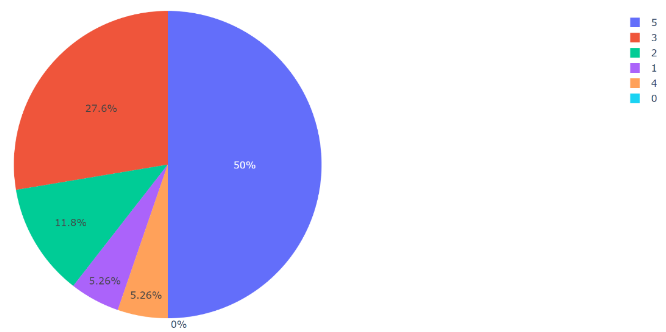
This scatter plot was guided by our heatmap. You can see the general pattern between BIO GPA, indicative that a linear regression would be moderately successful. That's what the research paper did. We can do better with our regressor.

- Pie chart

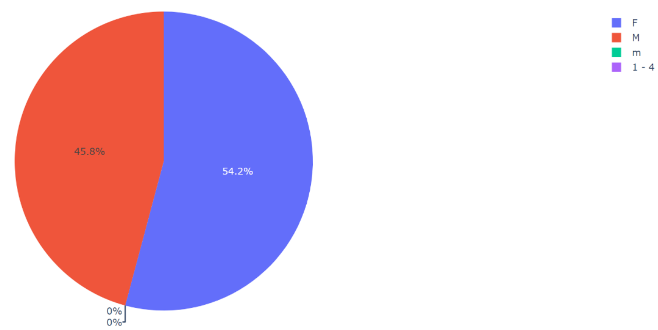
```
fig = px.pie(df, values='GPA', names='Ethnicity')
fig.show()
```



```
fig = px.pie(df, values='Economic', names='Ethnicity')
fig.show()
```




```
fig = px.pie(df, values='LEP', names='Gender')
fig.show()
```



We can see we have some class imbalance we will need to keep an eye out for in our regressor. We also have some cleaning of our gender variable to improve upon in increment 2.

4. Modeling the data

We used the XGBoost random forest regressor for our first model.
X=df.drop(columns=['BIO', 'Gender', 'Teacher', 'Period', 'Student ID', 'SORT', 'Ethnicity', 'SCICOUR', 'TOT', 'Gender.1']) #we might try one-hot encoding some categoricals
y=df['BIO'] #this doesn't have to be our dependent variable, but its sufficient for testing

We haven't performed hypertuning and will implement some improvements for increment 2.

Preliminary Results

We ended up with an RMSE of 9.24939117386846 which we found acceptable. We found this to be an acceptable result prior to hypertuning of parameters and engineering some of the features. We wanted to ensure that our model would accept our inputs and we'd have interpretable results. We will continue to refine our model for increment 2.

```
[179] import xgboost as xgb
      from sklearn.model_selection import train_test_split
      from sklearn.metrics import mean_squared_error
      X=df.drop(columns=['BIO', 'Gender', 'Teacher', 'Period', 'Student ID', 'SORT', 'Ethnicity', 'SCICOUR', 'TOT', 'Gender.1']) #we might try one-hot enco
      y=df['BIO'] #this doesn't have to be our dependent variable, but its sufficient for testing
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
      regressor = xgb.XGBRegressor(n_estimators=100,reg_lambda=1,gamma=0,max_depth=3)
      regressor.fit(X_train, y_train)
      y_preds=regressor.predict(X_test)
      RMSE = np.sqrt(mean_squared_error(y_test, y_preds))
      print("The RMSE is " + str(RMSE))

      [04:33:42] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
      The RMSE is 9.24939117386846

[ ]   """
      9.24939117386846 is an acceptable RMSE, but we can try to improve it
      at least we showed our model does work
      we will try to use the visualizations to guide our work on the features for increment 2
      """
```

Project Management

We choose the CRISP-DM methodology. However, we would simplify the CRISP-DM, which fits our project. Since our project is mainly understanding the insights of data and make the models, do some predictions. Below is the breakdown steps we would do for our project based on CRISP-DM method

- Step 1: Understand the topic, requirements
- Step 2: Collect and understand the data
- Step 3: Data preparation: Cleaning and perform visualizations
- Step 4: Modeling, select which models fit our data; Generate test and predictions
- Step 5: Interpret the results. We need the summary of insights data

Implementation status report

- Work completed

Task	Description	Contribution - Percentage
Cleaning the dataset	Drop null and duplicate data	Thoa
Implemented Exploratory Analysis		Thoa (50%) / David (50%)
Computed the RMSE	Modeling	David

- Work to be completed

Task	Description	Contribution - Percentage

Project Increment 1

Imputing	Try setting nulls to mean instead of dropping	Thoa
One Hot Encoding	On categoricals	Thoa / David
Hypertuning parameters	Improve model	David

References

1. Chandler, J. R. (2020). Predicting science literacy: A multiple regression model of factors that influence science literacy (Order No. 28031723). Available from ProQuest Dissertations & Theses Global. (2437410299). Retrieved from https://libproxy.library.unt.edu/login?url=https://www.proquest.com/dissertations-theses/predicting-science-literacy-multiple-regression/docview/2437410299/se-2
2. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
3. Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).
4. Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. CBE—Life Sciences Education, 11(4), 364-377.