C

**Empirical Analysis and Visualizations of Quantitative Data**

**from Student Literacy and State Assessment**

Project Increment 2

David L Downing

Thoa Nguyen

5310 Methods in Empirical Analysis

Dr Sayed Shah

Fall 2022

University of North Texas

Code:

<https://github.com/dldowning/2022-5310/>

Video:

<https://drive.google.com/file/d/1MnoGON535GCnSx6MU7pec_r5h7RsJ4ZD/view>

# 

### 

## 1. Introduction

High schools can better serve their students when they can see problems in advance. 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 (Tukey, J. W., 1977). 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. 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 (Chen, T., & Guestrin, C., 2016).

## 2. Background (Related Work)

The research that was done before used a multiple linear regression so we are looking to improve upon their results. The multiple linear regression variables and the coefficients are documented in the publication. (Chandler, J. R., 2020). We will implement their predictor model as a python function and then set to improve upon it with our own machine learning model.

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. 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.

## 3. Models

### Workflow diagram with explanation

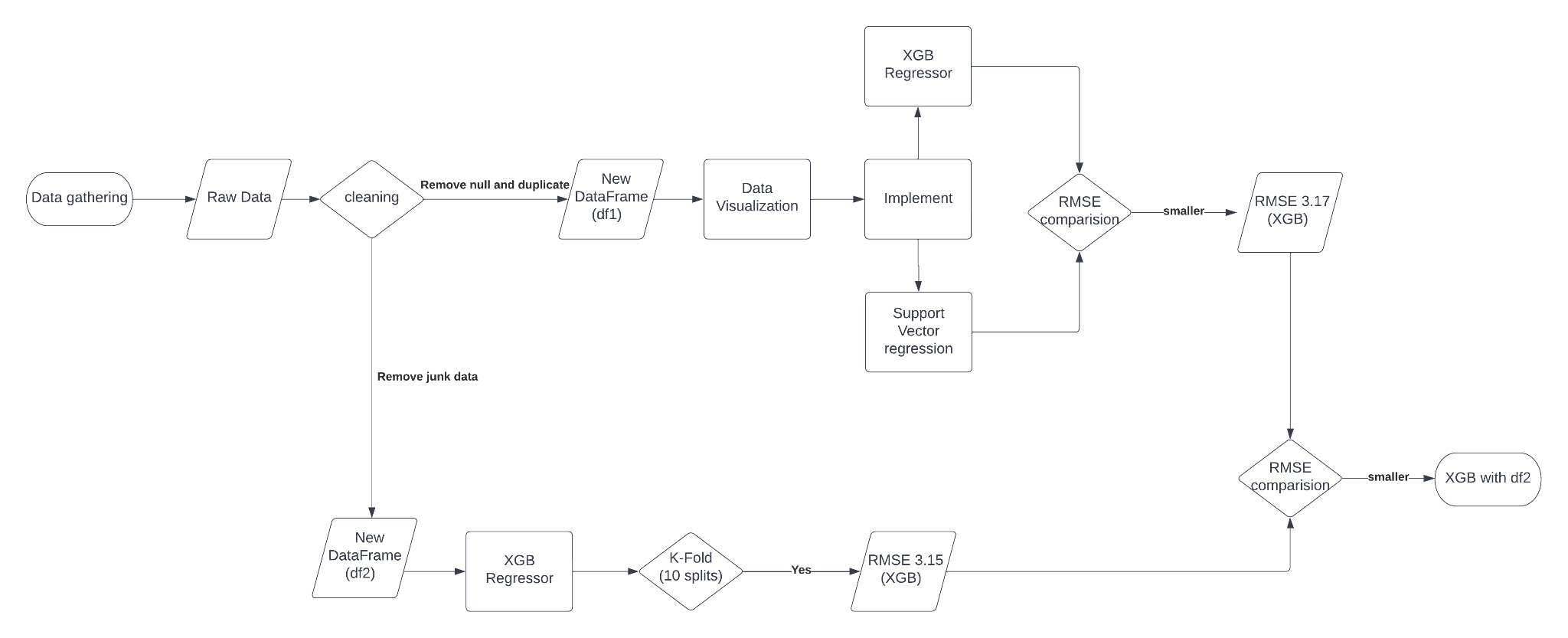


The above workflow demonstrates how we process the data and choose which model to make the best prediction. First, we collect the data from the research paper, which is used in Predicting Science Literacy: A Multiple Regression Model of Factors that Influence Science Literacy (Chandler, J. R., 2020). Then, we perform the EDA which includes the data cleaning and data visualization. For the data cleaning step, our first approach is to remove all the null values and duplicate data. We got the new dataset which has 360 rows. Then, we perform data visualizations such as heat map and some scatter plots, bar charts. Then, we construct the XGB Random Forest Regressor and Support Vector Regression. Next, we compare these two Root Mean Square Error (RMSE) to decide which model best fits our dataset.

### Complete workflow diagram

We’ve tried to improve the XGBoost Random Forest Regressor by any chance. Hence, our second approach is to generate the new dataset with the second cleaning method. Instead of removing the null values and duplicate values, we remove the junk data which occurs at the bottom of the .csv file. Our second approach is to try to avoid dropping as many rows as possible. The second dataset has 399 rows (39 more rows compared to the first dataset).

We construct the XGBoost with the new dataset to check if it can improve the RMSE.





With the second cleaning approach, the RMSE is 3.15 (decreased by 0.02 compared to the first cleaning approach). So far, it is the best model we could obtain.

It wasn’t clear which cleaning method would be superior so the answer was obtained through experimentation with a quantitative outcome.

## 4. Dataset

### Detailed description of 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.

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.

### Detail design of feature

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 featre. We also intend to check if applying a minmaxscaler or some normalization will boost our RMSE scoreus.

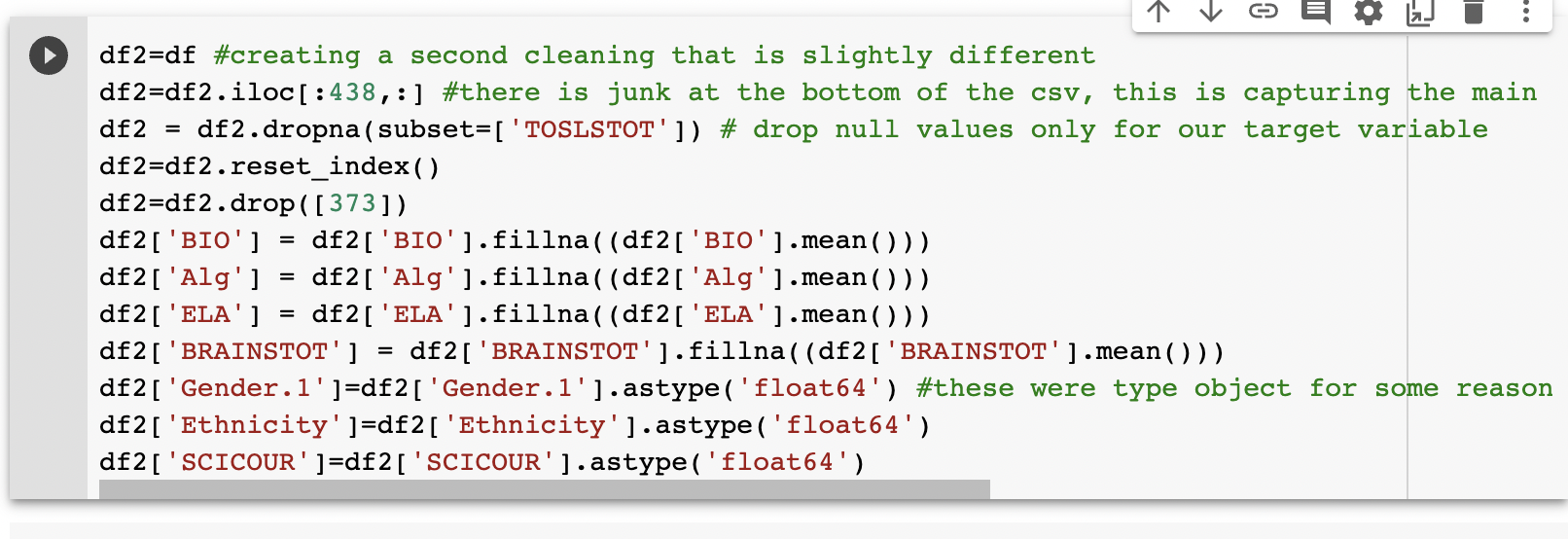
## 5. Analysis of data

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.

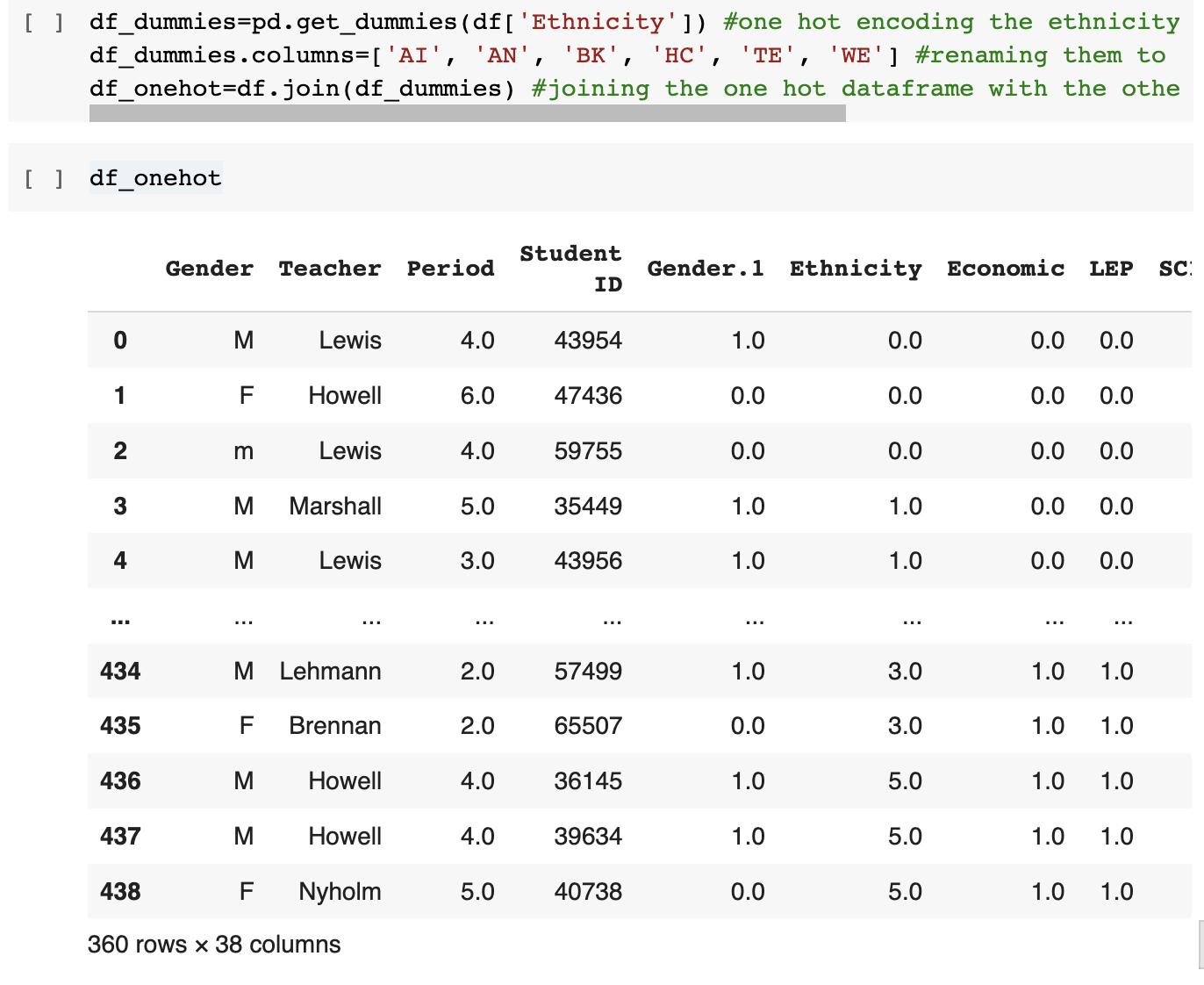
Unsurprisingly, the 3 standardized tests (BIO, ELA, ALG) and GPA are much more correlated than anything else. It's sad socially that LEP and Economics are somewhat correlated, but nice that LEP and GPA are not correlated.

## 6. Implementation

* We created the second cleaning to check if we would have the difference in RMSE. Instead of removing the null values and duplicate values, we remove the junk data at the bottom of the .csv file. Our approach is to avoid dropping as many as possible, and only drop null values for our target variable. With this approach, the new dataframe has 399 rows and 33 columns

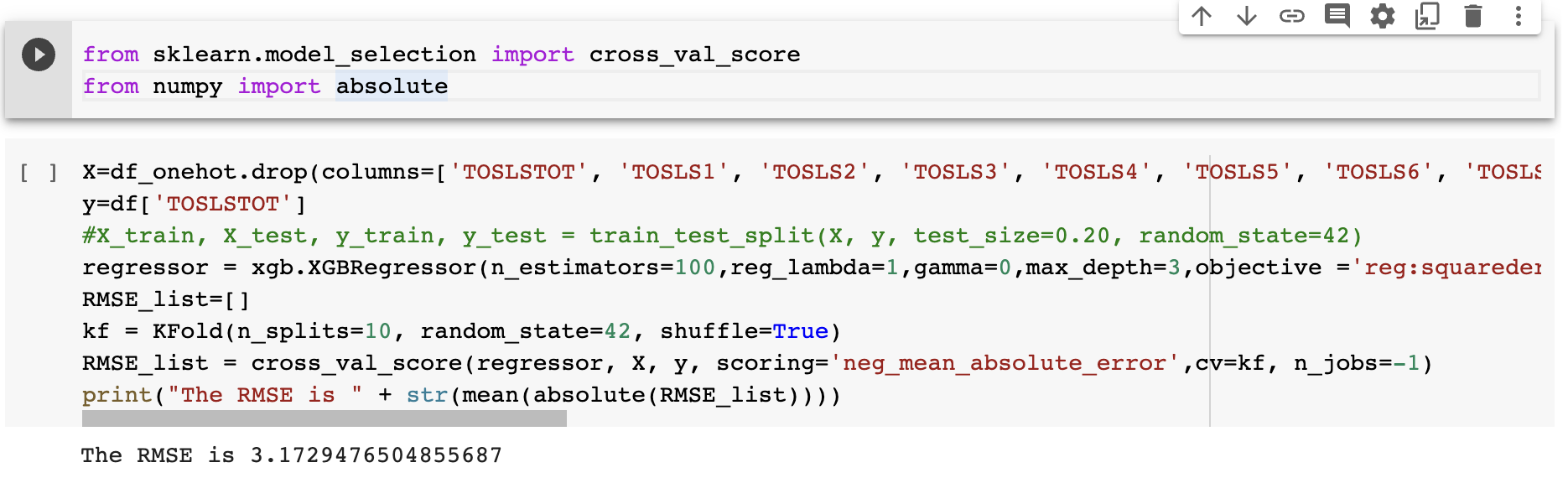


* Apply one hot encoding to Ethnicity variables with the 360 row dataframe

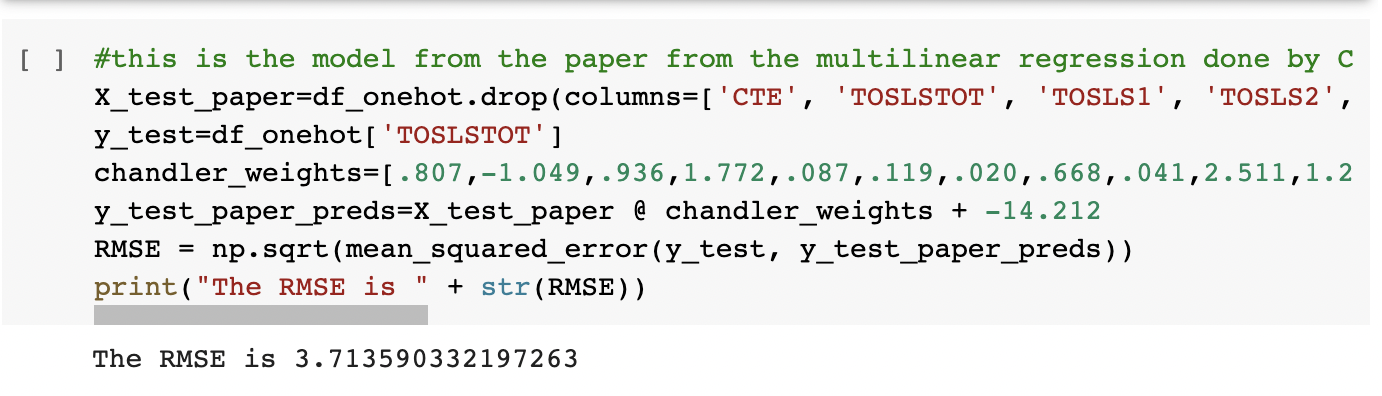


* Apply XGB Random Forest Regressor, K-fold cross validation and 10 split. We got Root Mean Square Error is 3.17

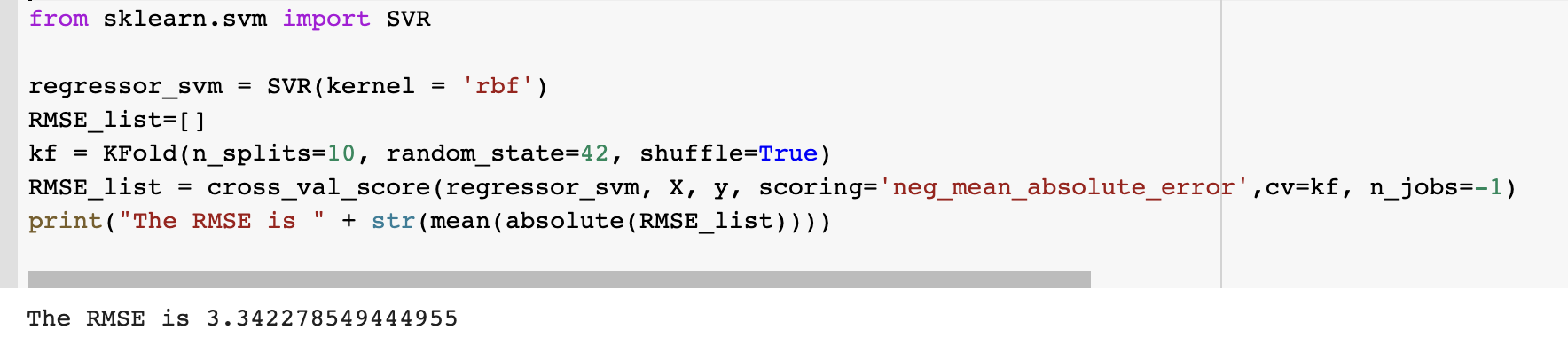
We added cross validation in increment 2. It had a dramatic impact on reducing the RMSE since we had so few training samples. Preserving more training was a significant boost in performance.



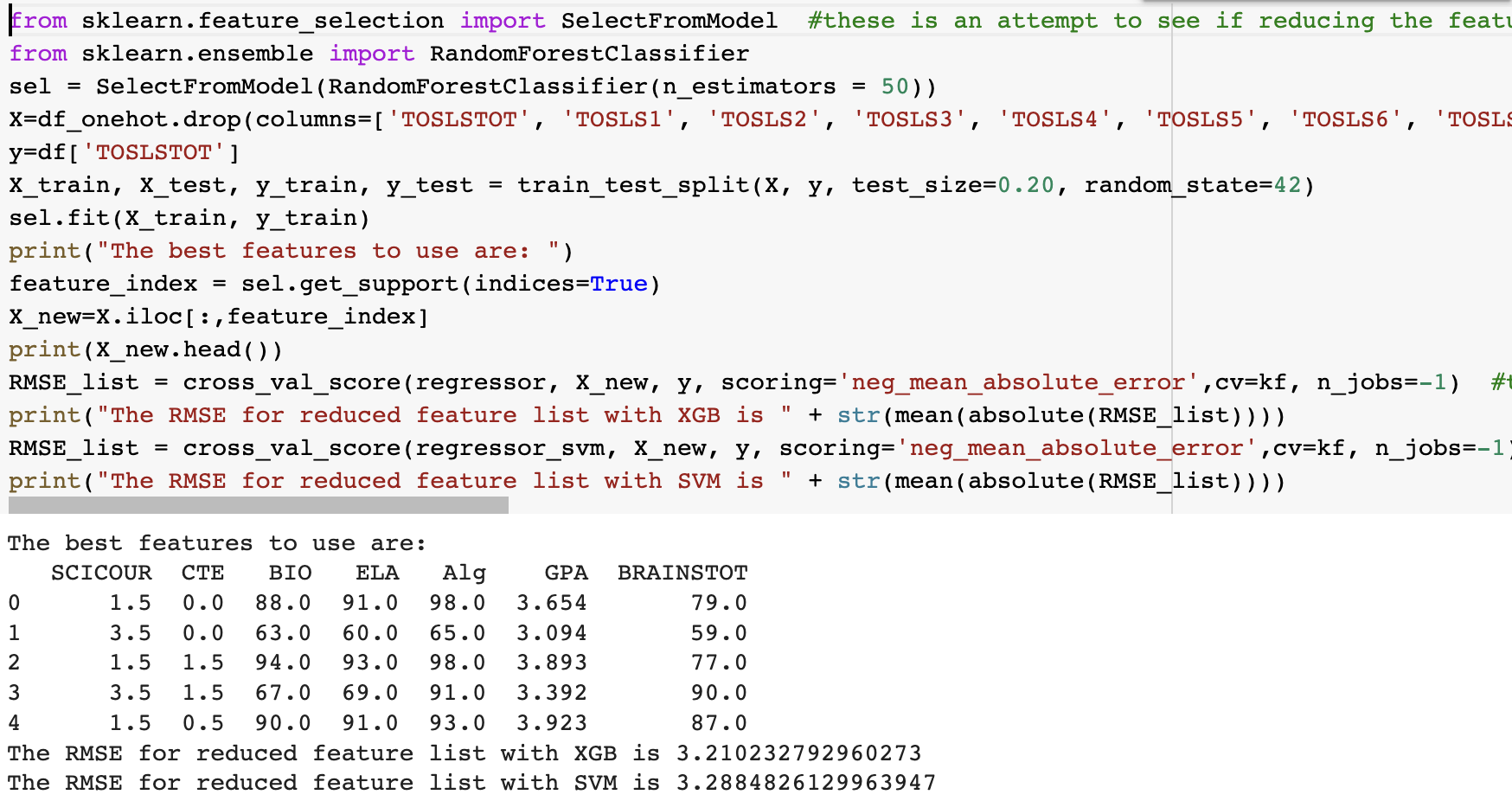
* We compare the XGB model vs the multilinear regression which was done in the paper. The multilinear regression model ended up with RMSE 3.71



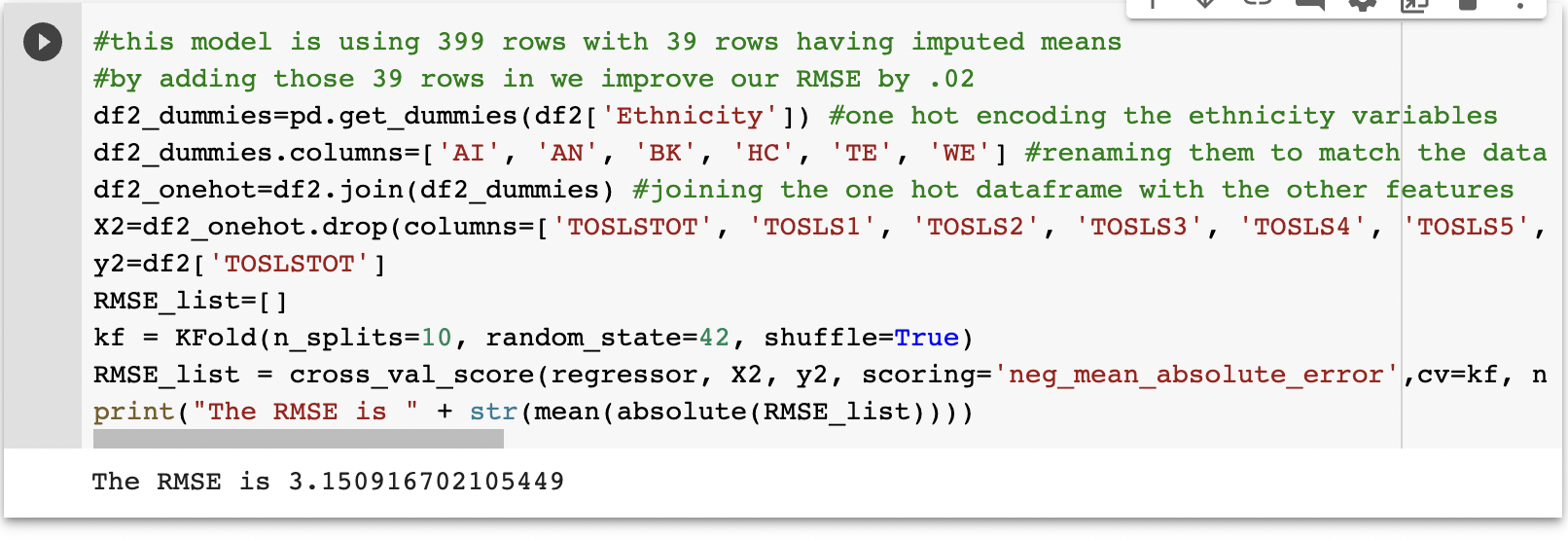
* We also use Support Vector regression as a comparison to our XGBoost Regressor, the RMSE is 3.34



* There is an attempt to see if reducing the feature list using SelectFromModel will improve the score. However, the improvement is not better than XGB. SVM has the RMSE is 3.28, while XGB has RMSE is 3.21 The improvement in SVM is likely because with fewer dimensions it was able to better slice the multidimensional space but the XGBoost was probably not increased because random forests are already robust to poor features in the way it divides the branches of the tree.



* Apply one hot coding, XGB and K - fold cross validation is 10 split for the new dataframe. This model is using 399 rows with 39 rows having imputed means. By adding those 39 rows, we improve the RMSE by 0.02 (the RMSE of the old data frame which has 360 rows is 3.17)



## 7. Results

* Our best model was the XGBoost regressor. It performed significantly better than the reference model that was developed by the published paper. They used simple statistical methods to develop a multiple linear regression. We implemented the method using the published coefficients and obtained an RMSE of 3.71 which was easily exceeded with our machine learning methods.
* This demonstrates that we were able to understand and make use of the feature space that we had available to us. We executed an experiment in increment 2 where we did the cleaning in a second way. Instead of dropping null values, we only dropped rows that had nulls for our target value and imputed the mean for any remaining nulls. This preserved 39 more samples and gave a slight boost to our RMSE. This is a quantitative proof that the cleaning method was effective and that imputing the mean was valid for the features we had available.

| DF1 (dataframe with 360 rows) | Model | RMSE |
| --- | --- | --- |
| XGB Regressor | 3.17 |
| Multilinear Regression (By Chandler et al. quoted) | 3.71 |
| Support Vector regression | 3.34 |
| DF2 (dataframe with 399 rows) | XGB Regressor | 3.15 |

## 8. Project Management

Below is the breakdown steps we would do for our project:

* 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 in increment 1

| 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 completed in increment 2

| Task | Description | Contribution - Percentage |
| --- | --- | --- |
| Imputing | Try setting nulls to mean instead of dropping | Thoa |
| One Hot Encoding | On categoricals | Thoa (50%)/ David (50%) |
| 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. *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.