**Empirical Analysis and Visualizations of Quantitative Data**

**from Student Literacy and State Assessment**

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/

## 

## 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. Literature Survey

Statistical modeling can help us check our assumptions that we made. We’ve learned about t-test Tukey’s test and ANOVA. We will check these assumptions mathematically but also visualize them. Methods for testing for significance and methods for analyzing assumptions have been well researched and documented in the literature.

We’ll use xgboost to see if what we ascertained from the exploratory data analysis and cleaning can be input into the xgboost algorithm that has a library to import. The literature shows this is a powerful technique. XGBoost can be used for classification but also for predicting continuous variables. We will compare its power to the linear regression techniques learned in class.

<https://www.lifescied.org/doi/full/10.1187/cbe.12-03-0026>

<http://www.ru.ac.bd/wp-content/uploads/sites/25/2019/03/102_05_01_Tukey-Exploratory-Data-Analysis-1977.pdf>

<https://dl.acm.org/doi/10.1145/2939672.2939785>

<https://xgboost.readthedocs.io/en/stable/>

## 6. Objectives

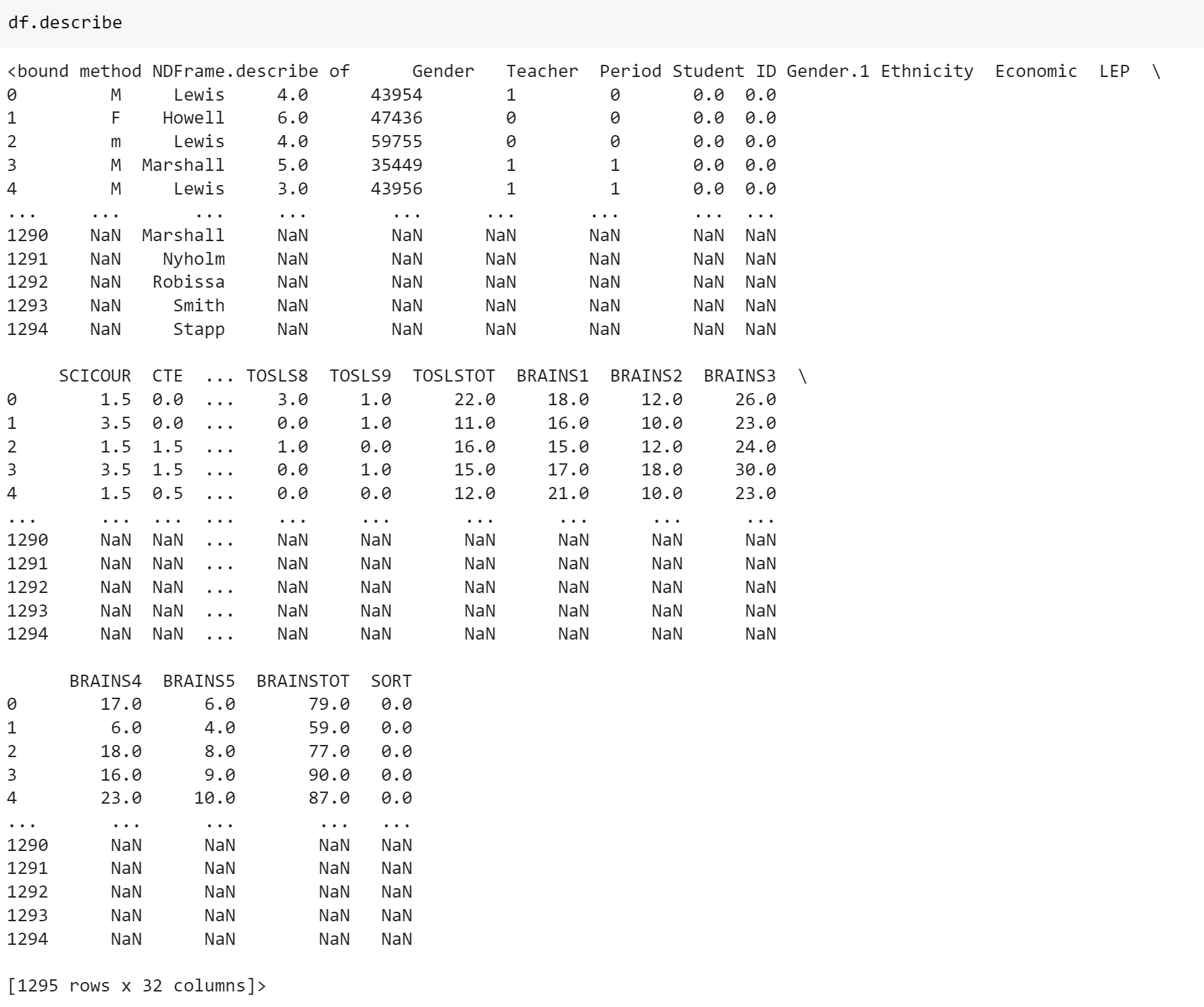
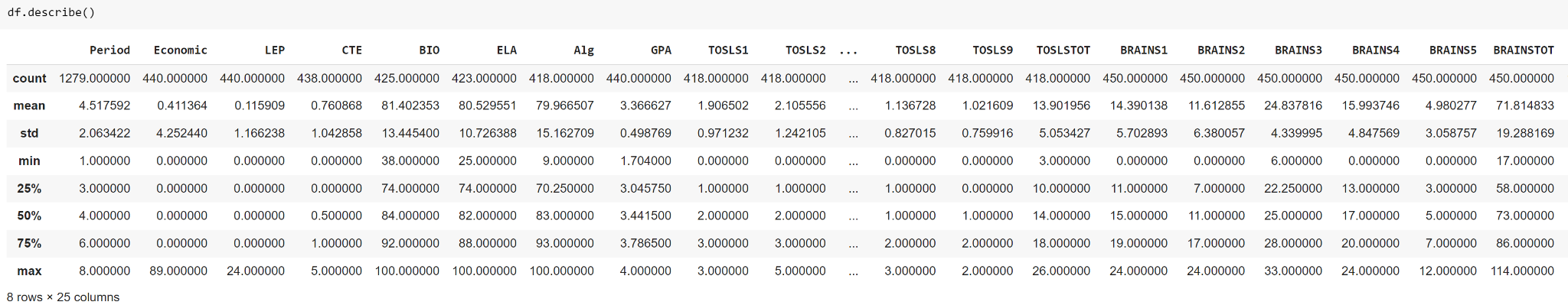
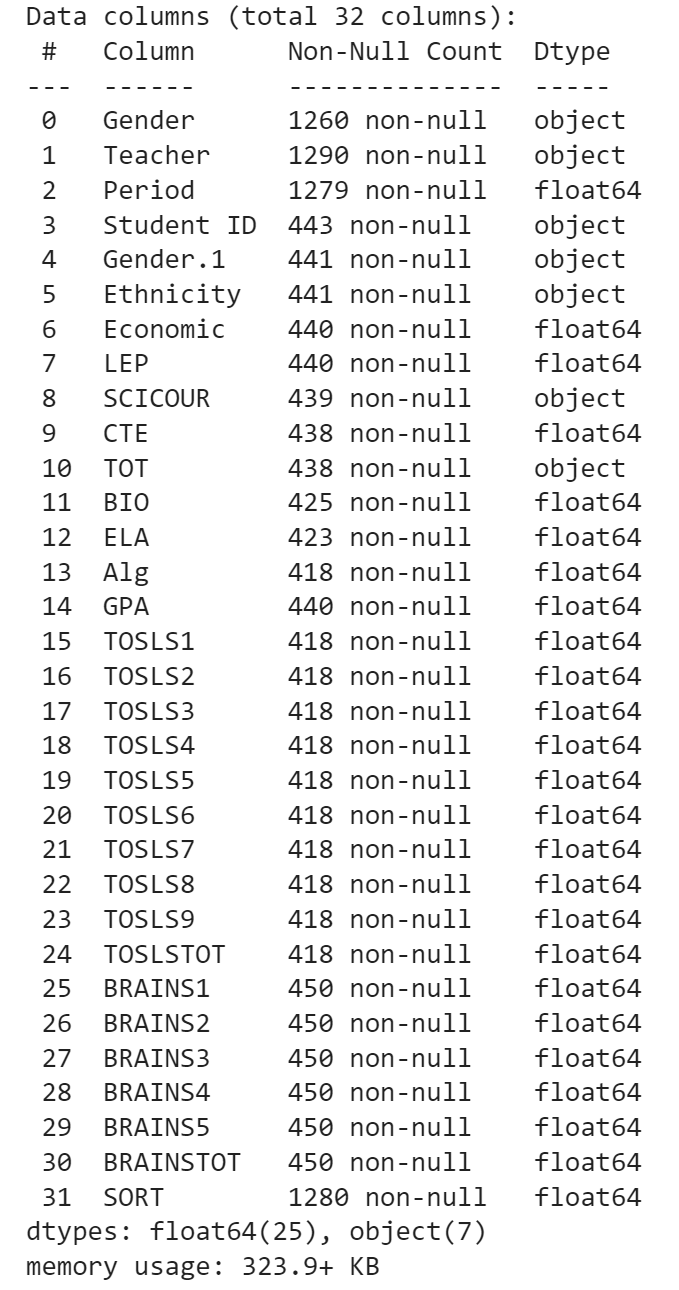
We want to be able to describe each of the feature variables in detail. We want a correlation matrix made between them with descriptive visualizations. We want to be able to compare that there are significant differences between populations. We want to be able to compare the linear regression of features 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.

## 7. Features

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

## 8. Expected outcome

We expect the exploratory data analysis to be easily interpretable by professionals outside of computer science. The expected outcome will be a report with visualizations to understand each of the independent variables that are used as predictors to get the value of TOSLSTOT. We expect to achieve a favorable metric in our prediction.

## 

## 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.