Background:

[Thomas Jefferson High School for Science and Technology](https://www.fcps.edu/registration/thomas-jefferson-admissions) consistently ranks as the nation’s top-performing high school. It is one of a few governor’s schools in Virginia and is located in Fairfax, Virginia. The community has often called for investigations into its admissions processes and even consideration of closure due to what many feel are unfair practices. Despite being a public school, there is limited transparency of its internal metrics and choices made.

High-Level Problem Statement:

I hope to analyze the variables influencing TJ admissions and identify their relationships on likelihood of entry.

Data Sources:

* Student applicant surveys that include:
  + Demographic information
  + Amount of test preparation
* Admissions test scores
* Scoring of teacher recommendations
* Scoring of student essays
* Average wealth/income of family (if attainable)

Plan to obtain data:

I have scheduled a meeting with the Assistant Superintendent who oversees TJ admissions soon and will be requesting all data sets. I hope to utilize my current position to overcome transparency barriers.

I anticipate all existing data to require a significant amount of data wrangling, at minimum to combine various data points and turn them into charts.

Methods planning to employ:

*Data processing*: First, I plan to utilize our data binding tools to put together the various spreadsheets I will obtain. I also anticipate that I may need to scrape data from various databases, possibly using for loops, to generate individual student data that is not housed in the same spreadsheets as admissions-related data (family income, for example).

I will then use data wrangling tools through dplyr to restructure the data to contain only the variables I am interested in studying. Given where I will be obtaining the spreadsheets (and that they have not been used for analysis before), I expect to clean up cells, utilize techniques to fill empty cells, and mutate entries to fit my needs.

*Data visualizations*: After conducting analyses that will reveal more information about the relationships among the variables, I plan to plot three to five variables of interest (e.g. race, income, test scores) on admission rate to reveal more information. I would hope to employ colors and other features in ggplot to plot three variables and produce at least one visualization that tells the broad TJ admissions story. I anticipate this may look like a grid.

Depending on bandwidth and what the data begins to tell, I may use topic models to understand the essay content and what subjects or themes, if any, lead to higher likelihood of admission.

*Machine learning*: First, I plan to use machine learning through recipe to impute values for empty data points. I would also have to form categorical variables of certain outcomes. I also anticipate using bootstrapping and tree diagrams/random forests to outline the predictive variables for TJ admissions, identifying the most prominent for further study.

Definition of Success:

A successful project would answer the question: what variables influence the likelihood of a student to be admitted into TJ? Or, more appropriately, what traits tended to be associated with students who were admitted into TJ?

It would accomplish the following:

* Compile data required to answer these questions from the school system
* Clean the data and consolidate it into one (or a few) file(s) that can be reused for future analysis
* Produce at least three graphs/plots/visualizations using variables of interest to reveal beneficial information about admissions trends for TJ
  + Beneficial here is defined as valuable for decision-making; has policy implications
* Run the most appropriate machine learning tool for statistical analysis (hypothesized here to be random trees) to reveal relationships of variables and likelihood of admission
* Collect all materials and information into a beneficial report

The minimum viability criteria for this project would be to clean the data, fill missing data using machine learning (imputation), and generate data visualizations that describe the relationships.