Preprocessing Data

- Selected a number of random samples from the entire dataset
  - Reservoir sampling to uniformly and randomly select samples
    - Used sample sizes of 10,000 and 500,000 samples
- Select a number of columns used as features to be used for machine learning models
  - Choose data that are related by job details and popularity
  - Average overall company rating, Description Length Chars, Clicks, Applies, Education Requirements, Cities, States
- Longitude and Latitude references for Cities
  - Use Google Maps to read in longitude and latitude strings
  - Show education requirement trends in United States regions

Machine Learning Models

- Decision Trees
  - Non ensemble learner for random forest classification. Mainly splits off one feature of the data. The split for children nodes is not always binary(20% accuracy)
- Random Forest Classification
  - Ensemble learning method consisting of many classification trees on subsets of the entire data set(10% accuracy)
- Support Vector Machines(Classification)
  - Minimize the distance between each data point with the line. Only able to run for the 10,000 samples(10% accuracy)

Data Visualization

- We tried to make a data visualization through querying google maps API
- By querying google maps API, we are able to get the longitude and latitude points of each major city
- Preprocessing the data through grabbing certain columns with querying proved to be too much
    - Columns queried: noEducationRequirementsJob, highSchoolEducationRequirementsJob, higherEducationRequirementsJob

For the Future

- Try grabbing more data and features to be able to distinguish between different classes
- The accuracy of the model increased with more data but the processing time also increased
- Use cluster to speed up computation to balance workload versus wait time
- Supply company activity information to users
- Analyze job type and date posted in relation to current events