CSE 545 Big Data Analytics

Project Proposal

SDG 4 - Quality Education

Team SPHS

What is it?

SDG 4 aims to “ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.”

Why should one care?

- 258 million children and adolescents remain out of school
- 617 million ie, around 55% lack minimum proficiency in reading and mathematics
- An estimated 50 per cent of out-of-school children of primary school age live in conflict-affected areas - crime rate. Income disparity, employment rate - analyse these features

What are we doing?

- In 2019, US cities were evaluated for SDG goal according in this report “The 2019 US Cities Sustainable Development Report”.
- Analyze factors like primary and secondary education rate, basic facilities that schools have, access to education within a region, youth skills and employment rate for different counties of USA.
- Gain interesting insights related to SDG 4.
- Find similarity between counties to identify factors that influence education quality.
- Project an outlook for the year 2030 based on time series analyses of the data.

Dataset

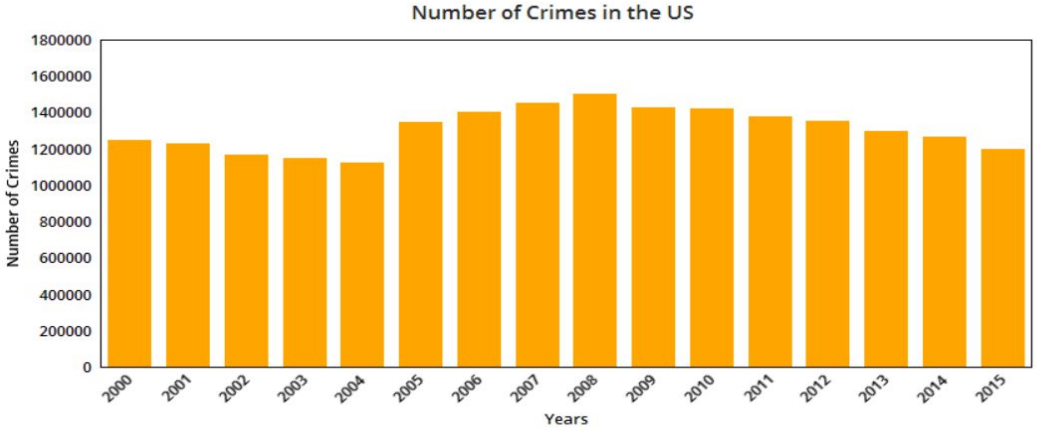
- **US College Scorecard data from 2010 - 2019**
  - Institution-level
  - 2000 features : course duration, pass % , tuition, demographic, grades, geographic location, stream of study, faculty salary, completion rate, student loan details etc.
  - **https://collegescorecard.ed.gov/data/**

- **US County level data**
  - Spanning employment rate ,crime rate and other economic indicators, demographics data
  - US county wise public school data https://www.kaggle.com/carlosaguayo/usa-public-schools
  - US county wise unemployment data https://data.ers.usda.gov/reports.aspx?ID=17828
  - US county wise crime related data https://www.kaggle.com/marshallproject/crime-rates
  - US county wise population data https://data.ers.usda.gov/reports.aspx?ID=17827
  - US Education Finances https://www.kaggle.com/noriuk/us-educational-finances
  - https://nces.ed.gov/fastfacts/display.asp?id=76

- Merge the datasets to obtain a richer set of features for hypothesis testing against quality of education.

- Use UN SDG 4 education quality indicators for reference

Exploratory Data Analysis



Analysis and Methodology

- Correlation analysis between various factors like crime, employment, student diversity etc and quality of education.

- Group data by county and aggregate the value for each feature.

  Data Frameworks:

- Wide data with more than 2000 features, requires distributed storage and parallel processing.
  - achieved by **HDFS**

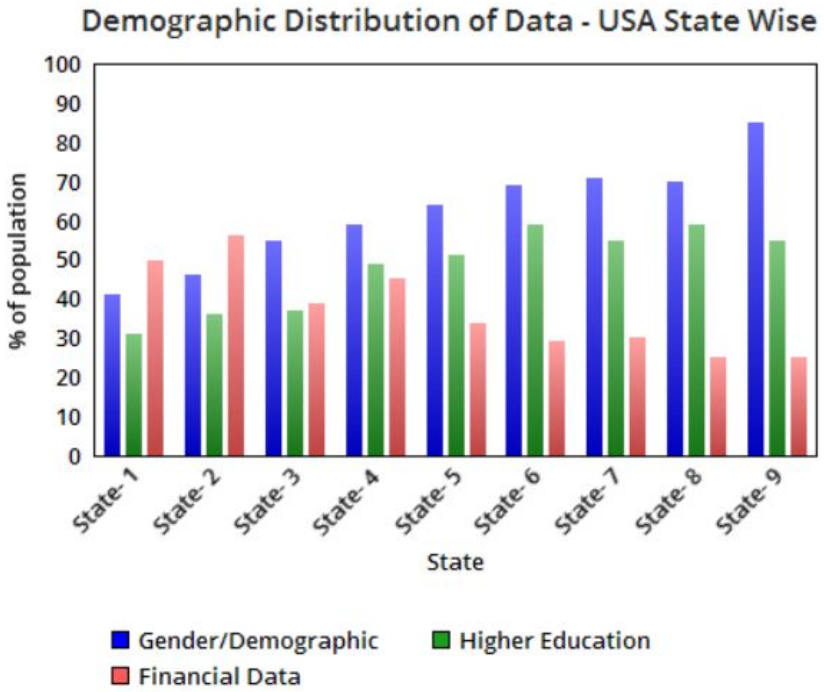
- Requires heavy transformations- data standardization, computations for correlation with multivariate regression analysis
  - achieved effectively by **Spark** transformations

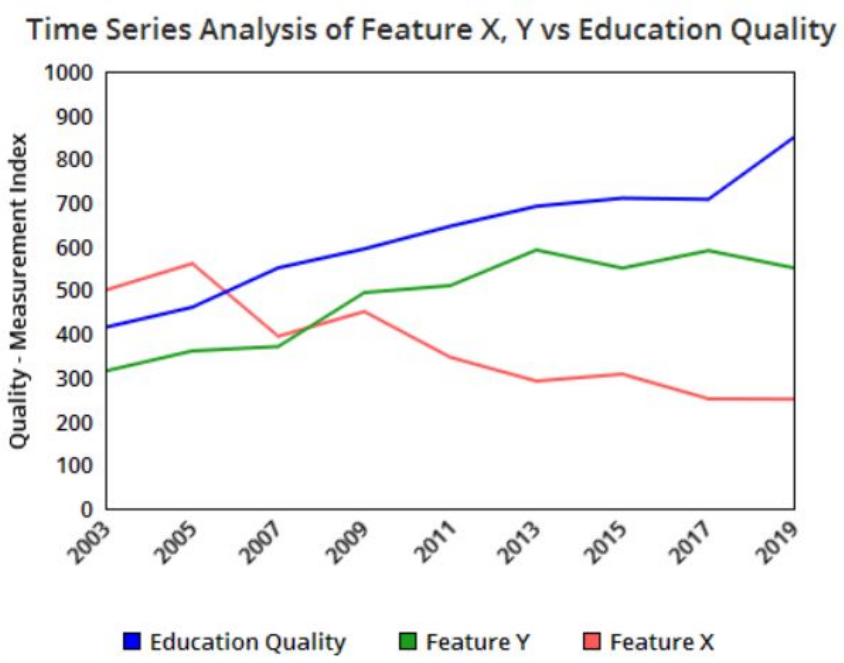
Analysis and Methodology

- **Hypothesis testing**
  Hypothesis: Feature X has no impact on education quality

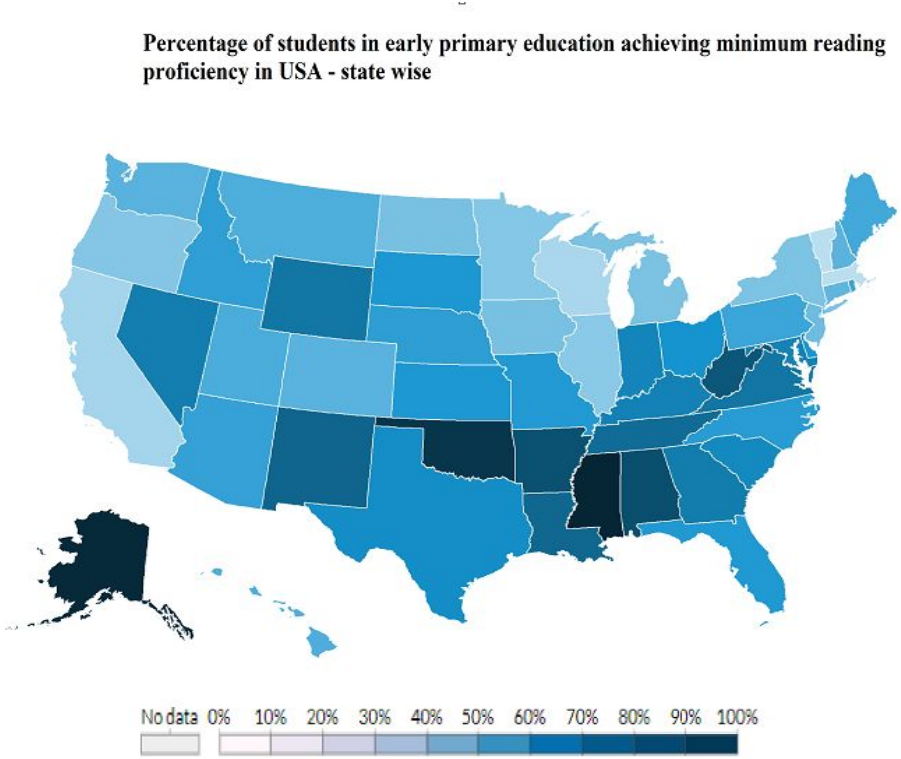
- **Similarity search**
  - Identify similar counties based on similarity metrics on factors like education quality, crime rate ex: Cosine Similarity
  - Identify or predict which factor is most influential for particular county to achieve the SDG.

Sample Projections





Sample Projections



Conclusion

- Infer influential factors from analysing dataset.
- Project an outlook for the year 2030 based on time series analyses of the data.
- Identify other factors like crime rate, population, income, etc that might affect education quality.