CSE 545 Big Data Analytics

Project

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.

Why big data?

- Study and analyze education quality at county level for USA
- High dimensional, wide data with rich set of features covering demographics, education streams, facilities
- Data covering colleges at zip-code level for every county
- Merged with census data for population
- Data spans years from 1997 to 2019 (we focus on years 2010-2018)
- Sparse data
- Inferences at granular levels, intuitive and non-intuitive findings
- Measure and quantify the current status of the education quality as defined by indicators

Goals

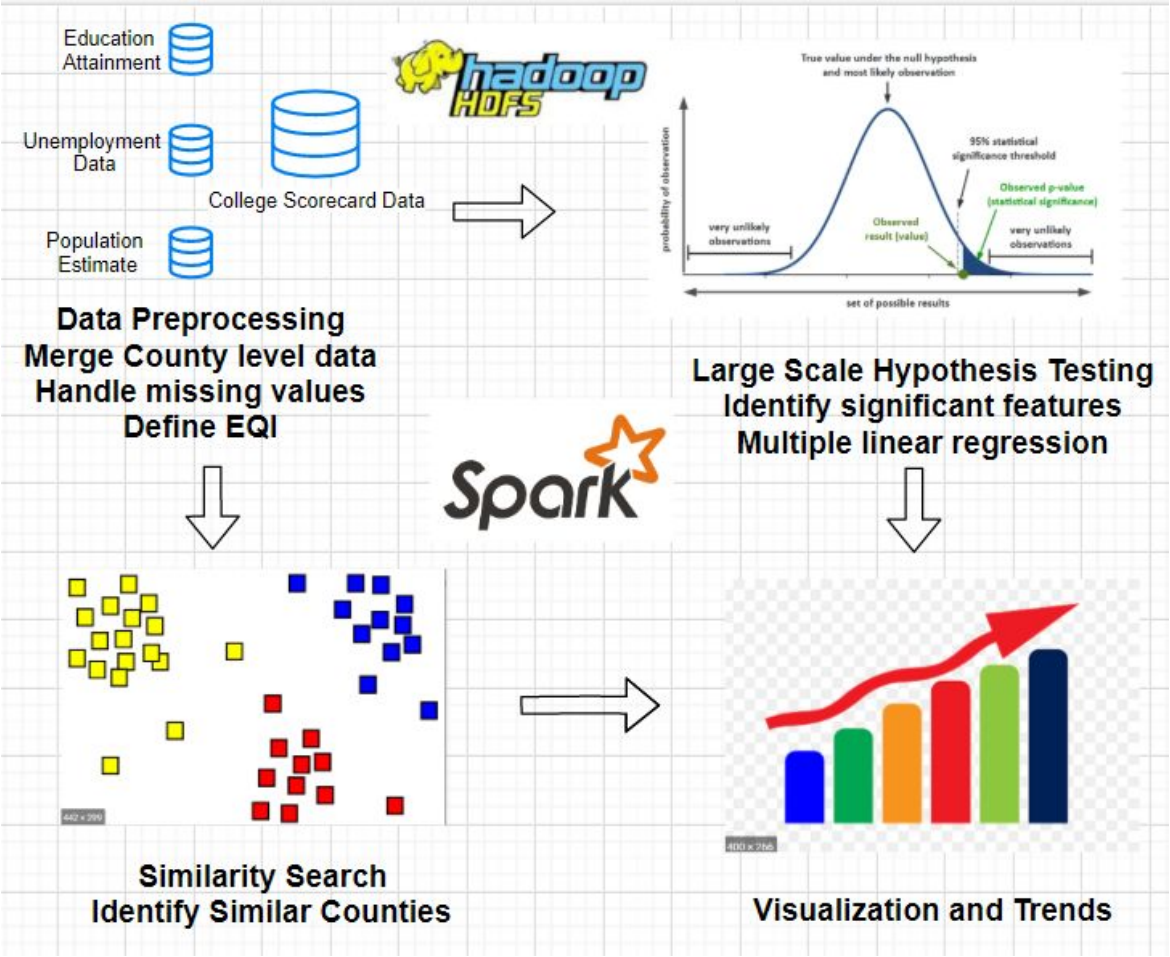
- Analyze and quantify progress towards SDG
- Study influence of multiple factors
- Draw meaningful inferences
- Education Quality at county level for USA

Background

- In 2019, US cities were evaluated for SDG goal according in this report “The 2019 US Cities Sustainable Development Report”.
  a. Analysis for different cities in USA
  b. Define Indices to quantify and measure the SDG
  c. Study encompasses all SDGs

- SDG Tracker : https://sdg-tracker.org/
  a. Analyzed proficiency at different levels of education
  b. Study for all countries in the world
  c. Mainly statistical analysis and comparisons

Process Flowchart



Dataset Description

Dataset	Size	Description	Source
College Scorecard data	11 GB	Dataset describing stats for all colleges in USA from 1997-2019 with a rich feature set of over **1983 features**. Key features: Admission rate, demographic distribution (gender/race), tuition fees, passing rate, federal loan and grants, grades, streams etc.	https://collegescorecard.ed.gov/data/
Unemployment data	50 MB	Dataset with county level statistics about count of employed, unemployed, civilian labor force and unemployment rate (over **6 key features**) spanning the years 2000-2018 (time series format data)	https://data.ers.usda.gov/reports.aspx?ID=17828
Population Estimate Data	120 MB	Dataset with county level population estimates, births, deaths (over **15 features**) for USA over the years 2010-2019 (time series format data)	https://data.ers.usda.gov/reports.aspx?ID=17827
Educational Attainment Data	140 MB	Dataset with county level educational attainment for adults age 25 and older, over **8 features** for indicators for USA spanning the years, 1970-2018 (time series format data)	https://www.ers.usda.gov/webdocs/DataFiles/48747/Education.xls?v=2752.9

Methods

- **Data Preparation:**
  - Multiple college data per county per year.
  - Aggregate data by grouping on year and county FIPS code.
  - Impute missing values by taking mean of attributes by grouping on year and county code.
  - Removed columns that had more than **10%** null values.

- **Google Cloud DataProc Cluster:**
  - 1 master node and 2 worker nodes
  - Image Version: 1.4 (Debian 9, Hadoop 2.9, Spark 2.4)
  - Configuration of Assignment 3 CSE 545

- **HDFS**
  - Distributed storage and parallel processing for the wide data with around 2000 features.

- **Spark and Map Reduce**
  - Data Preprocessing and Standardization - Spark RDDs and Dataframes
  - Heavy transformations - computations for correlation with multivariate regression analysis and finding similarities

Hypothesis Testing

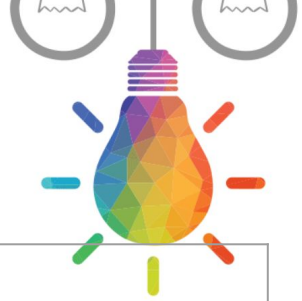
Goal: Determine most important features from over 2000 attributes

- Standardize the features
- Define Education Quality Index - feature engineering
- Linear regression with EQI as the target
- Multiple hypothesis testing using t-statistic
- Identify features with high beta values (high positive and negative correlation coefficients) with EQI

Some intuitive findings

Feature	Beta
Percentage of adults with a high school diploma only	-0.56163
Percentage of adults completing some college or associate's degree	0.52000
Count of undergraduate students enrolled during a 12 month period	0.12872
Net tuition revenue per full-time equivalent student	0.32836
Unemployment rate	-0.61060

Some non intuitive findings



Feature	Beta
Percentage of undergraduates who received a Pell Grant or federal student loan	-0.23067
Total share if enrollment of undergraduate degree-seeking students who are non-resident aliens	0.15718
Instructional expenditures per full-time equivalent student	0.13373
Bachelor's degree in multi/interdisciplinary studies	0.1591
Total Share of enrollment of undergraduate degree-seeking students who are two or more races	0.25427
NET_MIG (international and domestic migration)	0.16519

Similarity Search

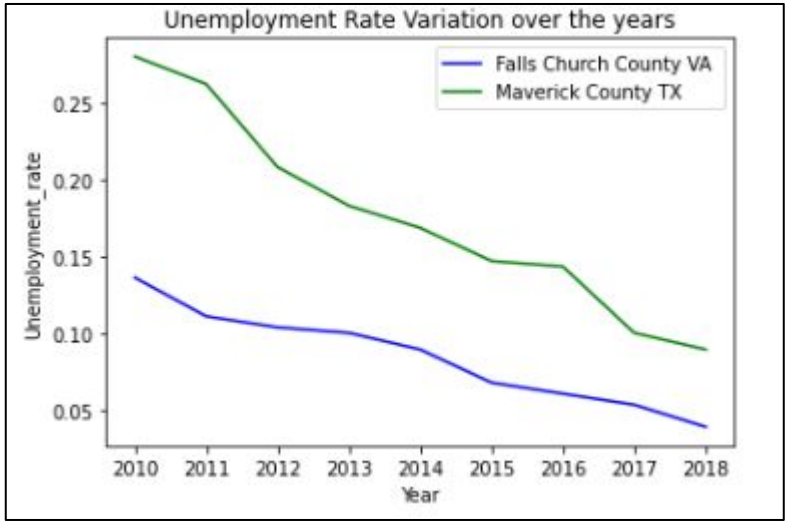
- Data normalization - Min-Max Norm (Features-wise)
- Cosine Similarity used.
- Calculated similarity matrix for all the counties based on their features.

Results

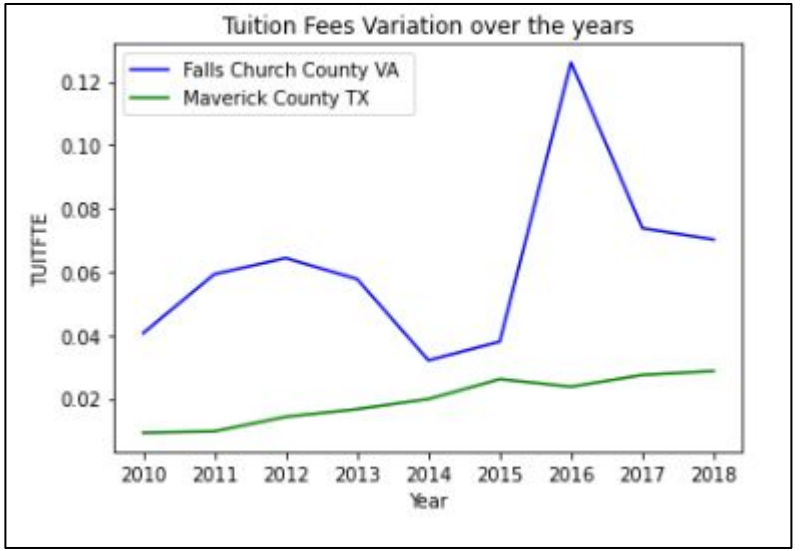
- We found the similarity matrix of counties.
- Verified results by:
  - Selecting 2 county which had maximum difference in their EQI values.
  - Finding their similarity values from Similarity matrix. (Ex: Similarity(Falls Church & Maverick County) -0.91)
  - Visualizing various features of those 2 counties.
  - The choice of features is done using the results from Hypothesis testing and choosing feature with different correlations

Feature variation over the years

Unemployment Rate



Tuition Fees

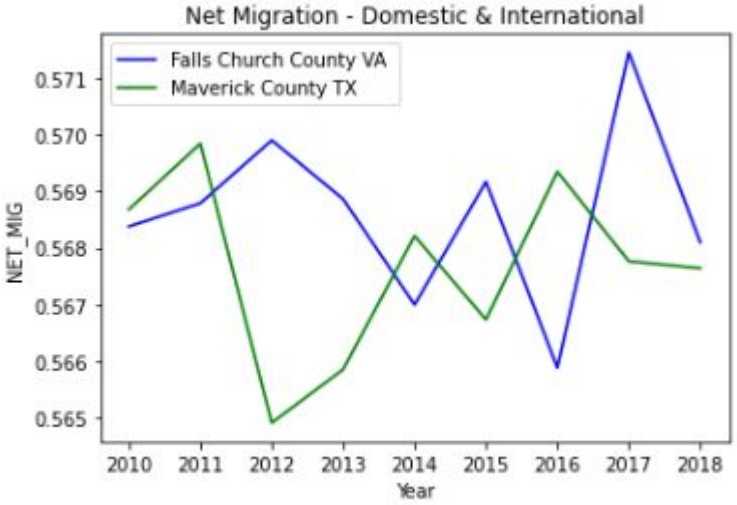


Falls Church County, VA EQI: 0.74373

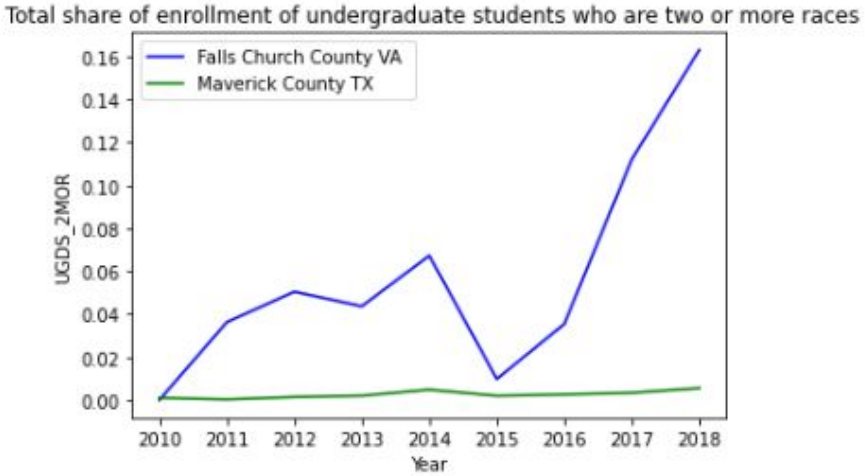
Maverick County, TX EQI: 0.34091

Feature variation over the years -

Net Migration



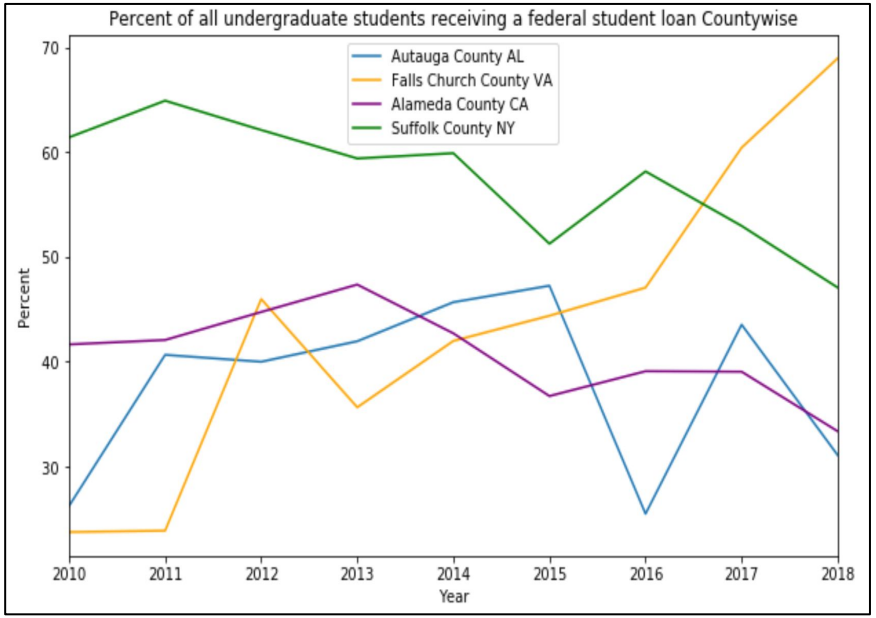
Undergraduate Enrollment Stats

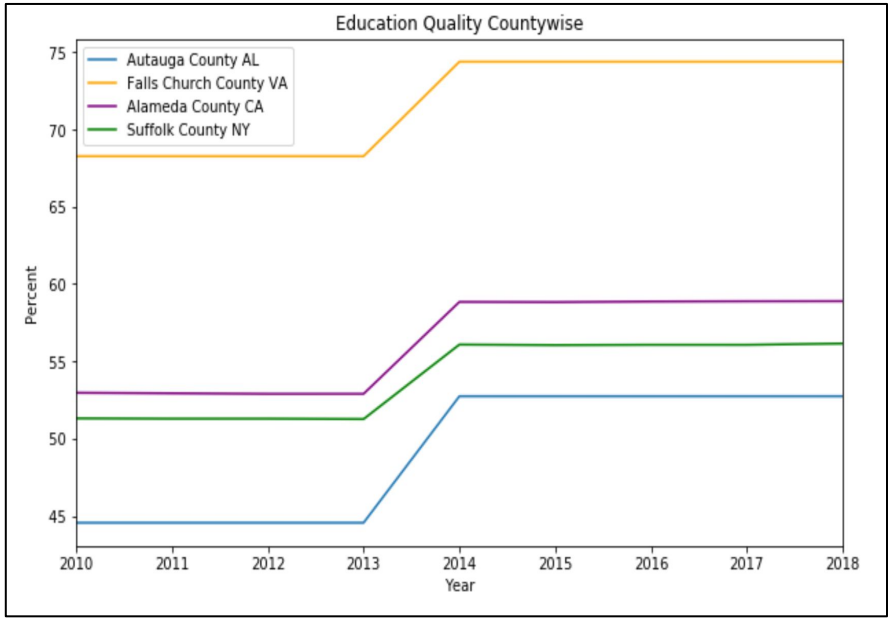


Falls Church County, VA EQI: 0.74373

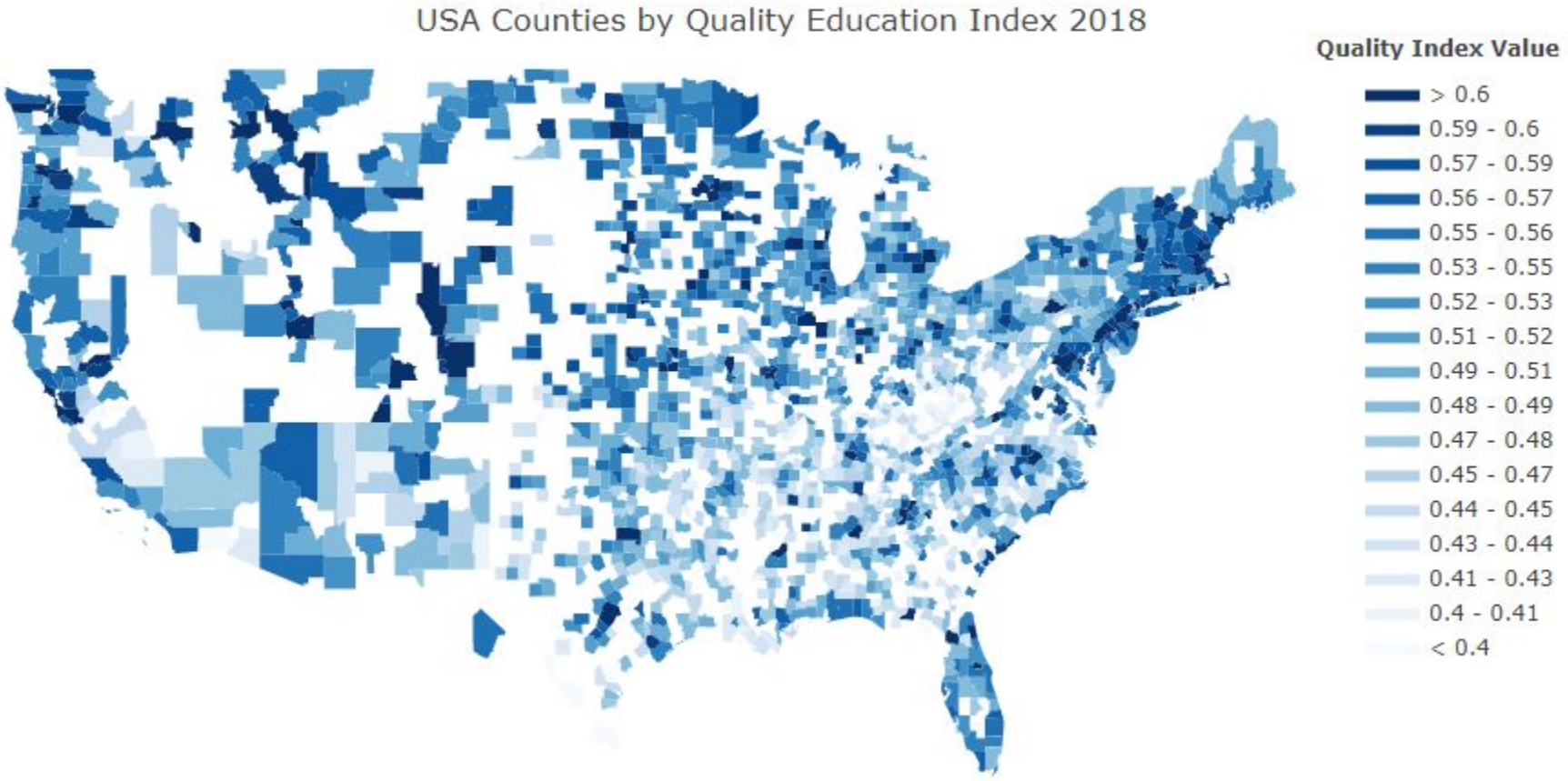
Maverick County, TX EQI: 0.34091

Results: Single Feature Different Counties





Results


USA Counties by Quality Education Index 2018
Quality Index Value
> 0.6
0.59 - 0.6
0.57 - 0.59
0.56 - 0.57
0.55 - 0.56
0.53 - 0.55
0.52 - 0.53
0.51 - 0.52
0.49 - 0.51
0.48 - 0.49
0.47 - 0.48
0.45 - 0.47
0.44 - 0.45
0.43 - 0.44
0.41 - 0.43
0.4 - 0.41
< 0.4

Conclusion

- Successfully identified factors significantly affecting Education Quality Index.
- Analyzed data over a decade to observe trends in features impacting education quality.
- Successfully identified similar counties and compared trends in features for such counties.
- Suggest improvements by contrasting counties with high and low EQI.

Thank You