CSE545 Project Report

SDG 4 Quality Education

Himanshu Agrawal

Introduction

Sustainable Development Goal 4 is an education goal set by the UN. Its objective is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all by 2030. “A quality education has the power to transform societies in a single generation, provide children with the protection they need from the hazards of poverty, labour exploitation and disease, and give them the knowledge, skills, and confidence to reach their full potential.”— Audrey Hepburn, the iconic American actor [3].

Why is it important?

According to UIS data for the school year ending in 2018, about 262 million children and youth are still out of school. It includes 59 million children of primary school age, 62 million of lower secondary school age and 138 million of upper secondary age. Also, more than half of all children and adolescents worldwide do not meet minimum proficiency standards in basic reading and mathematics[4]. In our project we aim to analyze factors like unemployment rate, population, education attainment, demographics and study their impact on the quality of education for the USA counties by drawing meaningful inferences.

Background

Since the launch of SDGs in September 2015, all countries have been working to analyse SDG4 indicators defined by the UN and determine where a country currently stands towards the path of goal attainment.

SDG Tracker is one such website that shows this analysis on world level, ie, for each country. It shows mainly statistical analysis and comparisons. It shows graphs for each indicator for SDG 4 for countries in the world.

We also studied ”The 2019 US Cities Sustainable Development Report” in which US cities were evaluated in the year 2019 for SDG goals. It defines Indices to quantify and measure the SDG goals for different cities in the USA. Taking motivation from the above two analysis, our goal is to analyse SDG 4 for all the counties in the USA.

Data Description

We built our master dataset by merging 4 individual datasets spanning both the academic and economic indicators of US counties.

1. College scorecard data (11 GB) : This dataset consists of statistics for all colleges in the USA, recorded from 1997 - 2019. It has a rich feature set of over 1983 features. Key features include admission rate, demographic distribution (gender/race), tuition fees, passing rate, federal loan and grants, grades, streams etc.

2. Unemployment data (50 MB) : This dataset describes county level employment statistics recorded from 2000-2018 in time-series format.
3. Population Estimate data (120 MB): The dataset describes county level population estimates for the US from years 2010-2019 curated in time-series format.
4. Education attainment data (140 MB): Education attainment data contains county level educational attainment information for adults age 25 and older, from years 1970-2018 in time series format.

Below are some of the important features selected from each dataset:

Dataset	*Important features used for Analysis*
College scorecard data	*% of students finishing a degree, diversity of student body, number of PhDs in faculty and 300 other features*
Unemployment data	*% of unemployed people, employed people and 6 other features*
Population estimates data	*Number of births, deaths and 15 other features*
Education attainment data	*% of students with a college degree, % of students with a high school degree etc.*

Table 1: Dataset and important features

Data Preprocessing

1. Data merging and aggregation: The above 4 datasets have been merged on county FIPS code. College scorecard data has statistics for each college. It has been aggregated at county level for each year.
2. Null value imputation : Null values have been replaced with the average value of the feature for the particular county and year to which the data point belongs. Columns with more than 10% null values have been filtered out. The final dataset after cleaning and preprocessing is 8.6 GB in size.

Methods

Hypothesis Testing:

In large scale hypothesis testing, we try to test the significance of the findings based on a data driven decision. In this study, we performed multiple hypothesis testing to decide whether a feature is significant towards the target feature - Education Quality Index. We begin with the null hypothesis being "the attribute is not significant towards the target Education Quality Index".
The Education Quality Index (EQI) has been derived by applying the following formula for selected education attainment features.

$$EQI = \frac{0.8*(CD)+0.6*(LCD)+0.4*(HSD)+PERCUG}{100}$$

PERC_UG = % undergraduate enrollment in the county over total population estimate.
CD : % of population with a 4 year college degree or higher.
LCD : % of population with some college degree less than 4 years.

HSD : % of population with a High School degree.
LHSD : % of population with less than a high school degree.
We used t-statistic for significance testing and as the data was standardized. The betas obtained from the results of multiple linear regression, were identified as the correlation coefficients with the target EQI. Using this approach, we identified the top positively and negatively correlated features with EQI so that potential attributes affecting the education quality can be focused upon. Also, bonferroni correction factor had to be applied to get accurate p-values for multiple hypothesis testing. We concluded with some intuitive and non-intuitive findings.

Results:

Feature	Correlation Coefficient	p-value
Unemployment Rate	-0.610	0.0
Net tuition revenue per full time student	0.328	0.0
Percentage of adults with high school diploma only	-0.561	0.0

Table 2: Intuitive Findings

Feature	Correlation Coefficient	p-value
Percentage of undergraduates who received a Pell Grant	-0.230	1.02e-202
Total Share of enrollment of undergraduate degree-seeking students who are two or more races	0.25427	2.47e-247

Table 3: Non Intuitive Findings

Interpretation of Results : It can be observed that Unemployment rate is highly negatively correlated with EQI as expected. Contrary to what one would expect, the percentage of students who received a Pell Grant is negatively correlated. It must be noted that The Pell Grant or federal student loans are granted to students with extreme need of financial aid and implies that such regions have lower accessibility to quality education, explaining the negative correlation. Net tuition revenue however is positively correlated to EQI, implying that higher tuition results in better infrastructure and facilities, thereby improving the quality of education.

Similarity Search

Similarity search is a concept used to identify similarity between 2 objects. This object can be documents, counties, cities etc. There are many methods to find the similarity like using the approximate distance method which includes the Locality Sensitive Hashing family. To compute the distance measure there are many options like Jaccard Similarity, Cosine Similarity, Euclidean distance etc.
For our project we are finding the similarity between two counties based on their feature values. For this, we are using cosine distance as our similarity measure as the data we have is mostly numerical. We are using Spark framework on HDFS to efficiently compute similarity.

Process:

1. The input data is the processed data from above.
2. Remove the header row and filter the county records for the year 2018
3. Create key value pairs of form (column_name,(county,row_value))
4. Using groupByKey() group all values of same column (column_name,[(county1, row_value),(county2, row_value)...])

5. The range of values of different features is not the same, so we applied data normalization using Min-Max Normalization technique. $X_{i_{scaled}} = \frac{X_{ij} - X_{i_{min}}}{X_{i_{max}} - X_{i_{max}}}$
6. Create key value pairs of form(county, (column_name, row_value))
7. Using groupByKey() group all values of same column (county,[(column_name1, row_value),(column_name2, row_value)...])
8. To compute our similarity matrix , the cartesian join of county_rdd with itself to get all the pairs of counties which forms our similarity matrix.
9. Filtered duplicate records for county pairs of the form (a,b) and (b,a) and kept only (a,b) so that our similarity matrix is an upper triangular matrix.
10. Cosine similarity is calculated as:

$Similarity(A,B) = \frac{A.B}{||A|| * ||B||} = \frac{\Sigma_i^n A_i}{\sqrt{\Sigma_i^n (A_i)^2} * \sqrt{\Sigma_i^n (B_i)^2}}$. Here, A, B are 2 counties vectors and A_i B_i are their columns(Features).

Frameworks and Technologies Used

HDFS - Distributed storage and parallel processing for the wide data with around 2000 features.
Spark and Map Reduce framework - Data Preprocessing and Standardization, Heavy transformations - computations for correlation with multivariate regression analysis and finding similarities
Google Cloud Platform - Apache Spark and Hadoop hosted on GCP Cluster

Process Flowchart

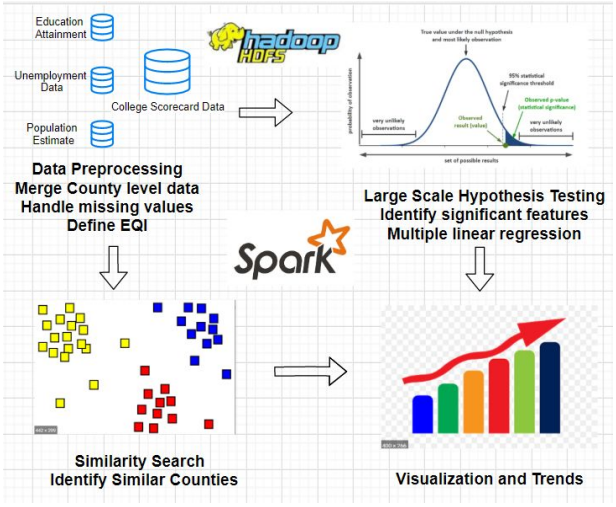

Figure 1 : Overview and flowchart of Implementation

Results

After calculating the similarity matrix, we wanted to find interesting insights from the data. So, we used the results from our hypothesis testing to select the feature with different β (beta) values (represents correlation) and also taking into account the EQI(Education Quality Index values).
We identified two counties with maximum difference in their EQI values. They are Falls Church County(VA): EQI- 0.74373 and Maverick County(TX): EQI- 0.34091.

From our similarity matrix we found their similarity to be -0.91. It suggests that the counties are very dissimilar. We can see the same from their EQI values also. Now we analyze how various features vary in these counties.

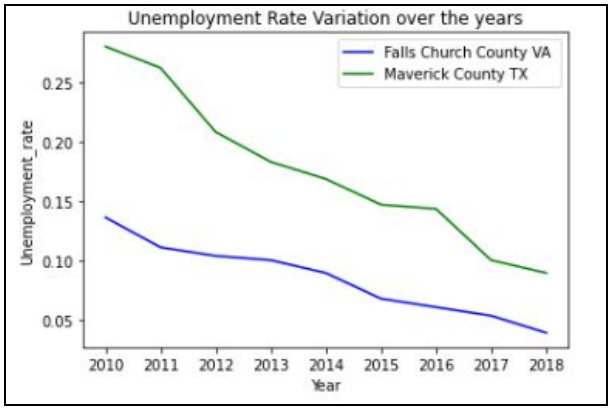

Figure 1: Unemployment rate variation

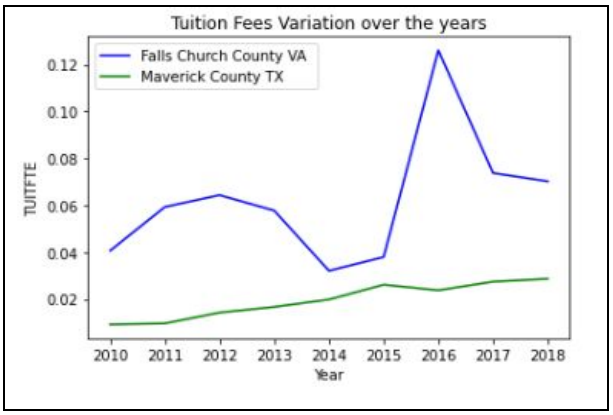

Figure 2: Tuition Fees variation

In **Figure 1** we have taken Unemployment rate(β = -0.61060, p-value = 0.0). We see that the trend of unemployment is the same for both counties but the value is more Maverick county. So improving unemployment will help the county to improve its EQI.

In **Figure 2** we have taken Tuition Fees(β = 0.328365055, p-value = 0.0). We see that the trend is consistent in maverick county but varies more for Falls Church county. This might be an indicator that Falls church has better college since we usually relate college fees with the college education quality(IVY League colleges cost more as compared to state colleges).

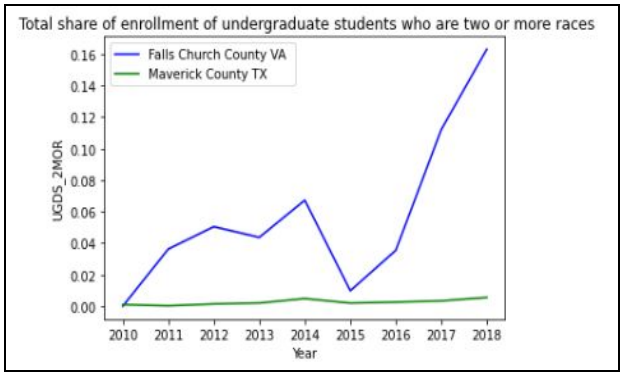

Figure 3:Undergraduate Enrollment variation by race

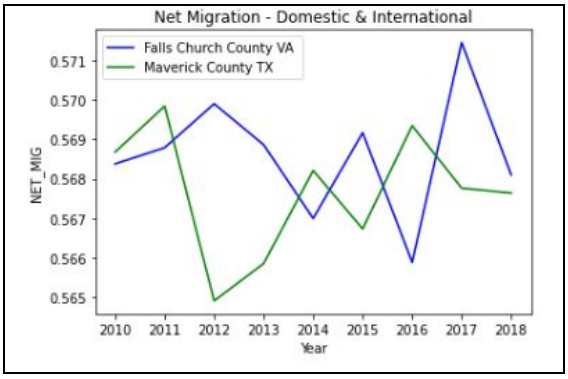

Figure 4: Net Migration of Students

Similarly in Figure 3 & 4, for two more features Percentage of Undergraduate Enrollment (β = 0.254276678 , p-value = 2.47e-247) and net_migration (β = 0.165198602 , p-value = 1.422e-103). We see that Maverick county performs lower to Falls church. So we can give suggestions to Maverick county about these analyses and they can work towards these attributes to raise their EQI.

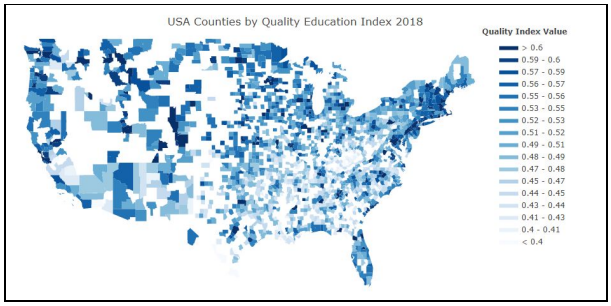

Figure 5: Unemployment rate variation

In **Figure 5** we have made a choropleth map showing the EQI for the counties in the USA. The color scale ranges over the values of EQI dark blue representing highest value and light blue representing lowest value. We observed that EQI is higher in the East coast while its lower in the Middle and northern USA. Some counties whose data is missing are represented by white color(Ex- Nevada State).

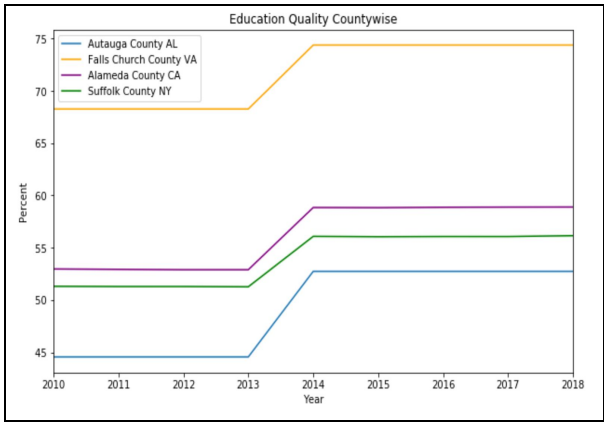

Figure 6: Unemployment rate variation

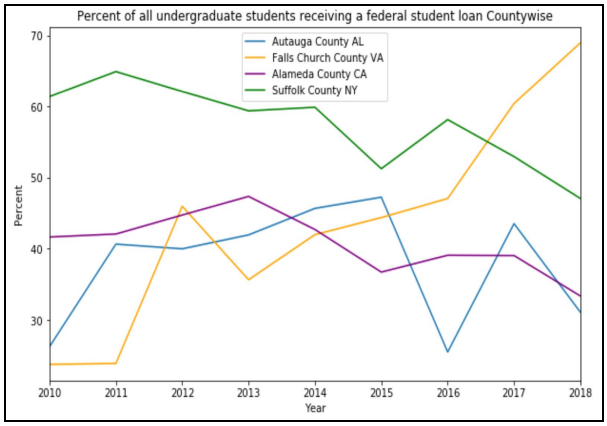

Figure 7: EQI variation in different counties

In **Figure 6**, we compare 4 counties and see how they vary for a particular feature. We have taken Percentage of undergraduate students receiving a Federal Loan(β =0.207755998 , p-value = 1.02e-202). Suffolk and Alameda county are on a decline trend while it's increasing for Falls Church. Having more students receiving federal loans is a good thing as it indicates more students are enrolling and education quality increases.

In **Figure 7**, we see that these counties are evolving with respect to their education quality.

Conclusion

We have successfully identified factors significantly affecting Education Quality Index using hypothesis testing. We have analyzed data over a decade to observe trends in features impacting education quality. We successfully identified similar counties and compared trends in features for such counties using similarity search. Using similarity search we can suggest improvements by contrasting counties with high and low EQI.

References:

[1]https://www.un.org/sustainabledevelopment/education/

[2]http://uis.unesco.org/en/topic/out-school-children-and-youth
[3]https://gulfnews.com/uae/education/why-quality-education-matters-1.1221277
[4]https://unstats.un.org/sdgs/report/2019/goal-04/
[5]https://sdg-tracker.org/
[6]https://www.sustainabledevelopment.report/reports/2019-us-cities-sustainable-development-report/
[7]https://www.timeshighereducation.com/university-impact-rankings-2019-sdg-quality-education-methodology
[8]https://datatopics.worldbank.org/education/wQueries/qachievement