Project 3- DBMS

Submitted by

Zainab Rizwan (2019-CE-36)

Submitted to

Ma'am Darakhshan Abdul Ghaffar

CS-363L Database Systems

Spring 2022

18th May, 2022

Department of Computer Engineering

University of Engineering and Technology, Lahore

Dataset Details

Dataset Name: world_bank_intl_education

Data set ID: bigquery-public-data: world_bank_intl_education

Description:

The world_bank_intl_education dataset is a part of the bigquery public dataset collection. This dataset combines key education statistics from a variety of sources to provide a look at global literacy, spending and access. It also provides information regarding the total population getting education on primary, secondary and higher-levels w.r.t age and gender.

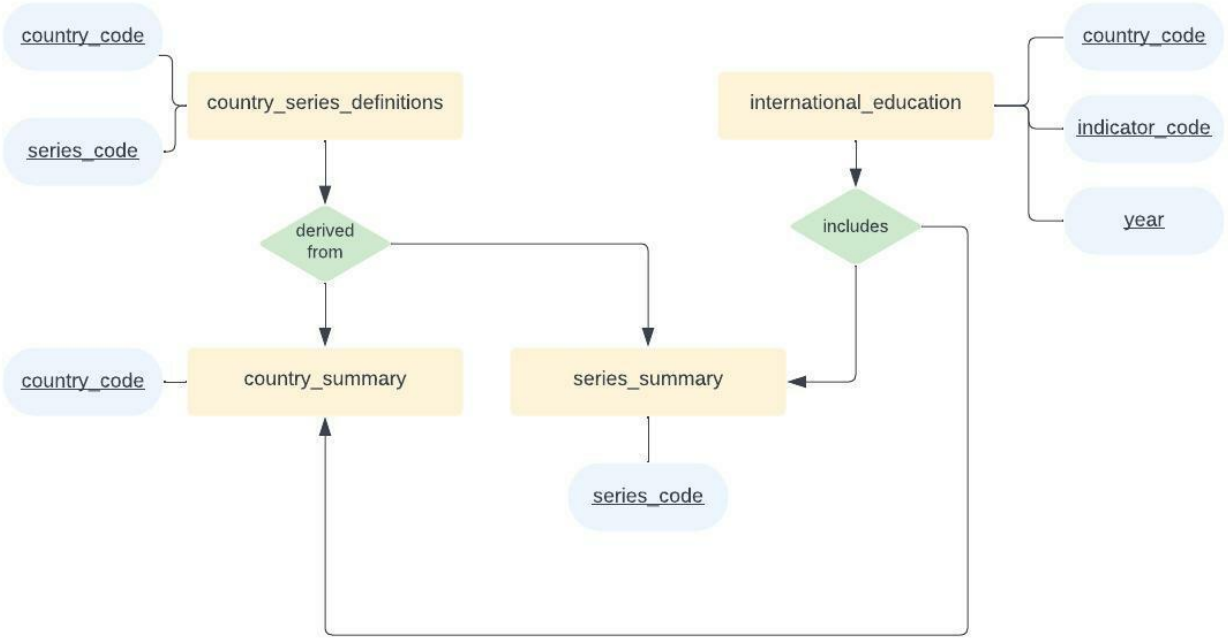
Analysis of Dataset

Is there redundant data?

Yes. The country_series_defintions, country_summary and series_summary tables contain little to no repetitions. But there is serious data redundancy in the international_education table.

What are the relationships between tables?

Table relations have been described in the ER diagram.



Functional Dependencies

1. **country_series_definitions**

   series_code → description

2. **country_summary**

   country_code → short_name, table_name, long_name, two_alpha_code,
   currency_unit, special_notes, region, income_group, wb_two_code,
   national_accounts_base_year, national_accounts_reference_year, sna_price_valuation,
   lending_category, other_groups, system_of_national_accounts,
   alternative_conversion_factor, ppp_survey_year,
   balance_of_payments_manual_in_use, external_debt_reporting_status,
   system_of_trade, government_accounting_concept,
   imf_data_dissemination_standard, latest_population_census

3. **international_education**

   country_code → country_name
   indicator_code→ indicator_name

4. **series_summary**

   series_code → topic, indicator_name, short_definition, long_definition,
   unit_of_measure, periodicity, base_period, other_notes, aggregation_method,
   limitations_and_exceptions, notes_from_original_source, general_comments, source,
   statistical_concept_and_methodology, development_relevance, related_source_links,
   other_web_links, related_indicators, license_type

Table keys

1. **country_series_definitions**

   Composite keys: country_code, series_code

```
SELECT country_code, series_code, COUNT(*) as x FROM
`bigquery-public-data.world_bank_intl_education.country_series_definit
ions`
GROUP BY country_code, series_code
HAVING x>1
```

   Foreign key: series_code

2. **country_summary**

   Primary key: country_code

3. **international_education**

   Composite keys: country_code, indicator_code, year

```
SELECT  country_code, indicator_code, year, COUNT(*) AS x FROM
`bigquery-public-data.world_bank_intl_education.international_educatio
n`
GROUP BY country_code, indicator_code, year
HAVING x>1
```

Foreign key: country_code, indicator_code

4. **series_summary**

   Primary key: series_code

Why do you think the original authors of the dataset chose this particular structure?
The authors of this dataset chose this particular structure as it was the most efficient, logical and normalised design.

Exploring questions with appropriate visualizations

Plotting library used: Matplotlib

Central question: What is the percentage of government spending spent on education

Questions to explore

1. Comparison of percentage of government spending spent on education in different countries
2. Global government spending throughout the years
3. Government spending trends in South-Asia

Indicator code to explore question: `SE.XPD.TOTL.GB.ZS`

Indicator Definition: General government expenditure on education (current, capital, and transfers) is expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.). It includes expenditure funded by transfers from international sources to the government. General government usually refers to local, regional and central governments.

Engineering feature used: Standard deviation

Reason: Standard deviation helps us understand just how spread out the data values are in a given dataset. It is a measure of how far each observed value is from the mean.

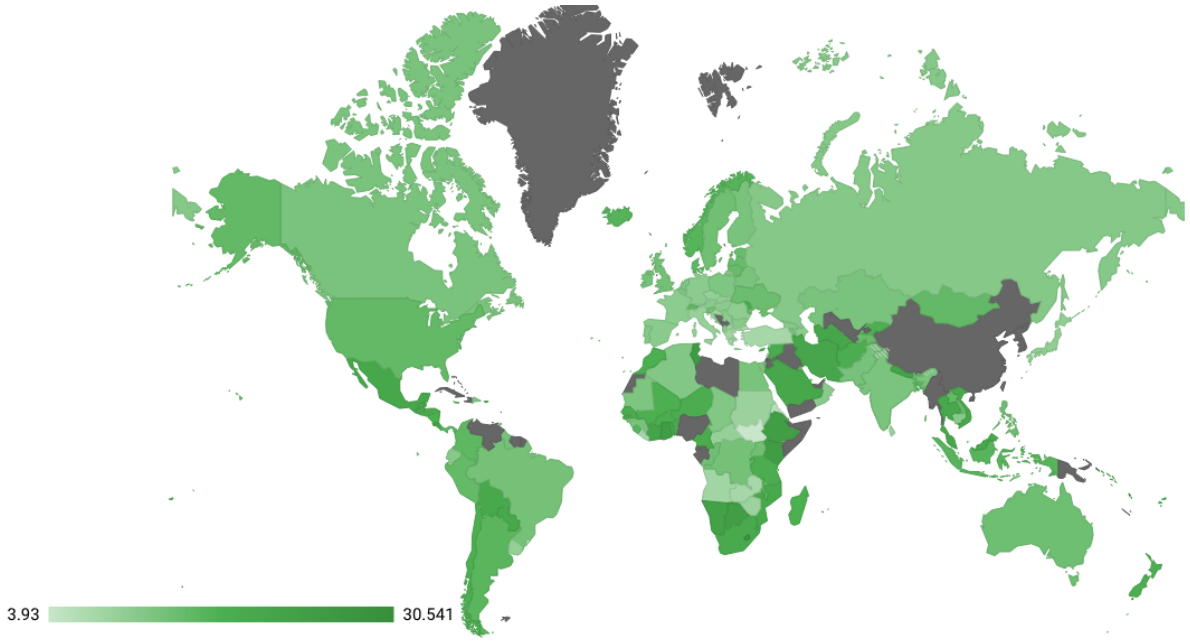
Here it has been used to better understand whether most government expenditures on education are close to the global average (14.79) or if there is a wide spread in these expenditures.

Visualisations

The visuals represent how economic growth and stability plays an important role in determining the education expense of a country

Visual representation of expenditures in various countries

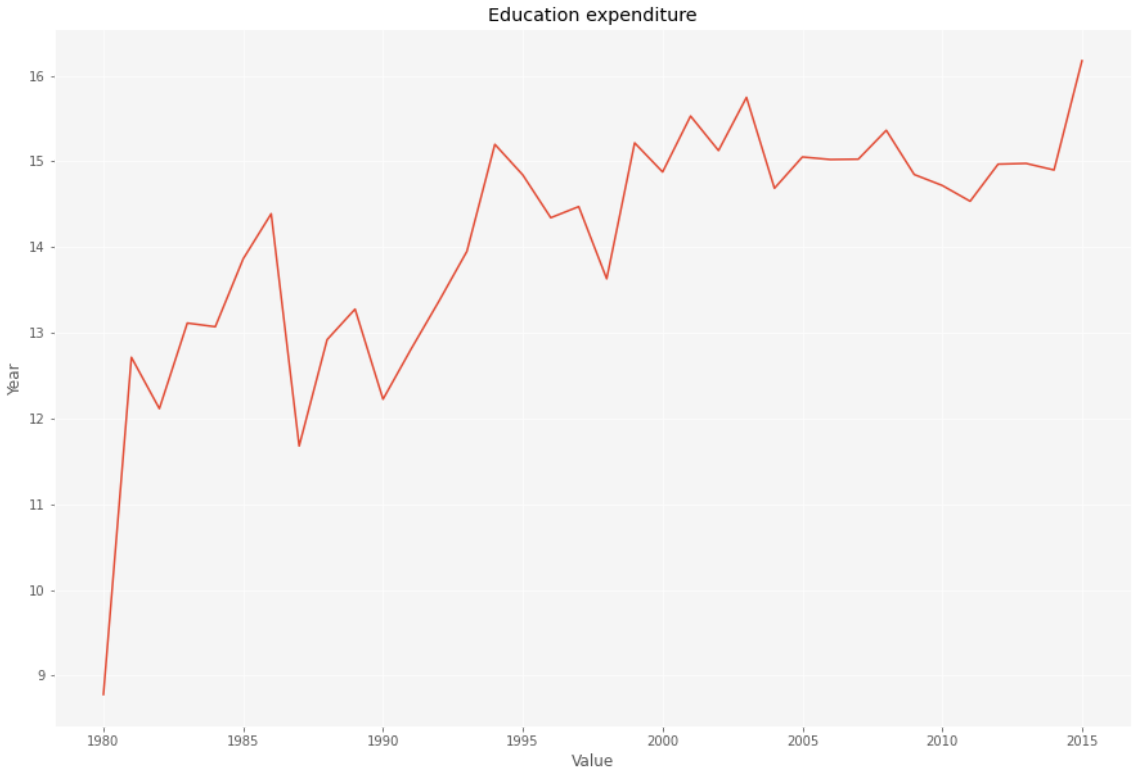
The following map provides a comparison of average government expenditures on education throughout the world using a green colored gradient. The lightest green acts as the lowest value assigned and the darkest green represents the highest value; in this case 30.541. The grey colored areas have been excluded from the representation as there are no records of their values in the dataset provided.



Trends of education expenditure throughout the years

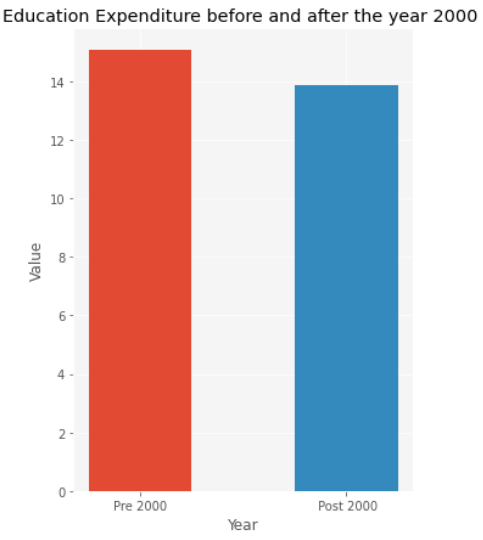
The following line map provides a graphical representation of trends of education expenditure throughout the globe recorded from the year 1980 till 2015. Some of the

noticeable downward dips in the graph in the years 1987, 1990 and 2001 highlight the years of global recession.



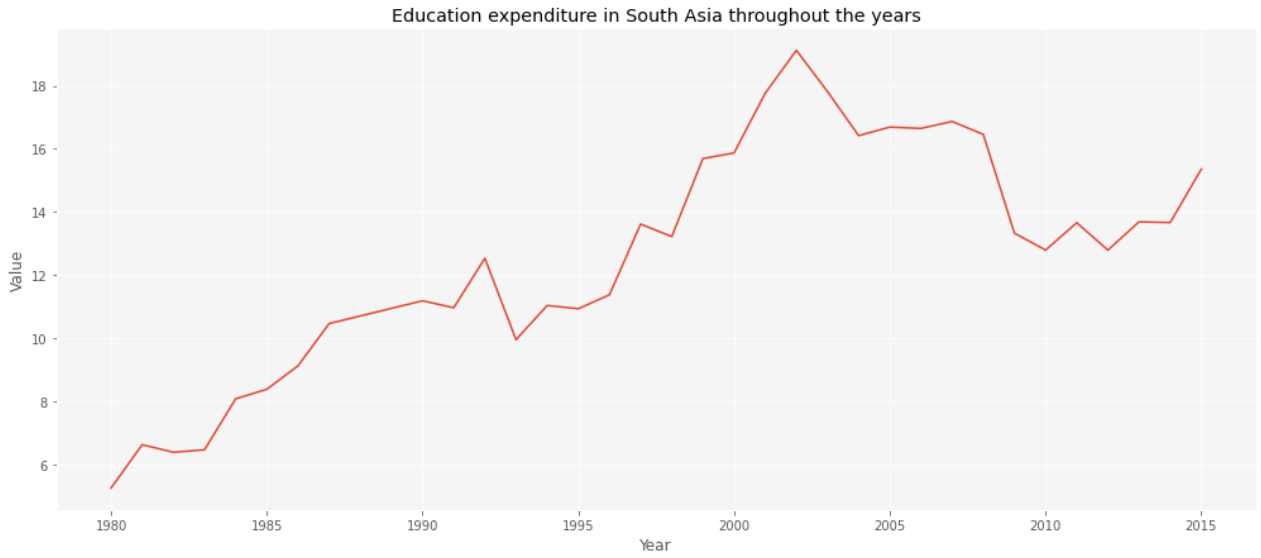
Comparison of education expenditure before and after the dawn of the 21st century

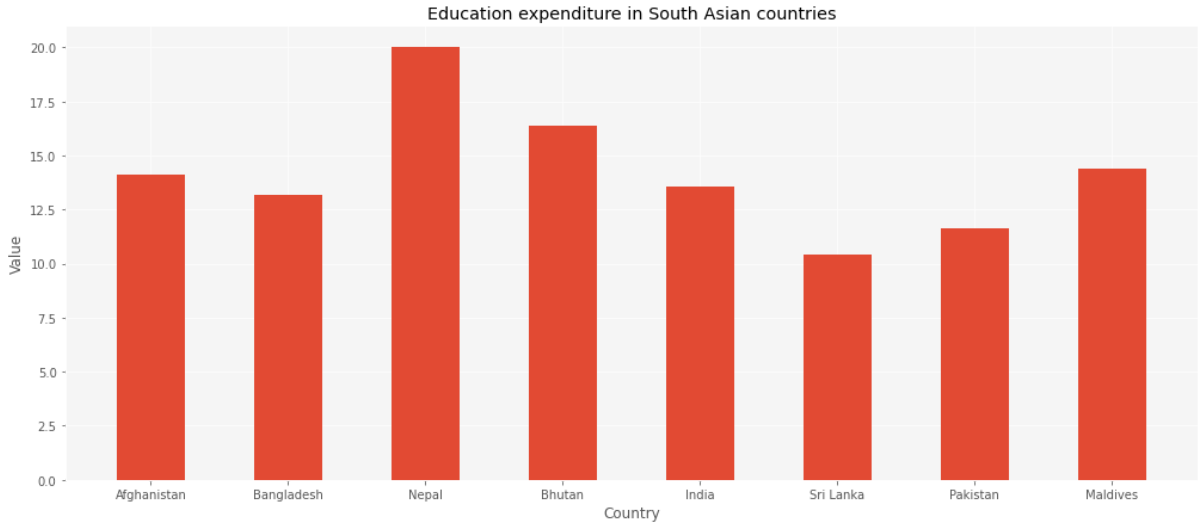
A simple comparison displays how the average expenditure of government on education has noticeably decreased with the dawn of the 21st century.



Trends of education expenditure in South Asia

The following visuals provide a brief overview of trends of expenditure on education in South Asia. The first graph provides a depiction of education expenses throughout the years with noticeable growth in years **2002** and **2003**. The second graph provides a comparison of government expenditure of South Asian countries provided in the dataset with **Nepal** having the highest and **Sri Lanka** having the lowest value assigned.

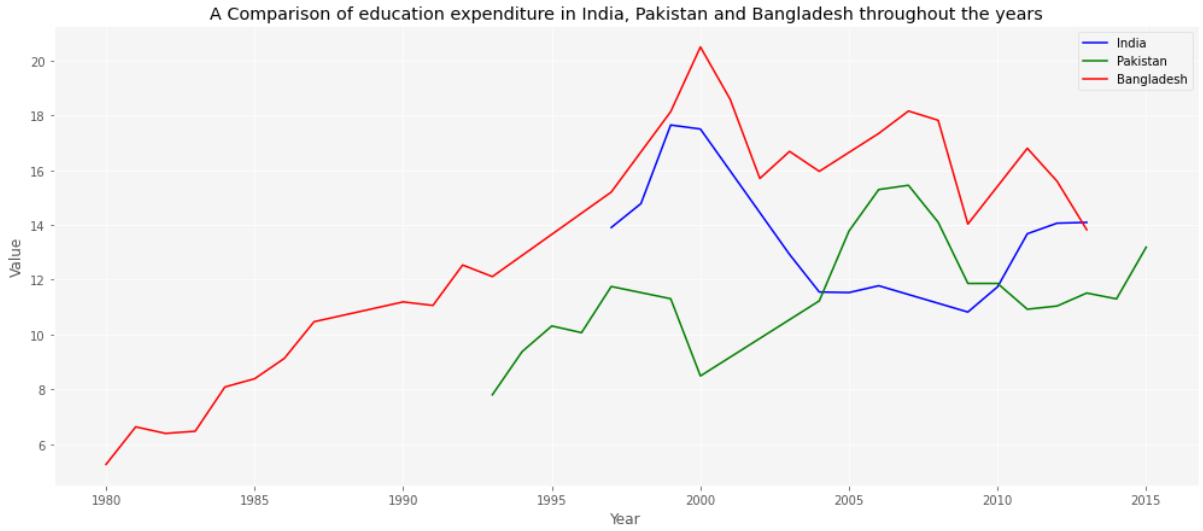




Course of education expenses by the government in India, Pakistan and Bangladesh

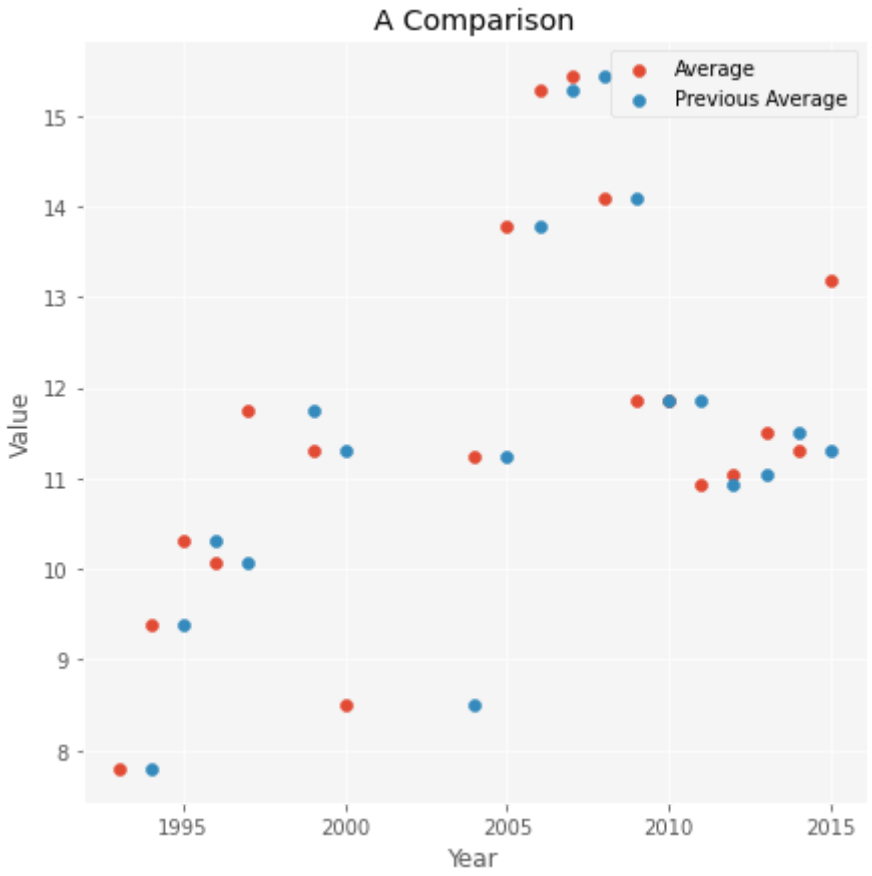
The line graph below is a pictorial representation to portray the course of education expenses by the government in India, Pakistan and Bangladesh. We can see Bangladesh is assigned the

highest value in the year 2000, namely the year in which there was 5.1% growth in the country's GDP.



A comparisons of Pakistan's expenditure with the previous years

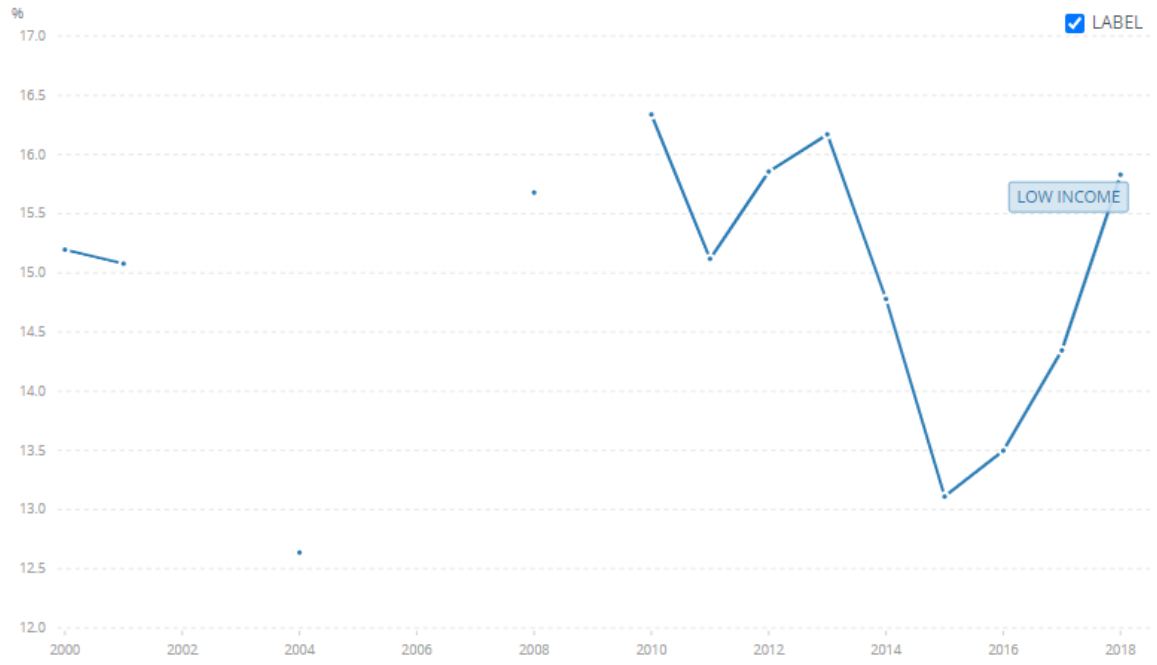
The visual below depicts how there has been minimal growth in expenditure when it comes to education.



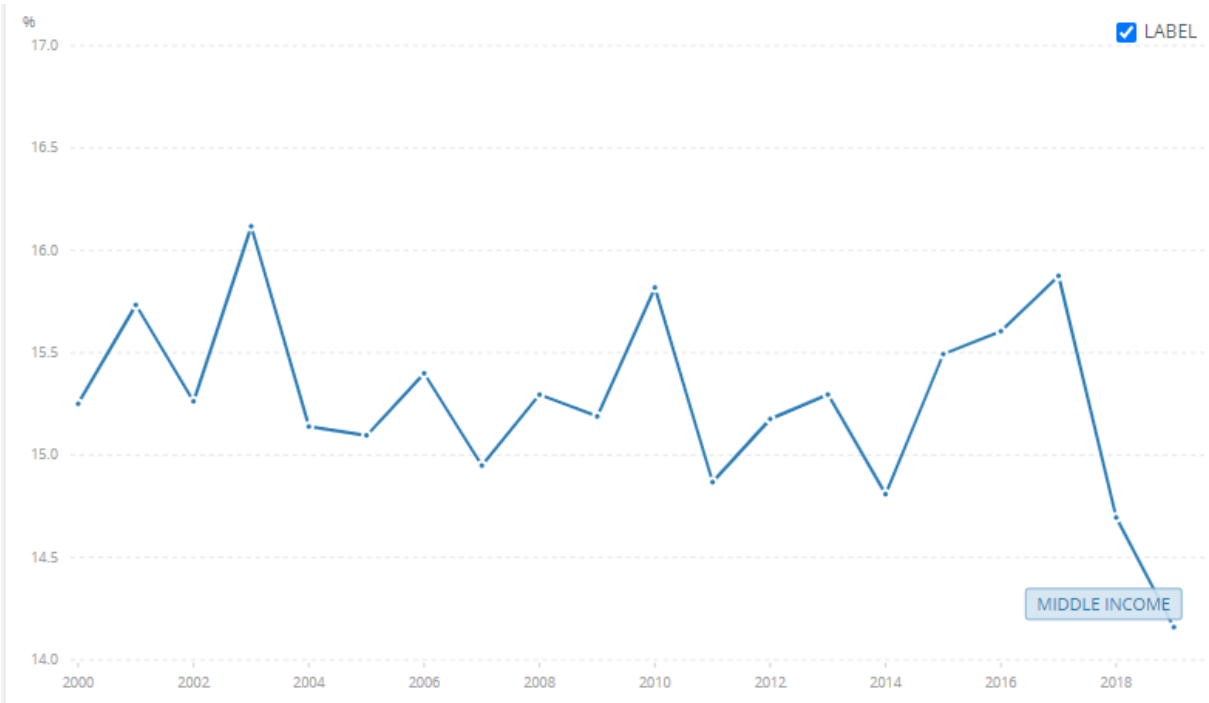
Income wise Visualisations

The following visualisations show the average government expenditure on education in low, middle and high income countries. From the graphs, as of 2018, the middle income countries have an average value of 15.8, the middle income countries have an average value of 14.7 and the high income countries have an average value of 11.8

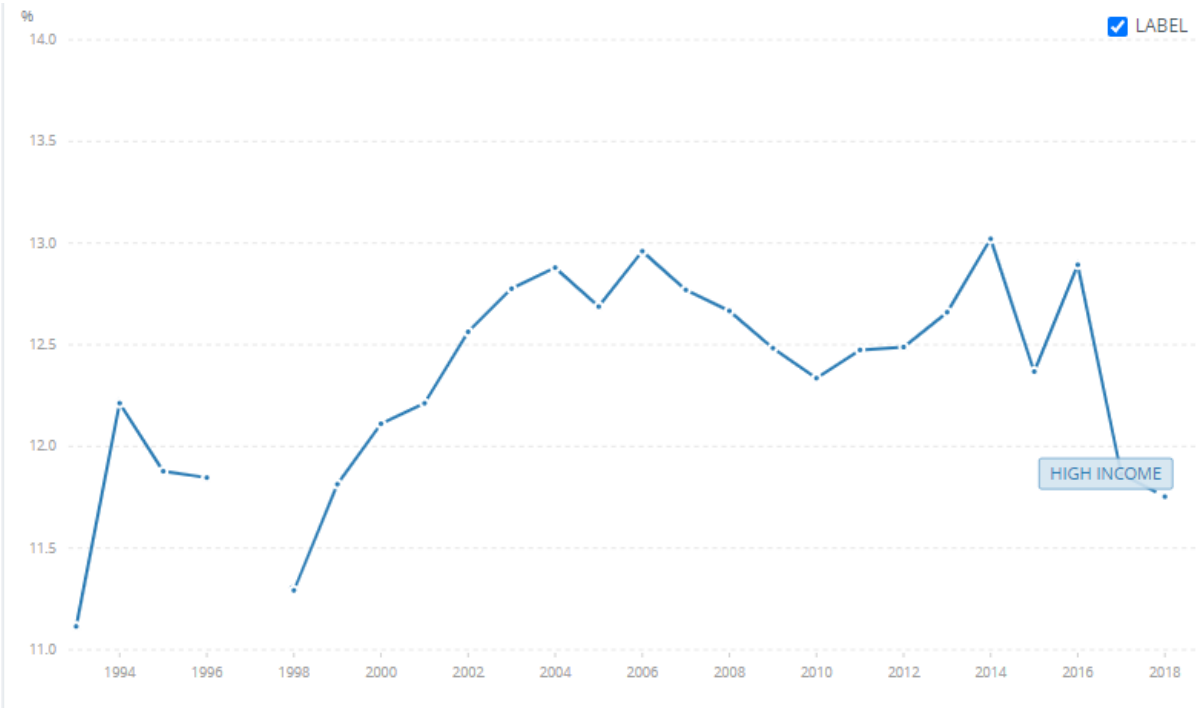
Low income countries



Middle income countries



High income countries



Predictions

Data prior to the year 2000 was used to train the model and data after the year 2000 has been used as testing data. A linear regression model has been used to generate predictions.

1. The model was first evaluated using test data post year 2000

	mean_absolute_error	mean_squared_error	mean_squared_log_error	median_absolute_error	r2_score	explained_variance
0	1.771286	5.698343	0.019847	1.451223	0.639005	0.63939

	training_run	iteration	loss	eval_loss	learning_rate	duration_ms
0	0	5	5.698343	None	0.4	2017
1	0	4	5.729577	None	0.4	2201
2	0	3	5.809847	None	0.4	2145
3	0	2	6.029726	None	0.4	2155
4	0	1	7.117667	None	0.2	2230
5	0	0	17.630813	None	0.2	1820

2. Average value of government spending on education in in all countries throughout the years

	predicted_average	actual_average
0	14.928571	15.027507

3. Overall average value government spending on education in Pakistan throughout the years

	predicted_average	actual_average
0	12.5	12.627891

4. Overall average value government spending on education in of Pakistan in the year 2012

	predicted_average	actual_average
0	12.0	11.04063

5. Average value of government spending on education in South Asian countries throughout the years

	predicted_label	label	country_name
0	14.477765	8.489880	Pakistan
1	14.234615	17.496500	India
2	18.182610	13.772010	Bhutan
3	17.767018	20.490419	Bangladesh
4	21.598350	19.113590	Nepal

Conclusion

Google BigQuery is a great place to explore public datasets and practice data analytics skills. BigQuery ML makes machine learning more accessible for more audiences by abstracting away many of the highly mathematical aspects into simple SQL queries.

The world_bank_international_education table was explored with prime focus on the central question being the value of government spending on education. There are several limitations

in the models. Although linear regression was used to train the data, the data itself is not perfectly linear since there are a lot of noisy data points lying around which affect the training result. Also, when we increase the number of features in training, the mean absolute error decreases, but the variance also increases by relatively the same amount. This means that although the overall trend of data may be linear, since the data points are spreaded fairly away from the model, the individual prediction may not be very accurate.