Project_Report_x18127339

by Punit Lohani

Submission date: 12-Apr-2019 09:53PM (UTC+0100)
Submission ID: 1111391632
File name: 17723_Punit_Lohani_Project_Report_x18127339_54897_373029279.pdf (995.41K)
Word count: 7975
Character count: 43498

National College of Ireland

Data Warehousing and Business Intelligence Project

on

Economic Indicators of European Union

Punit Lohani
18127339

MSc/PGDip Data Analytics – 2018/9

Submitted to: Sean Heeney

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

National College of Ireland

Student Name:	Punit Lohani
Student ID:	18127339
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Heeney
Submission Due Date:	12/04/2019
Project Title:	Economic Indicators of European Union

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	April 12, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Economic Indicators of European Union

Punit Lohani
18127339

April 12, 2019

Abstract

Economic indicators are important because they contribute and provides an information about the economic state of a country i.e. whether the economy is expanding or contracting. Wilkinson (2013) By going through the various economic indicators the analysts and the economists can predict and present the outcomes. The objective of this project is to go through some of the indicators that can be considered important and their relationship with the economy. Factors like education attainment, employment rate are very important as they can play a crucial role in boosting up the economy as if more people are educated then the chances of getting employment will also improve. Also we are going to see that what are the factors that affect the Gross domestic product. The other indicators like government revenue, government expenditure are also taken into consideration and how they are related to the GDP growth rate, which are described in detail in the below section. The datasets have been chosen by keeping the above factors in mind and these are used in establishing and analyzing the relationships.

1 Introduction

The Economic indicators play a significant role in the analysis of the economic growth of a nation. These indicators not only provide the trends over a period of time but also helps to get an insight whether there is an expansion or the contraction in the economy of a nation. Gross domestic product and the GDP growth rate are considered to be the best way for the measurement of the economic growth Amadeo (2019).The main objective behind choosing this topic is to get an idea about the vision of 20 European countries on the basis of various indicators. We are going to compare that how upper secondary education attainment in the age bracket of 25-64 years in the respective countries affects the employment rate among the individuals of same age group. With the help of this we are going to find out whether a country having minimum of upper secondary education have chances of having improved employment rate or not and how does population influence both these factors. Further with the comparison of employment rate with consumer spending we are going to find that how these two factors affects the Gross domestic product of a country. Finally the focus will be on the comparison between the government revenue with the government total expenditure to find out how these two can be related to the gdp growth rate. By comparing all of these factors we will be able to analyze about the trends that the 20 countries are adopting. In order to proceed with the analysis, the datasets have been chosen from seven different data sources. Automation of scripts are performed with

Source	Type	Brief Summary
Statista	Structured	Public spending ratio made during the year 2017
UNECE	Structured	Information about the Final consumption expenditure of households 2008 to 2017
Eurostat	Structured	Total employment rate from the year from 2006 to 2017
European Fiscal Board Report	Unstructured	GDP growth rate from 2000 to 2019
OECD	Structured	Government revenue from 2000 to 2017, Education attainment 2000 to 2017
International Monetary Fund	Structured	Gross domestic product from 2008 to 2017, General government total expenditure from 2008 to 2016
World Bank	Structured	Population in millions from 1960 to 2017

Table 2: Summary of sources of data used in the project

the help of R code which involves the cleaning activity and loading the data into the data warehouse. The fact and the dimension tables are represented with the help of Star schema. Three non trivial BI queries are supported by the data warehouse and the visualization is done with the help of Tableau.

2 Data Sources

The collection of data has been done from 7 sources out of which eight data are structured data and one is unstructured.

2.1 Source 1: Statista

The Statista dataset has been downloaded from: `https://www.statista.com/statistics/263220/public-spending-ratio-in-eu-countries/` comprises of 27 countries of the European Union representing the information about the Public spending ratio made during the year 2017.Basically Public spending is the ratio which provides an insight of the general government total expenditure as a percentage of GDP. The Release date of the data is April 2018. The data was downloaded in the XLS format. After removing the sheet containing the description, the final file was saved in csv format and then the cleaning of data was performed in R which is mentioned in the report under the appendix section. The final file obtained is having the below columns:

- Country
- Year
- Government expenditure

2.2 Source 2: UNECE (United Nations Economic Commission for Europe)

The dataset has been downloaded from: https://w3.unece.org/PXWeb2015/pxweb/en/STAT/STAT__20-ME__2-MENA/10_en_MECCGDPExpY_r.px/ The dataset comprises of 59 countries representing the information about the Final consumption expenditure of households which provides the information about the amount in millions of US dollars spent by the people for meeting their day to day requirements. The latest data present in the source is for the year 2018. The use of R script is done for the purpose of cleansing of data, which is mentioned in the report under the appendix section. The final file obtained consists of the below columns:

- Country
- Year
- Household expenditure

2.3 Source 3: Eurostat

https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=t2020_10&plugin=1 The dataset comprises of 39 countries representing the information about the total employment rate from the year 2006 to 2017, focusing on the age group between 20-64 years. This data is helpful to figure out the percentage of people employed in the particular country out of the total population. The release date of the data is September 2018. The file was downloaded and then the cleansing of data as per requirement is done using R which is mentioned in the report under the appendix section. The final file obtained consists of below columns:

- Country
- Year
- Employment rate

2.4 Source 4: European Fiscal Board Report

https://ec.europa.eu/info/sites/info/files/2018_efb_annual_report_en.pdf The data is unstructured in form and has been web scraped from the website using R language. This is a European Fiscal Board Annual Report 2018 in the form of PDF. From this report, GDP growth rate has been picked that has an importance while forming one of the BI query. The report consists of 109 pages and the data has been scraped from page no. 100. The table consists of 30 countries and data present is from the year 2000 to 2019 but for the analysis purpose cleaning of data has been done and selected the required countries and GDP growth rate values from the year 2008 to 2017. There were two challenges faced while obtaining the data from the pdf. Firstly, the country codes were in Alpha 2 format, so after going through the website nationsonline (2019) , identification of the Alpha 3 codes were performed. Another challenge was the picking the complete table from the pdf without any missing information. For this R code has been written in such a way that picks all the information from the particular source page. The code is mentioned in the report under the appendix section. The final file obtained is having the below columns:

- Country
- Location
- Year
- GDP growth rate

2.5 Source 5: OECD (Organization for Economic Co-operation and Development)

Two data sources were picked from this website and the details are mentioned below:https://data.oecd.org/eduatt/adult-education-level.htm#indicator-chart
below:https://data.oecd.org/tax/tax-revenue.htm

Education attainment

In this dataset the information is presented about the adult aged between 25-64 years of age who have attained the upper secondary education. In total there were 47 countries and the data available was from the year 2000 to 2017. As per the requirement 20 countries were selected out of these 47 and the years from 2008 to 2017 were taken into account. The data cleansing has been performed using R code and the same has been mentioned in the report. This dataset is helpful in determining how important is upper secondary education attainment for the people in different countries and what are the prospects of getting an employment. The detail has been discussed in the BI query later in the report. From this data set six attributes were obtained that are Country, Indicator, subject, measure, year and education attainment value in percentage. The dataset information is the Education at a Glance publication of 2018. The final file obtained consists of the below columns:

- Country
- Indicator
- Subject
- Measure
- Education attainment

Government Revenue

From this data source, dataset on Government Revenue has been picked. Basically Government Revenue is the money that the Government receives from the public in the form of taxes. This dataset has been used later in the BI query which is discussed in the report under BI query section. In total there were 37 countries and the data present was from year 2000 to 2017. As per the requirement 20 countries were chosen and the values from the year 2008 to 2017. The data is published by Revenue statistics publication (2018).The final file obtained consists of the below columns:

- Country
- Year
- Government revenue

2.6 Source 6: IMF (International Monetary Fund)

Two data sources were picked from this website and the details are mentioned below:

https://www.imf.org/external/datamapper/NGDPD@WEO/OEMDC/ADVEC/WEOWORLD https://www.imf.org/external/pubs/ft/weo/2018/01/weodata/weorept.aspx?sy=2008&ey=2016&scsm=1&ssd=1&sort=country&ds=.&br=1&pr1.x=55&pr1.y=9&c=512%2C946%2C914%2C137%2C612%2C546%2C614%2C962%2C311%2C674%2C213%2C676%2C911%2C548%2C193%2C556%2C122%2C678%2C912%2C181%2C313%2C867%2C419%2C682%2C513%2C684%2C316%2C273%2C913%2C868%2C124%2C921%2C339%2C948%2C638%2C943%2C514%2C686%2C218%2C688%2C963%2C518%2C616%2C728%2C223%2C836%2C516%2C558%2C918%2C138%2C748%2C196%2C618%2C278%2C624%2C692%2C522%2C694%2C622%2C142%2C156%2C449%2C626%2C564%2C628%2C565%2C228%2C283%2C924%2C853%2C233%2C288%2C632%2C293%2C636%2C566%2C634%2C964%2C238%2C182%2C662%2C359%2C960%2C453%2C423%2C968%2C935%2C922%2C128%2C714%2C611%2C862%2C321%2C135%2C243%2C716%2C248%2C456%2C469%2C722%2C253%2C942%2C642%2C718%2C643%2C724%2C939%2C576%2C644%2C936%2C819%2C961%2C172%2C813%2C132%2C726%2C646%2C199%2C648%2C733%2C915%2C184%2C134%2C524%2C652%2C361%2C174%2C362%2C328%2C364%2C258%2C732%2C656%2C366%2C654%2C734%2C336%2C144%2C263%2C146%2C268%2C463%2C532%2C528%2C944%2C923%2C176%2C738%2C534%2C578%2C536%2C537%2C429%2C742%2C433%2C866%2C178%2C369%2C436%2C744%2C136%2C186%2C343%2C925%2C158%2C869%2C439%2C746%2C916%2C926%2C664%2C466%2C826%2C112%2C542%2C111%2C967%2C298%2C443%2C927%2C917%2C846%2C544%2C299%2C941%2C582%2C446%2C474%2C666%2C754%2C668%2C698%2C672&s=GGX_NGDP&grp=0&a=

Gross Domestic Product (GDP)

The GDP values corresponding to countries are given in billions of US dollars from the year 1980 till 2018 along with the forecast till 2023 for 228 countries. In order to proceed downloaded the data in the XLS format and using R code filtered the data for the required 20 countries from the year 2008 to 2017. The dataset is present under World Economic Outlook (October 2018) in the IMF website. The dataset has been chosen since it is very important for the analysis in one of the BI query. The R code used for the cleansing of data is mentioned later in the report under the appendix section. The final file obtained consists of the below columns:

- Country
- Year
- GDP billions

General government total expenditure

The data is representing the information about the government expenditure as a percentage of GDP. The information in the dataset was present for 198 countries. The countries that were required were filtered from the complete list for the year 2008 to 2016. All the cleansing in this respect has been performed using R code which is mentioned in the report. The publication date for this data is April 2018. The final file obtained consists of the below columns:

- Country

- Year
- Government expenditure

2.7 Source 7: The World Bank

https://data.worldbank.org/indicator/sp.pop.totl?end=2017&start=1960&view=chart From this data source, dataset on population has been taken. The data is required for getting the information about the country and its population. The original data from the source comprises of 264 countries and year ranging from 1960 to 2017. The data has been cleaned using R and the countries and year has been taken as per requirement. All the cleansing activity has been performed using R code. In order to make the population of the country to make it easy to read, the population of the country has been converted to millions using R code. The code is mentioned in the report under appendix section. The final file obtained consists of the below columns:

- Country
- Country code
- Subject
- Year
- Population millions

3 Related Work

The research performed by OECD OECD (2015) on the advantages of upper secondary education attainment states that in order to enter the labour market, attainment of atleast upper secondary education is required. The people who do not complete the upper secondary education have low skill set and face the problem when they enter the market in search of job.Upper secondary education is also a necessity for the people who wish to attain the further education and training. As per OECD, in the year 2013 around 40 percent of the adults in the age bracket of 25-64 years have attained upper secondary education as their highest level. The attainment of upper secondary education as per the article makes the person insured against the risk of getting unemployed. Although upper secondary education level has an open room for the adults in the ground of employment but having tertiary education can have an add-on advantage for the individual as well as the country in which they are residing.

The Paper Daniel (2005) deals on how the economic growth is impacted by the government spending. Explanation on the government spending as per the paper suggests that some of the government spending proves to be successful for the growth of the economy. The Government spending comes either from taxes or by borrowing from the market. There are no positive effect of increased government spending can be observed in the OECD countries (Connolly & Li (2016)). On a concluding node, an inference can be made from the paper that the there is a need to reduce the government spending as it has negative impact on the economic growth. If the government spending is done in an efficient manner then it would result in the great economic output.

As a part of the project as an add-on, government revenue is taken as a data source to make an effective comparison with the government expenditure. With this the comparison will be made that what happens if the government revenue is higher than what the government is spending. With this the rise or fall in the gdp growth rate will be observed in the countries that has been taken in the data source.

As per the paper Toossi (2005), consumer spending plays a crucial and important role in creating the employment opportunity. As the needs of the consumer increases, the production of goods and services also increases in order to satisfy the requirements which provide employment to the individuals of the country. As a part of this project, Gross domestic product is added as one of the data source in which the relation between the consumer spending, employment rate and GDP will be checked. The European union countries are taken into consideration and the BI query is made accordingly. The visualization will be through tableau and check the relation between these three.

A system that extract, transform and load the data into the dimensional data store which can be used for the analysis and decision making is known as the data warehouse.Kimball (2005). A data warehouse is capable to intensify the business output by efficiently collecting the information for the analysis purpose and tracking a trend over the particular time intervals Mankad (2013). Since the data used in this project is for 20 countries and for 10 years so there is a need to create a data warehouse for an effective as well efficient analysis and decision making. The bottom up approach is given by Ralph Kimball and the same has been applied in this data warehousing project.

The reason behind choosing this approach is that it is easier for the extension of data warehouse and less time consuming and relatively fasterdwbimaster (2017).

4 Data Model

For the design of the data warehouse, star schema is used. In star schema the data is arranged into facts and dimensions. A fact contains the information that can be counted while the dimension provides the information about the fact table. The diagram in which the fact is surrounded by the dimensions is known as the star schema Margaret (2017) In a data warehouse the connection of the dimension and fact table is made using the foreign key relation. The dimension table contains one primary key while the fact table contains the foreign key which is the primary key of each dimension. The description about the fact and the dimension tables are explained below:

Dimension Table

In the star schema, Dim-Country and Dim Year are the two dimensions. Here Country ID and Year ID are the two primary keys which are serving as the foreign keys in the FactTable. Dim Country contains all the distinct countries that are obtained from the third data source. Dim Year is the another dimension which contains the distinct years that are obtained from sixth data source.

Fact Table

The FactTable is having one primary key and two foreign keys that were the primary keys in the dimension tables and they are Country ID and Year ID. The various fields in

the fact table are GDP billions,GDP growth rate,Government expenditure,Government revenue, Household expenditure, Employment rate, Education attainment and Population millions that are obtained from the sources of data as mentioned as mentioned in the above section. In the BI query which will be discussed later in the report, the use of all the columns in the fact table has been taken care. In the first query country, year, Government reveneue and Government expenditure and gdp growth rate are used. While forming the second query involve Education attainment,Employment from the fact table and Country and Year from the dimension tables. The third and last query involve the use of Employment rate, Household expenditure, GDP billions along with Country and Year from the dimension tables.

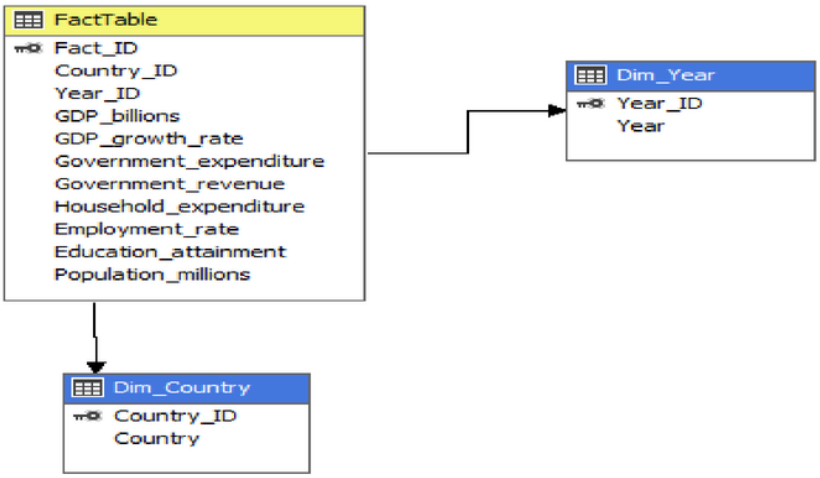

Figure 1: Star Schema

5 Logical Data Map

Logical data map has been represented below:

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure ??

Source	Column	Destination	Column	Type	Transformation
1	Country	Dim_Country	Country	Dimension	Filtering done on the basis of countries
2	Year	Dim_Year	Year	Dimension	Filtering done on the basis of year
6	GDP billions	FactTable	GDP billions	Fact	Rounded off to two decimal places
4	GDP Growth Rate	FactTable	GDP Growth Rate	Fact	Filtering done on the basis of Country
1	Government Expenditure	FactTable	Government Expenditure	Fact	Rounded off to two decimal places
5	Government Revenue	FactTable	Government Revenue	Fact	Rounded off to two decimal places
2	Household Expenditure	FactTable	Household Expenditure	Fact	Conversion to billion US dollars
3	Employment Rate	FactTable	Employment Rate	Fact	Filtering done on the basis of countries and year
5	Education Attainment	FactTable	Education Attainment	Fact	Rounded to two decimal places
7	Population millions	FactTable	Population millions	Fact	Rounded to make into readable form

6 ETL Process

ETL stands for Extract, Transform and Load. On elaborating further, it can be defined as extracting or pulling the data from the source, transforming the data which means performing any cleansing activity on the data so as to make it ready for the loading into the database.

The complete process in detail is explained below:

Extraction of data

The extraction of data is done from total seven data sources out of which eight are structured and one is unstructured. For extracting the unstructured data from annual report 2018 of European Fiscal Board, tabulizer package in R is used. For the structured sources the extraction of data is done from the source using execute process task in the SSIS.

Transformation of data

Cleaning of data is considered to be an important part in the process as with the help of this the redundant as well as the incorrect records are removed from the data. Data cleaning is done so as to ensure that the correct data flows throughout the entire process. For the data cleansing part R programming language is used and explanation of all the data source is mentioned below:

Source 1: The data downloading was done from Statista and after that the data comprising of 27 countries for the year 2017 was obtained. R code is used for data cleansing which comprises of removal of unwanted columns, renaming of columns. After this the required countries were filtered. In addition the government expenditure was rounded off to 2 digits and finally the file obtained was saved in the csv format.

Source 2: This data was downloaded from UNCECE in which there were 59 countries. R code is used for data cleansing which comprises of removing the unwanted columns, renaming the columns, and filtering on the basis of countries, in addition used gather function which comes under tidyr package to gather all the years under the column name Year and then finally arrangement of the data was done on the basis of countries. The expenditure on household spending then converted into billions of US dollars. The file obtained was finally saved in the csv format.

Source 3: This dataset was downloaded from Eurostat consisting of 39 countries. The use of R code is made to clean the data that involves removing of the unwanted column from the data, renaming of the columns, vector creation for the required countries and then filtering those countries as per the requirement. Used gather function to gather all the years under the column name year and finally arranging the data on the basis of countries. The cleaned file obtained was finally saved in the csv format.

Source 4: This is an unstructured data source in which web scraping was performed. From this report, GDP growth rate has been picked that has an importance while forming one of the BI query. The report consists of 109 pages and the data has been scraped from page no. 100. The table consists of 30 countries and data present is from the year

2000 to 2019 but for the analysis purpose cleaning of data has been done and selected the required countries and GDP growth rate values from the year 2008 to 2017. The cleansing part includes removing the unwanted columns, renaming the columns, vector creation for countries and filtering the countries as required. The use of gather function was made to gather all the years under Year column and finally arranging the data on the basis of countries. In this part of cleansing total of three R packages were used i.e. tabulizer, dplyr and tidyr. The final file obtained was saved in the csv format.

Source 5: In total two datasets were extracted from OECD. The data sets are education attainment and government revenue. There were extra countries in both the datasets so cleansing of data performed to select the countries as per the requirement. For the education attainment dataset filtering was also performed on the year. As a part of cleaning removal of redundant columns, followed by renaming of columns and gathering on the basis of years and placing under the year column and then finally arranging the complete data on the basis of countries was performed. Government revenue is rounded off to two decimal places and then the cleaned data obtained was saved in the csv format.

Source 6:In this two data sources were extracted from IMF namely government total expenditure and gross domestic product. As a part of cleaning removal of redundant columns, followed by renaming of columns and gathering on the basis of years and placing all the years under the year column and then finally arranging the complete data on the basis of countries was performed. The total government expenditure is rounded off to two decimal places and the cleaned data obtained was saved in the csv format.

Source 7: The data set was extracted from World Bank consisting of the information about the population of countries corresponding to the years. In total there were 264 countries and years were from 1960 to 2017. Since, as per the requirement 20 countries and 10 years data were needed so cleansing was performed. All the cleansing activity has been performed using R code which includes operations related to removing unwanted columns, renaming of columns and gathering the data on the basis of years and arranging the complete data in the according to the countries. In order to make the population of the country to make it easy to read, the population of the country has been converted to millions using R code.

7 Application

Rationale and evaluation approach with respect to addressing the business requirements noted in Section 1, i.e. how have you used the case studies / BI Queries to address and demonstrate the attainment of your business requirements.

7.1 BI Query 1: How does an increased household expenditure affects the employment rate and Gross domestic product?

For the generation of this BI query, three structured datasets i.e. household expenditure, employment rate and Gross domestic product are compared with each other. As it can be seen from the above visualization that Germany having household expenditure of 1741 Billion US dollars in the year 2016 and it increased and reached to 1772 billion US dollars

in 2017. With an increase in the household expenditure, the employment rate can also be seen to increase from 78.60percent to 79.20 percent from the year 2016 and 2017 respectively. With the increase in both the household expenditure and employment rate it can observed that the gross domestic product also got increased from 3497 billion US dollars to 3701 billion US dollars and the same has been observed in case of Ireland and other countries as well. Thus it can be inferred that with an increase in the household expenditure, consumers will buy more goods and services which in turns increases the demand in the market that ultimately creates an employment opportunity and thus in this way contributes to the increase in the gross domestic product.

Figure 2: BI Query 1

7.2 BI Query 2: How does upper secondary education attainment is related to employment rate and population?

This BI query includes structured datasets i.e. education attainment, employment rate and population. For the countries like Greece, Luxembourg, Netherlands, Spain and United Kingdom with an increase in the upper secondary education attainment it can be seen that the increment in the employment rate can be observed. As in case of Greece the upper secondary education attainment increased from 41.42percent to 41.68percent from the year 2016 to 2017, which in turns to be increment in case of employment rate as the employment rate increased from 56.20percent to 57.80percent from 2016 to 2017.And in the case of Luxembourg the upper secondary education attainment increment can be observed and the same with the employment rate. This trend can be seen in the other countries like Portugal, Spain, United Kingdom, Hungary, and Italy. But on comparing the countries like Sweden, Austria, Germany and Finland on the same parameters it can inferred that with a decrease in the upper secondary education attainment among the

people , the employment rate is not that much getting impacted. Thus it is presenting the information that the countries like Greece, Luxembourg, Netherlands, United Kingdom, Hungary upper secondary education acts like a gateway to the employment market and removes the risk of unemployment. Also, it has been observed that although the population of the countries like Czech Republic and Slovak Republic are less as compared to the higher population country like Germany but the skill set and the education attainment are making them competent enough to compete with the higher population country like Germany, Italy and United Kingdom.

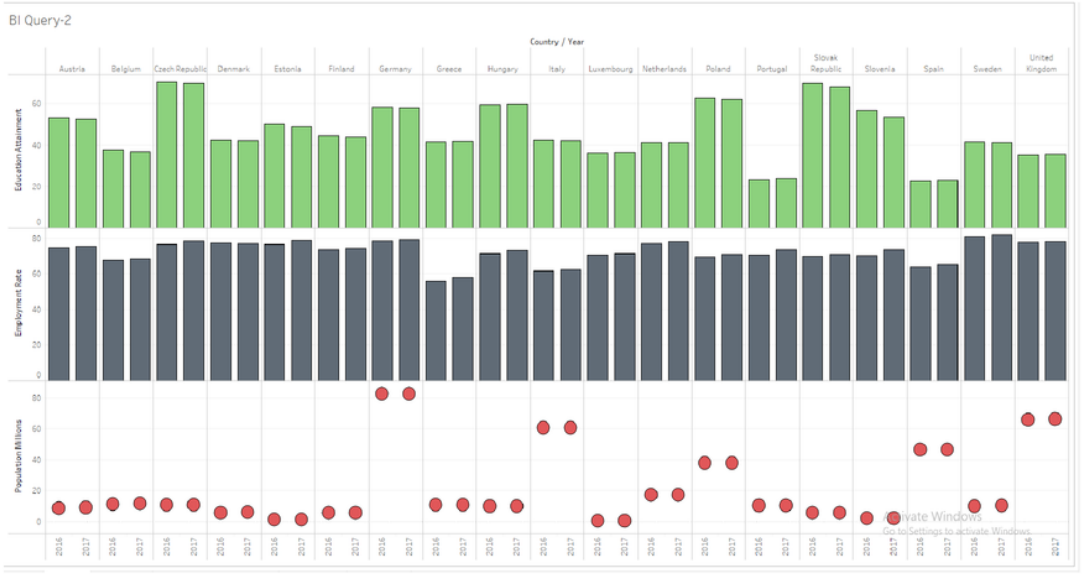

Figure 3: BI Query 2

7.3 BI Query 3: How does Government revenue and Government expenditure affects the GDP growth rate?

For the generation of this BI query two structured data i.e. government revenue, general government total expenditure and GDP growth rate which is the unstructured dataset have been used. The Government expenditure and the Government revenue as a percentage of GDP has been taken for the year 2015 and 2016 respectively and comparison of these has been performed with the GDP growth of the 2016 and 2017 as the positive or the negative effects can be seen in the succeeding year only. As it can been from the above visualization, in Czech Republic during the year 2015 the government expenditure as a percentage of GDP was slightly higher as compared to the government revenue, but in 2016 the expenditure became less as compared to revenue, so in the GDP growth rate 2017 there is an increase as compared to the previous year. The same can be seen in the countries like Finland, Germany, Greece and Portugal. Also on comparing the trend it has been inferred that even though the expenditure in some scenarios are higher than the revenue but if in the coming year it has been reduced to some extent then also gdp

is growing , this can be seen in the countries like Italy, Poland and Denmark. So it can be said that if the balance is maintained between the Government expenditure and Government revenue then there are chances of the stable economy.

The general findings are that . . . as illustrated in Figure ??.

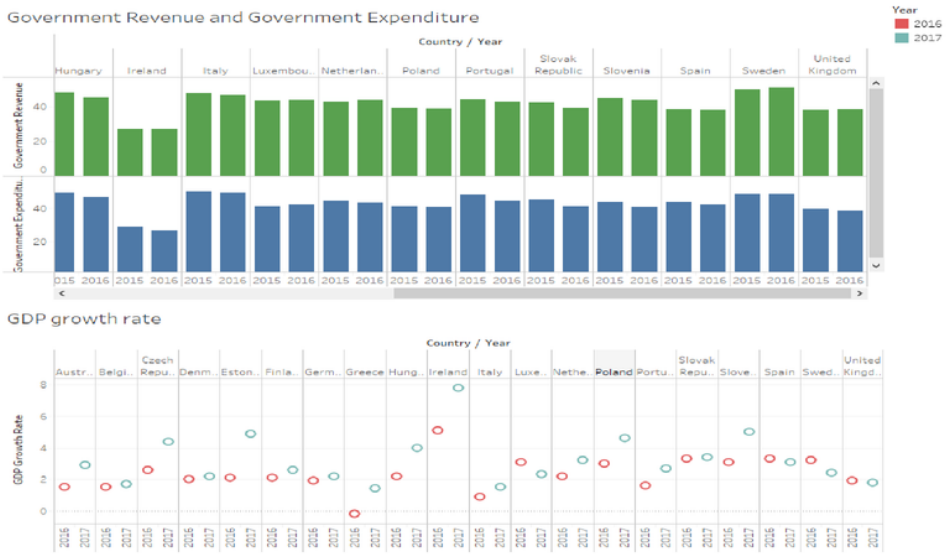

Figure 4: BI Query 3

7.4 Discussion

The research performed by OECD OECD (2015) states that the attainment of the upper secondary education is important for the entry into the labor market. The attainment of upper secondary education as per the article makes the person insured against the risk of getting unemployed. The same have been analyzed in this project. The trends were observed over a period of time and it is therefore inferred that the individuals in the country with at least upper secondary education attainment are eligible for the employment opportunity. So minimum of upper secondary education must be attained by the individual in the age bracket of 25-64 years of age.

As per the paper Toossi (2005), consumer spending plays a crucial and important role in creating the employment opportunity. As a part of the analysis the same has been analysed in the project for the European Union countries and it was inferred that the factors are related to each other. The countries that were taken into consideration shows the trends that with the increase of consumer spending or the household expenditure, employment rate is also increasing and which are ultimately proving to be useful for the nation as it is contributing to the gross domestic product and making the economy strong and healthier.

8 Conclusion and Future Work

On a concluding note it can be said that the economic indicators plays a significant role in determining the economic state of a nation. The indicators discussed in this project presented the overall idea about their importance. The relationship between the household expenditure and the employment rate presented the information that when these two indicators are combined with each other then they can help the economy to boom. Apart from this upper secondary education attainment for the particular age groups not only helps them to secure the job in the competitive market but also it is required for the people who wish to attain higher education and the related training.

On the other hand for the analysis of trends over a period of time, gross domestic product and gdp growth rate are crucial because on comparing them with the different factors the health of economy can be analyzed i.e. whether the economy of the country is increasing or decreasing. However apart from the indicators that are discussed in this project, there are many other factors as well that can be taken into consideration, so that when they are included can enhance the level of discussion of the topic.

9 Youtube Link of the Video

https://www.youtube.com/watch?v=8AHSv_WePVM&feature=youtu.be

References

Amadeo, K. (2019), 'Economic growth, its measurements, causes, and effects', https://www.thebalance.com/what-is-economic-growth-3306014.

Connolly, M. & Li, C. (2016), 'Government spending and economic growth in the oecd countries', *Journal of Economic Policy Reform* **19**(4), 386395.

Daniel (2005), 'The impact of government spending on economic growth', https://www.heritage.org/budget-and-spending/report/the-impact-government-spending-economic-growth.

dwbimaster (2017), 'dwbimaster', http://dwbimaster.com/bottom-up-vs-top-down-approach-in-data-warehouse/.

Kimball (2005), 'The data warehouse etl toolkit', http://users.itk.ppke.hu/~szoer/DW/Kimball%20&%20Caserta%20-The%20Data%20Warehouse%20ETL%20Toolkit%20%5BWiley%202004%5D.pdf.

Mankad, M. D. (2013), 'The study on data warehouse design and usage for u.s. job growth', http://www.ijsrp.org/research-paper-0313/ijsrp-p1597.pdf.

Margaret, R. (2017), 'searchdatamanagement.techtarge', https://searchdatamanagement.techtarget.com/definition/star-schema.

nationsonline (2019), 'Country codes list', https://www.nationsonline.org/oneworld/country_code_list.htm.

OECD (2015), 'Education indicators in focus', https://www.oecd-ilibrary.org/docserver/5jrw5p4jn426-en.pdf?expires=1554769113&id=id&accname=guest&checksum=4130FDD8148CBCF0CB0CA26D2C0F55D3.

Toossi, M. (2005), 'Consumer spending: an engine for u.s. job growth', https://www.bls.gov/opub/mlr/2002/11/art2full.pdf.

Wilkinson, J. (2013), 'Economic indicator', https://strategiccfo.com/economic-indicators/.

Appendix

R code Source 5

```
# Education attainment in Europe

# Installing the required packages

#install.packages("tidyr")
library(tidyr)
#install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("C:/Users/MOLAP/Documents/R/csv files")

# reading the input file and store in initial

initial<- read.csv("education_attainment_oecd.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial <- initial %>% select (-c(5,8))

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "Indicator"
colnames(initial)[3] <- "Subject"
colnames(initial)[4] <- "Measure"
colnames(initial)[5] <- "Year"
colnames(initial)[6] <- "Education_attainment"

# vector creation for the required countries
```

```
store<-c('Austria','Belgium','Czech Republic','Denmark','Estonia','Finland',

# filtering the condition for COuntry_code,Year and Subject

final<- filter(initial,Country %in% store)

yr<- as.numeric(2008:2017)

final <- filter(final, Year %in% yr)

sub<- c('UPPSRY')

final<-filter(final,Subject %in% sub)

# Rounding off the value to 2 digits

final$Education_attainment<-round(final$Education_attainment,digits=2)

# writing the final csv file

write.csv(final,file="Cleaned_data/education_attainment.csv",row.names = F)


# R code for Government revenue

# Installing the required packages
install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("E:/DWBI Project Backup")

# reading the input file and store in initial

initial<- read.csv("GOVT_REV_DATA.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[2:5]]<- NULL
initial[store[8]]<- NULL

# renaming the columns
```

```
colnames(initial)[1] <- "Country_code"
colnames(initial)[2] <- "Year"
colnames(initial)[3] <- "Government_revenue"

# vector creation for the required country_code

store<-c('AUT','BEL','CZE','DNK','EST','FIN','FRA','DEU','GRC','HUN','IRL','

# filtering the condition for Country_code

final<- filter(initial,Country_code %in% store)

# vector creation for the required Country

Country<- c('Austria','Belgium','Czech␣Republic','Denmark','Estonia','Finlan

# filtering the condition for Country_code

final<-cbind(Country,final)

# removing the unwanted column

final[2]<- NULL

# arranging on the basis of Country

final<-final%>%arrange(Country)

# rounding off to two digits

final$Government_revenue<-round(final$Government_revenue,digits=2)

# writing the final csv file

write.csv(final,file="Cleaned_data/Government_revenue.csv",row.names = F)
```

R code Source 6

```
# R code for GDP

# Installing the required packages

install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("C:/Users/MOLAP/Documents/R/csv␣files")
```

```r
# reading the input file and store in initial

initial<- read.csv("gdp.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[2:29]]<- NULL
initial[store[40:45]]<- NULL

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "2008"
colnames(initial)[3] <- "2009"
colnames(initial)[4] <- "2010"
colnames(initial)[5] <- "2011"
colnames(initial)[6] <- "2012"
colnames(initial)[7] <- "2013"
colnames(initial)[8] <- "2014"
colnames(initial)[9] <- "2015"
colnames(initial)[10] <- "2016"
colnames(initial)[11] <- "2017"

# vector creation for the required countries

store<-c('Austria','Belgium','Czech␣Republic','Denmark','Estonia','Finland',

# filtering the condition for Country

final<- filter(initial,Country %in% store)

# writing the csv file

write.csv(final,"intermediate/GDP_value.csv",row.names = F)

#using gather function

gdp <- read.csv("intermediate/GDP_value.csv", stringsAsFactors = FALSE, chec
gdp1<- gather(gdp,Year,GDP_billions,-Country)

# arranging on the basis of Country

gdp1<-gdp1%>%arrange(Country)

# Writing the final csv file

write.csv(gdp1,"Cleaned_data/GDP.csv",row.names = F)
```

```r
# R code for General government expenditure

# Installing the required packages

install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("C:/Users/MOLAP/Documents/R/csv files")

# reading the input file and store in initial

initial<- read.csv("public_spend_ratio.csv")


# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[2:5]]<- NULL
initial[store[16]]<- NULL
initial[store[15]]<- NULL

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "2008"
colnames(initial)[3] <- "2009"
colnames(initial)[4] <- "2010"
colnames(initial)[5] <- "2011"
colnames(initial)[6] <- "2012"
colnames(initial)[7] <- "2013"
colnames(initial)[8] <- "2014"
colnames(initial)[9] <- "2015"
colnames(initial)[10] <- "2016"

# vector creation for the required countries

store<-c('Austria','Belgium','Czech Republic','Denmark','Estonia','Finland',

# filtering the condition for Country

final<- filter(initial,Country %in% store)
```

```r
# writing the csv file

write.csv(final,file="intermediate/spending_ratio_i.csv",row.names = F)

# reading the csv file

spend <- read.csv("intermediate/spending_ratio_i.csv", stringsAsFactors = FA

# using gather function

spend1 <- gather(spend,Year,Government_expenditure,-Country)

# reading another csv file

initial1<- read.csv("public_spending_ratio_s.csv")

# get the column names

store1<- colnames(initial1)

# removing unwanted columns

initial1[store1[3]]<- NULL

# renaming the columns

colnames(initial1)[1] <- "Country"
colnames(initial1)[2] <- "2017"

# vector creation for the required countries

store1<-c('Austria','Belgium','Czech␣Republic','Denmark','Estonia','Finland'

# filtering the condition for Country

final1<- filter(initial1,Country %in% store1)

# writing the csv file

write.csv(final1,file="intermediate/spending_ratio_s.csv",row.names = F)

# using gather function

sp <- read.csv(file="intermediate/spending_ratio_s.csv", stringsAsFactors =
sp1<- gather(sp,Year,Government_expenditure,-Country)

# arranging on the basis of country

spend1<-spend1%>%arrange(Country)

#write.csv(spend1,file="Cleaned_data/spend_ratio.csv",row.names = F)
```

```r
# using rbind to bind spend1 and sp1

comb<-rbind(spend1,sp1)

# arranging on the basis of country

comb<- comb%>% arrange(Country)

# rounding off to 2 digits

comb$Government_expenditure<-round(comb$Government_expenditure,digits=2)

# writing the final csv file

write.csv(comb,file="Cleaned_data/general_govt_exp_data.csv",row.names = F)
```

R code Source 7

```r
# Installing the required packages

install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("C:/Users/MOLAP/Documents/R/csv␣files")

# reading the input file and store in initial

initial<- read.csv("population.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[4:52]]<- NULL
initial[store[63]]<- NULL

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "Country_code"
colnames(initial)[3] <- "Subject"
colnames(initial)[4] <- "2008"
colnames(initial)[5] <- "2009"
colnames(initial)[6] <- "2010"
```

```
colnames(initial)[7] <- "2011"
colnames(initial)[8] <- "2012"
colnames(initial)[9] <- "2013"
colnames(initial)[10] <- "2014"
colnames(initial)[11] <- "2015"
colnames(initial)[12] <- "2016"
colnames(initial)[13] <- "2017"

# vector creation for the required countries

store<-c('Austria','Belgium','Czech␣Republic','Denmark','Estonia','Finland',

# filtering the condition for Country

final<- filter(initial,Country %in% store)

# writing the csv file

write.csv(final,"intermediate/pop.csv",row.names = F)

#using gather function

popu <- read.csv("intermediate/pop.csv", stringsAsFactors = FALSE, check.nam
popu1<- gather(popu,Year,Population_millions,-Country,-Country_code,-Subject

# Converting the population into millions and rounding off to 2 digits

popu1<-popu1%>%arrange(Country)
popu1$Population_millions<-(popu1$Population_millions/1000000)
popu1$Population_millions<-round(popu1$Population_millions,digits=2)

# Writing the final csv file

write.csv(popu1,"Cleaned_data/population.csv",row.names = F)
```

R code Source 1

```
install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)
setwd("C:/Users/MOLAP/Documents/R/csv␣files")
initial<- read.csv("public_spending_ratio_s.csv")
store<- colnames(initial)
initial[store[3]]<- NULL
colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "2017"
store<-c('Austria','Belgium','Czech␣Republic','Denmark','Estonia','Finland',
final<- filter(initial,Country %in% store)
write.csv(final,file="intermediate/spending_ratio_s.csv",row.names = F)
sp <- read.csv(file="intermediate/spending_ratio_s.csv", stringsAsFactors =
sp1<- gather(sp,Year,Government_expenditure,-Country)
```

```
sp1<-sp1%>%arrange(Country)
write.csv(sp1,file="Cleaned_data/Spending_ratio_data.csv",row.names = F)
```

R code Source 4

```
# Installing the required packages
install.packages("tidyr")
install.packages("tabulizer")
install.packages("dplyr")
library(tabulizer)
library(tidyr)
library(dplyr)
initial<- 'https://ec.europa.eu/info/sites/info/files/2018_efb_annual_report
extract <-extract_tables(initial,pages=100)
info <-do.call(rbind,extract)
info<- as.data.frame(info)
info <- info %>% select (-c(2,13,14))
colnames(info)[1] <- "Location"
colnames(info)[2] <- "2008"
colnames(info)[3] <- "2009"
colnames(info)[4] <- "2010"
colnames(info)[5] <- "2011"
colnames(info)[6] <- "2012"
colnames(info)[7] <- "2013"
colnames(info)[8] <- "2014"
colnames(info)[9] <- "2015"
colnames(info)[10] <- "2016"
colnames(info)[11] <- "2017"
count <- c('AT','BE','CZ','DK','EE','FI','FR','DE','EL','HU','IE','IT','LU',
final<- filter(info,Location %in% count)
Country_code <- c('Belgium','Czech Republic','Denmark','Germany','Estonia','

# Country_code <- c('BEL','CZE','DNK','DEU','EST','IRL','GRC','ESP','FRA','I
final<-cbind(Country_code,final)
write.csv(final,"gdp_rate.csv",row.names = F)
gdp <- read.csv("gdp_rate.csv", stringsAsFactors = FALSE, check.names=FALSE)
gdp1 <- gather(gdp,Year,GDP_growth_rate,-Location,-Country_code)
gdp1<-gdp1%>%arrange(Country_code)
write.csv(gdp1,"GDP_growth_rate.csv",row.names = F)
```

R code Source 3

```
# Installing the required packages

install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory
```

```r
setwd("C:/Users/MOLAP/Documents/R/csv files")

# reading the input file and store in initial

initial<- read.csv("Employment_rate.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[2:31]]<- NULL
initial[store[33]]<- NULL
initial[store[35]]<- NULL
initial[store[37]]<- NULL
initial[store[39]]<- NULL
initial[store[41]]<- NULL
initial[store[43]]<- NULL
initial[store[45]]<- NULL
initial[store[47]]<- NULL
initial[store[49]]<- NULL
initial[store[51:52]]<- NULL

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "2008"
colnames(initial)[3] <- "2009"
colnames(initial)[4] <- "2010"
colnames(initial)[5] <- "2011"
colnames(initial)[6] <- "2012"
colnames(initial)[7] <- "2013"
colnames(initial)[8] <- "2014"
colnames(initial)[9] <- "2015"
colnames(initial)[10] <- "2016"
colnames(initial)[11] <- "2017"

# vector creation for the required countries

store<-c('Austria','Belgium','Czech Republic','Denmark','Estonia','Finland',

# filtering the condition for Country

final<- filter(initial,Country %in% store)

# writing the csv file

write.csv(final,file="intermediate/Employment_rate.csv",row.names = F)
```

```r
#using gather function

Emp <- read.csv(file="intermediate/Employment_rate.csv", stringsAsFactors =
Emp1 <- gather(Emp,Year,Employment_rate,-Country)

# arranging on the basis of Country

Emp1<-Emp1%>%arrange(Country)

# Writing the final csv file

write.csv(Emp1,file="Cleaned_data/Employment_rate.csv",row.names = F)
```

R code Source 2

```r
# Installing the required packages

install.packages("tidyr")
library(tidyr)
install.packages("dplyr")
library(dplyr)

# Setting the working directory

setwd("C:/Users/MOLAP/Documents/R/csv␣files")

# reading the input file and store in initial

initial<- read.csv("House_expend.csv")

# get the column names

store<- colnames(initial)

# remove the unwanted columns

initial[store[1:2]]<- NULL

# renaming the columns

colnames(initial)[1] <- "Country"
colnames(initial)[2] <- "2008"
colnames(initial)[3] <- "2009"
colnames(initial)[4] <- "2010"
colnames(initial)[5] <- "2011"
colnames(initial)[6] <- "2012"
colnames(initial)[7] <- "2013"
colnames(initial)[8] <- "2014"
colnames(initial)[9] <- "2015"
colnames(initial)[10] <- "2016"
```

```r
colnames(initial)[11] <- "2017"

# vector creation for the required countries

store<-c('Austria','Belgium','Czech Republic','Denmark','Estonia','Finland',

# filtering the condition for Country

final<- filter(initial,Country %in% store)

# writing the csv file

write.csv(final,file="intermediate/house_exp_data.csv",row.names = F)


#using gather function

Exp <- read.csv("intermediate/house_exp_data.csv", stringsAsFactors = FALSE,
Exp1 <- gather(Exp,Year,Household_expenditure,-Country)

# performing calculation and arranging on the basis of Country

Exp1$Household_expenditure<- (Exp1$Household_expenditure/1000)
Exp1<- Exp1%>% arrange(Country)

# Writing the final csv file

write.csv(Exp1,"Cleaned_data/Household_expenditure_data.csv",row.names = F)
```

Project_Report_x18127339

ORIGINALITY REPORT

21%	10%	2%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to National College of Ireland Student Paper	12%
2	Submitted to Aberystwyth University Student Paper	3%
3	excelkingdom.blogspot.com Internet Source	1%
4	Submitted to Staffordshire University Student Paper	<1%
5	Submitted to South University Student Paper	<1%
6	www.tandfonline.com Internet Source	<1%
7	JOHN GURLAND, ROGER S. McCULLOUGH. "Testing equality of means after a preliminary test of equality of variances", Biometrika, 1962 Publication	<1%
8	Submitted to Upper Iowa University Student Paper	<1%

9	Submitted to Glasgow Caledonian University Student Paper	<1%
10	ec.europa.eu Internet Source	<1%
11	trap.ncirl.ie Internet Source	<1%
12	impulse-science.ru Internet Source	<1%
13	Submitted to University of Hong Kong Student Paper	<1%
14	dspace.cuni.cz Internet Source	<1%
15	digilib.uin-suka.ac.id Internet Source	<1%
16	www.sgi-network.org Internet Source	<1%
17	aei.pitt.edu Internet Source	<1%
18	Submitted to Kaplan University Student Paper	<1%
19	Submitted to University of Iowa Student Paper	<1%
20	Submitted to University College London Student Paper	

		<1%
21	Submitted to Franklin High School Student Paper	<1%
22	Submitted to Institute of Technology, Nirma University Student Paper	<1%
23	Submitted to Deakin University Student Paper	<1%
24	gall.dcinside.com Internet Source	<1%
25	juncharees.blogspot.com Internet Source	<1%
26	www.cesifo.de Internet Source	<1%
27	Brunello, G.. "Education and earnings growth: evidence from 11 European countries", Economics of Education Review, 200402 Publication	<1%
28	Zhai Dongsheng, Ma Wenshan. "Design data warehouse architecture of English learning behavior", 2010 International Conference on Information, Networking and Automation (ICINA), 2010 Publication	<1%

29	lirias.kuleuven.be Internet Source	<1%
30	www.opendemocracy.net Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		