Data Warehousing and Business Intelligence Project

on

Econometrics in a Data Warehouse

Abhishek Angne
x18136923

MSc Data Analytics – 2019/20

Submitted to: Sean Heeney

https://youtu.be/ohM9p_aE9as

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

Student Name:	Abhishek Angne
Student ID:	x18136923
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Heeney
Submission Due Date:	26/11/2018
Project Title:	Econometrics in a Data Warehouse

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

Econometrics in a Data Warehouse

Abhishek Angne
x18136923

April 12, 2019

Abstract

Econometrics is actually a statistical method of analyzing the different factors related to the economic growth of a nation. In this project, we will try to analyze the relation between different economic factors. The factors that we are going to discuss in detail about are Population, the Gross Development Product(GDP) Per Capita PPS of a country, the Human Development Index(HDI), Foreign direct investment, World Happiness Report, unemployment rate and education of a particular country in a year wise format.

1 Introduction

They say the strength of an economy is to provide every person the right amount of education. Economic growth is a multidimensional concept and it depends on a lot of factors like education, trade, GDP, employment and much more. This data model aims to answer three specific queries or business requirements on various measures taken into consideration while looking at the various economic factors(7). In-depth research and various projects on economic factors and demographics leads to a lot of data generation on the basis of time series. Modern economists are aware that conventional measures and factors of calculating economic values and data do not always give the best of results.

(Req-1) What is the effect of a change in the Population and percentage of education on the rate of unemployment, and thus on the GDP Per Capita for a set of European nations?
This BI query looks at the relationship and impact of overall growth of population on the total percentage of education, unemployment rate and thus on the GDP Per Capita as per the Purchasing Power Standards.

(Req-2) Is there a significant dependency between the Human Development Index and the World Happiness Report score?
This query will look at the two economic indices namely the World Happiness Report and the Human Development Index and try finding out their inter-relativity and the reason why two separate indices for measuring a particular set of factors.

(Req-3) How impacting is the inflow of Foreign Direct Investment on a positive trade balance and thus on the Exchange Rate of Ireland from 2014-2017?
Through this query we are going to see if the Foreign Direct Investments inflows have an effect on the trade balance and thus the exchange rate of Ireland.

Source	Type	Brief Summary
UNDP Human Development Reports	Structured	This data source provides 2 datasets which provide year wise of the values of Human Development Index, the various factors that affect the index and the foreign direct investment
Eurostat	Structured	This data source provides data about GDP Per Capita as per the purchasing power standards of all the countries taken into consideration on a yearly basis
Statista	Structured	Two datasets have been extracted from this source. Data about the exchange rate between Euro to Pound and the Trade Balance of Ireland has been used from these data sources
Worldbank	Structured	Three datasets have been extracted from this source. Data about the Total population, male and female population has been extracted from this source.
International Monetary Fund	Structured	Data extracted from this source gives us information about the Unemployment Rate of various country. Data is represented on a gender wise and year wise basis
GitHub	Structured	2 datasets on the World Happiness Indices for the years 215 and 2016 have been extracted from GitHub
Wikipedia	Unstructured	Data for the World Happiness Index has been scraped from Wikipedia for the years 2017 and 2018
OECD.org	Structured	The dataset extracted from OECD has gender and year-wise data for 23 countries

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

2 Data Sources

This project has been developed for analyzing the various economic indices. A total of 9 data sets have been used to understand where a particular country needs to work on to improve their economic and social conditions. A total of 26 European countries have been considered for the analysis.

2.1 Source 1: United Nations Development Programme Human Development Reports

Dataset name: Human Development Index(HDI)
Release year: 2018
`http://hdr.undp.org/en/indicators/137506`

About: The data set on Human Development Index gives us an overall measure HDI for all the countries in the world. HDI is calculated as an aggregate of three factors namely life expectancy, standard of living and availability of knowledge which are three separate indices namely Life Expectancy index, GNI index and Education index respectively.

Dataset name: Foreign Direct Investment, net inflows(percentage of GDP)
Release year: 2018
`http://hdr.undp.org/en/data`
About: The data set on the Foreign Direct Investment, net inflows give us information about the amount of inward direct investments made by non-residential(external) investors in the reporting economy.

2.2 Source 2: Eurostat

GDP Per Capita PPS:
Release year: 2018
`https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00114&plugin=1`
About: This dataset consists of data on the Gross Domestic Product, per capita as per purchasing power standard on a yearly and country-wise basis. GDP is computed as the value of the goods and services produced subtracted by the amount of goods and services utilized in their production.

2.3 Source 3: Statista

Ireland: Trade balance from 2007 to 2017 (in billion U.S. dollars):
Release year: 2018
`https://www.statista.com/statistics/376998/trade-balance-of-ireland/`
About: Trade balance of Ireland in billion USD is calculated as exported goods subtracted from imported goods. A positive trade balance signifies trade surplus and negative trade balance signifies trade deficit.
Usage: This dataset was downloaded in a .xls format and it comprises of year and country dimensions from my data model. It addresses the business requirements listed in Section1(Requirement 3)

Euro (EUR) to British pound (GBP) monthly exchange rate from December 2014 to February 2019:
Release year: 2019
`https://www.statista.com/statistics/438077/euro-to-british-pound-monthly-exchange-rat`
About: This dataset displays the euro to british pound monthly exchange rate as of the last day of the month.
Usage: Data was downloaded in a .xls format and it comprises of month-wise data which was then converted to year to match the year dimension by considering the exchange rate changes from December to December basis as this data was taken only on the last day of the month. Section1(Requirement 3)

2.4 Source 4: Worldbank

Dataset name: Population
Release year: 2018
`https://data.worldbank.org/indicator/SP.POP.TOTL`
`https://data.worldbank.org/indicator/SP.POP.TOTL.MA.IN`
`https://data.worldbank.org/indicator/SP.POP.TOTL.FE.IN`
About: Three datasets were extracted from Worldbank. These datasets give information on the population in numbers of different countries in the world in a year-wise and gender-wise formats
Usage: Data was downloaded in a .csv format and it was loaded on to R for cleaning purposes. Section1(Requirement 1)

2.5 Source 5: International Monetary Fund

Dataset name: Population
Release year: 2018
`https://www.imf.org/external/pubs/ft/weo/2018/01/weodata/index.aspx`
About: Data on unemployment was extracted from this source. This dataset consists of the overall employment rate on a country-wise and year-wise dimension level.

2.6 Source 6: GitHub

Dataset name: World Happiness Report
Release year: 2018
`https://github.com/bkumar080/World-Happiness-Report`
About: Data for 2 years(2015 and 2016) was extracted from this source for the years 2015 and 2016. The World Happiness Report is a landmark(survey) report of the state of global happiness that analyses the rating by looking at how happy the citizens feel they are.

2.7 Source 7: Wikipedia

Dataset name: World Happiness Report(WHR)
Release year: 2018
`https://en.wikipedia.org/wiki/World_Happiness_Report`
AboutData for the years 2017 and 2018 was extracted from this Wikipedia page. This dataset includes fields like the WHR score itself, country rankings by happiness, healthy life expectancy and freedom of expression which gives us an individual and overall insight about a particular country.

2.8 Source 8: OECD

Dataset name: Education, gender-wise
Release year: 2018

https://stats.oecd.org/Index.aspx?QueryId=70670
About: This dataset discusses the male and female education rates of a particular country in a year-wise format.

3 Related Work

Econometrics is an area which requires continuous research on a regular basis. A lot of research papers are available which have plenty of good ideas and information. The BI queries in this project are based on the previous related research work presented in the form of undermentioned technical papers.

3.1 Usage of similar data in previous work:

1 In Appiah (2017), the authors are testing whether an increase in the education expenditure in developing countries affects the GDP Per Capita in a positive manner.

2 In Blanchflower et al. (2006), the authors are analyzing the relationship between HDI and the Happiness where they are analyzing as to why Australia ranks high in the development index and comparatively low in the happiness index.

3 In Hong (2014) the author is explaining whether FDI promotes economic growth in China and if yes, how? The study is based on dynamic panel data from 254 prefecture-level cities in China from the year 1994 to 2010.

4 In Hall & Helliwell (2014), the authors are analyzing the relationship and differences between the World Happiness Report and the Human Development Index. Their idea is to understand the difference between these two despite the vast amount of similarities.

5 In Comim (2016), the authors are researching and providing measures or solutions about a measure beyond the Human Development Index for economic growth.

6 In Bahmani-Oskooee & Baek (n.d.), the authors are trying to find whether trade balance has a symmetric or asymmetric relationship with the trade balance of a country.

3.2 Domain knowledge from previous findings:

1 In Appiah (2017), the authors come to a conclusion that the education expenditure has a positive impact on the GDP Per Capita of a nation.

2 In Blanchflower et al. (2006), the authors say that there is a high relationship between Happiness and HDI, the researchers found that people from Australia are a little more satisfied than the countrys HDI ranking or GNI per capita would predict.

3 In Hong (2014), the author definitely feels that FDI definitely promotes the economic growth in China. However, in China, being a pre-factored economy, the openness in trade and FDI have a less stronger bond than in other economies.

4 In Hall & Helliwell (2014), the authors say that Happiness and HDI are like complementary lenses for analyzing development related aspects on happiness and the human development share, to an extent, a recurrent lineage.

5 In Comim (2016), the author says that HDI is not the most perfect measure of calculating development and this cannot be solved perfectly by increasing the measures but by improving the quality of gathering and analyzing the data, preferably moving back to its capability origins.

6 In Bahmani-Oskooee & Baek (n.d.), the authors conclude by accepting their hypothesis saying that incorporating differentiated responses for appreciation against depreciation demystifies a higher effect of the exchange rate on trade between Korea and the U.S

4 Data Model

4.1 Dimensions:

4.1.1 Dimension Country

The foremost dimension of this data model is a spatial one which is bifurcated as CountryID, CountryName and Region. This dimension was created to to use location based data and analyze economic performance in country-wise manner from all the sources considered for the analysis have this spatial dimension. This dimension has drill down up such that the countries have also been segregated in the form of region to have a region-wise understanding wherever necessary.

4.1.2 Dimension Year

The year dimension is a temporal dimension and it helps in getting data from all the data sets to analyze a business requirement by looking at its temporal aspects. It consists of YearId and Year. The year dimension in this data model uses data from all the data sources that have utilized for performing analysis. For this dimension, a drill down or roll up is not possible. However, filtering the data from all the sources for all the twelve measures is possible.

4.1.3 Dimension Gender

The final dimension of this data model is the gender dimension which encompasses GenderId and Gender. This dimension is generated by combining data from OECD.org and Worldbank. For this particular dimension drill down or roll up is not possible. However, you can filter parameters on the basis of gender.

4.2 Fact:

The fact table of our data model consists of various measures like Population, GDP_PerCapita, FDI, HDI, Unemployment_Rate, HALE(healthy life expectancy), Freedom_to_make_Life_Choices, WHR_Score, Education_in_Percent, Trade_Balance_in_Billion_USD and Euro_to_Pound. Apart from these, we have FactId as the primary key, and YearId, CountryID, GenderId as foreign keys. All of these together will help our fact table form a relationship with the dimension tables of the data model.

Population, GDP_PerCapita, Education_in_Percent and Unemployment_Rate are utilized in 1(Requirement 1). Here, Population and Education Rate can be further drilled down or rolled up using the Country dimension and therefore get a to look at the effect of Education on the Unemployment Rate and thus on the GDP. This can be done in a country-wise or even region-wise manner.

The WHR_Score, HDI, are taken into consideration for analyzing the relationship between these two indices1(Requirement 2). A detailed understanding and inter-dependency between these two on year and country levels is the fact being analyzed.

The measures FDI, Trade_Balance_in_Billion_USD and Euro_to_Pound are being compared to study the effect of these factors on the GDP_PerCapita of Ireland and the how Brexit has affected these factors from 2014 to 2017 is being analyzed 1(Requirement 3)

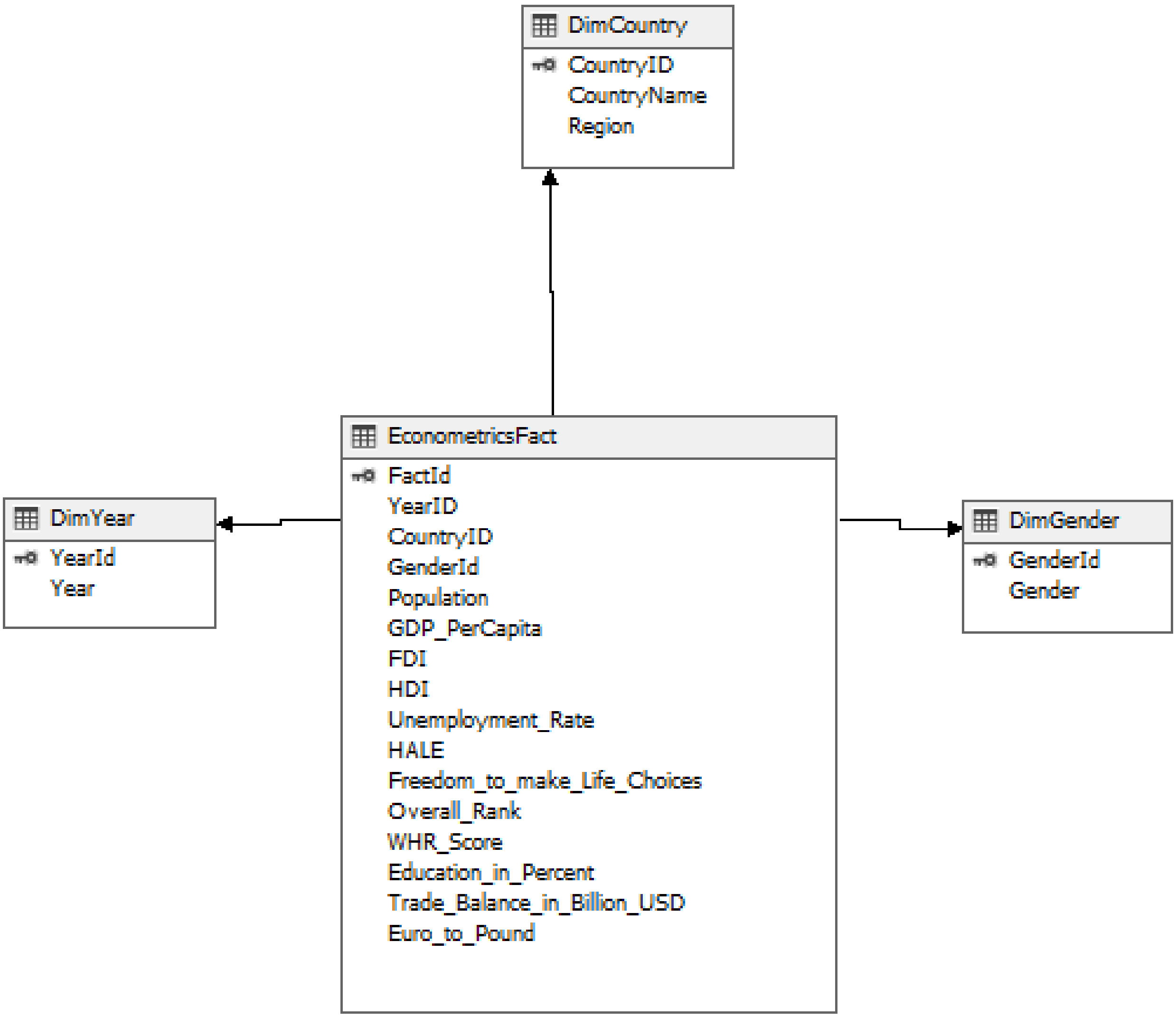

Figure 1: Star Schema

5 Logical Data Map

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	Column	Destination	Column	Type	Transformation
UNDP Human Development Reports	Country	DimCountry	Country	Dimension	The dataset was remodelled by applying transformation on the Country column such that only 28 countries from the EU get selected. The datatype of the Country column was changed to char and data was melted.
UNDP Human Development Reports	Years from 1991-2017	DimYear	Year	Dimension	The years from 1991 to 2011 were removed and a melt function was applied to normalize the overall dataset
UNDP Human Development Reports	HDI Value under every year	FactTable	HDI	Fact	The HDI value that was under the year columns was then added to a fresh new column named HDI Value and named as HDI in the Fact table and data was transformed using melt
UNDP Human Development Reports	Column not present originally	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension
Eurostat	geo/time	DimCountry	Country	Dimension	The dataset was remodelled by applying transformation on the Country column such that only 28 countries from the EU get selected. The data type of the Country column was changed to char. Data was transformed using melt

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Eurostat	Years from 2006-2017	DimMYear	Year	Dimension	The years from 2006 to 2011 were removed and a melt function was applied to normalize the overall dataset and data was transformed using melt.
Eurostat	GDP Value under every year	FactTable	GDP _PerCapita	Fact	The GDP value that was under the year columns was then added to a fresh new column named GDP Per capita PPs and named as GDP_PerCapita in the Fact table
Eurostat	Column not present originally	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension and data was transformed using melt
Statista	Column name absent for the year column	DimYear	Year	Dimension	The dataset had years from 2007 to 2017 and post clearning the years 2011 to 2017 were retained and others were deleted.
Statista	Trade balance in billion U.S. dollars	FactTable	Trade _Balance _in _Billion _USD	Dimension	Trade balance for years before 2014 and after 2017 was removed from the dataset.
Statista	Column not present	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension
Statista	Column name not present	DimYear	Year	Dimension	The exchange rate was considered only for the years 204-2018

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Statista	Exchange rate	FactTable	Euro to Pound	Fact	Data was available in a monthly fashion which was then converted to a yearly fashion by considering the values of December for all the years. The name in the fact table was later changed in the SQL stage by adding an underscore between two words
Statista	Column not present	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension
Worldbank	Country Name	DimCountry	Country	Dimension	Only the 26 countries required for the analysis were extracted from three datsets consisting of total, male and female population. Data was transformed using melt.
Worldbank	Years from 1960-2017	DimYear	Year	Dimension	Years between 2011-2017 were considered and extracted from three datasets consisting of total, male and female population. Data was transformed using the melt function.
Worldbank	No column name	FactTable	Total Population	Fact	Total population for 28 countries was considered from the years 2011-2017. This gets merged with the gender dimension and goes under a single fact named Population. Data was transformed using the melt function.
Worldbank	No column name	FactTable	Population	Fact	Male population for 28 countries was considered from the years 2011-2017. This gets merged with the gender dimension and goes under a single fact named Population
Worldbank	No column name	FactTable	Population	Fact	Female population for 28 countries was considered from the years 2011-2017. This gets merged with the gender dimension and goes under a single fact named Population

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Worldbank	No column name	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension. Data was transformed using the melt function.
Worldbank	No column name	DimGender	Gender	Dimension	Gender-wise data was extracted from Population male and Population female datasets from Worldbank
International Monetary Fund	Country	DimCountry	Country	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function.
International Monetary Fund	Years from 2011-2016	DimYear	Year	Dimension	Data for 28 countries was extracted from this source and added to a column year. Data was transformed using the melt function.
International Monetary Fund	Values under the year columns	FactTable	Unemployment _Rate	Fact	Unemployment rate data for 28 countries was extracted from this source. Data was transformed using the melt function.
International Monetary Fund	Column not present	DimCountry	Region	Dimension	Region column was created by looking at the geographical location of the specific country. This facilitated a drill down and look up for the Country Dimension. Data was transformed using the melt function.
GitHub	Country	DimCountry	Country	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
GitHub	Year	DimYear	Year	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018
GitHub	Country	DimCountry	Country	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018
GitHub	Happiness Score	FactTable	WHR Score	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018
GitHub	Health (Life Expectancy)	FactTable	HALE	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018
GitHub	Freedom	FactTable	Freedom to make life choices	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
GitHub	Happiness Rank	FactTable	Overall Rank	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function for the years 2015 and 2016 from 2 separate datasets and merged with Unstructured data from Wikipedia for the years 2017 and 2018
Wikipedia	Overall Rank	FactTable	Overall Rank	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
Wikipedia	Country	DimCountry	Country	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
Wikipedia	Year	DimYear	Year	Dimension	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
Wikipedia	Health (Life Expectancy)	FactTable	HALE	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
Wikipedia	Freedom	FactTable	Freedom to make life choices	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
Wikipedia	Happiness Rank	FactTable	Happiness Rank	Fact	Data for 28 countries was extracted from this source. Data was transformed using the melt function and then merged with data from GitHub
OECD.org	Country	DimCountry	Country	Dimension	Data for 28 countries was extracted from this source
OECD.org	SEX	DimGender	Gender	Dimension	Data for 28 countries was extracted from this source
OECD.org	Year	DimYear	Year	Dimension	Data for 28 countries was extracted from this source
OECD.org	Value	FactTable	Education _in _Percent	Fact	Data for 28 countries was extracted from this source

6 ETL Process

6.1 Extraction:

This project utilizes data from eight different sources to construct a data model on Econometrics.

6.1.1 UNDP Human Development Reports

Human Development Index: Data on the Human Development Index from 2011 - 2017 for the countries present in the European Union is extracted from hdr.undp.org.
Usage: The dataset was downloaded in a .csv format and comprises of year and country dimensions from my data model. This dataset addresses the business requirements listed in Section 1(Requirement 2)
Foreign Direct Investment: Data on the Foreign Direct Invest inflows as a percentage of GDP from 2011-2017 for the countries present in the EU is extracted from hdr.undp.org.
Usage: The dataset was downloaded in a .csv format and it comprises of year and country dimensions from my data model. It addresses business requirements listed in Section 1(Requirement)

6.1.2 Eurostat

GDP Per Capita,PPS: Data on GDP per Capita as per the purchasing power standard of a country for the years 2011 - 2017 for the 28 countries in the European Union is extracted from Eurostat.
Usage: The dataset was downloaded in a .csv format and it comprises of year and country dimensions from my data model. It addresses the business requirements listed in Section1(Requirement 1)

6.1.3 Statista

Ireland: Trade balance from 2007 to 2017 (in billion U.S. dollars):
Usage: This dataset was downloaded in a .xls format and it comprises of year and country dimensions from my data model. It addresses the business requirements listed in Section1(Requirement 3)

Euro (EUR) to British pound (GBP) monthly exchange rate from December 2014 to February 2019: This data displays the euro to british pound monthly exchange rate as of the last day of the month of December.
Usage: Data was downloaded in a .xls format and it comprises of month-wise data which was then converted to year to match the year dimension by considering the exchange rate changes from December to December basis as this data was taken only on the last day of the month. Section1(Requirement 3)

6.1.4 Worldbank

Population(Total): Three datasets were extracted from Worldbank on male, female and total population respectively. These datasets give information on the population in numbers of different countries in the world in a year-wise and gender-wise formats
Usage: Data was downloaded in a .csv format and it was loaded on to R for cleaning purposes. Section1(Requirement 1)

6.1.5 International Monetary Fund

The International Monetary Fund is one of the leading data collection organization and for this project, the rate of unemployment is a parameter which is very well explained throig this extraction.
Usage: Data was downloaded in a .csv format and it was loaded on to R for cleaning purposes. Section1(Requirement 1)

6.1.6 GitHub

Data for the World Happineess Report for the years 2015 and 2016 was downloaded from this source. **Usage**: Data from this resource was downloaded as a .csv file and was loaded onto R for cleaning and staging purposes.Section1(Requirement 2)

6.1.7 Wikipedia

Data for two tables is extracted from Wikipedia. Wikipedia is one of well accessible open source unstructured data source. **Usage**: Unstructured data was loaded into R for cleaning purposes. The data is arranged in terms of year and country dimensions. Section1(Requirement 2)

6.1.8 OECD.org

Data on the education of a particular country based on gender, year and country bifurcations was downloaded from OECD.org. **Usage**: Data was loaded into R for cleaning purposes which will thus arrange the data in a gender-wise format to get an insight of the overall educational growth. 1(Requirement 1)

6.2 Cleaning:

Data was cleaned using R by loading the data either as a .csv file in R or as a table in terms of Wikipedia using the 'htmltab' package in R. For purposes of data cleaning, the most utilized package is 'dplyr' in R followed by the 'readxl' package for reading an excel file. The last and the second most utilized package from R is 'reshape2'. Almost, all the data downloaded from the structured or unstructured datasets was melted to be brought to the best normalized form using the melt function in R from the 'reshape2' package.

Automated Extraction, transformation and loading of data is done by executing Rscript files in to the SQL Service Integration Services by using the Execute process task option.

6.3 Truncating and Staging the data:

In order to eliminate the risk of getting duplicated entries into the database, tables were truncated by executing an SQL query. Data is automatically cleaned from the tables by executing SQL in SSIS. This is done by utilizing the Execute SQL task. All relationships between the data are dropped before clearing the entire database. The reason for this is that truncation cannot be performed if a relationship exists in the database, for example a primary key - foreign key relationship.

Datasets that are in the .csv formats are loaded in to the database by taking aid of a flat file source and OLE DB destination components from SSIS.

6.4 Loading:

First Dimensions are created followed by facts into the database is performed with the help of an OLE DB destination component. This helps in establishing and strengthening the relationship between the dimension and fact by performing an SQL task.

6.5 OLAP Cube:

To create an OLAP cube, we first create a connection with the SQL Database Server to fetch Dimension tables followed by the Fact table in SSAS. A connection is then established and initiated, the populated dimensions and facts from the database are loaded into SSAS. The deployment starts with dimensions followed by the facts. The automation to redeploy the OLAP cube is done using Analysis Services Processing Task in SSIS. ETL process is automated in SSIS starting from extracting the data to deploying the cube.

7 Application

Looking at the undermentioned BI queries, we can describe our results and demonstrate the output as visualizations using Tableau.

7.1 BI Query 1: What is the effect of a change in the Population and percentage of education on the rate of unemployment, and thus on the GDP Per Capita for a set of European nations?

For this query, the contributing sources of data are: OECD, Worldbank, Eurostat, and IMF.

The general findings are that as there's an increase in the population of the countries as illustrated in Query 1 **??**, more people are getting educated and thus the rate of unemployment is going down. So, all of these three values are very much interrelated.

However, in terms of GDP Per Capita PPS, the values do not necessarily increase or decrease by a significant margin for let's say France in terms of Unemployment versus GDP. Therefore, the relationship of all these 3 factors over the GDP Per Capita is strong but not extremely causal. This query was analyzed by taking into consideration 9 European countries from the years 2014-2016.

QUERY 1 backup.PNG QUERY 1 backup.PNG

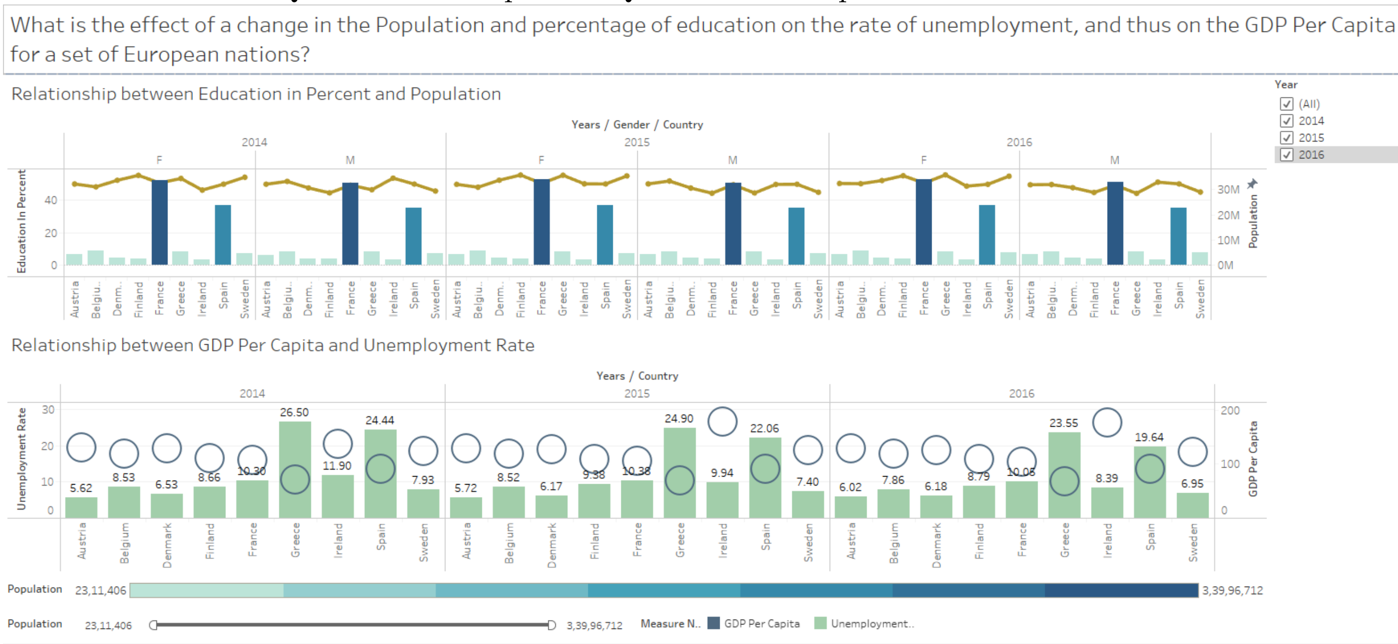

Figure 2: Results for BI Query 1

7.2 BI Query 2: Is there a significant dependency between the Human Development Index and the World Happiness Report score?

For this query, the contributing sources of data are: Wikipedia, UNDP HDR, and GitHub.

The general findings are that there is a lot of correlation between the as illustrated in the Human Development Index and the World Happiness Report. However, this relationship is not a perfectly causal one but related. We can refer to our research papers taken into consideration in the previous section: Blanchflower et al. (2006). Figure 3.

7.3 BI Query 3: How impacting is the inflow of Foreign Direct Investment on a positive trade balance and thus on the Exchange Rate of Ireland from 2014-2017?

For this query, the contributing sources of data are: Statista for Euro to Pound and Trade Balance Data and UNDP HDR for the foreign direct investment values.The increase was consistent in the Trade balance as well. For explanation purposes, trade balance is value calculated by imports subtracted from exports. A positive value indicates better economic growth.

QUERY 2.PNG QUERY 2.PNG

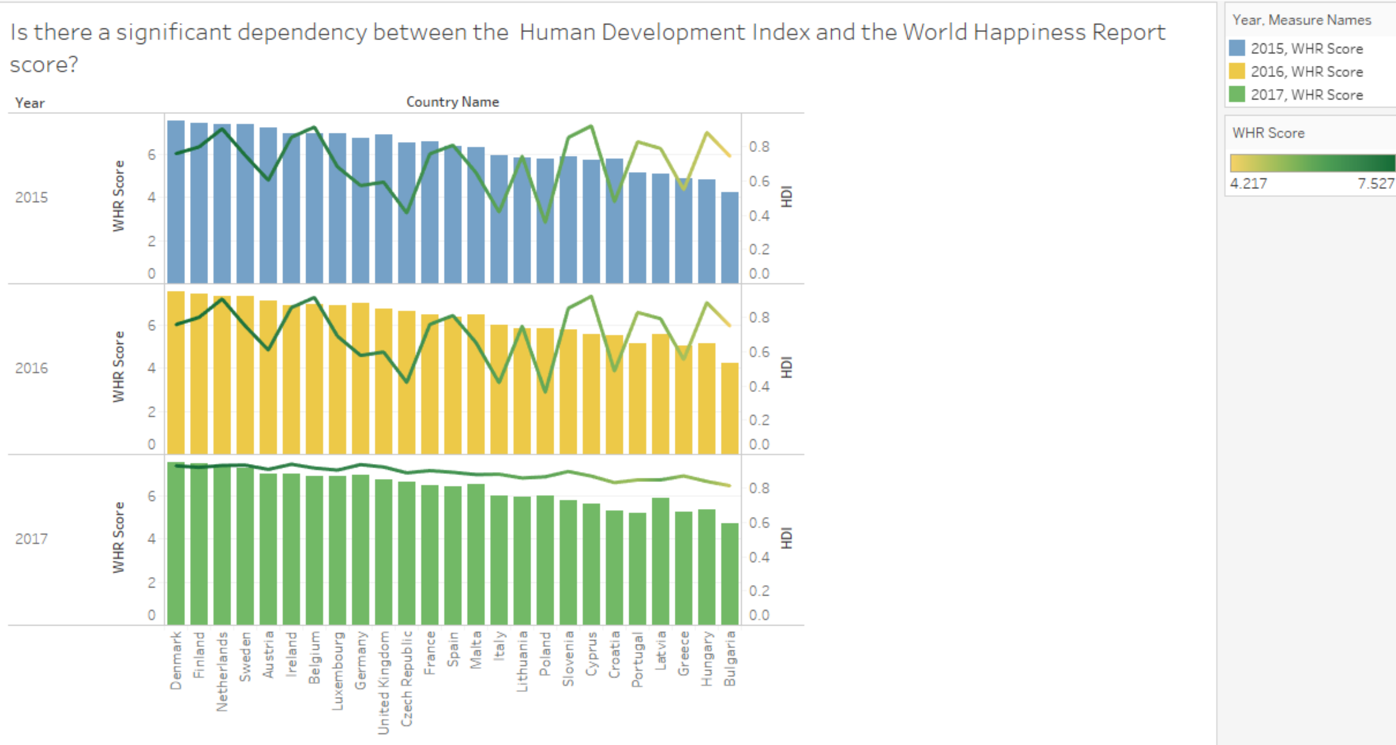

Figure 3: Results for BI Query 2

However, the Euro to Pound rate did fall from 0.78 to 0.73. The surge in the foreign direct investment in Ireland was due to the announcement of Brexit in the year 2014 which resulted a large number of foreign investment in the Irish soil.

The general findings are that the foreign direct investment did increase in the year 2015 to 81.00 from 33.60(2014) as illustrated in Figure 4. Therefore, we can conclude that an increase in trade balance is significant in the growth in the power of Euro against Pound. However, it's not significantly related to the FDI inflows(as percentage of GDP). But the growth in FDI does directly relate to a growth in GDP. The growth in GDP is directly proportional to Trade Balance. Therefore, FDI is indirectly related to trade balance.

7.4 Discussion

According to BI query 1, population, education and rate of unemployment are strongly related and they definitely contribute the growth in the GDP of a nation individually and as a set of three. This can also be seen in Appiah (2017) where the author mentions how the expenditure on education affects the growth in the GDP per capita.

According to BI query 2, there is a strong correlation between HDI and WHR Score as there a lot of parameters that are common in the calculation of these two factors. However, both the factors have some differences between the 2 as per the article by Blanchflower et al. (2006) where the authors talk about Australia being a paradox. We can also refer to other research references like the paper written by Comim (2016) wherein their hypothesis of finding factors beyond the Human Development Index for development does make it a point that a report like the World Happiness report is a good parallel comparative study to understand economic and social development. The paper written

QUERY 3.PNG QUERY 3.PNG

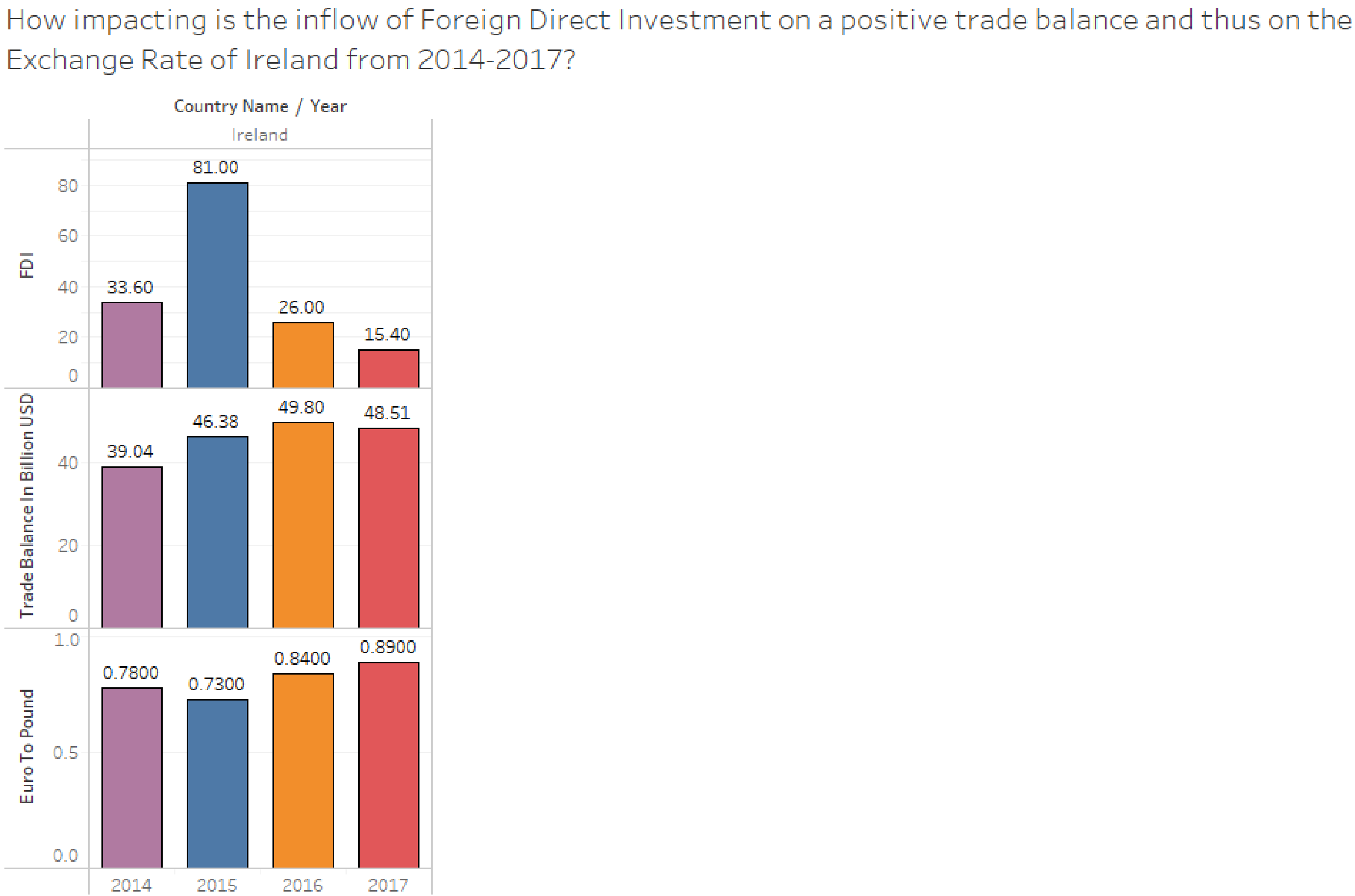

Figure 4: Results for BI Query 3

by Hall & Helliwell (2014) on 'Happiness and Human Development', whereim the authors conclude by saying that Happiness and HDI are complementary to each other for analyzing development and growth.

According to BI query 3, we can conclude that trade balance is directly related to growth in the exchange rate and indirectly related to the changes in the Foreign Direct Investment.

This can be understood from the paper Hong (2014).

8 Conclusion and Future Work

The conclusion of this study is that the dependency of various of economic factors considered in econometric analysis using statistical methods was successfully analyzed and understood using a data warehouse model.

Correlations between Population, Education, Unemployment and GDP Per Capita PPS were understood. Similarly other economic indices were also compared for research purposes and it could be found that there's a strong relationship between the Human Development Index and World Happiness Report. Also, an Ireland focused business query helped us understand the effect of Brexit on a surge in the Foreign direct investment in the year following the Brexit announcement. We also established an understanding of the relationship between Foreign Direct Investment, Euro to Pound Exchange Rate and

Trade Balance.

As a future scope of this project more number of countries and more factors could be taken into calculation for getting even better insights. More number of years should also be taken into considerations and similar data model could be created for analysis. A better Brexit analysis could be added by taking a sentiment analysis on the same whenever Brexit actually gets executed and the profits and losses to the EU and UK could be understood. We could also get an understanding of the effects post Brexit on the Irish Trade Balance by building a data warehouse.

References

Appiah, E. N. (2017), 'The Effect of Education Expenditure on Per Capita GDP in Developing Countries', *International Journal of Economics and Finance* **9**(10), 136.

Bahmani-Oskooee, M. & Baek, J. (n.d.).

Blanchflower, D., Carroll, N., Easterlin, R., Frijters, P., Norton, A. & Oswald, A. (2006), 'Happiness and the Human Development Index :', *The Australian Economic Review* **39**(2), 176–184.

Comim, F. (2016), 'Beyond the HDI? Assessing Alternative Measures of Human Development from a Capability Perspective', *Human Development Report* .
URL: *http://hdr.undp.org/sites/default/files/comim_template_0.pdf*

Hall, J. & Helliwell, J. F. (2014), 'Happiness and Human Development', *UNDP Human Development Report Office* **5**(1), 67–72.
URL: *http://hdr.undp.org/sites/default/files/happiness_and_hd.pdf*

Hong, L. (2014), 'Does and How does FDI Promote the Economic Growth? Evidence from Dynamic Panel Data of Prefecture City in China', *IERI Procedia* **6**, 57–62.
URL: *http://dx.doi.org/10.1016/j.ieri.2014.03.010*

Appendix

R code

```
#Population Total
install.packages("reshape2")
install.packages("dplyr")
install.packages("readxl")
install.packageso("htmltab")

#Loading the required libraries
library(dplyr)
library(readxl)
library(htmltab)
library(reshape2)
```

```
#Reading a CSV file
Popf <- read.csv("C:\\Users\\Abhi\\Desktop\\Project␣files\\Structured␣datase

#Deleting a particular set of rows and columns
Popf <- Popf [c(3, 16, 19, 23, 101, 55, 56, 60, 73, 77, 79, 57, 91, 103, 113
Popf <- Popf [-c(2, 4, 5:55)]
Popf <- Popf [-c(2)]
Popf <- Popf [-c(9,10)]

#Renaming Columns
names(Popf)[1] <- "Country"
names(Popf)[2] <- "2011"
names(Popf)[3] <- "2012"
names(Popf)[4] <- "2013"
names(Popf)[5] <- "2014"
names(Popf)[6] <- "2015"
names(Popf)[7] <- "2016"
names(Popf)[8] <- "2017"
Popf <- Popf[-c(1),]

#Using the melt function
Popf<-melt(data=Popf, id.vars= "Country" , measure.vars = c("2011","2012","2

#Renaming Columns
names(Popf)[2] ="Year"
names(Popf)[3] ="Population"

str(Popf)

Popf$Country <- as.character(Popf$Country) #Changing the type of Country col

Popf$Year <- as.character(Popf$Year)
Popf$Year <- as.numeric(Popf$Year)

#Adding a Region column

Popf$Region <- rep("␣", nrow(Popf))

#Adding a Gender column

Popf$Gender <- rep("F", nrow(Popf))
Df1 <- Popf
Df1[Df1$Country == "Austria", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Denmark", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Finland", "Region"] <- "Northern␣Europe"
```

```
Df1[Df1$Country == "France", "Region"] <- "Western␣Europe"
Df1[Df1$Country == "Germany", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Greece", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Hungary", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Italy", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Malta", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df1[Df1$Country == "Poland", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Romania", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df1[Df1$Country == "Spain", "Region"] <- "Southern␣Europe"
Df1[Df1$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df1[Df1$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

Popf <- Df1
#Checking data types
str(Popf)

#Reading a CSV file
Popm <- read.csv("C:\\Users\\Abhi\\Desktop\\Project␣files\\Structured␣datase

#Keep these rows
Popm <- Popm [c(3, 16, 19, 23, 101, 55, 56, 60, 73, 77, 79, 57, 91, 103, 113

#Deleting unwanted rows
Popm <- Popm [-c(2, 4, 5:55)]
Popm <- Popm [-c(2)]
Popm <- Popm [-c(9,10)]

#Renaming Columns

names(Popm)[1] <- "Country"
names(Popm)[2] <- "2011"
names(Popm)[3] <- "2012"
names(Popm)[4] <- "2013"
names(Popm)[5] <- "2014"
names(Popm)[6] <- "2015"
names(Popm)[7] <- "2016"
names(Popm)[8] <- "2017"
Popm <- Popm[-c(1),]

#Using the Melt Function
Popm<-melt(data=Popm, id.vars= "Country" , measure.vars = c("2011","2012","2
names(Popm)[2] ="Year"
```

```
names(Popm)[3] ="Population"

Popm$Country <- as.character(Popm$Country) #Changinng the type of Country co

Popm$Year <- as.character(Popm$Year)
Popm$Year <- as.numeric(Popm$Year)

#Adding a Region column

Popm$Region <- rep("␣", nrow(Popm))

#Adding a Gender column

Popm$Gender <- rep("M", nrow(Popm))

#Adding a region column
Df2 <- Popm
Df2[Df2$Country == "Austria", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Denmark", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Finland", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "France", "Region"] <- "Western␣Europe"
Df2[Df2$Country == "Germany", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Greece", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Hungary", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Italy", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Malta", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df2[Df2$Country == "Poland", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Romania", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df2[Df2$Country == "Spain", "Region"] <- "Southern␣Europe"
Df2[Df2$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df2[Df2$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

#Assigning Df2 to the object Popm
Popm <- Df2

#Checking data types
str(Popm)
```

```
P1 <- Popf
P2 <- Popm

#Using the reduce and merge functions to add the male and female populations
PopT = Reduce(function(x, y) merge(x, y, all=TRUE), list(P1, P2))

View(PopT)
#Grouping the data by country and year
Population_Total <- PopT %>% group_by(Country,Year) %>% summarise(total = su

PopulationX <- Reduce(function(x,y) merge(x,y, all = T), list(Population_Tot

#renaming a column
names(PopulationX)[3] <- "Total␣Population"

PopulationX<- PopulationX[!duplicated(PopulationX),]

#Writing the .csv file
write.csv(PopulationX, "PopulationX.csv",row.names=FALSE)#, na="", col.names

#### Education


#getwd()

setwd("C:/Users/Abhi/Desktop/Project␣files/Econometrics␣and␣the␣effects␣of␣B

#Reading the csv file from the local directory
Education <- read.csv(file = "C:/Users/Abhi/Desktop/Project␣files/Structured
Education <- Education[-c(1), -c(1, 4, 5, 6,8, 9, 10, 11, 13, 14, 15, 16, 17

#Removing blank values
Education <- Education[!(Education$V19 == ""),]

#Checking for missing values for verification purposes
#Education1 <- filter(Education, V20 == "x")
#Education2 <- filter(Education, (V20 == 'z'))
#Education3 <- filter(Education, (V20 == 'w'))

#Renaming column names
names(Education)[1] <- "Country"
names(Education)[4] <- "Year"
names(Education)[3] <- "Level"
names(Education)[5] <- "Education_in_Percent"
names(Education)[2] <- "Gender"

#Checking the data types
str(Education)

#Changing the data types
```

```
Education$Year <- as.character(Education$Year)
Education$Year <- as.numeric(Education$Year)

Education$Gender <- as.character(Education$Gender)
Education$Level <- as.character(Education$Level)

Education$Country <- as.character(Education$Country)

Education$Year <- as.character(Education$Year)
Education$Year <- as.numeric(Education$Year)

Education$Education_in_Percent <- as.character(Education$Education_in_Percen
Education$Education_in_Percent <- as.numeric(Education$Education_in_Percent)

#Checking if Education is a  data frame
is.data.frame(Education)
str(Education)

Education<- Education[!duplicated(Education),]


#Writing the csv to your local directory
write.csv(Education, file = "Education.csv", row.names=FALSE) #na="", col.na


#Reading an excel file
GDP_cleaned <- read.csv("C:\\Users\\Abhi\\Desktop\\Project␣files\\Structured

GDP_cleaned <- GDP_cleaned[-c(1,2,4, 5, 6,7, 36:56),] #DeleteRows
GDP_cleaned <- GDP_cleaned[-c(2:11,13,15,17,19,21,23,25)] #DeleteColumns

#Renaming columns
names(GDP_cleaned)[1] <- "Country"
names(GDP_cleaned)[2] <- "2011"
names(GDP_cleaned)[3] <- "2012"
names(GDP_cleaned)[4] <- "2013"
names(GDP_cleaned)[5] <- "2014"
names(GDP_cleaned)[6] <- "2015"
names(GDP_cleaned)[7] <- "2016"
names(GDP_cleaned)[8] <- "2017"

#Removing white space from Country column:
GDP_cleaned$Country <- gsub("[[:space:]]", "", GDP_cleaned$Country)


#names(WHR_New)[3] <- "WHR Score"
GDP_cleaned[29,1] <- "United␣Kingdom"
GDP_cleaned[4, 1] <- "Czech␣Republic"
GDP_cleaned <- GDP_cleaned[-c(1),]

names(GDP_cleaned)[1] <- "Country"
```

```
names(GDP_cleaned)[2] <- "2011"
names(GDP_cleaned)[3] <- "2012"
names(GDP_cleaned)[4] <- "2013"
names(GDP_cleaned)[5] <- "2014"
names(GDP_cleaned)[6] <- "2015"
names(GDP_cleaned)[7] <- "2016"
names(GDP_cleaned)[8] <- "2017"


#Converting Chr to numeric
str(GDP_cleaned)
GDP_cleaned$'2011' <- as.character(GDP_cleaned$'2011')
GDP_cleaned$`2011` <-  as.numeric(GDP_cleaned$`2011`)

GDP_cleaned$'2012' <- as.character(GDP_cleaned$'2012')
GDP_cleaned$`2012` <-  as.numeric(GDP_cleaned$`2012`)

GDP_cleaned$'2013' <- as.character(GDP_cleaned$'2013')
GDP_cleaned$`2013` <-  as.numeric(GDP_cleaned$`2013`)

GDP_cleaned$'2014' <- as.character(GDP_cleaned$'2014')
GDP_cleaned$`2014` <-  as.numeric(GDP_cleaned$`2014`)

GDP_cleaned$'2015' <- as.character(GDP_cleaned$'2015')
GDP_cleaned$`2015` <-  as.numeric(GDP_cleaned$`2015`)

GDP_cleaned$'2016' <- as.character(GDP_cleaned$'2016')
GDP_cleaned$`2016` <-  as.numeric(GDP_cleaned$`2016`)

GDP_cleaned$'2017' <- as.character(GDP_cleaned$'2017')
GDP_cleaned$`2017` <-  as.numeric(GDP_cleaned$`2017`)

GDP_cleaned<-melt(data=GDP_cleaned, id.vars= "Country" , measure.vars = c("2

names(GDP_cleaned)[2]= "Year"
names(GDP_cleaned)[3]= "GDP␣Per␣capita␣PPS"


#Adding a Region columns

GDP_cleaned$Region <- rep("␣", nrow(GDP_cleaned))


#Assigning values to the Region columns
Df3 <- GDP_cleaned
Df3[Df3$Country == "Austria", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Denmark", "Region"] <- "Northern␣Europe"
```

```
Df3[Df3$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "Finland", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "France", "Region"] <- "Western␣Europe"
Df3[Df3$Country == "Germany", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Greece", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Hungary", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "Italy", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Malta", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df3[Df3$Country == "Poland", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Romania", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df3[Df3$Country == "Spain", "Region"] <- "Southern␣Europe"
Df3[Df3$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df3[Df3$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

#Changing the data to the appropriate data type
GDP_cleaned <- Df3
GDP_cleaned$'Year' <- as.character(GDP_cleaned$'Year')
str(GDP_cleaned)
GDP_cleaned$'Year' <- as.numeric(GDP_cleaned$'Year')

GDP_cleaned<- GDP_cleaned[!duplicated(GDP_cleaned),]

#Writing the .csv file to the local directory
write.table(GDP_cleaned, file = "GDP_cleaned.csv",row.names=FALSE, na="", co

#### Unemployment

#Reading CSV
Unemployment <- read.csv("C:/Users/Abhi/Desktop/Project␣files/Structured␣dat

Unemployment <- Unemployment[-c(1, 30, 31) ,-c(2, 3, 11)]

colnames(Unemployment) <- c("Country", "2011", "2012", "2013", "2014", "2015

#Setting Country as Year
Unemployment$Country <- as.character(Unemployment$Country)

#Renaming a cell for consistency purposes
Unemployment[24, 1] <- "Slovakia"

Unemployment <-melt(data=Unemployment, id.vars= "Country" , measure.vars = c
```

```
#Renaming Columns
names(Unemployment)[2] <- "Year"
names(Unemployment)[3] <- "Unemployment_Rate"

#Setting the right data types
str(Unemployment)
Unemployment$Year <- as.character(Unemployment$Year)
Unemployment$Year <- as.numeric(Unemployment$Year)

#Adding a region column
Unemployment$Region <- rep("␣", nrow(Unemployment))

Df4 <- Unemployment
#Assigning values to the Region as per the geographic location
str(Df4)
Df4[Df4$Country == "Austria", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Denmark", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Finland", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "France", "Region"] <- "Western␣Europe"
Df4[Df4$Country == "Germany", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Greece", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Hungary", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Italy", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Malta", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df4[Df4$Country == "Poland", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Romania", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df4[Df4$Country == "Spain", "Region"] <- "Southern␣Europe"
Df4[Df4$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df4[Df4$Country == "Bulgaria", "Region"] <- "Southern␣Europe"
Unemployment <- Df4

#Removing duplicate rows
Unemployment<- Unemployment[!duplicated(Unemployment),]

#Writing the csv on the working directory

write.table(Unemployment, file = "Unemployment_Rate.csv",row.names=FALSE, na
```

```
#### HDI

#Loading the necessary libraries

#Reading a CSV file
HDI <- read.csv("C:\\Users\\Abhi\\Desktop\\Project␣files\\Structured␣dataset
HDI <- HDI[c(2, 22, 19, 53, 48, 34, 29, 13, 32, 17, 26, 7, 33, 47, 6, 30, 43
HDI <- HDI[-c (3:44, 46, 48, 50, 52, 54, 56)]

#To check and set the required working directory
getwd()
setwd("C:\\Users\\Abhi\\Desktop\\Project␣files\\Econometrics␣and␣the␣effects

#write.table(HDI, file = "HDI_cleaned.csv",row.names=FALSE, na="", col.names
#data <- read.csv("C:\\Users\\Abhi\\Desktop\\Project files\\Econometrics and
#data <- data[-c(1), ]

#write.table(data, file = "HDI_01.csv",row.names=FALSE, na="")

names(HDI)[1] <- "HDI␣RANK"
names(HDI)[2] <- "Country"
names(HDI)[3] <- "2011"
names(HDI)[4] <- "2012"
names(HDI)[5] <- "2013"
names(HDI)[6] <- "2014"
names(HDI)[7] <- "2015"
names(HDI)[8] <- "2016"
names(HDI)[9] <- "2017"
HDI <- HDI[-c(1),]

str(HDI)

HDI$Country <- as.character(HDI$Country)

HDI[6, 2] <- "Czech␣Republic"

#Using the melt function
HDI<-melt(data=HDI, id.vars= "Country" , measure.vars = c("2011","2012","201

#Renaming Columns
names(HDI)[2] <- "Year"
names(HDI)[3] <- "HDI␣Value"
HDI$Country <- as.character(HDI$Country)

#Assigning the right data types
HDI$Year <- as.character(HDI$Year)
HDI$Year <- as.numeric(HDI$Year)
str(HDI)

#Ading a Region Column
```

```r
HDI$Region <- rep("␣", nrow(HDI))

#Adding values to the Region column
Df <- HDI

Df[Df$Country == "Austria", "Region"] <- "Central␣Europe"
Df[Df$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df[Df$Country == "Denmark", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Finland", "Region"] <- "Northern␣Europe"
Df[Df$Country == "France", "Region"] <- "Western␣Europe"
Df[Df$Country == "Germany", "Region"] <- "Central␣Europe"
Df[Df$Country == "Greece", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Hungary", "Region"] <- "Central␣Europe"
Df[Df$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Italy", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df[Df$Country == "Malta", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df[Df$Country == "Poland", "Region"] <- "Central␣Europe"
Df[Df$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Romania", "Region"] <- "Central␣Europe"
Df[Df$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df[Df$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df[Df$Country == "Spain", "Region"] <- "Southern␣Europe"
Df[Df$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df[Df$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df[Df$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

str(Df)
HDI_final <- Df


#Writing the csv on the working directory

write.table(HDI_final, file = "HDI_cleaned.csv",row.names=FALSE, na="", col.

####

FDI <- read.csv("C:\\Users\\Abhi\\Desktop\\Project␣files\\Structured␣dataset

FDI <- FDI [-c(1, 3:10, 12:17, 19:26, 28:42, 46, 48:55, 57:59, 62:64, 66, 68
FDI <- FDI [-c(3:12, 14, 16, 18, 20, 22, 24, 26)] #DeleteColumns
FDI <- FDI [-c(1)] #DeleteColumns

names(FDI)[1] <- "Country"
```

```r
names(FDI)[2] <- "2011"
names(FDI)[3] <- "2012"
names(FDI)[4] <- "2013"
names(FDI)[5] <- "2014"
names(FDI)[6] <- "2015"
names(FDI)[7] <- "2016"
names(FDI)[8] <- "2017"
FDI <- FDI[-c(1),]


#Removing white space from Country column:
FDI$Country <- gsub("[[:space:]]", "", FDI$Country)
str(FDI)

FDI$Country <- as.character(FDI$Country)
FDI[28,1] <- "United␣Kingdom"
FDI[6, 1] <- "Czech␣Republic"

#Using the melt function
FDI <- melt(data=FDI, id.vars= "Country" , measure.vars = c("2011","2012","2

#Renaming Columns
names(FDI)[2] ="Year"
names(FDI)[3] ="FDI"

str(FDI)

FDI$Country <- as.character(FDI$Country) #Changing the type of Country colum

FDI$Year <- as.character(FDI$Year)
FDI$Year <- as.numeric(FDI$Year)

#names(WHR_New)[3] <- "WHR Score"

#GDP_cleaned <- GDP_cleaned[-c(1),]


#Adding a new column
FDI$Region <- rep("␣", nrow(FDI))

Df5 <- FDI
#Adding appropriate values to the Region column
Df5[Df5$Country == "Austria", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Belgium", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Croatia", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Cyprus", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Denmark", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Estonia", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Finland", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "France", "Region"] <- "Western␣Europe"
```

```
Df5[Df5$Country == "Germany", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Greece", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Hungary", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Ireland", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Italy", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Latvia", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Lithuania", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Luxembourg", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Malta", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Netherlands", "Region"] <- "Western␣Europe"
Df5[Df5$Country == "Poland", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Portugal", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Romania", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Slovakia", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Slovenia", "Region"] <- "Central␣Europe"
Df5[Df5$Country == "Spain", "Region"] <- "Southern␣Europe"
Df5[Df5$Country == "Sweden", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
Df5[Df5$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

FDI <- Df5

#Writing the CSV file to your local directory
write.table(FDI, file = "FDI_cleaned.csv",row.names=FALSE, na="", col.names=

####Exchange Rate of Euro to Pound

#Loading the dataset from a .xlsx file

Exchange_Rate <- read_excel("C:\\Users\\Abhi\\Desktop\\Project␣files\\Struct

#Deleting unnecessary rows

Exchange_Rate <- Exchange_Rate[-c(1, 2),]


#Renaming columns
names(Exchange_Rate)[1] <- "Year"
names(Exchange_Rate)[2] <- "Euro␣to␣Pound"

#Keeping the necessary columns
Exchange_Rate <- Exchange_Rate [c(1, 13, 25, 37, 49),]

#Changing December 14 to 2014 and so on
Exchange_Rate$Year <- gsub(".*?'", "20", Exchange_Rate$Year)

#Converting Data to the right data types
#Exchange_Rate$Year <- as.list(Exchange_Rate$Year)
str(Exchange_Rate)
Exchange_Rate$Year <- as.numeric(Exchange_Rate$Year)
Exchange_Rate$`Euro to Pound` <- as.numeric(Exchange_Rate$`Euro to Pound`)
```

```
#Adding a country column
Exchange_Rate$Country <- rep("Ireland", nrow(Exchange_Rate))

#Adding a country column
Exchange_Rate$Country <- rep("Ireland", nrow(Exchange_Rate))


str(Exchange_Rate)
#Writing the csv file on the local disk
write.table(Exchange_Rate, file = "Exchange_Rate.csv",row.names=FALSE, na=""

#### Trade Balance

#library(readxl)
#Loading the dataset from a .xlsx file

Trade_Balance <- read_excel("C:/Users/Abhi/Desktop/Project␣files/Structured␣

#Deleting unnecessary rows

Trade_Balance <-Trade_Balance [-c(1, 2),]


#Renaming columns
names(Trade_Balance)[1] <- "Year"
names(Trade_Balance)[2] <- "Trade_Balance_in_Billion_USD"

#Keeping the necessary columns
Trade_Balance <- Trade_Balance [-c(1:7),]

#Converting Data to the right data types

str(Trade_Balance)
Trade_Balance$Year <- as.numeric(Trade_Balance$Year)
Trade_Balance$Trade_Balance_in_Billion_USD <- as.numeric(Trade_Balance$Trade


#Adding a country column
Trade_Balance$Country <- rep("Ireland", nrow(Trade_Balance))

#Writing the csv file on the local disk
write.table(Trade_Balance, file = "Trade_Balance.csv",row.names=FALSE, na=""

####WHR

#WHR15

#library(dplyr)
```

```
#Reading a csv
WHR_15 <- read.csv("C://Users//Abhi//Desktop//Project␣files//Econometrics␣an
WHR_15 <- WHR_15[c(9, 14, 21, 36, 37, 38, 39, 45, 47, 48, 51, 53, 59, 65, 67

#Creating a new row
WHR_15$Year <- rep("2015", nrow(WHR_15))
#str(W18)
#Changing the row names
names(WHR_15)[1] <- "Country"
names(WHR_15)[2] <- "Overall␣Rank"
names(WHR_15)[3] <- "WHR␣Score"
names(WHR_15)[4] <- "Healthy␣life␣expectancy"
names(WHR_15)[5] <- "Freedom␣to␣make␣life␣choices"
names(WHR_15)[6] <- "Year"

#Assigning the appropriate data types
#WHR_15$'Overall Rank' <- as.double(WHR_15$'Overall Rank')
#WHR_15$'WHR Score' <- as.numeric(WHR_15$'WHR Score')
#WHR_15$'Healthy life expectancy' <- as.double(WHR_15$'Healthy life expectan
#WHR_15$'Healthy life expectancy' <- as.double(WHR_15$'Healthy life expectan
#WHR_15$'Freedom to make life choices' <- as.double(WHR_15$'Freedom to make 

WHR_15 <- WHR_15[,c(6, 1, 2, 3, 4, 5)]

#WHR16

#library(dplyr)
WHR_16 <- read.csv("C://Users//Abhi//Desktop//Project␣files//Econometrics␣an
WHR_16 <- WHR_16[c(9, 14, 22, 36, 37, 38, 39, 44, 46, 47, 50, 52, 58, 64, 66

#Creating a new column
WHR_16$Year <- rep("2016", nrow(WHR_16))

#Assigning names to various columns
names(WHR_16)[1] <- "Country"
names(WHR_16)[2] <- "Overall␣Rank"
names(WHR_16)[3] <- "WHR␣Score"
names(WHR_16)[4] <- "Healthy␣life␣expectancy"
names(WHR_16)[5] <- "Freedom␣to␣make␣life␣choices"
names(WHR_16)[6] <- "Year"
str(WHR_16)
#Changing the data to appropriate data types
#WHR_16$'Overall Rank' <- as.character(as.numeric(WHR_16$'Overall Rank'))
#WHR_16$'WHR Score' <- as.numeric(WHR_16$'WHR Score')
#WHR_16$'Healthy life expectancy' <- as.character(as.numeric(WHR_16$'Healthy
#WHR_16$'Healthy life expectancy' <- as.character(as.numeric(WHR_16$'Healthy
#WHR_16$'Freedom to make life choices' <- as.character(as.numeric(WHR_16$'Fr


WHR_16 <- WHR_16[,c(6, 1, 2, 3, 4, 5)]
```

```
#library(htmltab)
WHR <- as.data.frame(htmltab(doc = "https://en.wikipedia.org/wiki/World_Happ
#library(dplyr)

########Adding WHR_New dataframe to remove white space from Country/Region c

#library('readxl')

WHR_New <- WHR
WHR_New <- WHR [c(12, 16, 100, 82, 61, 21, 3, 63, 1, 23, 16, 79, 69, 14, 47,
is.character(WHR_New$`Country/Region`)

#Change the column names of Country/Region to Country:
names(WHR_New)[2] <- "Country"
str(WHR_New)

#Remove extra character from the Country column:
WHR_New$Country <- trimws(gsub("  ", "␣", WHR_New$Country))

#Removing white space from Country column#
WHR_New$Country <- gsub("[[:space:]]", "", WHR_New$Country)

str(WHR_New)

#WHR_New$'Overall Rank' <- as.numeric(WHR_New$'Overall Rank')
#WHR_New$'Score' <- as.numeric(WHR_New$'Score')
#WHR_New$'GDP per capita' <- as.numeric(WHR_New$'GDP per capita')
#WHR_New$'Social support' <- as.numeric(WHR_New$'Social support')
#WHR_New$'Healthy life expectancy' <- as.numeric(WHR_New$'Healthy life expec
#WHR_New$'Freedom to make life choices' <- as.numeric(WHR_New$'Freedom to ma
#WHR_New$'Generosity' <- as.numeric(WHR_New$'Generosity')
#WHR_New$'Perceptions of corruption' <- as.numeric(WHR_New$'Perceptions of c

names(WHR_New)[3] <- "WHR␣Score"
WHR_New[28,2] <- "United␣Kingdom"
WHR_New[6, 2] <- "Czech␣Republic"

WHR_New[WHR_New$Country == "Austria", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Belgium", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Croatia", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Cyprus", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Czech␣Republic", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Denmark", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Estonia", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Finland", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "France", "Region"] <- "Western␣Europe"
WHR_New[WHR_New$Country == "Germany", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Greece", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Hungary", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Ireland", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Italy", "Region"] <- "Southern␣Europe"
```

```
WHR_New[WHR_New$Country == "Latvia", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Lithuania", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Luxembourg", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Malta", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Netherlands", "Region"] <- "Western␣Europe"
WHR_New[WHR_New$Country == "Poland", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Portugal", "Region"] <- "Southern␣Euarope"
WHR_New[WHR_New$Country == "Romania", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Slovakia", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Slovenia", "Region"] <- "Central␣Europe"
WHR_New[WHR_New$Country == "Spain", "Region"] <- "Southern␣Europe"
WHR_New[WHR_New$Country == "Sweden", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "United␣Kingdom", "Region"] <- "Northern␣Europe"
WHR_New[WHR_New$Country == "Bulgaria", "Region"] <- "Southern␣Europe"

WHR_New$Year <- rep("2018", nrow(WHR_New))

WHR_New <- WHR_New[, c(11, 2, 1, 3, 6, 7)]


#Write the file in csv format:
write.csv(WHR_New, file = "WHR_New.csv", row.names = F)


#WHR final
#WHR_2017


WHR_17 <- as.data.frame(htmltab(doc = "https://en.wikipedia.org/wiki/World_H


#Change the column names of Country/Region to Country:
names(WHR_17)[3] <- "Country"
str(WHR_17)

#Remove extra character from the Country column:
WHR_17$Country <- trimws(gsub("  ", "", WHR_17$Country))


#WHR_17$Country <- trimws(WHR_17$Country, which = c("left"))#

#Removing white space from Country column#
WHR_17$Country <- gsub("[[:space:]]", "", WHR_17$Country)

#Deleting unnecessary rows and columns

WHR_17 <- WHR_17[c(1, 2, 5, 6, 10, 13, 15, 16, 17, 18, 19, 23, 27, 31, 34, 4
WHR_17 <- WHR_17[-c(29),]

#Converting to the right data type
WHR_17$'Overall␣Rank' <- as.numeric(WHR_17$'Overall␣Rank')
```

```
WHR_17$'Score' <- as.numeric(WHR_17$'Score')
#WHR_17$'GDP per capita' <- as.numeric(WHR_17$'GDP per capita')
#WHR_17$'Social support' <- as.numeric(WHR_17$'Social support')
WHR_17$'Healthy␣life␣expectancy' <- as.numeric(WHR_17$'Healthy␣life␣expectan
WHR_17$'Freedom␣to␣make␣life␣choices' <- as.numeric(WHR_17$'Freedom␣to␣make␣
#WHR_17$'Generosity' <- as.numeric(WHR_17$'Generosity')
#WHR_17$'Perceptions of corruption' <- as.numeric(WHR_17$'Perceptions of cor

#Adding a year row
WHR_17$Year <- rep("2017", nrow(WHR_17))
names(WHR_17)[3] <- "WHR␣Score"

#Reordering the columns
WHR_17 <- WHR_17[,c(6, 1, 2, 3, 4, 5)]

#Renaming cells on R
WHR_17[11, 3] <- "United␣Kingdom"
WHR_17[12, 3] <- "Czech␣Republic"

#ASsigning the different objects to W values
W17 <- WHR_17
W17 <- W17[,c(1,3,2,4,5,6)]
W18 <- WHR_New
W16 <- WHR_16
W15 <- WHR_15

#binding rows
W <- merge(W17,W18,all=T)
W <- merge(W,W16,all = T)
W <- merge(W,W15,all = T)


WHR_final <- W
#Writing the cleaned file to a csv
write.table(WHR_final, file = "WHR_final.csv", row.names=FALSE, na="", col.n
```