

## Contents

Research.....	3
World Development.....	4
Correlation .....	4
World Bank.....	5
Interview with Josie Gadsby (Teacher of World Development and Politics).....	6
What does world development cover at a level? .....	6
Which time periods of data are most useful?.....	6
Is it useful to use recent data to predict the future and how is it relevant? .....	6
What correlations would be interesting to you personally? .....	7
Are you aware of the data available on the World Bank website? .....	7
Analysis .....	8
Summary of Interview.....	8
Data Flow Diagram.....	8
Document Specification .....	8
Data from World Bank .....	9
IPSO Chart .....	16
Ways of calculating correlation .....	16
Linear and Nonlinear Correlation .....	17
Useful Indicators .....	18
Requirements.....	19
Design.....	21
Prototype .....	21
Process .....	21
Input data from CSV.....	21
Parsing of data .....	21
User selection.....	23
Calculations to be carried out.....	23
Results plotted on a graph.....	25
Class Diagram.....	26
Pseudocode.....	26
Pearson’s product-moment correlation calculation.....	26
Bubble sort algorithm .....	26
Quick sort algorithm .....	27
Storage.....	27

HCI (Human Computer Interface) .....	28
Input.....	31
Output.....	31
Graphing.....	32
Technical Solution.....	34
Testing.....	45
Evaluation .....	48
Evaluation against objectives.....	48
Feedback from Economics Department .....	51
Further development.....	52

## Research

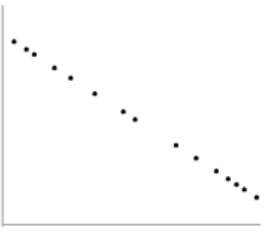

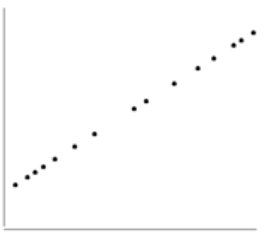
Type	Source	Date Accessed	Summary	Useful	Reliable
Website	<a href="https://www.mathsisfun.com/data/correlation.html">https://www.mathsisfun.com/data/correlation.html</a>	14/09/2016	Method for calculating a correlation	Y	Y
Website	<a href="http://databank.worldbank.org/data/home.aspx">http://databank.worldbank.org/data/home.aspx</a>	5/09/2016	Data source for correlation. Lots of data on world development and indicators of poverty for countries all over the world	Y	Y
Website	<a href="https://msdn.microsoft.com/en-us/library/system.windows.forms.datavisualization.charting.statisticformula.correlation(v=vs.110).aspx?cs-save-lang=1&amp;cs-lang=vb#code-snippet-1">https://msdn.microsoft.com/en-us/library/system.windows.forms.datavisualization.charting.statisticformula.correlation(v=vs.110).aspx?cs-save-lang=1&amp;cs-lang=vb#code-snippet-1</a>	14/09/2016	Coding method for data correlation		
Website	<a href="http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442">http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442</a>	15/9/16	Data correlation examples Pearson's correlation coefficient	Y	
Website	<a href="http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/">http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/</a>	21/09/16	Page with three different types of correlation that are explained in detail.		
Website	<a href="http://www.statisticssolutions.com/point-biserial-correlation/">http://www.statisticssolutions.com/point-biserial-correlation/</a>	21/9/16	Point-Biserial correlation		
Website	<a href="http://www.statisticssolutions.com/conduct-interpret-partial-correlation/">http://www.statisticssolutions.com/conduct-interpret-partial-correlation/</a>	21/9/16	Partial correlation		
Website	<a href="http://www.statisticssolutions.com/bivariate-correlation/">http://www.statisticssolutions.com/bivariate-correlation/</a>	21/9/16	Bivariate correlation		
Website	<a href="http://data.worldbank.org/data-catalog/world-development-indicators">http://data.worldbank.org/data-catalog/world-development-indicators</a>	21/9/16	Data source for every country for lots of world development indicators	Y	Y
Website	<a href="https://www.rutc.ac.uk/courses/social-sciences/333-level-3/225-world-development-as-a2.html">https://www.rutc.ac.uk/courses/social-sciences/333-level-3/225-world-development-as-a2.html</a>	21/9/16	World development definition and summary	Y	Y
Website	<a href="http://www.globalissues.org/article/26/poverty-facts-and-stats">http://www.globalissues.org/article/26/poverty-facts-and-stats</a>	21/9/16	General facts and statistics about poverty levels in the world	Y	Y

The aim of the project is to identify a correlation between two indicators of poverty from world banks data source in order to meet the needs of the end user. There are currently thousands of pieces of data available online however there is so much of it that it is difficult to find a correlation due to the large quantity therefore the aim is to clearly identify whether there is a correlation or not, what correlation it is and how strong the correlation may or may not be.

## World Development

World Development is about developing an understanding of our role as global citizens and our responsibilities toward global development and sustainability. Currently almost 50% of the world's population lives on \$2.50 a day and world development focuses on how that figure can be decreased so that humanity can live comfortably without large differentiating inequality.

## Correlation

Values of Pearson's correlation coefficient		
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:		
r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

There are three ways of calculating a correlation which I have found through conducting my research: Pearson, Spearman and Kendall. The most common method used is Pearson correlation. Pearson r correlation is used in statistics to measure the degree of the relationship between linear related variables. The example give in the source is stock markets and calculating the correlation between two goods. The formula is given as  $r = \frac{N \Sigma xy - \Sigma (x)(y)}{\sqrt{N \Sigma x^2 - \Sigma (x^2)} [N \Sigma y^2 - \Sigma (y^2)]}$  where r = Pearson r correlation coefficient, N = number of value in each data set,  $\Sigma xy$  = sum of the products of paired scores,  $\Sigma x$  = sum of x scores,  $\Sigma y$  = sum of y scores,  $\Sigma x^2$  = sum of squared x scores,  $\Sigma y^2$  = sum of squared y scores. For the Pearson r correlation, both variables should be normally distributed. Other assumptions include

linearity and homoscedasticity. Linearity assumes a straight-line relationship between each of the variables in the analysis and homoscedasticity assumes that data is normally distributed about the regression line.

Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables.

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. Spearman rank correlation makes no assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. And scores on one variable must be monotonically related to the other variable. Ordinal scales rank orders the items that are being measured to indicate if they possess more, less or the same amount of the variable being measured. Ordinal data does not define the magnitude of the relationship between units only whether they are less than, greater than or equal to.

Point-Biserial Correlation is a correlation measure of the strength of association between a continuous-level variable and a binary variable. Binary variables are variables of nominal scale with only two values. They are also dichotomous variables or dummy variables in Regression Analysis. Binary variables are commonly used to express the existence of a certain characteristic or the membership in a group of observed specimen. Typically Point-Biserial Correlation would be used to answer questions on biology, medicine, sociology, social psychology and economics.

Spurious correlations occur when two effects have clearly no causal relationship whatsoever in real life but can be statistically linked by correlation. Spurious correlations are caused by not observing a third variable that influences the two analyse variables. Partial correlation is the method to correct for the overlap of the moderating variable.

The correlation coefficient between two continuous-level variables is also called Pearson's  $r$ . A positive  $r$  values expresses a positive relationship between the two variables while a negative  $r$  values indicates a negative relationship. A correlation coefficient of zero indicates no relationship between the variables at all. However, correlations are limited to linear relationships between variables. Even if the correlation coefficient is zero, a non-linear relationship might exist.

## **World Bank**

The World Bank is an international financial institution that provides loans to developing countries for capital programs. The World Bank is a component of the World Bank Group, which is part of the United Nations system. Their official goal is the reduction of poverty to work for a world free of poverty.

The data which is available to use in this project will come from World Bank who have an extensive list of world development indicators for virtually every country as well as aggregates. The table below shows a list of some of the countries and aggregates as well as a selection of the world development indicators that are included in the data set. The limitation of this data is that there is not data for every country for every indicator. The data

is updated quarterly in April, July, September and December although the periodicity of the data is annual.

Countries	Indicators
Aruba	GDP growth (annual %)
Andorra	GDP (current US\$)
Afghanistan	GDP per capita (current US\$)
Angola	Exports of goods and services (% of GDP)
Albania	Foreign direct investment, net inflows (BoP, current US\$)
Arab World	GNI per capita, PPP (current international \$)
United Arab Emirates	GINI index
Argentina	Inflation, consumer prices (annual %)
Armenia	Population, total
American Samoa	Life expectancy at birth, total (years)
Antigua and Barbuda	Internet users (per 100 people)
Australia	Imports of goods and services (% of GDP)
Austria	Unemployment, total (% of total labour force)
Azerbaijan	Agriculture, value added (% of GDP_
Burundi	CO2 emissions (metric tons per capita)
Belgium	Literacy rate, adult total (% of people ages 15 and above)
Benin	Central government debt, total (% of GDP)

## Interview with Josie Gadsby (Teacher of World Development and Politics)

### What does world development cover at a level?

In A level, we look at two different themes. Theme 3 looks at development theory so it's what we mean by development so part of its economic development part of its social and political and looks at different kinds of application theories to that and we look at a few key case studies. We look at India in quite a lot of detail and look at its social political economic development. And then the other thing is very much economic development so we look at global inequality so this is probably where it fits in best. We look at the impact of technology on development and the impacts of globalisation and FDI so it's much more economics focused. So, I think it fits in quite nicely with the global inequality section quite nicely.

### Which time periods of data are most useful?

I think for both politics and world development we tend to try and keep it up to date so modern but again documenting it over time is really useful so you start to see the dips and changes. But I'd say anything from the 1980s onwards is probably more useful

### Is it useful to use recent data to predict the future and how is it relevant?

It certainly can be but I think where it becomes most useful is when students can begin to identify when its most relevant for example with the kind of thing you're looking at because it can be really useful to do with healthcare social development as well as economic development there is a direct correlation between the two. But what it won't highlight is that not everyone is going to get access to that there is distinct inequality even in quite developed countries. You'll probably find china's got quite a high GDP per capita but not

everyone is going to be accessing that in terms of life expectancy. So, I think it can be very useful but it needs to be looked at with an analytical eye.

**What correlations would be interesting to you personally?**

So, I don't think it matters too much which bits of data you pick. For example, I think the ones you chose would be fine [life expectancy and GDP per capita] so I think something like that would be fine. I think you might find it gives you quite a skewed result because the GDP per capita is just an average and you'll find most people in those countries don't have anywhere near that kind of per capita income. So, one thing that might be more interesting to look at would be distribution of wealth so if you're thinking about within a country what level of inequality. What economic inequality there is would be quite an interesting other one. You might find that some countries don't have any data and that's usually because those countries are corrupt and lack the government and governance you need for that kind of data

**Are you aware of the data available on the World Bank website?**

In terms of world development, we tend to use the World Bank more for looking at different aid programs rather than data. So, we do a bit but because you're looking at differences between countries its more for what aid programmes they use. With politics, it's more geared towards things like debt relief as well.

## Analysis

<https://datahelpdesk.worldbank.org/knowledgebase/articles/906531-methodologies>

### Summary of Interview

From the interview that I carried out with world development and politics teacher Josie Gadsby, it is clear to me that my chosen topics to correlate between would be relevant to her subject in her teaching. A recent study published by HM Revenue and Customs has shown that there has been an increase in the number of children living in poverty within the UK and identifying a correlation between world development indicators would help to be able to identify why this is the case, not just in the UK, but around the rest of the world.

One point Josie made was that some data sets will be missing data from countries for various reasons, the most prominent being that many smaller countries, particularly islands, are corrupt in the way they are run politically and economically and therefore the data is missing from these countries as they either do not have it or they do not wish to share it outside of their boundaries. This is an issue I am aware of having previously looked at the data available from World Bank and I must now decide which sets to use to provide me with the most accurate and relevant results. I am planning to use indicators which the majority of countries have provided data for so that my results will have more relevance and use.

Another issue raised by Josie was that over time data loses relevance to them as world development is a fast-changing subject and therefore only modern data is relevant to them. Josie recommended not using data from before the 1980s and as a department they prefer data from the past 20 years as it is most relevant, showing current trends and changes over the years as well as consistencies.

### Data Flow Diagram

World Bank collects data in this way.



### Document Specification

Below is a document specification sheet containing information on the data that is going to be used to correlate between and how the data is currently stored within the document.

Volumetrics			
Document description	System	Document	Name
World Bank Data Catalog	Poverty Indicators	1	WDI Data
Primary Table Size	File Size	Number of sheets	Method of preparation
372241 x 60	92.2MB	13	Compilation by World Bank
Medium		Prepared by	



Spreadsheet			World Bank		
Frequency of preparation		Retention period		Location of file	
Annual		Forever		World Bank Data Website	
Volume	Minimum	Maximum	Av/Abs	Growth rate/fluctuations	
	1	4	1 per year		
<b>Data Dictionary</b>					
Ref	Name	Data Type	Regex	Occurrence	Source of data / description
1	Country Name	String		264	World Bank
2	Country Code	String		264	World Bank
3	Indicator Name	String		264	World Bank
4	Indicator Code	String		264	World Bank
5	Year	Integer	$\wedge \backslash d \{ 4 \} \$$	56	World Bank

	A	B	C	D	E	F	G	H	I
1	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964
2	Arab World	ARB	2005 PPP conversion factor, GDP (LCU per international \$)	PA.NUS.PPP.05					
3	Arab World	ARB	2005 PPP conversion factor, private consumption (LCU per international \$)	PA.NUS.PRVT.PP.05					
4	Arab World	ARB	Access to electricity (% of population)	EG.ELC.ACCS.ZS					
5	Arab World	ARB	Access to electricity, rural (% of rural population)	EG.ELC.ACCS.RU.ZS					
6	Arab World	ARB	Access to electricity, urban (% of urban population)	EG.ELC.ACCS.UR.ZS					
7	Arab World	ARB	Access to non-solid fuel (% of population)	EG.NSF.ACCS.ZS					

## Data from World Bank

Below are a series of screenshots showing the data which is to be collected for the investigation. The screenshots will show the series of webpages to go through in order to be able to access and download the data files. The circled part of the image shows the link that needs to be selected in order to proceed closer to downloading the data files.

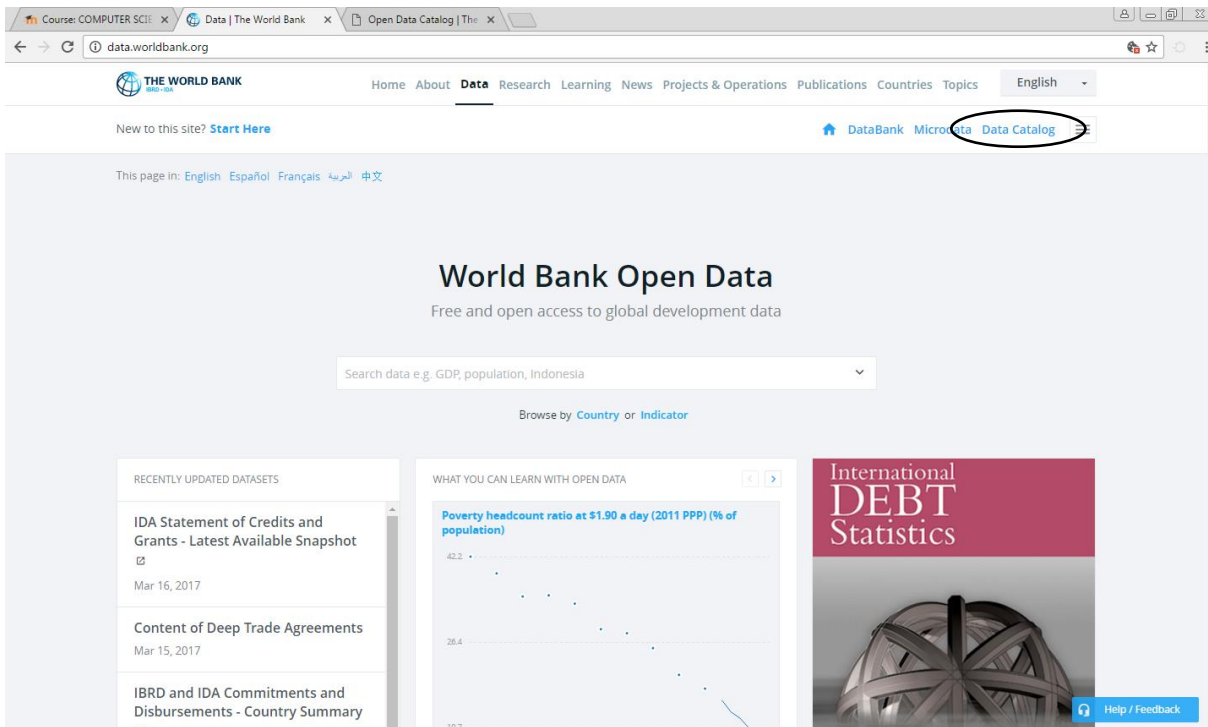


Figure 1

Figure 1 above shows the World Bank Data homepage which allows access to data collected, stored and maintained by World Bank.

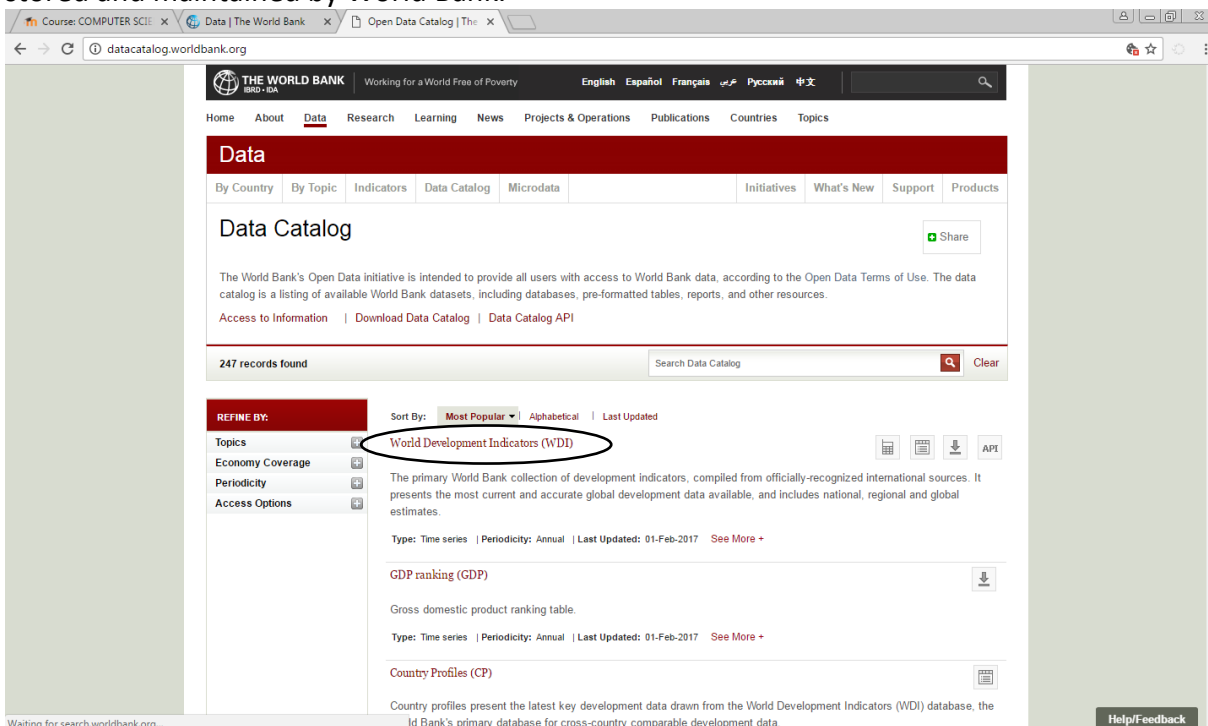


Figure 2

Figure 2 above shows the Data Catalog webpage for World Bank. World Development Indicators (WDI) is the first option for datasets out of 247 which are available. World Development Indicators is also the most popular option on this list and generally appears at the top of the webpage.

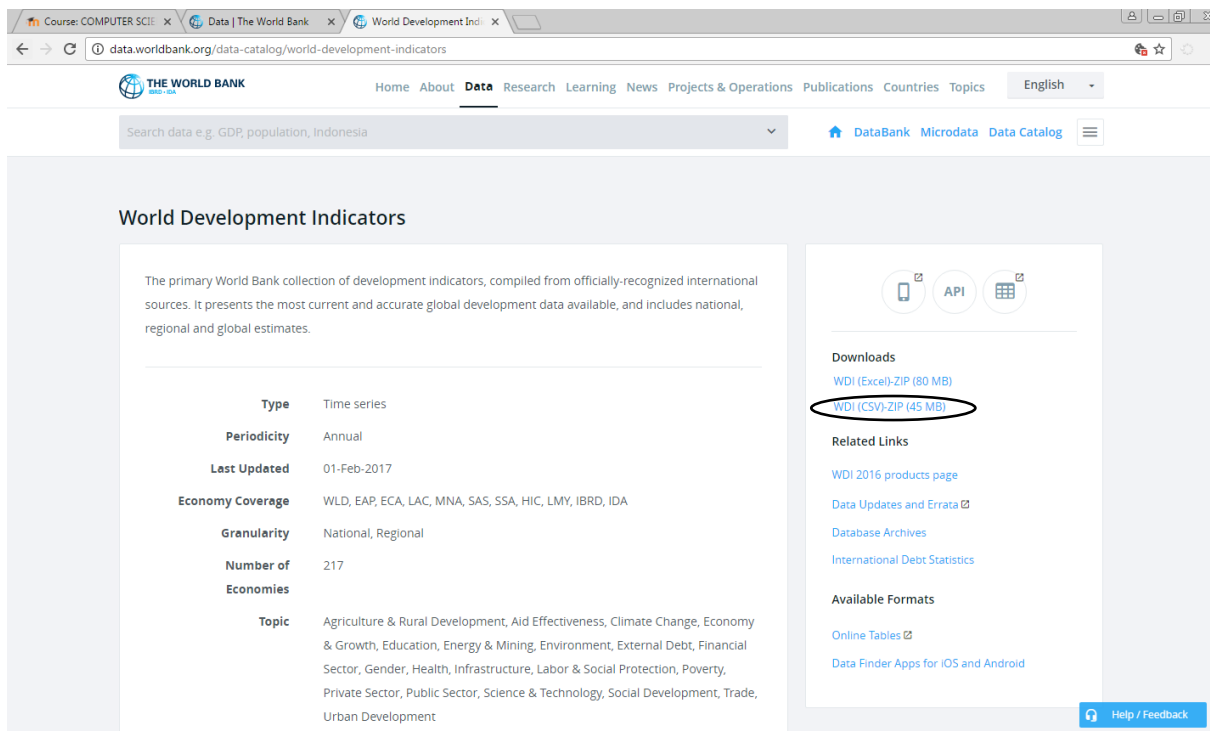


Figure 3

Figure 3 above shows a summary of the World Development Indicators dataset. It contains information about what is included in the file, when it was last updated, and how many countries/economies are covered. On the side there are several options for how to download the files. CSV files are smaller in size and there is less formatting so the information is easier to extract and analyse. The files are also downloaded as zips due to their size.

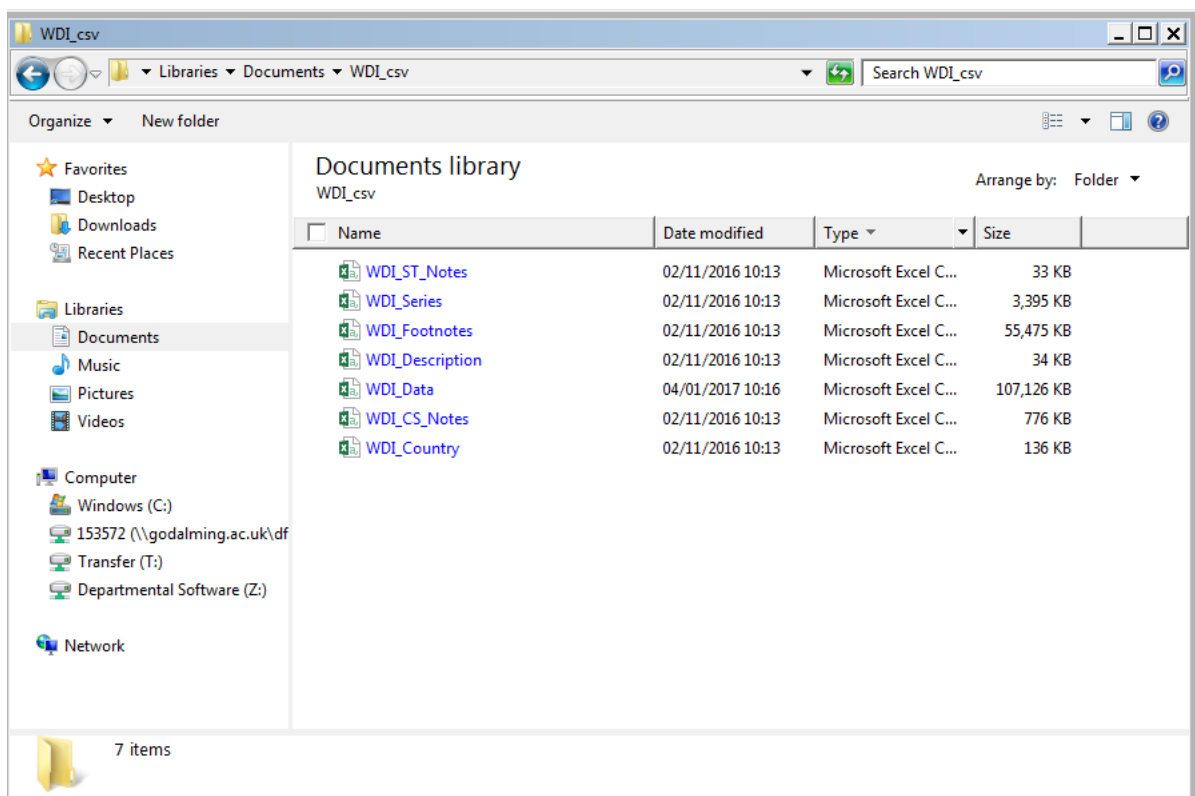


Figure 4

Figure 4 above shows the contents of the downloaded file once it has been unzipped. There are 7 csv files all of which contain information on world development indicators and they each vary in size.

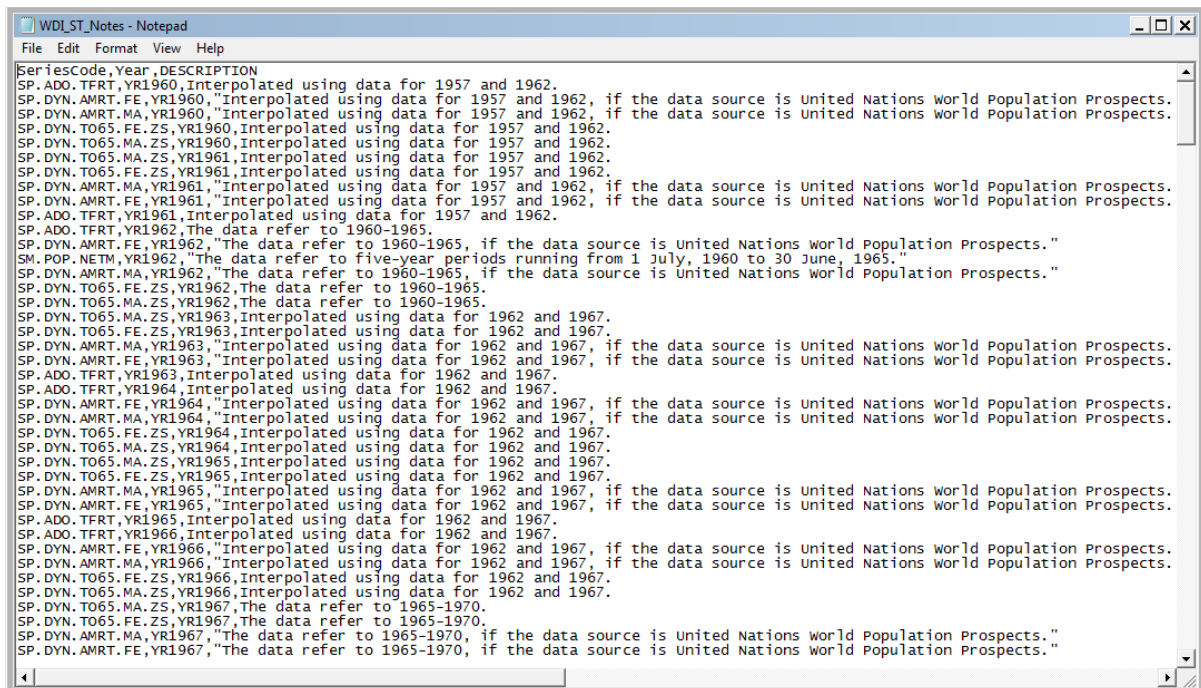


Figure 5

Figure 5 above shows the top of the first file in the folder, WDI\_ST\_Notes. This file contains notes about the data which has been collected.

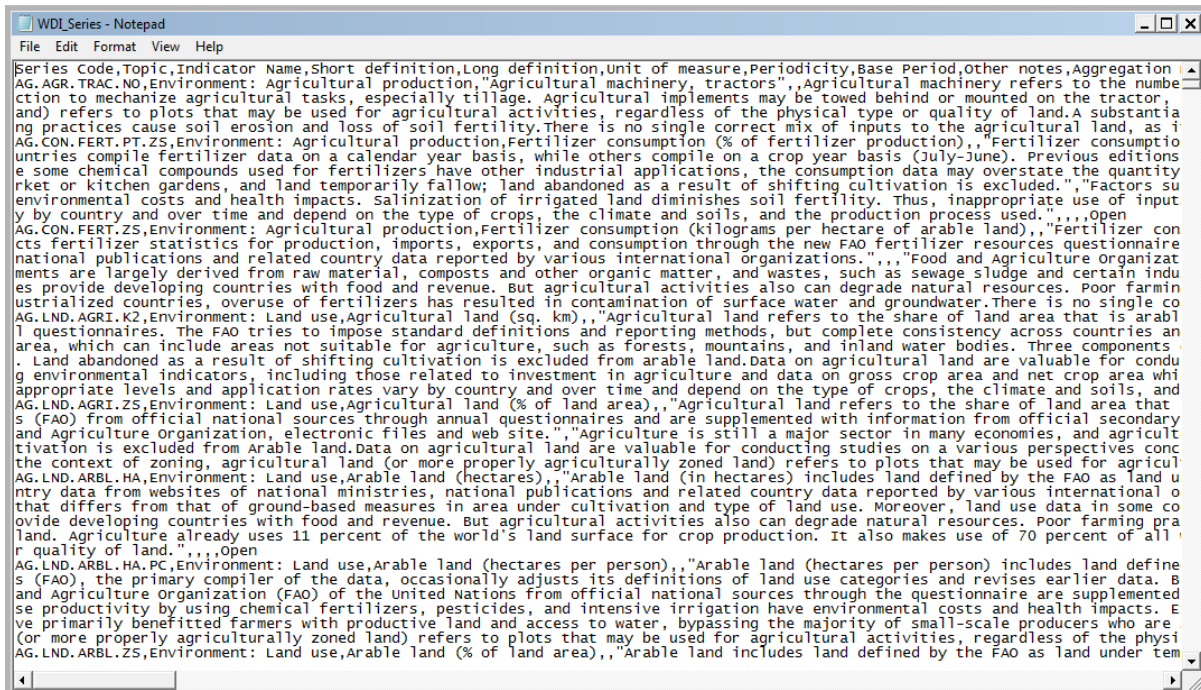


Figure 6

Figure 6 above shows the top of the file called WDI\_Series. This csv file contains various details about the indicators including, name, topic, indicator, definition and unit of measure.

Each indicator only appears in the list once and on each line has all of the information about it. In this file it shows that there are over 1400 indicators.

```

CountryCode, SeriesCode, Year, DESCRIPTION
ABW, AG, LND, FRST, K2, YR1990, Not specified
ABW, AG, LND, FRST, K2, YR2000, Not specified
ABW, AG, LND, FRST, K2, YR2005, Not specified
ABW, BM, KLT, DINV, CD, WD, YR1988, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1989, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1990, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1991, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1992, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1993, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1994, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1995, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BM, KLT, DINV, CD, WD, YR1987, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BX, KLT, DINV, CD, WD, YR1988, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BX, KLT, DINV, CD, WD, YR1989, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, BX, KLT, DINV, CD, WD, YR2013, "Source: United Nations Conference on Trade and Development, Foreign Direct Investment Online databases
ABW, DC, DAC, AUSL, CD, YR2000, Data are classified as official aid.
ABW, DC, DAC, AUSL, CD, YR2001, Data are classified as official aid.
ABW, DC, DAC, AUSL, CD, YR2002, Data are classified as official aid.
ABW, DC, DAC, AUSL, CD, YR2003, Data are classified as official aid.
ABW, DC, DAC, AUSL, CD, YR2004, Data are classified as official aid.
ABW, DC, DAC, AUTL, CD, YR2000, Data are classified as official aid.
ABW, DC, DAC, AUTL, CD, YR2001, Data are classified as official aid.
ABW, DC, DAC, AUTL, CD, YR2002, Data are classified as official aid.
ABW, DC, DAC, AUTL, CD, YR2003, Data are classified as official aid.
ABW, DC, DAC, AUTL, CD, YR2004, Data are classified as official aid.
ABW, DC, DAC, BELL, CD, YR2001, Data are classified as official aid.
ABW, DC, DAC, BELL, CD, YR2002, Data are classified as official aid.
ABW, DC, DAC, BELL, CD, YR2003, Data are classified as official aid.
ABW, DC, DAC, BELL, CD, YR2004, Data are classified as official aid.
ABW, DC, DAC, CANL, CD, YR2000, Data are classified as official aid.
ABW, DC, DAC, CANL, CD, YR2001, Data are classified as official aid.
ABW, DC, DAC, CANL, CD, YR2002, Data are classified as official aid.
ABW, DC, DAC, CANL, CD, YR2003, Data are classified as official aid.
ABW, DC, DAC, CANL, CD, YR2004, Data are classified as official aid.
ABW, DC, DAC, CECL, CD, YR2000, Data are classified as official aid.
ABW, DC, DAC, CECL, CD, YR2001, Data are classified as official aid.
ABW, DC, DAC, CECL, CD, YR2002, Data are classified as official aid.
ABW, DC, DAC, CECL, CD, YR2003, Data are classified as official aid.
ABW, DC, DAC, CECL, CD, YR2004, Data are classified as official aid.

```

Figure 7

Figure 7 above shows the top of the file, WDI\_Footnotes, which contains notes on the source and how the data is classified of the data included and references the country code, series code and year.

```

World Development Indicators
The world Bank
"14 October, 2016"

world Development Indicators (WDI) is the primary world Bank database for development data from officially recognized internationa
See other worksheets in this Excel file for a list of series changes since previous versions.

"14 October, 2016: An update was processed for national poverty and child mortality series, and to add 2016 data for Enterprise Su
"4 October, 2016: Data have been updated for international poverty and shared prosperity indicators, balance of payments series, m
"10 August, 2016: An update was processed for all monetary series, including corrections for reserves including gold and domestic
"22 July, 2016: An update was processed to revise Zambia's GDP growth rates and related national accounts and PPP data from 2011 t
"19 July, 2016: An update was processed to correct balance of payments data that were misaligned by one year for all countries and
"12 July, 2016: 2015 world values for gross capital formation, gross domestic savings, gross fixed capital formation, gross saving
"8 July, 2016: An update was processed to correct the scale of national accounts GDP components for Iceland, Ireland, Japan, Switz
"5 July, 2016: 2015 data (plus revised historical data, where necessary) for all countries and groups for population-, GDP- and GN
"14 June, 2016: A minor update was processed to correct 2015 statistical capacity indicator data for Brazil; military expenditure
"1 June, 2016: A minor update was processed to correct 1994 poverty data for Philippines; and 2012 births attended by skilled heal
"10 May, 2016: A minor update was processed to correct stock market index data for 2015 for France, Germany, and the United Kingdo
"2 May, 2016: A minor update was processed to revise and update national poverty, adjusted savings, and aggregate forest indicator
"14 April 2016: A minor update was processed to add grants and health expenditure data for 2014, and forest area as a percentage o
"11 April 2016: Full update of development data to coincide with publication of the world Development Indicators 2016 book. New in
"17 February 2016: Data have been updated for grants indicators and forestry indicators. External debt data for Malaysia, Romania,
"29 December, 2015: National accounts data for Eritrea from 2012 to 2014 have been removed pending further review. Data for series
"22 December, 2015: Data have been updated for malnutrition, national poverty indicators, and threatened species. Revisions and co

```

Figure 8

Figure 8 above shows the top of the file, WDI\_Description, which hold basic descriptions of the updates that have been made to the data set and when those changes have been made. This is one of the smallest files in the folder as it does not contain a particularly large amount of information unlike some of the other files.

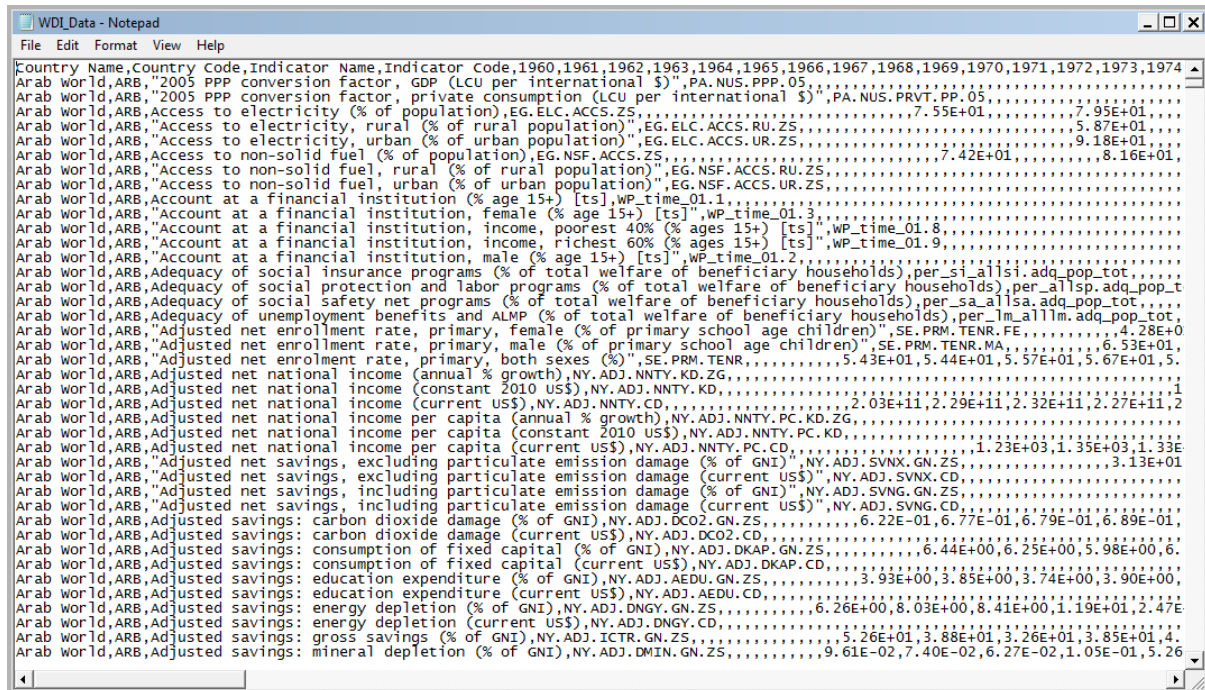


Figure 9

Figure 9 above shows the main file which is in the data set. WDI\_Data is the largest file in the folder and contains all the numerical values relating to the countries and indicators in the given years. As it can be seen in the figure, there are many gaps within the data where the data does not exist for those particular countries and indicators in certain years. It can also be seen that the numbers are stored in exponent form so will therefore need to be converted before they are used in the program.

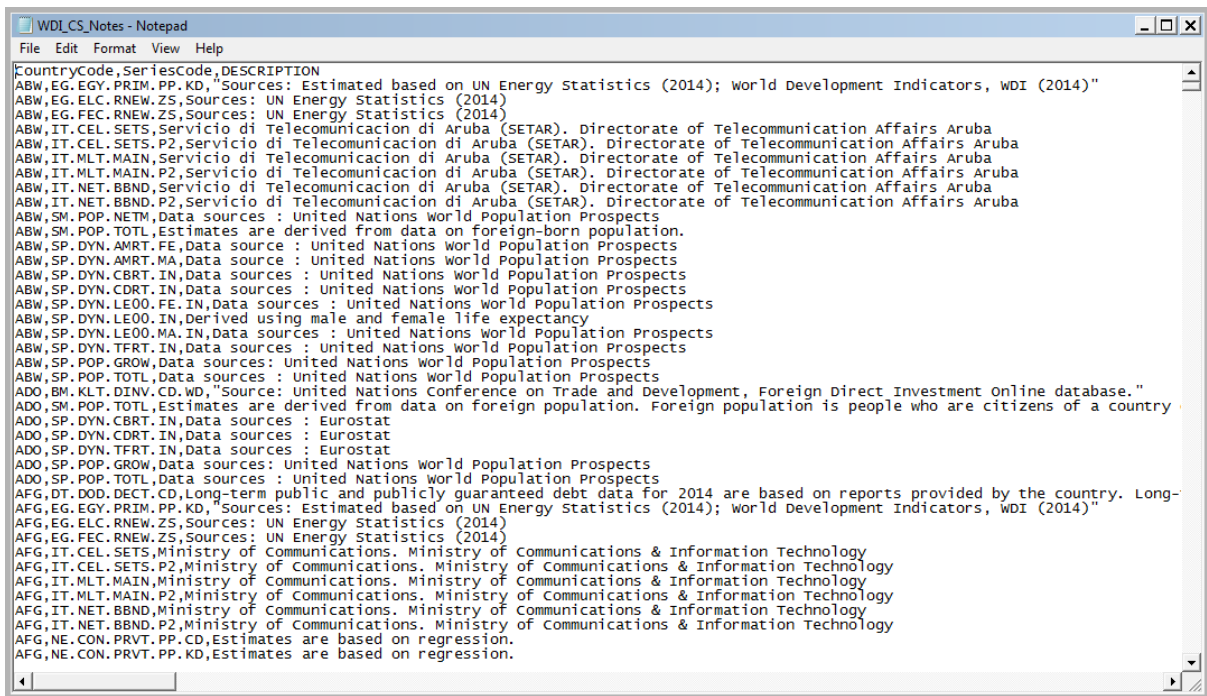


Figure 10

Figure 10 above shows WDI\_CS\_Notes which is a file with notes on the indicators for some of the countries which is referenced by country code and series code. This is one of the files which is not relevant and would have no use in the project.

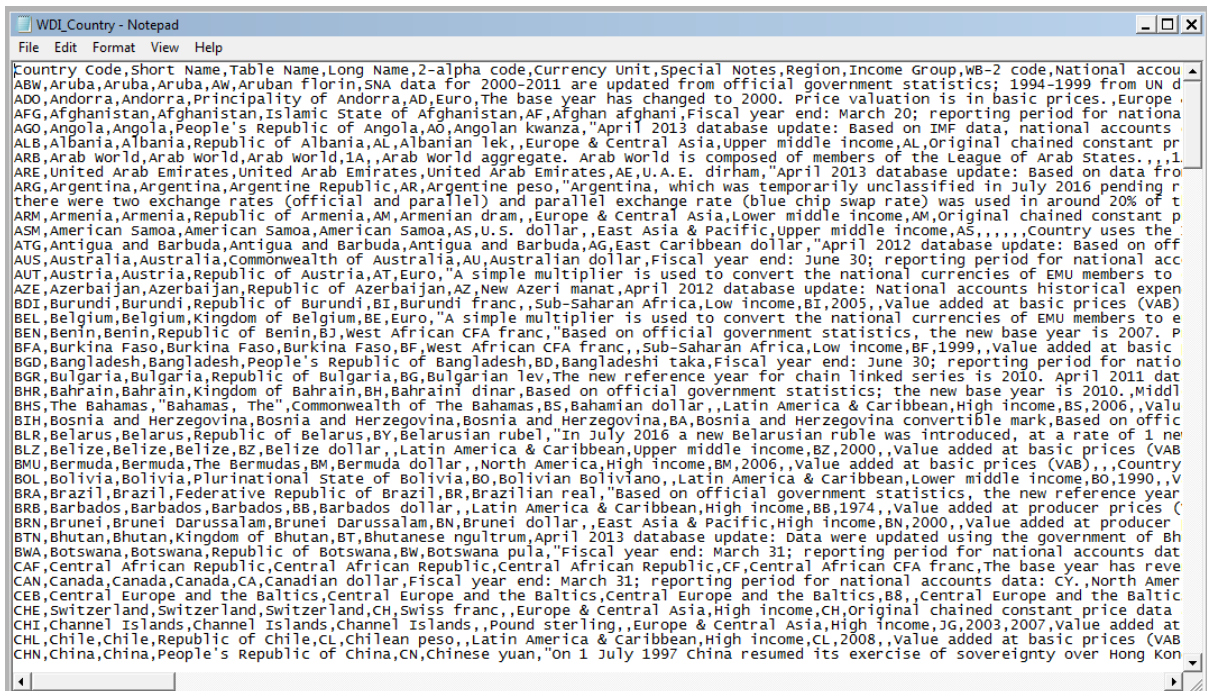


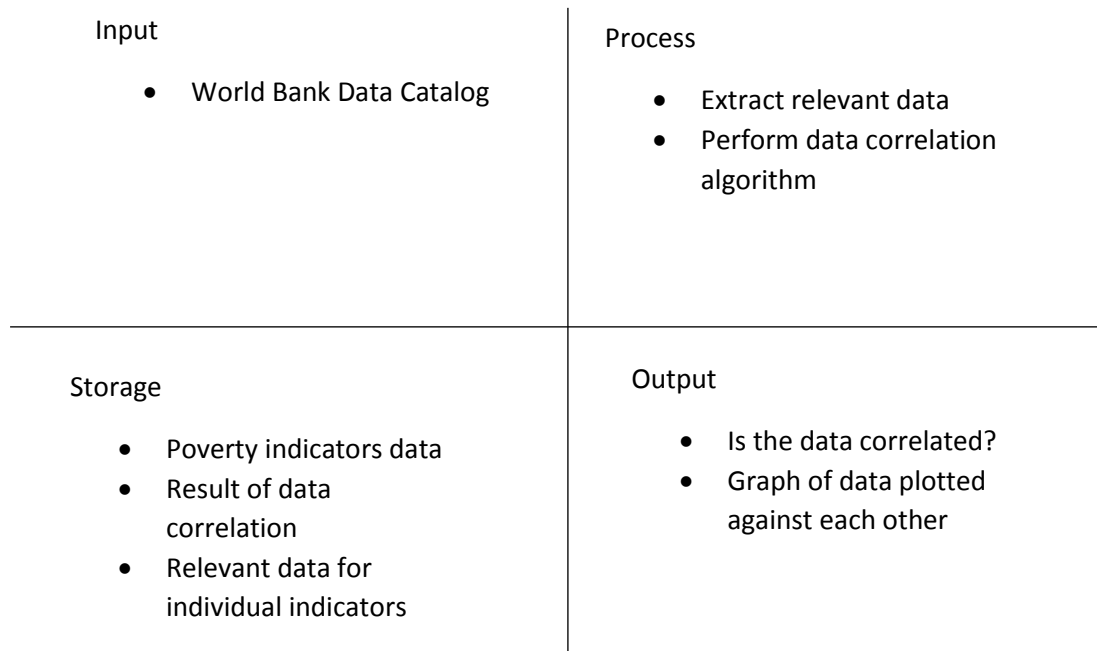
Figure 11

Figure 11 above shows the final file in the downloaded folder, WDI\_Country. This file contains a list of all of the countries and aggregates which are included in the data set. There are over 200 countries and indicators which are used. Each country has a shortened country code which, as shown in the figures above, is referred to from within other files.

There is also a short name, table name, long name, 2-alpha code, currency unit as well as lots of other information about the country's population and economy. This file is particularly useful as it contains a list of all of the countries which is repeated only once unlike in figure 9, the data file, where each country is repeated for each indicator before moving on to the next country.

## IPSO Chart

Here is an IPSO chart showing how the data will move around throughout the system.



## Ways of calculating correlation

There are multiple ways of calculating correlation and I plan on using multiple formulae to verify each of the correlations. The methods/formulae displayed below are taught in A level maths as part of statistics 1 which I am currently learning.

1. Product-moment correlation coefficient is calculated using the formula

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

where

$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$



$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

2. Pearson's product-moment correlation coefficient, denoted  $r$ , will give values between +1 and -1. When  $r \approx 1$  ( $x, y$ ) will lie close to a point with a positive gradient, when  $r \approx -1$ , points will lie close to a line with a negative gradient.
- $r \approx 1$  strong positive correlation
  - $r \approx 0$   $x$  and  $y$  unconnected
  - $r \approx -1$  strong negative correlation

Spearman's rank-order correlation involves ranking the data into order and then uses the tied ranks for the two sets of data to perform the calculation. Potentially ranks could be tied and there is a variation in the formula when this occurs.

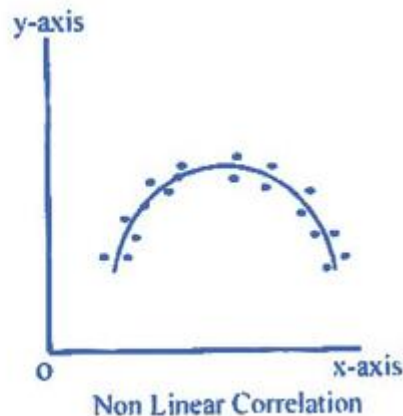
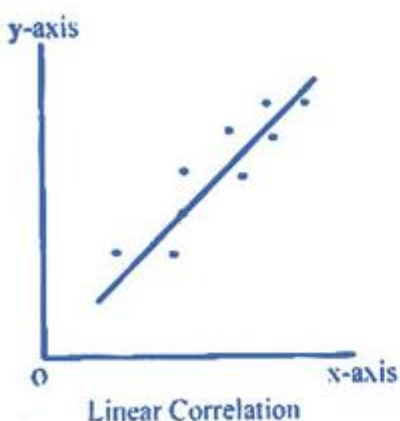
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where  $d_i$  = difference in paired ranks and  $n$  = number of cases

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Where  $i$  = paired score

### Linear and Nonlinear Correlation



Correlation is said to be linear if the ratio of change is constant. If all the points on the scatter diagram tends to lie near a line which look like a straight line, the correlation is said to be linear.

Correlation is nonlinear is the ration

of change is not constant. Often in this case the correlation on a scatter diagram would be curved.

In this project, non-linear correlation would not be picked up as the Pearson's correlation coefficient would be most likely to result in a number close to 0 showing that there is no clear correlation.

## **Useful Indicators**

From the research, I have done I can see that the most useful indicators to use include one economic indicator and another that is non-economic. This way it enables a comparison to be made between the economic state of a country and whether they have necessities such as electricity, clean water or have access to an education. I also think that the user should be able to choose which indicators they would like to correlate between as there are many to choose from and only some of them the user will be interested in. This will also enable it to be used by other users potentially for other purposes.

## Requirements

1. Data file is accessed
  - 1.1. Text file holds the memory locations of data, series and country csv files
  - 1.2. File locations text file is iterated over to match corresponding files
  - 1.3. Countries file is accessed
    - 1.3.1. File containing countries information is parsed
    - 1.3.2. Each country is extracted from the file and input into an abstract data type
  - 1.4. Series file is accessed
    - 1.4.1. File containing information about indicators is parsed
    - 1.4.2. Each indicator is extracted from the file and input into an abstract data type
2. User selection
  - 2.1. User is shown a list of options for what can be correlated
    - 2.1.1. The user must only be able to select one of these options
  - 2.2. User is shown a list of every country and every indicator in check lists
  - 2.3. User is able to select a country/countries and an indicator/indicators to correlate together
    - 2.3.1. For option one the user is able to select one country and two indicators only
    - 2.3.2. For option two the user is able to select two countries and one indicator only
3. Extraction of data points
  - 3.1. The data file is opened from the memory location held in the files location text file
  - 3.2. The file is iterated over until the selected country and indicator pairing is found, this will occur twice to obtain both data sets
  - 3.3. The values for each of the years are stored in a list for that specified data set
    - 3.3.1. The values are converted from an exponent into the correct double data type
  - 3.4. The data sets are compared and empty values are removed so that only pairs of data that exist are kept in the data set
4. Calculate correlation
  - 4.1. Correlation calculations stored as individual functions
  - 4.2. Pearson's product-moment correlation coefficient is able to be calculated
    - 4.2.1. The values are input into the correct part of the formula
    - 4.2.2. Calculation gives out a valid answer
  - 4.3. Spearman's rank-order correlation is able to be calculated
    - 4.3.1. The values are input into the correct part of the formula
    - 4.3.2. Calculation gives out a valid answer
  - 4.4. Data input from the data sets from extraction of csv file into formula
  - 4.5. Calculates strength of correlation
  - 4.6. Numerical values from each calculation are stored for future use in the final report
5. Output report to user
  - 5.1. Data sets extracted from the file are plotted on the graph
  - 5.2. Graph is plotted accurately using the coordinates
  - 5.3. Graph is displayed to the user with fully labelled axis with the countries and indicators
  - 5.4. Description of the graph is presented along with the graph

- 5.4.1. Key features of the graph and correlation calculation are noted
- 5.4.2. Gaps filled in for specific data
- 5.4.3. Sentences for each feature are output to the user

## Design

VB Forms has been chosen as the language for producing the program. This is because forms can be easily manipulated for their graphical output without much difficulty. VB.net can be used to control, edit and maintain forms and also allows for the reading and writing of text files which is necessary in order to be able to read

### Prototype

Use data that is in the world development indicator data set to confirm the calculation works by testing a known correlation. The prototype will need to test all the types of calculation which will be used. The prototype will need to have much of the functionality that is required of the final solution. This is so that each part of the solution can be tested and each part of the solution needs to work because subsequent elements are reliant on previous results. The interface for the prototype need not be too complex as long as it is functional.

Key features of the prototype are that it must be able to produce an accurate result for the Pearson's correlation calculation to prove a known correlation and it must be able to plot an accurate graph for these results.

### Process

#### Input data from CSV

File to be downloaded and stored on the computer locally so that the data can then be extracted from the file. The original file come from this url:(<http://data.worldbank.org/data-catalog/world-development-indicators>) and is downloaded as a zipped folder which contains 7 csv files. The primary file that is used for the actual data is called "WDI\_data.csv". As it is shown in the series of figures on the World Bank Data it can be seen that in figure 9, the WDI\_Data file, to iterate over the entire file to extract an entire list of countries and indicators would be inefficient as it contains over 300,000 lines. Therefore two other files are to be used in order to generate lists of countries and indicators without having to iterate over the entirety of the data file. The file for countries is "WDI\_country" and the file for indicators is "WDI\_series". Both of these files are in the original zipped file which is downloaded from the World Bank website. In each of the files, each indicator and country name exists only once on each line and it is therefore much quicker and more efficient to iterate over these files rather than the main data file.

The data file locations are initially retrieved from .txt file where they each exist on different lines and then are matched up to their corresponding variable.

#### Parsing of data

There are parsing libraries that are available to use in VB.net which will be made use of to extract the data. In order to do this the TextFieldParser class in VB.net will be used. This allows delimiter types to be set as opposed to splitting lines of text individually at the commas. An example line of text from the csv file shows that there are commas within the items and therefore that item is surrounded by quotation marks. The string.split() method would not allow for this as easily as the TextFieldParser would.

The data will be read line by line and split as it is delimited by commas. The structure of each line is the same for each and therefore can be easily stored in abstract data types. The first line of the file shows that the lines are split as follows:

CountryName,CountryCode,IndicatorName,IndicatorCode,1960,1961,1962,1963,1964,1965,1966,1967,1968,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016. Each line of text can easily be split into a list of strings and then matched by index to a list of the first line to know what that particular element is and then how to put it into the abstract data types.

There are to be two main abstract data types as classes: Country and Indicator both of which will inherit common variables in order to make the program as efficient as possible. Country and Indicator will both inherit from Value and Value will hold variables for the name and the code associated with either the instance of country or the instance of indicator. Inheritance enables Country and Indicator to not be identical and in addition, extra variables and procedures are able to be added if necessary.

They could be stored in two lists of lists called country and indicator. Each list will contain lists of their corresponding abstract data types. For example, country would have approximately 300,000 lists each of the lists would contain 56 elements, one for each year and each of the elements being an instance of country holding the name, year and most importantly value. The list of lists for indicator will essentially mirror that of country however instead of holding the country name it will hold the indicator name. This will enable indexing to be used to connect the lists together when retrieving the values to plot for the graph.

While the file is initially iterated over, each of the country names and indicator names will be stored in separate lists in order to keep track of how many there are and to make it easier to put them into the checkedlistboxes for user selection.

Alternatively, the values are not all extracted and the only values that are extracted aren't retrieved from the data file until the user has selected which data they would like to view. This would be much more efficient and the entire file would not necessarily have to be iterated over. The user selection would be carried out (as described below) and then the data file would be iterated over until the pairing of indicator and country are found. The length of time this would take would vary dramatically. For example, if the user had selected two indicators from the Arab World this would be very quick as the Arab World is the first country/aggregate that appears in the file, however if the user were to select two indicators from Zimbabwe this would be much slower as virtually the entire file would have to be iterated over in order to retrieve that selected data. Once the pairing has been found, the values for each year are inserted into a list. This process is repeated for the second pairing so that the length of the lists are the same and the indexes match the years for the data.

There are many data points missing from the data set as the values do not exist in the file. This will have to be dealt with in order to ensure that a graph can still be plotted or

alternatively a message is displayed that the data does not exist for the selection which has been made by the user. In order to deal with missing data points, a flag value can be used to represent the missing value or else nothing will be input into the data set list. For example, -99.99 could be used. This way both lists of data points will initially be the same length. These can then be iterated over together to check if the either or both values held in the same index of each list are -99.99 and that way it is known that they can be removed from the lists. This ensure that only valid data points are plotted on the graph and used in the correlation calculations.

### **User selection**

There are three ways to correlate within the data so the user initially will be shown a welcome screen form and then depending on which type of correlation they would like to view, they are shown a corresponding form window. The forms will have to be different as they give instructions as to how many selections are to be made. Each form will also allow the correct amount of selections to be made in each checkedlistboxlist. Validation must be used to ensure that the correct number of boxes are selected at any one time. The best way to do this would be to use a Try Catch which would ensure that the correct number of selection are made before the next form is opened.

The initial form will only take one input of the user selection for which type of correlation they would like to view. The best method to do this would be to use radio buttons as opposed to checkedlistboxes. Radio buttons only allow one selection to be made at any one time and as soon as another radio button is selected, the button that was originally selected is deselected.

This process is displayed visually in the HCI section of design in figures 14-20.

### **Calculations to be carried out**

Calculations should be stored as functions so that they can be called multiple time for various sets of data. These will then be called when the user has selected which inputs to use in the calculation. The data sets which have been retrieved from the data file will be passed into the function as lists of double and then the calculations will be carried out on the data sets.

Pearson's and Spearman's have been chosen as they are the most commonly used correlation calculations. The other calculations that have been researched could be added at a later date however the chosen two are used widely at a high level and are therefore appropriate for this project.

When calculating Pearson's correlation, the result will be verified as the result of the calculation must be between  $-1$  and  $+1$  so if the result given from the calculation is outside these boundaries there has been an error in the process.

Spearman's rank correlation is more complex than Pearson's product-moment correlation as the data needs to be sorted into ascending order before any correlation calculation can be carried out. This could be done in a number of ways. Firstly there could be two functions: one which holds the spearman's rank correlation calculation formula and another function

which sorts the data into order. Alternatively they could both be combined together within one function which carries out the correlation and the sorting in one go. The data must be sorted either way and this can be done in a number of ways as well. There are a number of sorting algorithms two of which are taught in the decision one module of A Level Maths.

Bubble sort makes multiple passes through a list. It compares adjacent items and exchanges those which are out of order. Each pass through the list places the next largest value where it should be located.

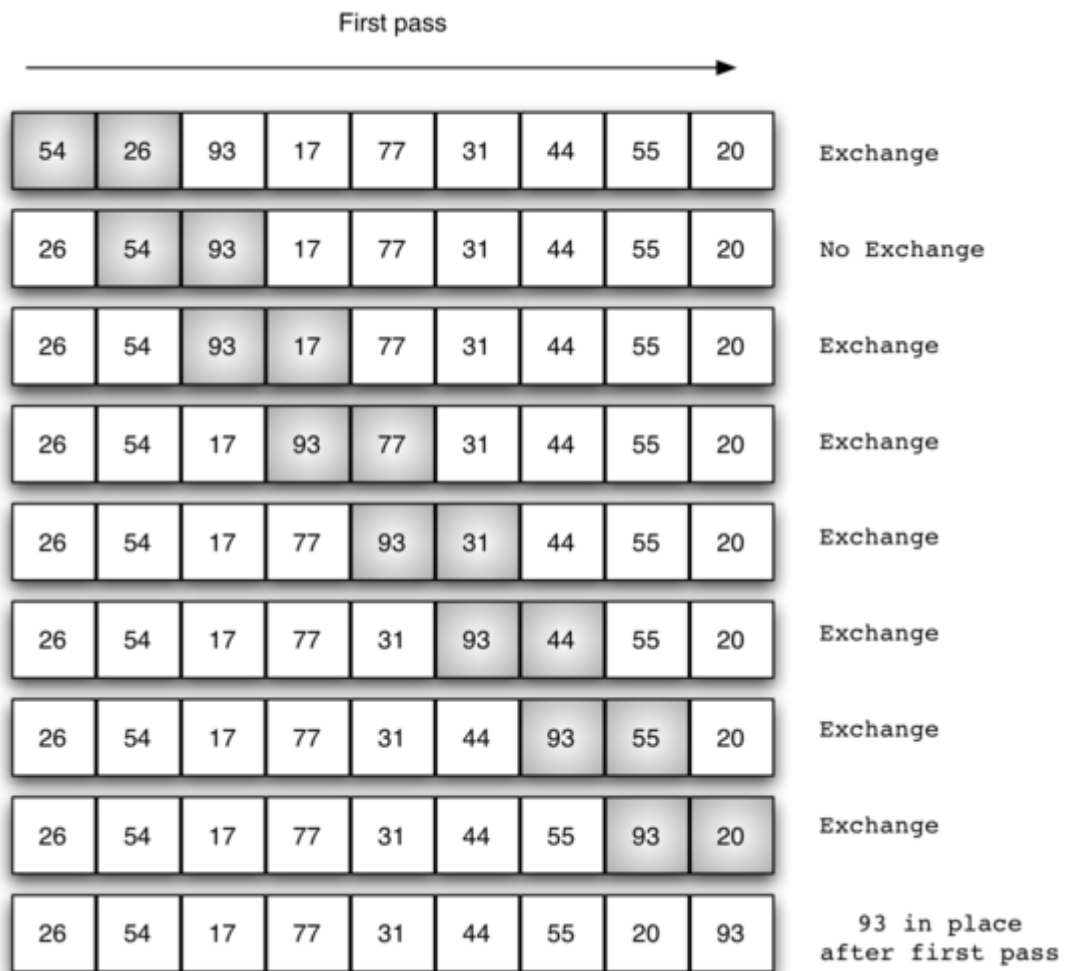


Figure 12

Figure 12 shows the first pass of an example bubble sort and shows how the largest number, 93, is placed at the end of the row as the highest value.

Bubble sort is often considered to be the most inefficient sorting method since it must exchange items before the final location is known. The sort can take a long time as the algorithm must pass over every item in the list unless it is modified so that if on one pass no exchanges are made the sort must be complete as no changes have been made.

A more efficient alternative sorting algorithm to use as opposed to using the bubble sort algorithm would be the quick sort algorithm. The quicksort algorithm uses pivots to split the



list into successively smaller sublists. Each element in the sub list is then compared to the limit and then put on the correct side of the pivot depending on whether the list is being sorted into ascending or descending order. The pivot is always the middle element if the number of items in the list is odd however if the number of items in the list is even, out of the middle two elements in the list, select the element on the right to use as the pivot. The algorithm is stopped once all of the elements in the list have been chosen as pivots.

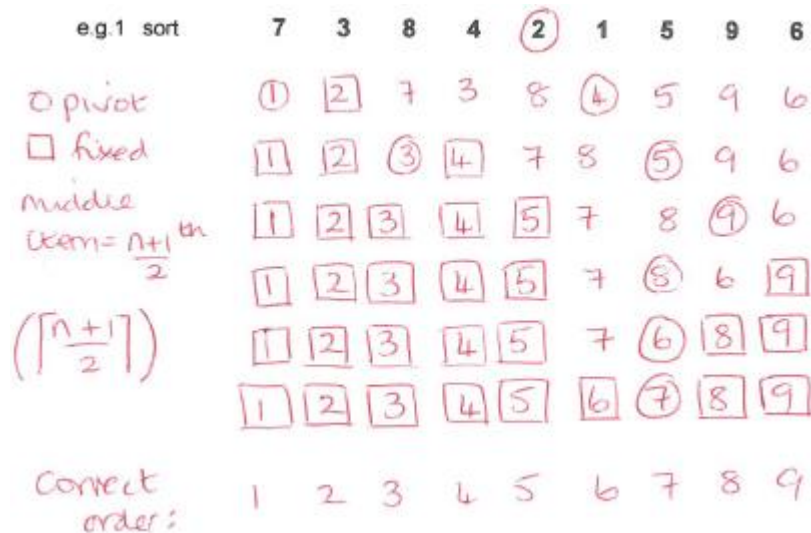


Figure 13

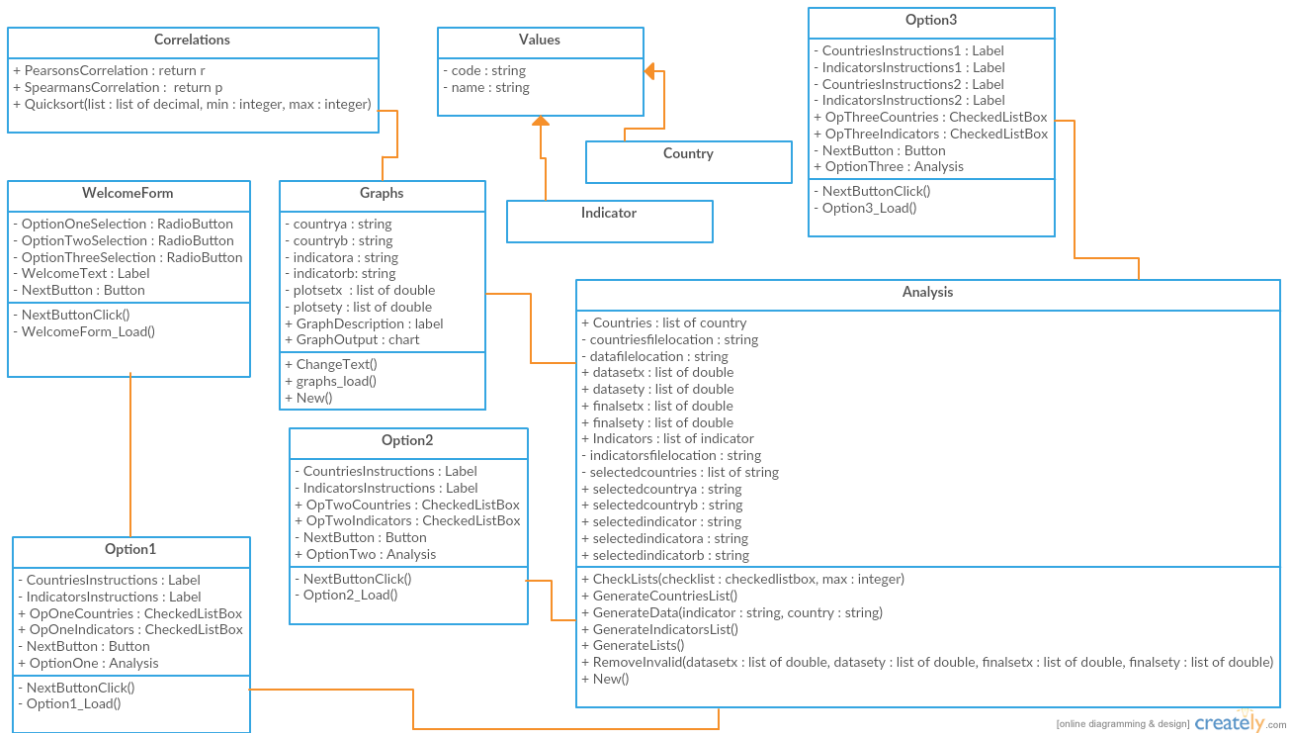
Figure 13 above shows an example of an entire quick sort. Both this example in figure 13 and the example in figure 12 shows sorts of lists containing 9 elements. The bubble sort example in figure 12 shows one pass which only sorts one element into the correct location where as figure 13 shows the entire sort and that the quick sort is much more efficient and quicker to carry out.

Quick sort algorithm is considered to be much more efficient than bubble sort as multiple elements in the list can be placed in the correct location on any one pass therefore the list is not iterated over as many times, thus making it more efficient to sort.

### Results plotted on a graph

The data sets, once they have been filtered through, are then used to represent coordinates to be plotted on a graph. These are to be plotted on a scatted graph in order to best show any correlation visually and then output to the user on the final form. There are more details on how this can be carried out in the Output and Graphing sections below.

## Class Diagram



## Pseudocode

### Pearson's product-moment correlation calculation

```

sum_sq_x = 0
sum_sq_y = 0
sum_coproduct = 0
mean_x = x[1]
mean_y = y[1]
for i in 2 to N:
    sweep = (i - 1.0) / i
    delta_x = x[i] - mean_x
    delta_y = y[i] - mean_y
    sum_sq_x += delta_x * delta_x * sweep
    sum_sq_y += delta_y * delta_y * sweep
    sum_coproduct += delta_x * delta_y * sweep
    mean_x += delta_x / i
    mean_y += delta_y / i
pop_sd_x = sqrt( sum_sq_x )
pop_sd_y = sqrt( sum_sq_y )
cov_x_y = sum_coproduct
correlation = cov_x_y / (pop_sd_x * pop_sd_y)
  
```

### Bubble sort algorithm

Function bubblesort (var a as array)

```

    For i from 1 to N
        For j from 0 to N -1
            If a[j] > a[j+1]
                Swap (a[j], a[j+1])
End function

```

### Quick sort algorithm

```

Sub Quicksort(A as array, low as int, high as int)
    If (low<high)
        Pivot_location = Partition(A as array, low as int, high as int)
        Quicksort(A,low,pivot_location)
        Quicksort(A, pivot_location,high)
End Sub

```

```

Function Partition(A as array, low as int, high as int)
    Pivot= A[low]
    Leftwall = low
    For i = low + 1 to high
        If (A[i] < pivot) then
            Swap(A[i], A[leftwall])
            Leftwall = leftwall + 1
    Swap(pivot,A[leftwall])
    Return leftwall
End function

```

### Storage

The data which will be used for the correlation calculations will be extracted from World Bank as a CSV file so one way of storing this data would be to put it into a database. I don't believe that using a database would be the best option because there is not an excessive amount of data to store. The CSV file size is 51,194 KB.

The data in the CSV file will need to be parsed into a list of lists so that it can be stored within the program. The structure I intend to use is below. This is a good way to store the data so that it can be retrieved easily to be put into calculations.

List [Years]

    [(Abstract Data Type) Country]

    [Indicator, Value]

The calculation formulas will be stored as functions so that they can be called upon multiple times throughout the code.

The program would also need to store the generic text output for the report at the end which will depend on the calculation values.

## HCI (Human Computer Interface)

There are two main options for displaying the user interface, both of which use Windows Forms.

The user will need to be able to see the options for the indicators to select and options for which type of correlation they want to see. One option for this process could be to use VB forms so that the user presses buttons that clearly display what each option will do. In this method, there is the potential for dropdown menus to be used to display all of the indicators. VB Forms would provide a simple interface for the user to be able to easily select exactly what they would like to see correlated. As shown in the figure below, the user can select check boxes to be select specifically which indicators, years and countries they wish to view correlated.

I believe that creating a graphical interface will be the best way for the user to be able to select which indicators they wish to view. There are a limited number of options for what can be correlated together so visually these can be easily represented.

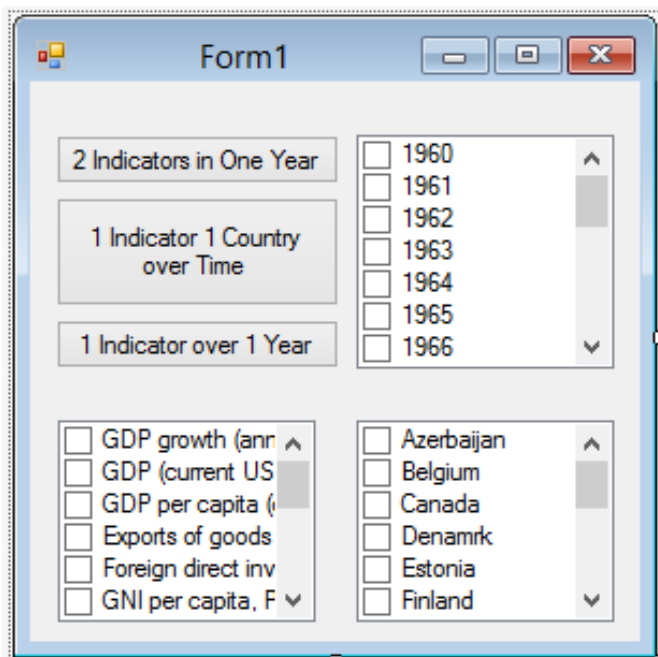


Figure 14

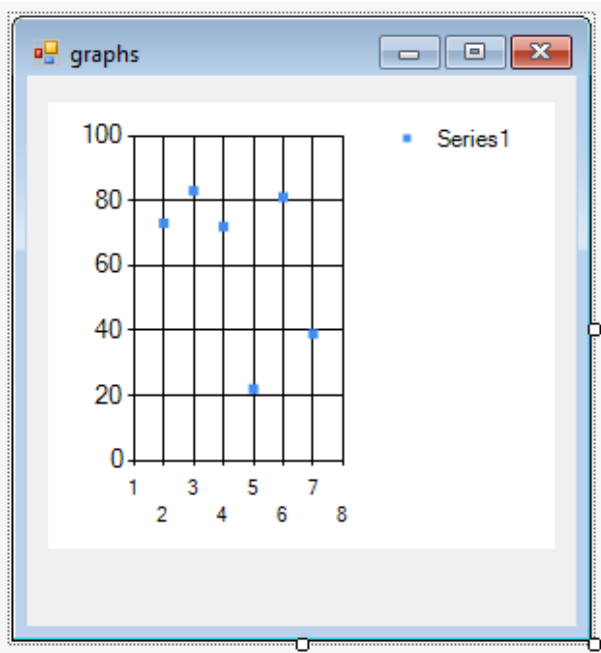


Figure 15

An alternative to only having one form with all the option on would be to have a series of forms that follow a welcome screen. The welcome screen would allow the user to select which type of correlation they would like to carry out and the following form would then be able to show a full list of countries and indicators and then only allow the user to select the correct number of each from their original selection from the first form. This method is more structured and guides the user more clearly through the process. It would be easier as less verification would be needed to know what the user is trying to do. The figures below show examples of how the series of forms would look depending on the user's original selection.



Figure 16

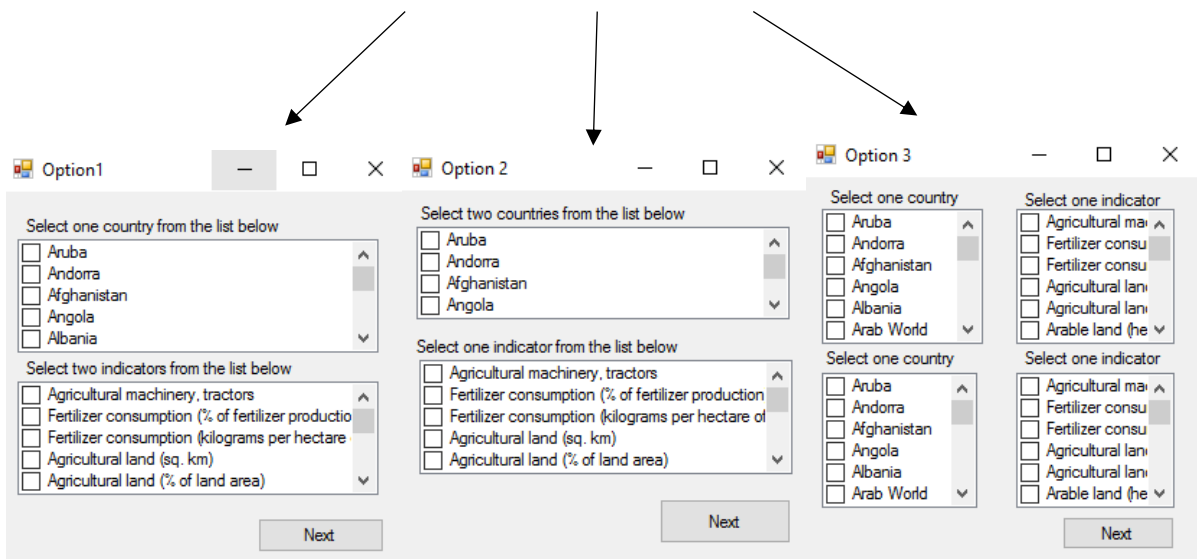


Figure 17

Figure 18

Figure 19

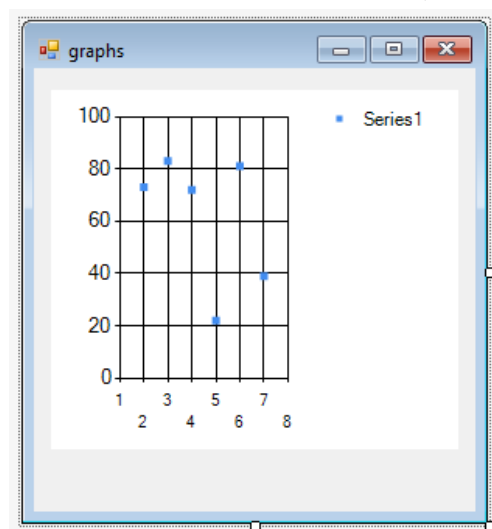


Figure 18

## Input

The CSV file will need to be extracted from World Bank. The file will be downloaded locally to ensure that it can always be accessed. An alternative way to retrieve the file would be to download it by the user inputting the URL of the webpage where the file is stored and then the file being stored and accessed from where it has been downloaded to. However, the file is updated regularly and there is the potential that it could be relocated to another URL so it may not be as easy to access it. Another issue with this method is that the file is downloaded as a zipped folder and therefore would have to be unzipped before it could be used to access the data. This process is time consuming for the program as the zipped folder is still large and would take time to download and unzip so it is much more efficient to have the file already downloaded. The user will also be able to select specifically which countries they would like to see. They may not wish to view every country all the time if they are viewing one indicator over time so they will need to be able to select which countries they wish to view specifically and there will be an option for showing all of the countries together.

The first user input will be made in the first form where they select one of the radio buttons depending on which correlation they would like to view. This process is easy to validate as only one radio button can be selected at any one time so as long as one radio button is selected the program will be able to continue.

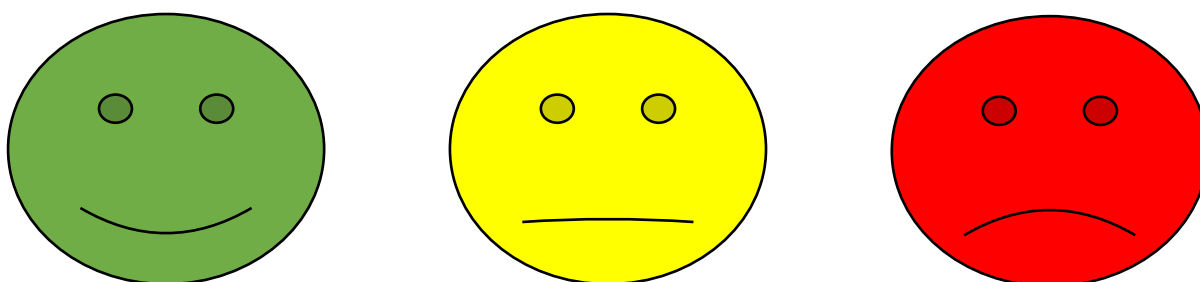
The users input for country and indicator will need to be verified to ensure that they are selecting the correct number of checkedlistboxes. The best way to do this would be to use a try catch which would enable exceptions and the next form will not be displayed until the correct number of selection have been made.

## Output

There will be a visual output to the user to indicate the type of correlation to the user. The type of correlation will either be positive or negative and the strength of correlation should also be shown.

For example, the faces below could indicate the strength of positive correlations. The larger smile on the left would indicate a very strong, or perfect, correlation and the smaller smile on the right would indicate a weaker, yet still positive, correlation.

Alternatively, to indicate negative correlation, faces such as the ones below could be used to show the strength.



A key must be displayed to the user so that they know what each of the faces mean and what strength they indicate in relation to other faces.

The strength of the correlation could be summarised into a description which is output to the user on the final screen. Within the description should be the numerical value from the Pearson's correlation calculation as well as a value for Spearman's rank correlation as well.

The main output after the user selection process will be a final form which displays a graph that plots all the points for the selected countries and indicators to give a graphical display of how the results are related. The graph could either be a straightforward scatter graph or alternatively it could be a line graph. However, a line graph may not be the best way to show the relationship between the two sets of data particularly if the data is unrelated then the lines will be all over the graph showing no distinct relationship.

A chart can easily be produced in VB.net and it is simple to manipulate the chart type and to input the points in order for the actual graph to be produced and displayed. The chart can exist in its own form and displayed after the other forms that have preceded it.

## Graphing

One method which could be used to display a result to the user would be to plot the data on a graph. This would enable the user to see a visualisation of the potential correlation. The

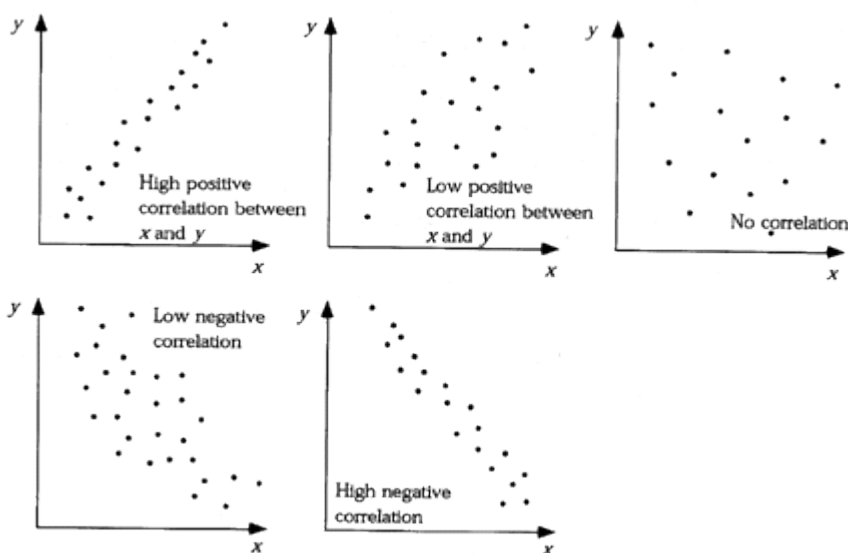


Figure 19

output would potentially look like the graphs below with each of the types of correlation being shown with individual points plotted.

The graphs could either be plotted as a single line or as a collection of points. I believe that a collection of points would be better as if there is no correlation, the points would not join up to

form a line.

A graph can easily be produced in VB.net particularly in forms where it can easily be manipulated to add points and change the chart type. The graph can exist on its own within its own form and the only change that needs to be made from initialisation is the plotting of the points which are input as two lists or arrays with a set of values that can be plotted.

As part of the graphs, the axis will need to be labelled with the correct indicators, country or year depending on which ones are being correlated. These variables are easily accessed in order to be able to change them to the selected indicators and countries. Individual points



could be labelled however this would overcomplicate the appearance of the graph and distract away from the correlation that is being shown.

Along with the visualisation of a graph there will be a brief description of the correlation to go with it. This will be a set of generic sentences that are stored potentially as a list of lists or alternatively in a dictionary or in their own class. The sentences will describe how the graph looks visually and the types of correlation that are shown. Specific values are to be input into spaces so that they are relevant to the specific graph and correlation shown and they will be varied depending on what correlation is being shown.

## Technical Solution

```
Imports System
Imports Microsoft.VisualBasic.FileIO.TextFieldParser
Imports System.Globalization
Imports System.IO
Imports System.Globalization.NumberStyles
Public Class WelcomeForm
    Private Sub WelcomeForm_Load(sender As Object, e As EventArgs) Handles
MyBase.Load
    End Sub
    Private Sub Button1_Click(sender As Object, e As EventArgs) Handles
NextButton.Click
    Try
        If OptionOneSelection.Checked And OptionTwoSelection.Checked = False
And OptionThreeSelection.Checked = False Then
            Option1.Show()
            Me.Close()
        End If
        If OptionTwoSelection.Checked And OptionOneSelection.Checked = False
And OptionThreeSelection.Checked = False Then
            Option2.Show()
            Me.Close()
        End If
        If OptionThreeSelection.Checked And OptionTwoSelection.Checked = False
And OptionOneSelection.Checked = False Then
            Option3.Show()
            Me.Close()
        End If
        Catch ex As Exception
            Console.WriteLine(ex.Message)
            Console.ReadLine()
        End Try
    End Sub
End Class
Public Class Analysis
    Public Indicators As New List(Of Indicator)
    Public Countries As New List(Of Country)
    Public selectedindicator, selectedcountrya, selectedcountryb,
selectedindicatora, selectedindicatorb As String
    Private selectedcountries As New List(Of String)
    Public datasetX, datasetY, finalsetx, finalsety As New List(Of Double)
    Private previous As Integer
    Private datafilelocation, indicatorsfilelocation, countriesfilelocation As
String
    Public Sub New(ByVal fileLocationPath As String)
        Dim lines As New List(Of String)
        Using r As StreamReader = New StreamReader("filelocations.txt")
            Dim line As String
            line = r.ReadLine
            Do While (Not line Is Nothing)
                lines.Add(line)
                line = r.ReadLine
            Loop
        End Using
        Dim linecount As Integer = 0
        For Each item In lines
            If linecount = 1 Then
                Me.datafilelocation = item
```

```

        ElseIf linecount = 2 Then
            Me.countriesfilelocation = item
        ElseIf linecount = 3 Then
            Me.indicatorsfilelocation = item
        End If
        linecount += 1
    Next
End Sub
Public Sub GenerateCountriesList()
    Dim afile As FileIO.TextFieldParser = New
FileIO.TextFieldParser(countriesfilelocation)
    Dim CurrentLine As String()
    Dim firstline As String
    Dim firstlinearray As String()
    firstline = afile.ReadLine()
    firstlinearray = firstline.Split(New Char() {"", "c"})
    afile.TextFieldType = FileIO.FieldType.Delimited
    afile.Delimiters = New String() {"", ""}
    afile.HasFieldsEnclosedInQuotes = True
    Do Until afile.EndOfData = True
        CurrentLine = afile.ReadFields()
        Dim newcountry As New Country
        newcountry.name = CurrentLine(2)
        newcountry.code = CurrentLine(0)
        countries.Add(newcountry)
    Loop
End Sub
Public Sub GenerateIndicatorsList()
    Dim afile As FileIO.TextFieldParser = New
FileIO.TextFieldParser(indicatorsfilelocation)
    Dim CurrentLine As String()
    Dim firstline As String
    Dim firstlinearray As String()
    firstline = afile.ReadLine()
    firstlinearray = firstline.Split(New Char() {"", "c"})
    afile.TextFieldType = FileIO.FieldType.Delimited
    afile.Delimiters = New String() {"", ""}
    afile.HasFieldsEnclosedInQuotes = True
    Do Until afile.EndOfData = True
        CurrentLine = afile.ReadFields()
        Dim newindicator As New Indicator
        newindicator.name = CurrentLine(2)
        newindicator.code = CurrentLine(0)
        Indicators.Add(newindicator)
    Loop
End Sub
Public Sub CheckLists(ByVal checklist As CheckedListBox, ByVal maximum As
Integer)
    Try
        Dim count As Integer = 0
        For Each chkbox In checklist.Items
            If chkbox.selected Then
                count += 1
            End If
            If count > maximum Then
                Exit For
            End If
        Next
    Catch ex As Exception

```

```

        Console.WriteLine(ex.Message)
        Console.ReadLine()
    End Try
End Sub
Public Sub GenerateLists()
    GenerateCountriesList()
    GenerateIndicatorsList()
End Sub
Public Function GenerateData(ByVal indicator As String, ByVal country As
String)
    Dim afile As FileIO.TextFieldParser = New
FileIO.TextFieldParser(datafilelocation)
    Dim CurrentRecord As String()
    Dim firstline As String
    Dim firstlinearray As String()
    Dim values As New List(Of Double)
    firstline = afile.ReadLine()
    firstlinearray = firstline.Split(New Char() {"", "c"})
    afile.TextFieldType = FileIO.FieldType.Delimited
    afile.Delimiters = New String() {"", ""}
    afile.HasFieldsEnclosedInQuotes = True
    Dim found As Boolean = False
    Do Until afile.EndOfData = True Or found = True
        Try
            CurrentRecord = afile.ReadFields()
            If CurrentRecord(0) = country And CurrentRecord(2) = indicator
Then
                For x = 0 To 55
                    Dim newvalue As Double
                    If Double.TryParse(CurrentRecord(x + 5),
NumberStyles.AllowExponent Or NumberStyles.AllowDecimalPoint) Then
                        newvalue = Double.Parse(CurrentRecord(x + 5),
NumberStyles.AllowExponent Or NumberStyles.AllowDecimalPoint)
                    Else
                        newvalue = -99.99
                    End If
                    values.Add(newvalue)
                Next
                found = True
            End If
        Catch ex As Exception
            Console.WriteLine(ex.Message)
            Console.ReadLine()
            Continue Do
        End Try
    Loop
    Return values
End Function
Public Sub RemoveInvalid(ByRef datasetx As List(Of Double), ByRef datasety As
List(Of Double), ByRef finalsetx As List(Of Double), ByRef finalsety As List(Of
Double))
    For x = 0 To (datasetx.Count() - 1)
        If datasetx(x) <> (-99.99) And datasety(x) <> -99.99 Then
            finalsetx.Add(datasetx(x))
            finalsety.Add(datasety(x))
        End If
    Next
End Sub
End Class

```

```

Public Class Indicator
    Inherits Values
End Class
Public Class Country
    Inherits Values
End Class
Public Class Values
    Public name As String
    Public code As String
End Class
Public Class Correlations
    Function PearsonsCorrelation(ByRef firstset As List(Of Double), ByRef
secondset As List(Of Double))
        Dim sxx, sxy, syy, r, sx, sy, sxsq, sysq, ssxy As Double
        For x = 0 To firstset.Count() - 1
            sx += firstset(x)
            sxsq += firstset(x) * firstset(x)
            ssxy += firstset(x) * secondset(x)
        Next
        For y = 0 To secondset.Count() - 1
            sy += secondset(y)
            sysq += secondset(y) * secondset(y)
        Next
        sxy = ssxy - ((sx * sy) / firstset.Count())
        sxx = (sxsq - ((sx * sx) / firstset.Count()))
        syy = (sysq - ((sy * sy) / secondset.Count()))
        r = (sxy / (Math.Sqrt((sxx * syy))))
        Return r
    End Function
    Function SpearmansCorrelation(ByRef firstset As List(Of Double), ByRef secondset As
List(Of Double))
        Dim firstcopy, secondcopy As New List(Of Double)
        Dim firstrank, secondrank As New List(Of Integer)
        Dim d As New List(Of Double)
        Dim dsq, p, diff As Double
        firstcopy = firstset
        secondcopy = secondset
        Quicksort(firstcopy, 0, firstset.Count() - 1)
        Quicksort(secondcopy, 0, secondset.Count() - 1)
        For x = 0 To firstset.Count() - 1
            For y = 0 To firstcopy.Count() - 1
                If firstset(x) = firstcopy(y) Then
                    firstrank.Add(y)
                End If
            Next
        Next
        For x = 0 To secondset.Count() - 1
            For y = 0 To secondcopy.Count() - 1
                If secondset(x) = secondcopy(y) Then
                    secondrank.Add(y)
                End If
            Next
        Next
        For x = 0 To firstset.Count() - 1
            If firstrank(x) >= secondrank(x) Then
                diff = firstrank(x) - secondrank(x)
            Else
                diff = secondrank(x) - firstrank(x)
            End If
            d.Add(diff)
        Next
    End Function
End Class

```

```

    For x = 0 To d.Count() - 1
        dsq += (d(x) * d(x))
    Next
    p = 1 - ((6 * dsq) / (d.Count() * (d.Count() - 1)))
    Return p
End Function
Public Sub Quicksort(ByVal list As List(Of Double), ByVal min As Integer, ByVal
max As Integer)
    Dim random_number As New Random
    Dim med_value As Double
    Dim hi As Double
    Dim lo As Double
    Dim i As Double
    If min >= max Then Exit Sub
    i = random_number.Next(min, max + 1)
    med_value = list(i)
    list(i) = list(min)
    lo = min
    hi = max
    Do
        Do While list(hi) >= med_value
            hi = hi - 1
            If hi <= lo Then Exit Do
        Loop
        If hi <= lo Then
            list(lo) = med_value
            Exit Do
        End If
        list(lo) = list(hi)
        lo = lo + 1
        Do While list(lo) < med_value
            lo = lo + 1
            If lo >= hi Then Exit Do
        Loop
        If lo >= hi Then
            lo = hi
            list(hi) = med_value
            Exit Do
        End If
        list(hi) = list(lo)
    Loop
    Quicksort(list, min, lo - 1)
    Quicksort(list, lo + 1, max)
End Sub
End Class

Public Class Option1
    Dim optionone As New Analysis("filelocations.txt")
    Public Sub Option1_Load(sender As Object, e As EventArgs) Handles MyBase.Load
        optionone = New Analysis("filelocations.txt")
        LoadLists()
    End Sub
    Public Sub LoadLists()
        optionone.GenerateLists()
        For Each item In optionone.countries
            OpOneCountries.Items.Add(item.name)
        Next
        OpOneCountries.Update()
        For Each item In optionone.Indicators
            OpOneIndicators.Items.Add(item.name)
        Next
        OpOneIndicators.Update()
    End Sub
End Class

```

```

    End Sub
    Private Sub Button1_Click(sender As Object, e As EventArgs) Handles
NextButton.Click
        optionone.CheckLists(OpOneCountries, 1)
        optionone.CheckLists(OpOneIndicators, 2)
        Dim selectedindicators As New List(Of String)
        Dim selectedcountry As String
        selectedcountry = OpOneCountries.SelectedItem
        For Each item In OpOneIndicators.CheckedItems
            selectedindicators.Add(item)
        Next
        optionone.selectedcountrya = selectedcountry
        optionone.selectedcountryb = selectedcountry
        optionone.selectedindicatora = selectedindicators(0)
        optionone.selectedindicatorb = selectedindicators(1)
        optionone.datasetX = optionone.GenerateData(selectedindicators(0),
selectedcountry)
        optionone.datasetY = optionone.GenerateData(selectedindicators(1),
selectedcountry)
        optionone.RemoveInvalid(optionone.datasetX, optionone.datasetY,
optionone.finalsetx, optionone.finalsety)
        Dim oponegraphs As New graphs(optionone)
        oponegraphs.Show()
        Me.Close()
    End Sub
End Class

Public Class Option2
    Dim optiontwo As New Analysis("filelocations.txt")
    Private Sub Form7_Load(sender As Object, e As EventArgs) Handles MyBase.Load
        optiontwo = New Analysis("filelocations.txt")
        optiontwo.GenerateLists()
        For Each item In optiontwo.Countries
            OpTwoCountries.Items.Add(item.name)
        Next
        For Each item In optiontwo.Indicators
            OpTwoIndicators.Items.Add(item.name)
        Next
    End Sub
    Public Sub Button1_Click(sender As Object, e As EventArgs) Handles
NextButton.Click
        optiontwo.CheckLists(OpTwoCountries, 2)
        optiontwo.CheckLists(OpTwoIndicators, 1)
        Dim selectedindicator As String
        Dim selectedcountries As New List(Of String)
        selectedindicator = OpTwoIndicators.SelectedItem
        For Each item In OpTwoCountries.CheckedItems
            selectedcountries.Add(item)
        Next
        optiontwo.selectedcountrya = selectedcountries(0)
        optiontwo.selectedcountryb = selectedcountries(1)
        optiontwo.selectedindicatora = selectedindicator
        optiontwo.selectedindicatorb = selectedindicator
        optiontwo.datasetX = optiontwo.GenerateData(selectedindicator,
selectedcountries(0))
        optiontwo.datasetY = optiontwo.GenerateData(selectedindicator,
selectedcountries(1))
        optiontwo.RemoveInvalid(optiontwo.datasetX, optiontwo.datasetY,
optiontwo.finalsetx, optiontwo.finalsety)

```

```

        Dim optwographs As New graphs(optiontwo)
        optwographs.Show()
        Me.Close()
    End Sub
End Class

Public Class Option3
    Dim optionthree As New Analysis("filelocations.txt")
    Private Sub Form8_Load(sender As Object, e As EventArgs) Handles MyBase.Load
        optionthree = New Analysis("filelocations.txt")
        optionthree.GenerateLists()
        For Each item In optionthree.countries
            OpThreeCountries1.Items.Add(item.name)
            OpThreeCountries2.Items.Add(item.name)
        Next
        For Each item In optionthree.Indicators
            OpThreeIndicators1.Items.Add(item.name)
            OpThreeIndicators2.Items.Add(item.name)
        Next
    End Sub
    Private Sub Button1_Click(sender As Object, e As EventArgs) Handles
NextButton.Click
        optionthree.CheckLists(OpThreeCountries1, 1)
        optionthree.CheckLists(OpThreeIndicators1, 1)
        optionthree.CheckLists(OpThreeCountries2, 1)
        optionthree.CheckLists(OpThreeIndicators2, 1)
        Dim selectedindicatora, selectedcountrya, selectedindicatorb,
selectedcountryb As String
        selectedindicatora = OpThreeIndicators1.SelectedItem
        selectedcountrya = OpThreeCountries1.SelectedItem
        selectedcountryb = OpThreeCountries2.SelectedItem
        selectedindicatorb = OpThreeIndicators2.SelectedItem
        optionthree.datasetX = optionthree.GenerateData(selectedindicatora,
selectedcountrya)
        optionthree.datasetY = optionthree.GenerateData(selectedindicatorb,
selectedcountryb)
        optionthree.RemoveInvalid(optionthree.datasetX, optionthree.datasetY,
optionthree.finalsetx, optionthree.finalsety)
        Dim opthreegraphs As New graphs(optionthree)
        opthreegraphs.Show()
        Me.Close()
    End Sub
End Class

Public Class graphs
    Dim plotsetx, plotsety As New List(Of Double)
    Dim countrya, countryb, indicatora, indicatorb As String
    Public Sub graphs_Load(sender As Object, e As EventArgs) Handles MyBase.Load
        GraphOutput.Series("Series1").Points.DataBindXY(plotsetx, plotsety)
        ChangeText(plotsetx, plotsety)
        GraphOutput.ChartAreas("ChartArea1").AxisX.Title = countrya + " & " +
indicatora
        GraphOutput.ChartAreas("ChartArea1").AxisY.Title = countryb + " & " +
indicatorb
        GraphOutput.Update()
    End Sub
    Public Sub New(ByRef optiondata As Analysis)
        InitializeComponent()
        plotsetx = optiondata.finalsetx
        plotsety = optiondata.finalsety
    End Sub
End Class

```



```

countrya = optiondata.selectedcountrya
countryb = optiondata.selectedcountryb
indicatora = optiondata.selectedindicatora
indicatorb = optiondata.selectedindicatorb
End Sub
Public Sub ChangeText(ByRef setx As List(Of Double), ByRef sety As List(Of
Double))
Dim variation As New Correlations
Dim answer As Double
Dim strongpositive, strongnegative, unrelated As String
strongpositive = "There is a strong positive correlation."
strongnegative = "There is a strong negative correlation."
unrelated = "There is no clear correlation here."
answer = variation.PearsonsCorrelation(setx, sety)
If answer < 0.25 And answer > -0.25 Then
GraphDescription.Text = unrelated
ElseIf answer >= 0.25 Then
GraphDescription.Text = strongpositive
ElseIf answer <= -0.25 Then
GraphDescription.Text = strongnegative
End If
GraphDescription.Text = GraphDescription.Text + "The result of Pearson's
product moment correlation is " + Str(answer)
End Sub
End Class

```

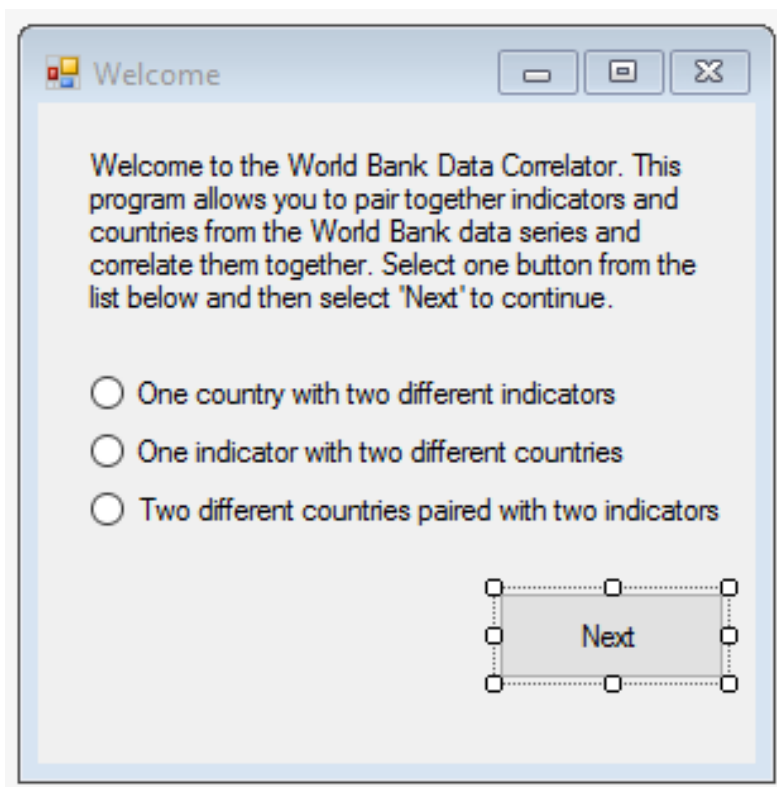


Figure 20

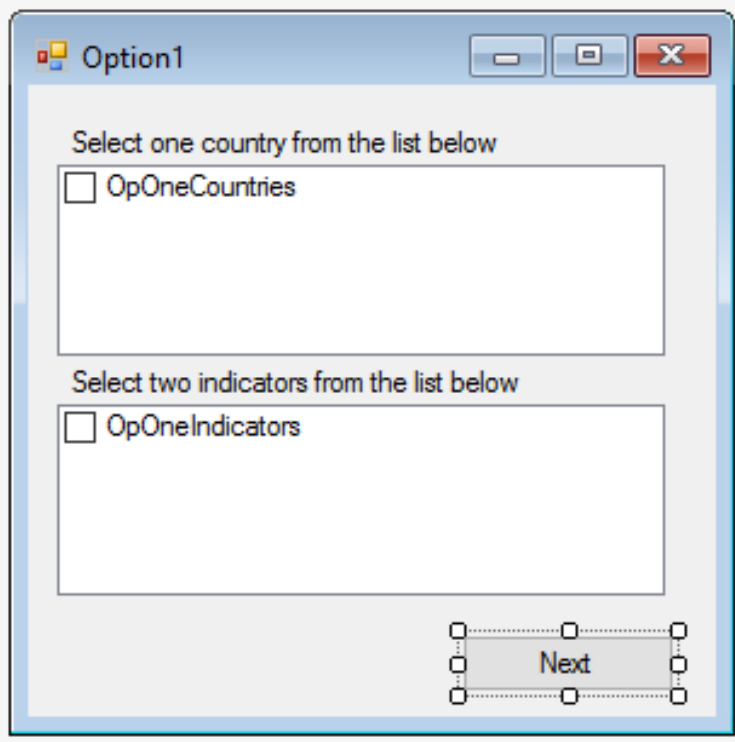


Figure 21

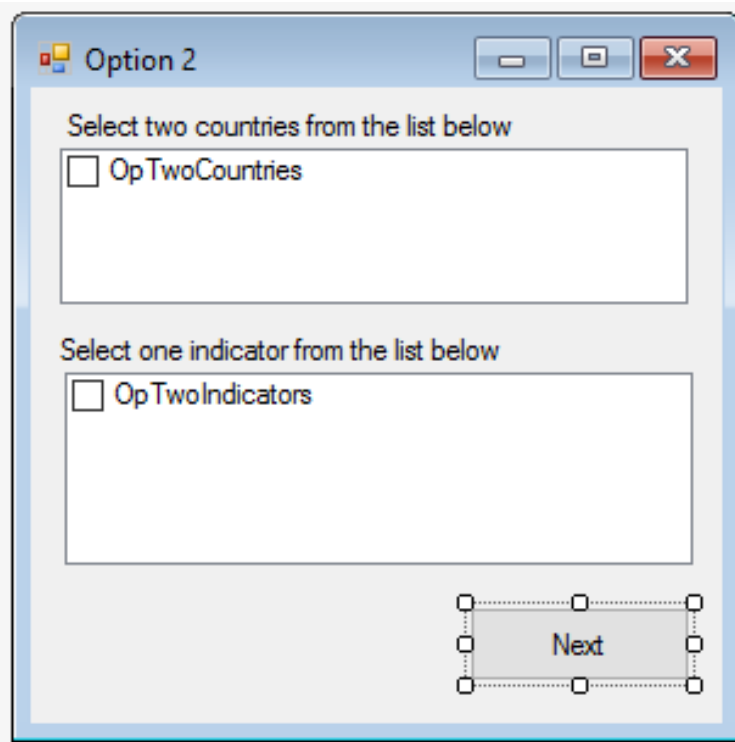


Figure 22

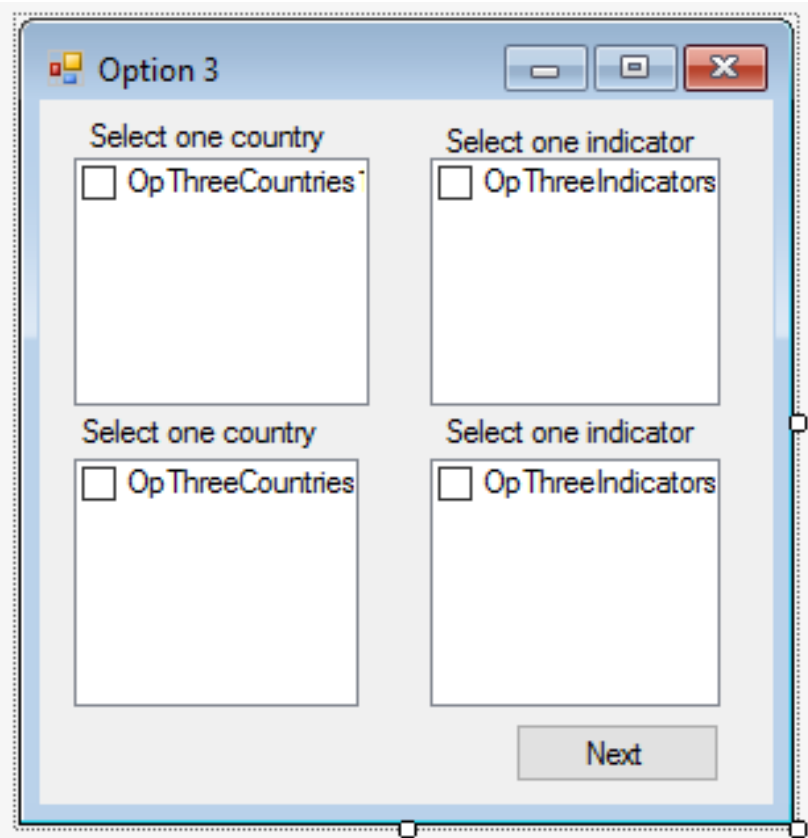


Figure 23

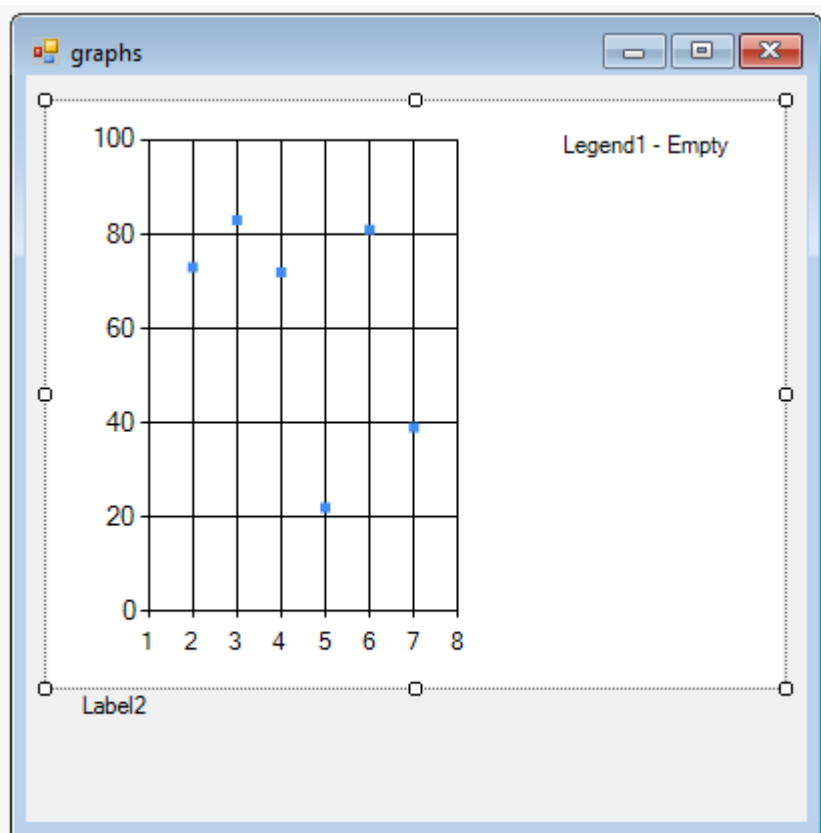


Figure 24

```
filelocations - Notepad
File Edit Format View Help
The first file holds the data, the second file holds the countries and the third file holds the indicators
C:\Users\New\Downloads\WDI_csv\WDI_data.csv
C:\Users\New\Downloads\WDI_csv\WDI_Country.csv
C:\Users\New\Downloads\WDI_csv\WDI_Series.csv
```

Figure 25

Figure 22 is the form design for the Welcome Form which opens when the program is initially run.

Figure 23 is the form design for option one which is the correlation of one country with two different indicators.

Figure 24 is the form design for option two which is the correlation of one indicator with two different countries.

Figure 25 is the form design for option three which is two different countries correlated against two different indicators.

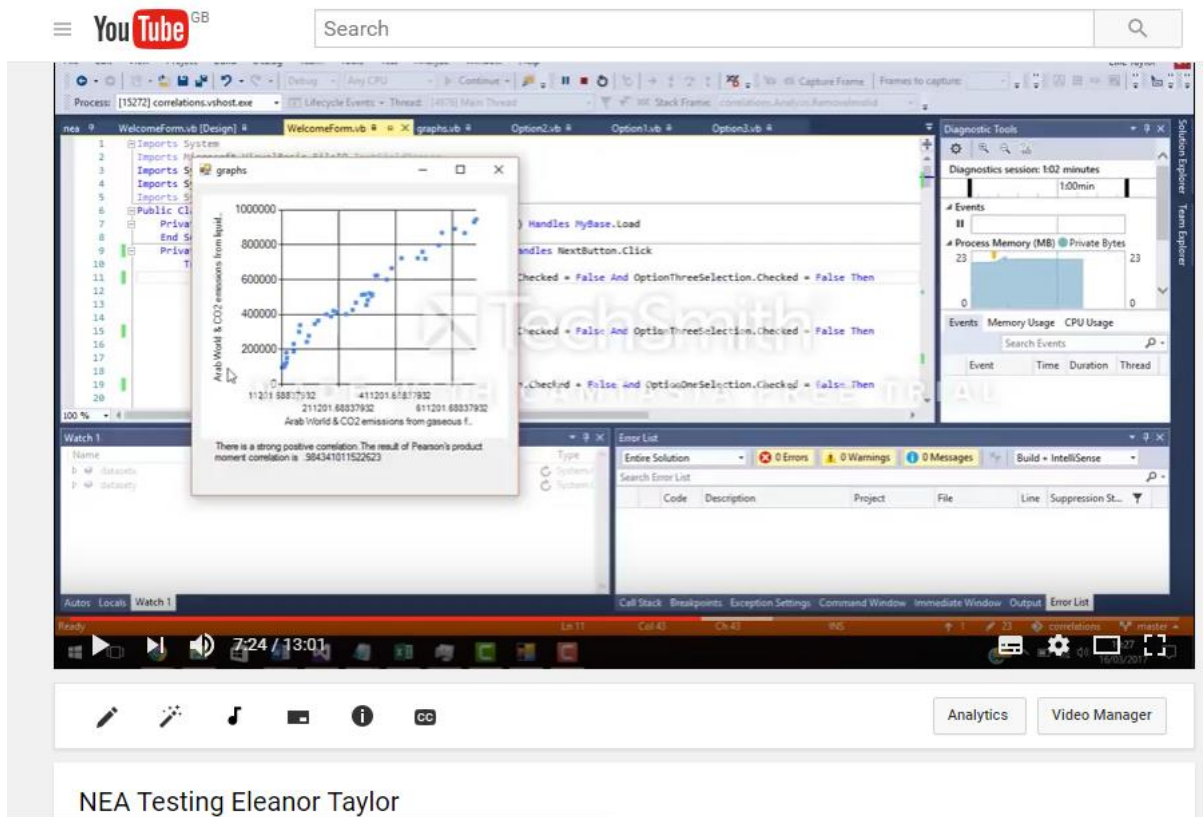
Figure 26 is the form design for the graph and report which is output to the user which is ultimately is displayed following option one, two or three depending on which one has been selected in the welcome form.

Figure 27 is the text file which holds the locations of the other files which are used in retrieving data.

## Testing

Testing for this project is in a video and is available to view online at this address:

<https://youtu.be/i0mEJVpFYgI>



NEA Testing Eleanor Taylor

Test number	Description	Expected outcome	Time
1	Data file locations are stored in a separate text file	The file contains four lines of text. The first describing what each line is and then three lines with three different file locations which correspond with the description in the first line.	2.21
2	When the program is run, an initial welcome form is displayed to the user	The initial form is displayed quickly with no errors	5.32
3	The user is only able to select one of the initial options	Only one option can be selected and the code doesn't break	6.08
4	The first option is selected and the second form is displayed while the original form closes	The open form closes and the next one opens in its place	6.22
5	Option one's first checked boxes lists contains a full list of countries	The full list of countries from the data file are in the list	6.22
6	Option one's second checked boxes lists contains a full list of indicators	The full list of indicators from the series data file are in the list	6.22

7	When the next button is pressed, the graphs form opens	The original form opens and the graphs form opens in its place	7.16
8	The graph is displayed with accurate points plotted	All the points are displayed on the graph in the correct place	7.16
9	The axis' are labelled correctly with the country and indicators selected	The axis labels are Arab World & CO2 emissions from gaseous fuel consumption and Arab World & CO2 emissions from liquid fuel consumption	7.16
10	A description of the correlation coefficient result is displayed	Text is displayed under the graph which accurately describes the correlation shown on the graph	7.16
11	Pearson's moment correlation coefficient is calculated accurately	The correlation calculation output by the program matches the result from the excel file with the inbuilt function	8.28
12	The second option is selected and the third form is displayed while the original form closes	The open form closes and the next one opens in its place	9.24
13	Option two's first checked boxes lists contains a full list of countries	The full list of countries from the data file are in the list	9.24
14	Option two's second checked boxes lists contains a full list of indicators	The full list of indicators from the series data file are in the list	9.24
15	When the next button is pressed, the graphs form opens	The original form closes and the graphs form opens in its place	10.33
16	The graph is displayed with accurate points plotted	All the points are displayed on the graph in the correct place	10.33
17	The axis' are labelled correctly with the country and indicators selected	The axis labels are with Arab World & CO2 emissions from gaseous fuel consumption and Euro Areas & CO2 emissions from gaseous fuel consumption	10.33
18	A description of the correlation coefficient result is displayed	Text is displayed under the graph which accurately describes the correlation shown on the graph	10.33
19	Correlation calculation is accurately displayed as part of text	The correlation calculations result is part of the description underneath the graph	10.33
20	The third option is selected and the fourth is displayed while the original form closes	The open form closes and the next one opens in its place	11.16
21	Option three's left two checked boxes lists contains a full list of countries	The full list of countries from the data file are in the list	11.16
22	Option three's right two checked boxes lists contains a full list of indicators	The full list of indicators from the series data file are in the list	11.16
23	When the next button is pressed, the graphs form opens	The original form closes and the graphs form opens in its place	12.02

24	The graph is displayed with accurate points plotted	All the points are displayed on the graph in the correct place	12.02
25	A description of the correlation coefficient result is displayed	Text is displayed under the graph which accurately describes the correlation shown on the graph	12.02
26	Correlation calculation is accurately displayed as part of text	The correlation calculations result is part of the description underneath the graph	12.02

## Evaluation

The purpose of this investigation was to identify whether it is possible to show correlation between two world development indicators computationally. The objective of this investigation has been met as the correlation between two indicators can be plotted on a graph and have correlation calculations carried out on it. Although the main objective has been met there are many ways in which this project could be extended further.

### Evaluation against objectives

Objective	Met?	Comments	Further development
Text file holds the memory locations of data, series and country csv files	Yes	Figure 27 shows a text file containing the locations of three text files and a description of what is contained within the file	
File locations text file is iterated over to match corresponding files	Yes	All of the other files are accessed when the program is run therefore this file must have been iterated over for this to occur	
Countries file is accessed	Yes	A full list of countries is output therefore the countries file must have been accessed	
File containing countries information is parsed	Yes	All of the countries appear in the correct format, i.e. without commas and speech marks, so the data has been parsed correctly	
Each indicator is extracted from the file and input into an abstract data type	Yes	A full list of indicators is output therefore the indicators have been extracted from the file and stored in the abstract data types before being output	
User is shown a list of options for what can be correlated	Yes	The initial form in figure 22 shows that there are three options for types of correlation that are able to be carried out therefore the objective has been met	
The user must only be able to select one of these options	Yes	As shown in test 3, only one option is able to be selected	
User is shown a list of every country and every indicator in check lists	Yes	There is a full list of countries and indicators that appear in each of the option forms that follow the initial welcome form	
User is able to select a country/countries and an indicator/indicators to correlate together	Yes	There are checkedlistboxes which enable the countries and indicators to be selected	To develop this further the user could select one pairing first and then the program removes all of the option which



			there is no data for and therefore would not produce any correlations, however this would take a long time at run time
For option one the user is able to select one country and two indicators only	Yes	The CheckLists procedure is used for each of the options and enables only the specified number of boxes to be checked before continuing	
For option two the user is able to select two countries and one indicator only	Yes	The CheckLists procedure is used for each of the options and enables only the specified number of boxes to be checked before continuing	
The data file is opened from the memory location held in the files location text file	Yes	The data is plotted on the graph as shown in text 8 and therefore the file must have been accessed for these data points to have been found	
The file is iterated over until the selected country and indicator pairing is found, this will occur twice to obtain both data sets	Yes	The correct data is plotted on the graphs as shown in test 8 which shows that the data has been found within the file and that the objective has been met	
The values for each of the years are stored in a list for that specified data set	Yes	When the file is iterated over and the values are extracted they are stored in lists	
The values are converted from an exponent into the correct double data type	Yes	The labels on the axis are not in exponent form and therefore they must have been converted to the correct data type or else they would not have plotted in the correct way	
The data sets are compared and empty values are removed so that only pairs of data that exist are kept in the data set	Yes	There are not always 56 points on the graphs as shown in testing which shows that the invalid points must have been removed	
Correlation calculations stored as individual functions	Yes	Spearman'sCorrelation and Pearson'sCorrelation exist as two separate functions within the Correlations class	The user could be asked whether or not they would like to view the correlation calculation results and which of them they would like to view

Pearson's product-moment correlation coefficient is able to be calculated	Yes	On the graphs form that is shown at the end of the series, in the description is a value which is the result of Pearson's correlation as shown in test 11	
The values are input into the correct part of the formula	Yes	The result of the correlation carried out in testing matches the result of the same calculation carried out in a separate excel file where the function is built in	
Calculation gives out a valid answer	Yes	Test 11 shows that this objective has been met as the result which is outputted is accurate	
Spearman's rank-order correlation is able to be calculated	Partial	Spearman's rank-order correlation has been included as one of the correlation functions however it does not produce an accurate value	
The values are input into the correct part of the formula	Partial	It is unclear whether the error at the moment lies within the quick sort algorithm or in the Spearman's rank algorithm however the values are not input in the right order and therefore this objective is only partially met	
Calculation gives out a valid answer	No	Although the calculation does produce a value it does not match the value that is produced from a verified Spearman's rank order calculator	The code should be developed so that the calculation returns the correct value
Data input from the data sets from extraction of csv file into formula	Yes	The values are input into the formula although the value that is produced is not correct	
Calculates strength of correlation	Partial	The value which is output is incorrect however a value is output	This should be developed so that the values are accurate
Numerical values from each calculation are stored for future use in the final report	Yes	The values from both Pearson's and Spearman's are shown on the output	
Data sets extracted from the file are plotted on the graph	Yes	A graph is output with valid points plotted on it as shown in test 8	
Graph is plotted accurately using the coordinates	Yes	This objective has been met as is shown in test 8	
Graph is displayed to the user with fully labelled	Yes	The labels on the axis are accurate as shown in test 9	

axis with the countries and indicators			
Description of the graph is presented along with the graph	Yes	There are labels containing text that are underneath the graph and therefore the objective has been met	The description could be developed further so that any outliers are identified and it could also be described in more detail
Key features of the graph and correlation calculation are noted	Yes	The strength of the correlation has been described and the values of the correlation calculations are within the text	
Gaps filled in for specific data	Yes	The values appear in the description underneath the graphs and therefore	
Sentences for each feature are output to the user	Partial	Although there is a description that is shown depending on the result of the Pearson's moment correlation, the description is limited and only describes the basic correlation.	This could be further developed so that a regression line is calculated and then outliers and anomalies that don't fit the pattern are found and either excluded or made aware to the user. This would make the correlation coefficients more reliable.

### Feedback from Economics Department

The program was shown to Courtney Anderson, Head of Economics, to see how useful it can be to the department and how it could be incorporated into their teaching.

Courtney's feedback was incredibly positive. He felt that the program would be incredibly useful in proving many economic theories to students in the classroom because of the visual display of the outcome. In particular, the data is useful for proving relationships between certain economic policies and economic development. Courtney felt that this program would allow students the gain a better understanding of how certain changes in an economy can impact the overall performance of the economy. From an economist's perspective, Courtney felt that this program showed a great visual model of many economic principles which exist primarily as theories and are difficult to prove verbally. The data from World Bank is a valuable tool to the economics department however Courtney wishes to integrate the data available more in lessons as it would contribute to application marks in economics examinations.

To develop this program further, Courtney believed that there would need to be cosmetic changes in making the interface more friendly and easy for students to use. He suggested the use of a global map which could be clicked on to narrow down the selection of countries. He also suggested breaking down the indicators into subcategories so that there aren't so many to scroll through to find the one that is being searched for. Alternatively, a search bar could be used to save time as well.

### **Further development**

This project can be developed in a number of ways to take it further than it has already been developed. First of all, it could be developed so that the entries in the text file with no data are removed so that it is not possible to select them to be correlated together. At the moment, this would currently show a blank graph and a correlation coefficient of 0. It could also be extended so that the entire file is iterated over and all the correlation coefficients are calculated and those within given ranges are removed perhaps as the user wishes through an input. They could input whether they want to view a strong positive or negative correlation or whether or not they want to see a proven correlation at all. Alternatively, the pairings which would only show a limited number of points could be removed as they are unlikely to show a strong correlation between the two sets of data.

There is the potential that a correlation could be non-linear and so the program could be developed further so that it is able to identify and non-linear correlation and it is able to be described as well.