Supporting big data analytics in computational biology and public health

A thesis submitted by

YAP YING HUI PRISCILLA

in partial fulfilment for the

Degree of Bachelor of Science with Honours

in

Computational Biology

Supervisors:

Professor CHOI KWOK PUI,

Professor WONG LIMSOON

Semester 1, 2013/2014

Acknowledgements

I would like to take this opportunity to thank my project supervisor Prof Choi from Faculty of Science and project co-supervisor Prof Wong, from School of computing. It is my honour to work with both of them; I would like to thank Prof Wong especially for coming up with such a fantastic open idea project idea that allows me much free space to think of how to handle the Data. Even though this project do not handle bio-Data, like genes or DNA related, however, it created a platform for me to use what was taught into analysing data. In fact data analysis is the core skill of what Computational Biology students are supposed to have. While working on this project, it allows me to re-visit many modules that I took up during my university coursework, mainly the Database Design, Linear Regression and Time Series Analysis, Data Structure Algorithms, Linear Programming and Optimization and Human Computer Interaction.

I would also like to thank other online forum users, from Stackoverflow and MSDN, for their contribution, to enable me to solve quite a number of technical questions, mainly the administrative part like setting up the MS SQL 2012 to MS VS 2012.

Abstract

The objective of this project is to develop a user friendly web application platform to perform analyse work on data from the internet so as those data could be make useful. The work was broken down into 3 portions, the data, the technical aspect of this project and the approach on handling the analysing work, which would be elaborated at the second section – materials and methods. This project placed more emphasis on the user friendly graphical than real statistical analysis work; as many prior statistical work has shown that many attributes are not correlated, which is shown in the result section. Also, it is not possible to predict all sort of statistical analysis that user will like to use, hence, user are able to download the data from the web application. At the last few weeks before submission, due to many technical issues that resulted in compatibility, so there are slight change in the software used within as short term solution. However, a set of proof of concept was drawn out at the discussion section.

I. Introduction

This is the big data age. Different industry of people defined it differently. After digesting what big data was defined by James Kobielus (2013) and James Taylor (2011), on Wikipedia (2013) and International Business Machines (IBM) website, here is how I would define big data – is the term used when referring to a collection of data sets from traditional and digital sources that was made available for on-going discovery and analysis, and is usually large in quantity and complex in format, that posed a challenge to current on-hand database management tools or traditional data processing applications to process it. Data, as share by Philip (2013), come in two flavours, structured like data stored in data warehouse with proper indexing for easy retrieval and unstructured, like data from social networks in the format of text, still images, motion graphic, video, sound or music and graphs. And according to Margaret (2012), out of these data, 80% of which are unstructured. It is also not easy to measure the exact volume of data stored electronically, however, according to Tom (2012) stated in his book, International Data Corporation (IDC) made a safe estimation of the data volume in 2006 of about 0.18 zettabytes*, and is forecasting a tenfold growth by 2011 to 1.8 zettabytes.

* zettabytes: is a quantity of 10^{21} bytes or equivalently to one thousand exabyte, one million petabytes, or one billion terabytes.

The flood of data comes from many sources, the internet, for instance is one true place that houses this vast amount of information. Sources like New York Stock Exchange generated about one terabytes of new trade data per day, Facebook a social network hosts approximately billions photos, now with the bought over of Instagram, could easily

take up petabytes of storage per cluster of users. There are also other cloud data storage providers like Amazon Web Service (AWS), IDC, IBM ... housing vast quantity of data for different industries. With this great deal of data made readily available increasing each year, it will be benefiting to the society if these data could aid in driving humanity profit. Also, Armonk (2013) responded that organisation no longer merely manage their own data, in fact the success in the future; is dictated to the large extend of by their ability to extract data from other organisation.

A little background of IBM and IDC and what they did; their direction on how to handle big data, as it did contributed to the drafting of the deployment plan used in this project. IBM has been the leading technological company with innovative approach. IBM is helping clients harness this Big Data to uncover valuable insights, and transform their business. Though AWS started off with by providing cloud storage services as stated by Tom (2012), it is now IBM the one leading in cloud storage and computing technologies and in coupled with business analysing tools.

Similarly to what IBM has done, IDC provides cloud storage services then draft up a set of methodology of analysis upon the request of their clients that provides useful insights. IDC in fact is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. As described on IDC (2013), it is a wholly-owned subsidiary of International Data Group (IDG), the world's leading technology media, events and Research Company. IDC has an Insights Research team which works with different industries to provide industry-focused analyses and advice for effective technology decision.

However, as mentioned, these data are near beyond the traditional way of processing or even analyzing, especially those unstructured data in format of video sound and picture. Looking at the complexity nature of these data, numeric data seems to be the easier to handle and available, where plentifully of websites provide tables and tables of figures and numbers. These numbers are seems just mere work of display unless it is converted into a more practical usage, example researchers could have better predict their direction of research to cater to the group of aging population if they could have done some trending work on the long available aging statistics earlier; and not only start researching on better aging healthcare upon the government policy. It is delightful to discover while researching work for this project, many companies and organization like IDC and IBM have went on large scale work in generating profit from tackling the issue of big data and their analyzing work. However, those available analyzing tools are usually business intelligence by nature, even if IDC do did work with healthcare industry, they are after all more tuned to business and profit. It would be good if only they could come up with more healthcare research related tool instead.

Thus, that derived this project – Big Data Analytics. The project is to create a web-based platform that allows users to easily perform simple analysing work on data extracted from the internet. For example, what are the pressing issue (poor healthcare availability as shown from the statistics) faced by the world and which countries need the attention most (high health-related illness, yet reflected low health expenditure, could it be due to low GDP), so the healthcare research could fine tune to solve the issue (cheaper treatment made available by mass production of certain therapy to bring down the cost).

There are four main issues when handling big data are how and where to store them, then what data is available, which part of the data is useful and how to use them.

So here is the definition of big data analytics with respect to this project as discussed during the mid-term presentation. It is the process of examining large amounts of data of a variety of types, to uncover hidden patterns, unknown correlations and other useful information to provide competitive advantages to help scientists make better decisions. This project works mainly on huge volumes of data that was left untapped by conventional business intelligence (BI) programs. However, due to many factors that will be explained in the next section, Materials and Methods, the choice of data chosen to work with are only partial data from the Central Intelligence Agency (CIA) world factbook, similarly for the choice of software used.

As the title of this project specified, to support, meaning it has to consider the user experience, hence more emphasis will be placed on the graphical effect and minimal options on the type of statistical work available for the user to choose. The prior work to confirm the final set of statistical ability that this project could offer has many of its limitations being justified at the Material and Methods sections. Also, it is not possible to predict all sort of statistical analysis that user will like to use, hence, user are able to download the data from the web application, as a newly add-ons feature after discussion with Prof Choi. Timeline of the work distribution was illustrated at table 1.1.

At the last few weeks before submission, due to many licensing problem that resulted in change of software, set the project to run into compatibility issue, so there are a decision

on the slight change in the software used within as short term solution. However, a set of proof of concept was drawn out at the future development, a subsection of the discussion section, and partial used and unused prep code was made available under the appendix for future reproduce purposes.

Before project description submission stage	
Brief Project:	Create Website
Data advised to use:	WHO, CIA
project planning stage	
Expectation draft:	Website, Statistical analysing list
Work done:	Data Selection; only choose from CIA source Database design, Website draft
project coding stage	
Data:	Data cleaning on Excel Run the data through the Statistical analysing list
Website:	Website design coding, Colour scheme selection, Website functionality coding
Add-ons:	Interactive features on website; mouseover

Table 1.1. Breakdown on the work timeline.

II. Materials & Methods

The three parts of approaching the project will be discussed; with each portion, the procedures will be listed in generic details. The data, the technical aspect of this project and the methods on handling the analysing work.

A. The Data

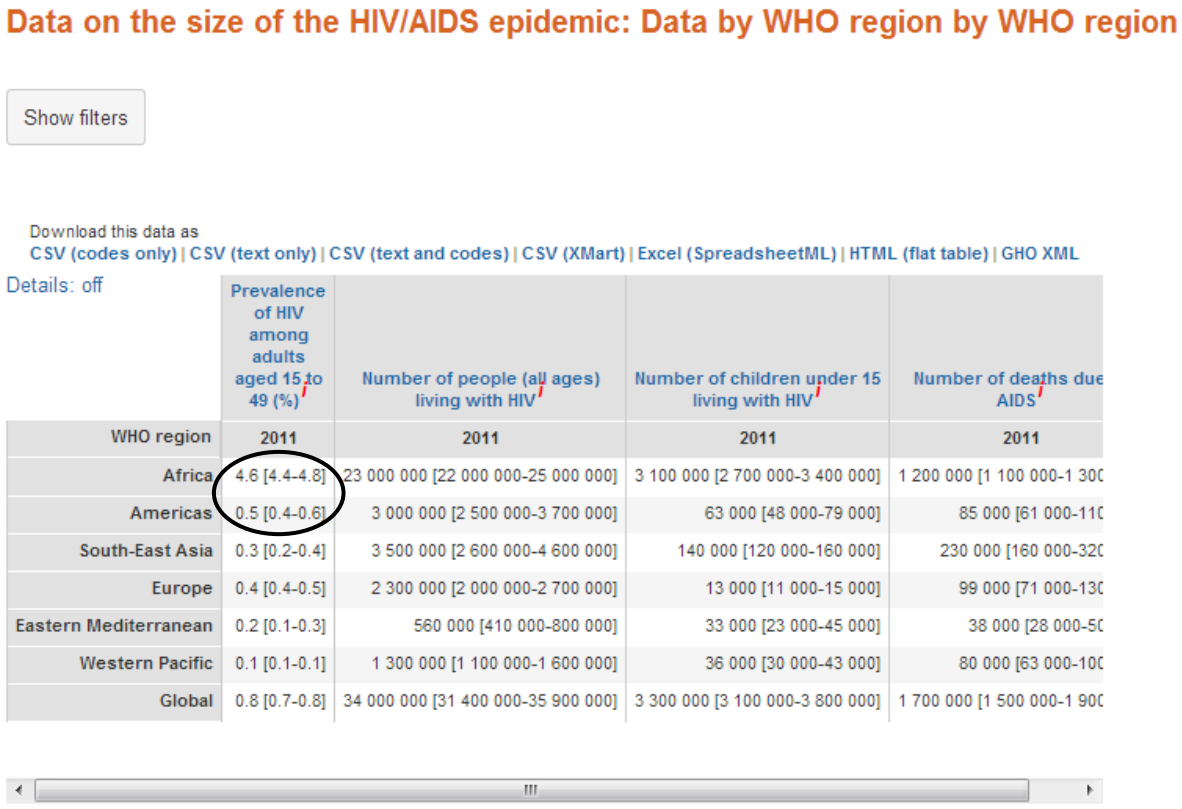
: Data used

First and foremost, is to discuss the choice of data and approach of using it. During the initial project discussion stage, data from Central Intelligence Agency (CIA) world factbook and World Health Organisation (WHO) was advised to be taken into consideration. At the studying stage of this project, it was to look at the data from both CIA and WHO website and draw a conclusion on the database design.

However, the display of the data on WHO website, as shown in figure 2.1 and figure 2.2, clearly shows the wide variant display. Let me explain what it means by wide variant display of data, it means that data between two rows do not have similar format or value of excess information. For example, taking the value in figure 2.1 as circled, under "Prevalence of HIV among adults aged 15 to 49 (%)" row of Africa and Americas; notice the value within the square bracket are different.

Thus, even if the data are made available to download, it is not possible to have generic algorithms to clean the wide display variant of data, as general cleaning of data may lost the original purpose of having that piece of information displayed. So a safer approach for

user who intended to use WHO's data, may consider download the excel format file and do a time-consuming data cleaning step. With limited time and resources available, the WHO data was then omitted, leaving to handle only the CIA data.



Data on the size of the HIV/AIDS epidemic: Data by WHO region by WHO region

Show filters

Download this data as
CSV (codes only) | CSV (text only) | CSV (text and codes) | CSV (XMart) | Excel (SpreadsheetML) | HTML (flat table) | GHO XML

Details: off

WHO region	Prevalence of HIV among adults aged 15 to 49 (%)	Number of people (all ages) living with HIV	Number of children under 15 living with HIV	Number of deaths due AIDS
	2011	2011	2011	2011
Africa	4.6 [4.4-4.8]	23 000 000 [22 000 000-25 000 000]	3 100 000 [2 700 000-3 400 000]	1 200 000 [1 100 000-1 300
Americas	0.5 [0.4-0.6]	3 000 000 [2 500 000-3 700 000]	63 000 [48 000-79 000]	85 000 [61 000-110
South-East Asia	0.3 [0.2-0.4]	3 500 000 [2 600 000-4 600 000]	140 000 [120 000-160 000]	230 000 [160 000-320
Europe	0.4 [0.4-0.5]	2 300 000 [2 000 000-2 700 000]	13 000 [11 000-15 000]	99 000 [71 000-130
Eastern Mediterranean	0.2 [0.1-0.3]	560 000 [410 000-800 000]	33 000 [23 000-45 000]	38 000 [28 000-50
Western Pacific	0.1 [0.1-0.1]	1 300 000 [1 100 000-1 600 000]	36 000 [30 000-43 000]	80 000 [63 000-100
Global	0.8 [0.7-0.8]	34 000 000 [31 400 000-35 900 000]	3 300 000 [3 100 000-3 800 000]	1 700 000 [1 500 000-1 900

Figure 2.1. Sample look of WHO data on size of HIV/AIDS epidemic, with variant display of data

Data on the size of the HIV/AIDS epidemic: Number of people (all ages) living with HIV by country

Static graph

Show filters

Download this data as
CSV (codes only) | CSV (text only) | CSV (text and codes) | CSV (XMart) | Excel (SpreadsheetML) | HTML (flat table) | GHO XML

Details: off

	Number of people (all ages) living with HIV[!]		
Country	2001	2006	2011
Afghanistan	2 000 [1 000 - 4 500]	3 200 [1 800 - 7 300]	5 800 [3 2
Algeria	[5 000 - 9 100]	[9 800 - 17 000]	[13 0
Angola	130 000 [92 000 - 200 000]	180 000 [130 000 - 260 000]	230 000 [160 00
Argentina	66 000 [51 000 - 82 000]	80 000 [65 000 - 99 000]	95 000 [79 00
Armenia	3 500 [1 500 - 12 000]	3 800 [2 200 - 7 400]	3 600 [2
Australia	13 000 [11 000 - 16 000]	17 000 [14 000 - 21 000]	22 000 [18 0
Austria	5 500 [4 000 - 7 300]	11 000 [8 700 - 14 000]	18 000 [13 0
Azerbaijan	3 000 [1 700 - 5 300]	5 100 [3 500 - 7 100]	6 700 [5
Bahamas	6 500 [6 100 - 6 900]	6 400 [5 900 - 6 800]	6 500 [6
Bangladesh	2 200 [1 300 - 4 700]	4 200 [2 800 - 8 700]	7 700 [4 9
Barbados	1 200 [<1000 - 1 500]	1 300 [1 100 - 1 600]	1 400 [1
Belarus	4 900 [2 200 - 10 000]	16 000 [12 000 - 23 000]	20 000 [15 0
Belgium	9 300 [7 100 - 12 000]	15 000 [12 000 - 19 000]	20 000 [16 0
Belize	3 400 [2 000 - 5 400]	4 100 [3 700 - 4 800]	4 600 [4
Benin	66 000 [52 000 - 79 000]	60 000 [52 000 - 67 000]	64 000 [56 0
Bhutan	<100 [<100 - <500]	<500 [<200 - <1000]	1 300 [<1

Figure 2.2. Second sample on WHO data on size of HIV/AIDS epidemic, with variant display of data

Since this project focus on the public health and computational biology context, partial healthcare related CIA data was extracted, meaning omitting those data records on geographical division, religions or second degree with impact to healthcare like type of governance. Also, those text data even with relevant to healthcare was also omitted as those texts are not convertible to any numeric figure for statistical representation. Since the CIA data has less variant type of displaying the data, this makes the manual data cleaning work on Microsoft Excel much easier. The need to do manual data cleaning was

due to excess information displayed, for instance the project only require the figure or actual value “14.35”, however referring to figure 2.3, on the CIA website “14.35 Deaths/1000 population (2013 est.)” was displayed instead and the excessive data of “Deaths/1000 population (2013 est.)” could not be converted to any numeric form to run statistical analysis. However, this piece of information is important to understand what the graphical solution. Hence, it was stored as a remarks text on a separate table (TB_Attribute, which will be mentioned at the later part) in the database.


THE WORLD FACTBOOK

Please select a country to view

ABOUT | REFERENCES | APPENDICES | FAQs | CONTACT

VIEW TEXT/LOW BANDWIDTH VERSION
DOWNLOAD PUBLICATION

FIELD LISTING :: DEATH RATE

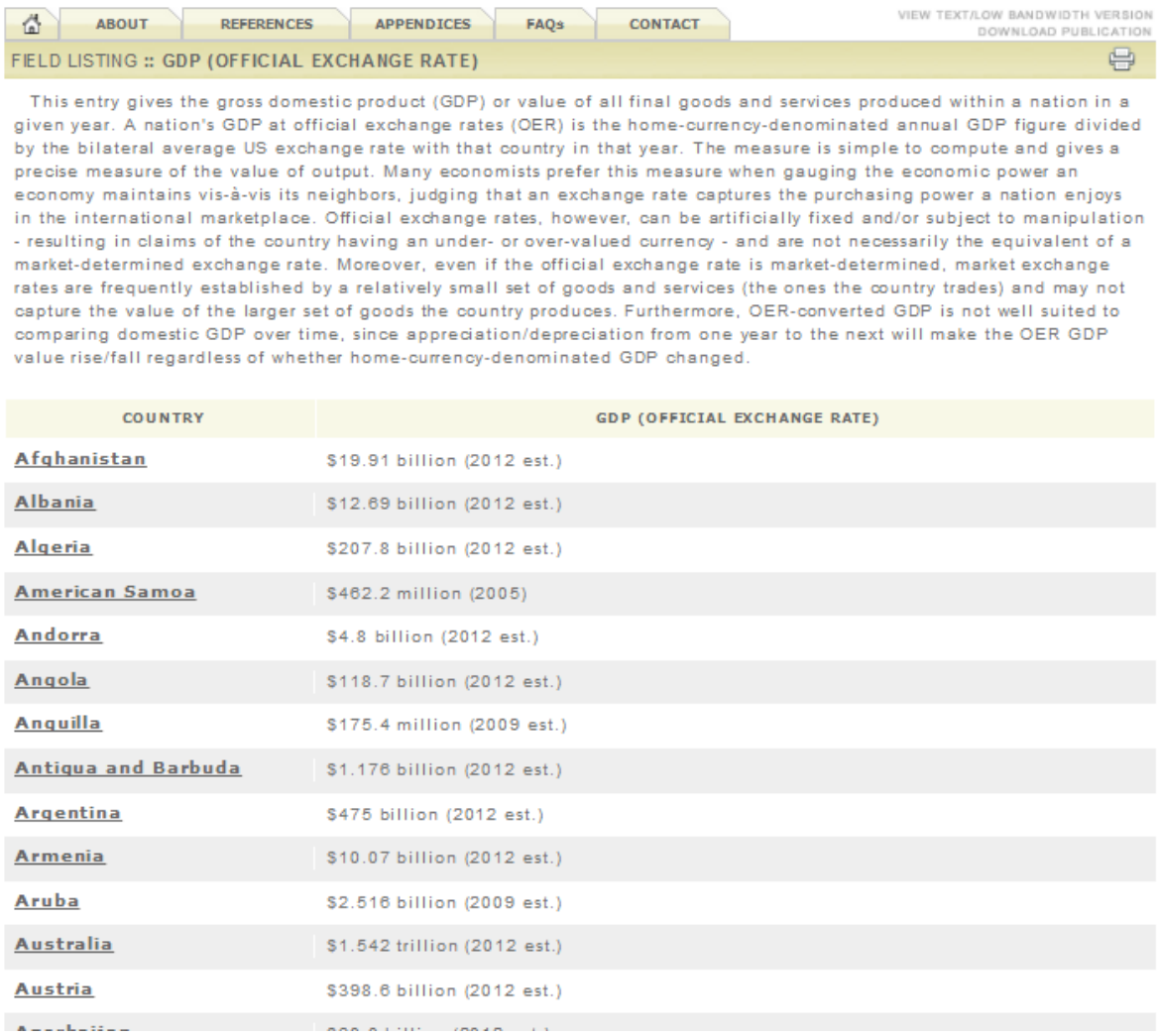
This entry gives the average annual number of deaths during a year per 1,000 population at midyear; also known as crude death rate. The death rate, while only a rough indicator of the mortality situation in a country, accurately indicates the current mortality impact on population growth. This indicator is significantly affected by age distribution, and most countries will eventually show a rise in the overall death rate, in spite of continued decline in mortality at all ages, as declining fertility results in an aging population.

Country Comparison to the World

COUNTRY	DEATH RATE(DEATHS/1,000 POPULATION)
Afghanistan	14.35 deaths/1,000 population (2013 est.)
Albania	6.36 deaths/1,000 population (2013 est.)
Algeria	4.31 deaths/1,000 population (2013 est.)
American Samoa	4.62 deaths/1,000 population (2013 est.)
Andorra	6.67 deaths/1,000 population (2013 est.)
Angola	11.86 deaths/1,000 population (2013 est.)
Anguilla	4.44 deaths/1,000 population (2013 est.)
Antigua and Barbuda	5.7 deaths/1,000 population (2013 est.)
Argentina	7.35 deaths/1,000 population (2013 est.)

Figure 2.3. Sample look of excess information display of data on CIA website

There is also another type of data that requires data cleaning – converting to the same unit. Figure 2.4 illustrate the GDP data as displayed on the CIA website, where the unit used for different countries, Andorra, Anguilla and Australia are billion, million and trillion respectively. With common unit ground, it ensures the graphical statistic result will not have much bias or error. Such step is therefore critical during data cleaning to obtain column C values of figure 2.5. To perform this step of data cleaning is not direct; it requires the splitting of the GDP value into numeric value and text value (ie. 'Billion'), as shown in figure 2.6. The Microsoft (MS) Excel function shown in figure 2.7, was assigned to every cell of column D of figure 2.6, has the following meaning; if the unit is equal to "billion", value will then multiply by 10^9 and if unit equal to "million" then value multiply by 10^6, else if the unit value equal to "trillion" then a multiply by 10^{11}, else the unit is not available thus a "0" will be automatically be displayed. The final else condition is to at least catch those non-numeric data, commonly displayed as "NA" on CIA data web.



ABOUT | REFERENCES | APPENDICES | FAQs | CONTACT

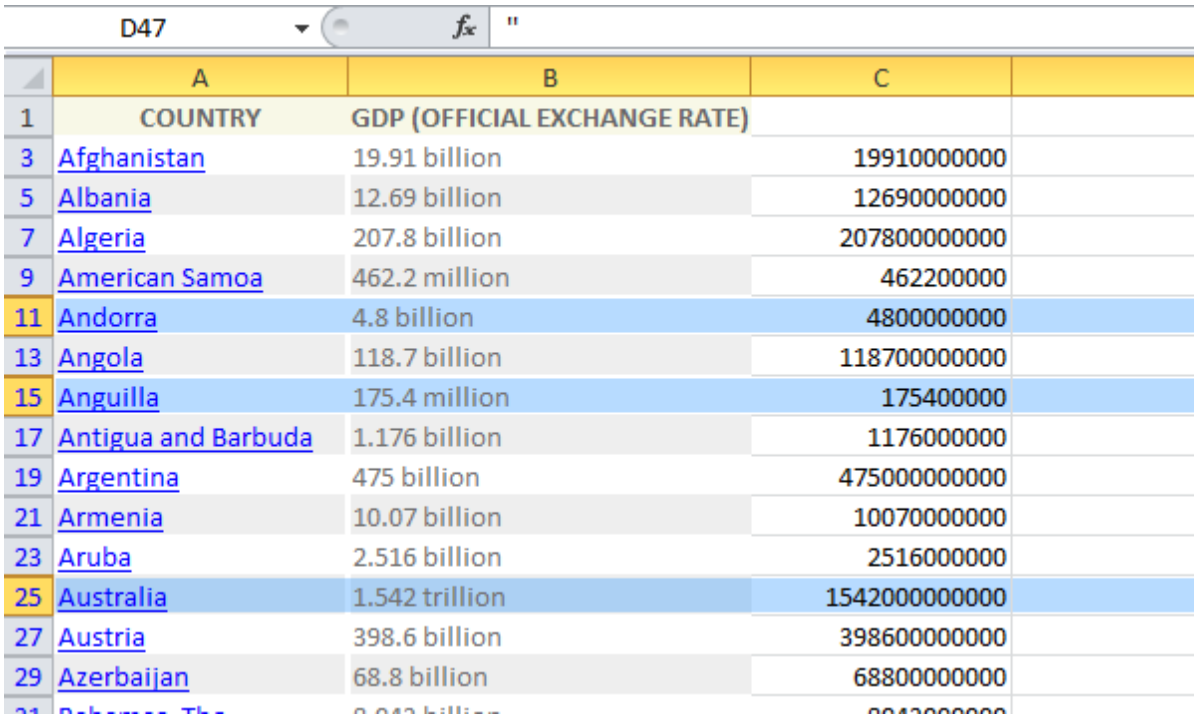
VIEW TEXT/LOW BANDWIDTH VERSION
DOWNLOAD PUBLICATION

FIELD LISTING :: GDP (OFFICIAL EXCHANGE RATE)

This entry gives the gross domestic product (GDP) or value of all final goods and services produced within a nation in a given year. A nation's GDP at official exchange rates (OER) is the home-currency-denominated annual GDP figure divided by the bilateral average US exchange rate with that country in that year. The measure is simple to compute and gives a precise measure of the value of output. Many economists prefer this measure when gauging the economic power an economy maintains vis-à-vis its neighbors, judging that an exchange rate captures the purchasing power a nation enjoys in the international marketplace. Official exchange rates, however, can be artificially fixed and/or subject to manipulation - resulting in claims of the country having an under- or over-valued currency - and are not necessarily the equivalent of a market-determined exchange rate. Moreover, even if the official exchange rate is market-determined, market exchange rates are frequently established by a relatively small set of goods and services (the ones the country trades) and may not capture the value of the larger set of goods the country produces. Furthermore, OER-converted GDP is not well suited to comparing domestic GDP over time, since appreciation/depreciation from one year to the next will make the OER GDP value rise/fall regardless of whether home-currency-denominated GDP changed.

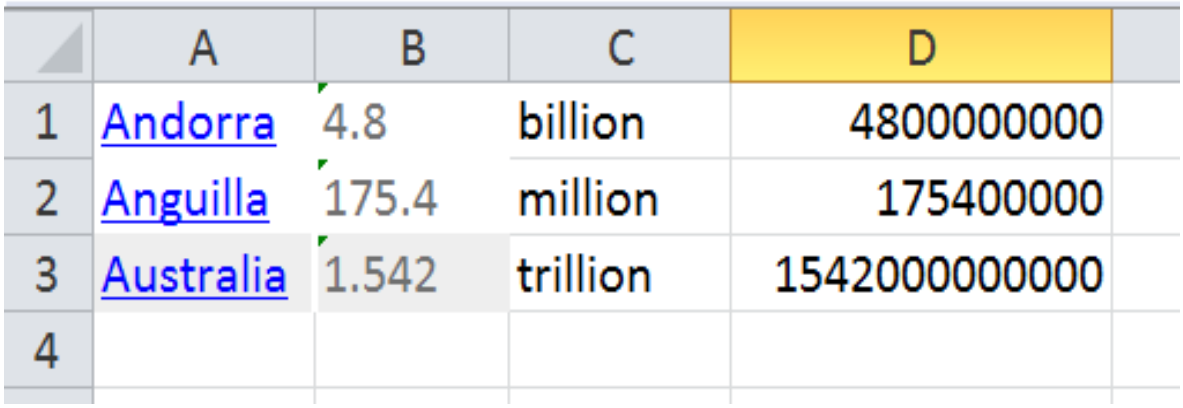
COUNTRY	GDP (OFFICIAL EXCHANGE RATE)
Afghanistan	$19.91 billion (2012 est.)
Albania	$12.69 billion (2012 est.)
Algeria	$207.8 billion (2012 est.)
American Samoa	$462.2 million (2005)
Andorra	$4.8 billion (2012 est.)
Angola	$118.7 billion (2012 est.)
Anguilla	$175.4 million (2009 est.)
Antigua and Barbuda	$1.176 billion (2012 est.)
Argentina	$475 billion (2012 est.)
Armenia	$10.07 billion (2012 est.)
Aruba	$2.516 billion (2009 est.)
Australia	$1.542 trillion (2012 est.)
Austria	$398.6 billion (2012 est.)

Figure 2.4. Sample look on CIA data – GDP



	A	B	C
1	COUNTRY	GDP (OFFICIAL EXCHANGE RATE)	
3	Afghanistan	19.91 billion	19910000000
5	Albania	12.69 billion	12690000000
7	Algeria	207.8 billion	207800000000
9	American Samoa	462.2 million	462200000
11	Andorra	4.8 billion	4800000000
13	Angola	118.7 billion	118700000000
15	Anguilla	175.4 million	175400000
17	Antigua and Barbuda	1.176 billion	1176000000
19	Argentina	475 billion	475000000000
21	Armenia	10.07 billion	10070000000
23	Aruba	2.516 billion	2516000000
25	Australia	1.542 trillion	1542000000000
27	Austria	398.6 billion	398600000000
29	Azerbaijan	68.8 billion	68800000000

Figure 2.5. Conversion to common unit done on MS Excel



	A	B	C	D
1	Andorra	4.8	billion	4800000000
2	Anguilla	175.4	million	175400000
3	Australia	1.542	trillion	1542000000000
4				

Figure 2.6. Break down steps on conversion done on MS Excel

```
D1 = IF(C2="billion", B2*1000000000,
                IF(C2="million", B2*1000000,
                                IF (C2=”trillion”, C2*1000000000000, 0))
```

Figure 2.7. Formula used on MS Excel

Database Design

The next focus will be on the database design; initial draft design of the CIA data before mid-term assessment, was shown in figure 2.8. However, after discussing with Prof Wong, he does have some convenience-related concern with the figure 2.8 database design, as there are many 0 (null or not applicable) entries, which is resource wastage especially when this project is working on high volume of data. These entries were assigned 0, firstly to make plotting of the statistical graph possible, secondly; because under certain attributes, the country did not release any figure to the CIA or the attribute case is not applicable to that particular country. An instance when the attribute could not be measured in a country is when that country does not have any population; there are only group of scientists who reside there due to a long term field research, country like: Antarctica. Refer to figure 2.9 for a screenshot of the CIA population data.

Prof Wong suggested having the extracted CIA data consolidate in a bigger table as represented in table 2.1. However, bearing in mind the MS SQL Server 2012 used in this project is an express version, the maximum row of data it could hold might be smaller than that of the total data intended to be used in this project. A safe estimation count, say 200 countries per attributes, and seventh bigger set of attributes that will go up to 3400 rows. A bigger set of attribute refers to attribute ‘sex ratio’, which was further classified into 'Sex Ratio 0 – 14’, 'Sex Ratio 15 - 24', 'Sex Ratio 25 - 54', 'Sex Ratio 55 - 64', 'Sex Ratio >= 65' and 'Sex Ratio Total Population'.

Hence, the finalised database design used was a much simpler database design as shown in figure 2.10, with an additional table – TB_Attributes, which schema could be shown in

figure 2.11. Also, with the aid of SQL, it is possible to get data from the figure 2.8 schema similar to that of figure 2.10, without having to worry of hitting the maximum row of data accepted by MS SQL Server 2012.

Country	HEALTH EXPENDITURES
Afghanistan	7.4% of GDP (2009)
Albania	6.9% of GDP (2009)
Algeria	5.8% of GDP (2009)

Country	PEOPLE LIVING WITH HIV/AIDS
Afghanistan	NA
Albania	NA
Algeria	18,000 (2009 est.)

Country	HIV/AIDS - ADULT PREVALENCE RATE (%)
Afghanistan	0.01% (2001 est.)
Albania	NA
Algeria	0.1% (2009 est.)

Country	HIV/AIDS - DEATHS
Afghanistan	NA
Albania	NA
Algeria	Fewer than 1,000 (2009 est.)

Country	HEALTH EXPENDITURES	PEOPLE LIVING WITH HIV/AIDS	HIV/AIDS - ADULT PREVALENCE RATE (%)	HIV/AIDS - DEATHS
Afghanistan	7.4	0	0.01	0
Albania	6.9	0	0	0
Algeria	5.8	18,000	0.1	1,000

Figure 2.8. Break down steps on conversion done on MS Excel

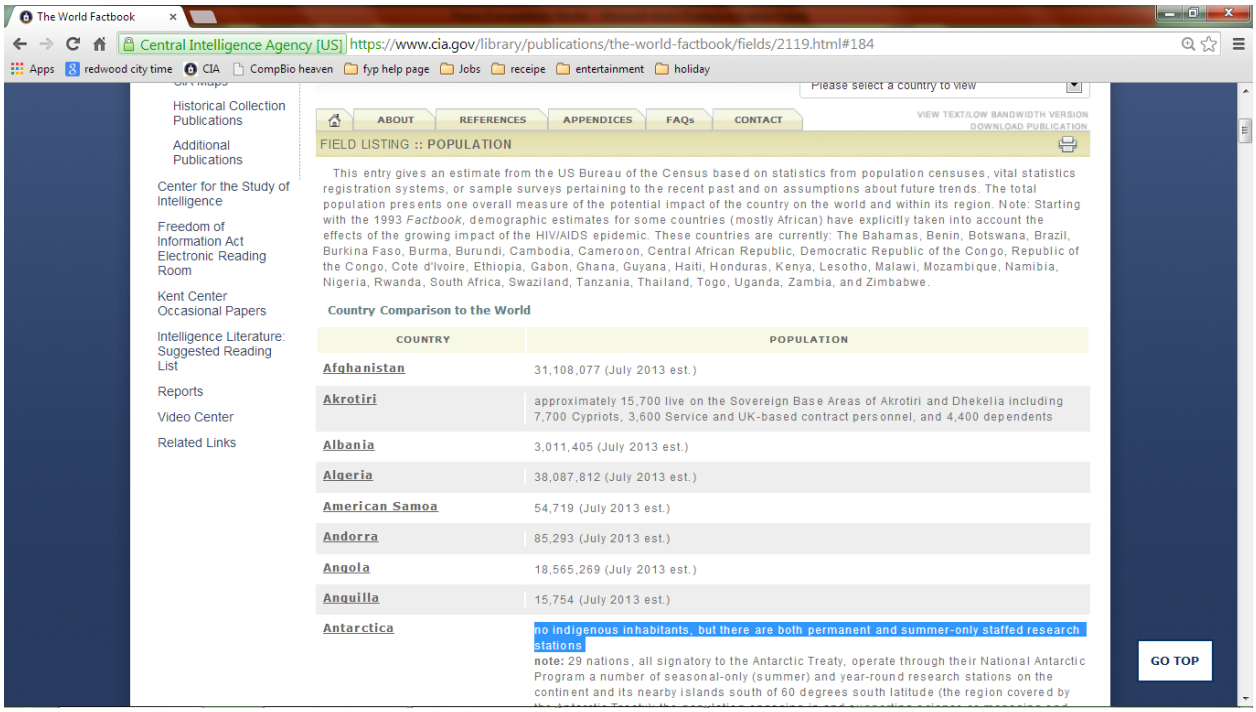


Figure 2.9. Screenshot of the CIA population data

Entity	Entity_type	Attribute	Attribute_value	Time_stamp	Data_source
Singapore	country	population	5000000	2013-03-15	CIA Fact Book

Table 2.1 Initial Database Design

Data from CIA Factbook:

Country	HEALTH EXPENDITURES
Afghanistan	7.4% of GDP (2009)
Albania	6.9% of GDP (2009)
Algeria	5.8% of GDP (2009)

Country	PEOPLE LIVING WITH HIV/AIDS
Afghanistan	NA
Albania	NA
Algeria	18,000 (2009 est.)

Country	HIV/AIDS - DEATHS
Afghanistan	NA
Albania	NA
Algeria	Fewer than 1,000 (2009 est.)

Country	HIV/AIDS - ADULT PREVALENCE RATE (%)
Afghanistan	0.01% (2001 est.)
Albania	NA
Algeria	0.1% (2009 est.)

Country	Data
Afghanistan	7.4
Albania	6.9
Algeria	5.8

TB_HealthExp

Country	Data
Afghanistan	0
Albania	0
Algeria	18,000

TB_HIVLiving

Country	Data
Afghanistan	0
Albania	0
Algeria	1,000

TB_HIVDeath

Country	Data
Afghanistan	0.01
Albania	0
Algeria	0.1

TB_HIVAdult

Figure 2.10. Actual database design deployed

PRISCILLA-PC\SQLE...dbo.TB_Attributes

Column Name	Data Type	Allow Nulls
Attributes	nchar(40)	☐
Unit	nchar(15)	☑
Remarks	nchar(90)	☑
TableName	nchar(30)	☑
		☐

Figure 2.11. TB_Attributes data schema and Attributes is the primary key.

B. The Technical Aspect

The project aim was to create a user-friendly web interface and the choice of Integrated Development Environment (IDE) that able to give a nice interface of say button. If one were to use html button, it probably looks like a greyish rectangular still picture box with a black border. However, an asp.net button will have mouse pointer hovering effect that triggers the colour darkening or brightening on the button. Also, there are the round edge finishes. As Microsoft Visual Studio Express 20xx (MS VS) is the main IDE from Microsoft, it is able to develop many web-based applications with MVC and WPF, coded in Visual Basic (VB) as well as Extensible Application Mark-up Language (XAML) for all platforms supported by Microsoft Windows .NET. However, in this project context, it does not need those, in fact it just need the MS VS Web development 2012, with or without the "express" behind it. Note that the express version will have certain limitation on the web controls.

The current version of MS VS is 2013, however, the project uses MS VS 2012, and do note that the 2010 version onwards, there are a slight modification with what needs to be downloaded in order to use the MS VS. From MS VS 2010 onwards, downloading web development is no longer just that web development express component, one need to download the web matrix package which contains a whole lot that comes with an installer; also, one must ensure they have the service park 1. MS VS web development 2012 was installed instead of the 2013 version, was purely because the project needs to use external extended libraries, Infragistics.WebGUI (which latest version is only for 2012 version) for the graph display.

Another component that comes in hand in hand will be the database storage. Download the Microsoft SQL Server 2012 refers to only server services, which was discovered at the initial data management stage on where the in-built SQL query could not execute the creation of dbo.Table. To manage the data stored on the conventional SQL Server, meaning having full control to create, add and delete tables and data, one needs to download the MS SQL Server Management studio, in this context, version 2012 was used.

Under table 2.2, shows the different combination of software considered. Before the tabulation of the result of mixing different software to use, there are in fact many trial and errors first. Actual coding of the web interface was roughly four months before the submission; the reason for not starting early was due to IDE licensing issue and other parts of the work of this project were carried out on a separate issue. For instance, many CIA data was taken to run much statistical analysis to find which data were useless in analysing.

MS VS Ver.	MS SQL Server	Statistical Software	Remarks	Conclusion
2010	2010	R	VS 2010 has licensing issue.	☹
2012	2010	R	R keeps showing error when data on loading.	☹
2012	2012	Extreme optimization	No graphical output	☹
2012	2012	NetAdvantage_Ultimate_20111	Simple analyses work	☺

Table 2.2. Trial and Errors of combination of software

The follow set of information will come in handy if one decided to reproduce this project and decided to:

- use R, there are a few important steps to look out:
    1. The version of R must be compatible (higher than 2.2.15)
    2. In order to use R in .Net, one needs to download the R(D)Com package
    3. Check that the R/bin contain sciproxy.dll (which has to be download separately)
    4. After downloading, under the project > Add Reference > COM > Type Libraries, add all sciproxy.dll related libraries, refer to figure 2.12
    5. At the source code page, do include the namespace as shown in figure 2.13
    6. At toolbox, right click add Items, refer to figure 2.14

- emphasis on the statistical ability, and would like to consider Extreme optimization (EO) software, similarly a couple of points to look out for:
    1. EO is like a built-on library or assembly
    2. EO only generate the analytical value but does not produce a graphical solution
    3. After downloading, under the project > Add Reference > Assemblies > Extension, add Extreme Numeric version 5.0.12317.0 (this matches version with VS 2012) depending on your MS VS version, refer to figure 2.15
    4. At source code page, do include the respective namespace as shown in figure 2.16
    5. Go to toolbox and add only the respective version of Extreme Numeric tools to prevent control tool item ambiguity, as shown in figure 2.17

- emphasis on the statistical ability, and would like to consider working on NetAdvantage (NetAdv), likewise to note:

    1. NetAdv does output a graphical solution, look at fourth point on how to get the UltraChart tool for display
    2. After downloading NetAdv, go, under the project > Add Reference > Assemblies > Extension, depending on your MS VS version. Add Infragistics4 (this matches VS 2012), refer to figure 2.18
    3. At source code page, do include the respective namespaces as shown in figure 2.19
    4. Go to toolbox and add only the required Infragistics tools as the list could be very long as shown in figure 2.20

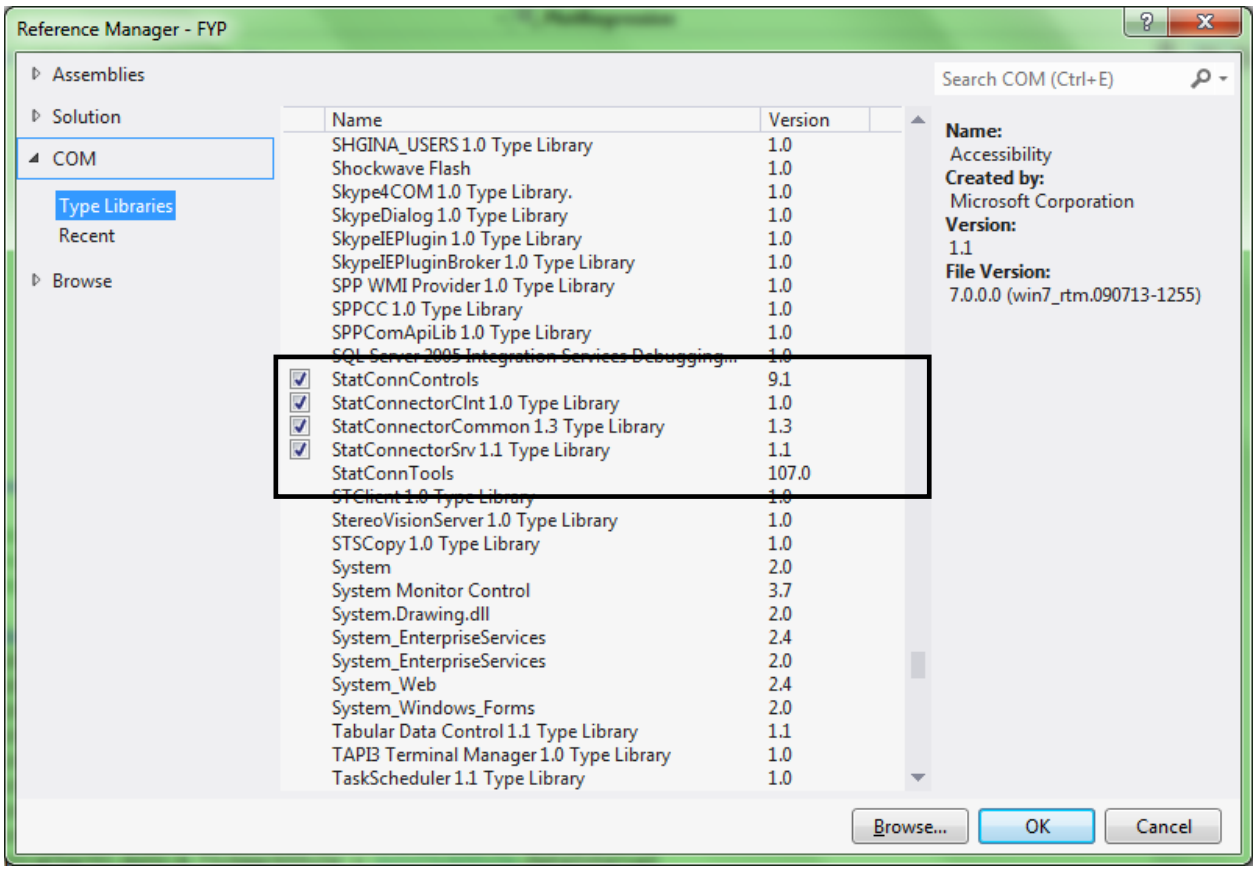


Figure 2.12. Sciproxy.dll related libraries to be included

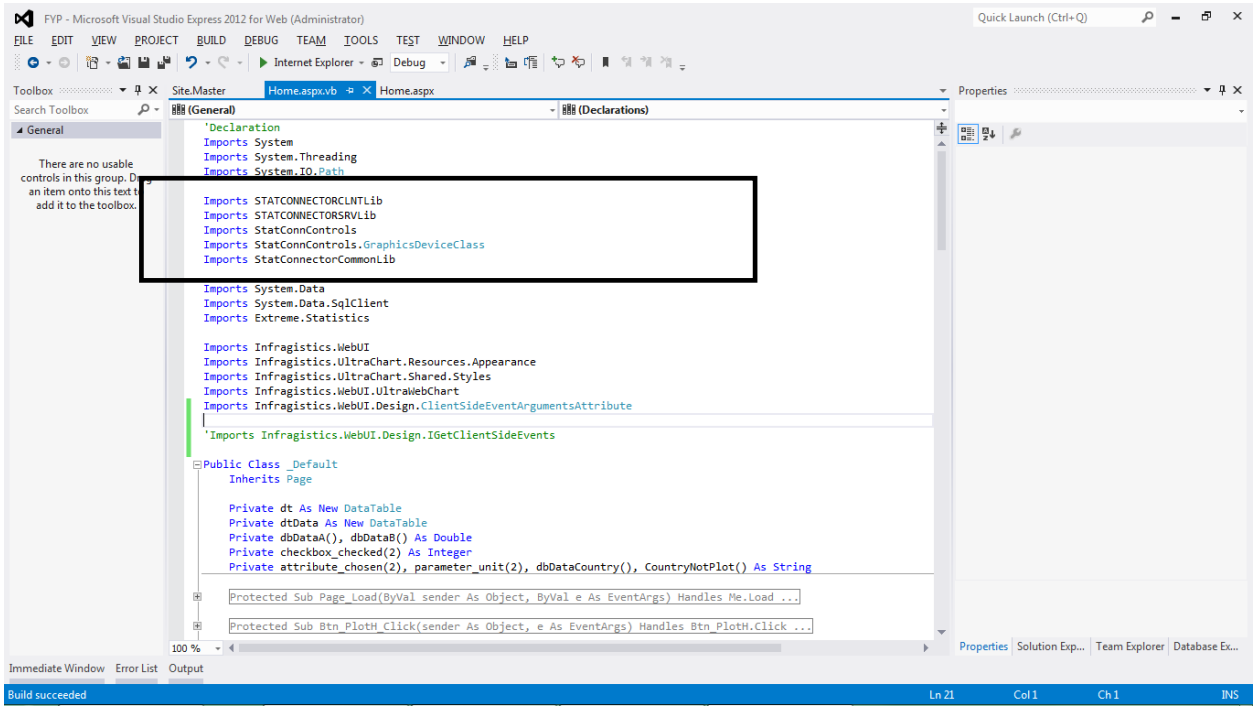


Figure 2.13. Sciproxy.dll related namespace to be included in the source code

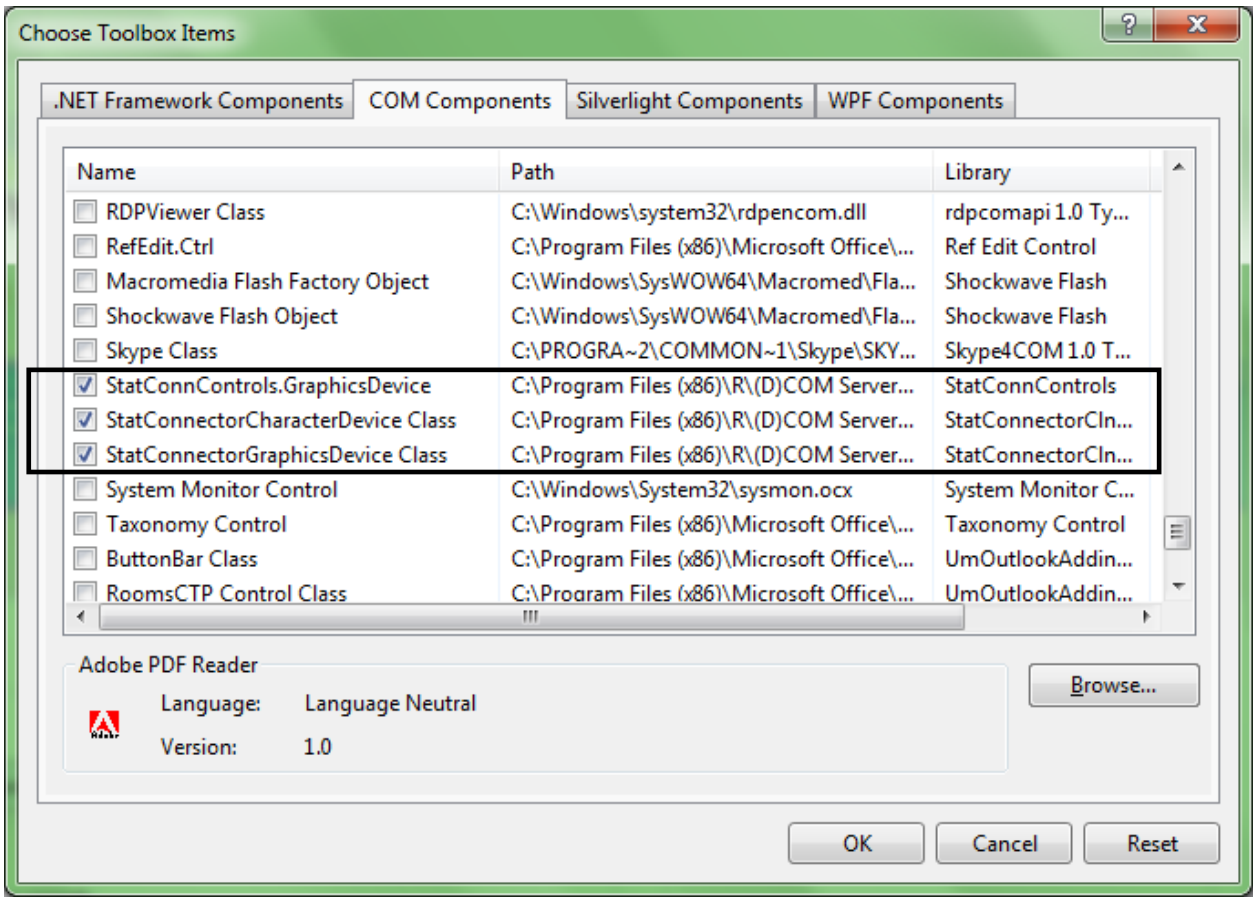


Figure 2.14. Sciproxy.dll related toolbox items to be included

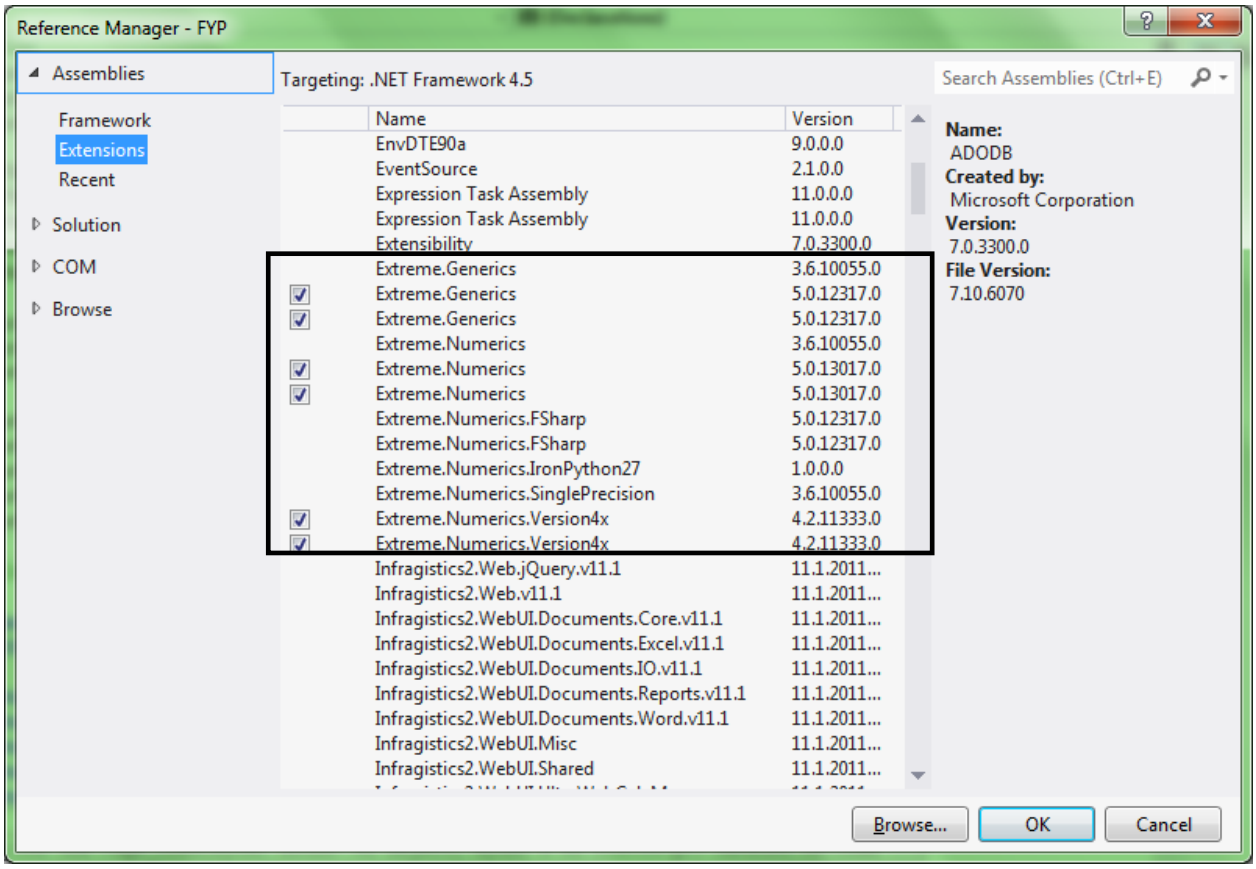


Figure 2.15. Extreme optimization (EO) related assemblies to be included

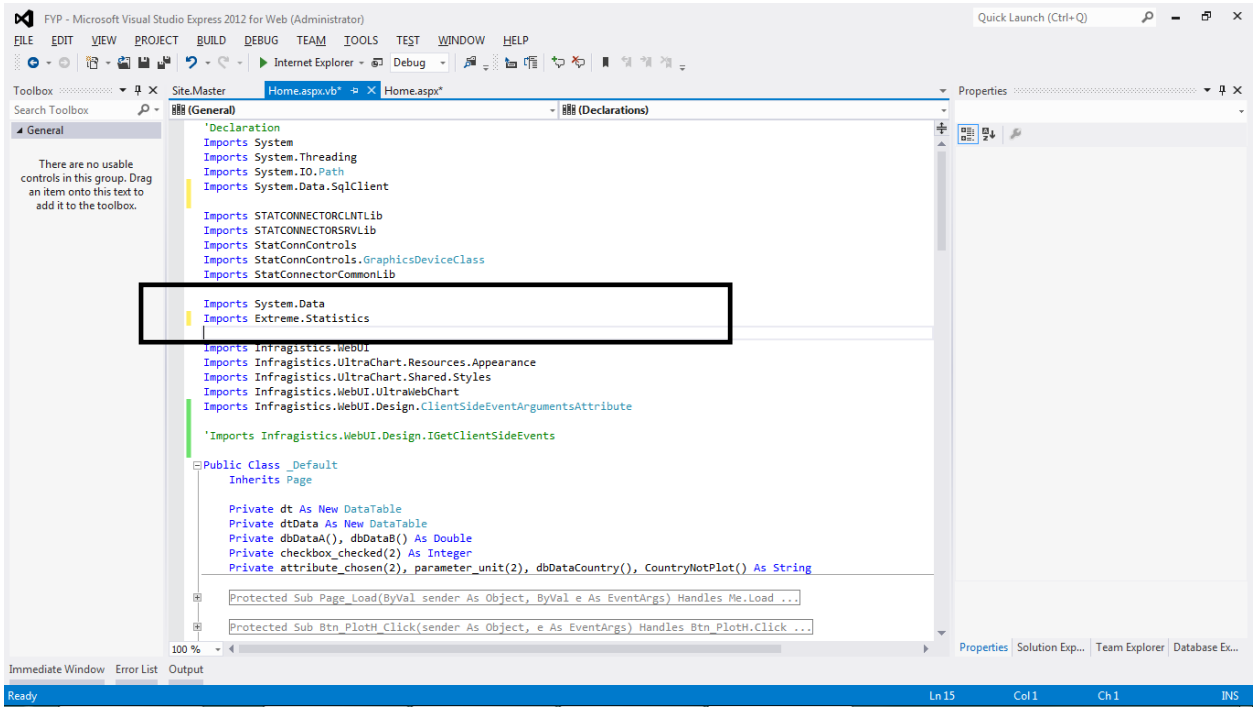


Figure 2.16. EO related namespace to be included in the source code

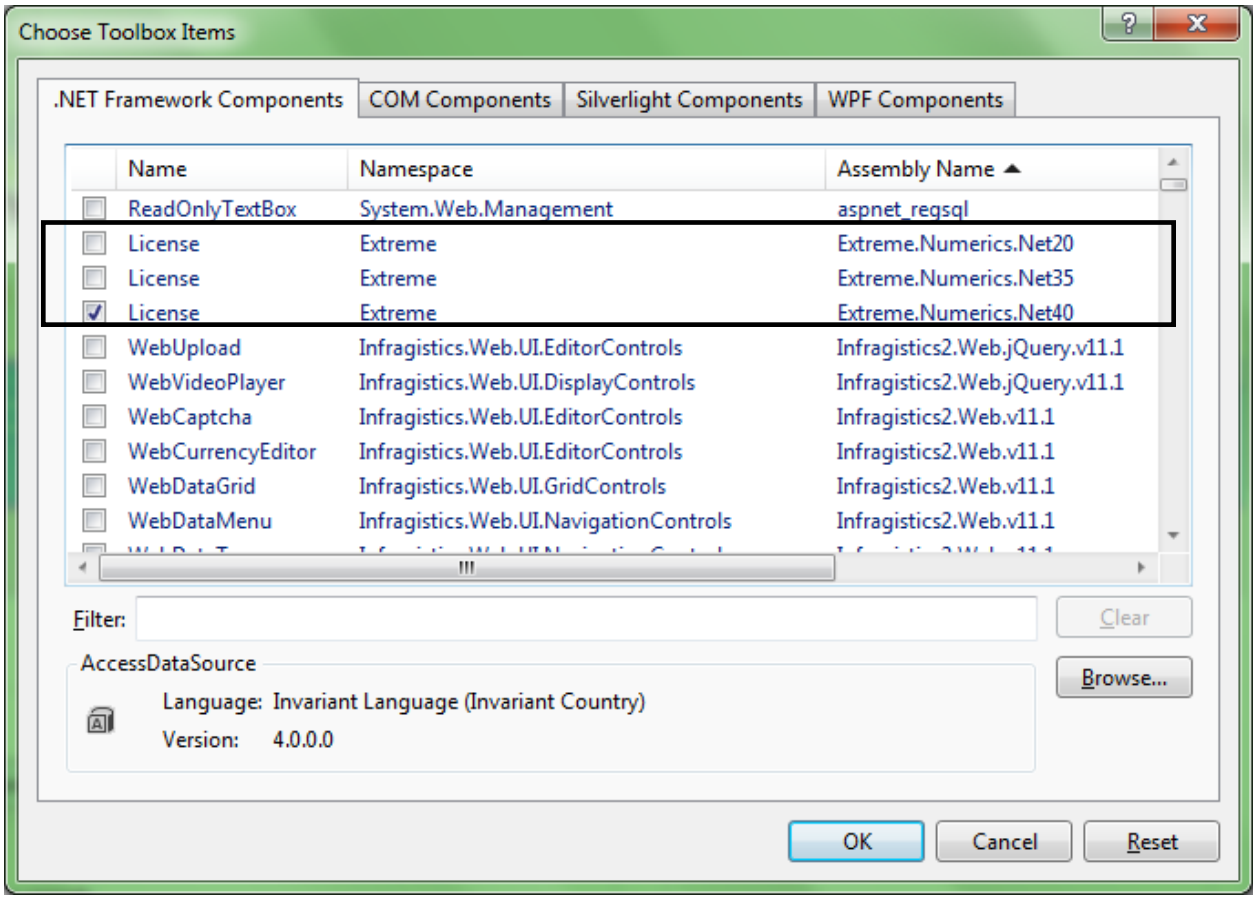


Figure 2.17. EO related toolbox items of correct version to be included

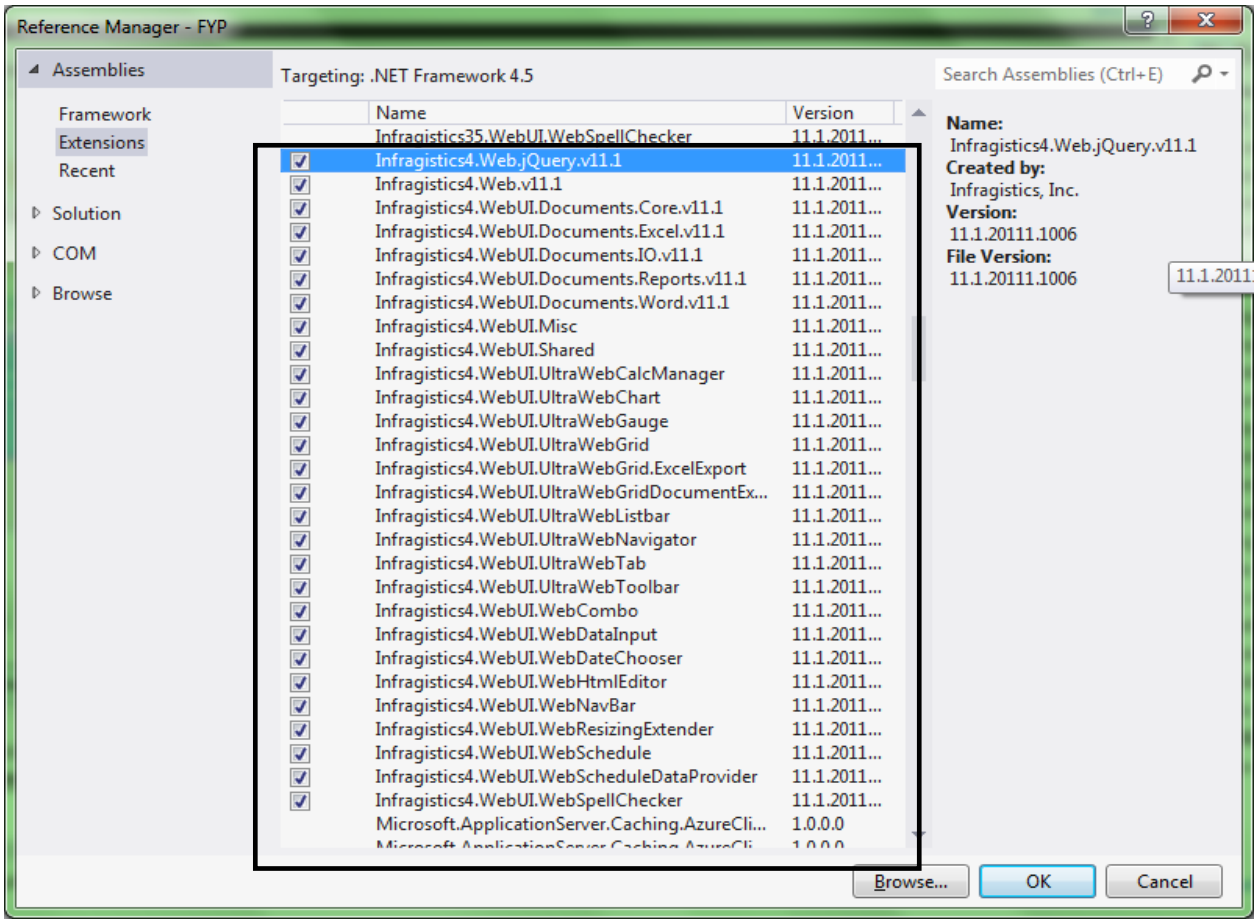


Figure 2.18. NetAdv - Infragistics related extensions assemblies to be included

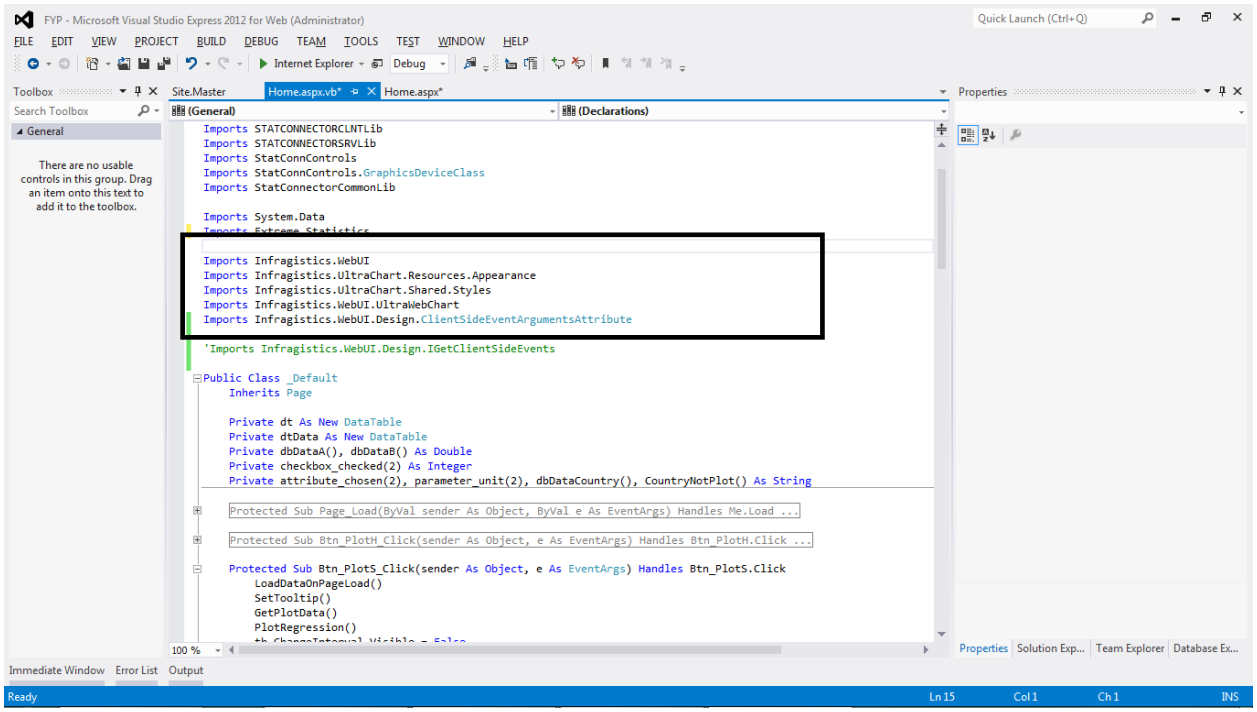


Figure 2.19. NetAdv – Infragistics namespace to be included in source code

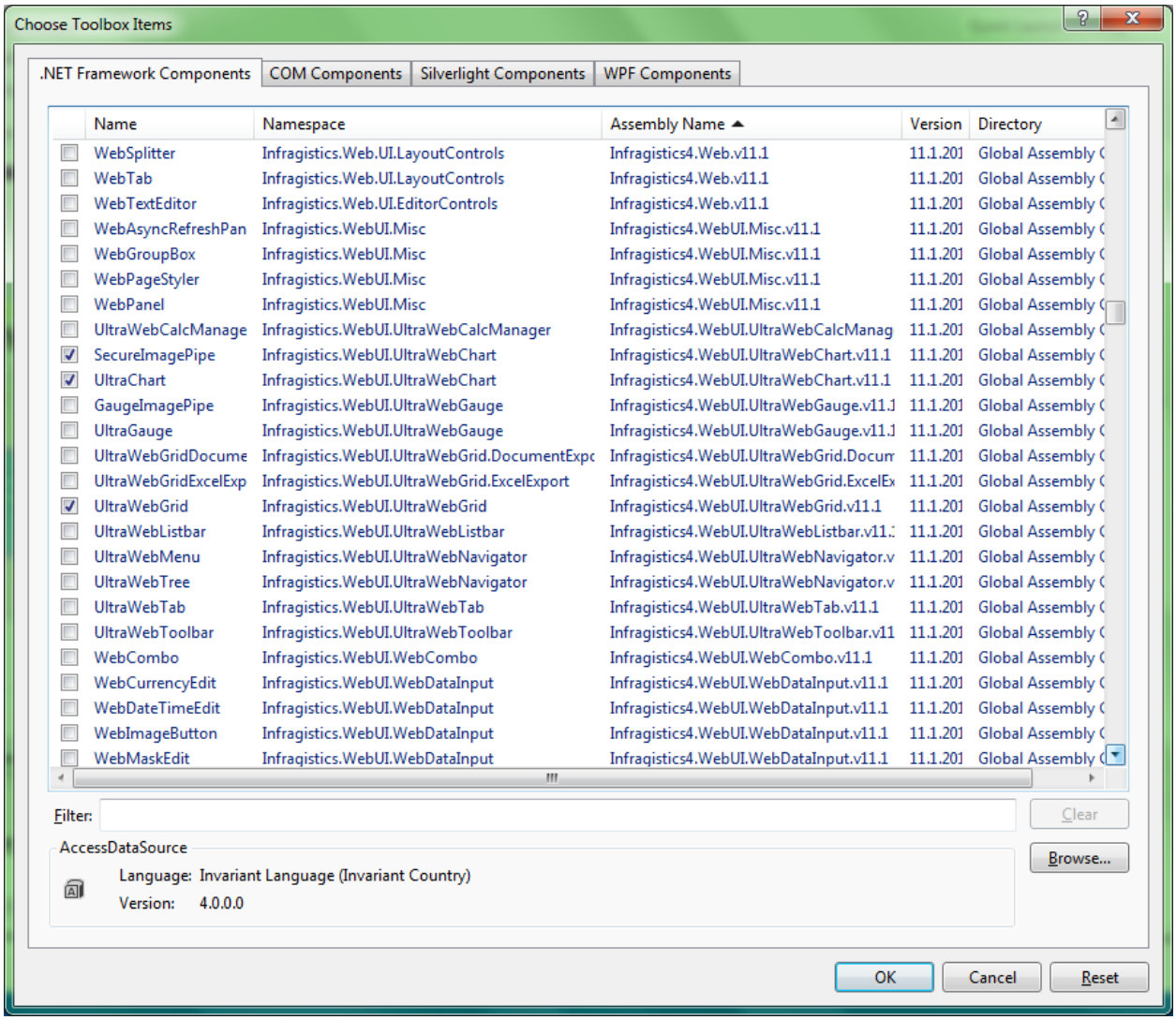


Figure 2.20. NetAdv – Infragistics related toolbox items of correct version 4 to be included

C. The Methods of Analysing

At the planning phase of the project, the data was manually put to run through a list of statistical analysis methods, like linear regression, seasonal trend reading, finding of standard deviation and correlation coefficient in R code. However, as R (D) Com version could not match with the latest R. Since R was not used eventually in this project, the preparatory work using R to decide on data selection from CIA factbook was appended under the appendix A and B.

Initial plan was to hard code a list of statistical modelling methods and let user select how they want the data to be handled. And due to time limitation on exploring the two newly discovered third party vendor software, EO and NetAdv, we will get down to the most basic analysing of data, just a mere histogram plotting and finding the average, with added features to be discussed under the next section.

.

III. Results & Discussion

In this section, we will discuss two main components of the project, the design and the functionality, together with the expectation of the project what was delivered and other limitations while executing the expectation and solutions proposed under the sub-section, future development.

A. The Design

As emphasised in previous sections, this project aims is to support the scientists, hence user experience is the first foremost important issue to address. Recalling on what was taught in one of the undergraduate module, CS3240 – human computer interaction, when designing the website interface, must always remember the user experience. User experience involves the person's behaviours, attitudes, and emotions about using a particular product, system or service.

As user experience is dynamic and it is constantly modified over time due to changing usage circumstances, however Shneiderman's "Eight Golden Rules of Interface Design" will never change.

The eight rules taken from Ben (2009) are:

1. Strive for consistency,
2. Enable frequent users to use shortcuts,
3. Offer informative feedback,
4. Design dialog to yield closure,
5. Offer simple error handling,

6. Permit easy reversal of actions,
7. Support internal locus of control, and
8. Reduce short-term memory load.

Applying all these eight rules will be too much in this project, so Mandel's Golden Rules was referred instead, it merely summed it up into three main points:

1. Place Users in Control
2. Reduce Users' Memory Load
3. Make the Interface Consistent

While drafting the outlook of the website, many thoughts for instance the positioning of the graph area, the attributes selections list, the buttons, the scrollbars and the text, and the choice of the display size, font size and their colour scheme were considered very thoroughly in hope it ties in with the Mandel's rules.

After a series of considerations, planning, sketching, the final draft of the website design was shown in figure 3.1, and the actual website interface could be referred to figure 3.2. In the upcoming discussion, we will often draw illustration reference back to these two figures.

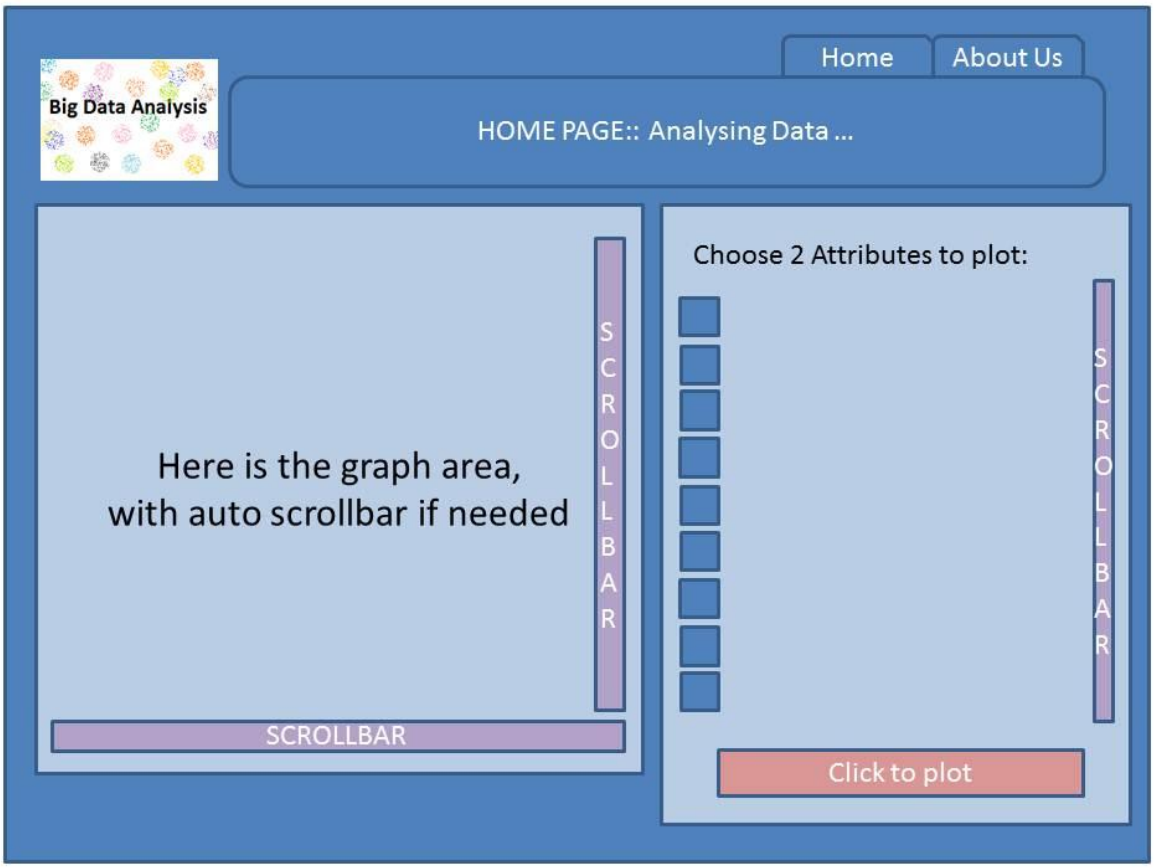


Figure 3.1. Final draft of the website Interface design

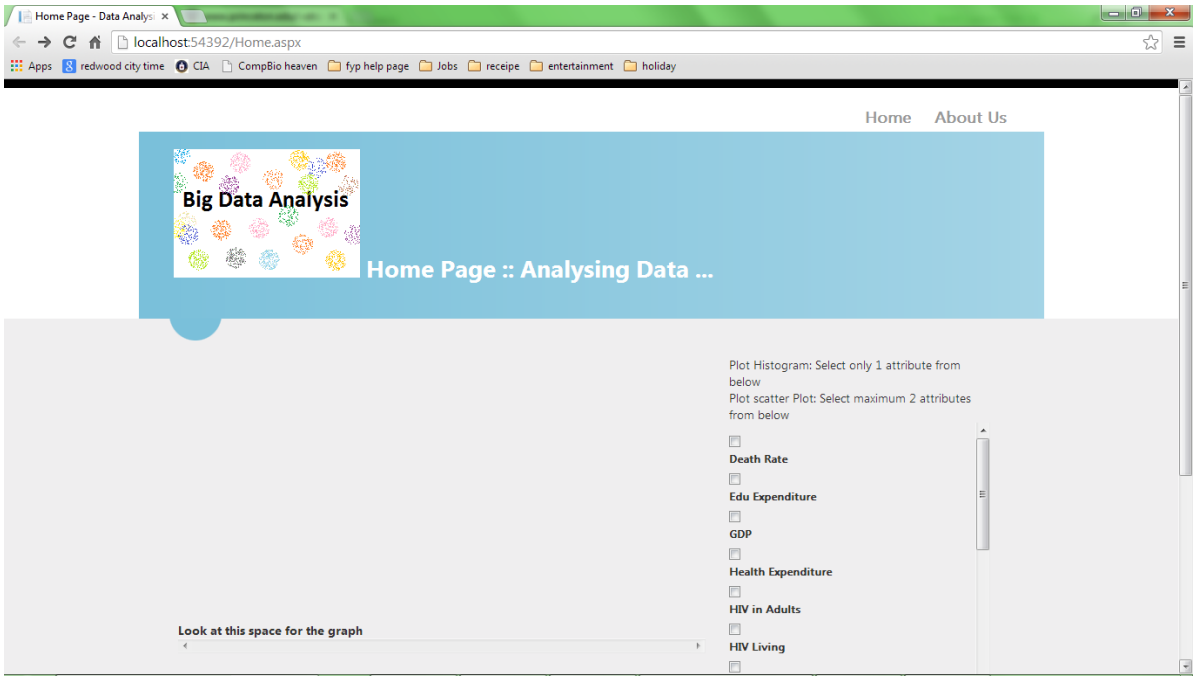


Figure 3.2. Implemented Website Interface – Home page

The Colour

Notice there is a difference in the choice of colour scheme, because in user experience definition, the mood of the user will contribute to the experience. Colour used in figure 3.1 tends to make the user feels a little blue in emotion, this is not a hear-say kind of blue is gloomy. It was in fact supported with quite sufficient scientific studies, and there is even colour therapy to cure certain emotional-linked-hormones-triggered illness. When one is exposed to a spectrum of lighter colour, one will tend to feel more positive and happier. Also, Samina (2005) mentioned that, the human body, according to the doctrine of chromotherapy*, is basically composed of colours. The body comes into existence from colours, the body is stimulated by colours and colours are responsible for the correct working of various systems that function in the body. All organs and limbs of the body have their own distinct colour.

As stated on the website of incredible Art (n.d), colour is categorised into warm colours, having like yellow, orange, red and cool colours like blue, green, black or anything of darker shade. And blue is an emotion colour, as of the words it was linked to like peace, tranquility, cold, calm, stability, harmony, unity, trust, truth, confidence, conservatism, security, cleanliness, order, loyalty, sky, water, technology, depression, appetite suppressant, mostly are emotional-related words. Hence the blue was chosen, in the actual coding of the website, but of a lighter shade.

* chromotherapy: defined in Wikipedia (2013), method of treatment that uses the visible spectrum (colours) of electromagnetic radiation to cure diseases.

The attributes selections list area

It was placed on the right of the website because the common position of the mouse. Most users' uses mouse on the right side of the keyboard, and thus it is for the convenient of the user to go the right. Taking a look at figure 3.2 and 3.5, the implement of scrollbar on the right, it was so that user need not have to use the external scrollbar to go through a long list of attributes, and then forget the instruction on attributes selection. That will be less memory load on the user, second rule of Mandel's was applied.

Additionally, the web interface was designed in such a way the user could have a very good view on what is expected on the page at one glance, by the help of colour used to split the area to focus. Looking back and forth on figure 3.2, the 'home' page and figure 3.3 the "about us" page demonstrate the consistency in design – the looks and colour scheme, with the aid of using a master site for page design and a common CSS style script. Apart from the two points mentioned in reducing user memory load, is the purposely positioning of 'home' and 'about us' links are on the page.

The interactive features

Figure 3.4, showing the draft sketch of hovering event of the mouse / cursor above the data point on the graph, just fitting, applied the first golden rule of Mandel's, allowing user to be in control, that means the interface of the website has to be as interactive as possible to provide flexibility, facilitative, forgiving and interruptible.

In this age, everything was expected to be interactive, that simple task actually took an unfamiliar web developer three good days trying luck with all sort of snip code found on

forum. Figure 3.5, illustrated the feature of mouseover value of the attributes, is exactly what I wanted in mind after looking at figure 3.4 draft sketch. But, the tooltip feature (in MS VS 2010 onwards, means mouseover), refers to the entire checkbox list, meaning all items mouseover will display similar value, that bad feature will totally mislead the user. So, from all these bulky and lengthy trial and error luck, a thought suddenly struck me, since they don't have individual tooltip, might as well I add a new feature to each item of the checkbox list of the attributes found at the right column of the web interface. Coding was made easy after having the thoughts, with just a simple three lines code in replace for a slow and lengthy code that could be found in appendix D. The partial code which I am very eager to share is shown in figure 3.6. The 'dt' is the data table that binds data from 'TB_Attributes' table, which contains all the attributes as primary key. For this particular code, it means for each item of the checkbox list name CB_Attributes, add a new title attribute with the 'Remarks' value found in the same row as the value as displayed on the checkbox list.

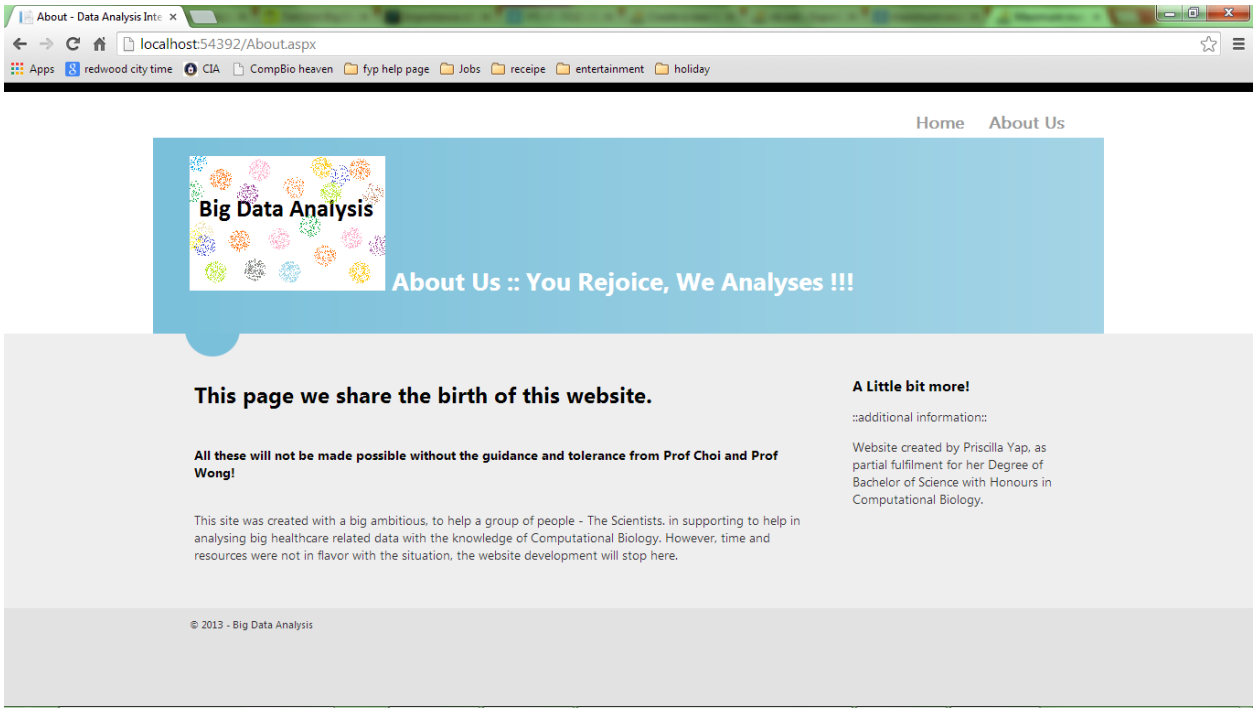


Figure 3.3 “About us” page display

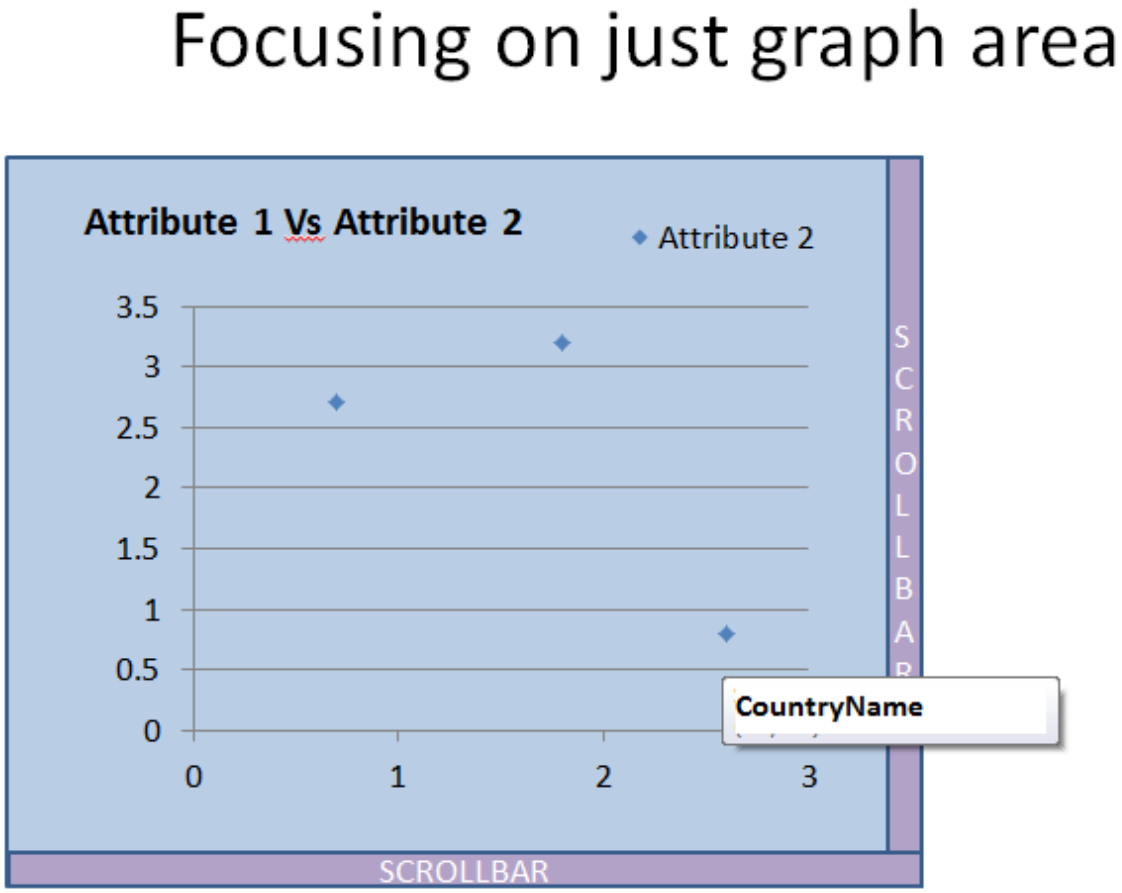


Figure 3.4. Draft sketch of Mouseover on the graph

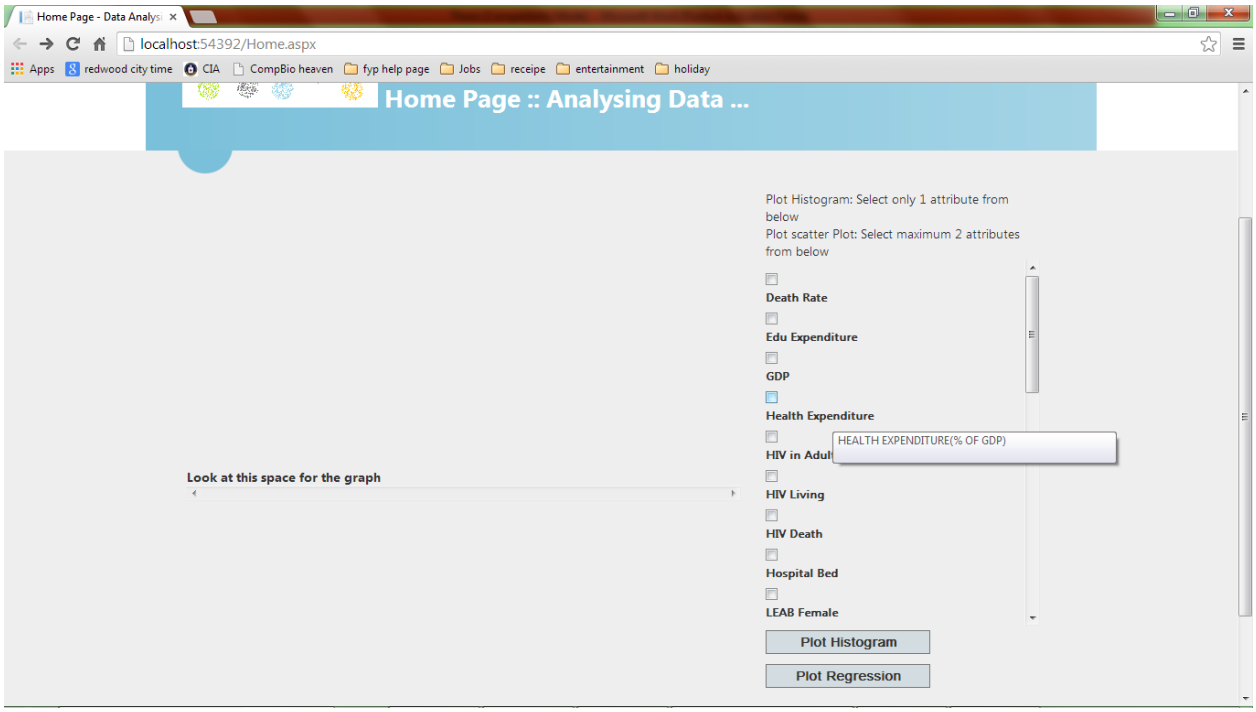


Figure 3.5. Mouseover each attribute selection will give a brief explanation

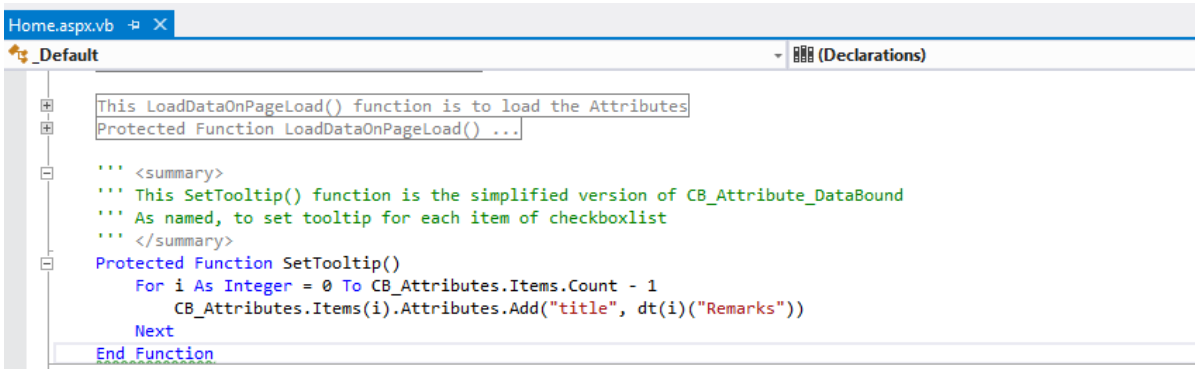


Figure 3.6. Codes that does the mouseover display for each checkbox item

B. The Functionality

With the entire nice and friendly interface well taken care, a website must be able to answer the needs of the user. Design is often the wants and functionality is the needs. Usually user comes to a website; they are often looking for information, not just look at how pleasant the graphical interface is. Factually, functionality was first planned during the drafting of the website design, as it is the main objective of the project. A quick run through on how to use the website to obtain useful information. Starting, user will be faced with image on figure 3.2, then select attributes by scrolling down the checkbox list on the right, select the attribute of interest, very common sensely, will look downward for any button – 'plot histogram' and lightly hit on it, user will get what is shown on figure 3.7.

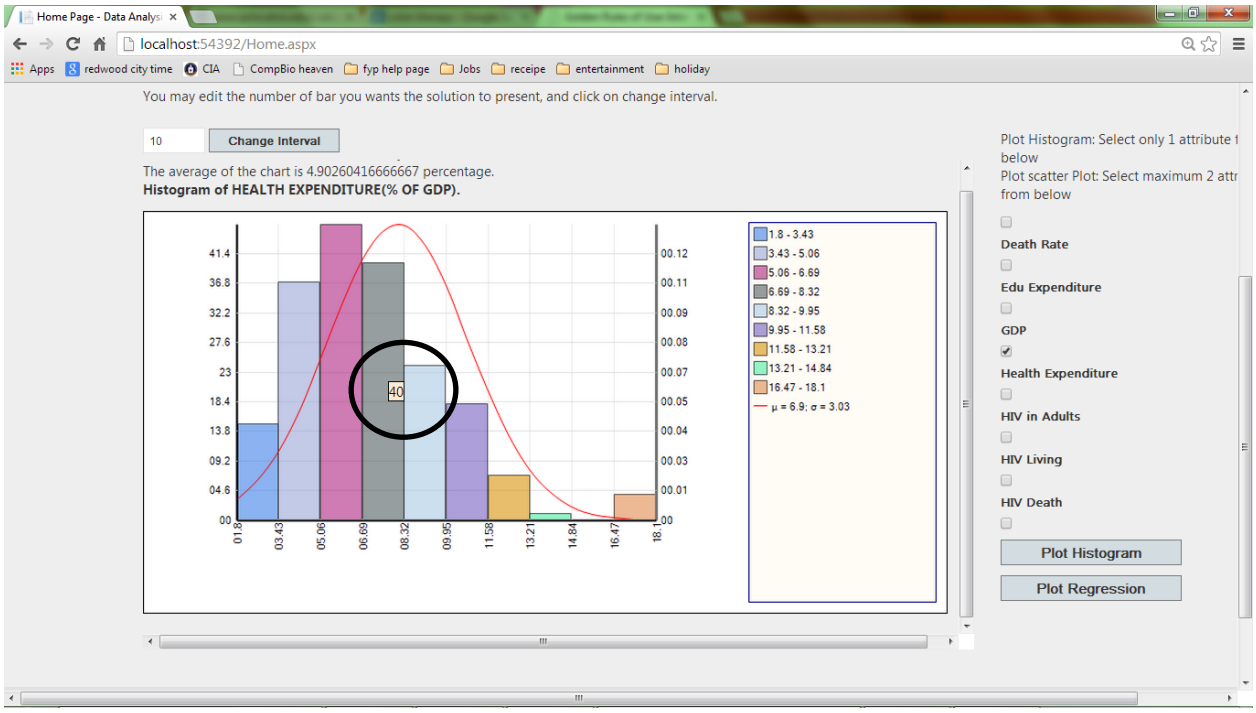


Figure 3.7 Histogram displays of Health Expenditure of CIA data with 10 bars

As mentioned under the Methods of Analysing of Materials and Methods, there are added feature on top of the simple histogram plotting. As NetAdv has quite a small range of statistical tools, Infragistics could only allow the display of simple histogram chart. However, the value circled in figure 3.7, was the total number of country, whose data happen to fall within the range of value represented by the bar. The value of total country would only be displayed when the mouse cursor hover above the bar chart. Apart from this interactive way providing information, the average was also calculated and display above the title of the histogram or could be obtained from the legend of with the symbol μ.

As the default number of bar chart display each time is 10, so here is the third feature for user to interact, is user could choose higher number of bar to be displayed, that will result in small interval per bar. Say change 10 to 15 and click on change interval, see figure 3.8 for the changed interval.

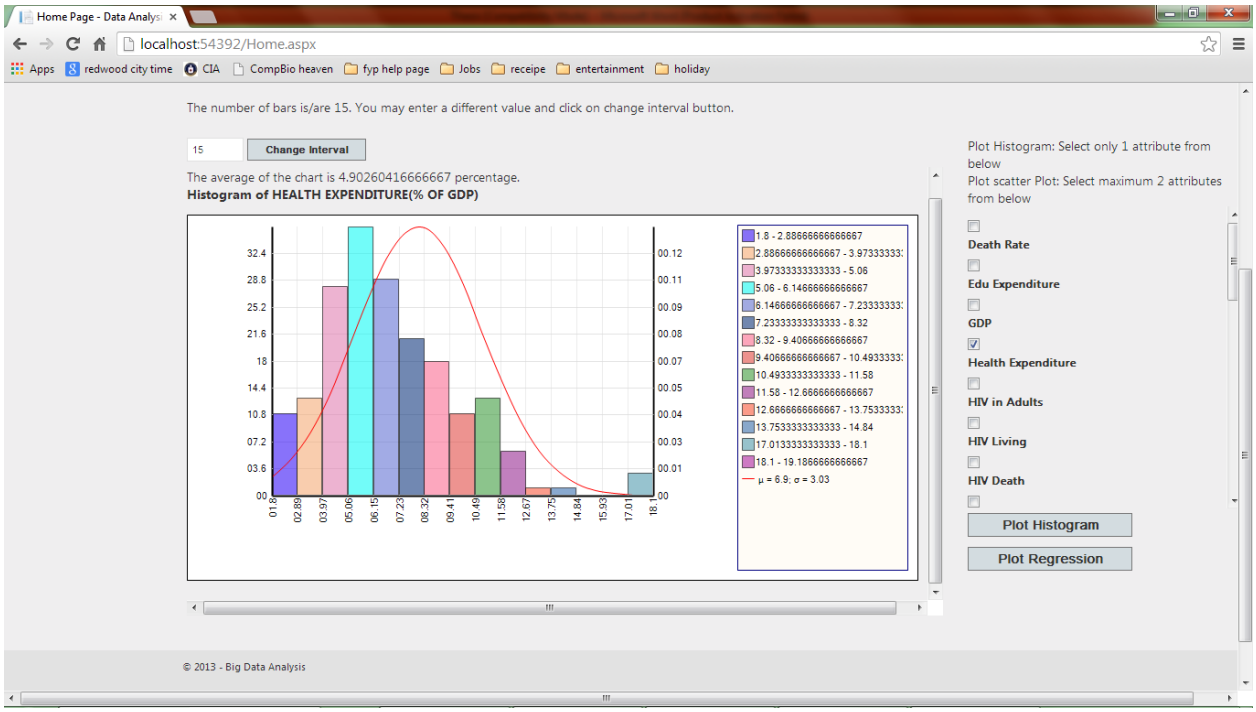


Figure 3.8. Histogram displays of Health Expenditure of CIA data with 15 bars

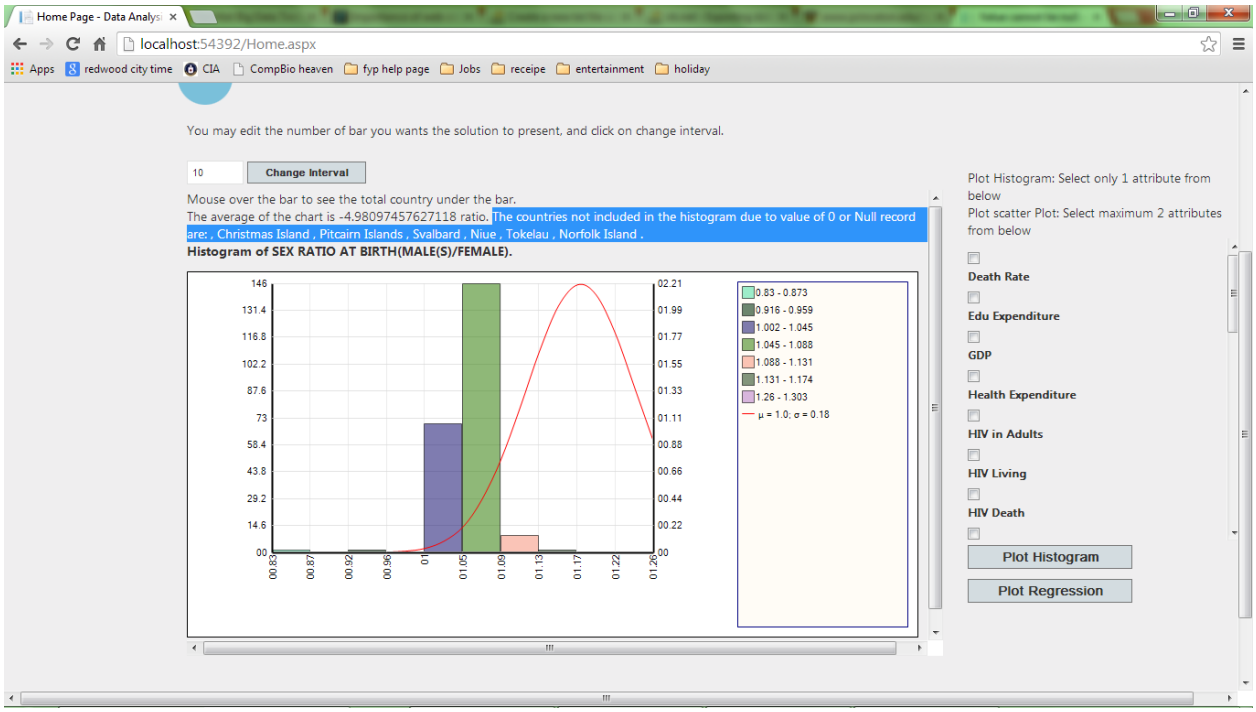


Figure 3.9. Histogram displays of sex ratio at birth based on CIA data with 10 bars and countries not involved in the plot displayed

Also, mentioned under the data of Materials and Methods, during data cleaning, those with true 0, null and NA were all assigned with a zero, however, was not taken into the plotting, else they will give a skewed histogram plot. Here is the fourth added feature to obtain more information, we use another attribute to illustrate in figure 3.9, similarly, also histogram plotting, however, a list of countries not involved in the plot was displayed as highlighted. Notice that for all case, the average calculated and the μ given are always different, this is because the μ is calculated by the smooth curve plot whereas the average displayed above the title is discrete value and with the consideration of removing the count of countries not involved in the plot. The partial code on plotting the histogram could be found at appendix E.

C. Suggestion for Future Development

A simpler suggestion on what to improve in this project is to have a set of pre-code statistical analysing methods, making use of EO and NetAdv together to generate more statistics options for the users to choose, this thus will meet the main driver reason of initialising this project. However, it will still be recommended to integrate R into the .Net platform as R is functional programming language, so if want to do the pre-code of statistical analysing methods will be much easier.

A tougher suggestion, also the ideal of this project is to have a 'wonder' button. That is to have dropdown list providing options to choose the data source, be it WHO or CIA factbook. After choosing, click on the button and it will do the magic of linking to the website of the chosen source and downloaded all the data and cleaned it. Then the data will be ready for user to use to check any sort of relationship or correlation.

IV. General Conclusion(s)

This project was so near to perfect completion, if from the start all the coding was not done separately. Also, if resources like MS VS 2010 registration key, if were requested earlier, probably that last minute, unforeseen technical issues would not have arisen. It was most likely the over-confident case, assuming with the online search with high respond of R and .Net works perfectly and event had the catch all exceptions troubleshooting website available, thinking that will eliminate the technical issue probability. However, after integrating R to .Net with the readiness of R(D)Com package and sciproxy.dll, did a couple of trials, R just do not work on MS VS 2010 or MS VS 2012. Every time when on load, it will catch and display this exceptional error “unable to load connector”. When refer this error to the R-installation help page that is only made available after installing the R(D)Com, C:\Program Files (x86)\R\(D)COM Server\doc\Install.html, which shows a list of possible errors and their solutions. However, many times I install and un-install, the R(D)Com just do not point to correct R.dll. I even went to the extent of changing the environment variables and path. Thus I concluded that probably the R(D)Com was not compatible with the latest R 3.0.2 version and my older R2.15.2 version, because on the R-installation webpage, it stated R2.7.0, which was the in between version.

However, it was a blessing to come across these two extension libraries, which will serves as a more powerful tools in complimentary to each other if more time permits to explore them. Meaning, using the strong statistical analysing abilities of Extreme Optimization and generate a set of values (because it is just a built-on library with no graphical solution) that

could be plotted into nice graphical illustration by NetAdv, the turn out interface will definitely be prettier and could work more hovering/tooltips/mouseover activity.

All in all, from this project, it allows me to revisits many learnt knowledge throughout the course of computational biology and the learning curve on software implementation is very challenging especially with all those namespaces, assemblies, libraries, binaries, … After this project, my learning will not end there, in fact I hope to actually finish up this project with the future development features mentioned in this write up.

V. References

Armonk, N.Y. (2013, October 29) IBM Analytics Study Reveals Big Data Equals Big Payoff. *IBM News Room*. Retrieved 2013, November 03, http://www-03.ibm.com/press/us/en/pressrelease/42326.wss

Ben Shneiderman and Catherine Plaisant 5th Ed (2009) Eight Golden Rules of Interface Design. *Designing the User Interface.* Published by Pearson Education

Incredible Art (n.d.) Color Symbolism and Culture. Incredible Art Department. Retrieved November 07, 2013, http://www.incredibleart.org/lessons/middle/color2.htm

International Business Machines, IBM (n.d.) About big data analytics. *International Business Machines, IBM website – What is Big Data Analytics?* Retrieved 2013, September 03, http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

International Data Corporation, IDC (2013) About IDC. *International Data Corporation, IDC website*. Retrieved 2013, September 03, http://www.idc.com/about/about.jsp

James Kobielus (2013, August 15) Big data means big challenges in lifecycle management. *InfoWorld – Big Data.* Retrieved November 01, 2013, http://www.infoworld.com/d/big-data/big-data-means-big-challenges-in-lifecycle-management-224828

James Taylor (2011, May) IBM's Big Data Platform and Decision Management Solutions. *International Business Machines, IBM website – Big Data at speed of Business.* Retrieved November 01, 2013, http://www-01.ibm.com/software/data/bigdata/

Margaret Rouse, Lisa Martinek (10 January 2012) Big data analytics. *Search Business Analytics*. Retrieved 2013, March 15, http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Philip Lee (2013, February 23) What is Big Data? The definition. TechRux: Trendies & Focus & Analysis. Retrieved 2013, September 03, http://techrux.net/big-data-definition/

QuickStart Samples (n.d.) Histograms QuickStart Sample (Visual Basic). *Extreme Optimization: Complexity made simple.* Retrieved November 01, 2013, http://www.extremeoptimization.com/QuickStart/VisualBasic/Histograms.aspx

Samina T. Yousuf Azeemi and S. Mohsin Raza (2005, December 02) A Critical Analysis of Chromotherapy and Its Scientific Evolution. *National Center for Biotechnology Information (NCBI) - PMC Journal website*. PMCID: PMC1297510. Volume 2(4), page 481–488.

Tom White (2012, May 10) Chapter 1: Meet Hadoop – Data. *Hadoop: The Definitive Guide* Published by O'Reilly Media, Inc.

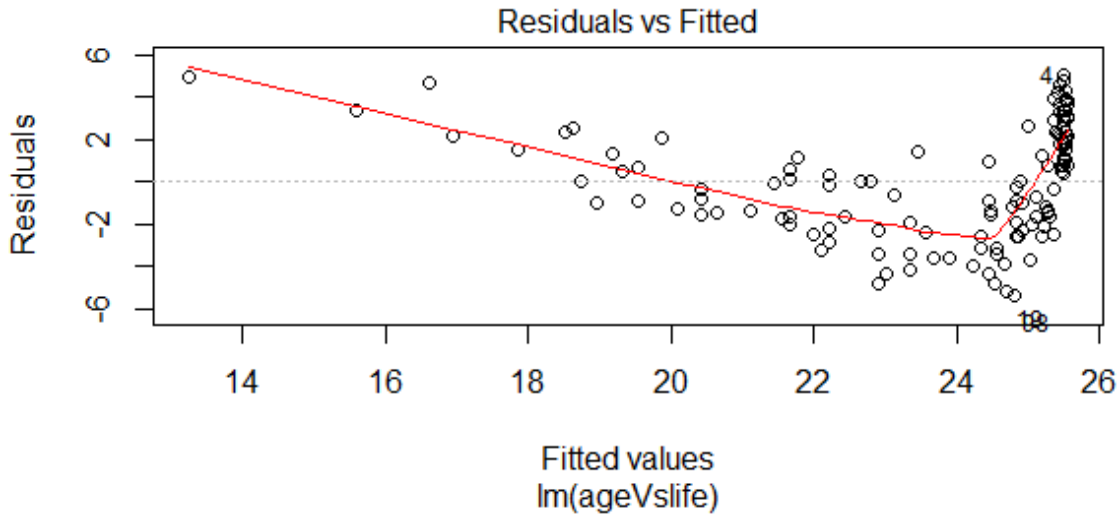
Wikipedia (last modified, 2013, October 31) Big Data. *Wikipedia, the free encyclopedia.* Retrieved November 01, 2013, http://en.wikipedia.org/wiki/Big_data

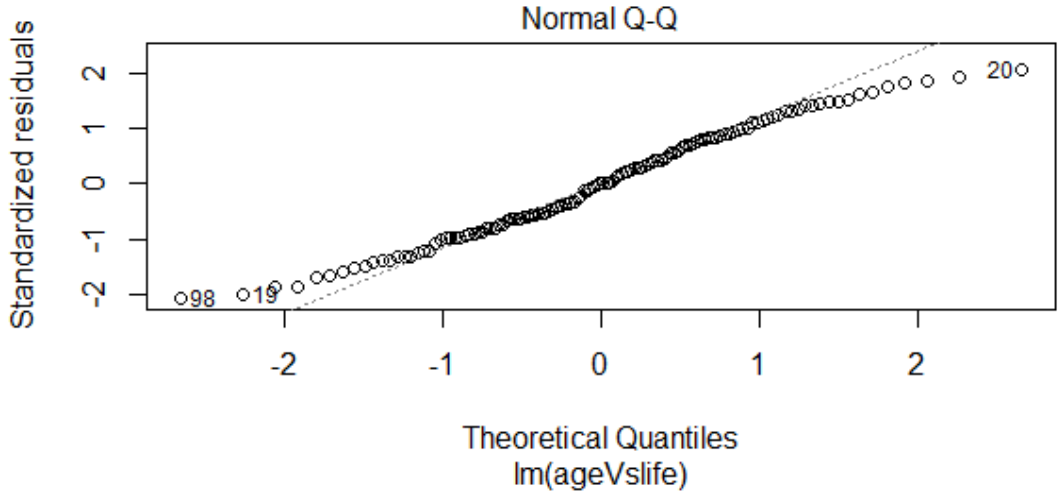
Wikipedia (last modified, 2013, October 30) Chromotherapy. *Wikipedia, the free encyclopedia.* Retrieved November 01, 2013, http://en.wikipedia.org/wiki/Chromotherapy

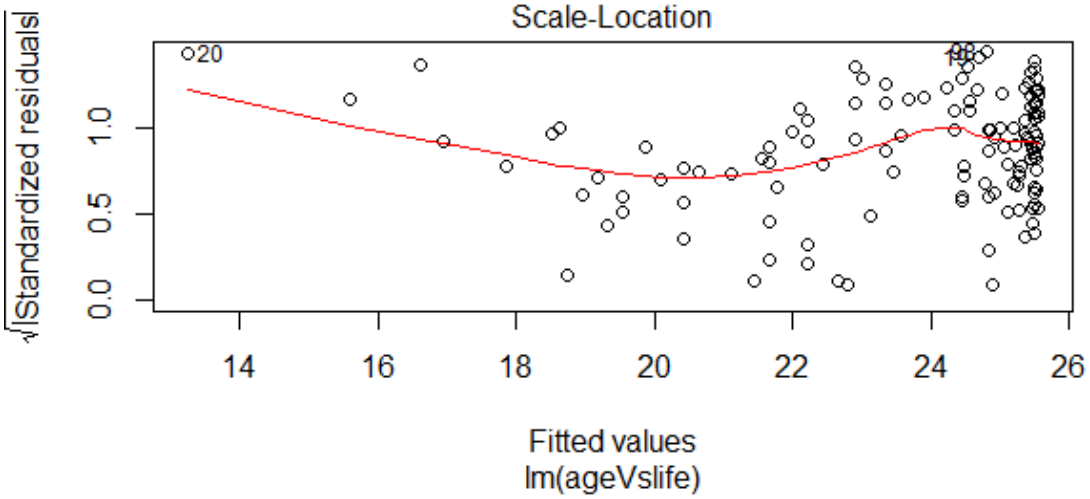
VI. Appendices

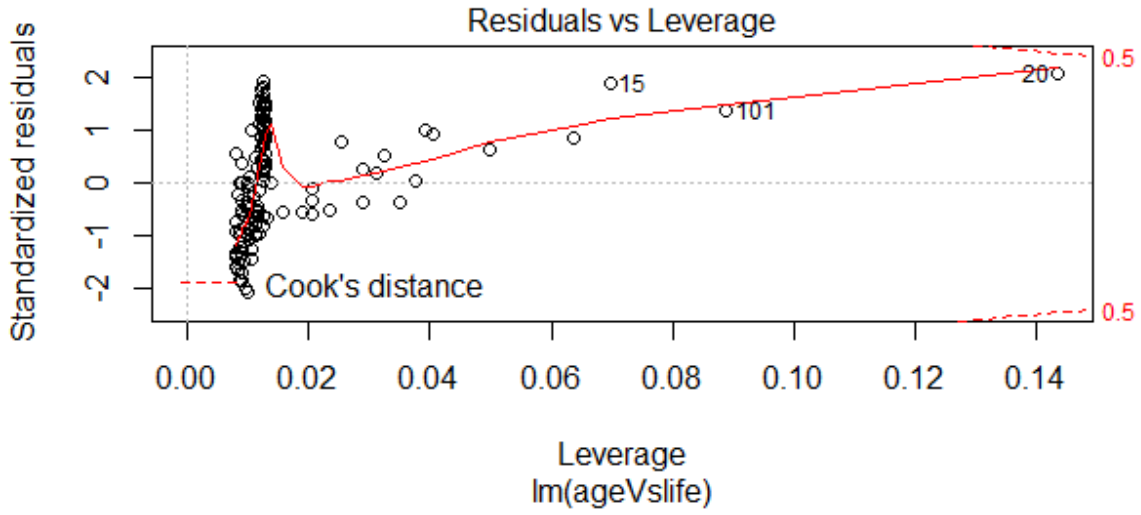
Appendix A: Data prep work, regression testing on MotherMeanAge and MotherMortalityRate data (from R)

```
ageVslife <- read.delim("C:/Users/Priscilla/Desktop/ageVslife.txt")

lm(ageVslife)

Call:
lm(formula = ageVslife)

Coefficients:
(Intercept)      MotherMortalityRate
   25.5810          -0.0112


plot(lm(ageVslife))
```









Appendix B: MotherMeanAge and MotherMortalityRate data

Mother Mean Age	Mother Mortality Rate
20.1	460
23.9	27
24.1	30
30.5	7
28.5	4
24.4	43
18.1	240
24.9	190
28	8
20	350
21.2	180
25.9	8
26.2	11
19.4	300
21.3	800
22.8	250
19.4	690
27.6	12
19.5	79
18.2	1100
23.7	25
21.4	92
20.2	540
19.8	560
19.7	400
27.7	17
27.5	10
27.6	5
29.1	12
20.3	150
21.8	110
22.9	66
20.8	81
20.6	240
26.3	2
19.6	350
27.9	5
28.6	8
18.7	230
23.9	67
28.9	7
21.8	350
29.2	3
20.3	120
18.8	610
20.8	280
22.2	350
20.1	100
28.2	21
27	5
19.9	200
22.5	220
29.8	6
27.3	7
27.7	4
21.2	110
29.4	5
24.9	63
27.6	51
19.8	360
29.6	16
23.6	71
26.4	34
21.2	620
19.1	770
26.6	8
29.3	20
26	10
19.5	240
18.9	460
23.9	60
18.6	540
26.5	8
21.9	510
21.3	50
23.5	41
26.3	8
25.4	100
18.8	490
21.4	200
20.1	170
28.9	6
27.7	15
19.7	95
18	590
20.9	630
28.4	7
22.7	260
21.1	92
22.9	99
22.3	67
23.1	99
26.6	5
27.4	8
26	27
22.9	340
23.6	100
19.4	70
21.4	370
27.2	12
19	890
29.4	3
27.3	6
28.7	12
22.5	300
29.3	6
22.6	35
19.5	320
28.6	4
30.2	8
22.3	65
19.6	460
23	48
22.1	300
20	300
22.9	20
24.6	67
18.9	310
23.1	32
30	12
25	21
23.8	28
22.6	59
19.2	200
19.2	440
20.5	570

Appendix C: Partial Codes on .vb script to get Database connection

```
Dim strCon As String
strCon = "Data Source=PRISCILLA-PC\SQLEXPRESS;Initial Catalog=FYPDB;Integrated
             Security=SSPI"
Dim conn As New SqlConnection(strCon)
Dim query As String = "SELECT * FROM TB_Attributes "
Try
       conn.Open()
       Dim cmd As New SqlCommand(query, conn)
       'MsgBox("Connection is successful")
       dt.Load(cmd.ExecuteReader())
       conn.Close()
       Catch ex As Exception
            MsgBox("Loading Data Connection is not successful")
End Try
```

Appendix D: Old lengthy code on mouseover event to each checkboxlist item

```
''' <summary>
    ''' This CB_Attribute_DataBound is to add tooltip to each item of the checkboxlist
    ''' however, has been simplified in SetTooltip()
    ''' </summary>
    Protected Sub CB_Attribute_DataBound(ByVal sender As Object, ByVal e As EventArgs)
        For Each item As ListItem In CB_Attributes.Items
            item.Attributes("id") = Guid.NewGuid().ToString()
            Dim decr As String = item.Text
            Dim rView As DataRow() = dt.[Select]("Attributes='" & decr & "'")
            Dim str As String = rView(0)("Remarks").ToString()_
            item.Attributes.Add("onmouseover", ("showtooltip('" &
item.Attributes("id") & "','") + str & "')")
        Next
    End Sub
```

Appendix E: Histogram plotting code (Using NetAdv package)

```
''' <summary>
    ''' This PlotHistogram(Integer) requires intervalValue
    ''' </summary>
    Protected Function PlotHistogram(intervalValue As Integer)
        lbl_plotGraph.Text = "Btn_PlotH"

        'find the extreme value
        Dim min_Avalue, max_Avalue, interval, sum, average As Double
        Dim z As Integer
        Dim CountriesNotPlotStr As String
        ReDim CountryNotPlot(dbDataA.Count - 2)

        min_Avalue = dbDataA.Min
        max_Avalue = dbDataA.Max
        sum = dbDataA.Sum
        z = 0

        If min_Avalue = 0 Then
            min_Avalue = max_Avalue
            For y As Integer = 1 To dbDataA.Count - 2
                'sum = sum + dbDataA(y)
                If dbDataA(y) = 0 And dbDataCountry(y) <> "" Then
                    CountryNotPlot(z) = dbDataCountry(y)
                    z = z + 1
                End If
                If dbDataA(y) < min_Avalue And dbDataA(y) > 0 Then
                    min_Avalue = dbDataA(y)
                End If
                If dbDataA(y) > max_Avalue Then
                    max_Avalue = dbDataA(y)
                End If
            Next
        End If


        If CountryNotPlot(0) <> "" Then
            average = sum / dbDataA.Count - z
            CountriesNotPlotStr = " The countries not included in the histogram due to
value of 0 or Null record are: "
            For i As Integer = 0 To z - 1
                CountriesNotPlotStr = CountriesNotPlotStr + ", " & CountryNotPlot(i)
            Next
            CountriesNotPlotStr = CountriesNotPlotStr + " . "
            lbl_GraphDataInfo.Text = "The average of the chart is " & average.ToString
& " " & parameter_unit(0) & ". " & CountriesNotPlotStr
        Else
            average = sum / dbDataA.Count - 2
            lbl_GraphDataInfo.Text = "The average of the chart is " & average.ToString
& " " & parameter_unit(0) & ". "
        End If

        'Count the interval on x-axis
        If max_Avalue - min_Avalue Then
            interval = (max_Avalue - min_Avalue) / intervalValue
        End If

        'Dim ultraChart1 As New UltraChart
        UltraChart1.Visible = True
        UltraChart1.ChartType = ChartType.HistogramChart
```

```
UltraChart1.Data.DataSource = dbDataA
UltraChart1.Data.DataBind()
UltraChart1.HistogramChart.ColumnAppearance.StringAxis = False
UltraChart1.HistogramChart.ColumnAppearance.ShowInLegend = True
UltraChart1.Width = 800
UltraChart1.Height = 400
UltraChart1.Axis.X.TickmarkStyle = AxisTickStyle.DataInterval
UltraChart1.Axis.X.TickmarkInterval = interval
UltraChart1.Axis.X.RangeMin = min_Avalue
UltraChart1.Axis.X.RangeMax = max_Avalue
UltraChart1.Axis.X.RangeType = AxisRangeType.Custom
UltraChart1.Axis.X.Labels.Orientation = TextOrientation.VerticalLeftFacing
UltraChart1.Legend.Visible = True

End Function
```