PROJECT STAGE 2 RESEARCH REPORT

Project Topic: Delineating the Causes Behind the Overall Socioeconomic Performance of Countries in the G20

DATA1002 – Informatics: Data & Computation

Report Overview

This report aims to provide detailed information to a variety of audiences based on the interests of each individual group. Throughout the report, all audiences will be able to identify the underlying theme of the socioeconomic performance of countries within the G20 as this concept is relevant to each of the three sections. Each section has been carefully selected to address the interests of a different type of audience, with the first section focusing on a general audience who is interested in the domain that is the socioeconomic performance of countries within the G20. The following section focuses on audiences who have a great interest in data analysis, more specifically through various tools such as tables and figures. Finally, the third and last section is an extension of the previous section as it still focuses on data analysis, but now through the use of predictive models.

Section 1 which has been produced for individuals with a general interest, will give insight into the relationship between various indicators such as diseases and government influence, personal health and freedom of trade and finally, population wellbeing and sound money in determining the socioeconomic performance of countries within the G20 (Argentina, Australia, Brazil, Canada, China, Germany, France, India, Indonesia, Italy, Japan, Mexico, Russia, Saudi Arabia, South Africa, South Korea, Turkey, UK & US – The EU is the 20th member of the G20 but has not been included in this report as the focus is on individual countries rather than a body which has multiple member states). The socioeconomic performance of a country has been measured through three distinct indicators which are the socioeconomic status (SES), GDP per capita (GDPpc) and years of education attained by the population of each country.

Section 2 which has been created for individuals with a specific interest in data analysis through various tools such as tables and figures, will provide individuals with an insight into the coding undertaken to produce final graphs that are vital in data analysis. This section will not only discuss the methodology but will also outline the thought process behind the selection of the specific processes that were utilised to create the figures utilised in the data analysis seen in the first section.

Section 3 is the final part of this report and focuses on the development and execution of predictive models including relevant settings and training data utilised. This section has been included for audiences with a specific interest in analysing data through the use of machine learning.

It is also important to note that while this report focuses on countries within the G20, the data analysis undertaken will also demonstrate the performance of countries around the world that are not members of the G20 to provide a deeper understanding of the relationship between different factors within countries in and out of a forum like the G20.

Section 1: General Audiences

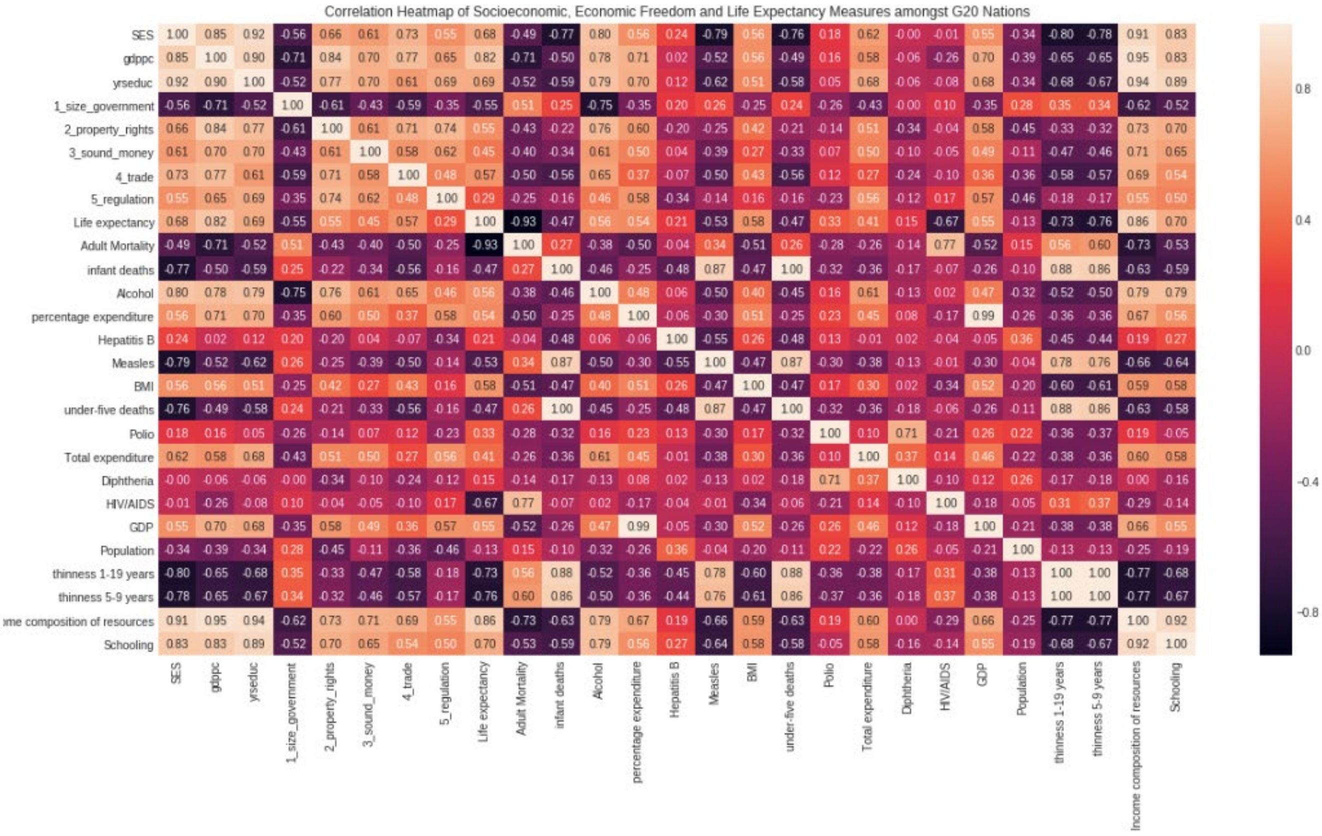
Socioeconomic Status in Relation to Diseases and Government Influence

Topic Overview

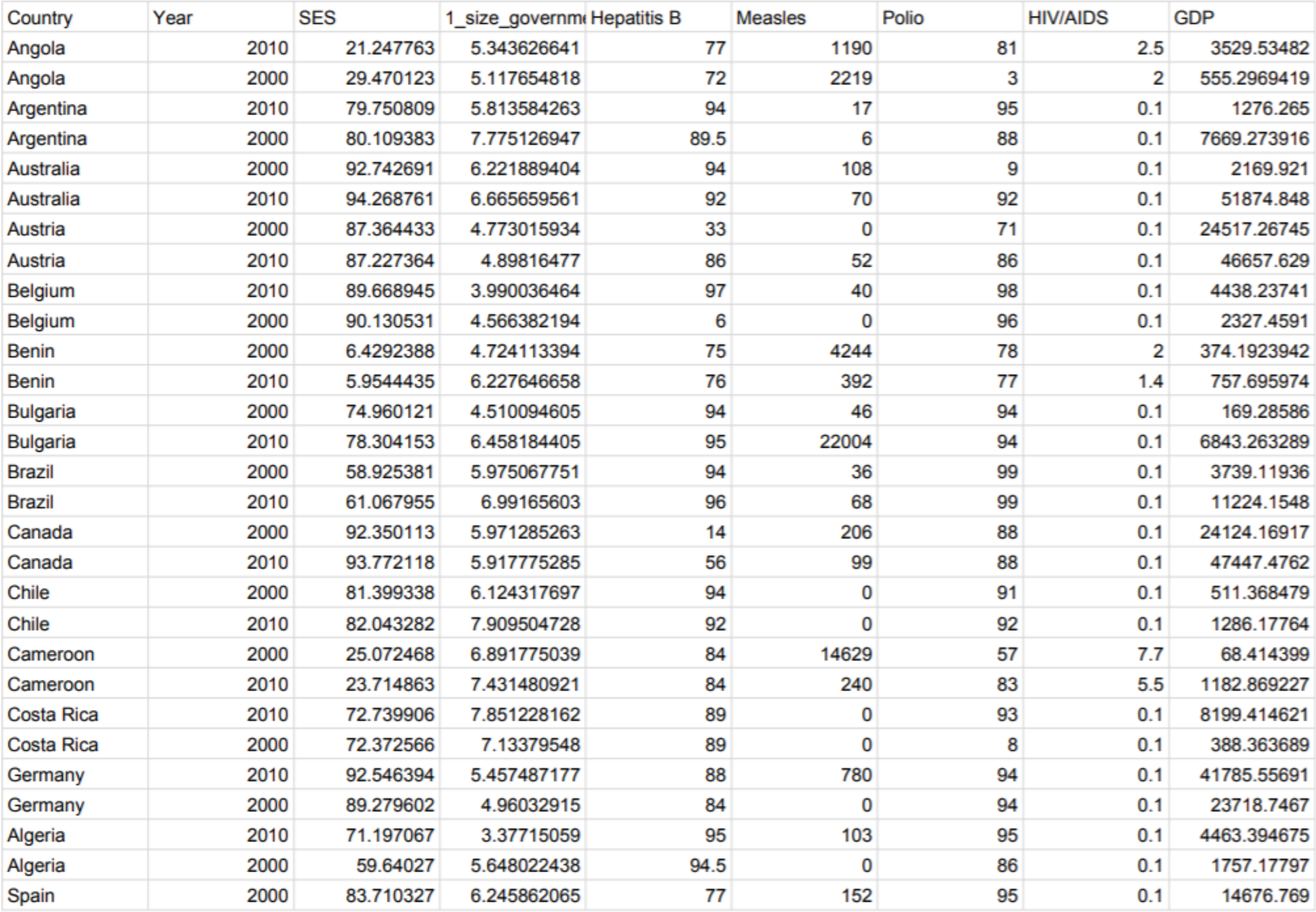
An analysis of the socioeconomic status of a country in relation to diseases and government influence was undertaken in an attempt to determine if there was any correlation between the socioeconomic status and government influence of countries and the prevalence of diseases in those nations specifically, the members of the G20. The indicator utilised to measure the socioeconomic status of a country is a score that is provided by various sources such as the OECD. While there are many different diseases currently affecting many individuals around the world, this report focuses on data relating to the prevalence of Hepatitis B, Measles and HIV/AIDS. The concept of government influence is described through an indicator of the size of the government of a country with this measure including various factors such as, government consumption, government transfers and subsidies, government enterprises and investment and top marginal tax rate. All the data discussed in this topic is from 2000 and 2010, with these years being selected to account for missing entries in the data available for other years.

The data analysed in this topic was obtained from three different datasets. For data on various diseases, a dataset from Kumar Rajarshi, who obtained the raw data from the World Health Organisation (WHO) was included. Data on the influence of government was accessed from Guillermina Sutter Schneider, who obtained the raw data from the Economic Freedom of the World: 2018 Annual Report which is published by the Fraser Institute. Finally, the dataset containing data points on the socioeconomic status of different countries was compiled and published by Shawn Dorius, who obtained the raw data from a variety of sources including the the 2004 publication of "The World Economy: Historical Statistics" by the Organisation for Economic Co-operation and Development (OECD).

For the analysis of the data, we utilised various different tables and figures including tables organising the raw data into categories, a correlation heatmap which outlined the correlation between all the indicators that were available in the dataset and scatter diagrams to represent the relationship between specific diseases, government influence and the socioeconomic status of countries. The correlation heatmap can be found below along with a small snapshot (due to a large amount of data points, it is inefficient to attach the entire table) of the table utilised to organise the data for both G20 and other countries around the world.



Correlation Heatmap of Socioeconomic, Economic Freedom and Life Expectancy Measures amongst G20 Nations

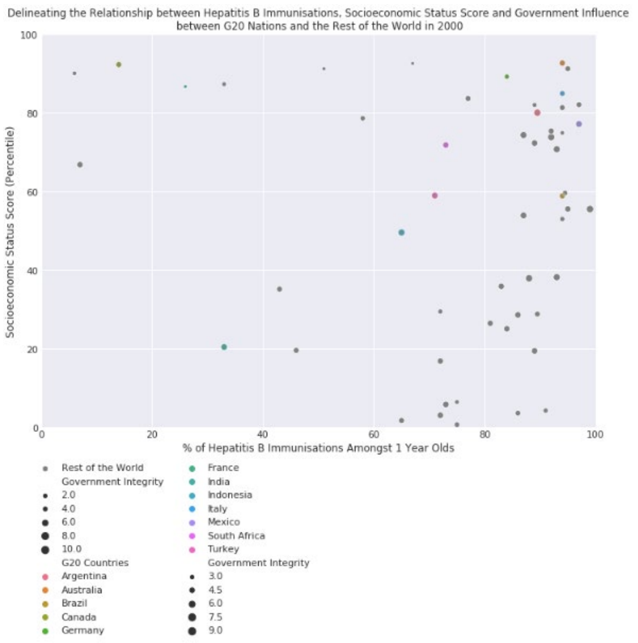


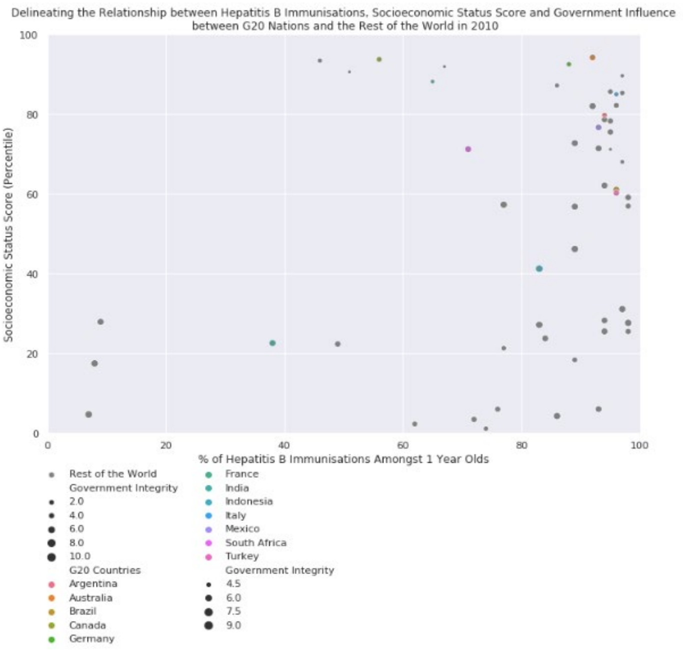
Country	Year	SES	1_size_governm	Hepatitis B	Measles	Polio	HIV/AIDS	GDP
Angola	2010	21.247763	5.343626641	77	1190	81	2.5	3529.53482
Angola	2000	29.470123	5.117654818	72	2219	3	2	555.2969419
Argentina	2010	79.750809	5.813584263	94	17	95	0.1	1276.265
Argentina	2000	80.109383	7.775126947	89.5	6	88	0.1	7669.273916
Australia	2000	92.742691	6.221889404	94	108	9	0.1	2169.921
Australia	2010	94.268761	6.665659561	92	70	92	0.1	51874.848
Austria	2000	87.364433	4.773015934	33	0	71	0.1	24517.26745
Austria	2010	87.227364	4.89816477	86	52	86	0.1	46657.629
Belgium	2010	89.668945	3.990036464	97	40	98	0.1	4438.23741
Belgium	2000	90.130531	4.566382194	6	0	96	0.1	2327.4591
Benin	2000	6.4292388	4.724113394	75	4244	78	2	374.1923942
Benin	2010	5.9544435	6.227646658	76	392	77	1.4	757.695974
Bulgaria	2000	74.960121	4.510094605	94	46	94	0.1	169.28586
Bulgaria	2010	78.304153	6.458184405	95	22004	94	0.1	6843.263289
Brazil	2000	58.925381	5.975067751	94	36	99	0.1	3739.11936
Brazil	2010	61.067955	6.99165603	96	68	99	0.1	11224.1548
Canada	2000	92.350113	5.971285263	14	206	88	0.1	24124.16917
Canada	2010	93.772118	5.917775285	56	99	88	0.1	47447.4762
Chile	2000	81.399338	6.124317697	94	0	91	0.1	511.368479
Chile	2010	82.043282	7.909504728	92	0	92	0.1	1286.17764
Cameroon	2000	25.072468	6.891775039	84	14629	57	7.7	68.414399
Cameroon	2010	23.714863	7.431480921	84	240	83	5.5	1182.869227
Costa Rica	2010	72.739906	7.851228162	89	0	93	0.1	8199.414621
Costa Rica	2000	72.372566	7.13379548	89	0	8	0.1	388.363689
Germany	2010	92.546394	5.457487177	88	780	94	0.1	41785.55691
Germany	2000	89.279602	4.96032915	84	0	94	0.1	23718.7467
Algeria	2010	71.197067	3.37715059	95	103	95	0.1	4463.394675
Algeria	2000	59.64027	5.648022438	94.5	0	86	0.1	1757.17797
Spain	2000	83.710327	6.245862065	77	152	95	0.1	14676.769

An in-depth analysis of various different figures produced from the original dataset was undertaken, which resulted in the discovery of key findings which are further discussed in the following report segments

SES in Relation to Hepatitis B and Government Influence

Hepatitis B (HBV) is a viral infection that has been identified to attack the liver with advance stages resulting in both acute and chronic disease. While the virus is most commonly transmitted genetically during birth and delivery, it can be prevented by vaccines that are not only safe but easily available. As a result, we have identified this disease as an indicator of a country's socioeconomic status as the number of immunisations for Hepatitis B is directly related to the influence of governments in raising awareness and implementing processes to allow easy access to vaccines for the general population. Secondary research through a variety of sources including the World Health Organisation has revealed that the reduction in the prevalence of chronic HBV infection in children under 5 to 1.3% in 2015 has been the result of the widespread use of the Hepatitis B vaccine which is the mainstay of Hepatitis B prevention. Subsequently, a direct link has been developed with the assumption being that countries with a greater government influence (indicated by the size of data points in the graphs below) will have higher Hepatitis B immunisations amongst infants and subsequently a higher socioeconomic status.



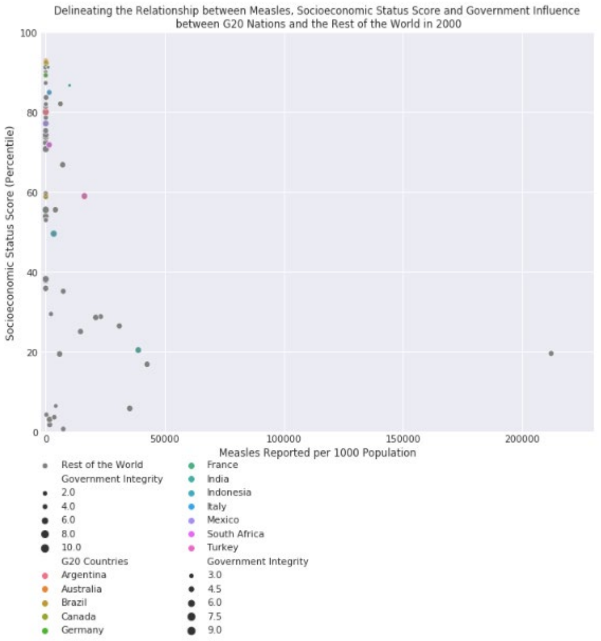


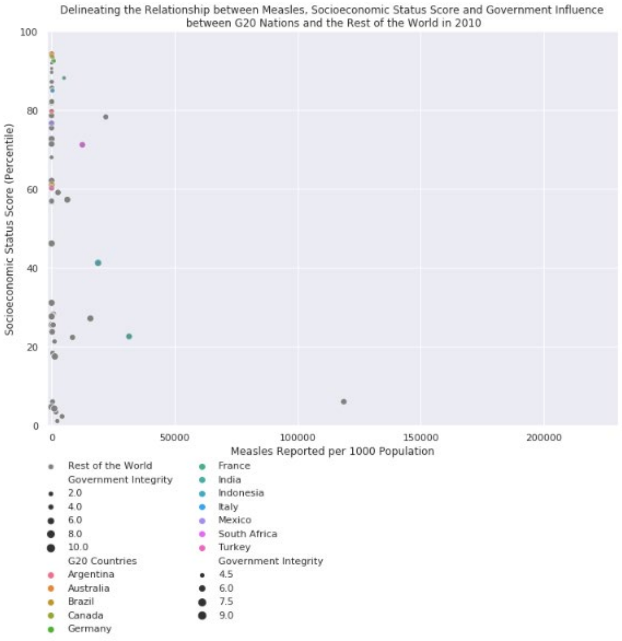
The scatter diagrams above indicate a relationship between Hepatitis B immunisations, socioeconomic status scores and government influence that is not in support of the original hypothesis that countries with greater government influence and Hepatitis B immunisations will also have higher SES scores. This can be seen through the wide range of data points on the graphs above which reveal no strong relationship between any of the factors as countries such as Argentina who have low government influence still have high percentages of Hepatitis B immunisations and SES scores. This analysis from the scatter graphs above is reinforced through an analysis of the correlation heatmap with the correlation between a country's SES score and Hepatitis B immunisations being only 0.24 and the correlation between government influence and Hepatitis B immunisations even lower at just 0.20. Subsequently, it is clear that whilst various studies such as Tosun, S., Aygün, O., Özdemir, H.Ö. et al (2018) suggest that there is a strong relationship between the factors discussed in this section, the data utilised in this report do not identify any such correlations between SES scores, Hepatitis B immunisations and government influence which is identified through the size of a government. It is also important to note that while the majority of countries have improved the percentage of Hepatitis B immunisations amongst 1-year old's, immunisations in some countries have actually declined. Through closer analysis it is evident that these countries are not actually a part of the G20 and rather, almost all the countries in the G20 have either improved or maintained the percentage of Hepatitis B immunisations between 2000 and 2010. While there are natural assumptions that can be made into the effectiveness of G20 nations in relation to Hepatitis B immunisations, further research and analysis must be undertaken to clearly identify the underlying factors that have allowed member states of the G20 to maintain or improve Hepatitis B immunisations.

In summary, the analysis of the relationship between Hepatitis B immunisations, SES scores and government influence has revealed that while there is no strong relationship between these three factors, nations in the G20 have been able to improve or maintain immunisation rates during the same 10-year period in which the percentages of immunisations in other countries around the world have declined resulting in the development of further points of research which can be undertaken by the audience of this section as they also hold a great interest in this general domain.

SES in Relation to Measles and Government Influence

Measles is a highly contagious viral disease which despite the availability of safe and effective vaccination options, continues to be a prevalent cause of death among young children globally. While measles is transmitted via droplets from various parts of an infected person such as their mouth or nose, initial symptoms do not appear till 10-12 days later which becomes more dangerous for poorly nourished children, who are also the most likely victim. Just like Hepatitis B, we identified measles as an indicator of a country's socioeconomic status as the deaths of individuals can be prevented through proper vaccination which is directly related to the influence of a country's government. Research conducted by the World Health Organisation has revealed that global deaths due to measles have decreased by 84% with 89,780 deaths reported in 2016 as opposed to 550,100 deaths reported in 2000. As a result, a hypothesis that countries with higher levels of government influence and greater SES scores will have lower reports of measles. Similar to the data discussed for Hepatitis B, government influence is indicated by the size of data points in the graphs below with SES scores and reports of measles per 1000 individuals are the axis of the graph.



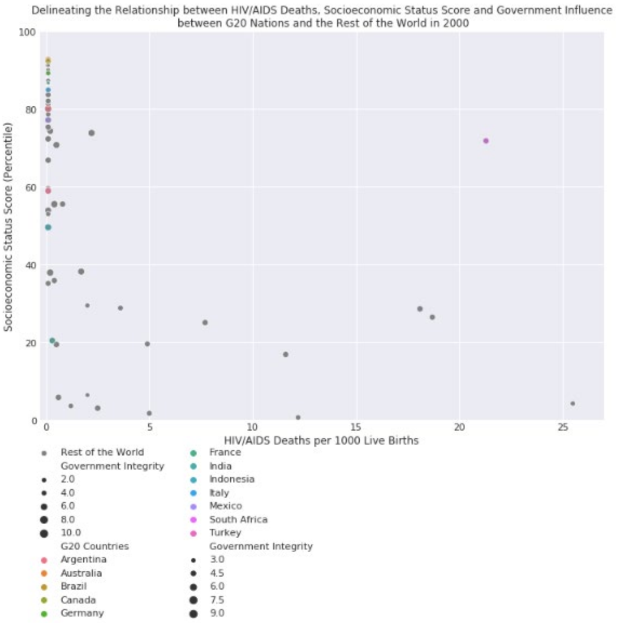


The scatter diagrams above indicate a relationship between reports of measles, socioeconomic status scores and government influence that is not in support of the original hypothesis that countries with greater government influence and Hepatitis B immunisations will also have higher SES scores. Through a close analysis of the graphs above, it is evident that while there isn't a large spread in the reporting of measles across countries in the G20, and around the world, the SES score varies greatly. This observation is reinforced through the correlation heatmap discussed earlier which reveals that there is in-fact a negative correlation of 0.79 between the SES score and reporting of measles. This clearly debunks the hypothesis as not all countries who have low reports of measles also have high SES scores. In contrast, the correlation between the size of government and measle reporting is 0.26 which means that measle reporting and government influence share a positive relationship which is much stronger than the relationship between SES scores and measle reports. As a result, similar to Hepatitis B, findings in studies conducted on the association between socioeconomic status and measles such as Minh, Giang, Mai, Kien, Tuan, Quam (2016) and Wilson, Ducharme, Hawken (2013) are not supported by the data analysed in this report. Further, there was no real clear disparity between G20 and other nations and so a point of further research could be trying to determine if there is a relationship between countries in specific geographical regions rather than in specific forums. As a result, there are various points that require greater research to help determine the true causes of the findings seen on the graphs above.

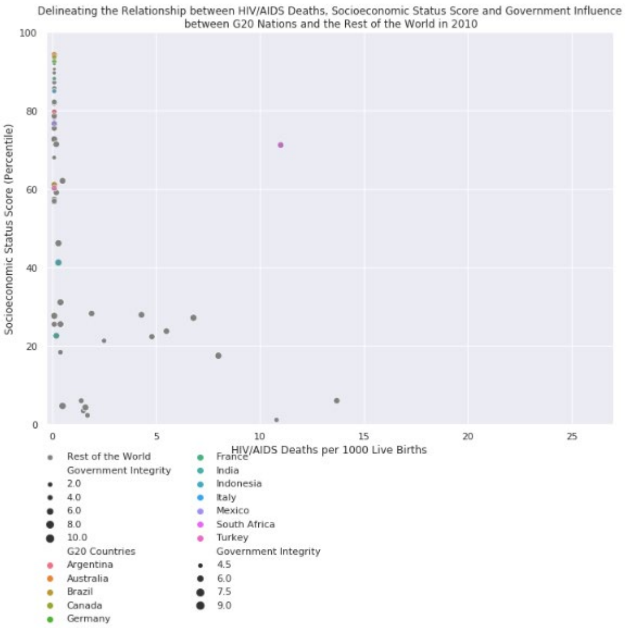
In summary, the analysis of the relationship between measle reports, SES scores and government influence has revealed that while there is no strong relationship between these three factors. Further while there is no clear evidence that countries in the G20 are dealing with measles better than nations not in the G20, it is clear that all countries around the world have either reduced or maintained the number of measle reports which means that the global community as a whole is effectively dealing with the issue of measles. However, due to no clear relationships and links being able to be identified, further research into areas such as the influence of geographical regions on the prevalence of measles can be undertaken to try understand and identify the negative relationship that was seen between SES scores and reports of measles.

SES in Relation to HIV/AIDS and Government Influence

The Human Immunodeficiency Virus (HIV) is a virus that can lead to the development of the condition, Acquired Immunodeficiency Syndrome (AIDS) which is why they are measured together as HIV/AIDS. HIV/AIDS targets the immune system, weakening individual's defence systems against infections such as cancers which can prove to be deadly for the person. While there is no cure/vaccination for HIV/AIDS, the risk of contracting the virus can be greatly reduced by limiting exposure to risk factors which comes through education and awareness. As the government influence of a country can impact the education and awareness of the population, we selected HIV/AIDS as a factor when measuring the relationship between government influence, diseases and SES scores. Through further research, we were able to identify that key populations often faced legal and social barriers that increased their vulnerability to HIV and also impeded their access to prevention, testing and treatment programmes, which resulted in the development of the hypothesis that countries with higher deaths due to HIV/AIDS would also have lower SES scores. This hypothesis is explored through an analysis of the graphs below which include data points for government influence (indicated by the size of data points), SES scores and HIV/AIDS deaths per 1000 live births.



Delineating the Relationship between HIV/AIDS Deaths, Socioeconomic Status Score and Government Influence between G20 Nations and the Rest of the World in 2000



Delineating the Relationship between HIV/AIDS Deaths, Socioeconomic Status Score and Government Influence between G20 Nations and the Rest of the World in 2010

By analysing the graphs above, it is clear that over the 10-year period between 2000 and 2010, there has been a decrease in the number of HIV/AIDS deaths reported around the world for countries in and out of the G20 which is a great positive for the wider global community. Further analysis of the two graphs and other figures such as the correlation heatmap reveal that the initial hypothesis is actually false as there is a large spread in the SES score of countries who have similar if not identical reports of deaths caused by HIV/AIDS. The graphs above visually represent this whilst the correlation heatmap reveals that the relationship between SES scores and the number of HIV/AIDS deaths is -0.01 which means that there is almost no correlation between the two factors. Furthermore, the relationship between HIV/AIDS and government influence is revealed to be 0.10 which is stronger than the relationship between HIV/AIDS and SES scores but is still very weak. Interestingly, these findings are echoed in various studies such as Bunyasi, Coetzee (2017) which state that there is a mixed association between SES scores and the prevalence of HIV/AIDS resulting in the deaths of individuals. With there being a wide spread between the G20 countries, it is also clear that there is little correlation between the membership of a country to a particular forum. As a result other points of research such as influences of SES in specific geographic regions need to be addressed before clear conclusions can be reached.

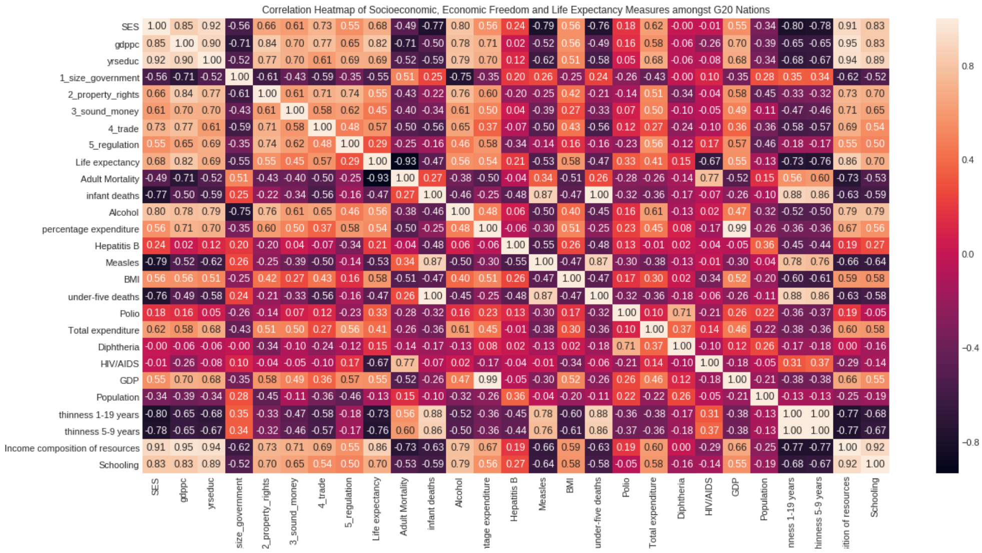
In summary, the analysis of the relationship between HIV/AIDS deaths, SES scores and government influence has revealed that while there is no strong relationship between HIV/AIDS deaths and government size and almost no correlation between HIV/AIDS deaths and SES scores. As a result audiences who are interested in this domain may look to further investigate the factors influencing SES scores to better understand the lack of correlation between SES scores and HIV/AIDS deaths.

GPD Per Capita in Relation to Personal Health and Freedom of Trade

Topic Overview

The timely series on trade and health provides an analysis to determine if the economic freedom or life indicators are strong determinants of socioeconomic health in G20 nations. Comparing the Indicators of Personal Health (Adult Mortality, Infant Deaths, BMI, Alcohol) with the Country's Freedom of Trade in relation to its Gross Domestic Product per capita (GDPpc), we witness the correlations between each indicator with respective trade and their policies. All data examined in this topic is within years 2000 to 2010 due to missing entries in data of other years.

The data used for this topic was acquired from three distinctive datasets. For data on a country's freedom of trade, a dataset from Guillermina Sutter Schneider, who assimilated the raw data from the Economic Freedom of the World 2018 Annual Report published by the Fraser Institute was accessed. On the other hand, the data used for the indicators of health was obtained from Kumar Rajarshi, who attained the raw data from the World Health Organisation (WHO) was utilised for this topic. Conclusively, the dataset encompassing measurements of GDP per capita published by Shawn Dorius, who acquired the raw data from sources comprising the 2004 publication of "The World Economy: Historical Statistics" by the Organisation for Economic Co-operation and Development (OECD) was utilised in scope of the topic.



Correlation Heatmap of Socioeconomic, Economic Freedom and Life Expectancy Measures amongst G20 Nations

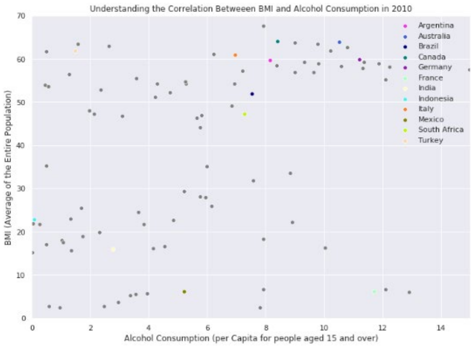
	SES	gdppc	yrseduc	size_government	2_property_rights	3_sound_money	4_trade	5_regulation	Life expectancy	Adult Mortality	infant deaths	Alcohol	ntage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	nness 1-19 years	hinness 5-9 years	ition of resources	Schooling
SES	1.00	0.85	0.92	-0.56	0.66	0.61	0.73	0.55	0.68	-0.49	-0.77	0.80	0.56	0.24	-0.79	0.56	-0.76	0.18	0.62	-0.00	-0.01	0.55	-0.34	-0.80	-0.78	0.91	0.83
gdppc	0.85	1.00	0.90	-0.71	0.84	0.70	0.77	0.65	0.82	-0.71	-0.50	0.78	0.71	0.02	-0.52	0.56	-0.49	0.16	0.58	-0.06	-0.26	0.70	-0.39	-0.65	-0.65	0.95	0.83
yrseduc	0.92	0.90	1.00	-0.52	0.77	0.70	0.61	0.69	0.69	-0.52	-0.59	0.79	0.70	0.12	-0.62	0.51	-0.58	0.05	0.68	-0.06	-0.08	0.68	-0.34	-0.68	-0.67	0.94	0.89
1_size_government	-0.56	-0.71	-0.52	1.00	-0.61	-0.43	-0.59	-0.35	-0.55	0.51	0.25	-0.75	-0.35	0.20	0.26	-0.25	0.24	-0.26	-0.43	-0.00	0.10	-0.35	0.28	0.35	0.34	-0.62	-0.52
2_property_rights	0.66	0.84	0.77	-0.61	1.00	0.61	0.71	0.74	0.55	-0.43	-0.22	0.76	0.60	-0.20	-0.25	0.42	-0.21	-0.14	0.51	-0.34	-0.04	0.58	-0.45	-0.33	-0.32	0.73	0.70
3_sound_money	0.61	0.70	0.70	-0.43	0.61	1.00	0.58	0.62	0.45	-0.40	-0.34	0.61	0.50	0.04	-0.39	0.27	-0.33	0.07	0.50	-0.10	-0.05	0.49	-0.11	-0.47	-0.46	0.71	0.65
4_trade	0.73	0.77	0.61	-0.59	0.71	0.58	1.00	0.48	0.57	-0.50	-0.56	0.65	0.37	-0.07	-0.50	0.43	-0.56	0.12	0.27	-0.24	-0.10	0.36	-0.36	-0.58	-0.57	0.69	0.54
5_regulation	0.55	0.65	0.69	-0.35	0.74	0.62	0.48	1.00	0.29	-0.25	-0.16	0.46	0.58	-0.34	-0.14	0.16	-0.16	-0.23	0.56	-0.12	0.17	0.57	-0.46	-0.18	-0.17	0.55	0.50
Life expectancy	0.68	0.82	0.69	-0.55	0.55	0.45	0.57	0.29	1.00	-0.93	-0.47	0.56	0.54	0.21	-0.53	0.58	-0.47	0.33	0.41	0.15	-0.67	0.55	-0.13	-0.73	-0.76	0.86	0.70
Adult Mortality	-0.49	-0.71	-0.52	0.51	-0.43	-0.40	-0.50	-0.25	-0.93	1.00	0.27	-0.38	-0.50	-0.04	0.34	-0.51	0.26	-0.28	-0.26	-0.14	0.77	-0.52	0.15	0.56	0.60	-0.73	-0.53
infant deaths	-0.77	-0.50	-0.59	0.25	-0.22	-0.34	-0.56	-0.16	-0.47	0.27	1.00	-0.46	-0.25	-0.48	0.87	-0.47	1.00	-0.32	-0.36	-0.17	-0.07	-0.26	-0.10	0.88	0.86	-0.63	-0.59
Alcohol	0.80	0.78	0.79	-0.75	0.76	0.61	0.65	0.46	0.56	-0.38	-0.46	1.00	0.48	0.06	-0.50	0.40	-0.45	0.16	0.61	-0.13	0.02	0.47	-0.32	-0.52	-0.50	0.79	0.79
percentage expenditure	0.56	0.71	0.70	-0.35	0.60	0.50	0.37	0.58	0.54	-0.50	-0.25	0.48	1.00	-0.06	-0.30	0.51	-0.25	0.23	0.45	0.08	-0.17	0.99	-0.26	-0.36	-0.36	0.67	0.56
Hepatitis B	0.24	0.02	0.12	0.20	-0.20	0.04	-0.07	-0.34	0.21	-0.04	-0.48	0.06	-0.06	1.00	-0.55	0.26	-0.48	0.13	-0.01	0.02	-0.04	-0.05	0.36	-0.45	-0.44	0.19	0.27
Measles	-0.79	-0.52	-0.62	0.26	-0.25	-0.39	-0.50	-0.14	-0.53	0.34	0.87	-0.50	-0.30	-0.55	1.00	-0.47	0.87	-0.30	-0.38	-0.13	-0.01	-0.30	-0.04	0.78	0.76	-0.66	-0.64
BMI	0.56	0.56	0.51	-0.25	0.42	0.27	0.43	0.16	0.58	-0.51	-0.47	0.40	0.51	0.26	-0.47	1.00	-0.47	0.17	0.30	0.02	-0.34	0.52	-0.20	-0.60	-0.61	0.59	0.58
under-five deaths	-0.76	-0.49	-0.58	0.24	-0.21	-0.33	-0.56	-0.16	-0.47	0.26	1.00	-0.45	-0.25	-0.48	0.87	-0.47	1.00	-0.32	-0.36	-0.18	-0.06	-0.26	-0.11	0.88	0.86	-0.63	-0.58
Polio	0.18	0.16	0.05	-0.26	-0.14	0.07	0.12	-0.23	0.33	-0.28	-0.32	0.16	0.23	0.13	-0.30	0.17	-0.32	1.00	0.10	0.71	-0.21	0.26	0.22	-0.36	-0.37	0.19	-0.05
Total expenditure	0.62	0.58	0.68	-0.43	0.51	0.50	0.27	0.56	0.41	-0.26	-0.36	0.61	0.45	-0.01	-0.38	0.30	-0.36	0.10	1.00	0.37	0.14	0.46	-0.22	-0.38	-0.36	0.60	0.58
Diphtheria	-0.00	-0.06	-0.06	-0.00	-0.34	-0.10	-0.24	-0.12	0.15	-0.14	-0.17	-0.13	0.08	0.02	-0.13	0.02	-0.18	0.71	0.37	1.00	-0.10	0.12	0.26	-0.17	-0.18	0.00	-0.16
HIV/AIDS	-0.01	-0.26	-0.08	0.10	-0.04	-0.05	-0.10	0.17	-0.67	0.77	-0.07	0.02	-0.17	-0.04	-0.01	-0.34	-0.06	-0.21	0.14	-0.10	1.00	-0.18	-0.05	0.31	0.37	-0.29	-0.14
GDP	0.55	0.70	0.68	-0.35	0.58	0.49	0.36	0.57	0.55	-0.52	-0.26	0.47	0.99	-0.05	-0.30	0.52	-0.26	0.26	0.46	0.12	-0.18	1.00	-0.21	-0.38	-0.38	0.66	0.55
Population	-0.34	-0.39	-0.34	0.28	-0.45	-0.11	-0.36	-0.46	-0.13	0.15	-0.10	-0.32	-0.26	0.36	-0.04	-0.20	-0.11	0.22	-0.22	0.26	-0.05	-0.21	1.00	-0.13	-0.13	-0.25	-0.19
thinness 1-19 years	-0.80	-0.65	-0.68	0.35	-0.33	-0.47	-0.58	-0.18	-0.73	0.56	0.88	-0.52	-0.36	-0.45	0.78	-0.60	0.88	-0.36	-0.38	-0.17	0.31	-0.38	-0.13	1.00	1.00	-0.77	-0.68
thinness 5-9 years	-0.78	-0.65	-0.67	0.34	-0.32	-0.46	-0.57	-0.17	-0.76	0.60	0.86	-0.50	-0.36	-0.44	0.76	-0.61	0.86	-0.37	-0.36	-0.18	0.37	-0.38	-0.13	1.00	1.00	-0.77	-0.67
Income composition of resources	0.91	0.95	0.94	-0.62	0.73	0.71	0.69	0.55	0.86	-0.73	-0.63	0.79	0.67	0.19	-0.66	0.59	-0.63	0.19	0.60	0.00	-0.29	0.66	-0.25	-0.77	-0.77	1.00	0.92
Schooling	0.83	0.83	0.89	-0.52	0.70	0.65	0.54	0.50	0.70	-0.53	-0.59	0.79	0.56	0.27	-0.64	0.58	-0.58	-0.05	0.58	-0.16	-0.14	0.55	-0.19	-0.68	-0.67	0.92	1.00

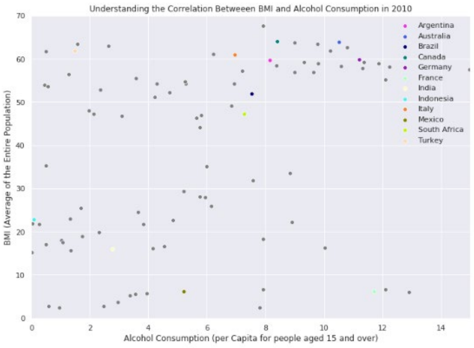
Organising the raw data and making the data more meaningful, a correlating heatmap between all the health indicators that were available in the dataset and scatter diagrams to represent the relationship between specific health indicators, trades and the GDP of countries was developed.

GDP in Relation to Personal Health

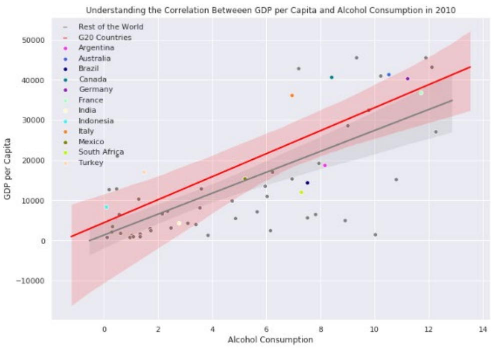
We place deeper focus on Alcohol as a major indicator of personal health as it indirectly relates to increases on other health indicators such as infant deaths, mortality and

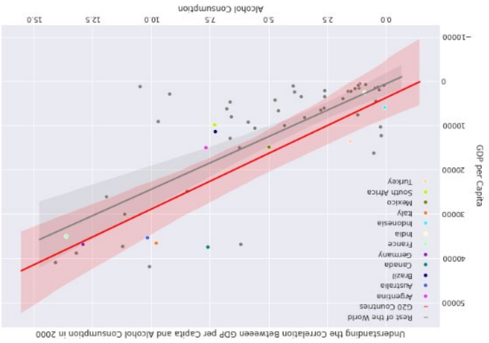
Body Mass Index (BMI). A study conducted by University of California by David Phillips, presented a **33%** increase in Sudden Infant Death Syndrome (SIDS). Further supporting the claim, Dr. Michael Malloy, a neonatologist at University of Texas Medical Branch said, "**when women are inebriated the attentiveness to the child is going to be reduced**" (News, 2019). This is verifiable as mothers may be creating unsafe environments for their children i.e. loss of consciousness. Correspondingly, alcohol has always been casually linked with injuries and diseases, resulting in future consequences of adult mortality (Jürgen Rehm, 2019). In relation, another study conducted by researchers at the University of California states that "**people who drink a small amount daily have a lower BMI**" (Nies et al., 2019). This is also true as supported by the scatter plot graph of the correlation between BMI and alcohol consumption of countries in the G20 in 2000, as there is a cluster within BMI below 20 and 2 Litres of pure alcohol. Furthermore, most of the data lies above BMI of 40. This is no difference to 2010, where the number of countries in the G20 have increased in significantly as there is a higher and more consistent cluster as more across BMI 50 and higher as more alcohol is consumed.





Yet, this infers that countries in G20 are spending more on alcohol, and thus increasing in sales and GDP. This is apparent in both scatter plots between GDP and Alcohol consumption in 2000 and 2010, as both have positive slope gradients, presenting that as alcohol consumed increases, GDP increases correspondingly.

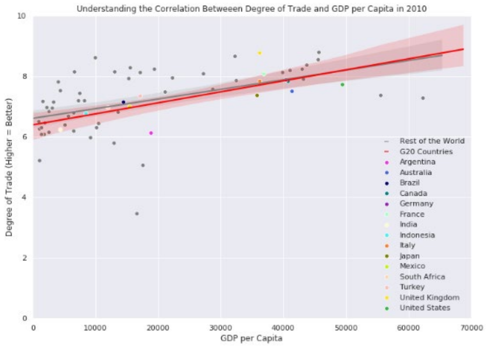


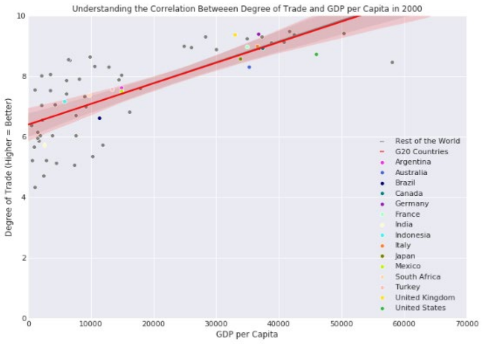


From a global scope, we witness that alcohol and GDP per capita have a 0.78 correlation, as presented in the correlation heatmap, highlighting how the consumption rate of alcohol with indirect reference with infant deaths and adult mortality, does indeed have a strong relationship with GDP.

GPD in Relation to Country's Freedom of Trade

A country's freedom of trade is buying, selling, making contracts, etc, essential to economic freedom and the growth. The two main organisations, WHO and WTO, aims to display the intertwinement between health and trade. World Health Organisation (WHO) aims to achieve a consistent policy between trade and health policies such that international trade and trade rules maximise health benefits and minimise health risks, mainly exclusive for poor and vulnerable countries. The World Trade Organisation (WTO) works to eliminate restrictions on free trade and trade liberalisation increases the sale and consumption of some goods that are harmful to health, depriving countries of taxes. In relation to previous discussion, limiting alcohol consumption through taxation, constraint on hours of sale and advertising restraints of alcohol are the key ingredients to uphold alcohol control policies. However, this spark debate between public health alcohol policies and free trade agreements, which remove constraints on the buying and selling of goods and reducing trade surplus. This is echoed in Article XIV of the General Agreement on Trade in Services (GATS), affirming that "**nothing in this Agreement shall be construed to prevent the adoption or enforcement by any Member of measures necessary to protect human, animal or plant life or health**" (Health Knowledge, 2019). Hence, the introduction of trade liberalisation aims to deprived countries of tariff revenue, brought upon by WTO. WTO does not guarantee trades to increase flow but rather to open up opportunities. Taking advantage of the WTO, we see how in 2000, the degree of trade was far less as more countries were closed off to trade in contrast to 2010, where countries were more open to trades.





However, this didn’t significantly impact the increase of GDP per capita but rather showing that there more opportunities for countries to make trades, regardless if they were vetoed or upheld, as the correlation between trades (substantially with respect with degrees/freedom of trade) and GDP is 0.36 from the correlation heatmap. Further reinforcing there is some essence and correlation, those wanting more understanding between GDP pc and trades are more inclined to delve into each individual trade rather as a whole.

In summary, the analysis of the relationship between trades and health indicators, there are gradual increases of GDP of each country in G20. It has revealed that there are some correlations between a country’s freedom of change and GDP whereas there is a strong correlation between health indicators, specifically alcohol, in comparison with GDP.

Years of Education in Relation to Population Wellbeing and Sound Money

Topic Overview

The general well being of a population are determined by numerous factors that requires through examination due to the ambiguity that's lies between causation and correlation. Through methods used within 'data science' this study will explore the level of influence education and economic stability has on the population's general wellbeing within a country throughout the 21st century. For the study, members of the G20 comity were chosen with the purpose of computing data from different parts of the world to accurately access the varying influence of Education and sound money in different regions of the world. All numerical data analysed within study are gathered from 2000 to 2010, however due to few missing entries for certain countries' data not all years will be used for a meaningful comparison.

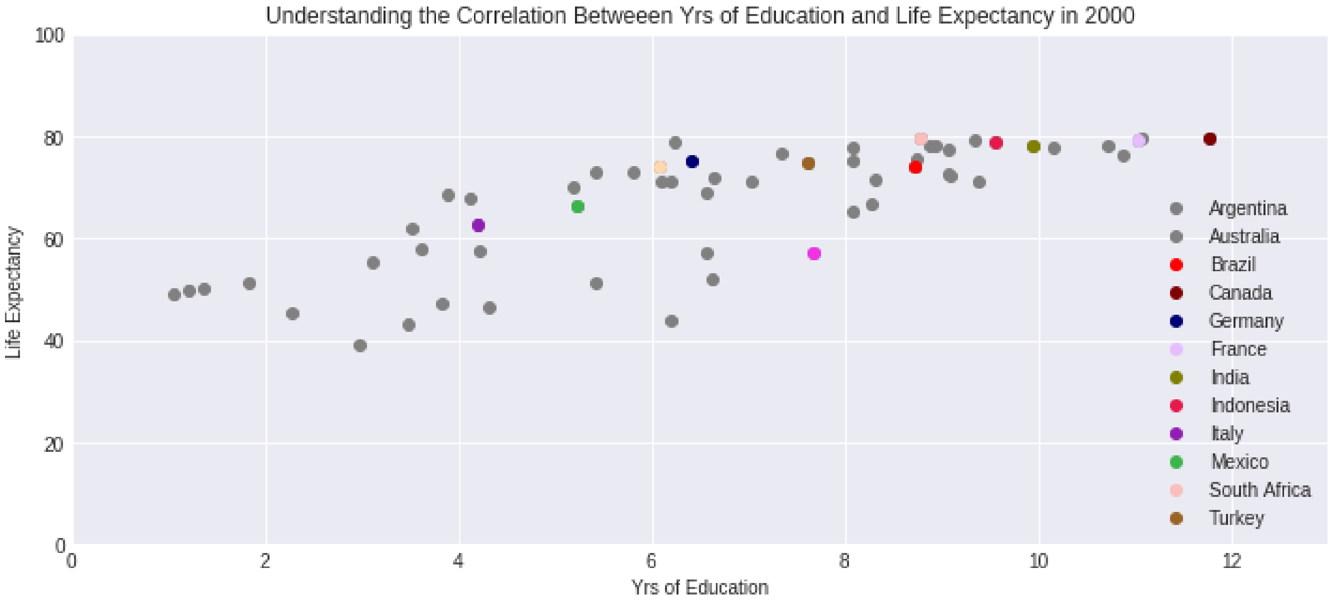
While there are numerous methods of accessing the 'general wellbeing' of a population, many studies like (Our World in Data, 2019) indicate that the average life expectancy of a population is an appropriate measure for the general wellbeing of the population. This idea stems from the fact that a longer, average life spans would often be linked to better healthcare, hygiene and overall expenditure (access to goods and services).

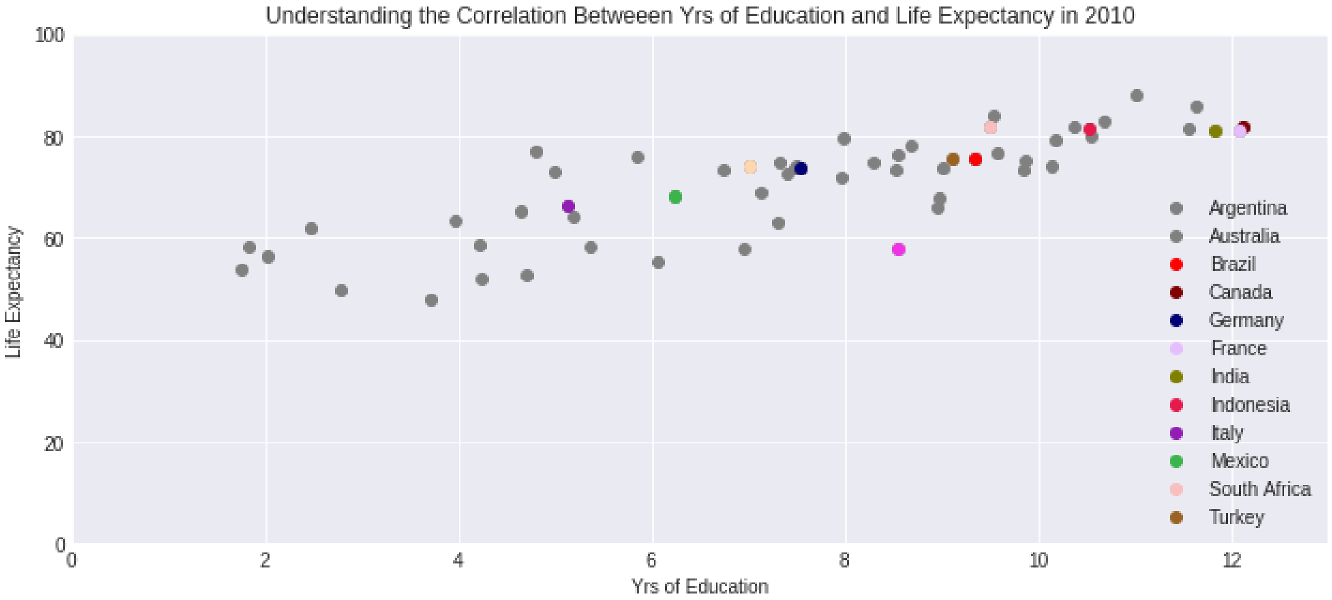
This study utilises information obtained from two different datasets. Data on the influence of government was accessed from Guillermina Sutter Schneider, who summarised raw data from the Economic Freedom of the World: 2018 Annual Report which is published by the Fraser Institute. Furthermore, the dataset containing data points on the socioeconomic status of different countries was compiled and published by Shawn Dorius, who extracted raw data from a variety of sources including the 2004 publication of "The World Economy: Historical Statistics" by the Organisation for Economic Co-operation and Development (OECD).

For the analysis of the data, we transformed the raw data into tables and a correlation heatmap to study the correlation between all the indicators were in the dataset and scatter diagrams to examine the possible relationships between different variables like education and the average lifespan within different countries. Furthermore, by incorporating 'data science' into the study, the report utilises different approaches to the raw data. For example within the study conducted to examine the possible influence of sound money to average life span, we compared the ratio yield by – (Life expectancy / sound money) to extract outliers for further examination.

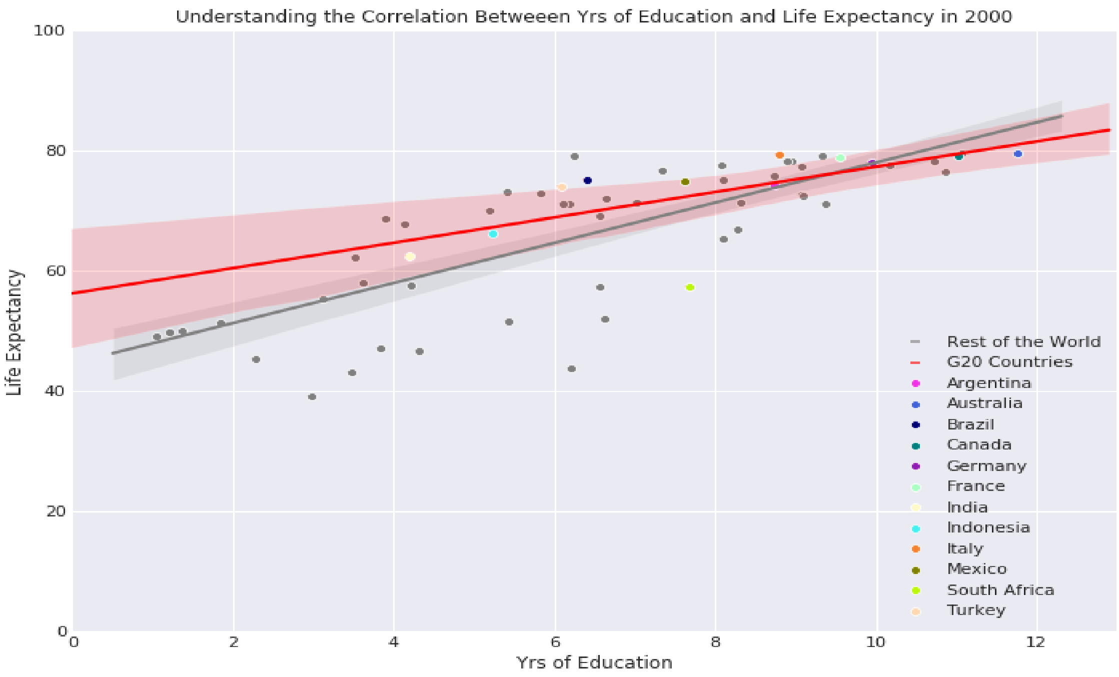
Relationships between average life span and level of education

The consensus belief between average life span and level of education can be outlined within the article (E Rogot, 2019). There is a general belief that a higher level of education would often lead to a longer life span within a specific region due to education being heavily linked with better infrastructure and high standards of societal functionality within a country. Within this report the level of education is a measure of the total number of years an average individual has spent from adolescence to tertiary education.





The above are scatters graphs outlining the relationship between the years of educations vs the life expectancy within the G20 countries. Both graphs summarised the set of data gathered in the year 2000 and 2010. Both graphs clearly indicate an increasing gradient of life expectancy which highlights a positive relationship between the years of total education and the life expectancy.



Like visible from the scatter graph above, this positive relationship can also be found with countries outside of the G20 comity. By utilising this data, the study can appropriately make a hypothesis that the years of education is a significant contributor that influences the average life expectancy of citizens world-wide. This finding suggests that as the years of education increases within a country, it is likely to expect a gradual increase of its citizen's life expectancy.

However, the exact influence of education to the citizen's life expectancy is unclear as this positive relationship is not constant between all countries. This idea stems from the numerous outliers visible within scatter graphs. These outliers suggest the idea that even though education is a significant factor of life expectancy, it is not the only the only influence. Furthermore, due to the nature of education, it is unclear how they affect the average life span of individuals as the result of educational benefits are not immediately shown within society but rather a long-term process. However, each countries' educational standards often reflect on the strength of the country's society (E Rogot, 2019). which could help translate the result gathered from this study.

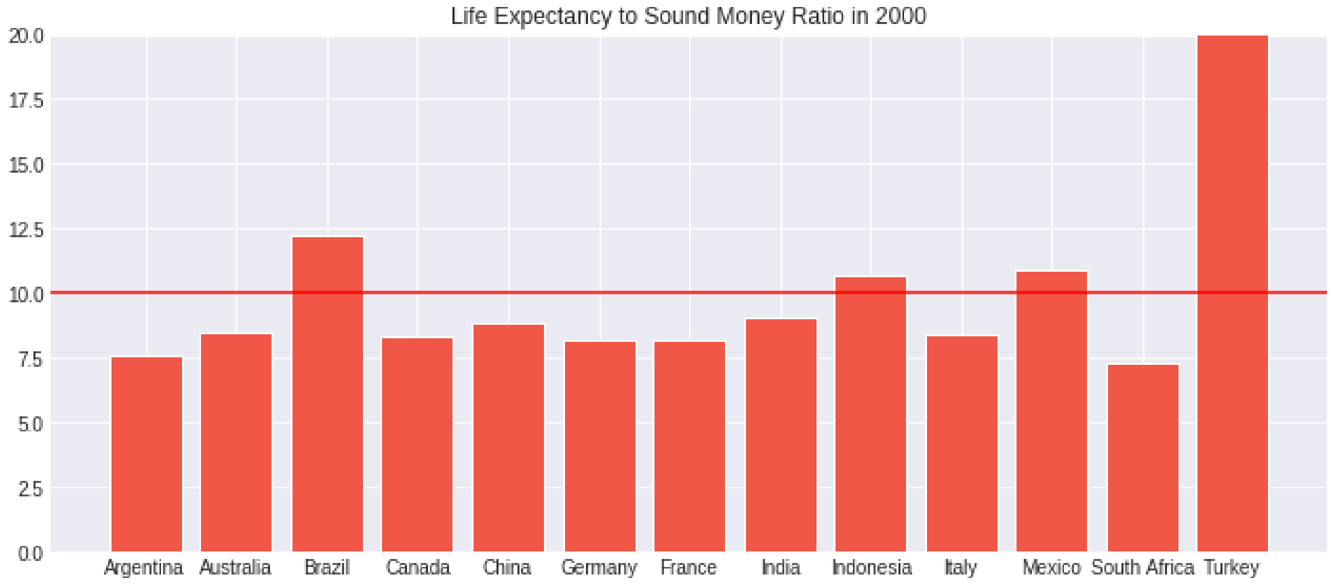
In conclusion, the study affirms the consensus that there is a positive relationship in between the level of education and the average life span. However, the study also showcases the idea that education's influence may just be a numerical causation which is heavily tied with the individual country's society strengths that defines their education standards.

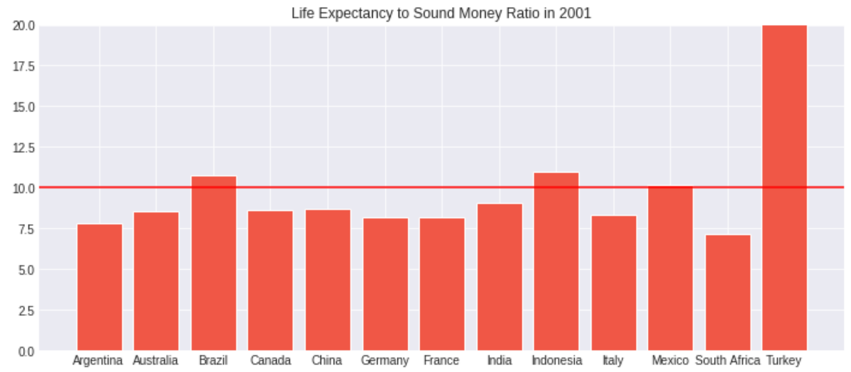
Relationship between sound money and life expectancy

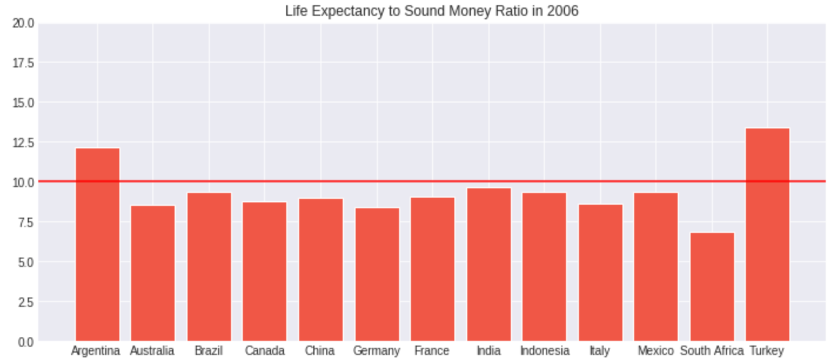
Sound money refers to the total liability of a currency. The sound money index assesses a currency's likelihood of appreciation or depreciation and also people's general trust and exchangeability for other goods. This section of the study will examine the level of influence sound money has on the life expectancy by utilising the sound money ratings countries received from 2000 ~ 2010.

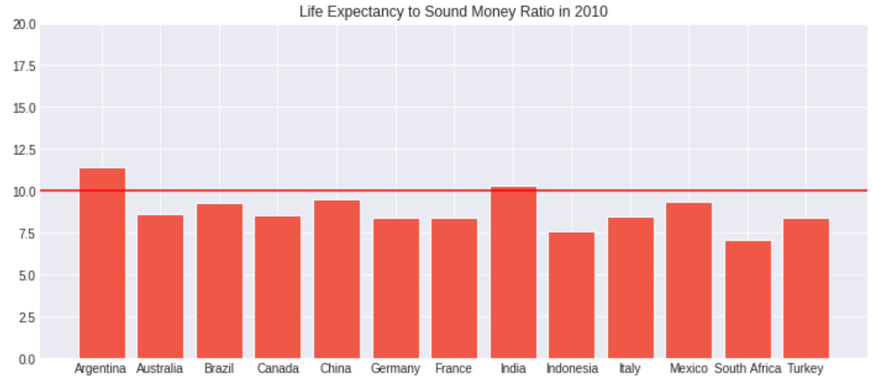
Numerous studies like (Shahbaz et al., 2015) reveals few insights into the importance of liability within a currency and the factors which determines a currency's liability. The article states that many factors including government mismanagement to different international and global crisis leads to the deprecation for a currency. During these events it is likely to expect the average individual's quality of life to decrease, thus during this study it will be hard to determine if sound money is a causation to life expectancy changes or a correlation.

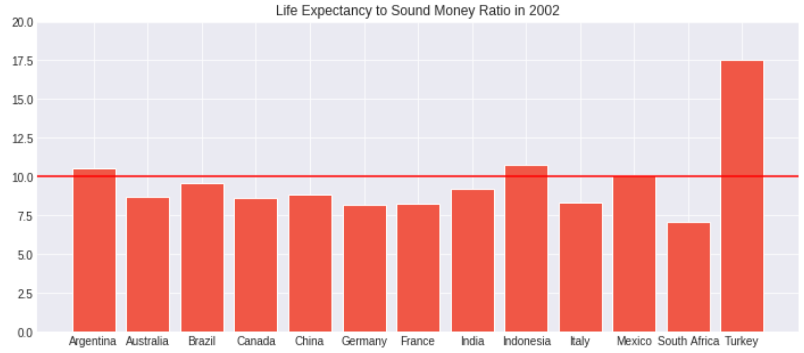
There are multiple issues revolving around this study as most of countries enrolled in the G20 comity are already well developed, first world countries. This means majority of these countries already have a history of well-established governments and society structures which translates to relatively high and steady average life expectancy and their sound money rating. Due to their relatively unchanging levels of life expectancy and sound money it is difficult to study their influences throughout time. To consider for this factor this analysis will be taking a different approach to the collection of data in hand. Thus, each countries' ratio of life expectancy divided by its sound money rating were displayed in a bar graph like below.











Through this comparison the study could find two obvious outliers, Argentina and turkey whom experienced noticeable changes in either their average expectancy or sound money in between 2000 - 2010. In addition, the bar graph above indicates that these two outliers eventually became 'stabilised' approaching closer to the norm throughout time.

A close-up study of both Turkey and Argentina's reveals that their estimated life expectancy shared great similarities with other G20 countries showing no significant change from year 2000 to 2010. Argentina started of the year 2000 with an estimated average life expectancy of 73 and Turkey with 70 and were gradually increased to 76 and 77 respectively in 2010. This numerical evaluation indicates the idea that the sound money rating was the only variable that had experienced great change over time. Furthermore, it established the idea that life expectancy and sound money are variables independent from one another highlighting the idea that there is very little evidence to support the direct influence of sound money in determining country's average life expectancy.

In summary, the study established between sound money and life expectancy revealed no numerical evidence to support the idea of strong relationship in between sound money and average life expectancy within a region. However, because this study was only conducted in the scope within the G20 countries, there still is a possibility that there might be more numerical evidence to support the idea that sound money and life expectancy are interlinked within the less developed regions of the world like the article (Shahbaz et al., 2015) suggests.

Conclusion

Throughout the report various themes of socioeconomical performance of 'G20' countries were carefully analysed utilising diverse programing techniques and fundamentals within 'Data Science'. Thus, the report is able to appropriately address the interests of varying types of audience by clearly outlining each study's methods, evidence and conclusive evaluations. Furthermore, by successfully providing insights into the development and execution of relevant predictive models, the overall project gained greater meaning and uniqueness to its findings.

Section 2: Audiences Interested in Data Analysis

Critical to substantiating the analysis from the previous section was the inclusion of various graphical elements (charts and tables). This section will detail the methodologies involved to produce them and is primarily intended for providing insights to individuals with an interest in IT approaches to data analysis. It is important to note that all data processing and chart creations were completed within the Google Colab python notebook environment. Thus, certain functions may refer specifically to its use exclusive to the environment only.

Loading the datasets was achieved through direct sourcing from Google Drive using PyDrive. This involves authorising google drive access to retrieve the clean datasets from where they had been stored for access.

```
# Code to read csv file into Colaboratory:
!pip install -U -q PyDrive
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

# Authenticate and create the PyDrive client.
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

Having authorised access from the Google Drive the datasets were then loaded to the Google Colab environment. Links from the files located on the Google drive were used for direct access and use. The relevant packages were imported also.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.optimize import curve_fit
sns.set_style('darkgrid')
link = 'https://drive.google.com/open?id=18dTY8lnkcUQLBVTF_gQ4wDSaxQpwU7uE'
fluff, id = link.split('=')
downloaded = drive.CreateFile({'id':id})
downloaded.GetContentFile('efw_cc_clean.csv')
link = 'https://drive.google.com/open?id=1BFJGh4VKsnHWhCrrOJ3sZp7nhxWy4fHA'
fluff, id = link.split('=')
downloaded = drive.CreateFile({'id':id})
downloaded.GetContentFile('GLOB-SES_clean.csv')
link = 'https://drive.google.com/open?id=19Rxhhpejbd_Apzzi-Mfj-3W7X-Esl-IC'
fluff, id = link.split('=')
downloaded = drive.CreateFile({'id':id})
downloaded.GetContentFile('Life Expectancy Data_clean.csv')
```

The loaded datasets were then converted to pandas dataframes for manipulation and construction of the charts.

```
efw = pd.read_csv('efw_cc_clean.csv')
gs = pd.read_csv('GLOB-SES_clean.csv')
le = pd.read_csv('Life Expectancy Data_clean.csv')
```

Some additional settings were applied to how many rows and columns would be displayed upon the printing of any pandas dataframes. These were to show at most 2500 rows and 500 columns of data.

```
pd.set_option('display.max_rows', 2500)
pd.set_option('display.max_columns', 500)
```

Since the primary focus of this research report revolved around an investigation for understanding a variety of measures relating to G20 nations these had to be filtered from each of the dataframes . While this was addressed these filtered data frames were then concatenated to form a single dataframe that contained these G20 nations. This process was repeated on the gs and le data frames.

```
#efw dataset
arg1 = efw.loc[efw.countries == 'Argentina']
aus1 = efw.loc[efw.countries == 'Australia']
brz1 = efw.loc[efw.countries == 'Brazil']
can1 = efw.loc[efw.countries == 'Canada']
Chn1 = efw.loc[efw.countries == 'China']
Grm1 = efw.loc[efw.countries == 'Germany']
Frn1 = efw.loc[efw.countries == 'France']
Ind1 = efw.loc[efw.countries == 'India']
Ina1 = efw.loc[efw.countries == 'Indonesia']
Ity1 = efw.loc[efw.countries == 'Italy']
Jpn1 = efw.loc[efw.countries == 'Japan']
Mex1 = efw.loc[efw.countries == 'Mexico']
Rus1 = efw.loc[efw.countries == 'Russia']
SA1 = efw.loc[efw.countries == 'Saudi Arabia']
Saf1 = efw.loc[efw.countries == 'South Africa']
SK1 = efw.loc[efw.countries == 'Korea, South']
Tky1 = efw.loc[efw.countries == 'Turkey']
UK1 = efw.loc[efw.countries == 'United Kingdom']
US1 = efw.loc[efw.countries == 'United States']
new_df1 = pd.concat([arg1, aus1, brz1, can1, Chn1, Grm1, Frn1, Ind1, Ina1, Ity1, 
Jpn1, Mex1, Rus1, SA1, Saf1, SK1, Tky1, UK1, US1])
```

```
#gs dataset
arg2 = gs.loc[gs.country == 'Argentina'].sort_values(by = ['year'])
aus2 = gs.loc[gs.country == 'Australia'].sort_values(by = ['year'])
brz2 = gs.loc[gs.country == 'Brazil'].sort_values(by = ['year'])
can2 = gs.loc[gs.country == 'Canada'].sort_values(by = ['year'])
Chn2 = gs.loc[gs.country == 'China'].sort_values(by = ['year'])
Grm2 = gs.loc[gs.country == 'Germany'].sort_values(by = ['year'])
Frn2 = gs.loc[gs.country == 'France'].sort_values(by = ['year'])
Ind2 = gs.loc[gs.country == 'India'].sort_values(by = ['year'])
Ina2 = gs.loc[gs.country == 'Indonesia'].sort_values(by = ['year'])
Ity2 = gs.loc[gs.country == 'Italy'].sort_values(by = ['year'])
Jpn2 = gs.loc[gs.country == 'Japan'].sort_values(by = ['year'])
Mex2 = gs.loc[gs.country == 'Mexico'].sort_values(by = ['year'])
Rus2 = gs.loc[gs.country == 'Russia'].sort_values(by = ['year'])
SA2 = gs.loc[gs.country == 'Saudi Arabia'].sort_values(by = ['year'])
Saf2 = gs.loc[gs.country == 'South Africa'].sort_values(by = ['year'])
SK2 = gs.loc[gs.country == 'Korea, South'].sort_values(by = ['year'])
Tky2 = gs.loc[gs.country == 'Turkey'].sort_values(by = ['year'])
UK2 = gs.loc[gs.country == 'United Kingdom'].sort_values(by = ['year'])
US2 = gs.loc[gs.country == 'United States'].sort_values(by = ['year'])
new_df2 = pd.concat([arg2, aus2, brz2, can2, Chn2, Grm2, Frn2, Ind2, Ina2, Ity2, 
Jpn2, Mex2, Rus2, SA2, Saf2, SK2, Tky2, UK2, US2])
```

```
#le dataset
arg3 = le.loc[le.Country == 'Argentina']
aus3 = le.loc[le.Country == 'Australia']
brz3 = le.loc[le.Country == 'Brazil']
can3 = le.loc[le.Country == 'Canada']
Chn3 = le.loc[le.Country == 'China']
Grm3 = le.loc[le.Country == 'Germany']
Frn3 = le.loc[le.Country == 'France']
Ind3 = le.loc[le.Country == 'India']
Ina3 = le.loc[le.Country == 'Indonesia']
Ity3 = le.loc[le.Country == 'Italy']
Jpn3 = le.loc[le.Country == 'Japan']
Mex3 = le.loc[le.Country == 'Mexico']
Rus3 = le.loc[le.Country == 'Russia']
SA3 = le.loc[le.Country == 'Saudi Arabia']
Saf3 = le.loc[le.Country == 'South Africa']
SK3 = le.loc[le.Country == 'Korea, South']
Tky3 = le.loc[le.Country == 'Turkey']
UK3 = le.loc[le.Country == 'United Kingdom']
US3 = le.loc[le.Country == 'United States']
new_df3 = pd.concat([arg3, aus3, brz3, can3, Chn3, Grm3, Frn3, Ind3, Ina3, Ity3,
Jpn3, Mex3, Rus3, SA3, Saf3, SK3, Tky3, UK3, US3])
```

Charting processing first began with the comparison made between years of education, sound money and wellbeing of the population measures. To prepare the data for effective analysis in graphical form the pandas join function was utilised, completed first between the le and efw data frames. Before this was achieved the keys that had to be joined on from required renaming to allow this.

```
# Renaming values so that tables dataframes can be joined
new_df1.rename(columns={'countries': 'Country', 'year': 'Year'}, inplace=True)
efw.rename(columns={'countries': 'Country', 'year': 'Year'}, inplace=True)
new_df2.rename(columns={'country': 'Country', 'year': 'Year'}, inplace=True)
gs.rename(columns={'country': 'Country', 'year': 'Year'}, inplace=True)
```

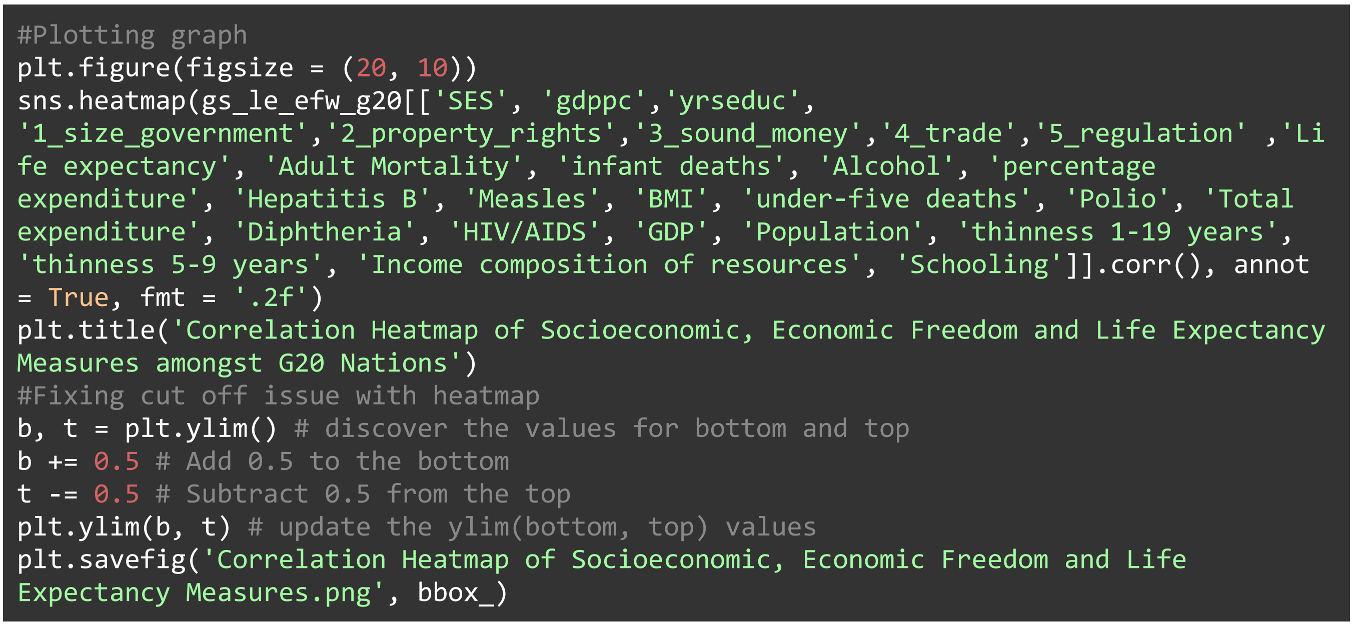
Then the following joined data frames were produced.

```
le_efw_g20 = pd.merge(new_df1, new_df3, on=['Country', 'Year'])
gs_le_efw_g20 = pd.merge(new_df2, le_efw_g20, on=['Country', 'Year'])
le_efw = pd.merge(efw, le, on=['Country', 'Year'])
gs_le_efw = pd.merge(gs, le_efw, on = ['Country', 'Year'])
```

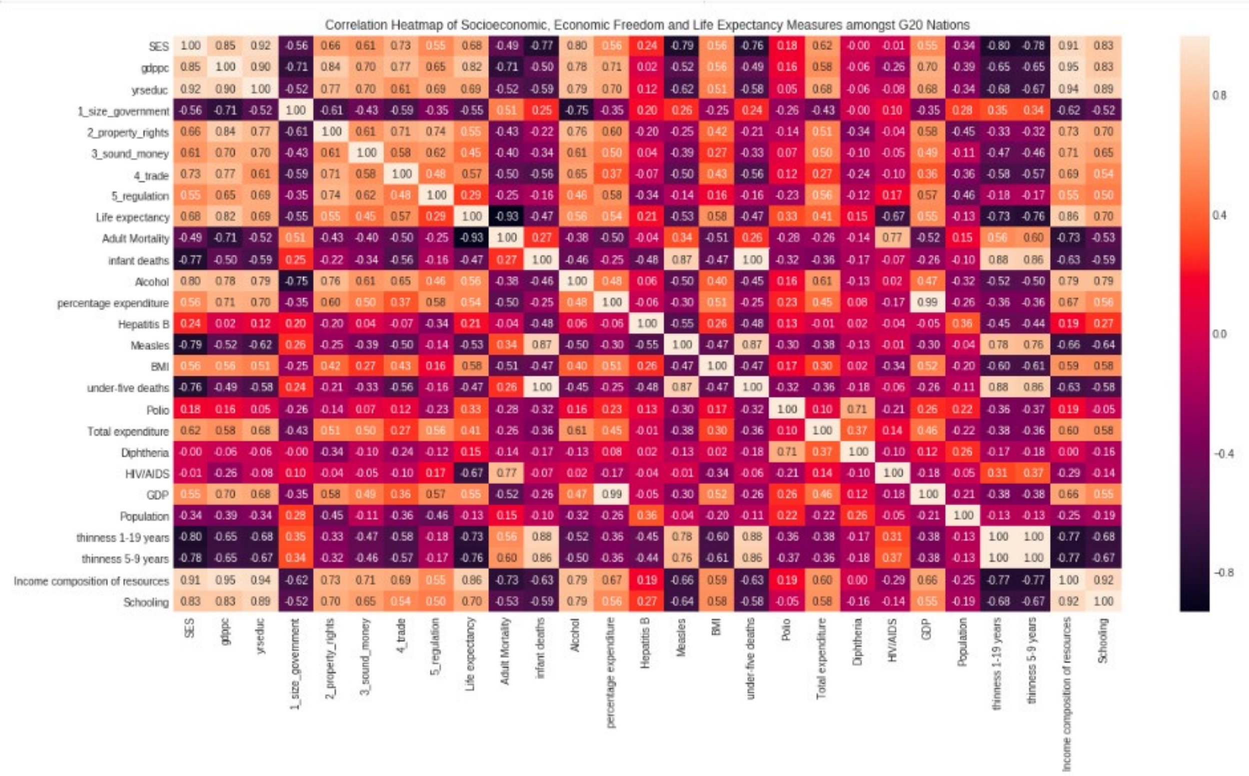
Although this was initially intended to further the findings of the topic mentioned above, the gs_le_efw data frame was also exported as a csv for use in table analysis. This allowed for a quick comparison amongst most of the G20 countries and all the countries in the years 2000 and 2010 (produced in different data frames). Note that some G20 countries were excluded by the fact that inconsistencies between the individual le and efw data frames were disallowed joining by the pandas function.

```
#Exporting as a table csv
gs_le_efw_g20.to_csv('gs_le_efw_g20.csv', index = None, header = True)
gs_le_efw.to_csv('gs_le_efw.csv', index = None, header = True)
```

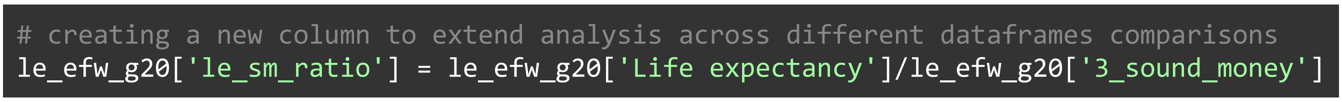
Again, slightly deviating from chart production of the above topic, a correlation heatmap utilising the seaborn library was produced. This showcased all the measures that had been joined from the gs_le_efw dataframe and allowed analysts of the report to gain a quick understanding of how particular measures were correlated. Moreover, initial issues arose from which the top and bottom of the heatmap had been cut off. This required a quick solution that involved a plt.ylim(). Finally, the chart was exported as a png file for external use within this report.



```
#Plotting graph
plt.figure(figsize = (20, 10))
sns.heatmap(gs_le_efw_g20[['SES', 'gdppc','yrseduc',
'1_size_government','2_property_rights','3_sound_money','4_trade','5_regulation' ,'Li
fe expectancy', 'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage
expenditure', 'Hepatitis B', 'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total
expenditure', 'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness 1-19 years',
'thinness 5-9 years', 'Income composition of resources', 'Schooling']].corr(), annot
= True, fmt = '.2f')
plt.title('Correlation Heatmap of Socioeconomic, Economic Freedom and Life Expectancy
Measures amongst G20 Nations')
#Fixing cut off issue with heatmap
b, t = plt.ylim() # discover the values for bottom and top
b += 0.5 # Add 0.5 to the bottom
t -= 0.5 # Subtract 0.5 from the top
plt.ylim(b, t) # update the ylim(bottom, top) values
plt.savefig('Correlation Heatmap of Socioeconomic, Economic Freedom and Life
Expectancy Measures.png', bbox_)
```

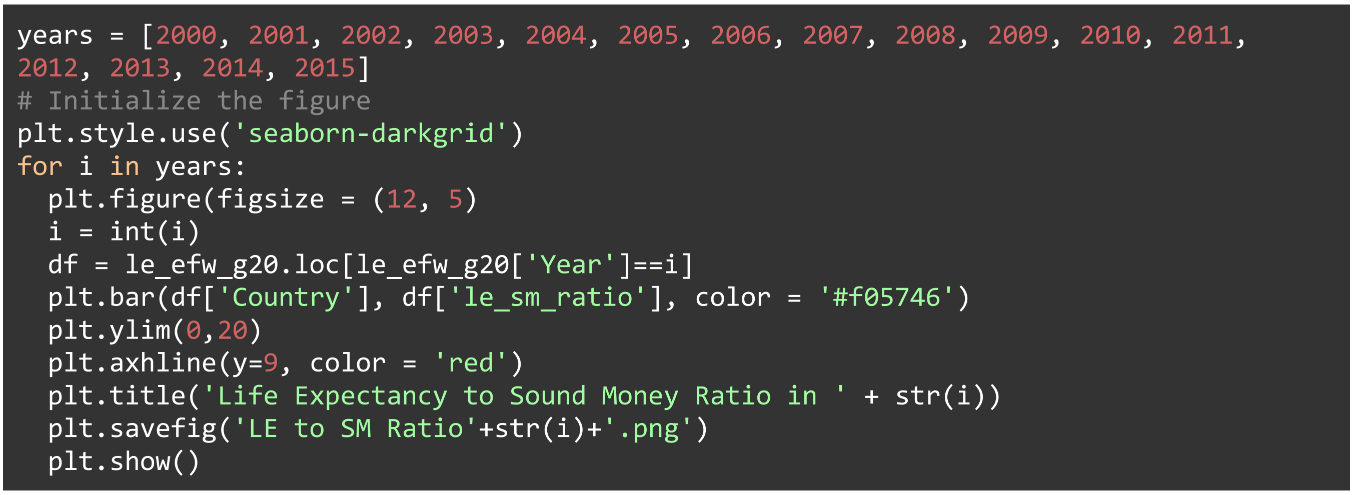


Returning to the intended focus of the topic, processing and manipulating of the dataframes was then addressed to produce the life expectancy to sound money ratio bar graphs. The ratio column was first produced through using the code shown below.



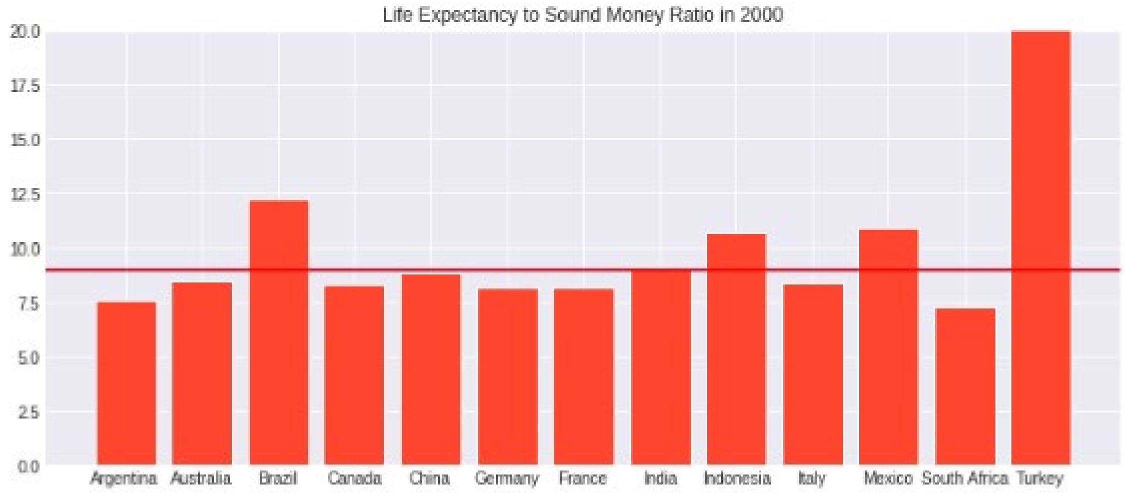
```
# creating a new column to extend analysis across different dataframes comparisons
le_efw_g20['le_sm_ratio'] = le_efw_g20['Life expectancy']/le_efw_g20['3_sound_money']
```

Having attained the appropriate measures, a for loop was utilised to efficiently produce each ratio graph to their respective year, 2000 to 2015. To do this a list names years was created so that the for loop would loop through each iterative year through which the pd.loc function would filter the rows relevant to the year specified in the loop. These filtered measures in the temporary df data frame were then displayed using matplotlib tools with an additional reference line running at y = 9.

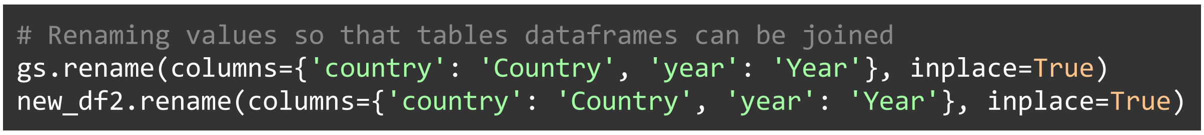


```
years = [2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011,
2012, 2013, 2014, 2015]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
for i in years:
  plt.figure(figsize = (12, 5)
  i = int(i)
  df = le_efw_g20.loc[le_efw_g20['Year']==i]
  plt.bar(df['Country'], df['le_sm_ratio'], color = '#f05746')
  plt.ylim(0,20)
  plt.axhline(y=9, color = 'red')
  plt.title('Life Expectancy to Sound Money Ratio in ' + str(i))
  plt.savefig('LE to SM Ratio'+str(i)+'.png')
  plt.show()
```

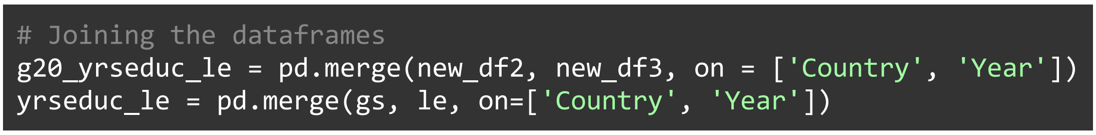
An example of one of the produced graphs is shown below:



Following this, to delineate the relationship between years of education and life expectancy the same joining process as detailed above was applied.



```
# Renaming values so that tables dataframes can be joined
gs.rename(columns={'country': 'Country', 'year': 'Year'}, inplace=True)
new_df2.rename(columns={'country': 'Country', 'year': 'Year'}, inplace=True)
```



```
# Joining the dataframes
g20_yrseduc_le = pd.merge(new_df2, new_df3, on = ['Country', 'Year'])
yrseduc_le = pd.merge(gs, le, on=['Country', 'Year'])
```

In preparation of the years of education and life expectancy seaborn scatter plot a list named countries was produced with the names of all the G20 nations contained within the g20_yrseduc_le data frame, its importance being described soon.

```
countries = list(g20_yrseduc_le.Country.unique())
print(countries)
```

Then a custom colour palette was made with a list of hex code colours, whilst the random package was imported to shuffle the order of the colours.

```
import random
nice_colours = ['#e6194b', '#3cb44b', '#ffe119', '#4363d8', '#f58231', '#911eb4',
'#46f0f0', '#f032e6', '#bcf60c', '#fabebe', '#008080', '#e6beff', '#fffac8',
'#800000', '#aaffc3', '#808000', '#ffd8b1', '#000075']
random.shuffle(nice_colours)
print(nice_colours[0])
```

Similarly to how the for-loop was applied in producing the ratio graphs the concept in which it would produce for different years was used here. Two temporary dataframes, df for the rest of the world and df1 for only G20 nations were created, with the intention of plotting a comparison between the distinguished group and the rest of the world.

```
years_adj = [2000, 2010]
```

```
# Initialize the figure
plt.style.use('seaborn-darkgrid')
_for i in years_adj:
  fig, ax = plt.subplots()
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  # Temporary data frames to compare the G20 nations and the rest of the world.
  df = yrseduc_le.loc[yrseduc_le['Year']==i]
  df1 = g20_yrseduc_le.loc[g20_yrseduc_le['Year']==i]
```

A linear regression model was then produced for each group with a generalised grey scatter of plots for the world specified by the sns.scatterplot and the sns.regplot function.

```
 sns.regplot(df['yrseduc'], df['Life expectancy'], color = 'grey', label = 'Rest of
the World',marker='_', ax=ax)
  sns.scatterplot(df['yrseduc'], df['Life expectancy'], color = 'grey',marker='o',
ax=ax)
```

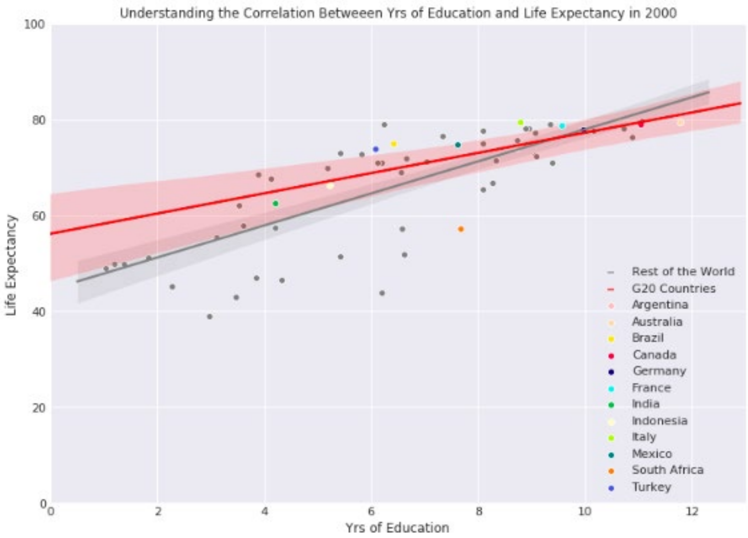
Then to plot each specific G20 nation a nested for loop was used that looped through the countries list mentioned above with the custom list nice_colours being applied so that each country could be differentiated.

```
 num1 = 2
  sns.regplot(df1['yrseduc'], df1['Life expectancy'], color = 'red', label = 'G20
Countries', marker='_',ax=ax)
  for k in countries:
    df2_ye = df1.loc[df1['Country'] == k]
    df2_le = df1.loc[df1['Country'] == k]
    g20_plot = sns.scatterplot(df2_ye['yrseduc'], df2_ye['Life expectancy'], color =
nice_colours[num1], label = k, marker='o', ax=ax)
    g20_plot
    sns.set(rc={'figure.figsize':(11.7,8.27)})
    num1+=1
```

Using the ax.get_legend_handles_labels() function this was utilised with the intention to produce a legend where the handles and labels parameters would be automatically acquired. Finally, finishing touches were applied with the labels, title, limits and a final png being exported for external use in this report.

```
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels)
  plt.xlabel('Yrs of Education')
  plt.ylabel('Life Expectancy')
  plt.ylim(0,100)
  plt.xlim(0,13)
  plt.title('Understanding the Correlation Betweeen Yrs of Education and Life 
Expectancy in ' + str(i))
  plt.savefig('Understanding the Correlation Between Yrs of Education and Life 
Expectancy in' + str(i))
  plt.show()
```

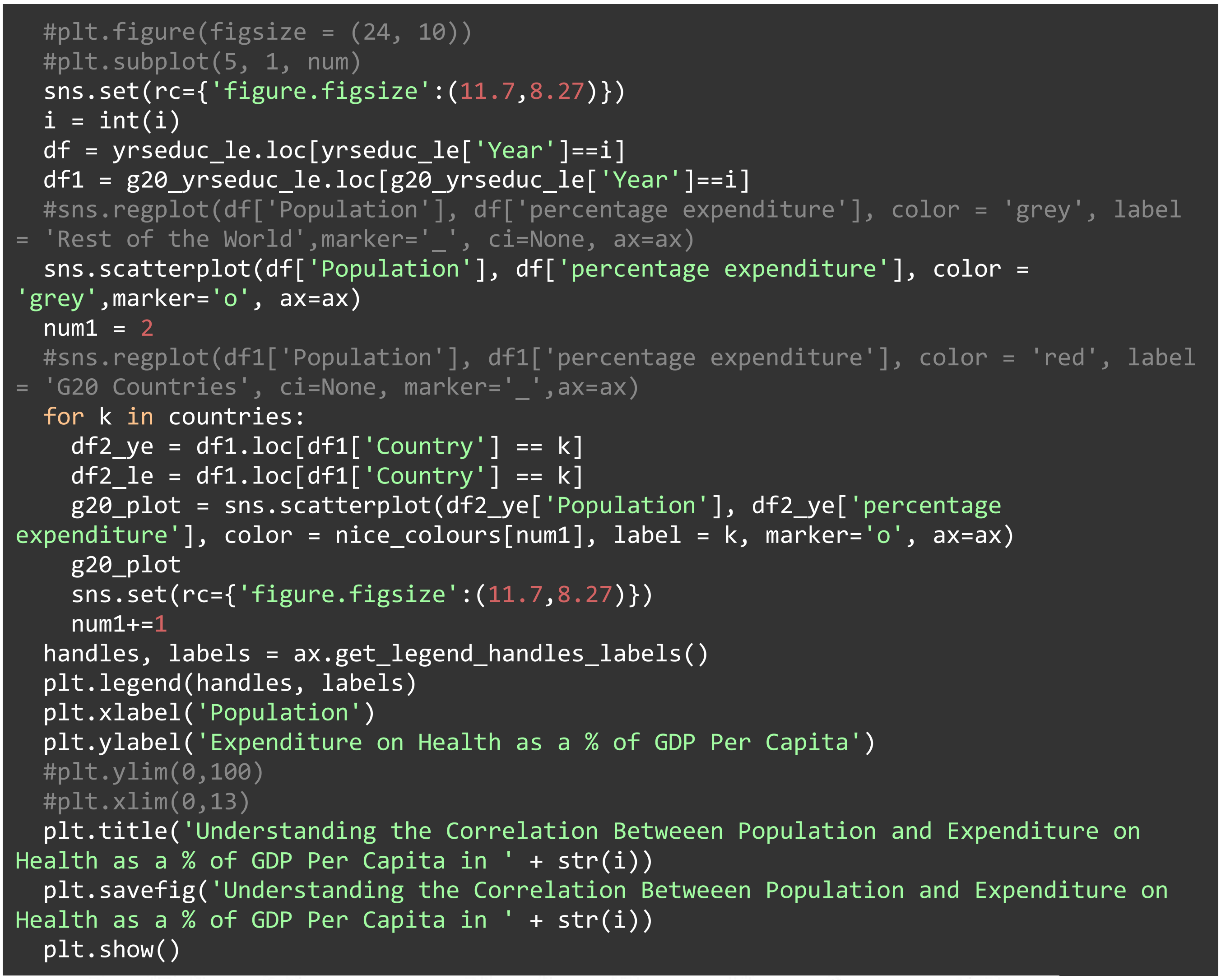
The one of the following graphs was produced as a result:



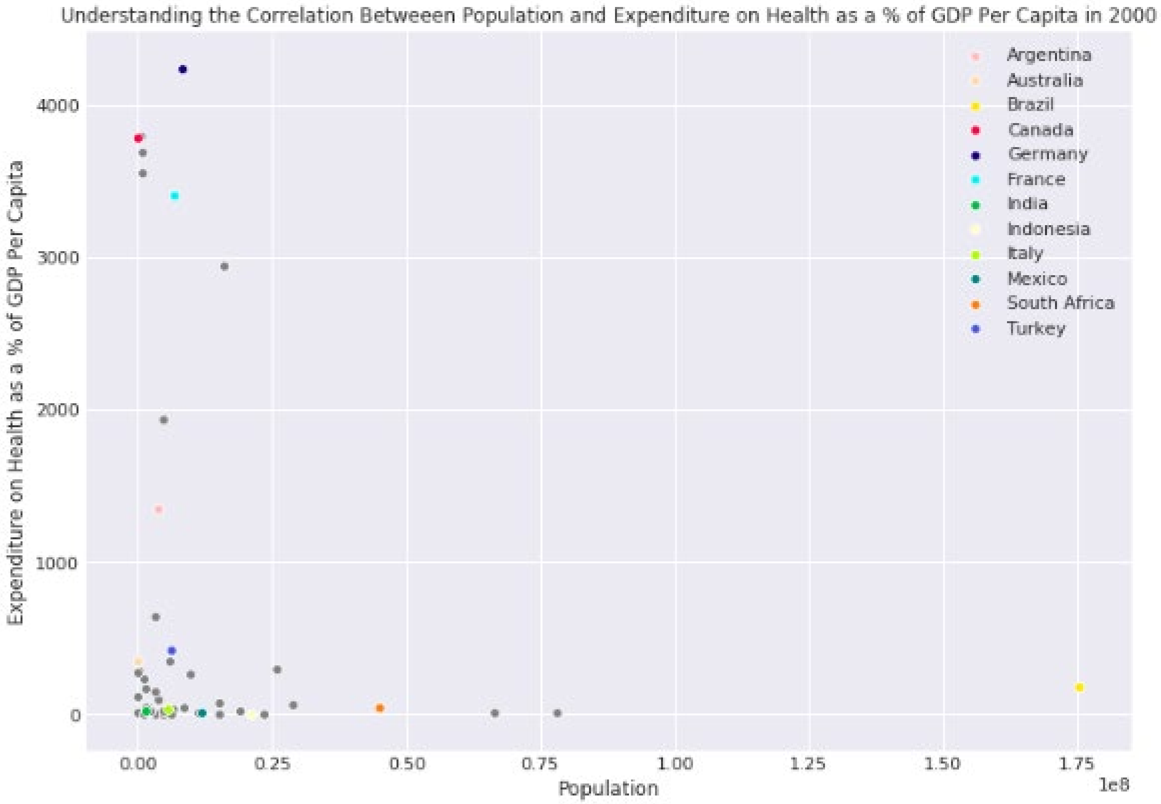
Similarly, this same process was utilised amongst the rest of the scatter plots only changing the data frames values for different measure comparisons, while limits were also adjusted to ensure no inaccuracies arose. The code and graphs are as follows:

Understanding the Correlation Between Population and Expenditure on Health as a % of GDP Per Capita

```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
```

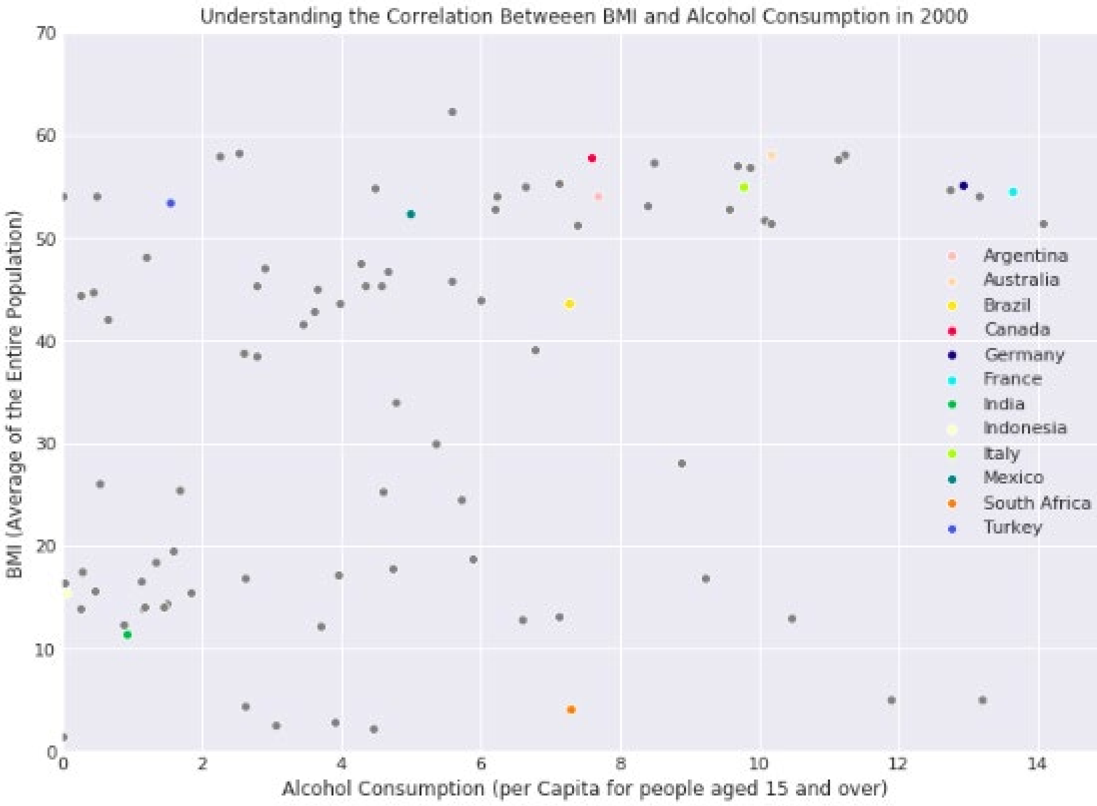


```
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = yrseduc_le.loc[yrseduc_le['Year']==i]
  df1 = g20_yrseduc_le.loc[g20_yrseduc_le['Year']==i]
  #sns.regplot(df['Population'], df['percentage expenditure'], color = 'grey', label
= 'Rest of the World',marker='_', ci=None, ax=ax)
  sns.scatterplot(df['Population'], df['percentage expenditure'], color =
'grey',marker='o', ax=ax)
  num1 = 2
  #sns.regplot(df1['Population'], df1['percentage expenditure'], color = 'red', label
= 'G20 Countries', ci=None, marker='_',ax=ax)
  for k in countries:
    df2_ye = df1.loc[df1['Country'] == k]
    df2_le = df1.loc[df1['Country'] == k]
    g20_plot = sns.scatterplot(df2_ye['Population'], df2_ye['percentage
expenditure'], color = nice_colours[num1], label = k, marker='o', ax=ax)
    g20_plot
    sns.set(rc={'figure.figsize':(11.7,8.27)})
    num1+=1
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels)
  plt.xlabel('Population')
  plt.ylabel('Expenditure on Health as a % of GDP Per Capita')
  #plt.ylim(0,100)
  #plt.xlim(0,13)
  plt.title('Understanding the Correlation Betweeen Population and Expenditure on
Health as a % of GDP Per Capita in ' + str(i))
  plt.savefig('Understanding the Correlation Betweeen Population and Expenditure on
Health as a % of GDP Per Capita in ' + str(i))
  plt.show()
```



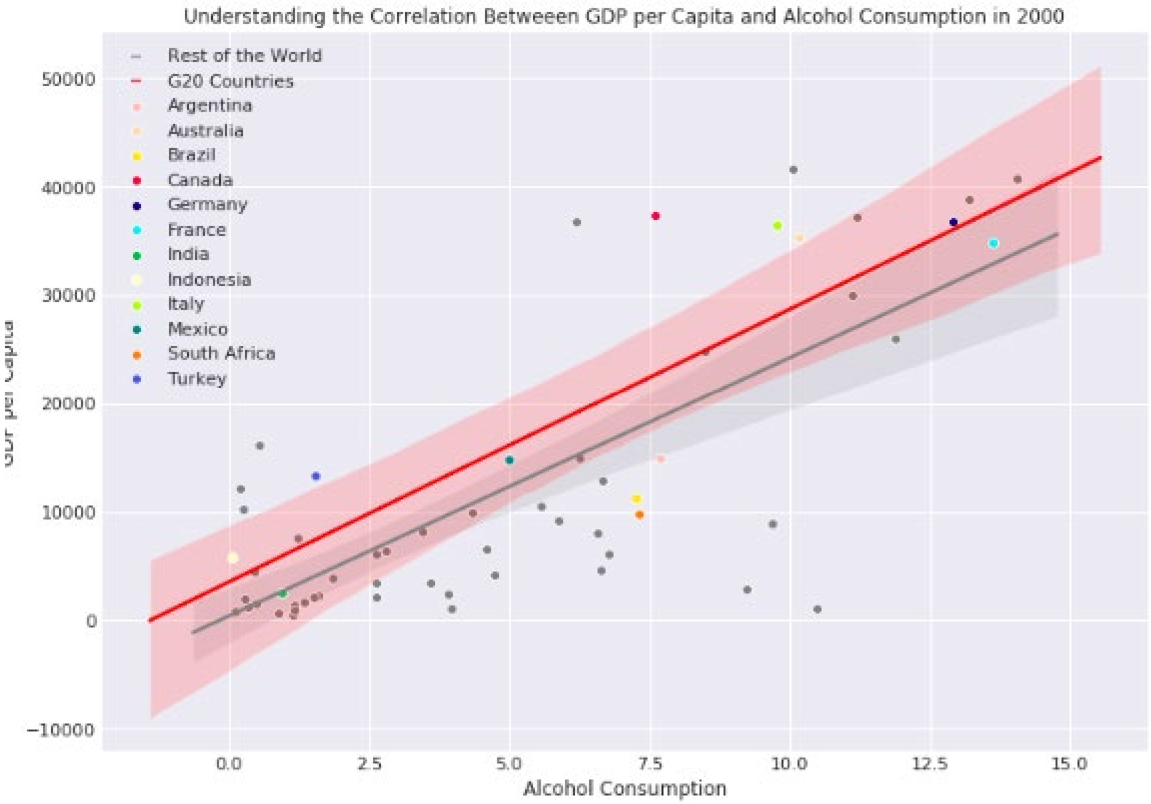
Understanding the Correlation Between BMI and Alcohol Consumption

```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = le_efw.loc[le_efw['Year']==i]
  df1 = g20_yrseduc_le.loc[g20_yrseduc_le['Year']==i]
  #sns.regplot(df['Alcohol'], df['BMI'], color = 'grey', label = 'Rest of the
World',marker='_', ci=None, ax=ax)
  sns.scatterplot(df['Alcohol'], df['BMI'], color = 'grey',marker='o', ax=ax)
  num1 = 2
  #sns.regplot(df1['Alcohol'], df1['BMI'], color = 'red', label = 'G20 Countries',
ci=None, marker='_',ax=ax)
  for k in countries:
    df2_ye = df1.loc[df1['Country'] == k]
    df2_le = df1.loc[df1['Country'] == k]
    g20_plot = sns.scatterplot(df2_ye['Alcohol'], df2_ye['BMI'], color =
nice_colours[num1], label = k, marker='o', ax=ax)
    g20_plot
    sns.set(rc={'figure.figsize':(11.7,8.27)})
    num1+=1
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels)
  plt.xlabel('Alcohol Consumption (per Capita for people aged 15 and over)')
  plt.ylabel('BMI (Average of the Entire Population)')
  plt.ylim(0,70)
  plt.xlim(0,15)
  plt.title('Understanding the Correlation Betweeen BMI and Alcohol Consumption in '
+ str(i))
  plt.savefig('Understanding the Correlation Betweeen BMI and Alcohol Consumption in
' + str(i))
  plt.show()
```



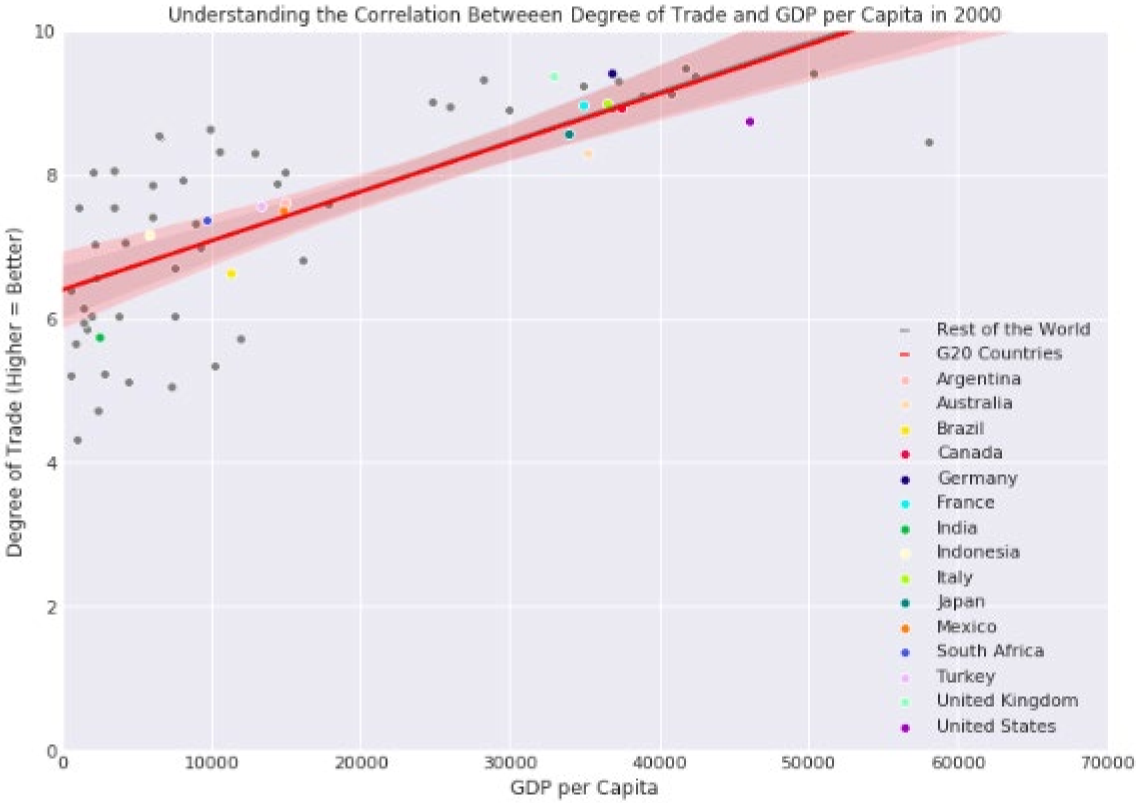
Understanding the Correlation Between GDP per Capita and Alcohol Consumption

```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = yrseduc_le.loc[yrseduc_le['Year']==i]
  df1 = g20_yrseduc_le.loc[g20_yrseduc_le['Year']==i]
  sns.regplot(df['Alcohol'], df['gdppc'], color = 'grey', label = 'Rest of the
World',marker='_', ax=ax)
  sns.scatterplot(df['Alcohol'], df['gdppc'], color = 'grey',marker='o', ax=ax)
  num1 = 2
  sns.regplot(df1['Alcohol'], df1['gdppc'], color = 'red', label = 'G20 Countries',
marker='_',ax=ax)
  for k in countries:
    df2_ye = df1.loc[df1['Country'] == k]
    df2_le = df1.loc[df1['Country'] == k]
    g20_plot = sns.scatterplot(df2_ye['Alcohol'], df2_ye['gdppc'], color =
nice_colours[num1], label = k, marker='o', ax=ax)
    g20_plot
    sns.set(rc={'figure.figsize':(11.7,8.27)})
    num1+=1
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels)
  plt.xlabel('Alcohol Consumption')
  plt.ylabel('GDP per Capita')
  #plt.ylim(0,100)
  #plt.xlim(0,13)
  plt.title('Understanding the Correlation Betweeen GDP per Capita and Alcohol
Consumption in ' + str(i))
  plt.savefig('Understanding the Correlation Betweeen GDP per Capita and Alcohol
Consumption in ' + str(i))
  plt.show()
```



Understanding the Correlation Between the Degree of Trade and GDP per Capita

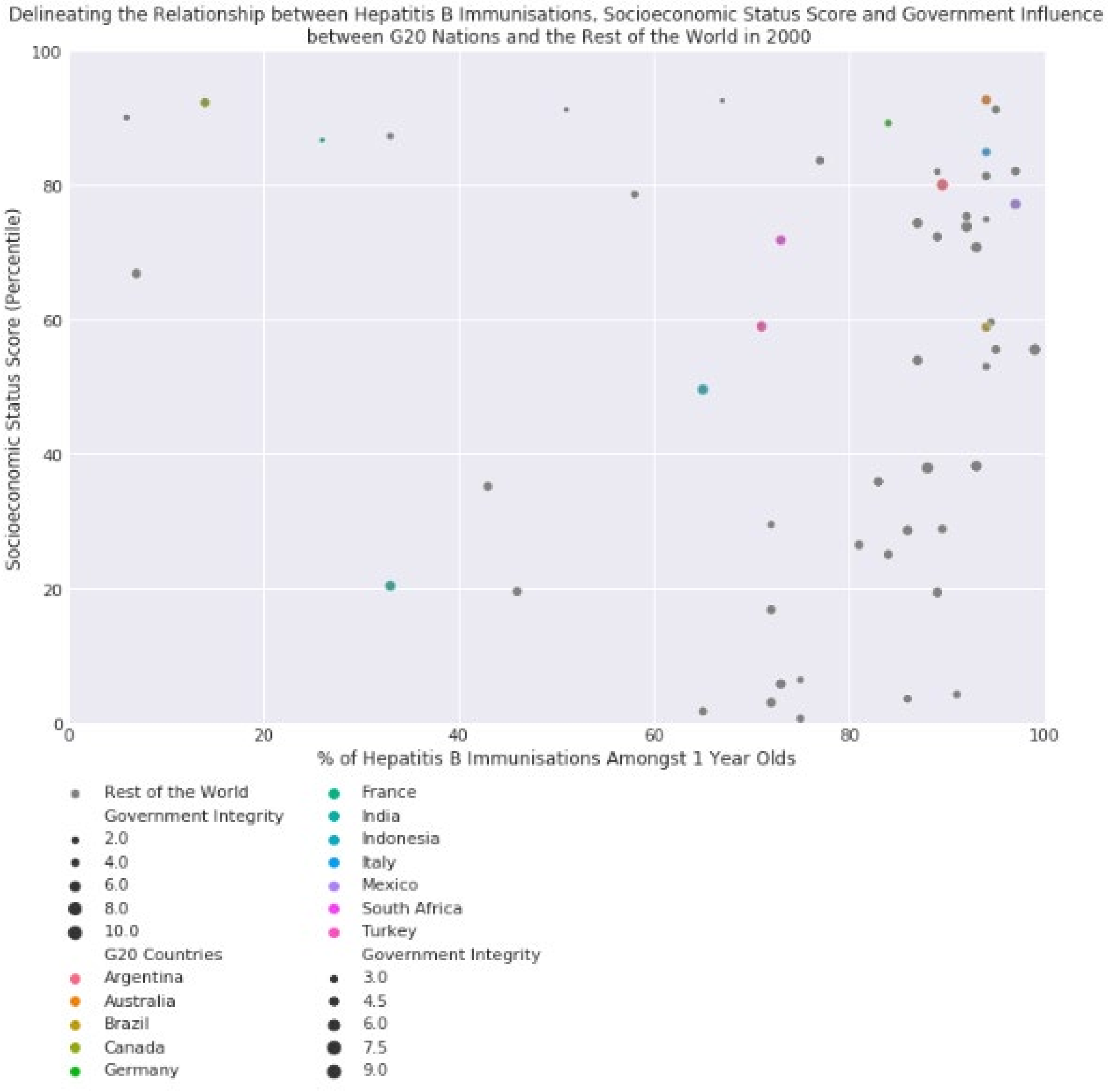
```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = gs_efw.loc[gs_efw['Year']==i]
  df1 = gs_efw_g20.loc[gs_efw_g20['Year']==i]
  sns.regplot(df['gdppc'], df['4_trade'], color = 'grey', label = 'Rest of the
World',marker='_', ax=ax)
  sns.scatterplot(df['gdppc'], df['4_trade'], color = 'grey',marker='o', ax=ax)
  num1 = 2
  sns.regplot(df1['gdppc'], df1['4_trade'], color = 'red', label = 'G20
Countries', marker='_',ax=ax)
  for k in countries2:
    df2_ye = df1.loc[df1['Country'] == k]
    df2_le = df1.loc[df1['Country'] == k]
    g20_plot = sns.scatterplot(df2_ye['gdppc'], df2_ye['4_trade'], color =
nice_colours[num1], label = k, marker='o', ax=ax)
    g20_plot
    sns.set(rc={'figure.figsize':(11.7,8.27)})
    num1+=1
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels)
  plt.xlabel('GDP per Capita')
  plt.ylabel('Degree of Trade (Higher = Better)')
  plt.ylim(0,10)
  plt.xlim(0,70000)
  plt.title('Understanding the Correlation Betweeen Degree of Trade and GDP per
Capita in ' + str(i))
  plt.savefig('Understanding the Correlation Betweeen Degree of Trade and GDP per
Capita in ' + str(i))
  plt.show()
```



Again, utilising the same techniques to produce the above scatter plots slight alterations were applied to the following graphs, some of these being the addition of a size attribute to make a four way comparison. To make use of this, the nested for loop that was used previously had to be removed while the hue attribute within the seaborn plot substituted for this. All other aspects and techniques remained the same otherwise.

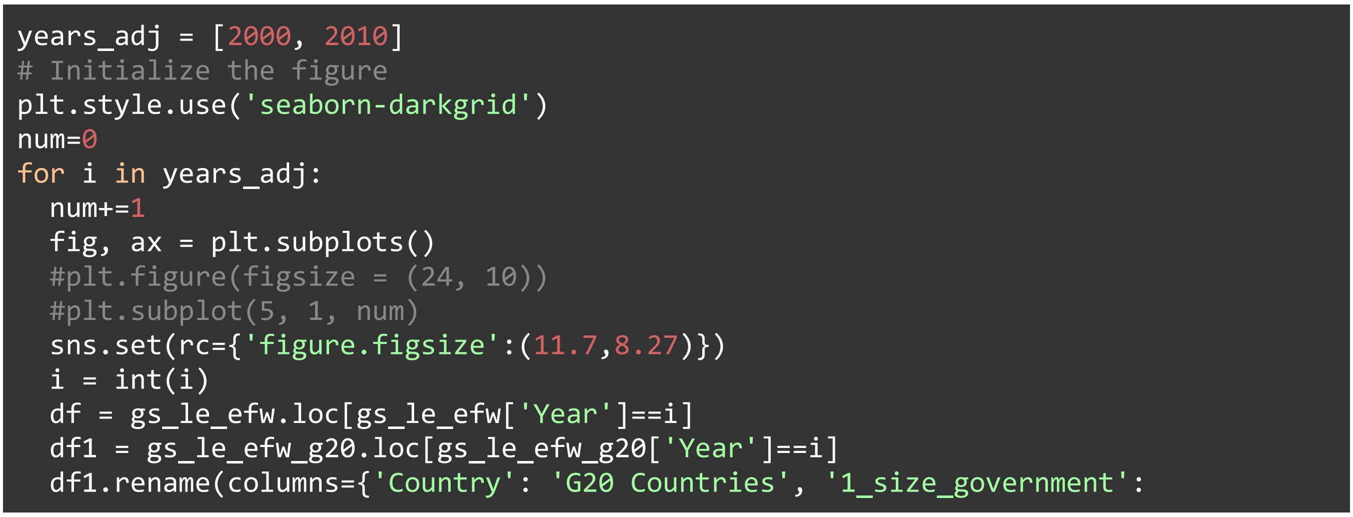
```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = gs_le_efw.loc[gs_le_efw['Year']==i]
  df1 = gs_le_efw_g20.loc[gs_le_efw_g20['Year']==i]
  df1.rename(columns={'Country': 'G20 Countries', '1_size_government': 'Government 
Integrity'}, inplace=True)
  df.rename(columns={'1_size_government': 'Government Integrity'}, inplace=True)
  sns.scatterplot(df['Hepatitis B'], df['SES'], color = 'grey', label = 'Rest of the 
World', size = 'Government Integrity', data = df, marker='o', ax=ax)
  #num1 = 2
  g20_plot = sns.scatterplot(df1['Hepatitis B'], df1['SES'], size = 'Government 
Integrity', data = df1, alpha = 0.5, hue = 'G20 Countries', marker='o', ax=ax)
  g20_plot
  #sns.set(rc={'figure.figsize':(11.7,8.27)})
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels, bbox_to_anchor=(0.5, -0.07), shadow=True, ncol=2)
  plt.xlabel('% of Hepatitis B Immunisations Amongst 1 Year Olds')
  plt.ylabel('Socioeconomic Status Score (Percentile)')
  plt.ylim(0,100)
  plt.xlim(0,100)
  plt.title('Delineating the Relationship between Hepatitis B Immunisations, 
Socioeconomic Status Score and Government Influence\n between G20 Nations and the 
Rest of the World in ' + str(i))
  plt.savefig('Delineating the Relationship between Hepatitis B Immunisations, 
Socioeconomic Status Score and Government Influence\n between G20 Nations and the 
Rest of the World in ' + str(i), bbox_inches='tight')
  plt.show()
```

The one of the following graphs were produced as a result:

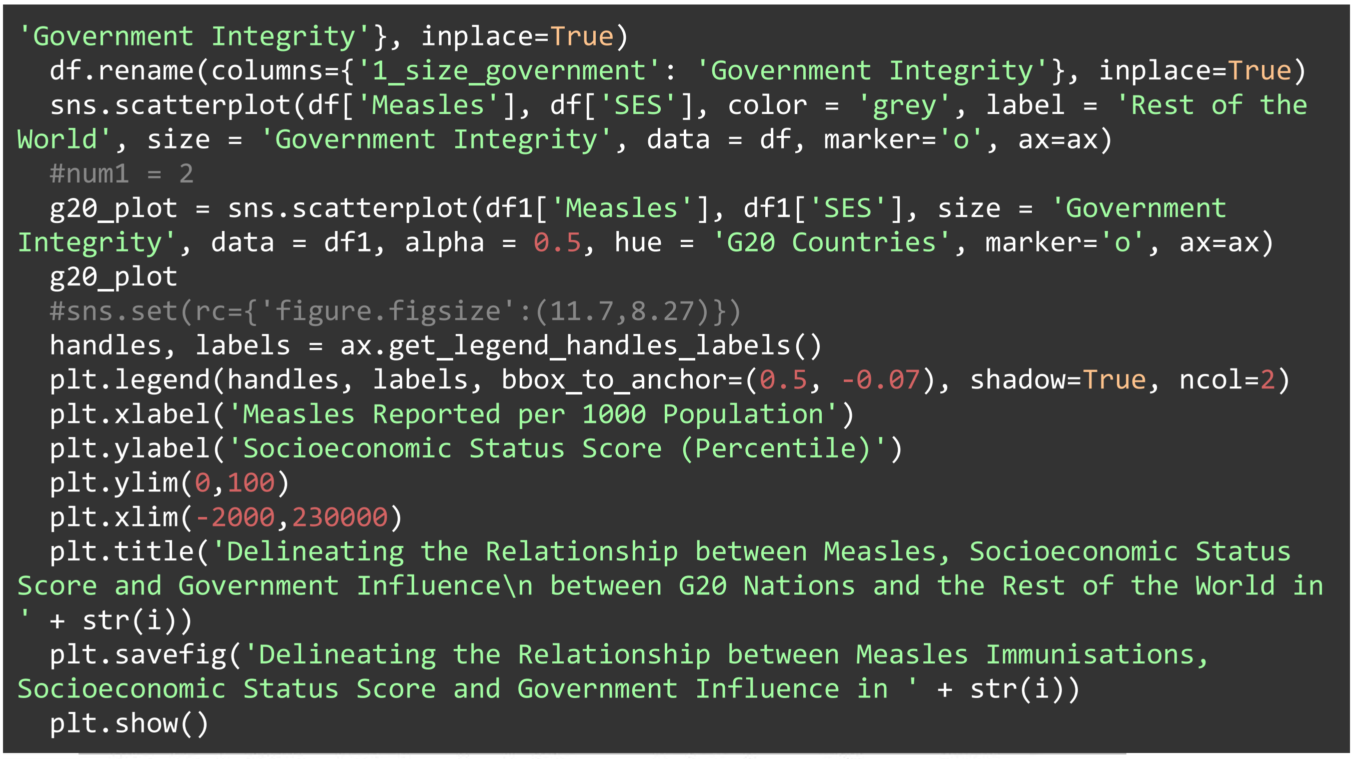


Similarly, this same new process was utilised amongst the rest of the scatter plots only changing the data frames values for different measure comparisons, while limits were also adjusted to ensure no inaccuracies arose. The code and graphs are as follows:

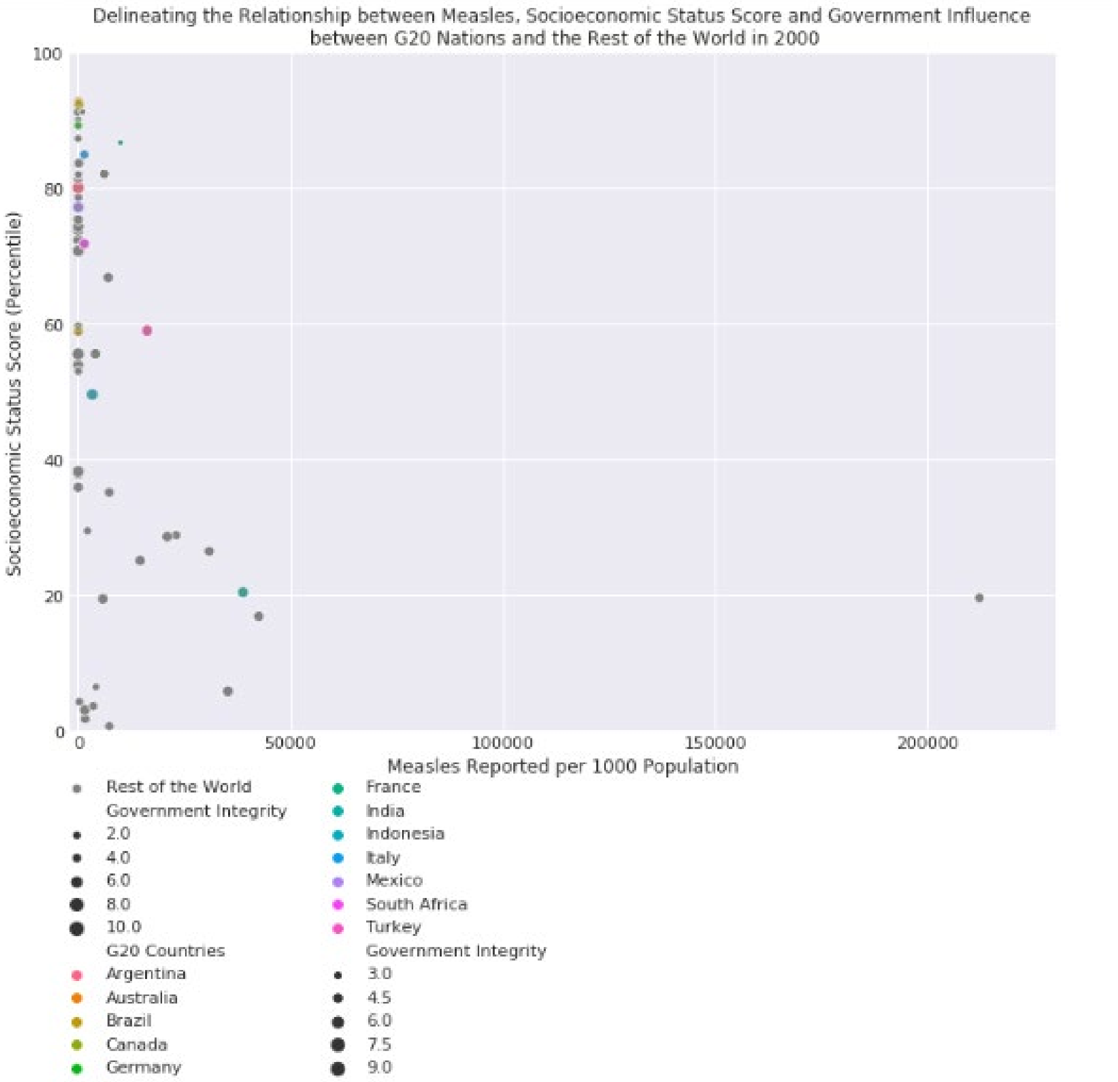
Delineating the Relationship between Measles, Socioeconomic Status Score and Government Influence between G20 Nations and the Rest of the World



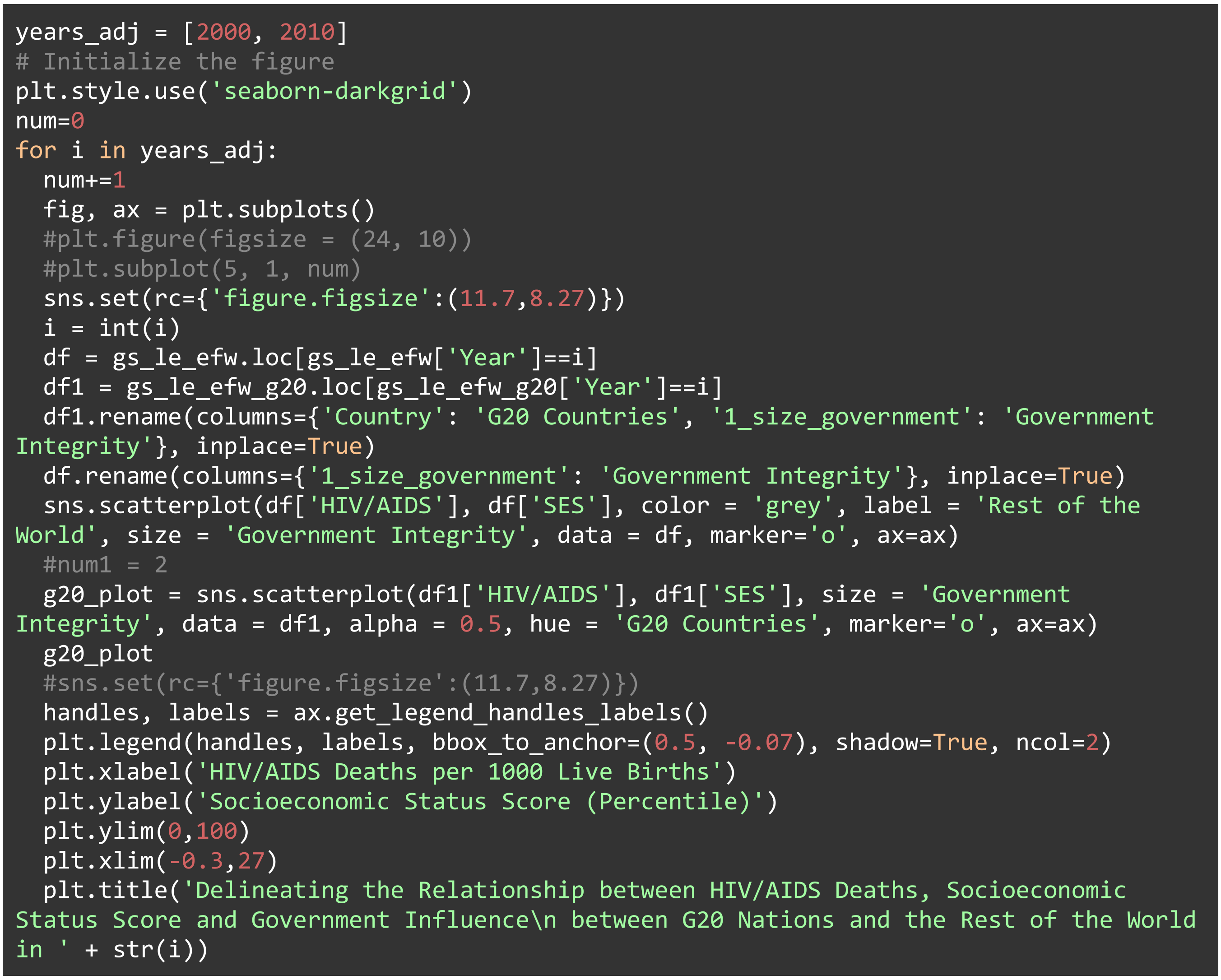
```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = gs_le_efw.loc[gs_le_efw['Year']==i]
  df1 = gs_le_efw_g20.loc[gs_le_efw_g20['Year']==i]
  df1.rename(columns={'Country': 'G20 Countries', '1_size_government':
```



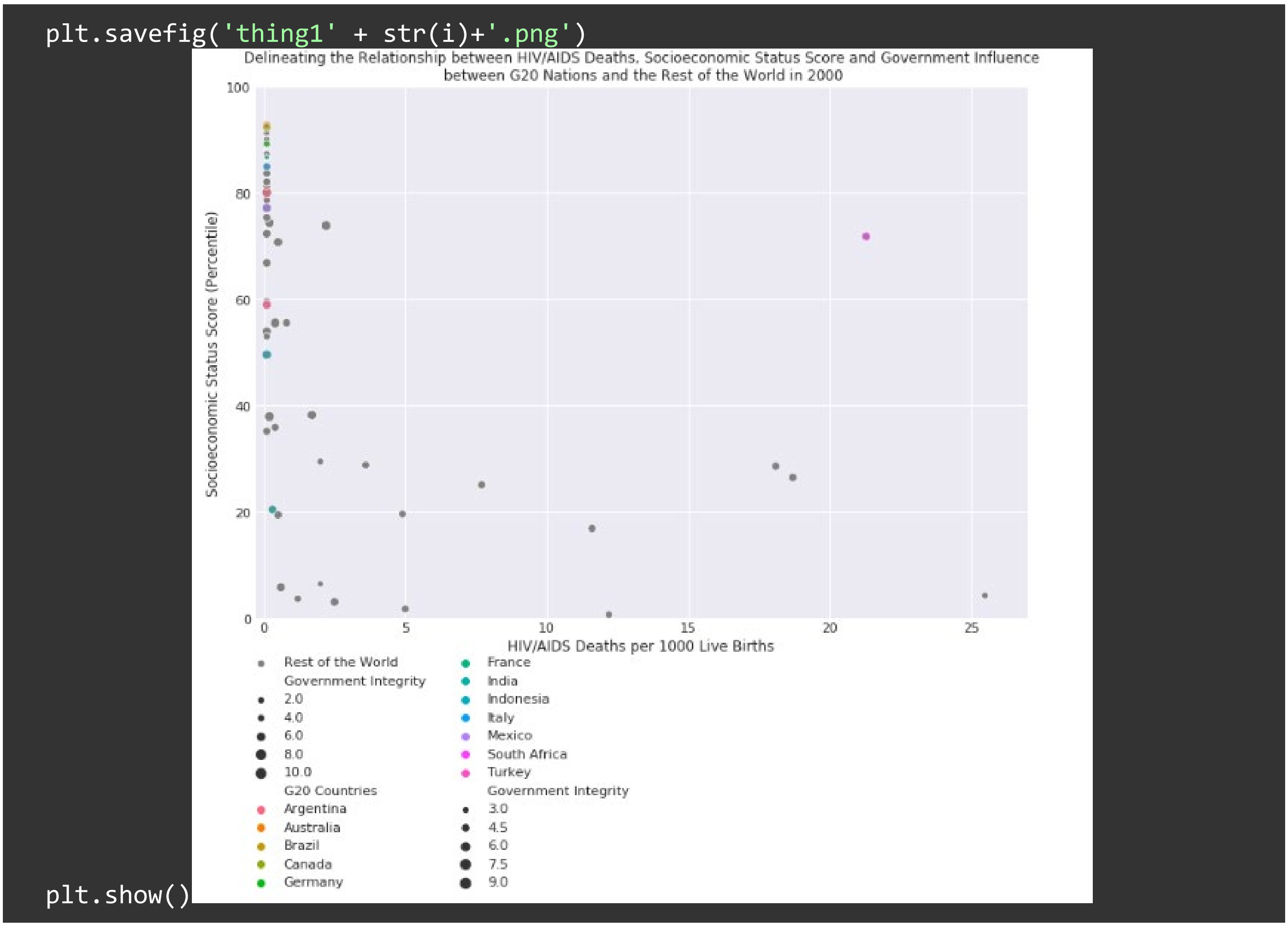
```
'Government Integrity'}, inplace=True)
  df.rename(columns={'1_size_government': 'Government Integrity'}, inplace=True)
  sns.scatterplot(df['Measles'], df['SES'], color = 'grey', label = 'Rest of the World', size = 'Government Integrity', data = df, marker='o', ax=ax)
  #num1 = 2
  g20_plot = sns.scatterplot(df1['Measles'], df1['SES'], size = 'Government Integrity', data = df1, alpha = 0.5, hue = 'G20 Countries', marker='o', ax=ax)
  g20_plot
  #sns.set(rc={'figure.figsize':(11.7,8.27)})
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels, bbox_to_anchor=(0.5, -0.07), shadow=True, ncol=2)
  plt.xlabel('Measles Reported per 1000 Population')
  plt.ylabel('Socioeconomic Status Score (Percentile)')
  plt.ylim(0,100)
  plt.xlim(-2000,230000)
  plt.title('Delineating the Relationship between Measles, Socioeconomic Status Score and Government Influence\n between G20 Nations and the Rest of the World in ' + str(i))
  plt.savefig('Delineating the Relationship between Measles Immunisations, Socioeconomic Status Score and Government Influence in ' + str(i))
  plt.show()
```



Delineating the Relationship between HIV/AIDS Deaths, Socioeconomic Status Score and Government Influence between G20 Nations and the Rest of the World



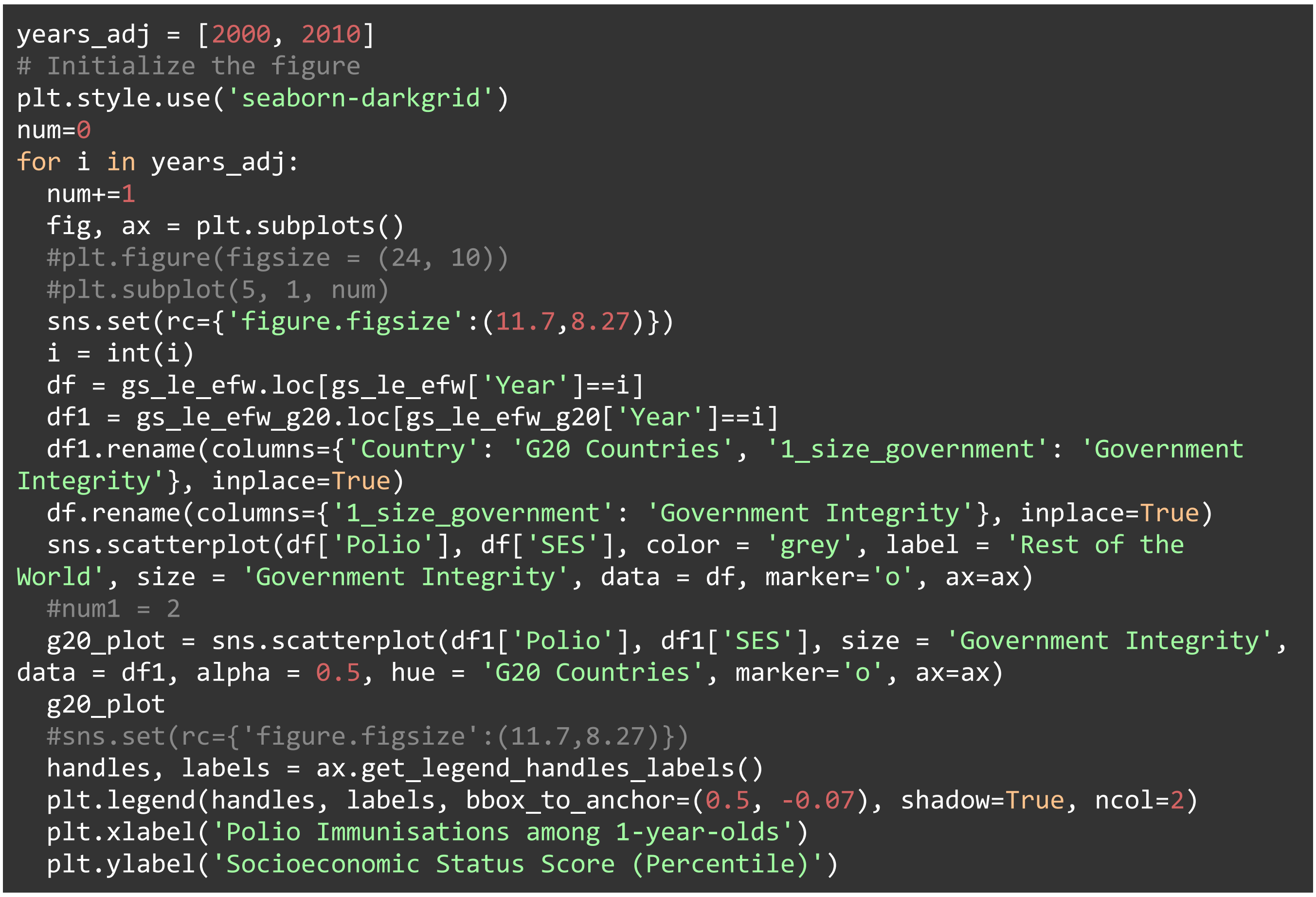
```python
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = gs_le_efw.loc[gs_le_efw['Year']==i]
  df1 = gs_le_efw_g20.loc[gs_le_efw_g20['Year']==i]
  df1.rename(columns={'Country': 'G20 Countries', '1_size_government': 'Government Integrity'}, inplace=True)
  df.rename(columns={'1_size_government': 'Government Integrity'}, inplace=True)
  sns.scatterplot(df['HIV/AIDS'], df['SES'], color = 'grey', label = 'Rest of the World', size = 'Government Integrity', data = df, marker='o', ax=ax)
  #num1 = 2
  g20_plot = sns.scatterplot(df1['HIV/AIDS'], df1['SES'], size = 'Government Integrity', data = df1, alpha = 0.5, hue = 'G20 Countries', marker='o', ax=ax)
  g20_plot
  #sns.set(rc={'figure.figsize':(11.7,8.27)})
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels, bbox_to_anchor=(0.5, -0.07), shadow=True, ncol=2)
  plt.xlabel('HIV/AIDS Deaths per 1000 Live Births')
  plt.ylabel('Socioeconomic Status Score (Percentile)')
  plt.ylim(0,100)
  plt.xlim(-0.3,27)
  plt.title('Delineating the Relationship between HIV/AIDS Deaths, Socioeconomic Status Score and Government Influence\n between G20 Nations and the Rest of the World in ' + str(i))
```



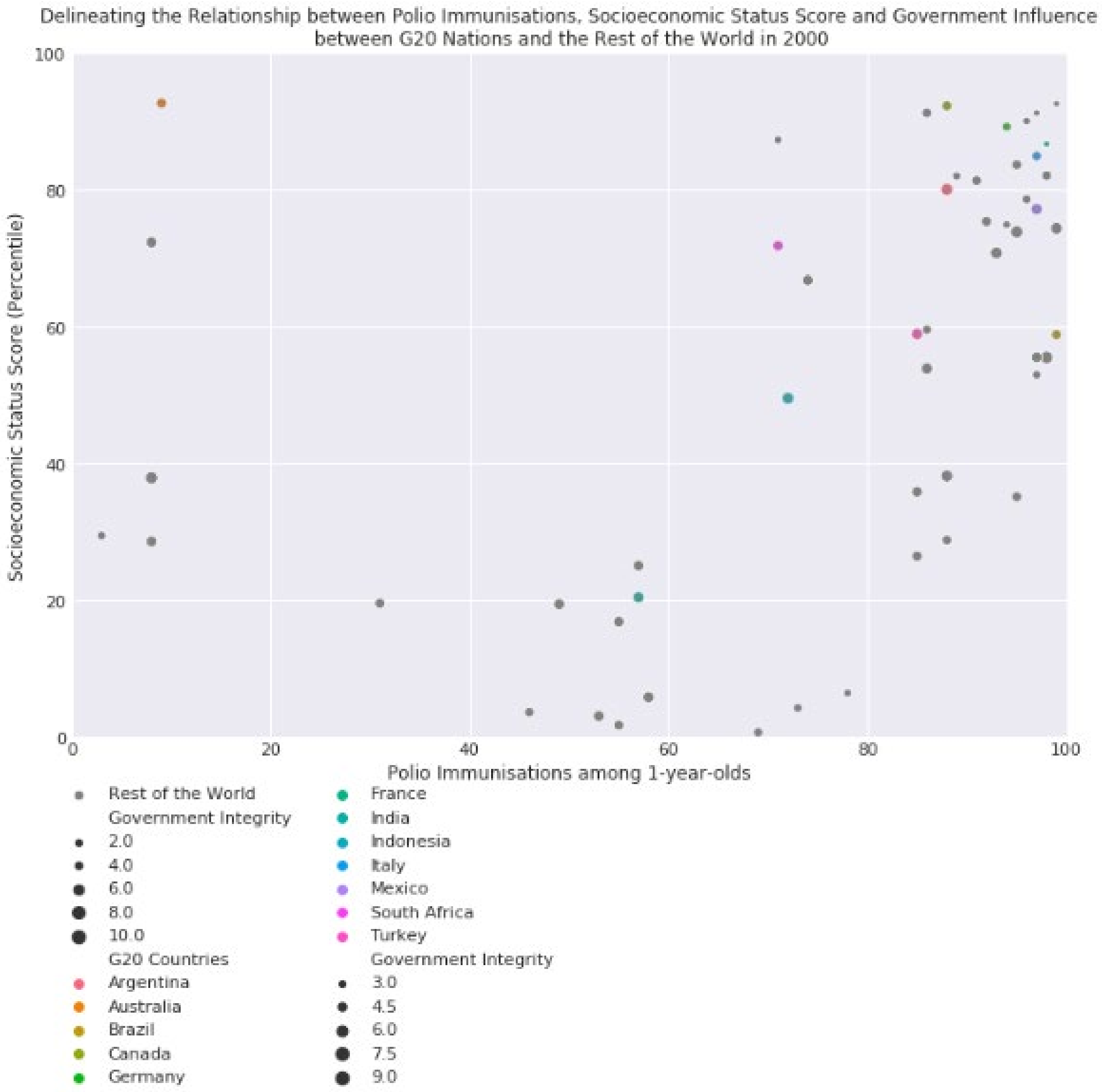
```
  plt.savefig('thing1' + str(i)+'.png')
  plt.show()
```

Delineating the Relationship between HIV/AIDS Deaths, Socioeconomic Status Score and Government Influence
between G20 Nations and the Rest of the World in 2000
Socioeconomic Status Score (Percentile)
100
80
60
40
20
0
0
5
10
15
20
25
HIV/AIDS Deaths per 1000 Live Births
Rest of the World
Government Integrity
2.0
4.0
6.0
8.0
10.0
G20 Countries
Argentina
Australia
Brazil
Canada
Germany
France
India
Indonesia
Italy
Mexico
South Africa
Turkey
Government Integrity
3.0
4.5
6.0
7.5
9.0

<u>**Delineating the Relationship between Polio Immunisations, Socioeconomic Status Score and Government Influence between G20 Nations and the Rest of the World**</u>



```
years_adj = [2000, 2010]
# Initialize the figure
plt.style.use('seaborn-darkgrid')
num=0
for i in years_adj:
  num+=1
  fig, ax = plt.subplots()
  #plt.figure(figsize = (24, 10))
  #plt.subplot(5, 1, num)
  sns.set(rc={'figure.figsize':(11.7,8.27)})
  i = int(i)
  df = gs_le_efw.loc[gs_le_efw['Year']==i]
  df1 = gs_le_efw_g20.loc[gs_le_efw_g20['Year']==i]
  df1.rename(columns={'Country': 'G20 Countries', '1_size_government': 'Government 
Integrity'}, inplace=True)
  df.rename(columns={'1_size_government': 'Government Integrity'}, inplace=True)
  sns.scatterplot(df['Polio'], df['SES'], color = 'grey', label = 'Rest of the 
World', size = 'Government Integrity', data = df, marker='o', ax=ax)
  #num1 = 2
  g20_plot = sns.scatterplot(df1['Polio'], df1['SES'], size = 'Government Integrity', 
data = df1, alpha = 0.5, hue = 'G20 Countries', marker='o', ax=ax)
  g20_plot
  #sns.set(rc={'figure.figsize':(11.7,8.27)})
  handles, labels = ax.get_legend_handles_labels()
  plt.legend(handles, labels, bbox_to_anchor=(0.5, -0.07), shadow=True, ncol=2)
  plt.xlabel('Polio Immunisations among 1-year-olds')
  plt.ylabel('Socioeconomic Status Score (Percentile)')
  plt.ylim(0,100)
  plt.xlim(0,100)
  plt.title('Delineating the Relationship between Polio Immunisations, Socioeconomic Status Score and Government Influence\n between G20 Nations and the Rest of the World in ' + str(i))
  plt.savefig('thing2' + str(i)+'.png')
  plt.show()
```

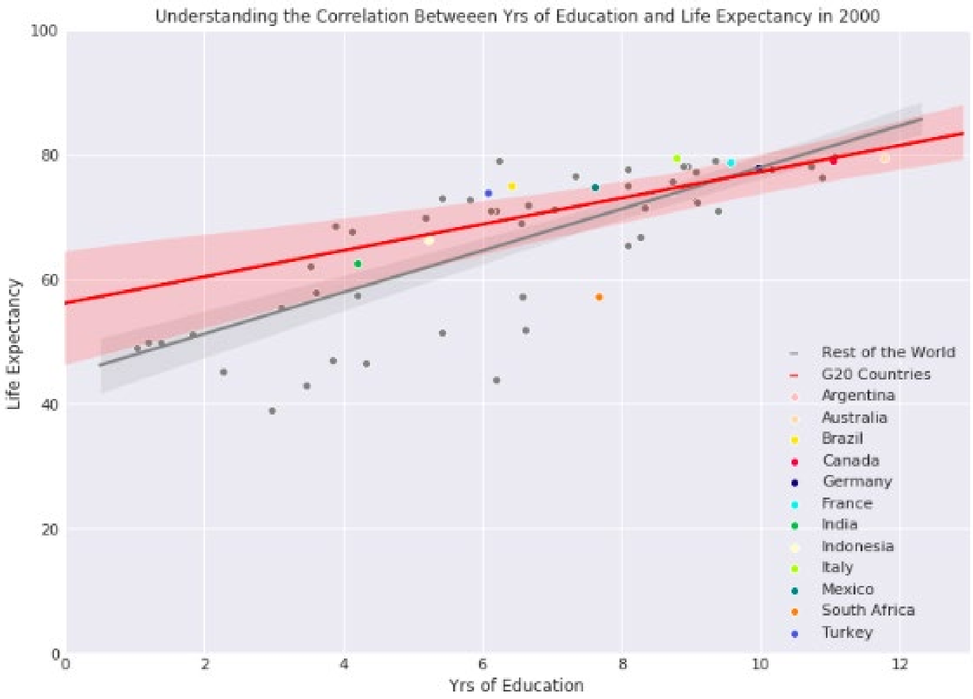


This concludes the methodologies and techniques utilised to produce the charts and tables mentioned above.

Section 3: Audiences Interested in Predictive Models

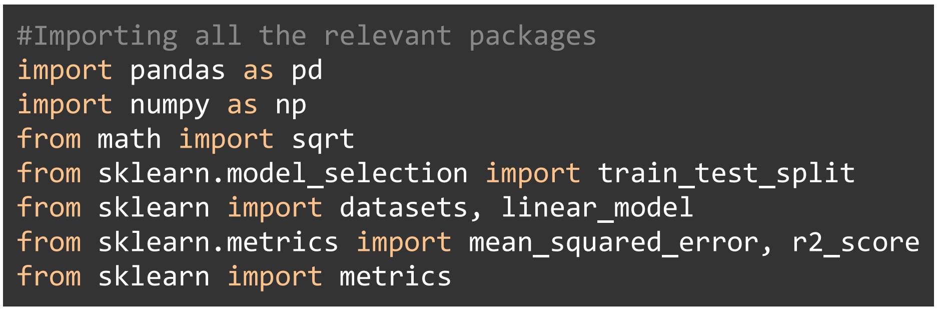
Furthering our understanding as gained from the creation and analysis of graphical elements, the development of predictive machine learning models was incorporated. Similar to section 2, this section will detail the methodologies involved to produce them and is primarily intended for providing insights to individuals with an interest in IT approaches to data analysis. Again, the code was executed within the Google Colab python notebook environment.

Initial preparation and direction was critical to positing the intended aim to be achieved with such machine learning tools. This involved a careful consideration of relationships depicted in the graphs created earlier. Having been convinced by the strong correlation posed between Years of Education and Life Expectancy, the decision was made to produce a linear regression model that uses input variables GDP per Capita and years of education to predict life expectancy.



Following the decision, an acknowledgement to recognising the potential of predicting a country's developing or developed status was realised. While there was the possibility to include a model that considered the prediction of this target variable with account of all measures from all three datasets this would extremely limit the data. This is due to the inconsistencies between datasets that causes particular rows to drop by the pd.merge function due to misalignment. Thus, the choice to delineate country status predictions was made with measures from the socioeconomic and life expectancy data set. Through a binary target variable approach, this was best achieved through applying a logistic regression model.

Delving into the methodologies, the relevant packages were first imported.



```
#Importing all the relevant packages
import pandas as pd
import numpy as np
from math import sqrt
from sklearn.model_selection import train_test_split
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score
from sklearn import metrics
```

Using the previously created yrseduc_le dataframe this contains all the required measures from both the socioeconomic and life expectancy dataset.

```
#Linear Regression across all countries using yrseduc_le dataframe
df1 = yrseduc_le.loc[yrseduc_le['Year'] == 2010]
df2 = yrseduc_le.loc[yrseduc_le['Year'] == 2000]
```

Having filtered the dataset with its respect years, the year 2015 was considered as the base year from which the predictions would be made as it is the most recent year available within the dataset. With the data frame, slicing was used to acquire both the input (independent) variables and target (dependent) variables. Effectively, columns in the data frame associated with 5 and 6 were selected, these were GDP per capita and years of education for the input and 9 as the target of which is life expectancy.

```
# Use only two features
X = df1.values[:, 5:7]
X1 = df1.values[:, 6]
y = df1.values[:, 9]
```

Utilising the train_test_split function offered by the pandas library the data was split for training and testing. The test size in this instance was made as 60% of all values.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.6,
random_state=42)
```

The input and target variables having been readied for modelling the linear regression model was prepared.

```
# Create linear regression object
regr = linear_model.LinearRegression().fit(X_train, y_train)
```

Therefore, with a linear model produced a sample can be inputted into the model to produce a prediction. For example below the GDP per capita of 40000 and 6 years of education predicted a 79 years life expectancy.

```
# Let's create one sample and predict the Life expectancy
sample = [40000, 6]        # a sample with a 40000 GDPpc and 6 yrs of education
print('----- Sample case -----')
for column, value in zip(list(df1)[5:7], sample):
    print(column + ': ' + str(value))
sample_pred = regr.predict([sample])
print('Predicted life expectancy:', int(sample_pred))
print('-----------------------')
```

OUTPUT:

```
----- Sample case -----
gdppc: 40000
yrseduc: 6
Predicted life expectancy: 79
-----------------------
```

Next, some manipulation was required before it would be ready for applying the logistic regression model. An adjusted le_efw data frame was produced excluding all columns that contained strings, except the status column of which was relocated to the end. The year was then filtered for 2010 and the final data frame having its indices reset to its new position and the year column dropped.

```
le_efw_adjusted = le_efw[['Year', 'ECONOMIC FREEDOM', 'rank', 'quartile',
'1a_government_consumption', '1b_transfers', '1c_gov_enterprises',
'1d_top_marg_tax_rate', '1_size_government', '2a_judicial_independence',
'2b_impartial_courts', '2c_protection_property_rights',
'2d_military_interference', '2e_integrity_legal_system',
'2f_legal_enforcement_contracts', '2g_restrictions_sale_real_property',
'2h_reliability_police', '2i_business_costs_crime', '2j_gender_adjustment',
'2_property_rights', '3a_money_growth', '3b_std_inflation', '3c_inflation',
'3d_freedom_own_foreign_currency', '3_sound_money', '4a_tariffs',
'4b_regulatory_trade_barriers', '4c_black_market',
'4d_control_movement_capital_ppl', '4_trade', '5a_credit_market_reg',
'5b_labor_market_reg', '5c_business_reg', '5_regulation', 'Life expectancy',
'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expenditure',
'Hepatitis B', 'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total
expenditure', 'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness 1-19
years', 'thinness 5-9 years', 'Income composition of resources', 'Status']]
df = le_efw_adjusted.loc[le_efw_adjusted['Year']==2010]
df = df.reset_index(drop=True)
df = df[['ECONOMIC FREEDOM', 'rank', 'quartile', '1a_government_consumption',
'1b_transfers', '1c_gov_enterprises', '1d_top_marg_tax_rate',
'1_size_government', '2a_judicial_independence', '2b_impartial_courts',
'2c_protection_property_rights', '2d_military_interference',
'2e_integrity_legal_system', '2f_legal_enforcement_contracts',
'2g_restrictions_sale_real_property', '2h_reliability_police',
'2i_business_costs_crime', '2j_gender_adjustment', '2_property_rights',
'3a_money_growth', '3b_std_inflation', '3c_inflation',
'3d_freedom_own_foreign_currency', '3_sound_money', '4a_tariffs',
'4b_regulatory_trade_barriers', '4c_black_market',
'4d_control_movement_capital_ppl', '4_trade', '5a_credit_market_reg',
'5b_labor_market_reg', '5c_business_reg', '5_regulation', 'Life expectancy',
'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expenditure',
'Hepatitis B', 'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total
expenditure', 'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness 1-19
years', 'thinness 5-9 years', 'Income composition of resources', 'Status']]
```

Utilising the shape method the independent and dependent variables was sliced respectively where all measures except the last column were included as inputs. The last column was selected as the target. Similarly to above, the train_test_split function was applied of which again 60% of the data was selected for testing.

```
#Building the Model
[num_row, num_var] = df.shape
X = df.values[:, 0:num_var - 1]       # slice dataFrame to extract input variables
y = df.values[:, num_var - 1]         # slice dataFrame to extract target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)
```

The LogisticRegression object was finally created where it uses the liblinear solver. It is essentially a large-linear classification model that supports logistic regression and linear support vector machines.

```
clf = linear_model.LogisticRegression(solver='liblinear').fit(X_train, y_train)
```

Retrieving such a model classification of the test data can be achieved, wherein the predict method is called. This allows a derivation of the probability that a country's variables may imply it being developed or developing; attained by the predict_proba method.

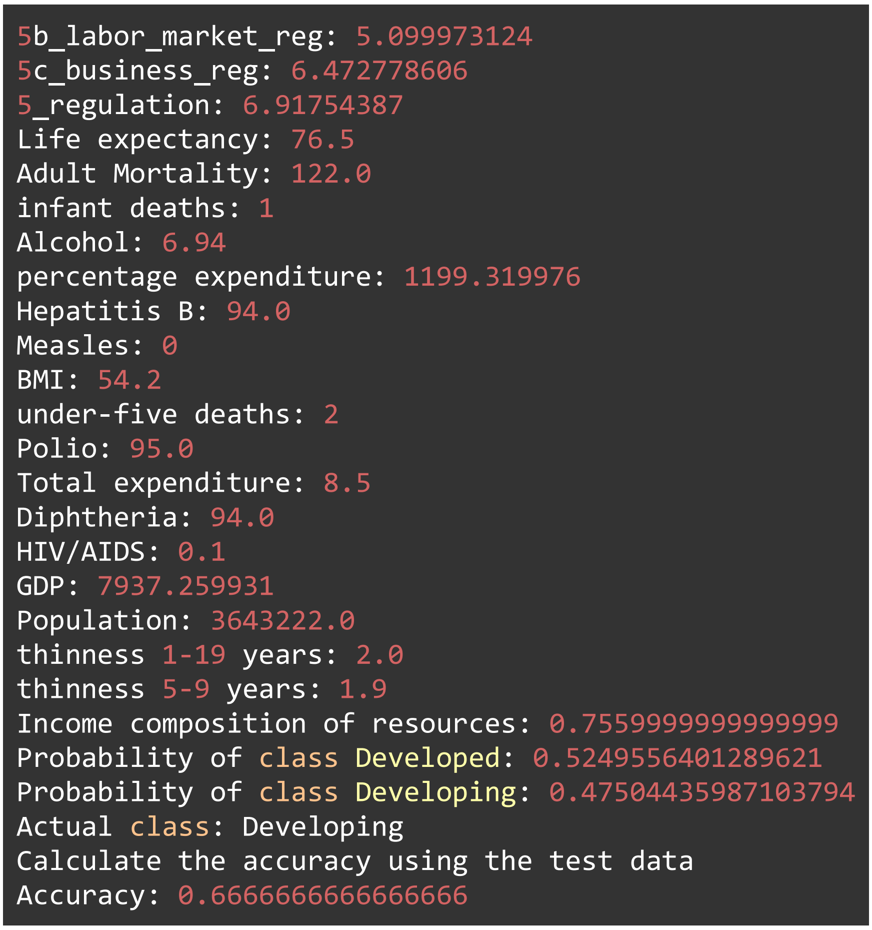
```
_pred = clf.predict(X_test)
y_pred_proba = clf.predict_proba(X_test)
```

Finally a sample was selected from the test data, this being the last row. While the for loop produces each variable and its value the probabilities of each class status is displayed. The actual status is included below.

```
# Let's get one sample and predict the probabilities
print('----- Sample case -----')
last_sample = X_test[-1]
for column, value in zip(list(df), last_sample):
    print(column + ': ' + str(value))
last_sample_proba = y_pred_proba[-1]
print('Probability of class Developed:', last_sample_proba[0])
print('Probability of class Developing:', last_sample_proba[1])
print('Actual class:', str(y_test[-1]))
print('Calculate the accuracy using the test data')
print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

OUTPUT:

```
----- Sample case -----
ECONOMIC FREEDOM: 7.22
rank: 52.0
quartile: 2.0
1a_government_consumption: 6.705882353
1b_transfers: 8.555858311
1c_gov_enterprises: 6.0
1d_top_marg_tax_rate: 6.0
1_size_government: 6.815435166
2a_judicial_independence: 1.848050606
2b_impartial_courts: 4.111011333
2c_protection_property_rights: 6.252152159
2d_military_interference: 8.333333332999999
2e_integrity_legal_system: 5.0
2f_legal_enforcement_contracts: 2.25658445
2g_restrictions_sale_real_property: 7.653054843
2h_reliability_police: 5.638990452000001
2i_business_costs_crime: 4.482692988
2j_gender_adjustment: 0.962962963
2_property_rights: 4.970208063
3a_money_growth: 8.287965518
3b_std_inflation: 8.919908286
3c_inflation: 9.301742255
3d_freedom_own_foreign_currency: 10.0
3_sound_money: 9.127404015
4a_tariffs: 7.878248317000001
4b_regulatory_trade_barriers: 7.577731788999999
4c_black_market: 10.0
4d_control_movement_capital_ppl: 7.690765802
4_trade: 8.286686477
5a_credit_market_reg: 9.17987988
```



```
5b_labor_market_reg: 5.099973124
5c_business_reg: 6.472778606
5_regulation: 6.91754387
Life expectancy: 76.5
Adult Mortality: 122.0
infant deaths: 1
Alcohol: 6.94
percentage expenditure: 1199.319976
Hepatitis B: 94.0
Measles: 0
BMI: 54.2
under-five deaths: 2
Polio: 95.0
Total expenditure: 8.5
Diphtheria: 94.0
HIV/AIDS: 0.1
GDP: 7937.259931
Population: 3643222.0
thinness 1-19 years: 2.0
thinness 5-9 years: 1.9
Income composition of resources: 0.7559999999999999
Probability of class Developed: 0.5249556401289621
Probability of class Developing: 0.47504435987103794
Actual class: Developing
Calculate the accuracy using the test data
Accuracy: 0.6666666666666666
```

This concludes the methodologies and techniques utilised to produce the machine learning models mentioned above.

References:

1. Who.int. (2019). *Hepatitis B*.
   Available at: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b [Accessed 3 Nov. 2019].
2. Who.int. (2019). *HIV/AIDS*.
   Available at: https://www.who.int/news-room/fact-sheets/detail/hiv-aids [Accessed 3 Nov. 2019].
3. Who.int. (2019). *WHO | Understanding the correlations between wealth, poverty and human immunodeficiency virus infection in African countries*.
   Available at: https://www.who.int/bulletin/volumes/88/7/09-070185/en/ [Accessed 3 Nov. 2019].
4. Igulot, P. and Magadi, M.A. (2018). Socioeconomic Status and Vulnerability to HIV Infection in Uganda: Evidence from Multilevel Modelling of AIDS Indicator Survey Data. *AIDS Research and Treatment*, 2018, pp.1–15.
   Available at: https://www.hindawi.com/journals/art/2018/7812146/ [Accessed 3 Nov. 2019].
5. Bunyasi, E.W. and Coetzee, D.J. (2017). Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa. *BMJ Open*, 7(11), p.e016232.
6. Burkey, M.D., Weiser, S.D., Fehmie, D., Alamo-Talisuna, S., Sunday, P., Nannyunja, J., Reynolds, S.J. and Chang, L.W. (2014). Socioeconomic Determinants of Mortality in

HIV. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, [online] 66(1), pp.41–47. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3981890/ [Accessed 3 Nov. 2019].

7. Perry, R.T. and Halsey, N.A. (2004). The Clinical Significance of Measles: A Review. *The Journal of Infectious Diseases*, [online] 189(Supplement_1), pp.S4–S16. Available at: https://academic.oup.com/jid/article/189/Supplement_1/S4/823958.
8. Hoes, J., Boef, A.G.C., Knol, M.J., de Melker, H.E., Mollema, L., van der Klis, F.R.M., Rots, N.Y. and van Baarle, D. (2018). Socioeconomic Status Is Associated With Antibody Levels Against Vaccine Preventable Diseases in the Netherlands. *Frontiers in Public Health*, [online] 6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6094970/ [Accessed 3 Nov. 2019].
9. Wilson, K., Ducharme, R. and Hawken, S. (2013). Association between socioeconomic status and adverse events following immunization at 2, 4, 6 and 12 months. *Human Vaccines & Immunotherapeutics*, [online] 9(5), pp.1153–1157. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899153/ [Accessed 3 Nov. 2019].
10. Measles. (2018). *World Health Organization*. [online] Available at: https://www.who.int/immunization/diseases/measles/en/.
11. Australian Government Department of Health (2019). *Measles*. [online] Australian Government Department of Health. Available at: https://www.health.gov.au/health-topics/measles.
12. Tosun, S., Aygün, O., Özdemir, H.Ö., Korkmaz, E. and Özdemir, D. (2018). The impact of economic and social factors on the prevalence of hepatitis B in Turkey. *BMC Public Health*, 18(1).
13. Shahbaz, M., Loganathan, N., Mujahid, N., Ali, A. and Nawaz, A. (2015). Determinants of Life Expectancy and its Prospects Under the Role of Economic Misery: A Case of Pakistan. *Social Indicators Research*, 126(3), pp.1299-1316.
14. Nies, M., Sun, L., Kazemi, D., Carriker, A. and Dmochowski, J. (2019). *Relationship of Body Mass Index to Alcohol Consumption in College Freshmen*.
15. News, A. (2019). *Holiday Hangover: Alcohol Increases SIDS Risk*. [online] ABC News. Available at: https://abcnews.go.com/Health/holiday-hangover-alcohol-increases-risk-sids-deaths/story?id=12495059 [Accessed 3 Nov. 2019].
16. Jürgen Rehm, K. (2019). *Alcohol and Mortality: Global Alcohol-Attributable Deaths From Cancer, Liver Cirrhosis, and Injury in 2010*. [online] PubMed Central (PMC). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908708/#b67-arcr-35-2-174 [Accessed 3 Nov. 2019].
17. Health Knowledge. (2019). *Health Effects of International Trade*. [online]Available at: https://www.healthknowledge.org.uk/public-health-textbook/medical-sociology-policy-economics/4c-equality-equity-policy/international-influence [Accessed 3 Nov. 2019].