Analysis on the Relationship between Population Growth Rate, Life Expectancy and the Gross National Income

Victory Chimamaka Uwaoma
MSc. Data Analysis
National College of Ireland
Dublin, Ireland
x19210931@student.ncirl.ie

Popoola Oyinlola Jadesola
MSc. Data Analysis
National College of Ireland
Dublin, Ireland
x19138202@student.ncirl.ie

Ajufo Kenechukwu Otito
MSc. Data Analysis
National College of Ireland
Dublin, Ireland
x19190174@student.ncirl.ie

***Abstract*— The study aims to examine the correlation between the total life expectancy, population growth rate in relation to gross national income of the countries from our data source and also to determine if there is significance using population growth rate and gross national income to predict the life expectancy of a country. It would be expected that as population grows, the gross national income grows, therefore, the analysis aims to check if when the life expectancy increases, the population growth rate and gross national income also increases. Exploratory data analysis and multiple linear regression and appropriate visualization tools would be used to check for patterns and relationship between each dataset. Therefore, appropriate visualizations would be done to easily understand these patterns and to visualize relationships properly. The result shows that there was no statistical significance of using the population growth rate and gross national income in predicting life expectancy.**

Keywords—Population Growth Rate, Life expectancy, Gross National Income, Exploratory Analysis, Visualization, MongoDB, PostgreSQL.

I. INTRODUCTION

Population growth rate is the rate at which the number of people in a geographical location rises within a period. It is measured as the ratio of the increase in population to the initial population over a given time span. Population can change in size with relation to time, there are several different factors that can influence this change. When births rise above deaths the population growth increases this is known as the birth rate, while the population reduces when the deaths surpasses the births, and this is known as the death rate. There are other factors that influence the population size and growth, such as net immigration, cost of education, economic growth, availability of contraceptive, social norms, government policy etc.

Life Expectancy is described as the time in which an organism is expected to live. It is commonly estimated using the organism's present age, it's year of birth and often demographic aspects like race, gender etc., as well as lifestyle habits. Life Expectancy can be used as a benchmark for checking the comprehensive health of a society, it is a thorough way of checking health levels across the age groups in the society, it can help understand mortality trends as an increase in the life expectancy in a particular society implies increase in population growth.

Gross national income (GNI), which was known formerly as gross national product (GNP) is accumulated value of the money received by the citizens of a country and that of their various businesses. It is the combination of the country's gross domestic product (GDP) and revenue gotten overseas. Basically, GNI can be used to calculate and track a country's bulk of finances annually as it is an effective means of tracking a country's resources.

Understanding Population growth rate, Life Expectancy and Gross National Income is crucial, it helps countries plan - it can be useful to know the population size for the provisions of the finite available resources and basic social amenities needed to support the society. Examples of these include housing, transportation, hospitals, electricity and demand of food and water etc. Furthermore, Life expectancy influences economic growth directly or indirectly - when the populace works optimally when they remain in relatively good health. Lastly, GNI, allows countries have a glimpse the structure of the economy, sense metrics such as inflation, standard of living etc.

Research questions

a. What is the relationship between Gross National Income and Population Growth Rate?
b. How significant is Gross National Income and Population Growth Rate in predicting Life Expectancy?
c. Is there a difference between the regions in the Life Expectancy?

II. RELATED WORKS

Despite common generalizations that females have higher life expectancy than males, studies show that there has been fluctuating differences in the life expectancy of females and males over the years and a recent convergence of this difference due to behavioral factors.[1]

Yan Le et al[2] study in China confirms the fluctuations of the gender gap and the contributions of factors that affects life expectancy in males and females thereby affecting the gender gap but fails to consider economic factors such as population growth rate and gross national income of China.

Using vector error correlation model, Mahyar Hami [3] linked life expectancy to GDP growth for Iran and showed that economic growth has a positive significant effect on life expectancy between 1966 and 2013 but the effect was negligible.

Cervellati & Sunde [4] tests the hypothesis that improvements in life expectancy adds to the average income across countries using OLS estimate and 2SLS regressions

and suggests that high life expectancy significantly triggers a transition to viable income growth.

Assessing the relationship between pollution, health and income in China from 1991-2012 using a regression method, there was a positive correlation between GDP and life expectancy. Though the results found no relationship between income and life expectancy at age five it predicts that a double income would lead to a 3.5-year gain in life expectancy.[5]

Using OLS estimates and 2SLS regression, Acemoglu and Johnson[6] documented that an increase in life expectancy results in twice as much increase in population but a smaller increase on total GDP. They also debunked the theory that a high increase in life expectancy raises income per capita.

Data mining involves graphical or numerical processing of large amounts data to extract information. The size of the data is the major difference between ordinary statistical applications and data mining and initial exploration of data is required to identify the most suitable model in data analysis, and diagnostic methods usually verifies the appropriateness of the choice of model to an extent but not the validity of a model. [7]

Statistical inference should be integrated with statistical concepts and exploratory data analysis in fields that evaluation of information must be based on data-based evidence[8]. Furthermore, an integrated description of inferential reasoning should include the ideas of informal inferential reasoning, formal inferential reasoning and activity of argumentation (production of statistical and contextual reasons). The logic of formal statistical appear highly abstract but it is easier to make sense of when accompanied with informal statistical inference [9]. Three principles essential to informal statistical inference are generalizations that extend beyond describing the given data, use of data as proof for the generalizations and use of probabilistic language in describing the generalizations including informal reference to levels of certainty about conclusions[10].

Glenn J. et al [11] outlines the steps in exploratory data analysis starting with preparation of the dataset(s) by cleaning, removing observations and variables such as variables with a lot of missing values, generating consistent scales across variables, defining new frequency distribution, conversion of text to numbers, combining variables by using mathematical operations to derived needed variables from existing ones and generating groups by creating subsets of the data. Afterwards, relationships between variables are visualized and the metrics about the relationships are calculated. Then the groups in the subset of data are identified and understood and finally the models are built from the data using approaches such as linear regression, logistic regression, k-nearest neighbors and classification and regression trees (CART) in order to understand and measure relationship between variables.

III. METHODOLOGY

This report implements the Knowledge Discovery Database methodology (KDD). The method provides us an appropriate description of the data to analyze and help us to visualize, then discover the connections between population growth rate, life expectancy and gross national income.
The flowchart below sheds light on the methodology process throughout all the stages from the raw data through the pre-processing stage, data mining and pattern evaluation/interpretation.

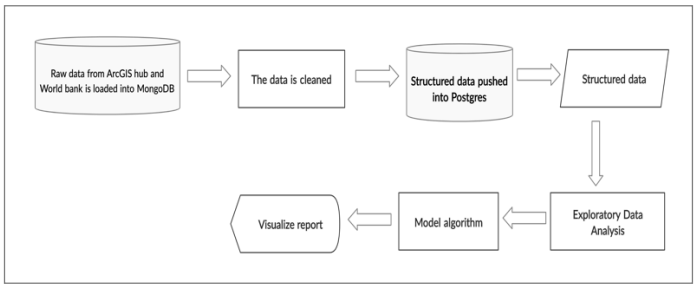

Figure 1: Flowchart showing the data pre-processing and visualization

The process carried out are:

- Data Source
- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation

This process will be carried out through the three related datasets.

a. Data Source

The data comprises of three unstructured datasets which were collected from both ArcGIS Hub and the world bank APIs, each having their domain in population, life expectancy and income of several countries. There were no size limitations in MongoDB since the data fulfilled all criteria. They were then loaded into MongoDB database directly from their APIs after several processing (data wrangling).

The population dataset gives information about the population growth rate of different countries, the square meter and years and regions of countries. The life expectancy data gives insight about the life expectancy at birth of males, females and also the total life expectancy of both, the square meters, years and regions of countries. Gross national income (GNI) contains the gross national income of countries, the different countries, and years.

Table 1: Population growth rate

Variable name	Type	Description
ObjectId	Number	Unique ID
Sovereign Territory	Text	Countries divided into sovereign states
Type	Text	Type of countries
Name	Text	Name of all the countries
Name-long	Text	Long name of the countries

Abbrev	Text	Abbreviations of the countries' names
Formal name	Text	Formal name of the countries
Region	Text	Regions of the countries
SqMI	Number	Square meter of the countries
Population growth rate	Number	Population growth rate of countries
Year	Number	From 1960-2014
Date start	Date or Time	The start date and time
Date end	Date or Time	The end date and time

Table 2: Life Expectancy

Variable name	Type	Description
ObjectId	Number	Unique ID
Sovereign Territory	Text	Countries divided into sovereign states
Type	Text	Type of countries
Name	Text	Name of all the countries
Name-long	Text	Long name of the countries
Abbrev	Text	Abbreviations of the countries' names
Formal name	Text	Formal name of the countries
Region	Text	Regions of the countries
SqMI	Number	Square meter of the countries
Life Expectancy (female)	Number	Life Expectancy at birth female (year)
Life Expectancy (male)	Number	Life Expectancy at birth male (year)
Life Expectancy (total)	Number	Total Life Expectancy at birth of male and female (year)
Year	Number	From 1960-2014
Date start	Date or Time	The start date and time
Date end	Date or Time	The end date and time

Table 3: Gross National income

Variable name	Type	Description
Country	Text	Names of countries
Item	Text	The GNI information
Year	Date	From 1960-2018
Value	Number	Value of the gross national income (billions of dollars)

b. Data Cleaning

Python and a library called Pandas in Python was used in data cleaning and organizing data points in the dataset. Firstly, columns in the python dataframe that did not add much importance to the analysis was dropped using the drop() function leaving only details that prove to be crucial to the analytics. Columns such as the object id, long name, abbrev, formal name, start date and end date all from population and life expectancy, then the id columns which automatically generates when the data is loaded into MongoDB was also dropped. Then there was a check for NAs and missing values, the NAs found where also dropped using the dropna() function because they did not amount to 3% of the data. All these was done in preparing the data for analysis.

c. Data Intergration

From the data cleaned, the structured dataframes were then loaded into PostgreSQL database using a remote connection. This same connection was used to create the tables for each dataset and populate the tables in the database. The libraries psycopg2 and sqlalchemy were utilized in performing these tasks.

d. Data Selection

In this stage, SQL queries were used to retrieve data from tables in PostgreSQL to visualize each dataset. Population, Life Expectancy and Gross National Income were retrieved and joined with their common factor, year. This was done to check the relationship between them.

e. Data Transformation

The data for the life expectancy and population were transformed thus, the response from APIs were gotten and saved into variables, and the loaded into a single MongoDB database and their respective collections. For the Gross National Income, the url returns an archive file that contains the XML document. The archive file is programmatically unzipped it, then looped through to get the values out of XML document, saved into a dataframe. Since only the values were saved into the dataframe without their corresponding fields, the dataframe was converted to a numpy array, then reshaped and then converted dataframe. To save the data in MongoDB, the dataframe was further converted into a dictionary. After all the documents were loaded the into MongoDB, the database was queried and response from MongoDB: a pymongo cursor object was gotten, this was transformed into a list and from the list to a dataframe which is structured data form. Consequently, the dataframes were inserted into PostgreSQL, a relational database.

f. Data Mining

- **Correlation Matrix plots:** The Correlation Matrix plots was used to view the relationships between the population growth rate, the total life expectancy and the gross national income. This was done using the corr() function found in the seaborn library.

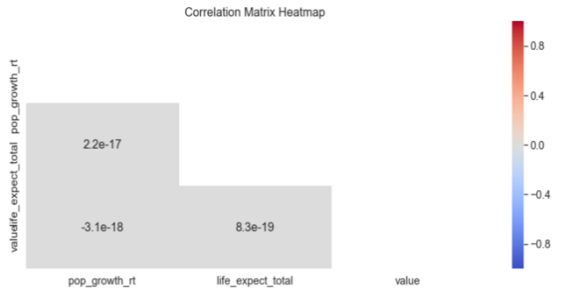

Figure 2: Correlation matrix heatmap

- **Multiple Linear Regression (MLR):** Multiple linear regression is a statistical method used to model a variable's value depending on two or more variables. The multiple linear regression was carried out using the total life expectancy as the target variable and population growth rate and gross national income as independent variables.

OLS Regression Results

Dep. Variable:	x	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	1.597e-07
Date:	Tue, 28 Apr 2020	**Prob (F-statistic):**	1.00
Time:	09:50:42	**Log-Likelihood:**	-3.0713e+07
No. Observations:	8215674	**AIC:**	6.143e+07
Df Residuals:	8215671	**BIC:**	6.143e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	67.0133	0.005	1.33e+04	0.000	67.003	67.023
y	-1.532e-28	2.86e-17	-5.36e-12	1.000	-5.6e-17	5.6e-17
z	2.733e-12	0.002	1.1e-09	1.000	-0.005	0.005

Omnibus:	790679.882	**Durbin-Watson:**	0.009
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	980297.579
Skew:	-0.827	**Prob(JB):**	0.00
Kurtosis:	2.639	**Cond. No.**	1.90e+14

Figure 3: Multiple linear regression output

- **Histogram:** Histograms is used to illustrate the distribution of variables. Figure 4 below shows the total life expectancy of 2013.

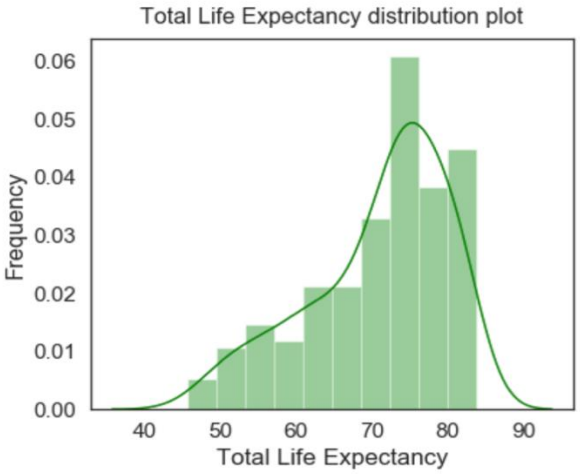

Figure 4: Total life expectancy using histogram

- **ANOVA:** ANOVA, shortened for Analysis of Variance, is a commonly used statistical tool for comparing means using statistical significance. The output of the ANOVA model used to estimate the variability of total life expectancy between the regions from the datasets gives the F statistic and p value ($F(6, 194) = 82.89$ $p < .005$, showing the model is significant. The coefficients show the difference in the means of the region. In *Fig: sums of squares (ANOVA)* the region row shows the between group effect. The sum of squares is for the model is how much variance explained by the model.

OLS Regression Results

Dep. Variable:	life_expect_total	**R-squared:**	0.719
Model:	OLS	**Adj. R-squared:**	0.711
Method:	Least Squares	**F-statistic:**	82.89
Date:	Tue, 28 Apr 2020	**Prob (F-statistic):**	7.42e-51
Time:	09:50:45	**Log-Likelihood:**	-623.69
No. Observations:	201	**AIC:**	1261.
Df Residuals:	194	**BIC:**	1284.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	70.0354	0.955	73.372	0.000	68.153	71.918
C(region_wb)[T.Europe & Central Asia]	3.5901	1.212	2.963	0.003	1.201	5.980
C(region_wb)[T.Latin America & Caribbean]	1.4210	1.330	1.068	0.287	-1.203	4.045
C(region_wb)[T.Middle East & North Africa]	1.3854	1.531	0.905	0.367	-1.633	4.404
C(region_wb)[T.North America]	7.8841	3.307	2.384	0.018	1.363	14.406
C(region_wb)[T.South Asia]	-6.3893	2.161	-2.957	0.003	-10.651	-2.127
C(region_wb)[T.Sub-Saharan Africa]	-18.1418	1.245	-14.568	0.000	-20.598	-15.686

Omnibus:	3.839	**Durbin-Watson:**	1.887
Prob(Omnibus):	0.147	**Jarque-Bera (JB):**	4.590
Skew:	0.068	**Prob(JB):**	0.101
Kurtosis:	3.728	**Cond. No.**	9.69

Figure 5: ANOVA output

	sum_sq	df	F	PR(>F)
C(region_wb)	14953.186618	6.0	82.889006	7.417316e-51
Residual	5832.937230	194.0	NaN	NaN

Figure 6: Sum of squares (ANOVA)

g. Pattern Evaluation

Correlation matrix plots (heatmap), choropleth maps, a scatter plot and histogram plots were used to show patterns in the datasets and the outcomes from the analysis are discussed in the results section.

IV. Results

The outcome of the analysis is showed below. Data gotten from population growth rate, life expectancy and gross national income were visualized using seaborn, plotly and Matlab, all from the python libraries.

The figure 7 below depicts a choropleth maps of the population growth rates in different countries. The color scale (legend) by the right of map shows the numerical values throughout the geographical region on the map. The scale has values ranging from -4 to 8 (in percentage). The darker colors depict lower values while the light colors depict high values of the population growth rate. For example, the country Latvia with a value of -1.113977% has a lower population growth rate while Oman with 8.088559% has a higher population growth rate.

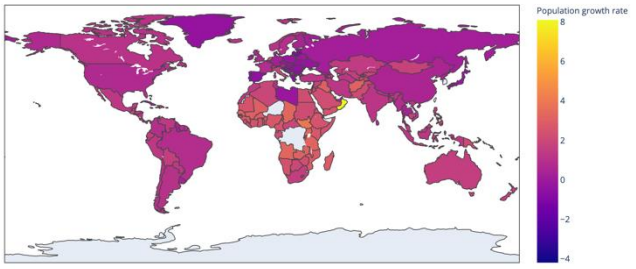

Figure 7: Choropleth map showing the population growth rate

The second figure 8 below shows the choropleth maps of the total life expectancy (male and female) in different countries. The color scale which shows values ranging from 50 – 80 (in years). Using the color scale, Botswana with a value of 47.40561 (in years) has a lower life expectancy for both male and female while a country like Canada with 81.40112 (in years) has a higher life expectancy for both the male and the female.

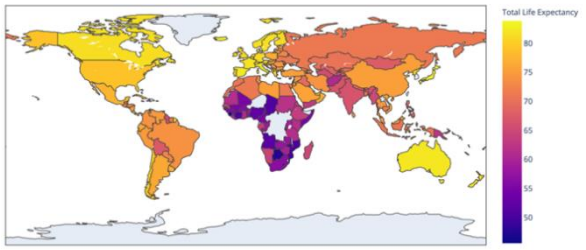

Figure 8: Choropleth map showing the total life expectancy

The figure 9 below shows gross national income of different countries. It is an animated Choropleth map which displays the gross income of countries (in billion dollars). China had a gross national income of 58814120000000.0 in 2013.

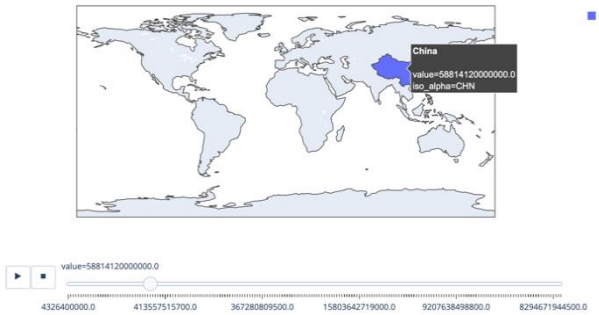

Figure 9: Choropleth map showing the gross national income

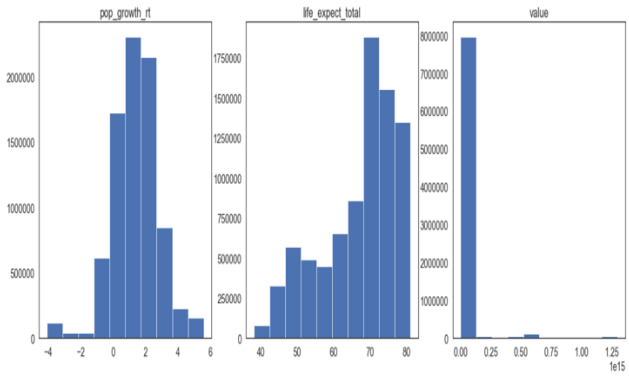

Figure 10: Visual distribution of winsorized data

V. Conclusion and Future Works

This report has explored and visualized population growth rate, life expectancy and gross national income to try to discover the relationship and has attempted to check the statistical significance of using the population growth rate and gross national income in predicting life expectancy and finally find the difference between the regions in the life expectancy.

In conclusion, the analysis carried out amongst population growth rate, life expectancy and gross national income showed that there was no significant relationship using these to determine Life Expectancy, but the ANOVA

model was significant as it should the variability among the regions. For future purpose, with consistent and subsequent years in the data, Time Series models can be fit on the individual datasets to predict future outcomes of the values of population growth, life expectancy and gross national income.

VI. References

[1] E. M. Crimmins, H. Shim, Y. S. Zhang, and J. K. Kim, 'Differences between Men and Women in Mortality and the Health Dimensions of the Morbidity Process', *Clin. Chem.*, vol. 65, no. 1, pp. 135–145, Jan. 2019, doi: 10.1373/clinchem.2018.288332.

[2] Y. Le, J. Ren, J. Shen, T. Li, and C.-F. Zhang, 'The Changing Gender Differences in Life Expectancy in Chinese Cities 2005-2010', *PLoS ONE*, vol. 10, no. 4, pp. 1–11, Apr. 2015, doi: 10.1371/journal.pone.0123320.

[3] B. Diene, 'A Closer Look at the Relationship Between Life Expectancy and Economic Growth', Accessed: Apr. 28, 2020. [Online]. Available: https://core.ac.uk/reader/6941430.

[4] '(PDF) Life Expectancy and Economic Growth: The Role of the Demographic Transition', *ResearchGate*. https://www.researchgate.net/publication/46443061_Life_Expectancy_and_Economic_Growth_The_Role_of_the_Demographic_Transition (accessed Apr. 28, 2020).

[5] A. Ebenstein, M. Fan, M. Greenstone, G. He, P. Yin, and M. Zhou, 'Growth, Pollution, and Life Expectancy: China from 1991–2012', *Am. Econ. Rev.*, vol. 105, no. 5, pp. 226–231, May 2015, doi: 10.1257/aer.p20151094.

[6] Daron Acemoglu and Simon Johnson, 'Disease and Development: The Effect of Life Expectancy on Economic Growth', *J. Polit. Econ.*, vol. 115, no. 6, p. 925, 2007, doi: 10.1086/529000.

[7] 'Data Analysis and Data Mining : An Introduction'. http://eds.a.ebscohost.com/eds/ebookviewer/ebook/bmxlYmtfXzQ1NDM4NV9fQU41?sid=a7c29326-a6d3-4014-ad4b-cc87189b9c1e%40sdc-v-sessmgr03&vid=43&format=EB&rid=1 (accessed Apr. 28, 2020).

[8] 'Recommendations for Teaching the Reasoning of Statistical Inference'. http://rossmanchance.com/papers/topten.html (accessed Apr. 28, 2020).

[9] K. Makar and A. Rubin, 'INFORMAL STATISTICAL INFERENCE REVISITED', p. 6, 2014.

[10] F. Jolliffe and I. Gal, 'Statistics Education Research Journal: Vol. 3, No.1, May 2004', *MSOR Connect.*, vol. 4, no. 3, pp. 58–58, Aug. 2004, doi: 10.11120/msor.2004.04030058.

[11] S. B. Online, '1 INTRODUCTION - Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, 2nd Edition'. https://learning.oreilly.com/library/view/making-sense-of/9781118422106/f_01.xhtml (accessed Apr. 28, 2020).