**LIFE EXPECTANCY PROJECT REPORT**

1. **BUSINESS UNDERSTANDING**

Life expectancy is a statistical measure of the average time a human being is expected to live. It is based on the year of birth, the current age and other demographic factors such as sex, income composition and, mortality rates. The issue is not an emerging one and studies have been done on it that did not take into account a lot of factors that could have an impact on life expectancy. As a result, conclusions that were arrived at were inaccurate. The World Health Organization has improved upon their recent studies on the issue by incorporating not only the population demographic, but also immunization factors, mortality factors, economic factors, and social factors. The organization also provides data that has been collected over a number of years, and this allows for further analysis on the overall trend of life expectancy in a selected period of time. This project will look into those factors that affect the life expectancy of people in different countries in the world, find a comparison and give a recommendation on what measures a country should take in order to improve the life expectancy of its people.

**1.1 BUSINESS OBJECTIVES**

Our main objective is to determine the global life expectancy over the last decade between 2005 and 2015. Our specific objectives include:

1. To determine the difference between life expectancy in developed countries and developing countries.
2. To determine which factors that have the greatest effect on life expectancy in developing countries.
3. To determine which factors that have the greatest effect on life expectancy in developed countries.
4. To determine whether the mortality rate was higher in adults or in kids.
5. To perform a time series analysis in order to understand the overall trend in life expectancy over the years.

Our success criteria will be pegged on our ability to determine the global life expectancy trend over the last decade between 2005 and 2015 in different continents and regions of the world. This will help provide insights to countries that have extremely low life expectancies on areas that need to be focused on to improve their life expectancy.

**1.2 ASSESSING THE SITUATION**

**1.2.1 Resource inventory**

For this study we will need personnel that are capable of performing data cleaning and Exploratory Data Analysis. Also we will require a data scientist that is proficient with data visualizations on Tableau software. We have capable personnel on our team to perform each of the aforementioned tasks so there is no need to outsource any personnel.

**1.2.2 Data sources**

The data for our analysis has been sourced from the World Health Organization. The data is sufficient and there will be no need for us to purchase any external data since all the columns possess the traits we deem important for the study. Also there are no security issues that can prevent us from accessing the data since the data has been made available to the public by WHO.

**1.2.3 Requirements, assumptions and constraints**

Our team comprises 6 members who have been assigned tasks to execute that align with the project. They are all aware of the scheduling and deadlines associated with each of the tasks. The results of our project require to be published on a Github repository and also on the Tableau public website for the management team to view. The assumptions on our data’s quality are that there are no missing values or duplicates of countries in the data. All members of the project team have access to passwords required for data access and there are no legal issues pertaining to the usage of the data since it has been made publicly available by WHO.

**1.2.4 Risks and contingencies**

| **RISKS** | **CONTINGENCY PLAN** |
| --- | --- |
| Data contains missing values and duplicates | We plan on filling the missing values with corresponding figures that are located in separate data sets published by the World Health Organization. We plan on dropping duplicates if necessary. |
| Project taking longer than expected | We plan on scheduling the work on JIRA software and having daily evening syncs to update each other on the progress made on the independent tasks to ensure the project schedule stays on track and is completed in 3 days. |

**1.2.5 Cost-Benefit analysis**

| **DETAILS** | **COSTS** |
| --- | --- |
| Data collection and external data | 0 KSHS was spent on purchasing/collecting the data sourcing for other external data since it is publicly available on WHO’s website. |
| Results deployment | 0 KSHS was spent in deploying the project results on both our Tableau public account and Github repository since both accounts are free to sign up and use |
| Operating costs | Costs for Data bundles used was provided by the management of Moringa School |

| **DETAILS** | **BENEFITS** |
| --- | --- |
| Primary objective has been met | We can tell what variables are primarily responsible for the difference in life expectancy in different continents and regions. |
| Additional insights generated from data exploration | We can predict the life expectancy trend from the data for future periods. We can also begin informed awareness campaigns in several regions against the factors that have a detrimental impact on their life expectancy. |

**1.3 DATA MINING GOALS**

1. To perform both univariate and bivariate analysis on all variables in order to identify their effects on life expectancy
2. To compute the correlation coefficient of different variables and life expectancy
3. To compare the life expectancies of North America, Europe, Asia, Africa
4. To compare and contrast the life expectancies of East Africa to that of West/North/South African regions.
5. To perform dimensionality reduction to determine the optimal number of factors accounting for the majority of variation in life expectancy
6. To train our model in order to be able to make predictions on the life expectancy of future periods

**1.4 PROJECT PLAN**

| **PHASE** | **TIME** | **RESOURCES** | **RISKS** |
| --- | --- | --- | --- |
| **Business understanding** | 3 hours | All 6 data scientists | Data problems and technology problems |
| **Data understanding** | 9 hours | All 6 data scientists | Data problems and technology problems |
| **Data preparation** | 9 hours | 3 data scientists | Data problems and technology problems |
| **Modeling** | 9 hours | 3 data scientists | Technology problems, inability to find adequate model |
| **Evaluation** | 3 hours | All 6 data scientists | Inability to implement results |
| **Deployment** | 3 hours | 2 data scientists | Inability to implement results |

**2. DATA UNDERSTANDING**

**2.1 COLLECTING DATA**

The data that will be used for the analysis has been sourced from the World Health Organization’s website. The data has been made publicly available to visitors of the website. It contains 22 columns all of which are relevant to our life expectancy analysis therefore none will be dropped based on our initial evaluation. Though the data does contain missing values in some of the columns that may affect our analysis and predictions. We have sourced for the respective missing values from separate data sets on WHO's website in order to fill them and negate this problem.

**2.2 DESCRIBING THE DATA**

Upon loading the data on colaboratory notebook we were able to analyze the shape of the data and found that it consists of 2,937 rows and 22 columns. There are 2 categorical columns and 20 numerical columns. Also the data types consist of floats, integers and objects. The columns in dataset are: country, status, infant deaths, percentage expenditure, Measles, under-five deaths, HIV/AIDS, life expectancy, adult mortality, alcohol, Hepatitis B, BMI, Polio, total expenditure, Diphtheria, GDP, population, thinness(1-19 years), thinness(5-9 years), income composition of resources, and schooling. The data includes all the characteristics that are relevant to the business question.

**2.3 EXPLORING THE DATA**

Data exploration is difficult in this phase since the data requires to be prepared first so that it can be meaningfully explored; but from initial exploration the columns relating to: adult mortality, income composition of resources, and schooling are showing promise due to their strong correlations with life expectancy i.e adult mortality(**strong negative**), income composition of resources(**strong positive**), and schooling(**strong positive**). Also the mean life expectancy across the world is 69.224. Further exploration is required in order to detect any errors that may have been made during data entry. We have developed 2 hypothesis statements to aid in our research:

| **NULL HYPOTHESIS** | **ALTERNATIVE HYPOTHESIS** |
| --- | --- |
| * North America and Europe have the same mean/median life expectancy | * NorthAmerica and Europe have the different mean/median life expectancy |
| * East Africa and West/South/North Africa have the same mean/median life expectancy(**African Regions**) | * East Africa and West/South/North Africa have different mean/median life expectancy(**African Regions**) |

Our initial exploration has neither altered our data mining goals nor our hypothesis statements.

**2.4 VERIFYING DATA QUALITY**

Missing data pertains to 14 of our columns i.e.: life expectancy, adult mortality, alcohol, Hepatitis B, BMI, Polio, Total expenditure, Diphtheria, GDP, Population, thinness(1-19 years), thinness(5-9 years), Income composition of resources, and Schooling. Other data errors pertain to the data types of certain columns being wrong and also wrongly named countries. It is important to note that we understand the need for valid and more recent data. The data available to us from 2000-2015 is from the World Health Organization and is the most recent one we could source to perform our analysis.

**3. DATA PREPARATION**

**3.1 SELECTING THE DATA**

Our analysis will be limited to 4 continents i.e. North America, Asia, Europe and African and a further analysis will be done comparing the East African region to the West, North and Southern Africa regions.

**3.2 CLEANING THE DATA**

| **DATA PROBLEM** | **SOLUTION** |
| --- | --- |
| **Missing data** | The blanks have been filled with corresponding values from separate data sets obtained from the World Health Organization’s website |
| **Outliers** | Detected outliers using the interquartile range which will not be dropped since they are essential figures for our analysis |
| **Duplicates** | No duplicates exist in our data |
| **Data errors in column names and country names** | To ensure uniformity, columns have been renamed and converted to lowercase for the study. Also wrongly named countries have been renamed. |
| **Irrelevant columns** | We have dropped the index 0 column that is irrelevant. Also the population and GDP columns have been dropped since they can not be imputed per country |

**3.3 CONSTRUCTING NEW DATA**

We created a copy of the original data set where we created a new column i.e the continent column. It has been populated with data obtained from the countries column in the original data set enabling each country to be matched to its respective continent. There are 6 continents in total i.e Africa, Europe, Asia, North America, South America, and Oceania.

**3.4 INTEGRATING DATA**

We were able to append separate data sets from WHO into our data set in order to fill in the missing values present

**3.5 SAMPLING**

Cluster sampling will be applied by dividing the continents into 4 clusters of interest via which we will conduct our analysis i.e North America, Africa, Asia, and Europe.

**4. DATA ANALYSIS**

There were several steps involved in conducting our analysis which shall be discussed below:

**4.1 UNIVARIATE ANALYSIS**

* Categorical variables were analyzed. The status column was plotted on a pie chart which displayed that the data set contains 83.7% developing countries and 16.3% developed countries.
* Frequency table was plotted displaying that 864 entries are African countries, 704 are Asian, 674 are European, and 338 are from North America. 2 bar graphs have been plotted i.e. one to display these figures and a separate one displaying the percentages of totals of these figures per continent.
* We are also able to perform descriptive statistics on each of the columns displaying their mean, count, standard deviation, minimum/maximum, median, lower quartile range, and upper quartile range. The mean life expectancy is 69.2.
* Also we are able to identify the skewness of each of the columns with 12 of our variables being positively skewed(**adult mortality, infant deaths, alcohol, percentage expenditure, measles, under five deaths, total expenditure, HIV/AIDS, GDP, population thinness 1-19yrs, and population thinness 5-9 yrs**) and 7 being negatively skewed(**year, life expectancy, bmi, polio, diphtheria, income composition of resources, and schooling**)
* We also plotted the samples taken from every year on a bar graph. Also several histograms were plotted such as life expectancy histogram that showed a steady increase in life expectancy up until it peaked at 75 yrs then it dropped, infant deaths histogram which showed a decrease in the frequency with an increase in the value of infant deaths etc.
* An area chart was plotted for thinness 1-19 yrs and thinness 5-9 yrs displaying that values between 0-4 occur more frequently and the frequency decreases as the value for thinness increases.

**4.2 BIVARIATE ANALYSIS**

**4.2.1 Categorical-Categorical**

We generated a stacked bar chart for the 4 clusters i.e. Africa, Europe, Asia, and North America. It displays that Africa is a purely developing continent and that Europe has the highest number of developed countries.

**4.2.2 Numerical-Numerical**

We were able to generate line scatter plots for several columns:

* Life expectancy vs year which shows life expectancy increasing steadily over the years. Upon computing the correlation coefficient between the 2 variables there is a very weak positive correlation between them of 0.174
* Life expectancy vs infant deaths shows no linear relationship between them when plotted. They have a very weak negative correlation of -0.194
* Life expectancy vs under five deaths shows no linear relationship. They have a very weak negative correlation of -0.22
* Life expectancy vs adult mortality shows a linear relationship between them with adult mortality decreasing with increase in life expectancy. They have a strong negative correlation of -0.702
* Life expectancy vs HIV/AIDS shows a negative linear relationship between them when plotted. They have a moderate negative correlation of -0.56
* Life expectancy vs BMI shows a positive linear relationship between them when plotted. They have a moderate positive correlation of 0.59
* Life expectancy vs alcohol shows a positive linear relationship between them when plotted. They have a weak positive correlation of 0.399
* Life expectancy vs thinness 5-9yrs shows a negative linear relationship between them when plotted. They have a negative correlation of -0.468
* Life expectancy vs thinness 1-19yrs shows a negative linear relationship between them when plotted. They have a negative correlation of -0.474
* Life expectancy vs measles shows no linear relationship between them when plotted. They have a weak negative correlation of -0.152
* Life expectancy vs polio immunization shows a positive linear relationship between them when plotted. They have a positive correlation of 0.494
* Life expectancy vs diphtheria immunization shows a positive linear relationship between them when plotted. They have a positive correlation of 0.52
* Life expectancy vs income composition of resources shows a positive linear relationship between them when plotted. They have a negative correlation of 0.732
* Life expectancy vs total expenditure shows a positive linear relationship between them when plotted. They have a weak positive correlation of 0.22
* Life expectancy vs percentage expenditure shows a positive linear relationship when the life expectancy is between 75-90 yrs. There is a weak positive correlation of 0.38 between them
* Life expectancy vs schooling shows a positive linear relationship between them when plotted. They have a strong positive correlation of 0.748

**4.2.3 Numerical-Categorical**

We were able to generate bar plots for several columns:

* Life expectancy vs status which shows developed countries have the higher life expectancy compared to developing countries
* Life expectancy vs continent which shows Europe has the highest life expectancy followed by Europe, North America, and lastly Africa
* Status vs infant deaths which shows developing countries have the highest number of infant deaths
* Status vs under five deaths which shows developing countries have the highest number of under five deaths
* Status vs adult mortality which shows developing countries have the highest adult mortality
* Status vs HIV/AIDS which shows developing countries have the highest deaths per 1000 live births attributable to HIV/AIDS
* Status vs alcohol shows developed countries have the highest alcohol consumed per capita
* Status vs BMI shows developed countries have the highest average body mass index for the entire population
* Status v thinness 5-9 yrs shows developing countries have the highest prevalence of thinness among children aged 5-9yrs
* Status v thinness 10-19 yrs shows developing countries have the highest prevalence of thinness among children aged 10-19yrs
* Status vs measles shows that developing countries have the highest immunization rate for measles
* Status vs polio shows that developed countries have the highest immunization rate for polio
* Status vs diphtheria shows that developed countries have the highest immunization rate for diphtheria
* Status vs income composition of resources shows developed countries have the highest income composition of resources
* Status vs total expenditure shows that developed countries have a higher total expenditure on health
* Status vs percentage expenditure shows that developed countries have a higher percentage expenditure on health
* Status vs schooling shows that developed countries have a higher amount of time spent by individuals in school

**4.3 HYPOTHESIS TESTING**

We set our significance level to 5 % in order to ensure a fair trade off between type 1 and type 2 errors since the probability of type 2 errors decreases with an increase in significance level.

| **NULL HYPOTHESIS** | **ALTERNATIVE HYPOTHESIS** |
| --- | --- |
| * North America and Europe have the same mean/median life expectancy | * NorthAmerica and Europe have the different mean/median life expectancy |
| * East Africa and West/South/North Africa have the same mean/median life expectancy(**African Regions**) | * East Africa and West/South/North Africa have different mean/median life expectancy(**African Regions**) |

Upon performing a parametric test we are able to ascertain that a Z-test is to be performed since the sample size is greater than 30, the data points are independent of each other, and the data is normally distributed.

* **Europe and North America**

Upon performing the Z-test we get a p-value of 0 which is less than 0.05 suggesting that we have sufficient evidence to reject our first null hypothesis. We further performed 2 non-parametric tests i.e the Mann Whitney U test and the Mood-Median test. Both of these confirmed the results of our Z-test by obtaining a p-value of less than 0.05 hence confirming that we should indeed reject our null hypothesis.

* **East Africa and West Africa**

The z test produces a p value of 0.00034 which is less than our significance level. The Mann Whitney U test and the Mood-Median test give us p-values of 0.000255 and 0.000472 respectively thus confirming we should reject the null hypothesis that East Africa and West Africa have the same mean/median life expectancy.

* **East Africa and North Africa**

The z test produces a p value of 0.00685 which is less than our significance level. The Mann Whitney U test and the Mood-Median test give us p-values of 0 and 0 respectively thus confirming we should reject the null hypothesis that East Africa and North Africa have the same mean/median life expectancy.

* **East Africa and Central Africa**

The z test produces a p value of 0.023 which is less than our significance level. The Mann Whitney U test and the Mood-Median test give us p-values of 0 and 0 respectively thus confirming we should reject the null hypothesis that East Africa and Central Africa have the same mean/median life expectancy.

* **East Africa and Southern Africa**

The z test produces a p value of 0.000127 which is less than our significance level. The Mann Whitney U test and the Mood-Median test give us p-values of 0 and 0 respectively thus confirming we should reject the null hypothesis that East Africa and Southern Africa have the same mean/median life expectancy.

**4.4 MULTIVARIATE ANALYSIS**

We created a duplicate dataset of our original data set and performed label encoding for our 3 categorical columns i.e. status, continent, and country. We then previewed the duplicate to confirm that the changes had been implemented.

We then began our analysis by splitting our data into the training and testing set. We also normalized our data set. Principal Component Analysis was also done for our data. Upon using all our components versus one of our components we were able to conclude that one principal component accounts for most of the variation.

We also used a random forest regressor and plotted a feature importance graph which displayed that 3 of our columns have the highest weighted importance i.e. HIV/AIDS, income composition of resources, and adult mortality.

**5. RECOMMENDATIONS**

From our analysis HIV/AIDS, income composition and adult mortality have the greatest negative effect on life expectancy more so in developing countries. Hence awareness campaigns on HIV/AIDS need to be increased in order to sensitize the population on ways of living better lives with the disease. Also contraceptives can continue to be made free to the public and sensitize them to the need for abstinence in developing nations.

The governments of developing countries should also intervene by providing opportunities to the unemployed in order to increase their income composition since we can find that life expectancy increases with increase in income composition.