MIDDLE EAST TECHNICAL UNIVERSITY

Factors Affecting Life Expectancy

Başak Uğurlu
Damla Başarmış
Omar Alagha
Zeynep Fenercioğlu

STAT112
Ceylan Yozgatlıgil

20/01/2023

Abstract

This report presents our life expectancy data cleaning, analyzing, visualizing, and interpreting process. The dataset related to life expectancy and health factors for 105 countries from three continents has been collected from the WHO data repository website. To examine the variables' relationships and distributions clearly, we must clean this data before starting to visualize since there were many errors in reaching accurate information. We examined the relationship between life expectancy and factors: Mortality Rate, Alcohol Consumption Rate, Developed or Developing Status, Average Body Mass Index, GDP (USD), and Total Expenditure depending on the Continent. Also, we examined the relationship between alcohol consumption with BMI and Status. We created six research questions. To visualize these graphs, we imported pandas, NumPy, matplotlib, and seaborn and wrote codes convenient to our graph. Then we changed their colors, writing font size, and font weight. Through this data cleaning, visualizing, and interpreting process, we have revealed that the country's GDP is the most important among all other factors on life expectancy. As GDP increases life expectancy also increases radically. We have observed that life expectancy is lower in Asia and America than in Europe. Life expectancy does not depend on alcohol consumption and BMI. As general government health expenditure increases, life expectancy decreases slightly in Europe, while it is the opposite for America and Asia, it slightly increases. We have revealed a positive relationship between mortality and life expectancy.

Keywords: Life expectancy, mortality, alcohol consumption, development status, body mass index, GDP, NumPy, Seaborn, Matplotlib, relationship, visualizing, data cleaning.

Introduction

This research has many steps to get information about life expectancy. First of all, the research started with the cleaning phase because the data (which are taken from WHO) had many errors so it prevents to reach accurate information. It is fixed by applying the methods in the "Checklist for The Data Cleaning and Tidying" through Google Collab. In this way, clean data is created and become prepared to represent correctly. This research's purpose is to give the relationship between Life Expectancy and factors: Mortality Rate, Alcohol Consumption Rate, Developed or Developing Status, Average Body Mass Index, GDP (USD), and Total Expenditure depending on the Continent. Also, informing about their distribution separately and the relationships of the variables with each other is important for this research. Thus, it shows their positive or negative effects on each other. Moreover, It gives information about central tendency (mode, median, mean), measures of position (outliers, percentiles, interquartile range), and measures of variability (range, standard deviation, variance). Also, this research finds an answer to 5 research questions. For instance, the Relationship between Life Expectancy and Mortality depending on Continent, the Relationship between Life Expectancy and GDP depending on Continent, the Relationship between Life Expectancy and Total Expenditure depending on Continent, Average Alcohol by Status, the Relationship between BMI and Alcohol Consumption for Developed Countries, distribution of Life Expectancy by Development Status and Continent. They are illustrated with many types of graphs.

Data Tidying and Cleaning Steps

1.We imported Pandas, NumPy, Matplotlib and Seaborn. Then we imported our ‘Life Expectancy’ data. We examined the non-null values’ count, variables, and their data types.

2.We examined the first and last five data. We made sure that we imported our data correctly. We didn’t have any separation argument problem. We had NAs. There were 5 NAs in X, 11 in life expectancy, 8 in status, 12 in the continent, 13 in mortality, 9 in alcohol, 11 in BMI, 15 in total expenditure, 12 in GDP, and 8 in the unit.

3.Our data’s column headers are variable names, not values. Multiple variables are not stored in one column. Variables are not stored in both rows and columns. Multiple types of observational units are not stored in the same table. A single observational unit is not stored in multiple tables. So, we didn’t have any problems in this step.

4.Firstly, we changed our column names into title form. Then we changed the ‘Life.Exp’ column name into ‘Life Expectancy’, ‘Mort’ into ‘Mortality’, ‘Total.Exp’ into ‘Total Expenditure’, ‘Bmi’ into ‘BMI’, ‘Gdp’ into ‘GDP’. Then we checked new column names.

5.The ‘X’ column was unnecessary, so we dropped it.

6.We didn’t have any duplicates.

7.We didn’t have unnecessary strings in the values.

8.We removed unnecessary white spaces in the ‘Status’, ‘Continent’, and ‘Unit’ columns.

9.In the ‘Status’ column, we had ‘DEVELOPed’ and ‘DEVELOPing’. We changed them into ‘Developed’ and ‘Developing’. In ‘Unit’ we had ‘usd’, ‘TL’, and ‘uSd’. We changed them into ‘USD’.

10.We checked the uniqueness of ‘Status’, ‘Continent’, and ‘Unit’. All of them were correct after the 9th step.

11.We do not have any date values.

12.We checked again if the data types were correct. ‘Life Expectancy’, ‘Mortality’, ‘Alcohol’, ‘BMI’, ‘Total Expenditure’, and ‘GDP’ are floats while ‘Status’, ‘Continent’, and ‘Unit’ are objects.

13.We examined that there were null values and that we should fill with whether mean, median, or mode. We examined that all life expectancy data are unexpected values; they should have been close to 70-75. Also, the maximum life expectancy value is unrelatedly high. The mean and median of life expectancy is quite far from each other; we will eliminate the outliers. Mortality does not have unexpected values; we will only eliminate null values. BMI has unexpected values such as -150. We should eliminate them. The maximum GDP value is relatively high. We filled null values with mean for ‘Alcohol’, ‘BMI’, ‘Total Expenditure’, and ‘GDP’. We filled null values with mode for ‘Unit’, ‘Status’, and ‘Continent’. We filled null values with the median for ‘Life Expectancy’ and ‘Mortality’.

14.'GDP' has one outlier, 48588.73. We changed it with mean. 'BMI' has two outliers, -150 and 78.56. We changed it with the mean. Alcohol does not have any outliers.

15.Only 'Unit' data had this problem. We changed all values into 'USD' in the 9th step.

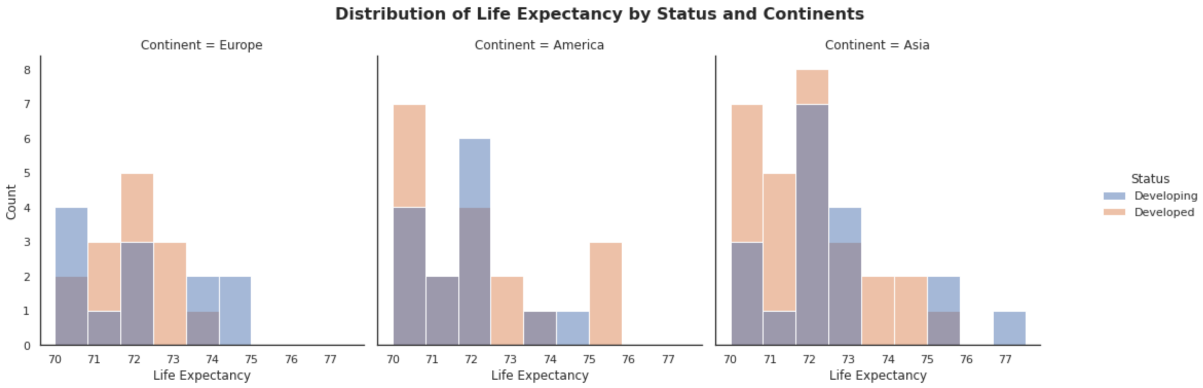
16.'Life Expectancy' has 10.5% missing values, 'Status' has 7.6%, 'Continent' has 11.4%, 'Mortality' has 12.4%, 'Alcohol' has 8.6%, and 'BMI' has 10.5%. 'Total Expenditure' has 14.3%, 'GDP' has 11.4%, and 'Unit' has 7.6%. Their percentage is below 60%. It means that we can fill them with mean/mode/median. We filled null values in the 13th step.

https://github.com/damlabasarmis/Life-Expectancy

'

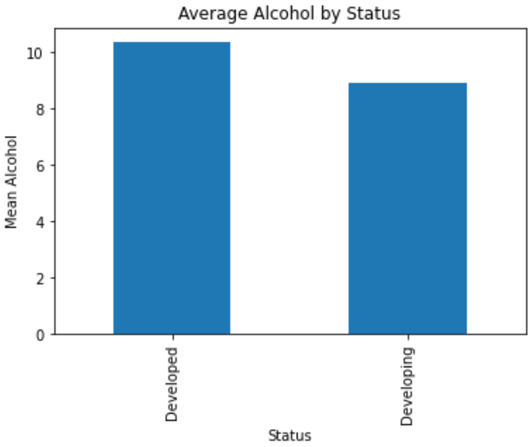
Exploratory Data Analysis

Research Question 1: How does life expectancy change by development status and continents Asia, America, and Europe?



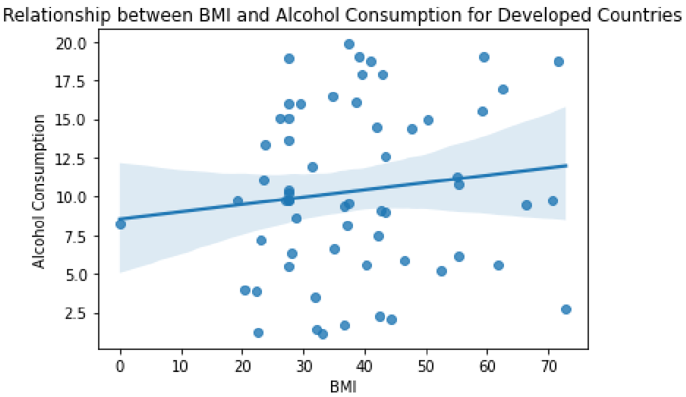
We see that the majority of our data is from Asia. In Asia, most frequent life expectancy is 72 for both developing and developed countries. We can say that the graph is right skewed for Asia. It is also right skewed for America continent. Most frequent life expectancy changes by the status. It is 72 for developing countries and 70 for developed countries. Europe life expectancy distribution changes by status. It is almost bell shaped for developed countries with the mode of 72, and right skewed for developing countries with the mode of 70. We can say that developed countries of Asia and America live approximately 70-71 years on average while developing countries of Asia and America live approximately 72 which is higher than developed countries. On the contrary in Europe, people in developed live approximately 72 years and people in developing countries live approximately 70 years which is less than developed countries.

Research Question 2: What's the relationship between average alcohol consumption and development status?



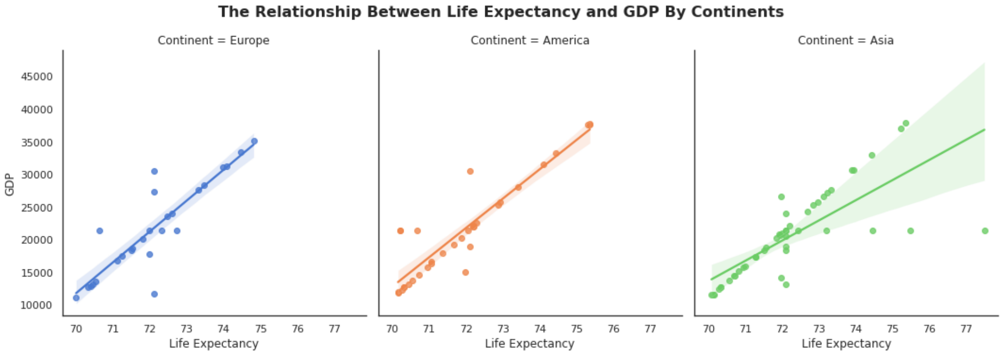
Based on the data provided, it can be observed that the mean level of alcohol consumption in developed countries is higher than the mean level of alcohol consumption in developing countries.

Research Question 3: Is there a correlation between BMI and alcohol consumption for developed countries?



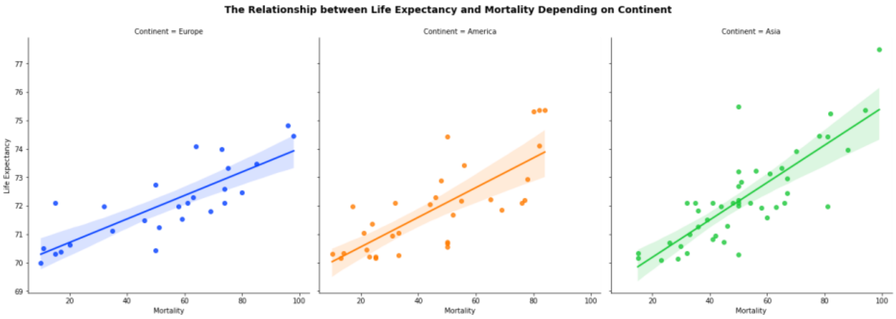
There is a weak positive correlation between alcohol consumption and BMI as alcohol consumption increases, BMI also increases. Which indicates that individuals who consume more alcohol Tend to have a higher body mass index (BMI).

Research Question 4: How does life expectancy get affected by GDP for Asia, America, and Europe?



We see that there is an almost positive linear correlation between GDP(USD) and Life Expectancy for all the continents. It means that as the country's GDP increases, the people's average lifespan also increases almost proportionally. Correlation is weaker in Asia.

Research Question 5: What is the relationship between life expectancy and mortality by continents Asia, Europe, and America?


The Relationship between Life Expectancy and Mortality Depending on Continent

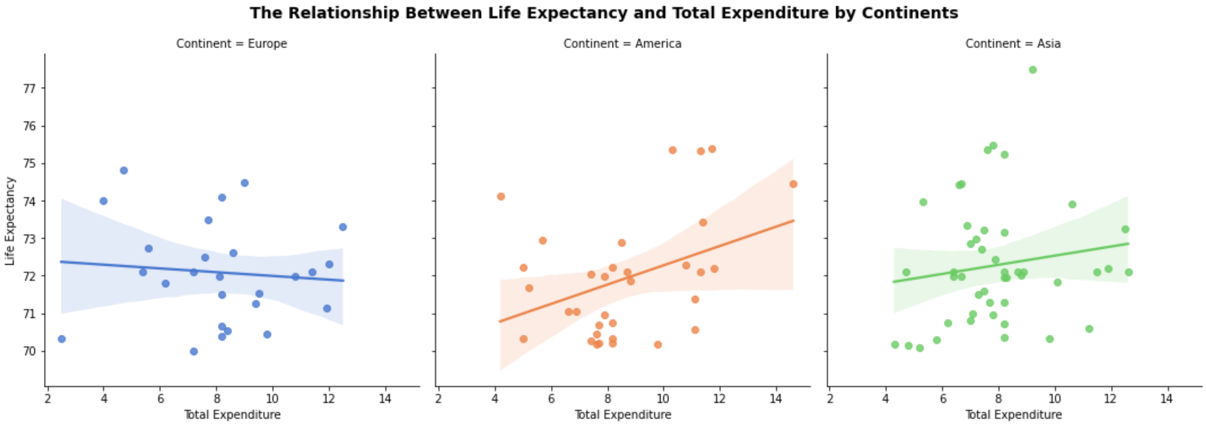
Continent = Europe

Continent = America

Continent = Asia

These three scatter plots show a strong positive correlation between mortality and life expectancy for all the continents. This means that the mortality rate increases if life expectancy increases. Also, if we compare the three scatter plots, Asia has a stronger correlation than others. However, Europe has a weaker correlation than others.

Research Question 6: Is there a relationship between life expectancy and total expenditure?


The Relationship Between Life Expectancy and Total Expenditure by Continents

Continent = Europe

Continent = America

Continent = Asia

There is a strong positive correlation between life expectancy and total expenditure in America. As the general government expenditure on health increases, life expectancy increases. There is a weak positive correlation between life expectancy and total expenditure in Asia. There is a weak negative correlation between life expectancy and total expenditure in Europe. We can say that it differs a lot between these continents.

Conclusion

As shown in the data provided, life expectancy depends on many variables, and it does change between continents. There is a right-skewed life expectancy distribution for Asia and America, but it is bimodal distributed for Europe. We can say that life expectancy is lower in Asia and America, than in Europe. While country development affects Asian and American people’ s life expectancy negatively, it affects European people’s lifespan positively. However, we could not see a pattern between life expectancy and consumption through variables BMI (body mass index) and alcohol consumption. Nevertheless, alcohol consumption is higher in developed countries and BMI increases as alcohol consumption rate increases in those countries. Life expectancy has the strongest correlation with GDP. Countries with higher GDP have higher life expectancy While the correlation is almost same for Europe and America, it is weaker in Asia. We have observed that as mortality increases, life expectancy also increases for all the continents. Correlation between life expectancy and general government expenditure on health changes a lot between continents. We see that as total expenditure increases, life expectancy decreases in Europe when it is the opposite for two other continents. Hence, through this data cleaning, visualizing and interpreting process, we have revealed that GDP of the country is the most important among all other factors on life expectancy.