Project Report

GitHub URL

<https://github.com/vinwalsh98/-UCDPA_vincentwalsh-.git>

Abstract

The aim of this project is to gain insights in to life expectancy data by using the techniques that were thought throughout this course. The datasets used were both large and relatively small, detailed and simple. The large sets were compiled of worldwide life expectancy data from hundreds of countries, as well as data from variable factors that contribute to an individual country’s life expectancy such as GDP per capita, disease, expenditure percentage, BMI and many others. The smaller set was compiled of life expectancy data of European countries based on gender. I manipulated these dataframes by using techniques such as merging, dropping duplicates and replacing missing values in order to improve the quality of the existing data and draw observations.

Introduction

Life expectancy data has always intrigued me. I wanted to see how the life expectancy of the world’s developing and developed countries have compared to each other since the beginning of the new century. By using datasets both large and relatively small, I was able to gain an extensive list of information to work with based on gender, wealth and multiple other variables. In my project, I paid particular attention to two variables in developed nations; GDP per capita and gender, and how they can effect a nation’s life expectancy.

Dataset

The first file used, ‘led 2.csv’, is a large dataset compiled of life expectancy data from hundreds of countries from the years 2000 to 2015. As well as life expectancy, it contains demographic variables that are contributing factors to a country’s life expectancy such as GDP, infant deaths, alcohol, Body Mass Index and diseases such as Hepatitis B and HIV. I chose this dataset as it gives us an extensive breakdown of many factors that influence the life expectancy of a country. As it is such a large dataset, techniques such as sorting, indexing, dropping duplicates and grouping were very useful in customising the layout and making the dataset easier and more informative to read. This dataset was found on Kaggle: <https://www.kaggle.com/augustus0498/life-expectancy-who>.

I could trust this data source to be reliable and accurate as the creator sourced his information from WHO and United Nations data.

The datasets that I used to implement merging were life expectancy statistics from hundreds of countries. “Data\_2010” contained life expectancy statistics from 2010 and “Data\_2015” contained life expectancy statistics from 2015. This dataset was retrieved from the official World Bank website:

<https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2018&start=2018>

I chose these to merge these datasets as given that they had the same column names, I wanted to compare and contrast their data side by side.

The final dataset I chose, “life\_expect\_eu.pd”, gave the life expectancy at birth of 33 EU countries.

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)>.

It was a smaller dataset than the others, containing just four columns; Country, Total, Males and Females. It was a useful dataset to identify the relationship between gender and life expectancy in Europe.

I was able to visualise how this data could be represented visually using MatPlotLib and Seaborn. These are official WHO statistics found on their website

Implementation Process

My first objective was to ensure that I had gathered good quality datasets from reliable sources. Hence, I used Official WHO statistics, United Nations statistics, and world bank statistics. These datasets were uploaded into a Pandas DataFrame. CSV, Pandas and Numpy were imported with their shortened names. This allowed me to read and manipulate the CSV dataframes. Following this, I inputted the instruction “print” to read the main dataset “led 2”.

Next, I installed the request package into Python. This was to get the data from the Kaggle API’s. A code of 200 ran when I printed this, proving that it ran successfully.

I wanted to adjust the width and size parameters of “led 2”. I inserted code to expand the amount of columns shown. I then indexed year by country. I then sorted the column life expectancy by years ascending. When I printed this, I could see that Haiti had the lowest life expectancy at 36.3 years in 2010 and Spain had the highest at 89.0 years in 2007. It was an excellent way of quickly identifying which countries had the lowest and highest life expectancy values.

Following on from this, I used the drop duplicates function to stop duplicates of numbers in the column “year”. When this was printed, the data for just Afghanistan was given from 2000 to 2015 as it was the first country on the dataset.

I used the iloc function to take a snippet of the dataset (rows one and rows six and columns zero to five). I printed this which gave me an output of the first five headings and statistics for Sierra Leone in the years 200 and 2005.

I looked to identify where the missing information was in the dataset. I printed the null function to locate where the missing information was. I wanted to replace “NaN” with “information currently unavailable”. When I printed the new variable “clean\_data”, it showed that no missing values were left.

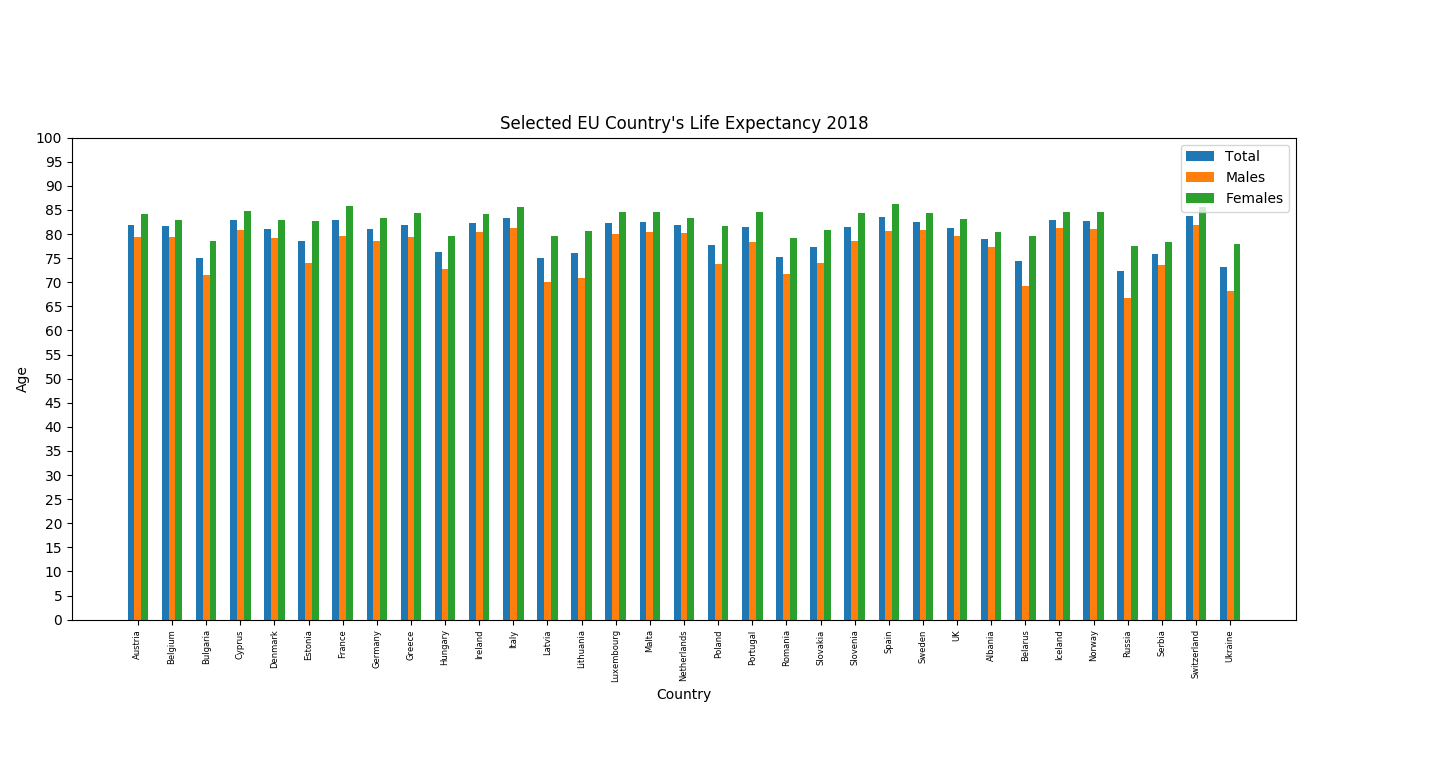
I performed the loop function. When printed, this gave me a breakdown the column names. Then I performed the iterrows function, which have me a breakdown of the first four rows and their corresponding column headings.

Two new csv files containing WHO life expectancy data from 2010 and 2015 were then inputted. I merged these datasets and printed the merged dataset.

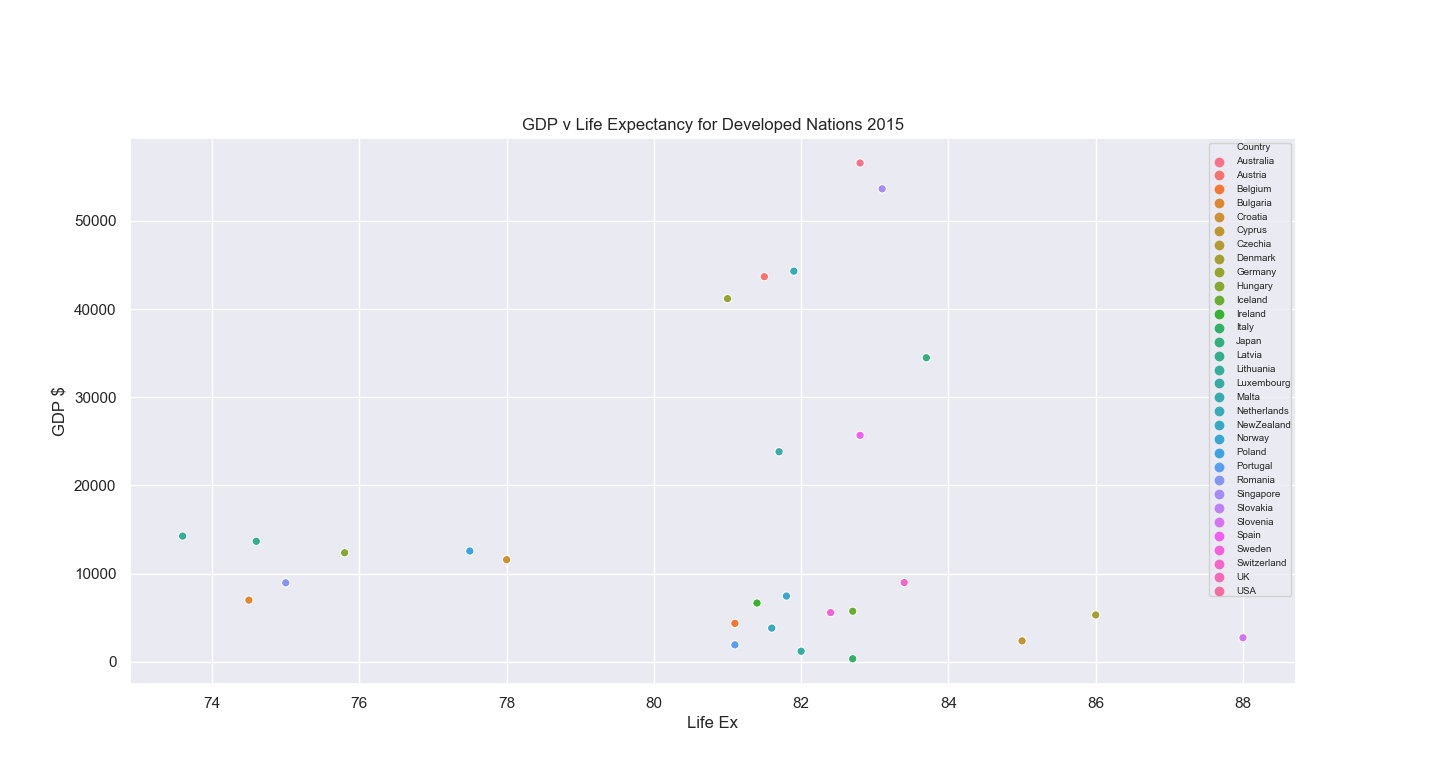
I used the dictionary function to display each column heading followed by its corresponding value in the first row in an array.

I calculated the mean function to show the mean age of each sub region from the file Data 2010. The information given was very interesting, showing that Africa has the lowest mean age at 38.29 and Europe has the oldest mean age at 48.48.

The CSV file “life\_expect\_eu” was imported and renamed as “eu’. This dataset was printed in order to run its contents through the programme. I wanted to use this dataset to generate a visualisation of a selected number of European countries based on gender. I used the dataset “led 2” to create a scatter plot displaying the relationship between GDP and life expectancy for developing nations. Using Seaborn and MatPlotLib, I created two visualisations of this dataset. Chart titles, length, width, axis titles and line colours were inserted to create a clear and precise visual.

Results

The above chart “Selected EU Country’s Life Expectancy 2018” demonstrates a selection of European Country’s Life Expectancy in the year 2018. The X-Axis shows the Average Age Life expectancy and the Y-Axis shows the country. The total, male and female life expectancies are broken down individually in colours blue, orange and green. Each can be identified using the legend in the top right corner of the chart. It is clear to see that each European nation has a life expectancy of between 70 and 90 years. Mediterranean countries such as Spain and Italy have high life expectancies. Switzerland has the highest life expectancy for women and men, at 85.7 and 83.8 respectively. Ukraine has the lowest life expectancy for women at 78.0 and also for men at 73.2.



The above scatter plot “GDP v Life Expectancy for Developed Nations 2015” demonstrates the correlation between Gross Domestic Product versus Life Expectancy of 33 countries. The X-Axis shows the Gross Domestic Product per year per capita. The y-axis gives the life expectancy. The different countries are represented by colour in the legend in the top right corner of the graph.

Insights

**1.** As can be seen “Selected EU Country’s Life Expectancy 2018”, females have a longer life expectancy than males in every European country.

**2.** From looking at “Selected EU Country’s Life Expectancy 2018”, Countries in the Baltic region such as Latvia and Lithuania have the lowest life expectancies in Europe.

**3.** From looking at “GDP v Life Expectancy for Developed Nations 2015”, it is clear that there is not a strong correlation between life expectancy and GDP per capita in developed nations. The three nations with the highest life expectancy in my sample (Slovenia, Denmark, Cyprus) all had a GDP per capita of less than $10000.

**4.** As seen in “GDP v Life Expectancy for Developed Nations 2015”, when we look at Australia and Singapore, the only two countries on our list of developed countries with a GDP per capita of over $50,000. This does not necessarily correlate to higher life expectancies. Perhaps this is something that nations may take in to consideration in the future (e.g. is a thriving economy really contributing to a nation’s wellbeing?).

**5.** As can be seen in “GDP v Life Expectancy for Developed Nations 2015”, the three countries with the lowest life expectancy (Latvia, Bulgaria, Lithuania) also have a low GDP per capita of under $15000. Again, these countries are located along the Baltic region. As these are all EU member states, perhaps assistance would be considered in the future to boost their economy which may in-turn boost their standards of living and life expectancy.

References

<https://www.kaggle.com/augustus0498/life-expectancy-who>

<https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2018&start=2018>

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)>.