**Brief Description of Data Used**

The data used for Stage 2 of the project is a modified version of the data in Stage 1, to fit more to the requirements. The ‘location’ (countries) and ‘life expectancy at birth’ were taken from Stage 1 data, and ‘life expectancy’ (our main subject of study) was taken from the group’s collective data. ‘Income groups’ were added in creating the new dataset.

The classification of countries based on income was taken from a PDF from the UN website. It was found in the form of a table, and the lists of countries were manually moved to a spreadsheet to be processed by Python in integrating it with the other parts.

Schema of the final dataset: location (country), life expectancy, life expectancy at birth (years), 'Medical doctors (per 10,000)', income group.

**Grouped-Aggregate Summary 1: Life Expectancy Based on Income Groups**

code

|  |
| --- |
| import csv  firstline = True  incomedic = {}  incomedic["High Income"] = {}  incomedic["Upper Middle Income"] = {}  incomedic["Lower Middle Income"] = {}  incomedic["Low Income"] = {}  with open("C:\\Users\\Kylie\\Desktop\\DATA1002\\stage 2 data summary.csv", "w", newline="", encoding="utf8") as outputfile:  writer = csv.writer(outputfile)  with open("C:\\Users\\Kylie\\Desktop\\DATA1002\\stage 2 data.csv") as file:  for line in file:  if firstline:  firstline = False  else:  elements = line.split(",")  country = elements[0]  lifeexp = elements[1]  lifeexpbirth = elements[2]  meddoc = elements[3]  income = elements[4].strip(" \n")  if (income == "High Income") and (len(lifeexp)>0):  tempkey = "High Income"  incomedic[tempkey][country] = lifeexp  if (income == "Upper Middle Income") and (len(lifeexp)>0):  tempkey = "Upper Middle Income"  incomedic[tempkey][country] = lifeexp  if (income == "Lower Middle Income") and (len(lifeexp)>0):  tempkey = "Lower Middle Income"  incomedic[tempkey][country] = lifeexp  if (income == "Low Income") and (len(lifeexp)>0):  tempkey = "Low Income"  incomedic[tempkey][country] = lifeexp  writer.writerow("LIFE EXPECTANCY BASED ON INCOME GROUPS".split(","))  writer.writerow("Income Group,Mean,Median,Range,Count,Maximum,Minimum".split(","))  for group in incomedic: #per income group  counter = 0  summed = 0  for key in incomedic[group]:  counter += 1  summed += float(incomedic[group][key])  avg = round((summed/counter),3)  sortedincomegroup = dict(sorted(incomedic[group].items(), key=lambda item: item[1]))  if (counter%2) != 0:  mid = int((counter+1)/2)-1  median = sortedincomegroup[list(sortedincomegroup.keys())[mid]]  else:  mid1 = int(counter/2)-1  mid2 = int((counter/2)+1)-1  median = (float(sortedincomegroup[list(sortedincomegroup.keys())[mid1]])+float(sortedincomegroup[list(sortedincomegroup.keys())[(mid2)]]))/2  median = round(float(median),3)  maxim = float(sortedincomegroup[list(sortedincomegroup.keys())[-1]])  minim = float(sortedincomegroup[list(sortedincomegroup.keys())[0]])  ranges = maxim - minim  maxim = round(maxim,3)  minim = round(minim,3)  ranges = round(ranges,3)  writer.writerow((group+","+str(avg)+","+str(median)+","+str(ranges)+","+str(counter)+","+str(maxim)+" ("+list(sortedincomegroup.keys())[-1]+")"+","+str(minim)+" ("+list(sortedincomegroup.keys())[0]+")").split(",")) |

Output

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Income Group | Mean | Median | Range | Count | Maximum | Minimum |
| High Income | 79.451 | 81.5 | 28.8 | 37 | 88.0 (Slovenia) | 58.2 (Equatoria) |
| Upper Middle Income | 73.126 | 75.0 | 27.2 | 35 | 79.6 (Costa Rica) | 52.4 (Angola) |
| Lower Middle Income | 67.45 | 67.9 | 21.2 | 30 | 74.9 (Sri Lanka) | 53.7 (Lesotho) |
| Low Income | 61.825 | 61.85 | 20.8 | 24 | 71.8 (Bangladesh) | 51.0 (Sierra Leone) |

**Grouped-Aggregate Summary 2: Life Expectancy at Birth (Based on Medical Doctor Bins)**

code

|  |
| --- |
| import csv  firstline = True  meddocdic = {}  meddocdict = {}  bylifeexpbirth = {}  with open("C:\\Users\\Kylie\\Desktop\\DATA1002\\stage 2 data summary.csv", "w", newline="", encoding="utf8") as outputfile:  writer = csv.writer(outputfile)  with open("C:\\Users\\Kylie\\Desktop\\DATA1002\\stage 2 data.csv") as file:  for line in file:  if firstline:  firstline = False  else:  elements = line.split(",") #list!  country = elements[0]  lifeexp = elements[1]  lifeexpbirth = elements[2]  meddoc = elements[3]  income = elements[4].strip(" \n")    if (len(meddoc) >0) and (len(lifeexpbirth)>0):  meddocdict[country] = float(meddoc)  meddocbin = int(float(meddoc)//10)  temp = country + "/" + meddoc + "/" + lifeexpbirth  if meddocbin in meddocdic:  meddocdic[meddocbin] += (","+ temp)  else:  meddocdic[meddocbin] = temp  writer.writerow("LIFE EXPECTANCY AT BIRTH (BASED ON MEDICAL DOCTOR BINS)".split(","))  writer.writerow("Bin,Medical Doctors (per 10000),Mean LE,Median LE,Range LE,Frequency,Maximum LE, Minimum LE".split(","))  for key in sorted(meddocdic.keys()): #per bin  counter = 0  summed = 0  meddocperbin = meddocdic[key].split(",")  bylifeexpbirth = {}    for x in meddocperbin:  values = x.split("/")  country = values[0]  meddoc = values[1]  lifeexpbirth = values[2]  bylifeexpbirth[country] = float(lifeexpbirth)  counter += 1  summed += float(lifeexpbirth)  avg = round((summed/counter),3)  sortedlifeexpbirth = dict(sorted(bylifeexpbirth.items(), key=lambda item: item[1]))  if (counter%2) != 0:  mid = int((int(counter)+1)/2)  medianlifeexp = bylifeexpbirth[list(sortedlifeexpbirth.keys())[mid-1]]  else:  mid1 = int(counter/2)  mid2 = int((counter/2)+1)  medianlifeexp = (float(sortedlifeexpbirth[list(sortedlifeexpbirth.keys())[mid1-1]])+float(sortedlifeexpbirth[list(sortedlifeexpbirth.keys())[mid2-1]]))/2  median = round(medianlifeexp,3)  maximlifeexp = float(bylifeexpbirth[list(sortedlifeexpbirth.keys())[-1]])  minimlifeexp = float(bylifeexpbirth[list(sortedlifeexpbirth.keys())[0]])  ranges = maximlifeexp - minimlifeexp  maxim = round(maximlifeexp,3)  minim = round(minimlifeexp,3)  ranges = round(ranges,3)  writer.writerow((str(key+1)+","+str(int(key)\*10)+"-"+str((int(key)+1)\*10)+","+str(avg)+","+str(median)+","+str(ranges)+","+str(counter)+","+str(maxim)+" ("+list(sortedincomegroup.keys())[-1]+")"+","+str(minim)+" ("+list(sortedincomegroup.keys())[0]+")").split(",")) |

Output

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Bin | Medical Doctors (per 10000) | Mean LE | Median LE | Range LE | Frequency | Maximum LE | Minimum LE |
| 1 | 0-10 | 63.808 | 63.04 | 27.248 | 81 | 75.653 (Thailand) | 48.405 (Lesotho) |
| 2 | 10-20 | 75.24 | 75.29 | 11.612 | 31 | 81.543 (Singapore) | 69.93 (Philippines) |
| 3 | 20-30 | 75.595 | 75.936 | 17.605 | 34 | 82.92 (Japan) | 65.315 (Turkmenistan) |
| 4 | 30-40 | 76.788 | 77.525 | 13.127 | 25 | 81.993 (Switzerland) | 68.865 (Ukraine) |
| 5 | 40-50 | 78.407 | 80.118 | 9.825 | 4 | 81.61 (Italy) | 71.785 (Georgia) |
| 7 | 60-70 | 77.52 | 77.52 | 0.0 | 1 | 77.52 (Cuba) | 77.52 (Cuba) |

**Chart 1: Health Statistics Based on Income Groups**

In making this chart, the columns were first rearranged: Location, Income Group, Life Expectancy, Medical Doctors, and Life Expectancy at Birth. The data was filtered per Income Group, then copy and pasted into another sheet. In the second sheet, a table was manually made to summarize the data per income group, using formulas =counta and =average. =round was then used to keep the final numbers to two decimal points. Then, the data was just made into a bar chart.

This chart shows the relationship between income groups and life expectancy, medical doctors, and life expectancy at birth. As the income of a country increases, the other elements also increase. This chart does not really show which element causes changes to which element. However, it can be concluded that the increase income is what causes the other elements to increase. With better incomes, the people in the country can afford better livelihood and therefore better life expectancies. Healthcare can also be improved in better economies.

**Chart 2: Impact of Medical Doctors on Life Expectancy at Birth**

The data for this chart was taken from the first sorted Excel sheet. A scatter plot diagram with a trendline was made based on the columns Medical Doctors and Life Expectancy at Birth.

It can be concluded that the number of medical doctors has a direct relationship with the life expectancy at birth. As there is more medical doctors, they have better healthcare. This can cause an increase in life expectancy at birth. However, the trend of points is more like a log function than a linear function like the trendline. The increase in life expectancy at birth slows down as the number of medical doctors increases further. It can be inferred that as the number of medical doctors get higher, the impact that it has on life expectancy at birth decreases.

**Chart 3: Life Expectancy at Birth by Life Expectancy**

This chart was also made from the first sorted Excel sheet. A scatter plot was made based on Life Expectancy at Birth and Life Expectancy.

This chart shows a strong direct relationship between life expectancy and life expectancy at birth. However, it cannot be concluded that this shows a causal relationship. This chart was mainly made to show that they were not the same thing (if they were, there would be a perfect linear relationship and the y-intercept would be at 0).

The data used for Stage 2 of the project is a modified version of the data in Stage 1, to fit more to the requirement of Stage 2. The ‘location’ (countries) and ‘life expectancy at birth’ were taken from my original Stage 1 data, and ‘life expectancy’ was taken from the group’s collective data. ‘Income groups’ was added in creating the new dataset.

**Provenance and Licenses**

The “location” and “life expectancy at birth” were taken from Kaggle (www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete) on 2 September 2022. The data is by Zeus on Kaggle and was last updated a year ago. The data is originally sourced from the World Health Organization (WHO), and filtered, cleaned, and put into subcategories to improve readability. The dataset has CC0: Public Domain licensing, no copyright, and is free for public use.

“Life expectancy” was sourced from Kaggle (//www.kaggle.com/datasets/augustus0498/life-expectancy-who?resource=download). The original dataset was collected from WHO and United Nations website with contribution of Deeksha Russell and Duan Wang. The data was obtained from KumarRajarshi on Kaggle before it was posted by Akhil.

The classification of countries based on income was taken from a PDF from the UN website, (www.un.org/en/development/desa/policy/wesp/wesp\_current/2014wesp\_country\_classification.pdf). The data was made by Development Policy and Analysis Division (DPAD) of the Department of Economic and Social Affairs of the United Nations Secretariat (UN/DESA). The document itself has license CC Attribution-NonCommercial License. The table of classifications was in the PDF, and the list of countries was manually moved to a spreadsheet to be processed by Python in integrating it with the other parts.

**Changes Made to Make the Final Dataset**

In data integration, there were some changes made to the names of the countries if the different datasets had slightly different names for them. This was so that Python can integrate them as one entity and not separate entities. The data were intended to be used in pairs: life expectancy based on income groups, and life expectancy at birth based on the number of medical doctors. In data cleaning, if the data of a country does not contain at least a pair of the required data, it would be removed.

**Data Summary**

|  |  |  |
| --- | --- | --- |
| Field Name | Description | File Type |
| Location | Name of country | String |
| Life Expectancy (years) | Life expectancy, country wise mentioned in age (years) | Float |
| Life Expectancy at Birth (years) | Life expectancy at birth, country wise mentioned in age (years) | Float |
| ‘Medical Doctors (per 10,000)’ | Number of medical doctors, per 10,000 population | Float |
| Income Group | Classification of the countries based on the income group | String |