WHO Life Expectancy Dataset Analysis:

A screenshot of a computer

Description automatically generated

This is the column headings from the CSV:

Id,IndicatorCode,SpatialDimType,SpatialDim,TimeDimType,ParentLocationCode,ParentLocation,Dim1Type,**Dim1,**TimeDim,Dim2Type,Dim2,Dim3Type,Dim3,DataSourceDimType,DataSourceDim,Value,NumericValue,**Low,**High,Comments,Date,TimeDimensionValue,TimeDimensionBegin,TimeDimensionEnd

This is a sample of what the CSV lines look like:

24614413,WHOSIS\_000001,COUNTRY,ALB,YEAR,EUR,Europe,SEX,**MLE,**2010,,,,,,,74.2,74.23331,**,**,,2020-12-04T16:59:45.957+01:00,2010,2010-01-01T00:00:00+01:00,2010-12-31T00:00:00+01:00

There is 25 variables in the file, however 9 of them are completely empty. The CSV is set up in a strange way,

The first variable is a unique ID per entry, the second is the indicator code, which is the same for every entry as it represents the dataset. Next is SpatialDimType, which is either country, region, or world, next is SpatialDim, which is usually a 3 letter code for the country. Next is TimeDimType, always YEAR. The next 2 columns are for the region, and are always the same, with one being a 3 letter code and the other the full title. DIM1Type is always SEX, with Dim1 being either MLE, FMLE, or BTSX. TimeDim is the year, with 582 entries for 2000, 2010, 2015, and 2019. Dim2 through DataSourceDim are all empty, so it is unclear why they are included. Value is the age rounded to 1 decimal, and NumericValue is the unrounded age. Low, high, and comments are all empty as well. Date appears to be the day they were added to the dataset, as there are only 2 different dates, 2020-12-04 and 2020-12-07. TimeDimension Value is again the years the data is from, same as TimeDim. TimeDimensionBegin and TimeDimensionEnd are the first and last days of their respective years.

A close-up of a computer screen

Description automatically generated

For SpatialDim we have 194 unique entries, however this is not all just countries, as it includes some oter things like regions.

So, out of the 25 columns I only really need about 7. SpatialDimType, SpatialDim, ParentLocationCode, Dim1, TimeDim, and NumericValue. Everything else is either unnecessary or just redundant.

A screenshot of a review

Description automatically generated

This is the column headings from the csv

Country Name,Code,1960,1961,1962,1963,1964,1965,**1966,**1967,1968,1969,1970,1971,1972,1973,1974,1975,**1976,**1977,1978,1979,1980,1981,1982,1983,1984,1985,**1986,**1987,1988,1989,1990,1991,1992,1993,1994,1995,**1996,**1997,1998,1999,2000,2001,2002,2003,2004,2005,**2006,**2007,2008,2009,2010,2011,2012,2013,2014,2015,**2016,**2017,2018,2019,2020,Unnamed: 65

This is an example entry from the csv. As you can see there are many null values, however most of them are very early on around the 60s, and my data from the other set starts at 2000 so this shouldn’t be too big of an issue.

Aruba,ABW,,,,,,,**,**,,,,,,,,,**,**,,,,,,,,,**6472.39870887477,**7885.15892730063,9765.90920740626,11392.2691499861,12306.7176791493,13495.79430054,14045.4745656146,14938.6815734932,16239.9878515928,16439.3563609282,**16583.8758560136,**17927.9594391833,19077.6874273513,19355.3335357274,20617.7504673528,20670.3670051627,20436.8871286309,20833.976357295,22568.3749910043,23300.7383630907,**24046.2263769373,**25834.1117784233,27083.6348596914,24631.1820522439,23513.527696809,24985.0139185797,24712.4932628907,26441.6199357378,26893.0115056581,28396.9084228616,**28452.1706146796,**29350.8050189137,30253.2793580195,,,

For the GDP dataset, there are 266 countries, and it features their full name, and the 3 letter code. So, I should be able to combine these datasets on the 3 letter code, assuming they are ISO 3166-1 alpha-3 codes, which they both appear to be. And as previously seen, the other dataset has less countries so that will be the limiting factor for how many countries are used. Below are screenshots of the years I plan to use. As you can see 3 to 7 percent of values are missing. But, over 70 of the countries in this dataset are not in the other one, so hopefully the ones missing data are also not in the other set, as we can assume the biggest countries would have the most complete data. As you can see the mean more than doubles between 2000 and 2019, which is rather impressive in only 19 years.

A screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated