REPORT

Data Analysis

CORRELATION BETWEEN LIFE EXPECTANCY AND SOCIO-ECONOMIC FACTORS WORLD BANK DATASET - 2000-2019

Prepared By: Jillian Lee

November 2023

Overview

1. **Background and Objective**
2. **Overview of the Dataset**
3. **Limitations of the Dataset**
4. **Data Download and Cleaning**
5. **Tiers 1-3 Analysis**
6. **Conclusion**

Background and Objective

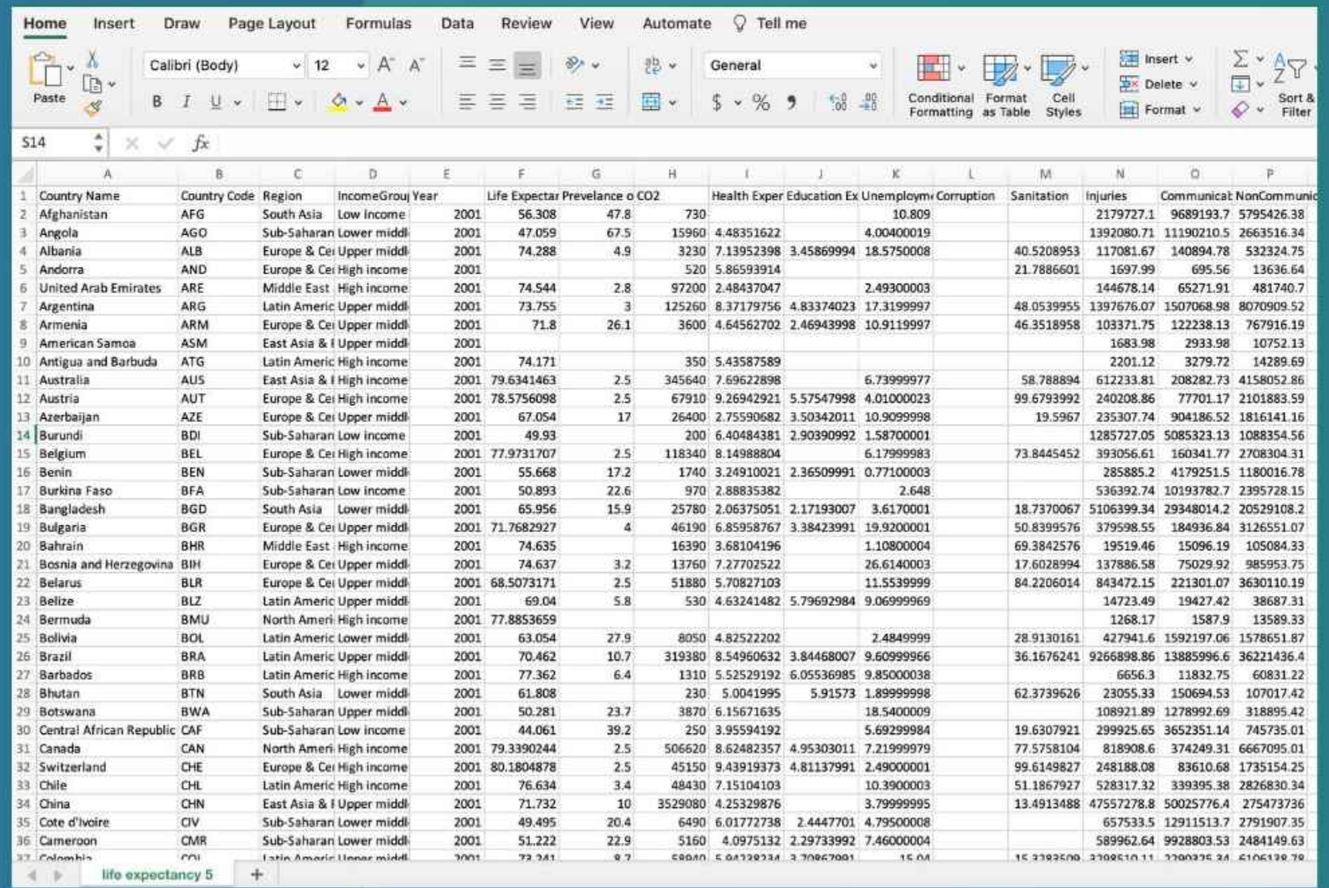
Life expectancy is a key metric used by countries across the world to assess their population health and overall well-being. Over the years, life expectancy has increased significantly, owing to factors such as improvements in medicine, sanitation, nutrition and education. Yet, despite these improvements, we still see inequity in life expectancy across and within countries.

The objective of this analysis is to determine if various socio-economic factors (such as income levels, CO2 emissions, health expenditure, etc.) are a useful indicator or predictor of life expectancy.

Overview of the Dataset

The dataset I have chosen is a World Bank CSV Dataset which covers a period of 20 years – from 2000 to 2019. For the purpose of this report, I would like to specifically focus on **2019** to present a more current analysis.

The table in the original CSV contains sixteen (16) columns containing the Country Names and corresponding attributes, such as Region, Income Group, Life Expectancy, etc. Apart from filtering for the year, I have additionally filtered out columns which I have chosen to not incorporate into my analysis including "Country Code" and "Injuries".



Home Insert Draw Page Layout Formulas Data Review View Automate Tell me

Calibri (Body) 12 General

Paste Conditional Formatting Format as Table Cell Styles Insert Delete Format Sort & Filter

S14

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Country Name	Country Code	Region	IncomeGrou	Year	Life Expectar	Prevelance o	CO2	Health Exper	Education Ex	Unemploym	Corruption	Sanitation	Injuries	Communicat	NonCommunic
2	Afghanistan	AFG	South Asia	Low income	2001	56.308	47.8	730			10.809			2179727.1	9689193.7	5795426.38
3	Angola	AGO	Sub-Saharan	Lower middl	2001	47.059	67.5	15960	4.48351622		4.00400019			1392080.71	11190210.5	2663516.34
4	Albania	ALB	Europe & Ce	Upper middl	2001	74.288	4.9	3230	7.13952398	3.45869994	18.5750008		40.5208953	117081.67	140894.78	532324.75
5	Andorra	AND	Europe & Ce	High income	2001			520	5.86593914				21.7886601	1697.99	695.56	13636.64
6	United Arab Emirates	ARE	Middle East	High income	2001	74.544	2.8	97200	2.48437047		2.49300003			144678.14	65271.91	481740.7
7	Argentina	ARG	Latin Americ	Upper middl	2001	73.755	3	125260	8.37179756	4.83374023	17.3199997		48.0539955	1397676.07	1507068.98	8070909.52
8	Armenia	ARM	Europe & Ce	Upper middl	2001	71.8	26.1	3600	4.64562702	2.46943998	10.9119997		46.3518958	103371.75	122238.13	767916.19
9	American Samoa	ASM	East Asia & I	Upper middl	2001									1683.98	2933.98	10752.13
10	Antigua and Barbuda	ATG	Latin Americ	High income	2001	74.171		350	5.43587589					2201.12	3279.72	14289.69
11	Australia	AUS	East Asia & I	High income	2001	79.6341463	2.5	345640	7.69622898		6.73999977		58.788894	612233.81	208282.73	4158052.86
12	Austria	AUT	Europe & Ce	High income	2001	78.5756098	2.5	67910	9.26942921	5.57547998	4.01000023		99.6793992	240208.86	77701.17	2101883.59
13	Azerbaijan	AZE	Europe & Ce	Upper middl	2001	67.054	17	26400	2.75590682	3.50342011	10.9099998		19.5967	235307.74	904186.52	1816141.16
14	Burundi	BDI	Sub-Saharan	Low income	2001	49.93		200	6.40484381	2.90390992	1.58700001			1285727.05	5085323.13	1088354.56
15	Belgium	BEL	Europe & Ce	High income	2001	77.9731707	2.5	118340	8.14988804		6.17999983		73.8445452	393056.61	160341.77	2708304.31
16	Benin	BEN	Sub-Saharan	Lower middl	2001	55.668	17.2	1740	3.24910021	2.36509991	0.77100003			285885.2	4179251.5	1180016.78
17	Burkina Faso	BFA	Sub-Saharan	Low income	2001	50.893	22.6	970	2.88835382		2.648			536392.74	10193782.7	2395728.15
18	Bangladesh	BGD	South Asia	Lower middl	2001	65.956	15.9	25780	2.06375051	2.17193007	3.6170001		18.7370067	5106399.34	29348014.2	20529108.2
19	Bulgaria	BGR	Europe & Ce	Upper middl	2001	71.7682927	4	46190	6.85958767	3.38423991	19.9200001		50.8399576	379598.55	184936.84	3126551.07
20	Bahrain	BHR	Middle East	High income	2001	74.635		16390	3.68104196		1.10800004		69.3842576	19519.46	15096.19	105084.33
21	Bosnia and Herzegovina	BIH	Europe & Ce	Upper middl	2001	74.637	3.2	13760	7.27702522		26.6140003		17.6028994	137886.58	75029.92	985953.75
22	Belarus	BLR	Europe & Ce	Upper middl	2001	68.5073171	2.5	51880	5.70827103		11.5539999		84.2206014	843472.15	221301.07	3630110.19
23	Belize	BLZ	Latin Americ	Upper middl	2001	69.04	5.8	530	4.63241482	5.79692984	9.06999969			14723.49	19427.42	38687.31
24	Bermuda	BMU	North Ameri	High income	2001	77.8853659								1268.17	1587.9	13589.33
25	Bolivia	BOL	Latin Americ	Lower middl	2001	63.054	27.9	8050	4.82522202		2.4849999		28.9130161	427941.6	1592197.06	1578651.87
26	Brazil	BRA	Latin Americ	Upper middl	2001	70.462	10.7	319380	8.54960632	3.84468007	9.60999966		36.1676241	9266898.86	13885996.6	36221436.4
27	Barbados	BRB	Latin Americ	High income	2001	77.362	6.4	1310	5.52529192	6.05536985	9.85000038			6656.3	11832.75	60831.22
28	Bhutan	BTN	South Asia	Lower middl	2001	61.808		230	5.0041995	5.91573	1.89999998		62.3739626	23055.33	150694.53	107017.42
29	Botswana	BWA	Sub-Saharan	Upper middl	2001	50.281	23.7	3870	6.15671635		18.5400009			108921.89	1278992.69	318895.42
30	Central African Republic	CAF	Sub-Saharan	Low income	2001	44.061	39.2	250	3.95594192		5.69299984		19.6307921	299925.65	3652351.14	745735.01
31	Canada	CAN	North Ameri	High income	2001	79.3390244	2.5	506620	8.62482357	4.95303011	7.21999979		77.5758104	818908.6	374249.31	6667095.01
32	Switzerland	CHE	Europe & Ce	High income	2001	80.1804878	2.5	45150	9.43919373	4.81137991	2.49000001		99.6149827	248188.08	83610.68	1735154.25
33	Chile	CHL	Latin Americ	High income	2001	76.634	3.4	48430	7.15104103		10.3900003		51.1867927	528317.32	339395.38	2826830.34
34	China	CHN	East Asia & I	Upper middl	2001	71.732	10	3529080	4.25329876		3.79999995		13.4913488	47557278.8	50025776.4	275473736
35	Cote d'Ivoire	CIV	Sub-Saharan	Lower middl	2001	49.495	20.4	6490	6.01772738	2.4447701	4.79500008			657533.5	12911513.7	2791907.35
36	Cameroon	CMR	Sub-Saharan	Lower middl	2001	51.222	22.9	5160	4.0975132	2.29733992	7.46000004			589962.64	9928803.53	2484149.63
37	Colombia	COL	Latin Americ	Upper middl	2001	73.241	8.2	58940	5.94238234	3.70867991	15.04		15.3383509	3298510.11	3290935.34	6106138.78

life expectancy 5 +

Dataset Source:
https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank/data

Limitations of the Dataset

Limitation 1

For some countries, there are missing values for columns such as “Prevalence of Undernourishment”, “CO2”, “Education Expenditure %” and “Corruption”. Caution will be exercised in interpreting the results, so that no invalid or biased conclusions be made.

Limitation 2

Ten (10) countries were missing values for Life Expectancy, and as such, were filtered out of the dataset for the analysis.

Automatic Data Download

IDLE Shell 3.12.0

```
Python 3.12.0 (v3.12.0:0fb18b02c8, Oct  2 2023, 09:45:56) [Clang 13.0.0 (clang-1300.0.29.30)]
on darwin
Type "help", "copyright", "credits" or "license()" for more information.
>>> import  requests
>>>
>>> url = 'https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank/
download?datasetVersionNumber=1'
>>>
>>> response=requests.get(url)
>>> with open('life_expectancy.csv', 'wb') as f:
...     if response.ok:
...         f.write(response.content)
...         print('finished writing')
...
...
4913
finished writing
>>>
```

Data Cleaning

```
>>>
>>> import pandas as pd
>>> df=pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df.drop(columns=['Country Code', 'Injuries'], inplace=True)
>>> df_2019=df.loc[df['Year'] == 2019]
>>> df_new=df_2019.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>> print(df_new)
              Country Name  ... NonCommunicable
3132           Afghanistan  ...      7601757.82
3133                Angola  ...      4176568.27
3134               Albania  ...       631629.88
3136  United Arab Emirates  ...      1637717.40
3137             Argentina  ...      9699014.80
...                    ...  ...             ...
3301               Vanuatu  ...        69213.56
3302                 Samoa  ...        43798.62
3303          South Africa  ...     10214261.89
3304                Zambia  ...      2649687.82
3305              Zimbabwe  ...      2364031.48

[164 rows x 14 columns]
>>>
```

Here we are filtering the dataset to account for data from 2019 only, and dropping the columns we won't be needing, as well as the rows with missing life expectancy data.

Tier 1 Analysis

For the first section, we will be seeking to extract the following data points:

1. Minimum and maximum life expectancy – by Region
2. Sorted list of life expectancy by lowest to highest
3. Range in terms of life expectancy – a) by Region; and b) by Income Group
4. Mean Life Expectancy – a) overall; b) by Region; and b) by Income Group
5. Scatterplot - Life expectancy vs. Health Expenditure %
6. Kernel Density Estimate (KDE) Plot - Distribution of observations in terms of life expectancy
7. Ranking of top 10 and bottom 10 countries in terms of life expectancy
8. Sqlite3 - Total number of countries with life expectancy lower than 50 years

1. Minimum and Maximum Life Expectancy - by Region

```
>>>
>>> result=df_new.groupby('Region').agg({'Life Expectancy World Bank': ['min', 'max']})
>>> print(result)
                              Life Expectancy World Bank
                                                     min        max
Region
East Asia & Pacific                            64.501000  84.356341
Europe & Central Asia                          68.191000  83.904878
Latin America & Caribbean                      64.001000  80.279000
Middle East & North Africa                     67.112000  82.858537
North America                                  78.787805  82.048780
South Asia                                     64.833000  78.921000
Sub-Saharan Africa                             53.283000  74.235854
>>>
```

Here, we are using the groupby statement and aggregate function to retrieve the minimum and maximum values of life expectancy for each region.

This table shows the minimum and maximum life expectancy across all regions.

- From this table, we can see that the region with the minimum life expectancy (53 years) is Sub-Saharan Africa.
- The region with the maximum life expectancy (84 years) is East Asia & Pacific.

2. Sorted List of Life Expectancy by lowest to highest

```
>>>
>>> result=df_new.groupby('Life Expectancy World Bank')['Country Name'].sum()
>>> print(result)
Life Expectancy World Bank
53.283000    Central African Republic
54.239000                        Chad
54.331000                     Lesotho
54.687000                     Nigeria
54.696000               Sierra Leone
                               ...
83.497561                       Italy
83.595122                   Singapore
83.831707                       Spain
83.904878                 Switzerland
84.356341                       Japan
Name: Country Name, Length: 164, dtype: object
>>>
```

We are also using the groupby statement on two columns to retrieve the life expectancy by country.

- From this table, we can see that the country with the lowest life expectancy is Central African Republic (53 years), and the country with the highest life expectancy is Japan (84 years).
- The data here is collapsed and is good for viewing the top/bottom values at a glance, but if we wanted to view the full table, we could input the following: pd.set_option('display.max_rows', 200)

3. Range in terms of Life Expectancy - (a) by Region

```
>>>
>>> df_new.groupby('Region').apply(lambda x: x['Life Expectancy World Bank'].max() - x['Life
Expectancy World Bank'].min())
Region
East Asia & Pacific           19.855341
Europe & Central Asia         15.713878
Latin America & Caribbean     16.278000
Middle East & North Africa    15.746537
North America                  3.260976
South Asia                    14.088000
Sub-Saharan Africa            20.952854
dtype: float64
>>>
```

- We will use the groupby statement to filter by Region and use the lambda function to perform subtraction to determine the range of life expectancy (maximum - minimum values).
- As we can see from the table, the region with the largest variance in range (i.e. greatest disparity in terms of life expectancy) is Sub-Saharan Africa (20.95), followed by East Asia & Pacific (19.86) and Latin America & the Caribbean (16.28).

3. Range in terms of Life Expectancy - (b) by Income Group

```
>>>
>>> df_new.groupby('IncomeGroup').apply(lambda x: x['Life Expectancy World Bank']
.max() - x['Life Expectancy World Bank'].min())
IncomeGroup
High income             12.30878
Low income              15.74100
Lower middle income     24.59900
Upper middle income     21.54400
dtype: float64
>>>
```

- We apply the same methodology as we used before to retrieve the range of life expectancy by Income Group.
- The income group with the largest variance in range is “Lower Middle Income” (24.60), followed by “Upper Middle Income” (21.54).

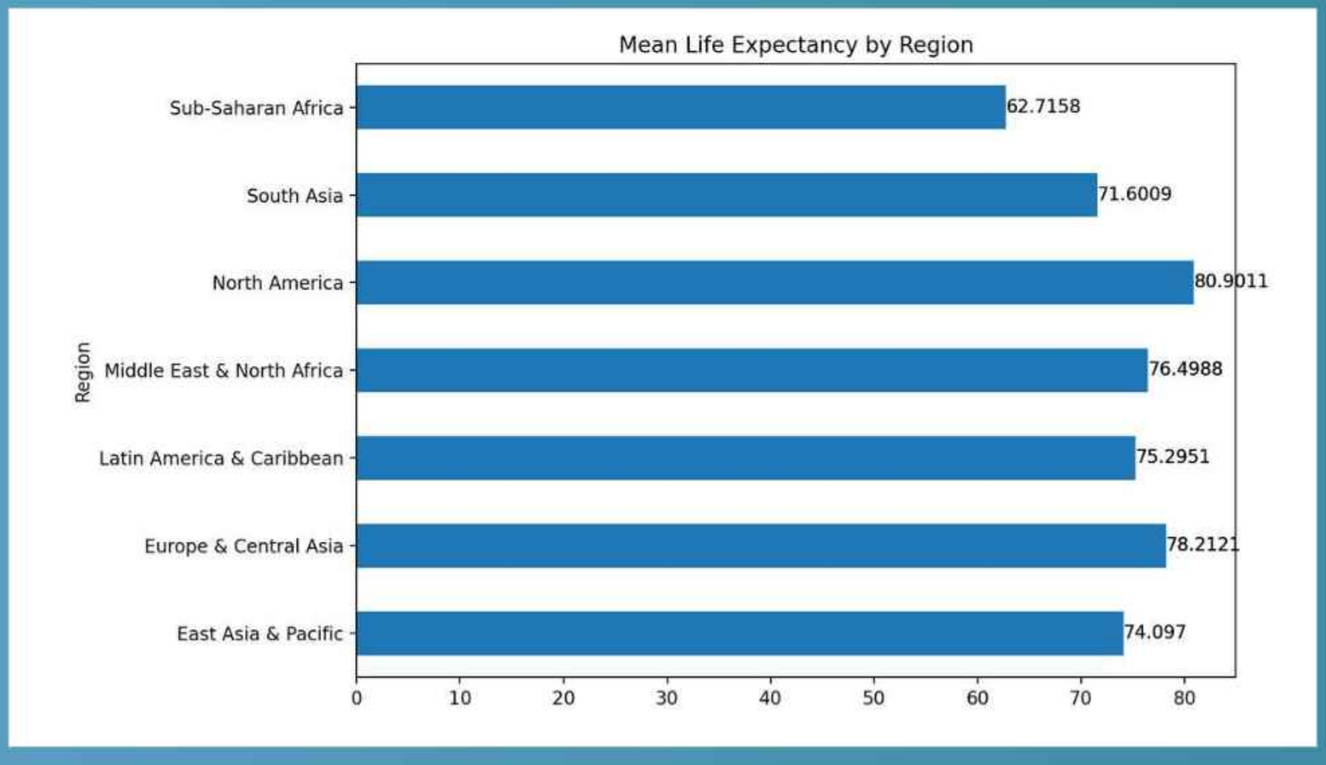
4. Mean Life Expectancy - (a) overall; (b) by Region

```
>>>
>>> df_new['Life Expectancy World Bank'].mean()
72.58911154074956
>>>
>>> region=df_new.groupby(['Region'])['Life Expectancy World Bank'].mean()
>>> print(region)
Region
East Asia & Pacific                 74.096990
Europe & Central Asia               78.212050
Latin America & Caribbean           75.295054
Middle East & North Africa          76.498838
North America                       80.901057
South Asia                          71.600875
Sub-Saharan Africa                  62.715777
Name: Life Expectancy World Bank, dtype: float64
>>>
```

- The mean life expectancy for the dataset is 72 years.
- We can use the groupby statement to retrieve the mean life expectancy by Region. This is shown in the table above.
- The region with the highest mean life expectancy is North America (80 years), followed by Europe & Central Asia (78 years). The region with the lowest mean life expectancy is Sub-Saharan Africa (62 years), followed by South Asia (71 years).

4. BAR CHART - Mean Life Expectancy - (b) by Region

```
>>> import matplotlib.pyplot as plt
>>>
>>> region_plot=region.plot.barh(x='Region', y='Mean Life Expectancy')
>>> region_plot.bar_label(region_plot.containers[0])
[Text(0, 0, '74.097'), Text(0, 0, '78.2121'), Text(0, 0, '75.2951'), Text(0,
0, '76.4988'), Text(0, 0, '80.9011'), Text(0, 0, '71.6009'), Text(0, 0, '62.7
158')]
>>> plt.title('Mean Life Expectancy by Region')
Text(0.5, 1.0, 'Mean Life Expectancy by Region')
>>> plt.show()
```



- We can visualize our data by creating a bar chart.
- To begin, we will need to import the matplotlib.pyplot library and then define our x and y axes.
- In the third statement, we access the containers in the axes to assign a label so that we can see the values for each bar.
- In the fourth statement, we set a title for our Figure, and lastly we enter plt.show() which will open an interactive window to display our bar chart.

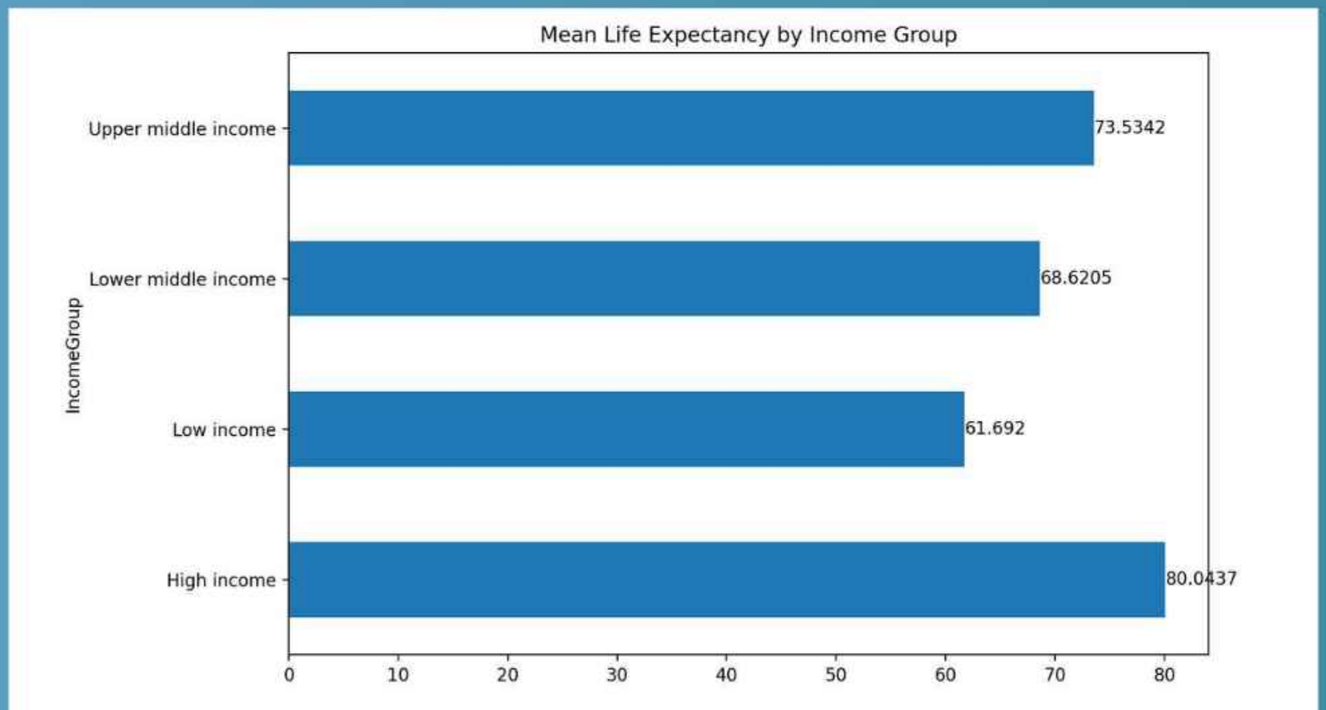
4. Mean Life Expectancy - (c) by Income Group

```
>>>
>>> income=df_new.groupby(['IncomeGroup'])['Life Expectancy World Bank'].mean()
>>> print(income)
IncomeGroup
High income             80.043734
Low income              61.692000
Lower middle income     68.620451
Upper middle income     73.534178
Name: Life Expectancy World Bank, dtype: float64
>>>
```

- We can use the groupby statement to retrieve the mean life expectancy by Income Group. The results are shown in the table above.
- The income group with the highest mean life expectancy is the High Income Group (80 years). The income group with the lowest mean life expectancy is the Low Income Group (61 years).

4. BAR CHART - Mean Life Expectancy - (c) by Income Group

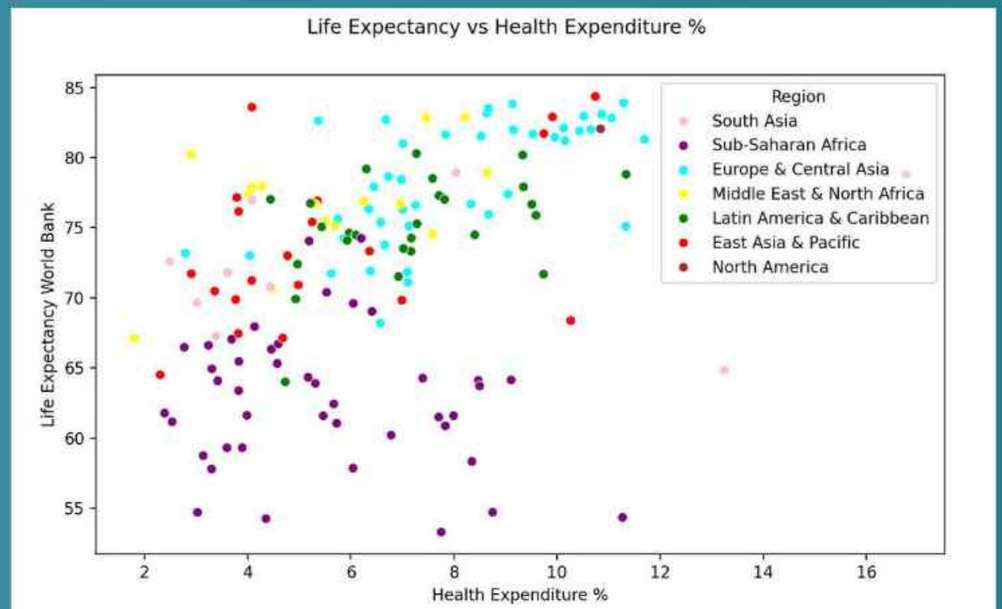
```
>>>
>>> income_plot=income.plot.barh(x='Income Group', y='Mean Life Expectancy')
>>> income_plot.bar_label(income_plot.containers[0])
[Text(0, 0, '80.0437'), Text(0, 0, '61.692'), Text(0, 0, '68.6205'), Text(
0, 0, '73.5342')]
>>> plt.title('Mean Life Expectancy by Income Group')
Text(0.5, 1.0, 'Mean Life Expectancy by Income Group')
>>> plt.show()
```



We apply the same methodology as before to create the bar chart to represent Mean Life Expectancy by Income Group.

5. SCATTERPLOT - Life Expectancy vs. Health Expenditure %

```
>>>
>>> import seaborn as sns
>>> fig, ax = plt.subplots(figsize=(6, 4))
>>> colors = {'East Asia & Pacific': 'red', 'South Asia': 'pink', 'Europe & Ce
    ntral Asia': 'cyan', 'Latin America & Caribbean': 'green', 'Middle East &
    North Africa': 'yellow', 'North America': 'brown', 'Sub-Saharan Africa': '
    purple'}
>>> sns.scatterplot(data=df_new, x='Health Expenditure %', y='Life Expectancy
    World Bank', hue='Region', palette=colors, ax=ax)
    <Axes: xlabel='Health Expenditure %', ylabel='Life Expectancy World Bank'>
>>>
>>> ax.set(xlabel='Health Expenditure %', ylabel='Life Expectancy World Bank')
    [Text(0.5, 29.444444444444432, 'Health Expenditure %'), Text(75.0694444444
    4443, 0.5, 'Life Expectancy World Bank')]
>>> fig.suptitle('Life Expectancy vs Health Expenditure %')
    Text(0.5, 0.98, 'Life Expectancy vs Health Expenditure %')
>>> plt.show()
```



Methodology:

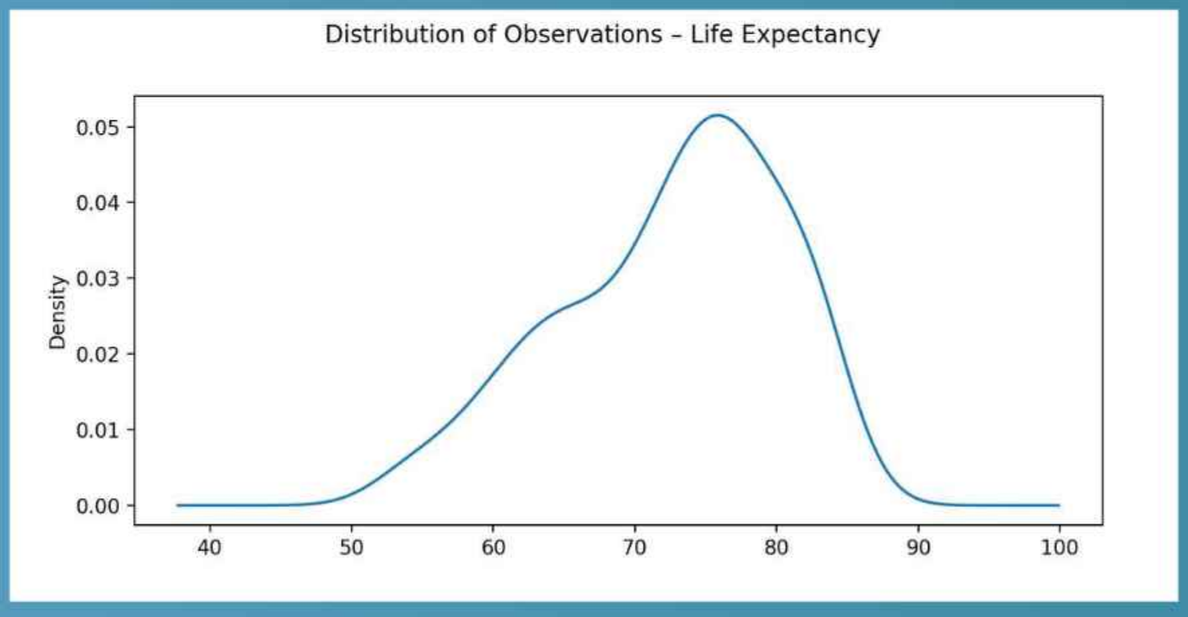
- To get a general idea of the relationship between Life Expectancy and Health Expenditure %, we can plot a scatterplot.
- We start by importing the seaborn library, which is a popular data visualization library.
- In the subsequent lines, we set the figure size of the plot, assign colors for each Region being graphed, set our x and y axes, and create a title.
- We enter plt.show() to display our scatterplot.

Analysis:

- From the plot, we can see that although East Asia & Pacific has the highest value in terms of life expectancy, the life expectancy for most countries within that region ranges between 60-80 years.
- Many countries in Europe & Central Asia have high life expectancy and report high levels of health expenditure% (light blue datapoints concentrated towards the top-right of the chart).
- Most countries within Sub-Saharan Africa have low life expectancies, and even those with high health expenditure still report some of the lowest levels of life expectancy.

6. KERNEL DENSITY ESTIMATE (KDE) PLOT - Distribution of Observations in Terms of Life Expectancy

```
>>>
>>> df_new['Life Expectancy World Bank'].plot(kind='kde')
    <Axes: ylabel='Density'>
>>> plt.suptitle('Distribution of Observations – Life Expectancy')
    Text(0.5, 0.98, 'Distribution of Observations – Life Expectancy')
>>> plt.show()
```



- Here, we are plotting the Kernel Density Estimate (KDE) to show the distribution of observations of life expectancy.
- The plot is negatively skewed which indicates to us that most countries in our dataset have higher life expectancies.

7. Ranking - Top 10/Bottom 10 Countries by Life Expectancy

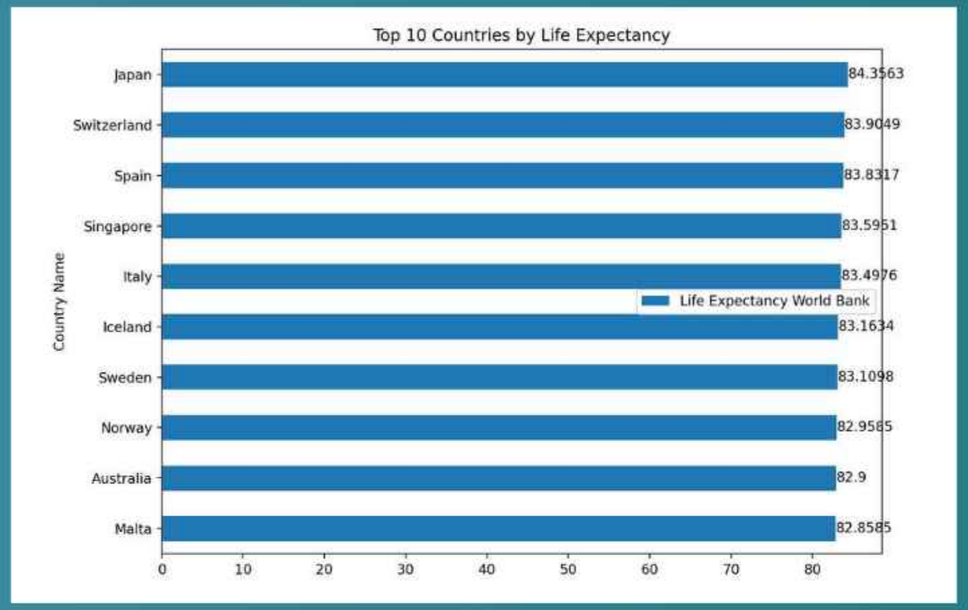
```
>>>
>>> top_10=df_new.nlargest(10,['Life Expectancy World Bank'])
>>> print(top_10)
     Country Name                     Region  ... Communicable  NonCommunicable
3212        Japan        East Asia & Pacific  ...   2030122.25      31250149.03
3162  Switzerland      Europe & Central Asia  ...     84137.81       1937774.53
3180        Spain      Europe & Central Asia  ...    494856.88      11233598.70
3271    Singapore        East Asia & Pacific  ...     84309.18        871899.33
3209        Italy      Europe & Central Asia  ...    552868.23      16313740.19
3207      Iceland      Europe & Central Asia  ...      3333.12         67186.17
3282       Sweden      Europe & Central Asia  ...    103748.70       2403126.02
3250       Norway      Europe & Central Asia  ...     59204.16       1182175.35
3141    Australia        East Asia & Pacific  ...    217234.75       5323684.11
3235        Malta  Middle East & North Africa  ...      5685.89        108042.18

[10 rows x 14 columns]
>>>
>>> bottom_10=df_new.nsmallest(10,['Life Expectancy World Bank'])
>>> print(bottom_10)
                  Country Name  ... NonCommunicable
3160  Central African Republic  ...      1000223.98
3285                      Chad  ...      2459705.38
3222                   Lesotho  ...       454034.40
3247                   Nigeria  ...     31050075.61
3273              Sierra Leone  ...      1399862.43
3276                   Somalia  ...      3026841.19
3165            Cote d'Ivoire  ...      3722986.86
3278               South Sudan  ...      1305031.76
3191             Guinea-Bissau  ...       319742.08
3192         Equatorial Guinea  ...       167488.68

[10 rows x 14 columns]
>>>
```

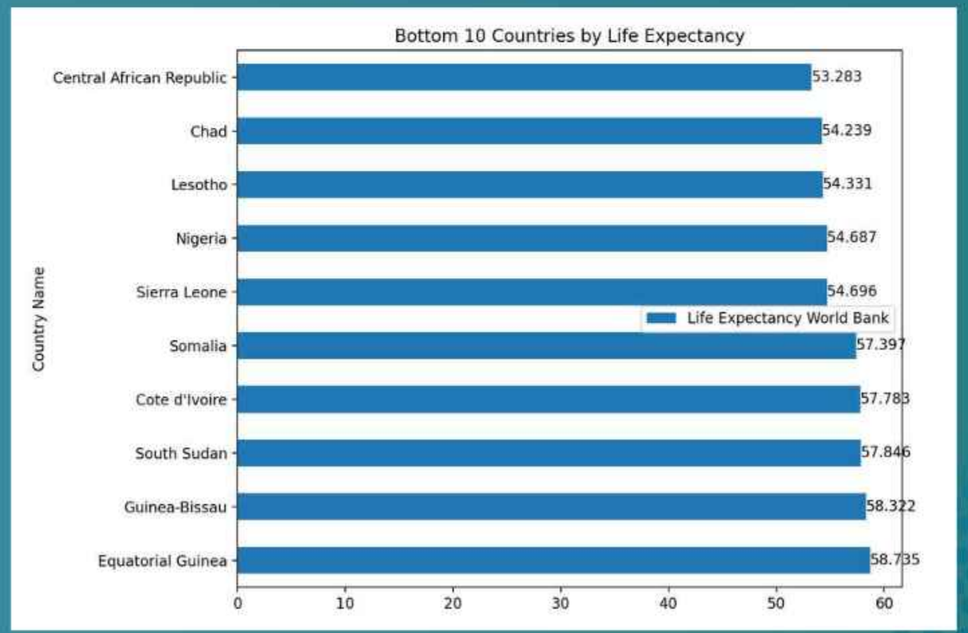
- We can find the top 10 and bottom 10 countries in terms of life expectancy by using the nlargest and nsmallest methods respectively.
- We specify that we are looking for the top 10/bottom 10 values as shown in the brackets.

7. BAR CHART - Ranking of Top 10/Bottom 10 Countries by Life Expectancy

```
>>>
>>> top_10_plot=top_10.plot.barh(x='Country Name', y='Life Expectancy World Bank')
>>> top_10_plot.bar_label(top_10_plot.containers[0])
    [Text(0, 0, '84.3563'), Text(0, 0, '83.9049'), Text(0, 0, '83.8317'), Text(0, 0, '83.595
    1'), Text(0, 0, '83.4976'), Text(0, 0, '83.1634'), Text(0, 0, '83.1098'), Text(0, 0, '82
    .9585'), Text(0, 0, '82.9'), Text(0, 0, '82.8585')]
>>> top_10_plot.invert_yaxis()
>>> plt.title('Top 10 Countries by Life Expectancy')
    Text(0.5, 1.0, 'Top 10 Countries by Life Expectancy')
>>> plt.show()
>>>
>>> bottom_10_plot=bottom_10.plot.barh(x='Country Name', y='Life Expectancy World Bank')
>>> bottom_10_plot.bar_label(bottom_10_plot.containers[0])
    [Text(0, 0, '53.283'), Text(0, 0, '54.239'), Text(0, 0, '54.331'), Text(0, 0, '54.687'),
    Text(0, 0, '54.696'), Text(0, 0, '57.397'), Text(0, 0, '57.783'), Text(0, 0, '57.846'),
    Text(0, 0, '58.322'), Text(0, 0, '58.735')]
>>> bottom_10_plot.invert_yaxis()
>>> plt.title('Bottom 10 Countries by Life Expectancy')
    Text(0.5, 1.0, 'Bottom 10 Countries by Life Expectancy')
>>> plt.show()
```

- We input the code above to plot horizontal bar graphs for the data, assign value labels, and add titles.
- We also take an extra step to invert the y-axes so that the data for the top 10/bottom 10 countries is in order (descending and ascending respectively).





8. Sqlite3 - Total Number of Countries with Life Expectancy <50 years

```
>>> df= pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df.drop(columns=['Country Code', 'Injuries'], inplace=True)
>>> df_2019=df.loc[df['Year'] == 2019]
>>> df_new=df_2019.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>> print(df_new)
              Country Name  ... NonCommunicable
3132           Afghanistan  ...      7601757.82
3133                Angola  ...      4176568.27
3134               Albania  ...       631629.88
3136  United Arab Emirates  ...      1637717.40
3137             Argentina  ...      9699014.80
...                    ...  ...             ...
3301               Vanuatu  ...        69213.56
3302                 Samoa  ...        43798.62
3303          South Africa  ...     10214261.89
3304                Zambia  ...      2649687.82
3305              Zimbabwe  ...      2364031.48

[164 rows x 14 columns]
>>> con = sqlite3.connect('life_expectancy.db')
>>> cursor=con.cursor()
>>> df_new.to_sql('MyTable', con, if_exists='replace')
164
>>> result = cursor.execute("SELECT COUNT(*) from 'Country Name' where 'Life Expectancy World Bank' < 50")
```

- **We can use the SELECT statement in Sqlite3 as shown above to extract the total number of countries with life expectancy <50 years.**

Tier 2 Analysis

For this section, we will be seeking to extract the following data points:

1. Median percentage spent on healthcare over the past 10 years (2010-2019) – group by (a) region and by (b) region & income group.
2. CO2 emissions (megatons) for countries with life expectancy lower than 60 years and over 75 years.

1. Median Percentage Spent on Healthcare - 2010-2019
Group By - (a) Region

```
>>>
>>> import pandas as pd
>>> df=pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df_2010_2019=df.loc[df['Year']>=2010].loc [df['Year']<=2019]
>>> df_new= df_2010_2019.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>>
>>> result=df_new.groupby(['Region'])['Health Expenditure %'].median()
>>> print(result)
Region
East Asia & Pacific             4.408508
Europe & Central Asia           8.079246
Latin America & Caribbean       6.647565
Middle East & North Africa      5.172563
North America                  13.543739
South Asia                      3.663719
Sub-Saharan Africa              5.007354
Name: Health Expenditure %, dtype: float64
>>>
```

- Although we filtered earlier for 2019, we are going to use a different filter here to extract data for the 10-year period being studied (2010-2019).
- Historically over the 10-year period, we see that the median is highest for North America (13.54) and lowest for South Asia (3.66), with East Asia & Pacific (4.41) having the second-lowest median.

1. Median Percentage Spent on Healthcare - 2010-2019
Group By - (b) Region & Income Group

```
>>>
>>> df_new.groupby(['Region','IncomeGroup'])['Health Expenditure %'].median()
Region                      IncomeGroup
East Asia & Pacific         High income            9.243402
                            Lower middle income    4.392903
                            Upper middle income    3.802595
Europe & Central Asia       High income            8.910776
                            Lower middle income    6.738881
                            Upper middle income    7.119706
Latin America & Caribbean   High income            6.866024
                            Lower middle income    7.251419
                            Upper middle income    6.071579
Middle East & North Africa  High income            4.180218
                            Lower middle income    6.099394
                            Upper middle income    5.263421
North America               High income           13.543739
South Asia                  Low income             9.817109
                            Lower middle income    3.454076
                            Upper middle income    8.311842
Sub-Saharan Africa          High income            4.868223
                            Low income             5.312203
                            Lower middle income    4.244715
                            Upper middle income    5.751601
Name: Health Expenditure %, dtype: float64
>>>
```

- We can go further by seeing which income segments of the population receive the most healthcare (in terms of healthcare expenditure %). We use the groupby statement to filter for two columns “Region” and “IncomeGroup” and find the corresponding “Health Expenditure %” values.

- As we can see from the table, for even some of the poorest segments of the population (i.e. Low Income Group) in South Asia and Sub-Saharan Africa, priority is placed on healthcare. In South Asia, for example, the figure is 9.81% for the low-income group, which is higher than the health expenditure % for other income segments in that region.

2. CO2 Emissions (Megatons) for Countries with Life Expectancy Lower than 60 Years

```
>>> import pandas as pd
>>> df=pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df.drop(columns=['Country Code', 'Injuries'], inplace=True)
>>> df_filtered=df.loc[(df['Year'] == 2019) & (df['Life Expectancy World Bank']<60)]
>>> df_new=df_filtered.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>>
>>> result=df_new.groupby(['Country Name'])['CO2'].sum()
>>> print(result//1000)
Country Name
Cameroon                      9.0
Central African Republic      0.0
Chad                          2.0
Cote d'Ivoire                10.0
Equatorial Guinea             5.0
Guinea-Bissau                 0.0
Lesotho                       0.0
Mali                          5.0
Nigeria                     115.0
Sierra Leone                  0.0
Somalia                       0.0
South Sudan                   1.0
Name: CO2, dtype: float64
>>>
>>> mean= (df_new['CO2'].mean())//1000
>>> print(mean)
12.0
```

Methodology:

- Here we are filtering for 2019 data, as previously done in our Tier 1 analysis. We use the ampersand (&) in our df.loc attribute to also filter for life expectancy < 60 years.
- We use the groupby statement to filter for CO2 by Country.
- The table shows the CO2 emissions for countries with life expectancy lower than 60 years. The numbers shown as zeroes reflect as such (but are not missing values or equal to 0) because all values have been converted from kilotons to megatons (division by 1000).

Analysis:

- The CO2 emissions (megatons) for countries with life expectancy lower than 60 years is 12 megatons. Despite life expectancy being low, there are not a high level of CO2 emissions. As we will see on the next slide, even countries with higher life expectancies have much higher levels of CO2. CO2 emissions by itself may therefore not be a very strong predictor of life expectancy.
- Before we make any definitive conclusions, we will explore the correlation between all variables in our dataset later in our Tier 3 analysis.

2. CO2 Emissions (Megatons) for Countries with Life Expectancy Over 75 Years

```
>>>
>>> import pandas as pd
>>> df=pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df.drop(columns=['Country Code', 'Injuries'], inplace=True)
>>> df_filtered=df.loc[(df['Year'] == 2019) & (df['Life Expectancy World Bank']>75)]
>>> df_new=df_filtered.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>> df_new2=df_new.dropna(axis=0, subset=['CO2'])
>>> df_new3=df_new2.groupby(['Country Name'])['CO2'].sum()
>>> print(df_new3//1000)
Country Name
Albania                    4.0
Algeria                  171.0
Antigua and Barbuda        0.0
Argentina                168.0
Armenia                    6.0
                          ...
United Arab Emirates     188.0
United Kingdom           348.0
United States           4817.0
Uruguay                    6.0
Vietnam                  336.0
Name: CO2, Length: 72, dtype: float64
>>>
>>> mean=(df_new3.mean())//1000
>>> print(mean)
338.0
>>>
```

Methodology:

We follow the same methodology as before, but now filter for life expectancy above 75 years (in our 4th statement).

Analysis:

The mean CO2 emissions (megatons) for countries with life expectancy over 75 years is 338 megatons. As noted before, even countries with higher life expectancies have much higher levels of CO2.

Tier 3 Analysis

For this section, we will be seeking to determine the Pearson's correlation coefficient between all of columns or variables in our dataset.

The Pearson's correlation coefficient, also known as Pearson's r, measures the strength of a relationship between two variables and their association with one another.

We will use the corr method in our Pandas Dataframe to calculate the Pearson's correlation coefficient between the variables. We will then use data visualization library, Seaborn, to generate a heatmap to visualize the data.

Based on our findings, we will answer the following types of questions:

(a) What are the strongest predictors of life expectancy?
(b) Do factors like corruption and unemployment rate impact life expectancy?
(c) How does the prevalence of undernourishment and communicable disease affect life expectancy?

1. Pearson Correlation Coefficient Matrix - Heatmap

```
>>> import numpy as np
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
>>> import seaborn as sns
>>>
>>> df=pd.read_csv('file:///Users/jillianlee/Downloads/life%20expectancy.csv')
>>> df.drop(columns=['Country Code', 'Injuries'], inplace=True)
>>> df_2019=df.loc[df['Year'] == 2019]
>>> df_new=df_2019.dropna(axis=0, subset=['Life Expectancy World Bank'])
>>>
>>> pd.set_option('display.max_rows', 100)
>>> pd.set_option('display.max_columns', 100)
>>>
>>> pearsoncorr=df_new.corr(method='pearson',numeric_only=True)
>>> pearsoncorr
                                Year  Life Expectancy World Bank  \
Year                             NaN                         NaN
Life Expectancy World Bank       NaN                    1.000000
Prevelance of Undernourishment   NaN                   -0.662132
CO2                              NaN                    0.108793
Health Expenditure %             NaN                    0.409401
Education Expenditure %          NaN                   -0.002775
Unemployment                     NaN                   -0.137671
Corruption                       NaN                    0.249774
Sanitation                       NaN                    0.696055
Communicable                     NaN                   -0.244994
NonCommunicable                  NaN                    0.033581

                                Prevelance of Undernourishment       CO2  \
Year                                                       NaN       NaN
Life Expectancy World Bank                           -0.662132  0.108793
Prevelance of Undernourishment                        1.000000 -0.101344
CO2                                                  -0.101344  1.000000
Health Expenditure %                                 -0.190699  0.099855
Education Expenditure %                              -0.066703 -0.040648
Unemployment                                          0.107375 -0.066340
Corruption                                           -0.313721 -0.010854
Sanitation                                           -0.536296  0.112290
Communicable                                          0.137999  0.264197
NonCommunicable                                      -0.046970  0.869825
```

```
                                Health Expenditure %  Education Expenditure %  \
Year                                             NaN                      NaN
Life Expectancy World Bank                  0.409401                -0.002775
Prevelance of Undernourishment             -0.190699                -0.066703
CO2                                         0.099855                -0.040648
Health Expenditure %                        1.000000                 0.347334
Education Expenditure %                     0.347334                 1.000000
Unemployment                                0.153309                 0.265657
Corruption                                 -0.262683                 0.264464
Sanitation                                  0.354135                -0.084213
Communicable                               -0.234420                -0.083296
NonCommunicable                            -0.037837                -0.067519

                                Unemployment  Corruption  Sanitation  \
Year                                     NaN         NaN         NaN
Life Expectancy World Bank         -0.137671    0.249774    0.696055
Prevelance of Undernourishment      0.107375   -0.313721   -0.536296
CO2                                -0.066340   -0.010854    0.112290
Health Expenditure %                0.153309   -0.262683    0.354135
Education Expenditure %             0.265657    0.264464   -0.084213
Unemployment                        1.000000   -0.196130   -0.169762
Corruption                         -0.196130    1.000000    0.318273
Sanitation                         -0.169762    0.318273    1.000000
Communicable                       -0.051707    0.032513   -0.174141
NonCommunicable                    -0.075774   -0.036266    0.025603

                                Communicable  NonCommunicable
Year                                     NaN              NaN
Life Expectancy World Bank         -0.244994         0.033581
Prevelance of Undernourishment      0.137999        -0.046970
CO2                                 0.264197         0.869825
Health Expenditure %               -0.234420        -0.037837
Education Expenditure %            -0.083296        -0.067519
Unemployment                       -0.051707        -0.075774
Corruption                          0.032513        -0.036266
Sanitation                         -0.174141         0.025603
Communicable                        1.000000         0.646415
NonCommunicable                     0.646415         1.000000
>>>
```

- **We apply the Pandas df_corr() function [df_new.corr() in our case] to find the correlation among the columns in our dataframe using the 'Pearson' method. We retrieve the values as shown in the table in blue font.**

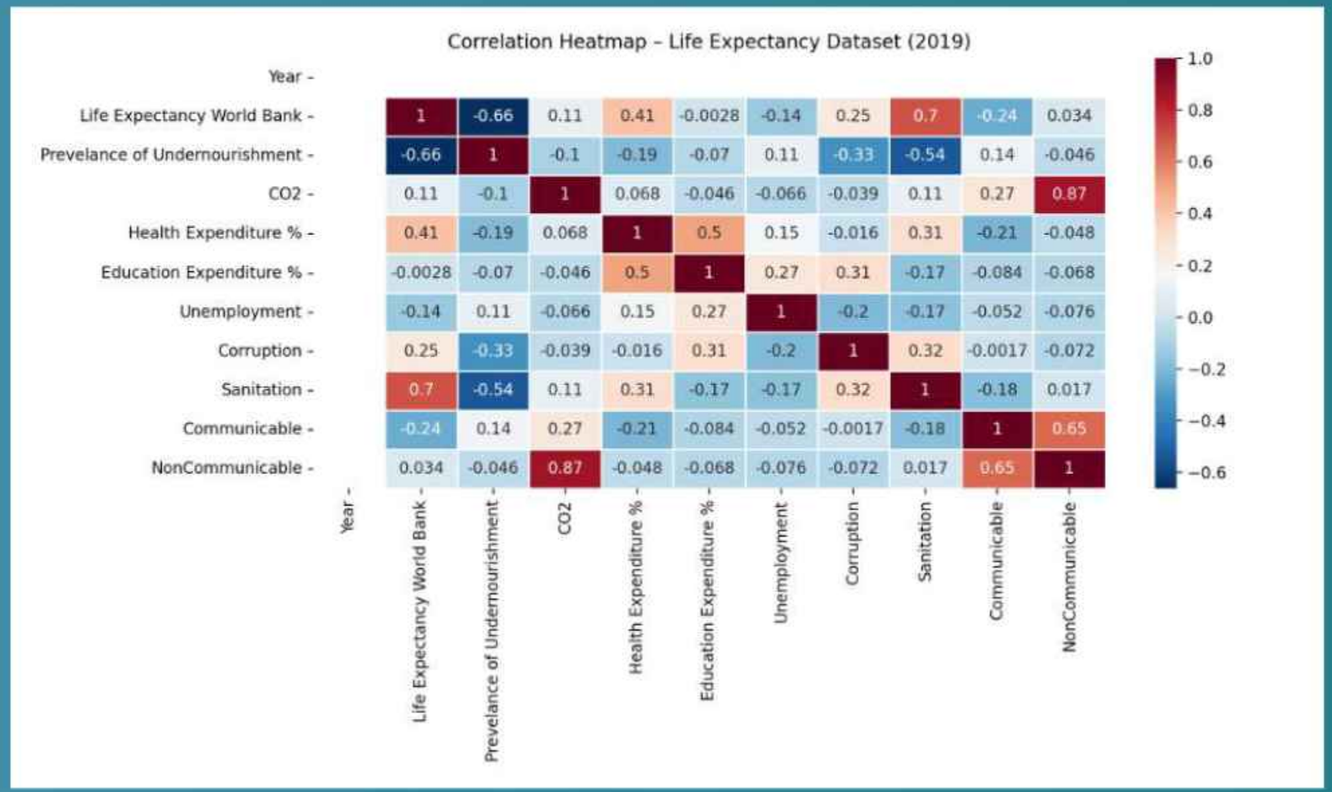
1. Pearson Correlation Coefficient Matrix (cont.) - Heatmap

- **We then use the function sns.heatmap() to plot the data as a color-encoded matrix.**

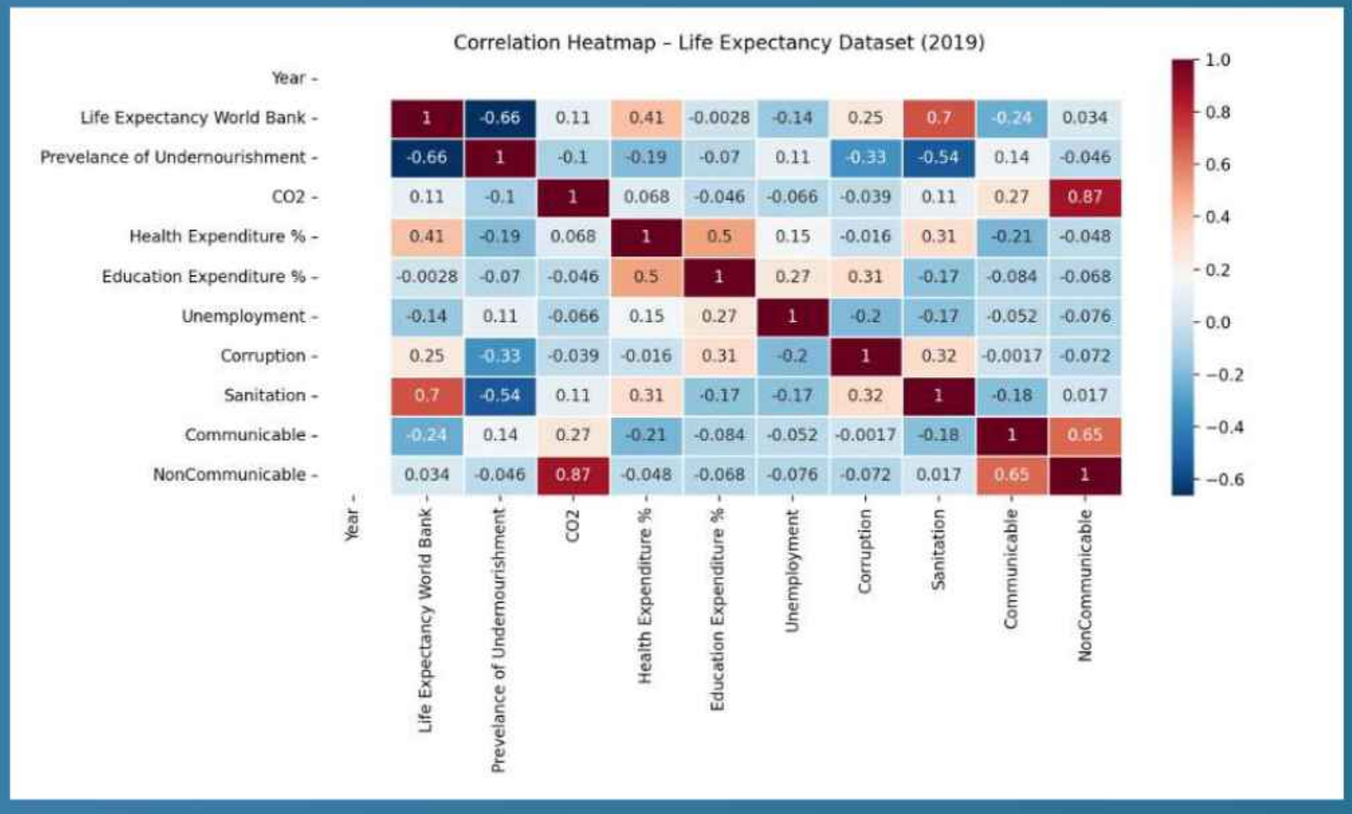
Interpreting the Pearson Correlation Matrix

Degree of Correlation	Interpretation
Perfect	If the value is near ±1, it is a perfect correlation – As one variable increases, the other tends to also increase (if positive) or decrease (if negative).
High Degree	If the coefficient value lies between ±0.5 and ±1, then there is a strong correlation.
Moderate Degree	If the value lies between ±0.30 and ±0.49, then there is a medium correlation.
Low Degree	When the value lies below ±0.29, then there is a small correlation.
No Correlation	When the value is zero.

```
>>>
>>> sns.heatmap(pearsoncorr, xticklabels=pearsoncorr.columns, yticklabels=pearsoncorr.colu
mns, cmap='RdBu_r', annot=True, linewidth=0.5)
<Axes: >
>>> plt.title('Correlation Heatmap – Life Expectancy Dataset (2019)')
Text(0.5, 1.0, 'Correlation Heatmap – Life Expectancy Dataset (2019)')
>>> plt.show()
```



1. Pearson Correlation Coefficient Matrix (cont.) - Heatmap Analysis



Based on the heatmap, we can see that the variable/column with the strongest degree of correlation (r) to life expectancy is Sanitation (0.7). The variables are positively correlated, and the higher the life expectancy, the higher the levels of sanitation (% of people in the population using safely managed sanitation services).

The variable with the second highest degree of correlation to life expectancy is undernourishment (-0.66). The variables are negatively correlated meaning that when one variable increases, the other decreases. This is expected with these two variables, as the lower the life expectancy, the higher the prevalence of undernourishment, and vice versa.

We can also see that there is a moderate degree of correlation (0.41) between life expectancy and health expenditure (%). These variables are positively correlated – the higher the health expenditure, the higher the life expectancy.

From our matrix, we also can see that the variable with the lowest correlation (-0.0028) relative to life expectancy is education expenditure (%). It is negatively correlated meaning that the higher the life expectancy, the lower the education expenditure (%).

Conclusion

Based on our findings, the variables with the strongest degree of correlation to Life Expectancy are “Sanitation” (r=0.7) and “Undernourishment” (r=-0.66). Variables with a small or negligible association to Life Expectancy are “Education Expenditure %” (r=-0.0028) and “Non Communicable” (r=0.034; i.e. disability-adjusted life years due to non-communicable diseases).

We also concluded that Health Expenditure has a moderate degree of correlation with Life Expectancy (r=0.41). An interesting finding from our analysis was that healthcare spend was found to be higher amongst the lowest income groups (relative to other income segments) in regions such as Sub-Saharan Africa and South Asia. These regions were also found to have the lowest mean life expectancy.

It is worthwhile to note that a degree of caution should be taken when interpreting the correlation of variables such as “Corruption” (r=0.25) since over half of the countries lacked data for that variable. Even if the data were filtered to omit the missing values, we would run the risk of misinterpreting the results and making biased conclusions.

The dataset provides a valuable opportunity for even further analysis using machine learning libraries such as scikit learn. We can use such tools to perform clustering to partition the data into logical groupings, or run a regression model to predict future trends in life expectancy.

Appendix

DEFINITION OF VARIABLES USED IN THE DATASET

Definition of Variables Used in the Dataset

Variable	Definition
Prevalence of Undernourishment	% of the undernourishment of the population
CO2	Kilotons of carbon dioxide emissions
Health Expenditure %	Level of current health expenditure expressed as a percentage of GDP
Education Expenditure %	General government expenditure on education expressed as a percentage of GDP
Unemployment	% of the share of the labor force that is without work but available for and seeking employment
Corruption	Corruption rating as obtained from the World Bank's Country Policy and Institutional Assessment (CPIA) survey (1=low to 6=high)
Sanitation	% of the population using safely managed sanitation facilities
Communicable	Disability-adjusted life years (DALYs) due to communicable diseases – One DALY represents the loss of the equivalent of one year of full health
Non Communicable	Disability-adjusted life years (DALYs) due to non-communicable diseases – One DALY represents the loss of the equivalent of one year of full health