Exploratory Data Analysis

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

--> "Life Expectancy Data.csv" information

In [2]:
```
df = pd.read_csv("Life Expectancy Data.csv")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   Year                             2938 non-null   int64  
 2   Status                           2938 non-null   object 
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64  
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64  
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64  
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
Country                             0
Year                                0
Status                              0
Life expectancy                    10
Adult Mortality                    10
infant deaths                       0
Alcohol                           194
percentage expenditure              0
Hepatitis B                       553
Measles                             0
 BMI                               34
under-five deaths                   0
Polio                              19
Total expenditure                 226
Diphtheria                         19
 HIV/AIDS                           0
GDP                               448
Population                        652
 thinness  1-19 years              34
 thinness 5-9 years                34
Income composition of resources   167
Schooling                         163
dtype: int64
```

In [4]:
```
df.head()
```

Out[4]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0

5 rows × 22 columns

In [5]:
```
df.tail()
```

Out[5]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatiti l
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.0	68.
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.0	7.
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.0	73.
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.0	76.
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.0	79.

5 rows × 22 columns

In [6]:
```
df.describe()
```

Out[6]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.00000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.94046
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.07001
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.00000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.00000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.00000
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.00000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.00000

In [7]:
```
df["Country"].describe()
df.Country.value_counts()
```

Out[7]:
```
Afghanistan              16
Peru                     16
Nicaragua                16
Niger                    16
Nigeria                  16
                         ..
Niue                      1
San Marino                1
Nauru                     1
Saint Kitts and Nevis     1
Dominica                  1
Name: Country, Length: 193, dtype: int64
```

In [8]:
```
df[["Country"]].describe(include="all")
```

Out[8]:

	Country
count	2938
unique	193
top	Afghanistan
freq	16

In [9]:
```
df[["Adult Mortality"]].describe(include="all")
```

Out[9]:

	Adult Mortality
count	2928.000000
mean	164.796448
std	124.292079
min	1.000000
25%	74.000000
50%	144.000000
75%	228.000000
max	723.000000

In [10]:
```
df['Status'].value_counts()
```

Out[10]:
```
Developing    2426
Developed      512
Name: Status, dtype: int64
```

In [11]:
```
df['Schooling'].value_counts()
```

Out[11]:
```
12.9    58
13.3    52
12.5    49
12.8    46
12.3    44
        ..
20.7     1
19.8     1
3.4      1
3.6      1
2.8      1
Name: Schooling, Length: 173, dtype: int64
```

In [12]:
```
df['Life expectancy '].value_counts()
```

Out[12]:
```
73.0    45
75.0    33
78.0    31
73.6    28
73.9    25
        ..
43.1     1
49.5     1
49.0     1
55.1     1
45.4     1
Name: Life expectancy , Length: 362, dtype: int64
```

--> Droping rows/coloumns with null values

In [13]:
```
df1 = df.dropna(axis = 0, how = 'any')
df1.shape
```

Out[13]:
```
(1649, 22)
```

In [14]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1649 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          1649 non-null   object
 1   Year                             1649 non-null   int64
 2   Status                           1649 non-null   object
 3   Life expectancy                  1649 non-null   float64
 4   Adult Mortality                  1649 non-null   float64
 5   infant deaths                    1649 non-null   int64
 6   Alcohol                          1649 non-null   float64
 7   percentage expenditure           1649 non-null   float64
 8   Hepatitis B                      1649 non-null   float64
 9   Measles                          1649 non-null   int64
 10   BMI                             1649 non-null   float64
 11  under-five deaths                1649 non-null   int64
 12  Polio                            1649 non-null   float64
 13  Total expenditure                1649 non-null   float64
 14  Diphtheria                       1649 non-null   float64
 15   HIV/AIDS                        1649 non-null   float64
 16  GDP                              1649 non-null   float64
 17  Population                       1649 non-null   float64
 18   thinness  1-19 years            1649 non-null   float64
 19   thinness 5-9 years              1649 non-null   float64
 20  Income composition of resources  1649 non-null   float64
 21  Schooling                        1649 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 296.3+ KB
```

In [15]:
```
df1 = df.dropna(axis = 1, how = "any")
df1
```

Out[15]:

	Country	Year	Status	infant deaths	percentage expenditure	Measles	under-five deaths	HIV/AIDS
0	Afghanistan	2015	Developing	62	71.279624	1154	83	0.1
1	Afghanistan	2014	Developing	64	73.523582	492	86	0.1
2	Afghanistan	2013	Developing	66	73.219243	430	89	0.1
3	Afghanistan	2012	Developing	69	78.184215	2787	93	0.1
4	Afghanistan	2011	Developing	71	7.097109	3013	97	0.1
...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	27	0.000000	31	42	33.6
2934	Zimbabwe	2003	Developing	26	0.000000	998	41	36.7
2935	Zimbabwe	2002	Developing	25	0.000000	304	40	39.8
2936	Zimbabwe	2001	Developing	25	0.000000	529	39	42.1
2937	Zimbabwe	2000	Developing	24	0.000000	1483	39	43.5

2938 rows × 8 columns

In [16]:
```
df1 = df.dropna(axis = 1, thresh = 2)
df1
```

Out[16]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepati
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	6
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	6
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	6
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	6
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	6
...	...	...	...	...	...	...	...	...	
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000	6
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000	
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000	7
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000	7
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000	7

2938 rows × 22 columns

--> Replacing null values with non-null values

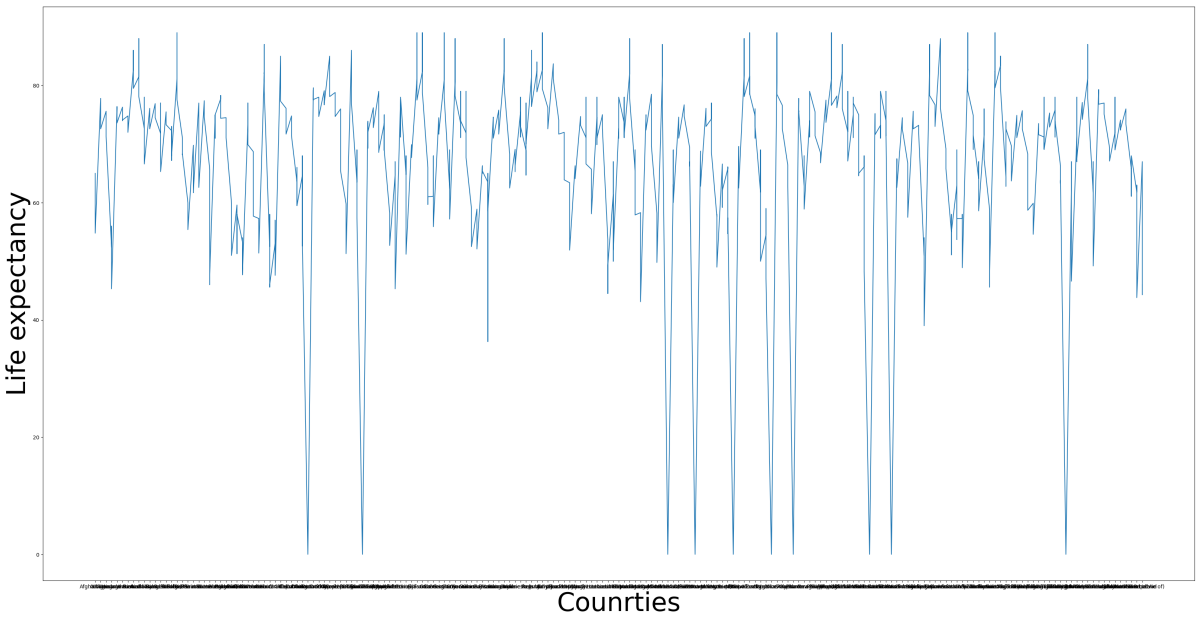
In [17]:
```
df1 = df.fillna(0)
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   Year                             2938 non-null   int64  
 2   Status                           2938 non-null   object 
 3   Life expectancy                  2938 non-null   float64
 4   Adult Mortality                  2938 non-null   float64
 5   infant deaths                    2938 non-null   int64  
 6   Alcohol                          2938 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2938 non-null   float64
 9   Measles                          2938 non-null   int64  
 10   BMI                             2938 non-null   float64
 11  under-five deaths                2938 non-null   int64  
 12  Polio                            2938 non-null   float64
 13  Total expenditure                2938 non-null   float64
 14  Diphtheria                       2938 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2938 non-null   float64
 17  Population                       2938 non-null   float64
 18   thinness  1-19 years            2938 non-null   float64
 19   thinness 5-9 years              2938 non-null   float64
 20  Income composition of resources  2938 non-null   float64
 21  Schooling                        2938 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

--> Visualization

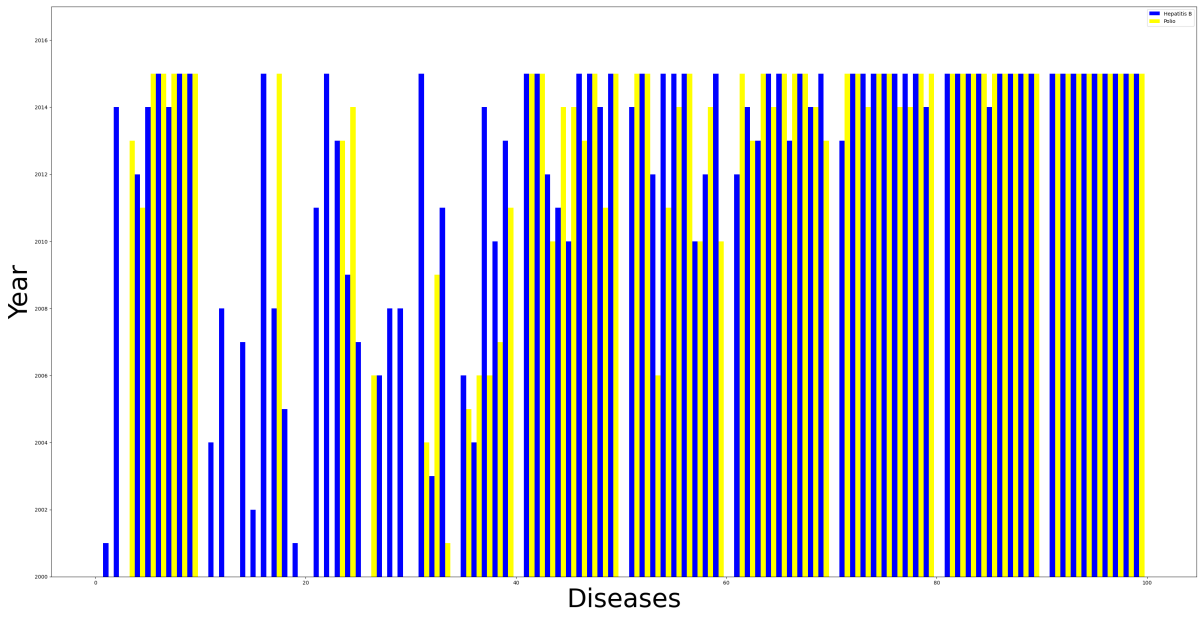
In [18]:
```
#LIFE EXPECTANCY WRT COUNTRY
x = df1['Country']
y = df1['Life expectancy ']

f = plt.figure()
f.set_figwidth(40)
f.set_figheight(20)

plt.xlabel('Counrties', fontsize = '50')
plt.ylabel('Life expectancy', fontsize = '50')

plt.plot(x,y)
```

Out[18]:
```
[<matplotlib.lines.Line2D at 0x1f1890be0b0>]
```



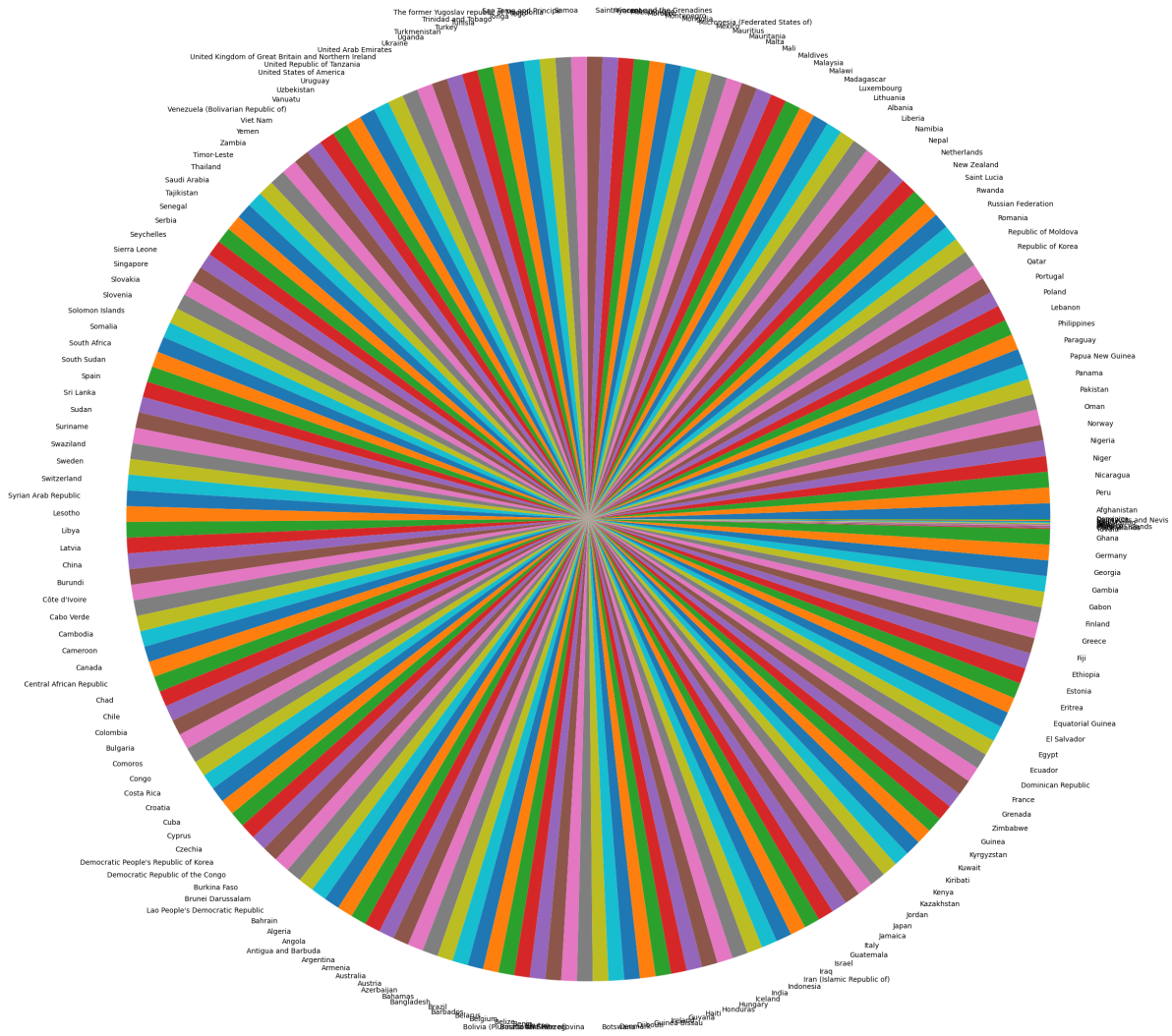
In [21]:
```
#DISEASES WRT YEARS
import warnings
warnings.filterwarnings('ignore')

x = df['Year']
y = df['Hepatitis B']
z = df['Polio']

f = plt.figure()
f.set_figwidth(40)
f.set_figheight(20)

plt.ylabel('Year', fontsize = '50')
plt.xlabel('Diseases', fontsize = '50')

plt.bar(y + 0.0, x, color = 'blue', width = 0.5)
plt.bar(z + 0.5, x, color = 'yellow', width = 0.5)

plt.ylim(2000,2017)
plt.legend(['Hepatitis B', 'Polio'])
```

Out[21]:
```
<matplotlib.legend.Legend at 0x1f193d57460>
```



In [22]:
```
#all the countries included in the dataset
fig = plt.figure(figsize = (30,30))
ax = fig.subplots()
df.Country.value_counts()[:200].plot(ax = ax, kind = 'pie')
ax.set_ylabel(" ")
plt.show
```

Out[22]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```


Afghanistan
Albania
Algeria
Angola
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belgium
Bolivia (Plurinational State of)
Botswana
Brazil
Brunei Darussalam
Bulgaria
Burkina Faso
Burundi
Cabo Verde
Cambodia
Cameroon
Canada
Central African Republic
Chad
Chile
China
Colombia
Comoros
Congo
Costa Rica
Côte d'Ivoire
Croatia
Cuba
Cyprus
Czechia
Democratic People's Republic of Korea
Democratic Republic of the Congo
Dominican Republic
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Fiji
Finland
France
Gabon
Gambia
Georgia
Germany
Ghana
Grenada
Guatemala
Guinea
Haiti
Honduras
Hungary
Iceland
India
Indonesia
Iran (Islamic Republic of)
Iraq
Israel
Italy
Jamaica
Japan
Jordan
Kazakhstan
Kenya
Kiribati
Kuwait
Kyrgyzstan
Lao People's Democratic Republic
Latvia
Lebanon
Lesotho
Libya
Lithuania
Luxembourg
Madagascar
Malawi
Malaysia
Maldives
Mali
Malta
Mauritania
Mauritius
Micronesia (Federated States of)
Nepal
Netherlands
New Zealand
Nicaragua
Niger
Nigeria
Norway
Oman
Pakistan
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Poland
Portugal
Qatar
Republic of Korea
Republic of Moldova
Romania
Russian Federation
Rwanda
Saint Lucia
Saint Vincent and the Grenadines
Samoa
Senegal
Serbia
Seychelles
Sierra Leone
Singapore
Slovakia
Slovenia
Solomon Islands
Somalia
South Africa
South Sudan
Spain
Sri Lanka
Sudan
Suriname
Swaziland
Sweden
Switzerland
Syrian Arab Republic
The former Yugoslav republic of Macedonia
Timor-Leste
Togo
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Uganda
Ukraine
United Arab Emirates
United Kingdom of Great Britain and Northern Ireland
United Republic of Tanzania
United States of America
Uruguay
Uzbekistan
Vanuatu
Venezuela (Bolivarian Republic of)
Viet Nam
Yemen
Zambia
Zimbabwe