Notebook7-checkpoint

May 22, 2020

[1]:
```
#Final research project Data-765
```

[2]:
```
#Importing couple of modules here
```

[30]:
```
import pandas as pd
import glob
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from scipy.stats import pearsonr, mannwhitneyu
import seaborn as sns
import statsmodels as sm
from scipy.stats import kendalltau
```

[31]:
```
Data_df = glob.glob("/home/haziz/Data/*.csv" ) #intialization
```

[32]:
```
Data_df
```

[32]:
```
['/home/haziz/Data/human-development-index-vs-corruption-perception-index.csv',
 '/home/haziz/Data/cancer-deaths-by-type-grouped.csv',
 '/home/haziz/Data/cancer-death-rates-by-age.csv',
 '/home/haziz/Data/death-rate-from-cancers-vs-average-income.csv',
 '/home/haziz/Data/cancer-incidence.csv',
 '/home/haziz/Data/share-of-cancer-deaths-attributed-to-risk-factors.csv']
```

[33]:
```
# I gathered all these data sets, so of them had a lot of missong data in them␣
↪so I had to drop them.
```

[34]:
```
#The data were pulled from https://ourworldindata.org/ and cancer atlas
#I will be using few data-sets for my analysis
```

Analysis on Neoplasm Incidence for Both the sex. A new and abnormal growth of tissue in some part of the body, especially as a characteristic of cancer. This data was pulled from https://ourworldindata.org/

New Cases per 100,000 for different countries around the globe. 1990 to 2017

[35]:
```
cancer_type_df = pd.read_csv('/home/haziz/Data/cancer-incidence.csv')
cancer_type_df
```

[35]:
```
           Entity Code  Year  \
0     Afghanistan  AFG  1990
1     Afghanistan  AFG  1991
2     Afghanistan  AFG  1992
3     Afghanistan  AFG  1993
4     Afghanistan  AFG  1994
…             …  …   …
6463     Zimbabwe  ZWE  2013
6464     Zimbabwe  ZWE  2014
6465     Zimbabwe  ZWE  2015
6466     Zimbabwe  ZWE  2016
6467     Zimbabwe  ZWE  2017

      Incidence - Neoplasms - Sex: Both - Age: Age-standardized (Rate) (new
cases per 100,000)
0                                                 169.700068
1                                                 169.607800
2                                                 169.857924
3                                                 170.170943
4                                                 172.203559
…                                                        …
6463                                              206.449353
6464                                              203.151575
6465                                              200.521929
6466                                              198.614810
6467                                              195.932929

[6468 rows x 4 columns]
```

[36]:
```
cancer_type_df.Entity.unique()
```

[36]:
```
array(['Afghanistan', 'Albania', 'Algeria', 'American Samoa',
       'Andean Latin America', 'Andorra', 'Angola', 'Antigua and Barbuda',
       'Argentina', 'Armenia', 'Australasia', 'Australia', 'Austria',
       'Azerbaijan', 'Bahamas', 'Bahrain', 'Bangladesh', 'Barbados',
       'Belarus', 'Belgium', 'Belize', 'Benin', 'Bermuda', 'Bhutan',
       'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil',
       'Brunei', 'Bulgaria', 'Burkina Faso', 'Burundi', 'Cambodia',
       'Cameroon', 'Canada', 'Cape Verde', 'Caribbean',
       'Central African Republic', 'Central Asia', 'Central Europe',
       'Central Europe, Eastern Europe, and Central Asia',
       'Central Latin America', 'Central Sub-Saharan Africa', 'Chad',
       'Chile', 'China', 'Colombia', 'Comoros', 'Congo', 'Costa Rica',
       "Cote d'Ivoire", 'Croatia', 'Cuba', 'Cyprus', 'Czech Republic',
```

```
       'Democratic Republic of Congo', 'Denmark', 'Djibouti', 'Dominica',
       'Dominican Republic', 'East Asia', 'Eastern Europe',
       'Eastern Sub-Saharan Africa', 'Ecuador', 'Egypt', 'El Salvador',
       'England', 'Equatorial Guinea', 'Eritrea', 'Estonia', 'Ethiopia',
       'Fiji', 'Finland', 'France', 'Gabon', 'Gambia', 'Georgia',
       'Germany', 'Ghana', 'Greece', 'Greenland', 'Grenada', 'Guam',
       'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti',
       'High SDI', 'High-income', 'High-income Asia Pacific',
       'High-middle SDI', 'Honduras', 'Hungary', 'Iceland', 'India',
       'Indonesia', 'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy',
       'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kenya', 'Kiribati',
       'Kuwait', 'Kyrgyzstan', 'Laos', 'Latin America and Caribbean',
       'Latvia', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Lithuania',
       'Low SDI', 'Low-middle SDI', 'Luxembourg', 'Macedonia',
       'Madagascar', 'Malawi', 'Malaysia', 'Maldives', 'Mali', 'Malta',
       'Marshall Islands', 'Mauritania', 'Mauritius', 'Mexico',
       'Micronesia (country)', 'Middle SDI', 'Moldova', 'Mongolia',
       'Montenegro', 'Morocco', 'Mozambique', 'Myanmar', 'Namibia',
       'Nepal', 'Netherlands', 'New Zealand', 'Nicaragua', 'Niger',
       'Nigeria', 'North Africa and Middle East', 'North America',
       'North Korea', 'Northern Ireland', 'Northern Mariana Islands',
       'Norway', 'Oceania', 'Oman', 'Pakistan', 'Palestine', 'Panama',
       'Papua New Guinea', 'Paraguay', 'Peru', 'Philippines', 'Poland',
       'Portugal', 'Puerto Rico', 'Qatar', 'Romania', 'Russia', 'Rwanda',
       'Saint Lucia', 'Saint Vincent and the Grenadines', 'Samoa',
       'Sao Tome and Principe', 'Saudi Arabia', 'Scotland', 'Senegal',
       'Serbia', 'Seychelles', 'Sierra Leone', 'Singapore', 'Slovakia',
       'Slovenia', 'Solomon Islands', 'Somalia', 'South Africa',
       'South Asia', 'South Korea', 'South Sudan', 'Southeast Asia',
       'Southeast Asia, East Asia, and Oceania', 'Southern Latin America',
       'Southern Sub-Saharan Africa', 'Spain', 'Sri Lanka',
       'Sub-Saharan Africa', 'Sudan', 'Suriname', 'Swaziland', 'Sweden',
       'Switzerland', 'Syria', 'Taiwan', 'Tajikistan', 'Tanzania',
       'Thailand', 'Timor', 'Togo', 'Tonga', 'Trinidad and Tobago',
       'Tropical Latin America', 'Tunisia', 'Turkey', 'Turkmenistan',
       'Uganda', 'Ukraine', 'United Arab Emirates', 'United Kingdom',
       'United States', 'United States Virgin Islands', 'Uruguay',
       'Uzbekistan', 'Vanuatu', 'Venezuela', 'Vietnam', 'Wales',
       'Western Europe', 'Western Sub-Saharan Africa', 'World', 'Yemen',
       'Zambia', 'Zimbabwe'], dtype=object)
```

[37]:
```
cancer_type_df
```

[37]:
```
          Entity Code  Year  \
0     Afghanistan  AFG  1990
1     Afghanistan  AFG  1991
2     Afghanistan  AFG  1992
```

```
3     Afghanistan  AFG  1993
4     Afghanistan  AFG  1994
…           …  …   …
6463     Zimbabwe  ZWE  2013
6464     Zimbabwe  ZWE  2014
6465     Zimbabwe  ZWE  2015
6466     Zimbabwe  ZWE  2016
6467     Zimbabwe  ZWE  2017

      Incidence - Neoplasms - Sex: Both - Age: Age-standardized (Rate) (new
cases per 100,000)
0                                                   169.700068
1                                                   169.607800
2                                                   169.857924
3                                                   170.170943
4                                                   172.203559
…                                                          …
6463                                                206.449353
6464                                                203.151575
6465                                                200.521929
6466                                                198.614810
6467                                                195.932929

[6468 rows x 4 columns]
```

[38]:
```
#checking for null values

cancer_type_df.isnull().sum()
```

[38]:
```
Entity
0
Code
980
Year
0
Incidence - Neoplasms - Sex: Both - Age: Age-standardized (Rate) (new cases per
100,000)      0
dtype: int64
```

[39]:
```
#Checking for length
len(cancer_type_df)
```

[39]:
```
6468
```

[40]:
```
cancer_type_df = cancer_type_df.dropna()
```

[41]:
```
cancer_type_df.isnull().sum()
```

[41]:
```
Entity
0
Code
0
Year
0
Incidence - Neoplasms - Sex: Both - Age: Age-standardized (Rate) (new cases per
100,000)    0
dtype: int64
```

top: Norway,Switzerland,Ireland,Germany, Australia,Iceland,United Kingdom,United States,Finland,Japan low: Pakistan, Yemen, Liberia, Guinea, Congo, Mozambique,Afghanistan,Zimbabwe,Syria,Iraq

[42]:
```
#cancer_type_df_new
cancer_type_df_new = cancer_type_df[(cancer_type_df['Year'] >=2002) &␣
↪(cancer_type_df['Incidence - Neoplasms - Sex: Both - Age: Age-standardized␣
↪(Rate) (new cases per 100,000)'] > 0)]
cancer_type_df_new
```

[42]:

	Entity	Code	Year	Incidence - Neoplasms - Sex: Both - Age: Age-standardized (Rate) (new cases per 100,000)
12	Afghanistan	AFG	2002	182.426026
13	Afghanistan	AFG	2003	182.696877
14	Afghanistan	AFG	2004	183.441469
15	Afghanistan	AFG	2005	182.421928
16	Afghanistan	AFG	2006	181.839963
…	…	…	…	…
6463	Zimbabwe	ZWE	2013	206.449353
6464	Zimbabwe	ZWE	2014	203.151575
6465	Zimbabwe	ZWE	2015	200.521929
6466	Zimbabwe	ZWE	2016	198.614810
6467	Zimbabwe	ZWE	2017	195.932929

```
[3136 rows x 4 columns]
```

[43]:
```
#Replacing the column name

cancer_type_df_new = cancer_type_df_new.rename(columns={"Incidence - Neoplasms␣
↪- Sex: Both - Age: Age-standardized (Rate) (new cases per 100,000)":
↪"Incidence_Neoplasms"})
```

[44]:
```
cancer_type_df_new
```

[44]:

	Entity	Code	Year	Incidence_Neoplasms
12	Afghanistan	AFG	2002	182.426026
13	Afghanistan	AFG	2003	182.696877
14	Afghanistan	AFG	2004	183.441469
15	Afghanistan	AFG	2005	182.421928
16	Afghanistan	AFG	2006	181.839963
…	…	…	…	…
6463	Zimbabwe	ZWE	2013	206.449353
6464	Zimbabwe	ZWE	2014	203.151575
6465	Zimbabwe	ZWE	2015	200.521929
6466	Zimbabwe	ZWE	2016	198.614810
6467	Zimbabwe	ZWE	2017	195.932929

```
[3136 rows x 4 columns]
```

[18]:
```
cancer_type_df_new.describe()
```

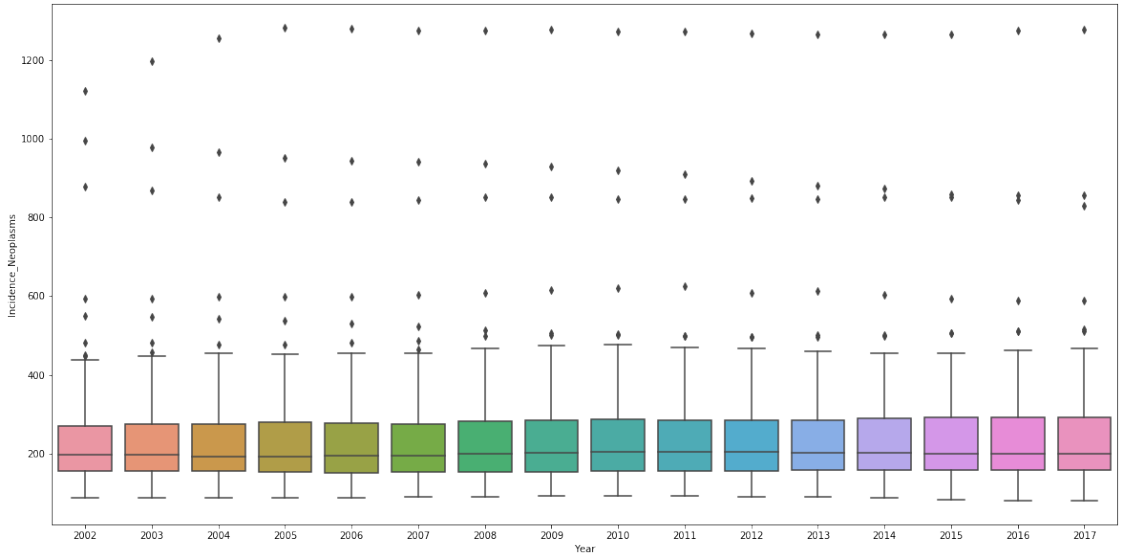
[18]:

	Year	Incidence_Neoplasms
count	3136.000000	3136.000000
mean	2009.500000	236.406334
std	4.610507	136.820072
min	2002.000000	79.712440
25%	2005.750000	155.296054
50%	2009.500000	198.616596
75%	2013.250000	283.005207
max	2017.000000	1282.921402

[19]:
```
import seaborn as sns
```

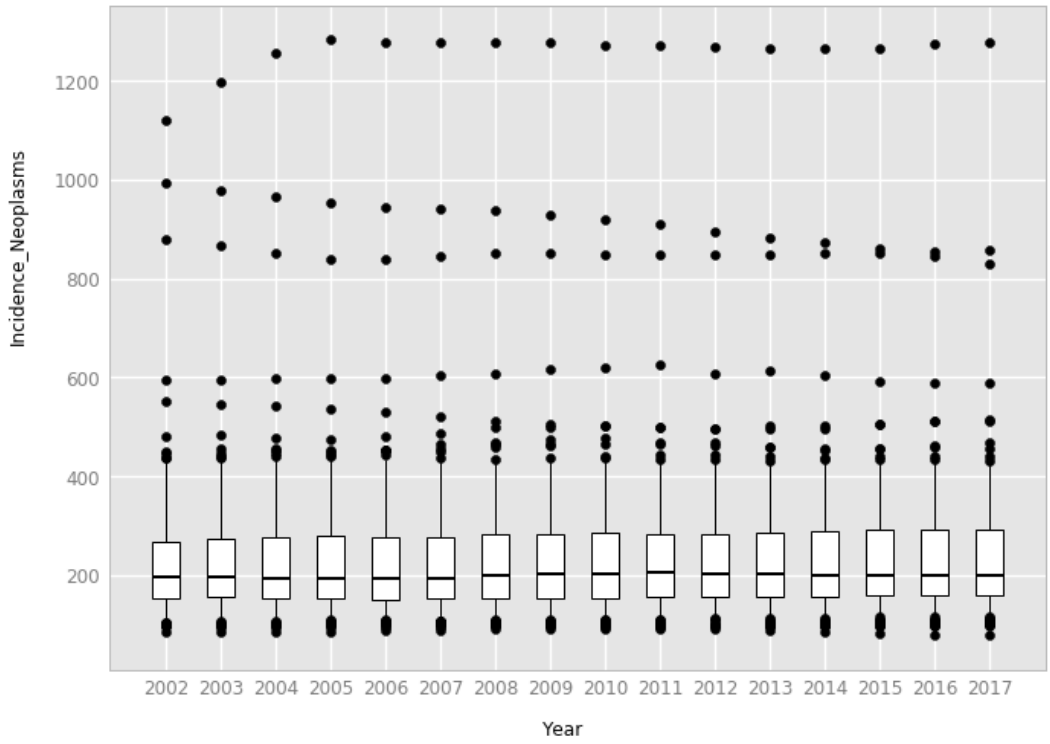
[20]:
```
plt.figure(figsize=(20,10))

sns.boxplot('Year', 'Incidence_Neoplasms',   data=cancer_type_df_new)
```

[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fca4c9577f0>
```



[21]:
```
from ggplot import *
```

[22]:
```
ggplot(cancer_type_df_new, aes(x='Year', y='Incidence_Neoplasms') ) + \
   geom_boxplot() + \
   theme(element_text(face = "bold", color = "black", size = 12))
```



[22]: <ggplot: (8781664016469)>

Here we can see that Incidence of Neoplasm (Age standardised for both sexes per 100,000 new cases) around the globe starting from 2002 to 2017. In the year 2002 there was a decline in number of cases i.e., the maximum number of new cases were 1000 and the minimum less than 200.

[23]:
```
#Picking the countries

cancer_type_df_new1 = cancer_type_df_new.query('Entity in ["Norway",␣
↪"Switzerland" , "Ireland" , "Germany" , "Australia" , "Iceland" , "United␣
↪Kingdom" , "United States" , "Finland" , "Japan", "Pakistan" , "Yemen" ,␣
↪"Liberia" , "Guinea", "Congo", "Mozambique" , "Afghanistan" , "Zimbabwe" ,␣
↪"Syria" , "Iraq"] ')
```

[24]:
```
cancer_type_df_new1
```

[24]:

	Entity	Code	Year	Incidence_Neoplasms
12	Afghanistan	AFG	2002	182.426026
13	Afghanistan	AFG	2003	182.696877
14	Afghanistan	AFG	2004	183.441469
15	Afghanistan	AFG	2005	182.421928
16	Afghanistan	AFG	2006	181.839963

```
…            …  …   …                    …
6463     Zimbabwe  ZWE  2013           206.449353
6464     Zimbabwe  ZWE  2014           203.151575
6465     Zimbabwe  ZWE  2015           200.521929
6466     Zimbabwe  ZWE  2016           198.614810
6467     Zimbabwe  ZWE  2017           195.932929

[320 rows x 4 columns]
```

[25]:
```
cancer_type_df_new1.max()
```

[25]:
```
Entity                 Zimbabwe
Code                        ZWE
Year                       2017
Incidence_Neoplasms     1282.92
dtype: object
```

[26]:
```
cancer_type_df_new1.min()
```

[26]:
```
Entity                 Afghanistan
Code                           AFG
Year                          2002
Incidence_Neoplasms        79.7124
dtype: object
```

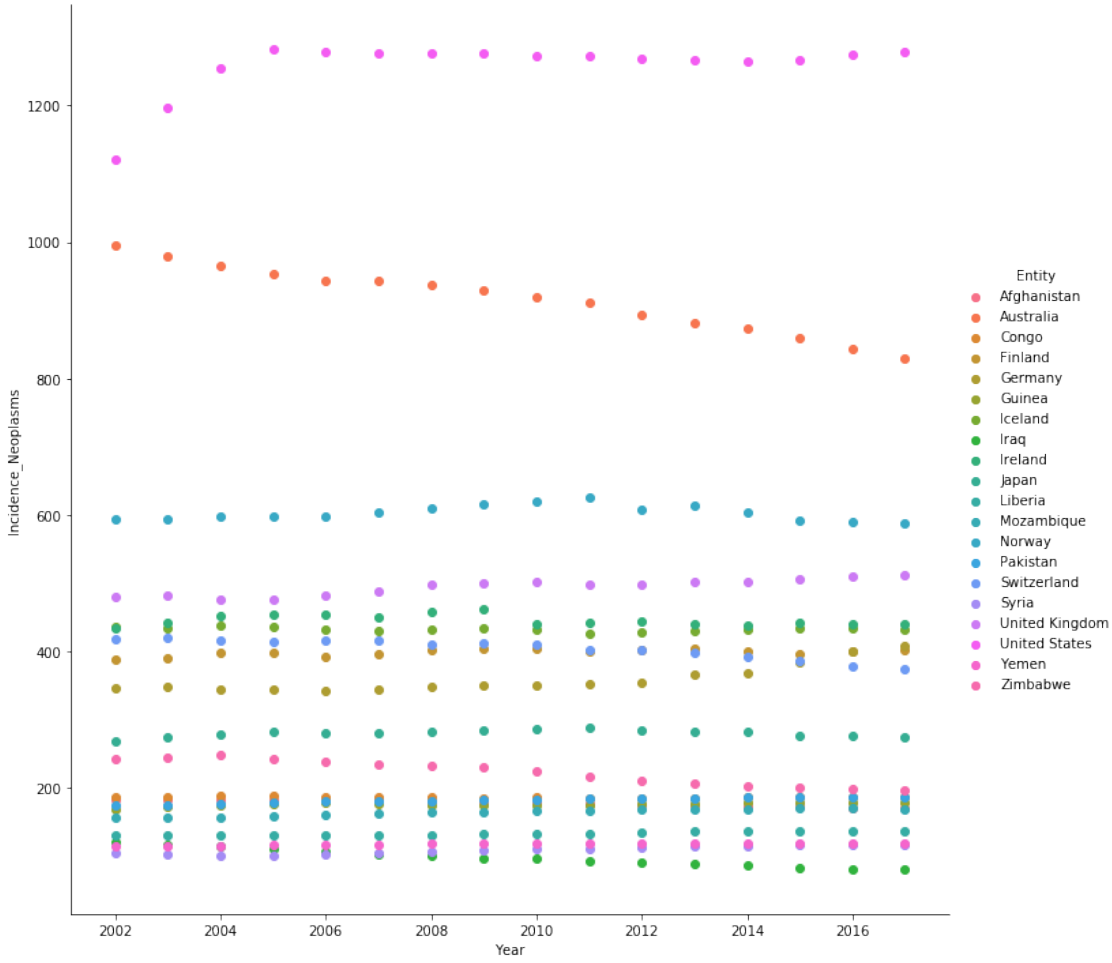
[27]:
```
cancer_type_df_new1.describe()
```

[27]:
```
              Year  Incidence_Neoplasms
count   320.000000           320.000000
mean   2009.500000           357.955722
std       4.616992           287.650895
min    2002.000000            79.712440
25%    2005.750000           169.739920
50%    2009.500000           258.523351
75%    2013.250000           436.955548
max    2017.000000          1282.921402
```

[28]:
```
#I picked 10 high and 10 low HDI countries based on UN data

sns.FacetGrid(cancer_type_df_new1, hue="Entity", size=10) \
   .map(plt.scatter, "Year", "Incidence_Neoplasms"  ) \
.add_legend()
```

```
/home/haziq/.local/lib/python3.6/site-packages/seaborn/axisgrid.py:243:
UserWarning: The `size` parameter has been renamed to `height`; please update
your code.
  warnings.warn(msg, UserWarning)
```

[28]: <seaborn.axisgrid.FacetGrid at 0x7fca3f56acf8>



[29]:
```
import plotly.express as px
fig = px.scatter(cancer_type_df_new1, x="Year", y="Incidence_Neoplasms",␣
 ↪color="Entity", color_discrete_sequence=["red", "green", "blue", "goldenrod",␣
 ↪"magenta", "cyan", "crimson", "darkblue", "darkcyan", "black", "coral",␣
 ↪"rosybrown", "purple", "chocolate", "cornsilk", "chartreuse", "turquoise",␣
 ↪"darkmagenta", "plum", "yellow"], title="Neoplasms Incidence from 2002-2017␣
 ↪per new 100,000 cases based on 20 countries ")
fig.update_traces(marker=dict(size=10,
                              line=dict(width=1,
                                        color='black')),
                  selector=dict(mode='markers'))
fig.show()
```

[ ]:

[140]:
```
#GDP
```

[45]:
```
cancer_type_df_2 = pd.read_csv('/home/haziz/Data/
↪death-rate-from-cancers-vs-average-income.csv')
```

[46]:
```
cancer_type_df_2
```

[46]:

	Entity	Code	Year	Cancers - age-standardized death rate (deaths per 100,000 individuals)	GDP per capita, PPP (constant 2011 international $) (constant 2011 international $)
0	Afghanistan	AFG	1800	NaN	NaN
1	Afghanistan	AFG	1820	NaN	NaN
2	Afghanistan	AFG	1870	NaN	NaN
3	Afghanistan	AFG	1913	NaN	NaN
4	Afghanistan	AFG	1950	NaN	NaN
…	…	…	…	…	…
22914	Zimbabwe	ZWE	2013	242.78	1929.765001
22915	Zimbabwe	ZWE	2014	245.74	1925.138698
22916	Zimbabwe	ZWE	2015	245.57	1912.280261
22917	Zimbabwe	ZWE	2016	245.77	1879.628119
22918	Zimbabwe	ZWE	2017	NaN	1899.774977

```
       Total population (Gapminder)
0                         3280000.0
1                         3280000.0
2                         4207000.0
3                         5730000.0
4                         8151455.0
…                               …
22914                    13327925.0
22915                           NaN
22916                           NaN
22917                           NaN
22918                           NaN

[22919 rows x 6 columns]
```

[47]:
```
cancer_type_df_2.isnull().sum()
```

[47]:
```
Entity
0
Code
1888
Year
0
Cancers - age-standardized death rate (deaths per 100,000 individuals)
16898
GDP per capita, PPP (constant 2011 international $) (constant 2011 international
$)    16512
Total population (Gapminder)
2845
dtype: int64
```

[48]:
```
cancer_type_df_2 = cancer_type_df_2.dropna()
```

[49]:
```
cancer_type_df_2.isnull().sum()
```

[49]:
```
Entity
0
Code
0
Year
0
Cancers - age-standardized death rate (deaths per 100,000 individuals)
0
GDP per capita, PPP (constant 2011 international $) (constant 2011 international
$)    0
Total population (Gapminder)
0
```

```
dtype: int64
```

[50]:
```
#columnn name chage
cancer_type_df_GDP = cancer_type_df_2.rename(columns={"GDP per capita, PPP␣
↪(constant 2011 international $) (constant 2011 international $)" : "GDP",␣
↪"Cancers - age-standardized death rate (deaths per 100,000 individuals)":
↪"Cancer_Deaths"})
```

[51]:
```
cancer_type_df_GDP.head()
```

[51]:
```
          Entity Code  Year  Cancer_Deaths          GDP  \
56  Afghanistan  AFG  2002         163.79  1063.635574
57  Afghanistan  AFG  2003         164.58  1099.194507
58  Afghanistan  AFG  2004         165.18  1062.249360
59  Afghanistan  AFG  2005         165.25  1136.123214
60  Afghanistan  AFG  2006         165.14  1161.124889

    Total population (Gapminder)
56                     24639841.0
57                     25678639.0
58                     26693486.0
59                     27614718.0
60                     28420974.0
```

[52]:
```
from scipy.stats import pearsonr, mannwhitneyu
import seaborn as sns
import statsmodels as sm
from scipy.stats import kendalltau
import numpy as np
```

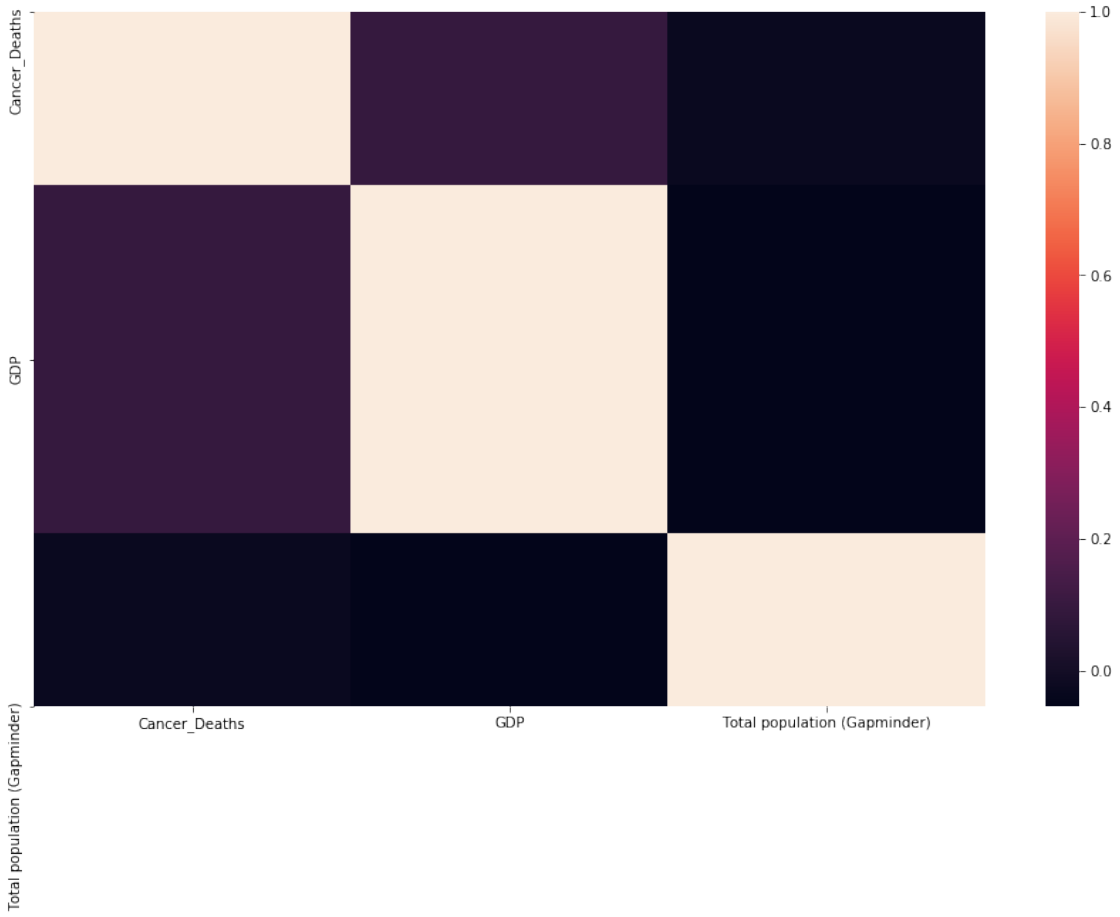
[53]:
```
corrmatrix = cancer_type_df_GDP[['Cancer_Deaths','GDP','Total population␣
↪(Gapminder)']].corr()
```

[54]:
```
corrmatrix
```

[54]:
```
                               Cancer_Deaths       GDP  \
Cancer_Deaths                        1.00000  0.094900
GDP                                  0.09490  1.000000
Total population (Gapminder)        -0.02961 -0.055501

                               Total population (Gapminder)
Cancer_Deaths                                     -0.029610
GDP                                               -0.055501
Total population (Gapminder)                       1.000000
```

[55]:
```
f,ax = plt.subplots(figsize=(15, 9))
sns.heatmap(corrmatrix)
```

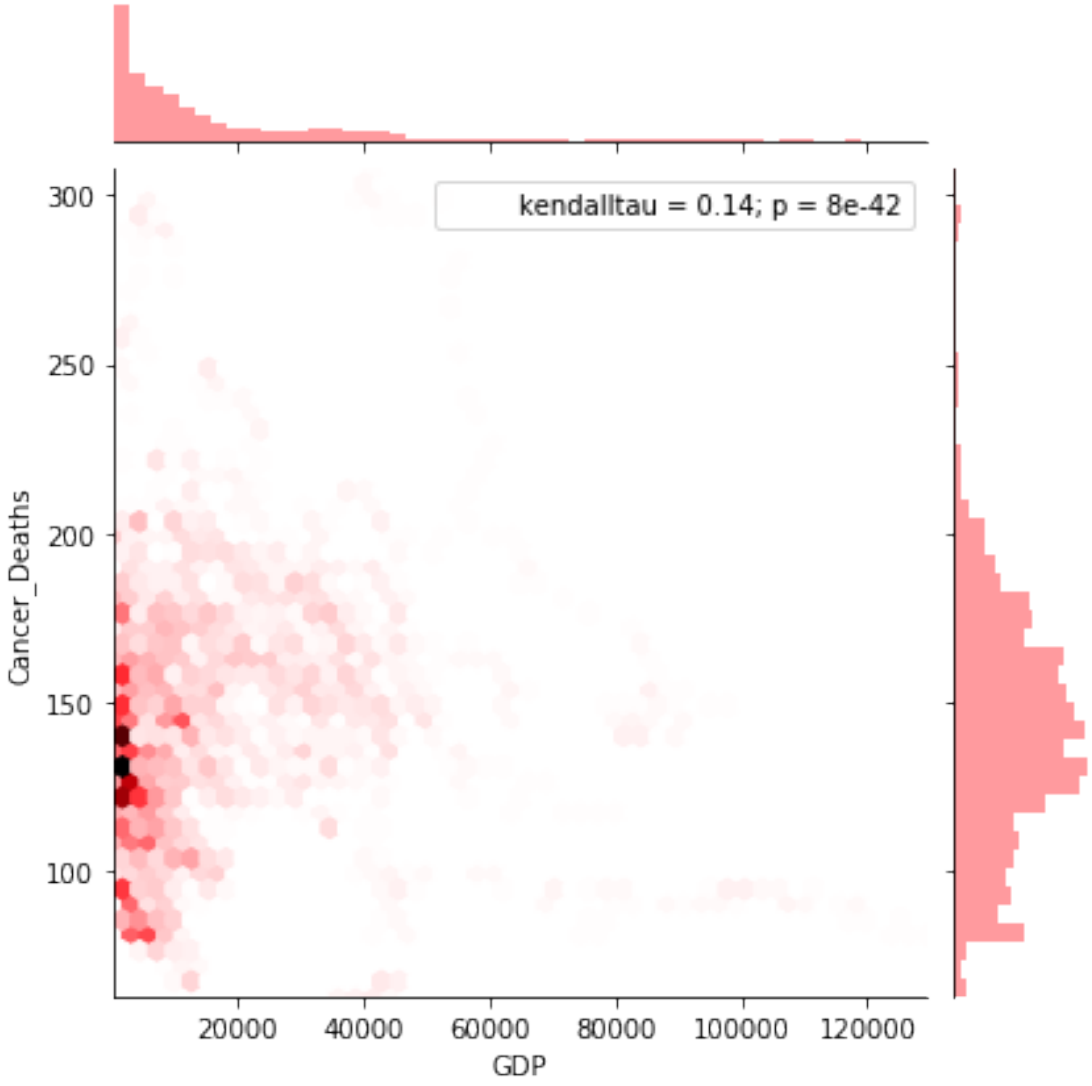
[55]: <matplotlib.axes._subplots.AxesSubplot at 0x7fca3704f898>



[56]:
```
sns.jointplot(cancer_type_df_GDP['GDP'], cancer_type_df_GDP['Cancer_Deaths'] ,

              kind="hex", stat_func=kendalltau, color="#FF030D")
```

```
/home/haziz/.local/lib/python3.6/site-packages/seaborn/axisgrid.py:1848:
UserWarning:

JointGrid annotation is deprecated and will be removed in a future release.
```

[56]: <seaborn.axisgrid.JointGrid at 0x7fca357dab38>



[ ]:

[ ]:

[39]:
```
#HDI
```

[57]:
```
cancer_type_df_6 = pd.read_csv('/home/haziz/Data/
 ↪human-development-index-vs-corruption-perception-index.csv')
cancer_type_df_6
```

[57]:
```
            Entity Code  Year  \
0      Afghanistan  AFG  1800
1      Afghanistan  AFG  1801
```

```
2      Afghanistan  AFG  1802
3      Afghanistan  AFG  1803
4      Afghanistan  AFG  1804
…            …  …   …
43388     Zimbabwe  ZWE  2015
43389     Zimbabwe  ZWE  2016
43390     Zimbabwe  ZWE  2017
43391     Zimbabwe  ZWE  2018
43392     Zimbabwe  ZWE  2019

       Human Development Index ((0-1; higher values are better))  \
0                                                    NaN
1                                                    NaN
2                                                    NaN
3                                                    NaN
4                                                    NaN
…                                                    …
43388                                              0.529
43389                                              0.532
43390                                              0.535
43391                                                NaN
43392                                                NaN

       Population by country
0                  3280000.0
1                  3280000.0
2                  3280000.0
3                  3280000.0
4                  3280000.0
…                        …
43388             15777451.0
43389             16150362.0
43390             16529904.0
43391             16913261.0
43392             17297495.0

[43393 rows x 5 columns]
```

[58]:
```
cancer_type_df_6.isnull().sum()
```

[58]:
```
Entity                                                    0
Code                                                    248
Year                                                      0
Human Development Index ((0-1; higher values are better))    38392
Population by country                                    53
dtype: int64
```

[59]:
```
cancer_type_df_6 = cancer_type_df_6.dropna()
```

[60]:
```
cancer_type_df_6.isnull().sum()
```

[60]:
```
Entity                                                  0
Code                                                    0
Year                                                    0
Human Development Index ((0-1; higher values are better))    0
Population by country                                   0
dtype: int64
```

[61]:
```
cancer_type_df_6
```

[61]:

	Entity	Code	Year	Human Development Index ((0-1; higher values are better))	Population by country
180	Afghanistan	AFG	1980	0.228	13248370.0
185	Afghanistan	AFG	1985	0.273	11783050.0
202	Afghanistan	AFG	2002	0.373	21979923.0
203	Afghanistan	AFG	2003	0.383	23064851.0
204	Afghanistan	AFG	2004	0.398	24118979.0
…	…	…	…	…	…
43386	Zimbabwe	ZWE	2013	0.516	15054506.0
43387	Zimbabwe	ZWE	2014	0.525	
43388	Zimbabwe	ZWE	2015	0.529	
43389	Zimbabwe	ZWE	2016	0.532	
43390	Zimbabwe	ZWE	2017	0.535	

```
43387             15411675.0
43388             15777451.0
43389             16150362.0
43390             16529904.0

[4955 rows x 5 columns]
```

[62]:
```
cancer_type_df_6_a = cancer_type_df_6[(cancer_type_df_6['Year'] >=2002)]
```

[63]:
```
cancer_type_df_6_a
```

[63]:
```
            Entity Code  Year  \
202    Afghanistan  AFG  2002
203    Afghanistan  AFG  2003
204    Afghanistan  AFG  2004
205    Afghanistan  AFG  2005
206    Afghanistan  AFG  2006
…              …  …   …
43386     Zimbabwe  ZWE  2013
43387     Zimbabwe  ZWE  2014
43388     Zimbabwe  ZWE  2015
43389     Zimbabwe  ZWE  2016
43390     Zimbabwe  ZWE  2017

       Human Development Index ((0-1; higher values are better))  \
202                                                    0.373
203                                                    0.383
204                                                    0.398
205                                                    0.408
206                                                    0.417
…                                                      …
43386                                                  0.516
43387                                                  0.525
43388                                                  0.529
43389                                                  0.532
43390                                                  0.535

       Population by country
202              21979923.0
203              23064851.0
204              24118979.0
205              25070798.0
206              25893450.0
…                      …
43386            15054506.0
43387            15411675.0
43388            15777451.0
```

```
43389             16150362.0
43390             16529904.0

[2948 rows x 5 columns]
```

[64]:
```
cancer_type_df_6_a = cancer_type_df_6[(cancer_type_df_6['Year'] >=2002)␣
↪|(cancer_type_df_6['Year'] >=2017)]
```

[65]:
```
cancer_type_df_6_a
```

[65]:
```
            Entity Code  Year  \
202    Afghanistan  AFG  2002
203    Afghanistan  AFG  2003
204    Afghanistan  AFG  2004
205    Afghanistan  AFG  2005
206    Afghanistan  AFG  2006
…              …  …   …
43386     Zimbabwe  ZWE  2013
43387     Zimbabwe  ZWE  2014
43388     Zimbabwe  ZWE  2015
43389     Zimbabwe  ZWE  2016
43390     Zimbabwe  ZWE  2017

       Human Development Index ((0-1; higher values are better))  \
202                                                    0.373
203                                                    0.383
204                                                    0.398
205                                                    0.408
206                                                    0.417
…                                                      …
43386                                                  0.516
43387                                                  0.525
43388                                                  0.529
43389                                                  0.532
43390                                                  0.535

       Population by country
202               21979923.0
203               23064851.0
204               24118979.0
205               25070798.0
206               25893450.0
…                       …
43386             15054506.0
43387             15411675.0
43388             15777451.0
43389             16150362.0
```

```
43390             16529904.0

[2948 rows x 5 columns]
```

[66]:
```
cancer_type_df_6_a = cancer_type_df_6_a.rename(columns={"Human Development␣
↪Index ((0-1; higher values are better))":"HDI", "Population by country":
↪"Population" })
```

[67]:
```
cancer_type_df_6_a
```

[67]:
```
            Entity Code  Year    HDI  Population
202    Afghanistan  AFG  2002  0.373  21979923.0
203    Afghanistan  AFG  2003  0.383  23064851.0
204    Afghanistan  AFG  2004  0.398  24118979.0
205    Afghanistan  AFG  2005  0.408  25070798.0
206    Afghanistan  AFG  2006  0.417  25893450.0
…              …  …   …     …          …
43386     Zimbabwe  ZWE  2013  0.516  15054506.0
43387     Zimbabwe  ZWE  2014  0.525  15411675.0
43388     Zimbabwe  ZWE  2015  0.529  15777451.0
43389     Zimbabwe  ZWE  2016  0.532  16150362.0
43390     Zimbabwe  ZWE  2017  0.535  16529904.0

[2948 rows x 5 columns]
```

[ ]:
```

```

[84]:
```
#Merge
```

[68]:
```
result = pd.merge(cancer_type_df_GDP,
                  cancer_type_df_6_a[['Entity',  'HDI', 'Population']],
                  on='Entity')
```

[69]:
```
result
```

[69]:
```
            Entity Code  Year  Cancer_Deaths          GDP  \
0      Afghanistan  AFG  2002         163.79  1063.635574
1      Afghanistan  AFG  2002         163.79  1063.635574
2      Afghanistan  AFG  2002         163.79  1063.635574
3      Afghanistan  AFG  2002         163.79  1063.635574
4      Afghanistan  AFG  2002         163.79  1063.635574
…              …  …   …              …           …
64590     Zimbabwe  ZWE  2013         242.78  1929.765001
64591     Zimbabwe  ZWE  2013         242.78  1929.765001
64592     Zimbabwe  ZWE  2013         242.78  1929.765001
64593     Zimbabwe  ZWE  2013         242.78  1929.765001
64594     Zimbabwe  ZWE  2013         242.78  1929.765001
```

```
       Total population (Gapminder)    HDI  Population
0                          24639841.0  0.373  21979923.0
1                          24639841.0  0.383  23064851.0
2                          24639841.0  0.398  24118979.0
3                          24639841.0  0.408  25070798.0
4                          24639841.0  0.417  25893450.0
…                                  …    …          …
64590                      13327925.0  0.516  15054506.0
64591                      13327925.0  0.525  15411675.0
64592                      13327925.0  0.529  15777451.0
64593                      13327925.0  0.532  16150362.0
64594                      13327925.0  0.535  16529904.0

[64595 rows x 8 columns]
```

[70]:
```
result.isnull().sum()
```

[70]:
```
Entity                          0
Code                            0
Year                            0
Cancer_Deaths                   0
GDP                             0
Total population (Gapminder)    0
HDI                             0
Population                      0
dtype: int64
```

[71]:
```
result1 = pd.merge(result,
                   cancer_type_df_new[['Entity', 'Incidence_Neoplasms']],
                   on='Entity'
                   )
```

[72]:
```
result1
```

[72]:
```
              Entity Code  Year  Cancer_Deaths          GDP  \
0        Afghanistan  AFG  2002         163.79  1063.635574
1        Afghanistan  AFG  2002         163.79  1063.635574
2        Afghanistan  AFG  2002         163.79  1063.635574
3        Afghanistan  AFG  2002         163.79  1063.635574
4        Afghanistan  AFG  2002         163.79  1063.635574
…                …  …    …              …           …
1033515     Zimbabwe  ZWE  2013         242.78  1929.765001
1033516     Zimbabwe  ZWE  2013         242.78  1929.765001
1033517     Zimbabwe  ZWE  2013         242.78  1929.765001
1033518     Zimbabwe  ZWE  2013         242.78  1929.765001
1033519     Zimbabwe  ZWE  2013         242.78  1929.765001
```

```
         Total population (Gapminder)    HDI  Population  Incidence_Neoplasms
0                           24639841.0  0.373  21979923.0           182.426026
1                           24639841.0  0.373  21979923.0           182.696877
2                           24639841.0  0.373  21979923.0           183.441469
3                           24639841.0  0.373  21979923.0           182.421928
4                           24639841.0  0.373  21979923.0           181.839963
…                                  …    …          …                    …
1033515                     13327925.0  0.535  16529904.0           206.449353
1033516                     13327925.0  0.535  16529904.0           203.151575
1033517                     13327925.0  0.535  16529904.0           200.521929
1033518                     13327925.0  0.535  16529904.0           198.614810
1033519                     13327925.0  0.535  16529904.0           195.932929

[1033520 rows x 9 columns]
```

[90]:
```
result1.isnull().sum()
```

[90]:
```
Entity                              0
Code                               32
Year                               32
Cancer_Deaths                      32
GDP                                32
Total population (Gapminder)       32
HDI                                80
Population                         80
Incidence_Neoplasms             58019
_merge                              0
dtype: int64
```

[73]:
```
result1 = result1.dropna()
```

[74]:
```
result1
```

[74]:
```
             Entity Code  Year  Cancer_Deaths          GDP  \
0       Afghanistan  AFG  2002         163.79  1063.635574
1       Afghanistan  AFG  2002         163.79  1063.635574
2       Afghanistan  AFG  2002         163.79  1063.635574
3       Afghanistan  AFG  2002         163.79  1063.635574
4       Afghanistan  AFG  2002         163.79  1063.635574
…               …  …    …              …          …
1033515    Zimbabwe  ZWE  2013         242.78  1929.765001
1033516    Zimbabwe  ZWE  2013         242.78  1929.765001
1033517    Zimbabwe  ZWE  2013         242.78  1929.765001
1033518    Zimbabwe  ZWE  2013         242.78  1929.765001
1033519    Zimbabwe  ZWE  2013         242.78  1929.765001
```

```
         Total population (Gapminder)    HDI  Population  Incidence_Neoplasms
0                          24639841.0  0.373  21979923.0           182.426026
1                          24639841.0  0.373  21979923.0           182.696877
2                          24639841.0  0.373  21979923.0           183.441469
3                          24639841.0  0.373  21979923.0           182.421928
4                          24639841.0  0.373  21979923.0           181.839963
…                                  …    …          …                    …
1033515                    13327925.0  0.535  16529904.0           206.449353
1033516                    13327925.0  0.535  16529904.0           203.151575
1033517                    13327925.0  0.535  16529904.0           200.521929
1033518                    13327925.0  0.535  16529904.0           198.614810
1033519                    13327925.0  0.535  16529904.0           195.932929

[1033520 rows x 9 columns]
```

[79]:
```
#here the subset value will keep all the years 2002 to 2017 for the countries,␣
↪while the drop duplicate will drop duplicate rows
result1 = result1.drop_duplicates(subset =("Year", "Entity"))
```

[80]:
```
result1
```

[80]:
```
             Entity Code  Year  Cancer_Deaths          GDP  \
0       Afghanistan  AFG  2002         163.79  1063.635574
256     Afghanistan  AFG  2003         164.58  1099.194507
512     Afghanistan  AFG  2004         165.18  1062.249360
768     Afghanistan  AFG  2005         165.25  1136.123214
1024    Afghanistan  AFG  2006         165.14  1161.124889
…               …  …    …              …           …
1032240    Zimbabwe  ZWE  2009         256.74  1336.212644
1032496    Zimbabwe  ZWE  2010         253.09  1474.877128
1032752    Zimbabwe  ZWE  2011         248.43  1667.137943
1033008    Zimbabwe  ZWE  2012         244.94  1871.366340
1033264    Zimbabwe  ZWE  2013         242.78  1929.765001

         Total population (Gapminder)    HDI  Population  Incidence_Neoplasms
0                          24639841.0  0.373  21979923.0           182.426026
256                        25678639.0  0.373  21979923.0           182.426026
512                        26693486.0  0.373  21979923.0           182.426026
768                        27614718.0  0.373  21979923.0           182.426026
1024                       28420974.0  0.373  21979923.0           182.426026
…                                  …    …          …                    …
1032240                    12473992.0  0.435  12500525.0           242.665680
1032496                    12571454.0  0.435  12500525.0           242.665680
1032752                    12754378.0  0.435  12500525.0           242.665680
1033008                    13013678.0  0.435  12500525.0           242.665680
1033264                    13327925.0  0.435  12500525.0           242.665680
```

```
[4108 rows x 9 columns]
```

[82]:
```
result1
```

[82]:
```
              Entity Code  Year  Cancer_Deaths          GDP  \
0        Afghanistan  AFG  2002         163.79  1063.635574
256      Afghanistan  AFG  2003         164.58  1099.194507
512      Afghanistan  AFG  2004         165.18  1062.249360
768      Afghanistan  AFG  2005         165.25  1136.123214
1024     Afghanistan  AFG  2006         165.14  1161.124889
…                …   …   …              …           …
1032240     Zimbabwe  ZWE  2009         256.74  1336.212644
1032496     Zimbabwe  ZWE  2010         253.09  1474.877128
1032752     Zimbabwe  ZWE  2011         248.43  1667.137943
1033008     Zimbabwe  ZWE  2012         244.94  1871.366340
1033264     Zimbabwe  ZWE  2013         242.78  1929.765001

         Total population (Gapminder)    HDI  Population  Incidence_Neoplasms
0                          24639841.0  0.373  21979923.0           182.426026
256                        25678639.0  0.373  21979923.0           182.426026
512                        26693486.0  0.373  21979923.0           182.426026
768                        27614718.0  0.373  21979923.0           182.426026
1024                       28420974.0  0.373  21979923.0           182.426026
…                                 …     …           …                    …
1032240                    12473992.0  0.435  12500525.0           242.665680
1032496                    12571454.0  0.435  12500525.0           242.665680
1032752                    12754378.0  0.435  12500525.0           242.665680
1033008                    13013678.0  0.435  12500525.0           242.665680
1033264                    13327925.0  0.435  12500525.0           242.665680

[4108 rows x 9 columns]
```

[256]:
```
#After removing null values from both the merged tables the out come data is␣
↪available from year 2002 to 2017
```

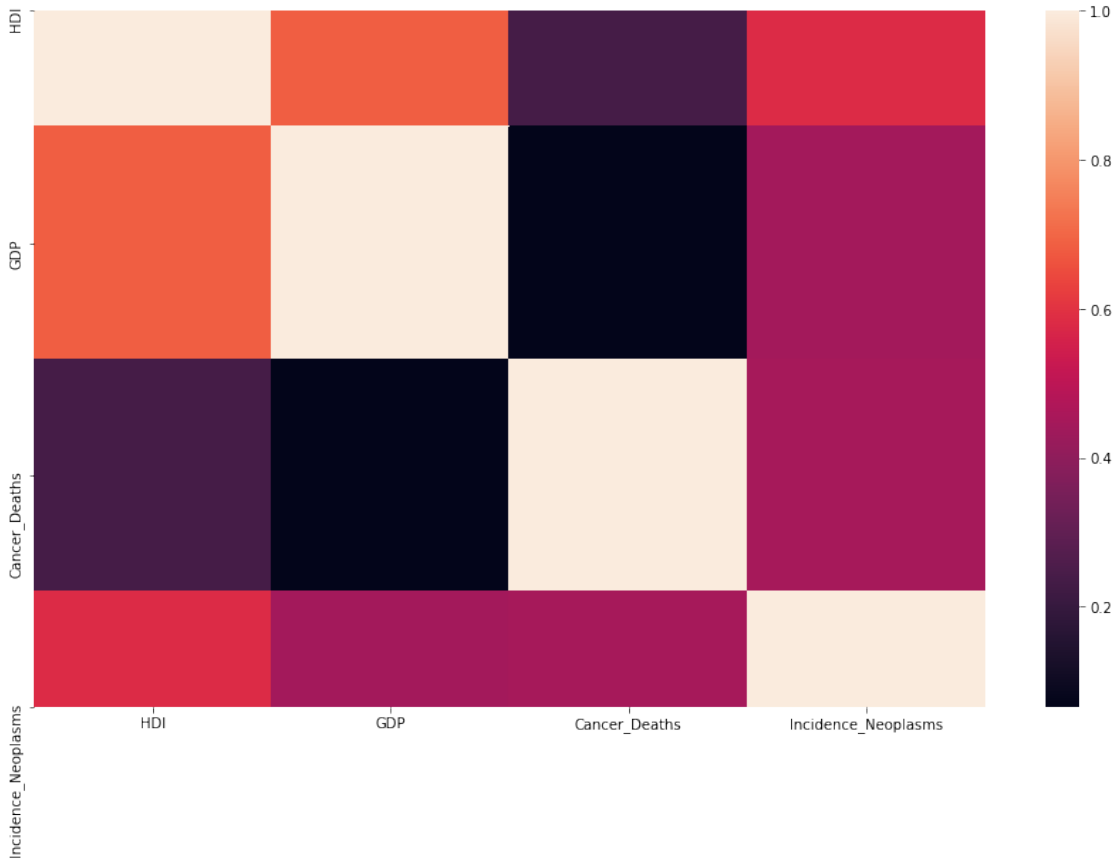
[83]:
```
corrmatrix = result1[['HDI','GDP','Cancer_Deaths','Incidence_Neoplasms']].corr()
```

[84]:
```
corrmatrix
```

[84]:
```
                          HDI       GDP  Cancer_Deaths  Incidence_Neoplasms
HDI                  1.000000  0.684399       0.234835              0.58190
GDP                  0.684399  1.000000       0.065559              0.44458
Cancer_Deaths        0.234835  0.065559       1.000000              0.44956
Incidence_Neoplasms  0.581900  0.444580       0.449560              1.00000
```

[85]:
```
f,ax = plt.subplots(figsize=(15, 9))
sns.heatmap(corrmatrix)
```

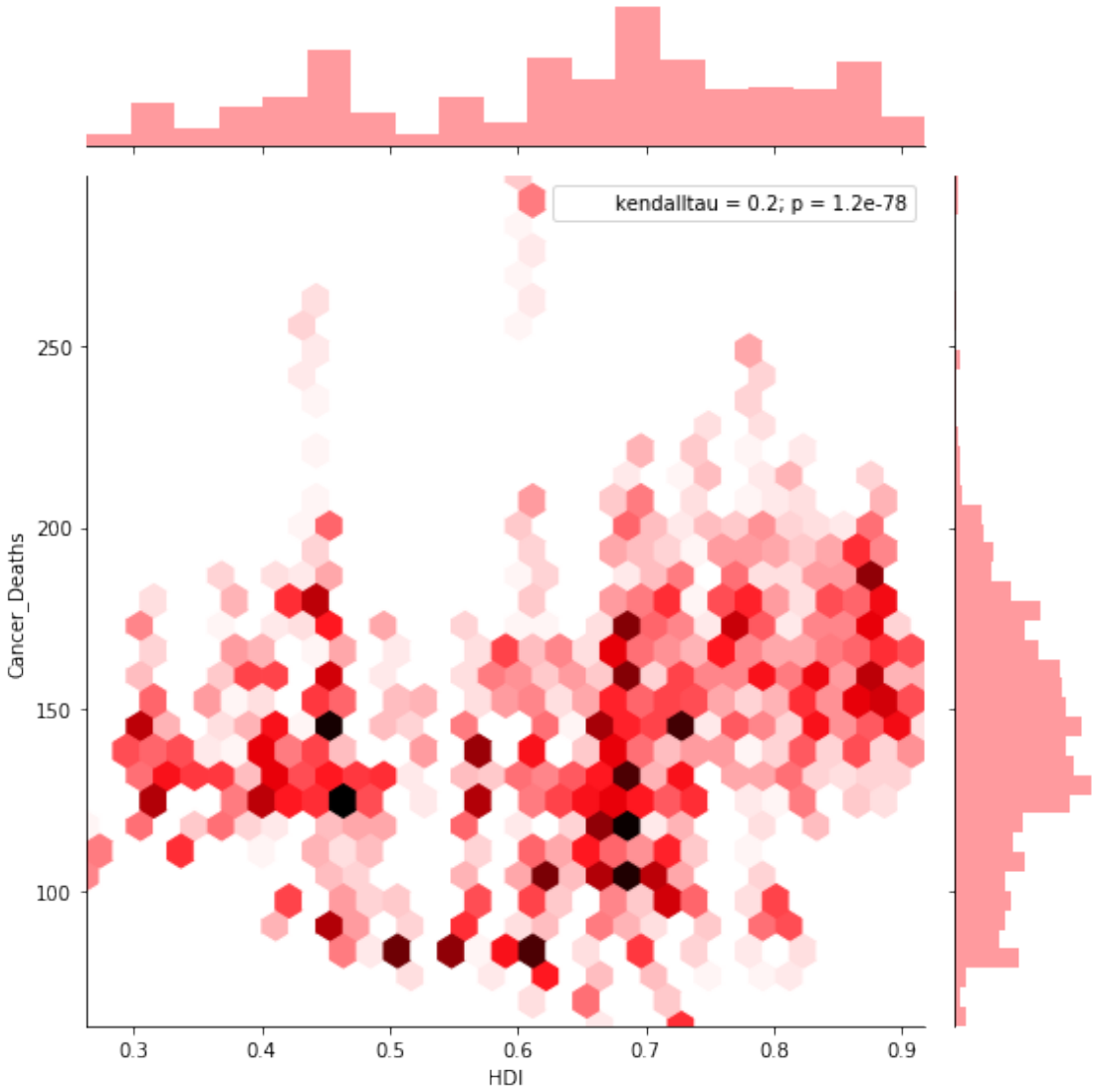
[85]: <matplotlib.axes._subplots.AxesSubplot at 0x7fca34541080>



[86]:
```
sns.jointplot(result1['HDI'], result1['Cancer_Deaths'] ,
              kind="hex", stat_func=kendalltau, color="#FF030D", size = 8 )
```

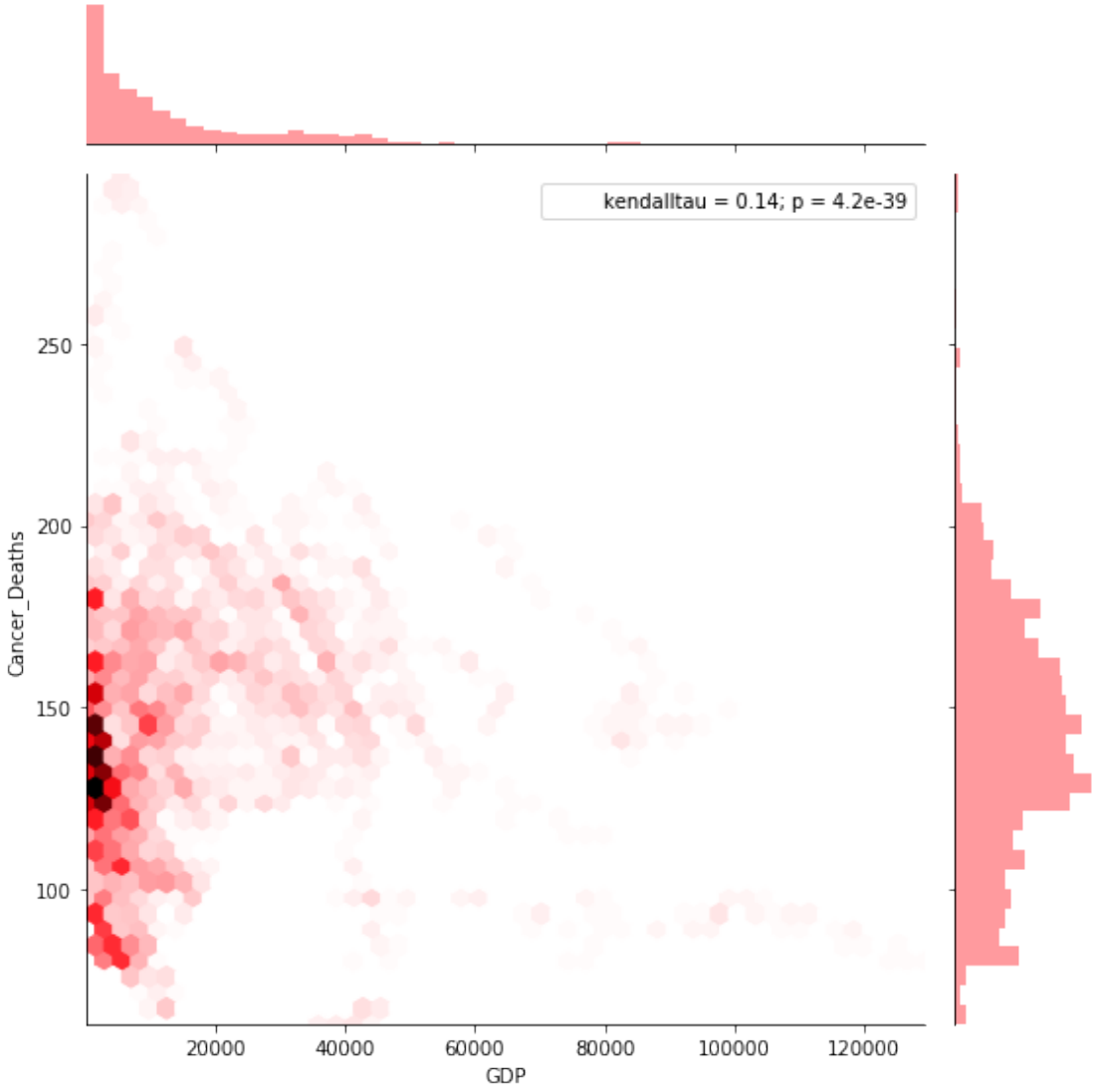
```
/home/haziq/.local/lib/python3.6/site-packages/seaborn/axisgrid.py:2272:
UserWarning:

The `size` parameter has been renamed to `height`; please update your code.
```

[86]: <seaborn.axisgrid.JointGrid at 0x7fca357df048>



[87]:
```
sns.jointplot(result1['GDP'], result1['Cancer_Deaths'] ,
              kind="hex", stat_func=kendalltau, color="#FF030D", size = 8 )
```

[87]: <seaborn.axisgrid.JointGrid at 0x7fca3427f908>



[89]:
```
#Picking the countries

result2 = result1.query('Entity in ["Norway", "Switzerland" , "Ireland" ,␣
↪"Germany" , "Australia" , "Iceland" , "United Kingdom" , "United States" ,␣
↪"Finland" , "Japan", "Pakistan" , "Yemen" , "Liberia" , "Guinea", "Congo",␣
↪"Mozambique" , "Afghanistan" , "Zimbabwe" , "Syria" , "Iraq"] ')

result2
```

[89]:

	Entity	Code	Year	Cancer_Deaths	GDP	\
0	Afghanistan	AFG	2002	163.79	1063.635574	
256	Afghanistan	AFG	2003	164.58	1099.194507	
512	Afghanistan	AFG	2004	165.18	1062.249360	

```
768       Afghanistan  AFG  2005          165.25  1136.123214
1024      Afghanistan  AFG  2006          165.14  1161.124889
…                 …  …    …             …       …
1032240      Zimbabwe  ZWE  2009          256.74  1336.212644
1032496      Zimbabwe  ZWE  2010          253.09  1474.877128
1032752      Zimbabwe  ZWE  2011          248.43  1667.137943
1033008      Zimbabwe  ZWE  2012          244.94  1871.366340
1033264      Zimbabwe  ZWE  2013          242.78  1929.765001

         Total population (Gapminder)    HDI  Population  Incidence_Neoplasms
0                          24639841.0  0.373  21979923.0           182.426026
256                        25678639.0  0.373  21979923.0           182.426026
512                        26693486.0  0.373  21979923.0           182.426026
768                        27614718.0  0.373  21979923.0           182.426026
1024                       28420974.0  0.373  21979923.0           182.426026
…                                …    …          …                    …
1032240                    12473992.0  0.435  12500525.0           242.665680
1032496                    12571454.0  0.435  12500525.0           242.665680
1032752                    12754378.0  0.435  12500525.0           242.665680
1033008                    13013678.0  0.435  12500525.0           242.665680
1033264                    13327925.0  0.435  12500525.0           242.665680

[414 rows x 9 columns]
```

[ ]:

[ ]: