CANCER DATASET

BUSINESS UNDERSTANDING Provided with the cancer dataset we are to determine the rates of the different cases of cancer and how cancer contributes to death rates

Start coding or generate with AI.

DATA LOADING

```
#loading dataset
import pandas as pd
df=pd.read_excel("/content/Cancer.xlsx")
df
```

	Country or Territory	Cancer deaths attributable to alcohol\nProportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Cancers attributable to infections\nProportion of cancers attributable to infections (%), by country	Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016	Obesity prev female\nInterna variation prevale obesity
0	Afghanistan	0.5	21.4	7	16	3.2	
1	Algeria	1.1	17.5	2.2	12.9	19.9	
2	Azerbaijan	3.2	36.7	0.3	15.4	15.8	
3	Albania	4.9	40.9	6.1	13.4	21.6	
4	Armenia	5	43.5	1.5	12.2	17.1	
...	...	...	...	...	...	...	
202	Réunion	No data	No data	No data	12.4	No data	I
203	French Polynesia	No data	No data	No data	6.6	No data	I
204	Guadeloupe	No data	No data	No data	10.2	No data	I
205	Guam	No data	No data	No data	8.5	No data	I
206	Martinique	No data	No data	No data	6.6	No data	I

207 rows × 27 columns

loading the dataset displays it in columns and rows

```
#dataset information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 207 entries, 0 to 206
Data columns (total 27 columns):
 #   Column
---  ------
 0   Country or Territory
 1   Cancer deaths attributable to alcohol
Proportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016   207 non-null    c
 2   Smoking prevalence male
Prevalence (%) of daily smoking for men                                                          207
 3   Smoking prevalence female
Prevalence (%) of daily smoking for women                                                        207 r
 4   Cancers attributable to infections
Proportion of cancers attributable to infections (%), by country                                 207 non-null
 5   Obesity prevalence male
International variation in the prevalence of obesity, 2016                                       207
 6   Obesity prevalence female
International variation in the prevalence of obesity, 2016                                       207 r
```

```
 7   Melanoma skin cancer incidence
Age-standardized rate (world) per 100,000, both sexes, 2018                                                    207 non-nu
 8   Breastfeeding at 12 months
Percent (%) of children who receive any breast milk at 12 months of age                                          207 nc
 9   Average births per woman
2010-2015                                                                                                          207
 10  Outdoor air pollution
Average annual population-weighted concentrations of PM2.5 (particulate matter of 2.5 µm diameter or less), measured in µg/m3, 2017  2
 11  Indoor air pollution
Proportion (%) of population using solid fuels in 2017
 12  Cancer rank as leading cause of death among 30-69
2016                                                                                  207 non-null    object
 13  Lung cancer incidence rates, male
Age-standardized rate (world) per 100,000, all ages, 2018                                                  207 non-null
 14  Lung cancer incidence rates, female
Age-standardized rate (world) per 100,000, all ages, 2018                                                207 non-null
 15  Breast most frequently diagnosed cancer in women
Countries where breast cancer is the most frequently diagnosed cancer in women, 2018                  207 non-null    object
 16  Human Development Index (HDI) levels
2017                                                                                                   207 non-null
 17  Most common cancer cases worldwide, females
2018                                                                                         207 non-null    object
 18  Most common cancer deaths worldwide, females
2018                                                                                       207 non-null    object
 19  Most common cancer cases worldwide, males
2018                                                                                          207 non-null    objec
 20  Most common cancer deaths worldwide, males
2018                                                                                        207 non-null    object
 21  Cancer survivors
Estimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
 22  Years lived with disability due to cancer
Both sexes, all ages, 2017                                                                                    207 non-null    objec
 23  Hepatitis B virus vaccination
Hepatitis B vaccination coverage (% of one-year-olds who have received three doses of hepatitis B vaccine), 2017             207 non-r
 24  Radiotherapy availability
Number of radiotherapy machines per 1,000 cancer patients                                                        207 r
 25  Cervical cancer incidence rates
Age-standardized rate (world) per 100,000, 2018                                                                207 non-nul
 26  HPV prevalence (%)
```

size of dataset is 43.8+ KB dataset has 27 columns datatype is object

```
#head
df.head()
```

	Country or Territory	Cancer deaths attributable to alcohol\nProportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Cancers attributable to infections\nProportion of cancers attributable to infections (%), by country	Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016	Obesity prevale female\nInternati variation in prevalenc obesity,
0	Afghanistan	0.5	21.4	7	16	3.2	
1	Algeria	1.1	17.5	2.2	12.9	19.9	
2	Azerbaijan	3.2	36.7	0.3	15.4	15.8	
3	Albania	4.9	40.9	6.1	13.4	21.6	
4	Armenia	5	43.5	1.5	12.2	17.1	

5 rows × 27 columns

shows first five rows

```
#tail
df.tail()
```

	Country or Territory	Cancer deaths attributable to alcohol\nProportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Cancers attributable to infections\nProportion of cancers attributable to infections (%), by country	Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016	Obesity prev female\nInterna variation prevale obesity
202	Réunion	No data	No data	No data	12.4	No data	I
203	French Polynesia	No data	No data	No data	6.6	No data	I
204	Guadeloupe	No data	No data	No data	10.2	No data	I
205	Guam	No data	No data	No data	8.5	No data	I
206	Martinique	No data	No data	No data	6.6	No data	I

5 rows × 27 columns

shows last 5 rows

```
#structure
df.shape
```

```
(207, 27)
```

there are 207 columns and 27 rows before cleaning

```
#datatypes
df.dtypes
```

```
Country or Territory
object
Cancer deaths attributable to alcohol\nProportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016
object
Smoking prevalence male\nPrevalence (%) of daily smoking for men
object
Smoking prevalence female\nPrevalence (%) of daily smoking for women
object
Cancers attributable to infections\nProportion of cancers attributable to infections (%), by country
object
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
object
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
object
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
object
Breastfeeding at 12 months\nPercent (%) of children who receive any breast milk at 12 months of age
object
Average births per woman\n2010-2015
object
Outdoor air pollution\nAverage annual population-weighted concentrations of PM2.5 (particulate matter of 2.5 µm diameter or less),
measured in µg/m3, 2017    object
Indoor air pollution\nProportion (%) of population using solid fuels in 2017
object
Cancer rank as leading cause of death among 30-69\n2016
object
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
object
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
object
Breast most frequently diagnosed cancer in women\nCountries where breast cancer is the most frequently diagnosed cancer in women, 2018
object
Human Development Index (HDI) levels\n2017
object
Most common cancer cases worldwide, females\n2018
object
Most common cancer deaths worldwide, females\n2018
object
Most common cancer cases worldwide, males\n2018
object
Most common cancer deaths worldwide, males\n2018
object
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
```

```
object
Years lived with disability due to cancer\nBoth sexes, all ages, 2017
object
Hepatitis B virus vaccination\nHepatitis B vaccination coverage (% of one-year-olds who have received three doses of hepatitis B
vaccine), 2017               object
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
object
Cervical cancer incidence rates\nAge-standardized rate (world) per 100,000, 2018
object
HIV prevalence (%)\nBoth sexes, 2017
object
dtype: object
```

datatype present is object only

```
from tabulate import tabulate

# Get all column names
column_names = df.columns.tolist()

# Convert column names to a list of lists for tabulate
column_names_table = [[i+1, name] for i, name in enumerate(column_names)]

# Print column names as a table
print(tabulate(column_names_table, headers=["Index", "Column Name"]))
```

```
  Index  Column Name
-------  ------------------------------------------------------------------------------------------------------------------------------------------------------
      1  Country or Territory
      2  Cancer deaths attributable to alcohol
         Proportion (%) of cancer deaths caused by alcohol drinking in men ages 15 years or older, 2016
      3  Smoking prevalence male
         Prevalence (%) of daily smoking for men
      4  Smoking prevalence female
         Prevalence (%) of daily smoking for women
      5  Cancers attributable to infections
         Proportion of cancers attributable to infections (%), by country
      6  Obesity prevalence male
         International variation in the prevalence of obesity, 2016
      7  Obesity prevalence female
         International variation in the prevalence of obesity, 2016
      8  Melanoma skin cancer incidence
         Age-standardized rate (world) per 100,000, both sexes, 2018
      9  Breastfeeding at 12 months
         Percent (%) of children who receive any breast milk at 12 months of age
     10  Average births per woman
         2010-2015
     11  Outdoor air pollution
         Average annual population-weighted concentrations of PM2.5 (particulate matter of 2.5 µm diameter or less), measured in µg/m3,
     12  Indoor air pollution
         Proportion (%) of population using solid fuels in 2017
     13  Cancer rank as leading cause of death among 30-69
         2016
     14  Lung cancer incidence rates, male
         Age-standardized rate (world) per 100,000, all ages, 2018
     15  Lung cancer incidence rates, female
         Age-standardized rate (world) per 100,000, all ages, 2018
     16  Breast most frequently diagnosed cancer in women
         Countries where breast cancer is the most frequently diagnosed cancer in women, 2018
     17  Human Development Index (HDI) levels
         2017
     18  Most common cancer cases worldwide, females
         2018
     19  Most common cancer deaths worldwide, females
         2018
     20  Most common cancer cases worldwide, males
         2018
     21  Most common cancer deaths worldwide, males
         2018
     22  Cancer survivors
         Estimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
     23  Years lived with disability due to cancer
         Both sexes, all ages, 2017
     24  Hepatitis B virus vaccination
         Hepatitis B vaccination coverage (% of one-year-olds who have received three doses of hepatitis B vaccine), 2017
     25  Radiotherapy availability
         Number of radiotherapy machines per 1,000 cancer patients
     26  Cervical cancer incidence rates
         Age-standardized rate (world) per 100,000, 2018
     27  HIV prevalence (%)
         Both sexes, 2017
```

shows the 27 columns in a list

```
#selecting relevant columns
selected_columns=[
                          'Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 populati
                          'Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018',
                          'Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018',
                          'Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018',
                          'Smoking prevalence female\nPrevalence (%) of daily smoking for women',
                          'Smoking prevalence male\nPrevalence (%) of daily smoking for men',
                          'Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients',
                          'Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018',
                          'Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016',
                          'Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016'
    ]
selected_df=df[selected_columns]
selected_df
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	Me inci s r p
0	151	3.2	9.4	0.3	7	21.4	0	
1	312.6	2.9	17.4	0.7	2.2	17.5	0.75	
2	228.1	2.6	25.5	0.53	0.3	36.7	0.86	
3	401.1	7.1	37.8	1.7	6.1	40.9	0.6	
4	402.2	8.5	58.5	1.6	1.5	43.5	0.34	
...	...	...	...	...	...	...	...	
202	590.1	9.6	34.6	3.3	No data	No data	No data	
203	661.9	23.4	55.7	6.5	No data	No data	No data	
204	776.5	6.5	13	1.5	No data	No data	0.88	
205	418	24.3	53.7	No data	No data	No data	2.58	
206	747	9.2	12.3	2	No data	No data	1.46	

207 rows × 10 columns

there are 10 relevant columns to be used in the *analysis*

```
#structure of selected columns
selected_df.shape
```

```
(207, 10)
```

207 rows and 10 columns shape after selecting relevant columns

```
#head of relevant cols
selected_df.head()
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	Mela incide sta rat per bo
0	151	3.2	9.4	0.3	7	21.4	0	
1	312.6	2.9	17.4	0.7	2.2	17.5	0.75	
2	228.1	2.6	25.5	0.53	0.3	36.7	0.86	
3	401.1	7.1	37.8	1.7	6.1	40.9	0.6	
4	402.2	8.5	58.5	1.6	1.5	43.5	0.34	

shows first five rows

```
#tail of relevant cols
selected_df.tail()
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	Me inci s r p
202	590.1	9.6	34.6	3.3	No data	No data	No data	
203	661.9	23.4	55.7	6.5	No data	No data	No data	
204	776.5	6.5	13	1.5	No data	No data	0.88	
205	418	24.3	53.7	No data	No data	No data	2.58	
206	747	9.2	12.3	2	No data	No data	1.46	

shows last 5 rows of selected columns

```
#datatypes in selected columns
selected_df.dtypes
```

```
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
object
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
object
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
object
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
object
Smoking prevalence female\nPrevalence (%) of daily smoking for women
object
Smoking prevalence male\nPrevalence (%) of daily smoking for men
object
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
object
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
object
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
object
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
object
dtype: object
```

object is the only datatype present

```
#change all columns to numeric types
selected_df=selected_df.apply(pd.to_numeric, errors='coerce')
selected_df
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	Me inci s r p
0	151.0	3.2	9.4	0.30	7.0	21.4	0.00	
1	312.6	2.9	17.4	0.70	2.2	17.5	0.75	
2	228.1	2.6	25.5	0.53	0.3	36.7	0.86	
3	401.1	7.1	37.8	1.70	6.1	40.9	0.60	
4	402.2	8.5	58.5	1.60	1.5	43.5	0.34	
...	...	...	...	...	...	...	...	
202	590.1	9.6	34.6	3.30	NaN	NaN	NaN	
203	661.9	23.4	55.7	6.50	NaN	NaN	NaN	
204	776.5	6.5	13.0	1.50	NaN	NaN	0.88	
205	418.0	24.3	53.7	NaN	NaN	NaN	2.58	
206	747.0	9.2	12.3	2.00	NaN	NaN	1.46	

207 rows × 10 columns

The data in the selected columns has been changed to numeric

```
#datatypes of columns
selected_df.dtypes
```

```
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
float64
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
float64
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
float64
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
float64
Smoking prevalence female\nPrevalence (%) of daily smoking for women
float64
Smoking prevalence male\nPrevalence (%) of daily smoking for men
float64
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
float64
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
float64
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
float64
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
float64
dtype: object
```

datatype confirmed to be float

```
#summary statistics
selected_df.describe()
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	in
count	185.000000	184.000000	185.000000	178.000000	191.000000	189.000000	182.00000	
mean	435.127027	9.440543	22.545189	3.881798	7.112565	23.939683	0.63000	
std	310.020328	8.777530	18.471358	6.407851	7.254766	11.825503	0.55368	
min	74.800000	0.040000	0.380000	0.040000	0.100000	1.300000	0.00000	
25%	202.300000	2.600000	6.400000	0.530000	1.500000	14.300000	0.08000	
50%	314.200000	6.400000	16.900000	1.200000	3.900000	21.600000	0.60000	
75%	591.800000	13.225000	35.500000	3.675000	12.400000	33.500000	1.00000	
max	1849.800000	41.400000	77.400000	33.600000	42.100000	66.600000	2.58000	

shows summary statistics of the selected columns before cleaning

DATA CLEANING

```
#missing values
missing_values=selected_df.isnull().sum()
missing_values
```

```
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
22
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
23
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
22
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
29
Smoking prevalence female\nPrevalence (%) of daily smoking for women
16
Smoking prevalence male\nPrevalence (%) of daily smoking for men
18
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
25
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
29
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
16
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
16
dtype: int64
```

shows missing values in each column

```
#remove missing values
selected_df.dropna(subset=selected_columns, inplace=True)
```

```
#checking dropped missing values
missing_values=selected_df.isnull().sum()
missing_values
```

```
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
0
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
0
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
0
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
0
Smoking prevalence female\nPrevalence (%) of daily smoking for women
0
```

```
Smoking prevalence male\nPrevalence (%) of daily smoking for men
0
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
0
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
0
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
0
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
0
dtype: int64
```

missing values confirmed dropped from the dataset

```
#duplicated
selected_df.duplicated().sum()
```

```
3
```

3 duplicates present in the datatset

```
#Drop rows with duplicated values
selected_df.drop_duplicates(inplace=True)
```

```
#check for duplicates
duplicates = selected_df.duplicated().sum()
print("number of duplicates", duplicates)
```

```
number of duplicates 0
```

successfully removed all duplicates

```
print(selected_df)
```

```
     Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2
0                                                151.0
1                                                312.6
2                                                228.1
3                                                401.1
4                                                402.2
..                                                 ...
189                                              743.2
190                                              593.7
191                                              975.6
192                                              729.5
193                                              432.1

     Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018  \
0                                                 3.2
1                                                 2.9
2                                                 2.6
3                                                 7.1
4                                                 8.5
..                                                ...
189                                               8.4
190                                              14.8
191                                              18.2
192                                              11.8
193                                               5.1

     Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018  \
0                                                 9.4
1                                                17.4
2                                                25.5
3                                                37.8
4                                                58.5
..                                                ...
189                                              41.0
190                                              31.3
191                                              40.0
192                                              28.2
193                                              21.8

     Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018  \
0                                                0.30
1                                                0.70
```

```
2                                                 0.53
3                                                 1.70
4                                                 1.60
..                                                 ...
189                                               4.50
190                                               1.80
191                                              16.50
192                                               8.00
193                                               0.58

     Smoking prevalence female\nPrevalence (%) of daily smoking for women  \
0                                                  7.0
1                                                  2.2
2                                                  0.3
```

```
#ckecking for completeness
completeness = selected_df.notnull().sum() / len(selected_df) * 100
print(completeness)
```

```
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
Smoking prevalence female\nPrevalence (%) of daily smoking for women
Smoking prevalence male\nPrevalence (%) of daily smoking for men
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
Obesity prevalence female\nInternational variation in the prevalence of obesity, 2016
Obesity prevalence male\nInternational variation in the prevalence of obesity, 2016
dtype: float64
```

the data is complete in all 10 columns

```
#structure after cleaning
selected_df.shape
```

```
(164, 10)
```

164 rows and 10 columns after cleaning

EXPLORATORY DATA ANALYSIS

```
#summary statistics
selected_df.describe()
```

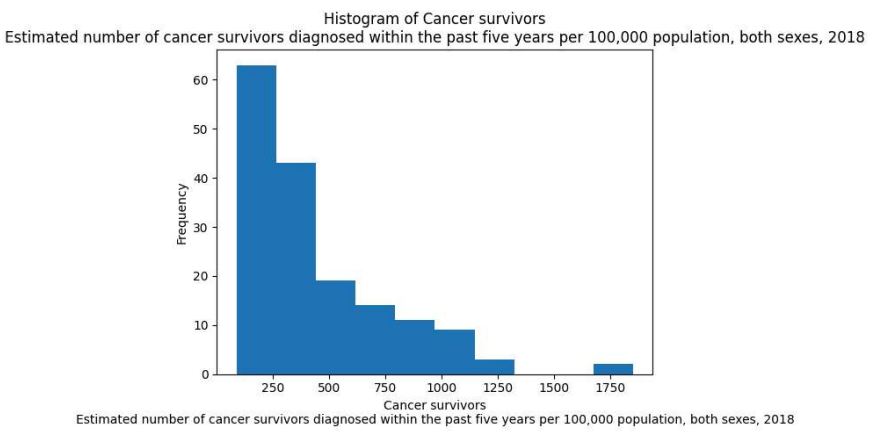
	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	Smoking prevalence female\nPrevalence (%) of daily smoking for women	Smoking prevalence male\nPrevalence (%) of daily smoking for men	Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	in
count	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	
mean	443.838415	9.476646	22.830366	4.030183	7.035976	24.107927	0.632744	
std	315.681201	8.779965	18.456194	6.624852	6.888346	11.502984	0.518112	
min	89.100000	0.040000	0.850000	0.040000	0.100000	4.400000	0.000000	
25%	213.525000	2.900000	6.400000	0.520000	1.500000	14.475000	0.160000	
50%	320.200000	6.150000	17.200000	1.200000	3.800000	21.750000	0.625000	
75%	592.275000	13.350000	35.975000	3.925000	12.650000	33.125000	0.980000	
max	1849.800000	41.400000	77.400000	33.600000	28.100000	66.600000	2.380000	

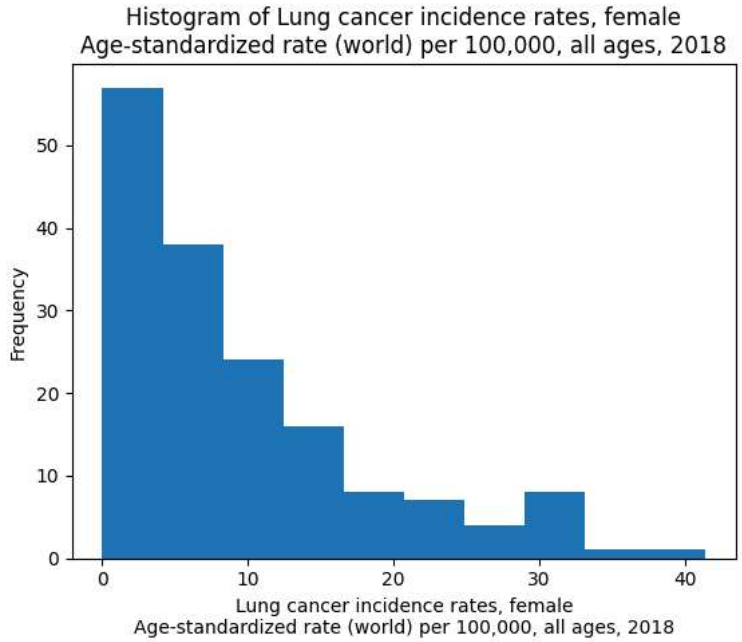
shows summary statistics after cleaning

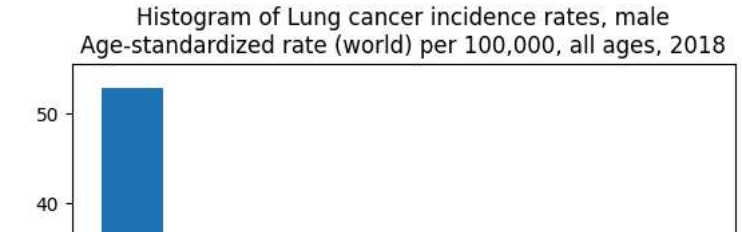
```
# Correlation analysis
import seaborn as sns
import matplotlib.pyplot as plt

# Calculate the correlation matrix
correlation_matrix = selected_df.corr()
correlation_matrix
```

	Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018
Cancer survivors\nEstimated number of cancer survivors diagnosed within the past five years per 100,000 population, both sexes, 2018	1.000000	0.804433	0.608835	0.874856
Lung cancer incidence rates, female\nAge-standardized rate (world) per 100,000, all ages, 2018	0.804433	1.000000	0.675371	0.705653
Lung cancer incidence rates, male\nAge-standardized rate (world) per 100,000, all ages, 2018	0.608835	0.675371	1.000000	0.377356
Melanoma skin cancer incidence\nAge-standardized rate (world) per 100,000, both sexes, 2018	0.874856	0.705653	0.377356	1.000000
Smoking prevalence female\nPrevalence (%) of daily smoking for women	0.695555	0.688884	0.712517	0.569839
Smoking prevalence male\nPrevalence (%) of daily smoking for men	0.026245	0.098676	0.537519	-0.113307
Radiotherapy availability\nNumber of radiotherapy machines per 1,000 cancer patients	0.498925	0.433936	0.382521	0.345231
Melanoma skin				

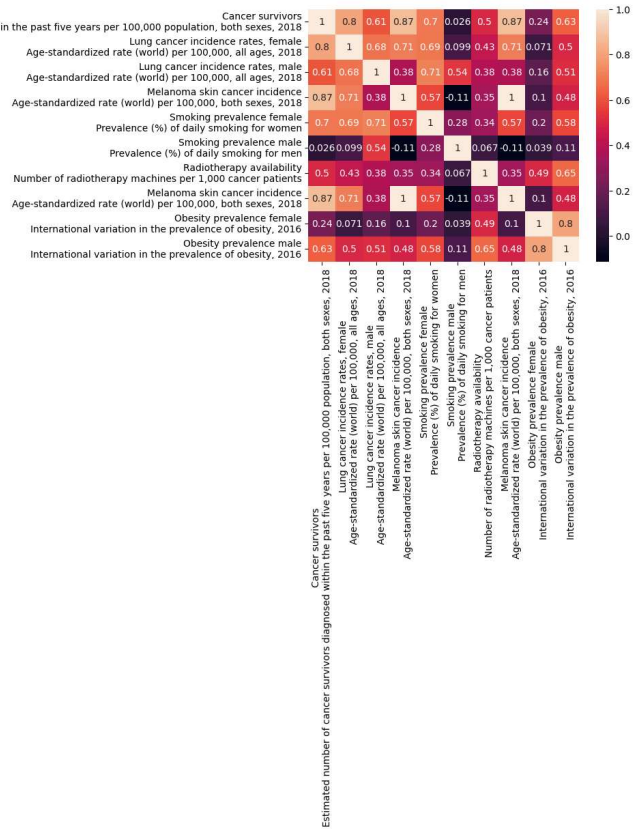
```
# plot histograms

import matplotlib.pyplot as plt

# Create a histogram for each selected column
for col in selected_df:
  plt.figure()
  plt.hist(selected_df[col], bins=10)
  plt.xlabel(col)
  plt.ylabel("Frequency")
  plt.title(f"Histogram of {col}")
  plt.show()
```



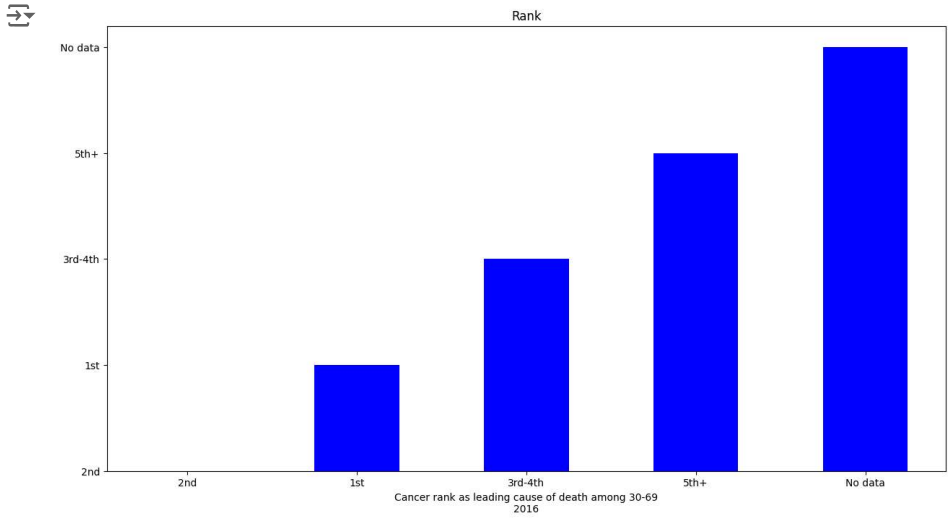





```
#heatmap of the correlation matrix
sns.heatmap(correlation_matrix, annot=True)
plt.show()
```

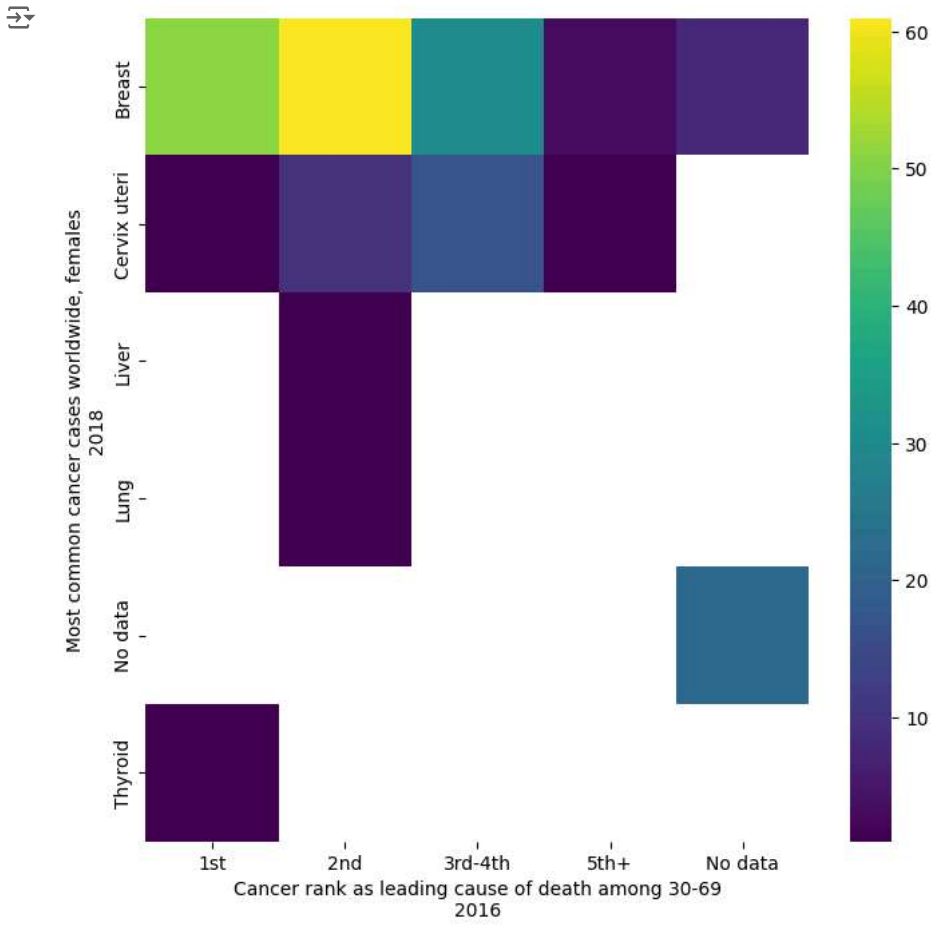



```
import matplotlib.pyplot as plt
import numpy as np
fig = plt.figure(figsize = (15,8))
values = df['Cancer rank as leading cause of death among 30-69\n2016']
#creating the bar plot
plt.bar(df['Cancer rank as leading cause of death among 30-69\n2016'], values, color ='blue', width=0.5)
plt.xlabel("Cancer rank as leading cause of death among 30-69\n2016")
plt.title("Rank")
plt.show()
```



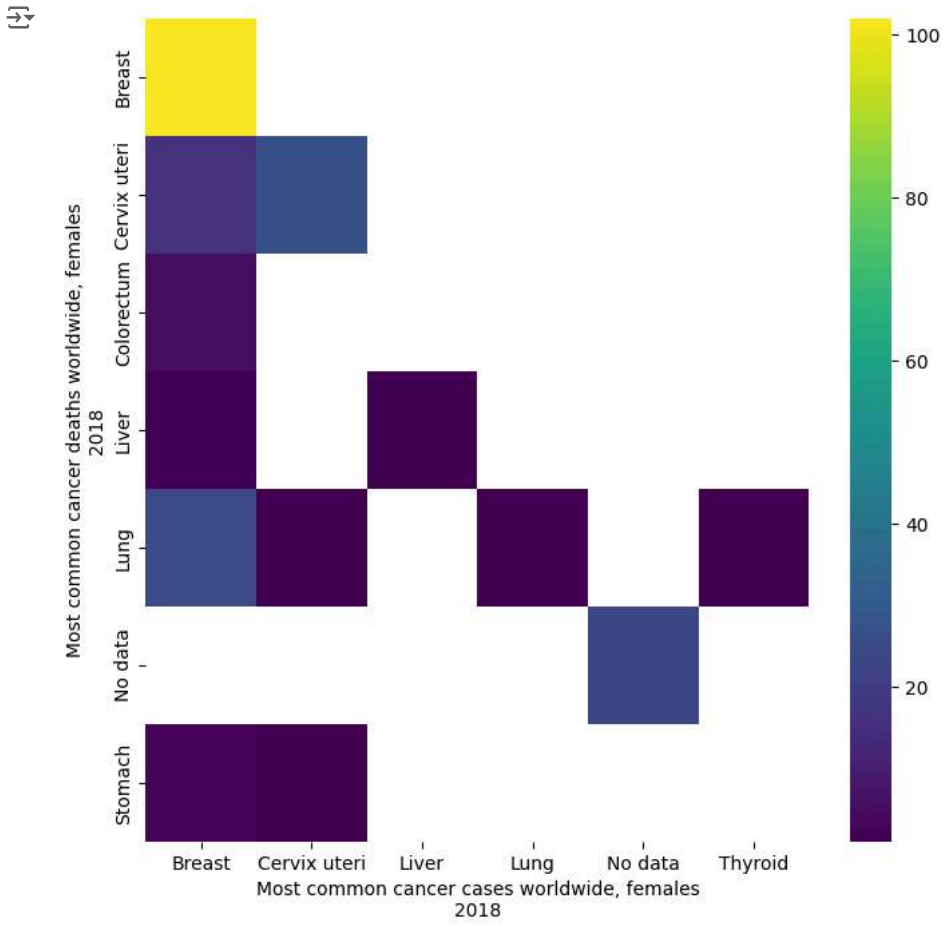
The bar graph shown displays how cancer ranks in different countries as the leading cause of death whereby in some countries, it ranked first as the leading cause of death. There was no data in most of the countries but it is evident in others where cancer ranked as 3rd-4th leading cause of death and even 5th in more countries

```
#Cancer rank as leading cause of death among 30-69
#2016 vs Most common cancer cases worldwide, females
#2018

from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
plt.subplots(figsize=(8, 8))
df_2dhist = pd.DataFrame({
    x_label: grp['Most common cancer cases worldwide, females\n2018'].value_counts()
    for x_label, grp in df.groupby('Cancer rank as leading cause of death among 30-69\n2016')
})
sns.heatmap(df_2dhist, cmap='viridis')
plt.xlabel('Cancer rank as leading cause of death among 30-69\n2016')
_ = plt.ylabel('Most common cancer cases worldwide, females\n2018')
```



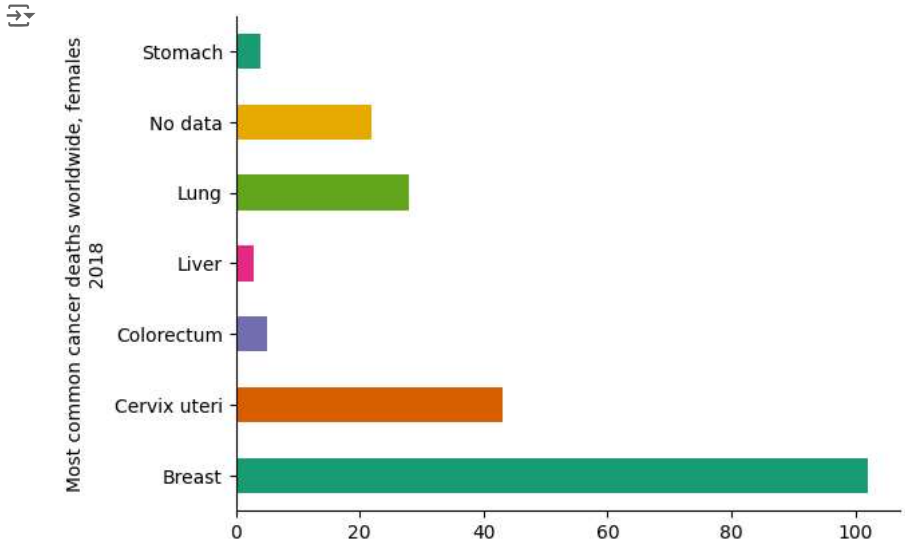
This graph portrays the rank of cancer as the leading cause of death among 30-69 vs the most common cases worldwide in females in 2016

```
# Most common cancer cases worldwide, females
#2018 vs Most common cancer deaths worldwide, females
#2018

from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
plt.subplots(figsize=(8, 8))
df_2dhist = pd.DataFrame({
    x_label: grp['Most common cancer deaths worldwide, females\n2018'].value_counts()
    for x_label, grp in df.groupby('Most common cancer cases worldwide, females\n2018')
})
sns.heatmap(df_2dhist, cmap='viridis')
plt.xlabel('Most common cancer cases worldwide, females\n2018')
_ = plt.ylabel('Most common cancer deaths worldwide, females\n2018')
```



The plot above displays the relation between the most common cancer cases worldwide in females in 2018

```
#Most common cancer deaths worldwide, females
2018

from matplotlib import pyplot as plt
import seaborn as sns
df.groupby('Most common cancer deaths worldwide, females\n2018').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```



The bar graph here shows the most common cancer deaths in females worldwide in 2018, whereby breast cancer was the leading cause of death and liver cancer was the least cause of death among females in 2018

```
#Most common cancer cases worldwide, females
2018
```

```
from matplotlib import pyplot as plt
import seaborn as sns
df.groupby('Most common cancer cases worldwide, females\n2018').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```