**Data Visualizations and Analysis on Cancer Dataset**

**California State University Los Angeles**

**By**

**Divya Pakhale**

**Neha Shashidhara Guli**

**Table of Contents**

**Dataset URL** ------------------------------------------------------------------------------------------- **3**

**Data Description -------------------------------------------------------------------------------------- 3**

**Data Cleaning ------------------------------------------------------------------------------------------4**

**Summary Statistics ----------------------------------------------------------------------------------- 7**

**Analytics & Visualization --------------------------------------------------------------------------- 9**

**References ---------------------------------------------------------------------------------------------14**

**DATASETURL:**

[**https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer**](https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer)

**DATA DESCRIPTION:**

Community Health Status Indicators (CHSI) to battle three kinds of disease are significant segments of the Community Health Data Initiative. This dataset gives key wellbeing pointers to neighborhood networks and empowers exchange about moves that can be made to enhance community health (e.g. Cancer). The CHSI report and dataset was planned for general wellbeing experts as well as for individuals from the community who are keen on the soundness of their community. The CHSI report contains more than 200 measures for every one of the 3,141 United States areas. In spite of the fact that CHSI presents pointers like deaths because of malignant growth, it is basic to comprehend that social factors, for example, heftiness, tobacco use, diet, physical action, liquor and medication use, sexual conduct and others significantly add to these deaths. From the CHSI dataset, we are just focusing on three different types of cancer such as breast cancer, colon cancer and lung cancer. Breast cancer starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. The tumor is malignant (cancer) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely in women, but men can get breast cancer, too. Colon cancer is malignant growth of the internal organ (colon), which is the last piece of your stomach related tract. Most instances of colon cancer growth start as little, noncancerous (benevolent) clusters of cells called adenomatous polyps. Lung cancer is a condition that makes cells isolate in the lungs wildly. This causes the development of tumors that diminish a man's capacity to relax. In our dataset we have 27 columns and about 3000 rows. Some of the main columns mentioned in our dataset are:

Breast Cancer - County data, death measures, breast cancer

Min\_Brst\_Cancer - Tenth percentile from peer counties, death measures, breast cancer

Max\_Brst\_Cancer - Nintieth percentile from peer counties, death measures, breast cancer

Col\_Cancer - County data, death measures, colon cancer

Lung\_Cancer - County data, death measures, lung cancer

Total\_Births - County data, total number of births

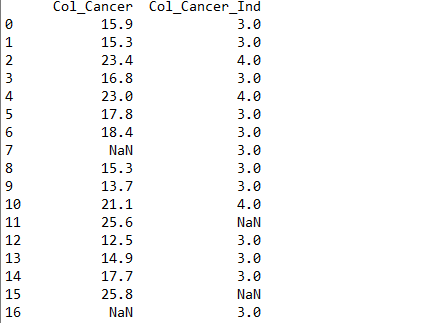
Total\_Deaths - County data, total number of deaths

MOBD\_Time\_Span - Time period of reported data for measures of birth and death

**DATA CLEANING:**

1. Removing “NULL” values

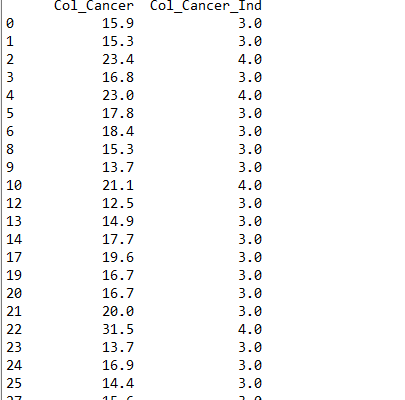
Before cleaning:



Code:

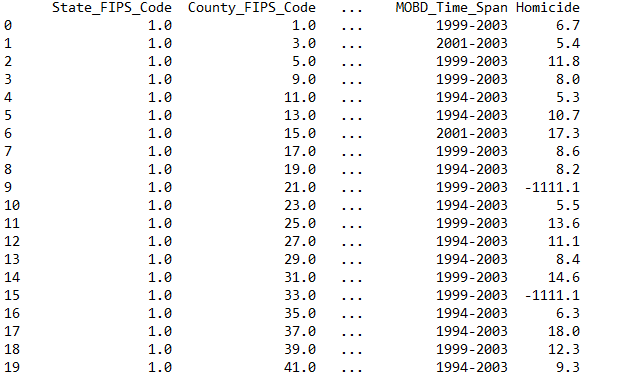
|  |
| --- |
| import pandas as pd  CHSI = pd.read\_csv('CHSI.csv')  df = CHSI[['Col\_Cancer','Col\_Cancer\_Ind']]  print(df.dropna(how='any')) |

After Cleaning:



2. Deleting Irrelevant columns

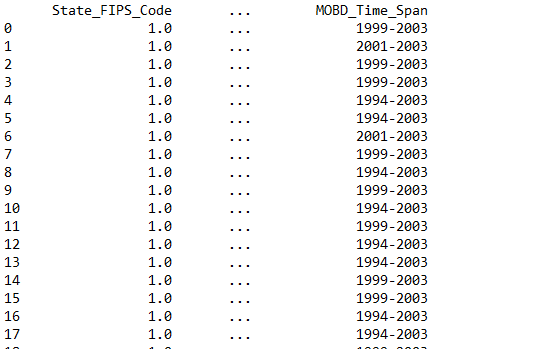
Before cleaning:



Code:

|  |
| --- |
| import pandas as pd  CHSI = pd.read\_csv('CHSI.csv')  del CHSI['Homicide']  print(CHSI) |

After Cleaning:



3. Applying Delimiters:

In our dataset, we don’t have any such column where we can apply delimiter.

**SUMMARY STATISTICS**

1. Minimum and Maximum value for Colon Cancer:

Code:

|  |
| --- |
| import matplotlib.pyplot as plt  import pandas as pd  df=pd.read\_csv('MeasuresOfBirth&Death.csv')  print('Minimum Value:', df['Col\_Cancer'].min())  print('Maximum Value:', df['Col\_Cancer'].max()) |

Output:



2. Mean and Standard Deviation:

Code:

|  |
| --- |
| import matplotlib.pyplot as plt  import pandas as pd  df=pd.read\_csv('MeasuresOfBirth&Death.csv')  print('Mean Value:',df['Col\_Cancer'].mean())  print('Standard Deviation Value:',df['Col\_Cancer'].std()) |

Output:

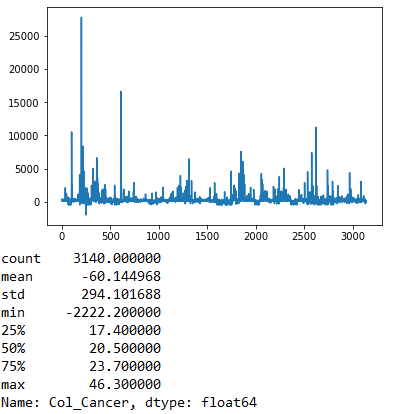


3. Describe function:

Code:

|  |
| --- |
| import matplotlib.pyplot as plt  import pandas as pd  df=pd.read\_csv('MeasuresOfBirth&Death.csv')  mean = df.mean(axis = 'columns')  mean.plot()  plt.show()  print(df['Col\_Cancer'].describe()) |

Output:



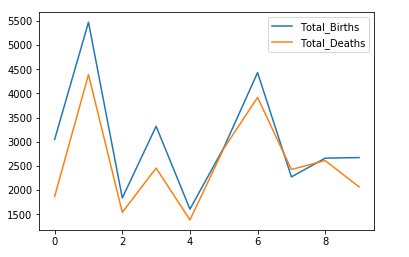
**ANALYSIS & VISULAIZATIONS**

**1. Total Death vs Total Birth**

**CODE:**

|  |
| --- |
| # total\_Deaths and Total\_Births  import matplotlib.pyplot as plt  import pandas as pd  df=pd.read\_csv('MeasuresOfBirth&Death.csv')  df2 = df.iloc[:,24:26].head(10)  labels = ['Total Births','Total Deaths']  plt.plot(df2)  plt.legend(('Total\_Births','Total\_Deaths'))  plt.show() |

**OUTPUT:**



Application used: [iloc, List]

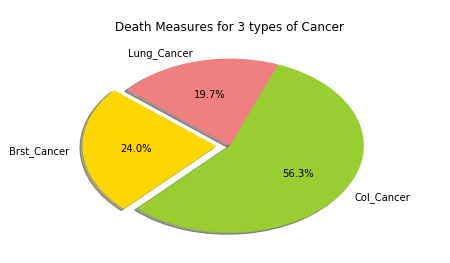
A line graph is commonly used to display change over time as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set. The above line chart describes about the total deaths due to breast, colon and lung cancer. And further to this, it also shows the ratio of birth rate during that period. As per the above chart, we can see that total number of births were more compared to total number of deaths hence, we can make out that less people were suffering due to three types of cancer. In this chart, we have used iloc function to select the two columns based on top 10 records.

**2. Death measures for 3 types of cancer**

**CODE:**

|  |
| --- |
| import matplotlib.pyplot as plt  import pandas as pd  cancer=pd.read\_csv('MeasuresOfBirth&Death.csv')  a=cancer.iloc[:,[6,12,18]].head(3)  Total\_Deaths = cancer['Total\_Deaths'].head(3)  colors = ['gold', 'yellowgreen', 'lightcoral']  explode = (0.1, 0, 0)  b = plt.pie(Total\_Deaths, explode=explode, labels=a, colors=colors,  autopct='%1.1f%%', shadow=True, startangle=140)  plt.title("Death Measures"+" for 3 types of Cancer")  plt.show() |

**OUTPUT:**



The above Pie-Chart, shows the percentage of total death measures due to breast, colon and lung cancer. The package used in this analysis are matplotlib and pandas. We have used head function to take top records. With the help of above analysis, we are able to determine which type of cancer has the highest number of deaths. As it can be clearly seen that maximum number of deaths were due to colon cancer which accounts for 56.3%. Further followed by percentage of people who died due to breast cancer which accounts for 24%. On the contrary, least deaths were due to lung cancer which was around 19.7%.

3. **Comparing death measure for colon and lung cancer**

Code:

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('MeasuresOfBirth&Death.csv')

col\_cancer = df['Col\_Cancer']

Lung\_Cancer = df['Lung\_Cancer']

plt.scatter(col\_cancer, Lung\_Cancer, color='g')

plt.scatter(Lung\_Cancer,col\_cancer, color='y')

plt.xlabel('colan\_cancer')

plt.ylabel('lung\_cancer')

plt.title('death measures for colan and lung cancer')

plt.xlim(9,45)

plt.ylim(9,46)

plt.show()

**OUTPUT:**

****

A scatter plot is a type of plot that shows the data as a collection of points. The position of a point depends on its two-dimensional value, where each value is a position on either the horizontal or vertical dimension. The above cod demonstrates the use of two libraries in python: matplotlib and pandas. The above scatter plot shows the death measure in all states within the range of 9-46 that are having lung and colon cancer. Scatter function is used and two colors green and yellow are used to differentiate the two entities. The green color dots represent lung cancer and yellow color represents colon cancer.

**4. Bar graph to show the 4 entities of measurement of Breast cancer**

Code:

import numpy as np

import pandas as pd

df = pd.read\_csv('MeasuresOfBirth&Death.csv')

df = pd.DataFrame(np.random.rand(10, 4), columns=['Min\_Brst\_Cancer', 'Max\_Brst\_Cancer', 'CI\_Min\_Brst\_Cancer', 'CI\_Max\_Brst\_Cancer'])

df.plot.bar();



A **bar chart** or **bar graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengthsproportional to the values that they represent. The bars can be plotted vertically or horizontally. Bar graphs are good if you to want to present the data of different groups that are being compared with each other. [3]

This visualization has four columns:

Min\_Brst\_Cancer- Tenth percentile from peer counties, death measures, breast cancer (female)

Max\_Brst\_Cancer- Nintieth percentile from peer counties, death measures, breast cancer (female)

CI\_Min\_Brst\_Cancer- Confidence interval lower limit, death measures, breast cancer (female)

CI\_Max\_Brst\_Cancer- Confidence interval upper limit, death measures, breast cancer (female)

Comparison is made between these measures with a common rage of reference. Here we have used two libraries – numpy and pandas. Calling a DataFrame’s **[plot.bar()](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.plot.bar.html" \l "pandas.DataFrame.plot.bar" \o "pandas.DataFrame.plot.bar)** method produces a multiple bar plot

**References:**

1. “Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer.” Datasets - Data.gov, Publisher Centers for Disease Control and Prevention, 26 Nov. 2018, catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer.

2. “Colorectal Cancer Risk Among Breast Cancer Survivors.” Medpage Today, MedpageToday, 21 Nov. 2017, [www.medpagetoday.com/resource-centers/breast-cancer-advances/colorectal-cancer-risk-among-breast-cancer-survivors/1626](http://www.medpagetoday.com/resource-centers/breast-cancer-advances/colorectal-cancer-risk-among-breast-cancer-survivors/1626).

3. Adnan. “Data Visualization in Python – Bar Graph in Matplotlib.” *Adnan's Random Bytes*, 14 Nov. 2017, blog.adnansiddiqi.me/data-visualization-in-python-bar-graph-in-matplotlib/.