Python Project

by Ansh Gulati

Submission date: 28-Apr-2023 02:09PM (UTC+0530)
Submission ID: 2078086277
File name: Final_Paper_Springer_Format.docx (5.61M)
Word count: 2079
Character count: 11870

Cancer Prediction Using Graph Database

Ansh Gulati[1], and Ameya Taneja[2]

[1]University of Petroleum & Energy Studies, Dehradun, India
anshgulati9817@gmail.com
[2]University of Petroleum & Energy Studies, Dehradun, India
ameyataneja2015@gmail.com

Abstract. This research paper aims to provide a comprehensive overview of cancer cases and rates in the various states of the United States. It explores the trends and patterns of cancer incidence and mortality in the country, as well as the factors such as age, sex/gender, type of cancer whether it is lung or breast cancer and its rates and also the factors that contribute to the development and progression of the disease. The paper reviews the latest statistics on cancer rates mainly breast and lung cancer in different population groups, including age, sex/gender, and geographical location/different states of USA. By analyzing the data, the project aims to provide insights and predictions related to the occurrence of cancer in the US. The Python code implements visualizations of cancer data for various states in the USA using Pandas and Matplotlib libraries. The dataset is read into a Pandas data frame and various types of visualizations are produced for the cancer data, including scatter plots, and bar graphs. The scatter plot represents the rate of lung and breast cancer in various states of the USA, and the bar graphs represent the total number of breast cancer and lung cancer, as well as the cancer rates in people of different age groups for each state. The visualizations allow for the comparison of cancer rates and total numbers between different states and age groups, aiding in identifying the states with higher cancer rates and potentially identifying any trends or patterns. The paper concludes by discussing the challenges and opportunities for cancer prevention, early detection, and treatment in the United States, and the implications for public health policy and practice. Potential applications of this analysis include informing strategies for cancer prevention and treatment in different states and age groups. The project could have implications for public health and policy, as well as for advancing our understanding of cancer and its impact on society. Overall, this paper aims to provide a comprehensive and up-to-date picture of the burden of cancer in the United States and to identify areas for further research and action.

Keywords: Cancer incidence, mortality, cancer prevention.

Table of Contents:

Theory

1 Introduction

1.1 About Cancer

Cancer is a complex and life-threatening disease that has become a significant health concern worldwide. It's a disease in which the human's cells grow out of control and multiply throughout the body. Cancer is a cluster of circumstances in which cells in the body start to reproduce and multiply in an unconstrained manner. The damage may succeed, or may be caused by errors that occur during normal cell regeneration. Cancer cells can go to different parts of the body, where they start to multiply and shape new tumors. This is called metastases. It happens when cells enter the bloodstream.

2.2 How common is Cancer

In the US, Cancer is the second major source of death, holding for nearly one in every four deaths. Cancer incidence and 5 types are affected by many factors, including age, gender, ethnicity, etc. Lung cancer is the most commonly diagnosed cancer in men, followed by oral cavity and laryngeal cancers whereas cervical cancer and breast cancer are most commonly diagnosed in women. Therefore, it's crucial to acknowledge the attributes that help to progress cancer growth and find solutions to predict its occurrence.

2.3 Aim of this Project

Machine learning and data analysis techniques have proven to be effective in identifying patterns and predicting outcomes in various fields, including cancer research. In this project, we aim to use machine learning and data analysis techniques to analyze cancer prediction rates in various states of the USA and predict the likelihood of breast and lung cancer development in the future. By analyzing this data, the project aims to provide insights and predictions related to the occurrence of cancer in the US.

2 Methodology of the Project

2.1 About the Dataset

We will use a publicly available dataset (a CSV file) that contains information on the total rates and numbers of breast cancer and lung cancer cases in different states of the US, as well as the cancer rates of different age groups. Using this dataset, we will perform data analysis and visualization techniques to identify any patterns or correlations between the different factors such as age, sex/gender, type of cancer whether it is lung or breast cancer and its rates.

2.2 Use of Python programming language and its specific libraries in the project

The project utilizes the pandas library in python to read and manipulate the data and the code also utilizes matplotlib library to create various graphs and visualizations (such as scatter plot and bar graphs) to better understand the trends and patterns in the data. Also the PrettyTable module is used to create a table that maps the full name of each state to its abbreviation. Specifically, the code creates scatter plots and bar graphs to represent the total rates and numbers of breast cancer and lung cancer cases in different states of the US, as well as the cancer rates of different age groups.

2.3 About the project

The code first creates a dictionary of states and their abbreviations, which is then used to create a pretty table for reference. The code then creates a scatter plot to represent the cancer rates of lung and breast cancer in the states of the USA. The scatter plot uses the state names from the Data Frame on the X-axis and the Total Rate of cancer on the Y-axis. Next, the code creates a bar graph to represent the total number of breast cancer and lung cancer in each state. The bar graph uses the state names from the Data Frame on the X-axis and the Total Number of cancer cases on the Y-axis.

Following this, four more bar graphs are created to represent the Cancer Rate of people based on different age groups i.e. below the age of 18, between 18-45, between 45-64 , and above 64 for both breast cancer and lung cancer in the various states of the USA.
The visualizations allow for the comparison of cancer rates (Breast and lung) between different states and age groups, aiding in identifying the states with higher cancer rates and potentially identifying any trends or patterns. Overall, the code aims to analyze and visualize the cancer rates and cases in different states of the USA through different types of plots.

2.4 Code Snippet

```
import pandas as pd
import matplotlib.pyplot as mpl
from prettytable import PrettyTable


df = pd.read_csv('cancer.csv')

from prettytable import PrettyTable

state_dict = {
  "Alabama": "AL",
  "Alaska": "AK",
  "Arizona": "AZ",
  "Arkansas": "AR",
  "California": "CA",
  "Colorado": "CO",
  "Connecticut": "CT",
  "Delaware": "DE",
  "District Of Columbia":"DOC",
  "Florida": "FL",
  "Georgia": "GA",
  "Hawaii": "HI",
  "Idaho": "ID",
  "Illinois": "IL",
  "Indiana": "IN",
  "Iowa": "IA",
  "Kansas": "KS",
  "Kentucky": "KY",
  "Louisiana": "LA",
  "Maine": "ME",
  "Maryland": "MD",
  "Massachusetts": "MA",
  "Michigan": "MI",
  "Minnesota": "MN",
  "Mississippi": "MS",
  "Missouri": "MO",
  "Montana": "MT",
  "Nebraska": "NE",
  "Nevada": "NV",
  "New Hampshire": "NH",
  "New Jersey": "NJ",
  "New Mexico": "NM",
  "New York": "NY",
```

```
    "North Carolina": "NC",
    "North Dakota": "ND",
    "Ohio": "OH",
    "Oklahoma": "OK",
    "Oregon": "OR",
    "Pennsylvania": "PA",
    "Rhode Island": "RI",
    "South Carolina": "SC",
    "South Dakota": "SD",
    "Tennessee": "TN",
    "Texas": "TX",
    "Utah": "UT",
    "Vermont": "VT",
    "Virginia": "VA",
    "Washington": "WA",
    "West Virginia": "WV",
    "Wisconsin": "WI",
    "Wyoming": "WY"
}

table = PrettyTable()
table.field_names = ["State", "Abbreviation"]

for state, abbr in states_dict.items():
    table.add_row([state, abbr])

print(table)

table2 = PrettyTable()
table2.field_names = ["Year", "Total Population", "Total Lung Cancer Cases","Total Breast
Cancer Cases",  "Cancer Rate"]
table2.add_row(["2014", "318386329", "224210", "232670", "143.99"])
table2.add_row(["2015", "320738994", "221200", "231840", "141.5"])
table2.add_row(["2016", "323071755", "224390", "246660", "145.8"])
table2.add_row(["2017", "325122128", "222500", "252710", "146.16"])
table2.add_row(["2018", "326838199", "234030", "266120", "153.02"])
table2.add_row(["2019", "328329953", "228150", "268600", "151.29"])
table2.add_row(["2020", "331501080", "228820", "276480", "156.34"])
table2.add_row(["2021", "331893745", "235760", "281550", "155.86"])
table2.add_row(["2022", "332403650", "236740", "287850", "157.81"])
print(table2)
print("To refer the graphs use the table above !")
#The rate of Lung and breast cancer in the states of USA using scatter plot
mpl.scatter(df['State'], df['Total.Rate'])
mpl.xlabel('State')
mpl.ylabel('Total Rate')
mpl.title('Cancer Rate In various states of USA in a Scatter Graph')
mpl.show()


#The total number of breast cancer and lung cancer in various states of USA represented in a
bar graph
mpl.bar(df['State'], df['Total.Number'])
mpl.xlabel('State')
mpl.ylabel('Total.Number')
mpl.title('Total number of breast cancer and lung cancer in various states of USA')
mpl.show()
```

```
#The Cancer Rate of people below the age of 18 of breast cancer and lung cancer in various
states of USA represented in a bar graph
mpl.bar(df['State'], df['Rates.Age.< 18'])
mpl.xlabel('State')
mpl.ylabel('Cancer Rates at Age< 18')
mpl.title('Cancer Rate of people below the age of 18')
mpl.show()

#The Cancer Rate of people at the ages of 18-45 of breast cancer and lung cancer in various
states of USA represented in a bar graph
mpl.bar(df['State'], df['Rates.Age.18-45'])
mpl.xlabel('State')
mpl.ylabel('Cancer Rates at Ages of 18-45')
mpl.title('Cancer Rate of people at the ages of 18-45')
mpl.show()

#The Cancer Rate of people at the ages of 45-64 of breast cancer and lung cancer in various
states of USA represented in a bar graph
mpl.bar(df['State'], df['Rates.Age.45-64'])
mpl.xlabel('State')
mpl.ylabel('Cancer Rates at Ages of 45-64')
mpl.title('Cancer Rate of people at the ages of 45-64')
mpl.show()

#The Cancer Rate of people at the ages above 64 of breast cancer and lung cancer in various
states of USA represented in a bar graph
mpl.bar(df['State'], df['Rates.Age.> 64'])
mpl.xlabel('State')
mpl.ylabel('Cancer Rates at Ages above 64')
mpl.title('Cancer Rate of people at the ages above 64')
mpl.show()

df2 =pd.read_csv('cancerrate2014-2022.csv')

mpl.plot(df2['Year'], df2['Cancer.Rate'])
mpl.xlabel('Year')
mpl.ylabel('Cancer Rate')
mpl.title('Plot graph showing average increase in Cancer Rate over the years')
mpl.show()
Usa_cancer_rate_yearwise = {
   "2007-2013":"190.65",
   "2014": "143.99",
   "2015": "141.50",
   "2016": "145.80",
   "2017": "146.16",
   "2018": "153.02",
   "2019": "151.29",
   "2020": "156.34",
   "2021": "155.86",
   "2022": "157.81",
}
a = 2.49 #2014-2015
b = 4.3  #2015-2016
c = 0.36 #2016-2017
d = 6.86 #2017-2018
e = 1.73 #2018-2019
f = 5.05 #2019=2020
```

```
g = 0.48 #2020-2021
h = 1.92 #2021=2022

average_2014_2022 = (a+b+c+d+e+f+g+h)/8
y = round(average_2014_2022 , 3)

print("The average increase in Cancer Rate in USA is" + ' ' + str(y))
cr_2023_high = 157.81 + y
cr_2023_low = 157.81 - y
print("So in the coming years the change in the rate of cancer is")
print('Peak value: ' + ' ' + str(cr_2023_high))
print('Least value: ' + ' ' + str(cr_2023_low))
```

3 Conclusions

Based on the project, we can conclude that the cancer incidence varies greatly between states in the USA. The scatter plot shows the overall cancer incidence (breast and lung) for each state, indicating that some states have much higher rates than others. The bar graphs for the total number of breast and lung cancer cases further highlights this variability, with some states having significantly more cases than others whereas the bar graphs for cancer rates at different age groups shows that the incidence rates also vary by age group. States with high rates of cancer overall do not always have high rates in all age groups. This information can be used to identify states with high rates of cancer in specific age groups and aid in targeted public health interventions. Also these visualizations can inform public health policies and interventions aimed at reducing cancer rates and improving access to cancer care and could have implications for public health and policy, as well as for advancing our understanding of cancer and its impact on society. Overall, this code provides a convenient way to visualize and compare cancer incidence across different states and age groups. It helps identify areas that need more attention and resources related to cancer prevention and treatment.

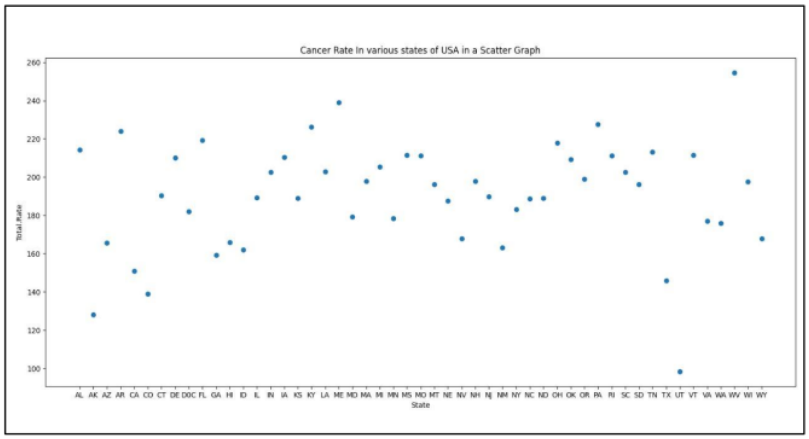

Fig. 1. Cancer rates in various states of USA in a Scatter Graph (Generated by MatPlotLib)

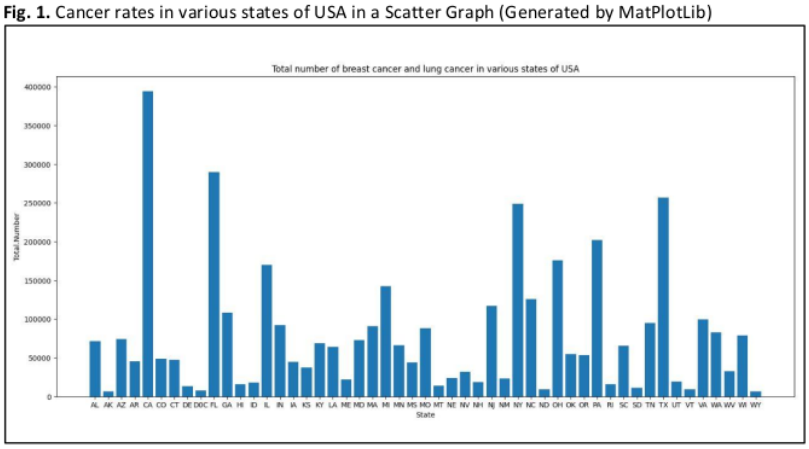

Fig. 2. Total number of breast cancer and lung cancer in USA (Generated by MatPlotLib)

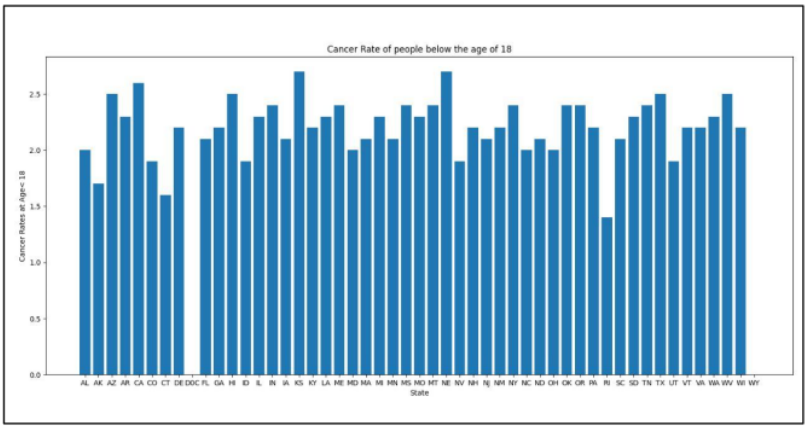

Fig. 3. Cancer rates of people below the age of 18 (Generated by MatPlotLib)

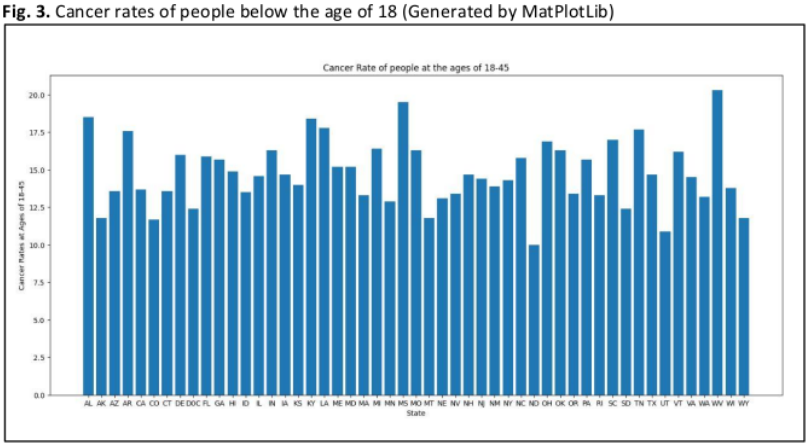

Fig. 4. Cancer rates of people at the ages of 18-45 (Generated by MatPlotLib)

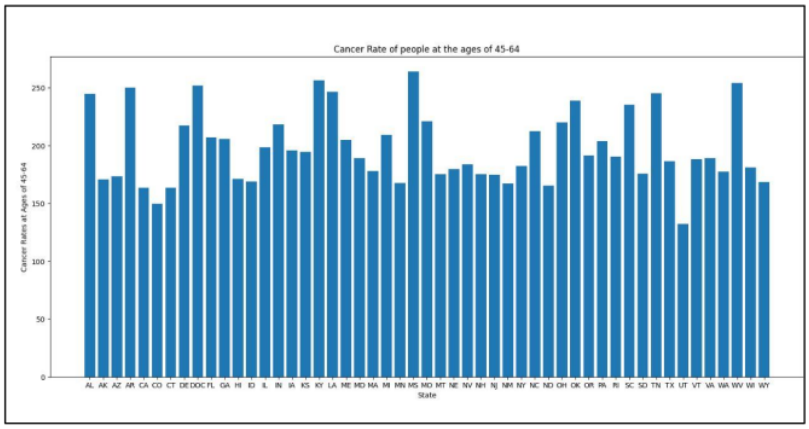

Fig. 5. Cancer rates of people at the ages of 45-64 (Generated by MatPlotLib)

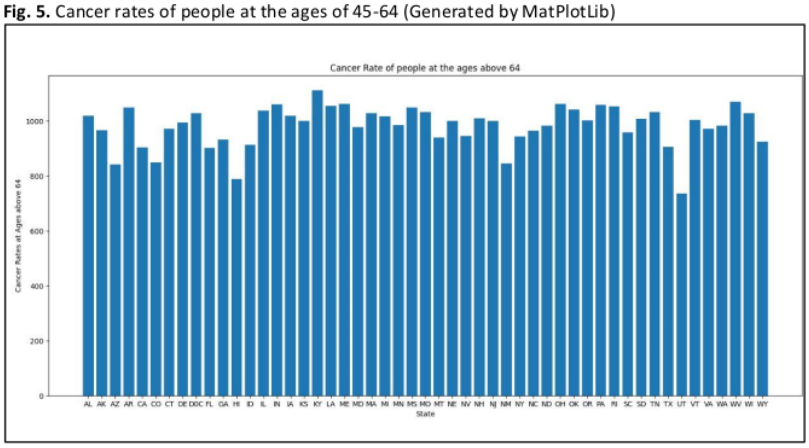

Fig. 6. Cancer rates of people above the age of 64 (Generated by MatPlotLib)

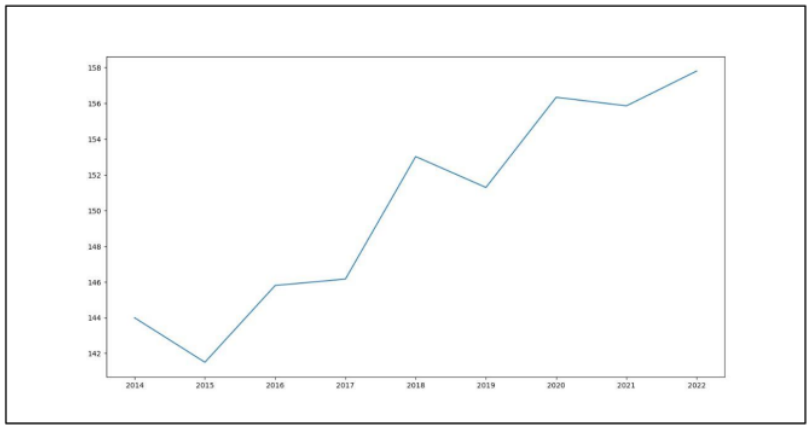

Fig. 7. A Graph showing average increase in Cancer Rates year by year(Generated by matplotlib)

4 References/Bibliography

1. https://www.sciencedirect.com/science/article/pii/B9780323909518000102
2. https://www.docplayer.net
3. https://www.academic.oup.com
4. https://www.maxhealthcare.in/blogs/introduction-to-cancer
5. https://www.cancer.org
6. https://data.worldbank.org

Python Project

ORIGINALITY REPORT

7%	7%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.diablogold.co.uk Internet Source	5%
2	pubmed.ncbi.nlm.nih.gov Internet Source	1%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	kandi.openweaver.com Internet Source	<1%
5	Brian F. Issell, Gertraud Maskarinec, Ian Pagano, Carolyn C. Gotay. "Breast Cancer Treatment Among Women of Different Ethnicity in Hawaii", Cancer Investigation, 2009 Publication	<1%
6	www.firstsydney.com.au Internet Source	<1%

Exclude quotes	Off	Exclude matches	< 2 words

Exclude bibliography On