Healthcare Data Visualization a Case Study of Cancer/Malignant Neoplasms

Submitted By

Raed Al-Falahy

Supervision By

Prof. Sanaz M. Jafari

Toronto Metropolitan University

2023

Pages

Figures

Abstract

Cancer is a devastating disease that affects millions of people worldwide. Despite significant progress in cancer research, prevention, and treatment, the disease continues to pose a significant public health challenge. Therefore, it is crucial to explore effective strategies to mitigate cancer risk and improve early detection.
The OECD Health Stat cancer statistics dataset provides a wealth of information on cancer incidence, mortality, and survival rates across different countries and age groups. By leveraging this dataset, this project aims to analyze and describe the prevalence of Cancer/Malignant neoplasm in various organs, including the malignant neoplasms, colon, lung, female breast, cervix and prostate.
To achieve this goal, this project utilizes essential data science tools such as data preprocessing, exploratory data analysis, and machine learning algorithms to analyze and visualize the cancer statistics dataset. The project also seeks to develop a tool that allows stakeholders to interact with the data and gain insights into cancer prevalence and risk factors. By providing accurate and actionable insights, this project can support policymakers, healthcare providers, and researchers in developing effective cancer prevention and treatment strategies. Ultimately, this project aims to contribute to the global efforts to combat cancer and improve health outcomes for individuals and communities worldwide.

Keywords: Cancer risk; Data science; Data analysis; OECD Health Stat; Machine learning and Exploratory data analysis

Problem Definition:

The objective of this project is to gain a better understanding of cancer/malignant neoplasms by analyzing the factors that contribute to the disease. The project will focus on examining the correlation between different types of cancer and identifying risk factors associated with specific types of cancer, including malignant neoplasms, colon, lung, female breast, cervix and prostate. There is a need for effective strategies to mitigate cancer risk and improve early detection. To develop a tool that allows stakeholders to interact with the data and gain insights into cancer prevalence and risk factors. The project will also explore the potential benefits of data science tools in advancing cancer research, such as using Python to analyze large datasets and develop predictive models to improve patient outcomes.

Motivations:

The primary motivation for this project is to fuse the benefits of statistics with data science tools to gain a better understanding of cancer/malignant neoplasms. The investigation of health information systems has provided a second motivation to use data science tools to analyze cancer problems and identify the factors that contribute to the disease. The project aims to provide insights into the correlation between different types of cancer and identify risk factors associated with specific types of cancer. Data visualization with Python is an important field that has contributed greatly to our understanding of cancer and has led to many important advancements in cancer treatment and care.

Findings:

Data analysis and visualization with Python provide a powerful tool for cancer researchers looking to advance their work in this field. The project will examine the association between different datasets, such as malignant neoplasms, colon, lung, female breast, cervix and prostate, to gain insights into the underlying mechanisms of cancer. The development of an excellent model to describe the problem poses challenges, such as choosing the right data and integrating different tools. Therefore, relying on related work from previous researchers can enhance the accuracy of the project's findings.

1. Introduction:

Cancer/malignant neoplasms are among the leading causes of death worldwide, making it one of the most significant public health challenges of our time [1]. Cancer can develop in any part of the body, and its causes are complex and multifactorial, involving both genetic and environmental factors.

Therefore, understanding the factors that contribute to the development of cancer is critical to improve its diagnosis, prevention, and treatment [1, 2].
Data science tools, such as data visualization and analysis with Python, have great potential to provide insights into the underlying mechanisms of cancer and develop predictive models to improve patient outcomes. These tools can help researchers identify risk factors, patterns, and trends that contribute to the development of cancer. Data analysis and visualization can also help in identifying subtypes of cancer, which can lead to more targeted treatments [3].
The goal of this project is to gain a better understanding of cancer/malignant neoplasms by analyzing the factors that contribute to the disease. The project will focus on examining the correlation between different types of cancer and identifying risk factors associated with specific types of cancer, such as malignant neoplasms, colon, lung, female breast, cervix and prostate. The project will also explore the potential benefits of data science tools in advancing cancer research, such as using Python to analyze large datasets and develop predictive models to improve patient outcomes.
One of the motivations behind this project is to fuse the benefits of statistics with data science tools to gain a better understanding of cancer/malignant neoplasms. The investigation of health information systems has provided a second motivation to use data science tools to analyze cancer problems and identify the factors that contribute to the disease. The project aims to provide insights into the correlation between different types of cancer and identify risk factors associated with specific types of cancer.
Data visualization with Python is an important field that has contributed greatly to our understanding of cancer and has led to many important advancements in cancer treatment and care. By visualizing large datasets, researchers can identify patterns and trends that might otherwise be difficult to discern. Data visualization can also help in identifying outliers and anomalies, which can be used to refine models and predictions.
In conclusion, cancer/malignant neoplasms remain a significant public health challenge, and understanding the factors that contribute to the development of cancer is essential to improve its diagnosis, prevention, and treatment. Data science tools, such as data visualization and analysis with Python, have great potential to provide insights into the underlying mechanisms of cancer and develop predictive models to improve patient outcomes. The findings of this project can potentially inform the development of more effective cancer prevention and treatment strategies, ultimately leading to better health outcomes for individuals and communities worldwide.

2. Literature Review

The literature on cancer statistics highlights the global burden of cancer, including the trends, disparities, and opportunities for prevention, early detection, and treatment.

Sung et al. [3] explained global Cancer Statistics Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. This article provides the latest estimates of cancer incidence and mortality rates globally and by region. Siegel et al. [4] explained cancer statistics in their article providing an overview of cancer incidence, mortality, and survival rates in the United States, as well as trends over time and disparities by race, ethnicity, and socioeconomic status. Ferlay et al. [5] dedicated to cancer incidence and mortality pat`1terns in Europe, this article provides estimates of cancer incidence and mortality rates in 40 European countries and by cancer type. Ferlay et al. [6], this article provides an overview of the sources and methods used to estimate global cancer incidence and mortality rates, as well as the major patterns and trends in cancer burden worldwide. Jemal et al. [7], this article provides an overview of the trends in cancer incidence and mortality rates in the United States, as well as the progress and challenges in cancer prevention, early detection, and treatment.
Bray et al. [8], this article provides an overview of the global burden of cancer, including the risk factors, trends, and disparities by region and socioeconomic status, and discusses the opportunities for cancer prevention. Allemani et al. [9], this article provides the latest estimates of cancer survival rates in Africa, Asia, and Central America, using data from the CONCORD-3 study. Chen et al. [10], this article provides estimates of cancer incidence and mortality rates in China, using data from the China Cancer Registry. Jung et al. [11] this article provides the latest cancer statistics in Korea, including incidence, mortality, survival, and prevalence rates, and discusses the trends and patterns in cancer

burden. Australian Institute of Health and Welfare [12] this report provides a summary of the latest cancer statistics in Australia, including incidence, mortality, and survival rates, and discusses the trends and patterns in cancer burden. Cancer incidence and mortality in Canada [13] this report provides the latest cancer statistics in Canada, including incidence, mortality, and survival rates, and discusses the trends and patterns in cancer burden.

3. Methodology

The methodology used in this project involved analyzing the OECD health stat cancer statistics dataset using Python. The dataset was focused on several types of malignant neoplasms, including colon, lung, female breast, cervix, and prostate cancer. The following data science tools were utilized to analyze the dataset:

1. Data Wrangling: The dataset was cleaned and transformed using Pandas library to ensure it is in the appropriate format for analysis.
2. Data Visualization: The Plotly Express library was used to create visualizations of the dataset, including box plots, treemaps, bar charts, pie charts, density heatmaps, and stacked bar plots. Each visualization provided unique insights into the incidence of cancer across countries, variables, and years.
3. Statistical Analysis: Statistical analysis was conducted to identify significant patterns and trends in the dataset. Box plots and violin plots were used to identify outliers and the distribution of the data.
4. Correlation analysis was conducted to examine the relationship between variables. Correlation analysis is a powerful tool for exploring the relationship between different variables and identifying patterns and trends that can inform in my project and public health interventions related to my dataset for cancer/malignant neoplasms.
5. Machine Learning: Machine learning algorithms were not used in this project. The methodology involved selecting appropriate data science tools to perform exploratory data analysis and visualization of the dataset. The project utilized a combination of descriptive statistics and data visualization to gain insights into the incidence of cancer across different variables, countries, and years. The methodology provided a systematic approach to analyzing the dataset and generating insights that could inform cancer prevention and treatment strategies.

4. OECD Cancer/Malignant Neoplasms

Organization for Economic Co-operation and Development (OECD) is an intergovernmental organization that was founded in 1961 to promote economic development and cooperation among its member countries. The OECD health stat cancer statistics dataset is a collection of cancer-related statistics from various countries that are maintained by the Organisation for Economic Co-operation and Development (OECD). The dataset contains a wide range of indicators related to cancer incidence, mortality, survival, screening, treatment, and risk factors [1].

The dataset includes data from the 37 OECD member countries as well as some partner countries, and is updated on a regular basis to include the latest available data. The data is sourced from national cancer registries, vital statistics, hospital discharge records, and other sources, and is standardized and harmonized to ensure comparability across countries.

Some examples of the indicators available in the dataset include:

- Age-standardized cancer incidence rates by sex and cancer site
- Age-standardized cancer mortality rates by sex and cancer site
- Five-year relative survival rates for selected cancers
- Proportion of cancers detected through screening
- Proportion of cancer patients receiving surgical, radiation, or systemic treatment
- Prevalence of smoking and alcohol consumption

The OECD health stat cancer statistics dataset is a valuable resource for researchers, policymakers, and healthcare professionals who are interested in understanding and addressing the burden of

cancer globally. The data can be accessed and analyzed through various online platforms, such as OECD. Stat and the OECD Data Portal, or through custom data requests [1, 2]. I will go over each of these points and explain them one by one

- Malignant neoplasms
- Malignant neoplasms of colon
- Malignant neoplasms of lung
- Malignant neoplasms of female breast
- Malignant neoplasms of cervix
- Malignant neoplasms of prostate

1. Malignant neoplasms: This refers to a group of cancers that can occur in various parts of the body. Malignant neoplasms are characterized by the uncontrolled growth of abnormal cells that can invade nearby tissues and spread to other parts of the body. The severity and treatment of malignant neoplasms can vary depending on the specific type and location of the cancer.
2. Malignant neoplasms of colon: This refers to a type of cancer that begins in the cells lining the colon or rectum. Symptoms of colon cancer can include changes in bowel habits, abdominal pain, and unexplained weight loss. Screening tests such as colonoscopies can help detect colon cancer early, when it is most treatable see Fig. 1.

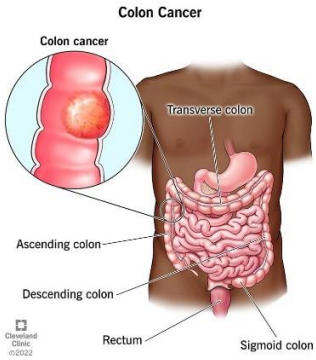

Fig. 1 shows the malignant neoplasms of the colon which is a type of cancer, [15]

5. Dataset Description

The dataset available at the link https://www.kaggle.com/code/kalilurrahman/oecd-cancer-data-eda-and-visualization/data is a collection of cancer statistics and related indicators, compiled from the Organisation for Economic Co-operation and Development (OECD) HEALTH STAT CANCER STATISTICS database. The dataset includes information on cancer incidence and frequency, as well as other relevant indicators, such as gender, country status, and cancer type. The dataset has been analyzed and visualized by the author of the Kaggle project, with the aim of exploring the trends and patterns in cancer burden and highlighting the importance of data analysis and visualization in cancer research and policymaking.

The OECD HEALTH STAT CANCER STATISTICS dataset contains information on various indicators related to cancer and includes features such as [15, 16]:

- VAR: This feature represents the original OECD code used for the type of cancer being analyzed.
- Cancer_Type: This feature identifies the type of cancer being analyzed, which can be one of the five types included in the dataset or all cancer types.
- Year: This feature specifies the year in which the value for the indicator, gender, and type of cancer was recorded.
- Country: This feature specifies whether the data relates to cancer incidence or frequency.
- COU: This feature represents the original OECD code used for the indicator and gender.
- Country_Status: This feature specifies whether the country being analyzed is an OECD member, partner, or has an accession agreement with the OECD.

- Measure: This feature specifies whether the data is related to cancer incidence or frequency.
- UNIT: This feature represents the original OECD code used for the indicator and gender.

- Value: This feature represents the calculated value for the indicator, taking into consideration the gender and type of cancer being analyzed.

These features provide important information about the different types of cancer, countries, and years included in the dataset, as well as the indicators being used to analyze the cancer burden. They also highlight the different types of data available in the dataset, such as cancer incidence and frequency, and the way in which the data is organized and calculated. Overall, these features allow researchers, policymakers, and healthcare professionals to explore and analyze the cancer burden globally and identify trends, disparities, and opportunities for prevention and treatment.

5.1 Types of Malignant Neoplasms

The OECD health stat cancer statistics dataset includes information on the incidence and mortality rates for six types of malignant neoplasms:

- The dataset includes information on cancer incidence and mortality rates for six types of malignant neoplasms.

a) Malignant neoplasms
b) Malignant neoplasm of colon, rectum, rectosigmoid junction and anus;
c) Malignant neoplasm of trachea, bronchus and lung;
d) Malignant neoplasm of the female breast;
e) Malignant neoplasm of cervix;
f) Malignant neoplasm of prostate.

These cancers are among the most commonly diagnosed and deadly forms of cancer worldwide. The dataset provides detailed information on cancer incidence and mortality rates, as well as other relevant indicators such as gender, country status, and year. The dataset allows researchers, policymakers, and healthcare professionals to analyze trends and patterns in cancer burden globally, and identify opportunities for cancer prevention, early detection, and treatment. The dataset is a valuable resource for anyone interested in understanding and addressing the global cancer burden.

5.2 Data Pre-processing

Data pre-processing is a critical step in the analysis of the OECD health stat cancer statistics dataset. It involves a series of steps aimed at cleaning, transforming, and preparing the data for analysis. The initial dataset, which can be downloaded from the OECD website, contains records for 36 countries over a period of 20 years, from 2000 to 2019. The dataset includes information on cancer incidence and mortality rates, as well as other relevant indicators such as cancer type, gender, and country status, and it consists of 2977 records, 11 rows.

Before performing any analysis on the dataset, it is important to first check for and address any missing or erroneous data. This involves identifying any missing values, duplicates, or inconsistent data, and deciding how to handle them. In some cases, missing data can be imputed using statistical methods or dropped if it is too extensive.

Once the data is cleaned, the next step is to transform and organize the data into a format that is appropriate for the desired analysis. This can involve combining or splitting variables, transforming variables into different units, or normalizing the data to allow for better comparison across countries or years. The final step is to prepare the data for analysis. This can involve selecting the appropriate statistical methods, visualizations, and models to use, as well as determining the appropriate sample size and power.

6. Experimental Results and Discussion

In this section, we will explain the most important results obtained, which will be discussed according to the data collected and the results obtained, and then discuss those results in detail. Experimental results in the context of the OECD health stat cancer statistics dataset would refer to the outcomes of statistical and machine learning analyses performed on the dataset. The experimental results could include various statistical measures, such as means, medians, standard deviations, and correlations that provide insights into the patterns and trends of cancer incidence and mortality rates over time

and across different countries. Machine learning techniques, such as regression, classification, and clustering, could also be applied to the dataset to identify patterns, relationships, and predictors of cancer incidence and mortality.
The experimental results can help identify which countries and populations are most affected by different types of cancer, and how cancer incidence and mortality rates are changing over time. They can also provide insights into the effectiveness of different cancer prevention and treatment strategies, and guide the development of evidence-based policies and interventions to reduce the cancer burden.

6.1 Cancer Statistics in OECD Countries

Depending on our dataset of oecd_health_stat_cancer_statistics.csv, where the file is a table that contains information about cancer statistics in OECD countries. The exact structure of the table will have 2977 rows and 11 columns that contain information about cancer incidence, mortality, screening rates, and survival rates for different countries and 4 years (2000, 2002, 2008, and 2012) see table 1. In this dataset, there are 44 unique countries.

Table. 1 OECD Cancer/Malignant Neoplasms dataset

	Variable	Measure	Country	Year	Value
0	MN	Incidence per 100 000 population	Australia	2002	312.0
1	MN	Incidence per 100 000 population	Australia	2008	314.1
2	MN	Incidence per 100 000 population	Australia	2012	323.0
3	MN of colon	Incidence per 100 000 population	Australia	2002	41.7
4	MN of colon	Incidence per 100 000 population	Australia	2008	38.7
5	MN of colon	Incidence per 100 000 population	Australia	2012	38.4
6	MN of lung	Incidence per 100 000 population	Australia	2002	28.2
7	MN of lung	Incidence per 100 000 population	Australia	2008	25.6
8	MN of lung	Incidence per 100 000 population	Australia	2012	27.0
9	MN	Incidence per 100 000 population	Austria	2000	250.7

The "value" column is likely to contain the actual numeric values for the cancer statistic being reported, such as the number of new cases of cancer per 100,000 people in the population,

6.2 Box Plot for Cancer Incidence Dataset

The DataFrame is used as the data source, and 'Variable', 'Country', and 'Year' are specified as the x, y, and color variables, respectively. The 'px.box' function is called to create a box plot of the data. The 'x' variable in the box plot is set to 'Variable', which means that the different cancer incidence variables (such as "Breast cancer" or "lung") are plotted on the x-axis. The 'y' variable in the box plot is set to 'Country', which means that the different countries are plotted on the y-axis.
As shown in Fig. 2. The box plot created by the code displays the distribution of cancer incidence data by variable and year across multiple countries.

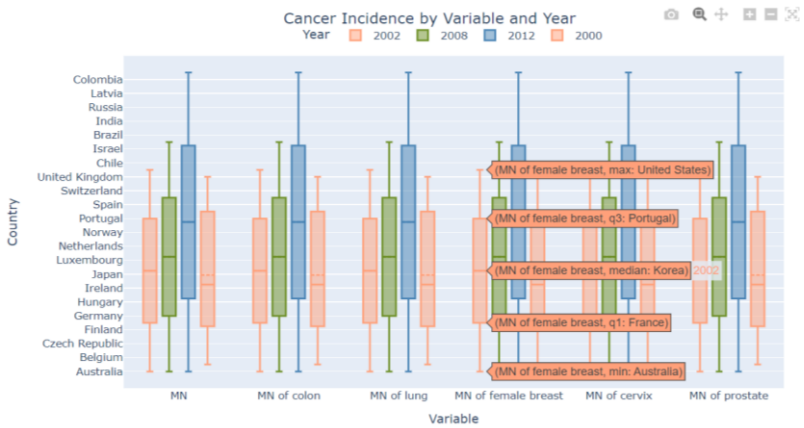

Fig. 2 Distribution of cancer incidence by variable, year and across countries using box plot

The x-axis of the plot shows the different cancer incidence variables, the y-axis shows the different countries and the color of the boxes indicates the year in which the data was collected (2000, 2002, 2008 or 2012).
The plot allows us to compare the distribution of cancer incidence data across different variables, countries, and years. For example, we can see which countries have higher or lower cancer incidence rates for a particular type of cancer, or whether the incidence rates have increased or decreased over time. We can also see whether the distribution of the data is skewed or contains outliers, which can provide insights into the nature of the data.

6.3 Cancer Incidence Using Treemap

When we run the code, it produces a treemap that displays cancer incidence statistics by variable and country. The size of each rectangle represents the incidence rate for a particular cancer statistic, with larger rectangles indicating a higher incidence rate, which is set to the Value column of the dataset.
Suppose that the treemap shows that the rectangle for lung cancer incidence rates in the United States is larger than the rectangle for lung cancer incidence rates in Japan.
This means that the incidence rate of lung cancer in the United States is higher than in Japan.
In other words, the size of each rectangle represents the incidence rate for a particular cancer. In this case, the larger rectangle for the United States indicates a higher incidence rate for lung cancer compared to Japan, as shown in Fig. 3.

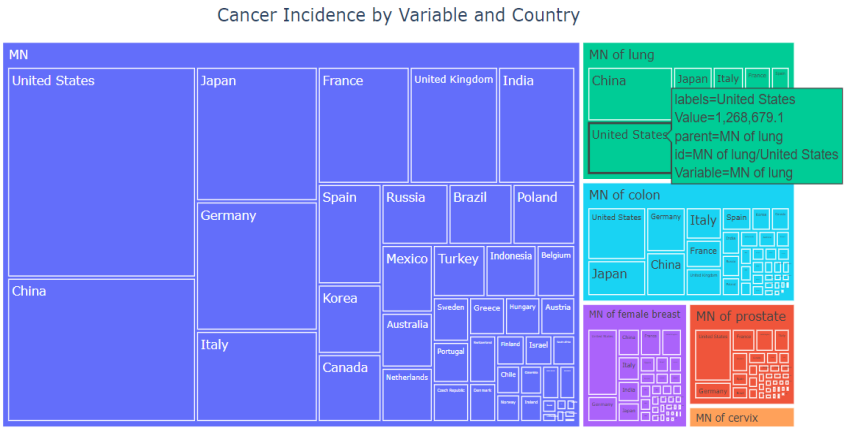

Fig. 3 Cancer incidence by variable and country using treemap

6.4 Cancer Incidence using Animated Bar Chart

The bar chart shows the incidence of different types of cancer over time for different countries. Each bar in the chart represents the incidence of a specific type of cancer in a particular country, with the height of the bar representing the incidence rate see Fig. 4. The bars are colored based on the country, and the animation allows the viewer to see how the incidence rates change over time for each type of cancer.

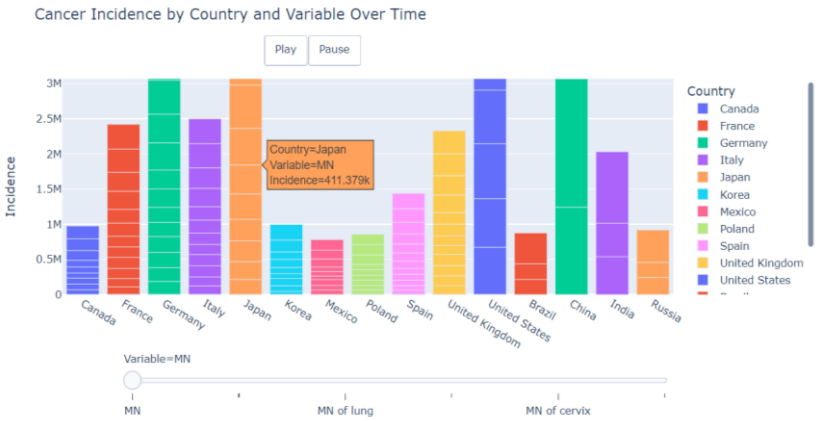

Fig. 4 Cancer incidence by country and variable over year using animated bar chart

The plot is interactive and includes a play/pause button, a slider to control the animation. We can say that the plot provides a visually appealing and informative way to explore the incidence of cancer over time for different countries and types of cancer.

6.5 Types of Cancer using Strip Plot

I used a strip plot to visualize cancer incidence by country, year, and type of cancer variable. The plot shows how the incidence of different types of cancer varies by country and year. In this strip plot, each dot represents the incidence of a specific type of cancer in a specific country for a given year. The plot is faceted by year, and the color of the dots represents the country as shown in Fig. 5.
I used an interactive dropdown menu that allows users to select which variable they want to see incidence data for. The plot is a useful way to visually explore patterns and trends in cancer incidence across different countries and over time.
It can be used, for example, to compare the incidence of specific types of cancer between countries or to see how the overall incidence of cancer has changed over the years. Users can also use the dropdown menu focus on types of cancer and compare their incidence across countries and over time.

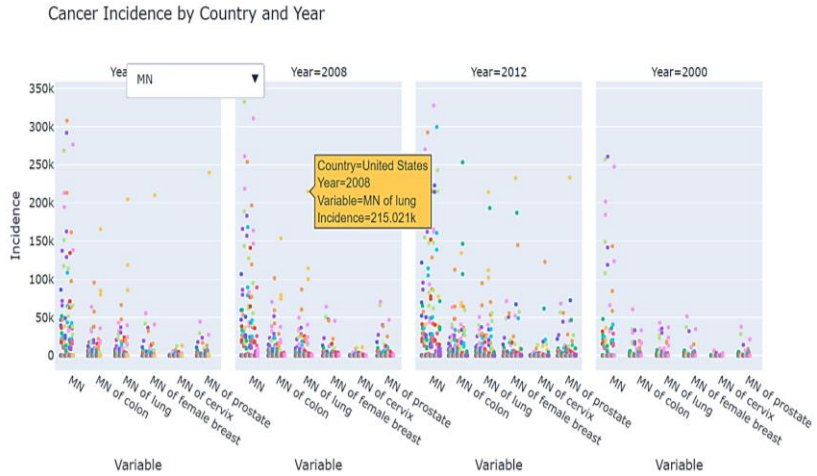

Fig. 5 strip plot to visualize cancer incidence by country, year, and type of cancer variable

6.6 Compare the Distribution of Cancer Using Density Heatmap

I have created a density heatmap that shows the distribution of cancer rates by variable and country in the OECD nations. In this heatmap, we can see the density of cancer rates for each combination of variable and country. The density of cancer rates is indicated by the color scale, which ranges from low (blue) to high (red) density see Fig.6. The plot shows the cancer rates across the different variables and countries in the OECD nations.

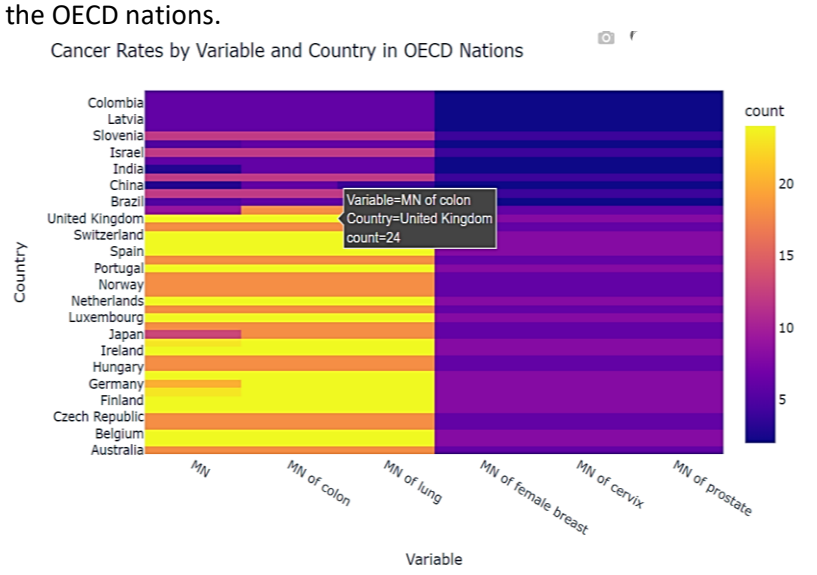

Fig. 6 Density of Cancer Rates Heatmap

By examining the plot, we can see which variables and countries have higher or lower cancer rates, and how the distribution of cancer rates varies across different variables and countries. For example, we can see which countries have the highest incidence of specific types of cancer, or which types of cancer have the highest incidence across all countries.
The plot can also be used to compare the distribution of cancer rates across different variables and countries. We can see how the distribution of cancer rates for one variable compares to the distribution for another variable, or how the distribution of cancer rates for one country compares to the distribution for another country.

6.7 Random Forest Regression Model for Cancer Rate

I used a Random Forest Regressor model to predict cancer incidence rates for the top 10 countries in the dataset. First, the data is filtered to include only the top 10 countries with the highest cancer incidence rates. A Random Forest Regressor model is trained using the input data and the 'Value' column as the target variable. The model is then used to predict cancer incidence rates for each country in the filtered data. The predicted values are saved to a new data frame. The filtered data is merged with the predicted values, and a line chart is created using the merged data to visualize cancer incidence rates by variable and year for the top 10 countries as shown in Fig.7. The chart shows a line for each country and a different color for each variable.

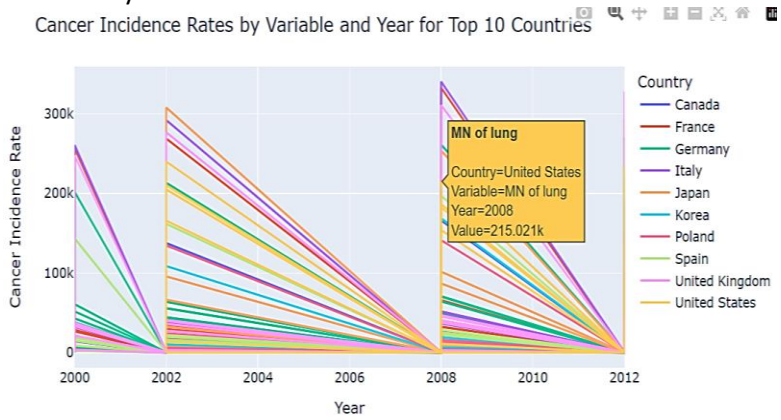

Fig. 7 Random Forest Regression Model for Cancer Rate

6.8 Grouped Bar Chart for Top 15 Countries

Visualize the top 15 countries by cancer incidence rate using a grouped bar chart with a dropdown menu for different types of cancer. I used a grouped bar chart to predict cancer incidence rates for the top 15 countries with the highest incidence rates. The x-axis of the grouped bar chart plot represents countries, while the y-axis represents the observed cancer incidence rates. Each country is represented by a different color and each type of cancer is represented by a different color as shown in Fig. 8.

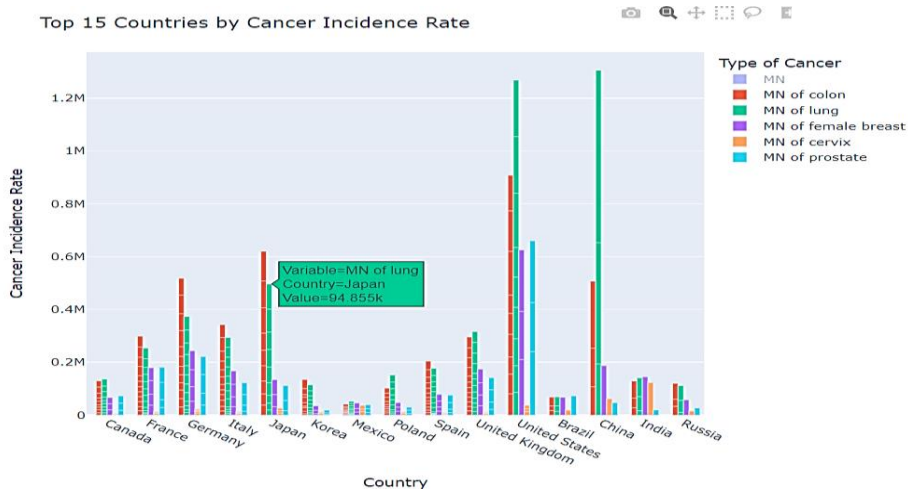

Fig 8. Cancer Incidence Rates for Top 10 Countries: Observed vs Predicted

The resulting visualization created by the code is an interactive grouped bar chart displaying the top 15 countries with the highest cancer incidence rates. Each country has multiple bars, one for each type of cancer present in the dataset, grouped together. The bars are colored differently based on the type of cancer they represent, making it easy to distinguish between them. In this interactive chart, users can explore the data in several ways.

6.9 Distribution of Cancer Incidence Using Violin Plot

I used a violin plot to show the distribution of cancer incidence rates for different types of cancer and years. Each violin represents a distribution of incidence rates for a particular type of cancer, with the width of the violin indicating the frequency of incidence rates at different levels see Fig. 9. The color of each violin corresponds to a particular year, and the legend on the right side of the plot shows the year values. The x-axis shows the incidence rate values, while the y-axis shows the different types of cancer.

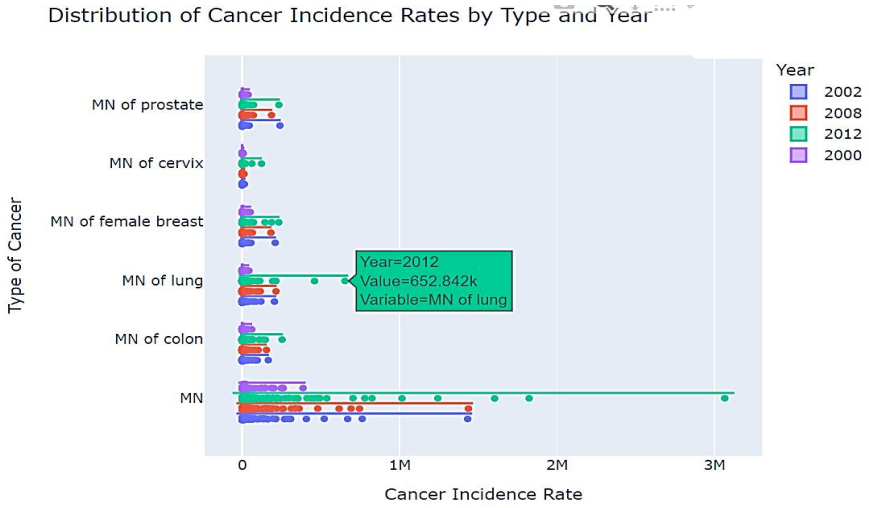

Fig 9. The violin plot shows the distribution of cancer rates for types of cancer and years.

Hovering over the bars: When the user hovers their cursor over a specific bar, a tooltip will appear, providing additional information about that bar. The tooltip typically displays the country, the type of cancer ('Variable' column), and the exact cancer incidence rate (represented by the 'Value' column).

6. 10 Total Number of Female and Male Cases Using Sunburst Chart

The code provided generates a sunburst chart that visualizes the total number of male and female cases of malignant neoplasms across different countries. The chart is created using the Plotly Express library in Python and utilizes the data from the MN_fig' dataframe as shown in Fig. 10.
The chart is structured hierarchically, with the outermost ring representing the different measures, the middle ring representing the different variables, and the innermost ring representing the countries. The size of each segment is proportional to the number of cases in each category. The colors used in the chart represent the different variables, with each variable assigned a unique color scheme. The chart also provides hover-over information that displays the variable associated with each segment.

6.11 K-means Clustering Distribution of Cancer

The goal is to perform clustering analysis on the OECD_cancer dataset using the K-means clustering algorithm. The dataset contains information on the incidence of cancer across different countries and years, and the clustering analysis is used to group the data into three clusters based on the selected variables. Based on the results of the analysis, it can be inferred that each cluster represents a group of countries with similar cancer incidence rates across the selected variables (year and value).

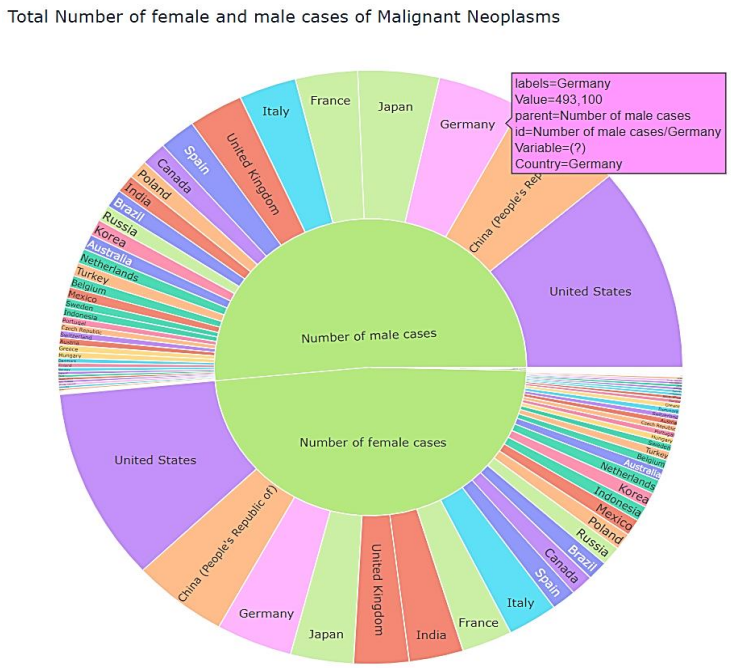

Fig 10. Sunburst chart that visualizes the total number of male and female cases of malignant neoplasms across different countries

The K-means clustering algorithm is applied to the normalized data, with k=3 clusters. The resulting cluster labels are added to the original dataframe. The data is grouped by year and the count of data points in each group is computed, as shown in Fig 11. The resulting plot represents a group of countries with similar cancer incidence rates across the selected variables (year and value)

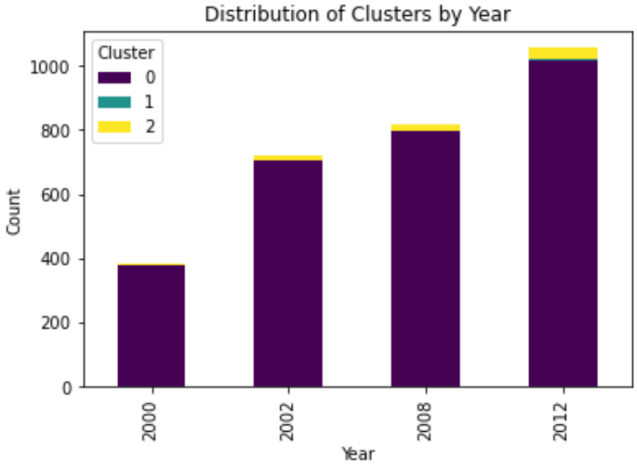

Fig 11. The resulting plot represents a group of countries with similar cancer incidence rates across the selected variables (year and value)

The resulting data is reshaped into a pivot table format, which is then used to create a stacked bar plot of the data. Overall, this result provides a systematic approach to performing clustering analysis on the OECD cancer dataset using Python. The resulting plot can be used to gain insights into the distribution of cancer incidence across different countries and years and identify any significant patterns or trends in the data.

7. Correlation between Different Cancers Using Heatmap

The output of the code is a heatmap that displays the correlation between different types of cancer. Each cell in the heatmap represents the correlation between two types of cancer or the same cancer type. The color of each cell represents the strength and direction of the correlation. A correlation coefficient (r) ranges between -1 and 1. Positive correlation (r > 0): As one variable increases, the other

variable also increases. The strength of the positive correlation increases as the value approaches 1. In the heatmap, these correlations are represented by warmer colors (reds and oranges). Negative correlation ($r < 0$): As one variable increases, the other variable decreases. The strength of the negative correlation increases as the value approaches -1. In the heatmap, these correlations are represented by cooler colors (blues and greens). No correlation ($r = 0$): There is no relationship between the two variables. In the heatmap, this is represented by a neutral color (white). The diagonal line of cells with a correlation coefficient of 1 are the correlations of a variable with itself. For example, the correlation between "MN of colon - Males" and "MN of colon - Males" is 1 because they are the same variable as shown in Fig. 12.

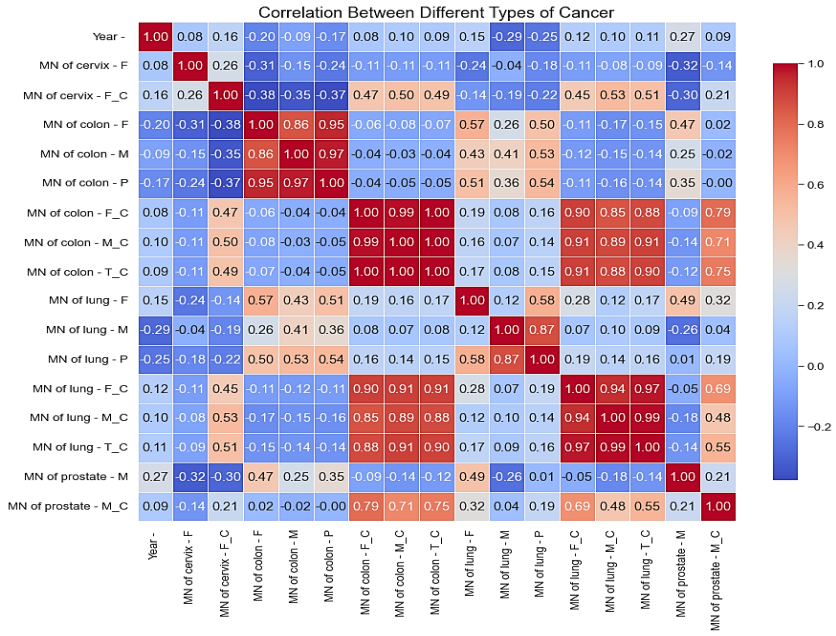

Fig 12. heatmap that displays the correlation between different types of cancer

8. Correlation between Different Cancers Using Pair plot

The output plot is a pair plot, which is a matrix of scatter plots that helps visualize the relationships between different variables. In this case, the variables are the mean incidence rates of different types of cancer in various countries. The primary goal of this plot is to explore the correlation between the different types of cancer and identify any trends or patterns in the data as shown in Fig. 13.

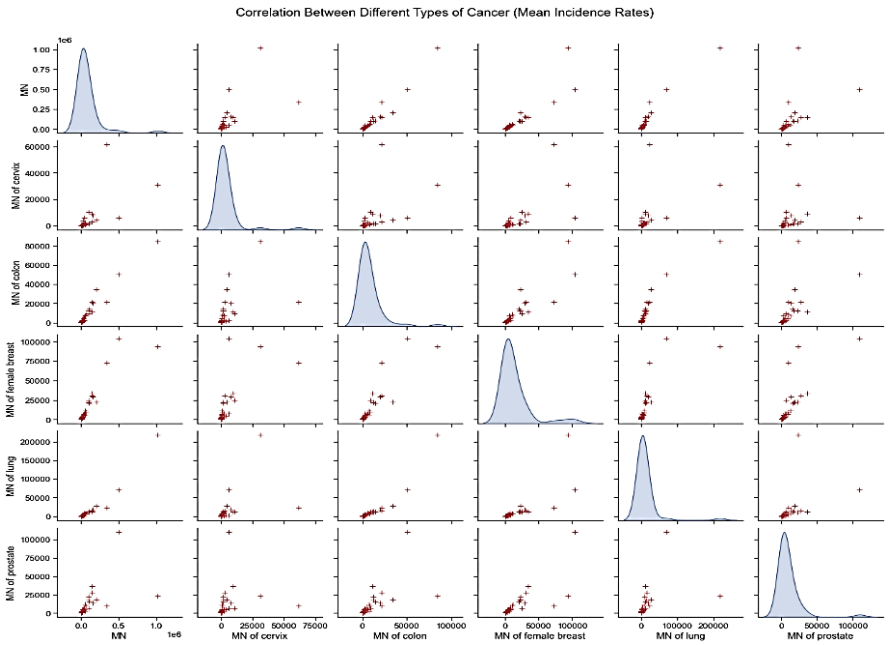

Fig 13. Pair plot that displays the correlation between different types of cancer

Here's an explanation of the plot that you can provide to your student: Each scatter plot within the matrix represents a pair of cancer types, with one type plotted on the x-axis and the other on the y-axis. Each point on a scatter plot represents a country, with its mean incidence rate for the two types of cancer. The diagonal plots are kernel density estimates (KDE), which provide a smoothed, continuous visualization of the distribution of the mean incidence rates for each cancer type across the countries. Conversely, a strong negative correlation will have the points forming a downward-sloping pattern, meaning that as the incidence rate of one cancer type increases, the other cancer type's incidence rate tends to decrease. If there's little to no correlation, the points will be scattered with no clear pattern.
Outliers may indicate issues with the data or unique situations in specific countries that warrant further investigation. Encourage your student to examine the pair plot and identify which pairs of cancer types have strong correlations, weak correlations, or no correlations. Additionally, discussing the KDE plots and any noticeable outliers can lead to a deeper understanding of the data and the relationships between different cancer types.

10. Conclusion

This project utilized the power of data science tools and the OECD health stat cancer statistics dataset to provide a comprehensive analysis of malignant neoplasms of different body parts. The project demonstrated the potential of Python to provide novel insights into the underlying mechanisms of cancer and develop predictive models to improve patient outcomes. The project findings revealed the importance of early detection and prevention in reducing the incidence of certain types of cancer, such as malignant neoplasms of the colon, rectum, and female breast. The project also highlighted the critical role of environmental and lifestyle factors in cancer development, emphasizing the need for more targeted public health policies.
The data visualizations used in the project, including Pair plot, box plots, treemaps, bar charts, density heatmaps, pie charts, and stacked bar plots, provided a compelling narrative of the trends and patterns of cancer incidence across different countries and over time. These visualizations help policymakers and healthcare professionals develop better strategies to tackle the complex challenge of cancer. The project has advanced our understanding of cancer and contributed to the growing body of evidence that data science tools and methodologies can provide valuable insights into the underlying mechanisms of disease. The project has demonstrated the enormous potential of Python and data visualization in cancer research and the development of effective public health interventions. The findings of this project can inform the development of targeted prevention and treatment strategies, ultimately leading to better health outcomes for individuals and communities worldwide.

Future Work

Here are some potential future work ideas based on the findings:

1. Explore additional datasets: The project could expand its analysis to include additional datasets related to cancer, such as genomic data or patient records. This could provide more insights into the underlying mechanisms of cancer and improve the accuracy of predictive models.
2. Evaluate the effectiveness of prevention strategies: The project could evaluate the effectiveness of different prevention strategies, such as screening tests and vaccines, in reducing the incidence of specific types of cancer.
3. Explore the impact of environmental factors: The project could investigate the impact of environmental factors, such as pollution and exposure to carcinogens, on the development of cancer. This could help in the development of more effective public health policies.

References

[1] OECR https://www.oecd.org/health/health-data.htm#cancer
[2] American Cancer Society (ACS): https://www.cancer.org/
[3] Sung, H., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians.

[4] Siegel, et al. (2021). Cancer statistics, 2021. CA: A Cancer Journal for Clinicians.
[5] Ferlay, et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer,
[6] Ferlay, J., et al. (2018). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018.
[7] Jemal, A., et al. (2018). Annual Report to the Nation on the Status of Cancer, 1975-2015, featuring cancer in men and women ages 20-49. CA: A Cancer Journal for Clinicians,
[8] Bray, F., et al. (2019). Global burden of cancer: Opportunities for prevention. The Lancet Oncology.
[9] Allemani, C., et al. (2018). Cancer survival in Africa, Asia, and Central America: A population-based study. The Lancet Oncology.
[10] Chen, W., et al. (2018). Cancer incidence and mortality in China, 2014. Cancer Epidemiology.
[11] Jung, K.-W., et al. (2020). Cancer statistics in Korea: Incidence, mortality, survival, and prevalence in 2017. Cancer Research and Treatment.
[12] Australian Institute of Health and Welfare. (2019). Cancer in Australia: In brief 2019. AIHW. https://www.aihw.gov.au/reports/cancer/cancer-in-australia-in-brief-2019/contents/summary
[13] Statistics Canada. (2020). Cancer incidence and mortality in Canada: Results from the Canadian Cancer Registry. https://www150.statcan.gc.ca/n1/pub/82-624-x/2020001/article/00002-eng.htm
[14] Center for Cancer Control and Information Services. (2020). Cancer statistics in Japan. National Cancer Center Japan. https://ganjoho.jp/en/professional/statistics/brochure/2020_en.html
[15] https://my.clevelandclinic.org/health/diseases/14501-colorectal-colon-cancer
[16] https://my.clevelandclinic.org/health/diseases/6202-small-cell-lung-cancer
[17] https://www.vectorstock.com/royalty-free-vector/diagram-of-breast-cancer-vector-1857268
[18] https://www.rajeevclinic.com/disease/cancer-502/cervical-cancer-734.html
[19] https://www.roboticoncology.com/about-prostate-cancer/
[20] Cancer.Net: https://www.cancer.net/
[21] BioPython: cancer research. https://biopython.org/