Data Visualization Dashboard for Cancer Death Data

Technical Report: Final Project DS 5110:
Introduction to Data Management and Processing

Team Members:
Brinton Cheng, Rich Goodier

Khoury College of Computer Sciences
Data Science Program
July 31, 2024

Abstract

This technical report presents the development and utilization of a data visualization dashboard designed to explore global cancer death trends by country and region from 1990 to 2016. Utilizing a dataset sourced from Kaggle, which includes records from the World Health Organization (WHO), this project aims to provide an intuitive platform for users to identify patterns and outliers in cancer mortality data across 27 cancer types in 184 countries and 38 regions. Our analysis reveals significant regional disparities, with Asia showing disproportionately high cancer mortality for lung and liver cancers, possibly influenced by genetic and lifestyle factors. Additionally, temporal trends indicate an overall increase in cancer deaths, although certain types like stomach cancer have seen a decrease, likely due to medical advancements. The report discusses the methodology, key findings, limitations due to data quality and undocumented collection methodologies, and suggests areas for future research, including enhanced data collection, integration of population data, and a deeper examination of lifestyle and genetic factors. Through this project, we aim to contribute to the ongoing efforts in understanding and combating cancer, supporting informed decision-making in public health.

Table of Contents

List of Figures and Tables

Introduction

Cancer remains one of the leading causes of death worldwide, with significant variations in incidence and mortality rates across different regions and types. Understanding these variations is crucial for healthcare stakeholders to formulate effective strategies for prevention, treatment, and resource allocation. This project aims to develop a feature-rich data visualization dashboard to explore trends and patterns in cancer deaths from 1990 to 2016 from 27 different cancer types from 184 countries around the globe. Leveraging a dataset from Kaggle, sourced from the World Health Organization (WHO), this project addresses the challenge of visualizing complex, multidimensional data to derive actionable insights. The dataset comprises records of total cancer deaths segmented by country, year, and type of cancer, offering a rich foundation for analysis.

Our objectives are twofold:

1) We aim to provide an intuitive platform for users to explore the data, identifying key patterns and outliers.
2) We seek to test the hypothesis that certain cancers exhibit distinct trends in specific regions, influenced by factors such as healthcare systems, lifestyle, and genetic predispositions. Additionally, we anticipate observing significant changes in cancer deaths over time, reflecting advancements in medical treatments and diagnostics.

The scope of this project encompasses data exploration, preprocessing, and the development of a Flask-based web application. The application will feature visualizations, including scatter plots and histograms, allowing users to filter and sort the

data dynamically. Through this dashboard, we aim to contribute to the ongoing efforts in understanding and combating cancer, ultimately supporting informed decision-making in public health.

Before embarking on the visualization dashboard, we consider the state of the research and the need for easy-to-use public visualization tools. We then delve into the methodology in creating such a tool, including data collection, pre-processing, and analysis. The analysis of the data will be conducted by utilizing the visualization dashboard itself, considering ten questions, and thus demonstrating the tool's effectiveness.

Literature Review

Summary of Relevant Existing Research in Cancer Deaths

Research on global cancer trends highlights significant variations in cancer incidence and mortality across regions and over time. Studies from the Global Burden of Disease (GBD) project show notable increases in non-melanoma skin cancer, thyroid cancer, and nasopharynx cancer, while the incidence of stomach, liver, and esophageal cancers has declined. These studies also emphasize disparities between high and low socio-demographic index (SDI) regions, influenced by healthcare access, lifestyle factors, and genetic predispositions.[1]

Additional research has focused on the impact of metabolic risk factors, such as high fasting plasma glucose (HFPG), on cancer mortality. This factor has become increasingly significant over the past three decades, alongside traditional risk factors like tobacco use, diet, and alcohol consumption.[2] Studies from the International Agency for Research on Cancer (IARC) provide further insights into specific cancer trends, such as the rising incidence of breast cancer in high-SDI countries and varying mortality rates for colorectal and prostate cancers based on regional healthcare quality.[3]

1 Lin L, Li Z, Yan L, Liu Y, Yang H, Li H. Global, regional, and national cancer incidence and death for 29 cancer groups in 2019 and trends analysis of the global cancer burden, 1990-2019. J Hematol Oncol. 2021 Nov 22;14(1):197. doi: 10.1186/s13045-021-01213-z. PMID: 34809683; PMCID: PMC8607714.
2 Ibid.
3 Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends--an update. Cancer Epidemiol Biomarkers Prev. 2016 Jan;25(1):16-27. doi: 10.1158/1055-9965.EPI-15-0578.

Discussion of Methodologies, Findings, and Gaps in the Literature

Data collection methodologies are difficult to unearth in the literature. For analyzing cancer trends, however, well-known epidemiological statistical models are commonly used. These models help differentiate the effects of age, period, and cohort on cancer incidence and mortality, drawing data from cancer registries, hospital records, and health surveys.

Findings:

- Higher cancer rates in high-income countries due to better detection and reporting.
- Major contributors to cancer deaths include smoking, diet, alcohol, and metabolic factors.
- Gender disparities. Men/Women experiencing higher cancer mortality rates due to genetic and environmental factors.

Gaps in the Literature:

- Inconsistent data collection methods across countries affect the reliability of global analyses.
- Limited detailed documentation on data collection methodologies.
- Sparse data from low- and middle-income countries introduce biases in global trends.

Need for Public Visualization Tools

There is a clear need for publicly available, easy-to-use platforms to view and visualize cancer death data. Such tools would enhance accessibility, allowing researchers, healthcare professionals, and policymakers to explore trends, identify risk factors, and make informed decisions to improve cancer outcomes globally. Furthermore, a publicly available visualization tool would increase awareness among the public and spur further action at the private and public level. Our project aims to address this need by developing an interactive visualization dashboard that provides valuable insights into cancer trends.

Methodology

This section outlines the methods and techniques used to create the visualization dashboard, detailing the data collection, preprocessing, and analysis processes, and the tools and software used.

1. Description of Methods and Techniques Used to Create the Visualization Dashboard

To provide an intuitive and interactive platform for visualizing trends in cancer deaths, we opted for a Flask web interface. Flask was chosen for its simplicity and flexibility, which benefit both the programmer and the end-user. Its lightweight nature makes it easy to set up and extend, while its robust framework supports the integration of various features needed for our project.

Task Distribution:

- Flask Skeleton: Rich was responsible for setting up the basic structure of the Flask application.
- Initial EDA, Preprocessing, Cleaning, and Importing to Flask Framework: Brinton and Rich collaborated on the initial exploratory data analysis (EDA), preprocessing, cleaning the dataset, and importing it into the Flask framework.
- Index.html Features: Rich added filtering and sorting functionalities to the index.html file.
- Custom Plotting Functions: Brinton developed custom plotting functions tailored to our dataset's needs.
- Integration of Plotting Functions: Rich integrated these plotting functions into the Flask framework, ensuring smooth interaction between the data and visualizations.
- Cleaning and Testing: Both Brinton and Rich worked on cleaning and testing the structure and code on different operating systems (Windows and Linux) to prepare it for production.

2. Details on Data Collection, Preprocessing, and Analysis

a) Data Collection

The dataset used in this project is sourced from Kaggle, titled "Cancer Deaths by Country and Type (1990-2016)." The data was collected by Albert Antony and authored by Jason Wu, with the primary source being the World Health Organization (WHO). The WHO compiles data on cancer deaths from various countries worldwide. However, the collection methodology is not detailed in the Kaggle description.

- License: Public Domain
- Expected Update Frequency: Annually (last updated a year ago)
- Provenance: Collected from the WHO

b) Preprocessing

The initial dataset was relatively clean, but we implemented several preprocessing steps to ensure consistency and usability, as outlined in our `clean_data` function.

Key preprocessing steps included:

- Removal of Redundant Columns: We removed the 'Code' column as it was redundant, opting to use country names as our keys.
- Exclusion of Non-Country Rows: The dataset included various regions and subregions. We focused on countries recognized by the UN, excluding these non-country entries.
- Conversion to Integer Values: All cancer death numbers were converted to integers to simplify further downstream tasks as well as analysis and visualization.

c) Analysis

The analysis of the dataset was conducted primarily through our visualization dashboard, demonstrating its effectiveness in viewing and interpreting the data. Using the "Final Questions" provided by Professor Nafa as a guide, we generated tables and plots to analyze the data and formulate hypotheses. Detailed analysis results are presented in the Results and Discussion sections.

3. Tools and Software Used

The following tools and software were used to create the visualization dashboard and perform the analysis:

- Pandas: For data manipulation and analysis.
- NumPy: For numerical operations.
- Matplotlib and mpl_toolkits.mplot3d: For plotting 2D and 3D visualizations.
- Seaborn: For advanced visualizations.
- Flask: For building web applications.
- Other Libraries: Various other basic Python libraries, such as `os`,were used to support the functionality of the application.

The combination of these tools allowed us to create an interactive and user-friendly dashboard, providing valuable insights into cancer death trends across different countries and types from 1990 to 2016.

Results

- The dataset that we studied contains data from 27 types of cancers, 184 countries, 38 regions (Western Europe, Southeast Asia, etc), and across 27 years (1990 ~ 2016). Showing all the data in one plot will certainly look ugly.
- The dataset looks like such:

Country	Year	Liver cancer	Kidney cancer	... cancer
Afghanistan	1990	243.6637	39.47049	...
Afghanistan	...	...	...	...
Afghanistan	2016	...	...	...
Western Europe	1990	...	...	...
Western Europe	2016			
World	1990	...	...	...
World	...	...	...	...
World	2016	...	...	...

- By analyzing data of "World" in the "Country" column, we were able to rank cancers by mortality numbers, and the top 5 cancers are
    1. Tracheal, bronchus, and lung cancer
    2. Stomach cancer
    3. Colon and rectum cancer
    4. Liver cancer
    5. Breast cancer
- We process and combine the dataset down to top 5 cancers and 5 continents for easier general data analysis. Data of continents is the sum of data of corresponding cancer and year from regions within the continent.

1. How do cancer mortality rates vary across different countries and regions from 1990 to 2016?

- It is generally increasing. Take “Tracheal, bronchus, and lung cancer” as an example. In figure-1, we can see the trend of increasing number of deaths in varies slope, mostly contributed by Asia area.

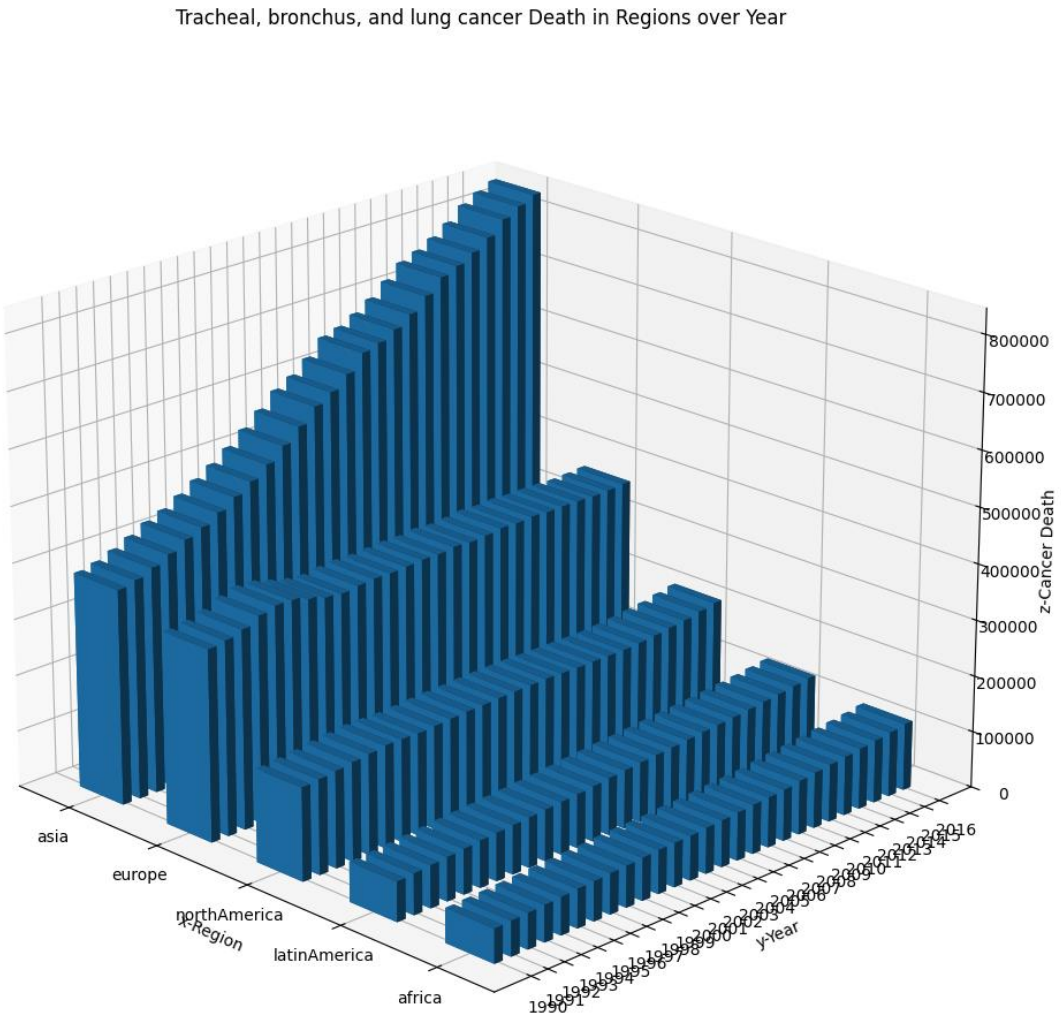

Figure 1. Lung cancer mortality number by Continents over years.

2. Which countries or regions show the highest and lowest cancer mortality rates, and what factors might contribute to these differences?
    - Based on the dataset, we are unable to answer questions about mortality **rate** due to the lack of data of total population of each country or regions over this period of year. From data analysis, top 5 cancers with highest mortality **numbers** are
        1. Tracheal, bronchus, and lung cancer
        2. Stomach cancer
        3. Colon and rectum cancer
        4. Liver cancer
        5. Breast cancer
    - According to figure-2 and figure-3, in 1996 and 2016, Asia has the highest mortality number in all the top 5 cancers.

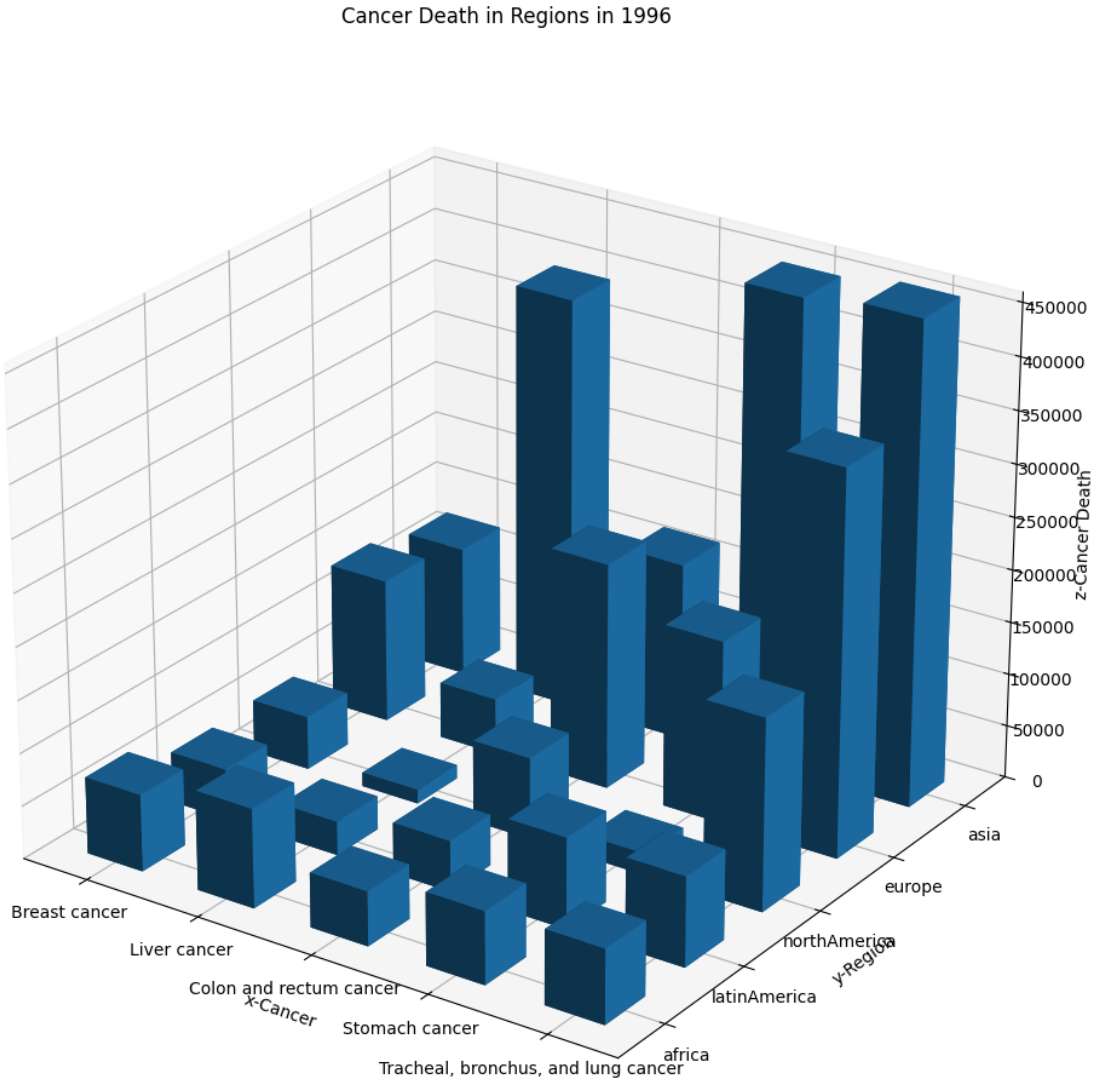

Figure 2. Mortality number by cancers and Continents in 1996

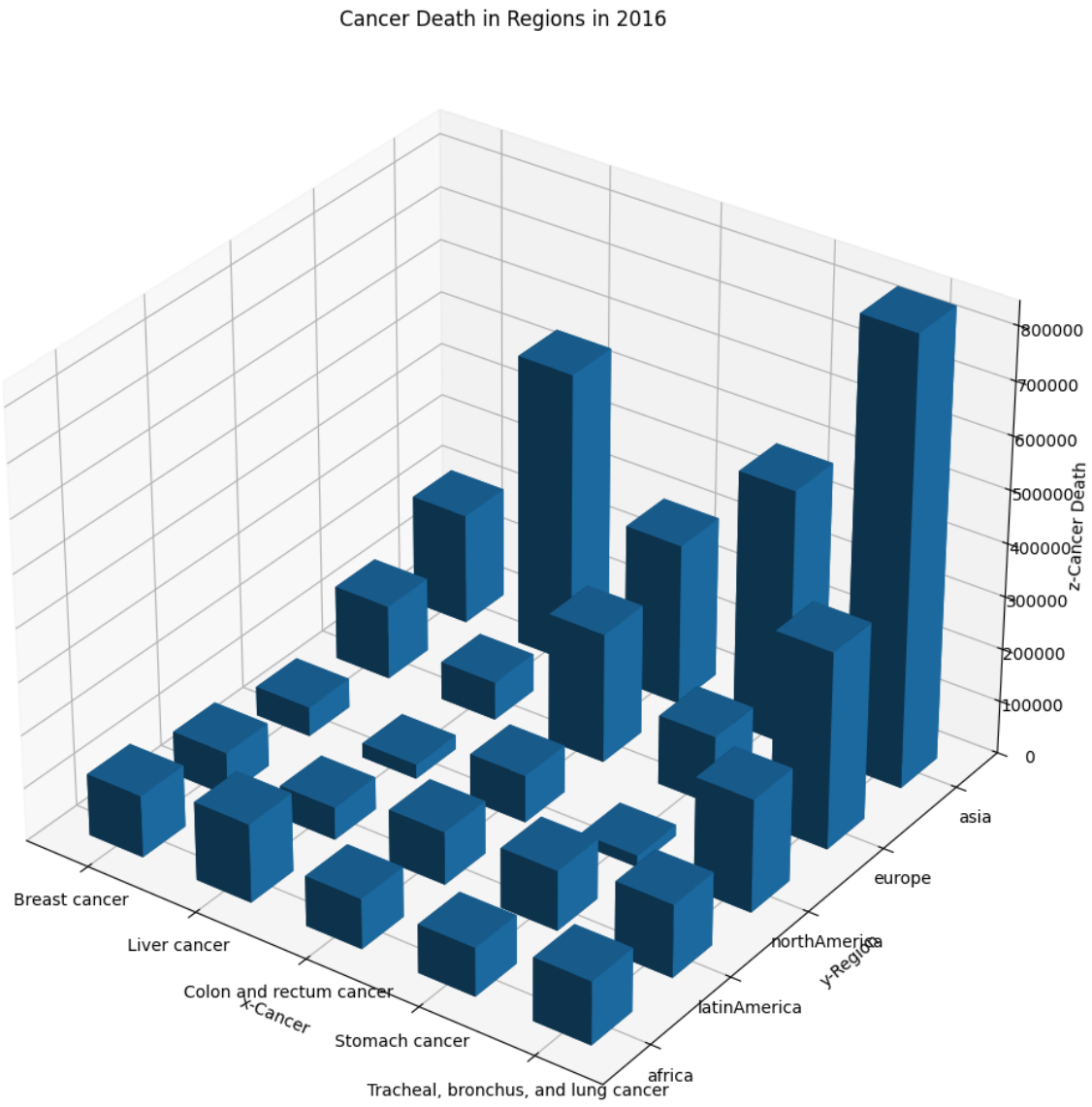

Figure 3. Mortality number by top 5 cancers and Continents in 2016

3. What are the trends in cancer mortality rates over time for various types of cancers from 1990 to 2016?
    - Figure-4 shows all 27 cancers data which is very messy.

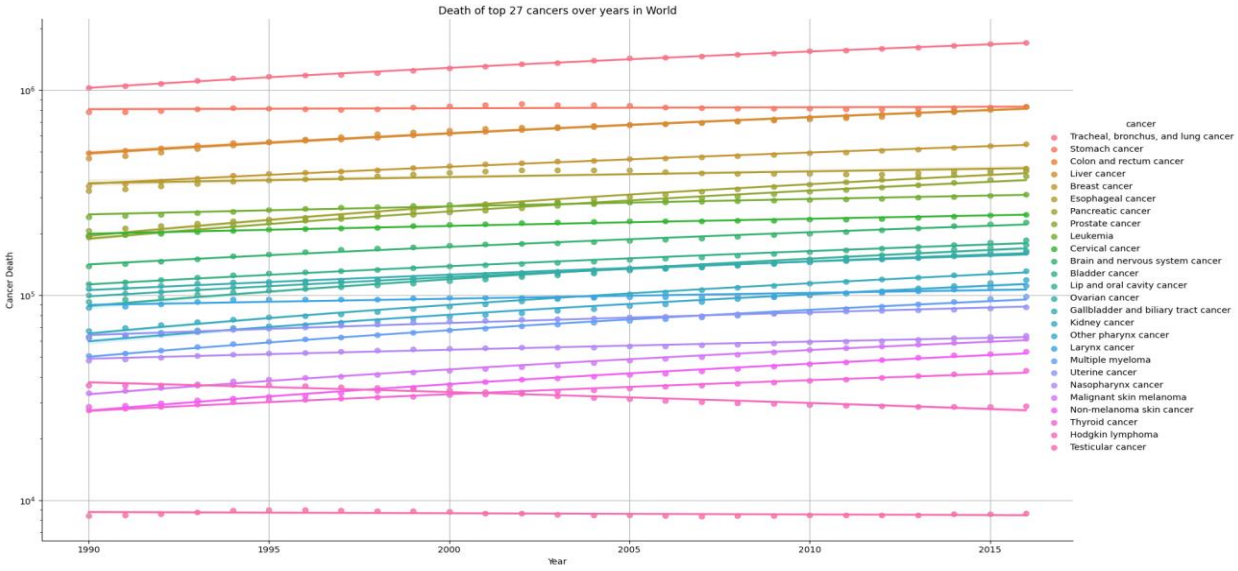

Figure 4. world mortality number of all cancers over years

- The number of mortalities increases over time in general. Figure-5 is the regression plot of the top 5 cancers mortality number over time. Unfortunately, Seaborn does not provide the slopes of regression lines.
- Google provides a (not so reliable) population growth rate of 30% between 1990 to 2010. If taking 30% as a reference, in figure-5, we can roughly conclude that lung cancer mortality rate is positive, liver, colon, and breast cancer mortality rate is close to 0, and stomach cancer mortality rate is negative.

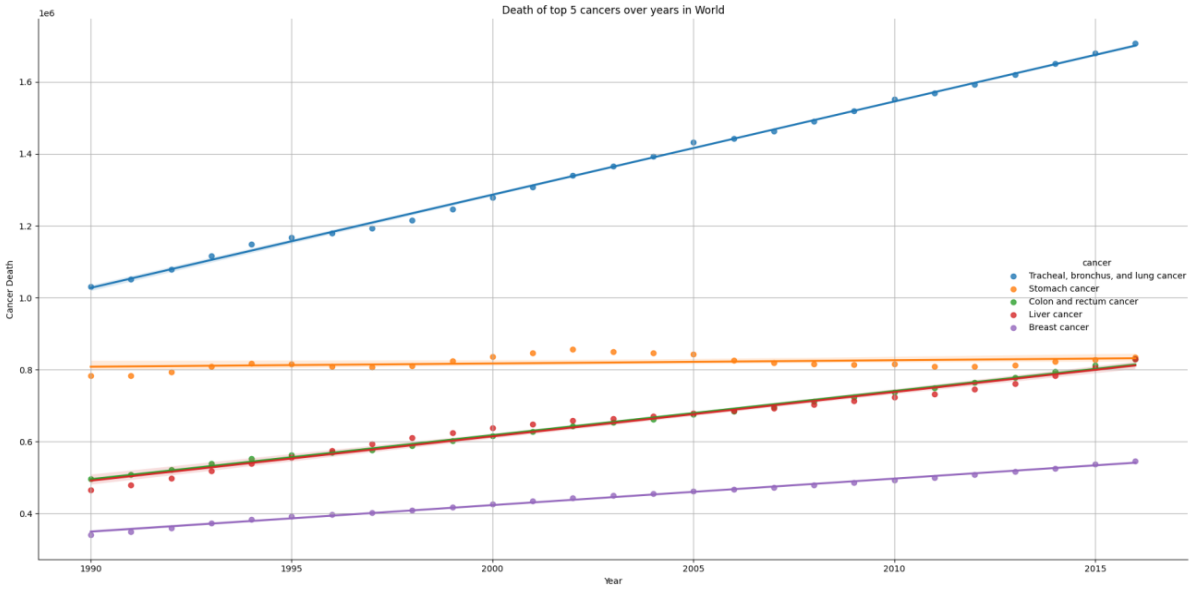

Figure 5. world mortality number of top 5 cancers over years

4. Are there any significant changes in mortality rates at certain time points, possibly due to advancements in medical treatments or changes in healthcare policies?
    - Referring to figure-5 again. The stomach cancer mortality rate in fact is negative (since we know world population growth rate is positive). This may indicate some degree of medical treatment or policy advancement.
    - On the other hand, lung cancer is growing faster than the population growth rate. Possible explanations are the worsening air pollution, growing cigarette users, and the increase in average life expectancy.

5. How do mortality rates differ among various types of cancers across different countries?

   According to figuer-3, we can see a few points.

   - Lung cancer death is number one across all continents except Africa.
   - Liver cancer death is unproportionally high in Asia compare with other continents.
   - Colon and rectum cancer death is the second highest in Europe and North America, but not as high by proportion in other continents.

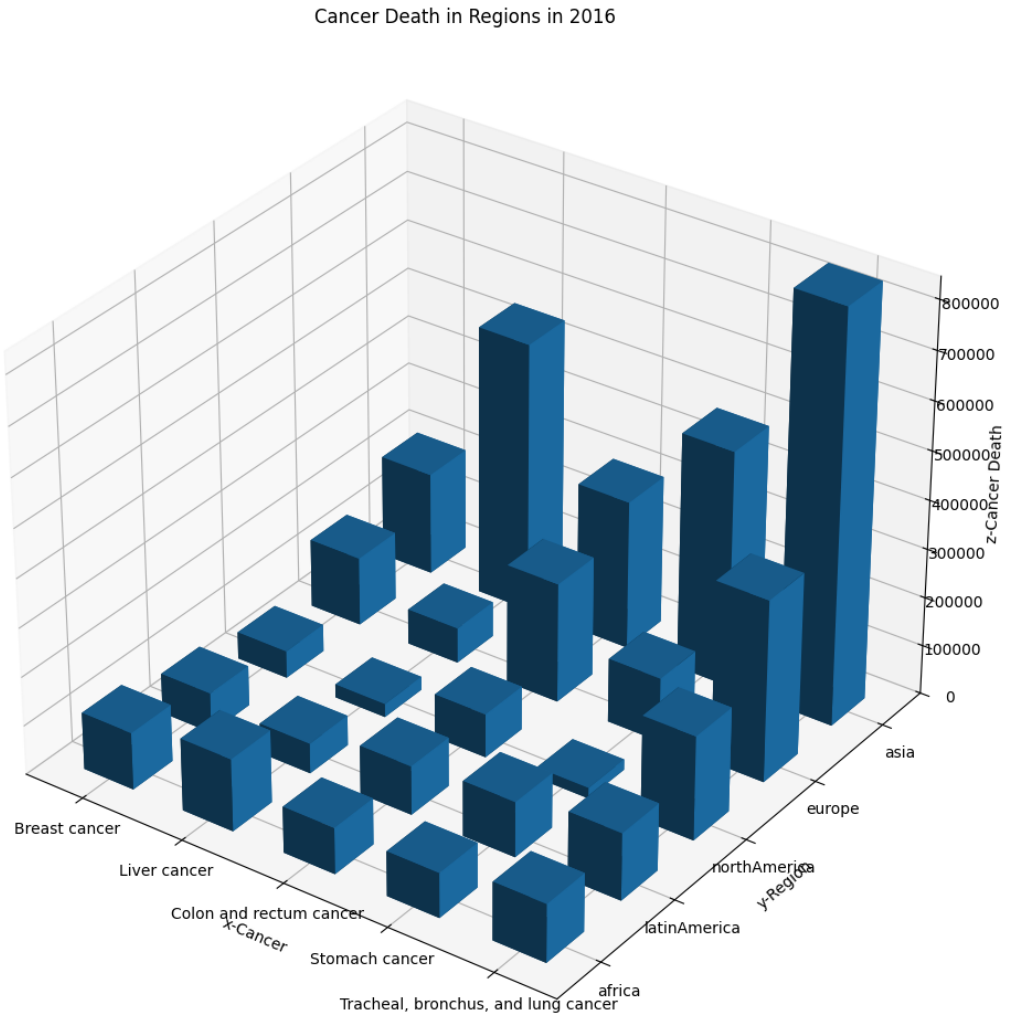

Figure 3. mortality number by top 5 cancers and continents in 2016.

6. Which types of cancers have the highest mortality rates, and are there any noticeable patterns or trends in these rates over the studied period?
    - Lung cancer mortality has always been significantly the highest and grows faster than other cancers. If looking at lung cancer alone, figure-6 shows a very large number of deaths are from Asia. It is a potential research topic in the field. (We cannot provide insights more than that due to the lack of data.)

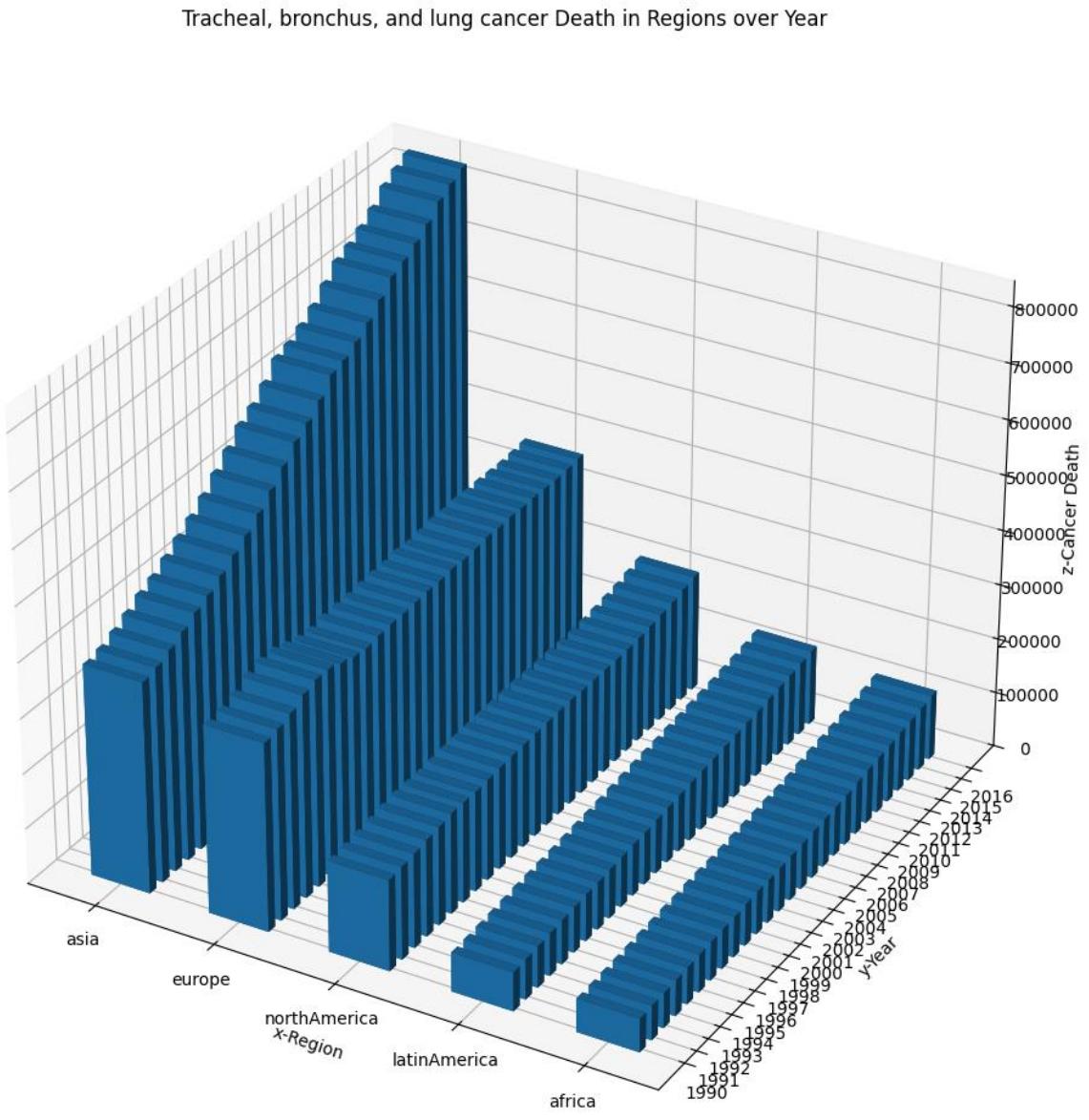

Figure 6. lung cancer mortality number by continents over years

7. How do variations in healthcare systems across countries affect cancer mortality rates?
    - We are not able to answer this question.
      First, due to the lack of total population data, comparing absolute death numbers doesn't really make sense. Because higher mortality numbers may indicate either higher population, or worse health care system, or both.
      Second, we do not have data which quantify the quality of health care systems across countries in the world. Data such as financial resources per citizen, hospital, clinic and doctor number per citizen can be helpful for analyzation.

8. Are there correlations between lifestyle factors (such as diet, smoking rates, and physical activity) and cancer mortality rates in different regions?
    - There are articles (Tai WP, 2017) about correlations between Esophageal cancer and the habit of drinking hot/warm water/fluid/soup in Asia. Figure-7 shows the top 7 cancer mortality numbers across continents in 2016, and it supports the article mentioned previously. Notice Esophageal cancer has the lowest to second lowest death number in all continents except Asia, where it's the fourth highest mortality number. This shows a correlation between the lifestyle of a certain region and cancer mortality.
    - It is hard to do more analyzation with current dataset alone, since it doesn't have data that reflects lifestyle factors. And without the data of populations of each country and region, it's even harder to see patterns.

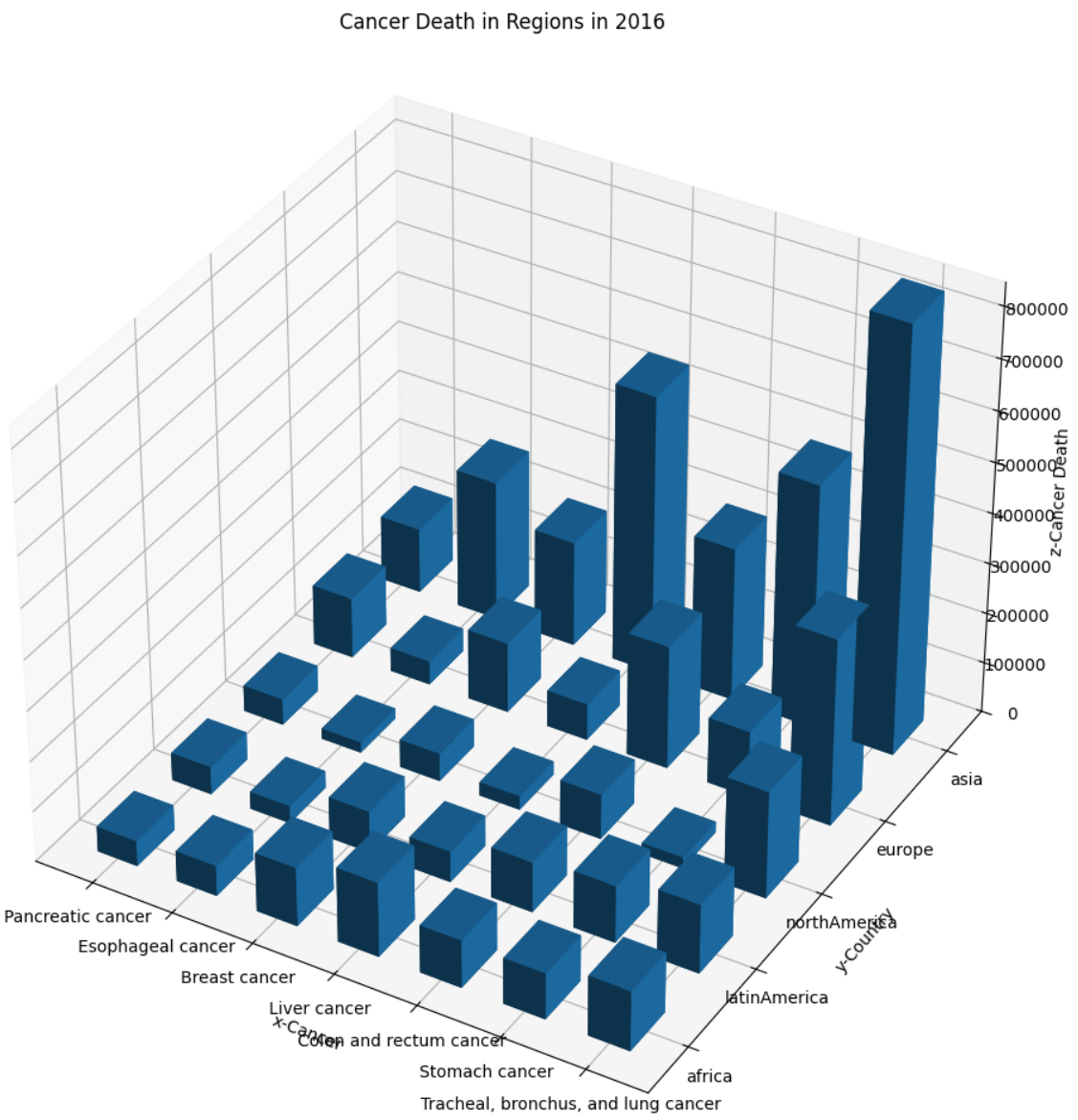

Figure 7. mortality number by top 7 cancers and continents in 2016 (Esophageal cancer)

9. How do genetic predispositions to certain types of cancers influence mortality rates in different regions?

- In figure-8, liver cancer is extremely high in Asia compared to other regions. Liver cancer is often associated with alcohol consumption. And Asian is known for having a particular genetic mutation in ALDH2 (aldehyde dehydrogenase 2) (Eng, M. Y.). This mutation reduces the activity of ALDH enzyme, which breaks down alcohol.
- Prostate Cancer is more common in populations of African descent (Odedina, F. T.). Figure-8 shows prostate cancer causes more mortality in Africa proportionally than other continents.

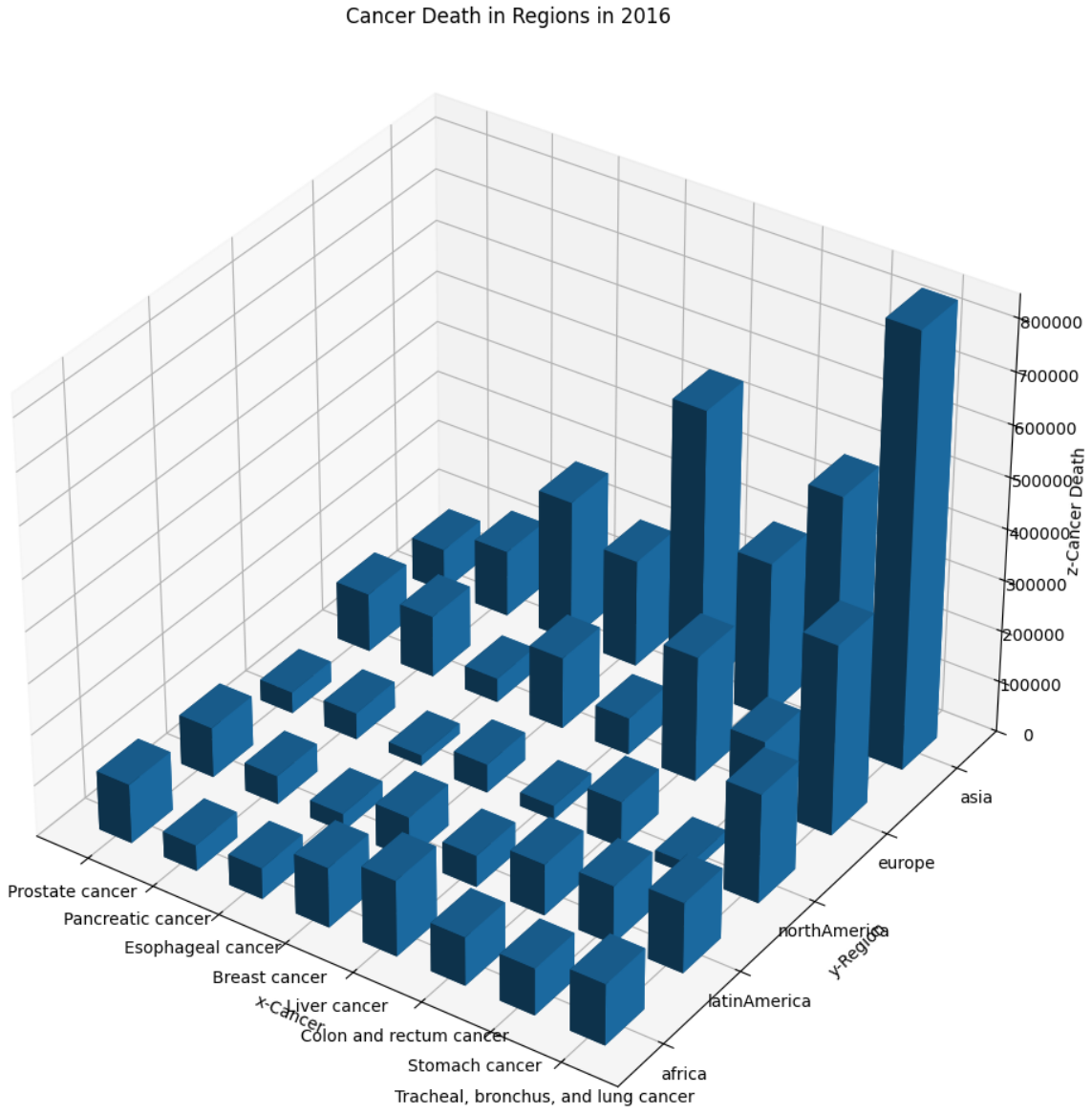

Figure 8. mortality number by top 8 cancers and continents in 2016. (prostate cancer)

10. Are there specific regions with higher mortality rates for cancers known to have strong genetic links, and how do these rates compare to regions with lower incidences of such cancers?
    - This question is similar to Q9, and we had answered as much as we could.

Discussion

1. Figure-4 shows the overview of mortalities of all cancers in the world.
    a. By analyzing the "World" data, between 1990 to 2016, the top 5 cancers (highest mortality numbers) had always been:
        1. Tracheal, bronchus, and lung cancer
        2. Stomach cancer
        3. Colon and rectum cancer
        4. Liver cancer
        5. Breast cancer
    b. Most of the slopes are positive, only a few are negative. Positive slopes make sense since we know the world population is increasing during the year of 1990 to 2016.
    c. Cancers with negative slopes are ones that we can study the reasons behind in the future.
    d. Possible factors of growing cancer mortality number: population growth, increase of average life expectancy, pollution, etc.

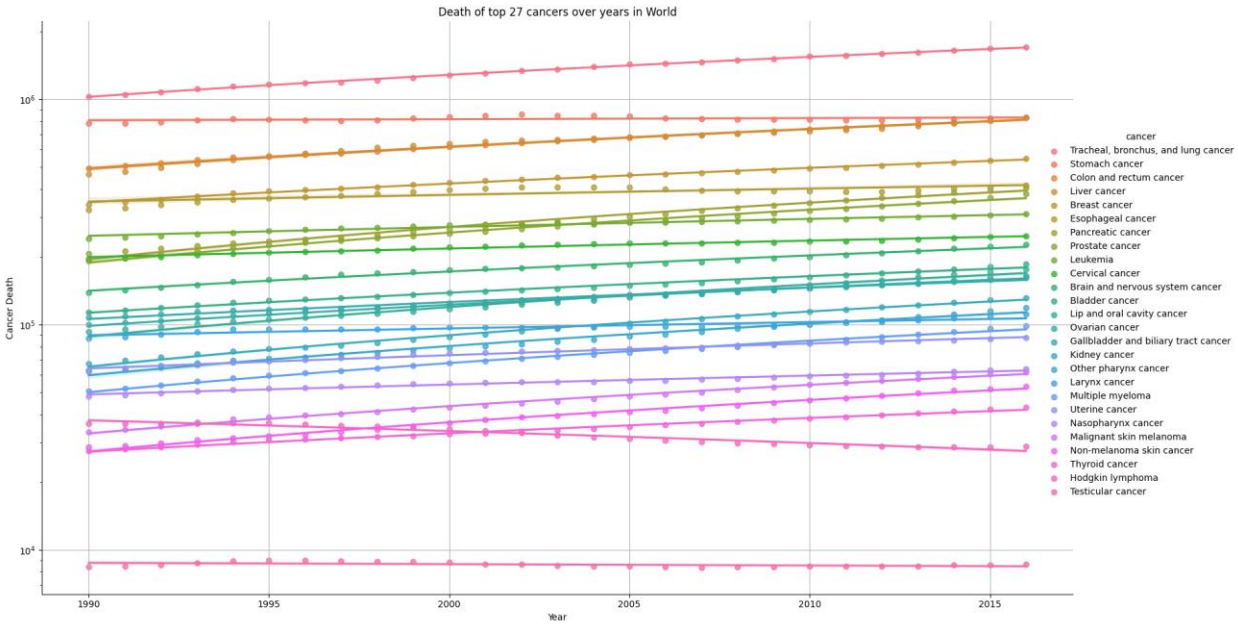

Figure 4. world mortality number of all cancers over years

2. Based on figure-4 and figure-5, we have concluded top 5 cancers in mortality.

a. As stated before, lung cancer shows a higher slope compared to other cancers.
b. On the other hand, stomach cancer mortality number remains, considering world population is growing over this period, this means stomach cancer has been treated effectively somehow. This is another research topic for the future.

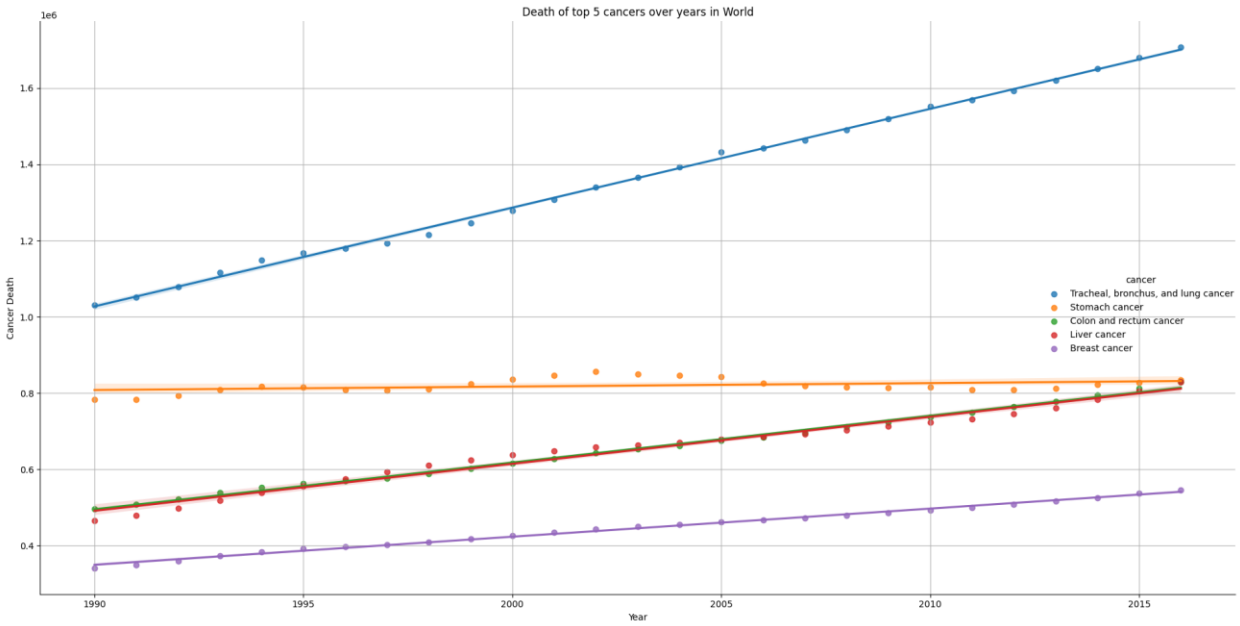

Figure 5. world mortality number of top 5 cancers over years

3. Figure-6 is a further analysis of lung cancer in terms of regions and time. Asia is the main boost of lung cancer mortality number over this period. And that is a potential research topic in the future.

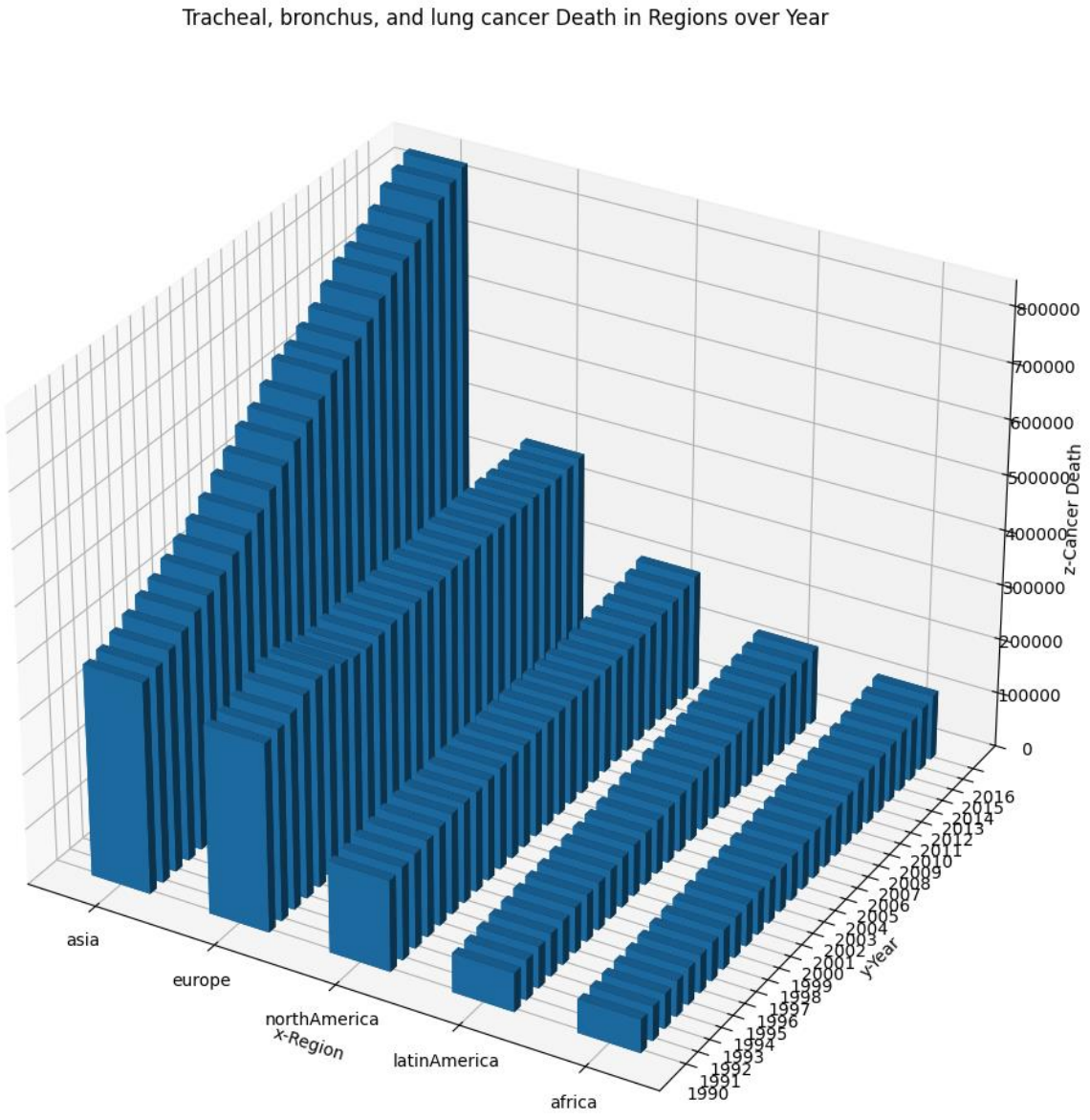

Figure 6. lung cancer mortality number by continents over years

4. If we focus on the top 5 cancers and divide countries into continents. We can see an unproportionally high liver cancer mortality number in Asia. Genetic mutations related to alcohol processing is a possible explanation as stated previously. But we will need more data and information to include or exclude other possible factors such as diet or lifestyle.

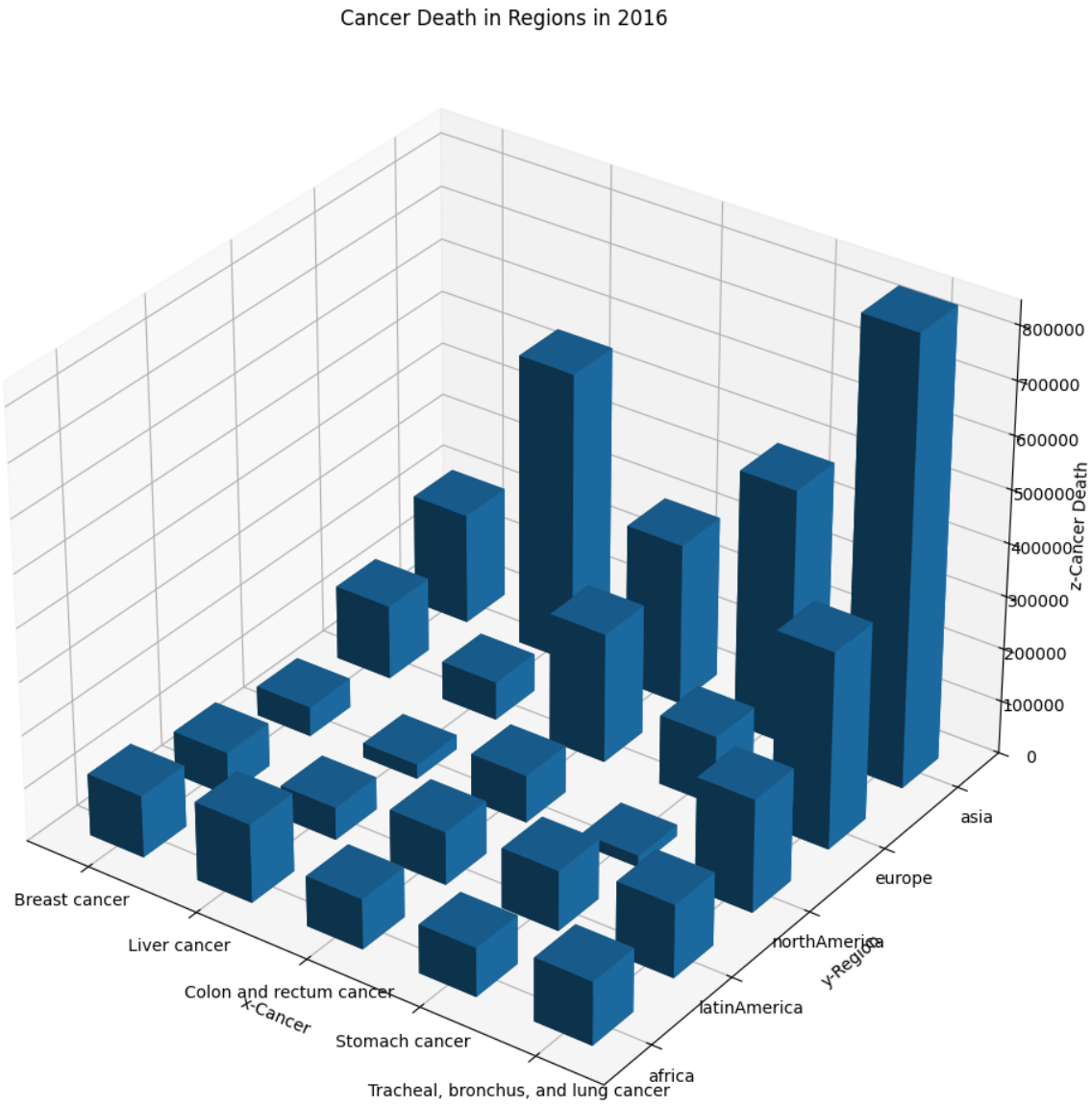

Figure 3. mortality number by top 5 cancers and continents in 2016.

5. Figure-9 is an analysis of liver cancer within Asia regions. East Asia holds 80% (if not 90%) of the liver cancer mortality number. Again, we need more data such as population of each region in Asia, related genetic mutation distribution, etc. to get a more accurate explanation or conclusion of this phenomenon.

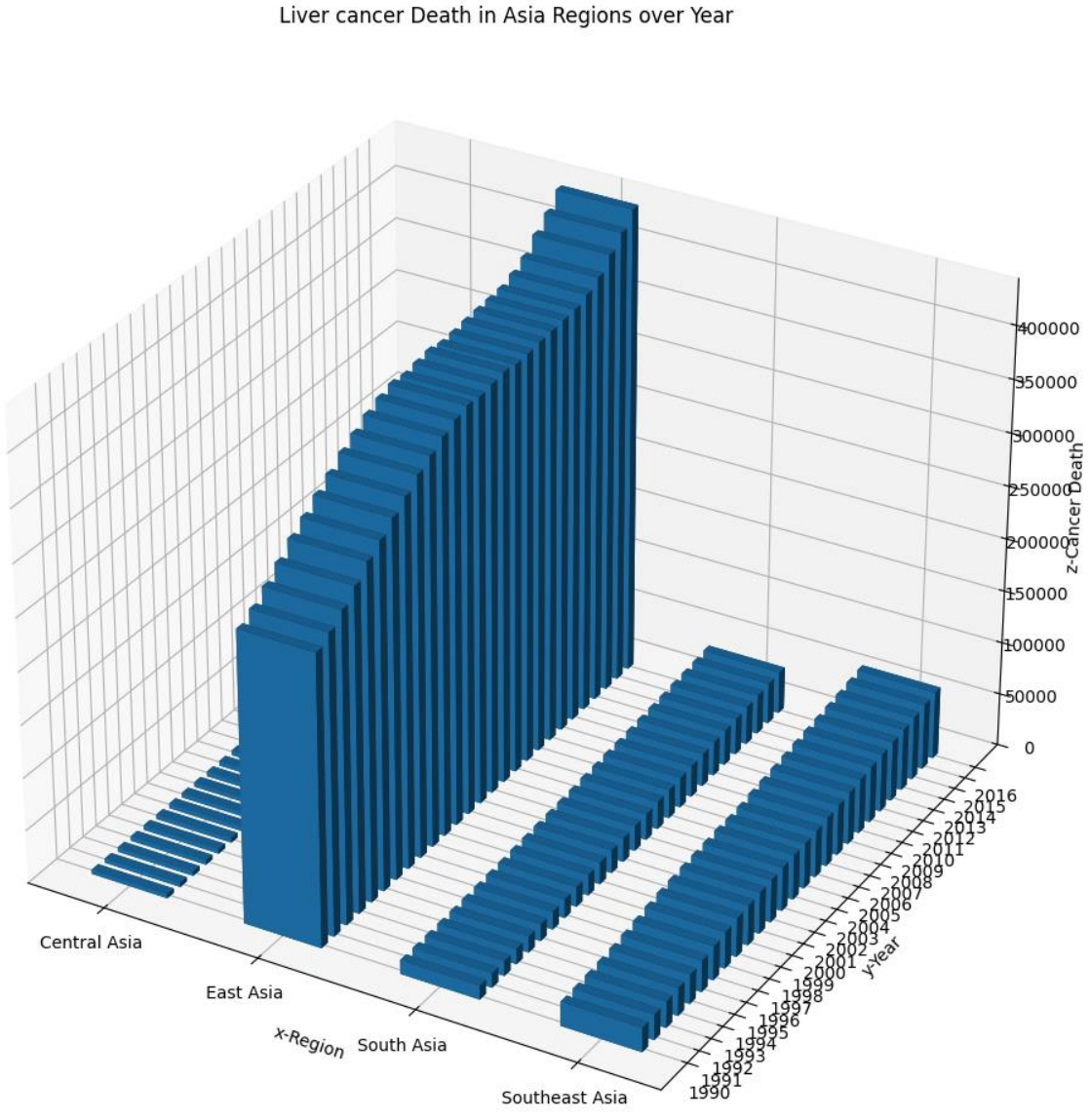

Figure 9. liver cancer mortality number by Asia regions over years.

Conclusion

Summary of Key Findings

This project successfully developed a data visualization dashboard to explore global cancer death trends from 1990 to 2016. Our key findings include:

1) Global Trends: The overall number of cancer deaths has increased over the period studied. Tracheal, bronchus, and lung cancer have consistently shown the highest mortality globally, followed by stomach, colon and rectum, liver, and breast cancers.
2) Regional Disparities: Asia has seen a disproportionate increase in lung and liver cancer deaths, likely influenced by genetic predispositions and lifestyle factors. Meanwhile, other regions such as North America and Europe showed high cancer deaths for colon and rectum cancers.
3) Temporal Trends: Some cancers, like stomach cancer, have shown a decrease in cancer deaths, potentially reflecting advancements in medical treatments and public health policies. In contrast, lung cancer mortality has risen faster than the population growth rate, suggesting external factors such as pollution and smoking rates play a significant role.

Discussion of Limitations

1) Data Quality: The dataset's completeness and accuracy are subject to the reporting standards of different countries, leading to potential biases. The lack of standardized data collection methods across regions further complicates global comparisons.
2) Undocumented Methodologies: The Kaggle dataset lacks detailed documentation on data collection and preprocessing methods, which limits our ability to validate and extend the findings.

3) Regional Disparities: Variations in diagnostic quality and healthcare reporting can skew mortality data, especially in low- and middle-income countries where resources are limited.

Suggestions for Future Research

To build upon our findings, future research should focus on:

1) Higher Quality Datasets: Efforts should be made to standardize data collection methodologies globally, ensuring consistent and accurate reporting across all regions.
2) Population Data: Integrating population data with cancer mortality rates would allow for more precise calculations of cancer mortality rates instead of absolute numbers, facilitating more meaningful comparisons.
3) Lifestyle and Genetic Factors: Further studies should investigate the correlations between lifestyle factors (diet, smoking, physical activity) and genetic predispositions with cancer mortality rates. This could help identify high-risk populations and inform targeted interventions. This would require integrating data in these domains.
4) Healthcare Systems Analysis: Research should examine how variations in healthcare systems impact cancer outcomes, using metrics such as healthcare expenditure per capita and accessibility to quality care as additional data to integrate.

By addressing these limitations and expanding the scope of our initial data, we can better inform public health decision-makers, ultimately combatting cancer globally.

References

Eng, M. Y., Luczak, S. E., & Wall, T. L. (2007). ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, 30(1), 22–27. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3860439/

Lin L, Li Z, Yan L, Liu Y, Yang H, Li H. Global, regional, and national cancer incidence and death for 29 cancer groups in 2019 and trends analysis of the global cancer burden, 1990-2019. J Hematol Oncol. 2021 Nov 22;14(1):197. doi: 10.1186/s13045-021-01213-z. PMID: 34809683; PMCID: PMC8607714.

Odedina, F. T., Ogunbiyi, J. O., & Ukoli, F. A. (2006). Roots of prostate cancer in African-American men. Journal of the National Medical Association, 98(4), 539–543. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2569237/

Tai WP, Nie GJ, Chen MJ, Yaz TY, Guli A, Wuxur A, Huang QQ, Lin ZG, Wu J. Hot food and beverage consumption and the risk of esophageal squamous cell carcinoma: A case-control study in a northwest area in China. Medicine (Baltimore). 2017 Dec;96(50):e9325. doi: 10.1097/MD.0000000000009325. PMID: 29390400; PMCID: PMC5815812. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815812/

Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends--an update. Cancer Epidemiol Biomarkers Prev. 2016 Jan;25(1):16-27. doi: 10.1158/1055-9965.EPI-15-0578.