Visualizing Bone and Joint Cancer Patterns: A Power BI Approach

Abstract:

In the field of cancer care, vast amounts of data are collected by institutions and registries, but accessing and analyzing this data can be challenging due to varying formats and limited accessibility. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute provides valuable cancer statistics in the United States; however, the data lacks consistency, making report generation a labor-intensive task. This thesis proposes an end-to-end process to cleanse, integrate, and present the SEER data through interactive dashboards with advanced reporting capabilities.

The goal is to offer a comprehensive view of 22 years of data, in a user-friendly and interactive format. The proposed approach allows for slicing and dicing the data along various dimensions, enabling the identification of hidden patterns and trends. These insights can be utilized to enhance treatment plans, allocate resources based on data-driven decision making, and ultimately improve patient care.

The developed dashboard is designed to be extensible, scalable, and capable of real-time updates as new data becomes available. This ensures that the visualizations and insights remain relevant and up to date, supporting ongoing research and decision-making processes.

By addressing the challenges of data accessibility, consistency, and reporting efficiency, this thesis contributes to the field of cancer care and research. The proposed end-to-end process provides a practical solution to transform complex and fragmented cancer data into actionable information. Ultimately, the aim is to leverage the power of interactive dashboards to facilitate informed decision-making, promote efficient resource allocation, and enhance the quality of care provided to cancer patients.

(Keywords:- Bone and joint cancer, Visualization, Patterns, Power BI, SEER database, Data analysis, Data visualization, ETL, Data extraction, Data preprocessing, Data transformation.)

Introduction:

Bone and joint cancer is a rare and complex disease that poses significant challenges in understanding its patterns and characteristics. The ability to visualize and analyze cancer data plays a crucial role in identifying trends, risk factors, and potential treatment strategies. This thesis focuses on the visualization of bone and joint cancer patterns using a Power BI approach, leveraging the extensive data available in the SEER (Surveillance, Epidemiology, and End Results) database spanning a period of 22 years. Additionally, data cleaning processes were conducted using Excel and MySQL to ensure data quality and reliability.

The field of data visualization offers powerful tools and techniques to explore and communicate complex information effectively. Power BI, a business intelligence tool developed by Microsoft, provides a comprehensive platform for data visualization, enabling the creation of interactive and dynamic visual representations of data. By harnessing the capabilities of Power BI, this research aims to enhance our understanding of bone and joint cancer and provide valuable insights into its patterns.

The SEER database is a widely recognized and reliable source of cancer data, encompassing a vast collection of patient records from various cancer registries in the United States. Extracting data from this extensive database allows for a comprehensive analysis of bone and joint cancer cases over a significant period.

Data cleaning is a critical step in ensuring the accuracy and consistency of the extracted data. In this thesis, a combination of Excel and MySQL was utilized to perform data cleaning procedures, including removing duplicate entries, standardizing variables, addressing missing data, and resolving inconsistencies.

By merging the power of Power BI's visualization capabilities with the extensive and cleansed data from the SEER database, this research aims to uncover patterns and trends in bone and joint cancer. The visualization process includes the development of interactive dashboards, charts, and maps that enable intuitive exploration and understanding of the temporal, and demographic aspects of the disease.

The outcomes of this research have the potential to contribute significantly to the field of oncology. By visualizing bone and joint cancer patterns using Power BI, this study provides healthcare professionals, researchers, and policymakers with valuable insights for informed decision-making. Moreover, the findings may aid in the development of targeted interventions, improved diagnosis strategies, and enhanced patient care.

In summary, this thesis focuses on visualizing bone and joint cancer patterns through a Power BI approach. Leveraging the SEER database and employing data cleaning techniques using Excel and MySQL, this research aims to provide comprehensive and reliable insights into the disease. By utilizing the power of data visualization, this study contributes to advancing our understanding of bone and joint cancer, ultimately facilitating improved treatment strategies and patient outcomes.

Material and methods:-

The following tools and techniques were employed in our research:

1. SEER\*Stat:- SEER\*Stat is a software utilized for selecting and extracting specific cancer data from the SEER database. This powerful tool enables researchers to access and retrieve relevant data subsets based on specific criteria and research objectives.
2. Excel:- Excel was utilized as a data storage tool in our study. It provided a convenient platform for organizing and managing the extracted data before further analysis.
3. MySQL Workbench:- MySQL Workbench was employed for data pre-processing and cleaning tasks. This comprehensive tool allowed us to perform necessary data transformations, cleansing, and filtering operations to ensure data quality and consistency.
4. Power BI Desktop:- Power BI Desktop served multiple purposes in our research. Firstly, it was used for data cleaning tasks through the Power Query Editor, enabling us to further refine and prepare the data for visualization. Additionally, Power BI Desktop provided robust capabilities for visualizing and analyzing the data, creating interactive dashboards, and generating insightful reports.

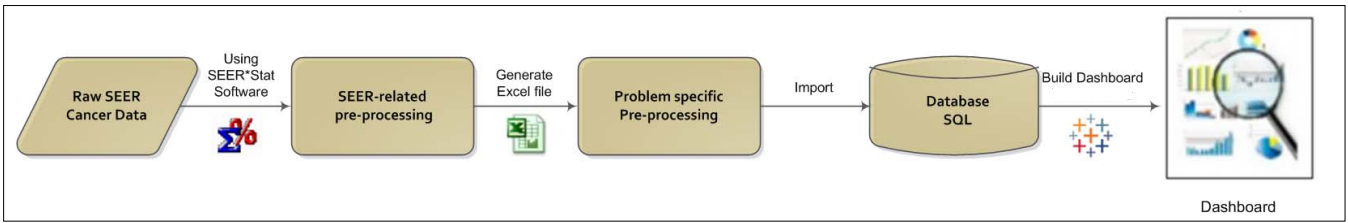
Data Overview :-

In our thesis, we utilized a dataset obtained from the Surveillance, Epidemiology, and End Results (SEER) database, specifically focusing on bone and joint cancer patients.

The dataset encompasses a time period spanning from 1992 to 2020, covering a total of 28 years. The dataset was accessed from the SEER database in November 2022, ensuring the inclusion of the most up-to-date and comprehensive information available.

Our dataset consists of 21 columns and 9,784 rows, representing individual patient records. The columns included in the dataset are as follows: PatientID, Sex, Race/Ethnicity, Year of Diagnosis, Type of Reporting, Source, Marital Status at Diagnosis, Median Household Income Inflation Adjusted to 2021, Year of Follow-up Recode, Year of Diagnosis Recode, Total Number of In Situ/Malignant Tumors for Patient, First Malignant Primary Indicator, Primary by International Rules, Cause of Death to Site Recode, Survival Months, Vital Status Recode (Study Cutoff Used), Radiation Recode, Chemotherapy Recode (Yes, No/Unknown), Scope of Regional Lymph Node Surgery (1998-2002), Site-Specific Surgery (1973-1997), Site Recode - Rare Tumors, and Primary Site – Labelled.

Methodology:-



The above diagram gives an overview of the method used for our study. It is primarily divided into three tasks: data extraction (from raw data), data analysis (including pre-processing), and data visualization.

1. Data Extraction:-

The SEER database maintains cancer statistics for the US and monitors annual cancer incidence progression of various types of cancer.

To extract data we have to gain access by registering for an Account: On the SEER website, look for the "Accessing SEER Data" or "SEER Research Data" section. Typically, you will need to create an account to gain access to the database. Register by providing the required information, such as your name, email address, affiliation, and research purpose.

Review the Data Use Agreement: As part of the registration process, you will be required to review and agree to the SEER Data Use Agreement. This agreement outlines the terms and conditions for accessing and using the data, including compliance with privacy and confidentiality regulations.

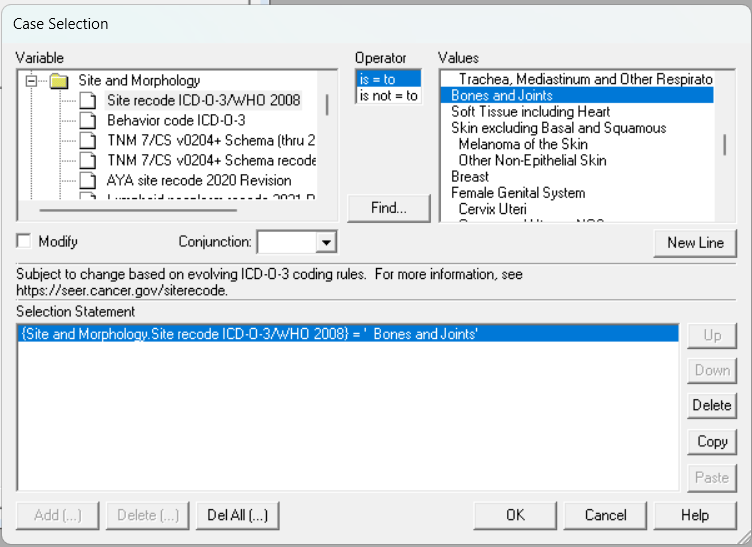
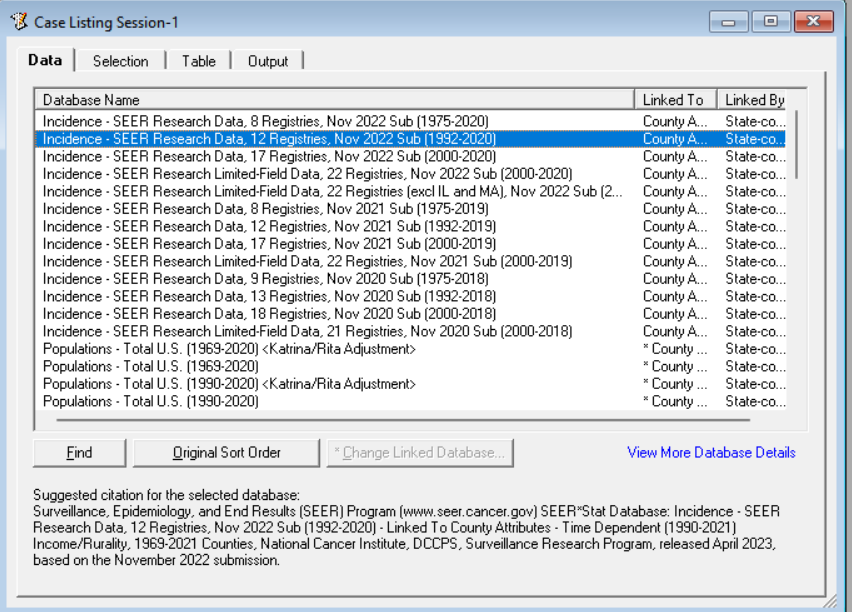
Complete the Training and Certification: Some institutions or organizations may require you to complete specific training or certification courses related to the use and handling of sensitive health data. Ensure that you comply with any training or certification requirements set by your institution or SEER.

Submit Data Request: After registering and completing any necessary training, you can submit a data request to access the specific data sets you need. The data request form will typically require you to provide details about your research project, including the variables, time period, and geographic scope of interest.

Data Use Agreement Approval: The SEER program will review your data request and determine if it aligns with their data usage policies and guidelines. Once approved, you will receive access to the requested data sets or instructions on how to access the data through their secure data portal.

Access the Data: Depending on the approved access method, you may be provided with login credentials to access the SEER database directly or be given instructions on how to download the requested data files.

After getting the Access and putting the login credentials on SEER\*Stat, We can get the data according to our needs,



We get a bunch of registries on SEER\*Stat were we choose “Incidence- SEER Research Data, 12 Registries, Nov 2022 Sub (1992-2020)”,

We choose Site and Morphology filter to select only Bones and Joints related cancer,

In Table we have certain column names, here we choose the types of column we need for our visualization, After choosing the needed Data we extracted that data in .CSV format in Excel,

1. Data Analysis :-

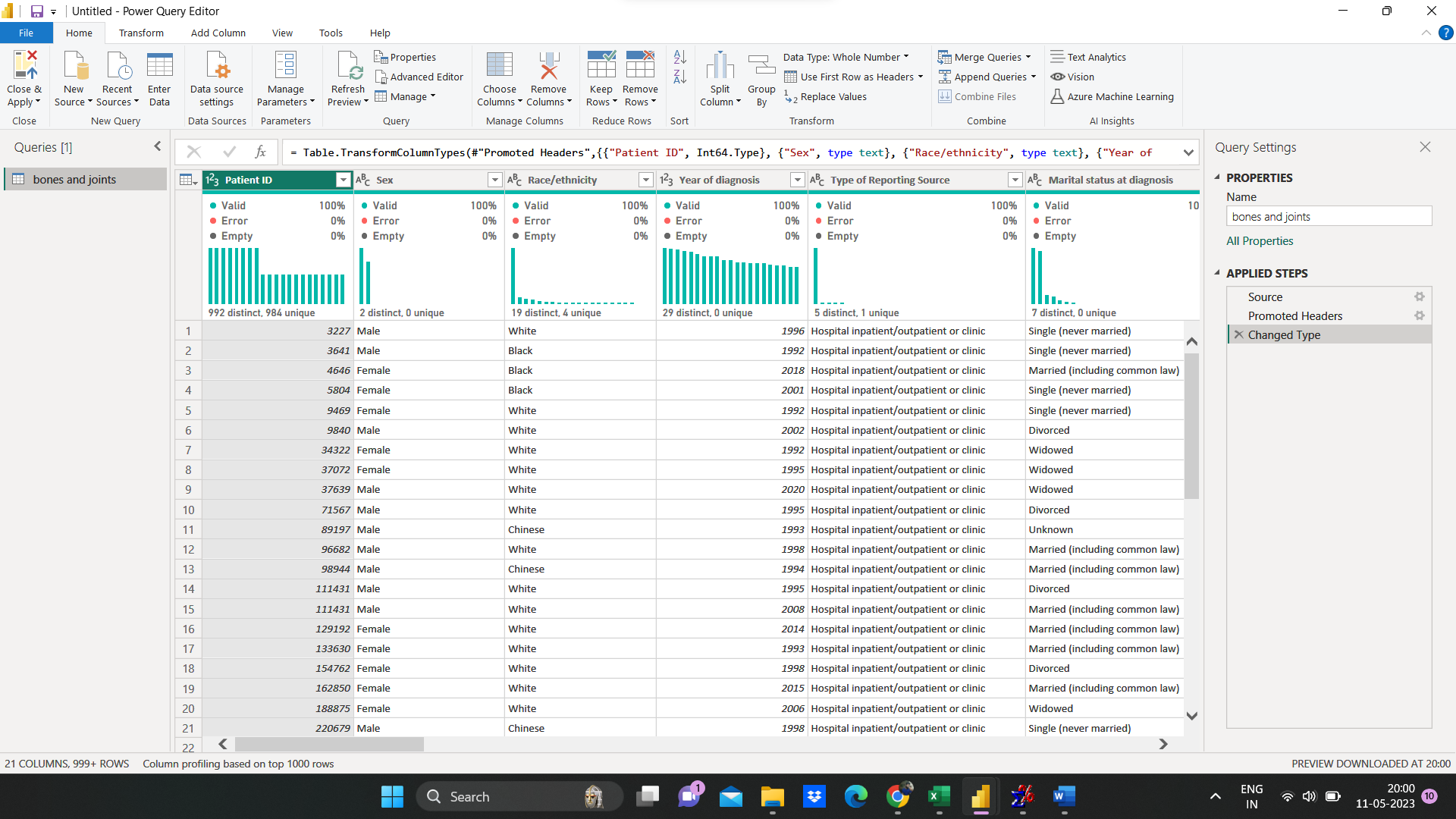
After extracting the raw records from the SEER database in excel, the data undergoes pre-processing to create a relevant subset. The pre-processed data is then imported into a SQL Server database. Analysis is performed on the data within the database, leading to the development of interactive dashboards and reports using Power BI. This enables users to explore and understand the bone and joint cancer data in a visually engaging and intuitive manner.

1. Data pre-processing

* SEER-related pre-processing involves using SEER\*stat software to normalize the data, converting text values to numeric representation. The derived data is then cleansed to remove redundant content, ensuring data consistency and quality.
* Problem-specific preprocessing Involves selecting relevant data records for a specific time period of significance and removing attributes that do not contribute significantly to predictive power, One of the steps is the removal of records which have data other then Bones and joint cancer.

1. Analysis

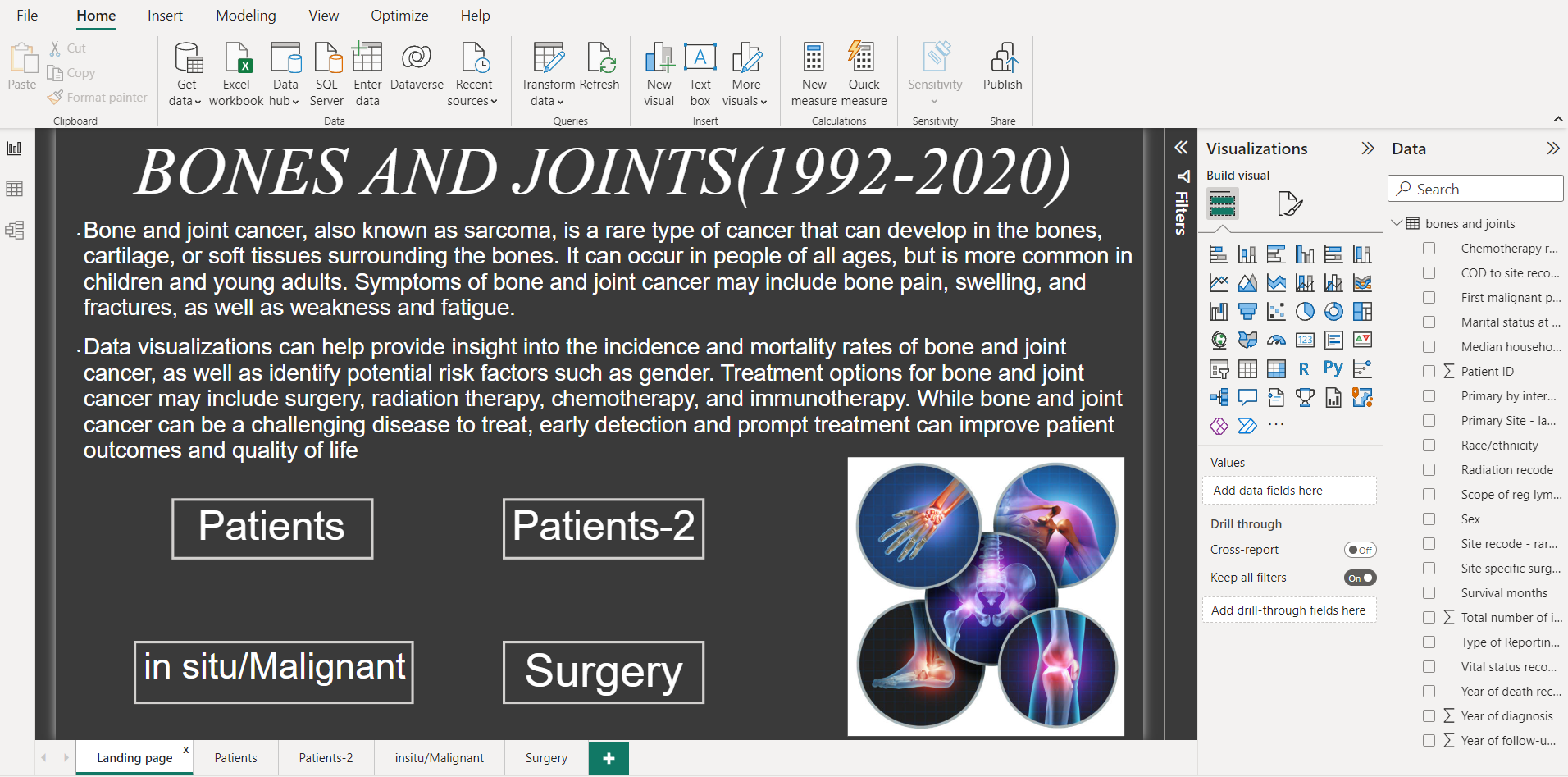
The pre-processed data is the filtered in Excel, then imported into MS SQL Server to create a database consisting of relevant dimensions and measures. Power BI is connected to the database and the tables are joined to create a view, extending horizontally by adding columns of data, as needed. The data is further cleansed (such as by changing data types, renaming & resetting fields) using Power Query Editor and prepared for analysis. Calculated fields, formulas, grouping and sets are added via SQL queries. The data is sliced and diced by using filters and parameters. The dissected data is then visualized using workbook, dashboards, and stories.



1. Data Visualization

After normalizing the data in Excel, MySQL, and Power Query Editor, the next step is to import the data into Power BI. In Power BI, we can observe the columns of our data in the data view. By utilizing the same dataset, we can create various graphs and charts, resulting in a dynamic and interactive dashboard.

Power BI is a powerful and commercially available data visualization tool that can transform large volumes of data into meaningful and actionable information. This is achieved through the creation of visually appealing and interactive dashboards. These dashboards contain components such as charts, graphs, tooltips, and drill-down/drill-through reports.



Dashboards in Power BI provide users with interactive access to informative data, enabling a better understanding of the vast amount of data generated by every cancer incidence. This, in turn, can significantly improve the quality of cancer care. Policymakers, health professionals, advisors, and planners can utilize this data to view and report bone and joint cancer statistics, gaining better insights into incidence and mortality trends.

Physicians can benefit from Power BI by identifying treatment options, designing wellness programs, and engaging patients effectively. Patients themselves can be empowered to make informed decisions about their care through interactive visualization of treatment cost, quality, and effectiveness.

Power BI offers connectivity to a variety of data sources, including Access, Excel, data warehouses, and web-based data. The reports generated using Power BI can be published to a shareable URL, making it easy to distribute and access the insights gained from the data.

By adopting a scientific approach, this thesis highlights the process of normalizing and importing data into Power BI. It emphasizes the significance of Power BI as a tool for transforming complex data into visual representations, ultimately leading to improved cancer care outcomes. The inclusion of policymakers, health professionals, advisors, planners, physicians, and patients as the target audience showcases the broad impact and potential benefits of utilizing Power BI for bone and joint cancer data analysis.

**5. RESULTS AND DISCUSSIONS**

5.1 First Dashboard:(Patients)

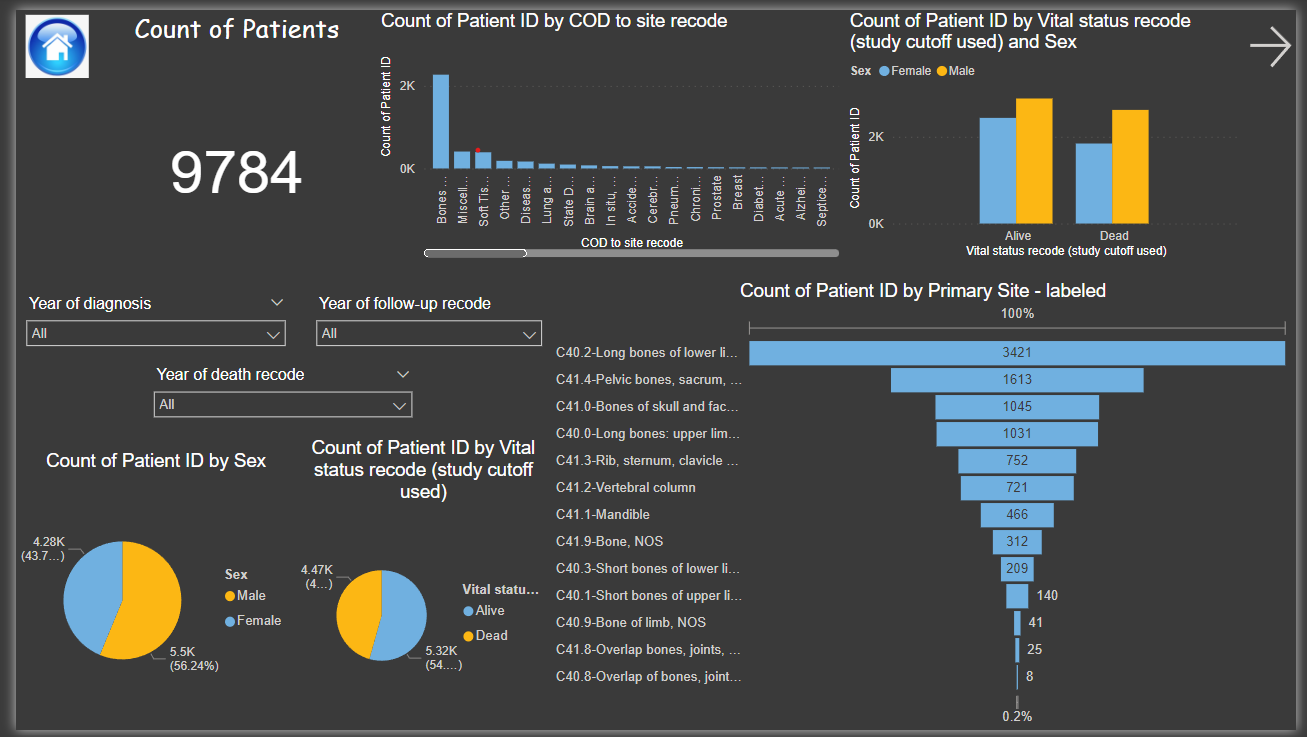


Figure 5.1 Patient Dashboard

In the first dashboard of our visualization, we present the following components:

Patient Count Card: This card displays the total count of patients in the dataset. The count dynamically updates based on the selected parameters.

Stacked Column Chart - Cause of Death: This chart visualizes the distribution of different causes of death among the patients. It provides insights into the major causes of death observed in the dataset.

Clustered Column Chart - Vital Status by Sex: This chart illustrates the vital status (alive or dead) of patients, segmented by their gender. It allows for a comparison of vital status between males and females.

Pie Chart - Gender Distribution: This chart represents the percentage of male and female patients in the dataset. It provides a visual understanding of the gender distribution among the recorded patients.

Pie Chart - Vital Status Distribution: This chart showcases the percentage of patients categorized by their vital status (alive or dead). It offers insights into the overall distribution of patient outcomes.

Funnel Graph - Types of Bone and Joint Cancer: This graph displays the different types of bone and joint cancer present in the dataset, arranged in descending order of patient count. It visualizes the relative prevalence of each cancer type.

Slicers - Year of Diagnosis, Year of Follow-up, and Year of Death: These slicers allow users to select a specific year of diagnosis, follow-up, or death. By choosing a particular year from the dropdown, the corresponding data for that year is displayed, enabling focused analysis and exploration.

Through this dashboard and its interactive components, users can gain insights into the patient count, cause of death, vital status, gender distribution, cancer types, and explore the data based on specific years of interest.

5.2 Second Dashboard:(Patients-2)

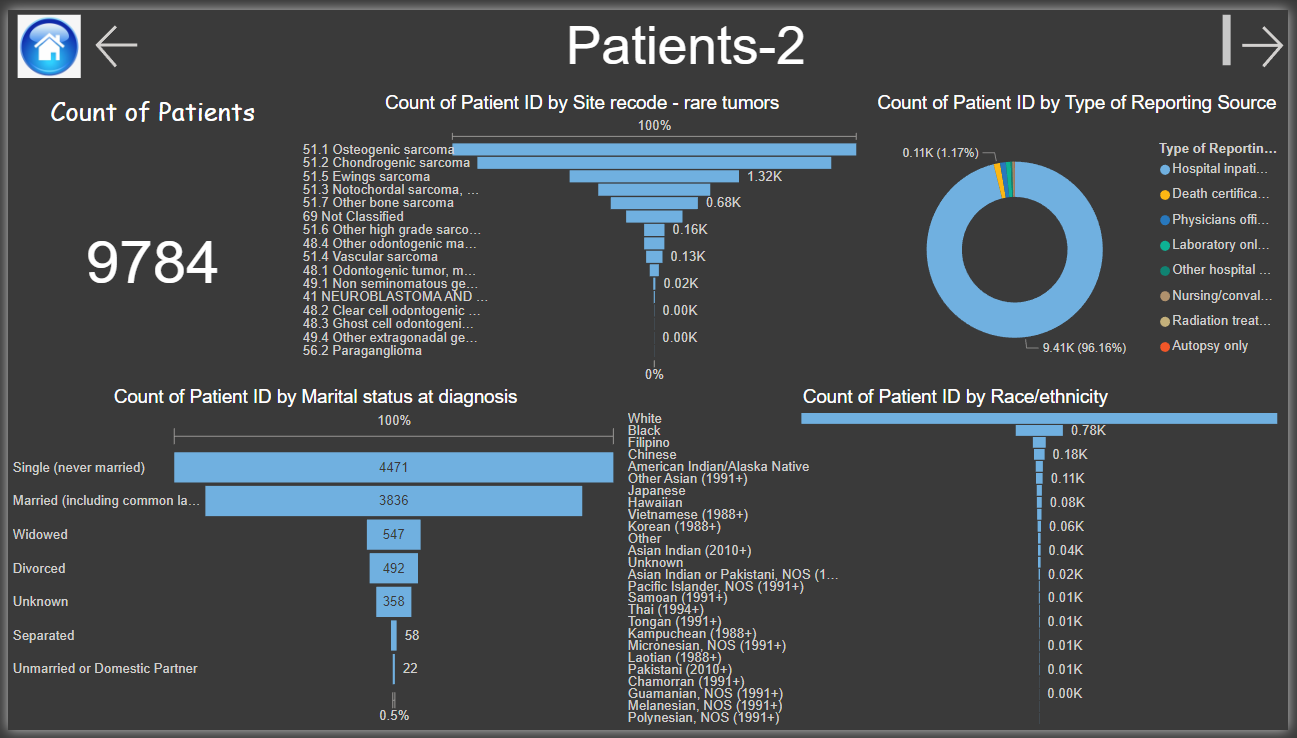


Figure 5.2 Patient-2 Dashboard

In the second dashboard of our visualization, we present the following components:

Patient Count Card: Similar to the first dashboard, this card displays the total count of patients in the dataset. The count dynamically updates based on the selected parameters.

Funnel Graph - Types of Tumors: This graph visualizes the different types of tumors present in the dataset, arranged in a funnel shape. It provides insights into the relative prevalence of each tumor type among the patients.

Donut Chart - Reporting Source: This chart represents the distribution of patients based on the reporting source. It showcases the proportion of patients reported from different sources.

Funnel Graph - Marital Status at Diagnosis: This graph depicts the marital status of patients at the time of diagnosis, arranged in a funnel shape. It allows for an understanding of the patient count based on their marital status.

Funnel Graph - Race/Ethnicity: This graph displays the distribution of patients across different race/ethnicity categories. It provides insights into the patient count based on their race/ethnicity.

Through this second dashboard, users can explore the patient count, tumor types, reporting sources, marital status at diagnosis, and race/ethnicity distribution. The interactive nature of the components enables users to analyze the data from different perspectives and gain a deeper understanding of the characteristics and demographics of the patients in the dataset.

5.3 Third Dashboard :(in situ/Malignant tumors)

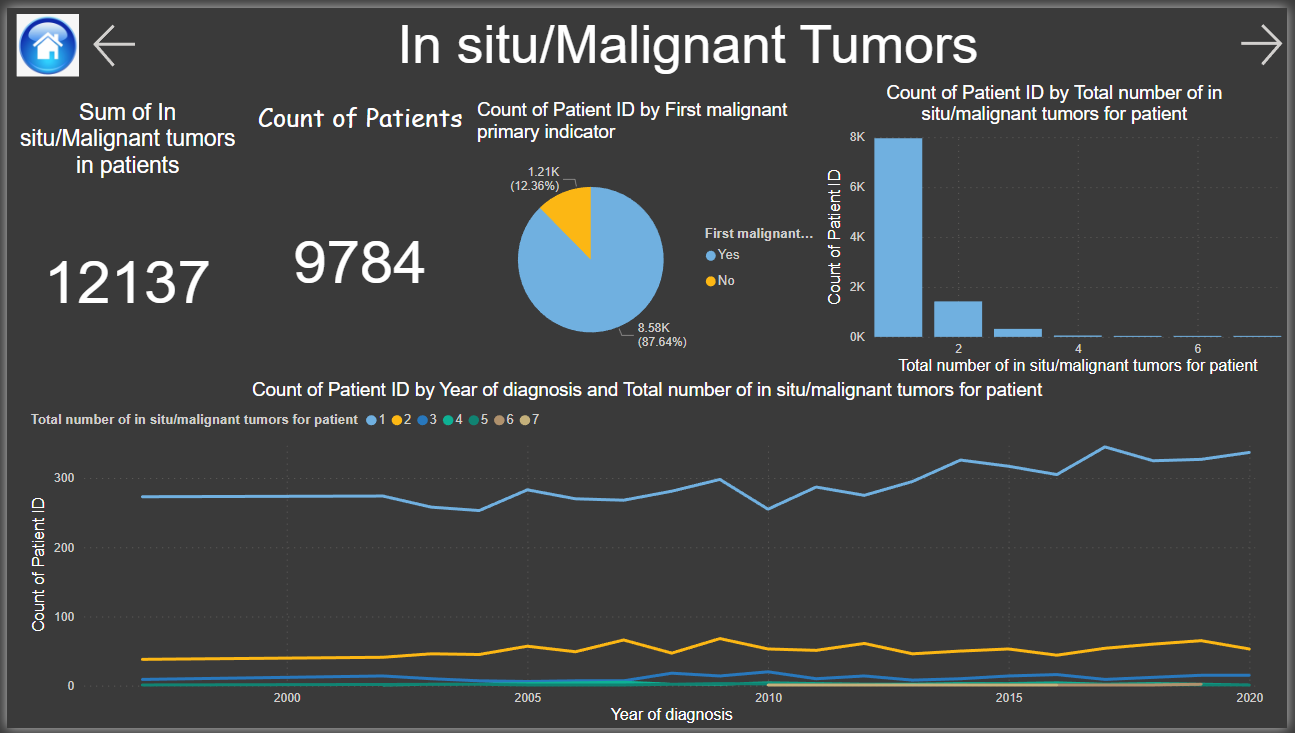


Figure 5.3 in situ/Malignant Tumors Dashboard

In the third dashboard of our visualization, we present the following components:

Patient Count Card: This card displays the total count of patients in the dataset. It provides an overview of the number of patients included in the analysis.

Sum of In Situ/Malignant Tumor Count Card: This card shows the sum of in situ/malignant tumor counts for all patients. It provides a cumulative count of tumors present in the dataset.

Pie Chart - First Malignant Primary Indicator: This chart represents the distribution of patients based on the first malignant primary indicator, categorized as "yes" or "no." It visualizes the proportion of patients with or without a first malignant primary tumor.

Clustered Column Chart - Total Number of In Situ/Malignant Tumors: This chart displays the count of patients based on the total number of in situ/malignant tumors they have (e.g., 1, 2, 3, 4, 5, 6). It allows for a comparison of patient counts across different tumor counts.

Line Chart - Total Number of In Situ/Malignant Tumors by Year of Diagnosis: This chart illustrates the trend of the total number of in situ/malignant tumors over the years of diagnosis. It helps identify any patterns or changes in tumor counts over time.

Through this third dashboard, users can explore the patient count, in situ/malignant tumor counts, the presence of a first malignant primary tumor, and the relationship between tumor counts and the year of diagnosis. The visual representations provide insights into the distribution, trends, and patterns of tumor data, allowing for a deeper understanding of the dataset.

5.4 Fourth Dashboard : (Treatment)

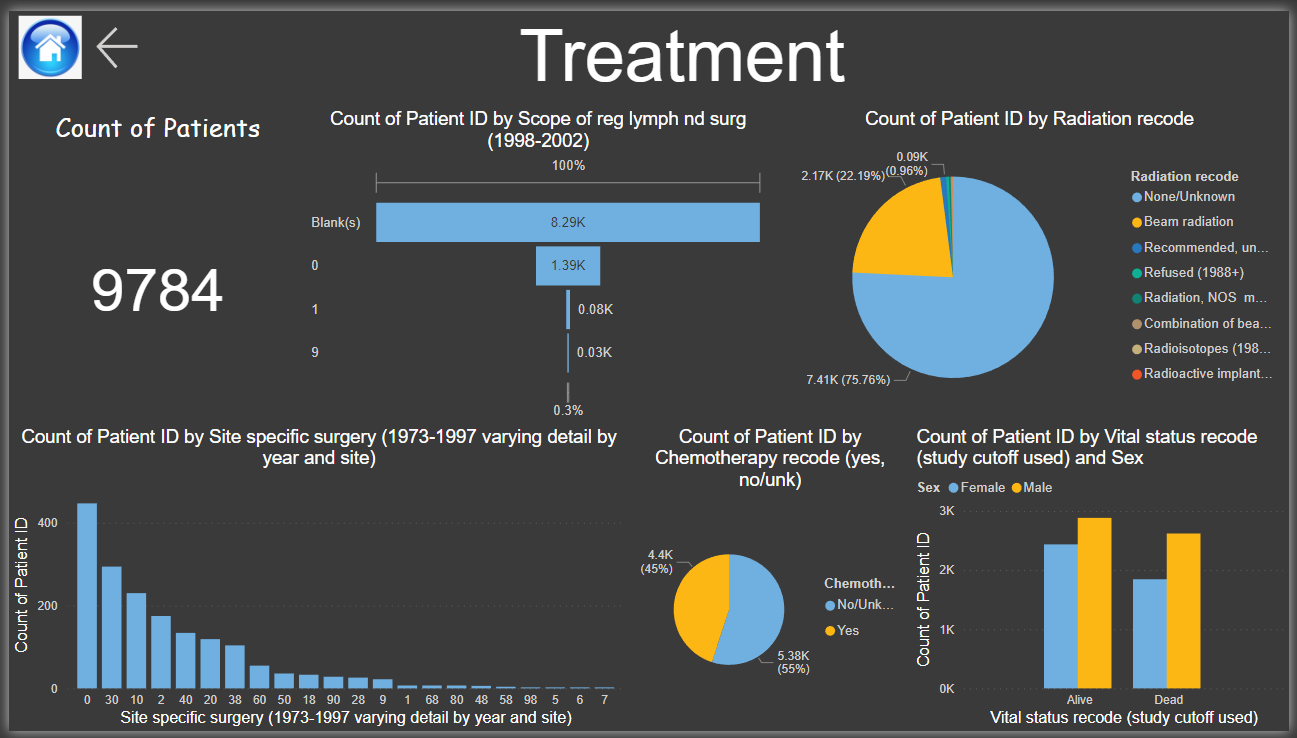


Figure 5.4 Treatment Dashboard

In the fourth dashboard of our visualization, we present the following components:

Patient Count Card: This card displays the total count of patients in the dataset. It provides an overview of the number of patients included in the analysis.

Funnel Graph - Lymph Node Surgery: This graph visualizes the distribution of patients based on the lymph node surgery they underwent. It categorizes surgeries as Blank(s), 0, 1, or 9 and provides insights into the relative frequency of each surgery type.

Pie Chart - Types of Radiation Therapy: This chart represents the distribution of patients based on the types of radiation therapy they received. It shows the proportion of patients who underwent different radiation therapy treatments.

Clustered Column Chart - Site-Specific Surgery (1973-1997): This chart displays the count of patients who underwent site-specific surgeries between the years 1973 and 1997. It allows for a comparison of the number of surgeries performed based on the patient count.

Pie Chart - Chemotherapy Status: This chart showcases the distribution of patients based on whether they underwent chemotherapy or not. It visualizes the proportion of patients who received chemotherapy versus those who did not.

Clustered Column Chart - Vital Status by Sex: This chart illustrates the vital status (alive or dead) of patients, segmented by their gender. It provides insights into the comparison of vital status between males and females.

Through this fourth dashboard, users can explore the patient count, lymph node surgery, radiation therapy types, site-specific surgeries, chemotherapy status, and the relationship between vital status and gender. The visualization components provide a comprehensive understanding of the treatment modalities, surgical interventions, and patient outcomes in the dataset.