Number of Cancer Surgeries (Volume) Performed in California Hospitals

GEORGE MASON UNIVERISTY

AIT-580-002-FINAL PROJECT REPORT

Amit Chowdary Kamma

akamma@gmu.edu

Abstract— The purpose of this study has been to analyse the number of cancer surgeries undertaken across California hospitals as entirely based on the dataset obtained from data.gov. Specifically, the dataset covers the surgeries data across counties, hospitals, surgery categories, and the cases performed within. Through Python, R, SQL, and AWS, and comprehensive data analysis competed, I have been able to address specific research questions related to cancer surgery trends in California. Key findings indicate notable differences in surgery volumes between various surgeries, hospitals, and regions. The strategies for treating cancer and distribution of healthcare resources are discussed as a result of the findings found. To improve the execution of cancer care and the distribution of resources in California hospitals, the research attempts to find trends and differences in the rates of cancer surgery.

Keywords— Cancer surgeries, data analysis, healthcare resource allocation, and Geographic variations.

I.INTRODUCTION

An ever-growing and increasingly powerful threat to public health, cancer requires constant monitoring and innovative ways to care. Knowledge of the nature of cancer surgery is essential in the context of medical care in order to devise strategies that effectively combat the disease. In order shed light on significant aspects of cancer care delivery in the area, this study sets out to understand the complex patterns and trends of cancer procedures throughout California hospitals.

Differences in access to cancer care still exist despite advances in medical research and technology, which presents problems for both patients and healthcare systems. It is important to understand the patterns, geographic distribution, changes in cancer surgery volumes, and disturbances in accessing cancer care among authorities in order for effective resource allocation and improve healthcare delivery.

This study is important because it can provide stakeholders, professionals, and policymakers in the healthcare industry with information regarding the state of cancer surgery in California today. This study intends to provide important insights to the ongoing efforts to enhance the delivery of cancer care and resource allocation in the area by clarifying trends, patterns, and disparities in surgical volumes and distributions.

Research Questions

1. How has the volume of cancer surgeries changed over the years for all types of cancer in California hospitals? [6]

Proposal: The research will study the trends of the number of cancer surgeries performed in California hospitals for different kinds of cancer over the years.

Results: An examination of past patterns in the number of cancer surgeries performed will show how different cancer kinds have changed over time.

2. Is there any geographical pattern of the distribution in cancer surgeries being performed across the counties in California? [6]

Proposal: This research will explore geographical trends in the spatial distribution of cancer procedures within the counties of California.

Results: Using spatial analysis, it will be demonstrated to show geographical patterns in the spread of cancer procedures over California counties.

3. Regarding the volume of cancer surgeries performed in various hospitals within the same country (USA), what patterns can be observed? (State wise data) [6]

Proposal: This research will look at differences in the number of cancer surgeries performed in various Californian hospitals, offering information about geographical differences in the availability of cancer care.

Results: An aim of this study was to identify variations in volumes of cancer surgery among different institutions in California and their implications on regional differences in access to cancer care.

4. Is there any specific type of cancer surgeries which are performed more in one place and less in another? [6]

Proposal: The purpose of this study is to compare the incidence of various cancer surgery kinds in various geographic areas of California.

Results: It will identify and examine differences in the frequency of cancer surgery kinds among the various areas of California.

5. Which cancer surgeries are most performed in California and how has this changed over the years? [6]

Proposal: This study will address the most common cancer surgeries held in California and further look at the frequency of these surgeries over the years.

Results: I will identify the most common cancer operations in California, track their evolution in frequency over time, and evaluate the results.

II. LITERATURE REVIEW

The paper "Safety in Numbers: Cancer Surgeries in California Hospitals" [2] by the California HealthCare Foundation provides information of how surgical volume affects patient outcomes in cancer surgeries. The research held indicates the relationship between the number of hospitals located in California and patient outcomes for different cancer types such as bladder, brain, breast, liver, pancreas, prostate, colon, esophagus, lung, rectum, and stomach. When comparing the hospitals with high volume capacity and low volume capacity hospitals, patients which undergo surgeries in low volume hospitals have a bigger risk of death and complications.

In reference to the study made in the report it is stated that there is a very less volume of cancer surgeries performed in California with 249 hospitals performing only one or two procedures for some specific cancer surgery in 2014. Even though the hospital leaders/groups do have a general idea of a higher surgical volume gives a better outcome, they still lack in knowledge of the negative impact which takes place when we have a low hospitals volume.

The paper concludes the significance of stakeholders in tackling the low hospital surgeries volume issue, including payers, policymakers, surgeons, referring physicians, and hospital administrators. To reduce the problem of California in having less volume of hospitals that conduct cancer surgeries a few recommendations are given such as increasing awareness, utilizing payer influence to do so, offering advice on the right number of surgeries, and enabling patients to choose their treatment facility location with insights based on available data in the report and other websites. This paper discusses my research questions in a very brief manner where it talks about the volume, distribution, and types of cancer surgeries performed in California hospitals.

The research paper is aimed to analyze the “Geographic Distribution of Adult Inpatient Surgery Capability in the US'' [3]. By using the data from American Hospital Association, the researchers found using geospatial analysis there are 3409 hospitals which can perform a cancer surgery. In reference to the American Hospital Association, Census Bureau, and the American Community Server, 1373 hospitals meet the requirements of being a major surgical hospital.

The relation of these hospitals and people living in the USA are shown and insights state that nearly 10% of the population lived outside of a 30-mile radius around hospitals that could perform adult inpatient surgery.

A few factors are linked to living the service regions which are race, age, work position, educational level, and insurance coverage. This paper focuses on the important aspects of having access to top surgical treatment as a vital component of public health. Geographical locations, demographic traits, and state Medicaid expansion status were found to be significant predictors of access disparities. The results emphasize how critical it is to bridge disparities in underprivileged people' access to surgical care.

In conclusion this study focuses on spread of surgical hospitals across USA and the need of these hospitals which has good access across the nation for every individual regardless of their geographic location. Also, the difficulties of certain location and populations of accessing these surgical services. With reference to my research questions, it does answer about the geographical locations in USA but not in particular to California and it also addresses the problems faced in the geographic locations.

The research paper “Impact of Hospital Volume on Operative Mortality for Major Cancer Surgery” [4] is an experiment carried out to investigate the impact of hospital volume on 30-day operative mortality for major cancer procedures. The experiment used the ‘SEER Medicare Linked database’ for their research. Begg et al conducted this in 1998. The investigation was limited to individuals who, between 1984 and 1993, received operations for different types of cancer and were 65 years of age or older. The surgeries which were kept an eye on was pneumonectomy, liver resection, pancreatectomy, esophagectomy, and pelvic exenteration. Specific factors were looked which was very vital to understand the influence those factors have on morality rates. Those factors were comorbidities, patient age, and cancer stage.

The findings in the experiment conducted was found that there is a relationship between the death/morality rate and hospital volume. Significantly there were lower death rates for a few types of cancer surgeries such as pancreatectomy, esophagectomy, liver resection, and pelvic exenteration. These surgeries belonged to the large hospital volumes. In reference to what they found they understood the significance of hospital proficiency and speciality in difficult surgical oncologic treatments, resulting in enhanced patient results.

In conclusion, the study went through the relationship between a hospital's surgical volume and cancer patient outcomes. With continuous monitoring, the researchers found that higher volume cancer surgery is associated with better outcomes for hospitals. By this outcome, people can prefer going to the hospitals that have a large volume of cancer surgeries and supports the idea of a specialized care can help patients live a better life. This paper does give me an idea about my 4th research question where there is a specific type of cancer surgery being performed more in one place than an another as we saw in this research paper that some specific hospitals are specialized in some type of cancers. So, some hospitals would be having a higher rate of a specific type of cancer treatment which they have less death rates in.

In this paper “Association of Distance Traveled for Surgery with Short- and Long-Term Cancer Outcomes” [5] the information from the National Cancer Data Base (NCDB) for the years 2003 through 2006, the study looked at how travel distance for surgical resection affected the course of therapy for patients with pancreatic, liver, colon, and esophageal cancer. Based on the centroids of zip codes, the distance traveled was calculated, and the results showed a mean of 30.0 miles and a median of 7.5 miles. Patients who were diagnosed with different types of cancer travelled different distances. Pancreatic and esophageal types of cancer had patients travel twice as far compared to patients diagnosed with colon cancer. Patients with liver cancer travelled three times far making it the highest. Several characteristics play a role in the deciding factor of how far patients travel.

The study found that on an average the longer the distance travelled better was the survival rate for liver, colon, and pancreatic cancer. Propensity scores were produced using Cox regression to compare 5-year mortality and mixed effects logistic regression to 90-day mortality. Possible variables consisted age, race, sex, tumor features, Charlson scores, type of treatment, and whether the patient lived in an urban or rural area. Hospital clustering was taken into consideration using a mixed-effect Cox proportional hazard model, which has a

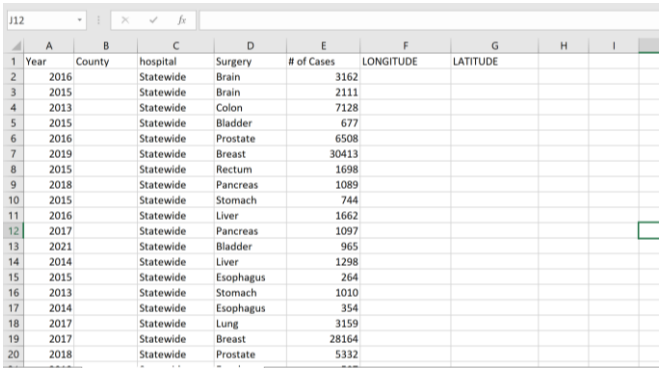
random intercept for every hospital. R 3.1.3 software was used.

In conclusion, the study aimed to talk about the impact travel distance can have in the health care sector biased towards cancers and talks about multiple variables such as clinical, demographic and geographic factors. My research question of the geographical pattern is not completely answered in the research paper but it does shed some light onto what a geographical location can do even in health care. It does talk about the distribution of certain types of cancer surgery being performed in further distances than other.

III. DATASET [1]

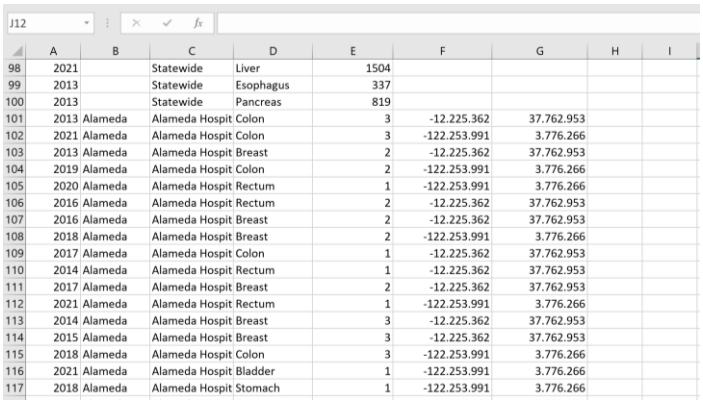
In my project, I utilized a dataset which I got from the data.gov website. The dataset includes details regarding the cancer treatments conducted in different California Hospitals. I have chosen the dataset "Number of Cancer Surgeries (Volume) Performed in California Hospitals" [1]. The dataset consists of seven columns of data containing data of type such as type of surgery, number of cases, county name, hospital name, and geographic coordinates of the surgery. The dataset has a good variety of data which can help in inspecting the trends, patterns, and differences in the number of cancer surgeries performed over a wide range of geographic locations and years within California. Using this dataset, the intent is to investigate patterns in surgical volumes across different hospitals and cancer types, as well as how surgical volumes have changed over time and geographically. This project aims to offer important insights on the state of cancer care delivery in California through thorough data analysis using Python, R, SQL, and AWS.

The screenshots below display the dataset [1], the first hundred lines of the dataset represents Statewide data which has no longitude and latitude data. The rest of the dataset which is about California has all the data. The dataset contains 19,482 rows of data and is of 1.52 MB.



	A	B	C	D	E	F	G
1	Year	County	hospital	Surgery	# of Cases	LONGITUDE	LATITUDE
2	2016		Statewide	Brain	3162		
3	2015		Statewide	Brain	2111		
4	2013		Statewide	Colon	7128		
5	2015		Statewide	Bladder	677		
6	2016		Statewide	Prostate	6508		
7	2019		Statewide	Breast	30413		
8	2015		Statewide	Rectum	1698		
9	2018		Statewide	Pancreas	1089		
10	2015		Statewide	Stomach	744		
11	2016		Statewide	Liver	1662		
12	2017		Statewide	Pancreas	1097		
13	2021		Statewide	Bladder	965		
14	2014		Statewide	Liver	1298		
15	2015		Statewide	Esophagus	264		
16	2013		Statewide	Stomach	1010		
17	2014		Statewide	Esophagus	354		
18	2017		Statewide	Lung	3159		
19	2017		Statewide	Breast	28164		
20	2018		Statewide	Prostate	5332		

Figure 1 – Dataset Preview in Excel



	A	B	C	D	E	F	G
98	2021		Statewide	Liver	1504		
99	2013		Statewide	Esophagus	337		
100	2013		Statewide	Pancreas	819		
101	2013	Alameda	Alameda Hospit	Colon	3	-12.225.362	37.762.953
102	2021	Alameda	Alameda Hospit	Colon	3	-122.253.991	3.776.266
103	2013	Alameda	Alameda Hospit	Breast	2	-12.225.362	37.762.953
104	2019	Alameda	Alameda Hospit	Colon	2	-122.253.991	3.776.266
105	2020	Alameda	Alameda Hospit	Rectum	1	-122.253.991	3.776.266
106	2016	Alameda	Alameda Hospit	Rectum	2	-12.225.362	37.762.953
107	2016	Alameda	Alameda Hospit	Breast	2	-12.225.362	37.762.953
108	2018	Alameda	Alameda Hospit	Breast	2	-122.253.991	3.776.266
109	2017	Alameda	Alameda Hospit	Colon	1	-12.225.362	37.762.953
110	2014	Alameda	Alameda Hospit	Rectum	1	-12.225.362	37.762.953
111	2017	Alameda	Alameda Hospit	Breast	2	-12.225.362	37.762.953
112	2021	Alameda	Alameda Hospit	Rectum	1	-122.253.991	3.776.266
113	2014	Alameda	Alameda Hospit	Breast	3	-12.225.362	37.762.953
114	2015	Alameda	Alameda Hospit	Breast	3	-12.225.362	37.762.953
115	2018	Alameda	Alameda Hospit	Colon	3	-122.253.991	3.776.266
116	2021	Alameda	Alameda Hospit	Bladder	1	-122.253.991	3.776.266
117	2018	Alameda	Alameda Hospit	Stomach	1	-122.253.991	3.776.266

Figure 2 – Dataset Preview in Excel

The NOIR data types in my dataset are.

Column Names(attributes)	**NOIR Datatypes**
Year	Ordinal
County	Nominal
Hospital	Nominal
Surgery	Nominal
No. of Cases	Ratio
Longitude	Interval
Latitude	Interval

Nominal: represents a label or any name to a thing.

Ordinal: represents groups of things that have a distinct ranking or order, but the gaps between them are not equal.

Interval: represents numbers which does not have a zero point, these values can go in negative and positive such as temperature.

Ratio: represents numbers which have a zero point making them not go under 0 such example would be age.

IV. METHODOLOGY

Data Collection: Get the dataset from data.gov that includes details on cancer treatments performed in hospitals throughout California.

Data Preprocessing: Handle missing values, eliminate duplicates, and format different data types to clean up the dataset.

Exploratory Data Analysis (EDA): In this step I planned out the visualizations that I will do in python,R,SQL, and aws.

Data Visualization: Geospatial maps, bar charts, line graph and stacked area graph.

Interpretation: Analyze the data and make conclusions about the patterns of cancer surgery in Californian hospitals.

The tools used in this project was the basic tools we use for python, R, SQL, and AWS. For python I used 'Jupyter Notebook' where for writing R scripts I used R studio and lastly 'mysql' for querying the database through SQL.

V. RESULTS

R

R Visualizations – Geospatial Map

Geospatial mapping techniques are very helpful in the evaluation and visualization of data in the health sector, for example, the trend and distributional of medical operations like cancer surgery. The integration of spatial data with statistics in healthcare by using geographic information systems allows researchers to investigate these relationships between geographic characteristics and healthcare outcomes..

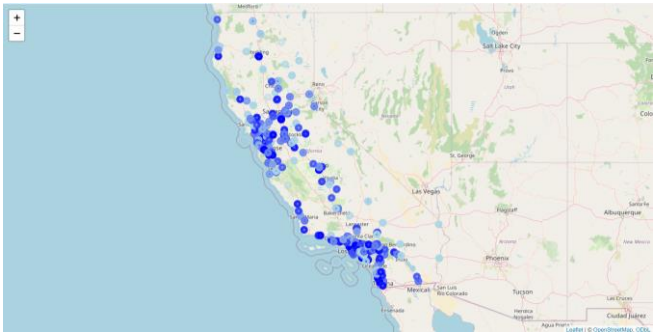


Figure 3- Geospatial Map

As we can see in figure 3, it is a geospatial map representing the California map. In the map we can see the distribution of data points such as dark blue and light blue. Dark blue data points indicate hospitals with a number of cases greater than or equal to the median number of cases whereas the light blue data points indicate hospitals with a number of cases less than the median number of cases. We can see two major hotspots in the California states, lets look into it closely.

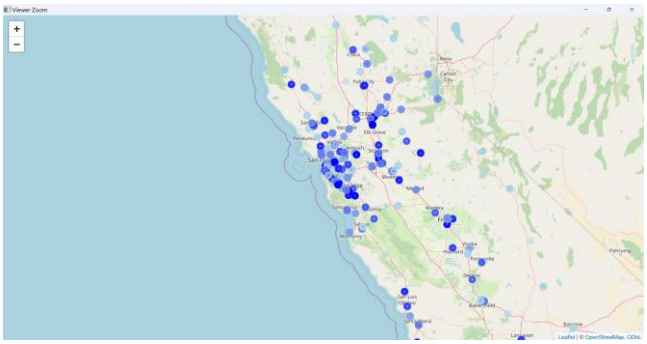


Figure 4 – Geospatial Map

In figure 4, we can see the San Francisco, San Jose counties is populated with datapoints indicating that there are many hospitals performing cancer surgeries. The strong clustering of data points in the counties of San Francisco and San Jose indicates a strong healthcare system serving urban areas. This concentration illustrates potential differences in healthcare access for underprivileged communities while also reflecting greater access to specialized cancer treatment services.

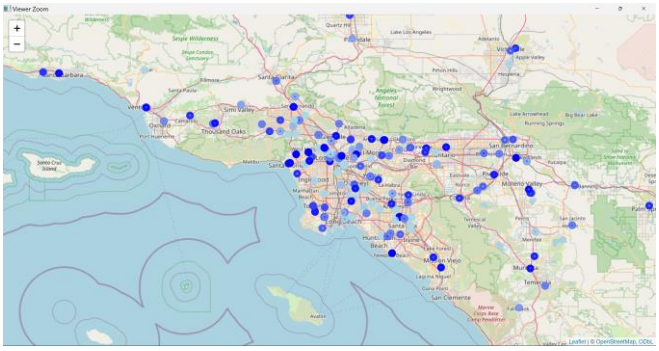


Figure 5 – Geospatial Map

Figure 5 shows that the Los Angeles County is home to many datapoints that show how many hospitals are doing cancer surgeries. The robust data point clustering in the Los Angeles counties suggests that urban areas are well-served by a robust healthcare system. This concentration reflects increased availability to specialized cancer treatment services while also illuminating potential disparities in healthcare access for poor communities.

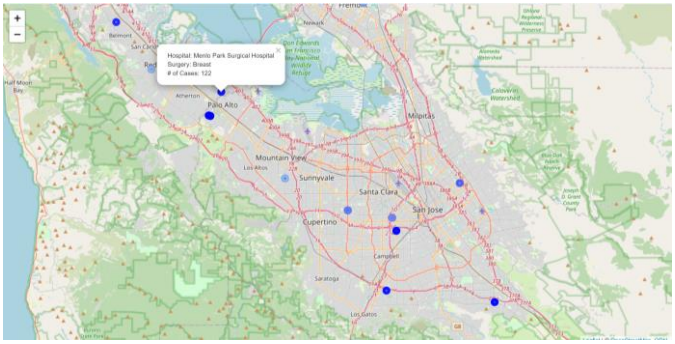


Figure 6 – Geospatial Map

In the above figure 6, when a datapoint a clicked on the map it shows the name of the hospital, the highest performed surgery in that hospital, and the total number of cases that hospital has performed.

Geospatial analysis in healthcare research benefits greatly from the Leaflet package in R, which offers a powerful framework for building interactive and specific maps. Researchers can produce interactive maps that illustrate spatial data, such as hospital locations, surgical volumes, and patient outcomes, by utilizing Leaflet's features. The use of interactive elements like pop-up data improves the maps' readability and makes it easier for users to study in-depth details about medical institutions and protocols.

The geospatial mapping of cancer procedures performed in California hospitals would help analyze a more efficient distribution and use of healthcare resources for cancer treatment. The analysis of surgery volumes across counties and hospitals can assess the effectiveness of healthcare delivery systems and establish regional inequalities in cancer care. Additionally, by observing high- and low-volume hospitals, politicians and healthcare professionals can take targeted actions in improving the standard and regulating the distribution of cancer care. The second research question of mine is found in this visualization as we can notice a geographical pattern in the distribution of cancer surgeries across the California state.

R Visualizations – Bar Chart

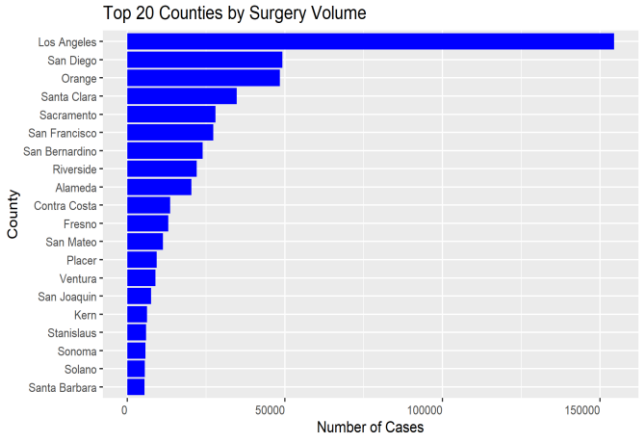


Figure 7 – Top 20 Counties by Surgery Volume

Based on the total number of treatments, the top 20 counties in California's surgery volume are shown visually in the bar chart. From the plot we can observe Los Angeles accounts for the maximum number of surgeries performed followed by San Diego and Orange County.

The code effectively converts raw data into a comprehensible graphical representation by utilizing R's ggplot2 package, which clarifies the differences in healthcare demand between various counties. In order to pick the best initiatives, allocate resources efficiently, and meet the unique needs of areas with greater surgery volumes, healthcare administrators and governments need to know this information.

Healthcare systems will improve quality by identifying the top 20 counties that have the highest number of surgeries as this would accord quality to the patients across the state. This will help in the perfect execution of services and aid in easing resource allocation.

This data can be of great help to the authorities as they get a idea of what type of surgery is most performed indicating the higher rate of the specific surgery. In this case, breast cancer is the highest where the officials can take note and bring more awareness for breast cancer and allocate more funds towards promoting it.

Python

Python Visualizations – Bar Chart

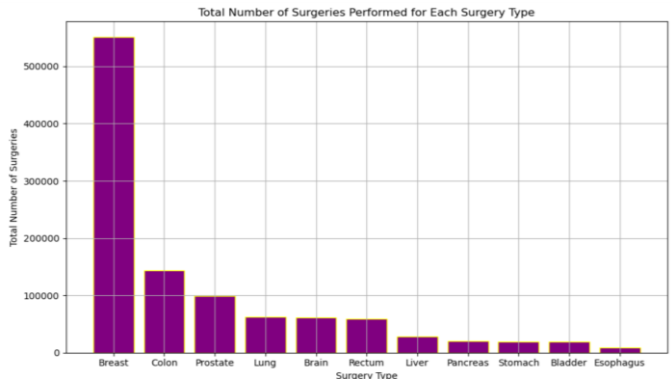


Figure 8 – Total Number of Surgeries Performed for Each Type of Surgery

With the bar chart, I have visualized the total number of surgeries for each type of surgery performed in California, where each bar constitutes one kind of surgery, and its height is the number of procedures performed. One of the general, but most common surgeries within the dataset are easily seen in this graph, thanks to bars related to the surgeries being sorted in descending order according to their frequency. The graph is well colored and easy to read due to the bright purple bars related to the yellow lines. In reference to the bar chart, we can see the most common type of surgery is breast cancer followed by colon and prostate.

Using matplotlib, a popular plotting package, the Python code displays the findings after processing the dataset effectively with pandas, an strong data manipulation library. A summary of surgery volumes is produced by the code by reading the dataset into a pandas DataFrame, classifying the data by types of surgery, and adding the total number of procedures for each type. The procedures that are most frequently performed are sure to be prominently focused on in the bar chart by sorting the surgeries in descending order.

Python Visualizations – Line Graph

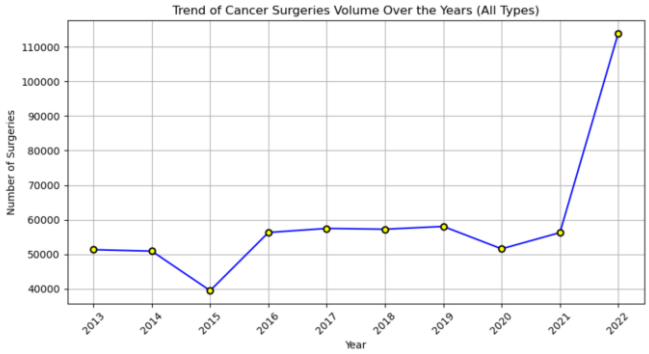


Figure 9 – Trends of Cancer Surgeries Volume Over the Years

The volume of cancer surgeries in California over time, including all types of surgeries, is depicted by a line graph. Plotting each data point against the number of surgeries performed represents an individual year. The graph provides insights into the shifting picture of cancer care in the region by displaying trends and variations in the number of surgeries over time. The viewer is directed through the annual trajectory of surgeries by the blue line, which has circular markers highlighted in yellow. Looking at the line graph it is possible to say there is a massive increase in the cancer surgeries between 2020-2022. We can see a slight drop in 2020 due to pandemic which avoids outside food which might have been the reason for a decrease in cancer.

The python code uses the pandas library to read and process data, to extract relevant information about the volume of surgeries that is done annually. By grouping the data according to years then summing procedures performed per year will produce a brief overview about trends over time regarding operation rates. For designing line graph, matplotlib is used and others including grids and markers customization.

Healthcare professionals, policymakers, and researchers engaged in managing cancer care and resource allocation will find this information important. When volumes of cancer surgeries are analyzed for trends over a certain period, the authorities or other groups are able to understand which direction it is taking thereby being able to predict or project future demands so that planned actions can be taken on specific issues relating to emerging needs of healthcare system. In conclusion we can give awareness to the public of what is to come and how to be safe in not being diagnosed. My first research question can be answered by this line graph as it tells the volume of all cancer surgeries that changed over the years in California.

Python Visualizations – Stacked Area Plot

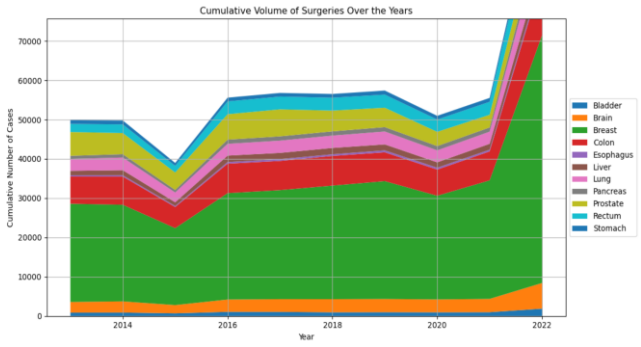


Figure 10 – Cumulative Volume of Surgeries Over the Years

With a different colored region for each type of surgery, the stacked area plot shows the total number of surgeries over time. The years are represented on the x-axis, while the total number of surgeries is displayed on the y-axis. The area under the curve for each surgery type shows its total volume over the years, and each surgery type adds to the overall trend. The plot makes it simple to see patterns in the surgeries carried out over various years and surgical specialties. From the plot we can see breast cancer has the bigger volume followed by colon.

The dataset is first read in the code given to classify the number of surgical cases by year and type of surgery. Next, using the stackplot() function in Matplotlib, a stacked area plot is made, with the years shown on the x-axis and the total amount of cases displayed on the y-axis. The y-axis limits are adjusted with the plt.ylim() method to improve visibility and make sure that even surgeries with lesser volumes are easily seen in the graphic.

It is important for those working in healthcare administration, policy making, and research to comprehend how many surgeries have been done since they began being performed because such knowledge helps them understand which types of operations will be required as time goes by hence allocation of resources, staffing levels as well as organizations planning itself. It can also point out those places where there may either be need for more or different kinds of supports leading to the creation of efficiencies in service delivery as well as performance enhancements. The fifth research question of mine is answered here as it tells the most performed surgeries in California and how it has changed over the years.

SQL

```
CREATE DATABASE IF NOT EXISTS akamma_ait_project;
USE akamma_ait_project;
```

With the help of CREATE statement in SQL, I created a database named ‘akamma_ait_project’ and then used it to work in it. I then imported the dataset into a table named ‘cancervolume1’ using the import wizard feature in MySQL.

SQL Query 1 - Regional Variations in Cancer Surgery Types: Determining Differences Between Counties

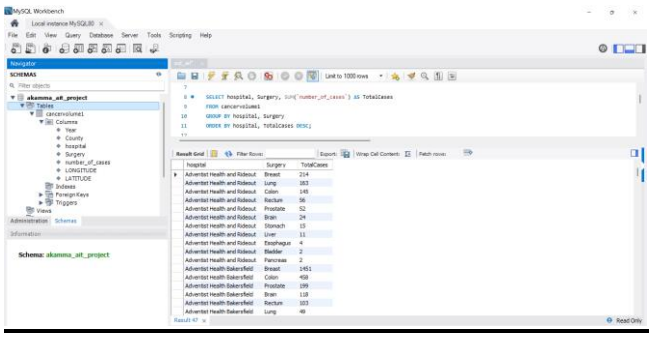


Figure 11

In this query it describes the total surgeries of all type of cancer for every hospital. The highest done surgeries of a type of cancer are first displayed for the hospital. The total number of cases of each kind of surgery performed in each hospital is outlined in

this SQL query. With such information, we can have the big picture of how cancer surgeries are shared among different health facilities by grouping by the facility and type of surgery. It is followed by hospitals' names then their respective numbers of surgeries and can therefore enable quick comparison among them. This query gives me my fourth research question answer as it shows a type of surgery being performed more in one place and less in another.

SQL Query 2 - Identifying Hospitals with the Highest Increase in Surgeries from 2013 to 2021:

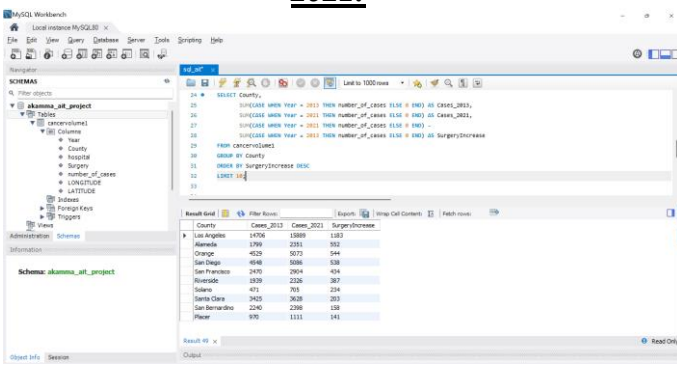
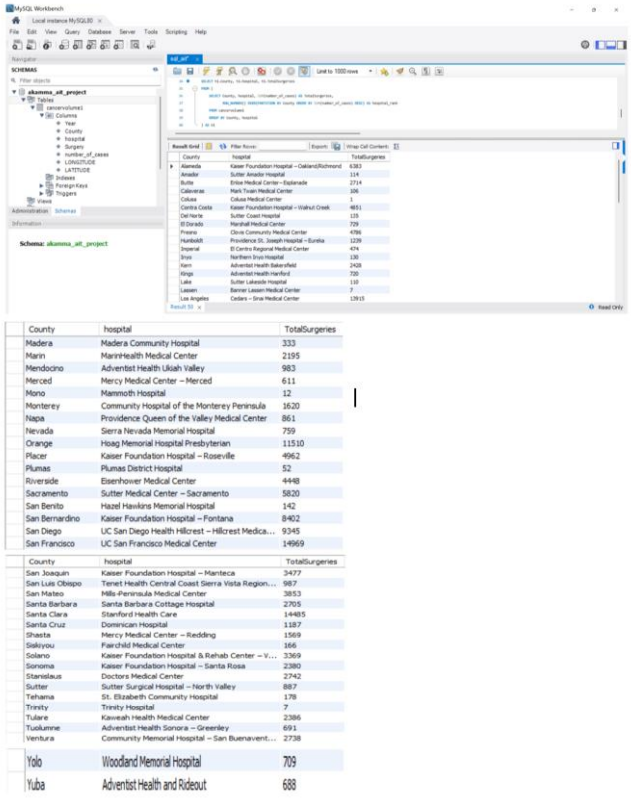


Figure 12

This SQL query looks at the difference in the number of cancer surgeries performed between 2013 and 2021 for each of the counties. It calculates the total number of cases for each year and takes the difference between cases performed in 2021 and those in 2013, which gives the increase in surgeries for that county. Then it presents the information grouped by county and ordered in descending order by the amount of increase in surgeries. This gives an insight into which counties had the most increased cases of cancer surgeries for the given time frame. This analysis is a good indication of where there is a significant change in the demand for cancer treatment and may indicate a place in need of more health resources or further action.

SQL Query 3 - Find the hospital with the highest number of surgeries in each county:

This SQL query below brings out information on the hospitals that have performed maximum cancer surgeries in respective counties. It ranks each hospital of a county according to the total number of cases by using the ROW_NUMBER() function. Therefore, the query selects those hospitals having a rank of 1 in a county, the top-performing hospital in the context of cancer surgery. This type of analysis helps to identify leading health care facilities according to the county in addressing the need for cancer treatment. Thus, such information is brought into light on the resource allocation and planning in the health care sector.



County	hospital	TotalSurgeries
Madera	Madera Community Hospital	333
Marin	MarinHealth Medical Center	2195
Mendocino	Adventist Health Ukiah Valley	983
Merced	Mercy Medical Center – Merced	611
Mono	Mammoth Hospital	12
Monterey	Community Hospital of the Monterey Peninsula	1620
Napa	Providence Queen of the Valley Medical Center	861
Nevada	Sierra Nevada Memorial Hospital	759
Orange	Hoag Memorial Hospital Presbyterian	11510
Placer	Kaiser Foundation Hospital – Roseville	4962
Plumas	Plumas District Hospital	52
Riverside	Eisenhower Medical Center	4448
Sacramento	Sutter Medical Center – Sacramento	5820
San Benito	Hazel Hawkins Memorial Hospital	142
San Bernardino	Kaiser Foundation Hospital – Fontana	8402
San Diego	UC San Diego Health Hillcrest – Hillcrest Medica...	9345
San Francisco	UC San Francisco Medical Center	14969

County	hospital	TotalSurgeries
San Joaquin	Kaiser Foundation Hospital – Manteca	3477
San Luis Obispo	Tenet Health Central Coast Sierra Vista Region...	987
San Mateo	Mills-Peninsula Medical Center	3853
Santa Barbara	Santa Barbara Cottage Hospital	2705
Santa Clara	Stanford Health Care	14485
Santa Cruz	Dominican Hospital	1187
Shasta	Mercy Medical Center – Redding	1569
Siskiyou	Fairchild Medical Center	166
Solano	Kaiser Foundation Hospital & Rehab Center – V...	3369
Sonoma	Kaiser Foundation Hospital – Santa Rosa	2380
Stanislaus	Doctors Medical Center	2742
Sutter	Sutter Surgical Hospital – North Valley	887
Tehama	St. Elizabeth Community Hospital	178
Trinity	Trinity Hospital	7
Tulare	Kaweah Health Medical Center	2386
Tuolumne	Adventist Health Sonora – Greenley	691
Ventura	Community Memorial Hospital – San Buenavent...	2738
Yolo	Woodland Memorial Hospital	709
Yuba	Adventist Health and Rideout	688

Figure 13

AMAZON WEB SERVICES (AWS)

Amazon S3 - It is Amazon Web Service's highly scalable, secure solution for storing data in the cloud. It is inherently very reliable, has high availability, and delivers great performance—high durability for storing and retrieving data from anywhere on the web.

Steps involved in creating a bucket in S3 Bucket:

Step 1: Launch the AWS Learner Lab.

Step 2: Launch your AWS Account and launch S3.

Step 3: Click on Create Bucket.

Step 4: Provide a Bucket name(Bucket name should be unique).

Step 5: Upload the dataset you want to visualize.

Step 6: Launch Glue DataBrew.
Step 7: Launch Dataset and provide a name to the dataset.
Step 8: Provide your source as the S3 bucket you created.
Step 9: Provide delimiter and select the role.
Step 10: Click on Create dataset.
Step 11: Create Project.
Step 12: Select the source of the dataset. Select your S3 bucket.
Step 13: Select the dataset and provide delimiter and Role
Step 14: Click on Create project.
Step 15: Click on Run job profile and the visualizations will get displayed.

Dataset Profile Overview

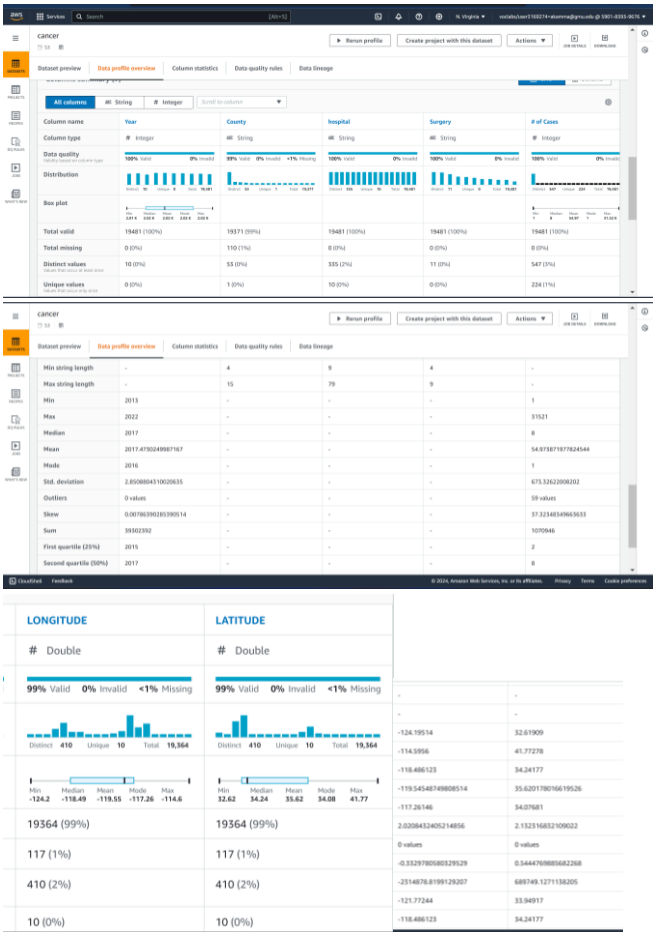


Figure 14

The Mean, Mode, Median, Min, Max, and Standard deviation values for each column in the dataset are displayed in the preview. Additionally, it displays the dataset's columns' distinct and unique values.

Correlation Matrix

The most common correlation coefficient, ranging from -1 to 1, is the main measure used to describe the correlation between two variables. The coefficient of zero reflects no correlation. The coefficient ranges from 0 up to 1 or even down to -1 to mark the strength of the relationship between variables as well as its direction. The numerical association of variables of correlation matrix presented in the dataset specifies and defines a correlation relationship numerically. A coefficient nearly equal to 1 or -1 indicates a very good positive or negative linear relationship, respectively. A zero coefficient indicates a weak or no linear relationship. The direction of correlation indicates whether variables vary together or in opposite directions with coefficients of +1 or -1 that indicate perfect correlation.

Figure 15

Dataset Statistics

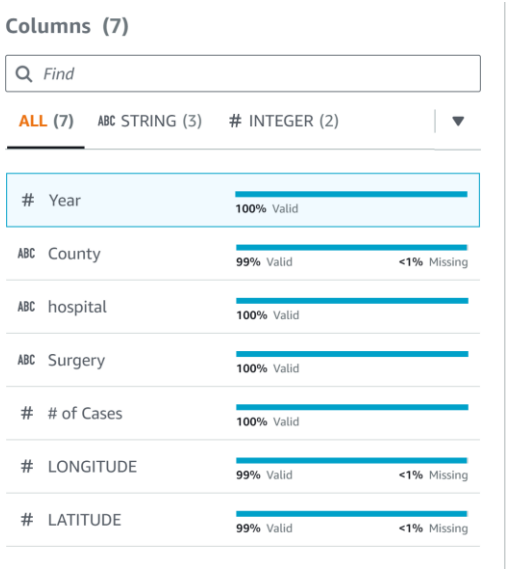


Figure 16

The image above shows the columns of the dataset and tells if any missing /null values are present. Since my dataset has the first 100 rows of statewide data the longitude, latitude, and county values are null and the statistics show it accurately.

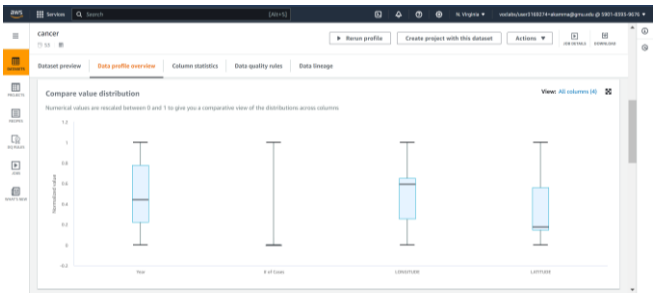


Figure 17 – Box Plot

VI. LIMITATIONS

The problems I faced using this dataset was the first 100 rows of data as it was representing the statewide data so when I had to create visualizations R and python were not working so smooth so I had to not include the first 100 lines of data in the code. But let alone the first 100 rows of data does talk about the statewide data but not in specific to a state name which does not give you much to do with it as you cannot visualize the statewide data. Hence I could not answer one of my research questions which was the third one. Lastly the '# of cases' column was a little tricky to work with in the start of the project as I was getting errors of a special character being used in the name but I tackled it.

VII. FUTURE WORK

Future work may extend the geographical scope of the data out of California and thus help better understand national trends in cancer surgery. Other variables added, such as patient demographics, surgical outcomes, and characteristics of the health facilities, could provide more depth in understanding the factors driving volume and outcomes of cancer surgery. Additionally the effectiveness of interventions to improve cancer screening rates and early detection, which would give an idea about the need for advanced surgical interventions, and longitudinal studies regarding patient outcomes related to different types of cancer surgeries could help in the evaluation of long-term efficacy and quality of care.

VIII. CONCLUSION

This dataset shows analysis in terms of surgery volume data across the counties of California. There is a significant difference in health demand in terms of the analysis between the top 20 counties with the highest surgery volumes. The analysis helps conclude that regions like Los Angeles, San Diego, and Orange face large healthcare needs that can be influenced by things like population density, healthcare infrastructure, and demographic characteristics. However, disparity in surgery volume across counties points out the need to allocate healthcare resources effectively to various counties and provide appropriate intervention strategies that can make the healthcare provision system better across the state. An understanding of these disparities can help the policymakers and providers in developing more effective strategies for addressing regional healthcare needs and delivering improved healthcare delivery systems. In conclusion the information/insights is a great way to learn about the cancer volume in California.

REFERENCES

[1] “Number of cancer surgeries (Volume) performed in California hospitals - catalog,” Dec. 02,2023. https://catalog.data.gov/dataset/number-of-cancer-surgeries-volume-performed-in-california-hospitals-a3f18

[2] M. O’Sullivan, “Safety in numbers: Cancer surgeries in California hospitals,”- “CALIFORNIA HEALTHCARE FOUNDATION,” 2015. https://www.chcf.org/wp-content/uploads/2017/12/PDFSafetyCancerSurgeriesHospitals.pdf

[3] A. Diaz, A. Schoenbrunner, J. Cloyd, and T. M. Pawlik, “Geographic Distribution of Adult Inpatient Surgery Capability in the USA,” Journal of Gastrointestinal Surgery, vol. 23, no. 8. Elsevier BV, pp. 1652–1660, Aug. 2019. doi: 10.1007/s11605-018-04078-9.

[4] C. B. Begg, “Impact of Hospital Volume on Operative Mortality for Major Cancer Surgery,” JAMA, vol. 280, no. 20. American Medical Association (AMA), p. 1747, Nov. 25, 1998. doi: 10.1001/jama.280.20.1747.

[5] N. Wasif, Y.-H. Chang, B. A. Pockaj, R. J. Gray, A. Mathur, and D. Etzioni, “Association of Distance Traveled for Surgery with Short- and Long-Term Cancer Outcomes,” Annals of Surgical Oncology, vol. 23, no. 11. Springer Science and Business Media LLC, pp. 3444–3452, Apr. 28, 2016. doi: 10.1245/s10434-016-5242-z.

[6] Amit Chowdary Kamma, “ackamma_ProjectAssignment2”,03,2024 “Blackboard Learn,” https://mymasonportal.gmu.edu/

[7] Amit Chowdary Kamma, “akamma_projectassignment3”,03,2024 “Blackboard Learn,” https://mymasonportal.gmu.edu/