# Navigating the Shifting Dynamics of Cancer Surgeries: A Comprehensive Analysis from ICD-9-CM to ICD-10-CM/PCS

George Mason University
AIT 580 | Abhishek Anish

**Abstract**

Cancer surgeries have witnessed significant transformations in reporting and coding practices, notably with the transition from ICD-9-CM to ICD-10-CM/PCS. This study delves into the nuanced trends in surgery volumes across diverse cancer types, leveraging a comprehensive analysis spanning the years 2013 to 2021. The impact of the coding transition on reported surgery volumes is scrutinized, revealing noteworthy fluctuations. Methodologically, the research employs a multifaceted approach incorporating data querying, regression analysis, and visualizations using R, Python, and SQL. The findings underscore a surge in the numbers of seven specific surgery types post-coding transition, while an evident dip in 2020 is attributed to the unprecedented challenges posed by the COVID-19 pandemic. Geographic variations in surgery volumes are explored, with Los Angeles consistently leading, followed by San Diego and Orange County, mirroring the population and healthcare infrastructure distribution. Furthermore, the study identifies a select group of 13 hospitals demonstrating remarkable consistency, performing at least one surgery for all 11 cancer types each year throughout the study period. Notably, a Poisson regression model proves reasonably effective in predicting surgery volumes for most types, albeit encountering challenges in stomach cancer predictions. In conclusion, this research sheds light on the dynamic landscape of cancer surgeries, elucidating trends pre- and post-coding transition. The findings contribute to a nuanced understanding of the evolving surgical landscape and hold implications for healthcare planning and resource allocation.

Keywords - Cancer Surgery, ICD Coding, Trends Analysis, Geographic Variations, COVID-19 Impact, Hospitals Consistency, Poisson Regression, R, Python, SQL

## I. INTRODUCTION

The landscape of cancer treatment has seen significant changes over the years, with advancements in surgical techniques, improved diagnostic tools, and the evolution of healthcare coding systems. One such pivotal change was the transition from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) to the Tenth Revision (ICD-10-CM/PCS). This transition, while aimed at improving the accuracy and specificity of disease reporting, raised questions about its impact on the reported volume of cancer surgeries.

Understanding the trends in different types of cancer surgeries before and after this transition is crucial for healthcare planning and resource allocation. Moreover, comparing the volumes of surgeries for each type of cancer can provide insights into the prevalence of certain procedures and the burden of different cancers on the healthcare system. Additionally, identifying the geographical distribution of these surgeries can highlight disparities in healthcare access and inform targeted interventions.

This study aimed to address these questions by analysing the trends in cancer surgeries from 2013 to 2021, comparing the volumes of different types of surgeries, identifying counties with the highest number of surgeries, and predicting future surgery volumes based on the new ICD-10-CM/PCS coding. The findings of this study have the potential to inform healthcare policies, improve resource allocation, and ultimately enhance patient care.

In this research paper, the aim is to address the following questions:

1. What are the trends in different types of cancer surgeries before and after the transition from ICD-9-CM to ICD-10-CM/PCS coding, and did this transition impact the reported surgery volume?
2. How do the volumes of surgeries for each type compare? Are certain types of surgeries more common than others?
3. Which counties consistently had the highest number of surgeries performed over the years?
4. Which hospitals consistently performed at least one surgery for all 11 types each year from 2013 to 2021?
5. Is it possible to predict the number of surgeries done per county according to the new coding (ICD-10-CM/PCS)?

This paper offers both the analytical methodologies and the derived conclusions to address the stated research problems. The structure of the paper is as follows: Section II delves into the relevant literature, followed by a comprehensive overview of the dataset in Section III. The overarching framework for

data processing and analysis is elaborated in Section IV. Section V presents the analytical insights pertaining to the trends in cancer surgeries conducted in California hospitals. Section VI introduces a Poisson regression model for forecasting the number of surgeries in a county for a specific year post the coding transition. Section VII encompasses a discussion on the research, its findings, and any intricacies unearthed during the research and analysis process. The paper culminates with suggestions for future research in Section VIII.

## II. RELATED WORK

The transition from ICD-9-CM to ICD-10-CM/PCS coding in the medical field has had a significant impact on the reporting of surgical procedures, including cancer surgeries. This transition has been the subject of several studies and reports, which have explored its effects on the volume of reported surgeries and the trends in different types of cancer surgeries.

In a report titled "Small Numbers Can Have Big Consequences," Laurence Baker and Maryann O'Sullivan discuss the dangerously low numbers of cancer surgeries performed at most California hospitals. They highlight the strong research linking low hospital surgery volume to an increased risk of mortality and complications. The report also emphasizes the persistence of this problem despite the transition to more recent data and coding systems.[1]

The Healthcare Cost and Utilization Project (HCAI) provides an annual report on the volume of cancer surgeries performed in California hospitals. This report offers valuable insights into the volumes of surgeries for each type of cancer and allows for comparisons between them. It also discusses hospitals performing one or two surgeries for 11 different types of cancer. In 2021, 84.7% of California hospitals performed one or two surgeries for at least one of these cancers. [2]

Another report by Maryann O'Sullivan, titled "Safety in Numbers: Cancer Surgeries in California Hospitals," further explores the link between the volume of surgeries and patient outcomes. This report reiterates the findings of the previous studies, emphasizing that hospitals performing a small number of cancer surgeries are more likely to have worse patient outcomes. [3]

These studies provide a comprehensive overview of the trends in different types of cancer surgeries before and after the transition from ICD-9-CM to ICD-10-CM/PCS coding. They highlight the impact of this transition on the reported surgery volume and underscore the need for further research to improve patient outcomes. However, predicting the number of surgeries done per county according to the new coding remains a complex task that requires more advanced statistical models and methodologies.

## III. DATASET

The dataset for this research was obtained from the "Number of Cancer Surgeries (Volume) Performed in California Hospitals" available on the Data.gov catalog. This dataset, which has a total of 17519 records, covers 395 hospitals across 53 counties in California. It contains the number (volume) for 11 types of cancer surgeries (bladder, breast, brain, colon, esophagus, liver, lung, pancreas, prostate, rectum, and stomach) performed in these hospitals.[4]

The data are reported for January – September 2015 due to coding changes from ICD-9-CM to ICD-10-CM/PCS for procedures, which began on October 1, 2015. Therefore, for the year 2015, the dataset only contains data for 9 months, deviating from the normal 12 months. The dataset includes procedures performed in both inpatient and outpatient settings for breast cancer surgeries, while for all other types of cancer surgeries, the dataset contains surgeries performed in the inpatient hospital setting.

Figure 1 shows the attributes in the dataset. The records consist of all four datatypes from nominal to ordinal, interval, and ratio. This comprehensive dataset provides valuable insights into the trends and patterns in cancer surgeries performed in California hospitals.

**Figure 1**

| Dataset | |
|---|---|
| Year | Interval |
| Year index (Derived from year) | Ordinal |
| County | Nominal (53 levels) |
| Hospital | Nominal (395 levels) |
| OSHPDID | Nominal – Hospital identifier ID |
| Surgery | Nominal (11 levels) |
| # of cases (ICD 9) | Ratio |
| # of cases (ICD 10) | Ratio |
| Latitude of the Hospital | Interval |
| Longitude of the Hospital | Interval |

## IV. HIGHER LEVEL FRAMEWORK

This research project relied on two programming languages and a relational database query language for data cleaning, data wrangling, data exploration, statistical summarization, and analysis, as well as graphical visualizations. The Python programming language and MS Excel served as the tools for data cleaning and wrangling, addressing issues such as duplicates and hospital name changes. In the data preprocessing phase, a crucial step involved merging the number of cases for ICD-9 and ICD-10 into a single, continuous column. This process ensured that no data was lost during the transition, as ICD-9 cases seamlessly transitioned into ICD-10 cases. This continuous representation facilitated a comprehensive analysis, maintaining the integrity of the dataset. Once the data was prepared, Python was used for graphical visualizations, MySQL for database querying, and R Studio for Poisson regression analysis.

The exploration, summarization, and analysis aimed to investigate the dataset from two perspectives. The macro-level approach considered statewide data, examining trends in various types of cancer surgeries throughout the state of California as a whole and how coding practices influenced

them. On the micro-level, the analysis delved into county-wise and hospital-wise data.

Exploration included plotting graphical visualizations for each surgery throughout the years, Querying and visualizations to find the top counties for each surgery for each year, Poisson regression to predict the number of surgeries done for a particular year in a particular county.

## V. ANALYSIS

In this project the number of cancer surgeries was analysed from various perspectives. First, the statewide data was taken into account where each surgery was plotted to visualize how the number of surgeries varied throughout the years. Considering how the coding change took place in 2015, focusing on the trends for each surgery before and after 2015 will reveal how the coding change impacted the number of surgeries performed.

**Figure 2**

```
df1.head()
```

| | Year | Year Index | County | Hospital | OSHPDID | Surgery | nCases |
|---|---|---|---|---|---|---|---|
| 0 | 2013 | Year1 | Alameda | Alameda Hospital | 106010735 | Stomach | 1 |
| 1 | 2013 | Year1 | Alameda | Alameda Hospital | 106010735 | Colon | 3 |
| 2 | 2013 | Year1 | Alameda | Alameda Hospital | 106010735 | Breast | 2 |
| 3 | 2013 | Year1 | Alameda | Alta Bates Summit Medical Center – Alta Bates ... | 106010739 | Brain | 8 |
| 4 | 2013 | Year1 | Alameda | Alta Bates Summit Medical Center – Alta Bates ... | 106010739 | Colon | 12 |

Figure 2 shows the first 5 records in the data frame.

**Figure 3.a**

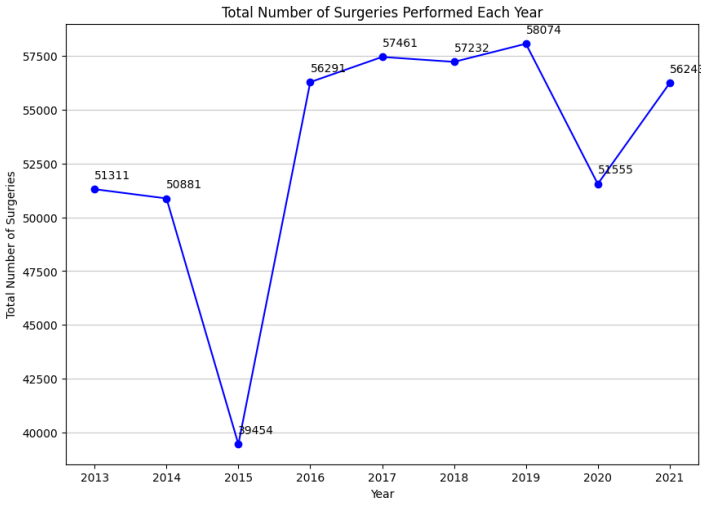



Figure 3.a illustrates the total number of surgeries performed each year, representing univariate analysis for the 'Year' variable, which is of interval data type.

**Figure 3.b**

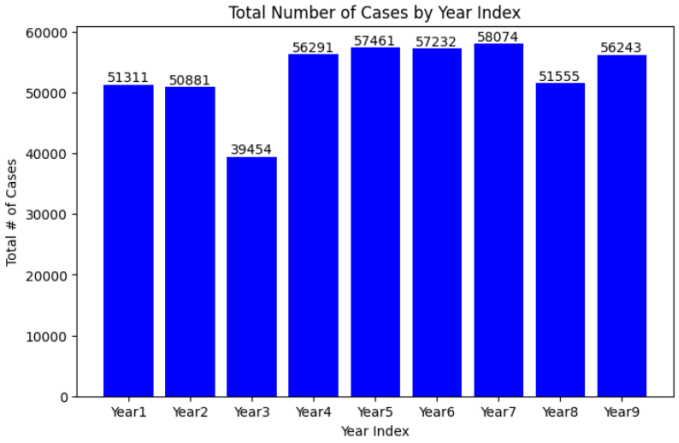



Figure 3.b displays the total number of surgeries performed for each year index, presenting univariate analysis for the 'Year Index' variable, which is of ordinal datatype.

Both Figure 3.a and Figure 3.b clearly depict an increase in the number of surgeries after the coding change, starting from 2016 or Year 4 onwards.

**Figure 4.a**

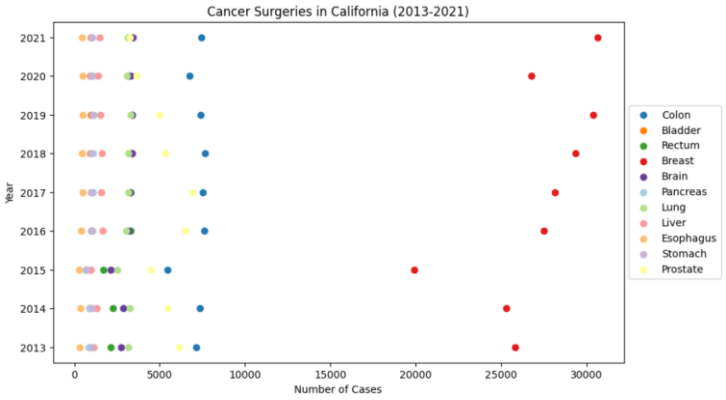



**Figure 4.b**

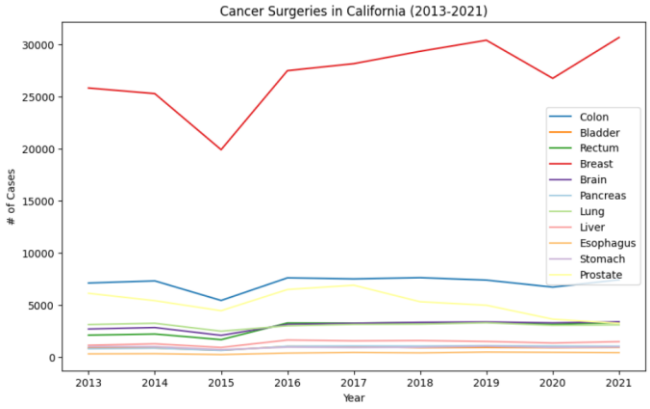



Figure 4.a and 4.b present a multivariate analysis of Year and the number of cases, offering an overview of how the volumes of each surgery have varied over the years. Breast surgery exhibits a wide range (around 30,000), in contrast to other surgeries (less than 10,000), influenced by considerations of both inpatient and outpatient settings. The figures highlight a dip in the number of surgeries for all procedures in 2020,

primarily due to the impact of COVID-19. Additionally, there is a dip in 2015, as the data collected for that year only spans 9 months due to the coding change. Breaking down this graph for each surgery effectively visualizes and quantifies how the volume of each surgery has changed over the years, enabling a detailed comparison of the impact of coding changes.

**Figure 5.a**

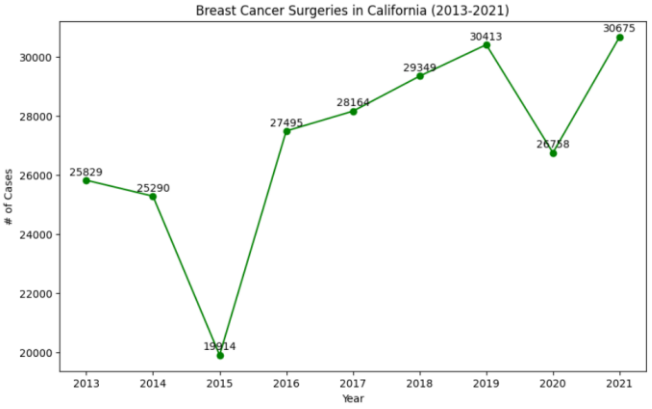



Figure 5.a efficiently illustrates a significant increase in the number of surgeries for breast cancer following the coding change. The count before the change was 25,290, and after the change, it rose to around 27,500, continuing to increase thereafter.

**Figure 5.b**

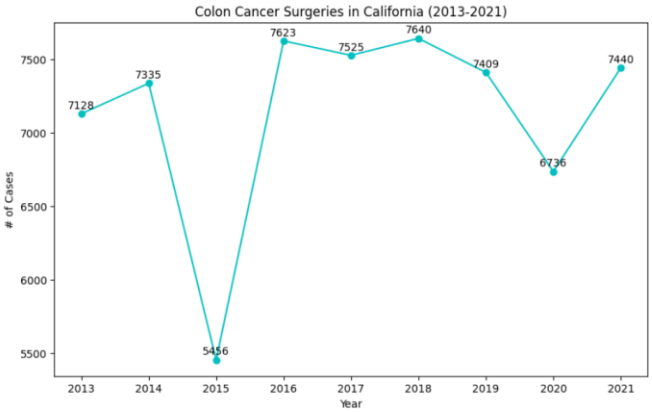



Figure 5.b indicates a slight increase in the number of surgeries for colon cancer after the coding change, with values of 7,335 before and 7,623 after. The post-change values remain relatively constant.

**Figure 5.c**

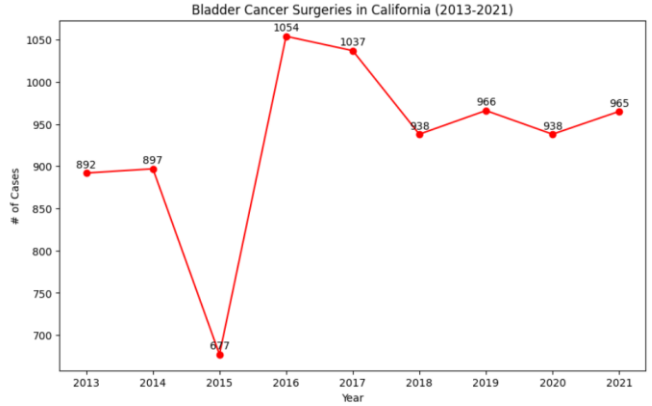



Figure 5.c illustrates an increase in the number of surgeries for bladder cancer after the coding change. The values before the coding change range around 890, while after the change, they show an increase to around 1,050 for the next two years, eventually stabilizing around 950 for the subsequent years.

**Figure 5.d**

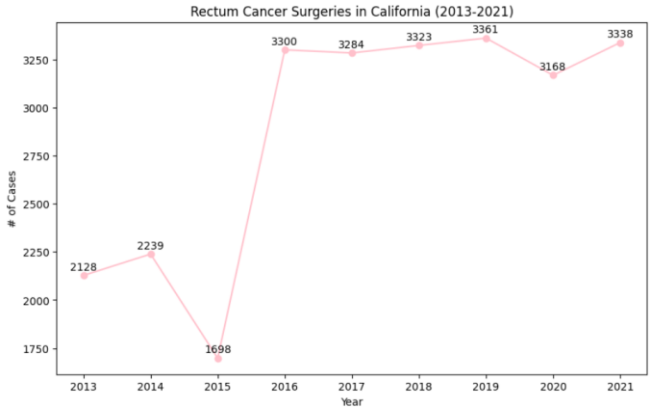



Figure 5.d demonstrates a significant increase in the number of surgeries for rectum cancer after the coding change. The values before the coding change range around 2,200, while after the change, they consistently hover around 3,300.

**Figure 5.e**

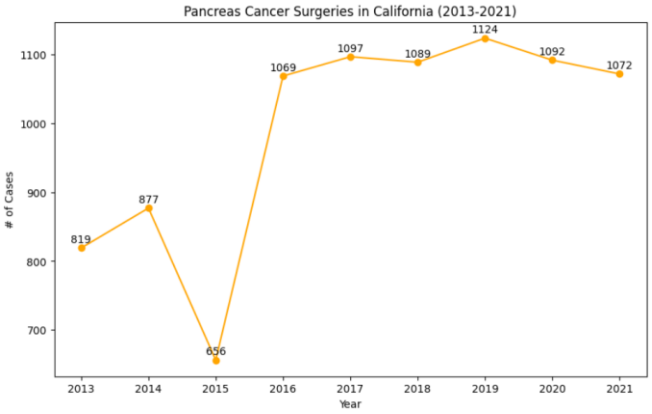



Figure 5.e indicates an increase in the number of surgeries for pancreatic cancer after the coding change. The values before the coding change range around 850, while after the change, they consistently remain around 1,100 for the subsequent years. Notably, this graph reveals that the number did not experience the occasional dip in 2020, unlike most other surgeries.

**Figure 5.f**

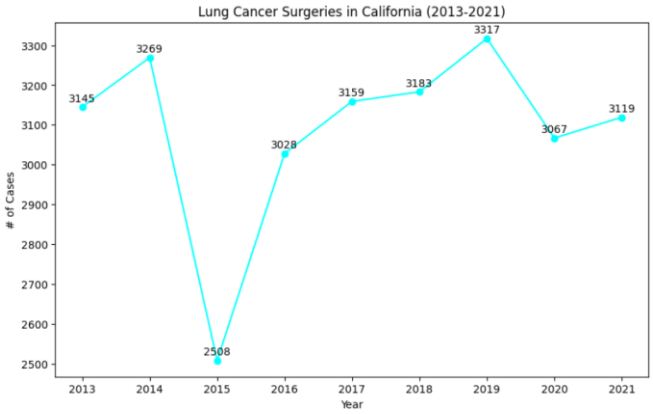



Figure 5.f indicates a slight decrease in the number of lung cancer surgeries after the coding change, despite recording the highest value in 2019. It is noteworthy that the coding changes haven't significantly affected the numbers.

**Figure 5.g**

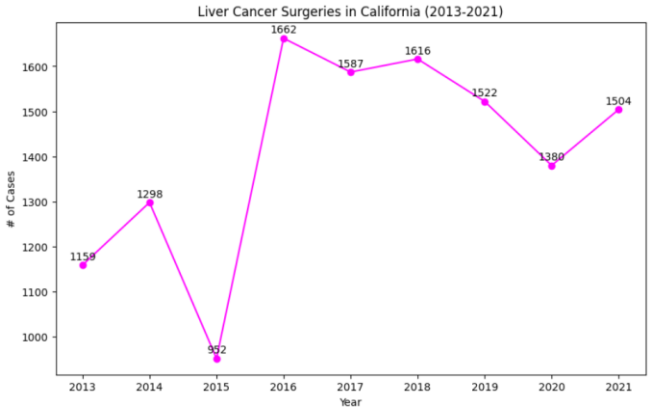



Figure 5.g shows that the numbers for liver cancer surgeries were in the range of 1,100-1,300 before the change and ranged from 1,500-1,650 after, indicating a significant increase due to the coding change.

**Figure 5.h**

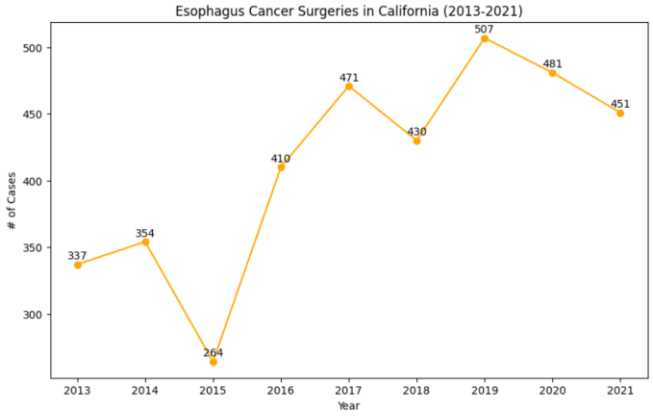



Figure 5.h shows an increase in the number of esophagus cancer surgeries after the coding change. It is noteworthy that esophagus surgeries have the smallest numbers compared to the rest of the surgeries analysed (Less than 500). Interestingly, this is one of the surgery types where the number did not dip during 2020.

**Figure 5.i**

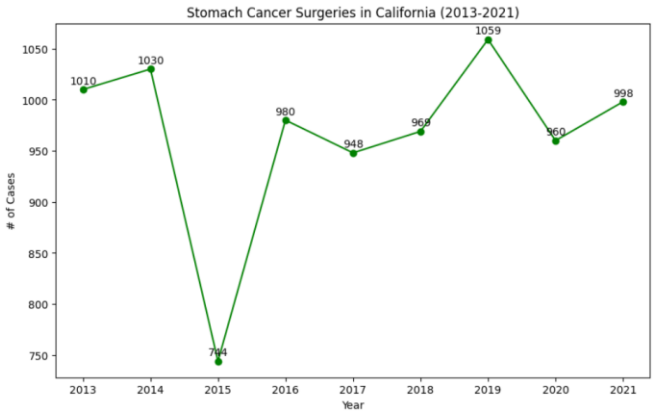



Figure 5.i indicates a slight decrease in the number of stomach cancer surgeries after the coding change, despite recording the highest value in 2019. It is noteworthy that the coding changes haven't significantly affected the numbers.

**Figure 5.j**

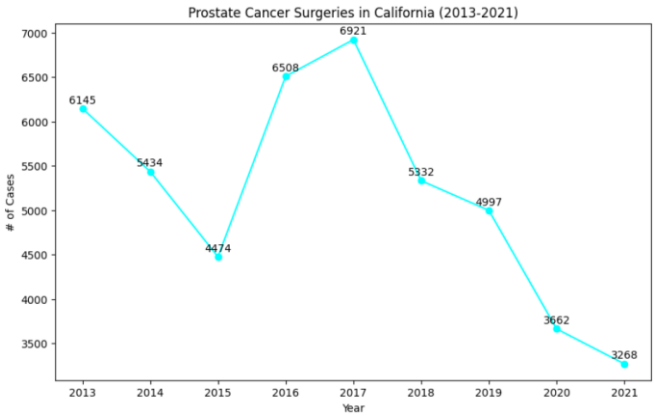



Figure 5.j displays the plot for prostate cancer surgeries, revealing an interesting trend. It is the only cancer type that showed values for certain years (2020 and 2021) less than the year 2015, despite 2015 only having data for 9 months. The number rose for two years after the coding change but then drastically decreased, reaching its lowest in the years 2020 and 2021.

**Figure 5.k**

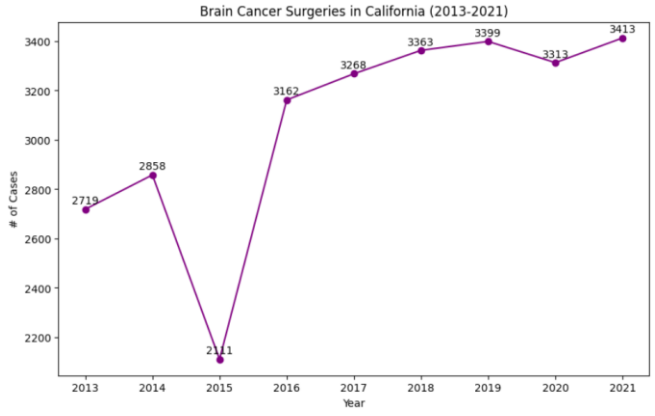



Figure 5.k shows the plot for brain cancer surgeries. The numbers before the coding changes were less than 3000 per

year, whereas the numbers after the change are above 3100. The numbers steadily increase, apart from the usual dip that occurred in 2020.

**Figure 6**

| Surgery | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Bladder | 8.0 | 960.875 | 59.016795 | 892.0 | 927.75 | 951.5 | 983.75 | 1054.0 |
| Brain | 8.0 | 3186.875 | 261.130202 | 2719.0 | 3086.00 | 3290.5 | 3372.00 | 3413.0 |
| Breast | 8.0 | 27996.625 | 2024.130497 | 25290.0 | 26525.75 | 27829.5 | 29615.00 | 30675.0 |
| Colon | 8.0 | 7354.500 | 299.327341 | 6736.0 | 7283.25 | 7424.5 | 7549.50 | 7640.0 |
| Esophagus | 8.0 | 430.125 | 60.328713 | 337.0 | 396.00 | 440.5 | 473.50 | 507.0 |
| Liver | 8.0 | 1466.000 | 173.207555 | 1159.0 | 1359.50 | 1513.0 | 1594.25 | 1662.0 |
| Lung | 8.0 | 3160.875 | 96.386777 | 3028.0 | 3106.00 | 3152.0 | 3204.50 | 3317.0 |
| Pancreas | 8.0 | 1029.875 | 114.558079 | 819.0 | 1021.00 | 1080.5 | 1093.25 | 1124.0 |
| Prostate | 8.0 | 5283.375 | 1294.028917 | 3268.0 | 4663.25 | 5383.0 | 6235.75 | 6921.0 |
| Rectum | 8.0 | 3017.625 | 518.899092 | 2128.0 | 2935.75 | 3292.0 | 3326.75 | 3361.0 |
| Stomach | 8.0 | 994.250 | 37.579440 | 948.0 | 966.75 | 989.0 | 1015.00 | 1059.0 |

Figure 6 shows the statewide summary statistics for number of cases by surgery type. The year 2015 has not been included to maintain consistency.

Through the analysis, it can be concluded that the volumes of 8 types of cancer surgeries increased after the coding change in 2015. These include breast, colon, rectum, bladder, pancreas, liver, brain, and esophagus. Both stomach and lung cancer surgery volumes slightly decreased after the coding change, apart from the sudden rise in 2019. Prostate cancer surgery volume showed a drastic increase for 2 years after the change and then fell significantly

It can also be concluded that breast cancer surgeries stand out as the most common, with an average of 27996, which is understandable as the data encompasses both outpatient and inpatient settings. Colon cancer surgeries and prostate cancer surgeries follow, with averages of 7354 and 5283 respectively. Lung cancer surgeries, rectum cancer surgeries, and brain cancer surgeries share similar averages around 3000 (3161 for lung, 3018 for rectum, and 3187 for brain). Liver cancer surgeries and pancreas cancer surgeries have averages of 1466 and 1030 respectively, while stomach cancer surgeries and bladder cancer surgeries share similar averages of 994 and 961 respectively. Esophagus cancer surgeries are the least common, with an average of 430. This analysis provides a clearer understanding of the varying frequencies of different cancer surgeries, highlighting breast, colon, and prostate surgeries as the most frequently performed procedures and esophagus surgeries as the least frequently performed one.

**Figure 7**

| Commonness ranking | Cancer surgery type | Average surgery performed per year |
|---|---|---|
| 1 | Breast | 27996 |
| 2 | Colon | 7354 |
| 3 | Prostate | 5283 |
| 4 | Brain | 3187 |
| 5 | Lung | 3161 |
| 6 | Rectum | 3018 |
| 7 | Liver | 1466 |
| 8 | Pancreas | 1030 |
| 9 | Stomach | 994 |
| 10 | Bladder | 961 |
| 11 | Esophagus | 430 |

Figure 7 shows the commonness ranking, the cancer surgery type and the average surgeries performed per year.

**Figure 8**

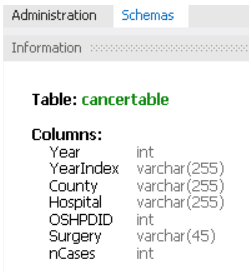



Figure 8 contains the schema of the table loaded into MySQL. The csv file was imported using the data import wizard and the corresponding columns of the dataset (csv file) and the table created were matched.

**Figure 9.a**

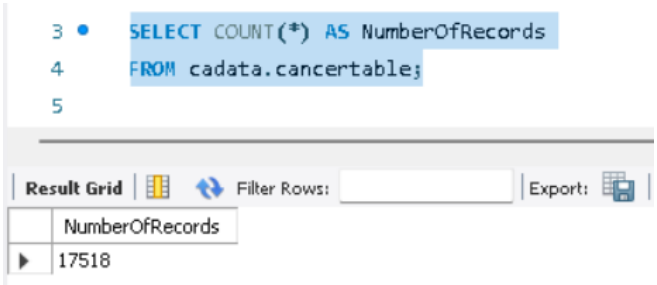



Figure 9.a shows the query and cross verification of total number of records imported; it matches with the total number of records of the dataset.

**Figure 9.b**

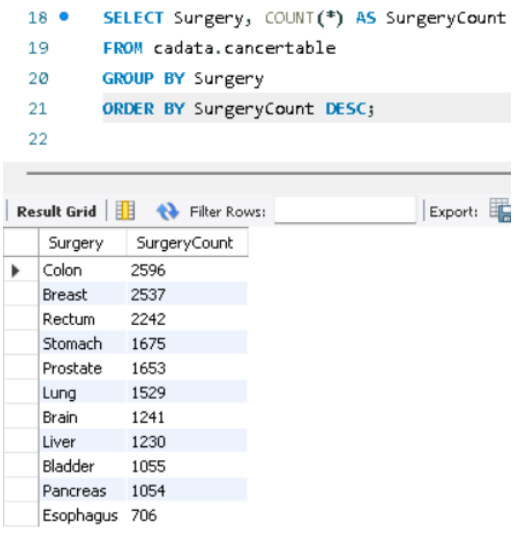


Figure 9.b displays both the query and output, counting the occurrences of each unique surgery type in the dataset. The results are presented in descending order based on the count. Colon has the highest occurrences, followed by Breast and Rectum, while Esophagus has the fewest occurrences.

The queries above (Figure 9.a and 9.b) are included to show how the querying works on the table or dataset.

**Figure 9.c**

```
SELECT Year, Surgery, County
FROM (
    SELECT
        Year, Surgery, County,
        ROW_NUMBER() OVER (PARTITION BY Year, Surgery ORDER BY nCases DESC) AS row_num
    FROM cadata.cancertable
) ranked
WHERE row_num <= 3;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| Year | Surgery | County |
|---|---|---|
| 2014 | Brain | San Francisco |
| 2014 | Brain | San Mateo |
| 2014 | Brain | Los Angeles |
| 2014 | Breast | Los Angeles |
| 2014 | Breast | Orange |
| 2014 | Breast | Santa Clara |
| 2014 | Colon | Los Angeles |
| 2014 | Colon | San Diego |
| 2014 | Colon | Orange |
| 2014 | Esopha... | Los Angeles |
| 2014 | Esopha... | Sacramento |
| 2014 | Esopha... | San Francisco |

Figure 9.c displays the SQL query output, revealing the top 3 counties for each type of surgery for each year. As the output comprises a total of 297 rows (Top 3 positions for 11 surgeries through 9 years), count plots were generated using Python after exporting the query result to a CSV file.

**Figure 10**

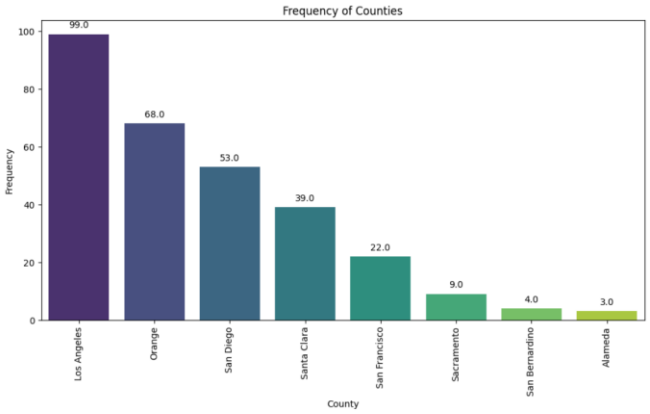



Figure 10 displays the frequency counts (Univariate analysis for county column – Nominal datatype) for the sql query result which is top counties with the highest number of surgeries performed. Out of the 297 total values, Los Angeles had the highest frequency of 99 (33.34%), followed by Orange County with 68 counts (22.89%), and then by San Diego County with 53 counts (17.84%).

**Figure 11**

```
Percentage of hospitals with at least one surgery for all 11 types throughout all the years: 3.25%

Total number of consistent hospitals: 13

The hospitals are as follows:-

Cedars Sinai Medical Center, City of Hope Helford Clinical Research Hospital

Hoag Memorial Hospital Presbyterian, Huntington Memorial Hospital

Kaiser Foundation Hospital - Orange County - Anaheim, Kaiser Foundation Hospital - Sunset

Keck Hospital of University of Southern California, Loma Linda University Medical Center

Ronald Reagan UCLA Medical Center, UC Davis Medical Center

UC Irvine Medical Center, UC San Diego Health System - Hillcrest Medical Center

UC San Francisco Medical Center
```

Figure 11 shows the percentage of hospitals that consistently performed at least one surgery for all 11 types throughout the years analysed, along with the list of hospital names. Only 3.25% (13 out of 395 hospitals analysed) meet this criterion.

## VI. POISSON REGRESSION

In this analysis, a Poisson regression model was developed in R Studio using R programming language to predict the number of cases based on ICD-10-CM coding per county for specific surgeries in each year. The dataset was filtered to include records from the year 2016 onwards, aligning with the effective start of coding transition. The 'County' and 'Surgery' variables were converted to factors, and the Poisson regression model was built on the filtered data. The initial model summary revealed significant coefficients for each county and surgery, indicating their impact on the expected number of cases. Subsequently, the model was trained on a subset of the data, and performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated for both the training and test sets.

**Figure 12**

```
set.seed(1234)
nobs <- nrow(df_filtered)
train <- sample(1:nobs, 2600)
test <- setdiff(1:nobs, train)

train_data <- df_filtered[train, ]
test_data <- df_filtered[test, ]
```

Figure 12 depicts the data split for training and testing sets. Approximately 80% (2600 out of 3267 records) of the dataset was used for model training, while the remaining portion was reserved for testing.

**Figure 13**

```
Call:
glm(formula = TotalCases ~ Year + County + Surgery, family = "poisson",
    data = df_filtered)
```

Figure 13 shows the Poisson regression model equation, the predictors used are the year, county and surgery.

**Figure 14**

```
SurgeryBrain          1.217011   0.014824  82.097  < 2e-16 ***
SurgeryBreast         3.377834   0.013241 255.097  < 2e-16 ***
SurgeryColon          2.018018   0.013859 145.606  < 2e-16 ***
SurgeryEsophagus     -0.763012   0.023091 -33.044  < 2e-16 ***
SurgeryLiver          0.452278   0.016656  27.155  < 2e-16 ***
SurgeryLung           1.163119   0.014918  77.970  < 2e-16 ***
SurgeryPancreas       0.103782   0.017955   5.780 7.46e-09 ***
SurgeryProstate       1.649258   0.014217 116.003  < 2e-16 ***
SurgeryRectum         1.209755   0.014836  81.539  < 2e-16 ***
SurgeryStomach        0.002709   0.018402   0.147    0.883
```

Figure 14 depicts the statistical significance of each surgery in the model.

In this analysis, it is noteworthy that while the majority of coefficients demonstrated high statistical significance, there were observed instances of comparatively lower significance. For instance, the coefficient associated with 'SurgeryStomach' yielded a p-value of 0.883, suggesting a lack of statistical significance in the impact of this specific surgery type on the expected number of cases.

**Figure 15**

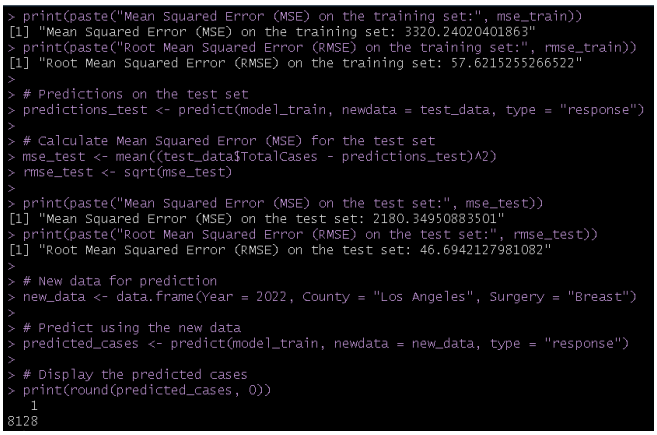



Figure 15 displays the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values for both the training and test sets. Additionally, it illustrates the predicted case counts for breast surgery in Los Angeles for the year 2022.

**Figure 16**

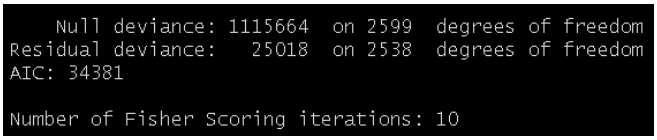



Figure 16 shows the summary of the Poisson regression model.

The model displayed reasonable predictive accuracy on both the training and test sets, with an MSE of 3320.24 and an RMSE of 57.62 for the training set, and an MSE of 2180.35 and an RMSE of 46.69 for the test set. Although the RMSE values indicate a certain level of effectiveness in capturing underlying patterns, they also suggest some degree of deviation between predictions and actual case counts. The AIC (Akaike Information Criterion) value, measuring the trade-off between model complexity and performance, was 34381, implying that the model's fit is acceptable. However, a lower AIC is typically desired, and this value suggests room for improvement. Despite these considerations, the model exhibited versatility by reasonably predicting 8128 cases for the year 2022 in Los Angeles County for Breast surgery, showcasing potential in making predictions for unseen data scenarios. The substantial difference between the null deviance (representing a model with no predictors) and the residual deviance indicates an ability to explain a noteworthy amount of variability in the data. In summary, while the Poisson regression model demonstrates reasonable performance, there is room for refinement to enhance its predictive capabilities.

## VII. DISCUSSION

The evolution of cancer surgeries, as explored in this study from 2013 to 2021, illuminates substantial shifts in reporting and coding practices, particularly with the transition from ICD-9-CM to ICD-10-CM/PCS. The investigation unveils nuanced trends in surgery volumes across diverse cancer types, employing a comprehensive methodology encompassing data querying, regression analysis, and visualizations with R, Python, and SQL. Noteworthy fluctuations in reported surgery volumes post-coding transition are observed, with a surge in seven specific surgery types, while 2020 records a dip attributed to the challenges posed by the COVID-19 pandemic. Geographic variations in surgery volumes reveal consistent leadership by Los Angeles, followed by San Diego and Orange County, aligning with population and healthcare infrastructure distribution. Remarkably, 13 hospitals demonstrate consistent performance across all 11 cancer types each year. While the Poisson regression model demonstrated reasonable performance in predicting surgery volumes for most types, it is important to note that it was not exceptional but rather reasonable, leaving room for improvement. This is particularly highlighted in the challenges encountered in stomach cancer predictions, emphasizing the need for nuanced interpretation. The frequency analysis underscores breast, colon, and prostate surgeries as the most common, with esophagus surgeries being the least frequent. Top counties with the highest surgery counts include Los Angeles, Orange County, and San Diego, while only 3.25% of hospitals analysed consistently perform surgeries for all types. The discussion encompasses the identified trends, implications for healthcare planning, and the model's effectiveness, emphasizing the multifaceted nature of cancer surgery dynamics and their relevance for resource allocation.

## VIII. FUTURE WORK

In contemplating future avenues of research, there are promising directions that can enhance our understanding of cancer surgery dynamics and contribute to improved healthcare planning. One potential avenue is the refinement and expansion of predictive models to forecast the number of surgeries performed by individual hospitals. Fine-tuning the existing Poisson regression model or exploring advanced machine learning techniques may provide more accurate predictions, accounting for intricate factors influencing surgery volumes. Additionally, delving into the socio-economic and demographic factors at a granular level can offer insights into variations in surgery rates across different populations. Collaborative efforts with healthcare institutions to gather real-time data and integrate it with historical records could further enhance the predictive capabilities of such models. Furthermore, investigating the long-term impacts of the COVID-19 pandemic on surgery volumes and patterns could be pivotal for healthcare resilience planning. This study's findings hold the potential to inform healthcare administrators, policymakers, and practitioners in optimizing resource allocation and strategic planning. By identifying trends and variations in surgery volumes, this research lays the foundation for targeted interventions and resource distribution, ultimately fostering a more efficient and responsive healthcare system.

## REFERENCES

[1] Laurence Baker & Maryann O'Sullivan. California Healthcare foundation (June 2017). Small Numbers Can Have Big Consequences. Accessed on: [10/24/2023]. Available at: https://www.chcf.org/wp-content/uploads/2017/12/PDF-SmallNumbersCancerSurgeries.pdf

[2] HCAI Number of Cancer Surgeries (Volume) Performed in California Hospitals, 2021. Volume of Cancer Surgeries Performed in California Hospitals. Accessed on: [10/23/2023]. Available at: https://hcai.ca.gov/visualizations/volume-cancer-surgery-reports/

[3] Maryann O'Sullivan. California Healthcare foundation (November 2015) Safety in Numbers: Cancer Surgeries in California Hospitals. Accessed on: [10/24/2023]. Available at: https://www.chcf.org/wp-content/uploads/2018/01/SafetyCancerSurgeriesHospitals.pdf

[4] Data.gov - Number of Cancer Surgeries (Volume) Performed in California Hospitals (August 12, 2023) - Department of Health Care Access and Information - State of California https://catalog.data.gov/dataset/number-of-cancer-surgeries-volume-performed-in-california-hospitals-a3f18

## APPENDIX

The appendix section contains few of the code snippets for visualizations, summary statistics and exploration done throughout the

- VISUALIZATIONS (Python code snippets)

I. Code snippet for figure 3.a - Total number of surgeries performed each year

```
total_surgeries_per_year = df.groupby('Year')['# of Cases'].sum().reset_index()

plt.figure(figsize=(10, 7))
plt.plot(total_surgeries_per_year['Year'], total_surgeries_per_year['# of Cases'], marker='o', color='Blue', linestyle='-')
plt.title('Total Number of Surgeries Performed Each Year')
plt.xlabel('Year')
plt.ylabel('Total Number of Surgeries')


for i, txt in enumerate(total_surgeries_per_year['# of Cases']):
    plt.annotate(txt, (total_surgeries_per_year['Year'].iloc[i], total_surgeries_per_year['# of Cases'].iloc[i] + 500))

plt.grid(axis='y', linestyle='-', alpha=0.7)
plt.show()
```

II. Code snippet for figure 3.b - Total number of cases per year index

```
total_cases_by_year_index = df.groupby('Year Index')['# of Cases'].sum().reset_index()


plt.figure(figsize=(8, 5))
bars = plt.bar(total_cases_by_year_index['Year Index'], total_cases_by_year_index['# of Cases'], color='blue')


for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval, round(yval, 1), ha='center', va='bottom')


plt.xlabel('Year Index')
plt.ylabel('Total # of Cases')
plt.title('Total Number of Cases by Year Index')


plt.show()
```

III. Code snippet for figure 4.a - Cancer Surgeries in California (2013-2021) (scatter plot)

```
# Defining a custom color palette with distinct colors
custom_palette = sb.color_palette(["#1f78b4", "#ff7f00", "#33a02c", "#e31a1c", "#6a3d9a", "#a6cee3", "#b2df8a", "#fb9a99", "#fdbf6f", "#cab2d6", "#ffff99"

plt.figure(figsize=(10, 6))


for i, surgery in enumerate(df['Surgery'].unique()):
    surgery_data = df[df['Surgery'] == surgery]
    plt.scatter(surgery_data['# of Cases'], surgery_data['Year'], label=surgery, color=custom_palette[i])

plt.xlabel('Number of Cases')
plt.ylabel('Year')
plt.title('Cancer Surgeries in California (2013-2021)')


plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))

plt.show()
```

IV. Code snippet for figure 5

The code below was edited and used for all the 11 types of surgery. (5.a – 5.k)

```
breast_data = df[df['Surgery'] == 'Breast']


plt.figure(figsize=(10, 6))
plt.plot(breast_data['Year'], breast_data['# of Cases'], marker='o', linestyle='-', color='g')


for i, txt in enumerate(breast_data['# of Cases']):
    plt.annotate(txt, (breast_data['Year'].iloc[i], breast_data['# of Cases'].iloc[i]), textcoords="offset points", xytext=(0,5), ha='center')


plt.xlabel('Year')
plt.ylabel('# of Cases')
plt.title('Breast Cancer Surgeries in California (2013-2021)')


plt.show()
```

V. Code snippet for figure 10 - Frequency of counties. The data frame 'df2' referred here is the result of the SQL query (figure 9.c) which was exported to a csv file.

```
county_freq = df2['County'].value_counts()


plt.figure(figsize=(12, 6))
bar_plot = sb.barplot(x=county_freq.index, y=county_freq.values, order=county_freq.index, palette='viridis')


plt.xlabel('County')
plt.ylabel('Frequency')
plt.title('Frequency of Counties')

plt.xticks(rotation=90)


for p in bar_plot.patches:
    bar_plot.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2., p.get_height()),
                      ha='center', va='center', xytext=(0, 10), textcoords='offset points')


plt.show()
```

- DATA EXPLORATIONS (Python code snippets)

VI. Code snippet for answering the 4th research question (Figure 11) which includes finding the percentage, name and count of hospitals which consistently did at least one surgery for each type throughout all the years. (Have included comments for easy interpretation)

```
# Grouping by Year, Hospital, and Surgery, then count the number of unique surgeries for each hospital in each year
hospital_surgeries_count = df1.groupby(['Year', 'Hospital', 'Surgery']).size().reset_index(name='Count')

# Filtering hospitals that have performed all 11 types of surgeries in each year
hospitals_with_all_surgeries = hospital_surgeries_count.groupby(['Year', 'Hospital']).filter(lambda x: x['Surgery'].nunique() == 11)

# Getting unique hospitals that have performed all 11 types of surgeries in each year
unique_hospitals_with_all_surgeries = hospitals_with_all_surgeries['Hospital'].unique()

# Filtering hospitals that have performed all 11 types of surgeries throughout all the years
consistent_hospitals = []
for hospital in unique_hospitals_with_all_surgeries:
    hospital_data = df1[df1['Hospital'] == hospital]
    if hospital_data['Year'].nunique() == (2021 - 2013 + 1):
        surgeries_per_year = hospital_data.groupby('Year')['Surgery'].nunique()colom
        if all(surgeries_per_year == 11):
            consistent_hospitals.append(hospital)

percentage_consistent_hospitals = (len(consistent_hospitals) / df1['Hospital'].nunique()) * 100

print(f"Percentage of hospitals with at least one surgery for all 11 types throughout all the years: {percentage_consistent_hospitals:.2f}%\n")
print(f"Total number of consistent hospitals: {len(consistent_hospitals)}")
print("\nThe hospitals are as follows:- \n")
for i in range(0, len(consistent_hospitals), 2):
    hospitals_line = consistent_hospitals[i:i+2]
    print(', '.join(hospitals_line))
    print()
```

VII. Code snippet for figure 6 – Summary statistics for number of cases grouped by each surgery

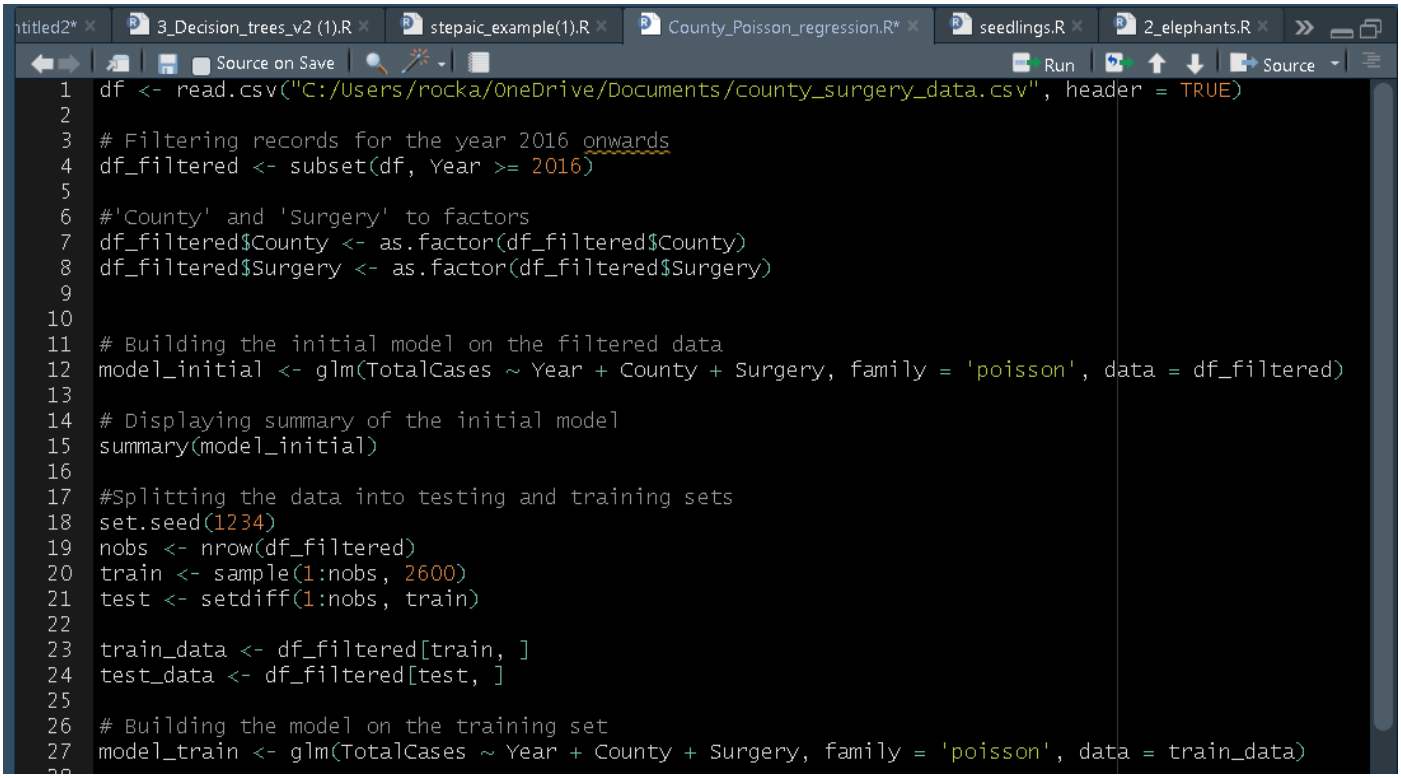
```
df_filtered = df[df['Year'] != 2015]

Statewide_statistics = df_filtered.groupby('Surgery')['# of Cases'].describe()

print(Statewide_statistics)
```

- SQL TABLE CREATION (MySQL workbench)

```
CREATE TABLE cancertable(
    Year INT,
    YearIndex VARCHAR(255),
    County VARCHAR(255),
    Hospital VARCHAR(255),
    OSHPDID INT,
    Surgery VARCHAR(255),
    nCases INT
);
```

- POISSON REGRESSION (R Programming in R Studio)

VIII. Code snippets for answering the 5th research question to predict the number of cases. (included comments for better readability)




```r
df <- read.csv("C:/Users/rocka/OneDrive/Documents/county_surgery_data.csv", header = TRUE)

# Filtering records for the year 2016 onwards
df_filtered <- subset(df, Year >= 2016)

#'County' and 'Surgery' to factors
df_filtered$County <- as.factor(df_filtered$County)
df_filtered$Surgery <- as.factor(df_filtered$Surgery)


# Building the initial model on the filtered data
model_initial <- glm(TotalCases ~ Year + County + Surgery, family = 'poisson', data = df_filtered)

# Displaying summary of the initial model
summary(model_initial)

#Splitting the data into testing and training sets
set.seed(1234)
nobs <- nrow(df_filtered)
train <- sample(1:nobs, 2600)
test <- setdiff(1:nobs, train)

train_data <- df_filtered[train, ]
test_data <- df_filtered[test, ]

# Building the model on the training set
model_train <- glm(TotalCases ~ Year + County + Surgery, family = 'poisson', data = train_data)
```

```r
model_train <- glm(TotalCases ~ Year + County + Surgery, family = "poisson", data = train_data)

#Displaying  Summary
summary(model_train)
# Predictions on the training set
predictions_train <- predict(model_train, newdata = train_data, type = "response")

# Calculate Mean Squared Error (MSE) for the training set
mse_train <- mean((train_data$TotalCases - predictions_train)^2)
rmse_train <- sqrt(mse_train)

print(paste("Mean Squared Error (MSE) on the training set:", mse_train))
print(paste("Root Mean Squared Error (RMSE) on the training set:", rmse_train))

# Predictions on the test set
predictions_test <- predict(model_train, newdata = test_data, type = "response")

# Calculate Mean Squared Error (MSE) for the test set
mse_test <- mean((test_data$TotalCases - predictions_test)^2)
rmse_test <- sqrt(mse_test)

print(paste("Mean Squared Error (MSE) on the test set:", mse_test))
print(paste("Root Mean Squared Error (RMSE) on the test set:", rmse_test))

# Testing new data for prediction by using an unknown year 2022
new_data <- data.frame(Year = 2022, County = "Los Angeles", Surgery = "Breast")

# Predict using the new data
predicted_cases <- predict(model_train, newdata = new_data, type = "response")

# Display the predicted cases
print(round(predicted_cases, 0))
```