FINAL PROJECT REPORT

Analytics: Big data to Information

California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery

Charan Teja Kurakula

George Mason University – Spring-2024

ckurakul@gmu.edu

Abstract— This dataset contains information and metrics that relate to CABG surgery conducted in California hospitals between 2018 and 2021 regarding locations of hospitals, number of patients, various performance metrics, and adverse events that make up a performance appraisal in consideration of case volumes. Techniques that were applied included measured performance. This dataset is of significant importance to researchers and healthcare professionals, allowing them critical information for the analysis of the performance of hospitals and the strategies of improving the quality of the cardiovascular surgical procedure. The dynamic state of the condition of CABG surgical outcomes is illuminated by this dataset from the different regions of California and changes over time.

Keywords— Coronary Artery Bypass Graft (CABG), hospital performance, risk-adjusted outcomes, healthcare quality, and California.

I.INTRODUCTION

The CABG (Coronary Artery Bypass Graft) is a very vital procedure that attempted to restore the blood flow to the heart in patients with the most advanced forms of coronary artery disease. Considerable variability in the success and quality of CABG surgeries depends on surgical expertise available, patient demographics, and other resources in hospitals. Various monitoring and analysis of surgical outcomes are definitely needed, which are indispensable in improving patient care and optimal healthcare delivery.

This paper analyses a complete CABG dataset carried out in hospitals in California from the year 2018 up until the year 2021. It contains data on vital interest for locations of hospitals in operative mortality rates, postoperative stroke incidences, and 30-day all-cause readmission rates. It is important to understand the differential outcomes in these metrics among hospitals. Identifying areas of improvement is critical to assuring consistent high-quality care.

In this study, it discusses whether CABG surgery is distributed geographically in California, whether to assess hospital performance in risk-adjusted outcomes, and to identify when and where the best outcomes are realized in the study period. This dataset can help researchers and health professionals add to insights into the efficacy of CABG surgery across regions of California, making evidence-based decisions on further improvements in the quality of care in cardiovascular treatment and patient safety. The study guides the targeted interventions to further optimize patient outcomes in California's health care system towards improving the quality of health care.

It provides CABG surgeries performed in the California hospitals for the years ranging from 2018 to 2021, columns containing the hospital name, county, performance rating, total cases, and adverse events, besides the performance rate. Geographic coordinates in the form of longitude and latitude are also available for data on location. This dataset provides a platform for the evaluation of hospital performance and outcome analysis, enabling the identification of quality disparities and the planning of intervention for improvisations in cardiovascular care. This data will enable stakeholders to analyze trends in the surgery performed, find areas of improvement, and provide treatment and patient care that is most effective in the healthcare system of California.

The "NOIR" framework categorizes data types based on their core attributes to improve data analysis and understanding. Nominal (N) data involves categorical values without a specific order, like colors. Ordinal (O) data represents categories with a defined sequence, such as performance ratings. Interval (I) data consists of numeric values where differences are meaningful but lack a true zero point, such as temperatures. Ratio (R) data types have a true zero point and allow for meaningful ratios, like measurements. Utilizing the NOIR framework assists in selecting appropriate analytical methods and comprehending data intricacies across diverse fields, enabling better decision-making based on data characteristics.

NOIR data types in our dataset:

Column Name	**Datatype**	**Description**
Year	Interval	According to the dataset, Year is Interval
County	Nominal	Name of the Counties
Hospital	Nominal	Name of the Hospitals that performs CABG Surgery.
Hospital ID	Ratio	Hospital ID is a number and it is Ratio.
Performance Measure	Nominal	Measures Performance Measure of Hospitals.
Total Statewide CABG Cases	Ratio	Count of total CABG Cases.
Total Statewide CABG Cases for Performance Measure Reporting	Ratio	Total cases for measuring performance of Hospitals.
Statewide Adverse Events	Ratio	Adverse Events across the state of California.
Statewide Adverse Events Rate (%)	Ratio	Percentage of Statewide Adverse Events Rate.
Total Hospital CABG Cases	Ratio	Number of CABG cases in Hospitals.
Total Hospital CABG Cases for Performance Measure Reporting	Ratio	CABG Cases for measuring Hospital Performances.
Hospital Adverse Events	Ratio	Adverse Events of Hospitals
Hospital Observed Adverse Events Rate (%)	Ratio	Adverse event rates of Hospitals in Percentage.
Hospital Expected Adverse Events Rate (%)	Ratio	Expected Adverse Rate of Hospitals.
Hospital Risk-Adjusted Adverse Events Rate (%,RAAER)	Ratio	Risk Adverse Event Rate of Hospitals.
Hospital Lower 95% CI for RAAER	Ratio	Confidence Interval Less than 95% of Adverse Events.
Hospital Upper 95% CI for RAAER	Ratio	Confidence Interval greater than 95% of Adverse Events.

Performance Rating	Ordinal	Rating of Hospitals based on their performances.
Longitude	Interval	Location of the County
Latitude	Interval	Location of the County.

Table 1: Data Types(NOIR).

II.RESEARCH QUESTIONS

The Main aim of this dataset is to examine the Coronary Artery Bypass Graft(CABG) Surgery. It provides the data about the Hospital performances between the years 2018 and 2021. The Research Questions which are related to the dataset are as follows:

1. What is the Correlation between different performance metrics at the hospitals?

2. Is there any correlation between Geographic location and Hospital performance for CABG Surgeries?

3. What is the state-wide variation in the rate of adverse events following coronary artery bypass grafting (CABG) surgeries between the various counties and hospitals?

4. Determine if there is any correlation between Total Hospital CABG Cases and Hospital Risk-Adjusted Adverse Events Rate (%,RAAER)?

5. Is there any correlation between Hospital Upper 95% CI for RAAER and Performance Rating?

These Research Questions can be answered through the visualizations between the columns of the dataset. Summaries of the columns of the dataset can also provide the vital information related to the Research Questions.

III. LITERATURE SURVEY

The "California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery" provides essential information regarding the outcomes and performance of CABG surgeries[1]. It enables geographical comparisons, adverse event rates, and even correlations between performance indicators—so, it supports the research objectives of performance ratings. The dataset and research touch upon CABG surgery at various levels and perspectives. CABG surgical technique improvements and the algorithmic controls for robotic surgical assistance are areas of interest for the research; the performance of hospitals in California is compared, and performance and outcomes of CABG surgery are evaluated using the dataset. They are complementary to each other and give insights into different facets of CABG functioning. In general, the second report relates much better to the topics of the requested study by providing relevant information to hospital performance in the context of CABG surgery, while the first report deals with surgical technique improvement at a technical level[2].

Comparison of Clinical and Administrative Data Sources for Hospital Coronary Artery Bypass Graft Surgery Report Cards" compares clinical and administrative data' sources for reliability and validity for the purpose of assessing hospital performance of coronary artery bypass graft surgery[3]. To examine the relationship between many hospital performance measures and the reliability of assessments that come from the sources of data, a study is conducted. The "California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery" report presents performance ratings for CABG surgery in California hospitals. Research offers deep insight into the hospital performance and surgical results regarding CABG surgery, which can be used as research on the themes of quality outcomes and quality of hospital performance in the procedure[1]. The two reports offer viable insight into the hospital performance and quality outcomes in CABG surgery, together with the peculiarities of research questions that dominate on the theme of hospital performance ratings, geographic variations, rates of adverse events, and the correlations with the peculiarities of performance metrics. The former, however, takes a stance of comparing two sources of data, while the latter focuses on the peculiarities of the performance ratings of the CABG operation in California.

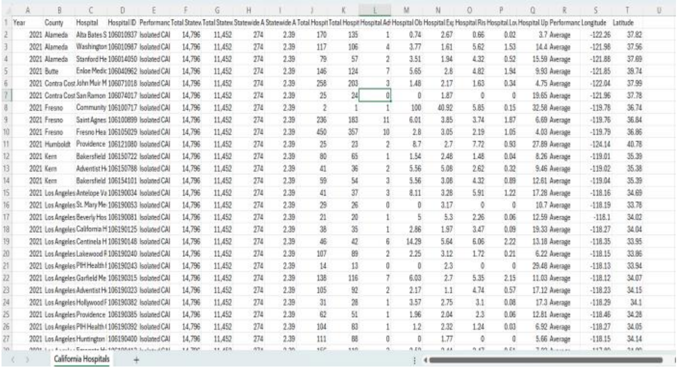
The paper "Coronary artery bypass graft surgery complications: A review for emergency clinicians" provides a comprehensive assessment of complications associated with CABG procedure to assist emergency physicians in quickly diagnosing and treating patients. While not directly addressing research issues about hospital performance ratings, it offers valuable insights into adverse events that follow CABG surgery, which is critical to understanding variations in hospital performance[4]. However, "California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery" explores quality outcomes specific to CABG surgery as well as hospital performance. It immediately aligns with the study objectives on hospital performance by facilitating evaluations of adverse event rates, regional disparities, and correlations between performance indicators through the provision of performance ratings[1].

IV. DATASET

The Dataset that I have selected is about 'Coronary Artery Bypass Graft Surgery' in California. The dataset provides deep insights into the Hospitals which perform CABG Surgeries in California. The data is collected in between the year 2018 to 2021. For better understanding about the dataset we need to analyse the data in the dataset. The dataset contains various types of data on the hospitals and also with the performance metrics that effect the Performance Rating of the hospitals. The Dataset almost contained every County mentioned and the Number of CABG Cases per County and Hospital along with their Co-ordinates(Latitude & Longitude). The dataset contains columns that mentioned each hospital's risk-adjusted adverse events rate; it is denoted by the hospital's lower and upper 95% confidence intervals for the rate. These columns give RAAER's 95% confidence interval's lower and upper bounds, respectively.

Fig.2 : Dataset

The relevant performance indicators are accompanied by the pertaining confidence intervals, namely, observed adverse event rates, predicted adverse event rates, and risk-adjusted adverse event rates. This dataset allows for comparing hospitals both within and between counties and for a fully-fledged analysis of hospital performance. It also makes it easier to find trends over time, outliers, and possible regional patterns in hospital performance. Using this dataset, we will be able to visualize and find solutions for the research questions so as to make sense of the data much better. Aside from the research questions, this dataset is helpful in many ways for the statistics about the CABG Surgery in California.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Year	County	Hospital	Hospital ID	Performanc	Total Statew	Total Statew	Statewide A	Statewide A	Total Hospit	Total Hospit	Hospital Ad	Hospital Ob	Hospital Ex	Hospital Ris	Hospital Lo	Hospital Up	Performanc	Longitude	Latitude	
2	2021	Alameda	Alta Bates S	106010937	Isolated CAI	14,796	11,452	274	2.39	170	135	1	0.74	2.67	0.66	0.02	3.7	Average	-122.26	37.82	
3	2021	Alameda	Washington	106010987	Isolated CAI	14,796	11,452	274	2.39	117	106	4	3.77	1.61	5.62	1.53	14.4	Average	-121.98	37.56	
4	2021	Alameda	Stanford He	106014050	Isolated CAI	14,796	11,452	274	2.39	79	57	2	3.51	1.94	4.32	0.52	15.59	Average	-121.88	37.69	
5	2021	Butte	Enloe Medic	106040962	Isolated CAI	14,796	11,452	274	2.39	146	124	7	5.65	2.8	4.82	1.94	9.93	Average	-121.85	39.74	
6	2021	Contra Cost	John Muir M	106071018	Isolated CAI	14,796	11,452	274	2.39	258	203	3	1.48	2.17	1.63	0.34	4.75	Average	-122.04	37.99	
7	2021	Contra Cost	San Ramon	106074017	Isolated CAI	14,796	11,452	274	2.39	25	24	0	0	1.87	0	0	19.65	Average	-121.96	37.78	
8	2021	Fresno	Community	106100717	Isolated CAI	14,796	11,452	274	2.39	2	1	1	100	40.92	5.85	0.15	32.58	Average	-119.78	36.74	
9	2021	Fresno	Saint Agnes	106100899	Isolated CAI	14,796	11,452	274	2.39	236	183	11	6.01	3.85	3.74	1.87	6.69	Average	-119.76	36.84	
10	2021	Fresno	Fresno Hea	106105029	Isolated CAI	14,796	11,452	274	2.39	450	357	10	2.8	3.05	2.19	1.05	4.03	Average	-119.79	36.86	
11	2021	Humboldt	Providence	106121080	Isolated CAI	14,796	11,452	274	2.39	25	23	2	8.7	2.7	7.72	0.93	27.89	Average	-124.14	40.78	
12	2021	Kern	Bakersfield	106150722	Isolated CAI	14,796	11,452	274	2.39	80	65	1	1.54	2.48	1.48	0.04	8.26	Average	-119.01	35.39	
13	2021	Kern	Adventist H	106150788	Isolated CAI	14,796	11,452	274	2.39	41	36	2	5.56	5.08	2.62	0.32	9.46	Average	-119.02	35.38	
14	2021	Kern	Bakersfield	106154101	Isolated CAI	14,796	11,452	274	2.39	59	54	3	5.56	3.08	4.32	0.89	12.61	Average	-119.04	35.39	
15	2021	Los Angeles	Antelope Va	106190034	Isolated CAI	14,796	11,452	274	2.39	41	37	3	8.11	3.28	5.91	1.22	17.28	Average	-118.16	34.69	
16	2021	Los Angeles	St. Mary Me	106190053	Isolated CAI	14,796	11,452	274	2.39	29	26	0	0	3.17	0	0	10.7	Average	-118.19	33.78	
17	2021	Los Angeles	Beverly Hos	106190081	Isolated CAI	14,796	11,452	274	2.39	21	20	1	5	5.3	2.26	0.06	12.59	Average	-118.1	34.02	
18	2021	Los Angeles	California H	106190125	Isolated CAI	14,796	11,452	274	2.39	38	35	1	2.86	1.97	3.47	0.09	19.33	Average	-118.27	34.04	
19	2021	Los Angeles	Centinela H	106190148	Isolated CAI	14,796	11,452	274	2.39	46	42	6	14.29	5.64	6.06	2.22	13.18	Average	-118.35	33.95	
20	2021	Los Angeles	Lakewood F	106190240	Isolated CAI	14,796	11,452	274	2.39	107	89	2	2.25	3.12	1.72	0.21	6.22	Average	-118.15	33.86	
21	2021	Los Angeles	PIH Health I	106190243	Isolated CAI	14,796	11,452	274	2.39	14	13	0	0	2.3	0	0	29.48	Average	-118.13	33.94	
22	2021	Los Angeles	Garfield Me	106190315	Isolated CAI	14,796	11,452	274	2.39	138	116	7	6.03	2.7	5.35	2.15	11.03	Average	-118.12	34.07	
23	2021	Los Angeles	Adventist H	106190323	Isolated CAI	14,796	11,452	274	2.39	105	92	2	2.17	1.1	4.74	0.57	17.12	Average	-118.23	34.15	
24	2021	Los Angeles	Hollywood F	106190382	Isolated CAI	14,796	11,452	274	2.39	31	28	1	3.57	2.75	3.1	0.08	17.3	Average	-118.29	34.1	
25	2021	Los Angeles	Providence	106190385	Isolated CAI	14,796	11,452	274	2.39	62	51	1	1.96	2.04	2.3	0.06	12.81	Average	-118.46	34.28	
26	2021	Los Angeles	PIH Health (	106190392	Isolated CAI	14,796	11,452	274	2.39	104	83	1	1.2	2.32	1.24	0.03	6.92	Average	-118.27	34.05	
27	2021	Los Angeles	Huntington	106190400	Isolated CAI	14,796	11,452	274	2.39	111	88	0	0	1.77	0	0	5.66	Average	-118.15	34.14	

V. VISUALIZATIONS

Visualizations play a major role in understanding the relationship between the metrics. We can use many types of Programming languages for visualizing the data but here we use the most common programming languages like Python, R. SQL is used for Manipulating, truncating, Deletion of data in the table. Apart from visualization I have also implemented the Amazon Web Services S3 and AWS Glue DataBrew for realizing the data in the dataset. The Research Questions asked in the previous sections may be answered here.

Python:

Python is a very popular language due to its ease of use and versatility. Python is used in many different sectors such as web development, data science, artificial intelligence, and automation, and it is also known for its readability and ease of use. Due to its huge collection of libraries and frameworks, worldwide developers use this language.

This code analyses and graphs the data that are extracted from a CSV file of hospitals in California and their performance in CABG surgeries using pandas, seaborn, and matplotlib. First, the script loads the CSV data into a panda DataFrame by the name "df," and the encoding issues are handled by the 'latin1' encoding parameter. Further, a scatter plot is created with the coordinates of the longitude and latitude extracted from the DataFrame, where each hospital's performance rating determines its color. This helps represent the possible correlation between the performance of hospitals in CABG surgery and their location in the state of California. Customizing the plot with a title and labels at the label makes for better visualization.

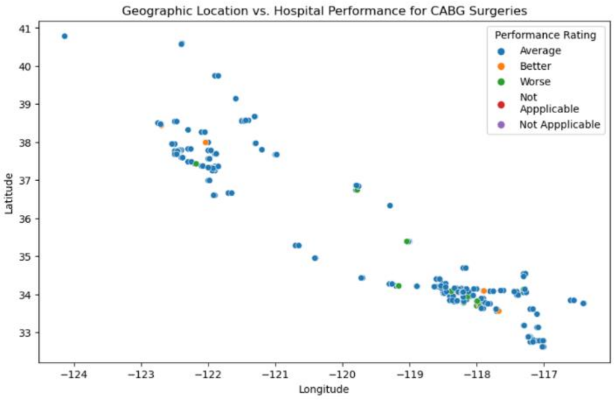


Fig.3: Scatter plot between Performance Rating and Location

**Research Question 2 can be answered from the scatter plot.

Analyzing data from California hospitals, the scatter plot tries to find whether there could be some kind of relation between hospital performance in CABG surgery and location. It shows the position of the hospitals based on latitude and longitude, with different colors representing the performance rating. The scatter plot is unable to find any possible relationship between the location and the hospital performance rating in the CABG surgical procedures. However, though there are some hospital groupings with similar performance rating in the same geographical areas, the general trend reveals that other factors contribute to the relationship between location and performance. In fact, lots of variation in performance is noted across every geographical area, which means that patient demographic, hospital policies, and available resources likely play a more substantial role in performance levels.

The aim of this study is to find whether there is a relationship between hospital performance in Coronary Artery Bypass Grafting (CABG) surgical procedures in California and location. This study is aimed at investigating the possible relationships between geographic distribution of the hospitals in different counties in California and the success rates in conducting CABG surgical procedures. This, therefore, will try to determine geographical factors that will have an impact on the results of the hospitals in cardiac surgical procedures.

The code documents the information of handling of the observed adverse event rate % in hospital settings. The information in the 'Hospital Observed Adverse Events Rate (%)' column was first converted into a numerical data type, and the found error during the process was transformed to the NaN values. All the rows of rows containing NaN values in the column 'Hospital Observed Adverse Events Rate (%)' were removed for keeping the validity of the dataset.

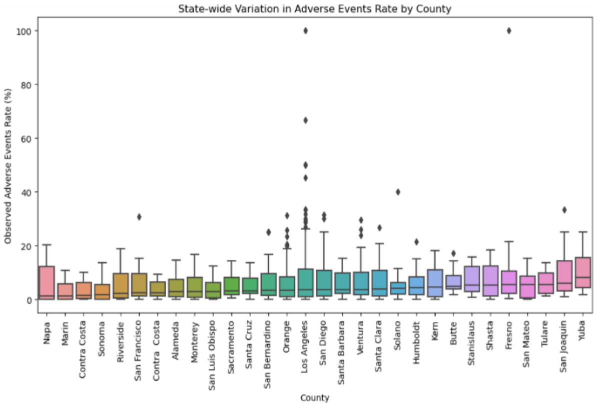


Fig.4: Box plot of County and Adverse Events Rate.

**Research Question 3 can be answered from the box plot.

A visualization was created for a county variation of negative events rates. On the x-axis, the county and on the y-axis, the observed adverse event rate %. In every box plot, the observed adverse event rate for the hospitals in a particular county is presented. To make easy-to-digest data, the counties are ordered based on the median observed adverse event occurrence rates. This box plot provides information about the variation in adverse event rates among the counties in California. It makes it easier to compare rates of adverse occurrences across counties and identify any outliers by providing data on the range, median, and dispersion of these rates. Additionally, it helps identify patterns or trends in the frequencies of adverse occurrences in different parts of the state.

The Next Visualization focuses on the Correlation between performance metrics and the Counties and Count of Hospitals.

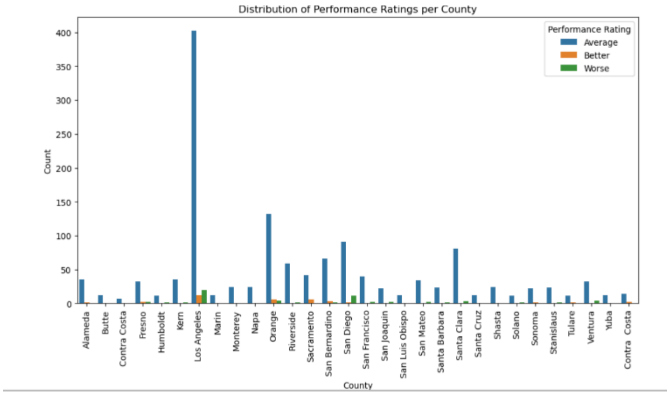


Fig.5: Representation of Count and Counties based on performance rating.

**Research Question 1 can be answered from the bar graph.

This analysis will attempt to make out the distribution of hospital performance ratings across the various counties of California. It will show, through the use of seaborn and matplotlib, the preponderance of Hospitals around the region, differently rated.

In this graph, one can view how each county in California spreads out on the performance ratings, with a color-coded bar for each rating performance level. By just looking at the height of the bars and the color patterns, insights may be gained into which ratings prevail at what proportion in each area.

R

R will been used very frequently in statistical analysis and data handling. There are numerous R packages available for meeting any type of analytical requirement. An active community guarantees collaboration and code sharing. Primarily, R is a key tool in data science due to its flexibility in syntax for statistical modeling and exploration.

The purpose of this analysis is to explore whether there is a potential correlation between the total volume of Coronary Artery Bypass Grafting (CABG) procedures conducted by hospitals and the Hospital Risk-Adjusted Adverse Events Rate (%RAAER) associated with those procedures. Our objective is to visually represent these variables to determine if there is any connection between the quantity of CABG surgeries performed by hospitals and the rate of adverse events adjusted for risk.

R code forms the density map of the Coronary Artery Bypass Grafting in the hospitals by using the ggplot2 library. This will convert the column 'Total.Hospital.CABG.Cases'(In R Spaces are read as .) to a numeric format after importing the data set from a CSV file. This is then followed by the building of data frame, df, and mappings with a dark and colorful appearance using the ggplot function. That is the x-axis reflects the total number of CABG cases of a hospital. The graphed visualization features a layer of the density plot that holds 50% opacity and a light blue color using the geom_density function. Furthermore, the labs function labels the x- and y-axes as "Density" and "Total Hospital CABG Cases" respectively, giving the plot and its axes titles. Last but not least, xlim limits the x-axis range from 0 to 1000 for better visualization.

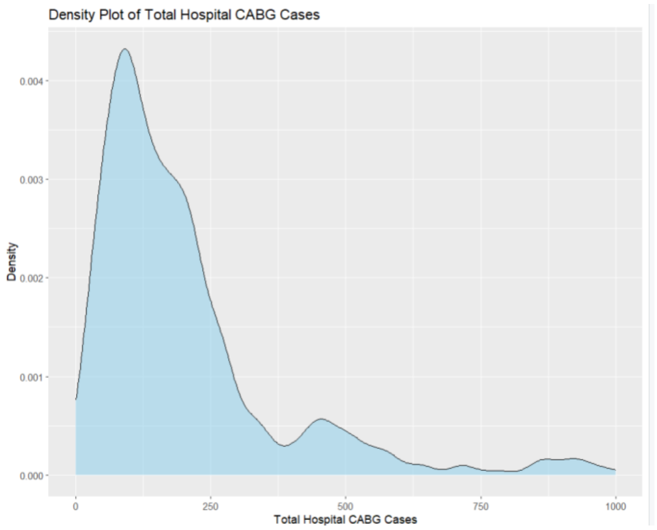


Fig.6: Density plot between Total CABG Cases.

**Research Question 4 can be analyzed from the density plot.

The density plot depicts the distribution of Coronary Artery Bypass Grafting (CABG) cases among hospitals. Its curve illustrates the probability of observing different levels of total CABG instances as a smooth, continuous curve. It contributes to our understanding of the shape and spread of the distribution, so we can tell if hospitals are likely to have low, moderate, or high numbers of CABG patients. A 50% opacity sky blue fill color improves the visibility without loss of clarity. This density plot, therefore, provides clarity to the different spreads of the total number of CABG cases among institutions and shows the general trend of CABG in the sample.e

This study aims to find a possible relationship between performance ratings and the risk-adjusted adverse events rate (RAAER). It investigates the correlation between the two variables by using the Upper 95% Confidence Interval (CI). It aims to find out whether hospitals with higher Upper 95% CI values for RAAER have different performance ratings compared to those with lower CI values. This study tries to bring light to the sense of how the performance rating of hospitals relates to the precision of RAAER estimates.

The R code is written to display a hexbin plot using the ggplot2 package, which in definition is a scatter plot where data points get placed into hexagonal bins. The hexbin plot specifically focuses on the performance rating y-axis against the Hospital Upper 95% Confidence Interval for Risk-Adjusted Adverse Events Rate on the x-axis. Each hexagonal bin is colored white for clarity for data points that have similar values. Broadly speaking, it is a tight graphical summary of the relationship between Performance Rating and Hospital Upper 95% CI RAAER.One type of two-dimensional histogram is the hexbin plot which can be very useful in showing the relationship between two continuous variables. A hexbin plot calculates the number of data points that are assigned to each of the hexagon-shaped bins that are created within the plot region. The color intensity or shading of each bin-darker hues meaning higher densities-indicates the density of data points within that region. In general, graphs are especially useful when viewing patterns and trends within a dense distribution of data.

Output:

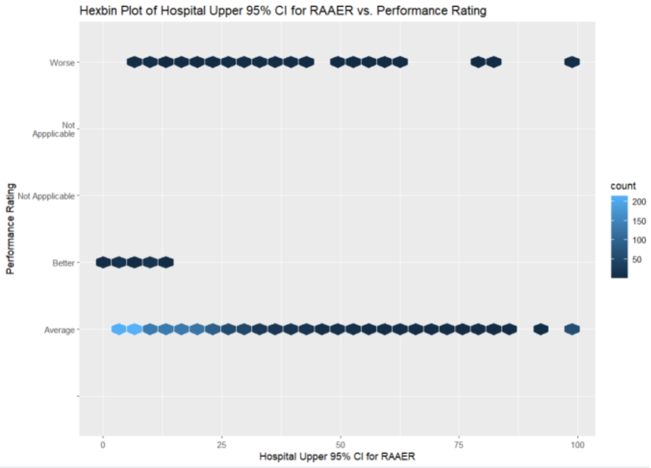


Fig.7: Hexbin between Performance Rating and CI.

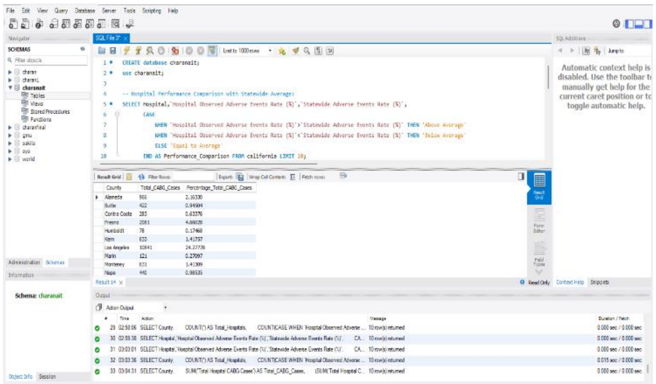
**Research Question 5 can be analyzed from the above graph.

VI.SQL

Structured Query Language, or SQL, is a computer language designed specifically to work with the relational database environment. SQL provides users with the ability to view, alter, and manage information stored in databases. With its versatility and ease of use, SQL is used in almost every industry for managing databases, producing reports, and analyzing data.

Using SQL we can Manipulate the data and we can also derive the summaries or any calculations we need.

Code:



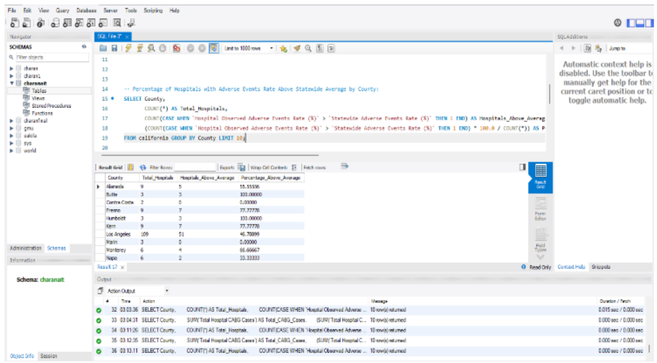
```sql
CREATE database charanait;
use charanait;

-- Hospital Performance Comparison with Statewide Average:
SELECT Hospital,`Hospital Observed Adverse Events Rate (%)`,`Statewide Adverse Events Rate (%)`,
        CASE
            WHEN `Hospital Observed Adverse Events Rate (%)`>`Statewide Adverse Events Rate (%)` THEN 'Above Average'
            WHEN `Hospital Observed Adverse Events Rate (%)`<`Statewide Adverse Events Rate (%)` THEN 'Below Average'
            ELSE 'Equal to Average'
        END AS Performance_Comparison FROM california LIMIT 10;
```

County	Total_CABG_Cases	Percentage_Total_CABG_Cases
Alameda	966	2.16330
Butte	422	0.94504
Contra Costa	283	0.63376
Fresno	2081	4.66028
Humboldt	78	0.17468
Kern	633	1.41757
Los Angeles	10841	24.27778
Marin	121	0.27097
Monterey	631	1.41309
Napa	440	0.98535

```
#   Time      Action                                                                                          Message              Duration / Fetch
29  02:58:06  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.000 sec / 0.000 sec
30  02:59:38  SELECT Hospital,`Hospital Observed Adverse Events Rate (%)`,`Statewide Adverse Events Rate (%)`,   CA...  10 row(s) returned   0.000 sec / 0.000 sec
31  03:03:01  SELECT Hospital,`Hospital Observed Adverse Events Rate (%)`,`Statewide Adverse Events Rate (%)`,   CA...  10 row(s) returned   0.000 sec / 0.000 sec
32  03:03:36  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.015 sec / 0.000 sec
33  03:04:31  SELECT County,   SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,   (SUM(`Total Hospital C...  10 row(s) returned   0.000 sec / 0.000 sec
```

The provided SQL code first creates the "charanait" database and then retrieves data from the "california" table. It gets the information regarding hospitals, that is the names and rates of adverse events and compares each hospital's observed rate with the average for the state. Hospitals with rates above average are marked above the average and the ones below average are marked below the average, whereas the ones equal to average are marked accordingly. This comparison is shown in the output for the first 10 records.

Code:



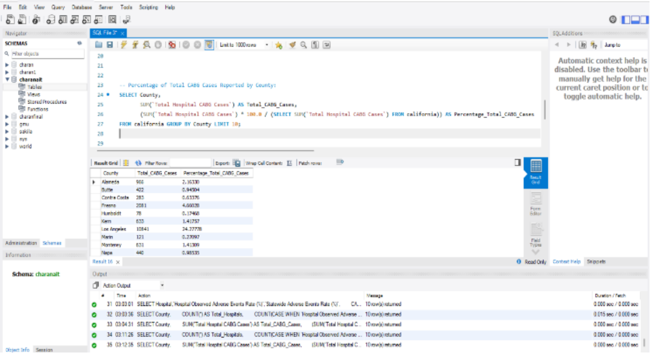
```sql
-- Percentage of Hospitals with Adverse Events Rate Above Statewide Average by County:
SELECT County,
        COUNT(*) AS Total_Hospitals,
        COUNT(CASE WHEN `Hospital Observed Adverse Events Rate (%)` > `Statewide Adverse Events Rate (%)` THEN 1 END) AS Hospitals_Above_Averag
        (COUNT(CASE WHEN `Hospital Observed Adverse Events Rate (%)` > `Statewide Adverse Events Rate (%)` THEN 1 END) * 100.0 / COUNT(*)) AS P
FROM california GROUP BY County LIMIT 10;
```

County	Total_Hospitals	Hospitals_Above_Average	Percentage_Above_Average
Alameda	9	5	55.55556
Butte	3	3	100.00000
Contra Costa	2	0	0.00000
Fresno	9	7	77.77778
Humboldt	3	3	100.00000
Kern	9	7	77.77778
Los Angeles	109	51	46.78899
Marin	3	0	0.00000
Monterey	6	4	66.66667
Napa	6	2	33.33333

```
#   Time      Action                                                                                          Message              Duration / Fetch
32  03:03:36  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.015 sec / 0.000 sec
33  03:04:31  SELECT County,   SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,   (SUM(`Total Hospital C...  10 row(s) returned   0.000 sec / 0.000 sec
34  03:11:26  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.000 sec / 0.000 sec
35  03:12:35  SELECT County,   SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,   (SUM(`Total Hospital C...  10 row(s) returned   0.000 sec / 0.000 sec
36  03:13:11  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.000 sec / 0.000 sec
```

This SQL query creates a percentage of hospitals in each county where the observed adverse event rate exceeds the statewide average. It gets data from the "california" dataset by counting the totals of the hospitals and the ones with the adverse event rates above the state average through the group of counting by the county. Then, it computes the percentage of hospitals above the average in every county. The output reveals the name of a county, the total hospitals in the county, the number of hospitals that are above the average, and the corresponding percentage, truncated within the first ten records.

Code:



```sql
-- Percentage of Total CABG Cases Reported by County:
SELECT County,
        SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,
        (SUM(`Total Hospital CABG Cases`) * 100.0 / (SELECT SUM(`Total Hospital CABG Cases`) FROM california)) AS Percentage_Total_CABG_Cases
FROM california GROUP BY County LIMIT 10;
```

County	Total_CABG_Cases	Percentage_Total_CABG_Cases
Alameda	966	2.16330
Butte	422	0.94504
Contra Costa	283	0.63376
Fresno	2081	4.66028
Humboldt	78	0.17468
Kern	633	1.41757
Los Angeles	10841	24.27778
Marin	121	0.27097
Monterey	631	1.41309
Napa	440	0.98535

```
#   Time      Action                                                                                          Message              Duration / Fetch
31  03:03:01  SELECT Hospital,`Hospital Observed Adverse Events Rate (%)`,`Statewide Adverse Events Rate (%)`,   CA...  10 row(s) returned   0.000 sec / 0.000 sec
32  03:03:36  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.015 sec / 0.000 sec
33  03:04:31  SELECT County,   SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,   (SUM(`Total Hospital C...  10 row(s) returned   0.000 sec / 0.000 sec
34  03:11:26  SELECT County,   COUNT(*) AS Total_Hospitals,   COUNT(CASE WHEN `Hospital Observed Adverse ...  10 row(s) returned   0.000 sec / 0.000 sec
35  03:12:35  SELECT County,   SUM(`Total Hospital CABG Cases`) AS Total_CABG_Cases,   (SUM(`Total Hospital C...  10 row(s) returned   0.000 sec / 0.000 sec
```

This, in turn, calculates the percentage of Coronary Artery Bypass Grafting cases reported for each county. It returns the name of the county, the number of CABG cases, and percentage for the first ten entries. It aggregates the CABG cases by county and calculates the percentage of the number of CABG cases relative to the total CABG cases in the dataset.

The ways these sorts of hospital performance metrics vary across the California counties, it does show that there is a significant difference in the degree of hospital performance. It is known to compare the hospital-specific rates of adverse events to the state average. The comparison revealed the performance diversities. Additionally, pointing out the hospitals whose adverse event rates are higher than the state average, the remaining concerns are followed. In addition, the analysis of the distribution of total CABG cases by county also revealed the healthcare utilization disparities. Resulting in such analyses, there lies a constant need to monitor and target intervention to improve the statewide outcomes within healthcare.

VII.AMAZON WEB SERVICES

Later in 2006, Amazon.com announced the launch of its cloud computing platform, now named Amazon Web Services. AWS has made available a broad variety of services to its customers, ranging from databases to machine learning storage, and processing power. Its global network of data centers is guaranteed to give accessibility, scalability, and reliability, with low downtimes and redundancy of data within backups.Combined with pay-as-you-go pricing, it meant businesses paid for only what they used, hence making it very cost-effective. Variety of services, along with its cost and scalability, made it de-facto standard for cloud computing solutions across the world.

The cloud storage solution that became very popular was Amazon S3. It is scalable, durable, and secure, with redundant storage of data in multiple facilities that make it very resilient. It has a bucket-based architecture, thereby providing flexible storage and retrieval of data over the internet. S3 delivers several storage classes that handle the various needs, and its deep integration with other AWS services simplifies the data processing and transfer workflows. AWS Glue Data Brew is, on the other hand, a managed data preparation service that makes data cleaning and transformation intuitive by using visual tools and pre-built transformations to increase the efficiencies of overall data preparation.

Steps involving in creating a bucket in S3 Bucket:

Step 1: Start the AWS Learner Lab.

Step 2: Open your AWS Account and open S3.

Step 3: Click on Create Bucket.

Step 4: Give a unique Bucket name because Bucket name should be unique and should not contain any spaces and special characters.

Step 5: Upload the dataset you need to visualize.

Step 6: Open Glue DataBrew.

Step 7: Open Dataset and give a name to dataset.

Step 8: Enter your source as the S3 bucket you created.

Step 9: Enter delimeter and select the role.

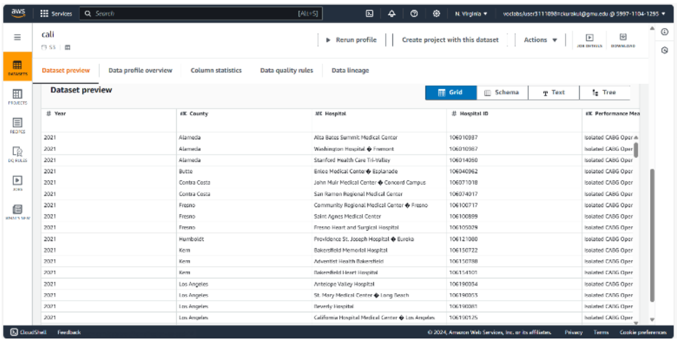
Step 10: Click on Create dataset.

Step 11: Create Project.

Step 12: Select the source of the dataset. Select your S3 bucket .

Step 13: Select the dataset and give delimeter and Role and Click on Create project.

Step 14: Click on Run job profile and the visualizations will be presented.

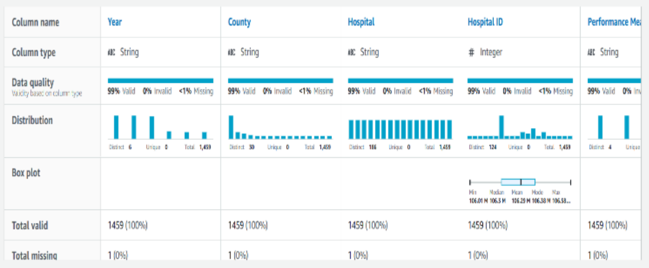


Dataset preview

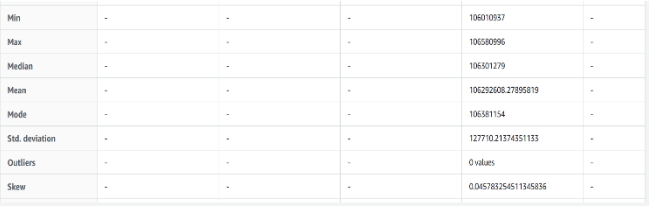
Year	County	Hospital	Hospital ID	Performance Mea
2021	Alameda	Alta Bates Summit Medical Center	106010937	Isolated CABG Oper
2021	Alameda	Washington Hospital � Fremont	106010987	Isolated CABG Oper
2021	Alameda	Stanford Health Care Tri-Valley	106014050	Isolated CABG Oper
2021	Butte	Enloe Medical Center� Esplanade	106040962	Isolated CABG Oper
2021	Contra Costa	John Muir Medical Center � Concord Campus	106071018	Isolated CABG Oper
2021	Contra Costa	San Ramon Regional Medical Center	106074017	Isolated CABG Oper
2021	Fresno	Community Regional Medical Center � Fresno	106100717	Isolated CABG Oper
2021	Fresno	Saint Agnes Medical Center	106100899	Isolated CABG Oper
2021	Fresno	Fresno Heart and Surgical Hospital	106105029	Isolated CABG Oper
2021	Humboldt	Providence St. Joseph Hospital � Eureka	106121080	Isolated CABG Oper
2021	Kern	Bakersfield Memorial Hospital	106150722	Isolated CABG Oper
2021	Kern	Adventist Health Bakersfield	106150788	Isolated CABG Oper
2021	Kern	Bakersfield Heart Hospital	106154101	Isolated CABG Oper
2021	Los Angeles	Antelope Valley Hospital	106190034	Isolated CABG Oper
2021	Los Angeles	St. Mary Medical Center � Long Beach	106190053	Isolated CABG Oper
2021	Los Angeles	Beverly Hospital	106190081	Isolated CABG Oper
2021	Los Angeles	California Hospital Medical Center � Los Angeles	106190125	Isolated CABG Oper

© 2024, Amazon Web Services, Inc. or its affiliates.

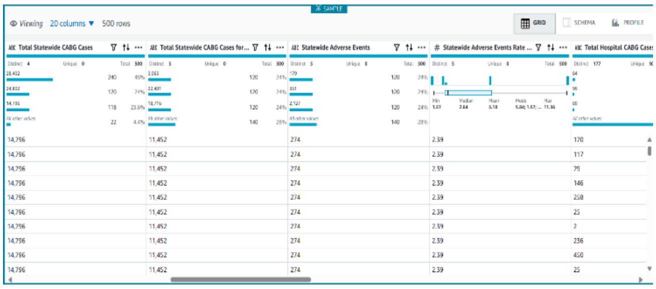
Fig.8: Dataset Preview.



Column name	Year	County	Hospital	Hospital ID	Performance Me
Column type	String	String	String	Integer	String
Data quality Validity based on column type	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% In
Distribution	Distinct 6 Unique 0 Total 1,459	Distinct 50 Unique 0 Total 1,459	Distinct 186 Unique 0 Total 1,459	Distinct 124 Unique 0 Total 1,459	Distinct 4 Unique
Box plot					
Total valid	1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)
Total missing	1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)



Min	-	-	-	106010937	-
Max	-	-	-	106580996	-
Median	-	-	-	106301279	-
Mean	-	-	-	106292608.27895819	-
Mode	-	-	-	106381154	-
Std. deviation	-	-	-	127710.21374351133	-
Outliers	-	-	-	0 values	-
Skew	-	-	-	0.045783254511345836	-



Viewing 20 columns ▼ 500 rows

Total Statewide CABG Cases: Distinct 4 Unique 0 Total 500; 28,402 240 48%; 24,802 120 24%; 14,796 118 23.6%; All other values 22 4.4%

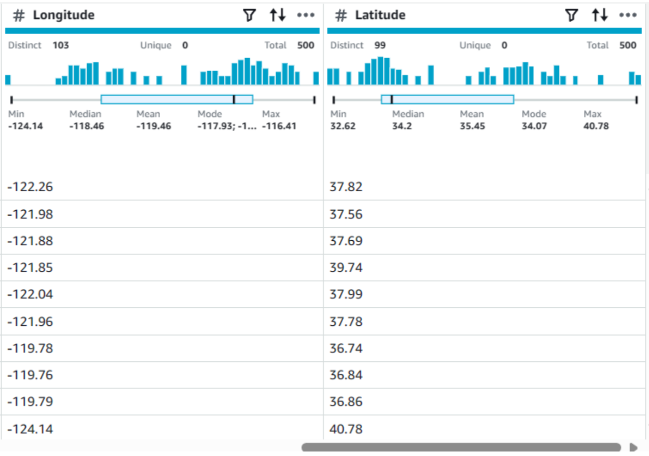
Total Statewide CABG Cases for...: Distinct 5 Unique 0 Total 500; 3,063 120 24%; 22,401 120 24%; 18,716 120 24%; All other values 140 28%

Statewide Adverse Events: Distinct 5 Unique 0 Total 500; 179 120 24%; 551 120 24%; 2,127 120 24%; All other values 140 28%

Statewide Adverse Events Rate ...: Distinct 5 Unique 0 Total 500; Min 1.57 Median 2.64 Mean 5.18 Mode 5.84; 1.57; ... Max 11.56

Total Hospital CABG Cases: Distinct 177 Unique 50; 84; 99; 55; All other values

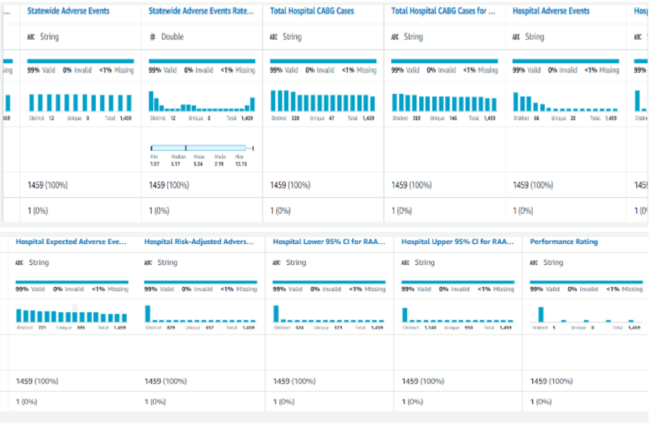
Total Statewide CABG Cases	Total Statewide CABG Cases for...	Statewide Adverse Events	Statewide Adverse Events Rate ...	Total Hospital CABG Cases
14,796	11,452	274	2.39	170
14,796	11,452	274	2.39	117
14,796	11,452	274	2.39	79
14,796	11,452	274	2.39	146
14,796	11,452	274	2.39	258
14,796	11,452	274	2.39	25
14,796	11,452	274	2.39	2
14,796	11,452	274	2.39	236
14,796	11,452	274	2.39	450
14,796	11,452	274	2.39	25



Longitude: Distinct 103 Unique 0 Total 500; Min -124.14 Median -118.46 Mean -119.46 Mode -117.93; -1... Max -116.41

Latitude: Distinct 99 Unique 0 Total 500; Min 32.62 Median 34.2 Mean 35.45 Mode 34.07 Max 40.78

Longitude	Latitude
-122.26	37.82
-121.98	37.56
-121.88	37.69
-121.85	39.74
-122.04	37.99
-121.96	37.78
-119.78	36.74
-119.76	36.84
-119.79	36.86
-124.14	40.78



...	Statewide Adverse Events	Statewide Adverse Events Rate...	Total Hospital CABG Cases	Total Hospital CABG Cases for ...	Hospital Adverse Events	Hos
	ABC String	# Double	ABC String	ABC String	ABC String	ABC
ing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99%
459	Distinct 12 Unique 0 Total 1,459	Distinct 12 Unique 0 Total 1,459	Distinct 328 Unique 47 Total 1,459	Distinct 369 Unique 146 Total 1,459	Distinct 66 Unique 28 Total 1,459	Distin
		Min 1.57 Median 5.17 Mean 5.34 Mode 2.19 Max 12.15				
	1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)	145
	1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0

Hospital Expected Adverse Eve...	Hospital Risk-Adjusted Advers...	Hospital Lower 95% CI for RAA...	Hospital Upper 95% CI for RAA...	Performance Rating
ABC String	ABC String	ABC String	ABC String	ABC String
99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing	99% Valid 0% Invalid <1% Missing
Distinct 721 Unique 395 Total 1,459	Distinct 829 Unique 557 Total 1,459	Distinct 524 Unique 321 Total 1,459	Distinct 1,140 Unique 938 Total 1,459	Distinct 5 Unique 0 Total 1,459
1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)	1459 (100%)
1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)

The preview shows the Mean, Median, Min,Max values of the columns in the dataset.

It also shows the Distinct and Unique values of the columns in the dataset. It shows the unique values, float values in the columns of the dataset.

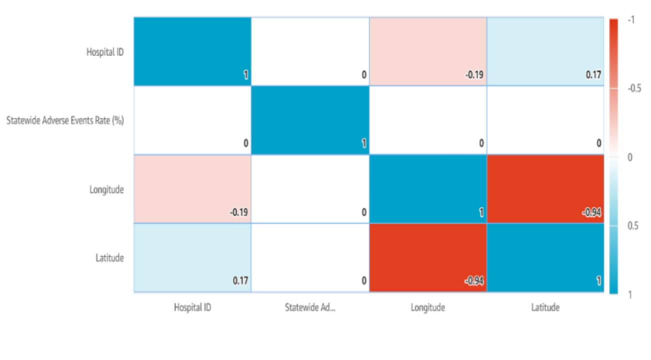


Fig.9: Correlation Matrix.

The primary determinant of correlation is the correlation coefficient, which spans from -1 to 1.

A coefficient of zero denotes no correlation between variables.

It gauges the intensity and orientation of the relationship between variables. The correlation matrix presented above pertains to numerical values within the dataset.

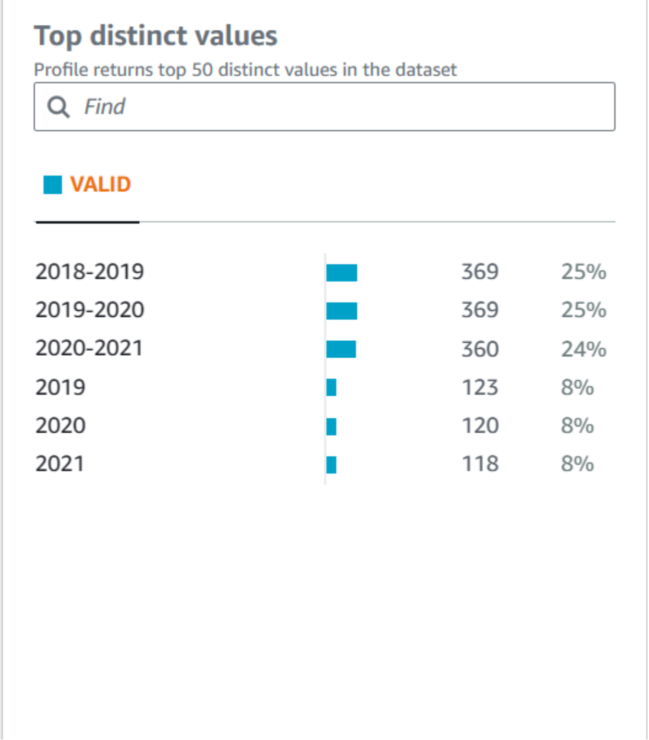
The correlation coefficient defines two important things:

1.Strength of relationship.

2. Direction of correlation.

A strength coefficient is about 1 or −1, which can indicate a strong positive or negative linear relationship. Otherwise, it is 0, meaning weak or no linear relationship.

Direction means that they vary together and in opposite directions, with coefficients of +1 or -1, which means perfect correlation.

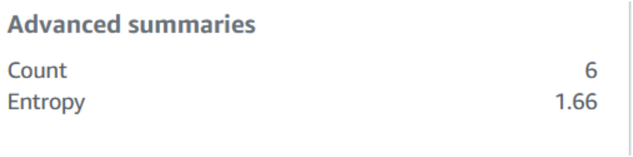


Top distinct values

Profile returns top 50 distinct values in the dataset

2018-2019	369	25%
2019-2020	369	25%
2020-2021	360	24%
2019	123	8%
2020	120	8%
2021	118	8%

Fig.10: Top Distinct values.



Advanced summaries

Count	6
Entropy	1.66

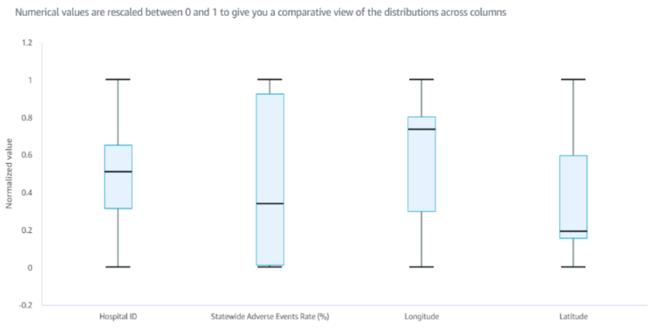


Fig.11: Box Plot.

A box plot is an almost graphical statistical way of showing the distribution of a given set of numerical data. The box in the middle of the plot shows the middle 50% of the data, with the line in the middle showing the median. The lines extending from the box, called whiskers, show the rest of the data, up to a certain point.

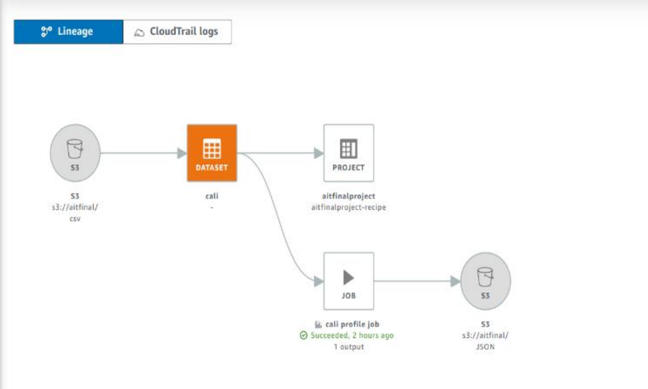


Fig.12: Methodology.

The steps involved in analyzing the dataset using S3 bucket and Glue DataBrew is clearly explained in Fig.12.

This implies that Amazon S3 and AWS Glue DataBrew are important tools in the Amazon Web Services environment and offer robust solutions in data storage, processing, and preparation. S3 provides security, scalability, and durability of high-volume datasets with customizable storage options so that every case will fit a diversity of needs. On the other hand, Glue DataBrew offers the possibility to generate data in a simple way with the end-to-end transformation and in an easy-to-use environment. Both these services empower organizations to efficiently manage, process, and analyze data and make decisions based on that information. Security, scalability, and user-friendliness are the reasons AWS is considered one of the leading providers of cloud computing solutions.

IX.FUTURE SCOPE

As the project progresses, some machine learning and advanced analytics could provide predictive insights on the rates of adverse events and the patterns of performance at the hospital. This could enable risk management and decision-making about the issue proactively. Better real-time data streaming capabilities may provide a continuous monitoring of hospital metrics so as to react quickly to emerging issues. Textual data analysis about patient feedback and medical records could be automated to analyze deeper insights into causes of adverse events through natural language processing. With better data visualization tools in the hands of users, it would become more accessible to explore complex healthcare data. Such project extension should consider dynamic legal frameworks, such as HIPAA—Health Insurance Portability and Accountability Act—that will ensure that data security and privacy are maintained.

X. LIMITATIONS

Several constraints affecting the project's outcomes and breadth were encountered during the project's execution. Poor quality of reporting from hospitals and health care facilities, in the form of low reliability and incompleteness, would have influenced the quality of data as the basis for analyses. Data analysis focused on the analysis of California-specific data, thus limiting the extent to which findings could be generalized to other areas with diverse demographics, regulations, or healthcare systems. History data would not be able to be extended to recent advances or changes in hospital performance metrics; thus, its elaboration, updating with new real-time data, would be necessary to derive more accurate analysis. Finally, limitations in time, budget, and computational capacity would make sure that depth or complexity in the analysis could not be obtained for subtle patterns or connections between data.

XI. CONCLUSION

Healthcare data analysis and visualisation The analysis of healthcare data would not be possible without data analysis tools and libraries in R and Python. The RStudio is used for this task.As RStudio is loaded to connect to the dataset in R, Python is used to program the visualizations using Python libraries like seaborn and pandas. Apart from the selection and filtration of data from a large data set, the case of Python selection and filtration of data from a large data set. In the seaborn visualization, cases involving CABG operation with performance rating data were explored. There were a total of 7,059 cases observed, with no CABG operations conducted in the state. We have used AWS S3 Bucket and Glue DataBrew for analyzing the dataset in Amazon Web Services.

REFERENCES

[1] California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery - Catalog. (2023, September15).
Retrieved from https://catalog.data.gov/dataset/california-hospital-performance ratings-for-coronary artery-bypass-graft-cabg-surgery-b9e65.
Publisher: Department of Health Care Access and Information
Maintainer: Department of Health Care Access and Information, Information Services Division, Healthcare Analytics Branch, Quality and Performance Section

[2] Bebek, O., & Cavusoglu, M. C. (2007). Intelligent control algorithms for robotic assisted beating heart surgery. In IEEE Transactions on Robotics (Vol. 23, Issue 3, pp. 468–480). Institute of Electrical and Electronics Engineers(IEEE). https://doi.org/10.1109/tro.2007.895077

[3] Shahian, D. M., Silverstein, T., Lovett, A. F., Wolf, R. E., & Normand, S.-L. T. (2007). Comparison of Clinical and Administrative Data Sources for Hospital Coronary Artery Bypass Graft Surgery Report Cards. In Circulation (Vol. 115, Issue 12, pp. 1518 1527). Ovid Technologies (Wolters Kluwer Health). https://doi.org/10.1161/circulationaha.106.633008

[4] Montrief, T., Koyfman, A., & Long, B. (2018). Coronary artery bypass graft surgery complications: A review for emergency clinicians. In The American Journal of Emergency Medicine (Vol. 36, Issue 12, pp. 2289–2297). Elsevier BV. https://doi.org/10.1016/j.ajem.2018.09.014

[5] Tu, J. V., Sykora, K., Naylor, C. D., & Steering Committee of the Cardiac Care Network of Ontario 1. (1997). Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough?. Journal of the American College of Cardiology, 30(5), 1317-1323.

[6] Akintoye, O. O., Fasina, O. P., Adiat, T. S., Nwosu, P. U., Olubodun, M. O., Adu, B. G., ... & Adu, B. (2023). Outcomes of Coronary Artery Bypass Graft Surgery in Africa: A Systematic Review and Meta-Analysis. Cureus, 15(10).