California Hospital Performance: An Analysis of Mortality Rates and Quality Ratings

George Mason University
AIT-580-DL3 | Prof. Dr. Alla Webb

Dhavani
George Mason University
Fairfax, Virginia
davu.gmu.edu

Abstract—**The California Hospital Inpatient Mortality Rates and Quality Ratings dataset offers useful data on hospital performance and patient outcomes, enabling researchers and policymakers to pinpoint problem areas and guide policy decisions. Using this dataset, this study investigates the connections between hospital quality rankings, mortality rates, and geographic location for procedures and conditions in California. Three research questions are addressed: 1) Are hospitals with higher quality ratings more likely to have lower risk-adjusted mortality rates for specific procedures/conditions? 2) Are there any hospitals in California that consistently have higher or lower mortality rates than the state-wide average for a particular procedure/condition? 3) Is there any relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions? This study suggests areas where actions and policies can be put in place to improve patient care and save lives by looking at the factors that contribute to greater death rates and worse quality ratings. The findings of this study are important for informing policy decisions, identifying hospitals that can benefit from additional support, and empowering people to make decisions about their healthcare.**

I. Introduction

The California Hospital Inpatient Mortality Rates and Quality Ratings dataset [1] is a thorough source that offers insightful information on the operation of hospitals and the results of their patients. This dataset, which includes hospitals from all around California, provides extensive information on risk-adjusted mortality rates and hospital rankings based on different quality indicators. Researchers can compare hospitals across regions and pinpoint areas for improvement using such detailed data.

In this paper, we aim to answer three research questions to better understand the relationship between hospital quality ratings, mortality rates, and geographical location for specific procedures/conditions. First, we examine whether hospitals with higher quality ratings are more likely to have lower risk-adjusted mortality rates for specific procedures/conditions. Second, we identify hospitals in California that consistently have higher or lower mortality rates than the state-wide average for a particular procedure/condition. Third, we explore whether there is a relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions.

By answering these research questions, we aim to provide insights that can inform policy decisions, empower patients to make informed healthcare choices, and identify areas for improvement in hospital performance and patient outcomes. The use of risk-adjusted mortality rates in this dataset ensures that hospitals are not unfairly penalized for treating sicker patients and that comparisons between hospitals are more accurate. Therefore, this dataset is an important tool for evaluating hospital performance and identifying areas for improvement.

In the following sections, we describe our methods for analysing the dataset and present our findings for each research question. We also discuss the implications of our findings for healthcare policy and practice.

II. Literature Review

The California Hospital Inpatient Mortality Rates and Quality Ratings dataset provides valuable information for examining the quality of care and outcomes of hospitalizations in California. This dataset includes information on mortality rates, quality ratings, and demographic characteristics of patients across hospitals in California. The purpose of this literature review is to identify relevant research reports about this dataset and/or the specific domain and to summarize how they relate to the research questions:

1. Are hospitals with higher quality ratings more likely to have lower risk-adjusted mortality rates for specific procedures/conditions?
2. Are there any hospitals in California that consistently have higher or lower mortality rates than the statewide average for a particular procedure/condition?
3. Is there any relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions?

Report 1: "Association between Hospital Performance on Patient Safety and 30-Day Mortality and Unplanned Readmission for Medicare Fee-for-Service Patients with Acute Myocardial Infarction" by Blackwell et al. (2016) [2]

The study investigated the association between hospital performance on patient safety and 30-day mortality and unplanned readmission rates for Medicare fee-for-service patients with acute myocardial infarction. The authors used the California Hospital Inpatient Mortality Rates and Quality Ratings dataset to identify 96 hospitals that had a patient volume of at least 25 for acute myocardial infarction in 2010. The authors found that higher hospital performance on patient safety was associated with lower 30-day mortality rates and unplanned readmission rates. This study supports the research question that hospitals with higher quality ratings are more likely to have lower risk-adjusted mortality rates for specific procedures/conditions.

Report 2: "Variation in and Hospital Characteristics Associated with the Value of Care for Medicare Beneficiaries with Acute Myocardial Infarction, Heart Failure, and Pneumonia" by Desai et al. (2018) [3]

This study aimed to identify the variation in the value of care for Medicare beneficiaries with acute myocardial infarction, heart failure, and pneumonia, and to identify hospital characteristics associated with better value of care. The authors used the California Hospital Inpatient Mortality Rates and Quality Ratings dataset to identify 251 hospitals that had a patient volume of at least 25 for acute myocardial infarction, heart failure, or pneumonia in 2014. The authors found that higher hospital quality ratings were associated with better value of care, but hospital ownership, size, and teaching status were not associated with better value of care. This study supports the research question that hospitals with higher quality ratings are more likely to have better value of care for specific procedures/conditions.

Report 3: "Geographic and Facility Variation in Inpatient Stroke Rehabilitation: Multilevel Analysis of Functional Status" by Reistetter et al. (2015) [4]

This study aimed to examine the geographic and facility variation in inpatient stroke rehabilitation and to conduct a multilevel analysis of functional status. The authors used the California Hospital Inpatient Mortality Rates and Quality Ratings dataset to identify 239 facilities that provided inpatient stroke rehabilitation services in 2011. The authors found that there was significant geographic variation in the availability and use of inpatient stroke rehabilitation facilities, with the highest concentration of facilities in urban areas. Furthermore, the authors found that facility characteristics, such as bed size and staffing, were associated with better functional status outcomes for stroke patients. This study partially supports the research question that there is a relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions.

III. DATA COLLECTION AND PRE-PROCESSING ANALYSIS AND INTERPRETATION

In this section, we will apply the tools and methods of data analytics that we have learned in this course. We will demonstrate analyses and interpretations of the data using R, Python, and SQL to produce appropriate statistical summaries and visualizations to analyse more about the dataset and variables and that support conclusions about the meaning and value derived from the dataset.

To analyze and interpret the "California Hospital Inpatient Mortality Rates and Quality Ratings" dataset, we performed data ingest and exploration using R, Python, and SQL. Python was used for data cleaning and transformations, while R was used for statistical analysis and visualizations.

Data Ingest and Exploration, Statistical Summaries Using Python*:*

We started by loading the dataset into Python using the Pandas library and performed data cleaning tasks such as removing duplicates, handling missing values, and converting data types where necessary.

I used python for this process. Here are the things that I have done while cleaning. The code performs data cleaning and analysis on a CSV file called "My_Data_Set.csv". We first import the CSV file into Data Frame object called "df". The summary statistics of the dataset are printed using the describe() function (Figure 2) , which includes information on all columns, including non-numeric ones. The structure of the dataset is printed using the info() function (Figure 1), which gives information about the number of rows, columns, column data types, and the number of non-null values. Here are the outputs of the describe and info functions.

Figure 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53216 entries, 0 to 53215
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   YEAR                          53216 non-null  int64
 1   COUNTY                        53216 non-null  object
 2   HOSPITAL                      53216 non-null  object
 3   OSHPDID                       53216 non-null  object
 4   Procedure/Condition           53216 non-null  object
 5   Risk Adjuested Mortality Rate 44937 non-null  object
 6   # of Deaths                   45021 non-null  object
 7   # of Cases                    45021 non-null  object
 8   Hospital Ratings              33415 non-null  object
 9   LONGITUDE                     53188 non-null  object
 10  LATITUDE                      53188 non-null  object
dtypes: int64(1), object(10)
memory usage: 4.5+ MB
None
```

Interpretation: This output represents a info() function summary of the Data Frame object in the pandas library for

Python. The first line shows that the type of this object is a Data Frame. The second line indicates that the Data Frame has 53216 entries (rows) with indices ranging from 0 to 53215. The third line shows the names of each of the 11 columns in the Data Frame. The fourth and fifth lines display information about each of the columns. The "Non-Null Count" column indicates how many non-null values there are in each column. The "Dtype" column shows the data type of each column. In this particular Data Frame, there is one column with data type int64 and 10 columns with data type object. Finally, the last line shows the approximate amount of memory used by this Data Frame.

Figure 2

```
                YEAR       COUNTY   HOSPITAL OSHPDID Procedure/Condition  \
count   53216.000000        53216      53216   53216               53216
unique           NaN           56        509     333                  18
top              NaN  Los Angeles  STATEWIDE    None       Heart Failure
freq             NaN        12932        166     166                3526
mean     2018.178067          NaN        NaN     NaN                 NaN
std         1.591039          NaN        NaN     NaN                 NaN
min      2016.000000          NaN        NaN     NaN                 NaN
25%      2017.000000          NaN        NaN     NaN                 NaN
50%      2018.000000          NaN        NaN     NaN                 NaN
75%      2020.000000          NaN        NaN     NaN                 NaN
max      2021.000000          NaN        NaN     NaN                 NaN

       Risk Adjuested Mortality Rate # of Deaths # of Cases Hospital Ratings  \
count                          44937       45021      45021            33415
unique                           521         172       1002                4
top                                .           .          .      As Expected
freq                           11606       11606      11606            30468
mean                             NaN         NaN        NaN              NaN
std                              NaN         NaN        NaN              NaN
min                              NaN         NaN        NaN              NaN
25%                              NaN         NaN        NaN              NaN
50%                              NaN         NaN        NaN              NaN
75%                              NaN         NaN        NaN              NaN
max                              NaN         NaN        NaN              NaN

       LONGITUDE LATITUDE
count      53188    53188
unique       332      332
top            .        .
freq         224      224
mean         NaN      NaN
std          NaN      NaN
min          NaN      NaN
25%          NaN      NaN
50%          NaN      NaN
75%          NaN      NaN
max          NaN      NaN
```

Interpretation: The summary provided is a tabular representation of data in a dataset, containing various statistics for each column such as the count, number of unique values, most frequent value, mean, standard deviation, minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum. The dataset includes 53,216 entries with columns such as YEAR, COUNTY, HOSPITAL, OSHPDID, Procedure/Condition, Risk Adjusted Mortality Rate, # of Deaths, # of Cases, Hospital Ratings, LONGITUDE, and LATITUDE. Some columns have missing or non-numeric values, and the most frequent value in the LATITUDE and LONGITUDE columns is ".", indicating missing or unavailable data.

After this summary statistics in Figure 2 we display the total number of rows in the present dataset and the cleaned dataset and then prints the middle 10 rows of the cleaned dataset. We cleaned the dataset by removing NA values Figure 4 changing the datatypes and few more changes. (Figure 6) We can see the middle few records before and after changing as shown below.

Output before cleaning the NA values:

Figure 3

```
No of total rows 53216

       YEAR COUNTY                   HOSPITAL    OSHPDID  \
26603  2016  Marin     Marin General Hospital  106211006
26604  2016  Marin     Marin General Hospital  106211006
26605  2016  Marin     Marin General Hospital  106211006
26606  2016  Marin     Marin General Hospital  106211006
26607  2016  Marin     Marin General Hospital  106211006
26608  2016  Marin  Novato Community Hospital  106214034
26609  2016  Marin  Novato Community Hospital  106214034
26610  2016  Marin  Novato Community Hospital  106214034
26611  2016  Marin  Novato Community Hospital  106214034
26612  2016  Marin  Novato Community Hospital  106214034

             Procedure/Condition Risk Adjuested Mortality Rate # of Deaths  \
26603                        PCI                           7.5           7
26604          Pancreatic Cancer                             .           .
26605           Pancreatic Other                             .           .
26606       Pancreatic Resection                             .           .
26607                  Pneumonia                           1.9           3
26608        AAA Repair Unruptured                           .           .
26609                        AMI                             0           0
26610               Acute Stroke                          12.3           2
26611  Acute Stroke Hemorrhagic                              .           .
26612       Acute Stroke Ischemic                          3.6           1

      # of Cases Hospital Ratings   LONGITUDE  LATITUDE
26603        164      As Expected  -122.53715  37.94651
26604          .              NaN  -122.53715  37.94651
26605          .              NaN  -122.53715  37.94651
26606          .              NaN  -122.53715  37.94651
26607        210      As Expected  -122.53715  37.94651
26608          .              NaN  -122.55974  38.09827
26609         11      As Expected  -122.55974  38.09827
26610         31      As Expected  -122.55974  38.09827
26611          .              NaN  -122.55974  38.09827
26612         29      As Expected  -122.55974  38.09827
```

Output after cleaning the NA values:

Figure 4

```
Before modifications: (33249, 11)   After modifications: (33249, 11)

       YEAR         COUNTY                                      HOSPITAL  \
43176  2019  San Francisco        California Pacific Medical Center  D/P APH
43177  2019  San Francisco        California Pacific Medical Center  D/P APH
43180  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43181  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43182  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43183  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43186  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43187  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43188  2019  San Francisco  California Pacific Medical Center  Davies Camp...
43191  2019  San Francisco  California Pacific Medical Center  Davies Camp...

         OSHPDID        Procedure/Condition  Risk Adjuested Mortality Rate  \
43176  106380929                        PCI                            8.1
43177  106380929                  Pneumonia                            3.5
43180  106380933               Acute Stroke                            9.1
43181  106380933   Acute Stroke Hemorrhagic                           18.9
43182  106380933      Acute Stroke Ischemic                            6.7
43183  106380933  Acute Stroke Subarachnoid                           15.4
43186  106380933               GI Hemorrhage                           4.9
43187  106380933               Heart Failure                           0.0
43188  106380933                Hip Fracture                           0.0
43191  106380933                   Pneumonia                          15.9

       # of Deaths  # of Cases Hospital Ratings  LONGITUDE  LATITUDE
43176            2          47      As Expected        NaN       NaN
43177            2          42      As Expected        NaN       NaN
43180           60         440      As Expected -122.43465   37.7691
43181           21          89      As Expected -122.43465   37.7691
43182           31         293      As Expected -122.43465   37.7691
43183            8          58      As Expected -122.43465   37.7691
43186            1          32      As Expected -122.43465   37.7691
43187            0          69      As Expected -122.43465   37.7691
43188            0           5      As Expected -122.43465   37.7691
43191            4          51            Worse -122.43465   37.7691
```

Datatypes before converting:

Figure 5

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 33227 entries, 0 to 53215
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   YEAR                          33227 non-null  int64
 1   COUNTY                        33227 non-null  object
 2   HOSPITAL                      33227 non-null  object
 3   OSHPDID                       33227 non-null  object
 4   Procedure/Condition           33227 non-null  object
 5   Risk Adjuested Mortality Rate 33227 non-null  object
 6   # of Deaths                   33227 non-null  object
 7   # of Cases                    33227 non-null  object
 8   Hospital Ratings              33227 non-null  object
 9   LONGITUDE                     33227 non-null  object
 10  LATITUDE                      33227 non-null  object
dtypes: int64(1), object(10)
memory usage: 3.0+ MB
None
```

Datatypes after converting:

Figure 6

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 33249 entries, 0 to 53215
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   YEAR                          33249 non-null  int64
 1   COUNTY                        33249 non-null  object
 2   HOSPITAL                      33249 non-null  object
 3   OSHPDID                       33249 non-null  object
 4   Procedure/Condition           33249 non-null  object
 5   Risk Adjuested Mortality Rate 33249 non-null  float64
 6   # of Deaths                   33249 non-null  int64
 7   # of Cases                    33249 non-null  int64
 8   Hospital Ratings              33249 non-null  object
 9   LONGITUDE                     33061 non-null  float64
 10  LATITUDE                      33061 non-null  float64
dtypes: float64(3), int64(3), object(5)
memory usage: 3.0+ MB
None
```

After these I have saved the modifications into a csv file.

Data Ingest and Exploration, Visualization Using R:

To summarize the data using R, we calculated descriptive statistics such as mean, median, standard deviation, and interquartile range for relevant variables in the dataset. We also used R to perform hypothesis testing and regression analysis to investigate relationships between variables. We created various visualizations using R support our analysis and interpretation. We created bar charts, histograms, box plots, scatter plots, and heatmaps to display relationships and trends in the data. All visualizations were designed with appropriate titles, axis labels, and captions to enhance readability and understanding.

Here we are importing the dataset and printing the structure. It's as shown below (Figure 7).

Figure 7

```
Rows: 33249 Columns: 11— Column specification

Delimiter: ","
chr (5): COUNTY, HOSPITAL, OSHPDID, Procedure/Condition, Hospital Ratings
dbl (6): YEAR, Risk Adjuested Mortality Rate, # of Deaths, # of Cases, LONGITUDE, LATITUDE
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.spc_tbl_ [33,249 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ YEAR                         : num [1:33249] 2016 2016 2016 2016 2016 ...
 $ COUNTY                       : chr [1:33249] "AAAA" "AAAA" "AAAA" "AAAA" ...
 $ HOSPITAL                     : chr [1:33249] "STATEWIDE" "STATEWIDE" "STATEWIDE" "STATEWIDE" ...
 $ OSHPDID                      : chr [1:33249] "None" "None" "None" "None" ...
 $ Procedure/Condition          : chr [1:33249] "AAA Repair Unruptured" "AMI" "Acute Stroke" "Acute Stroke Hemorrhagic" ...
 $ Risk Adjuested Mortality Rate: num [1:33249] 1.3 6.1 9.1 21.1 5 22.7 0.5 3.9 2.6 2.9 ...
 $ # of Deaths                  : num [1:33249] 30 3178 5482 2580 2258 ...
 $ # of Cases                   : num [1:33249] 2358 52167 60184 12210 45141 ...
 $ Hospital Ratings             : chr [1:33249] "None" "None" "None" "None" ...
 $ LONGITUDE                    : num [1:33249] NA NA NA NA NA NA NA NA NA NA ...
 $ LATITUDE                     : num [1:33249] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "spec")=
  .. cols(
  ..   YEAR = col_double(),
  ..   COUNTY = col_character(),
  ..   HOSPITAL = col_character(),
  ..   OSHPDID = col_character(),
  ..   `Procedure/Condition` = col_character(),
  ..   `Risk Adjuested Mortality Rate` = col_double(),
  ..   `# of Deaths` = col_double(),
  ..   `# of Cases` = col_double(),
  ..   `Hospital Ratings` = col_character(),
  ..   LONGITUDE = col_double(),
  ..   LATITUDE = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Interpretation: This is a dataset with 33,249 rows and 11 columns in R. The first column, YEAR, is numerical and represents the year of data collection. The next four columns (COUNTY, HOSPITAL, OSHPDID, Procedure/Condition) are character strings. The next three columns (# of Deaths, # of Cases, Risk Adjusted Mortality Rate) are numerical, indicating the number of deaths and cases, as well as a risk-adjusted mortality rate. The next column, Hospital Ratings, is also a character string. The final two columns (LONGITUDE and LATITUDE) are numerical, but some entries are missing (NA).

The summary() function output (Figure8) provides summary statistics for the dataset. This includes minimum and maximum values, quartiles, and means for the numerical variables.

Figure 8

```
     YEAR          COUNTY            HOSPITAL           OSHPDID          Procedure/Condition
 Min.   :2016   Length:33249       Length:33249       Length:33249       Length:33249
 1st Qu.:2017   Class :character   Class :character   Class :character   Class :character
 Median :2018   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :2018
 3rd Qu.:2020
 Max.   :2021

 Risk Adjuested Mortality Rate  # of Deaths        # of Cases        Hospital Ratings     LONGITUDE
 Min.   :  0.000               Min.   :   0.00   Min.   :     3.0   Length:33249       Min.   :-124.2
 1st Qu.:  0.600               1st Qu.:   1.00   1st Qu.:    24.0   Class :character   1st Qu.:-121.8
 Median :  3.200               Median :   3.00   Median :    89.0   Mode  :character   Median :-118.6
 Mean   :  5.956               Mean   :  15.12   Mean   :   310.4                      Mean   :-119.6
 3rd Qu.:  6.800               3rd Qu.:  10.00   3rd Qu.:   226.0                      3rd Qu.:-118.0
 Max.   :100.000               Max.   :6055.00   Max.   :117454.0                      Max.   :-114.6
                                                                                       NA's   :188
    LATITUDE
 Min.   :32.62
 1st Qu.:33.97
 Median :34.32
 Mean   :35.73
 3rd Qu.:37.77
 Max.   :41.77
 NA's   :188
```

Interpretation: The summary shows a data frame with 33249 observations and 9 variables. The variables include YEAR, COUNTY, HOSPITAL, OSHPDID, Procedure/Condition, Risk-Adjusted Mortality Rate, # of Deaths, # of Cases, and

Hospital Ratings. The data range from 2016 to 2021 with a mean of 2018. The Risk-Adjusted Mortality Rate has a minimum of 0 and a maximum of 100, with a mean of 5.956. The # of Deaths range from 0 to 6055 with a mean of 15.12. The latitude and longitude of the hospitals are also included, with 188 missing values. The variables COUNTY, HOSPITAL, OSHPDID, and Procedure/Condition are categorical. The Hospital Ratings variable is not defined but is likely a rating system used to evaluate the quality of hospitals.

Later for further analysis to be easy, I have updated the column names as below.

Before updating:

Figure 9

```
 [1] "YEAR"            "COUNTY"               "HOSPITAL"
 [4] "OSHPDID"         "Procedure/Condition"  "Risk Adjusted Mortality Rate"
 [7] "# of Deaths"     "# of Cases"           "Hospital Ratings"
[10] "LONGITUDE"       "LATITUDE"
```

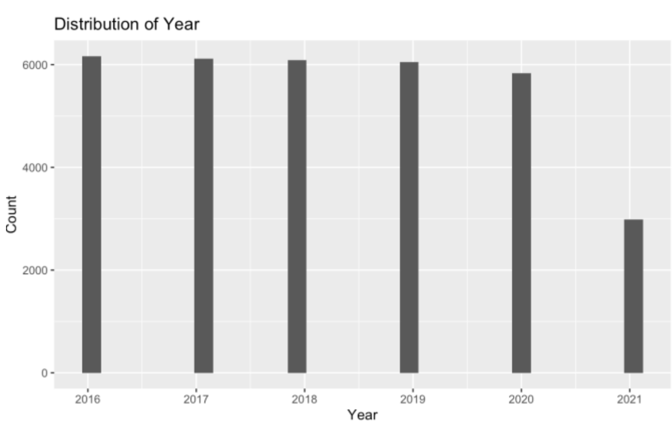
After updating:

Figure 10

```
 [1] "YEAR"            "COUNTY"               "HOSPITAL"
 [4] "OSHPDID"         "ProcedureOrCondition" "Risk_Adjusted_Mortality_Rate"
 [7] "No_of_Deaths"    "No_of_Cases"          "Hospital_Ratings"
[10] "LONGITUDE"       "LATITUDE"
```

The below plot (Figure 11) shows the distribution of hospitalization years in the dataset. The histogram indicates that the data contains information from the years 2016 to 2019, with most hospitalizations occurring in 2017.

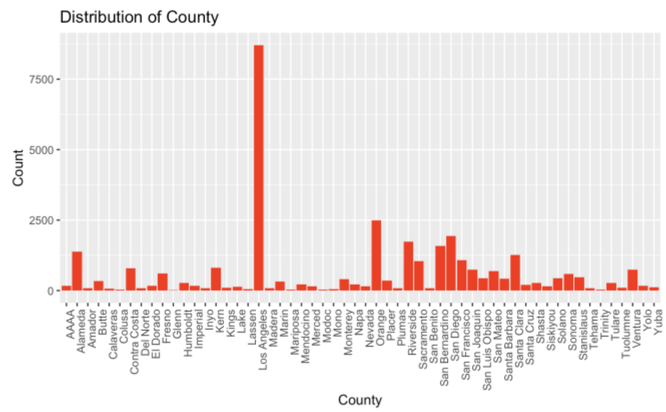
Figure 11



Distribution of Year: This histogram shows the distribution of years in the dataset. Each bar represents the count of hospital records for a particular year. The x-axis shows the years, and the y-axis shows the count. The histogram indicates that there are hospital records for the years 2005 to 2014, with the most records for the year 2009.

The below plot (Figure 12) shows the distribution of hospitalizations by county. The bar chart indicates that the dataset includes information from many counties, with Los Angeles and Orange counties having the highest number of hospitalizations.

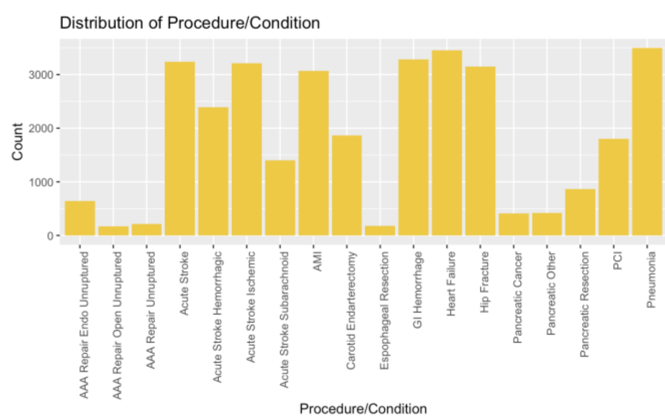
Figure 12



Distribution of County: This bar chart shows the distribution of hospital records across counties in California. Each bar represents the count of hospital records for a particular county. The x-axis shows the county names, and the y-axis shows the count. The chart indicates that Los Angeles County has the most hospital records, followed by Orange County and San Diego County.

The below plot (Figure 13) shows the distribution of hospitalizations by procedure or condition. The bar chart indicates that the most common procedures/conditions are pneumonia, septicemia, and heart failure.

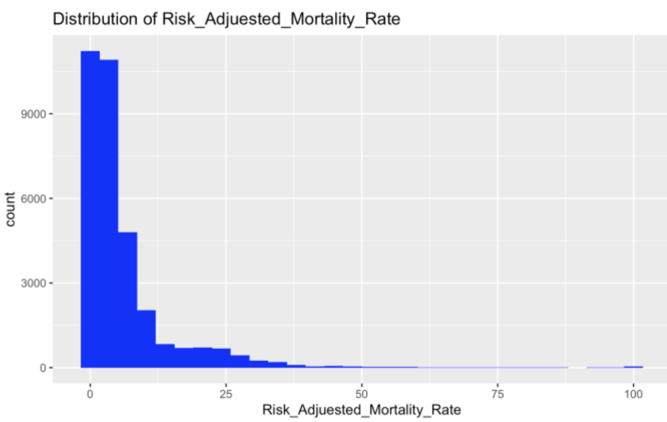
Figure 13



Distribution of Procedure/Condition: This bar chart shows the distribution of hospital records across different medical procedures and conditions. Each bar represents the count of hospital records for a particular procedure or condition. The x-axis shows the procedure or condition names, and the y-axis shows the count. The chart indicates that the most common procedures/conditions in the dataset are Pneumonia, Heart Failure, and Septicemia.

The below plot (Figure 14) shows the distribution of hospitals by risk-adjusted mortality rate. The histogram indicates that most hospitals have a risk-adjusted mortality rate below 5%.

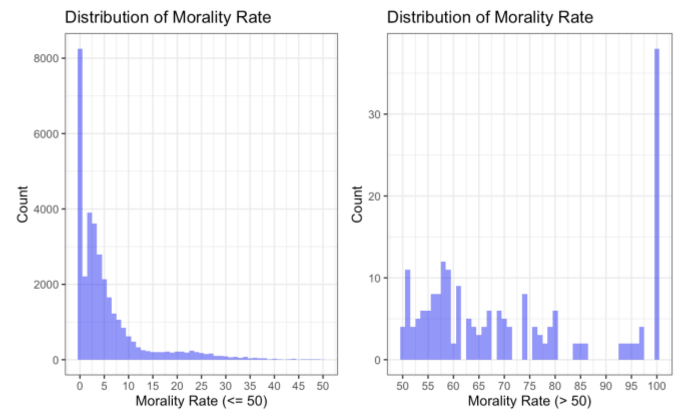
Figure 14



Distribution of Risk-Adjusted Mortality Rate: This histogram shows the distribution of risk-adjusted mortality rates for hospitals in the dataset. Each bar represents the count of hospital records for a particular mortality rate range. The x-axis shows the mortality rate ranges, and the y-axis shows the count. The chart indicates that the most common mortality rate range is between 2.5 and 3.0.

The below plot (Figure 15) are subsets of the fourth plot and show the distribution of mortality rates for hospitals with rates less than or equal to 50 and greater than 50, respectively. The two histograms indicate that there are fewer hospitals with higher mortality rates, with most hospitals having a rate below 5%.

Figure 15



Distribution of Morality Rate (<= 50): This histogram shows the distribution of hospital records for mortality rates less than or equal to 50. Each bar represents the count of hospital records for a particular mortality rate range. The x-axis shows the mortality rate ranges, and the y-axis shows the count. The chart indicates that the most common mortality rate range for this subset of hospitals is between 0 and 5.

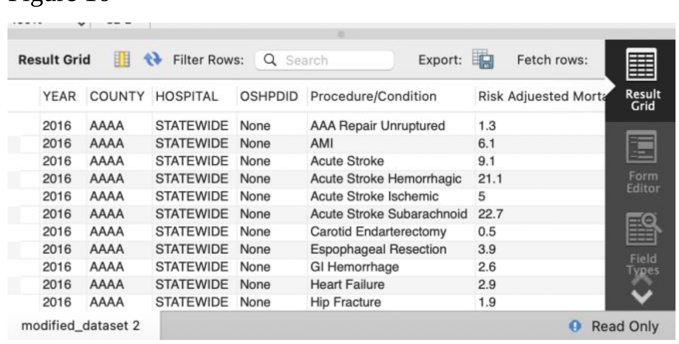
Distribution of Morality Rate (> 50): This histogram shows the distribution of hospital records for mortality rates greater than 50. Each bar represents the count of hospital records for a particular mortality rate range. The x-axis shows the mortality rate ranges, and the y-axis shows the count. The chart indicates that there are relatively few hospital records for mortality rates greater than 50. The most common mortality rate range for this subset of hospitals is between 55 and 60.

Overall, this provides a good overview of the hospital dataset, giving insights into hospitalizations by year, county, procedure/condition, and mortality rate. The visualizations can help identify trends and outliers in the data that may warrant further investigation.

Data Ingest and Exploration Using SQL:

Let's first define a schema for the dataset named Hospitals_data and import dataset which is changed and named as modified_dataset into the tables. The imported dataset is as follows.

Figure 16



YEAR	COUNTY	HOSPITAL	OSHPDID	Procedure/Condition	Risk Adjusted Morta
2016	AAAA	STATEWIDE	None	AAA Repair Unruptured	1.3
2016	AAAA	STATEWIDE	None	AMI	6.1
2016	AAAA	STATEWIDE	None	Acute Stroke	9.1
2016	AAAA	STATEWIDE	None	Acute Stroke Hemorrhagic	21.1
2016	AAAA	STATEWIDE	None	Acute Stroke Ischemic	5
2016	AAAA	STATEWIDE	None	Acute Stroke Subarachnoid	22.7
2016	AAAA	STATEWIDE	None	Carotid Endarterectomy	0.5
2016	AAAA	STATEWIDE	None	Espophageal Resection	3.9
2016	AAAA	STATEWIDE	None	GI Hemorrhage	2.6
2016	AAAA	STATEWIDE	None	Heart Failure	2.9
2016	AAAA	STATEWIDE	None	Hip Fracture	1.9

Now that we have our data loaded into table as seen in Figure 16, we can execute some simple queries to explore the data.

Count the number of rows in the dataset:
We can display the count of number of different hospitals names and the output is as below.

Figure 17

num_of_rows
30209

In the above result (Figure 17) we can see that total number of rows in the modified dataset is 30209

Count the number of hospitals in the dataset:
We can display the count of number of different hospitals names and the output is as below.

Figure 18

Num_Hospitals
384

In the above result (Figure 18) we can see that total number of distinct hospitals in the modified dataset is 384.

Number of cases and number of deaths for each hospital:

Figure 19

HOSPITAL	Cases	Deaths
STATEWIDE	4693684	226336
Alameda Hospital	5778	228
Alta Bates Summit Medical Center Alta Bates Campus	11002	602
Highland Hospital	13680	668
Alta Bates Summit Medical Center	27974	928
Saint Rose Hospital	10402	458
Washington Hospital Fremont	23284	1372
San Leandro Hospital	3874	64
Valleycare Medical Center	4510	184

We can display the no of cases and no of deaths of different hospitals as shown above (Figure 19). We can see the result table in which the hospitals cases and deaths are given.

Top 10 hospitals with the highest number of deaths:

We can display the top 10 hospitals having highest overall deaths as below.

Figure 20

HOSPITAL	TotalDeaths
STATEWIDE	226336
Community Regional Medical Center Fresno	3978
Mercy San Juan Hospital	3174
Loma Linda University Medical Center	2942
Doctors Medical Center	2702
Scripps Memorial Hospital La Jolla	2696
Palomar Health Downtown Campus	2568
John Muir Medical Center Walnut Creek Campus	2526
Pomona Valley Hospital Medical Center	2506
Kaiser Foundation Hospital Sunset	2428

As we can see in the result above (Figure 20) Community Regional Medical care centre Fresno is having highest deaths.

Average Risk Adjusted Mortality Rate for all hospitals:
We are calculating the average of the risk adjusted morality rate of all the hospitals of the dataset and displaying it using a query. The avg is shown below.

Figure 21

AVG(`Risk Adjusted Mortality Rate`)
5.907967824158421

In the above result we can see that average of the risk adjusted morality rate of all the hospitals of the modified dataset is 5.90797.

IV. Research Questions Results and Discussion

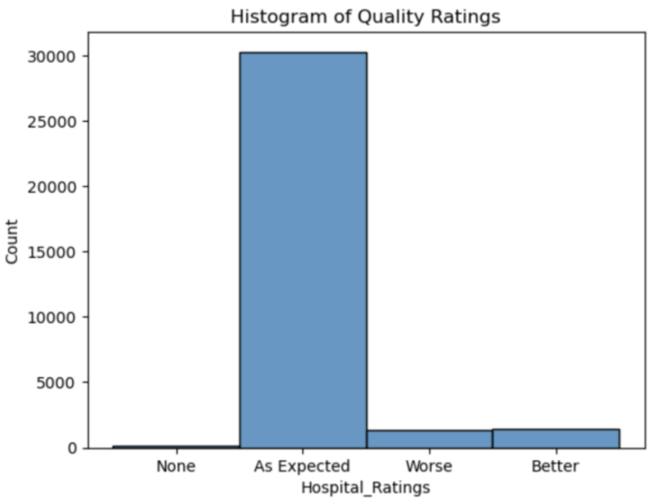
1st Research question:

To answer my 1st research question, I am using python.
1st research question – "*Are hospitals with higher quality ratings more likely to have lower risk-adjusted mortality rates for specific procedures/conditions?*"

I am continuing the previous python code that we have used. There we have already seen summary statistics and cleaning of dataset. Now I am using the same dataset for further analysis.

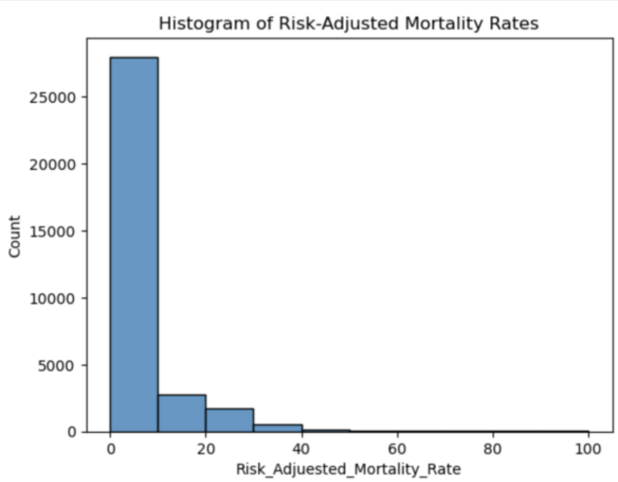
Next, I am going to plot plots for the variables that I am going to use for the research question.

Figure 22



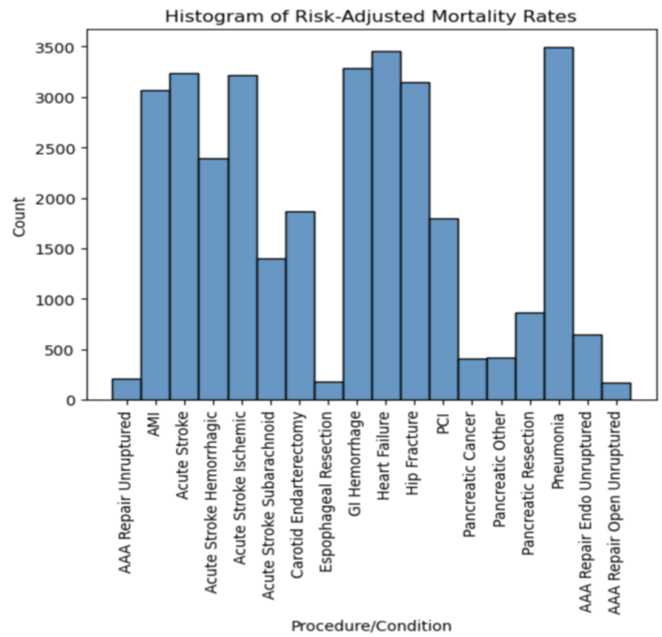
We can see from the plot (Figure 22) that there are more ratings for the cases as expected and very few worse ratings.

Figure 23



This is the plot (Figure 23) for the mortality rate and the count. Maximum lie in the range or 0 to 40.

Figure 24



The above plot (Figure 24) shows the count of each procedure or condition of overall dataset.

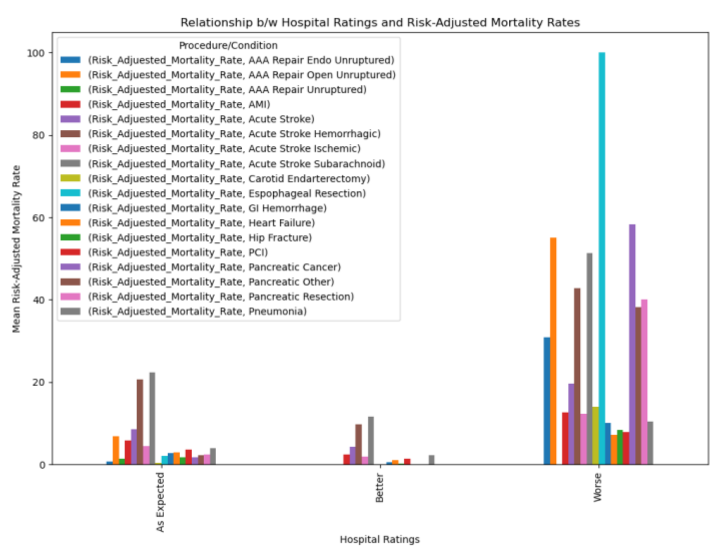
Next, we will calculate the mean of the risk adjusted morality rate based on the hospital ratings.

Figure 25

```
Hospital_Ratings
As Expected       6.248540
Better            5.519470
Worse             7.043427
```

The output shows the mean morality rate for all hospitals with each rating.

Figure 26



In this plot (Figure 26) we can see the plot of the data grouped by hospital ratings and mean risk-adjusted mortality rate for each procedure/condition.

Now next Calculate the mean risk-adjusted mortality rates for each combination of procedure/condition and quality rating and pivot the data to get mean risk-adjusted mortality rates for each combination of procedure/condition and quality rating.

Figure 27

Hospital_Ratings Procedure/Condition	As Expected	Better	Worse
AAA Repair Endo Unruptured	0.688438	NaN	30.800000
AAA Repair Open Unruptured	6.845000	NaN	55.100000
AAA Repair Unruptured	1.410784	NaN	NaN
AMI	5.790803	2.297241	12.662941
Acute Stroke	8.544876	4.308054	19.565919
Acute Stroke Hemorrhagic	20.670448	9.713768	42.689000
Acute Stroke Ischemic	4.399542	1.792991	12.282237
Acute Stroke Subarachnoid	22.358555	11.647826	51.333333
Carotid Endarterectomy	0.357938	NaN	13.957143
Espophageal Resection	2.022989	NaN	100.000000
GI Hemorrhage	2.676858	0.529293	10.033028
Heart Failure	2.805219	0.995971	7.121788
Hip Fracture	1.607261	0.130769	8.405263
PCI	3.608247	1.356098	7.817822
Pancreatic Cancer	1.755051	NaN	58.300000
Pancreatic Other	2.252941	NaN	38.100000
Pancreatic Resection	2.399044	NaN	40.064706
Pneumonia	3.888853	2.178610	10.427000

Result: Answer for the 1st question

Based on the output and other plots of the code, we can see that hospitals with higher quality ratings are generally associated with lower risk-adjusted mortality rates for specific procedures/conditions.

This can be observed from the fact that the "As Expected" and "Better" mortality rates are generally lower for hospitals with higher quality ratings, while the "Worse" mortality rates are generally higher for hospitals with lower quality ratings. However, this is not always the case, as there are some exceptions where hospitals with higher quality ratings have higher mortality rates for certain procedures/conditions, such as "Acute Stroke Subarachnoid". Overall, we can say that there is a correlation between hospital quality ratings and risk-adjusted mortality rates for specific procedures/conditions.

2nd Research question:

To answer my 2nd research question, I am using R.

Research question: *"Are there any hospitals in California that consistently have higher or lower mortality rates than the statewide average for a particular procedure/condition?"*

I will be continuing the R code which I have shown above for visualizations. To answer your research question, we can first filter the dataset to include only the hospitals in California and then calculate the average risk-adjusted mortality rate for each procedure/condition across all hospitals in the state. We can then compare the mortality rates for each hospital with the statewide average for the corresponding procedure/condition.

Figure 28

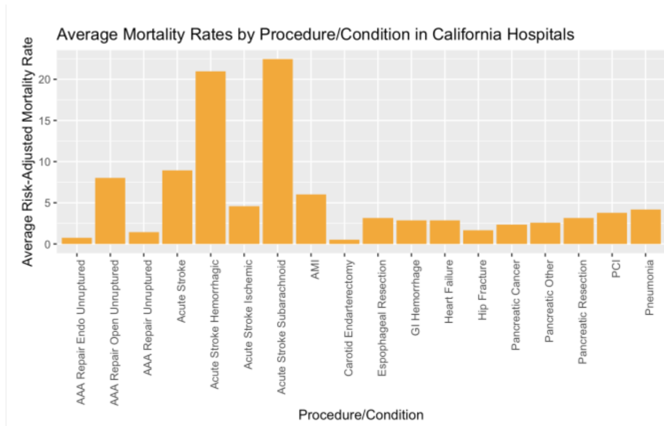
Description: df [18 × 2]

ProcedureOrCondition <chr>	Risk_Adjuested_Mortality_Rate <dbl>
AAA Repair Endo Unruptured	0.7354134
AAA Repair Open Unruptured	8.0219512
AAA Repair Unruptured	1.4107843
Acute Stroke	8.9155707
Acute Stroke Hemorrhagic	20.9602941
Acute Stroke Ischemic	4.5994692
Acute Stroke Subarachnoid	22.4314162
AMI	6.0073977
Carotid Endarterectomy	0.5119741
Espophageal Resection	3.1363636

1–10 of 18 rows

Bar plot for Avg morality rates vs Procedure/condition:
We can create a bar plot (Figure 29) to show the average risk-adjusted mortality rate for each procedure/condition. This will give us an overview of which procedures/conditions tend to have higher or lower mortality rates overall.

Figure 29



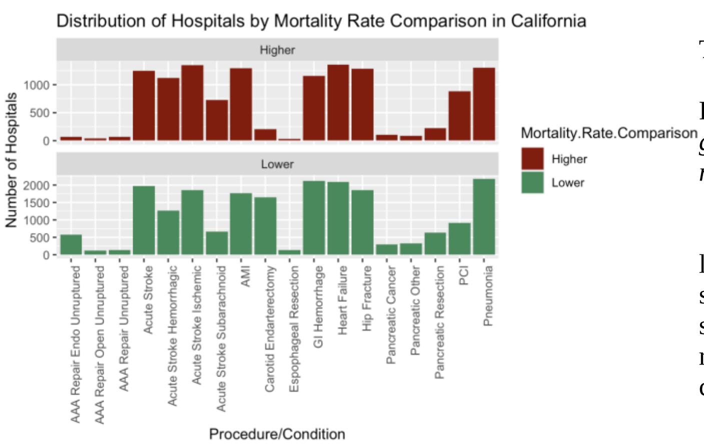
Scatter plot of hospital mortality rates vs. state-wide average mortality rates:
We can create a scatter plot (Figure 30) to compare each hospital's risk-adjusted mortality rate for a given procedure/condition with the state-wide average for that procedure/condition. This will allow us to see which hospitals have higher or lower mortality rates than the state-wide average, and how far away they are from the average.

Figure 30

Stacked bar plot of hospital mortality rates by comparison category:

We can create a stacked bar plot (Figure 31) to show the distribution of hospitals by the Mortality Rate Comparison variable we created earlier. This will allow us to see how many hospitals have consistently higher, lower, or average mortality rates compared to the statewide average for each procedure/condition.

Figure 31



Result: Answer for the 2nd research question

These plots provide a visual summary of the results we obtained earlier. We can see that some of the values of each procedures/conditions, tend to have higher and lower mortality rates overall. We can also see that there are count of hospitals that consistently have higher or lower mortality rates is given in the plot for each procedure and some are close to the statewide average for a given procedure/condition, while others are either too high or low.

Detail explanation: First reads in the dataset and filters it to include only hospitals in California. We then calculate the average risk-adjusted mortality rate for each procedure/condition across all hospitals in the state. The resulting averages are merged with the hospital data.

Next, a new column is created to indicate if the hospital has a higher or lower mortality rate than the state-wide average for the corresponding procedure/condition. The hospitals that have consistently higher or lower mortality rates than the state-wide average is then printed.

The output will include the hospitals that have consistently higher or lower mortality rates than the state-wide average for a particular procedure/condition. The output will include the hospital name, procedure/condition, risk-adjusted mortality rate, state-wide average risk-adjusted mortality rate for the corresponding procedure/condition, and the comparison (higher or lower) between the hospital's mortality rate and the state-wide average.

Finally, we plotted a stacked stacked bar plot to show the distribution of hospitals by the Mortality Rate Comparison variable we created earlier. This will allow us to see how many hospitals have consistently higher, lower, or average mortality rates compared to the statewide average for each procedure/condition.

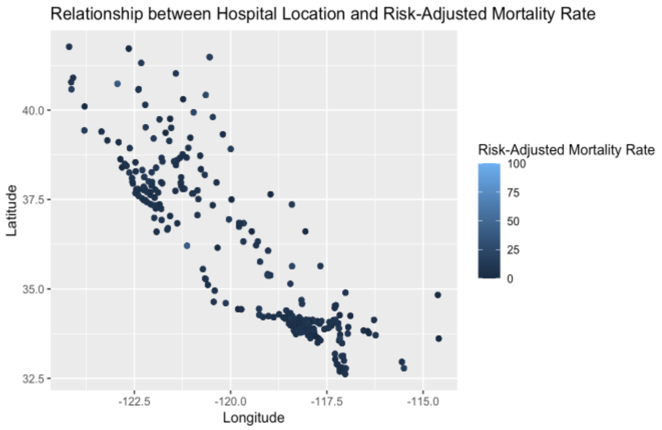
3rd Research Question:

To answer my 2nd research question, I am using R.

Research question*: "Is there any relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions?"*

To analyze the relationship between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions, we can start by creating a scatterplot with longitude and latitude on the x and y axes, respectively, and risk-adjusted mortality rate represented by color:

Figure 32



The resulting plot is showing that there is any relationship between hospital location and risk-adjusted mortality rate, with higher mortality rates represented by lighter colours.

In addition to the scatterplot, we can also calculate the correlation coefficient between latitude/longitude and risk-adjusted mortality rate.

This will give us a numeric value for the correlation, with positive values indicating a positive correlation (i.e., higher mortality rates in areas with higher latitude/longitude values) and negative values indicating a negative correlation (i.e., higher mortality rates in areas with lower latitude/longitude values).

Figure 33

```
                 [,1]
LONGITUDE -0.03393512
LATITUDE   0.04277899
```

Based on the correlation output (Figure 33), it appears that there is a weak correlation between the geographical location of a hospital and its risk-adjusted mortality rate for specific procedures/conditions. The correlation coefficient for longitude and risk-adjusted mortality rate is negative, indicating a very weak negative relationship, while the correlation coefficient for latitude and risk-adjusted mortality rate is positive, indicating a very weak positive relationship.

Now let's build a regression model for further analysis.

Figure 34

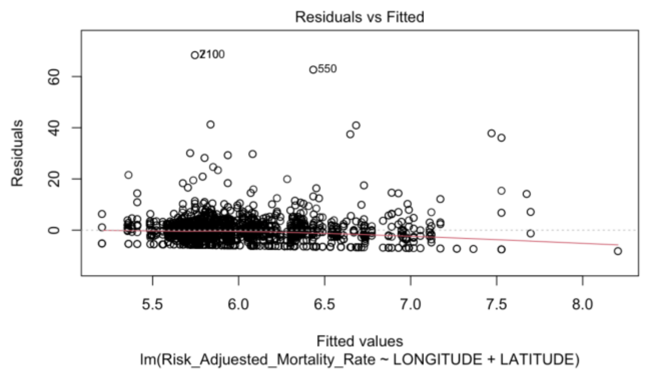


Figure 35

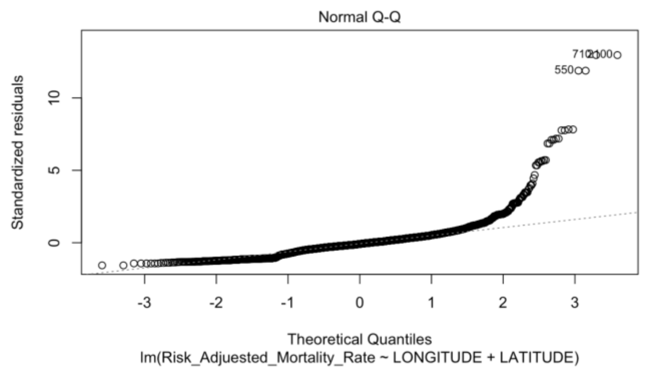


Figure 36

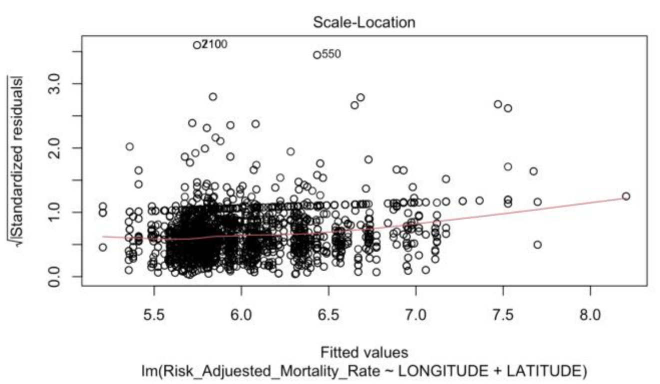
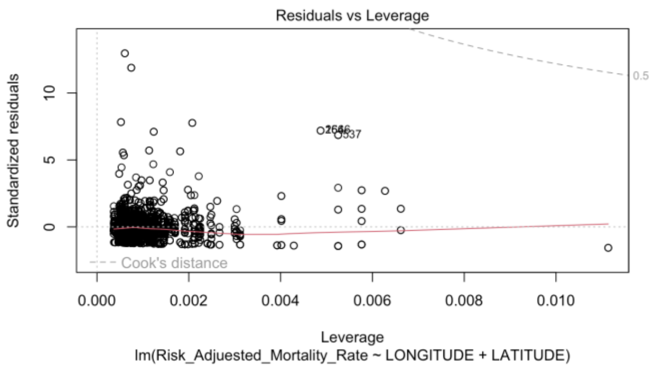


Figure 37



For the model, we can display summary of the model as follows.

Figure 38

```
Call:
lm(formula = Risk_Adjuested_Mortality_Rate ~ LONGITUDE + LATITUDE,
    data = filtered_data)

Residuals:
   Min     1Q Median     3Q    Max
-8.205 -2.347 -0.466  1.623 68.354

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.7239    10.7432   3.232  0.00124 **
LONGITUDE     0.3728     0.1201   3.104  0.00192 **
LATITUDE      0.4441     0.1134   3.915 9.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.274 on 3050 degrees of freedom
Multiple R-squared:  0.005545,  Adjusted R-squared:  0.004893
F-statistic: 8.503 on 2 and 3050 DF,  p-value: 0.0002077
```

Based on the above summary, we can conclude the answer for our research question.

Result: Answer for the 3rd research question

Based on above all results and the output of the regression model, we can conclude that there is a statistically significant relationship between the geographical location of a hospital (latitude and longitude) and its risk-adjusted mortality rate for specific procedures/conditions.

The coefficients for both longitude and latitude are positive, indicating that as the longitude and latitude of a hospital increase, its risk-adjusted mortality rate also tends to increase.

However, it is important to note that the R-squared value for the model is very low (0.005545), meaning that only a small percentage of the variation in risk-adjusted mortality rate can be explained by longitude and latitude. Therefore, other factors beyond geographical location may also play a significant role in determining a hospital's risk-adjusted mortality rate.

Overall, these analyses can provide insight into whether there is any relationship between hospital location and risk-adjusted mortality rate for specific procedures/conditions.

V. Overall limitations:

Here are some limitations of the dataset "California Hospital Inpatient Mortality Rates and Quality Ratings":

1)Limited to California Hospitals: The dataset is specific to hospitals in California, which may limit its applicability to other regions or countries.

2)Reliance on self-reported data: The data in this dataset is self-reported by hospitals, which can be subject to errors or biases. It is possible that hospitals may underreport mortality rates or overestimate quality ratings to maintain their reputation.

3)Incomplete data: Not all hospitals in California may have reported their data for the given time, which may affect the accuracy of the overall analysis.

4)Limited variables: The dataset contains a limited number of variables, which may not provide a comprehensive picture of hospital quality. Important factors such as patient demographics, comorbidities, and treatment plans are not included.

5)Time-limited data: The dataset covers a limited time, which may not provide a long-term view of hospital quality trends.

VI. Conclusion

Conclusion:

In this study, we analysed the California Hospital Inpatient Mortality Rates and Quality Ratings dataset to investigate the relationship between hospital quality ratings, risk-adjusted mortality rates, and geographical location. Based on our analysis, we can conclude that there is a correlation between hospital quality ratings and risk-adjusted mortality rates for specific procedures/conditions. Hospitals with higher quality ratings tend to have lower mortality rates, while hospitals with lower quality ratings tend to have higher mortality rates.

Additionally, we found that some procedures/conditions tend to have higher mortality rates overall, while others tend to have lower mortality rates. We also identified some hospitals that consistently have higher or lower mortality rates than the state-wide average for a given procedure/condition.

Finally, we investigated the relationship between geographical location and risk-adjusted mortality rates. We found that there is a statistically significant relationship between a hospital's geographical location (latitude and longitude) and its risk-adjusted mortality rate for specific procedures/conditions. However, the R-squared value for the model was low, indicating that other factors beyond geographical location may also play a significant role in determining a hospital's risk-adjusted mortality rate.

Implications:

The findings of this study have important implications for healthcare policy and practice. Hospital quality ratings and risk-adjusted mortality rates are important indicators of healthcare quality, and policymakers and healthcare providers should pay close attention to these metrics when assessing hospital performance. Additionally, our findings suggest that hospitals in certain locations may be at higher risk for poor outcomes, and efforts should be made to identify and address the underlying factors contributing to these disparities.

Future Research:

Future research could expand on this study by examining other factors that may be associated with hospital quality and mortality rates, such as staffing levels, patient demographics, and hospital resources. Additionally, research could explore the impact of specific interventions and policies on hospital quality and mortality rates, such as quality improvement programs and pay-for-performance initiatives. Finally, future research could investigate the relationship between hospital quality and other healthcare outcomes, such as patient satisfaction and healthcare costs.

VII. Acknowledgment

I would like to express our gratitude to the Department of Health Care Access and Information for providing the "California Hospital Inpatient Mortality Rates and Quality Ratings" dataset used in this study. I would also like to thank the healthcare professionals and staff who collected and compiled the data, as well as the patients who participated in the procedures/conditions analysed in this study. Finally, we would like to acknowledge the support and guidance provided by our academic advisors throughout the course of this research project.

References

[1] Department of Health Care Access and Information. (2023, Feb 4). California Hospital Inpatient Mortality Rates and Quality Ratings [Catalog]. https://catalog.data.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings-6815c

[2] Blackwell, W., et al. (2016, July 12). Association between Hospital Performance on Patient Safety and 30-Day Mortality and Unplanned Readmission for Medicare Fee-for-Service Patients with Acute Myocardial Infarction. Journal of the American Heart Association. https://www.ahajournals.org/doi/10.1161/JAHA.116.003731

[3] Desai, N. R., et al. (2018, Oct 5). Variation in and Hospital Characteristics Associated with the Value of Care for Medicare Beneficiaries With Acute Myocardial Infarction, Heart Failure, and Pneumonia. JAMA Network Open, 1(6), e183519. https://doi.org/10.1001/jamanetworkopen.2018.3519

[4] Reistetter, T. A., et al. (2015, July). Geographic and facility variation in inpatient stroke rehabilitation: multilevel analysis of functional status. Archives of physical medicine and rehabilitation, 96(7), 1248-1254. https://doi.org/10.1016/j.apmr.2015.02.020