Healthcare Capstone Project - Hospital Compare

SUBMITTED BY:

1. RICHA GOEL
2. ANNAMALAI GANAPATHY
3. ANKUR JOHARI
4. RAVI RANJAN KUMAR

Business Analysis Objectives

Problem Statement:

We need to develop an approach to calcualte hospital rating and using it to identify areas of improvement for certain hospitals.

Abstract:

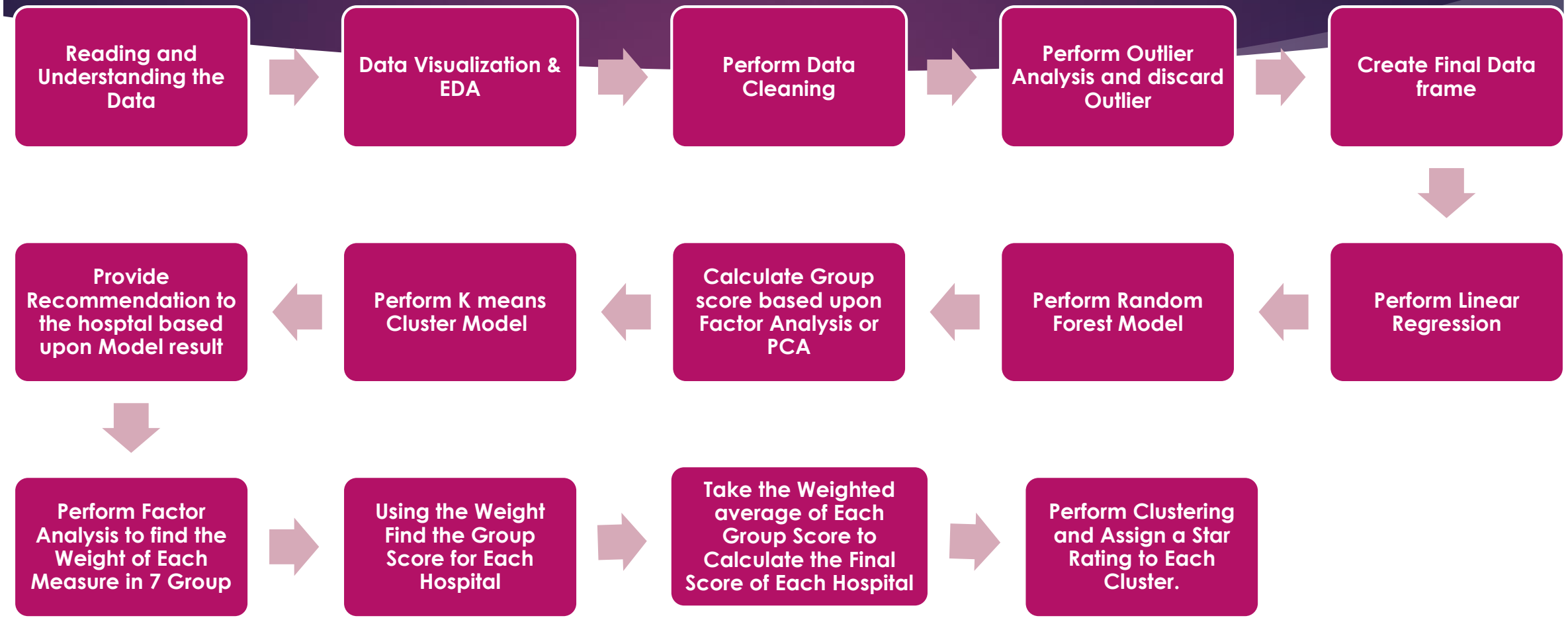
CMS rates hospitals in the US on a scale of 1-5, with the objective of making it easier for patients and consumers to compare the quality of services offered by hospitals.

The ratings directly influence the choice of hospitals made by consumers and may significantly impact hospitals revenues.

Project Goals

- We need to determine CMS rating as per below Methods :-
- **Part 1 - Supervised Learning-Based Rating :-**
  - We have to build two types of Model(Random Forest and Linear Regression) Model to predict the hospital rating (1-5) using the 64 measure id provided by CMS in different file .
- **Part 2 - Clustering-Based Rating (Unsupervised)**
  - We have to create an unsupervised Clustering model to assign hospital rating(1-5) using the measure specified by CMS.We can use either PCA or factor analysis to calcualte Group score for Clustering .
- **Part 3 - Recommendations for Hospitals**
  - Using your understanding of the star rating system, We have to recommend ways to improve the rating

Problem Solving Methodology-CRISP DM



Identification of relevant files and Measures

- We have been provided with raw data with the File Name: Hospital_Revised_FlatFiles_20161110
- We identified total 7 groups for available measures identified respective files for those measures.

- Other than measure group and their corresponding raw data files, we have **Hospital General Information.CSV** file which has various information related to provider along with their overall rating which would be useful in supervised learning modelling.

Identification of relevant files and Measures

We identified total 7 groups for available measures respective files for those measures

Group	Identified File
Mortality	Readmission and Deaths - Hospital
Safety of Care	Healthcare Associated Infections - Hospital
Readmission	Readmission and Deaths - Hospital
Complications	Complications - Hospital
HCAHPS - Hospital	Patient Experience
Timeliness of Care	Timely and Effective Care - Hospital
Efficient Use of Medical Imaging	Outpatient Imaging Efficiency - Hospital

Data quality checks

Hospital General Information.CSV :-

- It has total 4818 rows and 28 columns. Only three columns are numeric while 25 are non-numeric.
- Three types of Hospital Type are available: Acute Care Hospitals, Critical Access Hospitals, Children's. Acute Care Hospitals count is maximum out of all three -3382.
- Hospital overall rating has Not Available values for 1170 records.
- The Footnotes column for all 7 group has more than 60% missing data .
- Columns - Mortality national comparison, Safety of care national comparison, Readmission national comparison, Patient experience national comparison, Effectiveness of care national comparison, Timeliness of care national comparison , Efficient use of medical imaging national comparison. These have four types of distinct values - 'Same as the National average', 'Below the National average','Not Available' , 'Above the National average'. These are non-numeric values.

Data quality checks

File name	Columns Details	No of Rows and Columns Columns	Details Missing data	Additional Comments
Readmission and Deaths - Hospital	67452 Rows and 18 Columns	3 numeric and 15 non-numeric columns.	Footnote column has highest missing percentage.	Measure Id column has two types of prefixes – MORT and READM. The columns having 'Not Available' values are Compared to National,Denominator, Score, Lower Estimate, Higher Estimate.
Healthcare Associated Infection	231264 rows and 15 columns	3 numeric and 12 non-numeric columns	Compared to National, Footnote , Score and County Name has missing data.	The columns having 'Not Available' values are Compared to National, Score
Complications - Hospital	52998 rows and 18 columns	3 numeric and 15 non-numeric columns	The columns having 'Not Available' values are Compared to National, Denominator , Score, Lower Estimate, Higher Estimate - Compared to National, Footnote , County Name , Denominator , Score, Lower Estimate, and Higher Estimate have missing data.	The columns having 'Not Available' values are Compared to National, Denominator , Score, Lower Estimate, Higher Estimate
Timely and Effective Care	207174 rows and 16 column	3 numeric and 13 non-numeric columns.	Sample , Score, Footnote and County Name have missing data.	The column having 'Not Available' values are Score, Sample.

Data quality checks

File name	Columns Details	No of Rows and Columns Columns	Details Missing data	Additional Comments
HCAHPS - Hospital	42096 rows and 22 columns	3 numeric and 11 non-numeric columns	Number of Completed Surveys Footnote, HCAHPS Linear Mean Value, HCAHPS Answer Percent Footnote , HCAHPS Answer Percent , Patient Survey Star Rating Footnote , Survey Response Rate Percent Footnote ,County Name , Survey Response Rate Percent have missing data.	The columns having 'Not Available' values are Patient Survey Star Rating, HCAHPS Answer Percent, HCAHPS Linear Mean Value, Number of Completed Surveys, Survey Response Rate Percent
Outpatient Imaging Efficiency	28908 rows and 14 columns	3 numeric and 11 non-numeric columns	Footnote , Score, County Name have missing data	The column having 'Not Available' value is Score.
HCAHPS - Hospital	42096 rows and 22 columns	3 numeric and 11 non-numeric columns	Number of Completed Surveys Footnote, HCAHPS Linear Mean Value, HCAHPS Answer Percent Footnote , HCAHPS Answer Percent , Patient Survey Star Rating Footnote , Survey Response Rate Percent Footnote ,County Name , Survey Response Rate Percent have missing data.	The columns having 'Not Available' values are Patient Survey Star Rating, HCAHPS Answer Percent, HCAHPS Linear Mean Value, Number of Completed Surveys, Survey Response Rate Percent

Data preparation /Creation of Group Files

Hospital General Information.CSV:-

- DataFrame is filtered on Hospital Type equal to Acute Care Hospitals
- DataFrame is filtered to remove records having Hospital overall rating as Not Available
- 7 new columns are derived as per below logic :-

  Below the National average replaced with 0 , Same as the National average with 1, and Above the National average as 3 and all the null value columns are replaced with mean value of that columns.
- These columns are dropped due to high missing percentage: Readmission national comparison footnote , Effectiveness of care national comparison footnote, Timeliness of care national comparison footnote, Mortality national comparison footnote, Hospital overall rating footnote, Patient experience national comparison footnote, Safety of care national comparison footnote, Efficient use of medical imaging national comparison footnote.
- Final DataFrame has 3061 unique records and provider id.
- Hospital overall rating is converted to numeric values.

Data preparation /Creation of Group Files

Readmission and Deaths :-

- All the ‘Not Available’ values in all columns are replaced by ‘NaN’.
- Remove all the rows having ‘NaN’ in Score column.
- Two separate Group DataFrames for Mortality and Readmission are created based upon four characters prefixes (MORT and READ) in Measure Id column.
- All the Measure ID present in Mortality and Readmission DataFrames are converted from rows to columns (Long to Wide format) by using Pivot Table operation.
- In both newly created Group DataFrames only score column is included for all the Measure ID. All other columns are dropped.
- Mortality Group DataFrame has 4818 rows and 6 columns for Measure ID score.
- Readmission Group DataFrame has 4818 rows and 8 columns for Measure ID score.

Data preparation /Creation of Group Files

Complications - Hospital.CSV

- All the ‘Not Available’ values in all columns are replaced by ‘NaN’.
- Remove all the rows having ‘NaN’ in Score column.
- One Complication Group DataFrame is created by converting all Measure ID from rows to columns format(Long to Wide format) by using Pivot Table operation with Score column.
- In Complication Group DataFrames only score column is included for all the Measure ID. All other columns are dropped.
- Complication Group DataFrame has 3484 rows and 11 columns for Measure ID score.

Data preparation /Creation of Group Files

HCAHPS - Hospital.CSV

- All the ‘Not Available’ values in all columns are replaced by ‘NaN’.
- Remove all the rows having ‘NaN’ in Score column.
- Patient Experience Group DataFrame is created by converting all Measure ID from rows to columns format(Long to Wide format) by using Pivot Table operation with Score column.
- In Patient Experience Group DataFrames only score column is included for all the Measure ID. All other columns are dropped.
- Patient Experience Group DataFrame has 3508 rows and 12 columns for Measure ID score.

Data preparation /Creation of Group Files

Outpatient Imaging Efficiency - Hospital.CSV

- All the ‘Not Available’ values in all columns are replaced by ‘NaN’.
- Remove all the rows having ‘NaN’ in Score column.
- Efficient use of Medical Imaging Group DataFrame is created by converting all Measure ID from rows to columns format(Long to Wide format) by using Pivot Table operation with Score column.
- In Efficient use of Medical Imaging Group DataFrames only score column is included for all the Measure ID. All other columns are dropped.
- Efficient use of Medical Imaging Group DataFrame has 37836 rows and 8 columns for Measure ID score

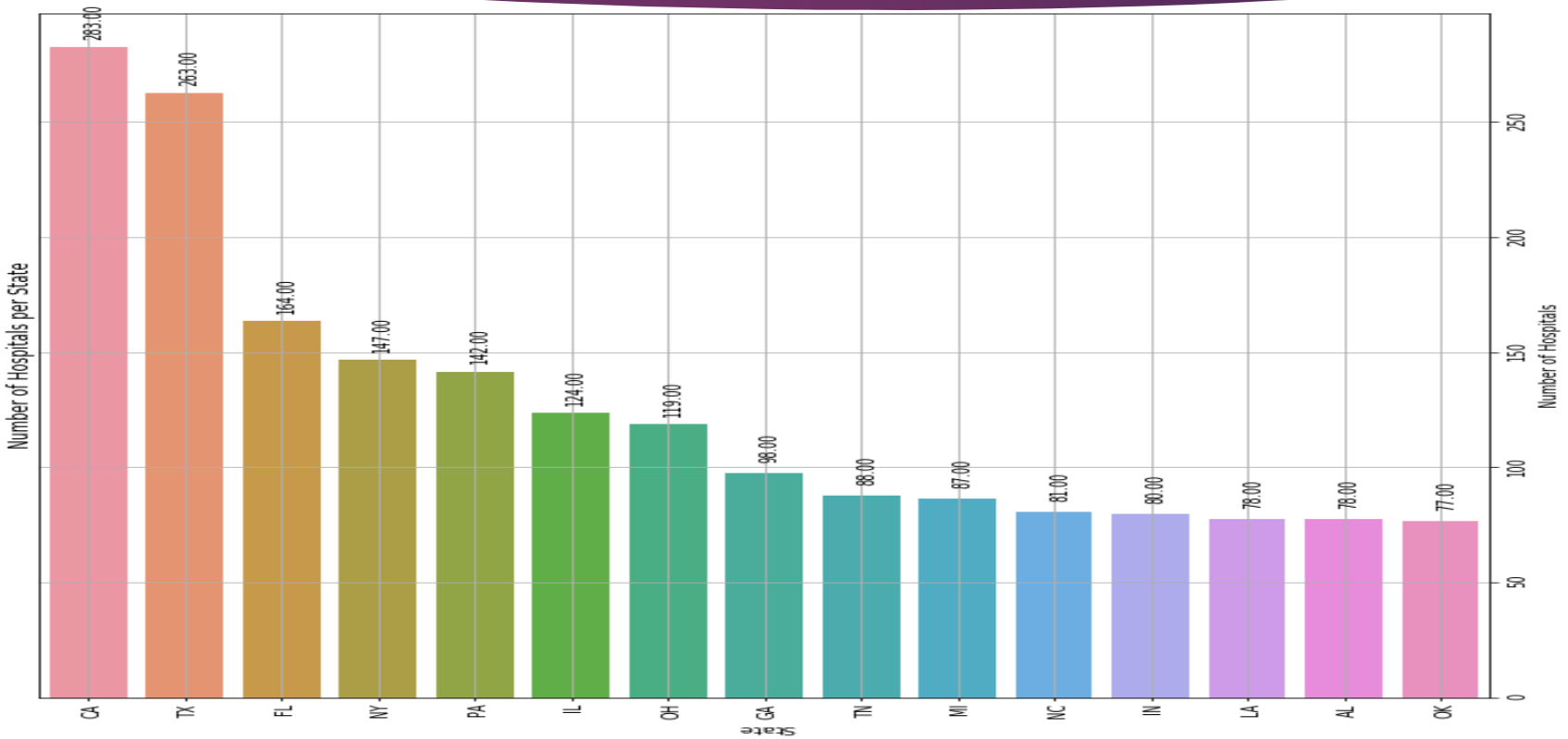
Data preparation /Creation of Group Files

Timely and Effective Care - Hospital.CSV

- All the ‘Not Available’ values in all columns are replaced by ‘NaN’.
- Remove all the rows having ‘NaN’ in Score column.
- Timeliness of Care Group DataFrame is created by converting all Measure ID from rows to columns format(Long to Wide format) by using Pivot Table operation with Score column.
- In Timeliness of Care Group DataFrames only score column is included for all the Measure ID. All other columns are dropped.
- Timeliness of Care Group DataFrame has 4463 rows and 43 columns for Measure ID score.

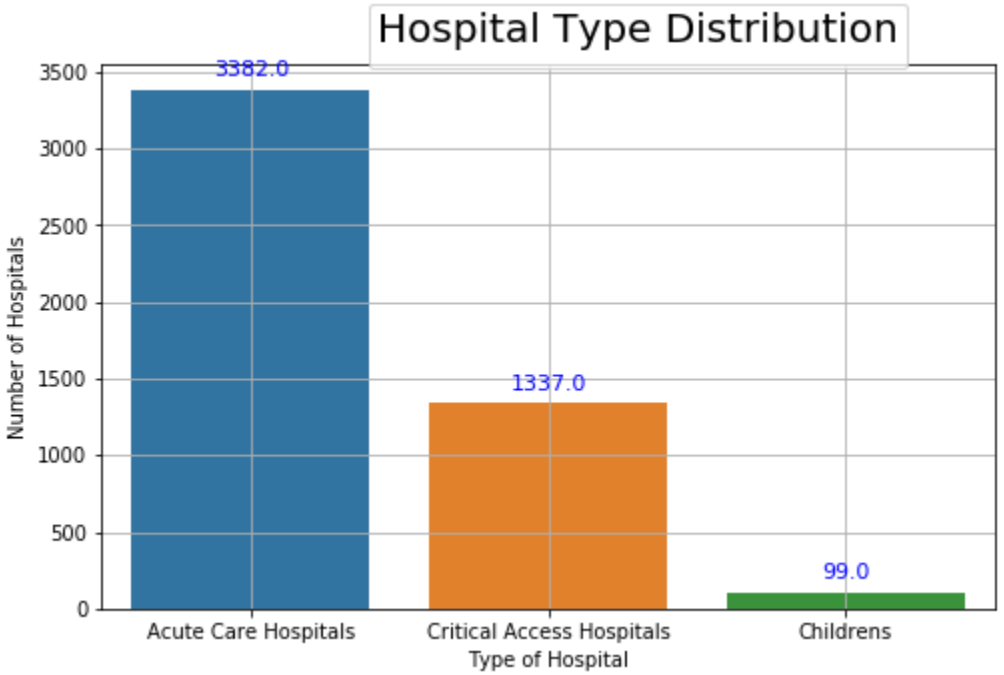
Exploratory Data Analysis (EDA)

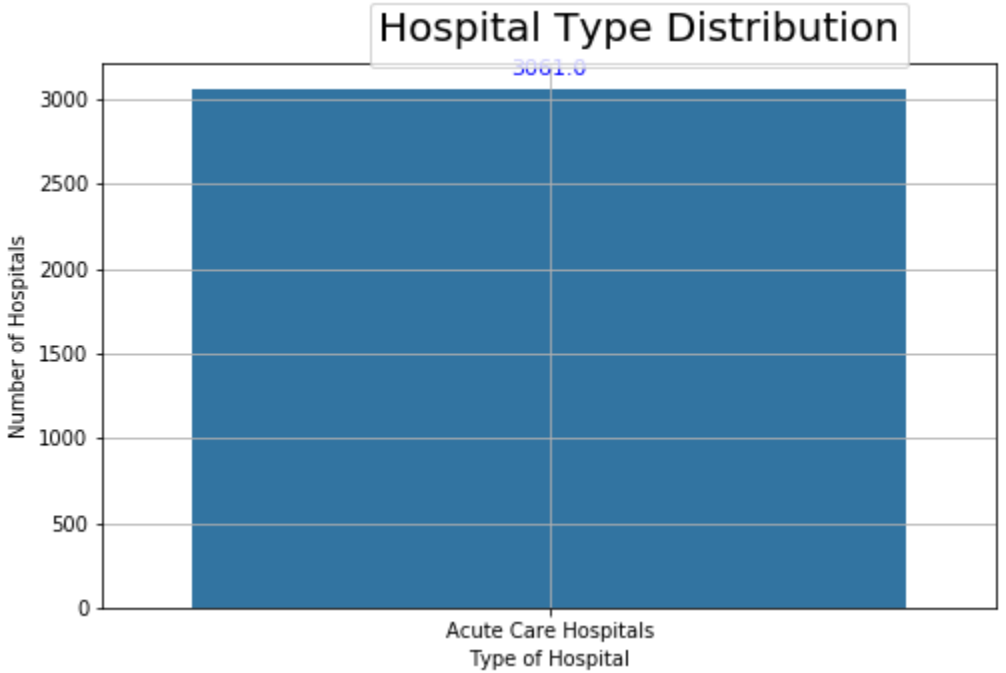
1. Hospitals per State



Insight: This is a bar plot between Number of Hospitals and State. We can see number of hospitals is maximum in California while minimum in OK.

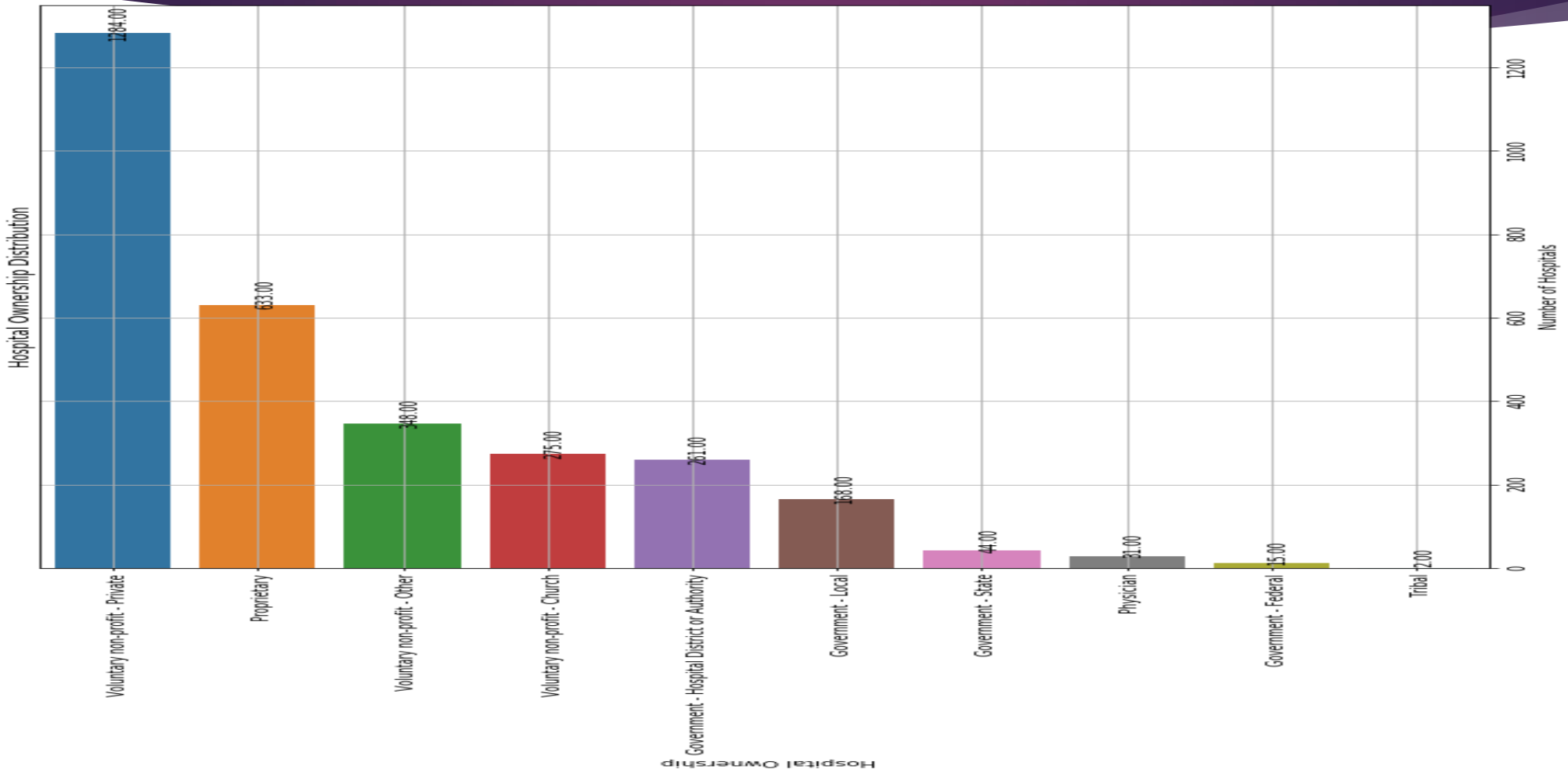
2. Hospital Type Distribution





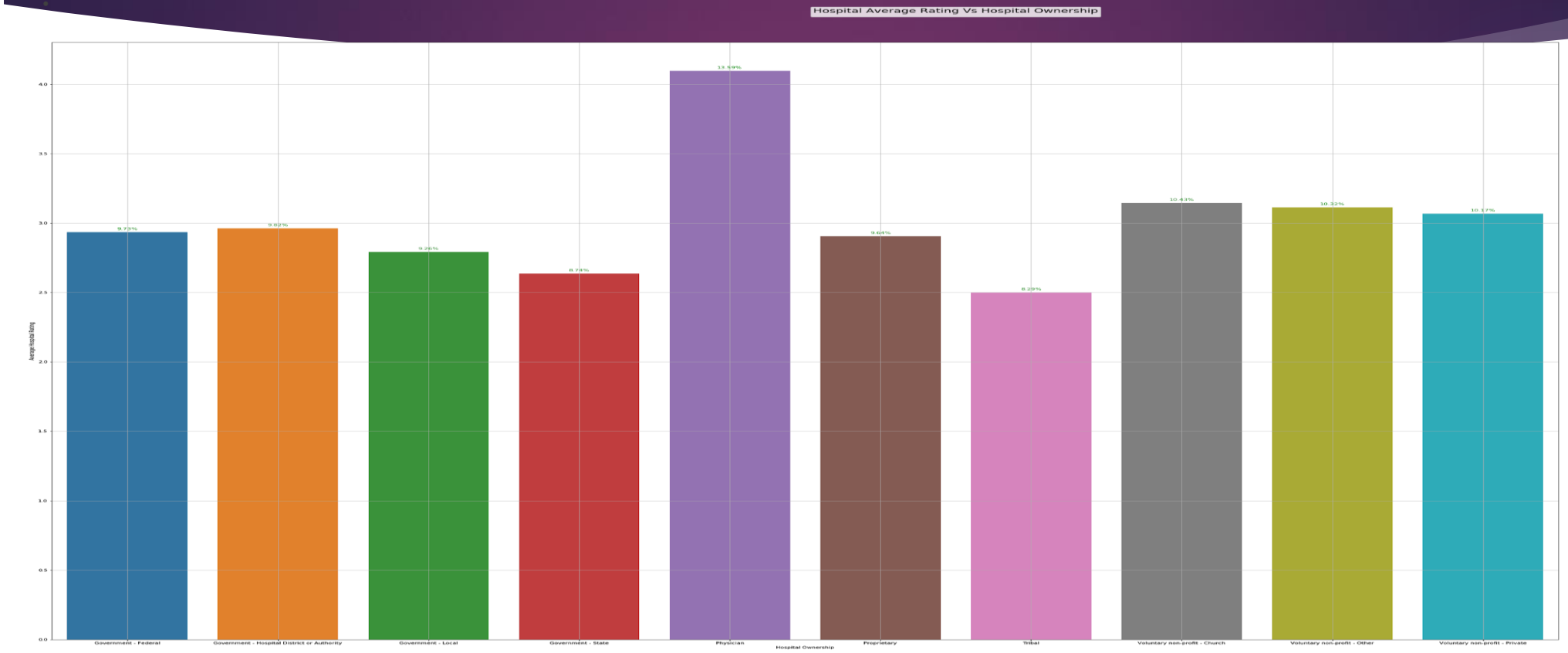
Insight: Initially we had highest hospitals count as Acute care hospitals and lowest as Childrens hospitals. After filtration on Not Available Scores we are left with a total of 3061 hospitals are Acute Care Hospitals.

3. Hospital Ownership Distribution



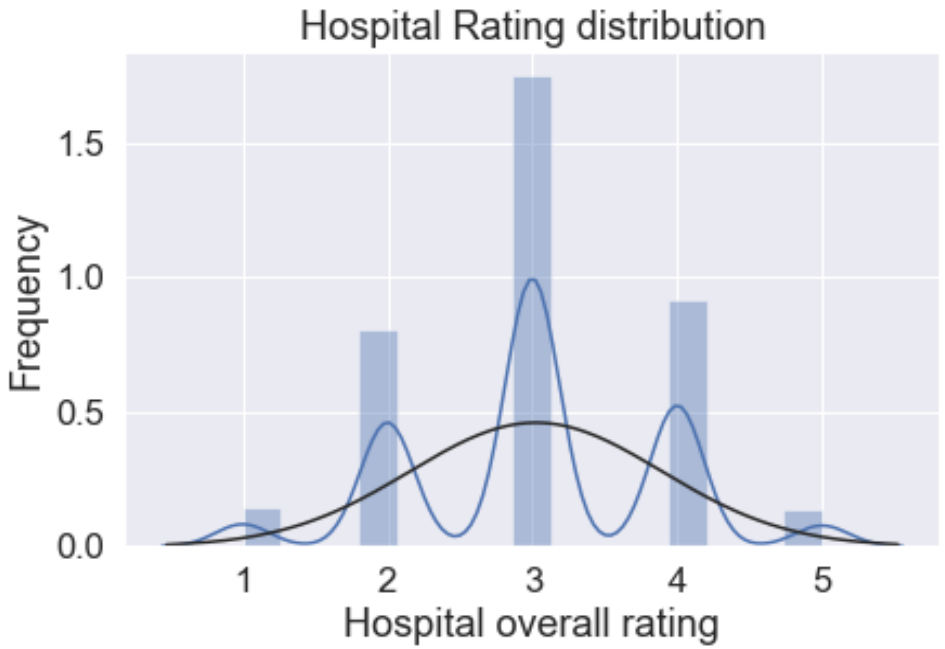
Insight: Voluntary non-profit Private hospitals are maximum in count while tribal hospitals are least.

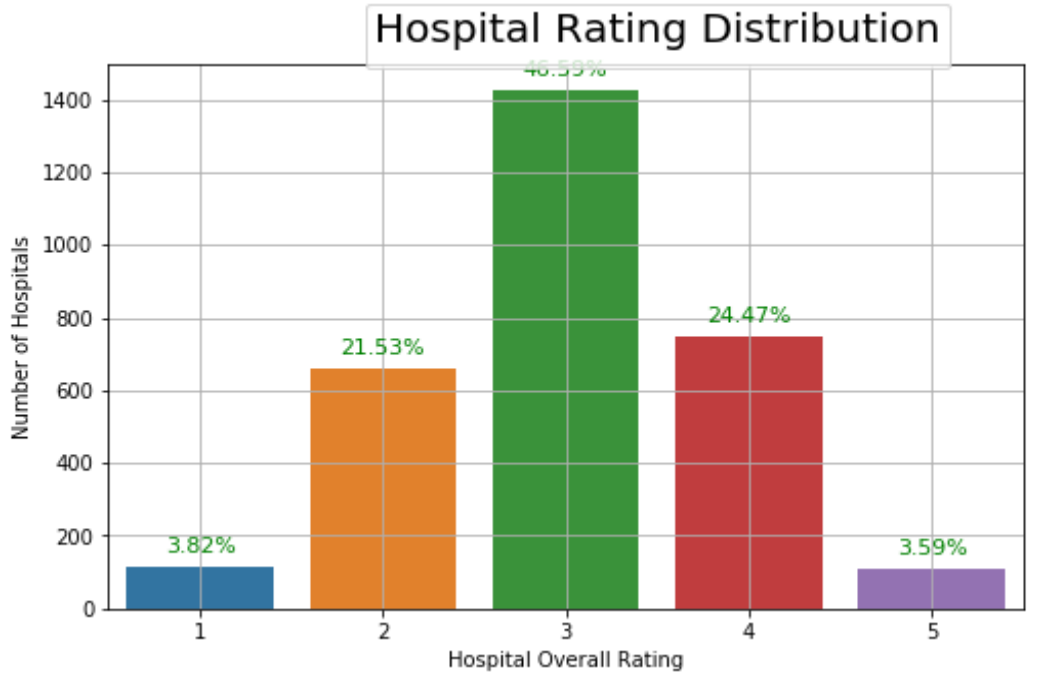
4. Hospital Ownership Distribution



Insight: This is a plot between hospital ownership and average overall rating. On average the hospitals owned by Physician has highest rating while owned by Tribal has lowest rating.

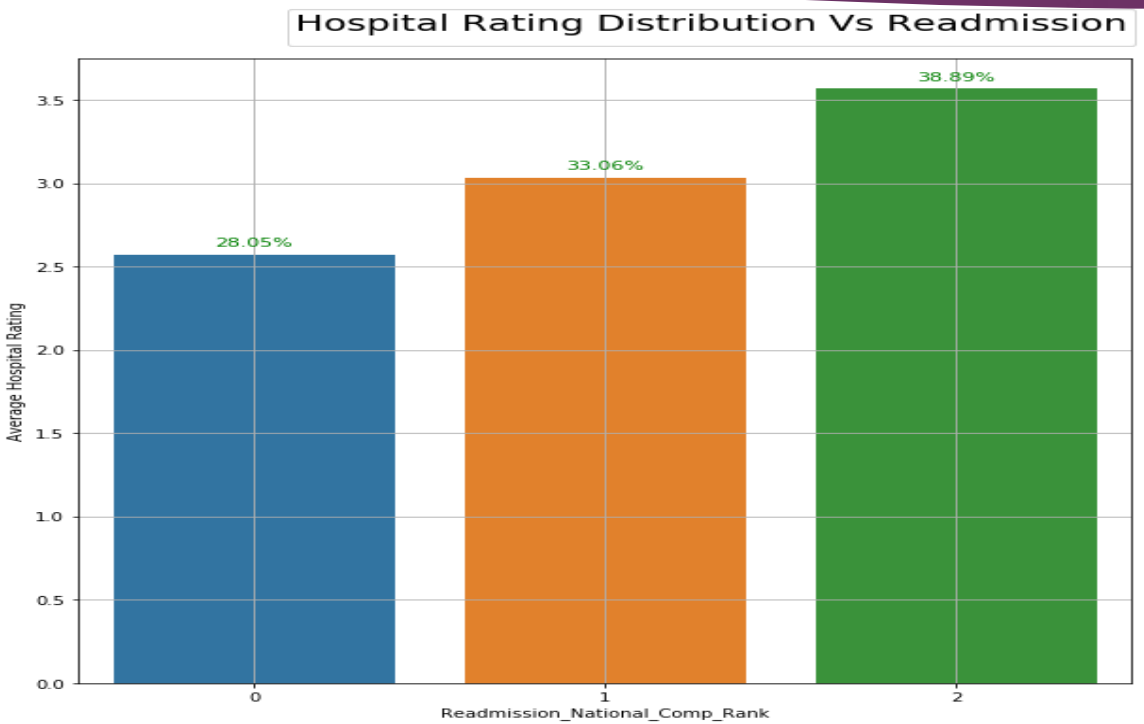
5. Hospital Rating Distribution

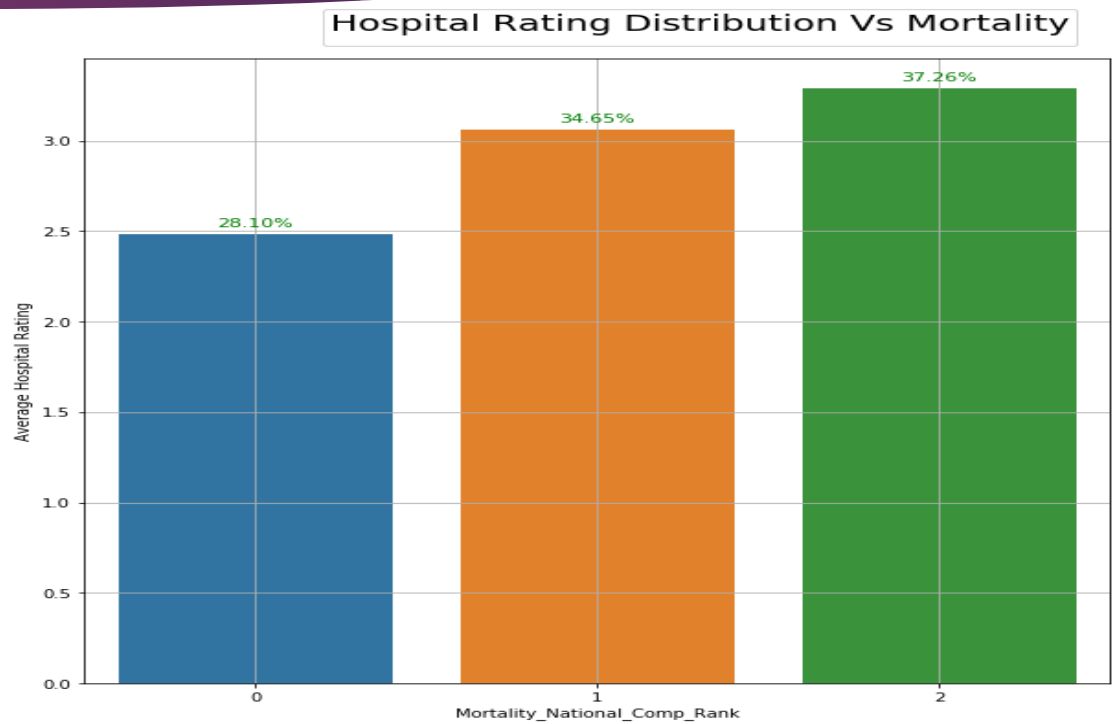




Insight: This is a plot between Hospital Overall Rating provided by CMS (1, 2 ,3 , 4, 5) and Number of Hospitals.Hospitals with Overall rating as 3 are highest in count while hospitals with overall rating as 5 are lowest in count. Hospital Overall Rating distribution is also a normal distribution.

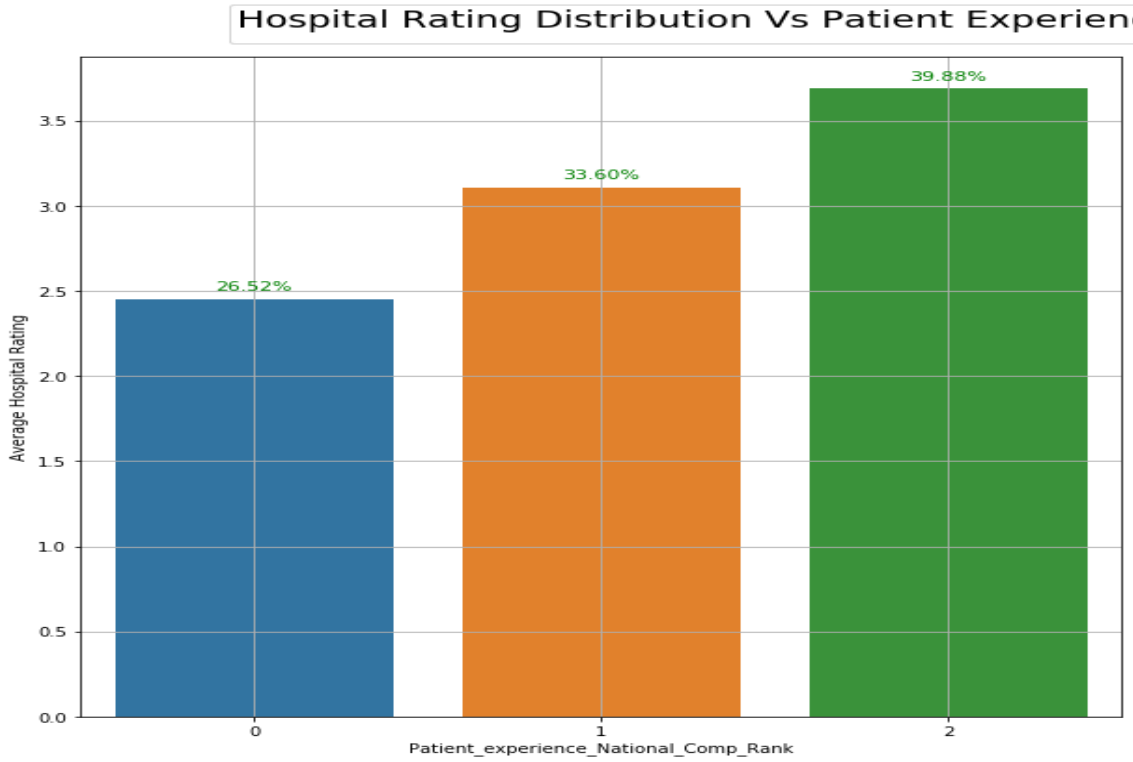
6. Readmission and Mortality Distribution

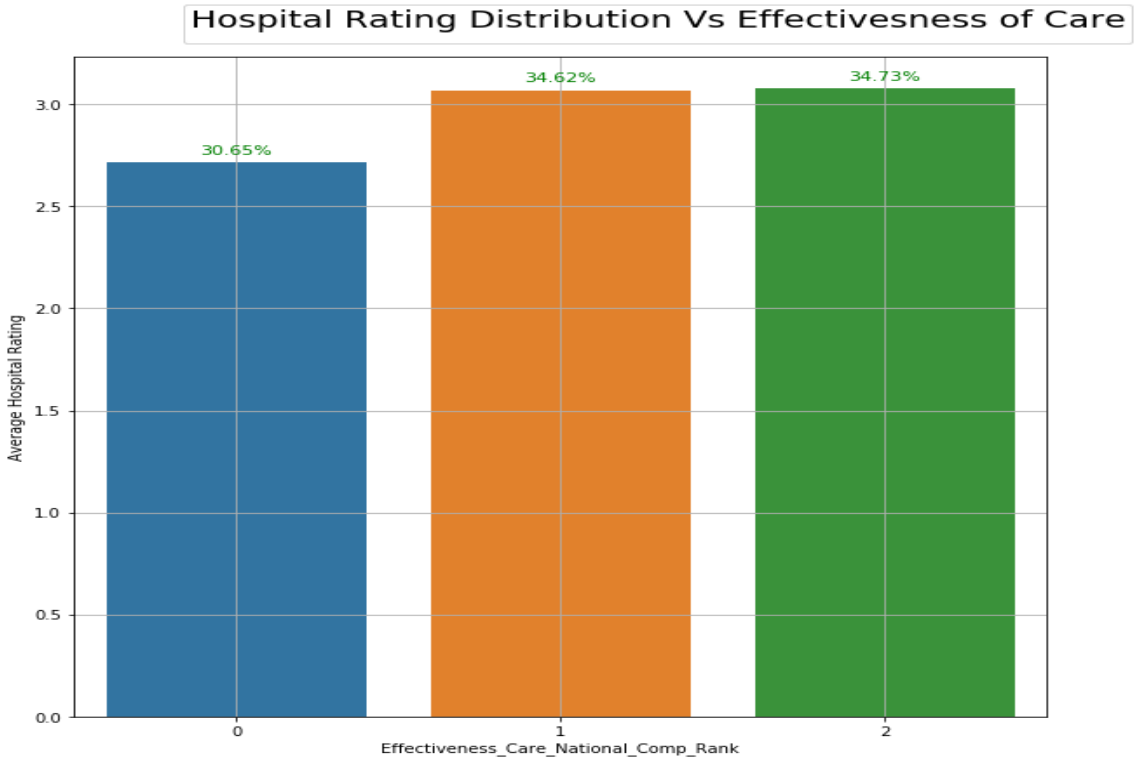




Insight: For Readmission and Mortality Group Distribution We have Higher Average Hospital Rating for Above the National Average Score.

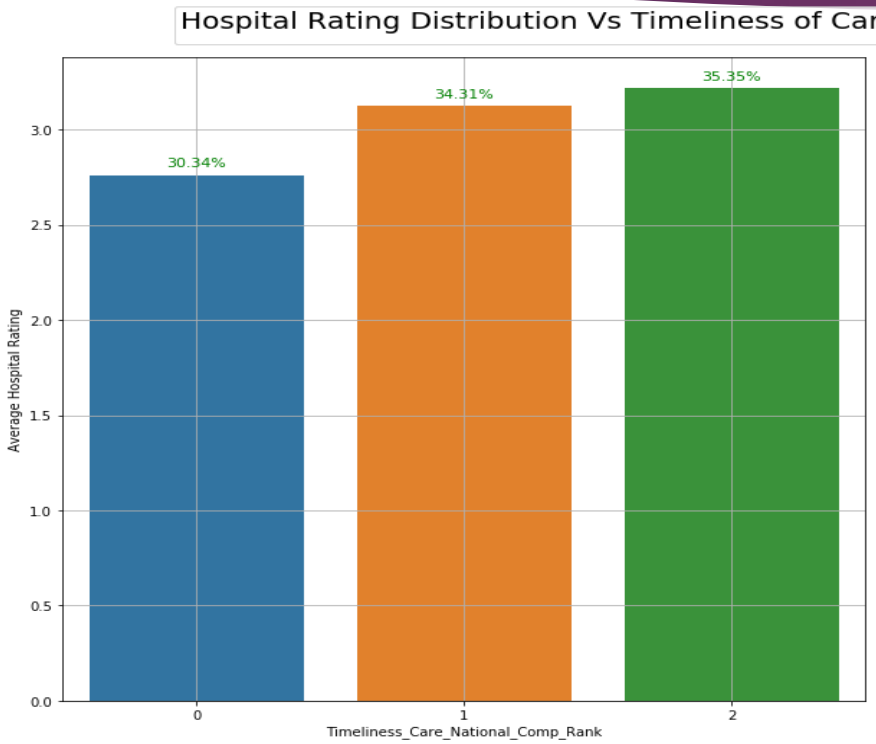
6. Patient Experience and Effectiveness of Care Distribution

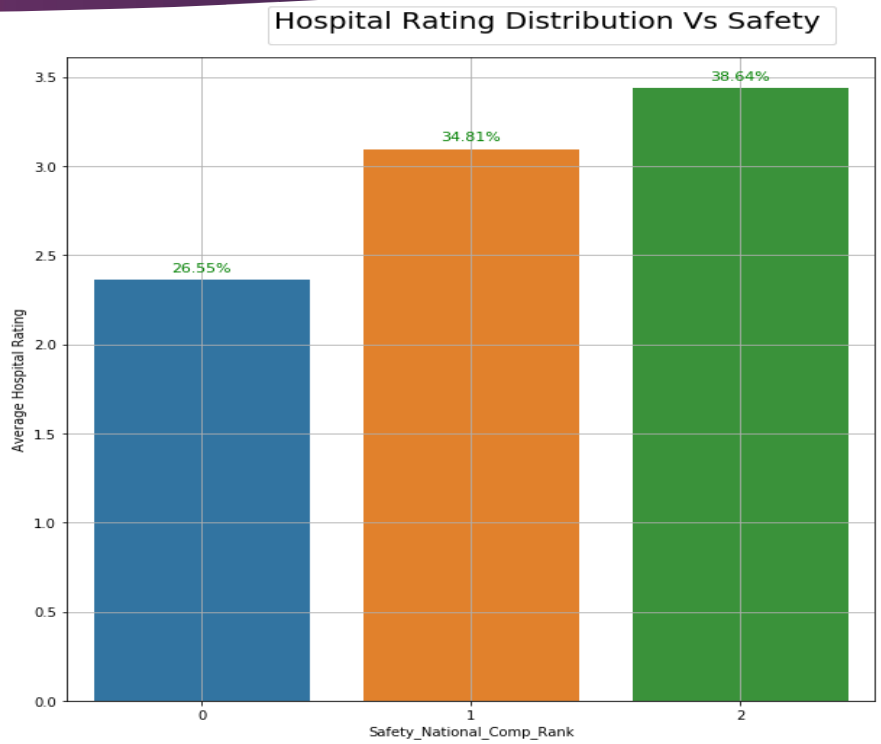




Insight: For Patient Experience and Effectiveness of Care Group Distribution We have Higher Average Hospital Rating for Above the National Average Score.

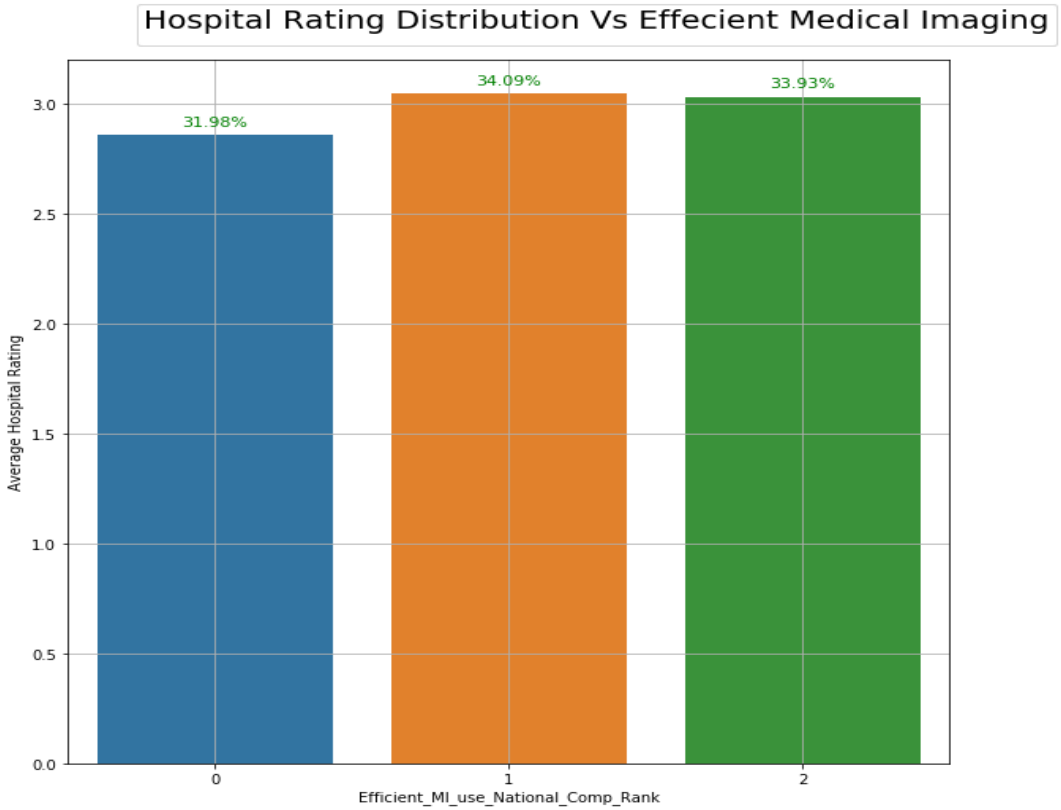
6. Safety and Timeliness of Care Distribution





Insight: For Safety and Timeliness of Care Group Distribution We have Higher Average Hospital Rating for Above the National Average Score.

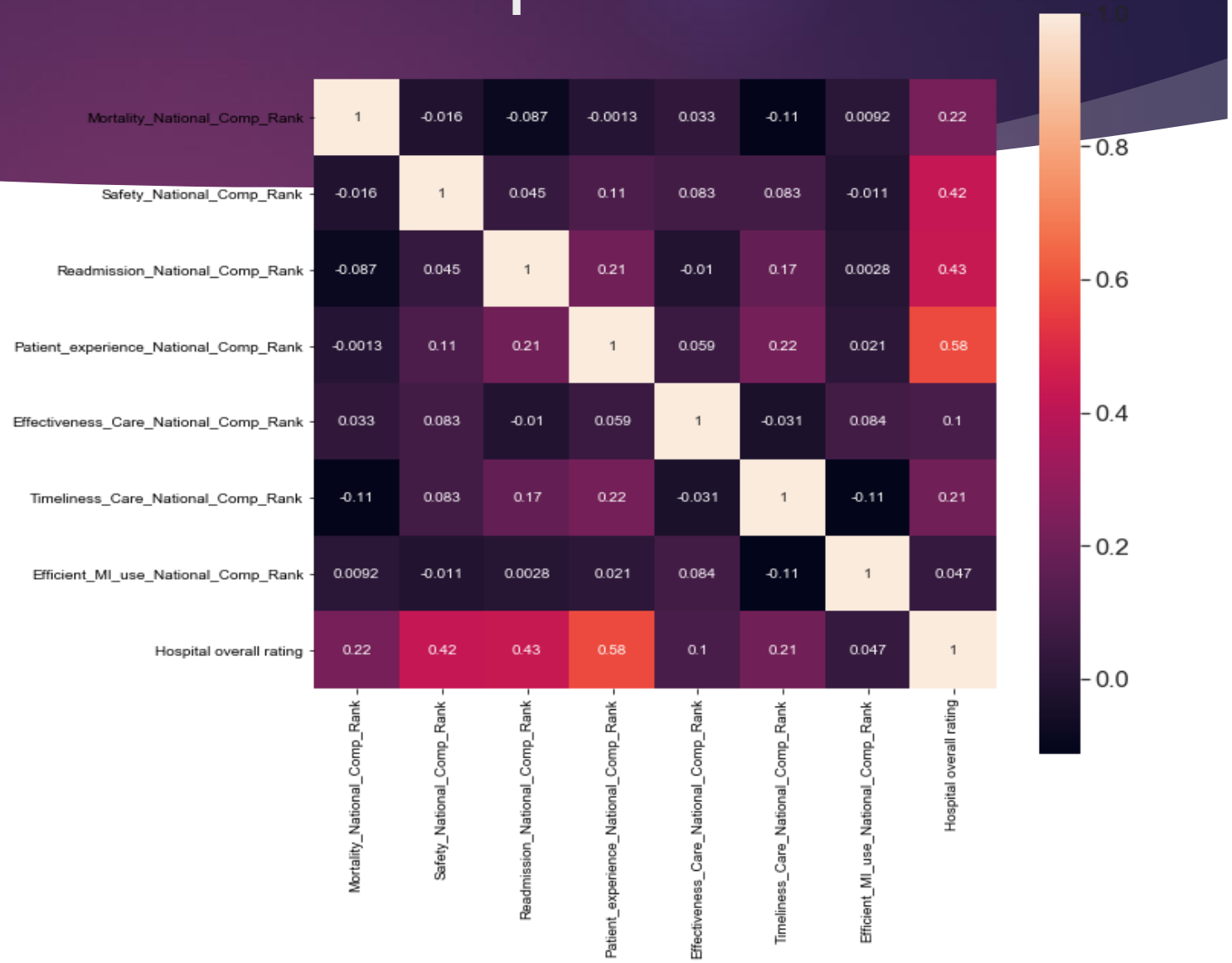
7. Effective use of Medical Imaging Distribution



Insight: For Effective use of Medical Imaging Group Distribution We have Higher Average Hospital Rating for Same as the National Average Score.

8. Correlation Matrix of Hospital General Information

Insight: This clearly shows we have a strong correlation between Hospital overall rating and Mortality, Safety , Readmission and Patient Experience groups and less correlation with Effectiveness of Care , Timeliness of Care and Efficient Use of Medical Imaging. This clearly explains why we have higher weightage for the first four groups and less weightage for other three.

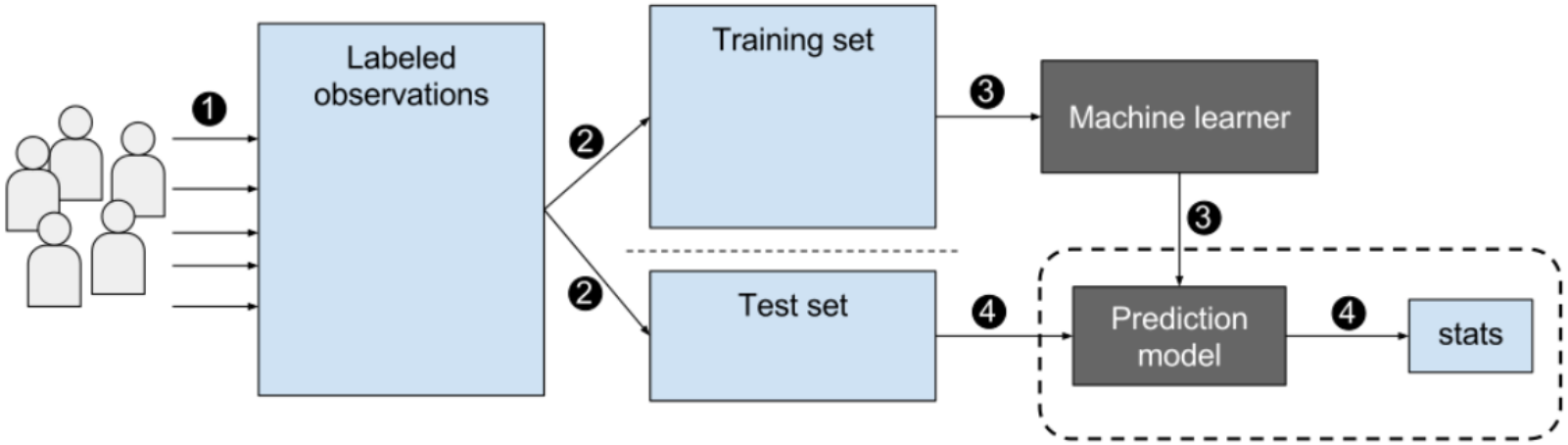


Outlier Treatment

We plotted Box plots for each Measure Groups and removed Extreme high and Low Outlies wherever possible with a maximum value more than 99.9 % of the Scores and minimum value less than 0.001% of Overall Scores.

Also after creating our final merged DataFrame , after standardization and normalization we again treated the Data for extreme outliers.

Supervised Learning



Supervised learning is useful in Classification and Regression problems where label is available for the dataset.
We can apply supervised learning with the actual hospital ratings as a labeled column in this scenario.

Supervised Learning – Methods

Supervised Learning:

- Regression
  - Linear Regression
  - Polynomial Regression.
- Decision Trees
- Random Forrest

If you look at the finalized dataset used for modeling, most of the independent variables are of numeric types, hence Linear regression is better suited than the logistic regression. Also the hospital ratings can be treated continuous for modeling purposes and derive predicted rating.

We will perform Linear Regression and Random Forrest to predict the hospital ratings and evaluate against the actual ratings provided by CMS.

Linear Regression – Model Accuracy

Linear Regression: With the Linear Regression modeling, we were able to achieve 90% accuracy.
Below are the classification report and accuracy.

Classification	Precision	Recall	F1-Score	Support
1	0.80	0.82	0.81	45
2	0.94	0.86	0.90	272
3	0.91	0.94	0.93	570
4	0.90	0.91	0.91	298
5	1.00	0.65	0.79	40
Avg/Total	0.91	0.90	0.91	1225

Linear Regression – Key Measures

Using the Model Coefficients and ELI5 library, following are the key measures driving he overall rating predictions as per our model results.

Positive Coefficients

	Features	Estimated_Coefficients
43	Patient Survey Star Rating_H_COMP_7_STAR_RATING	0.071054
37	Patient Survey Star Rating_H_COMP_1_STAR_RATING	0.070205
39	Patient Survey Star Rating_H_COMP_3_STAR_RATING	0.062002
44	Patient Survey Star Rating_H_HSP_RATING_STAR_R...	0.056954
46	Patient Survey Star Rating_H_RECMND_STAR_RATING	0.054368

Negative Coefficients

	Features	Estimated_Coefficients
28	Score_PSI_8_POST_HIP	-8.067355e-14
60	Score_STK_1	-3.889486e-05
48	Score_ED_2b	-1.203096e-04
56	Score_OP_30	-2.852027e-04
51	Score_OP_18b	-3.131915e-04

ELI5 - Weights

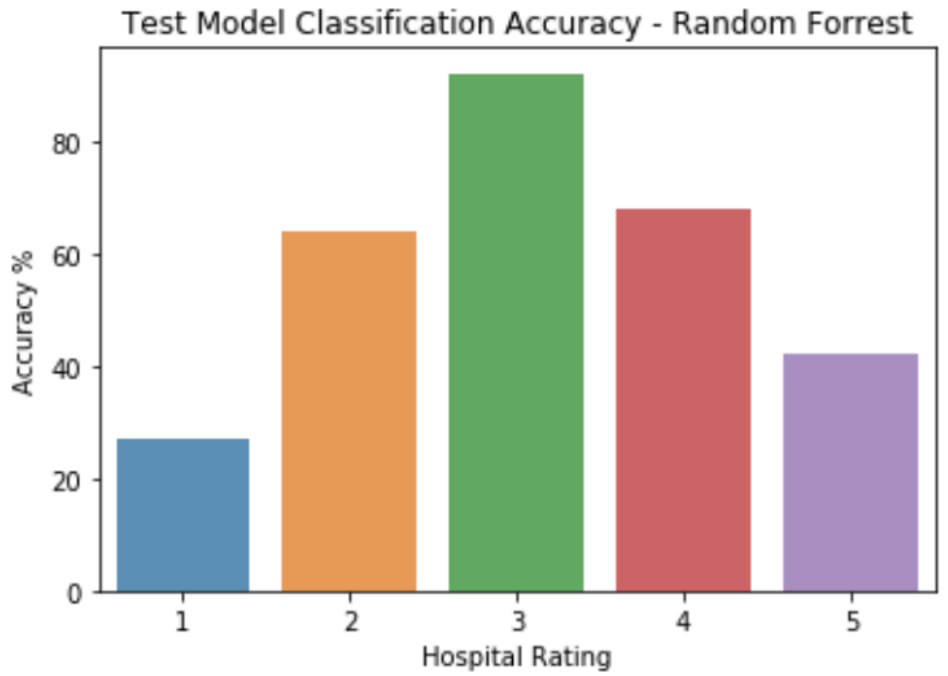
Weight[?]	Feature
+0.071	Patient Survey Star Rating_H_COMP_7_STAR_RATING
+0.070	Patient Survey Star Rating_H_COMP_1_STAR_RATING
+0.062	Patient Survey Star Rating_H_COMP_3_STAR_RATING
+0.057	Patient Survey Star Rating_H_HSP_RATING_STAR_RATING
+0.054	Patient Survey Star Rating_H_RECMND_STAR_RATING

Weight[?]	Feature
-0.000	Score_PSI_8_POST_HIP
-0.000	Score_STK_1
-0.000	Score_ED_2b
-0.000	Score_OP_30
-0.000	Score_OP_18b

Random Forrest – Model Accuracy

Random Forrest: With the Random Forrest modeling, we were able to achieve 77% accuracy, significantly lower than Linear Regression. Below are the classification report and accuracy.

Classification	Precision	Recall	F1-Score	Support
1	1.00	0.27	0.42	45
2	0.80	0.68	0.73	272
3	0.75	0.91	0.82	570
4	0.77	0.71	0.74	298
5	1.00	0.42	0.60	40
Avg/Total	0.78	0.77	0.76	1225



Un-Supervised Learning

Unsupervised learning is where we only have `input data (X)` and `no corresponding output variables`.
•The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
•These are called unsupervised learning because unlike supervised learning above there is `no correct answers` and there is no teacher.
•Algorithms are left to their own devises to discover and present the interesting structure in the data.

We have followed two method here

Method 1
PCA , followed by K Means & Hierarchical clustering

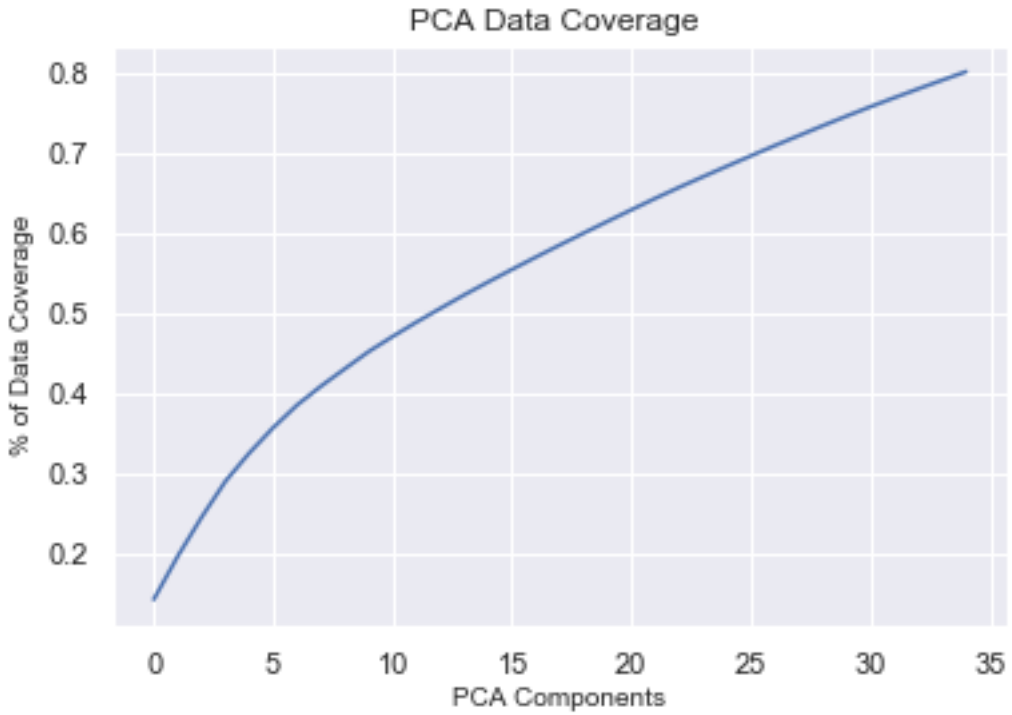
Method 2
Factor Analysis to find the weight of each measure in respective Group , Using factor analysis, find the group score for each hospital. You can then take the weighted average of group scores to calculate the final score of each hospital, perform clustering, and assign a star rating to each cluster

PCA – Principal Components Analysis

Method 1

PCA , followed by K Means & Hierarchical clustering

By Applying PCA , we do have taken 36 PCA Components to cover 81 % of our Data Set



Pro of PCA

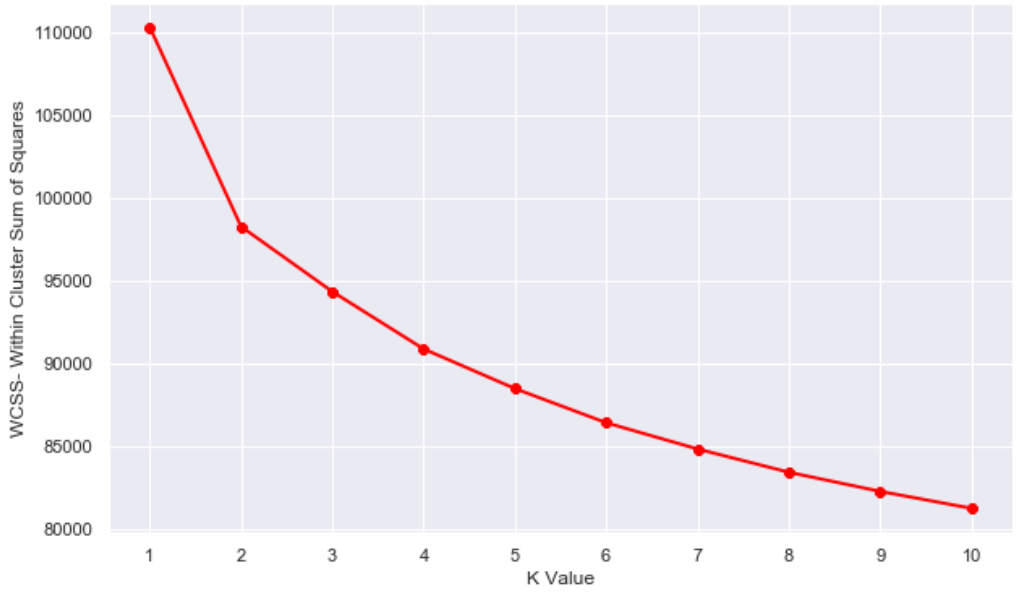
- Dimensionality Reduction can help learning . We are able to reduce of our Data set to **3061,35**
- Remove Noise
-Can deal with large datasets

Con of PCA

- hard to interpret
- sample dependent
- linear

K Means Clustering

Once we are done with our PCA , we use the same Data set for K – Means Clustering .
Using the Elbow Method to Determine the Optimal Number of Cluster.



Pro of K-Means Clustering

-Simple ,Understandable
-- Remove Noise
-items automatically assigned to clusters

Con of K – Means Clustering

-Must pick a no of clusters beforehand . With the help of Elbow Method optimum Cluster 5 is taken .
-- sensitive to noise, outlier points

We have got 67 % Accuracy by the help of k Means Clustering .

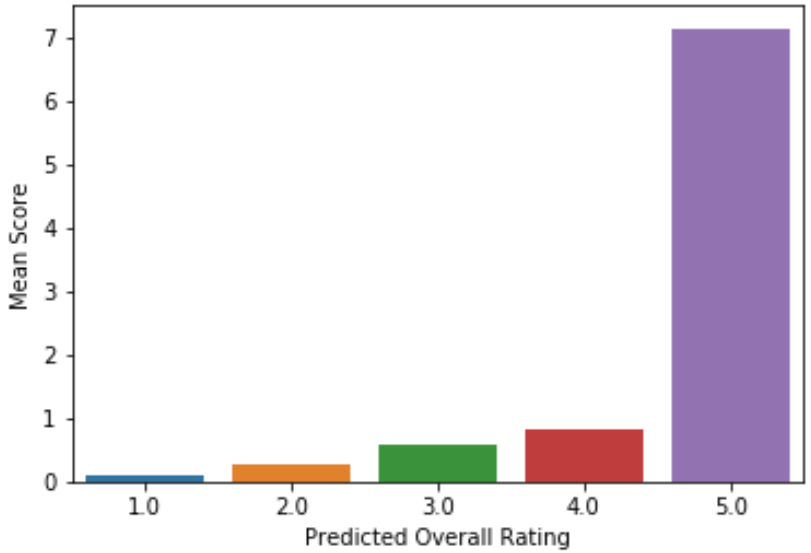
Factor Analysis

Method 2

With the help of Factor analysis , we are able to Determine the weightage of each measure ID in our 7 Groups
Final Score is calculated from group Score & the Weightage provided by CMS as below .

Example – 10001 (SOUTHEAST ALABAMA MEDICAL CENTER) has a Group score of 0.862549 with Ranking as 3

Factors #	Factor Details	Weight
1	Patient exp	22%
2	Timeliness of Care	4%
3	Effective Care	4%
4	Readmission	22%
5	Complication	4%
6	Mortality	22%
7	Safety of Care	22%



Recommendations for Hospitals

General Recommendations to Hospitals to Improving Star Rating:

- Review the impactful measures identified with the Linear Regression model - Highlight TOP 10 features that impacts the overall star rating and request the hospitals to review scores of those individual measures
- Hospitals to compare their specific measure scores and compare against the national average and focus on the areas where the measure score is lower than the National Average.
- Discuss with the domain experts & operational leaders to define processes and measure to improve the specific areas identified.

Case Study: (140010 - EVANSTON HOSPITAL):

H_COMP_7_STAR_RATING	H_COMP_1_STAR_RATING	H_COMP_3_STAR_RATING	H_HSP_STAR_RATING	H_RECMND_STAR_RATING
2	3	3	4	4
SCORE_PSI_8_POST_HIP	**SCORE_STK_1**	**SCORE_ED_2B**	**SCORE_OP_30**	**SCORE_OP_18B**
0.06	98	76	94	166

- H_Comp_7_Star_Rating is lower than the average of 3.2, the hospital need to work on getting the patients to strongly agree with the care they received when they leave the hospital. Likewise Comp_1 and _3 start rating are lower than the average of 4 and 3.8. This average is based on the other hospitals with overall rating of 4 or 5.

- Timely Effective Care measures are negatively impacting the hospital rating. This hospital needs to focus on the following:
  - Postoperative Hip Fracture Rate