Comparing Hospitals of USA

Final Report

Munazza Khan
Student of Data Analytics, Big Data, and Predictive Analytics
Chang School, Ryerson University, Canada

Supervisor
Dr.Uzair Ahmed

April 10, 2017

Contents

Overview

This research investigates quality of healthcare services in USA by using hospital compare dataset and applying different machine learning techniques over it.

Introduction

The objective of this project is to analyze "Hospital Compare Downloadable Dataset –USA" by using data mining and knowledge discovery techniques .The purpose is to identify difference in quality of healthcare services in different states of US in 2016. There will be two criteria for comparison

1. Comparing Hospitals in term of number of hospitals provided within a State
2. Comparing Hospitals in term of Hospital's Performance within a State

The following factors were taken into consideration when considering the Performance of a Hospital within a State:

- Emergency Services
- Mortality and Re-admissions
- Outpatient Imaging Efficiency
- Meaningful use of EHRs
- Safety and Effectiveness of care
- Patient experience
- Timeliness of care
- Ratings

Tool used during the process: RStudio

Literature Review

The six major government health care programs in USA are Medicare, Medicaid, the State Children's Health Insurance Program (SCHIP), the Department of Defense TRICARE and TRICARE for Life programs (DOD TRICARE), the Veterans Health Administration (VHA) program, and the Indian Health Service (IHS) program—provide health care services to about one-third of Americans.

The database used for this project covers Medicare and Medicaid services.

Medicare provides health insurance to all individuals eligible for social security who are aged 65 and over, those eligible for social security because of a disability, and those suffering from end-stage renal disease (ESRD). While Medicaid serves about 42 million people who are poor and who require health care services to achieve healthy growth and development goals or meet special health care needs. The program covers low-income people who meet its eligibility criteria, such as children, pregnant women, certain low-income parents, disabled adults, federal Supplemental Security Income (SSI) recipients (low-income children and adults with severe disability), and the medically needy (non-poor individuals with extraordinary medical expenditures who meet spend-down requirements generally for long-term care). There is a good deal of variability across states in the maximum income for eligible.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process. Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

Datasets

Five datasets are considered for analysis in this project; Hospital General Information and four other datasets taken from R-packages.

Primary Dataset

1. Hospital General Information

Hospital General Information dataset has be taken from Hospital Compare Database. Hospital Compare is a consumer-oriented website that provides information on the quality of care hospitals are providing to their patients. This information can help consumers make informed decisions about health care. The Centers for Medicare & Medicaid Services (CMS) created the Hospital Compare website to better inform health care consumers about a hospital's quality of care. Hospital Compare provides data on over 4,000 Medicare-certified hospitals, including acute care hospitals, critical access hospitals (CAHs), children's hospitals, and hospital outpatient departments. Hospital Compare is typically updated, or refreshed, each quarter in April, July, October, and December.
The data has been collected from the link given below
https://data.medicare.gov/data/hospital-compare
The Database is composed of 56 files, out of which 1 file have been selected for the purpose of this project. The description of data set is given below:

Hospital General Information- 4807 obs. of 28 variables							
S.no	**Attributes**	**Type**	**Descriptive Statistics**				**Used During Analysis- Yes/No**
			Observations	**Missing Values**	**Distinct Values**	**Distinct Values- Description**	
1	***Provider ID***	Qualitative-Nominal	4807	0	4807	Hospitals ID's, i.e.010001	Yes
2	***Hospital Name***	Qualitative-Nominal	4807	0	4608	Name of Hospital	Yes
3	***Address***	Qualitative-Nominal	Not Used				No
4	***City***	Qualitative-Nominal	4807	0	2947	Name of Cities i.e. New York	Yes
5	***State***	Qualitative-Nominal	4807	0	56	Names of States i.e. NY	Yes
6	***Zip Code***	Qualitative-Nominal	4807	0	4807	Zip codes of Hospitals	Yes
7	***County Name***	Qualitative-Nominal	Not Used				No
8	***Phone Number***	Qualitative-Nominal	Not Used				No
9	***Hospital Type***	Qualitative-Nominal	4807	0	3	Acute Care Hospitals Children's Critical Access Hospitals	Yes

10	***Hospital Ownership***	Qualitative-Nominal	4807	0	10	Government - Federal Government - Hospital District or Authority Government - Local Government - State Physician Proprietary Tribal Voluntary non-profit - Church Voluntary non-profit - Other Voluntary non-profit - Private	Yes
11	***Emergency Services***	Qualitative-Nominal	4807	0	2	Y|N	Yes
12	***Meet Criteria for meaningful use of EHRs***	Qualitative-Nominal	4373	434	1	Above National Average, Same as National Average, Below National Average	Yes
13	***Hospital overall Rating***	Quantitative-Distrete	3584	1223	5	1,2,3,4,5	Yes
14	***Hospital overall Rating footnote***	Qualitative-Nominal	Not Used				No
15	***Safety of care National comparision***	Qualitative-Ordinal	2654	2153	3	Above National Average, Same as National Average, Below National Average	Yes
16	***Safety of care National comparision footnote***	Qualitative-Nominal	Not Used				No
17	***Readmission national comparison***	Qualitative-Ordinal	3813	994	3	Above National Average, Same as National Average, Below National Average	Yes
18	***Readmission national comparison footnote***	Qualitative-Nominal	Not Used				No
19	***Patient experience National comparison***	Qualitative-Ordinal	3454	1353	3	Above National Average, Same as National Average, Below National Average	Yes
20	***Patient experience National comparison footnote***	Qualitative-Nominal	Not Used				No
21	***Effectiveness of care National comparison***	Qualitative-Ordinal	2790	2017	3	Above National Average, Same as National Average, Below National Average	Yes
22	***Effectiveness of care National comparison footnote***	Qualitative-Nominal	Not Used				No

23	***Timeliness of care National comparison***	Qualitative-Ordinal	3565	1242	3	Above National Average, Same as National Average, Below National Average	Yes
24	***Timeliness of care National comparison footnote***	Qualitative-Nominal	Not Used				No

Source :https://data.medicare.gov/data/hospital-compare

Secondary Datasets

2. us_states

The dataset is a part of R-Package USAboundaries. It's a Spatial Polygon Data Frame.

US_States- 9 Variables 52 Observations							
S.no	**Attributes**	**Type**	**Descriptive Statistics**				**Used During Analysis- Yes/No**
			Observations	**Missing Values**	**Distinct**	**Distinct Values- Description**	
1	***statefp***	Qualitative-Nominal	Not Used				No
2	***statens***	Qualitative-Ordinal	Not Used				No
3	***affgeoid***	Quantitative-Discrete	Not Used				No
4	***geoid***	Quantitative-Discrete	52	0	52	geoids of states	Yes
5	***stusps***	Qualitative-Ordinal	52	0	52	State names	Yes
6	***name***	Qualitative-Nominal	Not Used				No
7	***lsad***	Qualitative-Nominal	Not Used				No
8	***aland***	Quantitative-Discrete	52	0	56	land area of states	Yes
9	***awater***	Quantitative-Discrete	52	0	56	water area of states	Yes

Source:https://cran.r-project.org/web/packages/USAboundaries/USAboundaries.pdf

3. zip_codes

The dataset is a part of R-package noncensus. This data set considers each zip code throughout the U.S. and provides additional information, including the city and state, latitude and longitude, and the FIPS code for the corresponding county.

zip_codes-4 Variables 43524 Observations							
S.no	**Attributes**	**Type**	**Descriptive Statistics**				**Used During Analysis- Yes/No**
			Observations	**Missing Values**	**Distinct**	**Distinct Values- Description**	
1	***zip***	Qualitative-Nominal	43524	0	56	zip codes ,i.e. 90505	Yes
2	***latitude***	Qualitative-Nominal	43524	0	37890	latitude related to zip codes	Yes
3	***longitude***	Qualitative-Nominal	43524	0	37789	longitude related to zip codes	Yes
4	***fips***	Qualitative-Nominal	43524	0	3218		Yes

Source:https://cran.r-project.org/web/packages/noncensus/noncensus.pdf

4. u.s.cities

The dataset is a part of R-package maps.

u.s.cities- 6 Variables 1005 Observations							
S.no	Attributes	Type	Descriptive Statistics				Used During Analysis- Yes/No
			Observations	Missing Values	Distinct	Distinct Values- Description	
1	***name***	Qualitative-Nominal	1005	0	1005	city name i.e. Abilene TX	Yes
2	***country.etc***	Qualitative-Ordinal	Not Used				No
3	***pop***	Quantitative-Discrete	Not Used				No
4	***lat***	Quantitative-Discrete	Not Used				No
5	***lon***	Qualitative-Ordinal	Not Used				No
6	***capital***	Quantitative-Nominal	1005	0	2	0|2(0=No,2=Yes)	Yes

Source:http://svitsrv25.epfl.ch/R-doc/library/maps/html/us.cities.html

5. population

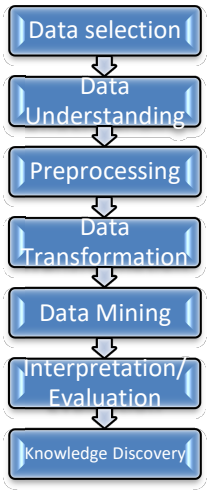
The dataset is a part of R-package ggcounty.

population- 2 Variables 3221 Observations							
S.no	Attributes	Type	Descriptive Statistics				Used During Analysis- Yes/No
			Observations	Missing Values	Distinct	Distinct Values- Description	
1	***FIPS***	Qualitative-Nominal	3221	0	56	federal information processing standards to locate geographic locations	Yes
2	***count***	Qualitative-Discrete	3221	0	56	population of cities	Yes

Source:devtools::install_github("hrbrmstr/ggcounty"

Approach

The project will follow the sequence given below



Step 1: Data Selection

The Hospital Compare Database is composed of 56 datasets, out of which one dataset named "*Hospital General Information* "has been selected for the purpose of this project. This dataset is the summary of all types of comparisons done within the database. There are 28 variables out of which 15 variables have been selected with 4806 observations.

In order to compare hospitals in term of how the hospitals are distributed over states and cities dataset from R-Packages are used to add more dimension to the dataset.

Step 2: Data Understanding

After selecting the dataset developing an understanding of data is very important by using following statistical description techniques such as data dispersion, central tendency measures, data visualization, data summarization. For data preprocessing to be successful, it is essential to have an overall picture of data. Basic statistical descriptions will be used to identify properties of the data and highlight which data values should be treated as noise or outliers. This step covers three areas of basic statistical descriptions. We will start with measures of central tendency, which measure the location of the middle or center of a data distribution. The mean, median, mode, and midrange are calculated to find out the location of the middle or center of a data distribution. In addition to assessing the central tendency of our data set, we also would like to have an idea of the dispersion of the data. That is, how are the data spread out? The dispersion measures are the range, quartiles, and inter-quartile range, box-plots, variance and standard deviation of the data.

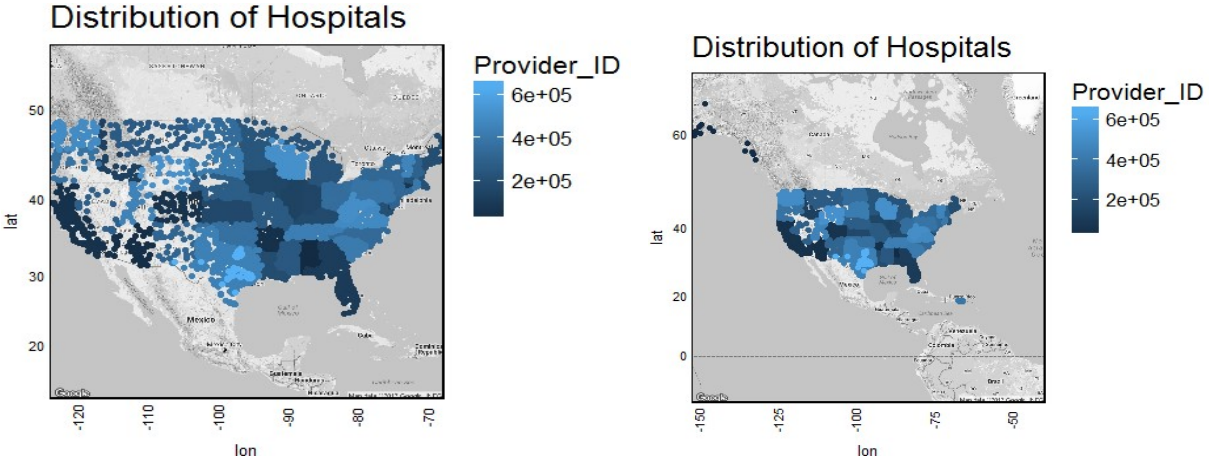
These measures are useful for identifying outliers. Finally, we can use many graphic displays of basic statistical descriptions to visually inspect our data by using histograms, scatter plots, and quintile plots.

Univariate Analysis

We will discuss each attribute here:

1. Provider_ID

This variable contains ID's of hospitals. There are 4806 hospitals in all.



2. *Hospital_Name*

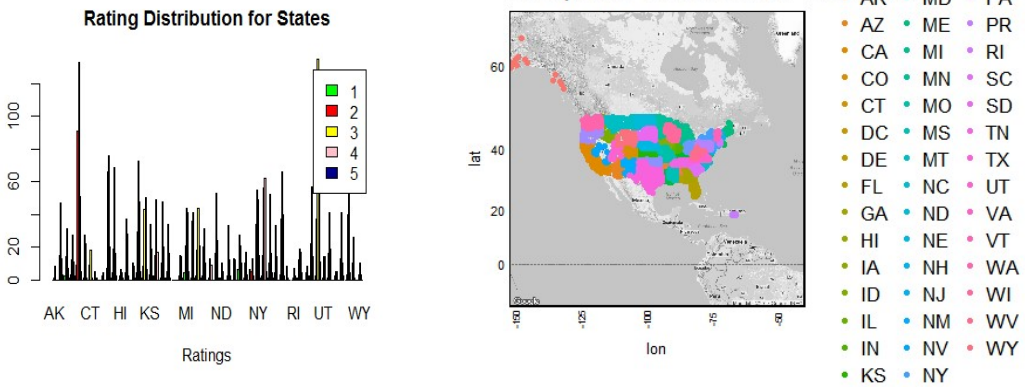
This variable contains names of hospitals. There are 4806 hospitals in all.

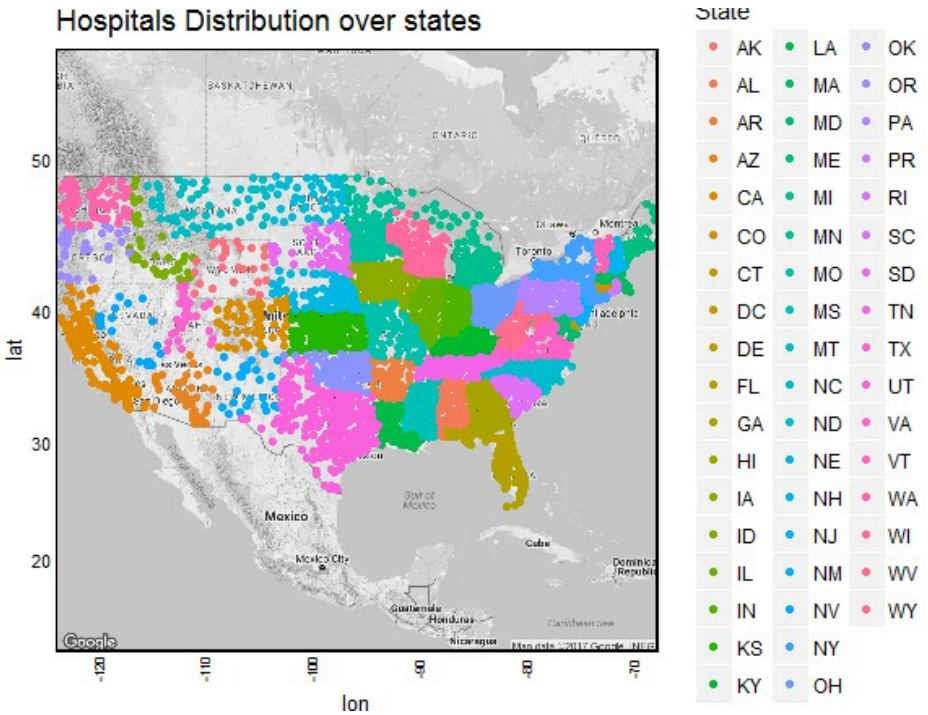
3. *City*

This variable contains names of cities where hospitals are located. There are 2947 cities in all.

4. *State*

This variable contains names of states where hospitals are located. There are 56 states.





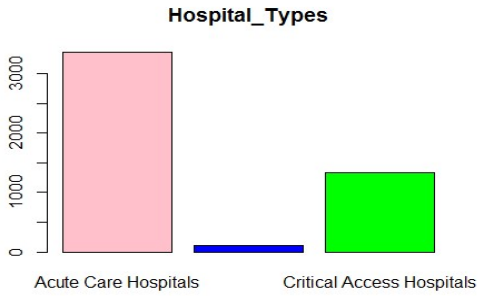
5. *zip*

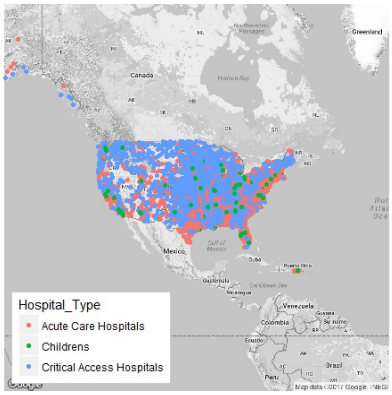
This variable contains zip codes

6. *Hospital_Type*

This variable contains categorical data. There are three categories:

- Acute Care Hospitals –with 3360 occurrences
- Children's- with 99 occurrences
- Critical Access Hospitals- with 1338 occurrences

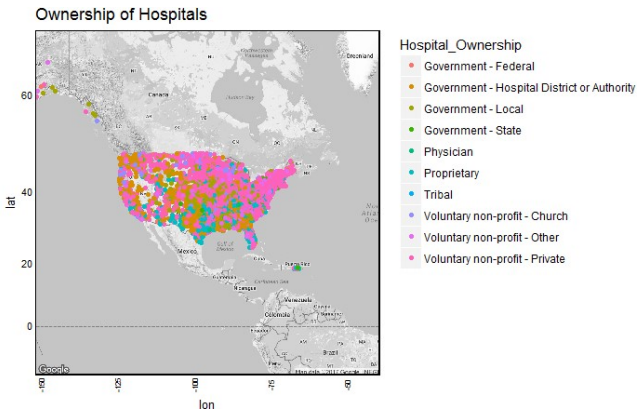




7. *Hospital_Ownership*

This variable contains categorical data. There are 10 categories

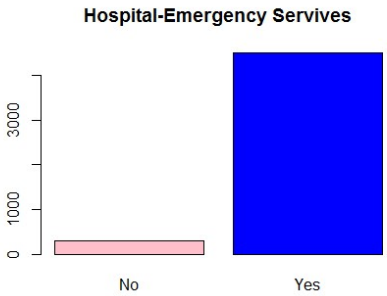
- Government – Federal- 46
- Government - Hospital District or Authority - 564
- Government – Local- 404
- Government – State- 66
- Physician- 64
- Proprietary - 795
- Tribal - 8
- Voluntary non-profit – Church- 343
- Voluntary non-profit – Other- 473
- Voluntary non-profit – Private- 2034



8. *Emergency_Services*

This variable contains factorial data. There are two categories.

- Yes - 4502 hospitals have emergency services
- No – 295 hospital don't have emergency services



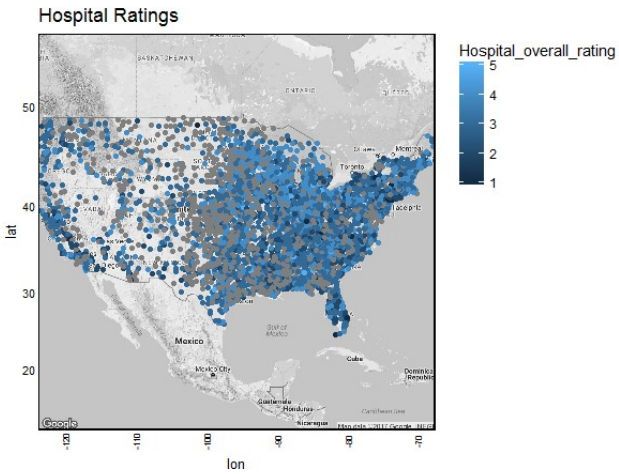
9. *Meets_criteria_for_meaningful_use_of_EHRs*

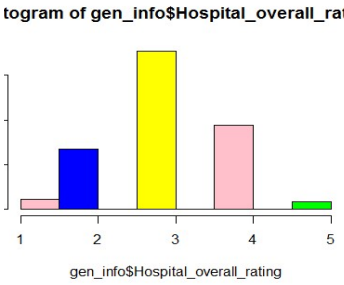
This variable has only one value i.e. yes. Which means that all hospitals meets criteria for meaningful use of HER.

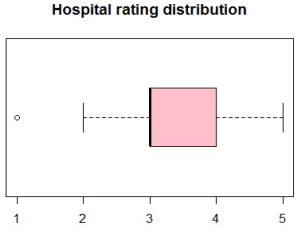
10. *Hospital_overall_rating*

This variable has ordinal data. The hospital overall ratings show the quality of care a hospital may provide compared to other hospitals based on the quality measures reported on Hospital Compare. The hospital overall ratings summarize more than 60 measures reported on Hospital Compare into a single rating .The hospitals can receive between one and five stars, with five stars being the highest ranking, and the more stars, the better the hospital performs on the quality measures. Most hospitals will display a three star rating.

1	2	3	4	5
108	677	1772	941	81



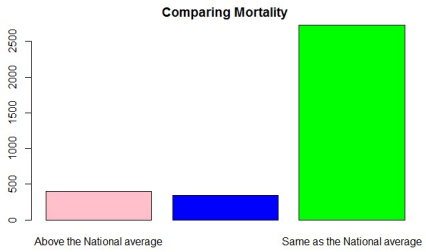


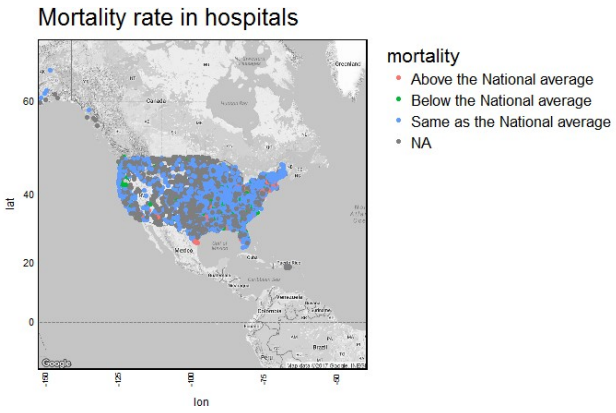


11. Mortality

This variable has categorical data with three categories

- Below average -400
- Same as Average-340
- Above Average-2731

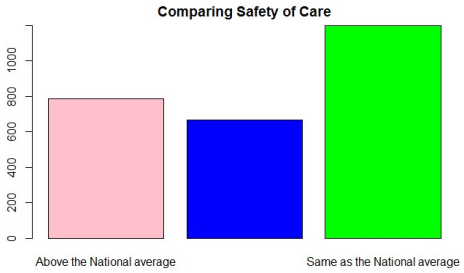


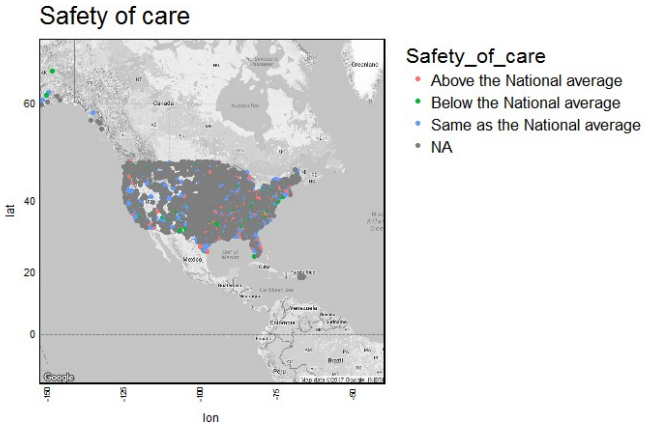


12. Safety_of_care

This variable has categorical data with three categories

- Below average-
- Same as Average
- Above Average

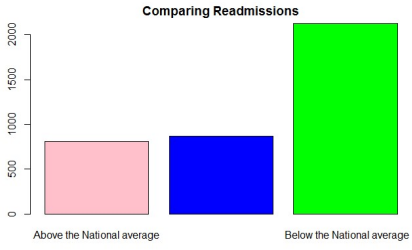


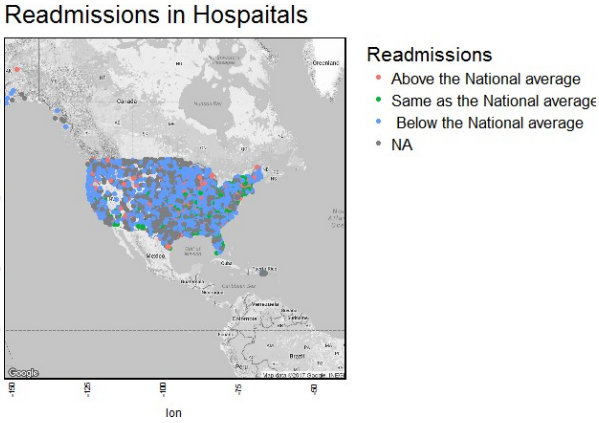


Readmissions

This variable has categorical data with three categories. The unplanned hospital-wide readmission measure focuses on whether patients who were discharged from a hospitalization were hospitalized again within 30 days.

- Below average
- Same as Average
- Above Average

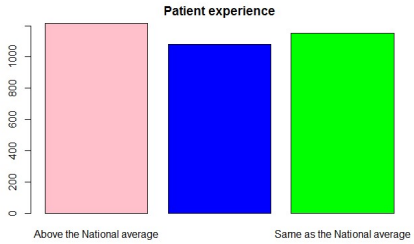




13. Patient_experience

This variable has categorical data with three categories

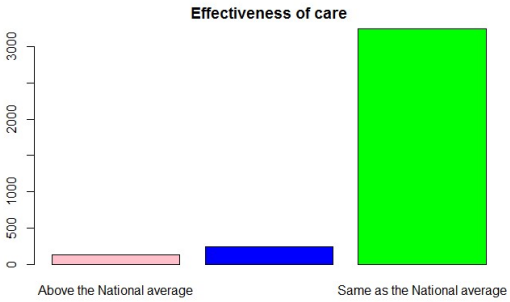
- Below average
- Same as Average
- Above Average

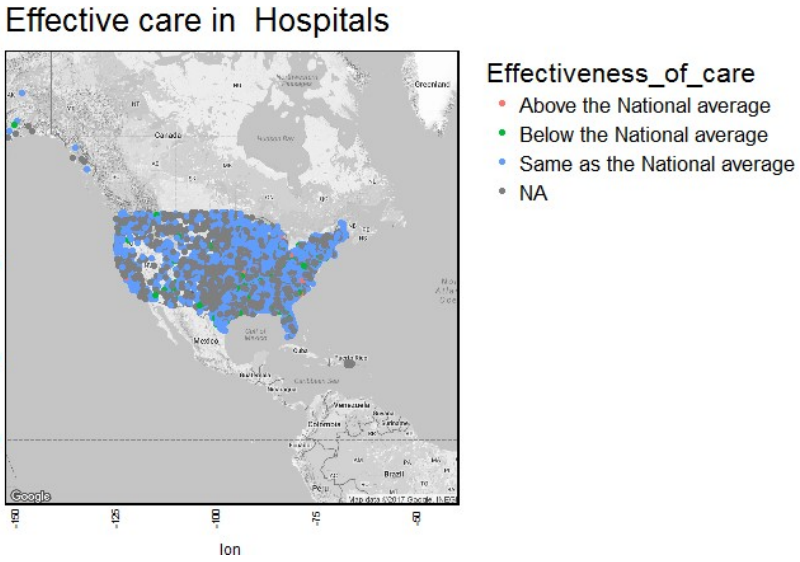


14. Effectiveness_of_care

This variable has categorical data with three categories

- Below average
- Same as Average
- Above Average



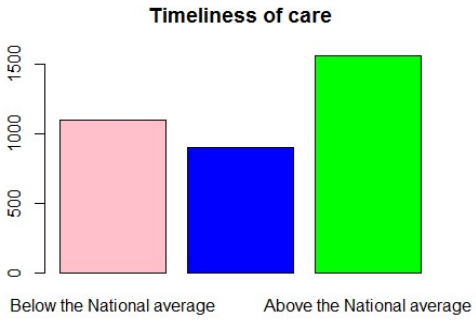


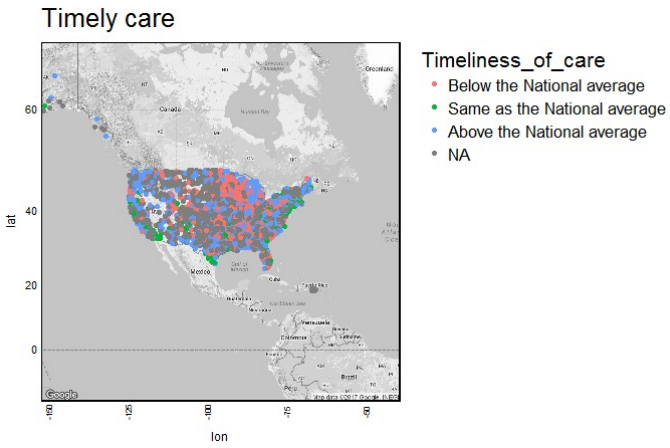
15. Timeliness_of_care

This variable has categorical data with three categories

- Below average
- Same as Average
- Above Average

Shows how quickly hospitals treat patients who come to the hospital with certain medical emergencies; and how well hospitals provide preventive services.

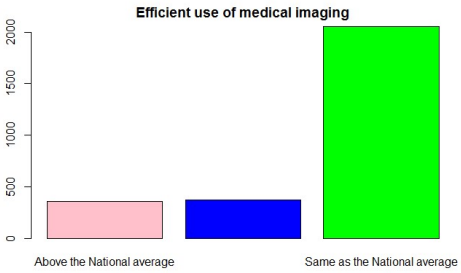


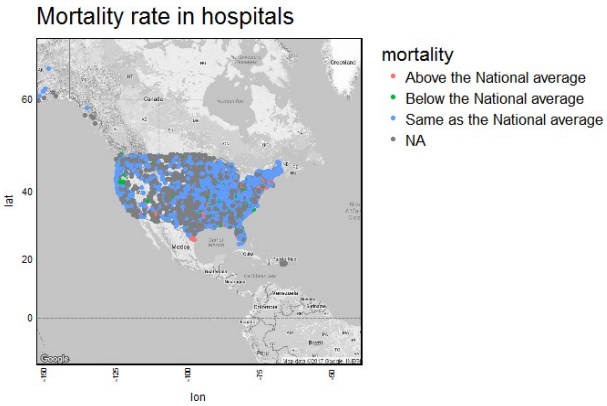


16. Efficient_use_of_medical_imaging

This variable has categorical data with three categories

- Below average
- Same as Average
- Above Average





17. latitude

This variable has discrete continuous data about the location of the hospital.

18. longitude

This variable has discrete continuous data about the location of the hospital.

19. fips

This variable has discrete continuous data about the location of the hospital. The FIPS county code is a five-digit Federal Information Processing Standard (FIPS) code (FIPS 6-4) which uniquely identifies

counties and county equivalents in the United States, certain U.S. possessions, and certain freely associated states.

20. population

This variable contains continuous data about the population of each city

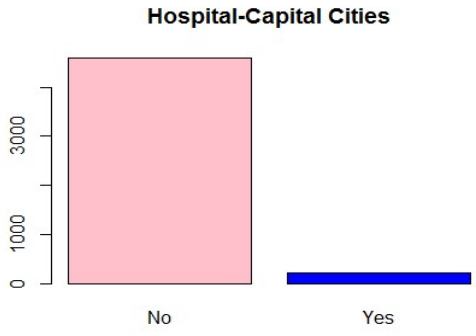
21. name

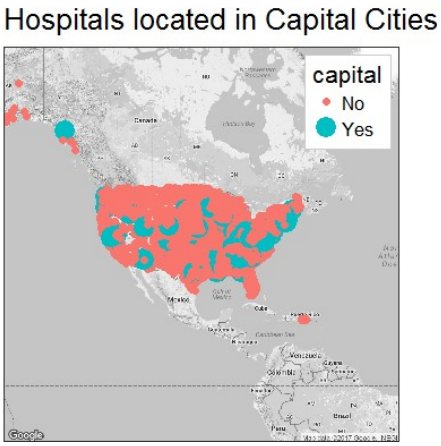
This variable contains city and state information as some states has same city names. eg:
`batavia ny` `batavia oh`

22. capital

This variable contains factorial data. There are two categories.

- 0 - Means not capital cities. There are 4591 hospitals which are not located in capital cities
- 2 -Means capital cities. There are 206 hospitals which are located in capital cities





23. geoid

This variable has geoids' of all states. There are 52 states with 52 geoids

24. land_area

This variable has area of all cities.

Bivariate Analysis

Chi Squared Test

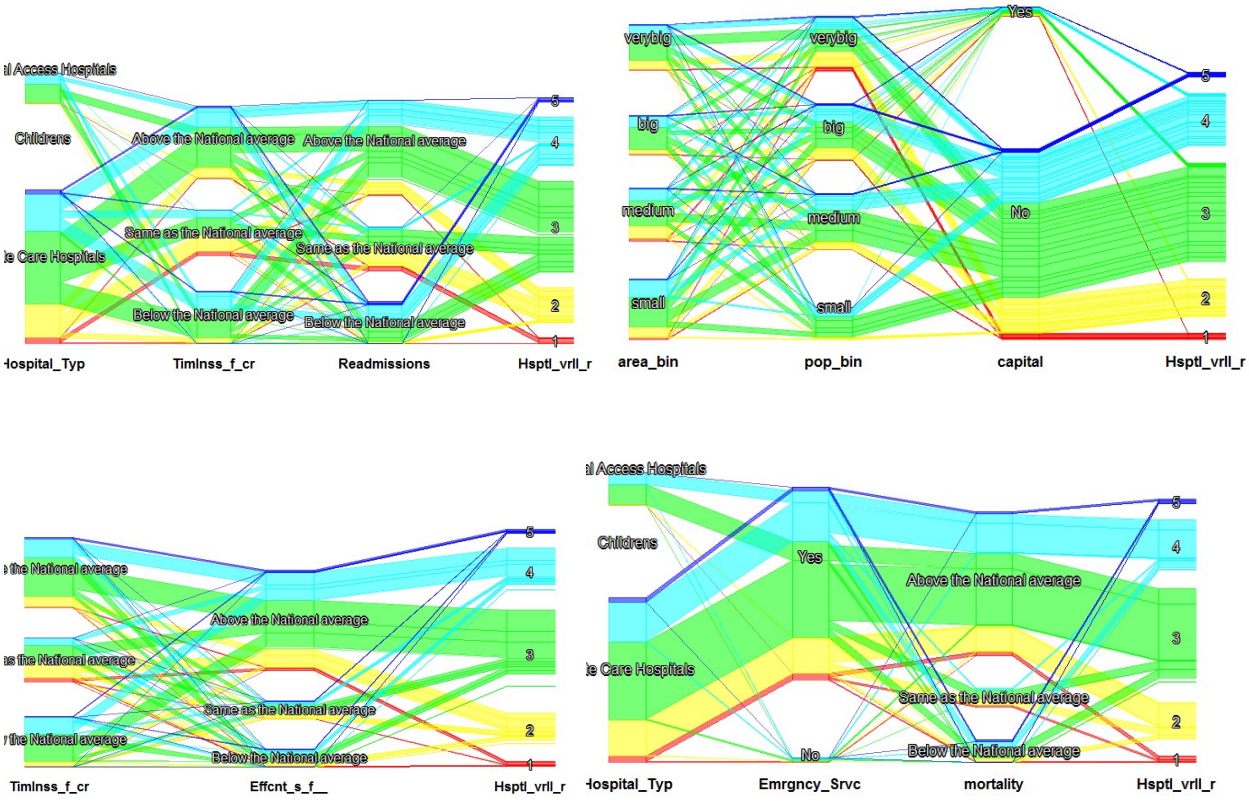
For bivariate analysis I have done Chi Squared Test for all variables against dependent variable overall rating. And the result shows that the p-value for all the variables is less than significance level (.05) which means that the null hypothesis that they are independent can be rejected. All variables are dependent on ratings

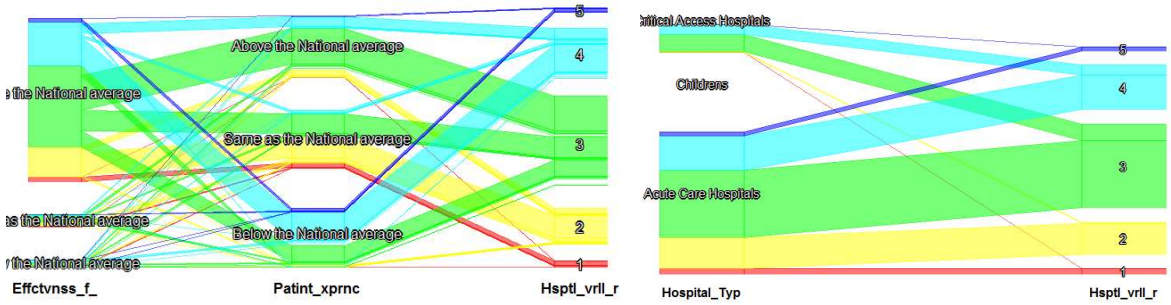
Variables	p-value	Chi-Squared Test Results
Hospital_Type	2.20E-16	p-value < 0.05 (dependent)
Hospital_Ownership	2.20E-16	p-value < 0.05 (dependent)
Emergency_Services	2.81E-05	p-value < 0.05 (dependent)
Mortality	2.81E-05	p-value < 0.05 (dependent)
Meets_criteria_for_meaningful_use_of_EHRs	2.20E-16	p-value < 0.05 (dependent)
Readmissions	2.20E-16	p-value < 0.05 (dependent)
Patient_experience	2.20E-16	p-value < 0.05 (dependent)
Effectiveness_of_care	2.20E-16	p-value < 0.05 (dependent)
Timeliness_of_care	2.20E-16	p-value < 0.05 (dependent)
Efficient_use_of_medical_imaging	1.72E-05	p-value < 0.05 (dependent)
gen_info$capital	0.00036	p-value < 0.05 (dependent)

Multivariate Analysis

Parallel Coordinates Analysis

Parallel coordinates are used to visualize the relationship between multiple variables.





Correlation-

	population	land_area	Hospital_overall_rating
population	1.00000000	0.19974438	-0.08545203
land_area	0.19974438	1.00000000	0.02075058
Hospital_overall_rating	-0.08545203	0.02075058	1.00000000

Step 3: Preprocessing

In the Data cleaning and preprocessing the main focus will be.

Attribute subset selection

Selecting subset of attributes by reducing the data set size by removing irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.
Meets criteria for meaningful use of EHR will be removed because there is no variance.

Removal of noise or outliers.

During parallel coordinate analysis it is diagnosed that there is no information available about the performance of Children's hospital. All the rows containing information about children hospital will be removed.

	Acute Care Hospitals	Childrens	Critical Access Hospitals	<NA>
1	107	0	1	0
2	653	0	24	0
3	1417	0	355	0
4	726	0	215	0
5	79	0	2	0
<NA>	378	99	741	0

```
> table(gen_info$Hospital_overall_rating,gen_info$State,useNA = "always")
```

	AK	AL	AR	AZ	CA	CO	CT	DC	DE	FL	GA	HI	IA	ID	IL	IN	KS	KY	LA	MA	MD	ME	MI	MN
1	0	0	2	3	9	0	0	5	0	7	4	0	0	0	4	1	0	2	1	1	0	0	4	0
2	1	15	14	12	91	1	9	1	0	66	19	2	4	0	29	8	3	15	10	5	0	2	21	1
3	8	47	31	27	133	27	18	1	2	76	69	6	37	10	73	43	34	49	48	34	0	15	44	36
4	0	13	7	11	51	22	1	0	4	20	16	4	28	8	48	50	19	17	20	16	0	14	41	41
5	0	3	1	2	5	1	0	0	0	1	2	1	2	1	5	6	2	0	1	1	0	1	5	1
<NA>	13	10	19	26	51	27	3	1	1	16	24	10	45	22	21	12	79	11	42	7	49	1	16	52

Finding out strategies for handling missing data fields.

1. Rating Imputation: Ratings will be imputed and predicted during data mining phase.
2. Rows containing missing values have assigned mode for their respective variables.

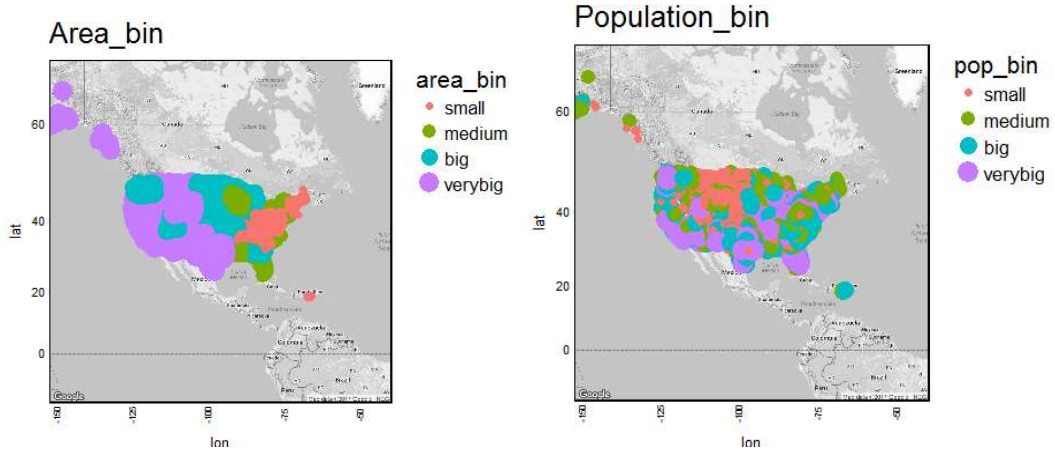
```
gen_info$Safety_of_care[is.na(gen_info$Safety_of_care)]="Same as the National average"

gen_info$mortality[is.na(gen_info$mortality)]="Same as the National average"

gen_info$Readmissions[is.na(gen_info$Readmissions)]="Same as the National average"

gen_info$Patient_experience[is.na(gen_info$Patient_experience)]="Above the National average"

gen_info$Effectiveness_of_care[is.na(gen_info$Effectiveness_of_care)]="Same as the National average"

gen_info$Timeliness_of_care[is.na(gen_info$Timeliness_of_care)]="Same as the National average"

gen_info$Efficient_use_of_medical_imaging[is.na(gen_info$Efficient_use_of_medical_imaging)]="Same as the
National average"
```

Step 4: Data Transformation

In data transformation, the data is transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

1. Smoothing works to remove noise from the data. Techniques include binning, regression, and clustering.

Attributes population and area of cities are binned into four classes.

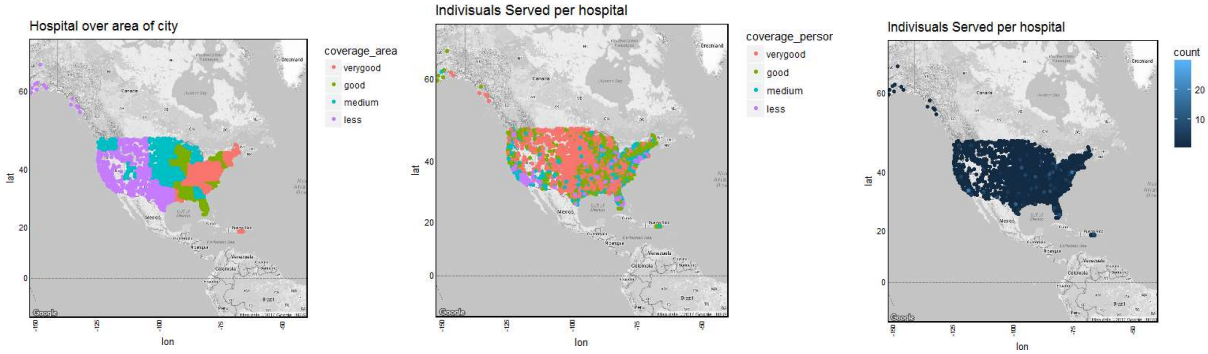


During parallel coordinate analysis it is determined that there is not enough information available for the variable Hospital Ownership for Tribal, Government and Physician. The rows containing these values are removed from dataset .

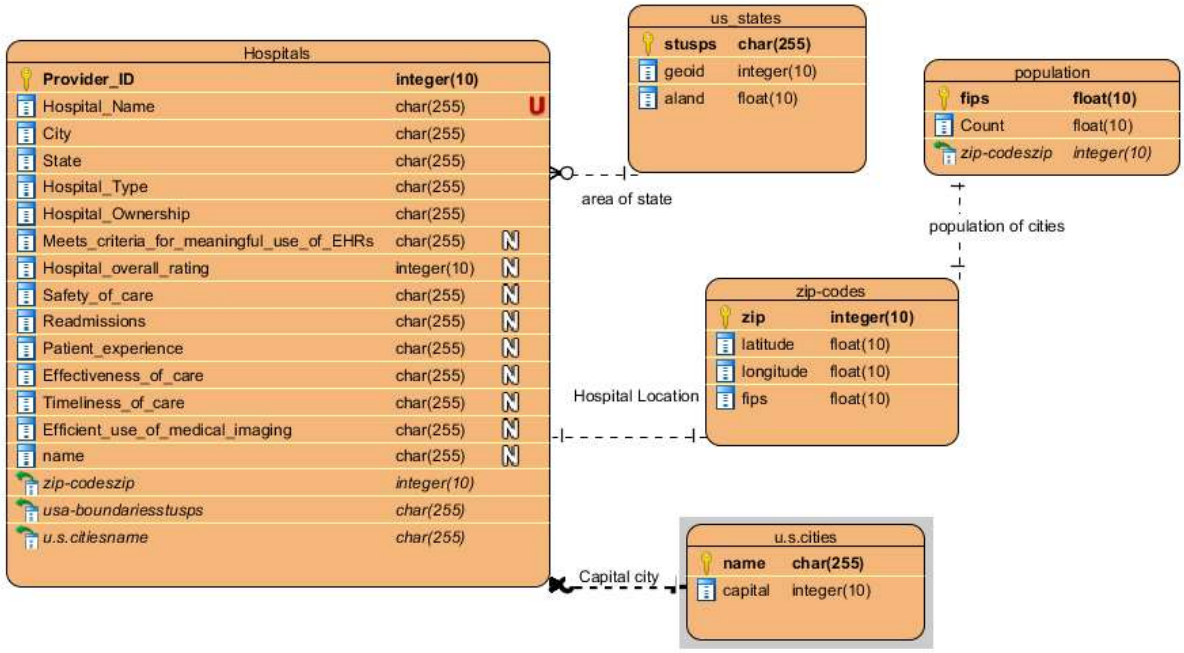
2. Attribute construction

New attributes constructed and added from the given set of attributes to help the mining process are:

Attributes	Description	Type
count	number of hospitals in each city	continuous
coverage_person	population/count-binned into 4 groups	categorical
area_person	area/count-binned into 4 groups	categorical

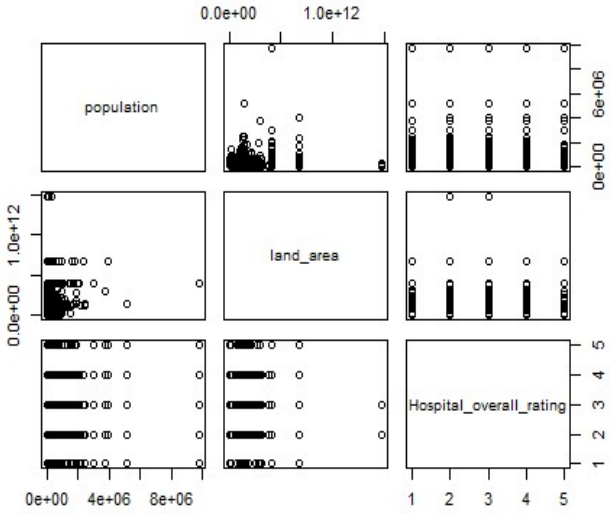


Creation of an ER-Diagram is also proposed in this step.

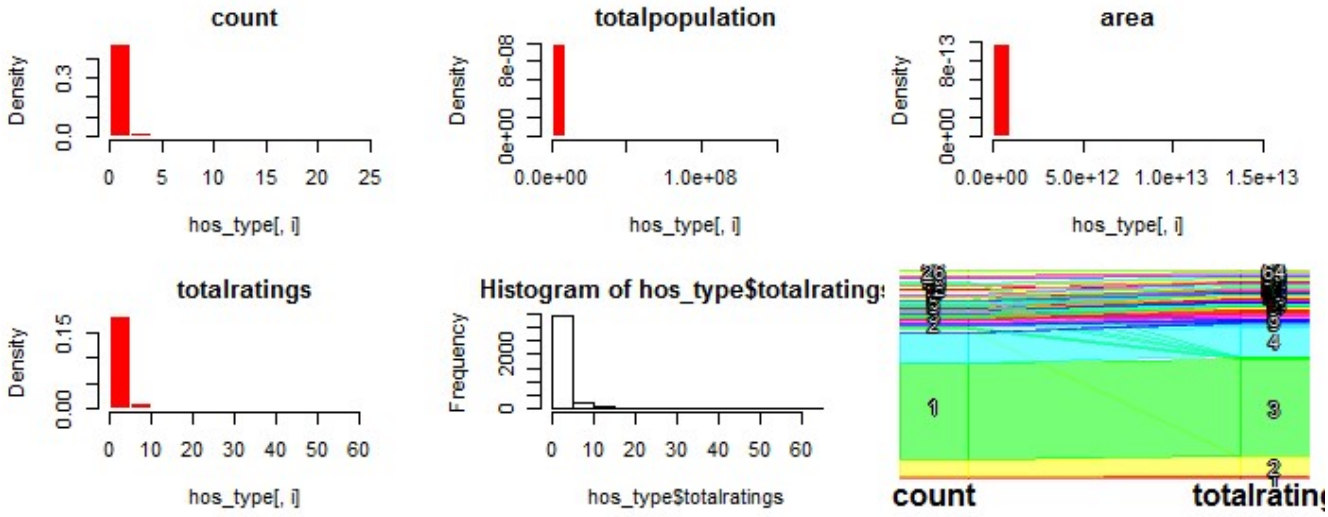


3. Aggregation, where summary or aggregation operations are applied to the data. For example, the number of hospitals for each state will be calculated so as to compute minimum or maximum hospitals for each state. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.

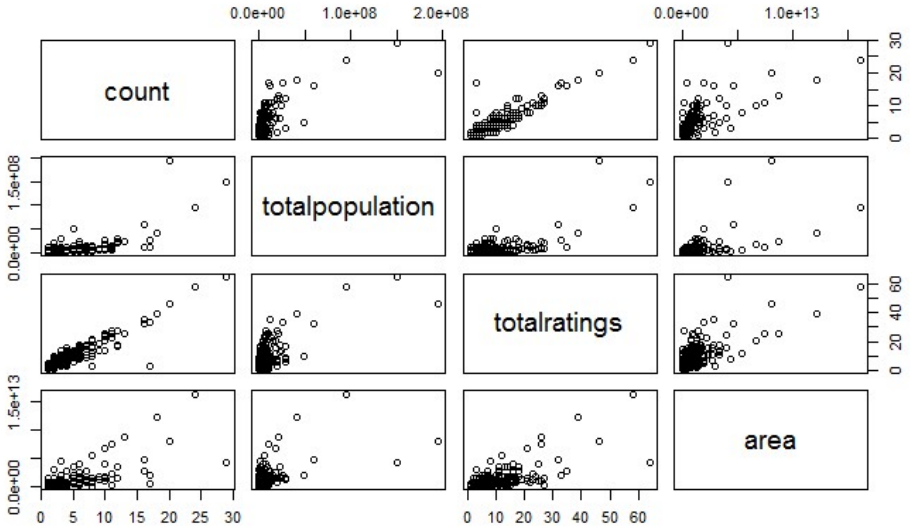
1. Aggregation on Hospital Level



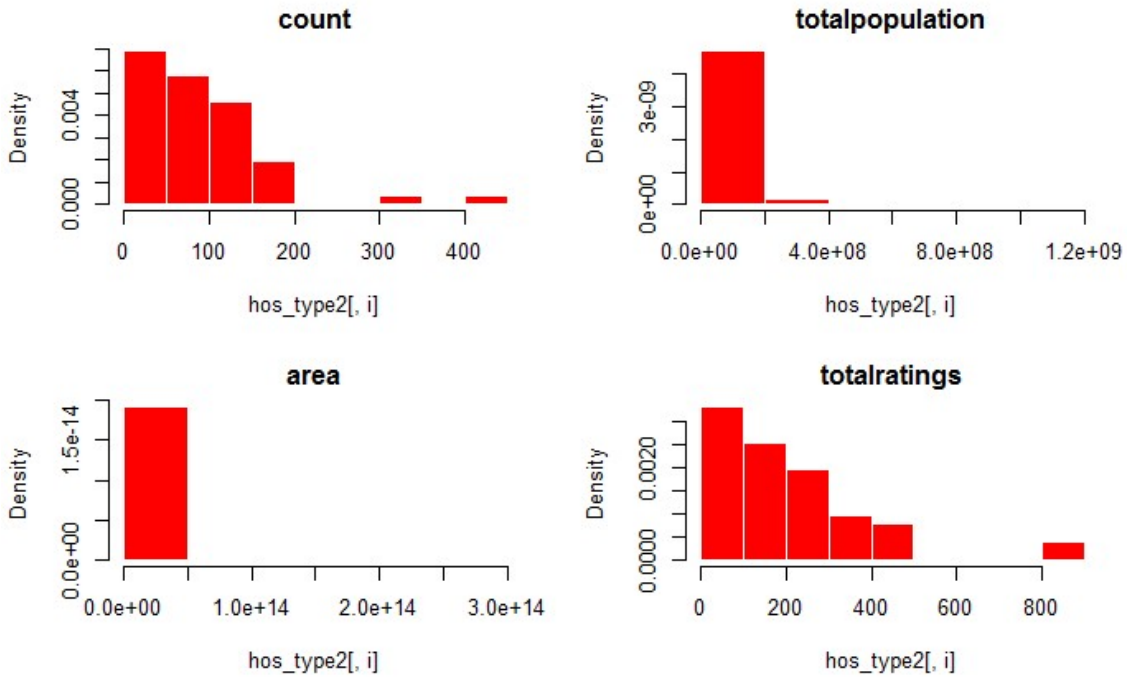
2. Aggregation on City Level



	count	totalpopulation	totalratings	area
count	1.0000000	0.6759701	0.9105381	0.6927141
totalpopulation	0.6759701	1.0000000	0.6563340	0.5848930
totalratings	0.9105381	0.6563340	1.0000000	0.6831173
area	0.6927141	0.5848930	0.6831173	1.0000000



3. Aggregation on State Level



5. Finding useful features to represent the data depending on the goal of the task.

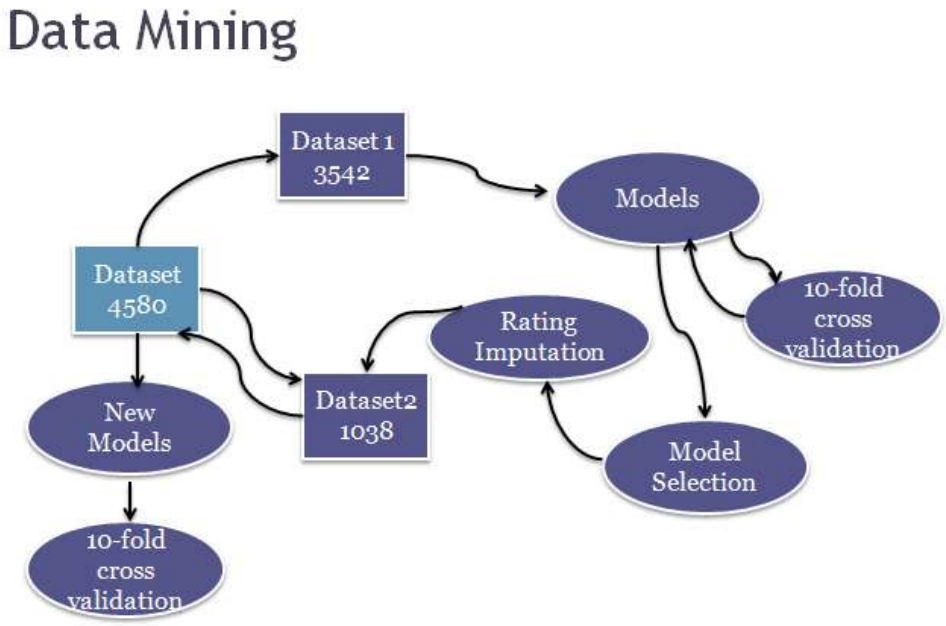
According to parallel coordinates and other data analysis the most useful features are:

- Hospital_overall_rating
- State
- Hospital_Type
- Emergency_Services
- Mortality
- coverage_area
- Safety_of_care
- coverage_person
- Effectiveness_of_care
- Timeliness_of_care

Step 5: Data Mining

After data transformation the main goals is data mining. Data mining process involves followings steps

1. Dividing the processed and transformed dataset into two datasets. Dataset 1 has 3542 rows with no missing values for ratings while dataset 2 has 1038 rows with missing values for ratings.
2. Used the data mining algorithms on dataset 1 to impute ratings. For classification decision tree and naïve bayes is used and for clustering decision tree is used.
3. Checked the accuracy, precision and recall to see how the models were performing
4. Picked 1 model which is the best based on performance matrix.
5. Used that model in rating imputation for dataset2
6. Used complete dataset with all labels for rating and build a model again and checked the accuracy(which increased)

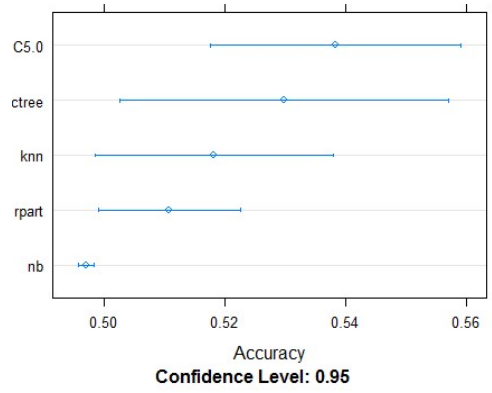


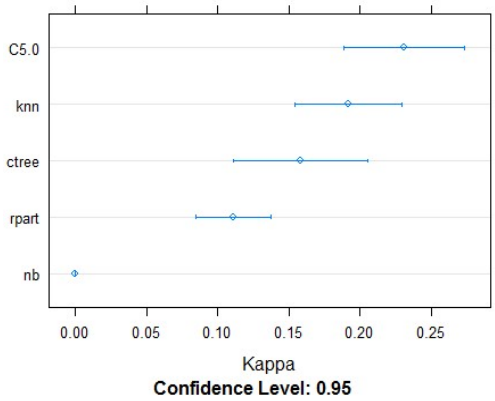
Data mining algorithms used for rating imputation and their accuracy is given below:

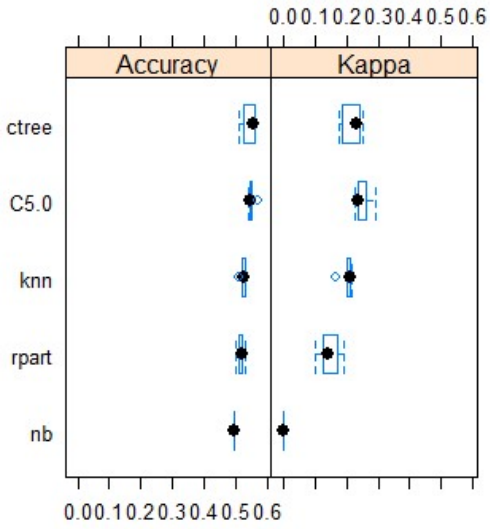
Performance Metrics

Algorithm	Factor Level	R-Package	Accuracy	Precision					Recall				
				Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Decision Tree	3	ctree	71%	-	-	-	-	-	-	-	-	-	-
Decision Tree	5	ctree	54%	0	48%	56%	49%	0	0	32%	80%	33%	0
Decision Tree	3	rpart	36%	-	-	-	-	-	-	-	-	-	-
Decision Tree	5	rpart	51%	0	47%	51%	0	0	0	25%	94%	0	0
Decision Tree	3	C50	72%	-	-	-	-	-	-	-	-	-	-
Decision Tree	5	C50	55%	36%	52%	58%	48%	0	18%	38%	76%	37%	0
Naive Bayes Classifier	3	naiveBayes	35%	-	-	-	-	-	-	-	-	-	-
Naive Bayes Classifier	5	naiveBayes	49%	0	0	100%	0	0	0	0	100%	0	0
KNN	3	knncat	84%	-	-	-	-	-	-	-	-	-	-
KNN	5	knncat	53%	50%	48%	56%	47%	0	18%	41%	75%	30%	0

10 fold cross validation is done to measure the performance of each model and C50 package for C4.5 tree has been selected as its performance is better than all models.







Results

Implementation of model on remaining data to find out missing rating is next step. After rating imputation the model is evaluated again to measure the performance. The overall accuracy of the model is 72.2% and other performance metrics of the model are given below

Confusion Matrix

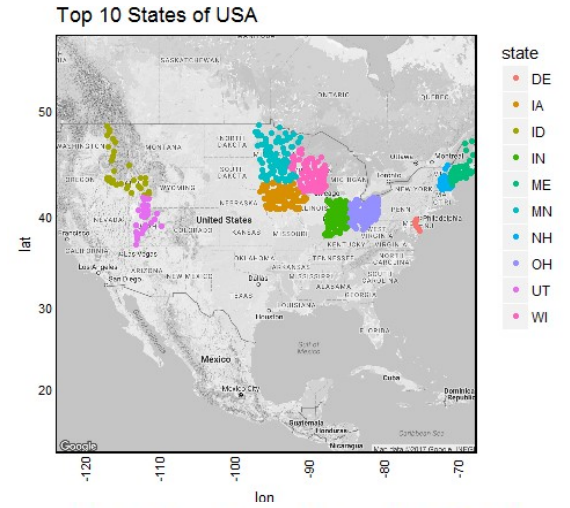
	References				
Predictions	1	2	3	4	5
1	7	7	1	0	0
2	16	67	25	0	0
3	3	54	289	40	0
4	0	5	82	276	9
5	0	1	4	7	23

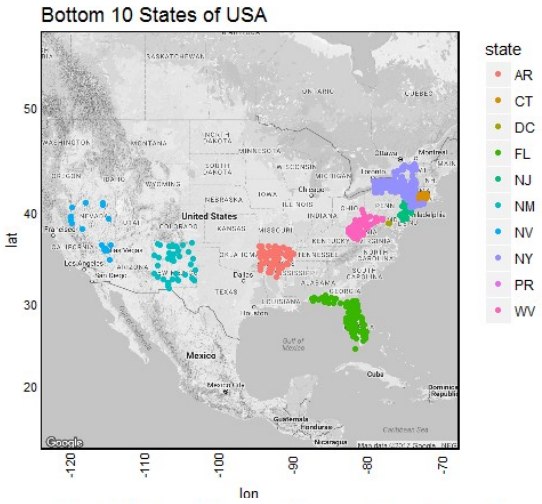
Performance Metrics	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Overall
Sensitivity	0.26	0.5	0.72	0.85	0.71	0.608
Specificity	0.99	0.94	0.81	0.83	0.98	0.91
Precision	0.46	0.62	0.78	0.74	0.65	0.65
Recall	0.26	0.5	0.72	0.85	0.71	0.608
Pos Pred	0.46	0.62	0.74	0.74	0.65	0.642
Neg Pred	0.97	0.91	0.78	0.91	0.98	0.91
Prevalencs	0.02	0.17	0.51	0.26	0.02	0.196
Detection Rate	0.007	0.1	0.43	0.14	0.003	0.136
Detection Prevalence	0.01	0.15	0.6	0.21	0.007	0.1954
Balanced Accuracy	0.63	0.72	0.76	0.84	0.85	0.76

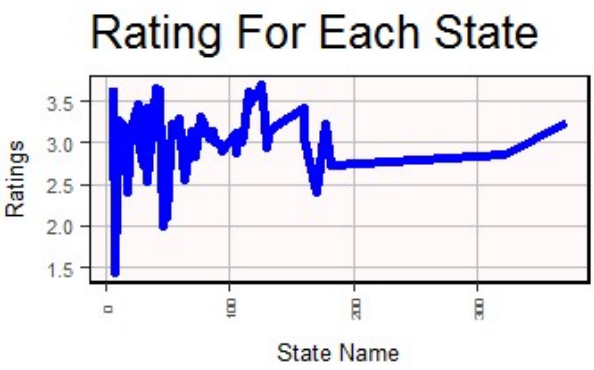
Most of the hospitals are predicted to get 3 rating, some get 2 and 4, and very few have been rated 1 and 5

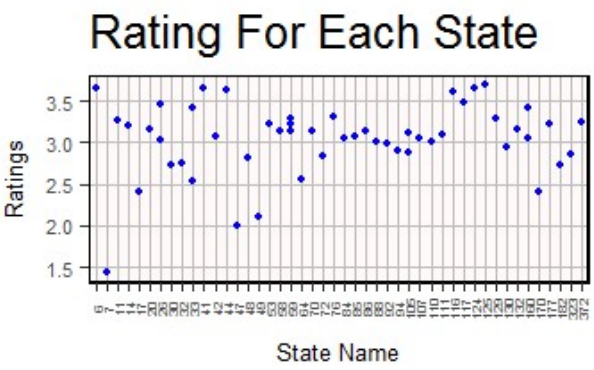
Conclusion

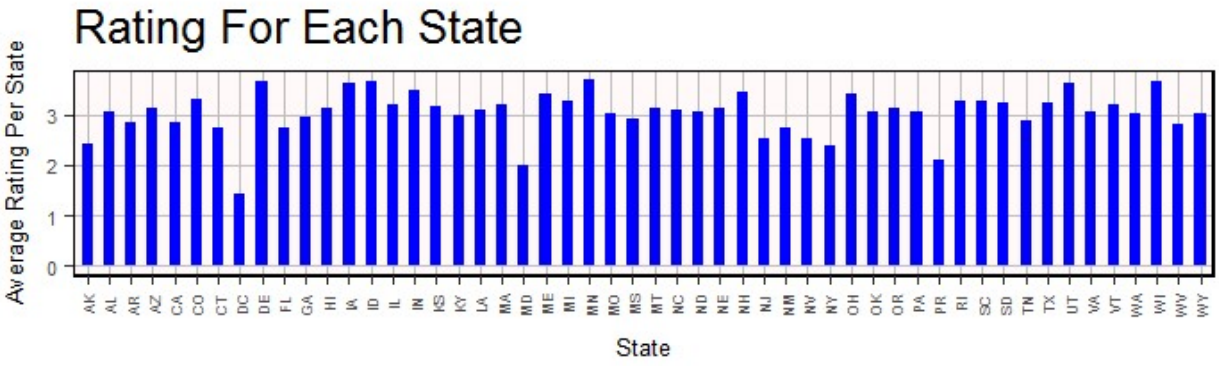
It is concluded that most of the hospitals are providing above national average services in different fields and are rated 3 mostly. If a hospital is getting rating 5 it doesn't means that it is performing above average in all fields. Health care services comparison for states shows the results given below:



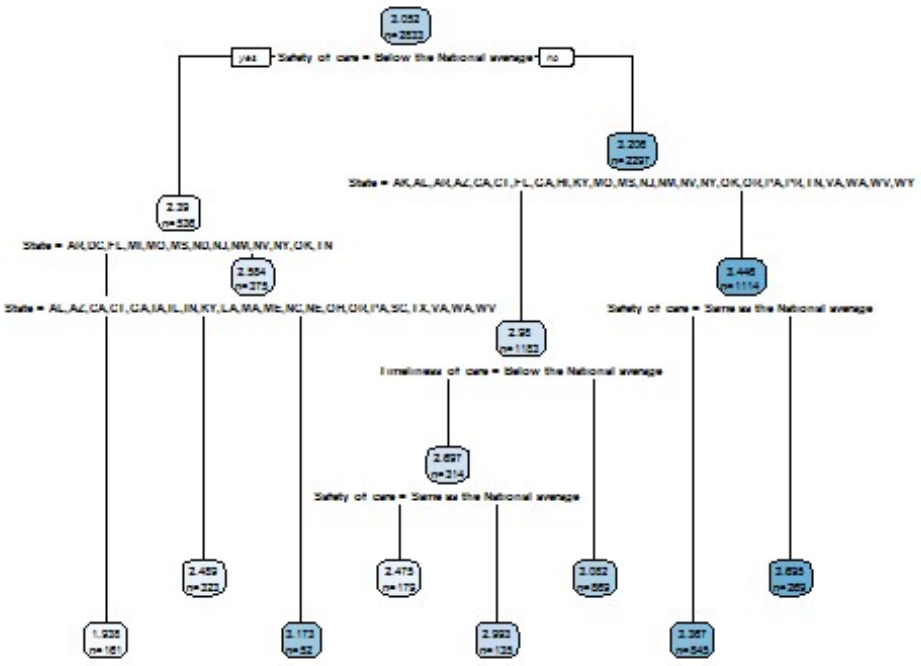




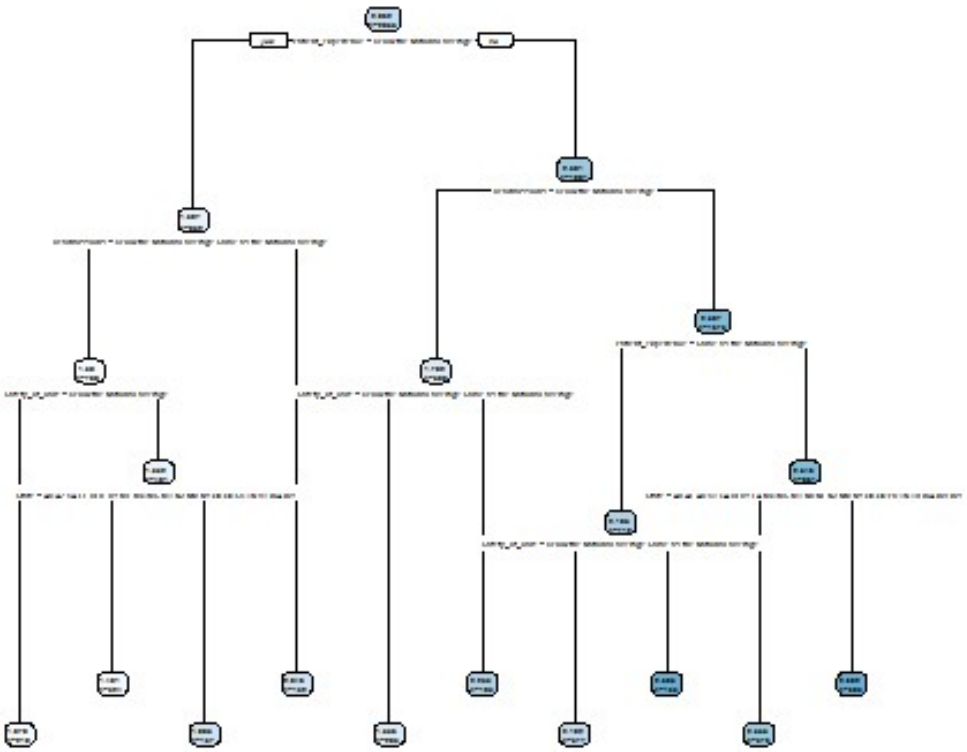




The above figure shows the distribution of average rating for each state.



Decision tree with 5 factor levels for dependent variable Hospital Overall ratings



Decision tree with 3 factor levels for dependent variable Hospital Overall ratings