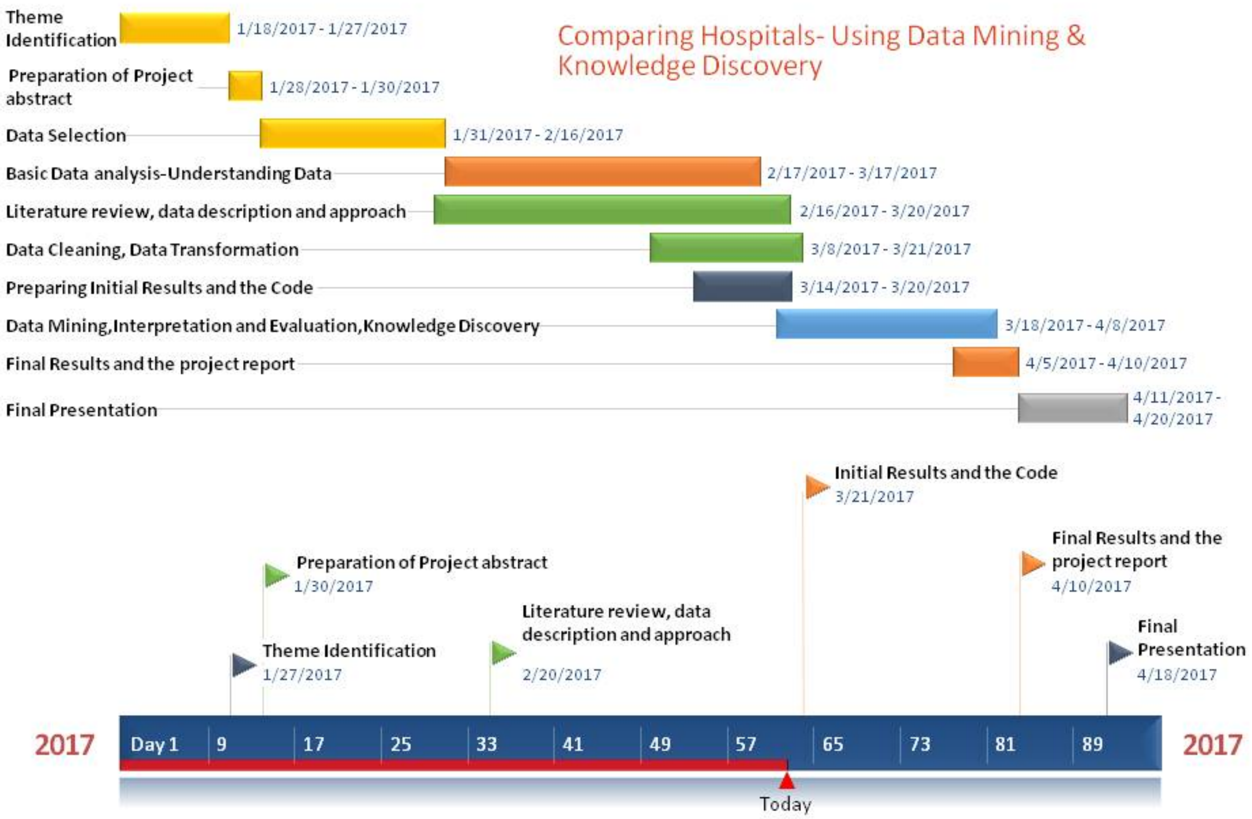
Comparing Hospitals of USA

Interim Report

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

Supervisor
Dr.Uzair Ahmed

March 21, 2017



Current Progress of Project

Contents

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:

- Healthcare associated Complications
- Healthcare associated Infections.
- Emergency wait time.
- Mortality and Re-admissions
- Outpatient Imaging Efficiency
- Payment and Value of Care

Literature Review

The term Knowledge Discovery refers to the process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub-sampling, and transformations of that database.

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

Dataset

Two data sets are considered for analysis in this project; Hospital General Information and US States-Population and Area

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 files 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	Disntinct	mean	median	
1	***Provider ID***	Qualitative-Nominal	4807	0	4807	NA	NA	Yes
2	***Hospital Name***	Qualitative-Nominal	4807	0	4608	NA	NA	Yes
3	***Address***	Qualitative-Nominal	Not Used					No
4	***City***	Qualitative-Nominal	4807	0	2947	NA	NA	Yes
5	***State***	Qualitative-Nominal	4807	0	56	NA	NA	Yes
6	***Zip Code***	Qualitative-Nominal	Not Used					No
7	***County Name***	Qualitative-Nominal	Not Used					No
8	***Phone Number***	Qualitative-Nominal	Not Used					No
9	***Hospital Type***	Qualitative-Nominal	4807	0	3	NA	NA	Yes
10	***Hospital Ownership***	Qualitative-Nominal	4807	0	10	NA	NA	Yes
11	***Emergency Services***	Qualitative-Nominal	4807	0	2	NA	NA	Yes
12	***Meet Criteria for meaningful use of EHRs***	Qualitative-Nominal	4373	434	1	NA	NA	Yes
13	***Hospital overall Rating***	Quantitative-Distrete	4807	0	6	3.8069	3	Yes
14	***Hospital overall Rating footnote***	Qualitative-Nominal	Not Used					No
15	***Safety of care National comparision***	Qualitative-Ordinal	2654	2153	3	NA	NA	Yes

16	***Safety of care National comparision footnote***	Qualitative-Nominal	Not Used					No
17	***Readmission national comparison***	Qualitative-Ordinal	3813	994	3	NA	NA	Yes
18	***Readmission national comparison footnote***	Qualitative-Nominal	Not Used					No
19	***Patient experience National comparison***	Qualitative-Ordinal	3454	1353	3	NA	NA	Yes
20	***Patient experience National comparison footnote***	Qualitative-Nominal	Not Used					No
21	***Effectiveness of care National comparison***	Qualitative-Ordinal	2790	2017	3	NA	NA	Yes
22	***Effectiveness of care National comparison footnote***	Qualitative-Nominal	Not Used					No
23	***Timeliness of care National comparison***	Qualitative-Ordinal	3565	1242	3	NA	NA	Yes
24	***Timeliness of care National comparison footnote***	Qualitative-Nominal	Not Used					No

2.US States-Population and Area

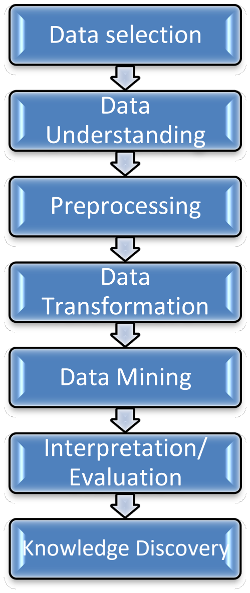
Apart from above tables another dataset is created to compare healthcare in different states by using population and area of states. The data has been collected from these websites.

http://www.enchantedlearning.com/usa/states/population.shtml
http://www.enchantedlearning.com/usa/states/area.shtml
http://www.infoplease.com/ipa/A0110468.html

US States-Population and Area - 8 Variables 56 Observations								
S.no	**Attributes**	**Type**	**Descriptive Statistics**					**Used During Analysis-Yes/No**
			Observations	**Missing Values**	**Distinct**	**mean**	**median**	
1	***State***	Qualitative-Nominal	56	0	56	NA	NA	Yes
2	***Area Ranking***	Qualitative-Ordinal	56	0	56	NA	NA	Yes
3	***Area(square miles)***	Quantitative-Distrete	56	0	56			Yes
4	***Area(square km)***	Quantitative-Distrete	56	0	56			Yes
5	***Population Ranking***	Qualitative-Ordinal	56	0	56	26		Yes
6	***Population***	Quantitative-Distrete	56	0	56			Yes
7	***Postal codes***	Qualitative-Nominal	56	0	56			Yes
8	***Capital City***	Qualitative-Nominal	56	0	56			Yes

Approach

The project will follow the sequence given below



Step 1: Data Selection

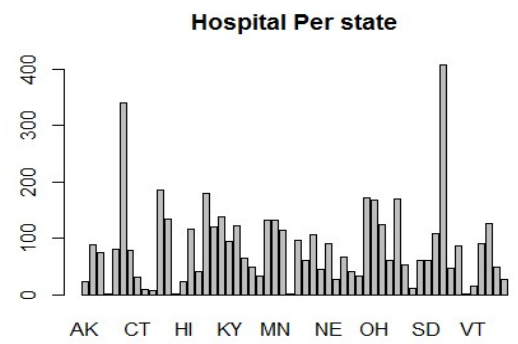
The Hospital Compare Database is composed of 56 datasets, out of which one dataset named "*Hospital General Information* "has been selected for the purpose of this project. This dataset is the summary of all types of comparisons done within the database. There are 28 variables out of which 15 variables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

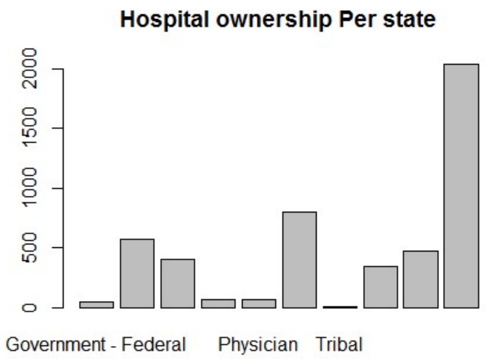
In order to compare hospitals in term of how the hospitals are distributed within a state a new dataset called *"US States-Population and Area "* has been created the details of the dataset is given under the section Dataset.

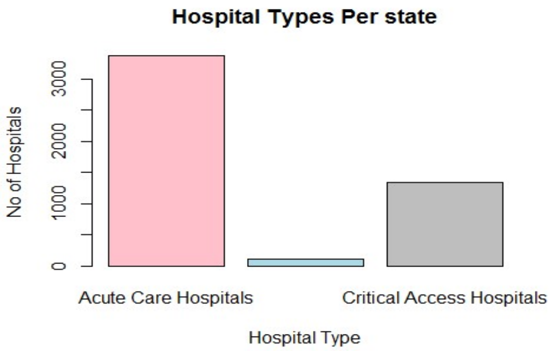
Step 2: Data Understanding

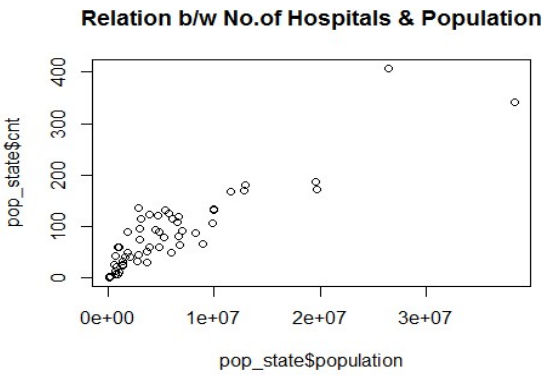
After selecting the dataset developing an understanding of data is very important by using following statistical description techniques such as data dispersion, central tendency measures, data visualization, data summarization. For data preprocessing to be successful, it is essential to have an overall 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.

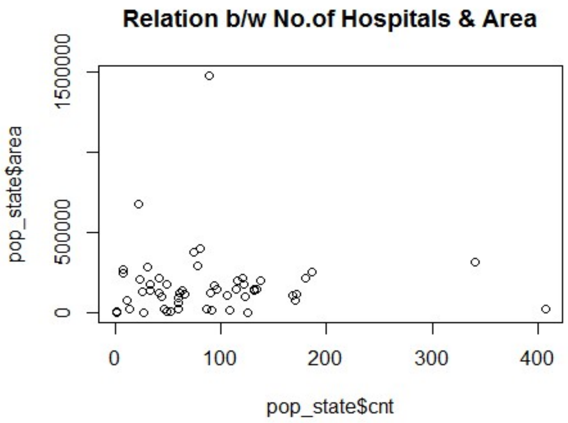
These measures are useful for identifying outliers. Finally, we can use many graphic displays of basic statistical descriptions to visually inspect our data by using histograms, scatter plots, and quintile plots.

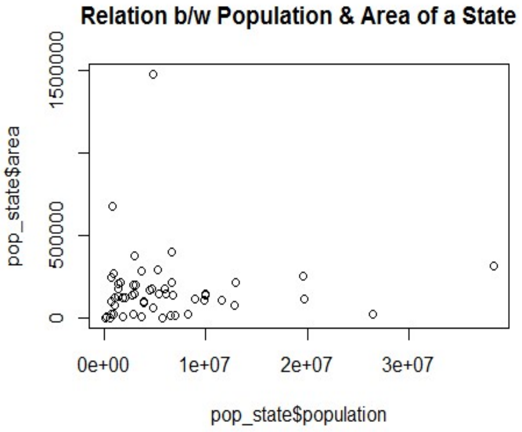












Step 3: Preprocessing

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

1. Removal of noise or outliers.
2. Collecting necessary information to model or account for noise.
3. Finding out strategies for handling missing data fields.
4. Attribute subset selection 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

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.
2. Attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process. 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.
4. Normalization to give all attributes an equal weight
5. Finding useful features to represent the data depending on the goal of the task.

Step 5: Data Mining

After data transformation the main goals of this step are

1. Choosing the data mining task. Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

2. Choosing the data mining algorithm(s).

    - Selecting method(s) to be used for searching for patterns in the data.
    - Deciding which models and parameters may be appropriate.
    - Matching a particular data mining method with the overall criteria of the KDD process.

2. Data mining. Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

Step 6: Interpretation or Evaluation

After the data has been mined the next focus is to

1. Interpreting mined patterns.

2. Consolidating discovered knowledge.

Step 7: Knowledge Discovery and Presentation of Knowledge

In this step the overall process of discovering useful knowledge from data will be presented. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.