"**Nursing Home Compare:** the data analysis of quality of care of over 15,000 nursing homes in the U.S"

INSTRUCTOR: Dr. Pamela Thompson
TEAM NAME: Dark Looters
MEMBERS:
Rashmi Ravindra Dasappa (800960696)
Sravani Kannelur (800967996)
Lakshmi Iswarya Kesari (800972717)
Virinchi Ande (800970447)
Gandhi Amarnadh Tadiparthi (800954124)

Table of Contents

Chapter 1: Data Set Introduction

1.1 Introduction:

Health has been the utmost important thing in a human's life. So, when the patients require assistance to improve their health conditions they opt for nursing homes.

Nursing home care provides help for a seriously ill care recipient. These facilities offer 24-hour supervision, nursing care, rehabilitation programs, and social activities.

When it comes to health we should always be cautious and when investigating nursing home options, find out whether the facility is government certified. If your loved one plans to use Medicare or Medicaid for payment, check the limited coverage of these as well by contacting the local Social Security Office.

Considering the importance and role being played by the nursing homes in the daily life of a person, any country's government will be interested in evaluating the nursing homes by the quality of service being provided by them and conducting inspections on yearly basis to identify the deficiencies and precautionary measures taken to rectify them.

So, all this prompted us to choose this dataset which has the real-time data and analysis can be improved by considering several other factors which the analyst may find useful apart from the ones considered by us in the project.

The main aim of this project to identify the nursing homes are outstanding and what are the causes for the ineffective performance of other nursing homes.

Objectives:
The project is aimed to analyse the state of nursing homes, their deficiencies and ratings. It is designed to achieve the following objectives:

- To determine what are the top 10 best and least nursing homes in the country.
- To know which states, have best level and least level of nursing homes.
- To learn the most common type of deficiencies and complaints.

1.2 Methodologies for Data Collection:

1.2.1 Primary data collection

- Raw data (also known as primary data) is a term for data collected from a source. Raw data has not been subjected to processing or any other manipulation, and are also referred to as primary data.
- Primary data is a type of information that is obtained directly from first-hand sources by means of surveys, observation or experimentation. It is data that has not been previously published and is derived from a new or original research study and collected at the source such as in marketing.
- Primary data collection is observed and recorded directly from respondents. The information collected is directly related to the specific research problem identified. All the questions that one asks the respondents must be totally unbiased and

formulated so that all the different respondents understand it.

1.2.2 Secondary data collection:

Secondary data is data collected by someone other than the user. Common sources of secondary data for social science include censuses, organizational records and data collected through qualitative methodologies or qualitative research. Primary data, by contrast, are collected by the investigator conducting the research.

- https://www.kaggle.com/medicare/nursing-home-compare

Kaggle is a great place to obtain the datasets related to various topics and explore the analysis work by others.

In this project all the information has been gathered from secondary sources that is on the kaggle website.

Chapter 2: CRISP-DM process

2.1 Introduction:

To approach the data mining, a cross-industry standard was clearly required, that is industry-neutral, tool-neutral, and application-neutral.

According to CRISP-DM, a given data mining project has a life cycle consisting of six phases which are adaptive:

- Business Understanding Phase
- Data Understanding Phase
- Data Preparation Phase
- Modelling Phase
- Evaluation and Deployment Phase

2.2 Phases

2.2.1 Phase 1: Business/Research Understanding Phase

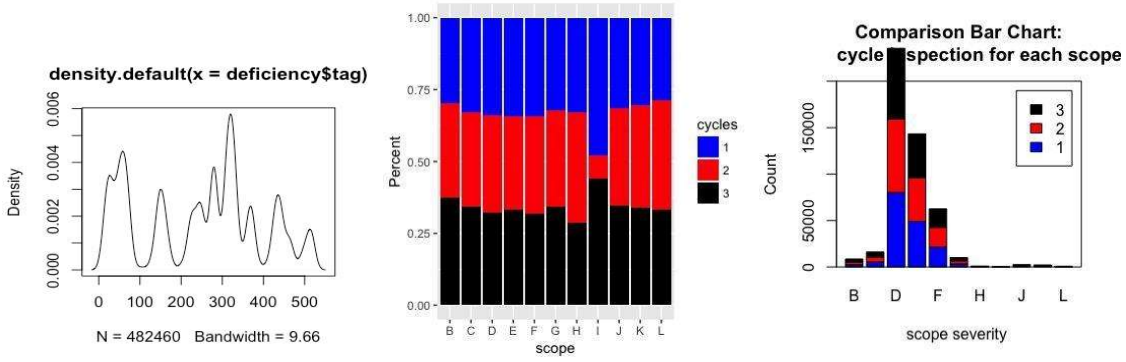
Data Mining Problem: How we can achieve to improve the ratings and quality of nursing homes by focusing on aspects that minimize the deficiencies/penalties?
Our objectives aim to do the following:

- Improves the facilities provided by the nursing homes by identifying the deficiencies.
- The nursing home can evaluate the performance of their staff based on the staff ratings.
- Helps the company to know which nursing home is performing good and which is not.
- Helps to the reason why some employees are not performing to the expected level.

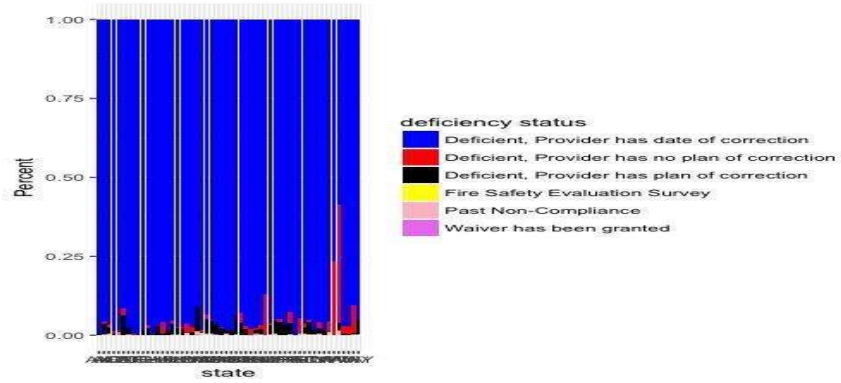
2.2.2 Phase 2: Data Understanding Phase

We have taken two datasets namely deficiencies dataset and provider information dataset. The deficiencies dataset contains 4,82,460 observations with 18 variables. Whereas, the provider information dataset contains 15,644 observations with 80 variables.

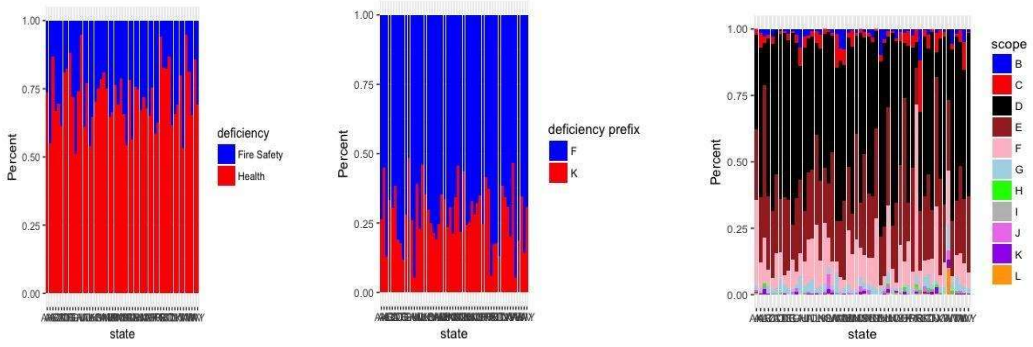
Graphs for understanding deficiencies dataset:



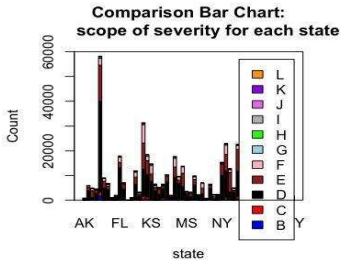
The graph is a density plot which represents the tag column under the deficiency dataset. The cycles of deficiency are compared as above through both general and normalized bar chart. At the first place, the density plot shows that data is not normalized and is to be carefully handled. The cycles 1,2 and 3 seems to have no dominating pattern. At scope of severity 'I' , an unusual behaviour is noted.

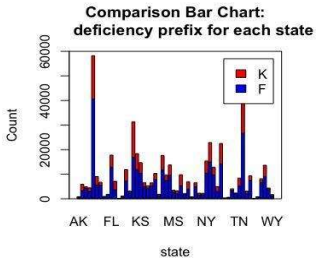


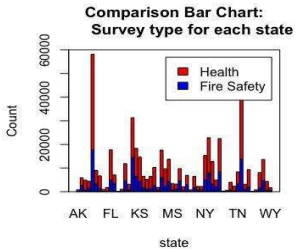
The plot shows the normalised deficiency status based on states.



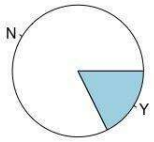
The plot shows the normalised deficiency based on states. The plot shows the normalised deficiency prefix based on states. The plot shows the normalised scope severity based on states.The health categorised deficiencies seems high in all the states. The deficiency prefix 'F' categorised seems occurring more. The scope severity code of 'D' is to be highly observed of all codes. We have to make sure that these dominating factors are closely observed in next phases.

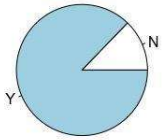


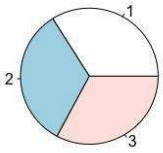




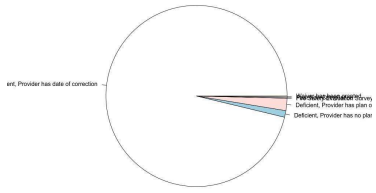
The mapping shows the scope severity based on states.The mapping shows the deficiency prefix based on states.The mapping shows the deficiency status based on states.Through these mappings, we try to compare states based on various factors which may help to analyse the behaviour of different states. A city in Arkansas and Tennessee is found to unnatural behavior. The states can be closely observed in next phases as conclude what might be the reason for such behavior. The reasons might be higher number of nursing homes in those states etc.

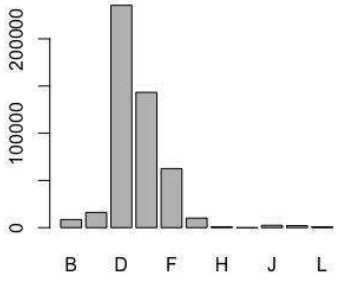


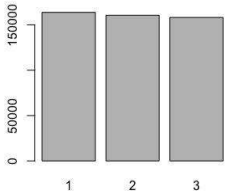




The above pie chart determines the number of complaints deficiency in the given deficiency dataset. The above pie chart determines the number of standard deficiencies based on the category taken from the deficiency dataset.The above pie chart determines how many cycles of inspection are more from the given deficiency dataset.From the above pie charts, we can observe that there are more number of complaints regarding deficiency from most of the nursing homes.

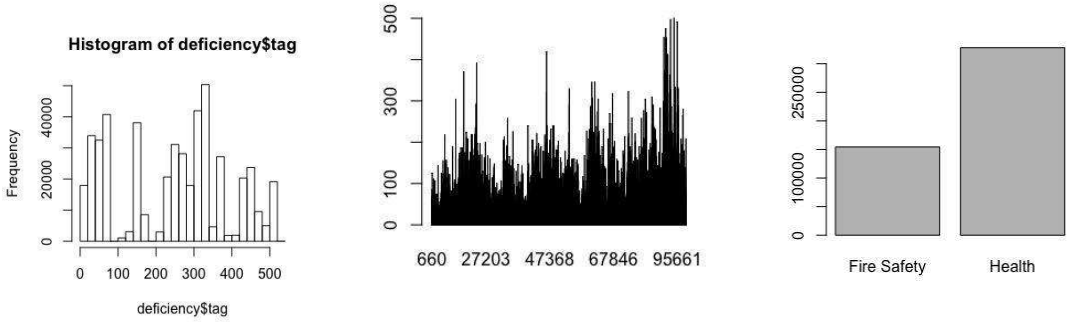




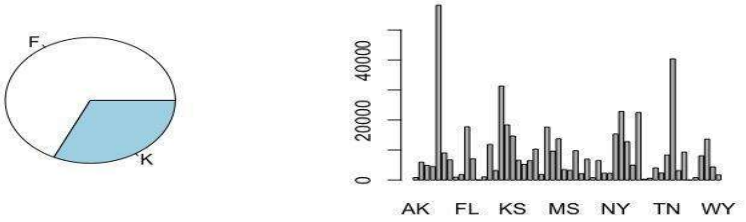


The above pie chart determines the status of deficiency corrected by the providers in the deficiency dataset.The above bar plot determines how severe a deficiency is in all providers (nursing homes) in the given deficiency dataset.The above bar plot helps to identify the cycles from the deficiency dataset.From the above graphs, we observe that there are many

number of nursing homes whose deficiencies have been corrected from time to time. As we can see from the bar plot, most of the nursing homes are under scope D severity.

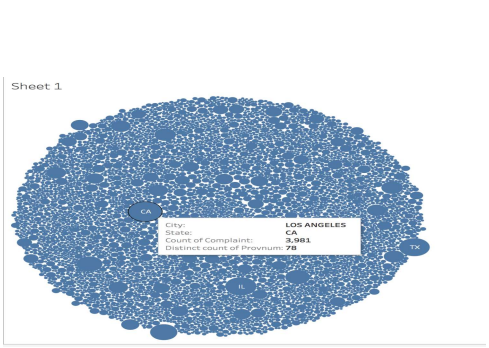


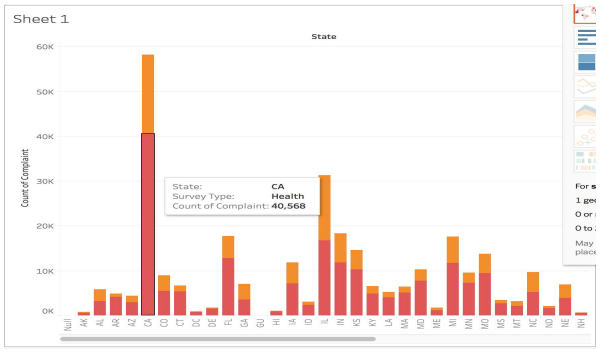
The above histogram helps in identifying the distribution of deficiency tag.
The above bar plot helps in identifying the frequency of zip code in the deficiency dataset.The above bar plot determines the types of survey done and their frequency from the deficiency dataset.From the given graphs, we can observe that there are more number of providers facing deficiencies in and around the zipcode of 95661. Also, there are more number of health deficiencies faced by nursing homes than fire safety deficiencies.



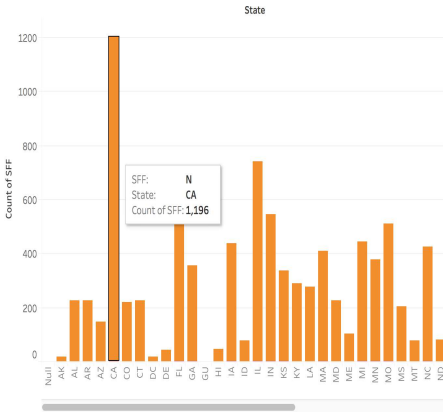
The above pie chart determines the type of deficiencies detected which are represented by the prefixes.The above bar graph determines the provider of medicare are distributed across which states.From the above graphs, we can observe that there are more number of F type deficiencies detected from most of the nursing homes. Also, there are more number of providers with medicare between states AK and FL which does not exactly provide which state has more number of providers indicating abnormal behavior.
Also, we have used Tableau visualizations to understand data better.

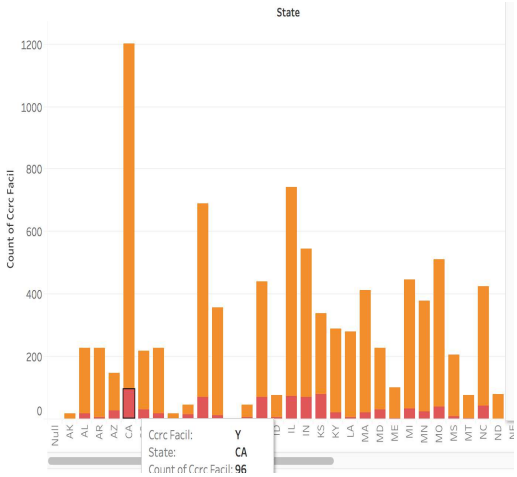
Tableau observations in terms of complaints:



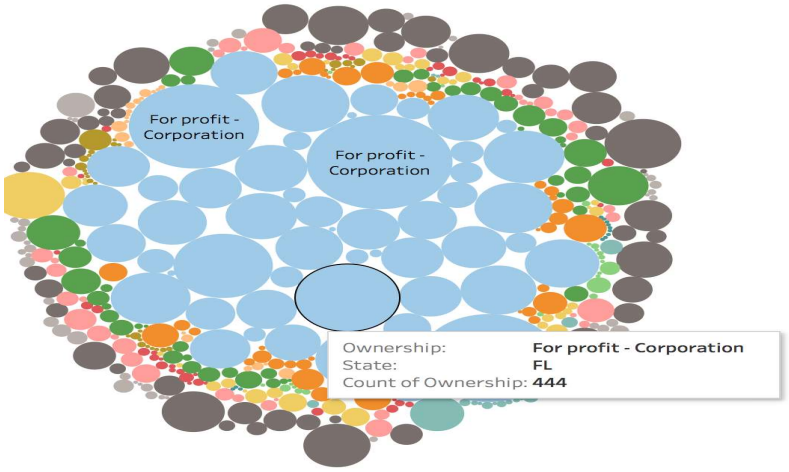


The nursing homes in California is observed to be with highest complaint counts and in particular Health. AK, DC, DE, HI, GU, ME, MS, MT, ND, NH are found to be recording less number of complaints. CA state conclusions in next stages are to be taken care of.

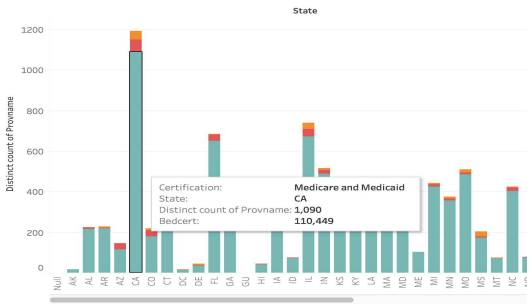


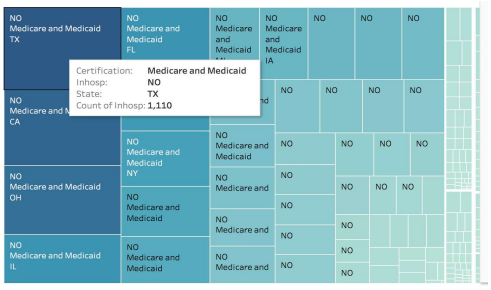


The SFF (Special Focus Facility) is almost unobserved in all the states. The continuing care retirement facility is found be least recorded and is found in notable number in few states like CA,IL,IN,KS and so on.



From the above bubble chart, we can observe that the number of nursing homes which have For profit - corporation type of ownership in Florida is 444. We observed that this type of ownership is high in number in California state.





We can observe that the highest number of certified beds in the nursing homes with both medicare and medicaid facilities is in the state of California. Also, the highest number of nursing homes who do not have inhospital facilities are in the state of Texas and there are very minute number of nursing homes who have such facilities.

Observations under Data Understanding Phase:

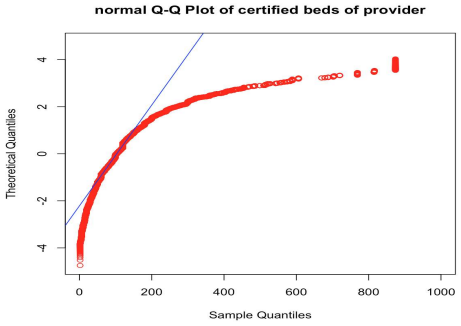
- When we checked for the columns with missing values. we found that statdate has missing values.
- But based on our observation, those missing values are not due to the data being missed but to represent a pattern.
- the fact that those providers has no date of correction of their deficiencies.
- In the dataset we have chosen, most of the columns are character type columns and we have tag column that is integer.
- But, those tags are predefined and are assigned to the providers based on the deficiency. So, it doesn't make any sense to find outliers for that tag variable.
- As per the dataset consider Inspection cycle attribute,
    - If it is 1 then the survey date and the file date fall between one year
    - If it is 2 then it fall between 12-24 months and the correction date is between the survey date and file date.

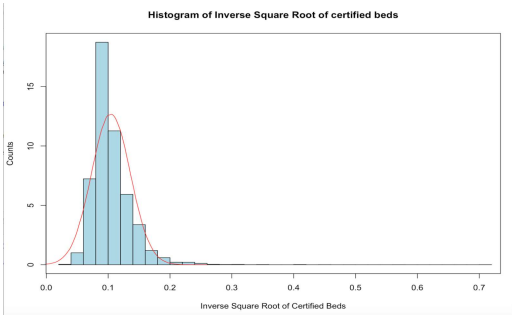
2.2.3 Phase 3: Data Preparation Phase

- The reason for considering provider_info.csv dataset is to consider the most influencing attributes like the quality_rating, staffing_rating etc to determine the overall_rating effectively for the new records.
- Merged the datasets deficiencies.csv and provider_info.csv using a library called **'sqldf'**.
- Performed Min-Max normalization and Z-score on some variables such as BEDCERT, RESTOT,etc.
- Divided the dataset into training and test datasets.
- Determining the missing values in the various rating variables by using **'mice'** package.
- Calculated skewness with z-score standardization and found that:
- The skewness is 0.4648 for 'BEDCERT' variable.
    - right skew mean > medium and skew is positive.
    - no skew is zero, left skew is negative.

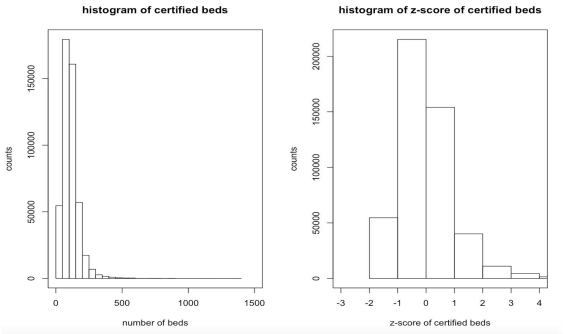
Exploratory Data Analysis on Provider information dataset:

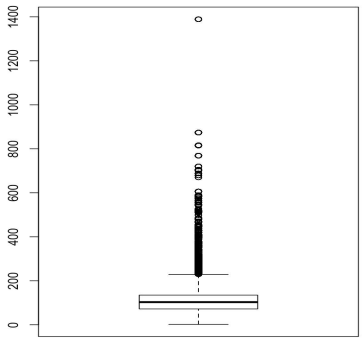
BEDCERT : Number of certified beds attribute





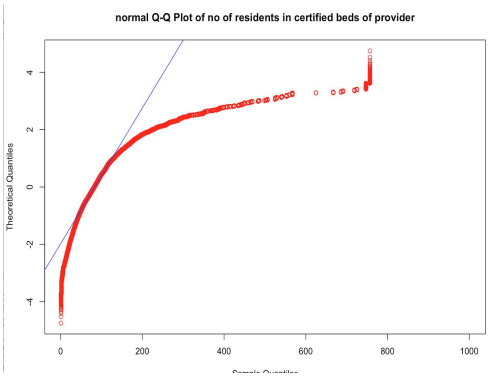
The distribution of certified beds of provider does not fall on the qqline when plotted, so, we conclude that certified beds distribution is not normalized. In order to remove skewness, we made inverse square root transformation and found the data to be normal.

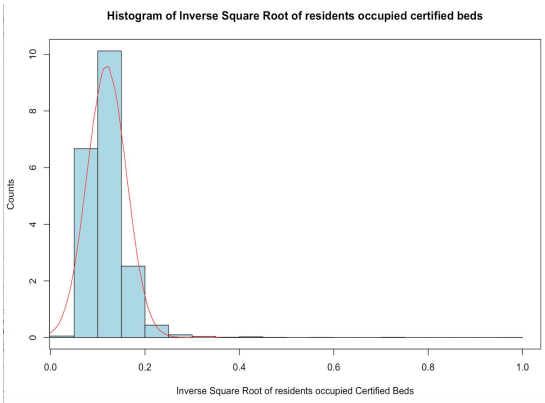




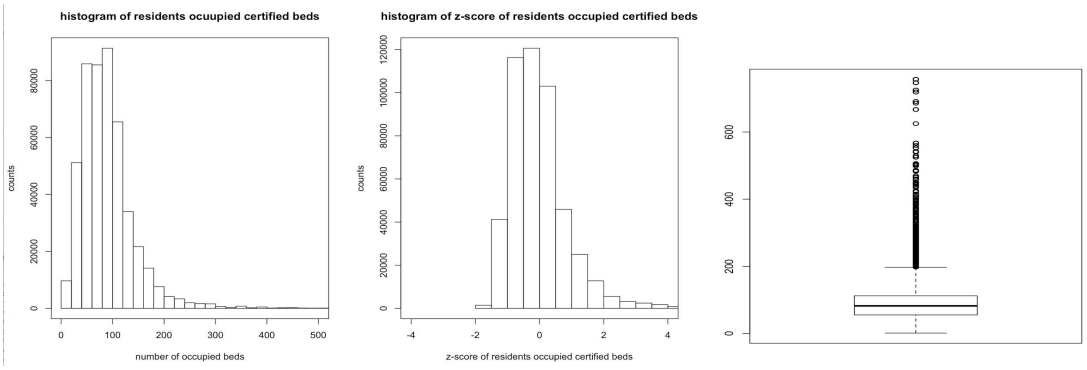
Z-score transformation is performed over certified beds, the z-score value of +4 is observed for some data. A box plot over certified beds also prove that outliers existence.

RESTOT: Number of residents in certified beds attribute



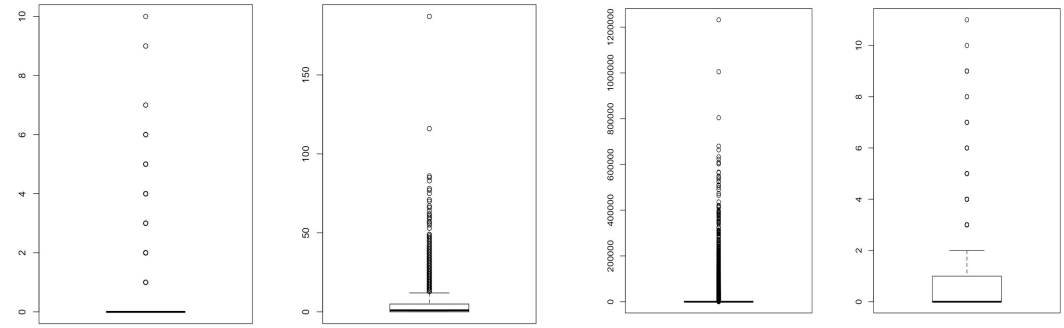


The distribution of number of residents at a provider does not fall on the qqline when plotted, so, we conclude it as not normalized. In order to remove skewness, we made inverse square root transformation and found the data to be normal.



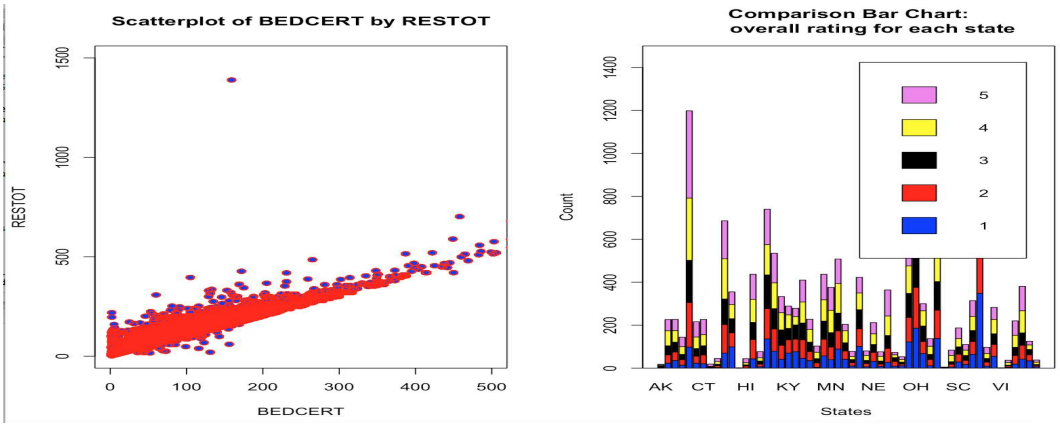
Z-score transformation is performed over residents occupying certified beds, the z-score value of +4 and further is observed for some data values. A box plot over occupied residents also prove that outliers existence.

An outlier check on other attributes such as: Fine count,Complaint count, Final total, Total penalty count



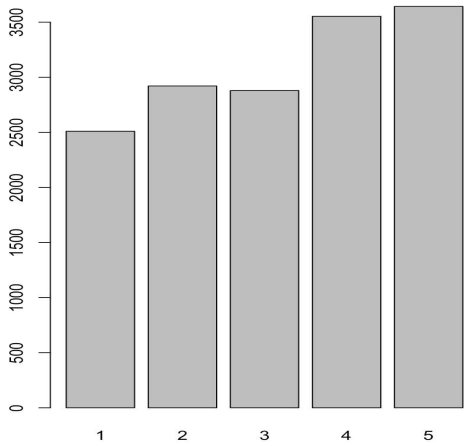
Fine count Complaint count Final count Total penalty count

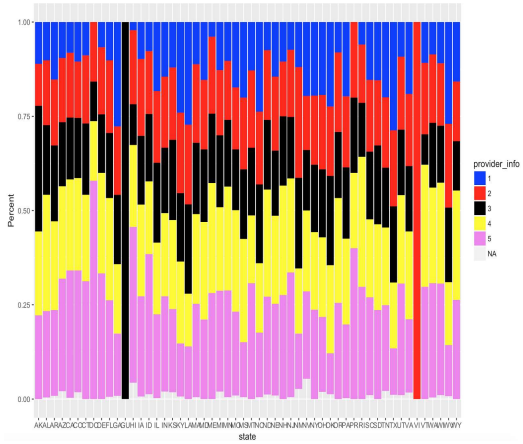
We may think that the above attributes consists of outliers from seeing the above box plots. But, as these variables are used in predicting the missing values, so these attributes should not be altered.



- The scatterplot helps in identifying the distribution of the number of certified beds in a nursing home over the number of beds occupied by the residents in the nursing home.

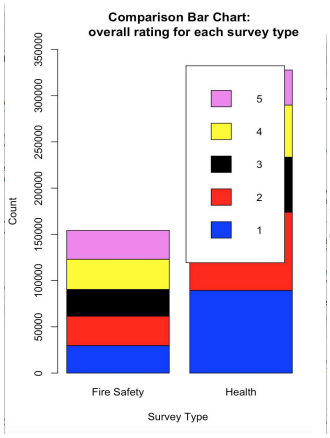
- The bar chart shows the overall rating in detail which ranges from 1 to 5 values for each state.

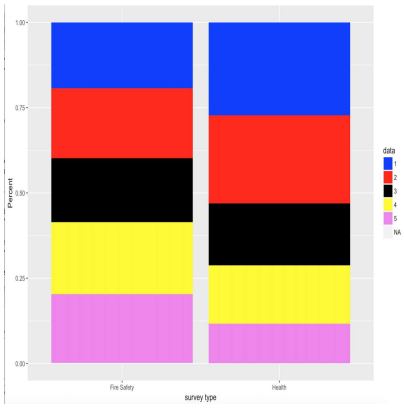


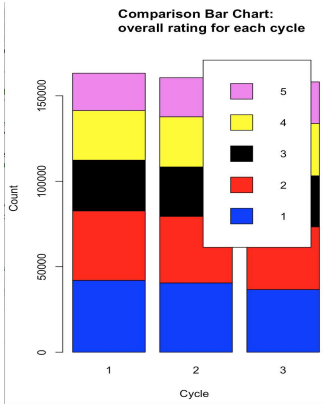


The above bar plot indicates number of nursing homes with their respective overall ratings. We can observe that most of the nursing homes around 3500 have an overall rating of 4 and 5. The normalised plot shows the overall rating for each state. We observed that most of the nursing homes in Guam state has an overall rating of 3. Also, most of the nursing homes in Washington DC state has an overall rating of 5. In addition to that, most of the nursing homes in Virgin Islands state has an overall rating of 2.

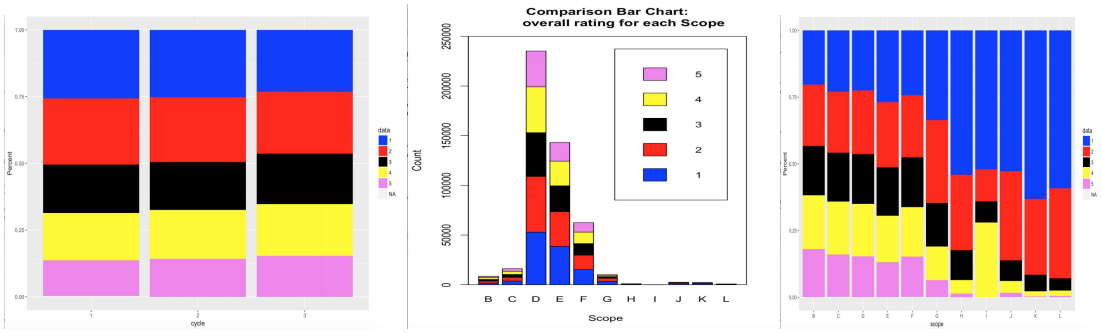
Performed Exploratory Data Analysis on merging the deficiencies dataset and provider information dataset:







The bar chart plots overall rating for each survey type. We observed that the most of the nursing homes around 1 million which have survey type as Health have an overall rating of 1 and 2. The bar chart plots overall rating for each survey type. We observed that less than 25% of the nursing homes have a rating of 5 for the Fire Safety survey type. Also, 50% to 100% of the nursing homes have a rating of 4 and 5 for the Health survey type. The bar chart plots overall rating for each cycle. Around 50000 nursing homes who had inspection in the last 24 months have a rating of 1.



The above bar chart plots overall rating for each cycle. 12% of the nursing homes which has inspection cycle as cycle 1 had a rating of 5 approximately with a slight difference in each of the cycles. Whereas, 50% to 75% of the nursing homes which had inspection cycle as cycle 3 had a rating of 2. The above bar chart plots overall rating for each scope. Approximately, 50,000 nursing homes which were under scope D severity had a rating of 1 and 2. Also, less than 50000 nursing homes under scope of severity 'E' had a rating of 1. The above bar chart plots overall rating for each scope. From 37.5% to 100% of the nursing homes had an overall rating of 1 under scope K severity. Very few of the nursing homes under scope K severity have a rating of 5.

Partition of Data into Training and Test datasets:

The merged dataset containing both deficiencies and provider information datasets are partitioned into training and testing data sets. This split is made as 75% : 25%

Conclusion from t-test on training and test data set:

The variable considered is: Number of certified beds "BEDCERT".
The p-value (0.9457523) is large, there is no evidence that the mean number of certified beds differs between the training data set and the test data set. For this variable at least, the partition seems valid.

Conclusion from Chi-Square test:

The variable considered is: Inspection Cycle "Cycle".

- The p-value (0.06679071) is large, there is no evidence that the observed frequencies represent proportions that are significantly different for the training and test data sets. In other words, for this variable, the partition is valid.
- However, the data that is being partitioned into training and test data varies based on the random values assigned to the variable part from 0 to 1. So as the values for part variable in the dataset changes, the training and test data varies which implies the corresponding results associated with the statistical tests also changes.

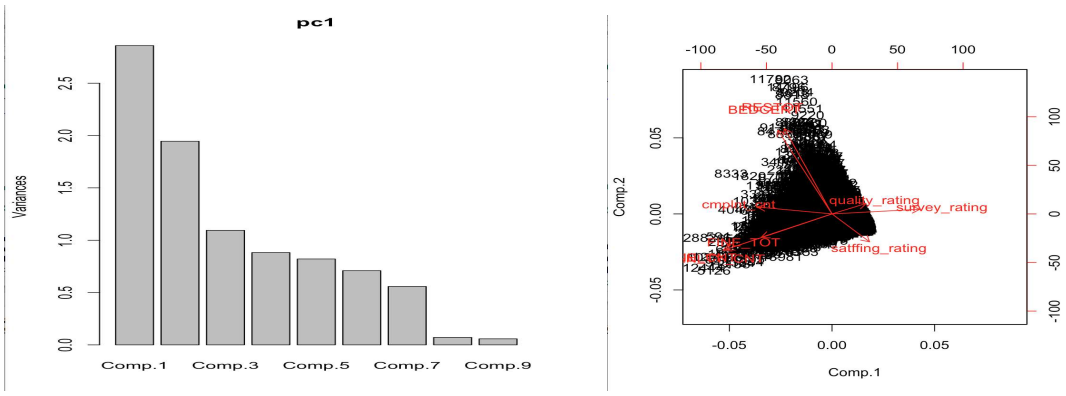
2.2.4 Phase 4: Data Modelling Phase

- Modelling techniques used are:
  - Principle Component Analysis,
  - Linear and Multiple Regression,
  - Visualisation through Decision Tree and
  - Neural Networks are performed.

- K-Means Clustering

Initially we have done the principal component analysis to identify the impact of each attribute on the target variable and identify the interesting patterns by reducing the dimensionality.

Principle Component Analysis:

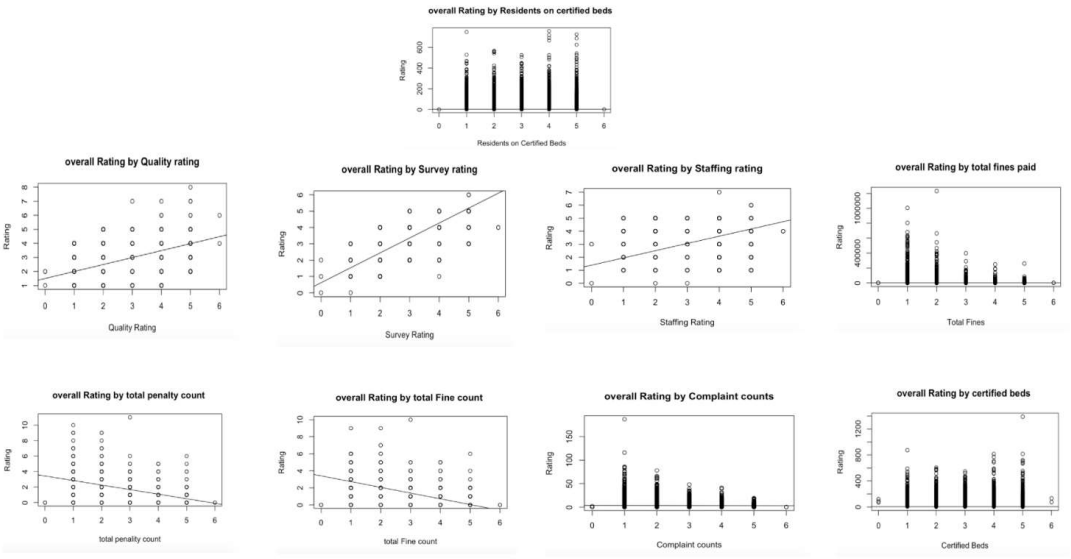


From the plot graph of PCA, the two components are found to have value greater than 1. When observed closely by biplot graph, In first component FINECNT and total penalty count play significant role. In second component, BEDCERT and RESTOT plays significant role. The **nine** attributes which are found to have impact over decision attribute are:

1. Survey Rating
2. Quality Rating
3. Staffing Rating
4. Total penalty count
5. Fine count
6. Number of Certified Beds
7. Residents on certified beds
8. Number of complaints
9. Total Fines paid

After Principal Component Analysis, we did the Linear regression considering the total penalty attribute to predict the target Variable overall_rating. Based on the results we decided to move to multiple regression considering various attributes which are influencing the overall_rating to improve the efficiency of the model.

Linear and Multiple Regression Analysis:



Residents on certified beds is 2%;Quality rating 3% ; Survey rating 7%; Staffing rating 21% ; Fine count is 13%; Total penalty count is 15% ; Total fines is 6%; Complaint count is 15%; Certified beds is 3% .

Output for Linear Regression :-
Residual standard error: 1.297 on 15642 degrees of freedom
Multiple R-squared: 0.1449, Adjusted R-squared: 0.1449
F-statistic: 2651 on 1 and 15642 DF, p-value: < 2.2e-16

Observation from Linear Regression :-
The linear regression performed over (overall_rating and total penalty count) is giving 15% accuracy in prediction. So, we will go with the multiple regression.

Output for Multiple Linear Regression :-
Residual standard error: 0.4544 on 15634 degrees of freedom
Multiple R-squared: 0.895, Adjusted R-squared: 0.895
F-statistic: 1.481e+04 on 9 and 15634 DF, p-value: < 2.2e-16

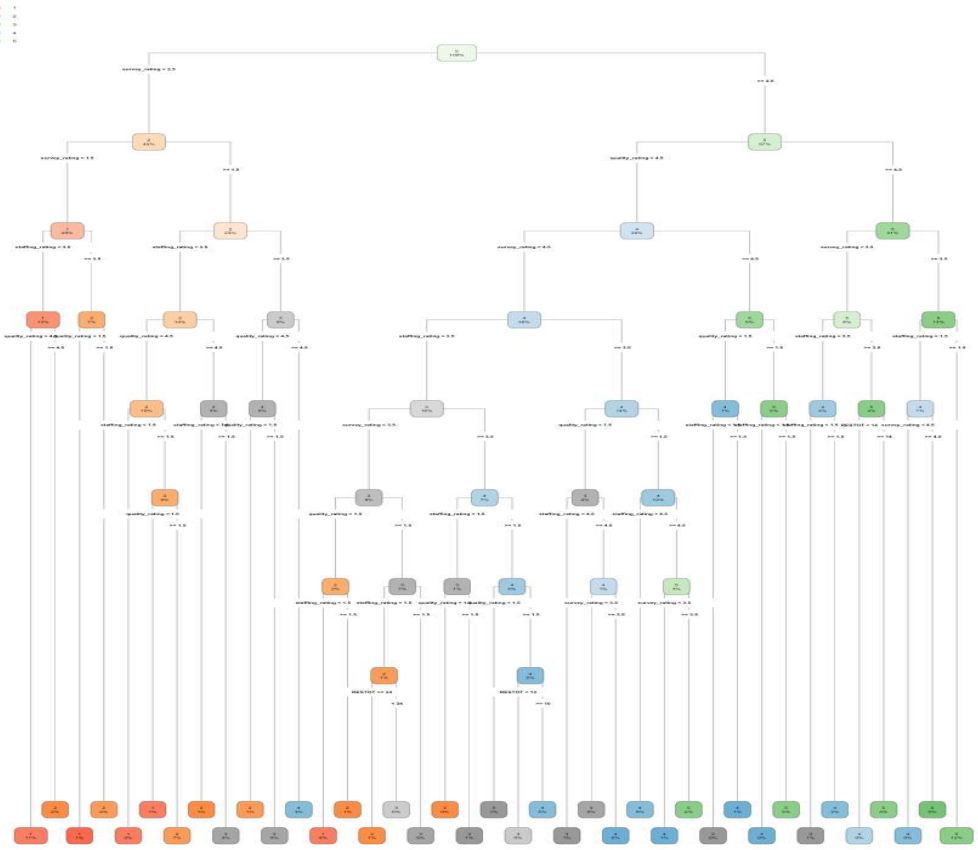
Observation from Multiple Linear Regression output :-
The multiple regression model predicts 89.5% accurately. The multiple linear regression equation of all variables with overall_rating is observed. We finally conclude that the model is good fit with multiple regression accuracy.

Supervised Learning:

In supervised Learning, the models try to learn the patterns and classify the new records to the categories in the target variable. Among several modelling techniques available, we have used the below techniques.

Decision Tree Analysis:

In Decision Tree analysis , Normalization of data is done prior to diving the data into training and testing dataset. The normalized training dataset attributes are fetched into the algorithm as input.



	A	B
1	provnum	overall_ratin
2	15009	5
3	15010	3
4	15012	4
5	15014	4
6	15015	3
7	15016	3
8	15019	4
9	15023	3
10	15024	5
11	15027	4
12	15028	4
13	15031	2
14	15032	1
15	15034	3
16	15035	4

ModellingPhase.R* | neuraldf1 | my_solution | comb | data | deficiency

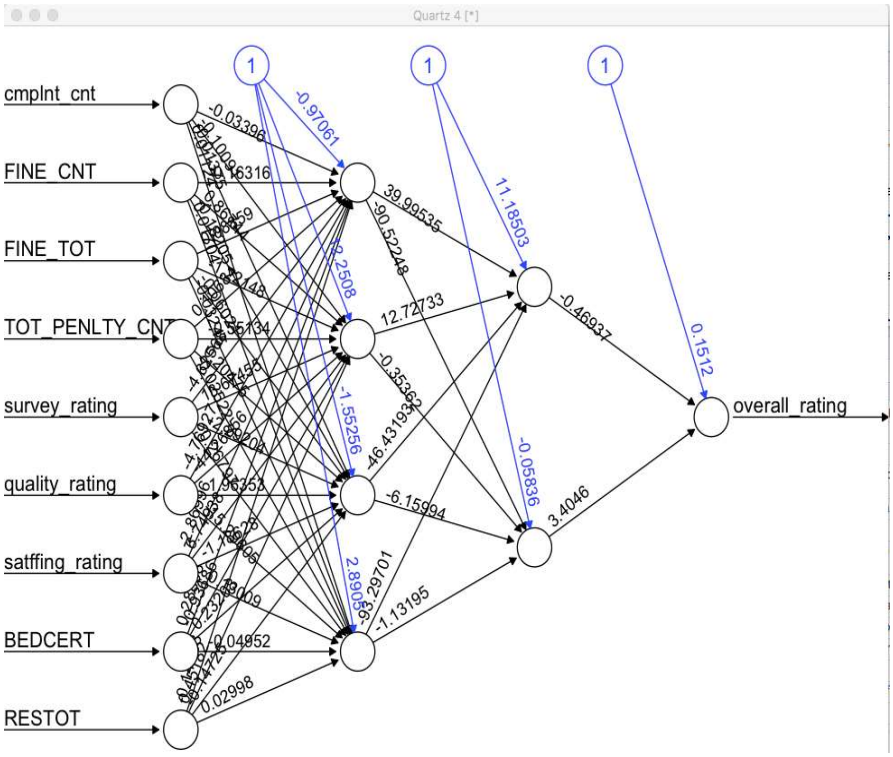
Filter

	provnum	overall_rating
10	015027	4
19	015043	3
20	015044	4
21	015045	2
22	015047	3
23	015048	1
24	015049	3
28	015063	4
32	015071	4
35	015076	2
36	015083	2
37	015084	5

Using the decision tree modeling technique, the prediction of overall rating is done based on the variables such as FINE_CNT , TOT_PENLTY_CNT, BEDCERT, RESTOT, cmplnt_cnt, FINE_TOT, survey_rating, quality_rating and staffing_rating. The prediction seems accurate by the values we compare.

Neural Networks:

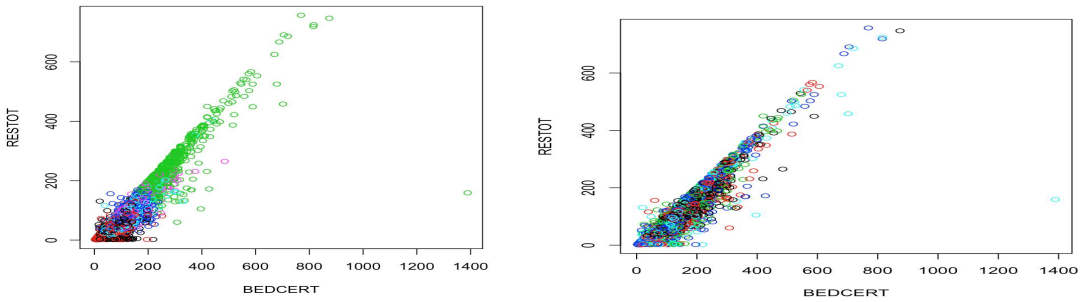
In Neural Network analysis , First data normalization is done prior to diving the data into training and testing dataset. The normalized training dataset attributes are fetched into the algorithm as input.



As we know in neural network, the weights are adjusted back and forth until the network is stable. Using this modeling technique, the mean squared error is about 1.3% which is very less for a very large dataset.

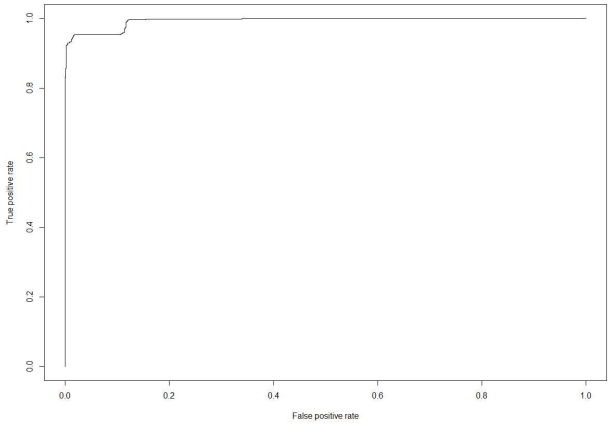
Unsupervised Learning: In the unsupervised learning the models try to identify the patterns in the dataset and there is no target variable. We have used the K-Means clustering method as part of this learning.

K-Means Clustering:



K-means clustering with 6 clusters of sizes 5247, 3475, 1217, 3591, 1483, 631 are formed. A plot for clustering data based on clusters and another plot for the actual data based on overall_rating are checked.

Logistic Regression:



The following Graph is ROCR curve for the logistic Regression. As the curve is more towards true positive rate the accuracy of prediction by logistic regression is high.

Note: We have tried to implement the association rules but we were not successful because all the important variables were continuous and the rules were redundant and we got zero rules when we pruned the rules. So the Association Rules is not suitable for our dataset.

2.2.5 Phase 5: Evaluation and Deployment Phase:

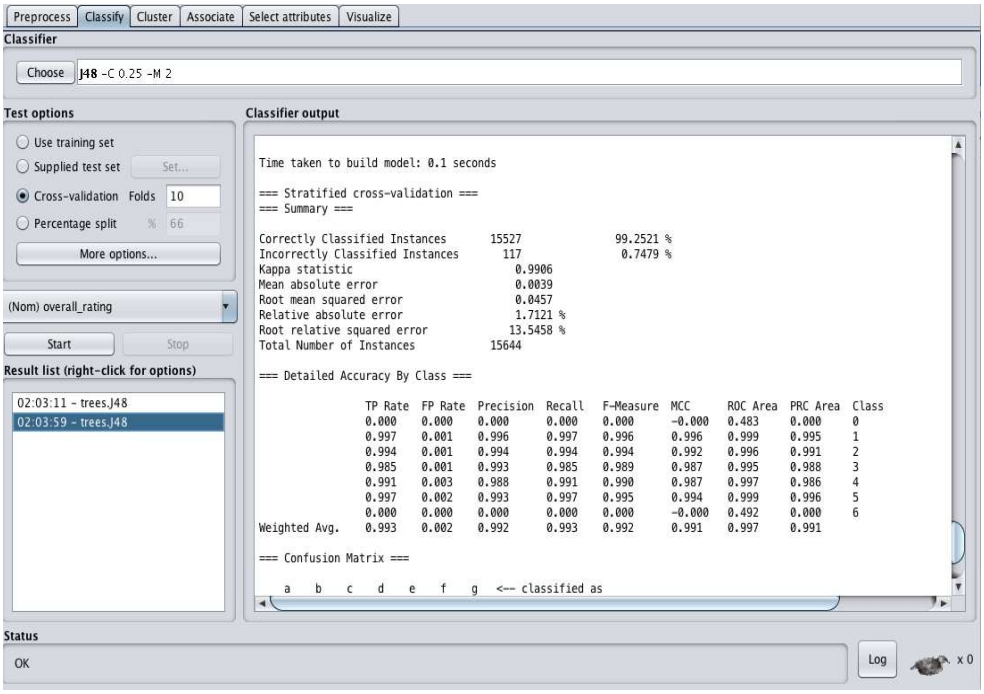
In this phase, we have evaluated the various models we trained to determine the quality and effectiveness of the model.

Accuracy of the models:

Decision tree Algorithm Accuracy:

```
Time taken to build model: 0.14 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances        4654               99.169  %
Incorrectly Classified Instances        39                0.831  %
Kappa statistic                          0.9896
Mean absolute error                      0.004
Root mean squared error                  0.0481
Relative absolute error                  1.7794 %
Root relative squared error             14.2746 %
Total Number of Instances             4693

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000     0.000     0.000       0.000   0.000      0.000    ?         ?         0
                 0.996     0.001     0.993       0.996   0.995      0.994    0.999     0.992     1
                 0.988     0.001     0.995       0.988   0.992      0.990    0.994     0.990     2
                 0.983     0.002     0.993       0.983   0.988      0.985    0.995     0.985     3
                 0.994     0.004     0.986       0.994   0.990      0.986    0.996     0.986     4
                 0.997     0.002     0.994       0.997   0.995      0.994    0.999     0.997     5
                 0.000     0.000     0.000       0.000   0.000      0.000    0.980     0.011     6
Weighted Avg.    0.992     0.002     0.992       0.992   0.992      0.990    0.997     0.990

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
    0    0    0    0    0    0    0 |    a = 0
    1  746    2    0    0    0    0 |    b = 1
    0    5  853    5    0    0    0 |    c = 2
    0    0    2  852   13    0    0 |    d = 3
    0    0    0    1 1099    6    0 |    e = 4
    0    0    0    0    3 1104    0 |    f = 5
    0    0    0    0    0    1    0 |    g = 6
```

- As we can see from the output, the root mean square error is 0.0481.
- We can see that 99.169 % it predicted correctly which is very effective.
- Based on the accuracy, we thought that overfitting might have happen and model might have tried to memorise the things. But we have achieved the same accuracy when we run the 10 fold cross-validation in Weka.



Neural Networks Algorithm Accuracy:

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances        3644               93.1731 %
Incorrectly Classified Instances       267                6.8269 %
Kappa statistic                          0.914
Mean absolute error                      0.0234
Root mean squared error                  0.1266
Relative absolute error                 10.2694 %
Root relative squared error             37.5586 %
Total Number of Instances             3911

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         0
                 1.000    0.002    0.989      1.000    0.994      0.993    0.999     0.992     1
                 0.973    0.013    0.940      0.973    0.956      0.946    0.982     0.894     2
                 0.880    0.003    0.983      0.880    0.929      0.916    0.935     0.917     3
                 0.907    0.040    0.876      0.907    0.891      0.856    0.955     0.845     4
                 0.921    0.029    0.909      0.921    0.915      0.888    0.996     0.989     5
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         6
Weighted Avg.    0.932    0.020    0.933      0.932    0.932      0.913    0.973     0.925

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
   0   0   0   0   0   0   0 |   a = 0
   0 623   0   0   0   0   0 |   b = 1
   0   7 672  10   2   0   0 |   c = 2
   0   0  43 640  44   0   0 |   d = 3
   0   0   0   1 848  86   0 |   e = 4
   0   0   0   0  74 861   0 |   f = 5
   0   0   0   0   0   0   0 |   g = 6
```

- As we can see from the output, the root mean square error is 0.1266.
- We can see that 93.17 % it predicted correctly.

Logistic Regression Algorithm Accuracy:

```
Time taken to build model: 9.19 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances        3123               79.8517 %
Incorrectly Classified Instances       788               20.1483 %
Kappa statistic                          0.7461
Mean absolute error                      0.0729
Root mean squared error                  0.1969
Relative absolute error                 32.0783 %
Root relative squared error             58.4397 %
Total Number of Instances             3911

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         0
                 0.838    0.026    0.857      0.838    0.847      0.819    0.992     0.958     1
                 0.725    0.068    0.696      0.725    0.710      0.647    0.969     0.875     2
                 0.662    0.051    0.747      0.662    0.702      0.640    0.959     0.855     3
                 0.843    0.078    0.773      0.843    0.806      0.743    0.955     0.844     4
                 0.889    0.029    0.905      0.889    0.897      0.865    0.988     0.961     5
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         6
Weighted Avg.    0.799    0.051    0.799      0.799    0.798      0.748    0.972     0.898

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
   0   0   0   0   0   0   0 |   a = 0
   0 522 101   0   0   0   0 |   b = 1
   0  87 501 103   0   0   0 |   c = 2
   0   0 118 481 128   0   0 |   d = 3
   0   0   0  60 788  87   0 |   e = 4
   0   0   0   0 104 831   0 |   f = 5
   0   0   0   0   0   0   0 |   g = 6
```

- As we can see from the output, the root mean square error is 0.1969.
- We can see that 80 % it predicted correctly.

K-Means Algorithm Accuracy:

```
Time taken to build model (full training data) : 0.18 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       2790 ( 18%)
1       2385 ( 15%)
2       2481 ( 16%)
3       3378 ( 22%)
4       2284 ( 15%)
5       2326 ( 15%)


Class attribute: overall_rating
Classes to Clusters:

    0    1    2    3    4    5  <-- assigned to cluster
    3    0    0    0    0    0 | 0
 1971   56  280    0   15  195 | 1
  778  296 1106    0  315  449 | 2
   37  602  630    2  866  785 | 3
    1 1035  452  512  901  689 | 4
    0  396   13 2862  187  208 | 5
    0    0    0    2    0    0 | 6

Cluster 0 <-- 1
Cluster 1 <-- 4
Cluster 2 <-- 2
Cluster 3 <-- 5
Cluster 4 <-- 3
Cluster 5 <-- No class

Incorrectly clustered instances :       7804.0   49.8849 %
```

The model is 50 % accurate. The clustering method was not able to predict the patterns in the data may be because of large data that is overlapping.

Observations from Evaluation and Deployment Phase:

Results:

- Most of the nursing homes in the states Guam and Virgin Islands have more number of deficiencies and this should be overcomed.
- The number of complaints coming from nursing homes are more in the states of California, Illinois and Texas.
- The top 10 deficiencies identified through the associated tag are :

	A	B	C
1	tag	tag_desc	count(*)
2	323	Ensure that a nursing home area is free from accident hazards and provide adequate supervision to prevent avoidable accidents.	19765
3	441	Have a program that investigates, controls and keeps infection from spreading.	19497
4	371	Store, cook, and serve food in a safe and clean way.	17367
5	309	Provide necessary care and services to maintain or improve the highest well being of each resident .	16695
6	62	Automatic sprinkler systems that have been maintained in working order.	13593
7	147	Properly installed electrical wiring and equipment.	13276
8	329	Ensure that each resident's 1) entire drug/medication regimen is free from unnecessary drugs; and 2) is managed and monitored to achieve highest level of well-being.	11110
9	279	Develop a complete care plan that meets all the resident's needs, with timetables and actions that can be measured.	10967
10	431	Maintain drug records and properly mark/label drugs and other similar products according to accepted professional standards.	10572
11	29	Special areas constructed so that walls can resist fire for one hour or an approved fire extinguishing system.	10343

- We determined the top 10 nursing homes in the country based on the overall_rating which has 5 rating.

provnum	PROVNAME
15009	BURNS NURSING HOME, INC.
15024	SENIOR REHAB & RECOVERY AT LIMESTONE HEALTH FACILI
15042	SUMTER HEALTH AND REHABILITATION, L L C
15066	TERRACE MANOR NURSING & REHABILITATION CENTER, INC
15073	HANCEVILLE NURSING & REHAB CENTER, INC
15084	PARK PLACE NURSING AND REHABILITATION CENTER, LLC
15089	EVERGREEN NURSING HOME
15098	ALLEN MEMORIAL HOME
15101	ALTOONA HEALTH & REHAB
15112	MAGNOLIA HAVEN HEALTH AND REHABILITATION CENTER

Reasons identified for the complaints and deficiencies:

- May be due to states have less funding for maintaining the nursing homes.
- May be the nursing homes are less spread and over crowded as the states have more population.

Reasons identified for Least performing nursing homes:

- The reason being deficiencies not corrected even though they have been reported to them to correct.
- After getting the certification, they might not maintain the nursing homes properly
- The nursing home staff might not be performing up to mark

At their best :

- The best nursing homes are mostly in the state of Washington DC.
- The number of certified beds are more in the state of California.

Reasons for best performing nursing homes:

- They being in the primary location of the state may be the reason to get the money they need to maintain the quality of nursing home.
- The deficiencies identified are rectified immediately before the next cycle inspection occurs
- As the number of complaints identified are less, the total amount of penalty paid is less

Chapter 3: Future Scope and References

Future Scope:

- We have considered only overall rating for determining the best nursing homes but we can determine the best nursing homes more effectively by considering the quality related data.

- The nursing home's performance might also depend on type of services they provide which can be considered.

References:

- R bloggers
  https://www.r-bloggers.com/

- To understand the column names associated with the dataset we referred
  https://www.medicare.gov/nursinghomecompare/Data/About.html

- For syntactical help in R
  https://cran.r-project.org/