HOSPITAL RATINGS PROJECT

Group Members:

1. Aman Patni
2. Roopa
3. Ranganath S
4. Thejo Kishan

Business Use Case – CMS Hospital Rating

- CMS, an US Federal Agency's responsibilities include administration of key public health programs, administrative simplification standards from the HIPAA, quality standards in long-term care facilities / nursing homes through its survey and certification process, clinical laboratory quality standards under the Clinical Laboratory Improvement Amendments, and oversight of HealthCare.gov.
- CMS launched "hospital compare" project to device a methodology that would
  - Rate providers on a scale of 1-5 both within multiple categories and overall
  - One-stop solution enabling end consumers / patients choose across providers that would suite their needs
  - Induce a healthy competition across providers and create value for public money
- CMS rates hospitals in US based on various categories including quality, effectiveness etc. The performance measures are organized into various groups as listed below:
  - ❑ General information
  - ❑ Survey of patients' experiences
  - ❑ Timely & effective care
  - ❑ Complications
  - ❑ Readmissions & deaths
  - ❑ Use of medical imaging
  - ❑ Payment & value of care

Our Approach

Objective: The objective of our analysis are two-fold:

- Identify and develop the approach to calculate the hospital ratings
- Deduce recommendations for a provider suggesting the areas of improvement to improve their CMS Ratings

Objective 1: To identify and develop the approach/model to calculate the hospital ratings is divided into multiple steps as listed below:

- Step 1: Data Understanding
  - Research on CMS methodology & Analyze various attributes within Source Data Files and capture our observations
- Step 2: Data Cleaning
  - Outlier Handling
  - Missing Value Elimination or Imputation
- Step 3: Exploratory Analysis
  - Identify key attributes impacting the scores or rating within each group
- Step 4: Data Preparation
  - Normalization data scales
  - Retain only key attributes impacting group scores for model building
- Step 5: Model Building:
  - Tried multiple models including supervised models (random forest & k-means) and unsupervised models (FA, PCF) to arrive at a best model
- Step 6: Model Evaluation:
  - Accuracy has been considered as one of the key metrics in choosing the best model
  - Confusion matrix & comparison of our ratings with CMS ratings have been carried out

Our Approach

Objective 2: To Deduce recommendations for a provider suggesting the areas of improvement to improve their CMS Ratings

- Compare specific provider's final score with median final scores of group 4 and group 5 providers to inspect the variance between them
- Compare it at group level with median values of each of these groups of Group 4 and Group 5 providers to inspect the variance between them
  - Visual Inspection: Box Plots have been leveraged for this step
  - Which groups are they having lower scores.
  - Then drill down in that group for important variables(random forest).
  - Start comparing their score with variable level with others
- Lagging variables based on scores can be recommendations
  - 5 to 6 important variable can be suggested

Data Understanding – Summary & Observations

- CMS Ratings depend on 64 variables grouped into 7 groups carrying specific weightage
- Group Weights:

Group	Weightage (%)
Readmission	22
Mortality	22
Safety of Care	22
Patient Experience	22
Timeliness	4
Imaging Efficiency	4
Effectiveness	4

Data Understanding – Summary & Observations (Contd.)

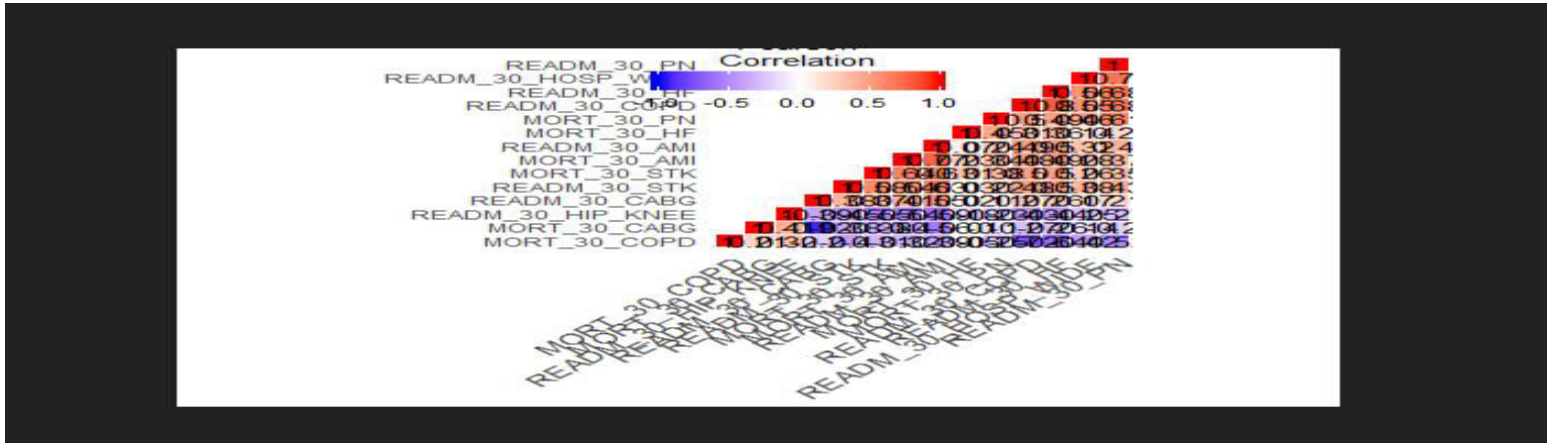
- Measures:
  - Positively correlated measures: Measures reflecting effective and timely treatment have a positive impact on rating
  - Negatively correlated measures: Measures reflecting mortality and readmission rates have a negative impact on rating

Data Cleaning & Preparation – Summary & Observations

- Missing value treatment:
  - Missing values have been imputed with median values rather than eliminating as most of the variable have this problem and eliminating them would reduce the dataset size drastically
- Standardization:
  - All attributes have been transformed so that higher values indicate higher rating (for both positive & negative co-related measures/ attributes)
- Formatting
  - Input data has been transformed from wide format to narrow format
  - And all the measures related to specific group have been extracted to same file
  - All groups have been merged to form the master file
  - All score columns have been made numeric

Exploratory Data Analysis –Group data

- EDA for each Group
  - ✓ Missing values in each attribute of the Group is taken care.
  - ✓ Correlation matrix is build among the attributes in each Group
  - ✓ The magnitude of correlation is identified for each group using the heat maps .
  - ✓ Plot for each group is identified as below.



Model selection – Supervised Approach

- ✓ Logistic Regression and Support Vector Machines(SVM) are considered for model building but observe the following pointers .
- ✓ Logistic Regression:

Cons:

- ❑ In the current data set we have 64 variables and it was understood LR couldn't have large number of Categorical variables

- ✓ Support Vector Machines(SVM)

Cons:

- ❑ Not very efficient with many variables in the data set.
- ❑ Difficult to find the appropriate kernels.

Model Building and Evaluation – Random Forest

- ✓ Random Forests is considered for the model building based on the following parameters
- ❑ Variable interactions can be handled.
- ❑ Problem of over fitting is handled by using the group of models(decision trees).

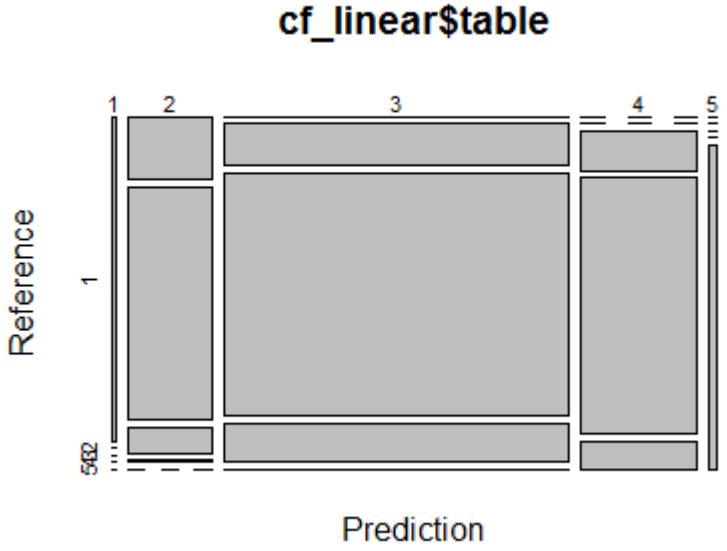
Model Building & Evaluation :

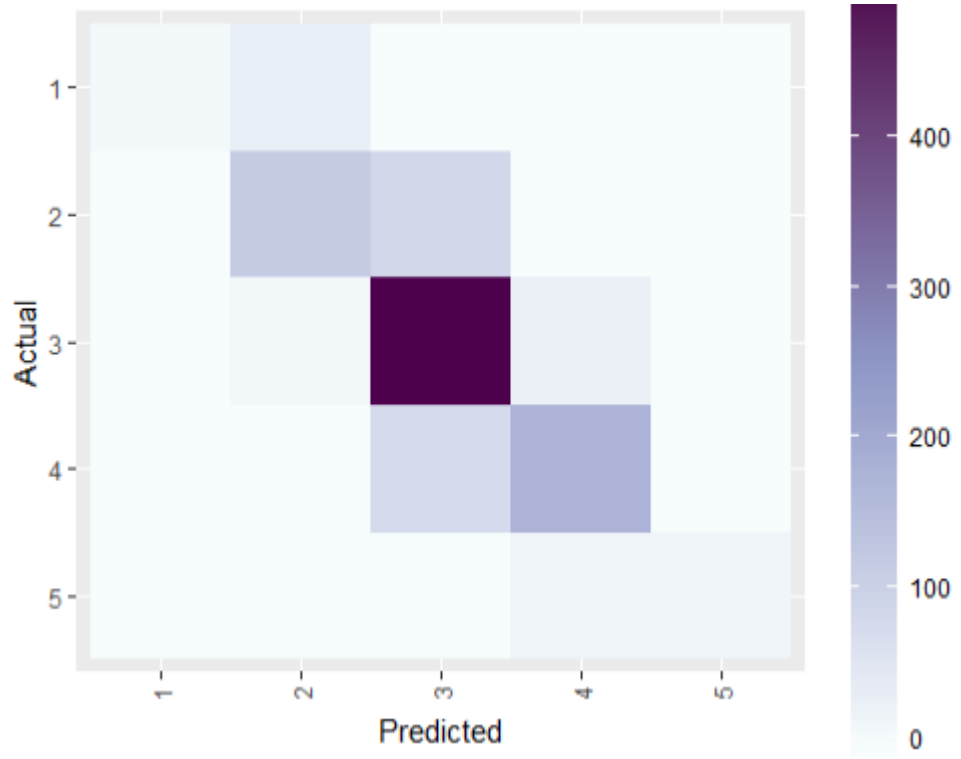
- ❑ Train & Test data set is split in certain ratio.
- ❑ The hyperparameters with the number of trees =800 and the number of variables at each split is defined as random.
- ❑ Cross validation is done using GBM .
- ❑ Out of bag estimate of error is 23.52 % which measures the error at each tree of the data set committed.
- ❑ Confusion Matrix gives that the smaller/larger classes tend to give high class.error.
- ❑ Simple Cross Validation and with GBM have been explored

Model Building and Evaluation – Random Forest Prediction Analysis

✓ Prediction Summary:

```
summary(predictdf)
  1   2   3   4   5
  9 164 678 227  16
```





Model Building and Evaluation - Random Forest Cross Validation Results

Confusion Matrix and Statistics

Reference

Prediction	1	2	3	4	5
1	5	0	0	0	0
2	25	126	9	0	0
3	1	81	470	113	0
4	0	0	34	193	18
5	0	0	0	3	16

Overall Statistics

Accuracy : 0.7404
95% CI : (0.7133, 0.7662)
No Information Rate : 0.4689
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5836
Mcnemar's Test P-Value : NA

Model Building and Evaluation - Random Forest Cross Validation Results Contd.

Statistics by Class:

#	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
#Sensitivity	0.16129	0.6087	0.9162	0.6246	0.47059
#Specificity	1.00000	0.9617	0.6644	0.9338	0.99717
#Pos Pred Value	1.00000	0.7875	0.7068	0.7878	0.84211
#Neg Pred Value	0.97612	0.9133	0.8998	0.8634	0.98326
#Prevalence	0.02834	0.1892	0.4689	0.2824	0.03108
#Detection Rate	0.00457	0.1152	0.4296	0.1764	0.01463
#Detection Prevalence	0.00457	0.1463	0.6079	0.2239	0.01737
#Balanced Accuracy	0.58065	0.7852	0.7903	0.7792	0.73388

Model Selection – Unsupervised Approach

✓ Unsupervised Learning algorithms approach :

- ❑ We explored K means clustering and hierarchical clustering models to assign hospital ratings
- ❑ The hospital ratings are calculated based on the weight scores assigned to each hospital from the 7 groups.
- ❑ Factor analysis/latent variable analysis are used to calculate the weight of each measure in the 7 groups.
- ❑ Final ratings derived based on the scores from each group and hospital

Model Building and Evaluation – KMeans

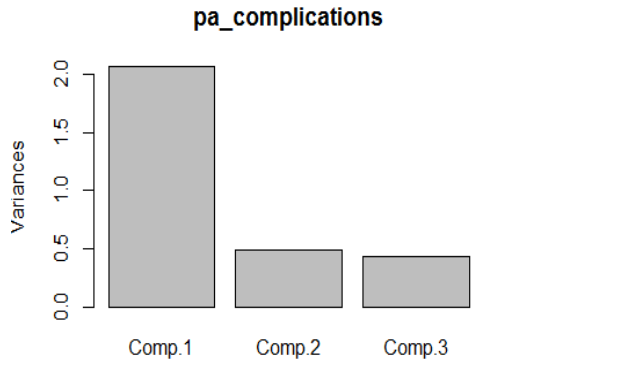
- ✓ Random Forests is considered for the model building based on the following parameters
- ❑ Variable interactions can be handled.
- ❑ Problem of over fitting is handled by using the group of models(decision trees).

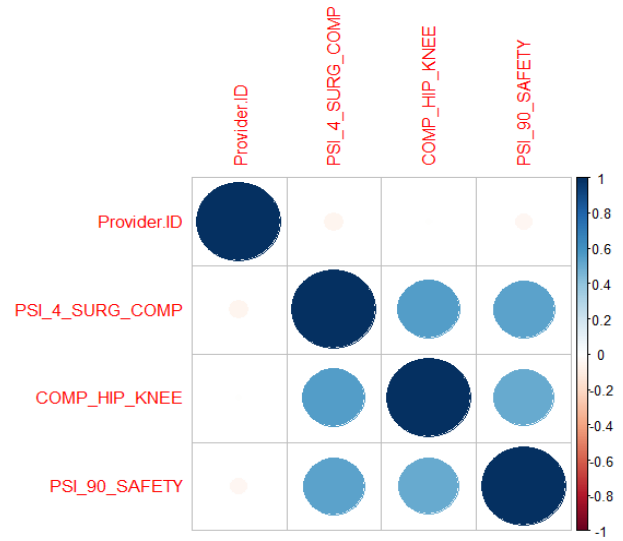
Model Building & Evaluation :

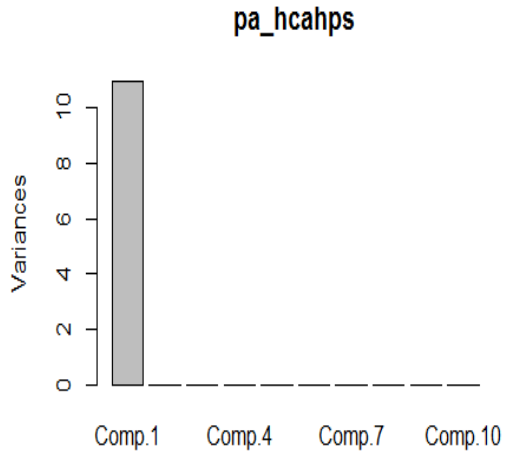
- ❑ Train & Test data set is split in certain ratio.
- ❑ The hyperparameters with the number of trees =800 and the number of variables at each split is defined as random.
- ❑ Cross validation is done using GBM .
- ❑ Out of bag estimate of error is 23.52 % which measures the error at each tree of the data set committed.
- ❑ Confusion Matrix gives that the smaller/larger classes tend to give high class.error.

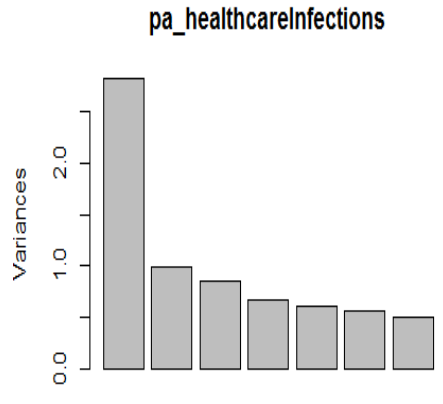
Model Building and Evaluation – Factor Analysis

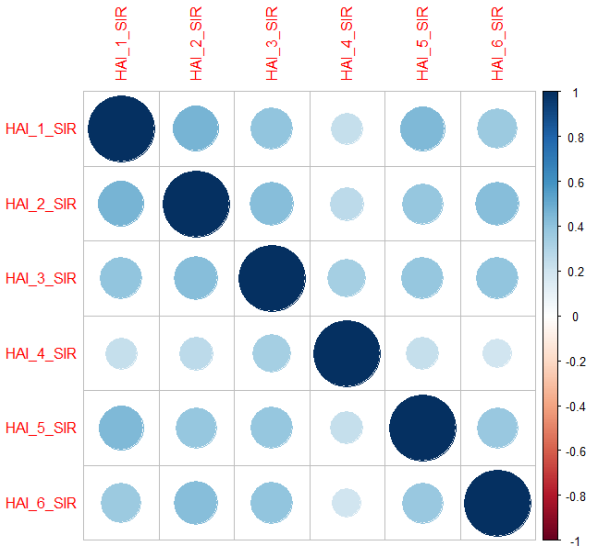
Visual inspection of principal components for different groups to find the important factors



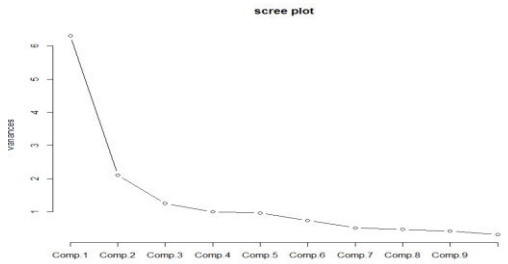


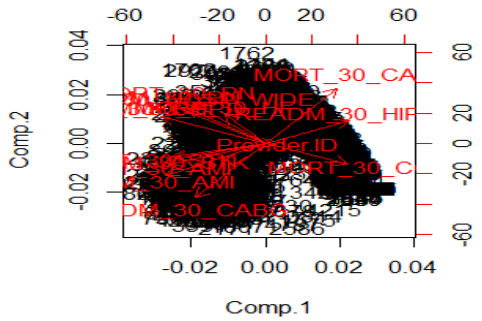


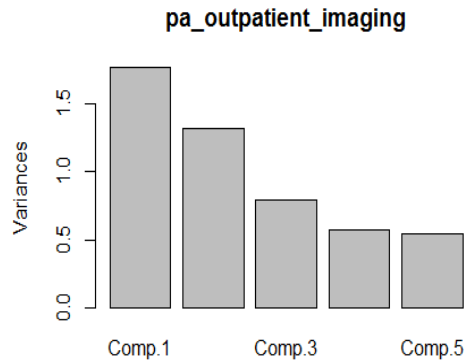


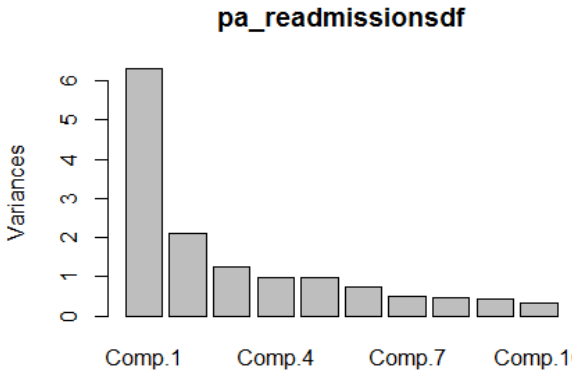


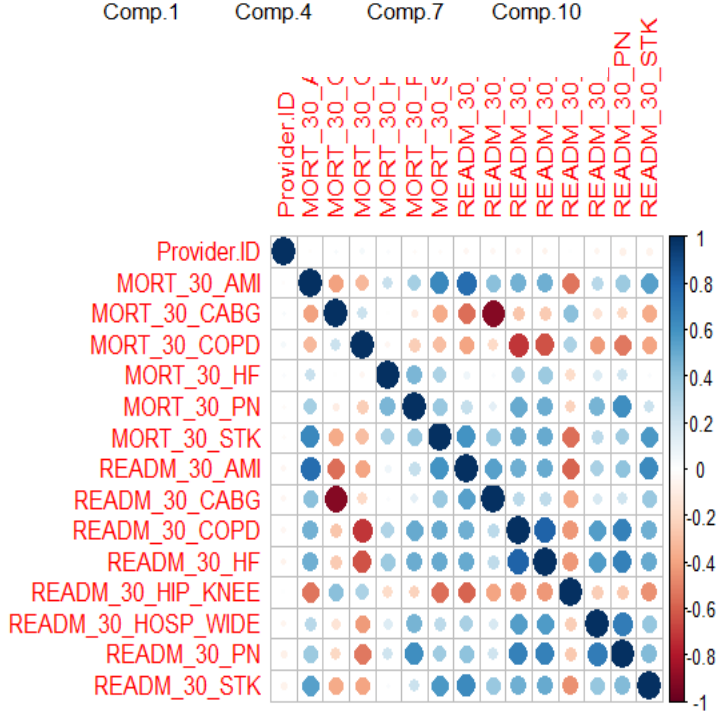
Model Building and Evaluation – Factor Analysis Contd.

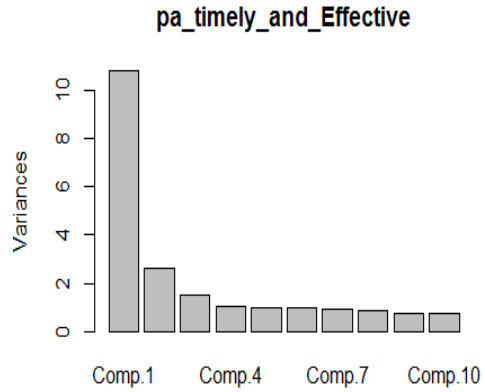


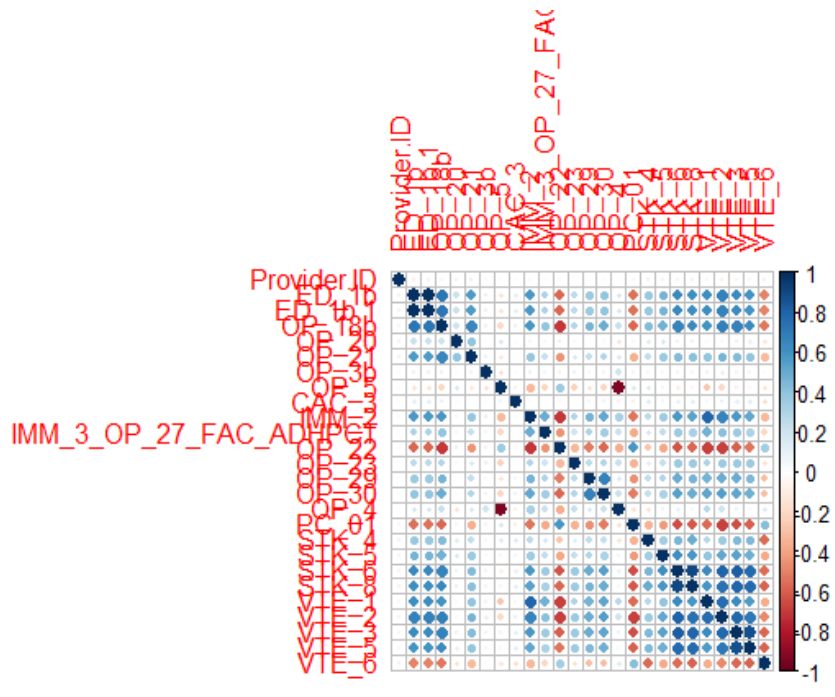




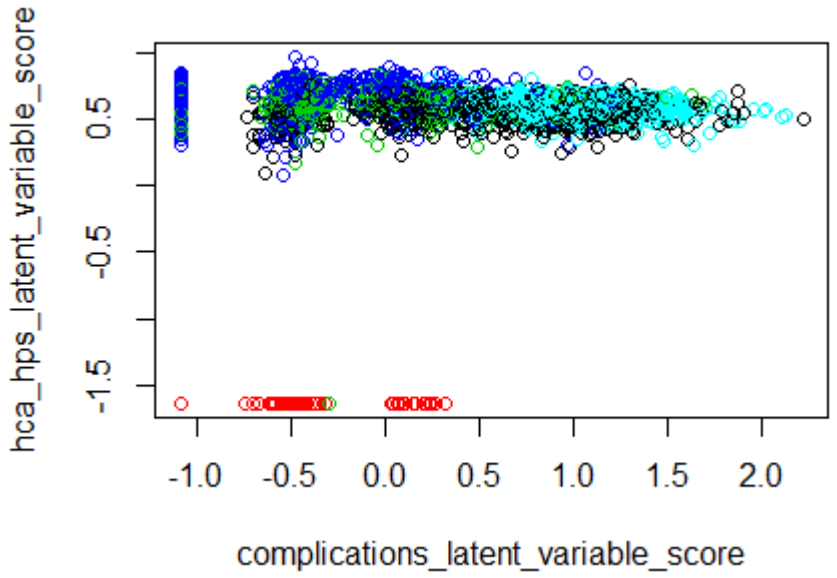


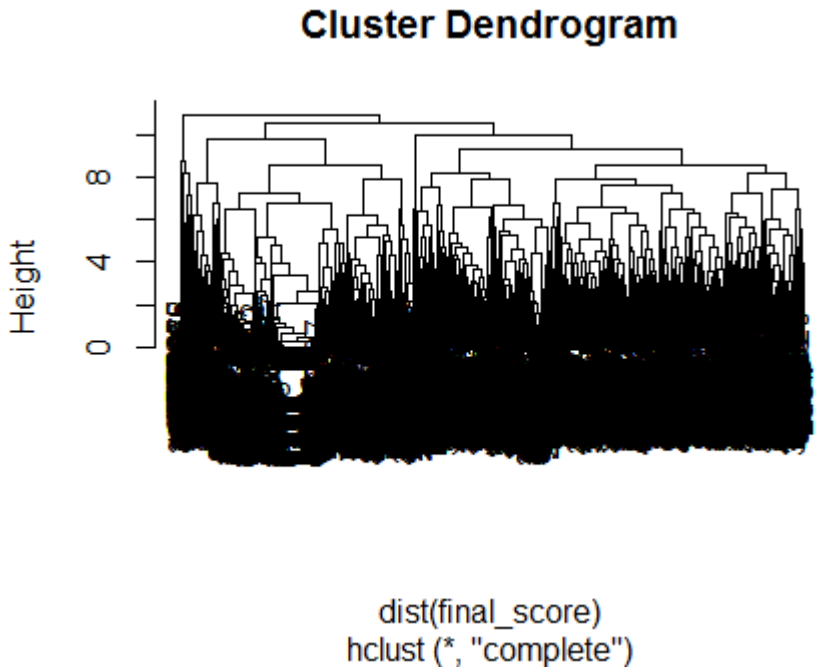






Model Building and Evaluation – KMeans & HClust Clustering Model Output





Recommendations for Hospitals

- Using Factor Analysis the attributes having same patterns or trend are extracted from each group .
- With the Current analysis on the given data set it was derived that certain measures under the Complications , hca_hps , healthcare infections, readmissions , time & effective and Outpatient Imaging helps improving the ratings of the hospitals .