

National Diabetes Inpatient Audit

Principal Component Analysis - Assignment

**Report by:**

Balagopal Unnikrishnan (E0267709)

Chokkalingam (E0267541)

Gananathan (E0267639)

Kenneth Rithvik (E0267759)



# Index

1. Introduction
2. Data Description
3. Data Understanding
4. Data Cleaning
   1. Missing Data
   2. Distribution and Outliers
   3. Variable Selection
5. Dimension Reduction
   1. Principal Component Analysis
   2. Eigen Values
   3. Scree Plots
   4. Component Selection
   5. Component Profiling
6. Clustering
   1. Number of clusters
   2. Description of clusters
7. Regression Analysis
8. Conclusion

# **Introduction**

The main purpose of this project is to do dimension reduction using principal component analysis/factor analysis to understand the key factors that can be improved to increase the overall satisfaction of diabetic patients. Hospitals were clustered using these factors and each resulting cluster was analyzed to find their characteristics. A prediction model was built to predict the overall satisfaction of diabetic patients for the care they receive at hospitals in England and Wales.

For the analysis, we have made use of a data set from the Government data foundation of United Kingdom. The data set, titled “**The National Diabetes Inpatient Audit (NaDIA)**” in total contains 47 variables and 1088 observations. Using Dimension Reduction techniques and clustering, the major variables that define the quality of care provided to diabetic patients in the UK were understood.

# **Dataset Description**

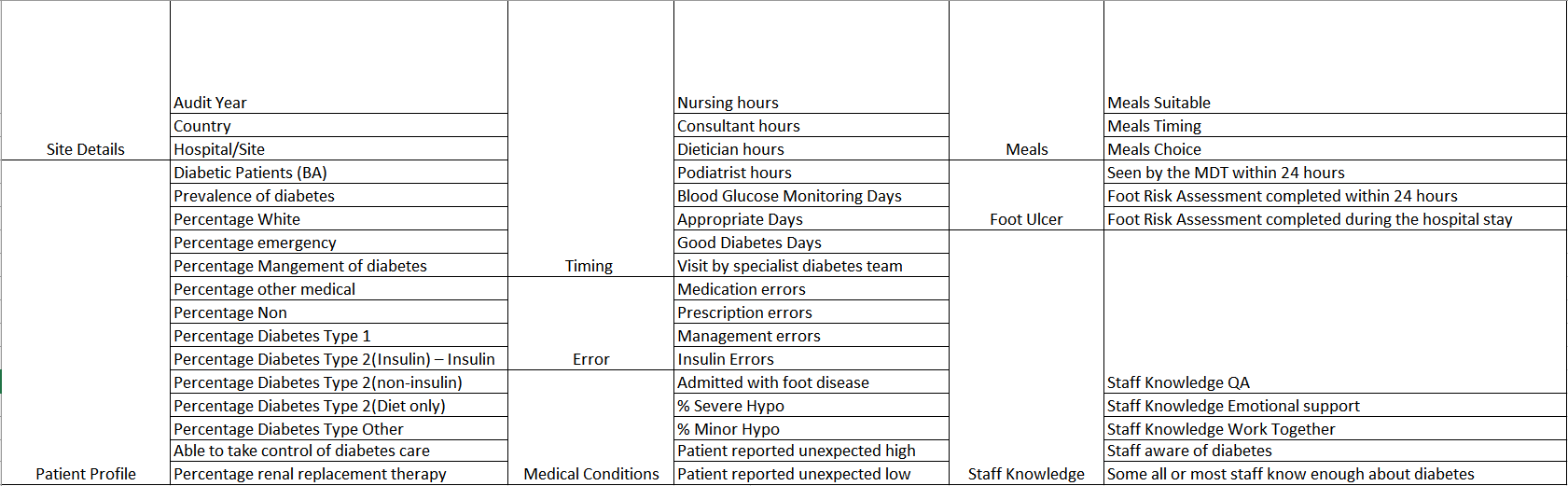
The National Diabetes Inpatient Audit (NaDIA) studies inpatient care for diabetes patients in the UK and Wales. It is commissioned by the Healthcare Quality Improvement Partnership (HQIP) and was done by NHS Digital along with Diabetes UK.

The data collected was on factors in the hospital like:

* Patient experience information
* Patient clinical data
* Staffing structures

There are data records available from 2011, 2012, 2015 and 2016. Around 209 hospital sites in England and Wales took part in the study. All these datasets were combined into one for analysis.

# **Data Understanding**



# **Data Cleaning**

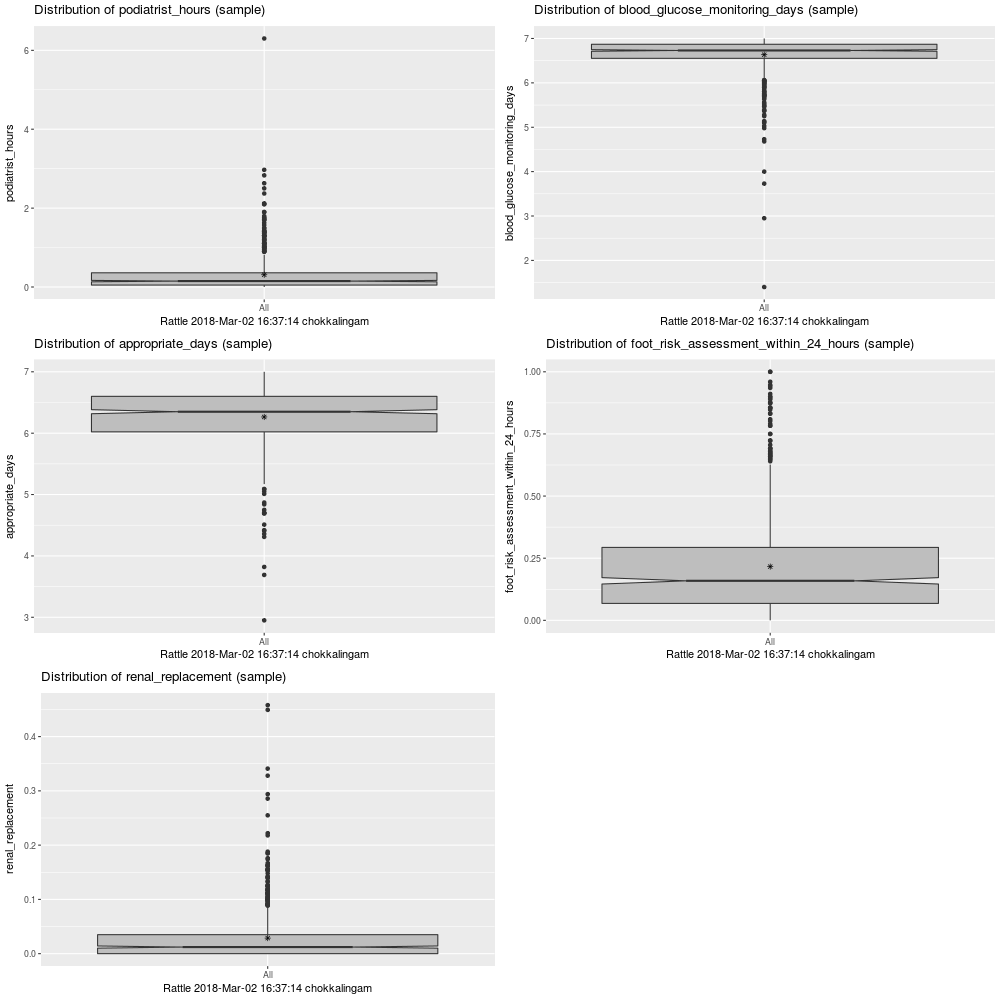
|  |  |
| --- | --- |
| ITEM | STATS |
| Number of variables | 47 |
| Number of samples | 1088 |
| Rows with partial data | 380 |
| Number of Target Variables | 1 |

## **Missing Data**

Initial Exploratory data analysis identified variables with high degree of missing data. Before performing principal component analysis on our dataset, it is necessary to impute the missing data values so that we can guarantee a higher degree of accuracy for our analysis. We had 380 missing values in our data set and we imputed these values using multivariate imputation technique which is “Multivariate Imputation by Chained Equations” (MICE). MICE gave better results compared to KNN imputation in preserving the underlying distribution of the data.

## **Distribution and outliers**

The inter quartile range of each variable was calculated and used to identify the outliers and remove them.

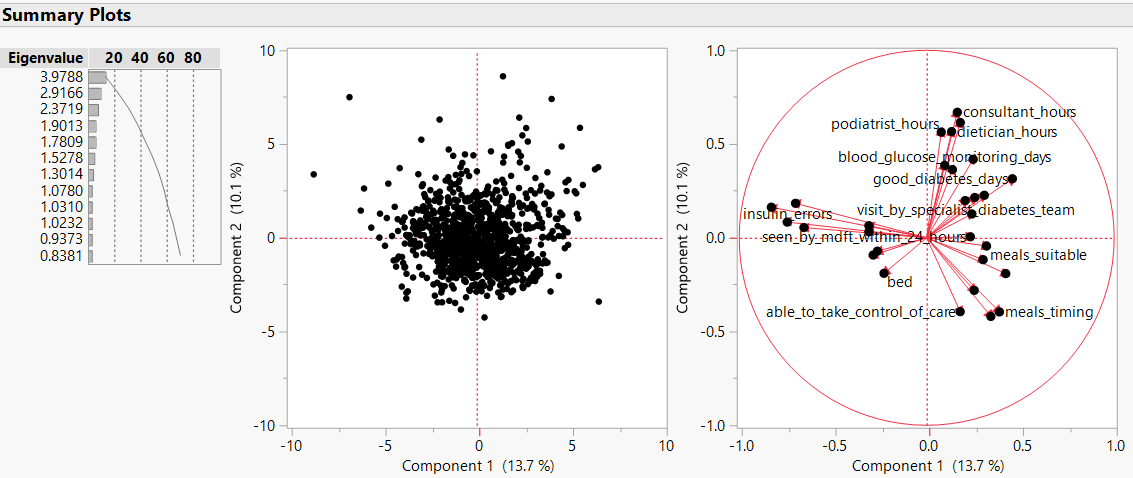
## **Variable Selection**

There were 48 variables present initially in the dataset. Since our objective is to analyze and improve the underlying factors influencing patient’s satisfaction, we selected variables on which the hospital does have control over. Hence factors like country, prevalence of diabetes, percentage distributions of emergency-care patients, white patients, severity of hypoglycemia were removed. This reduced the dataset to 26 attributes.

# **Dimension Reduction**

## **Principal Component Analysis**

We use principal component analysis to get reduced number of independent linear combinations (principal components) that can be used to project as much of the variability in the original variables as possible. These are the summary plots generated after applying the appropriate settings and carrying out PCA.



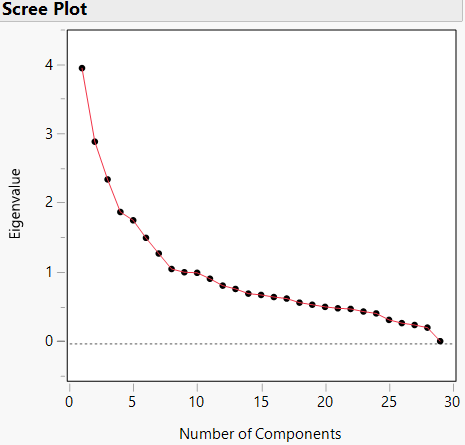
After performing PCA, we were able to analyze the total variance between the dimensions within our dataset. The Principal Component Analysis also helps us decide the number of principal components based on the largest eigenvalues obtained. This gives us insight about the direction in which the largest variance is present. To decide the number of total principal components, we searched the Eigenvalues table for the components with eigenvalue greater around 1.

## **Eigenvalues**

From the Eigenvalues table, we observed that the first 8 principal components were adequate in explaining 61.17 % of the cumulative variance. Hence, from the Eigenvalues Table, we have selected the total number of Principal Components to be 8. We can also double check with the findings from the corresponding Scree Plot generated.

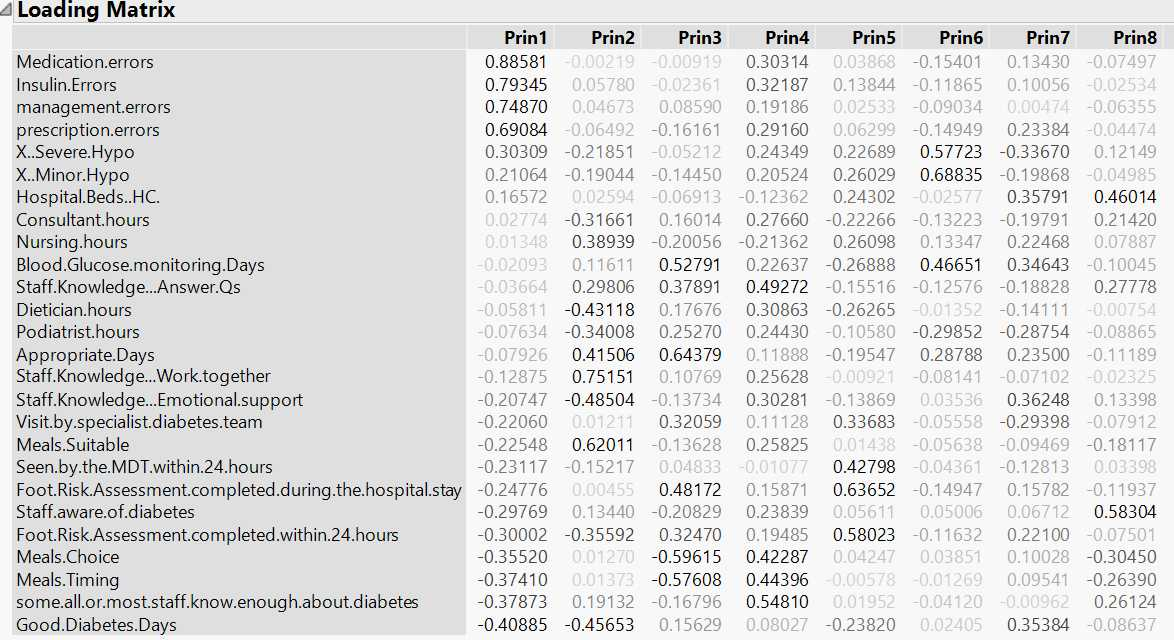
## **Scree plots**

We analyzed the Scree Plot to figure out the value at the elbow of the graph to verify the number of Principal Components required. The value at the elbow point coincides with the selected number from the Eigenvalues Table. We observed an elbow at the point corresponding to 8 principal components.



## **Component Selection**

Out of the resulting 26 principle components, we chose the top 8 principal components with Eigen values greater than 1 and this corresponds to 61.17 % coverage.



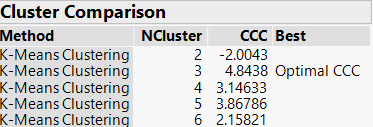
## **Component Profiling**

Based on the type of columns involved in each type of the clustering done, we have assigned names for all the selected components, and for the type of care that the diabetic patient received. We have:

|  |  |  |
| --- | --- | --- |
| PC | Assigned Name | Description |
| 1 | Errors in Medical care | Covers medical, insulin and management errors |
| 2 | In-ward convenience | Interpret how well ward service and meal plans are handled |
| 3 | Monitoring Frequency | Reflects how well the monitoring schedule was followed |
| 4 | Diabetic care specialization | Shows how well the staff were equipped to treat diabetics |
| 5 | Foot Ulcer assessment | Covers how soon and well the foot risk assessment was carried out |
| 6 | Hypo Glycaemia occurrence | Covers the preventive measure carried out to avoid glycaemia |
| 7 | Personal/Emotional space | Specifies how well the patient is put to ease during the stay |
| 8 | Staff awareness about Diabetic care | Specifies how well the general staff is aware about diabetic care |

# **Clustering**

## **Number of clusters**

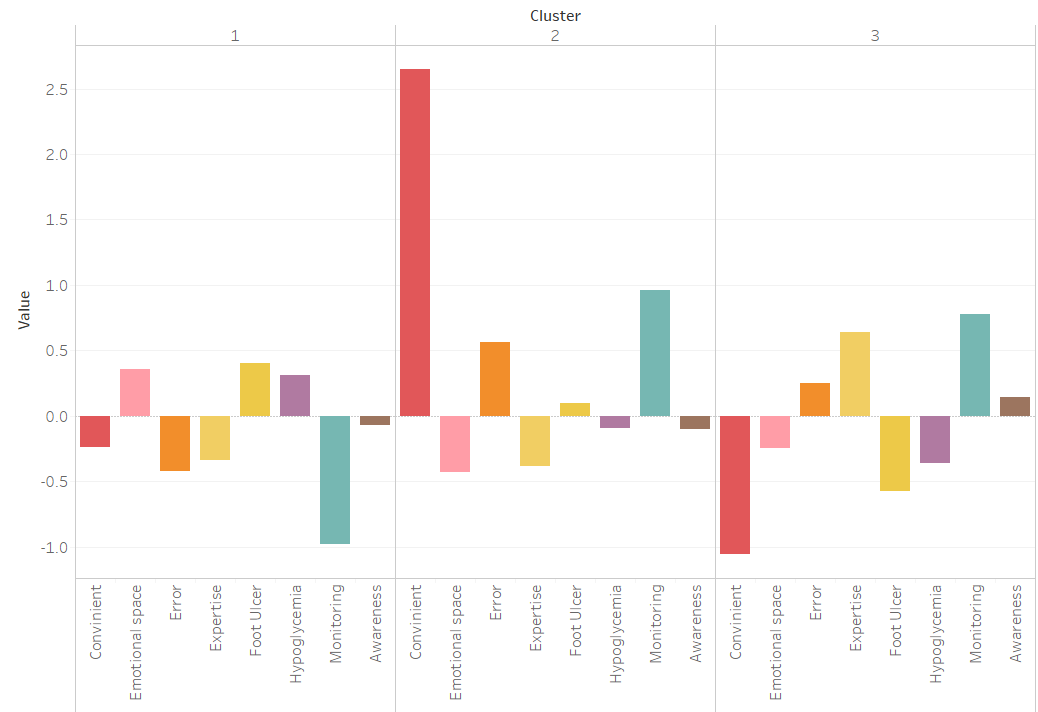


We consider the top 8 Principal components from above to use for clustering. K-means clustering is used for K values in the range of 2 to 6. Of this, number of clusters were fixed to 3 for the following reasons

The clusters were well distributed, and the cluster centers were not overlapping and also an optimal CCC was obtained for 3 clusters

## **Description of clusters**

Based on the clusters obtained and the summary using the principal components, the key characteristics of each were examined



## **Cluster 1- Feel Good Hospitals**

Even though the hospitals in this cluster were characterized by low rates of monitoring and a lack in expertise compared to the other two clusters, it still had a good overall experience rate. This is because these hospitals had a low error rate and high emotional connect with the patients, thereby counter balancing the above mentioned negative factors.

## **Cluster 2 – Convenience Hospitals**

These hospitals had a proper monitoring schedule and the patients felt the infrastructure and service to be very convenient. However, due to low expertise and awareness levels among the staff, there were high error rates.

## **Cluster 3 - Specialist Centers**

This cluster was characterized by a high-level of expertise, a regular monitoring system and high awareness among staff about diabetes. This resulted in lower cases of Foot ulcer and Hypo Glycemic conditions. The hospitals in this cluster tend to not give much room for emotional space for patients and compromises on their convenience.

# **Regression Analysis**

The regression analysis was done with Overall Satisfaction as the target variable. The quality of care can be thought of as the level of satisfaction that the diabetic patient feels he has received. Using the derived factors and linear regression, we will now test our model for its performance.

* The R2 Square and the Mean Square Error MSE was taken while taking Principal Components 1 to 10
* We can see that the model's performance stabilizes and does not improve much beyond the point where the first 4 principal components are taken.
* Initial R2 Score on the clean data: 0.330 and the R2 score on the model which uses the first 4 principal component as input: 0.345. We can see that after dimension reduction and PCA, representative nature of the variables was not lost.

|  |  |
| --- | --- |
|  |  |

# **Conclusion**

On the data set, titled “**The National Diabetes Inpatient Audit**” containing 47 variables and 1088 observations, we applied dimension reduction techniques and were able to correctly project the major part of our dataset onto the principal components and determine how they correlate with the quality of diabetic patient care in a particular medical institution. We reduced the total number of input variables from 47, to just 8 principal components.

Using these 8 principal components we were able to cluster our dataset into three distinct clusters that helped us in identifying the differences and grouping of our dataset. We were then able to apply linear regression to predict the patient satisfaction level based on the treatment given as well.