HOWARD UNIVERSITY

Identifying Subgroups of Minority Diabetes Type II Data Using Cluster Analysis

A Thesis
Submitted to the Faculty of the
Graduate School

Of

HOWARD UNIVERSITY

in partial fulfillment of
the requirements for the
degree of

MASTER OF COMPUTER SCIENCE

Department of Electrical Engineering
And Computer Science

by

Tacuma Kwabena Solomon

Washington, D.C.
May 2017

HOWARD UNIVERSITY
GRADUATE SCHOOL
DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE

THESIS COMMITTEE

Peter Keiller, D.Sc.
Chairperson

Legand Burge Ph.D.

Mugizi Robert Rwebangira, Ph.D.

Gloria Washington Ph.D.

Mugizi Robert Rwebangira, Ph.D.
Thesis Advisor

Candidate: Tacuma Solomon
Date of Defense: May 11, 2017

ACKNOWLEDGEMENTS

I would first like to thank Dr. Rwebangira, Dr. Gulriz Kurban, and everyone in the Center for Computational Biology and Bioinformatics for their help and support. I would also like to thank my committee - Dr. Keiller, Dr. Burge, and Dr. Washington - for their time and advice. Last, but not least, I would especially like to thank the staff of the Department of Computer Science for helping me through every issue I have had since arriving on campus almost a year and a half ago. I would not have gotten this far without you all.

ABSTRACT

Diabetes type 2 occurs in African Americans at a rate higher than Non-Hispanic whites. They are characterized by higher rates of the disease, with higher rates of mortality than other ethnic groups. With advancements in medical and computational technology, more information than ever exists in the form of medical data.

The objective of this project is to perform cluster analysis on anonymized diabetes type II data from Howard University Hospital's electronic health records.

The data was first extracted from SQL, cleaned, and preprocessed. It was then uploaded into R. Four algorithms were chosen to create two, three, four, and five clusters of the data, which was then subject to comparative analysis. It was then determined that DIANA (Divisive ANAlysis) clustered the data best, and from which results were extrapolated.

It was discovered that there were high correlations between type II diabetes, hypertension, hyperlipidia, and cholesterolemia, which validated existing knowledge about African Americans most at risk for diabetes. There was also evidence of higher rates of benign neoplasm of the colon; non-cancerous colon tumors. Distinctions about other chronic diseases were made by gender and marital status. There were significantly more cases of acquired hypothyroidism cases occurring in women who are black, female, and non-single. There were elevated incidences of prostate cancer (neoplasm, malignant, of the prostate) in men who are black and non-single. Incidences of Tobacco use disorder also had higher occurrences in clusters featuring mostly single men and women. Many of these relationships remain unexplored.

Performing cluster analysis on electronic health records has enormous potential as a method of research. With advances in computational power and the proliferation of data, there is huge opportunity in mining medical data for knowledge.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figures **Page**

LIST OF SYMBOLS AND ACRONYMS

FFSGCDB	Fibromyalgia and Chronic Fatigue Syndrome Spanish Genetic Data Bank
GML	Generalized Linear Model
ICD-9	The International Classification of Diseases, Ninth Revision
WHO	World Health Organization
SAS EM	Statistical Analysis System Enterprise Manager
WEKA	Waikato Environment for Knowledge Analysis
PAM	Partitioning Around Medoids
MST	Minimum Spanning Tree
PCA	Principal Component Analysis
I2B2	Informatics for Integrating Biology and the Beside
SQL	Structured Query Language
Hclust	An agglomerative hierarchical clustering algorithm
DIANA	Divisive ANAlysis
t-SNE	t-disbributed stochastic neighbor embossing

CHAPTER 1. INTRODUCTION

1.1 Background

Type II diabetes is a metabolic disease which affects more than 29 million Americans. It is the more common form of diabetes, accounting for 90% of diabetes cases, with occurrence rates even higher over the age of 45 [1]. It is typically characterized by high levels of blood sugar, and can lead to further complications, such as stroke, heart disease, amputations in the extremities, and death. Healthy treatment and disease management is critical.

Diabetes type II is especially prevalent in the African-American community. The ethnic group is characterized by the American diabetes Association as having a higher risk of diabetes [1]. African-Americans suffer from higher rates - 1.4 to 2.3-fold compared to White Americans - suffer across a wider age group, and suffer higher morbidity and mortality rates than other ethnic groups. [1]

With the proliferation of Electronic Health Records and advances in computing power, we can utilize new methods of research to uncover new insights about diabetes. What do certain groups of diabetes sufferers have in common? How are we able to predict the likelihood that someone may have diabetes? What new treatments can we suggest? These questions and more can be answered by machine learning.

Machine learning is defined as the ability to fit existing data to models which can be used to predict the outcomes of new instances of Data. Machine learning may allow us to predict the probability that someone may have diabetes, based on characteristics from their Health Records and can provide new insights into existing Records. It can also group data by similarity, allowing researchers to spot trends in the data. Using machine learning, clustering, specifically, is the aim of the research.

This research is exploratory. By performing clustering algorithms on anonymized Electronic Health Records, we may be able to learn more about how diabetes patients are grouped, and why.

A novel element of research is the data that is being worked on. The data is anonymized Electronic Health Records collected over the period 2009 – 2013, in the I2b2 system. There is unique opportunity since the Howard University Hospital serves a 90% minority population. This allows us to gather results specifically from minorities.

To maintain ethical standards, it is important to emphasize that all Electronic Health Records are anonymized. All records are stripped of anything that can identify a patient.

CHAPTER 2. STATEMENT OF THE PROBLEM

The objective is to find subgroups within Howard University's electronic health records using cluster analysis, and to discover what characteristics these subgroups may share. Four algorithms will be used, two partitioning, and two hierarchical. This research is done with the expectation that insights will be found, so that we may understand more about diabetes type II in minority populations.

CHAPTER 3. RELATED WORK

Finding subgroups of data using clustering algorithms is an analytic process that is often used to generate subgroups from data.

Researchers used clustering to find specialized patterns and insight by segmenting the data into smaller fragments, each with specialized attributes. Different approaches produced different results by leveraging different techniques in formatting the dataset, choosing features, collecting data, the types of clustering algorithms that were used, the methodology of determining the algorithm's effectiveness, and in interpreting the clusters that result.

3.1 Data Collection

Different researchers used different methods of data collection. Some researchers collected data from disparate sites, whereas other researchers collected data from one site only. In the paper 'Characteristic evaluation of diabetes data using clustering techniques', Padmaja, Vikkurty, et al. the researchers gathered their data from the National Institute of Diabetes, Digestive, and Kidney Diseases in India in their effort to evaluate characteristics of diabetes via clustering. The Fibromyalgia and Chronic Fatigue Syndrome Spanish Genetic and Clinical Data Bank (FFSGCDB), is an online dataset that was used to find subgroups of Fibromyalgia patients. Another common dataset used for diabetes research is the Pima Indian dataset, which was used in the paper "Clustering and Classifying Diabetic Data Sets using K-means Algorithm". Data sets can also range in size; the paper 'Hidden Patterns; Clustering Diabetes Data' uses a dataset with 185,000 observations ranging over 5 years, while the smallest found sample was 1,446 observations in the FFSGCDB.

3.2 Data Preprocessing

Before loading the data into the software package of choice, the data must first be formatted. This preprocessing ensures that the data is in a form that can be adequately used by software and algorithms. In many research papers, their work was assisted by software tools. In the paper 'Hidden Patterns: Clustering Diabetes Data', Hu and Cook had to consider both the problems of missing data, and variable conversion. They used a technique called 'Tree Surrogate Imputation', an effective and widely used method replacing missing values, to help format their CDC dataset. After solving the problem for formatting however, they came upon another problem; they had to deal with a mixed dataset. They had the task of converting categorical and ordinal variables to numeric variables, since the algorithms that they wanted to use only worked for interval data. They used rank ordering for the ordinal variables, and Generalized Linear Model (GLM) for nominal variables.

Specialized software can also assist in data preprocessing. In the paper "Mining Hospital Databases for Management Support", Freitas, Alberto, et al., in their efforts to analyze a hospital inpatient database, they prepared their data by using SPSS; IBM's predictive analytics software, before uploading it into R.

Choosing the most important attributes of this data is also of importance. This section of machine learning is called feature selection. The method depends on the researcher and can be as much of an art, as a science. The features chosen by a researcher can determine the quality of the clusters; careful consideration is required. Datasets generated by medical health records pose unique challenges; variables may be of many types, and can be sparse. They also tend to suffer from very high dimensionality, having to model real world objects and humans. Algorithmic tools can assist in making this task more manageable. The paper, 'Feature Selection for unsupervised learning through local learning' tries to rectify this by posing a way to best choose the optimum level of features from a high-dimensional dataset. It does this by revealing the intrinsic structures of a high dimensional space, and using gap statistics and for parameter estimation and asses the statistical significance of the structure through permutation tests.

In many medical databases, disease categories are stored in codes. The most popular of these is the ICD-9 code, developed by the World Health Organization (WHO). The code is associated with an encounter in a database, and is often used as a condition to filter for certain diseases. In the paper 'Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis', ICD-9 codes were aggregated into 802 categories. The researchers even counted the number of a certain code per patient, dropping an individual if the count was too high in a 6-month period, and only included categories that had at least 5% prevalence in the sample. In another paper titled, 'Identification of Type 2 diabetes subgroups through topological analysis of patient similarity' Li, Cheng, et.al used the technique of identifying individual records using these ICD-9 codes, and then aggregated the substantial number of codes to 281 single-level disease categories or 18 level 1 categories in multilevel disease categories. Once the data is processed, the data was then imported.

3.3 Software package Used

In the literature, many different software packages were used. The list includes

- 'Mining Hospital Databases for Management Support' – R
- 'Hidden Patterns: Clustering Diabetes Data' – SAS EM tool
- 'Clustering and Classifying Diabetic Data Sets using K-means Algorithm – WEKA tool

- ‘Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis’ – Matlab

While each paper uses software for their approach, each package contains similar algorithms that can find comparable results.

3.4 Clustering Algorithms Selected

In cluster analysis, many algorithms can cluster data, and researchers generally choose which is best for them through a variety of factors; the size of the dataset, the type of data used (numeric vs mixed), and runtime. Partitioning algorithms, like K-Means, tend to cluster faster than hierarchical algorithms, but their cluster centers are susceptible to false minima. In similar work, some researchers settle on a using single algorithm, while others use a more comparative approach. Different papers used different approaches depending on the data.

In ‘Characteristic evaluation of diabetes data using clustering techniques’, by Padmaja, Vikkurty, et al. used K-Means, Partitioning Around Medoids (PAM), Minimum Spanning Tree (MST), and Nearest Neighbor for generating clusters. At the end of the paper, they then evaluated each algorithm using an approach called Attribute Oriented Induction. When using this approach, they first identified the distinct counts of various features. The features with the maximum number of distinct values were then removed. The remaining maximum and minimum items were then grouped together using the set grouping. Set grouping can be found in the paper ‘Using Destination Set Grouping to Improve the Performance of Window-Controlled Multipoint Connections.’

Other approaches used singular algorithms. While the reasons for such were not stated, it can be speculated that the datasets that were chosen must have had some influence in their decision.

An example of this is in ‘Cluster Analysis of Clinical Data Identifies Fibromyalgia Subgroups’ where the dataset used is mixed. Docampo, Collado, et al. had a dataset that contained mostly continuous and dichotomous values. Dichotomous values are binary; taking the form of either 1, or 0. Due to the dataset being 75% dichotomous, they decided to convert their continuous variables to dichotomous ones. As a result, if they were to use a partitioning algorithm, it could not be K-Means, since the algorithm only functions on purely numeric data. Instead they used Partitioning around Medoids and Gower’s similarity measure. Gower’s general similarity measure is a technique that can calculate the distances between continuous variables, returning a distance matrix and contains the distances between each point of data in the dataset. As a parameter to PAM algorithm in R, the distance matrix can then be used to cluster the data.

In another example, ‘Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm’, the Kothianayaki and Thangaraj used the K-means algorithm and remarked that it was used both because of its popularity, and because it worked given a set of numeric objects.

Different algorithms were best given the discretion of the researcher, and the type of data that was used. Other approaches in related works include:

- ‘Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time Series Analysis’ – Hierarchical Clustering
- ‘Mining Hospital Databases for Management Support’ – Hierarchical Clustering (using Hclust, diana (Divisive Analysis) in R)
- ‘Hidden Patterns: Clustering Diabetes Data’ – Hierarchical Clustering (Ward method)

3.5 Visualization of Results

When the data had been processed, it was time to represent the data for interpretation. This is visualization. In some papers, researchers may not do this, and instead provide summary statistics – tables that encapsulate cluster information - while others may choose to both visualize and summarize their findings. Due to the large feature set size of the data, methods are needed to compress the dataset down to 2 dimensions for viewing.

A widely-used method for visualization is using Principal Component Analysis (PCA). PCA compresses the dimensionality of data without changing its structure. Figure 3.1 shows an example of a 3-dimenisonal visualization of Fibromyalgia patients, while Figure 3.2 shows how clusters - with their differing sizes - can be recorded in tabular form.

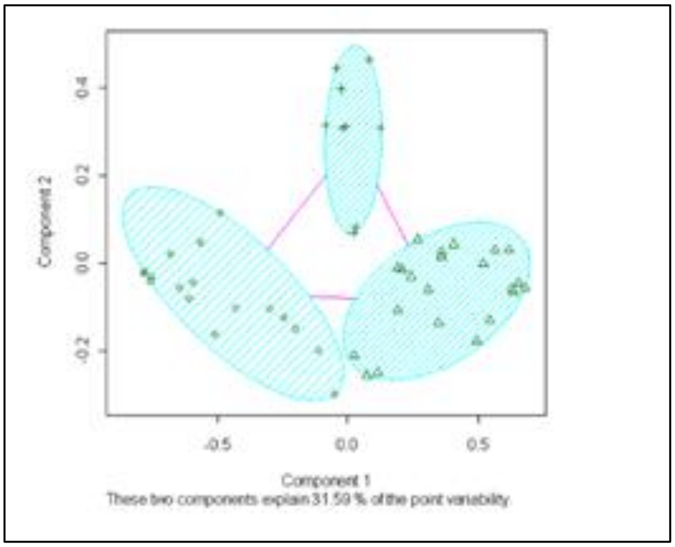


Figure 3.1 Clustering of variables into three dimensions (Docampo, Elisa, et al. "Cluster analysis of clinical data identifies fibromyalgia subgroups." ***PLoS One*** **8.9 (2013): e74873.)**

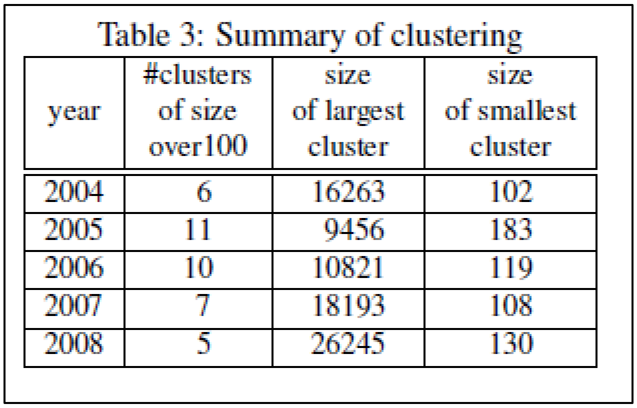


Table 3: Summary of clustering

year	#clusters of size over100	size of largest cluster	size of smallest cluster
2004	6	16263	102
2005	11	9456	183
2006	10	10821	119
2007	7	18193	108
2008	5	26245	130

Figure 3.2 Example of Cluster Summary (Cook, Rachel, and Gongzhu Hu. "Hidden Patterns: Clustering Diabetes Data." ***CAINE*****. 2010.)**

3.6 Discussion of Results

Discussion of results is where patterns and trends are extrapolated from the resulting clusters. Outlier clusters are often removed ("Hidden Patterns: Clustering Diabetes Data") when they are too small (in that case, smaller than 100).

Attributes of a cluster that is noticeable higher can indicate novel observations. In the paper 'Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time – Series Analysis, results were extrapolated from the number of codes, where one cluster contained over 5 times more codes than a larger subgroup, which lead to an observation. Sometimes proportions of features can lead to interesting findings. In the same paper, comparisons were made of the number of diagnoses of Asperger Syndrome in one subgroup versus the number of diagnoses of autism in another group.

Another study, in the paper 'Characteristic evaluation of diabetes data using clustering techniques' by Padmaja, Vikkurty, et al., researchers could predict the onset stage of diabetes by looking at the percentages of a factor (the number of women), and comparing them to already established data.

In conclusion, this section should provide a comprehensive overview of related works in the field of cluster analysis for medical data. It shows how other approaches chose to gather data, pre-process data, choose software, choose the appropriate algorithm, and to visualize and interpret their results. Papers in this field are relatively sparse, as the field is still new. This is written with the intention that the reader has an idea of goings on in the field, and the procedure of steps should he/she choose to conduct research in it.

CHAPTER 4. TECHNICAL APPROACH

4.1 Feature Selection

Generating clusters that best describe the data depend heavily on feature selection. This is particularly difficult in medical data due to the sheer size and variety of information that can either be in the form of categorical, nominal, or ordinal variables, images, handwriting, etc. For this research, the medical data was accessed through the I2B2 schema [12]. The database provided a small enough range of features that algorithmic approaches were not needed. Features instead were hand-picked.

The features that were agreed on primarily included demographic information, the number of inpatient encounters, the average time of an inpatient stay, and the mode diagnosis of a patient.

4.2 Data Collection

Having an idea of what features were needed for the dataset, proper procedure was used to create and extract the dataset. The data was stored in the I2B2 format in an Oracle database and was accessed using Oracle SQL Developer.

4.2.1 I2B2 Schema

The data set for extraction was stored in the i2b2 system. I2b2 stands for Informatics for Integrating Biology and the bedside, and is a star schema (see Figure 4.1) for representing, storing, and retrieving medical Information. I2b2 came about as a means for standardizing the data repository for electronic health data. The i2b2 schema as 5 main tables:

- Observation Fact
- Patient Dimension

- Concept Dimension
- Visit Dimension
- Observer Dimension
- Modifier Dimension

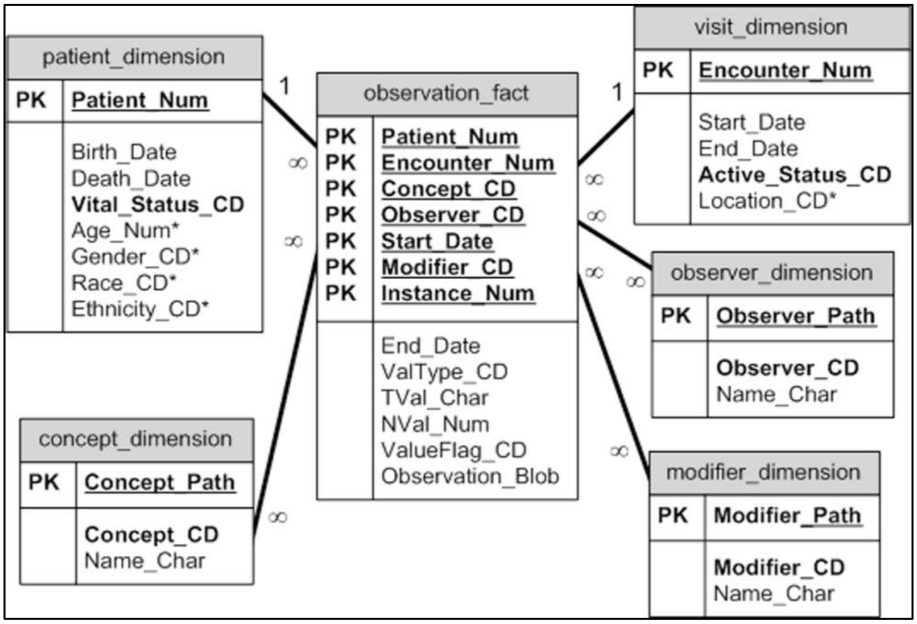
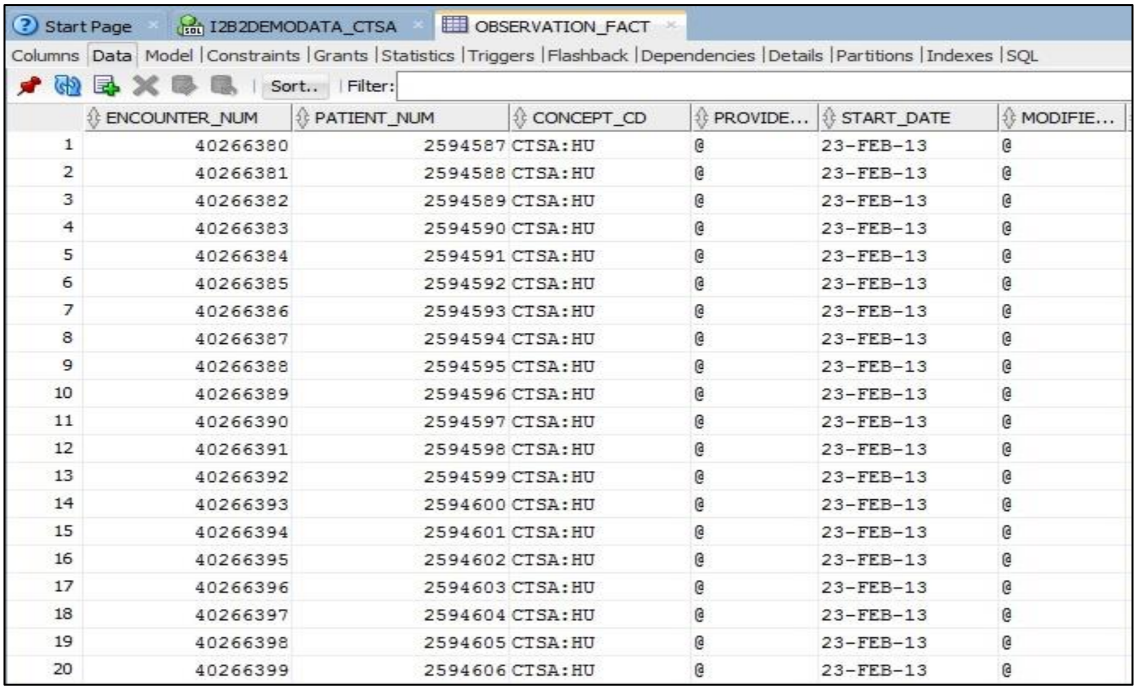


Figure 4.1 I2B2 Star Schema

4.2.1.1 Observation Fact

Observation fact is a table where each row is a record of an inpatient encounter. An inpatient encounter is any occurrence where a patient visits a hospital and is required to stay. In this case, a patient has an encounter number, which is the identification of the encounter and the primary key, the patient number, the patient's identification, a Concept CD, and a code that classifies lab tests or diagnoses (Figure 4.2). It is based on the ICD-9 codes [24]. A single patient can have multiple observations here.

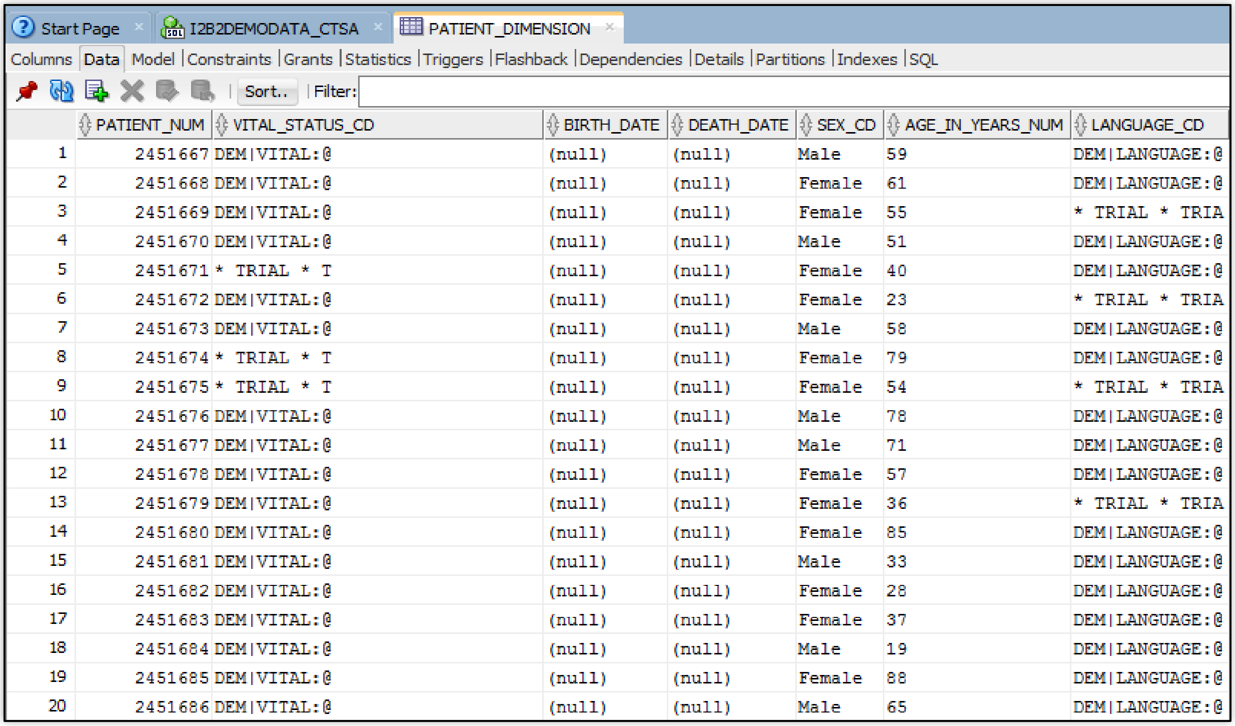

Start Page
I2B2DEMODATA_CTSA
OBSERVATION_FACT
Columns | Data | Model | Constraints | Grants | Statistics | Triggers | Flashback | Dependencies | Details | Partitions | Indexes | SQL
Sort..
Filter:

	ENCOUNTER_NUM	PATIENT_NUM	CONCEPT_CD	PROVIDE...	START_DATE	MODIFIE...
1	40266380	2594587	CTSA:HU	@	23-FEB-13	@
2	40266381	2594588	CTSA:HU	@	23-FEB-13	@
3	40266382	2594589	CTSA:HU	@	23-FEB-13	@
4	40266383	2594590	CTSA:HU	@	23-FEB-13	@
5	40266384	2594591	CTSA:HU	@	23-FEB-13	@
6	40266385	2594592	CTSA:HU	@	23-FEB-13	@
7	40266386	2594593	CTSA:HU	@	23-FEB-13	@
8	40266387	2594594	CTSA:HU	@	23-FEB-13	@
9	40266388	2594595	CTSA:HU	@	23-FEB-13	@
10	40266389	2594596	CTSA:HU	@	23-FEB-13	@
11	40266390	2594597	CTSA:HU	@	23-FEB-13	@
12	40266391	2594598	CTSA:HU	@	23-FEB-13	@
13	40266392	2594599	CTSA:HU	@	23-FEB-13	@
14	40266393	2594600	CTSA:HU	@	23-FEB-13	@
15	40266394	2594601	CTSA:HU	@	23-FEB-13	@
16	40266395	2594602	CTSA:HU	@	23-FEB-13	@
17	40266396	2594603	CTSA:HU	@	23-FEB-13	@
18	40266397	2594604	CTSA:HU	@	23-FEB-13	@
19	40266398	2594605	CTSA:HU	@	23-FEB-13	@
20	40266399	2594606	CTSA:HU	@	23-FEB-13	@

Figure 4.2 Screenshot of Observation Fact

4.2.1.2 Patient Dimension

In the Patient Dimension table, each row is a record of a patient (Figure 4.3). The row has an ID number, and categories for age, gender, religion, and ethnicity. The information on this table is anonymized (there are no identifiers that can in a row that can identify an individual).


Start Page
I2B2DEMODATA_CTSA
PATIENT_DIMENSION
Columns | Data | Model | Constraints | Grants | Statistics | Triggers | Flashback | Dependencies | Details | Partitions | Indexes | SQL
Sort..
Filter:

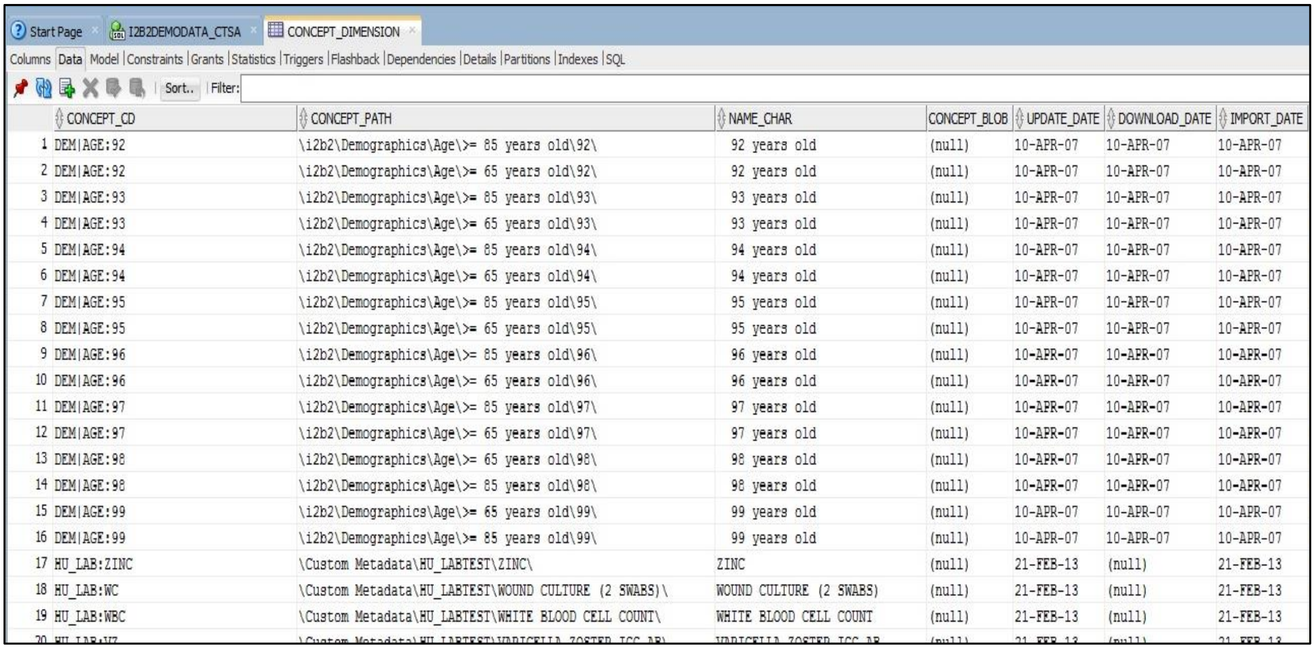
	PATIENT_NUM	VITAL_STATUS_CD	BIRTH_DATE	DEATH_DATE	SEX_CD	AGE_IN_YEARS_NUM	LANGUAGE_CD
1	2451667	DEM|VITAL:@	(null)	(null)	Male	59	DEM|LANGUAGE:@
2	2451668	DEM|VITAL:@	(null)	(null)	Female	61	DEM|LANGUAGE:@
3	2451669	DEM|VITAL:@	(null)	(null)	Female	55	* TRIAL * TRIA
4	2451670	DEM|VITAL:@	(null)	(null)	Male	51	DEM|LANGUAGE:@
5	2451671	* TRIAL * T	(null)	(null)	Female	40	DEM|LANGUAGE:@
6	2451672	DEM|VITAL:@	(null)	(null)	Female	23	* TRIAL * TRIA
7	2451673	DEM|VITAL:@	(null)	(null)	Male	58	DEM|LANGUAGE:@
8	2451674	* TRIAL * T	(null)	(null)	Female	79	DEM|LANGUAGE:@
9	2451675	* TRIAL * T	(null)	(null)	Female	54	* TRIAL * TRIA
10	2451676	DEM|VITAL:@	(null)	(null)	Male	78	DEM|LANGUAGE:@
11	2451677	DEM|VITAL:@	(null)	(null)	Male	71	DEM|LANGUAGE:@
12	2451678	DEM|VITAL:@	(null)	(null)	Female	57	DEM|LANGUAGE:@
13	2451679	DEM|VITAL:@	(null)	(null)	Female	36	* TRIAL * TRIA
14	2451680	DEM|VITAL:@	(null)	(null)	Female	85	DEM|LANGUAGE:@
15	2451681	DEM|VITAL:@	(null)	(null)	Male	33	DEM|LANGUAGE:@
16	2451682	DEM|VITAL:@	(null)	(null)	Female	28	DEM|LANGUAGE:@
17	2451683	DEM|VITAL:@	(null)	(null)	Female	37	DEM|LANGUAGE:@
18	2451684	DEM|VITAL:@	(null)	(null)	Male	19	DEM|LANGUAGE:@
19	2451685	DEM|VITAL:@	(null)	(null)	Female	88	DEM|LANGUAGE:@
20	2451686	DEM|VITAL:@	(null)	(null)	Male	65	DEM|LANGUAGE:@

Figure 4.3 Screenshot of Patient Dimension

4.2.1.3 Concept Dimension

The concept dimension table contains all the disease codes for every possible mappable illness. It contains fields for concept_cd, which is the code that is associated with an illness, name_char, which is the string attached to that code, and other characteristics (Figure 4.4).

These are the main three tables from which the data was extracted. The I2b2 schema was stored in oracle and accessible by Oracle SQL developer. Extracting the dataset that was needed required many SQL calls.


Start Page
I2B2DEMODATA_CTSA
CONCEPT_DIMENSION
Columns | Data | Model | Constraints | Grants | Statistics | Triggers | Flashback | Dependencies | Details | Partitions | Indexes | SQL
Sort..
Filter:

	CONCEPT_CD	CONCEPT_PATH	NAME_CHAR	CONCEPT_BLOB	UPDATE_DATE	DOWNLOAD_DATE	IMPORT_DATE
1	DEM|AGE:92	\i2b2\Demographics\Age\>= 85 years old\92\	92 years old	(null)	10-APR-07	10-APR-07	10-APR-07
2	DEM|AGE:92	\i2b2\Demographics\Age\>= 65 years old\92\	92 years old	(null)	10-APR-07	10-APR-07	10-APR-07
3	DEM|AGE:93	\i2b2\Demographics\Age\>= 85 years old\93\	93 years old	(null)	10-APR-07	10-APR-07	10-APR-07
4	DEM|AGE:93	\i2b2\Demographics\Age\>= 65 years old\93\	93 years old	(null)	10-APR-07	10-APR-07	10-APR-07
5	DEM|AGE:94	\i2b2\Demographics\Age\>= 85 years old\94\	94 years old	(null)	10-APR-07	10-APR-07	10-APR-07
6	DEM|AGE:94	\i2b2\Demographics\Age\>= 65 years old\94\	94 years old	(null)	10-APR-07	10-APR-07	10-APR-07
7	DEM|AGE:95	\i2b2\Demographics\Age\>= 85 years old\95\	95 years old	(null)	10-APR-07	10-APR-07	10-APR-07
8	DEM|AGE:95	\i2b2\Demographics\Age\>= 65 years old\95\	95 years old	(null)	10-APR-07	10-APR-07	10-APR-07
9	DEM|AGE:96	\i2b2\Demographics\Age\>= 85 years old\96\	96 years old	(null)	10-APR-07	10-APR-07	10-APR-07
10	DEM|AGE:96	\i2b2\Demographics\Age\>= 65 years old\96\	96 years old	(null)	10-APR-07	10-APR-07	10-APR-07
11	DEM|AGE:97	\i2b2\Demographics\Age\>= 85 years old\97\	97 years old	(null)	10-APR-07	10-APR-07	10-APR-07
12	DEM|AGE:97	\i2b2\Demographics\Age\>= 65 years old\97\	97 years old	(null)	10-APR-07	10-APR-07	10-APR-07
13	DEM|AGE:98	\i2b2\Demographics\Age\>= 65 years old\98\	98 years old	(null)	10-APR-07	10-APR-07	10-APR-07
14	DEM|AGE:98	\i2b2\Demographics\Age\>= 85 years old\98\	98 years old	(null)	10-APR-07	10-APR-07	10-APR-07
15	DEM|AGE:99	\i2b2\Demographics\Age\>= 65 years old\99\	99 years old	(null)	10-APR-07	10-APR-07	10-APR-07
16	DEM|AGE:99	\i2b2\Demographics\Age\>= 85 years old\99\	99 years old	(null)	10-APR-07	10-APR-07	10-APR-07
17	HU_LAB:ZINC	\Custom Metadata\HU_LABTEST\ZINC\	ZINC	(null)	21-FEB-13	(null)	21-FEB-13
18	HU_LAB:WC	\Custom Metadata\HU_LABTEST\WOUND CULTURE (2 SWABS)\	WOUND CULTURE (2 SWABS)	(null)	21-FEB-13	(null)	21-FEB-13
19	HU_LAB:WBC	\Custom Metadata\HU_LABTEST\WHITE BLOOD CELL COUNT\	WHITE BLOOD CELL COUNT	(null)	21-FEB-13	(null)	21-FEB-13
20	HU_LAB:VZ	\Custom Metadata\HU_LABTEST\VARICELLA ZOSTER IGG AB\	VARICELLA ZOSTER IGG AB	(null)	21-FEB-13	(null)	21-FEB-13

Figure 4.4 Screenshot of Concept Dimension

4.2.2. Extracting the Dataset

To create the dataset, a table was first created, where all the columns represented the features of the dataset. The final dataset needed to contain:

- Patient ID number
- Gender
- Race
- Age
- Religion
- Marital Status
- Number of Inpatient Encounters
- Average Length of Stay
- Mode Concept CD

The mode concept_cd represented the ICD-9 code that appeared the most times in that patient's encounter records, that was not diabetes related. This was done to determine the chief co-morbidity for that patient. Since the data was stored in Oracle SQL server, the database needed to query and create a new table to export the final feature set.

4.2.2.1. Creating a preliminary feature set (Without Average Length of stay)

Creating a dataset with average length of stay fields, and mode concept cd required more elaborate SQL calls, so first a preliminary dataset was built. This preliminary Dataset contains all the previously mentioned fields, except for average length of stay, and mode concept cd.

The SQL code to produce that table in ORACLE SQL:

```
create global temporary table feature_set
on commit preserve rows
as select patient_dimension.patient_num Patient_number, patient_dimension.SEX_CD gender,
patient_dimension.AGE_IN_YEARS_NUM Age, patient_dimension.race_cd race,
patient_dimension.religion_cd religion,
  patient_dimension.marital_status_cd marital_status, count(observation_fact.encounter_num)
Number_of_Inpatient_Encounters
  from patient_dimension
  join observation_fact
  on patient_dimension.patient_num = Observation_Fact.Patient_Num
  join concept_dimension
  on observation_fact.concept_cd = concept_dimension.concept_cd
  where (name_char like '%diabet%type%II%') or (name_char like '%type%II%diabet%') or
       (concept_path like '%diabet%type%II%') or (concept_path like '%type%II%diabet%')
  group by patient_dimension.patient_num, patient_dimension.SEX_CD,
patient_dimension.AGE_IN_YEARS_NUM, patient_dimension.race_cd,
patient_dimension.religion_cd,
```

The code pulled the patient number, the gender, age, race, religion, marital status, and places it into a temporary table called feature_set.

4.2.2.2. Creating the Average Length of Inpatient Stay

To obtain the average length of stay for each patient, the end date was subtracted from the start date for all a patient’s encounters in the encounter_dimension table, where they were all averaged and associated with a Patient Number. This result of this was then placed in table called average_length_of_stay.

The ORACLE SQL code appears in the following text:

```
 create global temporary table average_length_of_stay
on commit preserve rows


as select patient_num, Round(avg(end_date - start_date),4) Average_length_of_stay
from observation_fact
join concept_dimension
on observation_fact.concept_cd = concept_dimension.concept_cd
where (name_char like '%diabet%type%II%') or (name_char like '%type%II%diabet%') or
     (concept_path like '%diabet%type%II%') or (concept_path like '%type%II%diabet%')
group by patient_num;
```

4.2.2.3. Joining the feature_set and average_length_of_stay

The next set of SQL queries then merged the feature_set and average_length_of_stay into one table:

```
create global temporary table final_feature_set
on commit preserve rows


as select patient_number, gender, race, age, religion, marital_status, Feature_Set.Zip_Code,
  Feature_Set.Number_Of_Inpatient_Encounters,
Average_Length_Of_Stay.Average_Length_Of_Stay from feature_set
join average_length_of_stay
on Feature_Set.Patient_Number = Average_Length_Of_Stay.Patient_Num;
```

4.2.2.4. Creating mode concept_cd

The last step in creating the complete dataset was to append the mode concept_cd to the dataset. This showed the most frequent co-disease that each patient has.

The first thing that was done was to create a table called concept_count. In this temporary table, the database was queried so that for every patient, there would be a Patient Number, Concept_cd, Name_car, Max count of concept cds.

```
/* Creates a temporary able listing the Patient Number, Concept Cd, Name Char, Count, Max
count of concept cds */

create global temporary table concept_count

on commit preserve rows

as select observation_fact.patient_num, observation_fact.concept_cd,
concept_dimension.name_char, Count(observation_fact.concept_cd) cnt,
max(count(observation_fact.concept_cd)) over (partition by observation_fact.patient_num)
max_count

    from observation_fact

    join concept_dimension

    on observation_fact.concept_cd = concept_dimension.concept_cd

    where observation_fact.concept_cd not in (select observation_fact.concept_cd from
observation_fact where (concept_cd like 'DEM%') or (concept_cd like 'CTSA:%') or
(concept_cd like 'HU_LAB:%') or

    (name_char like '%diabet%type%II%') or (name_char like '%type%II%diabet%') or
(concept_path like '%diabet%type%II%') or (concept_path like '%type%II%diabet%') )

    group by observation_fact.patient_num, observation_fact.concept_cd,
Concept_Dimension.Name_Char;
```

This SQL code created a temporary table in oracle called concept_count. The table contains the mode ICD-9 codes for every patient in encounter fact. It is not complete however; there was still the problem of repeat rows per patient. What was needed was only one mode. This was solved with this code:

```
create global temporary table final_concept_count

on commit preserve rows

as  select t.patient_num, t.concept_cd, t.name_char

from (select concept_count.patient_num, concept_count.concept_cd,
concept_count.name_char, ROW_NUMBER() OVER (PARTITION BY patient_num
ORDER BY patient_num ) as rnum

from concept_count) t

where t.rnum = 1;
```

This only returned one instance of a patient, the concept_cd / ICD-9 code of only one disease, and its associated name_char value.

4.2.2.5 Joining last_feature_set and final_concept_cd

Joining these two tables generated the final dataset for export. This is done with the following SQL code:

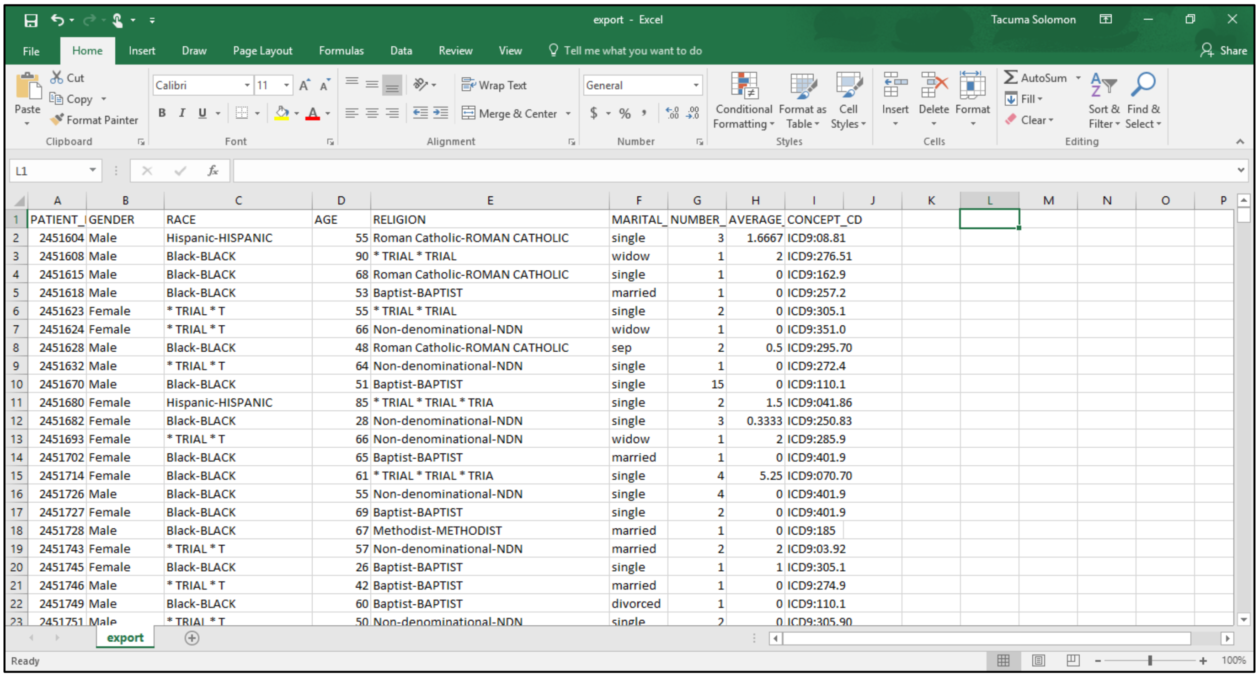
```
create global temporary table last_feature_set

on commit preserve rows

as select patient_number, gender, race, age, religion, marital_status,
Number_Of_Inpatient_Encounters, Average_Length_Of_Stay,
final_concept_count.concept_cd

from final_feature_set

join final_concept_count

on Final_Feature_Set.Patient_Number = Final_Concept_Count.Patient_Num;
```

4.3 Software Used

The software chosen for cluster analysis was R. From the related work section, and looking at requirements for cost, ease of obtainability, ease of use, results, popularity and support, it was determined that R would be the best choice. As a result, all data processing tasks were to make the data suitable for processing by R.

4.4 Preprocessing

Before the data can be uploaded into R, it first was made ready for use [13]. The dataset was exported to a csv file, and the changes made in Microsoft Excel. The data freshly exported from Oracle SQL developer is displayed in Figure 4.5:



	A	B	C	D	E	F	G	H	I
1	PATIENT_I	GENDER	RACE	AGE	RELIGION	MARITAL_	NUMBER_	AVERAGE_	CONCEPT_CD
2	2451604	Male	Hispanic-HISPANIC	55	Roman Catholic-ROMAN CATHOLIC	single	3	1.6667	ICD9:08.81
3	2451608	Male	Black-BLACK	90	* TRIAL * TRIAL	widow	1	2	ICD9:276.51
4	2451615	Male	Black-BLACK	68	Roman Catholic-ROMAN CATHOLIC	single	1	0	ICD9:162.9
5	2451618	Male	Black-BLACK	53	Baptist-BAPTIST	married	1	0	ICD9:257.2
6	2451623	Female	* TRIAL * T	55	* TRIAL * TRIAL	single	2	0	ICD9:305.1
7	2451624	Female	* TRIAL * T	66	Non-denominational-NDN	widow	1	0	ICD9:351.0
8	2451628	Male	Black-BLACK	48	Roman Catholic-ROMAN CATHOLIC	sep	2	0.5	ICD9:295.70
9	2451632	Male	* TRIAL * T	64	Non-denominational-NDN	single	1	0	ICD9:272.4
10	2451670	Male	Black-BLACK	51	Baptist-BAPTIST	single	15	0	ICD9:110.1
11	2451680	Female	Hispanic-HISPANIC	85	* TRIAL * TRIAL * TRIA	single	2	1.5	ICD9:041.86
12	2451682	Female	Black-BLACK	28	Non-denominational-NDN	single	3	0.3333	ICD9:250.83
13	2451693	Female	* TRIAL * T	66	Non-denominational-NDN	widow	1	2	ICD9:285.9
14	2451702	Female	Black-BLACK	65	Baptist-BAPTIST	married	1	0	ICD9:401.9
15	2451714	Female	Black-BLACK	61	* TRIAL * TRIAL * TRIA	single	4	5.25	ICD9:070.70
16	2451726	Male	Black-BLACK	55	Non-denominational-NDN	single	4	0	ICD9:401.9
17	2451727	Female	Black-BLACK	69	Baptist-BAPTIST	single	2	0	ICD9:401.9
18	2451728	Male	Black-BLACK	67	Methodist-METHODIST	married	1	0	ICD9:185
19	2451743	Female	* TRIAL * T	57	Non-denominational-NDN	married	2	2	ICD9:03.92
20	2451745	Female	Black-BLACK	26	Baptist-BAPTIST	single	1	1	ICD9:305.1
21	2451746	Male	* TRIAL * T	42	Baptist-BAPTIST	married	1	0	ICD9:274.9
22	2451749	Male	Black-BLACK	60	Baptist-BAPTIST	divorced	1	0	ICD9:110.1
23	2451751	Male	* TRIAL * T	50	Non-denominational-NDN	single	2	0	ICD9:305.90

Figure 4.5 Screenshot of Preprocessed csv file

4.4.1 Error Rates for Features

R requires that the data be clean when inserted, with null values appearing strictly as NA. Before this however, the task remained in finding the error counts of each dimension. In our method, if number of missing values was less than 40% it would be admissible for use in R. Each error rate was found by dividing the erroneous values in a dimension by the total number of elements in that dimension.

The error rates for 3 elements in the dataset were as follows:

- Error rates for race = 33.29%
- Error rates for religion = 31.76%
- Error rates for marital status = 2.56%

These errors were well below the upper threshold of availability, so they remained in the dataset.

4.4.2. Formatting Titles and Values

The dataset was formatted to remain in line with R standards and naming conventions. Rules must be followed with rows and column names:

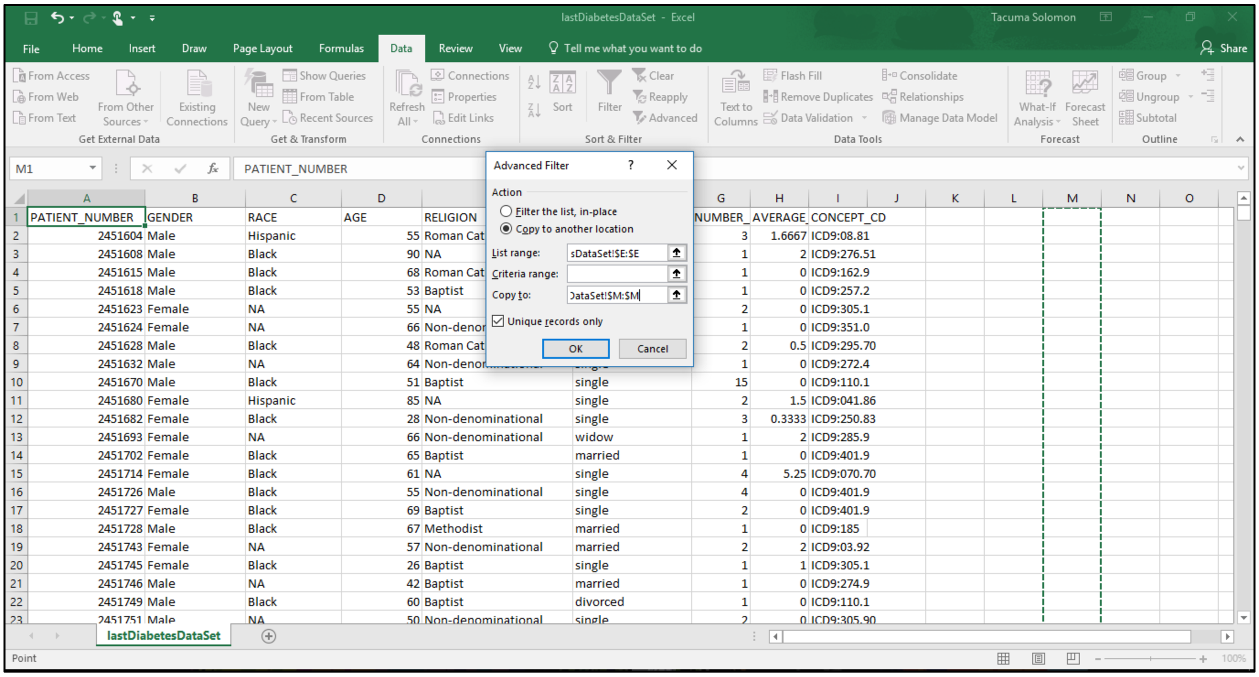
- The first row must be used for headers. They generally represent variables.
- The first column should be used as row names; they represent observations
- Each row name should be unique. Remove duplications.
- Names with blank spaces should be avoided. 'First_name' or 'first.name' is acceptable, 'first name', is not.
- Avoid names with special symbols. Only underscore can be used.
- Variable names must not begin with a number. Letters should be used instead.

- Column names must be unique.
- R is case sensitive.
- Blank rows in data should be avoided
- Blank values should be replaced by NA (for not available)

To meet those requirements, erroneous values in the dataset were first be removed [13]. Values with '*Trial*' – a value from the I2b2 database – were replaced with NA. Variable names were standardized for uniformity and readability. For example, values such as 'Black – BLACK', were changed to 'Black' in the race column, and values such as "Non-denominational – NDN" are changed to 'Non-Denominational'. Blank spaces were changed to NA, falling in like with R's policy of empty data cells.

4.4.3. Checking for Errors

To validate the changes, each column was checked to ensure that values were within the stipulated guidelines. This was done by looking at the distinct values of every feature, using the advanced filter in the data tab of excel. The method for this is shown in Figure 4.6 and Figure 4.7:



M1 | PATIENT_NUMBER

Advanced Filter

Action

○ Filter the list, in-place

● Copy to another location

List range: sDataSet!$E:$E

Criteria range:

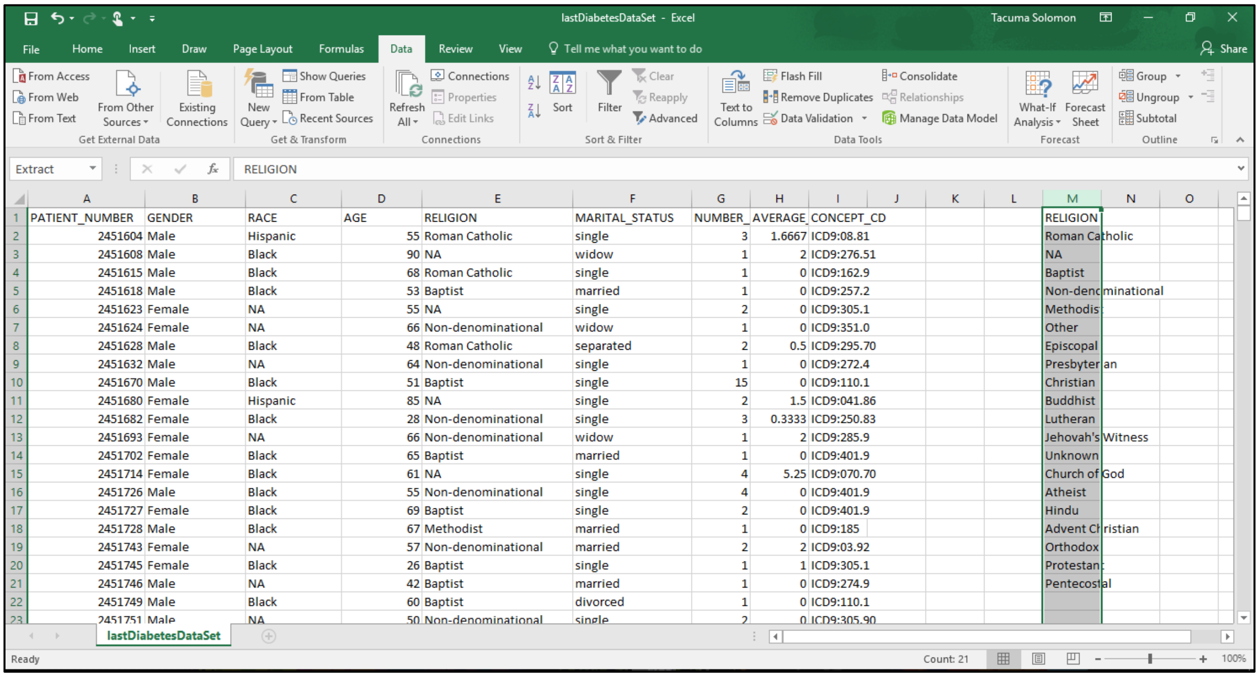
Copy to: DataSet!$M:$M

☑ Unique records only

OK | Cancel

	A	B	C	D	E	F	G	H	I
1	PATIENT_NUMBER	GENDER	RACE	AGE	RELIGION		NUMBER_	AVERAGE_	CONCEPT_CD
2	2451604	Male	Hispanic	55	Roman Cat		3	1.6667	ICD9:08.81
3	2451608	Male	Black	90	NA		1	2	ICD9:276.51
4	2451615	Male	Black	68	Roman Cat		1	0	ICD9:162.9
5	2451618	Male	Black	53	Baptist		1	0	ICD9:257.2
6	2451623	Female	NA	55	NA		2	0	ICD9:305.1
7	2451624	Female	NA	66	Non-denor		1	0	ICD9:351.0
8	2451628	Male	Black	48	Roman Cat		2	0.5	ICD9:295.70
9	2451632	Male	NA	64	Non-denominational		1	0	ICD9:272.4
10	2451670	Male	Black	51	Baptist	single	15	0	ICD9:110.1
11	2451680	Female	Hispanic	85	NA	single	2	1.5	ICD9:041.86
12	2451682	Female	Black	28	Non-denominational	single	3	0.3333	ICD9:250.83
13	2451693	Female	NA	66	Non-denominational	widow	1	2	ICD9:285.9
14	2451702	Female	Black	65	Baptist	married	1	0	ICD9:401.9
15	2451714	Female	Black	61	NA	single	4	5.25	ICD9:070.70
16	2451726	Male	Black	55	Non-denominational	single	4	0	ICD9:401.9
17	2451727	Female	Black	69	Baptist	single	2	0	ICD9:401.9
18	2451728	Male	Black	67	Methodist	married	1	0	ICD9:185
19	2451743	Female	NA	57	Non-denominational	married	2	2	ICD9:03.92
20	2451745	Female	Black	26	Baptist	single	1	1	ICD9:305.1
21	2451746	Male	NA	42	Baptist	married	1	0	ICD9:274.9
22	2451749	Male	Black	60	Baptist	divorced	1	0	ICD9:110.1
23	2451751	Male	NA	50	Non-denominational	single	2	0	ICD9:305.90

Figure 4.6 Screenshot of unique validation of data



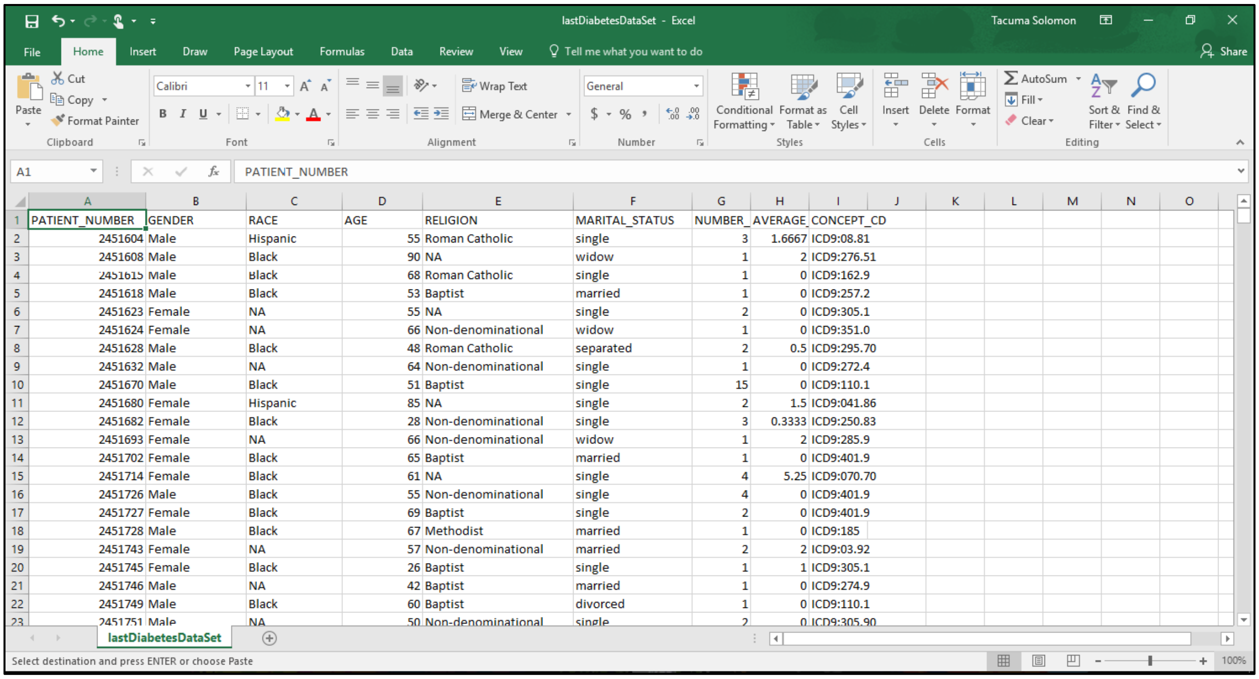
Extract | RELIGION

	A	B	C	D	E	F	G	H	I	M
1	PATIENT_NUMBER	GENDER	RACE	AGE	RELIGION	MARITAL_STATUS	NUMBER_	AVERAGE_	CONCEPT_CD	RELIGION
2	2451604	Male	Hispanic	55	Roman Catholic	single	3	1.6667	ICD9:08.81	Roman Catholic
3	2451608	Male	Black	90	NA	widow	1	2	ICD9:276.51	NA
4	2451615	Male	Black	68	Roman Catholic	single	1	0	ICD9:162.9	Baptist
5	2451618	Male	Black	53	Baptist	married	1	0	ICD9:257.2	Non-denominational
6	2451623	Female	NA	55	NA	single	2	0	ICD9:305.1	Methodist
7	2451624	Female	NA	66	Non-denominational	widow	1	0	ICD9:351.0	Other
8	2451628	Male	Black	48	Roman Catholic	separated	2	0.5	ICD9:295.70	Episcopal
9	2451632	Male	NA	64	Non-denominational	single	1	0	ICD9:272.4	Presbyterian
10	2451670	Male	Black	51	Baptist	single	15	0	ICD9:110.1	Christian
11	2451680	Female	Hispanic	85	NA	single	2	1.5	ICD9:041.86	Buddhist
12	2451682	Female	Black	28	Non-denominational	single	3	0.3333	ICD9:250.83	Lutheran
13	2451693	Female	NA	66	Non-denominational	widow	1	2	ICD9:285.9	Jehovah's Witness
14	2451702	Female	Black	65	Baptist	married	1	0	ICD9:401.9	Unknown
15	2451714	Female	Black	61	NA	single	4	5.25	ICD9:070.70	Church of God
16	2451726	Male	Black	55	Non-denominational	single	4	0	ICD9:401.9	Atheist
17	2451727	Female	Black	69	Baptist	single	2	0	ICD9:401.9	Hindu
18	2451728	Male	Black	67	Methodist	married	1	0	ICD9:185	Advent Christian
19	2451743	Female	NA	57	Non-denominational	married	2	2	ICD9:03.92	Orthodox
20	2451745	Female	Black	26	Baptist	single	1	1	ICD9:305.1	Protestant
21	2451746	Male	NA	42	Baptist	married	1	0	ICD9:274.9	Pentecostal
22	2451749	Male	Black	60	Baptist	divorced	1	0	ICD9:110.1	
23	2451751	Male	NA	50	Non-denominational	single	2	0	ICD9:305.90	

Figure 4.7 Screenshot of unique validation of data output

This is done for all features, to validate correctness.

At the end of the formatting process, the final processed dataset is show in Figure 4.8:



A1 PATIENT_NUMBER

	A	B	C	D	E	F	G	H	I
1	PATIENT_NUMBER	GENDER	RACE	AGE	RELIGION	MARITAL_STATUS	NUMBER_	AVERAGE_	CONCEPT_CD
2	2451604	Male	Hispanic	55	Roman Catholic	single	3	1.6667	ICD9:08.81
3	2451608	Male	Black	90	NA	widow	1	2	ICD9:276.51
4	2451615	Male	Black	68	Roman Catholic	single	1	0	ICD9:162.9
5	2451618	Male	Black	53	Baptist	married	1	0	ICD9:257.2
6	2451623	Female	NA	55	NA	single	2	0	ICD9:305.1
7	2451624	Female	NA	66	Non-denominational	widow	1	0	ICD9:351.0
8	2451628	Male	Black	48	Roman Catholic	separated	2	0.5	ICD9:295.70
9	2451632	Male	NA	64	Non-denominational	single	1	0	ICD9:272.4
10	2451670	Male	Black	51	Baptist	single	15	0	ICD9:110.1
11	2451680	Female	Hispanic	85	NA	single	2	1.5	ICD9:041.86
12	2451682	Female	Black	28	Non-denominational	single	3	0.3333	ICD9:250.83
13	2451693	Female	NA	66	Non-denominational	widow	1	2	ICD9:285.9
14	2451702	Female	Black	65	Baptist	married	1	0	ICD9:401.9
15	2451714	Female	Black	61	NA	single	4	5.25	ICD9:070.70
16	2451726	Male	Black	55	Non-denominational	single	4	0	ICD9:401.9
17	2451727	Female	Black	69	Baptist	single	2	0	ICD9:401.9
18	2451728	Male	Black	67	Methodist	married	1	0	ICD9:185
19	2451743	Female	NA	57	Non-denominational	married	2	2	ICD9:03.92
20	2451745	Female	Black	26	Baptist	single	1	1	ICD9:305.1
21	2451746	Male	NA	42	Baptist	married	1	0	ICD9:274.9
22	2451749	Male	Black	60	Baptist	divorced	1	0	ICD9:110.1
23	2451751	Male	NA	50	Non-denominational	single	2	0	ICD9:305.90

Select destination and press ENTER or choose Paste

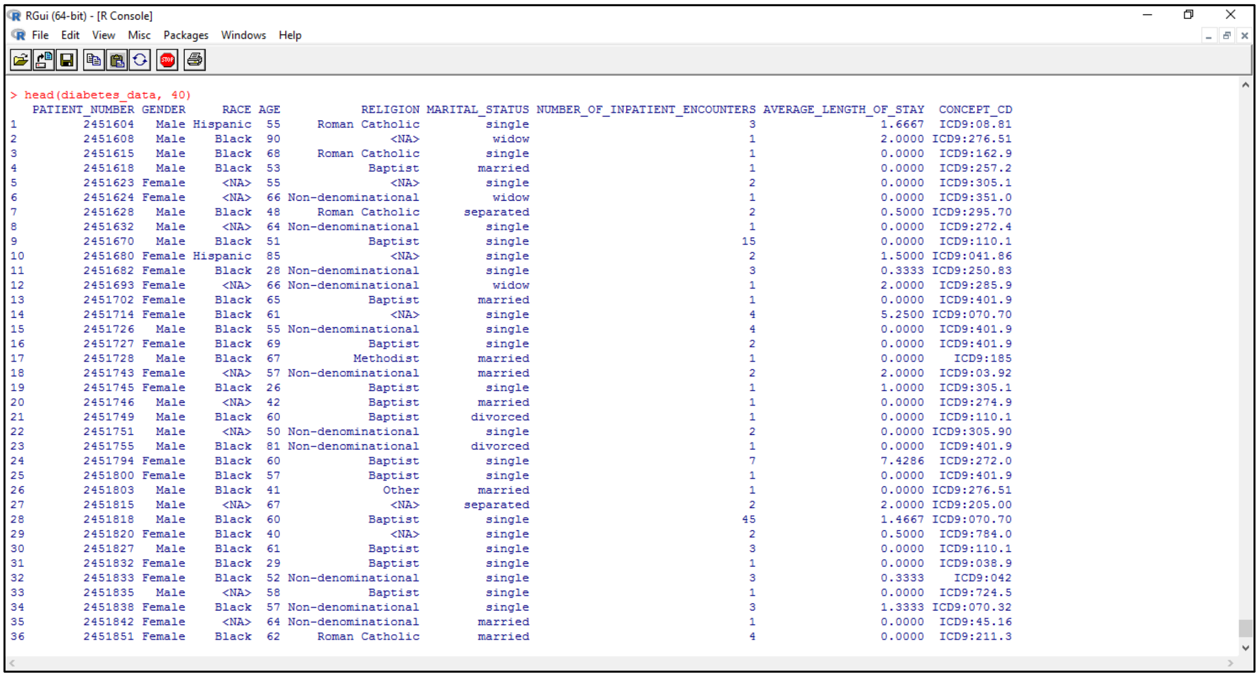
Figure 4.8 Screenshot of processed dataset

4.5 Importing the Data Into R

Importing the data into R involved using the read.csv () command, specifically [13]:

```
diabetes_data -> read_csv(file.choose())
```

The imported data takes the appearance of Figure 4.9 in R:


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

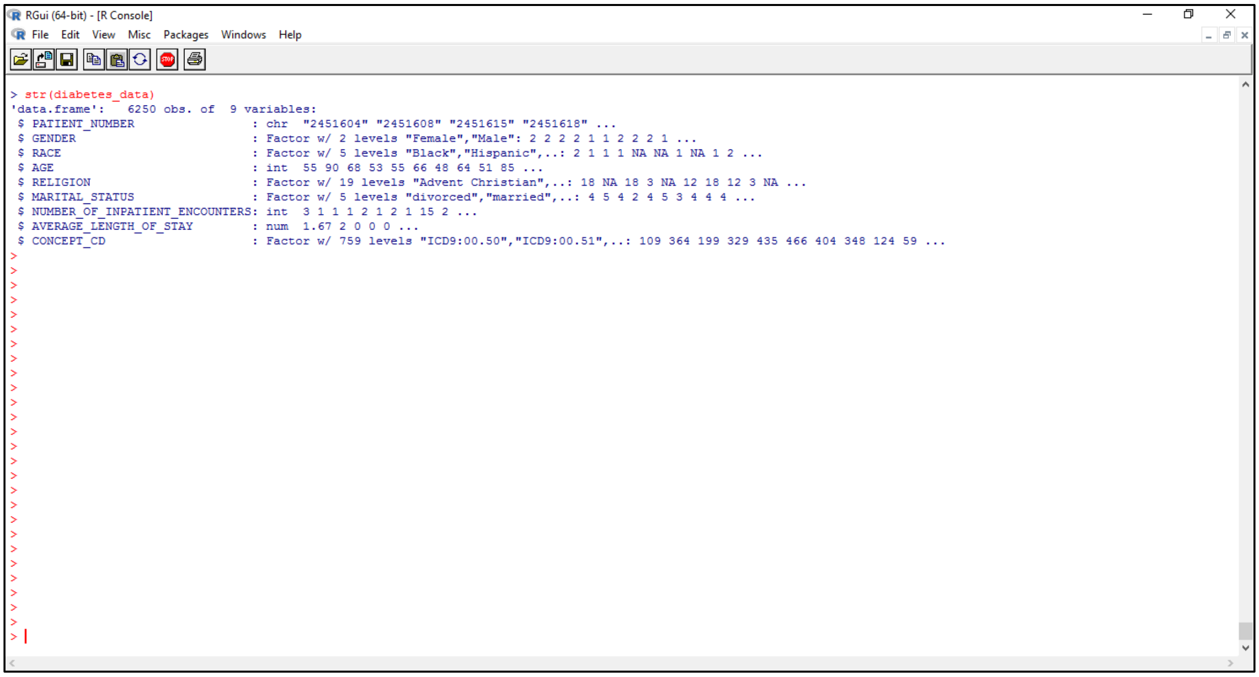
```
> head(diabetes_data, 40)
   PATIENT_NUMBER GENDER     RACE AGE           RELIGION MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY  CONCEPT_CD
1         2451604   Male Hispanic  55     Roman Catholic         single                              3                 1.6667  ICD9:08.81
2         2451608   Male    Black  90               <NA>          widow                              1                 2.0000 ICD9:276.51
3         2451615   Male    Black  68     Roman Catholic         single                              1                 0.0000  ICD9:162.9
4         2451618   Male    Black  53            Baptist        married                              1                 0.0000  ICD9:257.2
5         2451623 Female     <NA>  55               <NA>         single                              2                 0.0000  ICD9:305.1
6         2451624 Female     <NA>  66 Non-denominational          widow                              1                 0.0000  ICD9:351.0
7         2451628   Male    Black  48     Roman Catholic      separated                              2                 0.5000 ICD9:295.70
8         2451632   Male     <NA>  64 Non-denominational         single                              1                 0.0000  ICD9:272.4
9         2451670   Male    Black  51            Baptist         single                             15                 0.0000  ICD9:110.1
10        2451680 Female Hispanic  85               <NA>         single                              2                 1.5000 ICD9:041.86
11        2451682 Female    Black  28 Non-denominational         single                              3                 0.3333 ICD9:250.83
12        2451693 Female     <NA>  66 Non-denominational          widow                              1                 2.0000  ICD9:285.9
13        2451702 Female    Black  65            Baptist        married                              1                 0.0000  ICD9:401.9
14        2451714 Female    Black  61               <NA>         single                              4                 5.2500 ICD9:070.70
15        2451726   Male    Black  55 Non-denominational         single                              4                 0.0000  ICD9:401.9
16        2451727 Female    Black  69            Baptist         single                              2                 0.0000  ICD9:401.9
17        2451728   Male    Black  67          Methodist        married                              1                 0.0000    ICD9:185
18        2451743 Female     <NA>  57 Non-denominational        married                              2                 2.0000  ICD9:03.92
19        2451745 Female    Black  26            Baptist         single                              1                 1.0000  ICD9:305.1
20        2451746   Male     <NA>  42            Baptist        married                              1                 0.0000  ICD9:274.9
21        2451749   Male    Black  60            Baptist       divorced                              1                 0.0000  ICD9:110.1
22        2451751   Male     <NA>  50 Non-denominational         single                              2                 0.0000 ICD9:305.90
23        2451755   Male    Black  81 Non-denominational       divorced                              1                 0.0000  ICD9:401.9
24        2451794 Female    Black  60            Baptist         single                              7                 7.4286  ICD9:272.0
25        2451800 Female    Black  57            Baptist         single                              1                 0.0000  ICD9:401.9
26        2451803   Male    Black  41              Other        married                              1                 0.0000 ICD9:276.51
27        2451815   Male     <NA>  67               <NA>      separated                              2                 2.0000 ICD9:205.00
28        2451818   Male    Black  60            Baptist         single                             45                 1.4667 ICD9:070.70
29        2451820 Female    Black  40               <NA>         single                              2                 0.5000  ICD9:784.0
30        2451827   Male    Black  61            Baptist         single                              3                 0.0000  ICD9:110.1
31        2451832 Female    Black  29            Baptist         single                              1                 0.0000  ICD9:038.9
32        2451833 Female    Black  52 Non-denominational         single                              3                 0.3333    ICD9:042
33        2451835   Male     <NA>  58            Baptist         single                              1                 0.0000  ICD9:724.5
34        2451838 Female    Black  57 Non-denominational         single                              3                 1.3333 ICD9:070.32
35        2451842 Female     <NA>  64 Non-denominational        married                              1                 0.0000  ICD9:45.16
36        2451851 Female    Black  62     Roman Catholic        married                              4                 0.0000  ICD9:211.3
```

Figure 4.9 Screenshot of data imported into R

More processing was done on this data however. str(diabetes_data) displays the structure of the dataset, displaying the values of each type. R shows the 'PATIENT_NUMBER' is set as an int type. The following command changes the type:

```
diabetes_data$PATIENT_NUMBER<- as.character(diabetes_patient$PATIENT_NUMBER)
```

When the str(diabetes_data) command is entered, the structure of the dataset is shown in Figure 4.10:



```
> str(diabetes_data)
'data.frame':   6250 obs. of  9 variables:
 $ PATIENT_NUMBER                 : chr  "2451604" "2451608" "2451615" "2451618" ...
 $ GENDER                         : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 2 2 2 1 ...
 $ RACE                           : Factor w/ 5 levels "Black","Hispanic",..: 2 1 1 1 NA NA 1 NA 1 2 ...
 $ AGE                            : int  55 90 68 53 55 66 48 64 51 85 ...
 $ RELIGION                       : Factor w/ 19 levels "Advent Christian",..: 18 NA 18 3 NA 12 18 12 3 NA ...
 $ MARITAL_STATUS                 : Factor w/ 5 levels "divorced","married",..: 4 5 4 2 4 5 3 4 4 4 ...
 $ NUMBER_OF_INPATIENT_ENCOUNTERS: int  3 1 1 1 2 1 2 1 15 2 ...
 $ AVERAGE_LENGTH_OF_STAY         : num  1.67 2 0 0 0 ...
 $ CONCEPT_CD                     : Factor w/ 759 levels "ICD9:00.50","ICD9:00.51",..: 109 364 199 329 435 466 404 348 124 59 ...
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> |
```

Figure 4.10 Screenshot of structure of imported data

Representing the PATIENT_NUMBER in R ensures that the dimension is a factor or key, and not as a numeric value. The dataset was ready to be processed algorithmically.

4.6 Clustering Algorithms Used

To do the analysis, multiple algorithms were used. It was decided that to find the algorithm that can best cluster the dataset, partitioning and hierarchical algorithms will be implemented, with a comparative analysis on the results of each to determine the one must suitable to the task. From the research, it was decided that four algorithms will be used, two of a partitioning type, and two of a hierarchical type. Those algorithms were:

Partitioning:

- K-Modes algorithm [14] [16]
- PAM (Partitioning around Medoids) [16]

Hierarchical:

- Hclust – An agglomerative hierarchical clustering algorithm [17] [18]
- DIANA – A hierarchical algorithm that functions via Divisive Analysis [19]

To choose the algorithms, considerations were made based on the size and type of the dataset. K-means and K-means++, popular algorithmic approaches, were not used because the type of dataset gleaned from the database held both continuous and numeric data. K-means can only use numeric data. Before considering algorithms, much consideration was made on whether further processing on the data was needed. Should the data be converted to numeric? Were there algorithms that can effectively cluster mixed data? After trial and error, it was decided that mixed data would suffice. To further validate this reasoning, the dataset was modified in diverse ways and tested with PAM's package to determine what the visualizations of those clusters may look like. Figure 4.11 shows PAM clustering with different transformations of the data:

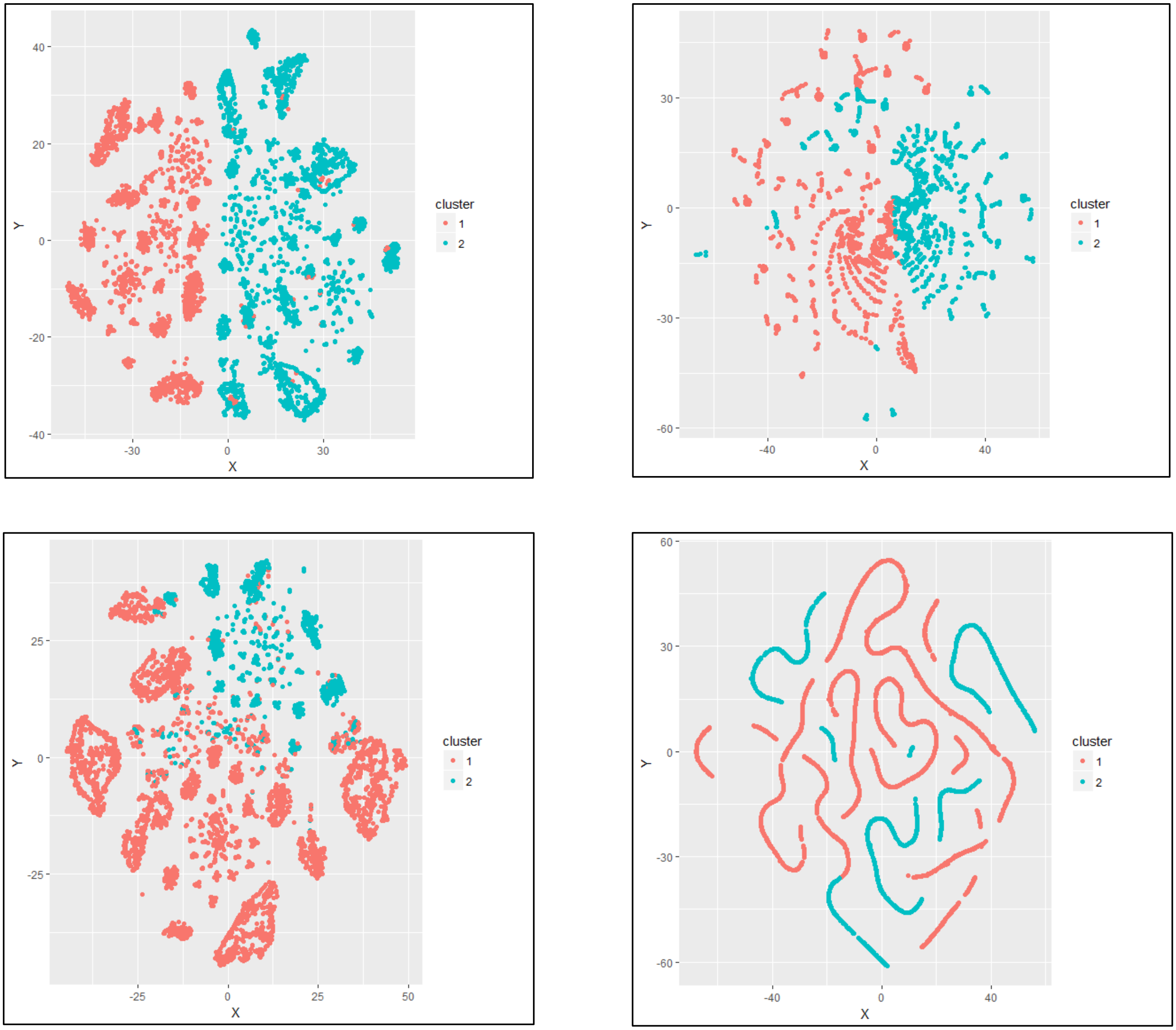


Figure 4.11 Figure of different visualizations from different data transformations

Eventually, the unaltered dataset was settled upon for further analysis. This however, presented a complication. How best to cluster a mixed dataset? Enter Gower dissimilarity measure [25]. Gower is a formula that measures the distance between two data points. It is contained in daisy, a R function that returns a dissimilarity matrix. While K-Means algorithms were unable to run mixed datasets, there existed partitioning algorithms that were able to do so, such as K-Modes.

4.6.1 K-Modes

K-modes is an algorithm which partitions a dataset into discrete clusters. K-Modes was first introduced in the 1997 paper, 'A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining', Huang. The algorithm overcomes the main limitation of the K-Means, which only manipulates numeric data. This is due to the algorithm using Euclidean Distance [25]. To prepare the dataset, the code na.omit was run to remove all of the rows that contained NA values in the dataset.

```
dd_for_kmodes <- na.omit(dd_random)
```

Removes rows with any NA values

```
cluster_fit <- kmodes(dd_for_kmodes[,-1], 2, iter.max = 10, weighted = FALSE )
```

4.6.2 Gower Dissimilarity Matrix

K-Modes can handle mixed datasets. The other algorithms in the list require a dissimilarity measure. A dissimilarity measure is a mathematical formula used to describe the distances between data points. Gower, unlike other distance measures, such as Euclidian, can calculate the distance values between data points in mixed datasets. The output of the Gower function in daisy is a dissimilarity matrix. The other algorithms in this list accept it as a parameter, and subsequently performs the requisite cluster analysis.

To produce the dissimilarity matrix in R for the dataset, the following code was used:

```
gower_dist <- daisy(diabetes_data[, -1],
                    metric = "gower",
                    type = list(symm = 1))
```

This creates a dissimilarity matrix, stored in the variable gower_dist

4.6.3 PAM – Partitioning Around Medoids

PAM, or Partitioning Around Medoids, is an algorithm that functions by picking points in a dataset of medoids. The algorithm then clusters on certain data points in the center and including the values that have points that are closest to it. Silhouette width can be used to determine the optimum number of clusters. To visualize PAM, like other partitioning algorithms, requires Principal Component Analysis or similar dimensionality compression algorithms to reduce the number of dimensions to either 2 or 3, which enables human-readable plots.

```
pam_fit <- pam(gower_dist, diss = TRUE, k = 8)
#where k = the number of clusters.
```

4.6.4 HClust

Hclust is an agglomerative hierarchical algorithm that works by treating each data point as a cluster. It groups individual clusters to the cluster nearest to them, using the dissimilarity measure created using gower. This process happens recursively, connecting larger and larger clusters until the entire dataset is connected. As it is a Hierarchical means of clustering, there is no need for transformations in order the view how the clusters converge. Hclust, like other hierarchical clustering algorithms, represent their results on a dendrogram. A dendrogram is a long, tree-like representation of the data, represented by levels. To derive cluster statistics, one must “cut” the tree at a certain level. That level corresponds to the number of clusters. Hclust, while generating more accurate results, can be sensitive to noise, and has a longer running time than Partitioning Methods.

To perform hclust on data. The dissimilarity matrix was used as a parameter.

```
hgroup <- cutree(d.hclust, 4)
```

To cut the tree, the first parameter is the hclust variable, and the second is the level, or number of clusters.

```
d.hclust = hclust(gower_dist)
```

Command to plot the dendrogram

```
plot(d.hclust)
```

4.6.5 DIANA - Divisible Analysis

The second hierarchical clustering algorithm that is being is DIANA, also known as Divisive Analysis [19]. Diana groups the entire dataset as a cluster, and does the opposite of HCLUST, splitting into sub clusters based on the distances of the points that are farther away. It does this by using the gower dissimilarity matrix that is fed to it as a parameter. In R, Diana has both a banner and dendrogram representation.

It’s banner representation enables the user to tell what the most distinct cluster groupings are in the dataset. Then as, with hclust, the user can “cut” the tree, ascertaining the number of nodes.

```
d.dclust = diana(gower_dist)
```

To perform DIANA on data. The dissimilarity matrix is used as a parameter.

```
dgroup <- cutree(d.dclust, 4)
```

To cut the tree, the first parameter is the DIANA variable, and the second is the level, or number of clusters.

4.6.6 Visualization

To make the results of cluster partitioning readable to the eye, compression was needed. Electronic health records typically suffer from high dimensionality, which makes them difficult to map visually, requiring dimensions of either 2, or 3. An algorithm is need to do that compression of dimensionality, without compromising the value and structure of the data. For this research, the algorithm chosen was t-distributed stochastic neighbor embossing (t-SNE). [22]

T-SNE works by creating a probability distribution among data in the high-dimensionality set. Similar points of data are given a higher probability, less similar points are given a lower probability. The algorithm does again to points in a low-dimensional map, and then proceeds to minimize the Kullback-Leibler divergence. This algorithm is used by R, which is utilized here for our visualizations. The R code is shown below.

```
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)


tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         name = diabetes_data$PATIENT_NUMBER)


ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```

CHAPTER 5. RESULTS

This section provides the result visualizations and clusters for each algorithm used on the data. Each algorithm clusters the data into 2 clusters, 3 clusters, 4 clusters, and 5 clusters. While silhouette analysis (shown in Figure 5.1) determined that optimum number of clusters to be two, the silhouette widths were still sufficient that more information can be gleaned for up to 5 clusters. In this section, Figure 5.2, Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.6 each show visualizations of K-Mode, PAM, Hclust, and DIANA respectively, along with results of each of their one, two, three, four, and five cluster sets.

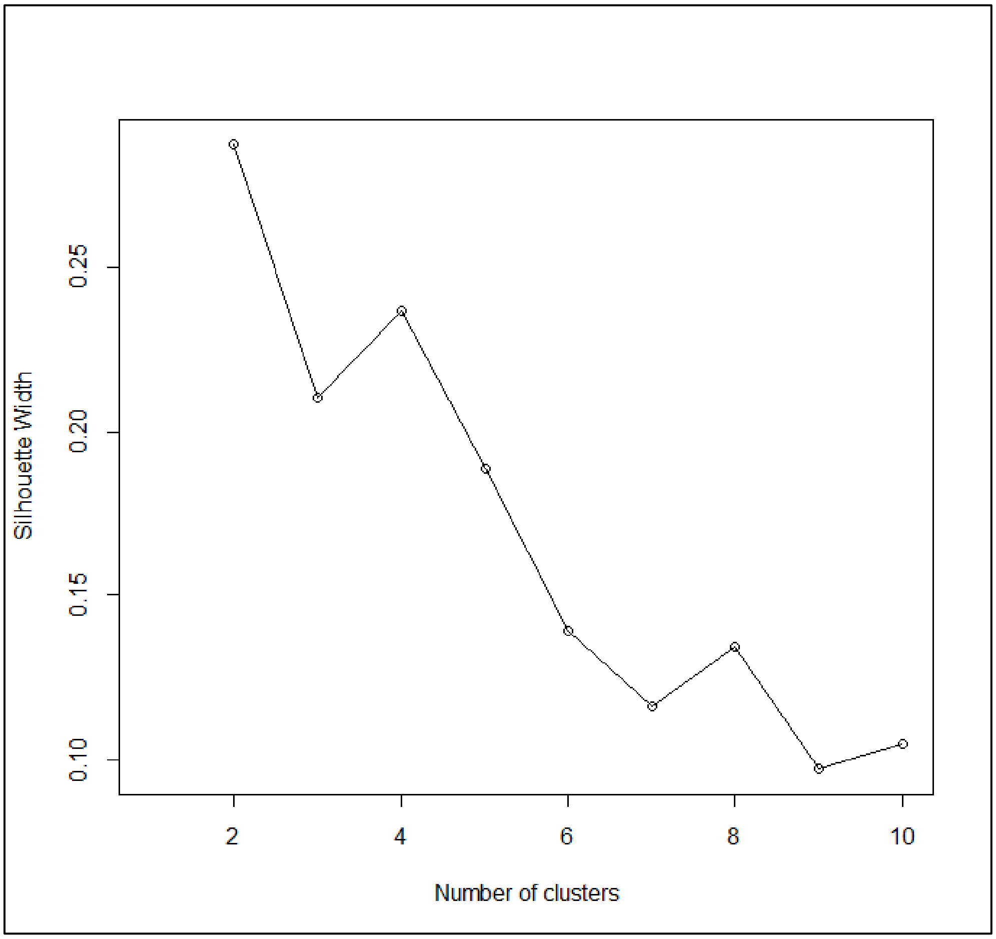
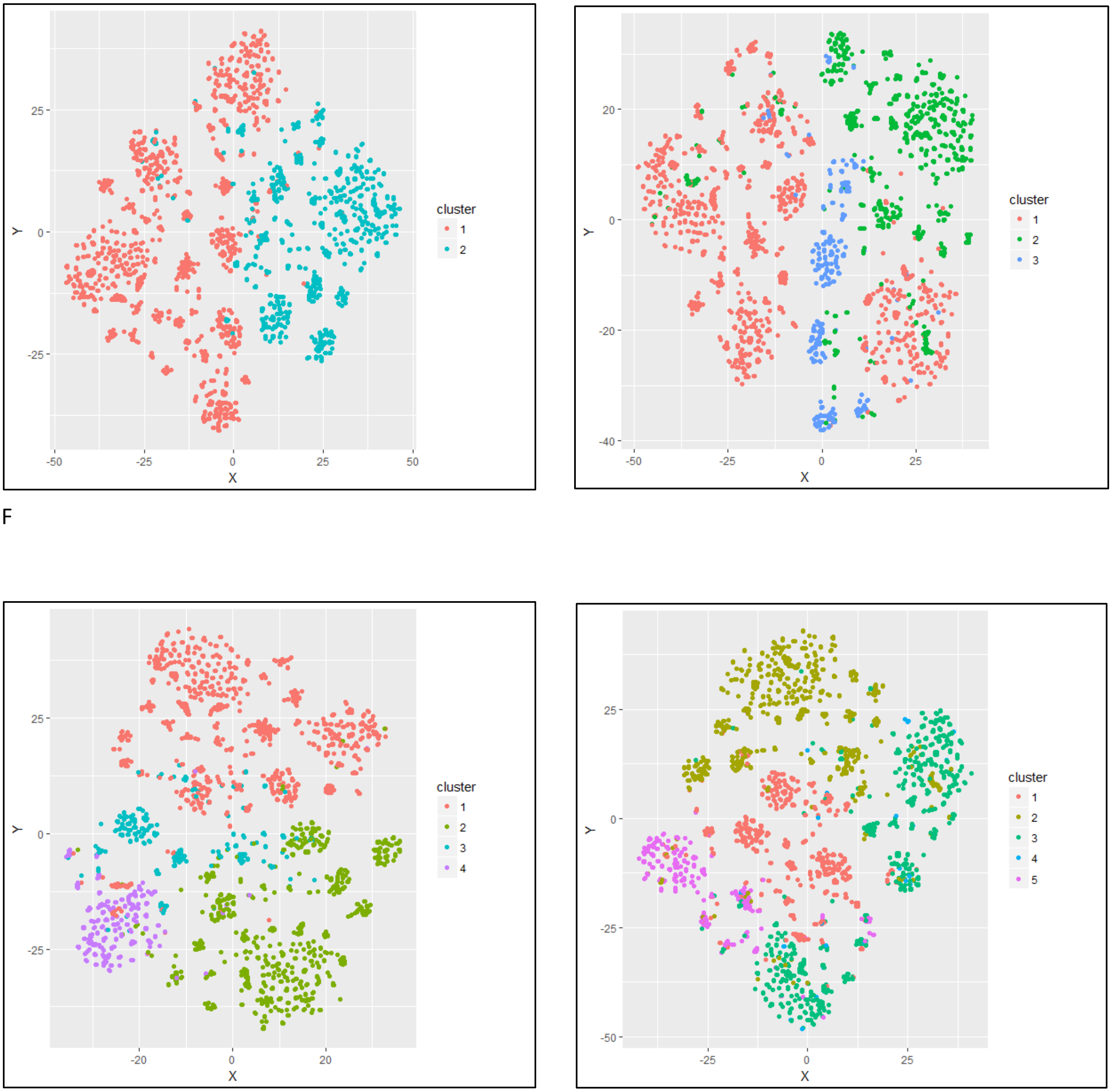


Figure 5.1 Silhouette width of PAM clusters

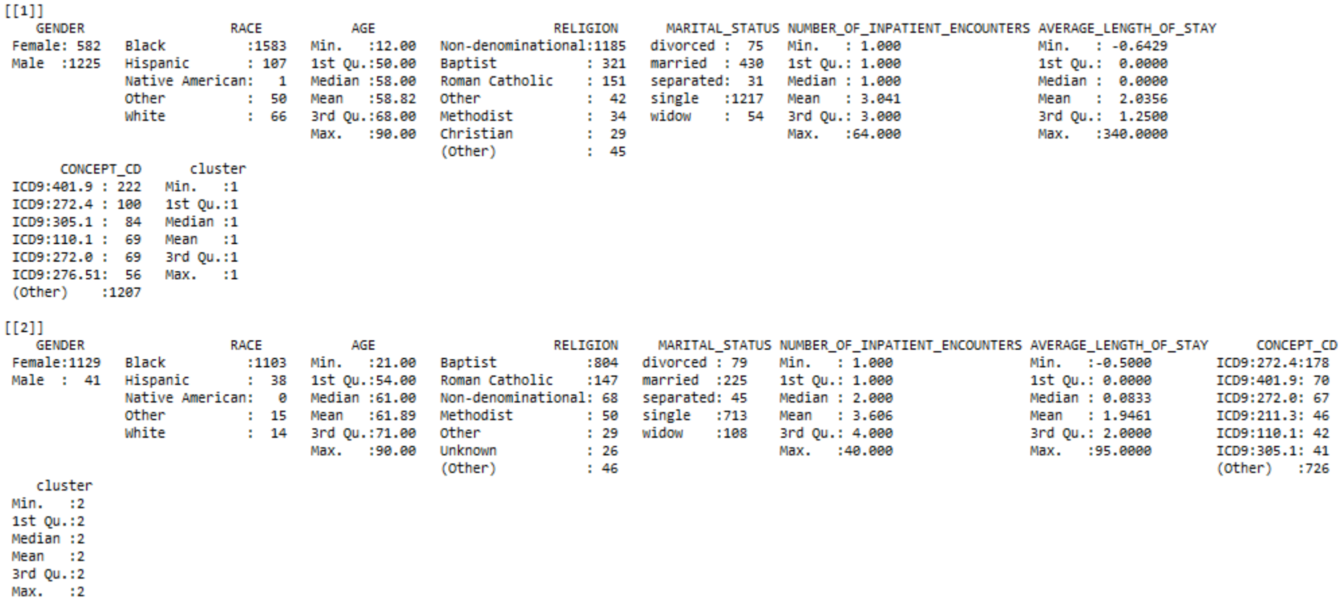
5.1 K-Modes Clusters



F

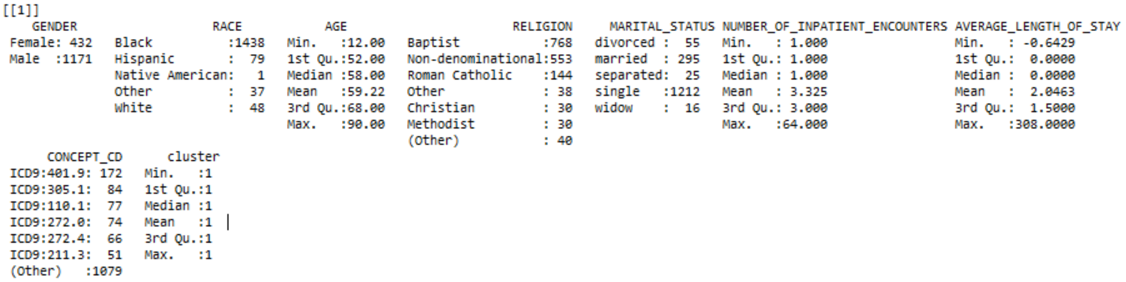
Figure 5.2 Visualizations of K-Modes clusters

2 Cluster Summary:

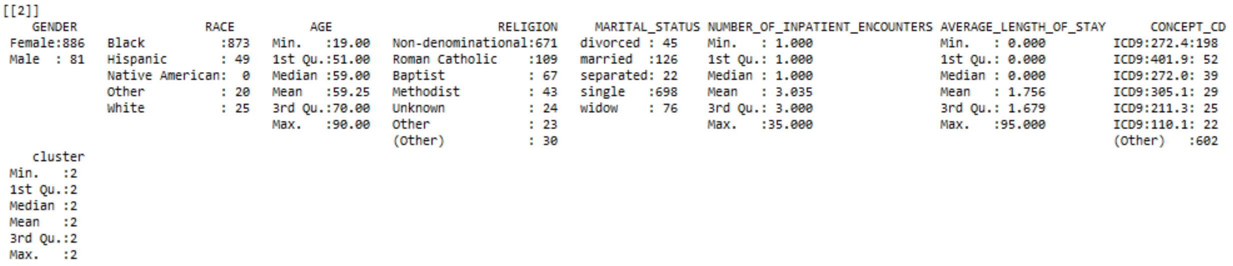


```
[[1]]
    GENDER                  RACE           AGE                      RELIGION      MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female: 582   Black           :1583   Min.   :12.00   Non-denominational:1185   divorced :  75   Min.   : 1.000                 Min.   : -0.6429  
 Male  :1225   Hispanic        : 107   1st Qu.:50.00   Baptist           : 321   married  : 430   1st Qu.: 1.000                 1st Qu.:  0.0000  
               Native American:   1   Median :58.00   Roman Catholic    : 151   separated:  31   Median : 1.000                 Median :  0.0000  
               Other           :  50   Mean   :58.82   Other             :  42   single   :1217   Mean   : 3.041                 Mean   :  2.0356  
               White           :  66   3rd Qu.:68.00   Methodist         :  34   widow    :  54   3rd Qu.: 3.000                 3rd Qu.:  1.2500  
                                       Max.   :90.00   Christian         :  29                    Max.   :64.000                 Max.   :340.0000  
                                                       (Other)           :  45                                                                    
      CONCEPT_CD      cluster 
 ICD9:401.9 : 222   Min.   :1  
 ICD9:272.4 : 100   1st Qu.:1  
 ICD9:305.1 :  84   Median :1  
 ICD9:110.1 :  69   Mean   :1  
 ICD9:272.0 :  69   3rd Qu.:1  
 ICD9:276.51:  56   Max.   :1  
 (Other)    :1207              

[[2]]
    GENDER                  RACE           AGE                      RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD 
 Female:1129   Black           :1103   Min.   :21.00   Baptist           :804   divorced : 79   Min.   : 1.000                 Min.   :-0.5000        ICD9:272.4:178  
 Male  :  41   Hispanic        :  38   1st Qu.:54.00   Roman Catholic    :147   married  :225   1st Qu.: 1.000                 1st Qu.: 0.0000        ICD9:401.9: 70  
               Native American:   0   Median :61.00   Non-denominational: 68   separated: 45   Median : 2.000                 Median : 0.0833        ICD9:272.0: 67  
               Other           :  15   Mean   :61.89   Methodist         : 50   single   :713   Mean   : 3.606                 Mean   : 1.9461        ICD9:211.3: 46  
               White           :  14   3rd Qu.:71.00   Other             : 29   widow    :108   3rd Qu.: 4.000                 3rd Qu.: 2.0000        ICD9:110.1: 42  
                                       Max.   :90.00   Unknown           : 26                   Max.   :40.000                 Max.   :95.0000        ICD9:305.1: 41  
                                                       (Other)           : 46                                                                         (Other)   :726  
    cluster 
 Min.   :2  
 1st Qu.:2  
 Median :2  
 Mean   :2  
 3rd Qu.:2  
 Max.   :2  
```

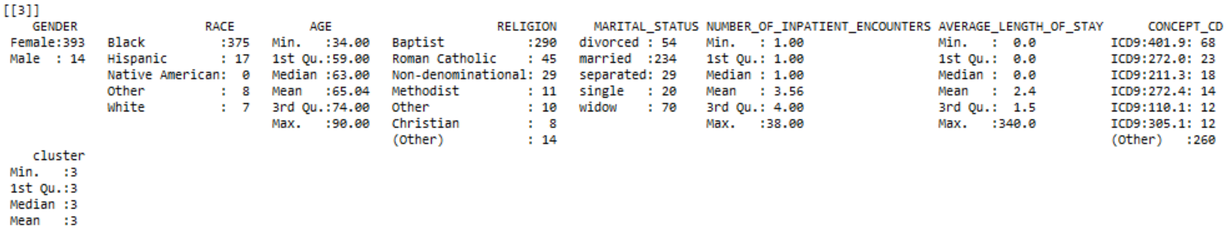
3 Cluster Summary:



```
[[1]]
    GENDER                  RACE           AGE                      RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female: 432   Black           :1438   Min.   :12.00   Baptist           :768   divorced :  55   Min.   : 1.000                 Min.   : -0.6429  
 Male  :1171   Hispanic        :  79   1st Qu.:52.00   Non-denominational:553   married  : 295   1st Qu.: 1.000                 1st Qu.:  0.0000  
               Native American:   1   Median :58.00   Roman Catholic    :144   separated:  25   Median : 1.000                 Median :  0.0000  
               Other           :  37   Mean   :59.22   Other             : 38   single   :1212   Mean   : 3.325                 Mean   :  2.0463  
               White           :  48   3rd Qu.:68.00   Christian         : 30   widow    :  16   3rd Qu.: 3.000                 3rd Qu.:  1.5000  
                                       Max.   :90.00   Methodist         : 30                    Max.   :64.000                 Max.   :308.0000  
                                                       (Other)           : 40                                                                    
      CONCEPT_CD      cluster 
 ICD9:401.9: 172   Min.   :1  
 ICD9:305.1:  84   1st Qu.:1  
 ICD9:110.1:  77   Median :1  
 ICD9:272.0:  74   Mean   :1  |
 ICD9:272.4:  66   3rd Qu.:1  
 ICD9:211.3:  51   Max.   :1  
 (Other)   :1079              
```

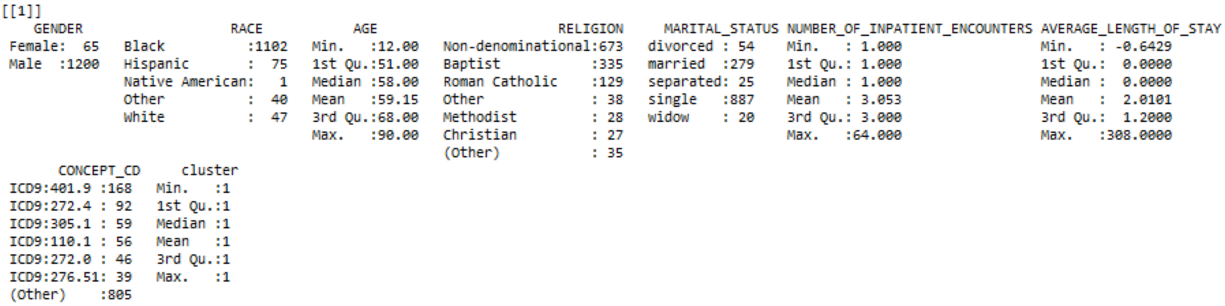


```
[[2]]
    GENDER                 RACE           AGE                      RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD 
 Female:886   Black           :873   Min.   :19.00   Non-denominational:671   divorced : 45   Min.   : 1.000                 Min.   : 0.000          ICD9:272.4:198  
 Male  : 81   Hispanic        : 49   1st Qu.:51.00   Roman Catholic    :109   married  :126   1st Qu.: 1.000                 1st Qu.: 0.000          ICD9:401.9: 52  
              Native American:  0   Median :59.00   Baptist           : 67   separated: 22   Median : 1.000                 Median : 0.000          ICD9:272.0: 39  
              Other           : 20   Mean   :59.25   Methodist         : 43   single   :698   Mean   : 3.035                 Mean   : 1.756          ICD9:305.1: 29  
              White           : 25   3rd Qu.:70.00   Unknown           : 24   widow    : 76   3rd Qu.: 3.000                 3rd Qu.: 1.679          ICD9:211.3: 25  
                                     Max.   :90.00   Other             : 23                   Max.   :35.000                 Max.   :95.000          ICD9:110.1: 22  
                                                     (Other)           : 30                                                                         (Other)   :602  
    cluster 
 Min.   :2  
 1st Qu.:2  
 Median :2  
 Mean   :2  
 3rd Qu.:2  
 Max.   :2  
```

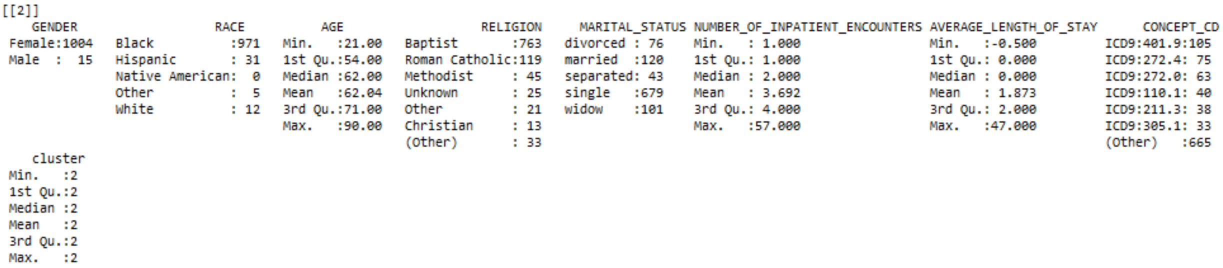


```
[[3]]
   GENDER                RACE          AGE                       RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY     CONCEPT_CD
 Female:393   Black          :375   Min.   :34.00   Baptist            :290   divorced : 54   Min.   : 1.00                  Min.   :  0.0         ICD9:401.9: 68
 Male  : 14   Hispanic       : 17   1st Qu.:59.00   Roman Catholic     : 45   married  :234   1st Qu.: 1.00                  1st Qu.:  0.0         ICD9:272.0: 23
              Native American:  0   Median :63.00   Non-denominational: 29   separated: 29   Median : 1.00                  Median :  0.0         ICD9:211.3: 18
              Other          :  8   Mean   :65.04   Methodist          : 11   single   : 20   Mean   : 3.56                  Mean   :  2.4         ICD9:272.4: 14
              White          :  7   3rd Qu.:74.00   Other              : 10   widow    : 70   3rd Qu.: 4.00                  3rd Qu.:  1.5         ICD9:110.1: 12
                                    Max.   :90.00   Christian          :  8                   Max.   :38.00                  Max.   :340.0         ICD9:305.1: 12
                                                    (Other)            : 14                                                                        (Other)   :260
    cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
```

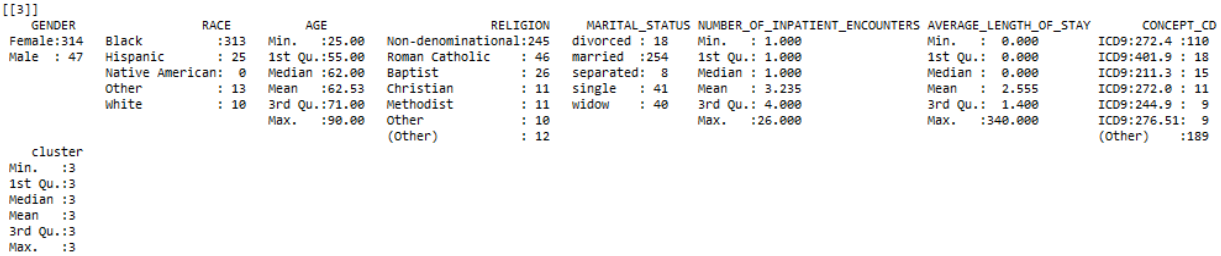
4 Cluster Summary:



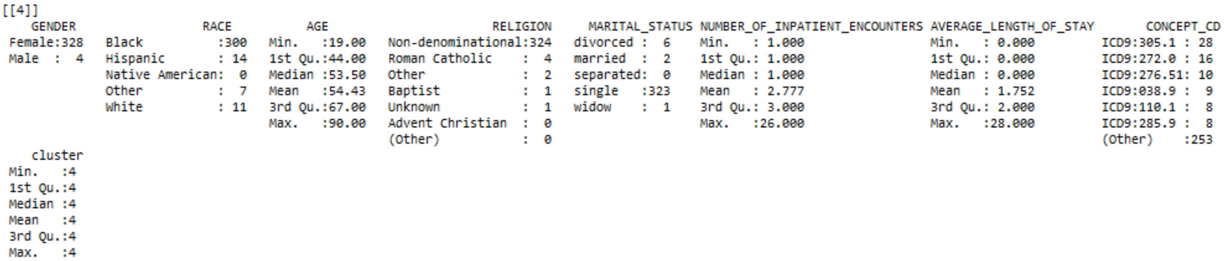
```
[[1]]
   GENDER               RACE            AGE                        RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:  65   Black          :1102   Min.   :12.00   Non-denominational:673   divorced : 54   Min.   : 1.000                 Min.   : -0.6429
 Male  :1200   Hispanic       :  75   1st Qu.:51.00   Baptist           :335   married  :279   1st Qu.: 1.000                 1st Qu.:  0.0000
               Native American:   1   Median :58.00   Roman Catholic    :129   separated: 25   Median : 1.000                 Median :  0.0000
               Other          :  40   Mean   :59.15   Other             : 38   single   :887   Mean   : 3.053                 Mean   :  2.0101
               White          :  47   3rd Qu.:68.00   Methodist         : 28   widow    : 20   3rd Qu.: 3.000                 3rd Qu.:  1.2000
                                      Max.   :90.00   Christian         : 27                   Max.   :64.000                 Max.   :308.0000
                                                      (Other)           : 35
      CONCEPT_CD      cluster
 ICD9:401.9 :168   Min.   :1
 ICD9:272.4 : 92   1st Qu.:1
 ICD9:305.1 : 59   Median :1
 ICD9:110.1 : 56   Mean   :1
 ICD9:272.0 : 46   3rd Qu.:1
 ICD9:276.51: 39   Max.   :1
 (Other)    :805
```



```
[[2]]
   GENDER                 RACE          AGE                 RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:1004   Black          :971   Min.   :21.00   Baptist       :763   divorced : 76   Min.   : 1.000                 Min.   :-0.500          ICD9:401.9:105
 Male  :  15   Hispanic       : 31   1st Qu.:54.00   Roman Catholic:119   married  :120   1st Qu.: 1.000                 1st Qu.: 0.000          ICD9:272.4: 75
               Native American:  0   Median :62.00   Methodist     : 45   separated: 43   Median : 2.000                 Median : 0.000          ICD9:272.0: 63
               Other          :  5   Mean   :62.04   Unknown       : 25   single   :679   Mean   : 3.692                 Mean   : 1.873          ICD9:110.1: 40
               White          : 12   3rd Qu.:71.00   Other         : 21   widow    :101   3rd Qu.: 4.000                 3rd Qu.: 2.000          ICD9:211.3: 38
                                     Max.   :90.00   Christian     : 13                   Max.   :57.000                 Max.   :47.000          ICD9:305.1: 33
                                                     (Other)       : 33                                                                          (Other)   :665
    cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```

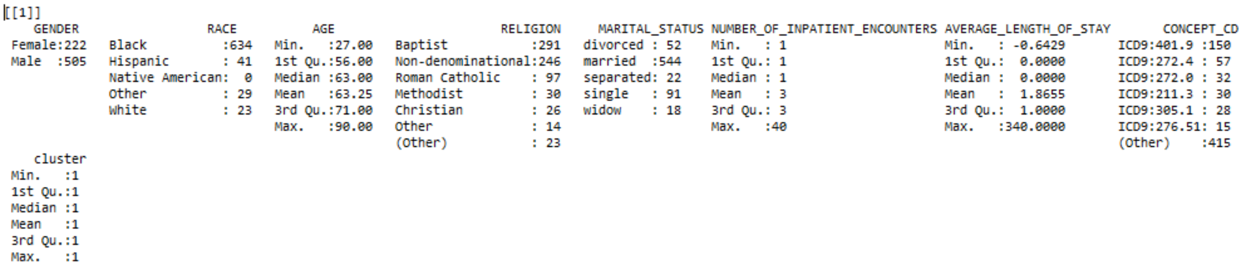


```
[[3]]
   GENDER               RACE          AGE                        RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:314   Black          :313   Min.   :25.00   Non-denominational:245   divorced : 18   Min.   : 1.000                 Min.   :  0.000         ICD9:272.4 :110
 Male  : 47   Hispanic       : 25   1st Qu.:55.00   Roman Catholic    : 46   married  :254   1st Qu.: 1.000                 1st Qu.:  0.000         ICD9:401.9 : 18
              Native American:  0   Median :62.00   Baptist           : 26   separated:  8   Median : 1.000                 Median :  0.000         ICD9:211.3 : 15
              Other          : 13   Mean   :62.53   Christian         : 11   single   : 41   Mean   : 3.235                 Mean   :  2.555         ICD9:272.0 : 11
              White          : 10   3rd Qu.:71.00   Methodist         : 11   widow    : 40   3rd Qu.: 4.000                 3rd Qu.:  1.400         ICD9:244.9 :  9
                                    Max.   :90.00   Other             : 10                   Max.   :26.000                 Max.   :340.000         ICD9:276.51:  9
                                                    (Other)           : 12                                                                          (Other)    :189
    cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
 3rd Qu.:3
 Max.   :3
```

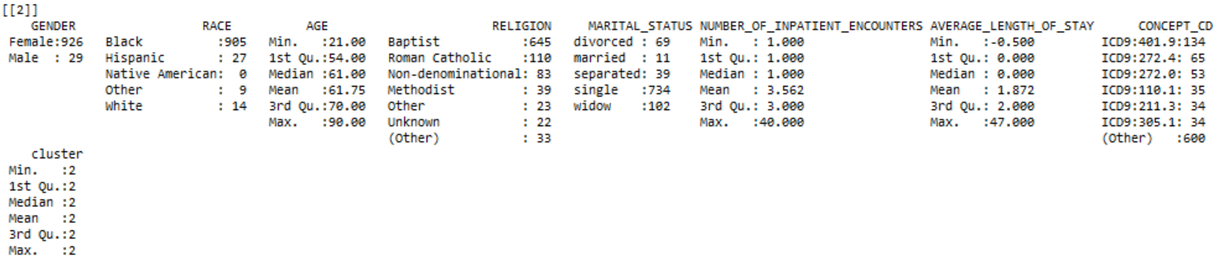


```
[[4]]
   GENDER               RACE          AGE                        RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:328   Black          :300   Min.   :19.00   Non-denominational:324   divorced :  6   Min.   : 1.000                 Min.   : 0.000          ICD9:305.1 : 28
 Male  :  4   Hispanic       : 14   1st Qu.:44.00   Roman Catholic    :  4   married  :  2   1st Qu.: 1.000                 1st Qu.: 0.000          ICD9:272.0 : 16
              Native American:  0   Median :53.50   Other             :  2   separated:  0   Median : 1.000                 Median : 0.000          ICD9:276.51: 10
              Other          :  7   Mean   :54.43   Baptist           :  1   single   :323   Mean   : 2.777                 Mean   : 1.752          ICD9:038.9 :  9
              White          : 11   3rd Qu.:67.00   Unknown           :  1   widow    :  1   3rd Qu.: 3.000                 3rd Qu.: 2.000          ICD9:110.1 :  8
                                    Max.   :90.00   Advent Christian  :  0                   Max.   :26.000                 Max.   :28.000          ICD9:285.9 :  8
                                                    (Other)           :  0                                                                          (Other)    :253
    cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```

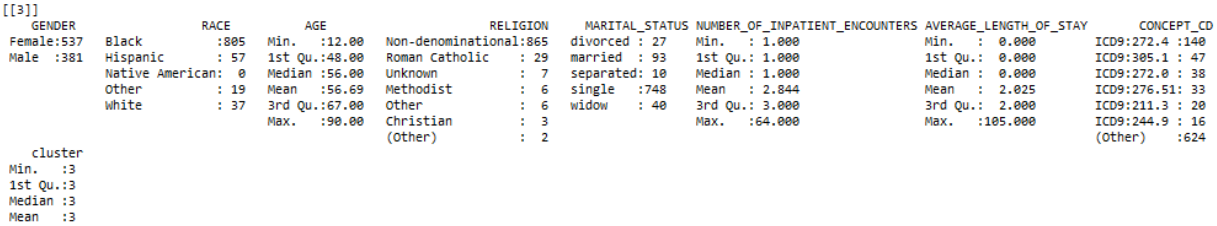
5 Cluster Summary:



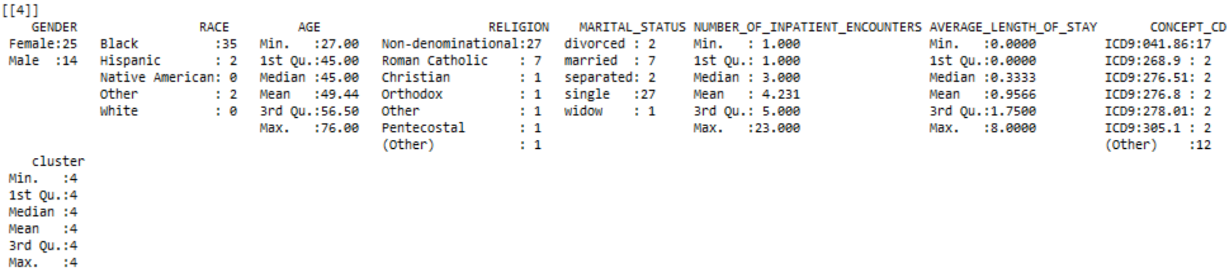
```
[[1]]
   GENDER                 RACE          AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female:222   Black           :634   Min.   :27.00   Baptist              :291   divorced : 52   Min.   : 1                      Min.   : -0.6429       ICD9:401.9 :150
 Male  :505   Hispanic        : 41   1st Qu.:56.00   Non-denominational:246   married  :544   1st Qu.: 1                      1st Qu.:  0.0000       ICD9:272.4 : 57
              Native American:  0   Median :63.00   Roman Catholic     : 97   separated: 22   Median : 1                      Median :  0.0000       ICD9:272.0 : 32
              Other           : 29   Mean   :63.25   Methodist          : 30   single   : 91   Mean   : 3                      Mean   :  1.8655       ICD9:211.3 : 30
              White           : 23   3rd Qu.:71.00   Christian          : 26   widow    : 18   3rd Qu.: 3                      3rd Qu.:  1.0000       ICD9:305.1 : 28
                                     Max.   :90.00   Other              : 14                  Max.   :40                      Max.   :340.0000       ICD9:276.51: 15
                                                     (Other)            : 23                                                                         (Other)    :415
   cluster
 Min.   :1
 1st Qu.:1
 Median :1
 Mean   :1
 3rd Qu.:1
 Max.   :1
```



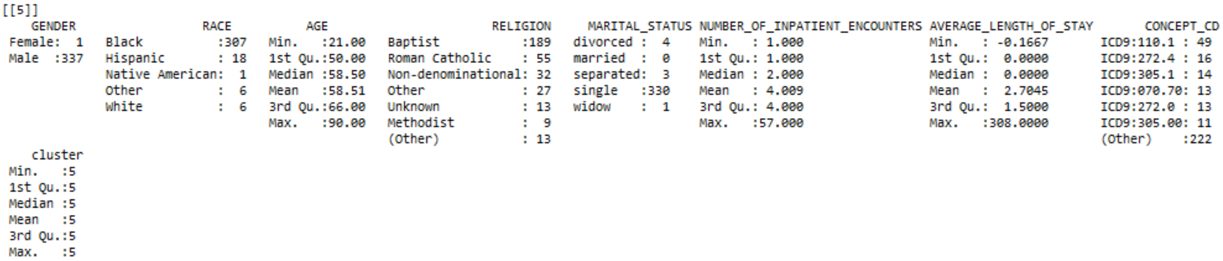
```
[[2]]
   GENDER                 RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:926   Black           :905   Min.   :21.00   Baptist             :645   divorced : 69   Min.   : 1.000                 Min.   :-0.500         ICD9:401.9:134
 Male  : 29   Hispanic        : 27   1st Qu.:54.00   Roman Catholic      :110   married  : 11   1st Qu.: 1.000                 1st Qu.: 0.000         ICD9:272.4: 65
              Native American:  0   Median :61.00   Non-denominational: 83   separated: 39   Median : 1.000                 Median : 0.000         ICD9:272.0: 53
              Other           :  9   Mean   :61.75   Methodist           : 39   single   :734   Mean   : 3.562                 Mean   : 1.872         ICD9:110.1: 35
              White           : 14   3rd Qu.:70.00   Other               : 23   widow    :102   3rd Qu.: 3.000                 3rd Qu.: 2.000         ICD9:211.3: 34
                                     Max.   :90.00   Unknown             : 22                   Max.   :40.000                 Max.   :47.000         ICD9:305.1: 34
                                                     (Other)             : 33                                                                         (Other)   :600
   cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```



```
[[3]]
   GENDER                 RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:537   Black           :805   Min.   :12.00   Non-denominational:865   divorced : 27   Min.   : 1.000                 Min.   :  0.000        ICD9:272.4 :140
 Male  :381   Hispanic        : 57   1st Qu.:48.00   Roman Catholic     : 29   married  : 93   1st Qu.: 1.000                 1st Qu.:  0.000        ICD9:305.1 : 47
              Native American:  0   Median :56.00   Unknown            :  7   separated: 10   Median : 1.000                 Median :  0.000        ICD9:272.0 : 38
              Other           : 19   Mean   :56.69   Methodist          :  6   single   :748   Mean   : 2.844                 Mean   :  2.025        ICD9:276.51: 33
              White           : 37   3rd Qu.:67.00   Other              :  6   widow    : 40   3rd Qu.: 3.000                 3rd Qu.:  2.000        ICD9:211.3 : 20
                                     Max.   :90.00   Christian          :  3                  Max.   :64.000                 Max.   :105.000        ICD9:244.9 : 16
                                                     (Other)            :  2                                                                        (Other)    :624
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
```



```
[[4]]
   GENDER                RACE          AGE                        RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female:25   Black           :35   Min.   :27.00   Non-denominational:27   divorced : 2    Min.   : 1.000                 Min.   :0.0000          ICD9:041.86:17
 Male  :14   Hispanic        : 2   1st Qu.:45.00   Roman Catholic     : 7   married  : 7    1st Qu.: 1.000                 1st Qu.:0.0000          ICD9:268.9 : 2
             Native American: 0   Median :45.00   Christian          : 1   separated: 2    Median : 3.000                 Median :0.3333          ICD9:276.51: 2
             Other           : 2   Mean   :49.44   Orthodox           : 1   single   :27    Mean   : 4.231                 Mean   :0.9566          ICD9:276.8 : 2
             White           : 0   3rd Qu.:56.50   Other              : 1   widow    : 1    3rd Qu.: 5.000                 3rd Qu.:1.7500          ICD9:278.01: 2
                                   Max.   :76.00   Pentecostal        : 1                   Max.   :23.000                 Max.   :8.0000          ICD9:305.1 : 2
                                                   (Other)            : 1                                                                          (Other)    :12
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```



```
[[5]]
   GENDER                 RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:  1   Black           :307   Min.   :21.00   Baptist             :189   divorced :  4   Min.   : 1.000                 Min.   : -0.1667       ICD9:110.1 : 49
 Male  :337   Hispanic        : 18   1st Qu.:50.00   Roman Catholic      : 55   married  :  0   1st Qu.: 1.000                 1st Qu.:  0.0000       ICD9:272.4 : 16
              Native American:  1   Median :58.50   Non-denominational: 32   separated:  3   Median : 2.000                 Median :  0.0000       ICD9:305.1 : 14
              Other           :  6   Mean   :58.51   Other               : 27   single   :330   Mean   : 4.009                 Mean   :  2.7045       ICD9:070.70: 13
              White           :  6   3rd Qu.:66.00   Unknown             : 13   widow    :  1   3rd Qu.: 4.000                 3rd Qu.:  1.5000       ICD9:272.0 : 13
                                     Max.   :90.00   Methodist           :  9                   Max.   :57.000                 Max.   :308.0000       ICD9:305.00: 11
                                                     (Other)             : 13                                                                         (Other)    :222
   cluster
 Min.   :5
 1st Qu.:5
 Median :5
 Mean   :5
 3rd Qu.:5
 Max.   :5
```

5.2 PAM Clusters

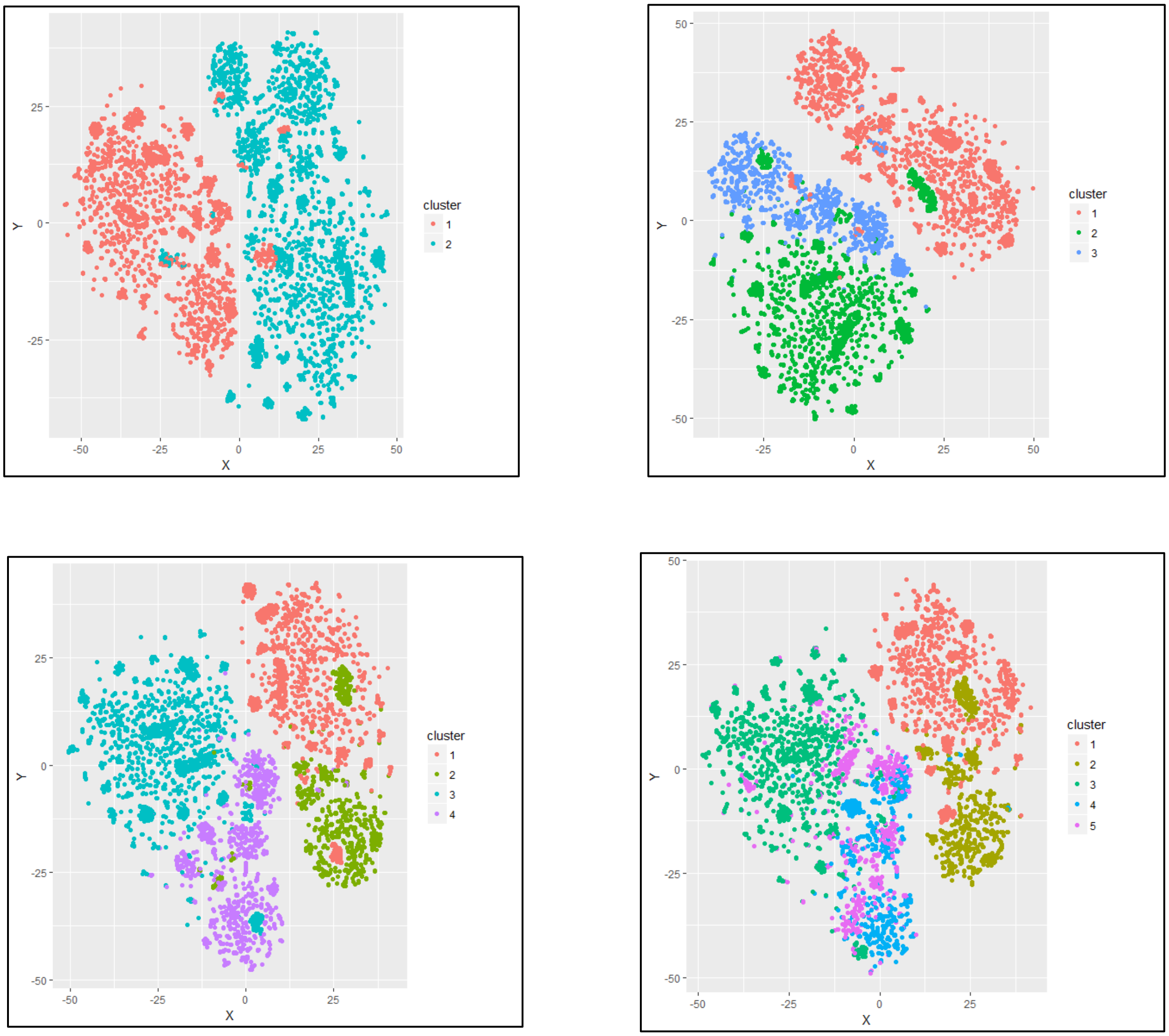
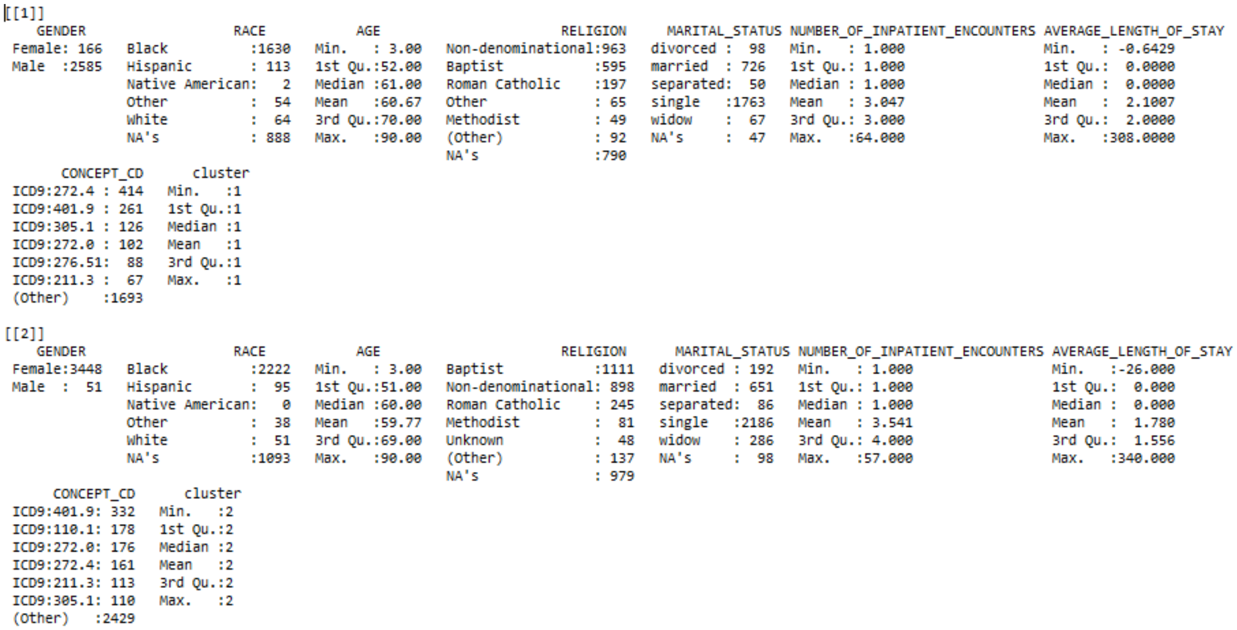


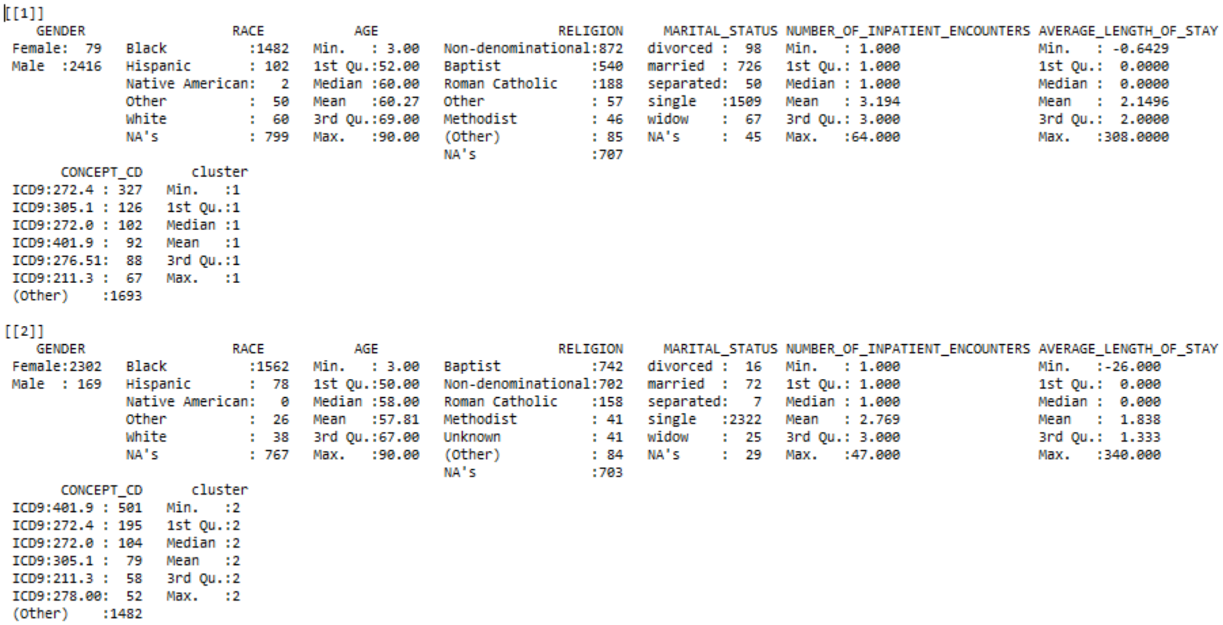
Figure 5.3 Visualizations of PAM clusters

2 Cluster Summary:

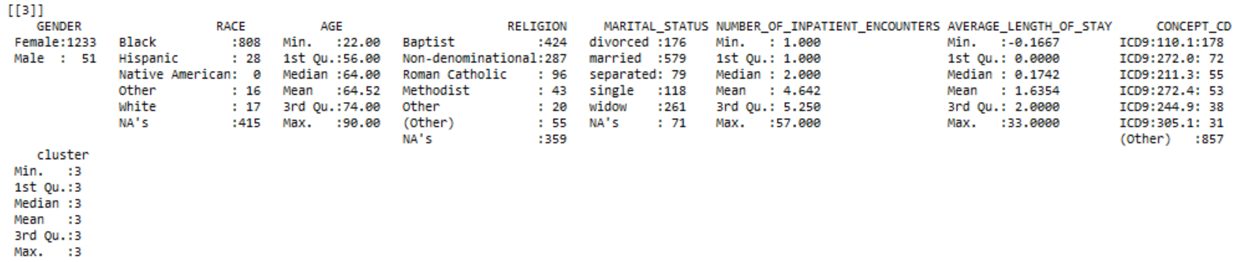


```
[[1]]
   GENDER                   RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female: 166   Black           :1630   Min.   : 3.00   Non-denominational:963   divorced :  98   Min.   : 1.000                  Min.   : -0.6429
 Male  :2585   Hispanic        : 113   1st Qu.:52.00   Baptist           :595   married  : 726   1st Qu.: 1.000                  1st Qu.:  0.0000
               Native American:   2   Median :61.00   Roman Catholic    :197   separated:  50   Median : 1.000                  Median :  0.0000
               Other           :  54   Mean   :60.67   Other             : 65   single   :1763   Mean   : 3.047                  Mean   :  2.1007
               White           :  64   3rd Qu.:70.00   Methodist         : 49   widow    :  67   3rd Qu.: 3.000                  3rd Qu.:  2.0000
               NA's            : 888   Max.   :90.00   (Other)           : 92   NA's     :  47   Max.   :64.000                  Max.   :308.0000
                                                       NA's              :790
      CONCEPT_CD      cluster
 ICD9:272.4 : 414   Min.   :1
 ICD9:401.9 : 261   1st Qu.:1
 ICD9:305.1 : 126   Median :1
 ICD9:272.0 : 102   Mean   :1
 ICD9:276.51:  88   3rd Qu.:1
 ICD9:211.3 :  67   Max.   :1
 (Other)    :1693

[[2]]
   GENDER                   RACE          AGE                        RELIGION      MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:3448   Black           :2222   Min.   : 3.00   Baptist            :1111   divorced : 192   Min.   : 1.000                  Min.   :-26.000
 Male  :  51   Hispanic        :  95   1st Qu.:51.00   Non-denominational: 898   married  : 651   1st Qu.: 1.000                  1st Qu.:  0.000
               Native American:   0   Median :60.00   Roman Catholic     : 245   separated:  86   Median : 1.000                  Median :  0.000
               Other           :  38   Mean   :59.77   Methodist          :  81   single   :2186   Mean   : 3.541                  Mean   :  1.780
               White           :  51   3rd Qu.:69.00   Unknown            :  48   widow    : 286   3rd Qu.: 4.000                  3rd Qu.:  1.556
               NA's            :1093   Max.   :90.00   (Other)            : 137   NA's     :  98   Max.   :57.000                  Max.   :340.000
                                                       NA's               : 979
     CONCEPT_CD      cluster
 ICD9:401.9: 332   Min.   :2
 ICD9:110.1: 178   1st Qu.:2
 ICD9:272.0: 176   Median :2
 ICD9:272.4: 161   Mean   :2
 ICD9:211.3: 113   3rd Qu.:2
 ICD9:305.1: 110   Max.   :2
 (Other)   :2429
```

3 Cluster Summary:

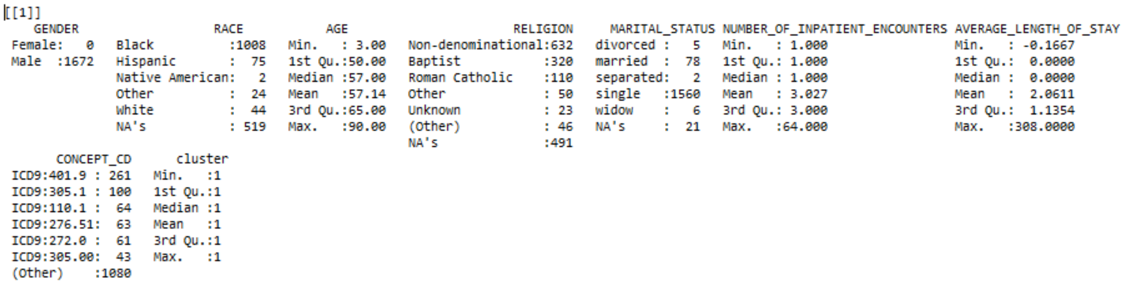


```
[[1]]
   GENDER                   RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:  79   Black           :1482   Min.   : 3.00   Non-denominational:872   divorced :  98   Min.   : 1.000                  Min.   : -0.6429
 Male  :2416   Hispanic        : 102   1st Qu.:52.00   Baptist           :540   married  : 726   1st Qu.: 1.000                  1st Qu.:  0.0000
               Native American:   2   Median :60.00   Roman Catholic    :188   separated:  50   Median : 1.000                  Median :  0.0000
               Other           :  50   Mean   :60.27   Other             : 57   single   :1509   Mean   : 3.194                  Mean   :  2.1496
               White           :  60   3rd Qu.:69.00   Methodist         : 46   widow    :  67   3rd Qu.: 3.000                  3rd Qu.:  2.0000
               NA's            : 799   Max.   :90.00   (Other)           : 85   NA's     :  45   Max.   :64.000                  Max.   :308.0000
                                                       NA's              :707
      CONCEPT_CD      cluster
 ICD9:272.4 : 327   Min.   :1
 ICD9:305.1 : 126   1st Qu.:1
 ICD9:272.0 : 102   Median :1
 ICD9:401.9 :  92   Mean   :1
 ICD9:276.51:  88   3rd Qu.:1
 ICD9:211.3 :  67   Max.   :1
 (Other)    :1693

[[2]]
   GENDER                   RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:2302   Black           :1562   Min.   : 3.00   Baptist           :742   divorced :  16   Min.   : 1.000                  Min.   :-26.000
 Male  : 169   Hispanic        :  78   1st Qu.:50.00   Non-denominational:702   married  :  72   1st Qu.: 1.000                  1st Qu.:  0.000
               Native American:   0   Median :58.00   Roman Catholic    :158   separated:   7   Median : 1.000                  Median :  0.000
               Other           :  26   Mean   :57.81   Methodist         : 41   single   :2322   Mean   : 2.769                  Mean   :  1.838
               White           :  38   3rd Qu.:67.00   Unknown           : 41   widow    :  25   3rd Qu.: 3.000                  3rd Qu.:  1.333
               NA's            : 767   Max.   :90.00   (Other)           : 84   NA's     :  29   Max.   :47.000                  Max.   :340.000
                                                       NA's              :703
      CONCEPT_CD      cluster
 ICD9:401.9 : 501   Min.   :2
 ICD9:272.4 : 195   1st Qu.:2
 ICD9:272.0 : 104   Median :2
 ICD9:305.1 :  79   Mean   :2
 ICD9:211.3 :  58   3rd Qu.:2
 ICD9:278.00:  52   Max.   :2
 (Other)    :1482
```

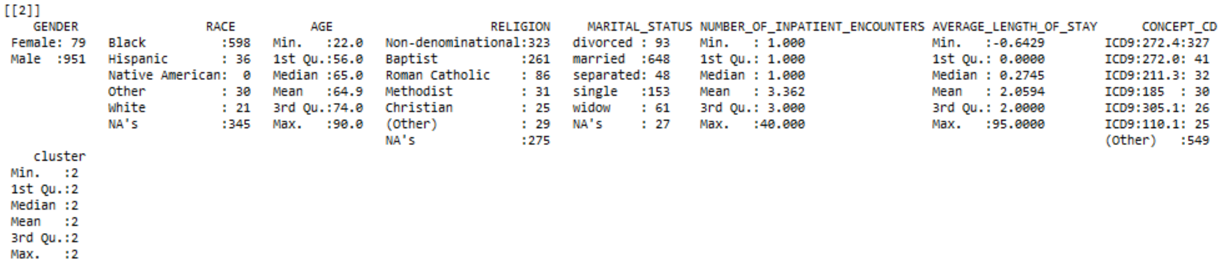


```
[[3]]
   GENDER                    RACE            AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:1233   Black           :808   Min.   :22.00   Baptist              :424   divorced :176    Min.   : 1.000                Min.   :-0.1667        ICD9:110.1:178
 Male  :  51   Hispanic        : 28   1st Qu.:56.00   Non-denominational:287   married  :579    1st Qu.: 1.000                1st Qu.: 0.0000        ICD9:272.0: 72
               Native American:  0   Median :64.00   Roman Catholic     : 96   separated: 79    Median : 2.000                Median : 0.1742        ICD9:211.3: 55
               Other           : 16   Mean   :64.52   Methodist          : 43   single   :118    Mean   : 4.642                Mean   : 1.6354        ICD9:272.4: 53
               White           : 17   3rd Qu.:74.00   Other              : 20   widow    :261    3rd Qu.: 5.250                3rd Qu.: 2.0000        ICD9:244.9: 38
               NA's            :415   Max.   :90.00   (Other)            : 55   NA's     : 71    Max.   :57.000                Max.   :33.0000        ICD9:305.1: 31
                                                      NA's               :359                                                                        (Other)   :857
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
 3rd Qu.:3
 Max.   :3
```

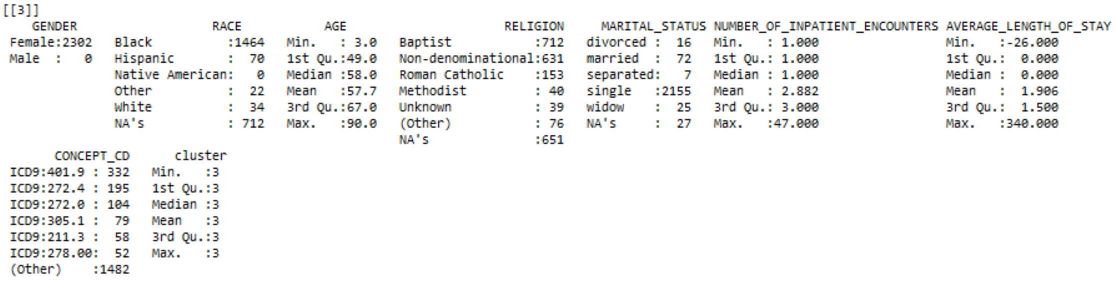
4 Cluster Summary:



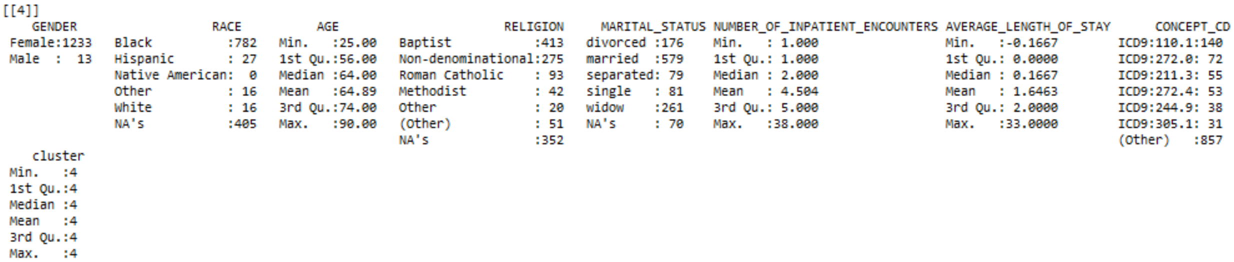
```
[[1]]
   GENDER                    RACE           AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1008   Min.   : 3.00   Non-denominational:632   divorced :    5   Min.   : 1.000                Min.   : -0.1667
 Male  :1672   Hispanic        :  75   1st Qu.:50.00   Baptist            :320   married  :   78   1st Qu.: 1.000                1st Qu.:  0.0000
               Native American:   2   Median :57.00   Roman Catholic     :110   separated:    2   Median : 1.000                Median :  0.0000
               Other           :  24   Mean   :57.14   Other              : 50   single   :1560   Mean   : 3.027                Mean   :  2.0611
               White           :  44   3rd Qu.:65.00   Unknown            : 23   widow    :    6   3rd Qu.: 3.000                3rd Qu.:  1.1354
               NA's            : 519   Max.   :90.00   (Other)            : 46   NA's     :   21   Max.   :64.000                Max.   :308.0000
                                                       NA's               :491
      CONCEPT_CD       cluster
 ICD9:401.9 : 261   Min.   :1
 ICD9:305.1 : 100   1st Qu.:1
 ICD9:110.1 :  64   Median :1
 ICD9:276.51:  63   Mean   :1
 ICD9:272.0 :  61   3rd Qu.:1
 ICD9:305.00:  43   Max.   :1
 (Other)    :1080
```



```
[[2]]
   GENDER                   RACE          AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female: 79   Black           :598   Min.   :22.0   Non-denominational:323   divorced : 93   Min.   : 1.000                Min.   :-0.6429        ICD9:272.4:327
 Male  :951   Hispanic        : 36   1st Qu.:56.0   Baptist            :261   married  :648   1st Qu.: 1.000                1st Qu.: 0.0000        ICD9:272.0: 41
              Native American:  0   Median :65.0   Roman Catholic     : 86   separated: 48   Median : 1.000                Median : 0.2745        ICD9:211.3: 32
              Other           : 30   Mean   :64.9   Methodist          : 31   single   :153   Mean   : 3.362                Mean   : 2.0594        ICD9:185  : 30
              White           : 21   3rd Qu.:74.0   Christian          : 25   widow    : 61   3rd Qu.: 3.000                3rd Qu.: 2.0000        ICD9:305.1: 26
              NA's            :345   Max.   :90.0   (Other)            : 29   NA's     : 27   Max.   :40.000                Max.   :95.0000        ICD9:110.1: 25
                                                    NA's               :275                                                                       (Other)   :549
   cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```

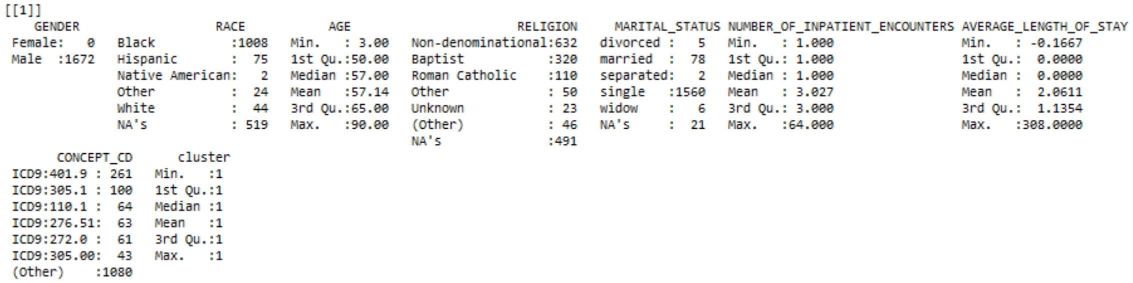


```
[[3]]
   GENDER                    RACE          AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:2302   Black           :1464   Min.   : 3.0   Baptist            :712   divorced :  16   Min.   : 1.000                Min.   :-26.000
 Male  :   0   Hispanic        :  70   1st Qu.:49.0   Non-denominational:631   married  :  72   1st Qu.: 1.000                1st Qu.:  0.000
               Native American:   0   Median :58.0   Roman Catholic     :153   separated:   7   Median : 1.000                Median :  0.000
               Other           :  22   Mean   :57.7   Methodist          : 40   single   :2155   Mean   : 2.882                Mean   :  1.906
               White           :  34   3rd Qu.:67.0   Unknown            : 39   widow    :  25   3rd Qu.: 3.000                3rd Qu.:  1.500
               NA's            : 712   Max.   :90.0   (Other)            : 76   NA's     :  27   Max.   :47.000                Max.   :340.000
                                                      NA's               :651
      CONCEPT_CD       cluster
 ICD9:401.9 : 332   Min.   :3
 ICD9:272.4 : 195   1st Qu.:3
 ICD9:272.0 : 104   Median :3
 ICD9:305.1 :  79   Mean   :3
 ICD9:211.3 :  58   3rd Qu.:3
 ICD9:278.00:  52   Max.   :3
 (Other)    :1482
```

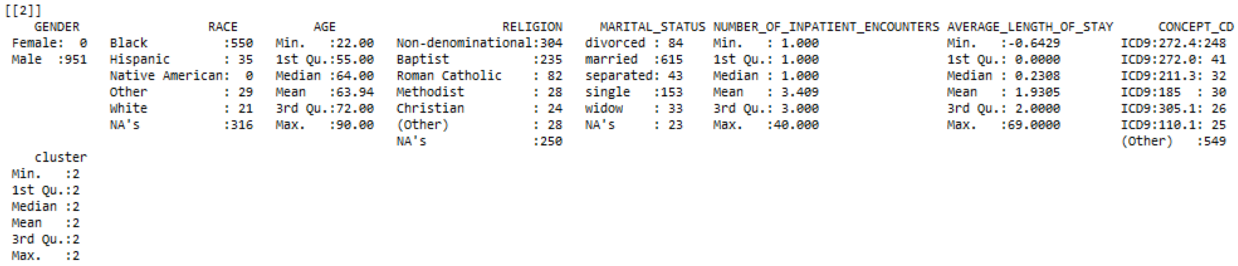


```
[[4]]
   GENDER                    RACE           AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:1233   Black           :782   Min.   :25.00   Baptist            :413   divorced :176    Min.   : 1.000                Min.   :-0.1667        ICD9:110.1:140
 Male  :  13   Hispanic        : 27   1st Qu.:56.00   Non-denominational:275   married  :579    1st Qu.: 1.000                1st Qu.: 0.0000        ICD9:272.0: 72
               Native American:  0   Median :64.00   Roman Catholic     : 93   separated: 79    Median : 2.000                Median : 0.1667        ICD9:211.3: 55
               Other           : 16   Mean   :64.89   Methodist          : 42   single   : 81    Mean   : 4.504                Mean   : 1.6463        ICD9:272.4: 53
               White           : 16   3rd Qu.:74.00   Other              : 20   widow    :261    3rd Qu.: 5.000                3rd Qu.: 2.0000        ICD9:244.9: 38
               NA's            :405   Max.   :90.00   (Other)            : 51   NA's     : 70    Max.   :38.000                Max.   :33.0000        ICD9:305.1: 31
                                                      NA's               :352                                                                        (Other)   :857
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```

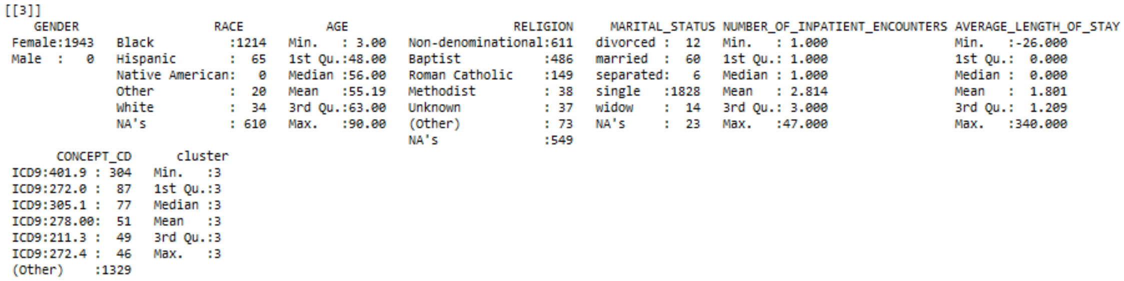
5 Cluster Summary:



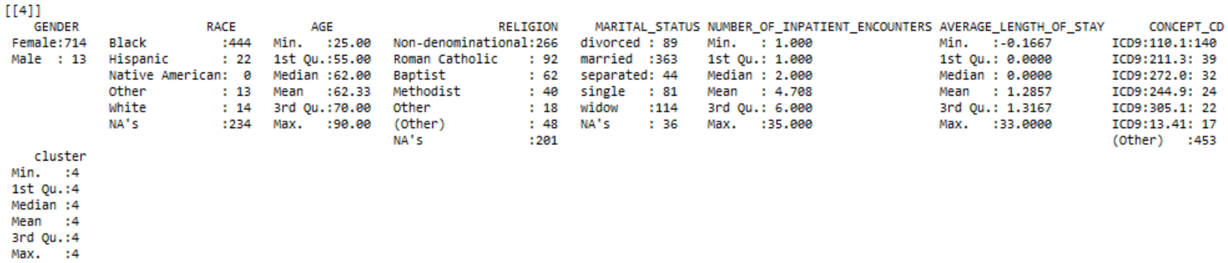
```
[[1]]
   GENDER                RACE          AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1008   Min.   : 3.00   Non-denominational:632   divorced :   5   Min.   : 1.000                  Min.   : -0.1667
 Male  :1672   Hispanic        :  75   1st Qu.:50.00   Baptist           :320   married  :  78   1st Qu.: 1.000                  1st Qu.:  0.0000
               Native American:   2   Median :57.00   Roman Catholic    :110   separated:   2   Median : 1.000                  Median :  0.0000
               Other           :  24   Mean   :57.14   Other             : 50   single   :1560   Mean   : 3.027                  Mean   :  2.0611
               White           :  44   3rd Qu.:65.00   Unknown           : 23   widow    :   6   3rd Qu.: 3.000                  3rd Qu.:  1.1354
               NA's            : 519   Max.   :90.00   (Other)           : 46   NA's     :  21   Max.   :64.000                  Max.   :308.0000
                                                       NA's              :491
      CONCEPT_CD      cluster
 ICD9:401.9 : 261   Min.   :1
 ICD9:305.1 : 100   1st Qu.:1
 ICD9:110.1 :  64   Median :1
 ICD9:276.51:  63   Mean   :1
 ICD9:272.0 :  61   3rd Qu.:1
 ICD9:305.00:  43   Max.   :1
 (Other)    :1080
```



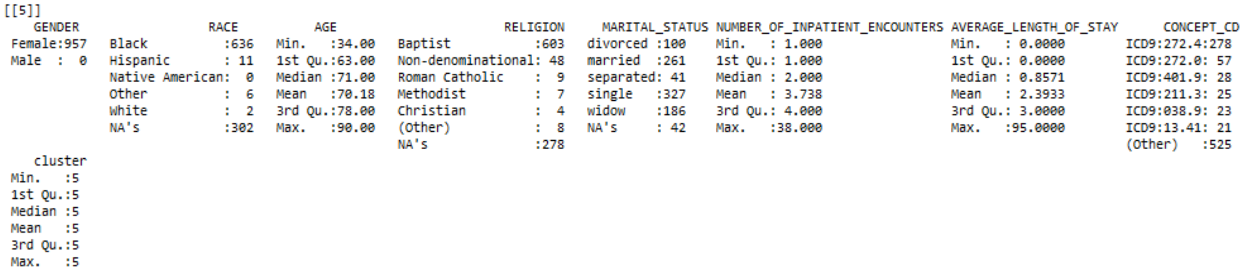
```
[[2]]
   GENDER               RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:  0   Black           :550   Min.   :22.00   Non-denominational:304   divorced : 84   Min.   : 1.000                  Min.   :-0.6429       ICD9:272.4:248
 Male  :951   Hispanic        : 35   1st Qu.:55.00   Baptist           :235   married  :615   1st Qu.: 1.000                  1st Qu.: 0.0000       ICD9:272.0: 41
              Native American:  0   Median :64.00   Roman Catholic    : 82   separated: 43   Median : 1.000                  Median : 0.2308       ICD9:211.3: 32
              Other           : 29   Mean   :63.94   Methodist         : 28   single   :153   Mean   : 3.409                  Mean   : 1.9305       ICD9:185  : 30
              White           : 21   3rd Qu.:72.00   Christian         : 24   widow    : 33   3rd Qu.: 3.000                  3rd Qu.: 2.0000       ICD9:305.1: 26
              NA's            :316   Max.   :90.00   (Other)           : 28   NA's     : 23   Max.   :40.000                  Max.   :69.0000       ICD9:110.1: 25
                                                     NA's              :250                                                                        (Other)   :549
   cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```



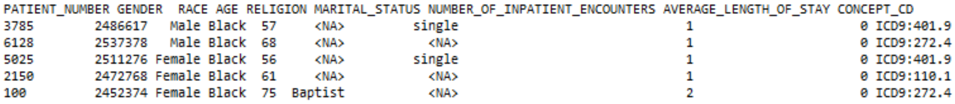
```
[[3]]
   GENDER                RACE          AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:1943   Black           :1214   Min.   : 3.00   Non-denominational:611   divorced :  12   Min.   : 1.000                  Min.   :-26.000
 Male  :   0   Hispanic        :  65   1st Qu.:48.00   Baptist           :486   married  :  60   1st Qu.: 1.000                  1st Qu.:  0.000
               Native American:   0   Median :56.00   Roman Catholic    :149   separated:   6   Median : 1.000                  Median :  0.000
               Other           :  20   Mean   :55.19   Methodist         : 38   single   :1828   Mean   : 2.814                  Mean   :  1.801
               White           :  34   3rd Qu.:63.00   Unknown           : 37   widow    :  14   3rd Qu.: 3.000                  3rd Qu.:  1.209
               NA's            : 610   Max.   :90.00   (Other)           : 73   NA's     :  23   Max.   :47.000                  Max.   :340.000
                                                       NA's              :549
      CONCEPT_CD      cluster
 ICD9:401.9 : 304   Min.   :3
 ICD9:272.0 :  87   1st Qu.:3
 ICD9:305.1 :  77   Median :3
 ICD9:278.00:  51   Mean   :3
 ICD9:211.3 :  49   3rd Qu.:3
 ICD9:272.4 :  46   Max.   :3
 (Other)    :1329
```



```
[[4]]
   GENDER               RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:714   Black           :444   Min.   :25.00   Non-denominational:266   divorced : 89   Min.   : 1.000                  Min.   :-0.1667       ICD9:110.1:140
 Male  : 13   Hispanic        : 22   1st Qu.:55.00   Roman Catholic    : 92   married  :363   1st Qu.: 1.000                  1st Qu.: 0.0000       ICD9:211.3: 39
              Native American:  0   Median :62.00   Baptist           : 62   separated: 44   Median : 2.000                  Median : 0.0000       ICD9:272.0: 32
              Other           : 13   Mean   :62.33   Methodist         : 40   single   : 81   Mean   : 4.708                  Mean   : 1.2857       ICD9:244.9: 24
              White           : 14   3rd Qu.:70.00   Other             : 18   widow    :114   3rd Qu.: 6.000                  3rd Qu.: 1.3167       ICD9:305.1: 22
              NA's            :234   Max.   :90.00   (Other)           : 48   NA's     : 36   Max.   :35.000                  Max.   :33.0000       ICD9:13.41: 17
                                                     NA's              :201                                                                        (Other)   :453
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```



```
[[5]]
   GENDER               RACE          AGE                       RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:957   Black           :636   Min.   :34.00   Baptist           :603   divorced :100   Min.   : 1.000                  Min.   : 0.0000       ICD9:272.4:278
 Male  :  0   Hispanic        : 11   1st Qu.:63.00   Non-denominational: 48   married  :261   1st Qu.: 1.000                  1st Qu.: 0.0000       ICD9:272.0: 57
              Native American:  0   Median :71.00   Roman Catholic    :  9   separated: 41   Median : 2.000                  Median : 0.8571       ICD9:401.9: 28
              Other           :  6   Mean   :70.18   Methodist         :  7   single   :327   Mean   : 3.738                  Mean   : 2.3933       ICD9:211.3: 25
              White           :  2   3rd Qu.:78.00   Christian         :  4   widow    :186   3rd Qu.: 4.000                  3rd Qu.: 3.0000       ICD9:038.9: 23
              NA's            :302   Max.   :90.00   (Other)           :  8   NA's     : 42   Max.   :38.000                  Max.   :95.0000       ICD9:13.41: 21
                                                     NA's              :278                                                                        (Other)   :525
   cluster
 Min.   :5
 1st Qu.:5
 Median :5
 Mean   :5
 3rd Qu.:5
 Max.   :5
```



```
PATIENT_NUMBER GENDER  RACE AGE RELIGION MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY CONCEPT_CD
3785        2486617   Male Black  57     <NA>         single                              1                      0 ICD9:401.9
6128        2537378   Male Black  68     <NA>           <NA>                              1                      0 ICD9:272.4
5025        2511276 Female Black  56     <NA>         single                              1                      0 ICD9:401.9
2150        2472768 Female Black  61     <NA>           <NA>                              1                      0 ICD9:110.1
100         2452374 Female Black  75  Baptist           <NA>                              2                      0 ICD9:272.4
```

5.3 HCLUST Summary

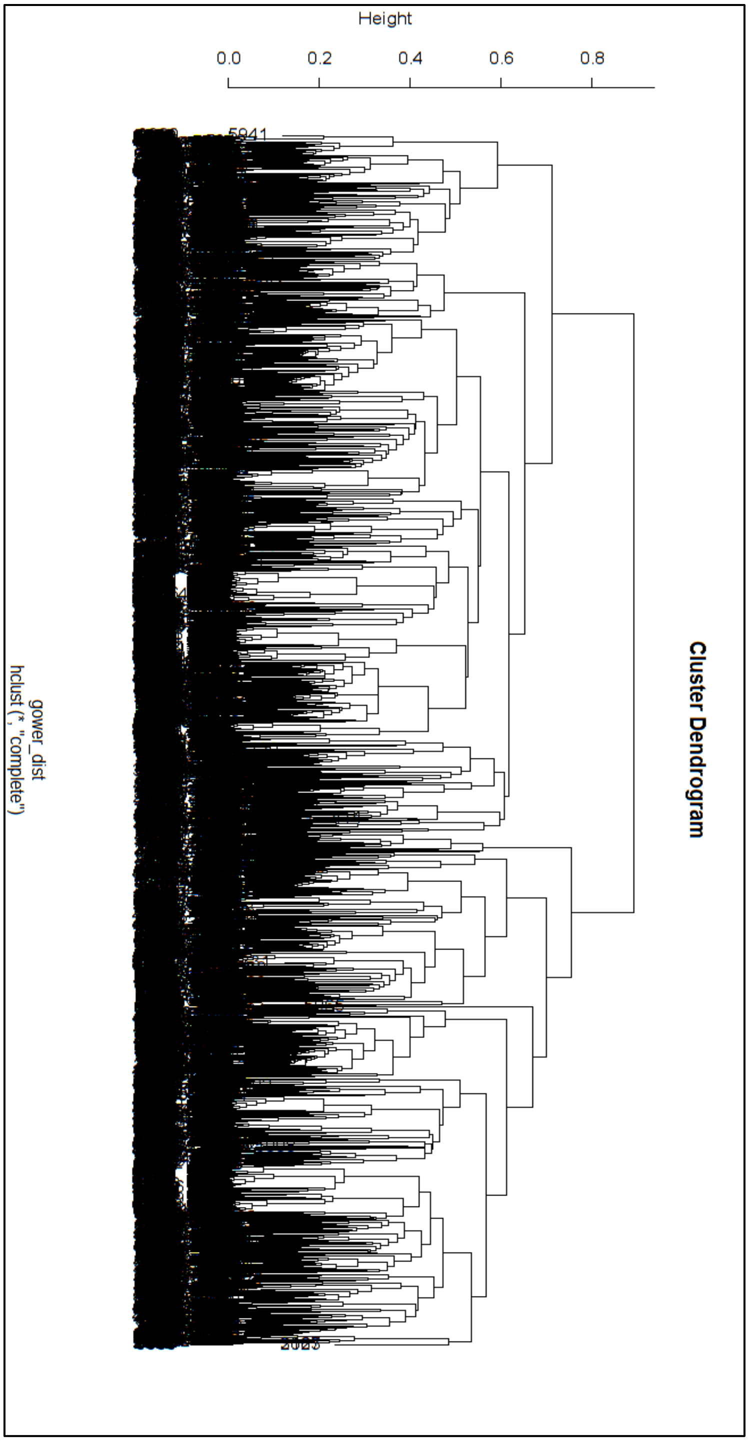
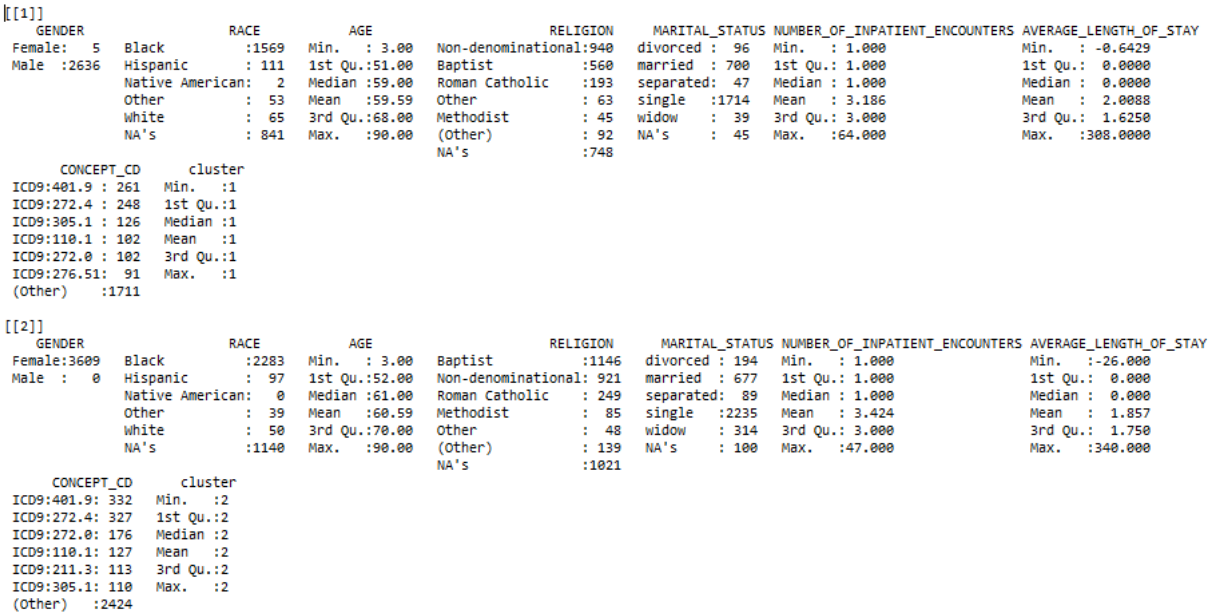


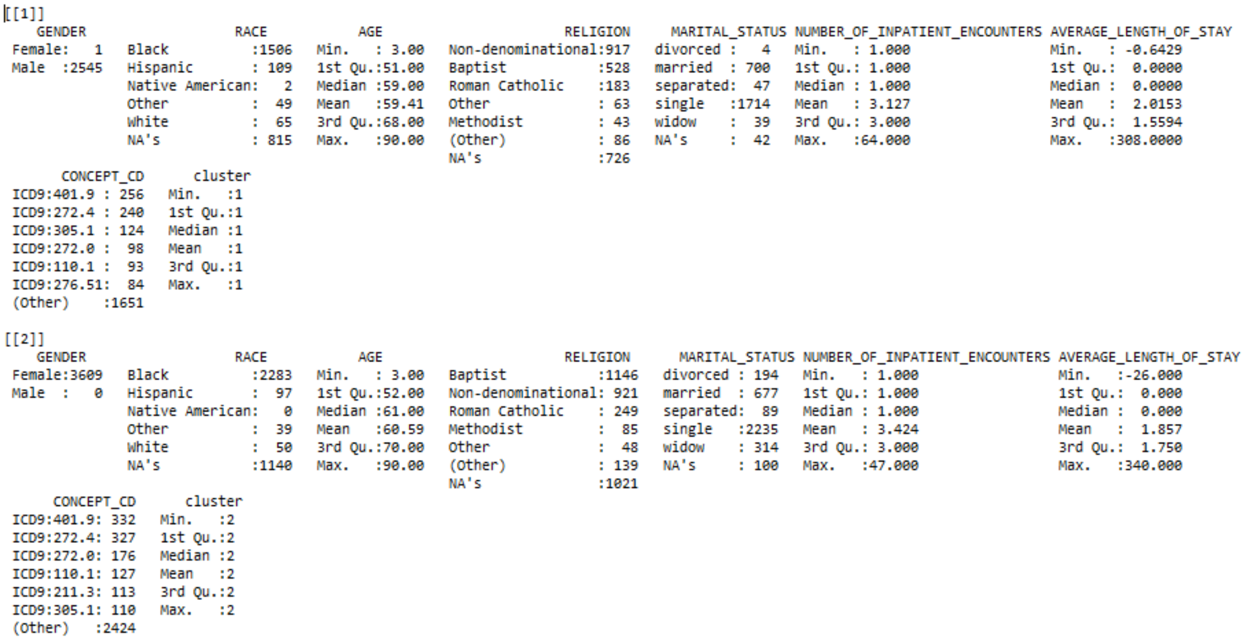
Figure 5.4 Visualization of clustered data through a Hclust dendrogram

2 Cluster Summary:

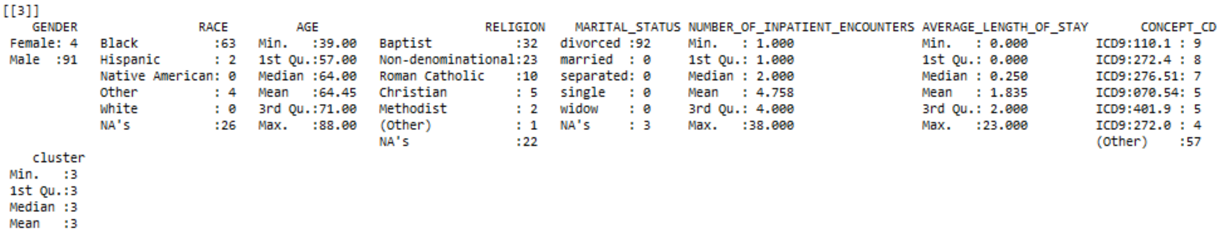


```
[[1]]
    GENDER                    RACE            AGE                         RELIGION      MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   5   Black           :1569   Min.   : 3.00   Non-denominational:940   divorced :  96   Min.   : 1.000                  Min.   : -0.6429
 Male  :2636   Hispanic        : 111   1st Qu.:51.00   Baptist           :560   married  : 700   1st Qu.: 1.000                  1st Qu.:  0.0000
               Native American:   2   Median :59.00   Roman Catholic    :193   separated:  47   Median : 1.000                  Median :  0.0000
               Other           :  53   Mean   :59.59   Other             : 63   single   :1714   Mean   : 3.186                  Mean   :  2.0088
               White           :  65   3rd Qu.:68.00   Methodist         : 45   widow    :  39   3rd Qu.: 3.000                  3rd Qu.:  1.6250
               NA's            : 841   Max.   :90.00   (Other)           : 92   NA's     :  45   Max.   :64.000                  Max.   :308.0000
                                                       NA's              :748
      CONCEPT_CD       cluster
 ICD9:401.9 : 261   Min.   :1
 ICD9:272.4 : 248   1st Qu.:1
 ICD9:305.1 : 126   Median :1
 ICD9:110.1 : 102   Mean   :1
 ICD9:272.0 : 102   3rd Qu.:1
 ICD9:276.51:  91   Max.   :1
 (Other)    :1711

[[2]]
    GENDER                    RACE            AGE                         RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:3609   Black           :2283   Min.   : 3.00   Baptist           :1146   divorced : 194   Min.   : 1.000                  Min.   :-26.000
 Male  :   0   Hispanic        :  97   1st Qu.:52.00   Non-denominational: 921   married  : 677   1st Qu.: 1.000                  1st Qu.:  0.000
               Native American:   0   Median :61.00   Roman Catholic    : 249   separated:  89   Median : 1.000                  Median :  0.000
               Other           :  39   Mean   :60.59   Methodist         :  85   single   :2235   Mean   : 3.424                  Mean   :  1.857
               White           :  50   3rd Qu.:70.00   Other             :  48   widow    : 314   3rd Qu.: 3.000                  3rd Qu.:  1.750
               NA's            :1140   Max.   :90.00   (Other)           : 139   NA's     : 100   Max.   :47.000                  Max.   :340.000
                                                       NA's              :1021
     CONCEPT_CD      cluster
 ICD9:401.9: 332   Min.   :2
 ICD9:272.4: 327   1st Qu.:2
 ICD9:272.0: 176   Median :2
 ICD9:110.1: 127   Mean   :2
 ICD9:211.3: 113   3rd Qu.:2
 ICD9:305.1: 110   Max.   :2
 (Other)   :2424
```

3 Cluster Summary:

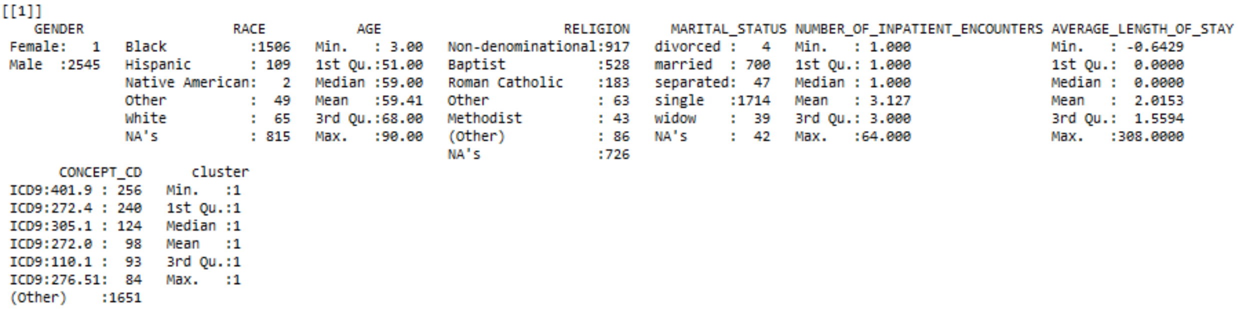


```
[[1]]
    GENDER                    RACE            AGE                         RELIGION      MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   1   Black           :1506   Min.   : 3.00   Non-denominational:917   divorced :   4   Min.   : 1.000                  Min.   : -0.6429
 Male  :2545   Hispanic        : 109   1st Qu.:51.00   Baptist           :528   married  : 700   1st Qu.: 1.000                  1st Qu.:  0.0000
               Native American:   2   Median :59.00   Roman Catholic    :183   separated:  47   Median : 1.000                  Median :  0.0000
               Other           :  49   Mean   :59.41   Other             : 63   single   :1714   Mean   : 3.127                  Mean   :  2.0153
               White           :  65   3rd Qu.:68.00   Methodist         : 43   widow    :  39   3rd Qu.: 3.000                  3rd Qu.:  1.5594
               NA's            : 815   Max.   :90.00   (Other)           : 86   NA's     :  42   Max.   :64.000                  Max.   :308.0000
                                                       NA's              :726
      CONCEPT_CD       cluster
 ICD9:401.9 : 256   Min.   :1
 ICD9:272.4 : 240   1st Qu.:1
 ICD9:305.1 : 124   Median :1
 ICD9:272.0 :  98   Mean   :1
 ICD9:110.1 :  93   3rd Qu.:1
 ICD9:276.51:  84   Max.   :1
 (Other)    :1651

[[2]]
    GENDER                    RACE            AGE                         RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:3609   Black           :2283   Min.   : 3.00   Baptist           :1146   divorced : 194   Min.   : 1.000                  Min.   :-26.000
 Male  :   0   Hispanic        :  97   1st Qu.:52.00   Non-denominational: 921   married  : 677   1st Qu.: 1.000                  1st Qu.:  0.000
               Native American:   0   Median :61.00   Roman Catholic    : 249   separated:  89   Median : 1.000                  Median :  0.000
               Other           :  39   Mean   :60.59   Methodist         :  85   single   :2235   Mean   : 3.424                  Mean   :  1.857
               White           :  50   3rd Qu.:70.00   Other             :  48   widow    : 314   3rd Qu.: 3.000                  3rd Qu.:  1.750
               NA's            :1140   Max.   :90.00   (Other)           : 139   NA's     : 100   Max.   :47.000                  Max.   :340.000
                                                       NA's              :1021
     CONCEPT_CD      cluster
 ICD9:401.9: 332   Min.   :2
 ICD9:272.4: 327   1st Qu.:2
 ICD9:272.0: 176   Median :2
 ICD9:110.1: 127   Mean   :2
 ICD9:211.3: 113   3rd Qu.:2
 ICD9:305.1: 110   Max.   :2
 (Other)   :2424
```

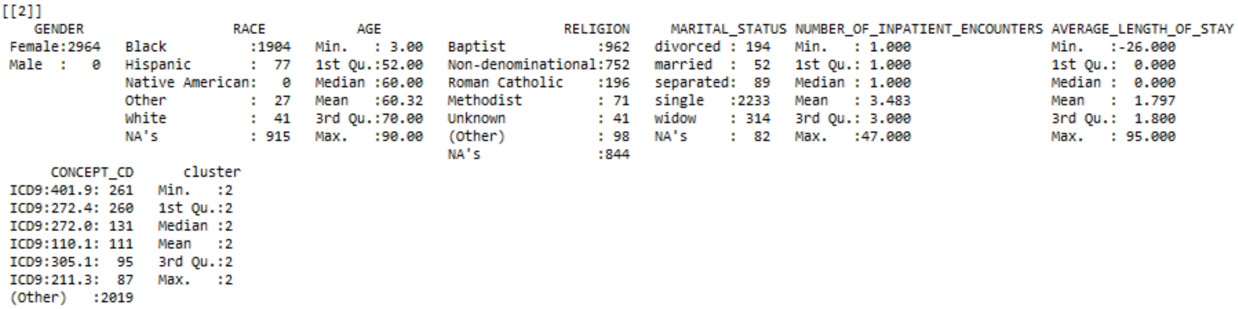


```
[[3]]
   GENDER                RACE          AGE                          RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female: 4   Black          :63   Min.   :39.00   Baptist            :32   divorced :92    Min.   : 1.000                 Min.   : 0.000         ICD9:110.1 : 9
 Male  :91   Hispanic       : 2   1st Qu.:57.00   Non-denominational:23   married  : 0    1st Qu.: 1.000                 1st Qu.: 0.000         ICD9:272.4 : 8
             Native American: 0   Median :64.00   Roman Catholic    :10   separated: 0    Median : 2.000                 Median : 0.250         ICD9:276.51: 7
             Other          : 4   Mean   :64.45   Christian         : 5   single   : 0    Mean   : 4.758                 Mean   : 1.835         ICD9:070.54: 5
             White          : 0   3rd Qu.:71.00   Methodist         : 2   widow    : 0    3rd Qu.: 4.000                 3rd Qu.: 2.000         ICD9:401.9 : 5
             NA's           :26   Max.   :88.00   (Other)           : 1   NA's     : 3    Max.   :38.000                 Max.   :23.000         ICD9:272.0 : 4
                                                  NA's              :22                                                                        (Other)    :57
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
```

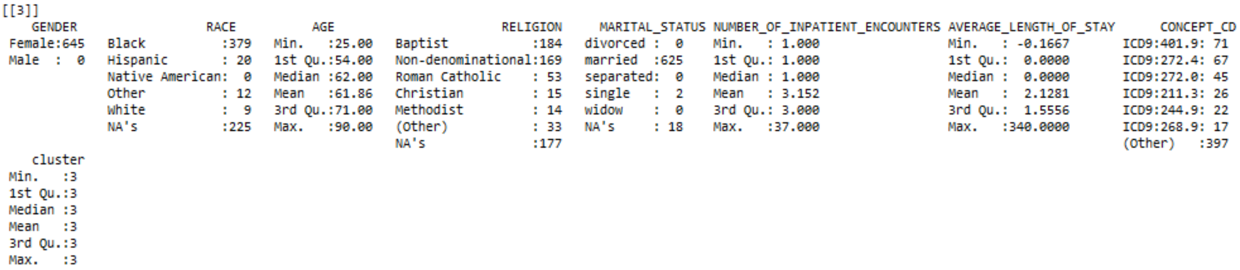
4 Cluster Summary:



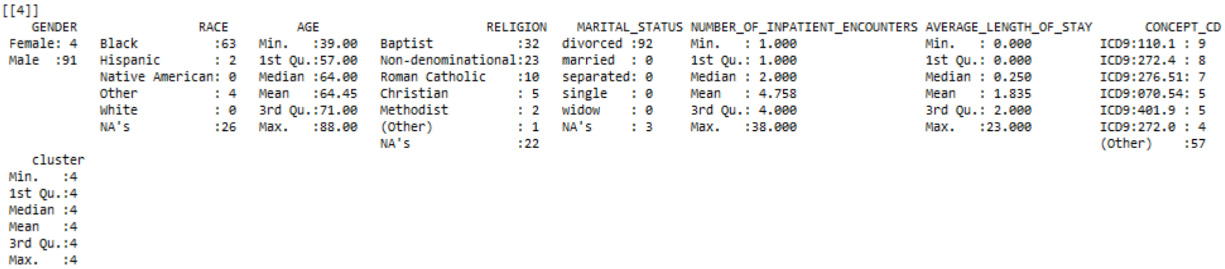
```
[[1]]
   GENDER                  RACE           AGE                         RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:    1   Black           :1506   Min.   : 3.00   Non-denominational:917   divorced :   4   Min.   : 1.000                 Min.   : -0.6429
 Male  :2545   Hispanic        : 109   1st Qu.:51.00   Baptist           :528   married  : 700   1st Qu.: 1.000                 1st Qu.:  0.0000
               Native American:   2   Median :59.00   Roman Catholic    :183   separated:  47   Median : 1.000                 Median :  0.0000
               Other          :  49   Mean   :59.41   Other             : 63   single   :1714   Mean   : 3.127                 Mean   :  2.0153
               White          :  65   3rd Qu.:68.00   Methodist         : 43   widow    :  39   3rd Qu.: 3.000                 3rd Qu.:  1.5594
               NA's           : 815   Max.   :90.00   (Other)           : 86   NA's     :  42   Max.   :64.000                 Max.   :308.0000
                                                      NA's              :726
      CONCEPT_CD       cluster
 ICD9:401.9 : 256   Min.   :1
 ICD9:272.4 : 240   1st Qu.:1
 ICD9:305.1 : 124   Median :1
 ICD9:272.0 :  98   Mean   :1
 ICD9:110.1 :  93   3rd Qu.:1
 ICD9:276.51:  84   Max.   :1
 (Other)    :1651
```



```
[[2]]
   GENDER                  RACE           AGE                         RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:2964   Black          :1904   Min.   : 3.00   Baptist           :962   divorced : 194   Min.   : 1.000                 Min.   :-26.000
 Male  :   0   Hispanic       :  77   1st Qu.:52.00   Non-denominational:752   married  :  52   1st Qu.: 1.000                 1st Qu.:  0.000
               Native American:   0   Median :60.00   Roman Catholic    :196   separated:  89   Median : 1.000                 Median :  0.000
               Other          :  27   Mean   :60.32   Methodist         : 71   single   :2233   Mean   : 3.483                 Mean   :  1.797
               White          :  41   3rd Qu.:70.00   Unknown           : 41   widow    : 314   3rd Qu.: 3.000                 3rd Qu.:  1.800
               NA's           : 915   Max.   :90.00   (Other)           : 98   NA's     :  82   Max.   :47.000                 Max.   : 95.000
                                                      NA's              :844
     CONCEPT_CD       cluster
 ICD9:401.9: 261   Min.   :2
 ICD9:272.4: 260   1st Qu.:2
 ICD9:272.0: 131   Median :2
 ICD9:110.1: 111   Mean   :2
 ICD9:305.1:  95   3rd Qu.:2
 ICD9:211.3:  87   Max.   :2
 (Other)   :2019
```

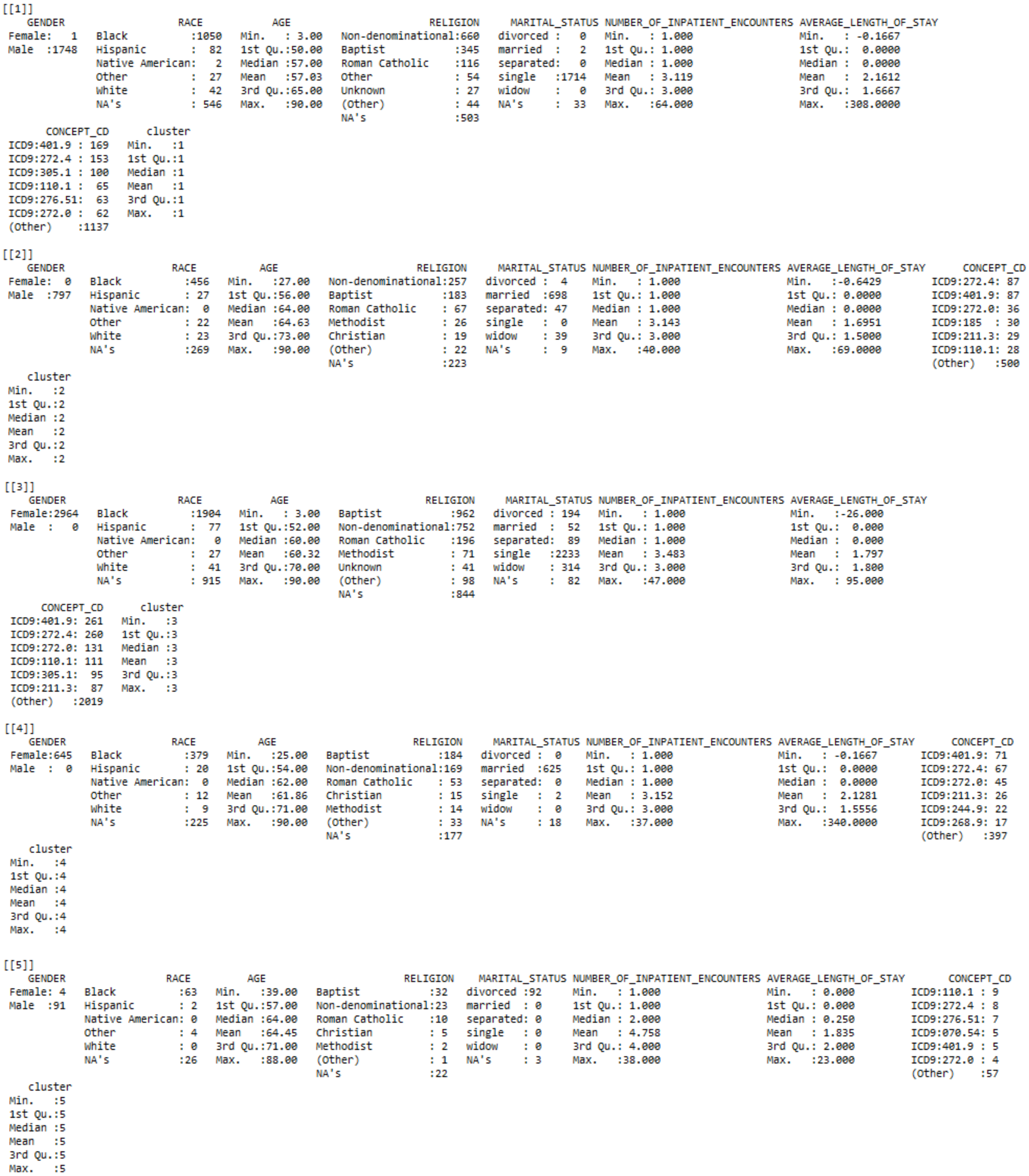


```
[[3]]
   GENDER               RACE          AGE                          RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female:645   Black          :379   Min.   :25.00   Baptist           :184   divorced :  0   Min.   : 1.000                 Min.   : -0.1667        ICD9:401.9: 71
 Male  :  0   Hispanic       : 20   1st Qu.:54.00   Non-denominational:169   married  :625   1st Qu.: 1.000                 1st Qu.:  0.0000        ICD9:272.4: 67
              Native American:  0   Median :62.00   Roman Catholic    : 53   separated:  0   Median : 1.000                 Median :  0.0000        ICD9:272.0: 45
              Other          : 12   Mean   :61.86   Christian         : 15   single   :  2   Mean   : 3.152                 Mean   :  2.1281        ICD9:211.3: 26
              White          :  9   3rd Qu.:71.00   Methodist         : 14   widow    :  0   3rd Qu.: 3.000                 3rd Qu.:  1.5556        ICD9:244.9: 22
              NA's           :225   Max.   :90.00   (Other)           : 33   NA's     : 18   Max.   :37.000                 Max.   :340.0000        ICD9:268.9: 17
                                                    NA's              :177                                                                          (Other)   :397
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
 3rd Qu.:3
 Max.   :3
```



```
[[4]]
   GENDER                RACE          AGE                          RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female: 4   Black          :63   Min.   :39.00   Baptist            :32   divorced :92    Min.   : 1.000                 Min.   : 0.000         ICD9:110.1 : 9
 Male  :91   Hispanic       : 2   1st Qu.:57.00   Non-denominational:23   married  : 0    1st Qu.: 1.000                 1st Qu.: 0.000         ICD9:272.4 : 8
             Native American: 0   Median :64.00   Roman Catholic    :10   separated: 0    Median : 2.000                 Median : 0.250         ICD9:276.51: 7
             Other          : 4   Mean   :64.45   Christian         : 5   single   : 0    Mean   : 4.758                 Mean   : 1.835         ICD9:070.54: 5
             White          : 0   3rd Qu.:71.00   Methodist         : 2   widow    : 0    3rd Qu.: 4.000                 3rd Qu.: 2.000         ICD9:401.9 : 5
             NA's           :26   Max.   :88.00   (Other)           : 1   NA's     : 3    Max.   :38.000                 Max.   :23.000         ICD9:272.0 : 4
                                                  NA's              :22                                                                        (Other)    :57
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```

6 Cluster Summary



```
[[1]]
   GENDER               RACE            AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   1   Black          :1050   Min.   : 3.00   Non-denominational:660   divorced :   0   Min.   : 1.000                  Min.   : -0.1667     
 Male  :1748   Hispanic       :  82   1st Qu.:50.00   Baptist           :345   married  :   2   1st Qu.: 1.000                  1st Qu.:  0.0000     
               Native American:   2   Median :57.00   Roman Catholic    :116   separated:   0   Median : 1.000                  Median :  0.0000     
               Other          :  27   Mean   :57.03   Other             : 54   single   :1714   Mean   : 3.119                  Mean   :  2.1612     
               White          :  42   3rd Qu.:65.00   Unknown           : 27   widow    :   0   3rd Qu.: 3.000                  3rd Qu.:  1.6667     
               NA's           : 546   Max.   :90.00   (Other)           : 44   NA's     :  33   Max.   :64.000                  Max.   :308.0000     
                                                      NA's              :503                                                                         
      CONCEPT_CD       cluster  
 ICD9:401.9 : 169   Min.   :1  
 ICD9:272.4 : 153   1st Qu.:1  
 ICD9:305.1 : 100   Median :1  
 ICD9:110.1 :  65   Mean   :1  
 ICD9:276.51:  63   3rd Qu.:1  
 ICD9:272.0 :  62   Max.   :1  
 (Other)    :1137              

[[2]]
   GENDER              RACE           AGE                         RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD 
 Female:  0   Black          :456   Min.   :27.00   Non-denominational:257   divorced :  4   Min.   : 1.000                 Min.   :-0.6429        ICD9:272.4: 87  
 Male  :797   Hispanic       : 27   1st Qu.:56.00   Baptist           :183   married  :698   1st Qu.: 1.000                 1st Qu.: 0.0000        ICD9:401.9: 87  
              Native American:  0   Median :64.00   Roman Catholic    : 67   separated: 47   Median : 1.000                 Median : 0.0000        ICD9:272.0: 36  
              Other          : 22   Mean   :64.63   Methodist         : 26   single   :  0   Mean   : 3.143                 Mean   : 1.6951        ICD9:185  : 30  
              White          : 23   3rd Qu.:73.00   Christian         : 19   widow    : 39   3rd Qu.: 3.000                 3rd Qu.: 1.5000        ICD9:211.3: 29  
              NA's           :269   Max.   :90.00   (Other)           : 22   NA's     :  9   Max.   :40.000                 Max.   :69.0000        ICD9:110.1: 28  
                                                    NA's              :223                                                                         (Other)   :500  
    cluster 
 Min.   :2  
 1st Qu.:2  
 Median :2  
 Mean   :2  
 3rd Qu.:2  
 Max.   :2  

[[3]]
   GENDER               RACE            AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:2964   Black          :1904   Min.   : 3.00   Baptist           :962   divorced : 194   Min.   : 1.000                 Min.   :-26.000      
 Male  :   0   Hispanic       :  77   1st Qu.:52.00   Non-denominational:752   married  :  52   1st Qu.: 1.000                 1st Qu.:  0.000      
               Native American:   0   Median :60.00   Roman Catholic    :196   separated:  89   Median : 1.000                 Median :  0.000      
               Other          :  27   Mean   :60.32   Methodist         : 71   single   :2233   Mean   : 3.483                 Mean   :  1.797      
               White          :  41   3rd Qu.:70.00   Unknown           : 41   widow    : 314   3rd Qu.: 3.000                 3rd Qu.:  1.800      
               NA's           : 915   Max.   :90.00   (Other)           : 98   NA's     :  82   Max.   :47.000                 Max.   : 95.000      
                                                      NA's              :844                                                                        
     CONCEPT_CD       cluster  
 ICD9:401.9: 261   Min.   :3  
 ICD9:272.4: 260   1st Qu.:3  
 ICD9:272.0: 131   Median :3  
 ICD9:110.1: 111   Mean   :3  
 ICD9:305.1:  95   3rd Qu.:3  
 ICD9:211.3:  87   Max.   :3  
 (Other)   :2019              

[[4]]
   GENDER              RACE           AGE                         RELIGION    MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD 
 Female:645   Black          :379   Min.   :25.00   Baptist           :184   divorced :  0   Min.   : 1.000                 Min.   : -0.1667       ICD9:401.9: 71  
 Male  :  0   Hispanic       : 20   1st Qu.:54.00   Non-denominational:169   married  :625   1st Qu.: 1.000                 1st Qu.:  0.0000       ICD9:272.4: 67  
              Native American:  0   Median :62.00   Roman Catholic    : 53   separated:  0   Median : 1.000                 Median :  0.0000       ICD9:272.0: 45  
              Other          : 12   Mean   :61.86   Christian         : 15   single   :  2   Mean   : 3.152                 Mean   :  2.1281       ICD9:211.3: 26  
              White          :  9   3rd Qu.:71.00   Methodist         : 14   widow    :  0   3rd Qu.: 3.000                 3rd Qu.:  1.5556       ICD9:244.9: 22  
              NA's           :225   Max.   :90.00   (Other)           : 33   NA's     : 18   Max.   :37.000                 Max.   :340.0000       ICD9:268.9: 17  
                                                    NA's              :177                                                                         (Other)   :397  
    cluster 
 Min.   :4  
 1st Qu.:4  
 Median :4  
 Mean   :4  
 3rd Qu.:4  
 Max.   :4  


[[5]]
   GENDER             RACE          AGE                        RELIGION   MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD 
 Female: 4   Black          :63   Min.   :39.00   Baptist           :32   divorced :92     Min.   : 1.000                 Min.   : 0.000          ICD9:110.1 : 9  
 Male  :91   Hispanic       : 2   1st Qu.:57.00   Non-denominational:23   married  : 0     1st Qu.: 1.000                 1st Qu.: 0.000          ICD9:272.4 : 8  
             Native American: 0   Median :64.00   Roman Catholic    :10   separated: 0     Median : 2.000                 Median : 0.250          ICD9:276.51: 7  
             Other          : 4   Mean   :64.45   Christian         : 5   single   : 0     Mean   : 4.758                 Mean   : 1.835          ICD9:070.54: 5  
             White          : 0   3rd Qu.:71.00   Methodist         : 2   widow    : 0     3rd Qu.: 4.000                 3rd Qu.: 2.000          ICD9:401.9 : 5  
             NA's           :26   Max.   :88.00   (Other)           : 1   NA's     : 3     Max.   :38.000                 Max.   :23.000          ICD9:272.0 : 4  
                                                  NA's              :22                                                                           (Other)    :57  
    cluster 
 Min.   :5  
 1st Qu.:5  
 Median :5  
 Mean   :5  
 3rd Qu.:5  
 Max.   :5  
```

5.4 DIANA Summary

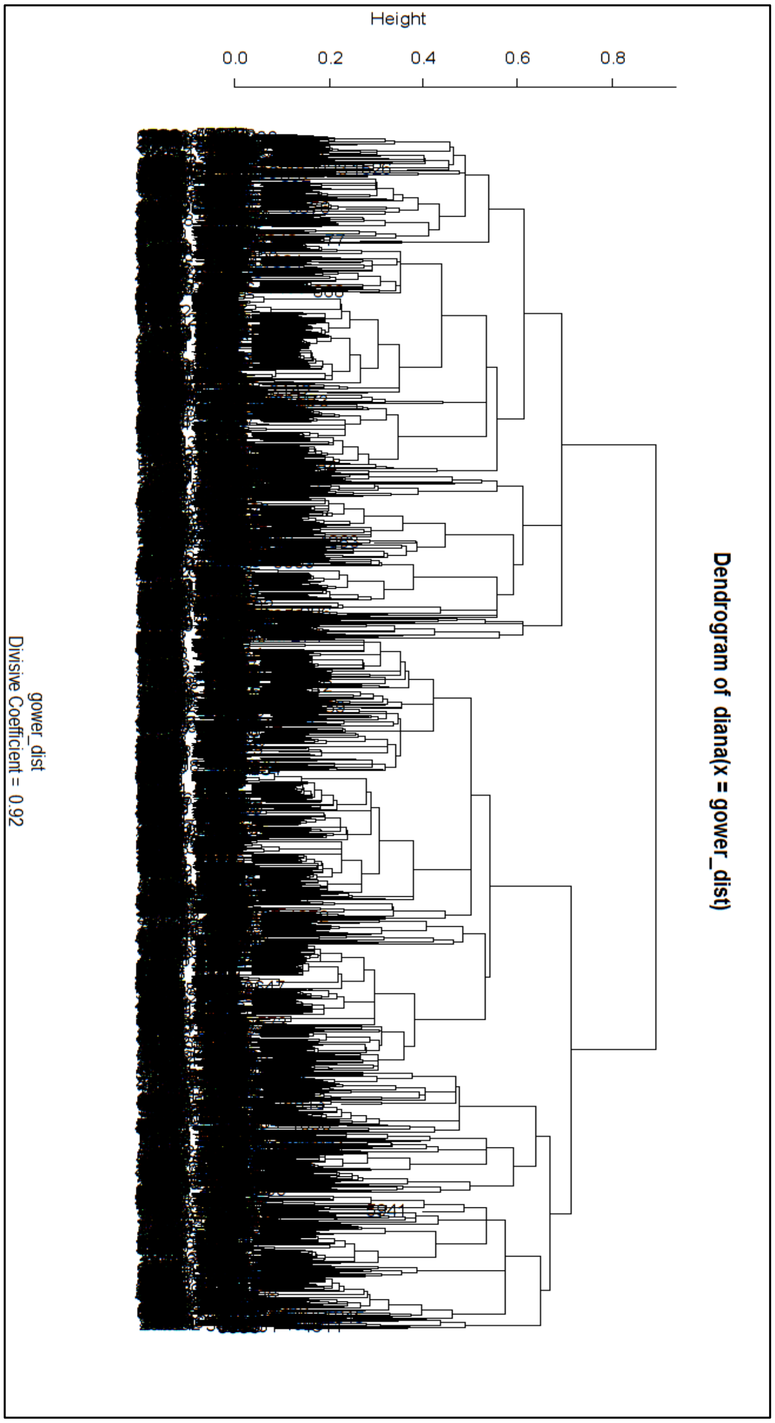


Figure 5.5 Visualization of clustered data through DIANA dendrogram

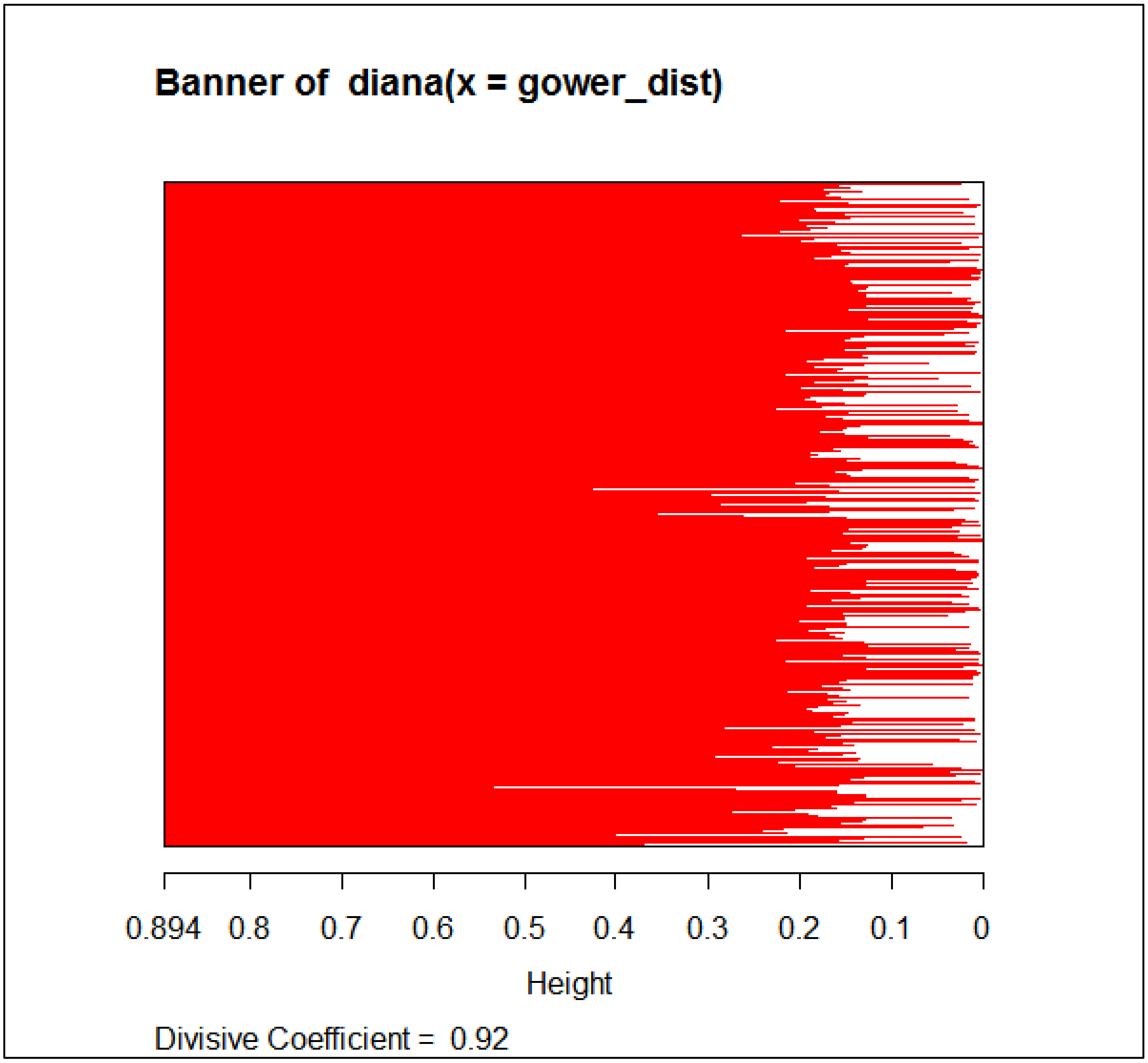
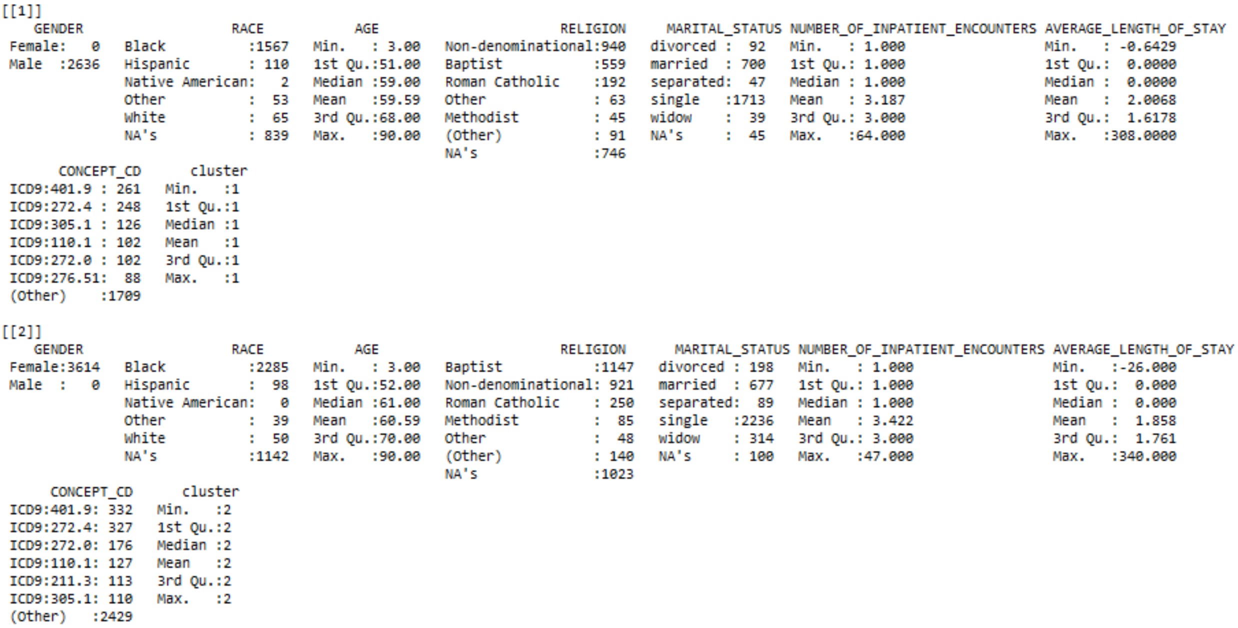


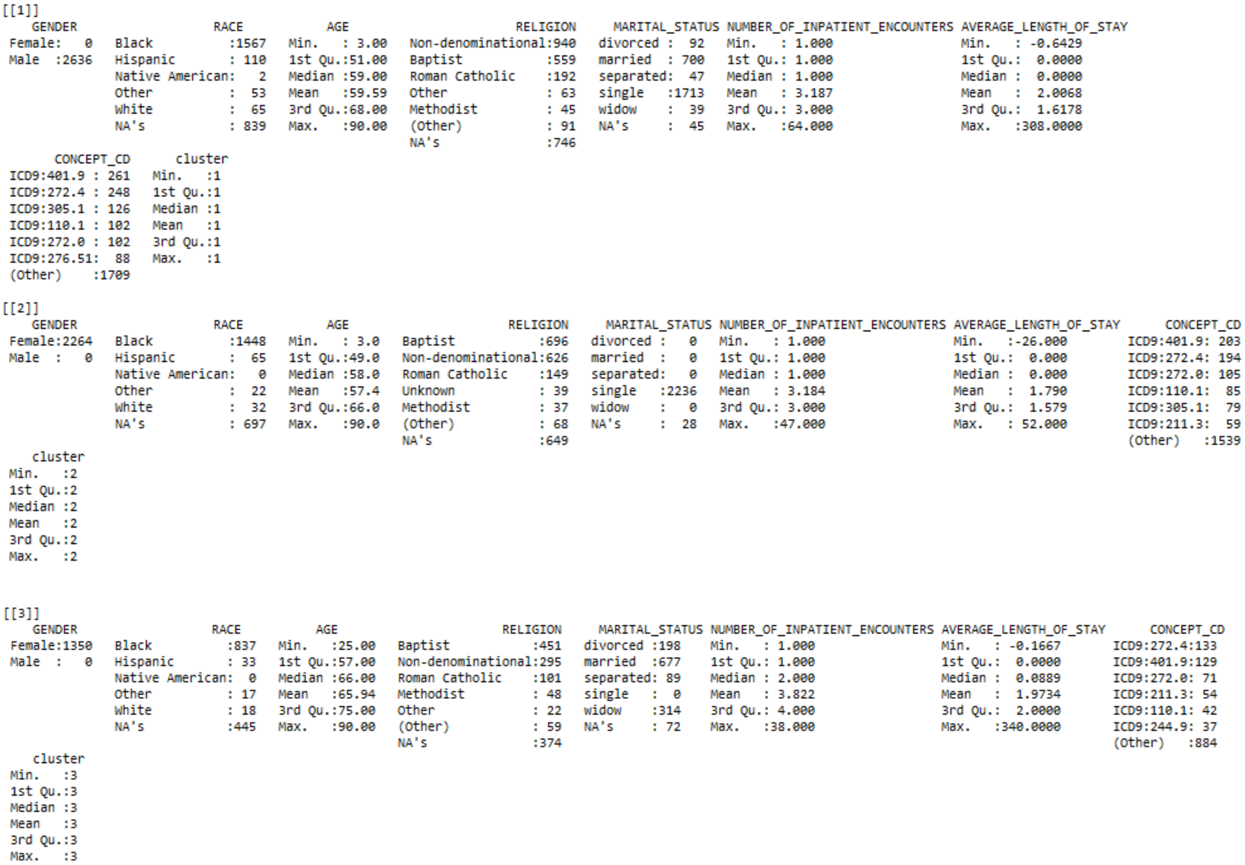
Figure 5.6 Visualization of clustered data though DIANA banner diagram

2 Cluster Summary:



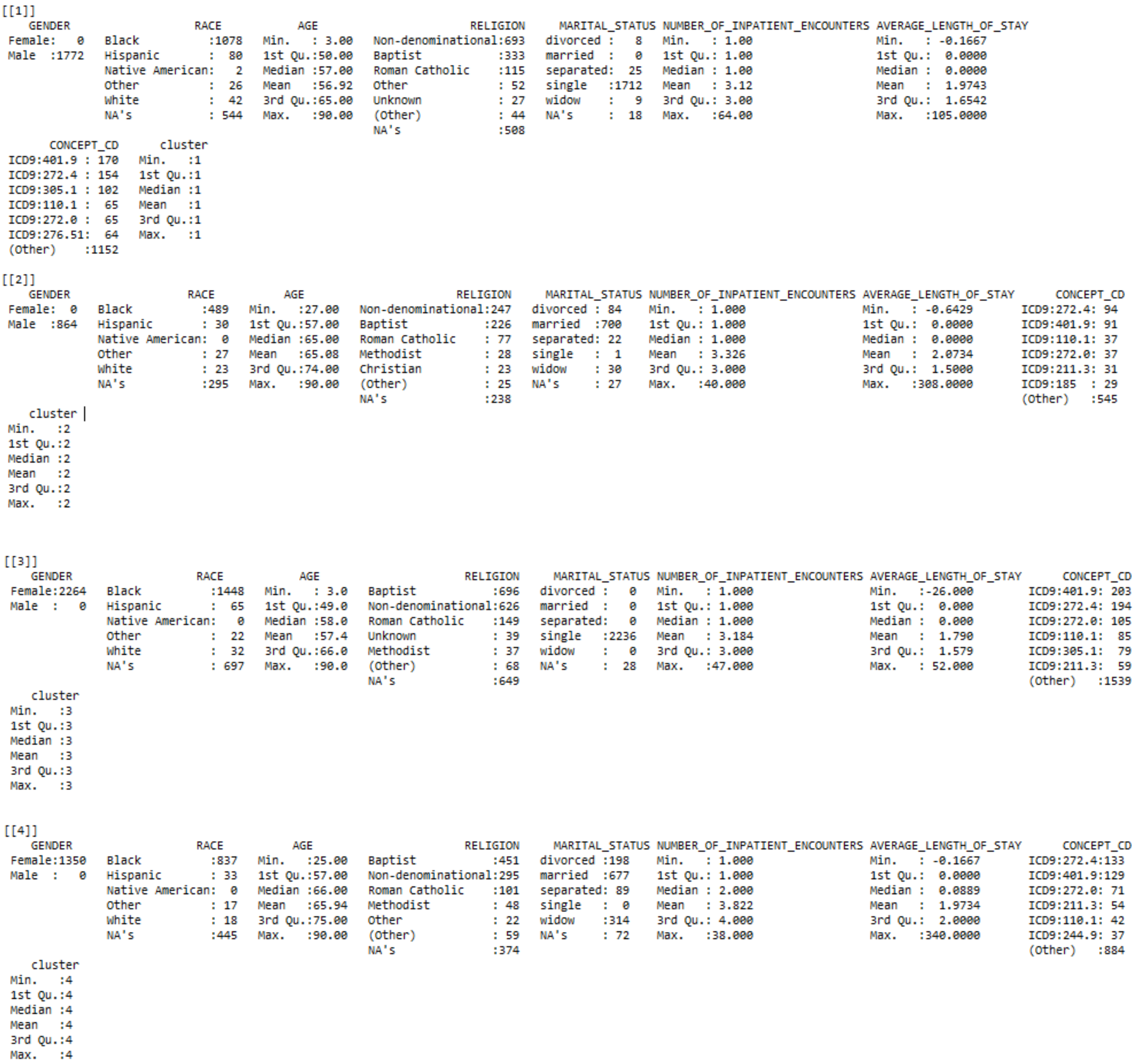
```
[[1]]
   GENDER                  RACE          AGE                     RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1567   Min.   : 3.00   Non-denominational:940   divorced :  92   Min.   : 1.000                 Min.   : -0.6429   
 Male  :2636   Hispanic        : 110   1st Qu.:51.00   Baptist           :559   married  : 700   1st Qu.: 1.000                 1st Qu.:  0.0000   
               Native American:   2   Median :59.00   Roman Catholic    :192   separated:  47   Median : 1.000                 Median :  0.0000   
               Other           :  53   Mean   :59.59   Other             : 63   single   :1713   Mean   : 3.187                 Mean   :  2.0068   
               White           :  65   3rd Qu.:68.00   Methodist         : 45   widow    :  39   3rd Qu.: 3.000                 3rd Qu.:  1.6178   
               NA's            : 839   Max.   :90.00   (Other)           : 91   NA's     :  45   Max.   :64.000                 Max.   :308.0000   
                                                       NA's              :746                                                                     
      CONCEPT_CD       cluster 
 ICD9:401.9 : 261   Min.   :1  
 ICD9:272.4 : 248   1st Qu.:1  
 ICD9:305.1 : 126   Median :1  
 ICD9:110.1 : 102   Mean   :1  
 ICD9:272.0 : 102   3rd Qu.:1  
 ICD9:276.51:  88   Max.   :1  
 (Other)    :1709              

[[2]]
   GENDER                  RACE          AGE                     RELIGION        MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:3614   Black           :2285   Min.   : 3.00   Baptist           :1147   divorced : 198   Min.   : 1.000                 Min.   :-26.000       
 Male  :   0   Hispanic        :  98   1st Qu.:52.00   Non-denominational: 921   married  : 677   1st Qu.: 1.000                 1st Qu.:  0.000       
               Native American:   0   Median :61.00   Roman Catholic    : 250   separated:  89   Median : 1.000                 Median :  0.000       
               Other           :  39   Mean   :60.59   Methodist         :  85   single   :2236   Mean   : 3.422                 Mean   :  1.858       
               White           :  50   3rd Qu.:70.00   Other             :  48   widow    : 314   3rd Qu.: 3.000                 3rd Qu.:  1.761       
               NA's            :1142   Max.   :90.00   (Other)           : 140   NA's     : 100   Max.   :47.000                 Max.   :340.000       
                                                       NA's              :1023                                                                         
     CONCEPT_CD       cluster 
 ICD9:401.9: 332   Min.   :2  
 ICD9:272.4: 327   1st Qu.:2  
 ICD9:272.0: 176   Median :2  
 ICD9:110.1: 127   Mean   :2  
 ICD9:211.3: 113   3rd Qu.:2  
 ICD9:305.1: 110   Max.   :2  
 (Other)   :2429              
```

3 Cluster Summary:



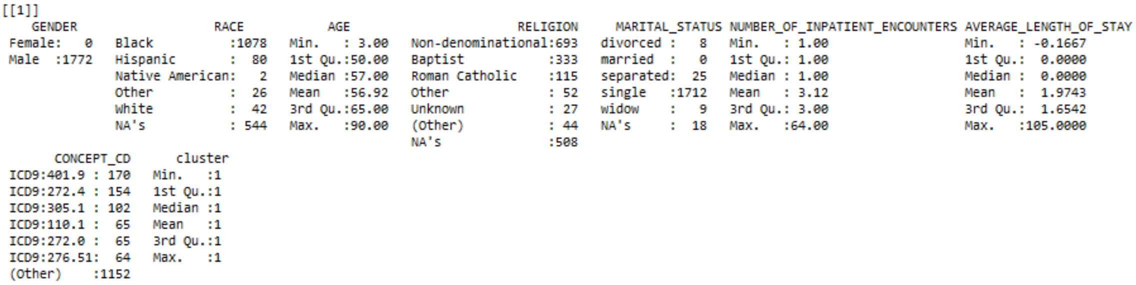
```
[[1]]
   GENDER                  RACE          AGE                     RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1567   Min.   : 3.00   Non-denominational:940   divorced :  92   Min.   : 1.000                 Min.   : -0.6429   
 Male  :2636   Hispanic        : 110   1st Qu.:51.00   Baptist           :559   married  : 700   1st Qu.: 1.000                 1st Qu.:  0.0000   
               Native American:   2   Median :59.00   Roman Catholic    :192   separated:  47   Median : 1.000                 Median :  0.0000   
               Other           :  53   Mean   :59.59   Other             : 63   single   :1713   Mean   : 3.187                 Mean   :  2.0068   
               White           :  65   3rd Qu.:68.00   Methodist         : 45   widow    :  39   3rd Qu.: 3.000                 3rd Qu.:  1.6178   
               NA's            : 839   Max.   :90.00   (Other)           : 91   NA's     :  45   Max.   :64.000                 Max.   :308.0000   
                                                       NA's              :746                                                                     
      CONCEPT_CD       cluster 
 ICD9:401.9 : 261   Min.   :1  
 ICD9:272.4 : 248   1st Qu.:1  
 ICD9:305.1 : 126   Median :1  
 ICD9:110.1 : 102   Mean   :1  
 ICD9:272.0 : 102   3rd Qu.:1  
 ICD9:276.51:  88   Max.   :1  
 (Other)    :1709              

[[2]]
   GENDER                  RACE         AGE                    RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD  
 Female:2264   Black           :1448   Min.   : 3.0   Baptist           :696   divorced :   0   Min.   : 1.000                 Min.   :-26.000        ICD9:401.9: 203  
 Male  :   0   Hispanic        :  65   1st Qu.:49.0   Non-denominational:626   married  :   0   1st Qu.: 1.000                 1st Qu.:  0.000        ICD9:272.4: 194  
               Native American:   0   Median :58.0   Roman Catholic    :149   separated:   0   Median : 1.000                 Median :  0.000        ICD9:272.0: 105  
               Other           :  22   Mean   :57.4   Unknown           : 39   single   :2236   Mean   : 3.184                 Mean   :  1.790        ICD9:110.1:  85  
               White           :  32   3rd Qu.:66.0   Methodist         : 37   widow    :   0   3rd Qu.: 3.000                 3rd Qu.:  1.579        ICD9:305.1:  79  
               NA's            : 697   Max.   :90.0   (Other)           : 68   NA's     :  28   Max.   :47.000                 Max.   : 52.000        ICD9:211.3:  59  
                                                      NA's              :649                                                                         (Other)   :1539  
    cluster 
 Min.   :2  
 1st Qu.:2  
 Median :2  
 Mean   :2  
 3rd Qu.:2  
 Max.   :2  


[[3]]
   GENDER                  RACE         AGE                      RELIGION       MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD 
 Female:1350   Black           :837   Min.   :25.00   Baptist           :451   divorced :198   Min.   : 1.000                 Min.   : -0.1667       ICD9:272.4:133  
 Male  :   0   Hispanic        : 33   1st Qu.:57.00   Non-denominational:295   married  :677   1st Qu.: 1.000                 1st Qu.:  0.0000       ICD9:401.9:129  
               Native American:  0   Median :66.00   Roman Catholic    :101   separated: 89   Median : 2.000                 Median :  0.0889       ICD9:272.0: 71  
               Other           : 17   Mean   :65.94   Methodist         : 48   single   :  0   Mean   : 3.822                 Mean   :  1.9734       ICD9:211.3: 54  
               White           : 18   3rd Qu.:75.00   Other             : 22   widow    :314   3rd Qu.: 4.000                 3rd Qu.:  2.0000       ICD9:110.1: 42  
               NA's            :445   Max.   :90.00   (Other)           : 59   NA's     : 72   Max.   :38.000                 Max.   :340.0000       ICD9:244.9: 37  
                                                      NA's              :374                                                                        (Other)   :884  
    cluster 
 Min.   :3  
 1st Qu.:3  
 Median :3  
 Mean   :3  
 3rd Qu.:3  
 Max.   :3  
```

4 Cluster Summary:

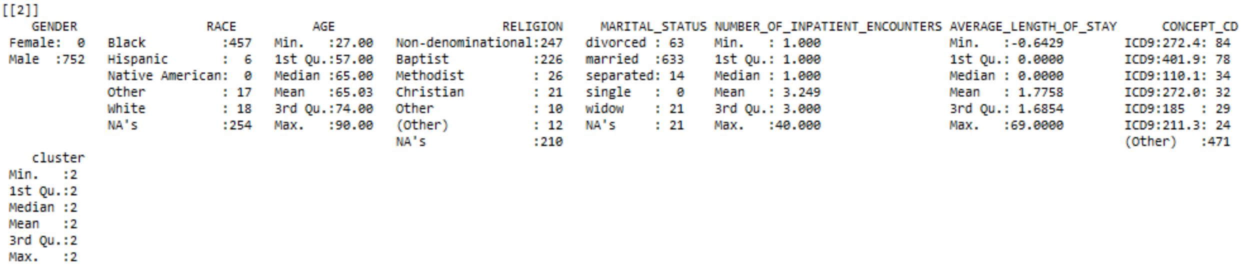


```
[[1]]
   GENDER                  RACE           AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1078   Min.   : 3.00   Non-denominational:693   divorced :   8   Min.   : 1.00                  Min.   : -0.1667
 Male  :1772   Hispanic        :  80   1st Qu.:50.00   Baptist           :333   married  :   0   1st Qu.: 1.00                  1st Qu.:  0.0000
               Native American:   2   Median :57.00   Roman Catholic    :115   separated:  25   Median : 1.00                  Median :  0.0000
               Other           :  26   Mean   :56.92   Other             : 52   single   :1712   Mean   : 3.12                  Mean   :  1.9743
               White           :  42   3rd Qu.:65.00   Unknown           : 27   widow    :   9   3rd Qu.: 3.00                  3rd Qu.:  1.6542
               NA's            : 544   Max.   :90.00   (Other)           : 44   NA's     :  18   Max.   :64.00                  Max.   :105.0000
                                                       NA's              :508
      CONCEPT_CD       cluster
 ICD9:401.9 : 170   Min.   :1
 ICD9:272.4 : 154   1st Qu.:1
 ICD9:305.1 : 102   Median :1
 ICD9:110.1 :  65   Mean   :1
 ICD9:272.0 :  65   3rd Qu.:1
 ICD9:276.51:  64   Max.   :1
 (Other)    :1152

[[2]]
   GENDER                 RACE          AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:  0   Black           :489   Min.   :27.00   Non-denominational:247   divorced : 84   Min.   : 1.000                  Min.   : -0.6429       ICD9:272.4: 94
 Male  :864   Hispanic        : 30   1st Qu.:57.00   Baptist           :226   married  :700   1st Qu.: 1.000                  1st Qu.:  0.0000       ICD9:401.9: 91
              Native American:  0   Median :65.00   Roman Catholic    : 77   separated: 22   Median : 1.000                  Median :  0.0000       ICD9:110.1: 37
              Other           : 27   Mean   :65.08   Methodist         : 28   single   :  1   Mean   : 3.326                  Mean   :  2.0734       ICD9:272.0: 37
              White           : 23   3rd Qu.:74.00   Christian         : 23   widow    : 30   3rd Qu.: 3.000                  3rd Qu.:  1.5000       ICD9:211.3: 31
              NA's            :295   Max.   :90.00   (Other)           : 25   NA's     : 27   Max.   :40.000                  Max.   :308.0000       ICD9:185  : 29
                                                     NA's              :238                                                                        (Other)   :545
   cluster |
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2


[[3]]
   GENDER                  RACE          AGE                        RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female:2264   Black           :1448   Min.   : 3.0   Baptist           :696   divorced :   0   Min.   : 1.000                  Min.   :-26.000        ICD9:401.9: 203
 Male  :   0   Hispanic        :  65   1st Qu.:49.0   Non-denominational:626   married  :   0   1st Qu.: 1.000                  1st Qu.:  0.000        ICD9:272.4: 194
               Native American:   0   Median :58.0   Roman Catholic    :149   separated:   0   Median : 1.000                  Median :  0.000        ICD9:272.0: 105
               Other           :  22   Mean   :57.4   Unknown           : 39   single   :2236   Mean   : 3.184                  Mean   :  1.790        ICD9:110.1:  85
               White           :  32   3rd Qu.:66.0   Methodist         : 37   widow    :   0   3rd Qu.: 3.000                  3rd Qu.:  1.579        ICD9:305.1:  79
               NA's            : 697   Max.   :90.0   (Other)           : 68   NA's     :  28   Max.   :47.000                  Max.   : 52.000        ICD9:211.3:  59
                                                      NA's              :649                                                                           (Other)   :1539
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
 3rd Qu.:3
 Max.   :3

[[4]]
   GENDER                  RACE         AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY       CONCEPT_CD
 Female:1350   Black           :837   Min.   :25.00   Baptist           :451   divorced :198   Min.   : 1.000                  Min.   : -0.1667       ICD9:272.4:133
 Male  :   0   Hispanic        : 33   1st Qu.:57.00   Non-denominational:295   married  :677   1st Qu.: 1.000                  1st Qu.:  0.0000       ICD9:401.9:129
               Native American:  0   Median :66.00   Roman Catholic    :101   separated: 89   Median : 2.000                  Median :  0.0889       ICD9:272.0: 71
               Other           : 17   Mean   :65.94   Methodist         : 48   single   :  0   Mean   : 3.822                  Mean   :  1.9734       ICD9:211.3: 54
               White           : 18   3rd Qu.:75.00   Other             : 22   widow    :314   3rd Qu.: 4.000                  3rd Qu.:  2.0000       ICD9:110.1: 42
               NA's            :445   Max.   :90.00   (Other)           : 59   NA's     : 72   Max.   :38.000                  Max.   :340.0000       ICD9:244.9: 37
                                                      NA's              :374                                                                          (Other)   :884
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```

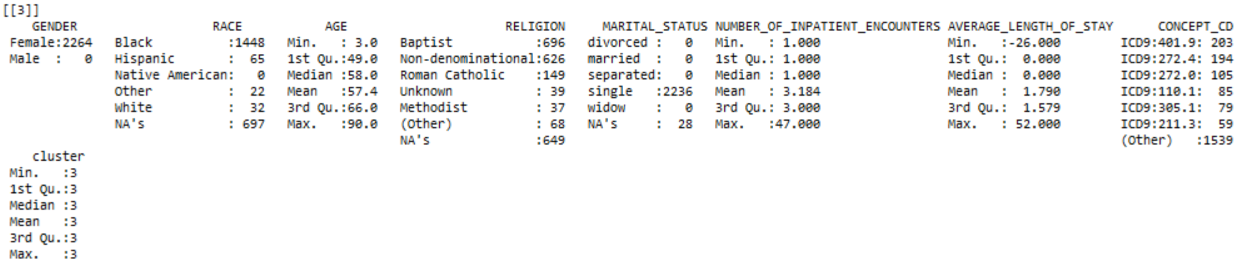
5 Cluster Summary:



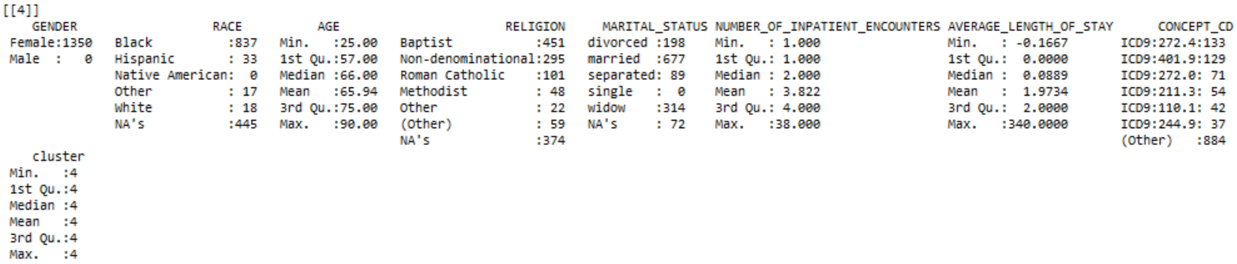
```
[[1]]
   GENDER                  RACE            AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY
 Female:   0   Black           :1078   Min.   : 3.00   Non-denominational:693   divorced :   8   Min.   : 1.00                  Min.   : -0.1667
 Male  :1772   Hispanic        :  80   1st Qu.:50.00   Baptist           :333   married  :   0   1st Qu.: 1.00                  1st Qu.:  0.0000
               Native American:   2   Median :57.00   Roman Catholic    :115   separated:  25   Median : 1.00                  Median :  0.0000
               Other           :  26   Mean   :56.92   Other             : 52   single   :1712   Mean   : 3.12                  Mean   :  1.9743
               White           :  42   3rd Qu.:65.00   Unknown           : 27   widow    :   9   3rd Qu.: 3.00                  3rd Qu.:  1.6542
               NA's            : 544   Max.   :90.00   (Other)           : 44   NA's     :  18   Max.   :64.00                  Max.   :105.0000
                                                       NA's              :508
      CONCEPT_CD      cluster
 ICD9:401.9 : 170   Min.   :1
 ICD9:272.4 : 154   1st Qu.:1
 ICD9:305.1 : 102   Median :1
 ICD9:110.1 :  65   Mean   :1
 ICD9:272.0 :  65   3rd Qu.:1
 ICD9:276.51:  64   Max.   :1
 (Other)    :1152
```



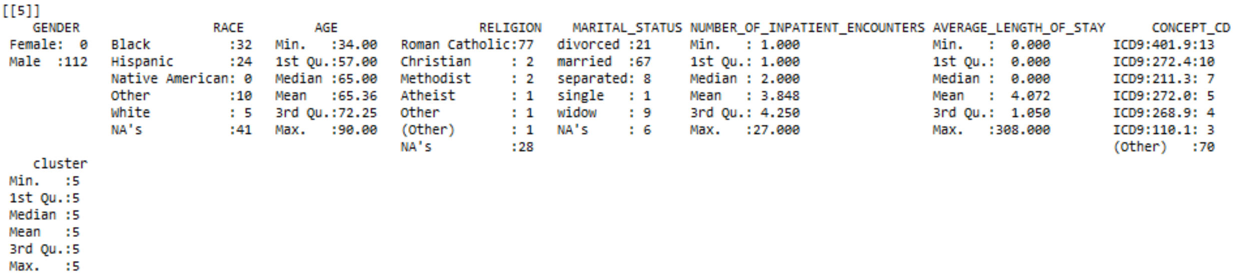
```
[[2]]
   GENDER                 RACE           AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:  0   Black           :457   Min.   :27.00   Non-denominational:247   divorced : 63   Min.   : 1.000                 Min.   :-0.6429         ICD9:272.4: 84
 Male  :752   Hispanic        :  6   1st Qu.:57.00   Baptist           :226   married  :633   1st Qu.: 1.000                 1st Qu.: 0.0000         ICD9:401.9: 78
              Native American:  0   Median :65.00   Methodist         : 26   separated: 14   Median : 1.000                 Median : 0.0000         ICD9:110.1: 34
              Other           : 17   Mean   :65.03   Christian         : 21   single   :  0   Mean   : 3.249                 Mean   : 1.7758         ICD9:272.0: 32
              White           : 18   3rd Qu.:74.00   Other             : 10   widow    : 21   3rd Qu.: 3.000                 3rd Qu.: 1.6854         ICD9:185  : 29
              NA's            :254   Max.   :90.00   (Other)           : 12   NA's     : 21   Max.   :40.000                 Max.   :69.0000         ICD9:211.3: 24
                                                     NA's              :210                                                                          (Other)   :471
   cluster
 Min.   :2
 1st Qu.:2
 Median :2
 Mean   :2
 3rd Qu.:2
 Max.   :2
```



```
[[3]]
   GENDER                  RACE           AGE                         RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:2264   Black           :1448   Min.   : 3.0   Baptist           :696   divorced :   0   Min.   : 1.000                 Min.   :-26.000         ICD9:401.9: 203
 Male  :   0   Hispanic        :  65   1st Qu.:49.0   Non-denominational:626   married  :   0   1st Qu.: 1.000                 1st Qu.:  0.000         ICD9:272.4: 194
               Native American:   0   Median :58.0   Roman Catholic    :149   separated:   0   Median : 1.000                 Median :  0.000         ICD9:272.0: 105
               Other           :  22   Mean   :57.4   Unknown           : 39   single   :2236   Mean   : 3.184                 Mean   :  1.790         ICD9:110.1:  85
               White           :  32   3rd Qu.:66.0   Methodist         : 37   widow    :   0   3rd Qu.: 3.000                 3rd Qu.:  1.579         ICD9:305.1:  79
               NA's            : 697   Max.   :90.0   (Other)           : 68   NA's     :  28   Max.   :47.000                 Max.   : 52.000         ICD9:211.3:  59
                                                      NA's              :649                                                                          (Other)   :1539
   cluster
 Min.   :3
 1st Qu.:3
 Median :3
 Mean   :3
 3rd Qu.:3
 Max.   :3
```



```
[[4]]
   GENDER                  RACE           AGE                          RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:1350   Black           :837   Min.   :25.00   Baptist           :451   divorced :198   Min.   : 1.000                 Min.   : -0.1667        ICD9:272.4:133
 Male  :   0   Hispanic        : 33   1st Qu.:57.00   Non-denominational:295   married  :677   1st Qu.: 1.000                 1st Qu.:  0.0000        ICD9:401.9:129
               Native American:  0   Median :66.00   Roman Catholic    :101   separated: 89   Median : 2.000                 Median :  0.0889        ICD9:272.0: 71
               Other           : 17   Mean   :65.94   Methodist         : 48   single   :  0   Mean   : 3.822                 Mean   :  1.9734        ICD9:211.3: 54
               White           : 18   3rd Qu.:75.00   Other             : 22   widow    :314   3rd Qu.: 4.000                 3rd Qu.:  2.0000        ICD9:110.1: 42
               NA's            :445   Max.   :90.00   (Other)           : 59   NA's     : 72   Max.   :38.000                 Max.   :340.0000        ICD9:244.9: 37
                                                      NA's              :374                                                                          (Other)   :884
   cluster
 Min.   :4
 1st Qu.:4
 Median :4
 Mean   :4
 3rd Qu.:4
 Max.   :4
```



```
[[5]]
   GENDER                  RACE          AGE                      RELIGION     MARITAL_STATUS NUMBER_OF_INPATIENT_ENCOUNTERS AVERAGE_LENGTH_OF_STAY      CONCEPT_CD
 Female:  0   Black           :32   Min.   :34.00   Roman Catholic:77   divorced :21   Min.   : 1.000                 Min.   :  0.000         ICD9:401.9:13
 Male  :112   Hispanic        :24   1st Qu.:57.00   Christian     : 2   married  :67   1st Qu.: 1.000                 1st Qu.:  0.000         ICD9:272.4:10
              Native American: 0   Median :65.00   Methodist     : 2   separated: 8   Median : 2.000                 Median :  0.000         ICD9:211.3: 7
              Other           :10   Mean   :65.36   Atheist       : 1   single   : 1   Mean   : 3.848                 Mean   :  4.072         ICD9:272.0: 5
              White           : 5   3rd Qu.:72.25   Other         : 1   widow    : 9   3rd Qu.: 4.250                 3rd Qu.:  1.050         ICD9:268.9: 4
              NA's            :41   Max.   :90.00   (Other)       : 1   NA's     : 6   Max.   :27.000                 Max.   :308.000         ICD9:110.1: 3
                                                    NA's          :28                                                                       (Other)   :70
   cluster
 Min.   :5
 1st Qu.:5
 Median :5
 Mean   :5
 3rd Qu.:5
 Max.   :5
```

5.5 Discussion

Given the four algorithms, the next step was to compare the cluster patterns between the algorithms. Did the cluster the data similarly? Comparative analysis was done to determine whether the data clustered similarly, and to which algorithm clusters the data the best.

Table 5.1, Table 5.2, Table 5.3, and Table 5.4 shows the frequency of males and females in the instance where the data was split into four clusters, via the K-modes, PAM, Hclust, and DIANA algorithms. What was being observed here is the degree of overlap. The better the data is clustered, the more segregated it is by column; more data in one column, vs. another.

Table 5.1

K-MODES Clusters

	Male	**Female**	**Mode Marital Status**
Cluster 1	875	60	Single
Cluster 2	62	955	Single
Cluster 3	328	38	Married
Cluster 4	1	658	single

Table 5.2

PAM Clusters:

	Male	**Female**	**Mode Marital Status**
Cluster 1	1672	0	single
Cluster 2	951	79	married
Cluster 3	0	2302	single
Cluster 4	13	1233	married

Table 5.3

HClust Clusters:

	Male	**Female**	**Mode Marital Status**
Cluster 1	2545	1	Single
Cluster 2	0	2964	Single
Cluster 3	0	645	Married
Cluster 4	91	4	divorced

Table 5.4

DIANA Clusters:

	Male	**Female**	**Mode Marital Status**
Cluster 1	1772	0	Single
Cluster 2	864	0	Married
Cluster 3	0	2264	Single
Cluster 4	0	1350	Married

In the small sample of features and their counts, of all the algorithms performed similarly except for HClust, clustering samples of single men, married men, single woman, married woman. Hclust presented a subgroup of married men.

The DIANA algorithm, however, stood out. It clustered data with very little overlap, as evidenced in the table. When splitting along the lines of gender, the algorithm delivered perfectly segregated gender findings. For this reason, all further analysis on the dataset was used by DIANA's interpretation of clustering.

5.6 Top Mode Diseases per Cluster

DIANA
2 CLUSTERS
1)
ICD9:401.9 - Hypertensive disease NOS – 9.9%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 9.4%
ICD9:305.1 - Tobacco use disorder – 4.78%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.87%
ICD9:272.0 – Cholesterolemia – 3.87%
ICD9:276.51 – Dehydration – 3.34%

2)
ICD9:401.9 - Hypertensive disease NOS – 9.19%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 12.4%
ICD9:272.0 – Cholesterolemia – 6.68 %
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.5%
ICD9:211.3 - Benign Neoplasm of Colon – 3.1%
ICD9:305.1 - Tobacco use disorder – 3.0%

3 CLUSTSERS

1)
ICD9:401.9 - Hypertensive disease NOS – 9.9%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 9.4%
ICD9:305.1 - Tobacco use disorder – 4.8%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.9%
ICD9:272.0 – Cholesterolemia – 3.9%

ICD9:276.51 – Dehydration – 3.3%

2)
ICD9:401.9 - Hypertensive disease NOS – 9.0%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.6%
ICD9:272.0 – Cholesterolemia – 4.6%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.8%
ICD9:305.1 - Tobacco use disorder – 3.5%
ICD9:211.3 - Benign Neoplasm of Colon – 2.6%

3)
ICD9:272.4 - Hyperlipidemia, other and unspecified – 9.9%
ICD9:401.9 - Hypertensive disease NOS – 9.6%
ICD9:272.0 – Cholesterolemia – 5.3%
ICD9:211.3 - Benign Neoplasm of Colon – 4.0%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.1%
ICD9:244.9 - Acquired hypothyroidism NOS – 2.7%

4 CLUSTERS

1)
ICD9:401.9 - Hypertensive disease NOS - 9.6%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.7%
ICD9:305.1 - Tobacco use disorder – 5.8%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.7%
ICD9:272.0 – Cholesterolemia- 3.7%
ICD9:276.51 – Dehydration – 3.65

2)
ICD9:272.4 - Hyperlipidemia, other and unspecified – 10.9%
ICD9:401.9 - Hypertensive disease NOS – 10.5%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 4.3%
ICD9:272.0 – Cholesterolemia – 4.3%
ICD9:211.3 - Benign Neoplasm of Colon – 3.5%
ICD9:185 - Neoplasm, malignant, of prostate – 3.4%

3)
ICD9:401.9 - Hypertensive disease NOS – 9.0%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.7%
ICD9:272.0 – Cholesterolemia – 4.6%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.8%
ICD9:305.1 - Tobacco use disorder – 3.5%
ICD9:211.3 - Benign Neoplasm of Colon – 2.6%

4)
ICD9:272.4 - Hyperlipidemia, other and unspecified – 9.8%
ICD9:401.9 - Hypertensive disease NOS – 9.6%
ICD9:272.0 – Cholesterolemia – 5.3%
ICD9:211.3 - Benign Neoplasm of Colon - 4.0%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.1%
ICD9:244.9 - Acquired hypothyroidism NOS – 2.7%

5 CLUSTERS

1)
ICD9:401.9 - Hypertensive disease NOS – 9.6%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.7%
ICD9:305.1 - Tobacco use disorder – 5.8%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.7%
ICD9:272.0 – Cholesterolemia – 3.7%
ICD9:276.51 – Dehydration – 3.6%

2)
ICD9:272.4 - Hyperlipidemia, other and unspecified – 11.2%
ICD9:401.9 - Hypertensive disease NOS – 10.4%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 4.5%
ICD9:272.0 – Cholesterolemia – 4.3%
ICD9:185 - Neoplasm, malignant, of prostate – 3.9%
ICD9:211.3 - Benign Neoplasm of Colon – 3.2%

3)
ICD9:401.9 - Hypertensive disease NOS – 9.0%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.6%
ICD9:272.0 – Cholesterolemia – 4.6%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 3.8%
ICD9:305.1 - Tobacco use disorder – 3.5%
ICD9:211.3 - Benign Neoplasm of Colon – 2.6%

4)
ICD9:272.4 - Hyperlipidemia, other and unspecified – 9.8%
ICD9:401.9 - Hypertensive disease NOS – 9.6%
ICD9:272.0 – Cholesterolemia – 5.3%
ICD9:211.3 - Benign Neoplasm of Colon - 4.0%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum - 3.1%
ICD9:244.9 - Acquired hypothyroidism NOS – 2.7%

5)
ICD9:401.9 - Hypertensive disease NOS – 11.6%
ICD9:272.4 - Hyperlipidemia, other and unspecified – 8.9%

ICD9:211.3 - Benign Neoplasm of Colon – 6.3%
ICD9:272.0 – Cholesterolemia – 4.5%
ICD9:268.9 - Ergosterol deficiency – 3.6%
ICD9:110.1 - Onychomycosis due to Trichophyton rubrum – 2.7%

5.7 Findings

Doing analysis on the data, trends and recurrences appeared. Creating two clusters indicated larger trends while creating smaller clusters (a maximum of 5), showed more minute patterns. Each patient in the dataset was assigned an ICD-9 code of the mode comorbidity. Comorbidity is the presence of other chronic diseases in a patient. In the summary of each cluster, the frequencies of the comorbidities of each patient were tallied, and the clusters were parsed for insights.

In total six observations were made:

1) In each subgroup, whether in clusters of 2, 3, 4 and 5, the number of comorbidities were dominated by
    a. Hypertensive Disease NOS – High Blood Pressure
    b. Hyperlipidemia – Abnormally elevated levels of fat in the blood
    c. Cholesterolemia – The presence of elevated levels of cholesterol in the blood.
2) Prevalence of Onychomycosis due to Trichophyton rubrum
3) Elevated rates of Benign Neoplasm of Colon
4) Acquired hypothyroidism cases showing up in women who are black, female, and non-single, with a median age of 66. (0 cases of single women).
5) Incidences of Neoplasm, malignant, of prostate in men who are black, non-single, with a mean age of 65.

6) Incidences of Tobacco use disorder showing up in clusters featuring mostly single men and women.

Chapter 6. CONCLUSION

Each of these observations either validate existing trends of diabetes patients or provide potential new areas of research on comorbidity, risk factors, and demographics. From using cluster analysis, we can use more specific demographic information to learn more about the disease.

Discussion of the conclusion can be described in two parts:

- Validation of existing knowledge of diabetes type II comorbidity,
- The effect of demographic features on co-morbidity, mainly gender and marital status.

Firstly, the data validates existing correlations of diabetes type II. These were the top three diseases in all of the clusters:

- Hypertensive disease – High blood pressure, and has a high correlation for African-Americans, and occurs more frequently among black than white Americans with diabetes [1]
- Hyperlipidemia - Is a result of heightened levels of fats and lipoproteins in the blood. It is commonly normally associated with diabetes and is the most common cause of diabetes death [23]
- Cholesterolemia - is the presence of elevated levels of cholesterol in the blood. According to the journal chapter, 'Diabetes In African Americans', by Eugene S. Tull and Jeffrey M. Roseman, they insist that 'Individuals who have insulin-resistant diabetes have higher levels of cardiovascular disease risk factors, including LDL- cholesterol and triglyceries.

These three diseases, and their prevalence in the data, only served to prove that the research was on the right track. The fact that the primary three diseases uncovered by cluster analysis matches pre-existing notions for diabetes type II risk factors is an indication that the results were good.

Onychomycosis due to Trichophyton rubrum is a common disease and showed up frequently in results. It occurs in toenails and is caused by the fungus Trichophyton rubrum. According to the paper ‘Prevalence of Toe Nail Onychomycosis in Diabetic Patients’, by Saunte, Holgersen, et al., ‘Male gender and old age are predisposing factors for fungal nail infection, as well as diabetes, psoriasis, peripheral arterial disease and immune suppression’. This shows that a correlation does exist, and the results validate this. What remains to be seen however, is the realationship that Onchomycosis has between age, diabetes type II, and African Americans. This relationship remains unexplored in research.

As more observations were made, more interesting correlations were discovered, mainly around marital status and gender. Hypothyroidism is the most common adult thyroid [28] disease in adults and is the most common in women. Results showed that the disease has a higher rate in women who are black, non-single, and with a mean age of 65. Diabetic patients have higher rates of thyroid disorders compared with the normal population, correlating with the results. This includes patients who are married, divorced, widowed or separated. They were all slightly at higher risk.

Men were not exempt. Non-single black men with a mean age of 35 showed elevated incidences of Neoplasm, malignant, of the prostate; prostate cancer. Again, this was also spread across men who are married, divorced, widowed or separated.

Marital status again played in a role in disease correlation. Clusters featuring mostly either single black men, or single black women, each at the mean ages of 56.92, and 57.4, respectively had higher incidences of Tobacco use disorder. What is important to note here for both previous examples is that there seemingly is a correlation between diseases and marital status. This is an indication of a potential pattern. Does marital status influence diabetes type II patients? Perhaps. Perhaps not. What matters is that a trend was spotted – a potential thread of investigation. This is the power of machine learning. Using cluster analysis, we were better able to observe that a seemingly innocuous demographic factor may affect what subsequent diseases a diabetes type II patient will suffer.

In conclusion, by using cluster analysis on minority health records, we were better able to understand comorbidities of diabetes type II in African Americans. Present correlations were validated, and new ones were found. Insights from demographic data can now spur further research on this disease and its effects.

Chapter 7. FUTURE WORK

Cluster analysis on minority health data is a field in its infancy. The work done in this thesis is only the start of what can be an established precedent of using machine learning algorithms on health data at Howard University.

Future work includes:

- Using cluster analysis to improve the accuracy rate in predicting whether someone has diabetes, by creating a “synthetic” feature for supervised learning.
- Using unused features in this project to provide additional information e.g. Clustering diabetes 2 patients around zip code data to find out the density of diabetes patients in different areas.
- Performing cluster analysis on established datasets (i.e. CDC) and comparing results to Howard University’s data to produce insights.
- Comparing cluster analysis results of people from other ethnic groups to Howard University’s results may uncover unforeseen relationships.

BIBLIOGRAPHY

[1] M C Marshall Jr. Diabetes in African Americans. www.postgradmedj.com.

[2] Padmaja, P., et al. "Characteristic evaluation of diabetes data using clustering techniques." *IJCSNS* 8.11 (2008): 244.

[3] Kothainayaki, M., and P. Thangaraj. "Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm." *Journal of Applied Information Science* 1.1 (2015).

[4] Cook, Rachel, and Gongzhu Hu. "Hidden Patterns: Clustering Diabetes Data." *CAINE*. 2010.

[5] Yao, Jin, et al. "Feature selection for unsupervised learning through local learning." *Pattern Recognition Letters* 53 (2015): 100-107.

[6] Doshi-Velez, Finale, Yaorong Ge, and Isaac Kohane. "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis." *Pediatrics* 133.1 (2014): e54-e63.

[7] Li, Li, et al. "Identification of type 2 diabetes subgroups through topological analysis of patient similarity." *Science translational medicine* 7.311 (2015): 311ra174-311ra174.

[8] Cheung, Shun Yan, and Mostafa H. Ammar. "Using destination set grouping to improve the performance of window-controlled multipoint connections." *Computer Communications* 19.8 (1996): 723-736.

[9] Docampo, Elisa, et al. "Cluster analysis of clinical data identifies fibromyalgia subgroups." *PLoS One* 8.9 (2013): e74873.

[10] Freitas, Alberto, Pavel Brazdil, and A. Costa-Pereira. "Mining hospital databases for management support." *IADIS Virtual Multi Conference on Computer Science and Information Systems*. 2005.

[11] Doshi-Velez, Finale, Yaorong Ge, and Isaac Kohane. "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis." *Pediatrics* 133.1 (2014): e54-e63.

[12] FAQ | i2b2 Research Data Warehouse

"FAQ | I2b2 Research Data Warehouse". *I2b2.cchmc.org*. N. p., 2017. Web. 16 Apr. 2017.

[13] "Best Practices In Preparing Data Files For Importing Into R - Easy Guides - Wiki - STHDA". *Sthda.com*. N. p., 2017. Web. 16 Apr. 2017.

[14] Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." *DMKD*. 1997.

[15] Huang, Zhexue. "Extensions to the k-means algorithm for clustering large data sets with categorical values." *Data mining and knowledge discovery 2.3* (1998): 283-304.

[16] Kaufman, Leonard, and Peter J. Rousseenw. "Patirionsing around medoids (program pam)." *Finding groups in data: an introduction to cluster analysis* (1990): 68-125.

[17] Kodali, Teja. "Hierarchical Clustering In R". *R-bloggers*. N. p., 2016. Web. 16 Apr. 2017.

[18] "Cluster Analysis". *Stat.berkeley.edu*. N. p., 2017. Web. 16 Apr. 2017.

[19] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009..

[20] "R: Dissimilarity Matrix Calculation". *Stat.ethz.ch*. N. p., 2017. Web. 16 Apr. 2017.

[21] "Clustering Mixed Data Types In R". *R-bloggers*. N. p., 2016. Web. 16 Apr. 2017.

[22] Python, Comprehensive, and Comprehensive Python. "Comprehensive Guide On T-SNE Algorithm With Implementation In R & Python". *Analytics Vidhya*. N. p., 2017. Web. 16 Apr. 2017.

[23] "Tips From Other Journals - American Family Physician". *Aafp.org*. N. p., 2017. Web. 16 Apr. 2017.

[24] "What Is ICD-9-CM (International Classification Of Diseases, Ninth Revision, Clinical Modification) ? - Definition From Whatis.Com". *SearchHealthIT*. N. p., 2017. Web. 16 Apr. 2017.

[25] Petchey, Owen L., and Kevin J. Gaston. "Dendrograms and measuring functional diversity." *Oikos* 116.8 (2007): 1422-1426.

[26] Saunte, Ditte Marie L., et al. "Prevalence of toe nail onychomycosis in diabetic patients." *Acta dermato-venereologica* 86.5 (2006): 425-428.