Finding Unique Patterns in Dialysis Facilities with Patients' Data using Unsupervised Learning Algorithms

Capstone project for Data Science Career Track bootcamp

Benhur Tedros

Summary

Kidney failure (ESRD: end stage renal disease) is one of the leading causes of death in the United States. According to USRDS 2013 Annual data report, this disease affects almost 650,000 people per year in US, and its rate is increasing by 5% each year. Today, ESRD patients have two treatment options, which are either kidney transplantation or dialysis. The best current treatment is the transplantation; however, the number of kidney donors to ESRD patients' ratio is 1 to 6. Moreover, the need for kidney is increasing at 8% per year while their availability has not grown to match up that number. Therefore, dialysis is the only alternative option that the patients on waiting list have. People undergoing dialysis often have multiple health concerns, which can have an adverse impact on their life expectancy, though dialysis may offer a better quality of life. According to the National Institute of Diabetes and Digestive and Kidney Diseases report, the two-year, five year and ten-year survival rates are around 64%, 33%, and 10% respectively. The survival or mortality rate varies from one dialysis facility to another. In light of this, "could we find groups of dialysis facilities with similar behavior given a dataset of the patients' health records that have been collected during their visits?" For example, would the effect be better if the facility use hemodialysis for patient A and peritoneal dialysis for patient B or vice versa.

There are several dialysis facilities registered with Medicare in the country where patients often visit. Besides the importance of the quality of care they provide, additional health data of their patients is collected. Some of the data include mortality rate (deaths), hospitalizations, blood transfusions, incidents of hyperkalemia (too much calcium in the blood), percentage of waste removed during hemodialysis in adults and children, percentage of waste removed in adults during peritoneal dialysis, percentage of AV fistulas, percentage of catheters in use over 90 days and others.

This capstone project will try to find a meaningful pattern within the health data of the patients and cluster the facilities with similar behavior into groups. The result of this work will primarily help the dialysis facilities to improve their services. The Medicare department of the Federal government will also be benefited from the result of this project.

Objective of the Project

The goal of this capstone project is:

- Organize the dialysis facilities into groups in which the facilities in each cluster are similar in some way

Data

The data for this project was published by Centers for Medicare & Medicaid Services and was downloaded from the DATA.MEDICARE.GOV. The dataset is comprised of data on anemia management, phosphorus levels, transfusion rate, dialysis adequacy, vascular access, mineral and bone disorder, hospitalization rate, readmission rate, infection ratio, scale rate of the facility and others. The data was collected from 2012 to 2015 and can be downloaded from:

https://data.medicare.gov/Dialysis-Facility-Compare/Dialysis-Facility-Compare-Listing-by-Facility/23ew-n7w9/data

The details of the data fields with their term definitions can be found at the following link by clicking the "get supporting documents tab":

https://data.medicare.gov/data/dialysis-facility-compare

Two additional dataset were also used to merge with the original one. The first dataset includes population size for each county, and the second one has information on household income based on counties. The dataset were downloaded from the websites which are hyperlinked below.

https://www.census.gov/data/datasets/2016/demo/popest/counties-detail.html

https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income

Methodology

This project will be treated as unsupervised learning classification problem. Data loading, data wrangling and cleaning, feature selection, exploratory data analysis, inferential statistics, matrix manipulation, data visualization, clustering, model evaluation were carried out one after another to achieve the objectives of the project. The libraries used in this project include numpy, pandas, matplotlib, and scikit-learn.

A. Data Wrangling and Cleaning

The dataset are stored in MS Excel spreadsheet in CSV format, which were easily loaded into pandas data frames. Several components of data wrangling were utilized to make the dataset ready to be used for clustering. The dataset was originally in the format of CSV and loaded to the python notebook. It contains 6810 observations with 98 data fields with some missing/null values. Many of the names of the data fields were too long and were shortened for our convenience. The dataset was subset into fewer data fields/features depending on their scalability nature and importance to our prediction. Some of the columns in the dataset contained categorical variables, but were converted into numerical nature in order to get them included in the analysis. As this dataset was short of information on population size and household income in each county where the dialysis facilities are located, two additional dataset with the same data were merged to the original dataset before conducting the exploratory data analysis.

B. Feature Selection

The dataset has 6810 observations with 98 data fields with some null values dataset and contains some non-scalable and categorical variables/ non-numerical labels. The non-scalable data fields were dropped, and the others with the categorical variables were converted into numerical variables in order not to miss important information from those attributes. Population size and household income data fields are extracted from other datasets.

Forty six data fields/features were selected to be utilized in the model prediction.

C. Exploratory data Analysis (EDA) and Inferential Statistics

Exploratory data analysis for this dataset is useful in determining relationships among the explanatory variables, preliminary selection of appropriate models and assessing the direction and rough size of relationships among explanatory variables. This section will try to answer if there is significance in terms of explaining unique pattern among the features given the health data of patients in the dialysis facilities across the States. Moreover, are there strong correlations between pairs of facilities or States? Before hand, it is vital to explore the distribution of the some of the features such as Mortality_rate, five_star (which is 1-5 scale reviews for the facilities) across the States. The mean and median of the Mortality_rate and five_star variables across the States will be explored (Fig.1, 2, 3). This will help us to explore the variations and similarities among the dialysis facilities with the same or different States.

Hypothesis:

Some of the health data such as readmission, hospitalization, standard infection rates are believed to show a strong positive correlation to the Mortality_rate. Moreover, patients with arteriovenous_fistulae have a lower risk of infection than patients with catheters, and so does their Mortality_rate. Therefore, exploring mortality rate across each States is vital, which is believed to show some trend. The five_star variable is also expected to display some variations.

Null Hypothesis,

- Ho: There is no correlation between the pairs of dialysis facilities across the States given the health data of patients in the dialysis facilities across the States.

Alternative Hypothesis,

- Ha: There is correlation between the pairs of the dialysis facilities

This hypothesis will be tested in this section.

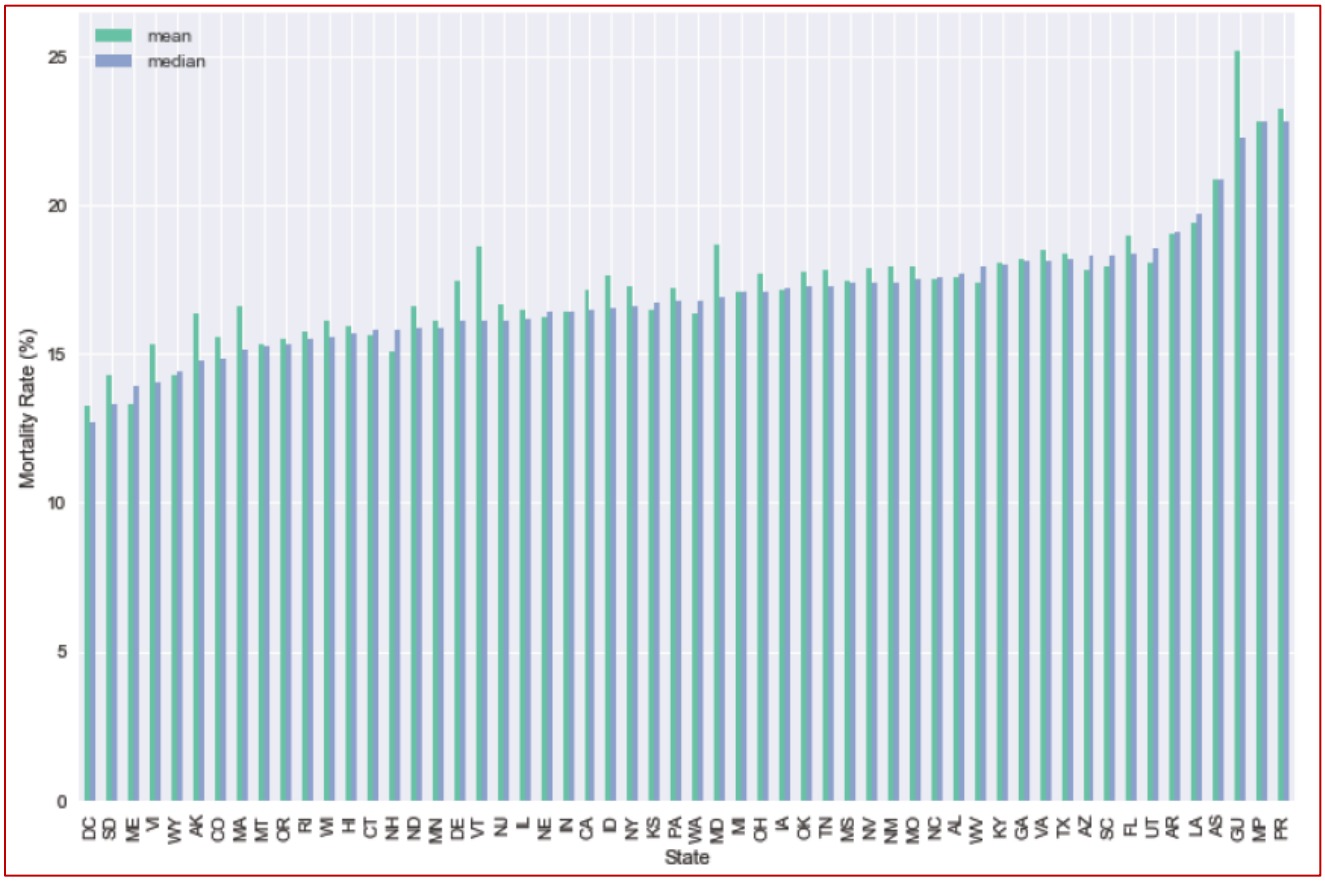

Fig 1: Mean and median distribution of mortality rate plotted against each States

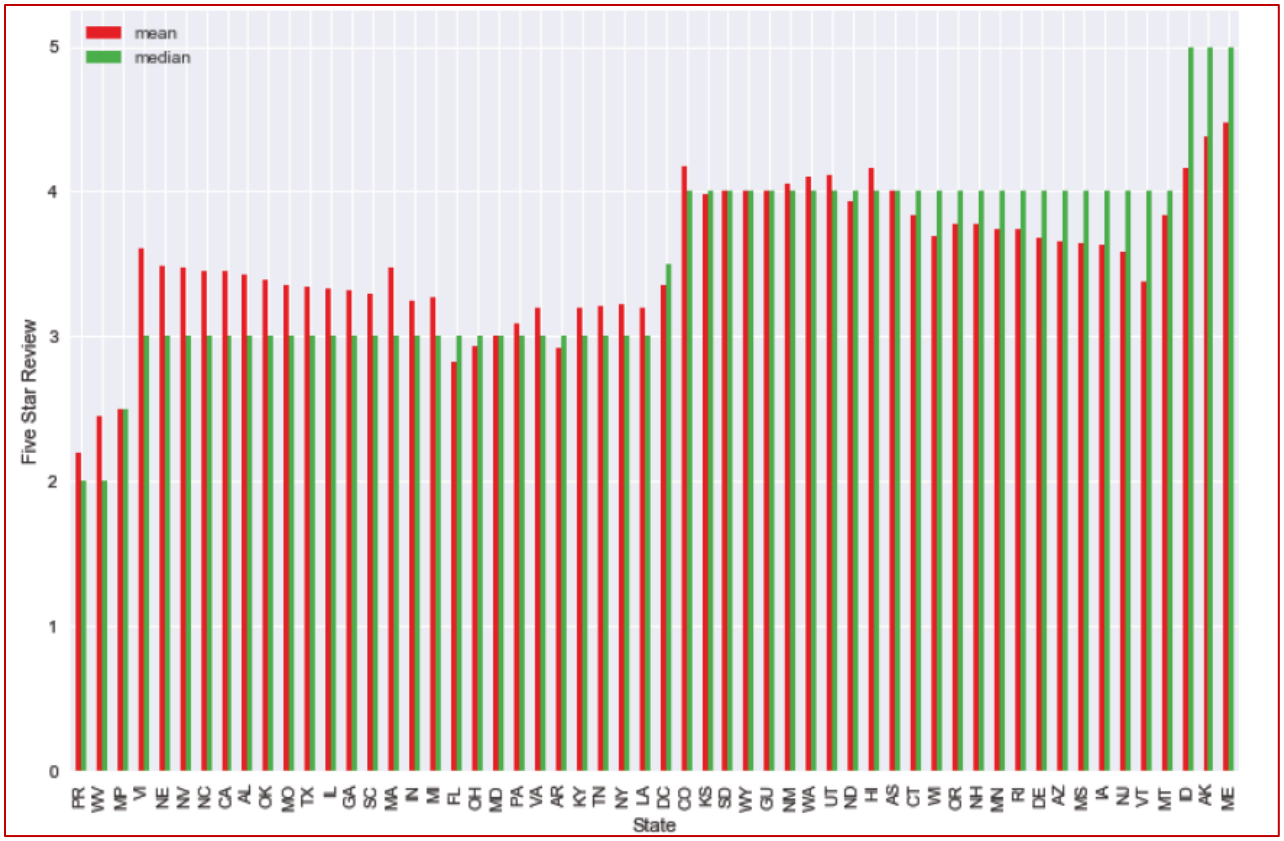

Fig 2: Mean and median distribution of five star reviews plotted against each States

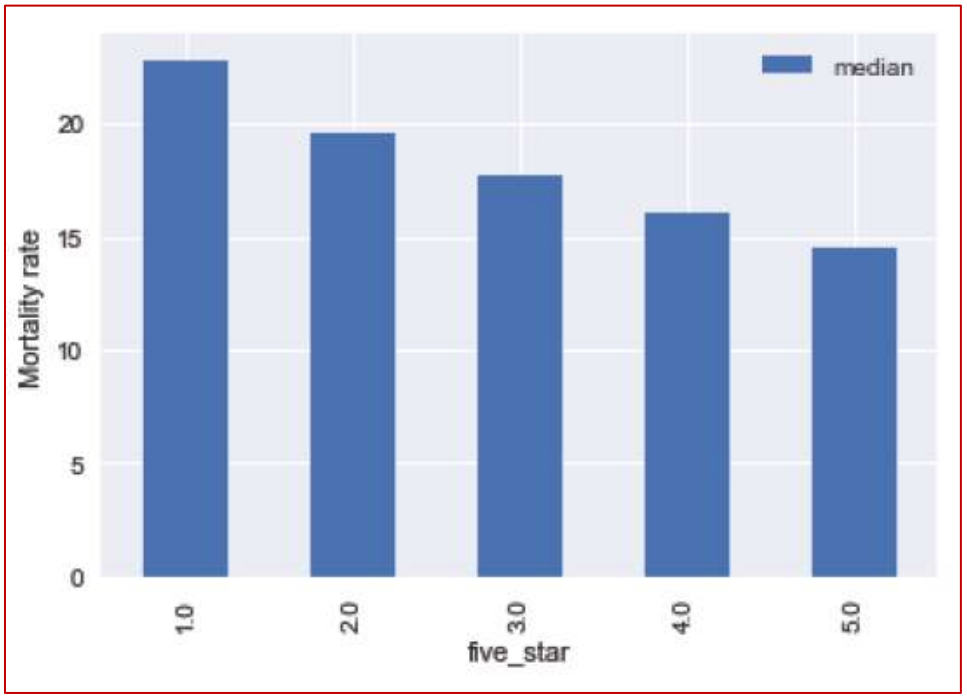

Fig 3: Mortality rate against Five Star

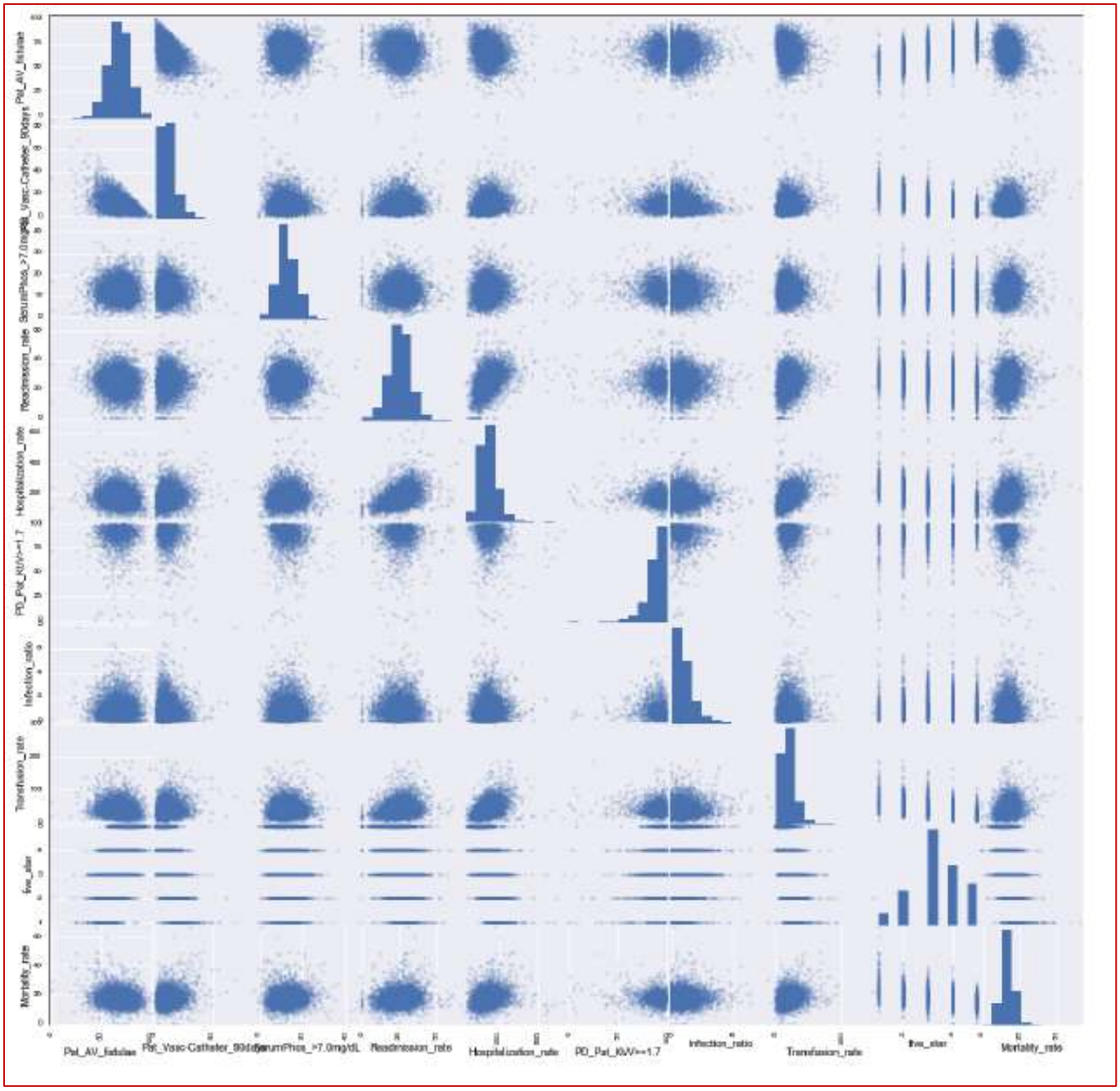

Fig 4: plot of mortality rate and five star variables against the other variable which show good trend

The above plots show that the trend of the Mortality_rate or five_star varies from one State to another. Both appeared to classify the States into four groups (Fig.1 and 2). When both are correlated with the other variables, it can be observed some trends which vary from insignificant to very significant. For example, the mortality rate indicates an increasing trend with hospitalization, readmission and transfusion rates, while infection ratio seems to show not a strong trend with the mortality rate (Fig.4).

However, the other features did not seem to show any significant trend against the mortality rate. The relationship between Mortality-rate and five_star is appeared to be negative (Fig.3). The variations observed across the States can indicate that there could be significant pattern in the facilities given the health data of the patients.

D. Model Fitting

Clustering algorithms were utilized for this project. Each algorithm implements fit method to learn the clusters on train data and returns an array of integer labels corresponding to the different clusters.

K-Means

This algorithm is simple and cluster data by trying to separate samples in n groups of equal variance (minimizing within-cluster sum-of-squares). Number of clusters has to be specified for this algorithm. A range of number of clusters (2 to 10) was used and their results were evaluated using the silhouette coefficient. From the EDA of the dataset, it can be observed that the dataset can be categorized into four groups. Therefore, number of clusters of 4 was selected, though the silhouette score were less than that of cluster-2 and 3.

Agglomerative clustering

Agglomerative clustering algorithms build nested clusters by merging or splitting them successively. This algorithm also takes an input for number of clusters. The same range in number of clusters was utilized as of the K-Means. The result of using each number in the algorithm was evaluated using silhouette coefficient.

Affinity propagation

Affinity propagation builds clusters by sending messages between pairs of samples until convergence. The interesting part of this algorithm is that it chooses the number of clusters based on the data provided. It does not need to specify the number of clusters.

DBSCAN

The DBSCAN is a density based algorithm and sees clusters as areas of high density separated by areas of low density. It extracts the dense clusters and leaves sparse background classified as 'noise'.

E. Model Evaluation

The dataset was split into 50% of training and 50% testing dataset. The goal to split into two dataset is to observe if their result can be similar and evaluate the model performance. The above four clustering algorithms are fit on the training data, and the performance of each model was evaluated by the silhouette score.

The model with better performance was selected to be utilized on the left out 50% of the dataset (testing data).

Results

Selected clustering models were utilized to achieve the objective of this project. These models were fit on the train dataset (50% of the whole dataset). Some of them such as K-Means and Agglomerative clustering require number of clusters to be specified. However, the Affinity Propagation and DBSCAN do not have an input for number of clusters. The silhouette coefficient for each model seemed to have different values. While K-Means and Agglomerative clustering were indicating a reasonable to strong structure/pattern in the train dataset, the Affinity Propagation and DBSCAN models were not effective in finding a substantial structure in the train dataset. As the result of K-Means and Agglomerative clustering was similar, K-Means model were selected for further analysis because of its simplicity.

The Silhouette score for different number of clusters used in the K-Means model shows values ranging from 0.85 to 0.61. These values can indicate that there is a reasonable to strong structure/pattern in the dataset. Though the number of cluster of 2 yielded the highest silhouette score (0.88), the number of cluster of 4 with a value of 0.70 was selected. The choice for that number was dependent on the exploratory data analysis carried out between the mortality rate/five star reviews and the States. On both cases, the States appeared to be categorized into four groups based on those data fields.

As part of model evaluation, the K-Means was fit on the 50 % left out dataset (testing dataset). Similar to the training dataset, the number of cluster of 4 with the silhouette score of 0.72 were estimated. The observations in each cluster were further analyzed based on their State level. From the clustering analysis, 5248, 1007, 384 and 171 observations (dialysis facilities) were grouped into cluster-0, cluster-3, cluster-2 and cluster-1 respectively (Fig.5 and 6). Table-1 also shows the number of dialysis facilities in each State clustered in each group.

Cluster - 0:

About 77 % of the dialysis facilities are clustered in this category. Each State contributed some of its facilities in this cluster.

Cluster - 1:

Only facilities from California were included into cluster-1, and it holds 3% of the dialysis facilities.

Cluster - 2:

About 6% of the facilities are part of this cluster. Cluster-2 only holds some facilities from California, Illinois, Texas and Arizona.

Cluster - 3:

This is the second cluster containing about 14% of the observations. In cluster-3, some dialysis facilities from California, Texas, Florida, Ohio, Pennsylvania, Minnesota, New York, Michigan, Washington, Utah, Virginia, Hawaii, and Montana are present.

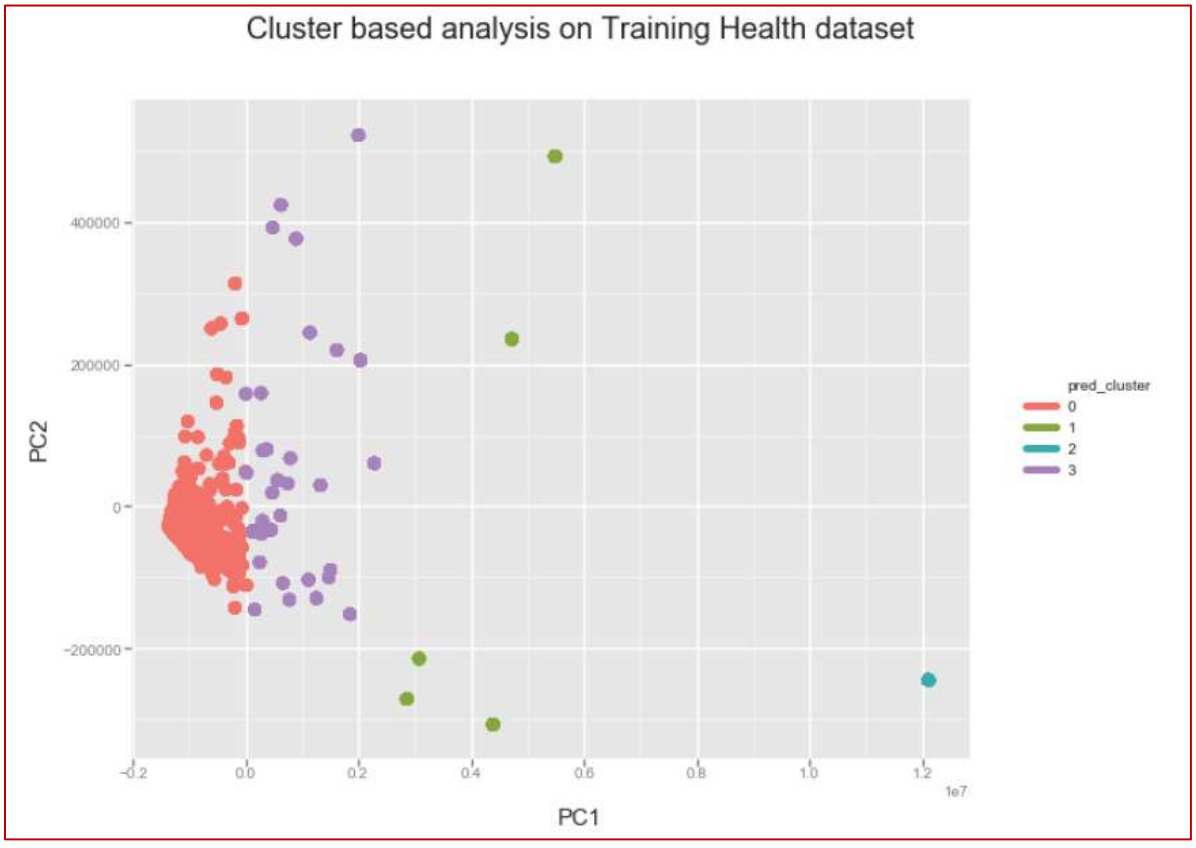

Fig 5: Plot principle component-1 and principle component-2 explained by predicted clusters in the training dataset

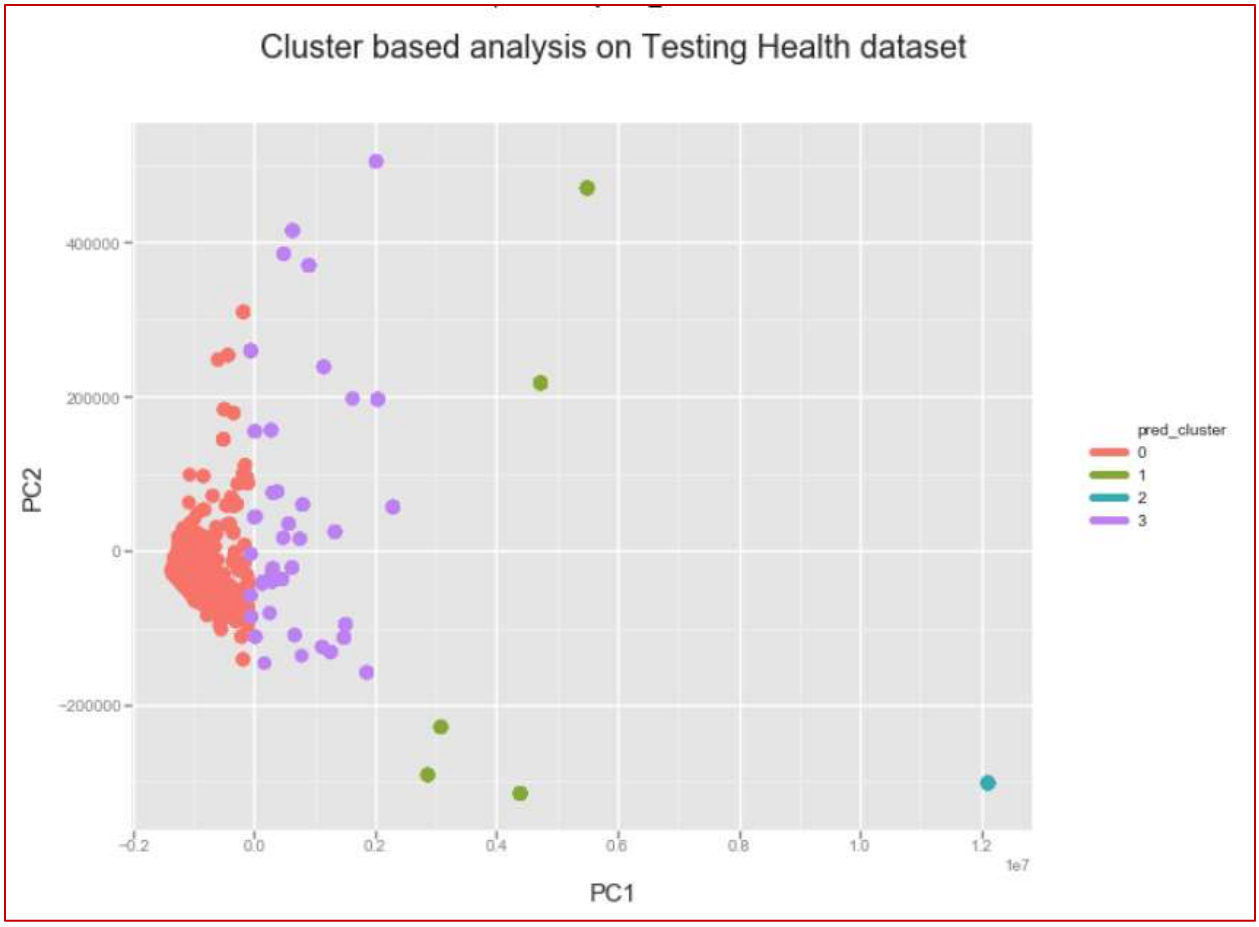

Fig 6: Plot principle component-1 and principle component-2 explained by predicted clusters in the testing dataset

Table 1: The number of dialysis facilities in each State clustered in each group from training dataset

	Cluster_0	State	count		Cluster_1	State	count		Cluster_2	State	count		Cluster_3	State	count
14	TRUE	GA	281	56	TRUE	CA	139	56	TRUE	IL	96	56	TRUE	FL	155
15	TRUE	TX	260					57	TRUE	TX	94	57	TRUE	CA	141
16	TRUE	FL	199					58	TRUE	CA	64	58	TRUE	TX	134
17	TRUE	OH	197					59	TRUE	AZ	58	59	TRUE	NY	111
18	TRUE	PA	162									60	TRUE	PA	71
19	TRUE	NC	159									61	TRUE	MI	55
20	TRUE	TN	147									62	TRUE	OH	53
21	TRUE	CA	145									63	TRUE	MN	19
22	TRUE	IL	140									64	TRUE	WA	14
23	TRUE	VA	140									65	TRUE	MA	13
24	TRUE	IN	137									66	TRUE	UT	10
25	TRUE	AL	136									67	TRUE	VA	9
26	TRUE	MD	136												
27	TRUE	LA	130												
28	TRUE	MO	129												
29	TRUE	NJ	129												
30	TRUE	SC	117												
31	TRUE	MI	115												
32	TRUE	NY	99												
33	TRUE	KY	93												
34	TRUE	WI	87												
35	TRUE	MS	78												
36	TRUE	MN	75												
37	TRUE	OK	65												
38	TRUE	CO	61												
39	TRUE	WA	60												
40	TRUE	AR	58												
41	TRUE	MA	57												
42	TRUE	OR	54												
43	TRUE	IA	53												
44	TRUE	KS	46												
45	TRUE	NV	40												
46	TRUE	AZ	37												
47	TRUE	CT	36												
48	TRUE	PR	36												
49	TRUE	NM	34												
50	TRUE	WV	34												
51	TRUE	NE	28												
52	TRUE	ID	26												
53	TRUE	UT	22												
54	TRUE	DE	20												
55	TRUE	DC	19												
56	TRUE	HI	19												
57	TRUE	SD	19												
58	TRUE	ME	15												
59	TRUE	ND	13												
60	TRUE	NH	13												
61	TRUE	RI	11												
62	TRUE	MT	10												
63	TRUE	AK	9												
64	TRUE	VT	7												
65	TRUE	WY	7												
66	TRUE	VI	5												
67	TRUE	GU	4												
68	TRUE	MP	2												
69	TRUE	AS	1												

Limitations

The result of this project does not answer why the dialysis facilities from different States clustered into one or the other. For example, the dialysis facilities from California are clustered into different groups, though they are from the same state. What made them differ or similar from each other? The clustering models do not answer these questions which are very important. The models simply examine for any significant structure/pattern in the dialysis facilities dataset which helps to learn more about the data and for further analysis.

Further Research

The results are a preliminary product to our virtual client. Further research is needed to hand over a final product which will have concrete and labeled variables. The dataset in each cluster should be analyzed separately using supervised learning algorithms.

Client Recommendations

As the model identified some structure in the dataset, some recommendations can be taken from the carried out analysis.

- The Federal government should investigate more the similarities and differences among the facilities clustered in one to the others respectively.
- The dialysis facilities in one State, but categorized into different clusters, should learn from each other.
- As this project is a preliminary analysis report, further research is needed based on the result of this project. The baseline for the next research should be the preliminary results of this project.
- The dialysis facilities should continue gathering more health data of the patients.

Appendix

Data fields (in order of appearance):

VARIABLE	DESCRIPTION
SUMLEV	Geographic Summary Level
STATE	State FIPS code
COUNTY	County FIPS code
STNAME	State Name
CTYNAME	County Name
YEAR	Year
AGEGRP	Age group
TOT_POP	Total population
TOT_MALE	Total male population
TOT_FEMALE	Total female population
WA_MALE	White alone male population
WA_FEMALE	White alone female population
BA_MALE	Black or African American alone male population
BA_FEMALE	Black or African American alone female population
IA_MALE	American Indian and Alaska Native alone male population
IA_FEMALE	American Indian and Alaska Native alone female population
AA_MALE	Asian alone male population
AA_FEMALE	Asian alone female population
NA_MALE	Native Hawaiian and Other Pacific Islander alone male population
NA_FEMALE	Native Hawaiian and Other Pacific Islander alone female population
TOM_MALE	Two or More Races male population
TOM_FEMALE	Two or More Races female population
WAC_MALE	White alone or in combination male population
WAC_FEMALE	White alone or in combination female population
BAC_MALE	Black or African American alone or in combination male population
BAC_FEMALE	Black or African American alone or in combination female population
IAC_MALE	American Indian and Alaska Native alone or in combination male population
IAC_FEMALE	American Indian and Alaska Native alone or in combination female population

AAC_MALE	Asian alone or in combination male population
AAC_FEMALE	Asian alone or in combination female population
NAC_MALE	Native Hawaiian and Other Pacific Islander alone or in combination male population
NAC_FEMALE	Native Hawaiian and Other Pacific Islander alone or in combination female population
NH_MALE	Not Hispanic male population
NH_FEMALE	Not Hispanic female population
NHWA_MALE	Not Hispanic, White alone male population
NHWA_FEMALE	Not Hispanic, White alone female population
NHBA_MALE	Not Hispanic, Black or African American alone male population
NHBA_FEMALE	Not Hispanic, Black or African American alone female population
NHIA_MALE	Not Hispanic, American Indian and Alaska Native alone male population
NHIA_FEMALE	Not Hispanic, American Indian and Alaska Native alone female population
NHAA_MALE	Not Hispanic, Asian alone male population
NHAA_FEMALE	Not Hispanic, Asian alone female population
NHNA_MALE	Not Hispanic, Native Hawaiian and Other Pacific Islander alone male population
NHNA_FEMALE	Not Hispanic, Native Hawaiian and Other Pacific Islander alone female population
NHTOM_MALE	Not Hispanic, Two or More Races male population
NHTOM_FEMALE	Not Hispanic, Two or More Races female population
NHWAC_MALE	Not Hispanic, White alone or in combination male population
NHWAC_FEMALE	Not Hispanic, White alone or in combination female population
NHBAC_MALE	Not Hispanic, Black or African American alone or in combination male population
NHBAC_FEMALE	Not Hispanic, Black or African American alone or in combination female population
NHIAC_MALE	Not Hispanic, American Indian and Alaska Native alone or in combination male population
NHIAC_FEMALE	Not Hispanic, American Indian and Alaska Native alone or in combination female population
NHAAC_MALE	Not Hispanic, Asian alone or in combination male population
NHAAC_FEMALE	Not Hispanic, Asian alone or in combination female population
NHNAC_MALE	Not Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination male population
NHNAC_FEMALE	Not Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination female population
H_MALE	Hispanic male population
H_FEMALE	Hispanic female population
HWA_MALE	Hispanic, White alone male population
HWA_FEMALE	Hispanic, White alone female population
HBA_MALE	Hispanic, Black or African American alone male population
HBA_FEMALE	Hispanic, Black or African American alone female population

HIA_MALE	Hispanic, American Indian and Alaska Native alone male population
HIA_FEMALE	Hispanic, American Indian and Alaska Native alone female population
HAA_MALE	Hispanic, Asian alone male population
HAA_FEMALE	Hispanic, Asian alone female population
HNA_MALE	Hispanic, Native Hawaiian and Other Pacific Islander alone male population
HNA_FEMALE	Hispanic, Native Hawaiian and Other Pacific Islander alone female population
HTOM_MALE	Hispanic, Two or More Races male population
HTOM_FEMALE	Hispanic, Two or More Races female population
HWAC_MALE	Hispanic, White alone or in combination male population
HWAC_FEMALE	Hispanic, White alone or in combination female population
HBAC_MALE	Hispanic, Black or African American alone or in combination male population
HBAC_FEMALE	Hispanic, Black or African American alone or in combination female population
HIAC_MALE	Hispanic, American Indian and Alaska Native alone or in combination male population
HIAC_FEMALE	Hispanic, American Indian and Alaska Native alone or in combination female population
HAAC_MALE	Hispanic, Asian alone or in combination male population
HAAC_FEMALE	Hispanic, Asian alone or in combination female population
HNAC_MALE	Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination male population
HNAC_FEMALE	Hispanic, Native Hawaiian and Other Pacific Islander alone or in combination female population

The key for SUMLEV is as follows:

050 = County and/or Statistical Equivalent

The key for the YEAR variable is as follows:

1 = 4/1/2010 Census population
2 = 4/1/2010 population estimates base
3 = 7/1/2010 population estimate
4 = 7/1/2011 population estimate
5 = 7/1/2012 population estimate
6 = 7/1/2013 population estimate
7 = 7/1/2014 population estimate
8 = 7/1/2015 population estimate
9 = 7/1/2016 population estimate

The key for AGEGRP is as follows:

0 = Total
1 = Age 0 to 4 years
2 = Age 5 to 9 years

3 = Age 10 to 14 years
4 = Age 15 to 19 years
5 = Age 20 to 24 years
6 = Age 25 to 29 years
7 = Age 30 to 34 years
8 = Age 35 to 39 years
9 = Age 40 to 44 years
10 = Age 45 to 49 years
11 = Age 50 to 54 years
12 = Age 55 to 59 years
13 = Age 60 to 64 years
14 = Age 65 to 69 years
15 = Age 70 to 74 years
16 = Age 75 to 79 years
17 = Age 80 to 84 years
18 = Age 85 years or older

Note: "In combination" means in combination with one or more other races. The sum of the five race groups adds to more than the total population because individuals may report more than one race. The estimates are based on the 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions. Hispanic origin is considered an ethnicity, not a race. Hispanics may be of any race. Responses of "Some Other Race" from the 2010 Census are modified. This results in differences between the population for specific race categories shown for the 2010 Census population in this file versus those in the original 2010 Census data. For more information, see https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/modified-race-summary-file-method/mrsf2010.pdf. For population estimates methodology statements, see http://www.census.gov/programs-surveys/popest/technical-documentation/methodology.html.

The 6,222 people in Bedford city, Virginia, which was an independent city as of the 2010 Census, are not included in the April 1, 2010 Census enumerated population presented in the county estimates. In July 2013, the legal status of Bedford changed from a city to a town and it became dependent within (or part of) Bedford County, Virginia. This population of Bedford town is now included in the April 1, 2010 estimates base and all July 1 estimates for Bedford County. Because it is no longer an independent city, Bedford town is not listed in this table. As a result, the sum of the April 1, 2010 census values for Virginia counties and independent cities does not equal the 2010 Census count for Virginia, and the sum of April 1, 2010 census values for all counties and independent cities in the United States does not equal the 2010 Census count for the United States. Substantial geographic changes to counties can be found on the Census Bureau website at http://www.census.gov/geo/reference/county-changes.html.