Results and methods:

As mentioned in the data analysis section , i have chosen to cluster patients based on their demographic and medical data , prior to the treatment they received in the hospital.

<u>Method 1 - K-means :</u>

For this i used the following features:

- Race
- Gender
- Age
- Admission_type_id
- Admission_source_id
- Payer_code
- Number_outpatient
- Number_emergency
- number_inpatient

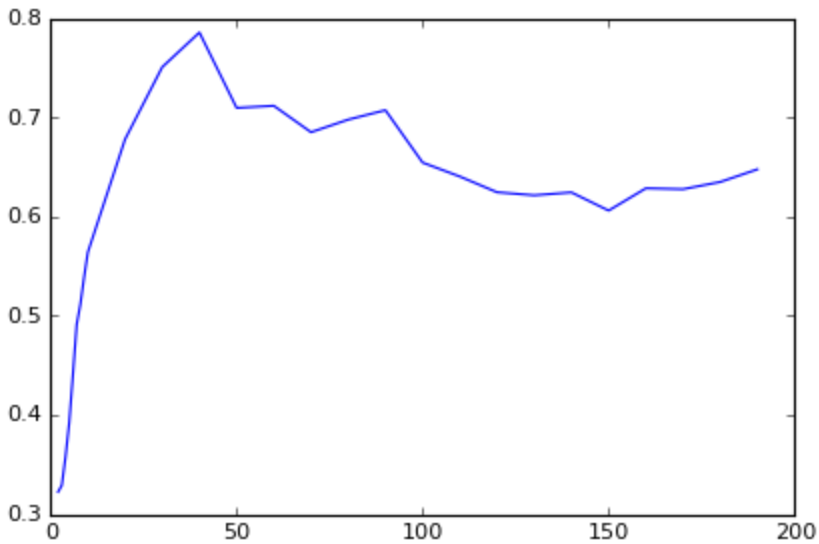
The data preparation was done in 2 ways:

1. For the nominal features use one hot encoding , and normalize each numerical feature to [0-1] (including age)
2. For the nominal features use one hot encoding , and perform dimension reduction for the numerical features using PCA.(except age which will be handled like in 1).

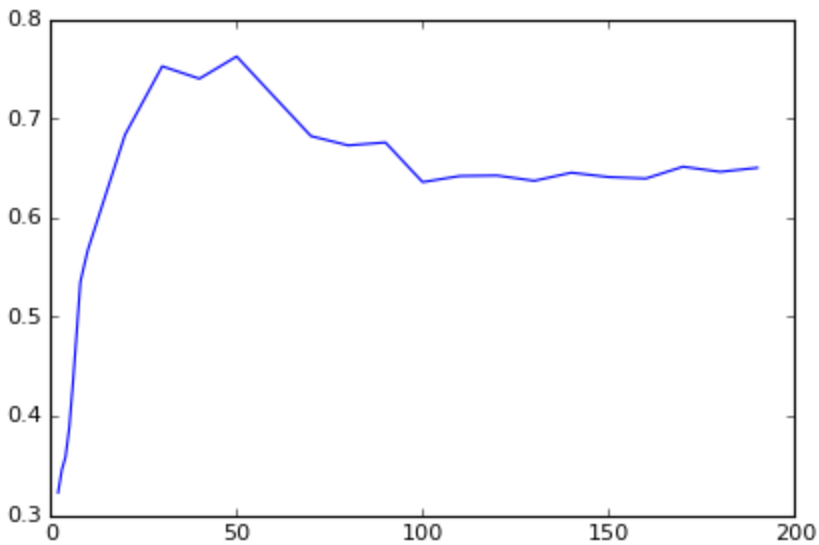
Then apply K-means algorithm , and find the best K using silhouette score.

Results :

This graph displays the average silhouette score for each k :



And this one after PCA :



A reasonable choice would be k = 30 (in both cases).

After examining the behaviour of other features (not used for clustering) , there are some interesting findings.

For example , cluster 7 which is mostly **Caucasian Females** , has a much smaller probability of **change ,** then cluster 3 which is mostly **Caucasian Males.**

Or cluster 26 which is mostly **African American** and **Hispanic** , has a higher probability to be **readmitted**.

I did not find any change in **time in the hospital** distribution , all clusters are with mean~4.2 days and std ~3 days

[In order to confirm the results were not random , i randomly assigned the dataset into 30 clusters , and checked to see if there are similar changes in probability , but the random clusters were all very similar to the general probability distribution.]

Method 2 - similarity/distance matrix :

For the following features :

- Race
- Gender
- Age
- Admission_type_id
- Admission_source_id
- Payer_code
- Number_outpatient
- Number_emergency
- number_inpatient

There will be a similarity function, that will receive 2 samples as input , and return the similarity score.

The similarity function will operate as follows :

1. For numeric features , it will be (1 - distance)
2. For nominal features :
    a. 1 if the same and zero if different
    b. If the same divide : 1 / probability of value (i.e : 2 Hispanic get a higher score than 2 Caucasian)

And then operate spectral clustering examine the results , and perform additional spectral clustering on the clusters already generated.

[Due to computational limitations , it will only be performed on a sample of 10000] .

Results : When creating a small number of clusters , i could not find what distinguishes each cluster , when i created 30 clusters (like the number of clusters) some clusters had a higher readmission rate , but it could be due to the fact that when performing k-means most clusters were pretty much

the same size , spectral clustering created a higher variance in cluster sizes.