Hospital Occupancy Prediction: Clustering

Justin Ledford
Eric Smith

October, 2018

Abstract

In this report we look at data from insurance health claims, and attempt to find meaningful clusters within the data. Although the data seems to have little clustering tendency, we were able to find some clusterings separated by demographic data and overall health. Using PCA and k-means clustering we were also able to develop a basic prediction model for number of days spent in the hospital by a person in the next year.

Contents

1 Data Preparation

In order to perform clustering by a member's claims, we have flattened the claim's into summary statistics, for example age, sex, pay delay, and a member's maximum Charlson Index. For sex, which is a binary value, we used a 0 for female and 1 for male. We also used the number of lab visits per year, and number of unique prescriptions filled. The number of unique prescriptions, or drug count, is in the range 1-7+. We chose these features as we thought they are most indicative of the number of days spent in the hospital the next year.

For any feature that contained more than one number, like 7+ for drug count, we replaced it with that number, because these values were chosen as they are the 95th percentile for that feature. We then standardized these features before feeding into our clustering algorithms. Because all variables are ratio we use Euclidean distance, the distance measurement used in k-means clustering.

1.1 Categorical Feature Counts

In order to create more features and use the categorical features for clustering, we have created columns containing the counts of unique values for Primary Condition Group, Specialty and Place of Service. For example, given a member, some of these features would be number of claims with a metastatic cancer, number of claims where the member visited an diagnostic imaging lab, and number of claims where the member visited an internal medicine specialist. By doing this we have created 50+ new features, so we will be using Principal Component Analysis to reduce the number of dimensions.

1.2 Principal Component Analysis

Principal Component Analysis is a technique to convert a set of correlated variables into a set of uncorrelated variables called principal components. This is very useful for reducing the number of dimensions when using clustering algorithms that can suffer from the curse of dimensionality.

After performing PCA on the features described above from a random sample of 8000 data points, we can look at the the variance explained by each component in Figure 1 and the cumulative variance in Figure 2. Component 1 and 2 make up about 10% variance, while the top 25 components make up about 50% variance. We have clustered using both the top 2 and top 25 components.

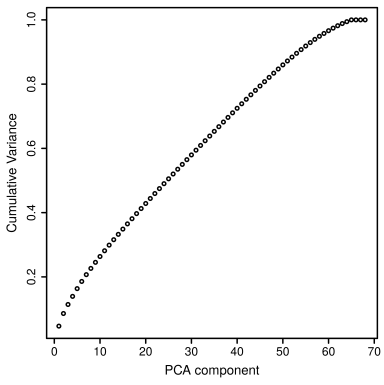


Figure 1: Principal component's cumulative variance

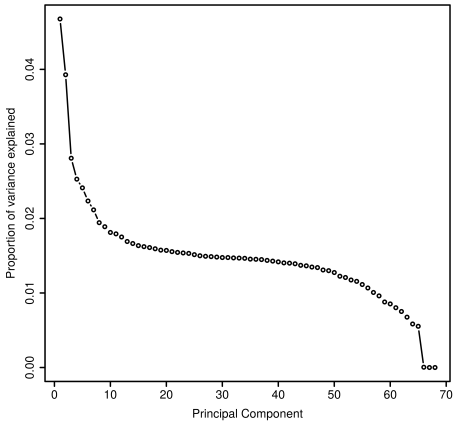
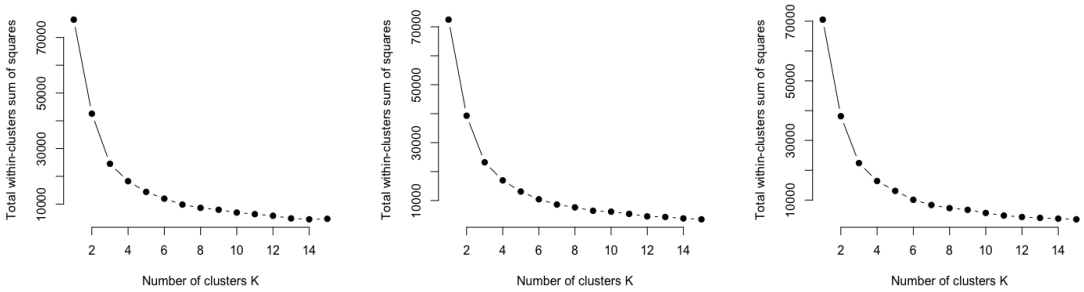


Figure 2: Variance explained by principal components

2 Modeling

2.1 Clustering on Initial Features

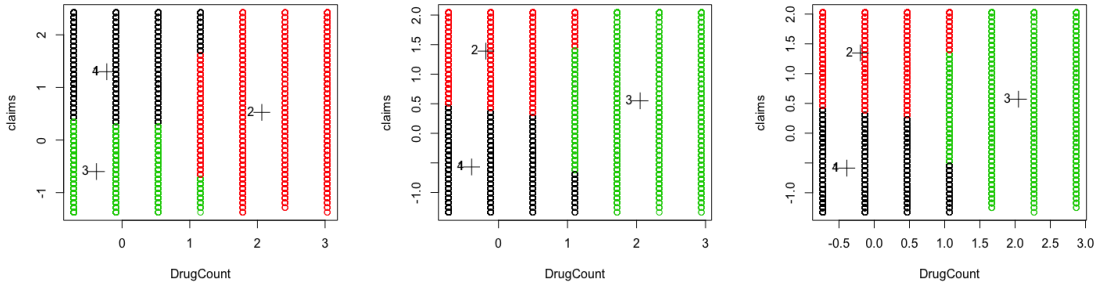
2.1.1 DrugCount and Claims



(a) Year 1 (b) Year 2 (c) Year 3

Figure 3: Elbow method for DrugCount and Claims Years 1-3.

We started with basic, two-variable clusters that we thought would produce the most obvious clusterings. For example, DrugCount was plotted against number of claims for each unique member for Years 1, 2, and 3. We used the elbow method (Figure 3) to determine the number of ideal clusters, which happened to be three for each year. The clusterings (shown in Figure 4) themselves look very similar and are all broken up three categories: less drugs and less claims, less drugs and more claims, and more drugs and more claims. The three clusters do not have distinct separations, but this is likely because of how dense this data is, as well as drug count only have 7 possible values. The clusters are very similar across all 3 years, which is expected, given that a year should not have any bias on the claim data.

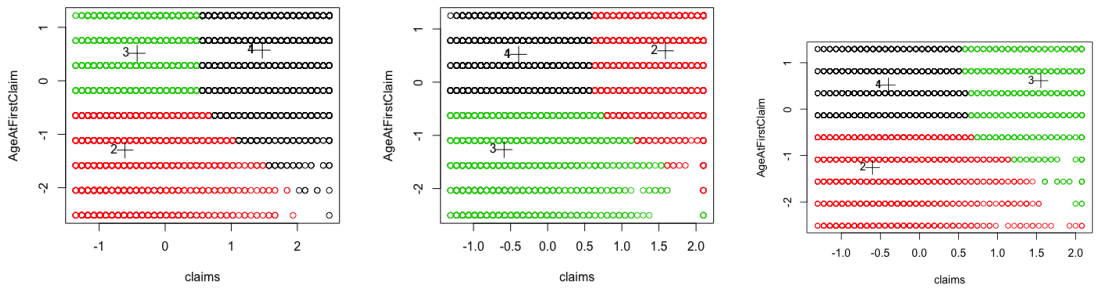


(a) Year 1 (b) Year 2 (c) Year 3

Figure 4: K-means clusters for DrugCount and Claims Years 1-3.

2.1.2 AgeAtFirstClaim and Claims

We then clustered with AgeAtFirstClaim and the number of claims using the same methods. This yielded similar results (shown in Figure 5) such that three non-distinct clusters formed: those with a high age at first claim and a high number of claims, those with a high age of first claim but a low number of claims, and those patients who had a low age of first claim who had the widest range of the number of claims. This is likely because there are more members with higher ages, and those can be split up by overall health.

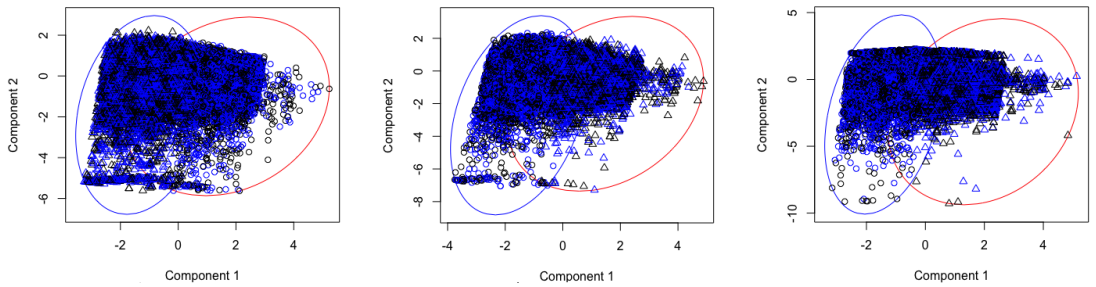


(a) Year 1 (b) Year 2 (c) Year 3

Figure 5: K-means clusters for AgeAtFirstClaim and Claims Years 1-3.

2.1.3 Additional Features

We then increased the number of attributes to cluster to try to get a better understanding of which components made up the most variance. Using the features AgeAtFirstClaim, PayDelay, CharlsonIndex, and the number of claims, we plotted each member again by year (Figure 7). When plotted by the two principal components, we see two clusters form rather than three. For all three years, the two principal components make up for about 69% of the point variance.

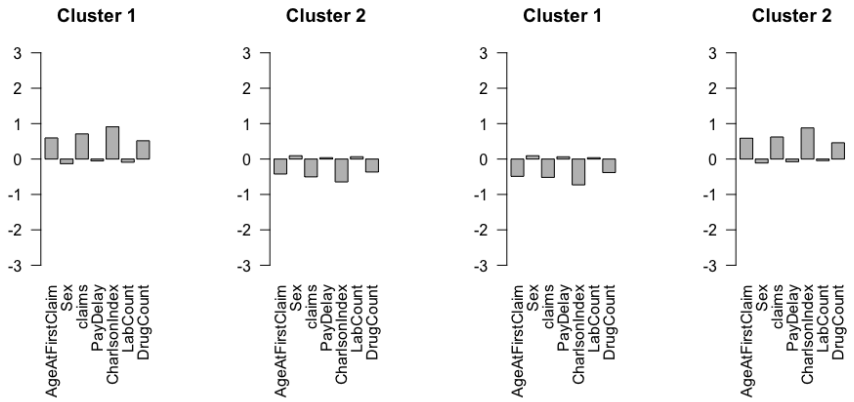


(a) Year 1 (b) Year 2 (c) Year 3

Figure 6: K-means clusters for additional features Years 1-3.

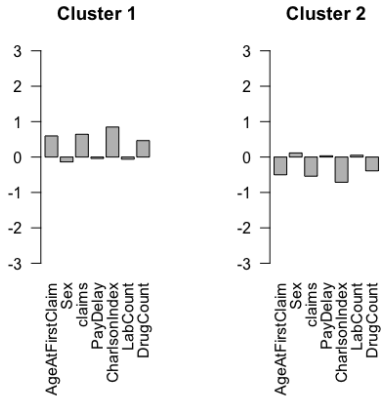
2.1.4 All Selected Features

Building on top of the four features, we clustered all members by each year like we did before. This time we included all of our selected member attributes: AgeAtFirstClaim, Sex, Claims, PayDelay, CharlsonIndex, LabCount, and DrugCount. Using the same elbow method as before, we clustered using k-means where k is equal to two. When analyzing the two clusters for all three years, we discovered the same pattern (shown in Figure 8a). One cluster had a high average AgeAtFirstClaim, CharlsonIndex, DrugCount, and number of claims while the other cluster had low averages of these variables. This most likely represents older people have more health problems, so they visit the doctor more, as well as need more medicine.



(a) Year 1

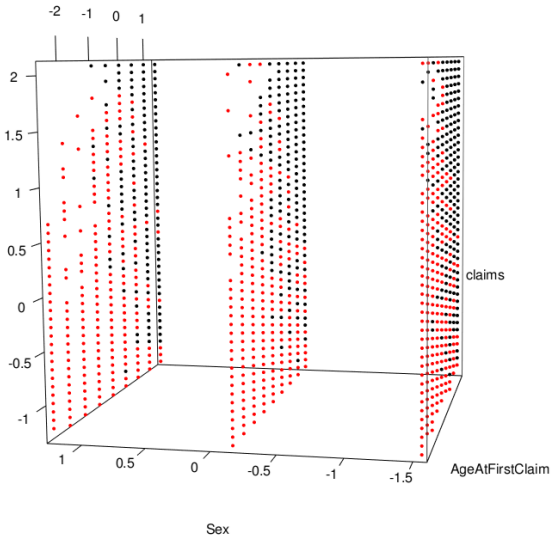
(b) Year 2



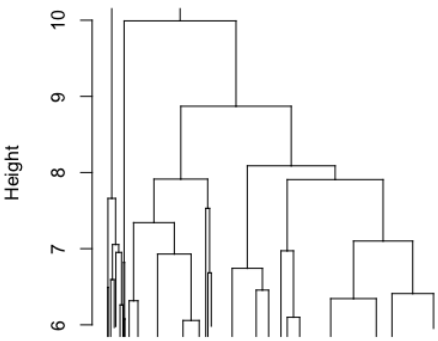
(c) Year 3

Figure 7: Average values for all features Years 1-3.

This k-means clustering was compared to a hierarchical clustering where the number of clusters was also equal to two. This clustering method produced almost exactly the same inter and intra-cluster distances among points (about 1.83 and 0.54, respectively).



(a) K-means Clustering



(b) Hierarchical Clustering

Figure 8: Clustering on all selected attributes.

2.2 Clustering with PCA

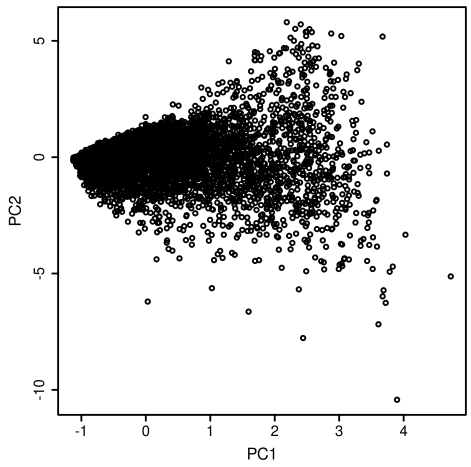


Figure 9: Distribution of points by top 2 PCs

Next we clustered using the principal components created from the count features.

Looking at Figure 9 we can see the distribution of data points has an inconsistent density throughout the plane, so we will choose not to use density based clustering algorithms such as DBSCAN.

2.2.1 k-means clustering

To find the optimal number of clusters for k-means clustering, we have plotted the within cluster sum squares measure for each k from 2-10, to look for the knee. In Figure 10 and 11 we can see that there is not really a knee at all, so it is possible that 2 would be an optimal k.

We also looked at the average silhouette width to get another measure. The average silhouette width is a metric combining the cohesion of a point within its own cluster and separation from points from other clusters. A higher average silhouette width is better, and we found $k = 2$ to have the maximum average silhouette width. In Figure 12 and 13 we can see the maximum is at 2.

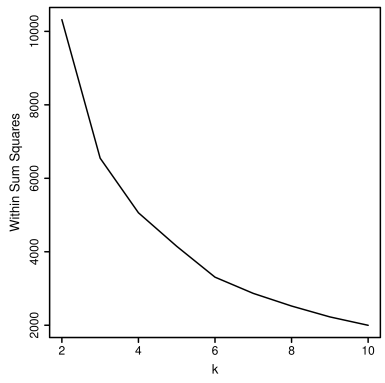


Figure 10: Within Sum Squares (Top 2 PCs)

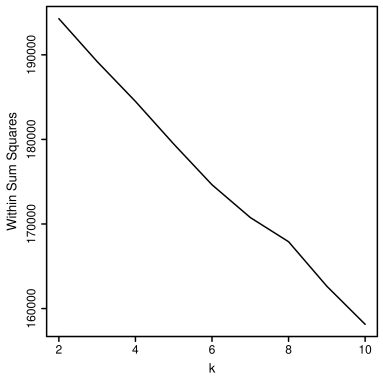


Figure 11: Within Sum Squares (Top 25 PCs)

n	**Cluster 1 Size**	**Cluster 2 Size**
2	2452	5548
25	2540	5460

Table 1: Cluster Sizes with top n principal components

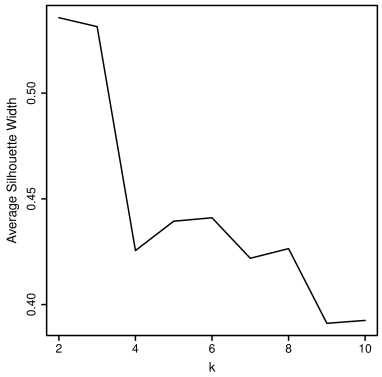


Figure 12: Average Silhouette Width (Top 2 PCs)

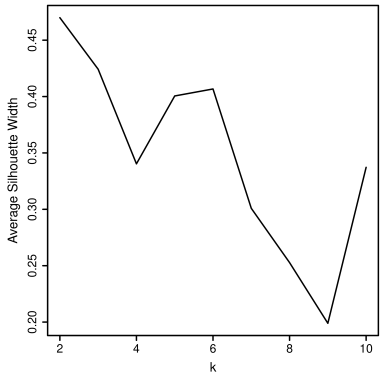


Figure 13: Average Silhouette Width (Top 25 PCs)

With $k = 2$, we visually compared the clusterings of the top 2 and top 25 principal components in Figure 14 and 15. Because the clustering with the top 25 components is in a dimension higher than 2, there are some overlaps in the plot in Figure 15. By plotting this in 3D in Figure 16 we can see how the points are clustered. This also shows that most of the variability within the clusters comes from these 3 components, are this is not much overlap in 3 dimensions.

We can see that there is one cluster made up of the dense points around the center of each component, and another cluster from the points with less density. This could be interpreted as the dense cluster containing points that follow an average of the dataset, and the other cluster with points outside of that average. In Table 1 we can see the cluster sizes are about equal. Because these clusterings are in different dimensions we can't directly compare internal validation measures, but we will compare the two with external validation.

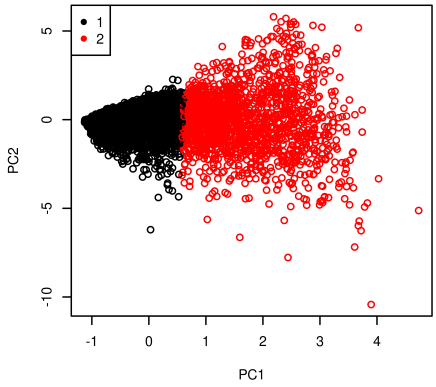


Figure 14: k-means clustering with $k = 2$ (Top 2 PCs)

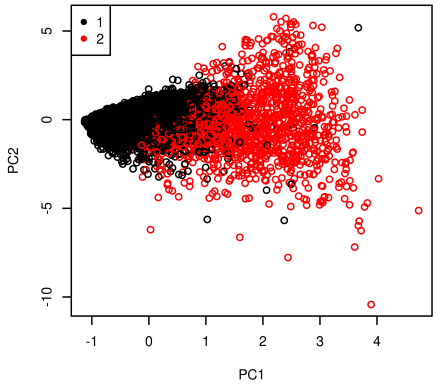


Figure 15: k-means clustering with $k = 2$ (Top 25 PCs)

	k-means	Hierarchial
Within Sum Squares	10,052	15,177
Average Distance Within	1.337	1.708
Average Distance Between	2.282	4.350

Table 2: Internal Validation Statistics on Top 2 Principal Component Clusterings

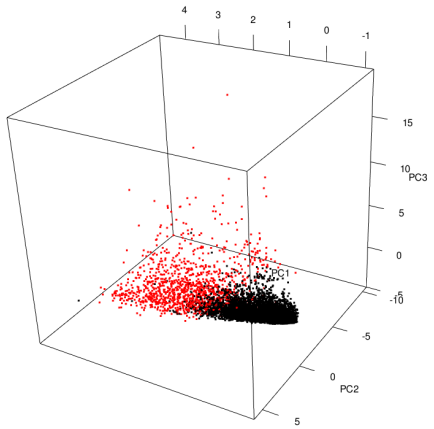


Figure 16: k-means clustering with $k = 2$ (Top 25 PCs)

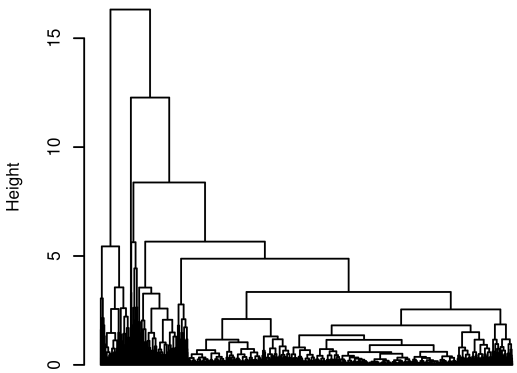


Figure 17: Dendrogram (Top 2 PCs)

By creating a dendrogram from the hierarchial clustering of the top 2 principal components with complete linkage, we can see in Figure 17 that 2 clusters may give us similar results to our k-means clustering, but after plotting in Figure 18 we see that a different area is clustered.

Comparing these 2 clusterings for the top 2 principal components, we can see in Table 2 that the k-means clustering has a smaller distances within the clusters than the hierarchial clustering, but also smaller distances between the clusters. Because the data is so dense and varied, we may prefer a smaller within cluster distance rather than a higher between cluster distance.

Cluster	Top 2 PCs (k-means)	Top 25 PCs (k-means)	Top 2 PCs (Hierarchial)
1	0.0918	0.0962	0.1192
2	0.2287	0.2429	0.1416

Table 3: Mean Days in Hospital by Cluster

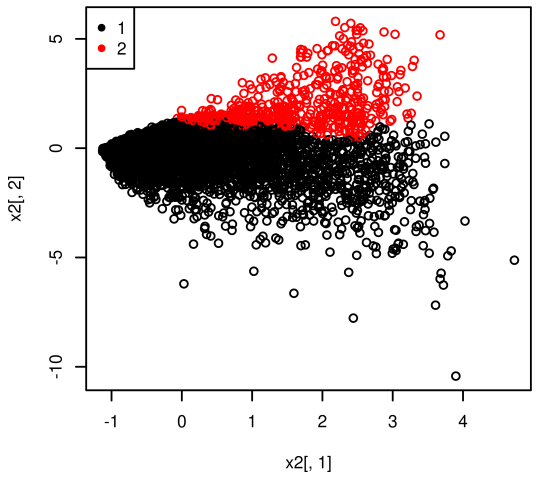


Figure 18: Hierarchial Clustering with $k = 2$ (Top 2 PCs)

To validate the quality of the clusters and the potential to use the clusters, we can compare the mean days in hospitals of each cluster. This can then be used to predict a person's days in hospital by seeing what cluster they belong to. Since there are two clusters, we have created a binary categorical variable for days in hospital, with 0 being 0 days, and 1 being 1+ days. From the sample used for the clusterings above, 12.12% of the rows have 1+ days in the hospital.

Looking at Table 3 we can see the mean of our binary days in hospital variable. We can see that for the k-means clusterings, cluster 1 has below average days in hospital, and cluster 2 has above average days in hospital, while both of the clusters from the hierarchial clustering are closer to the average.

3 Evaluation

Overall this data is very dense and at the same time has a high variability in density. Because of this, clustering may not be the best technique to further analyze this data. And although PCA was able to give us a better look, it is possible that non-linear transformation is needed to provide better insight.

From clustering the initial features, we were able to find that clusters tended to form by the demographic data and overall health of a member.

With the k-means clusterings of principal components from count features, we were able to find 2 clusters, for below average, and above average days in the hospital. We can use these clusterings to do some prediction on the number of days in hospital, and by looking at the amount of people in each cluster, we could then get an estimate of how many more people will be in the hospital the next year.