Clustering Methods

We log transformed the variables (plus 1 for those variables sets containing a value of 0). We ran a correlation/ heat map. There was a signficant correlation 0.94 between log transformed IP days and admissions. The correlation prior to the transformation was 0.74. We examined the dataset and found that the majority of patients fell into the 0-3 admissions range. We think that this, in combination with the removal of outliers, drove the very high correlation between the two variables.

There was a very low correlation between procedures and inpatient days. There were some patients who had 0 inpatient days and many procedures; this might reflect procedures for which the patient is admitted for a day.

Formula for range standardization: (x-min)/(max-min) .

First, we ran Ward’s agglomerative clustering. Based on our interpretation of the dendrogram, we chose to start with three clusters. We ran k-medoids to form 3 clusters. We compared the clusters formed by Ward’s with those formed by k-medoids and found that both methods clustered patients similarly (i.e. patients who belonged to clusters 1, 2, 3 via Ward’s method were, for the most part, placed into the same clusters using k-medoids).