IEMS 308
HW 1 – Clustering
Alex Merryman

The dataset initially contained a number of massive outliers that drastically misrepresented the general nature of the data, so the first step was to clean the data; for each numerical feature, any row containing any numerical value that was more than the feature's mean + 2*standard deviation, or less than the feature's mean – 2*standard deviation was removed from the dataset. This eliminated over 1.8 million datapoints.

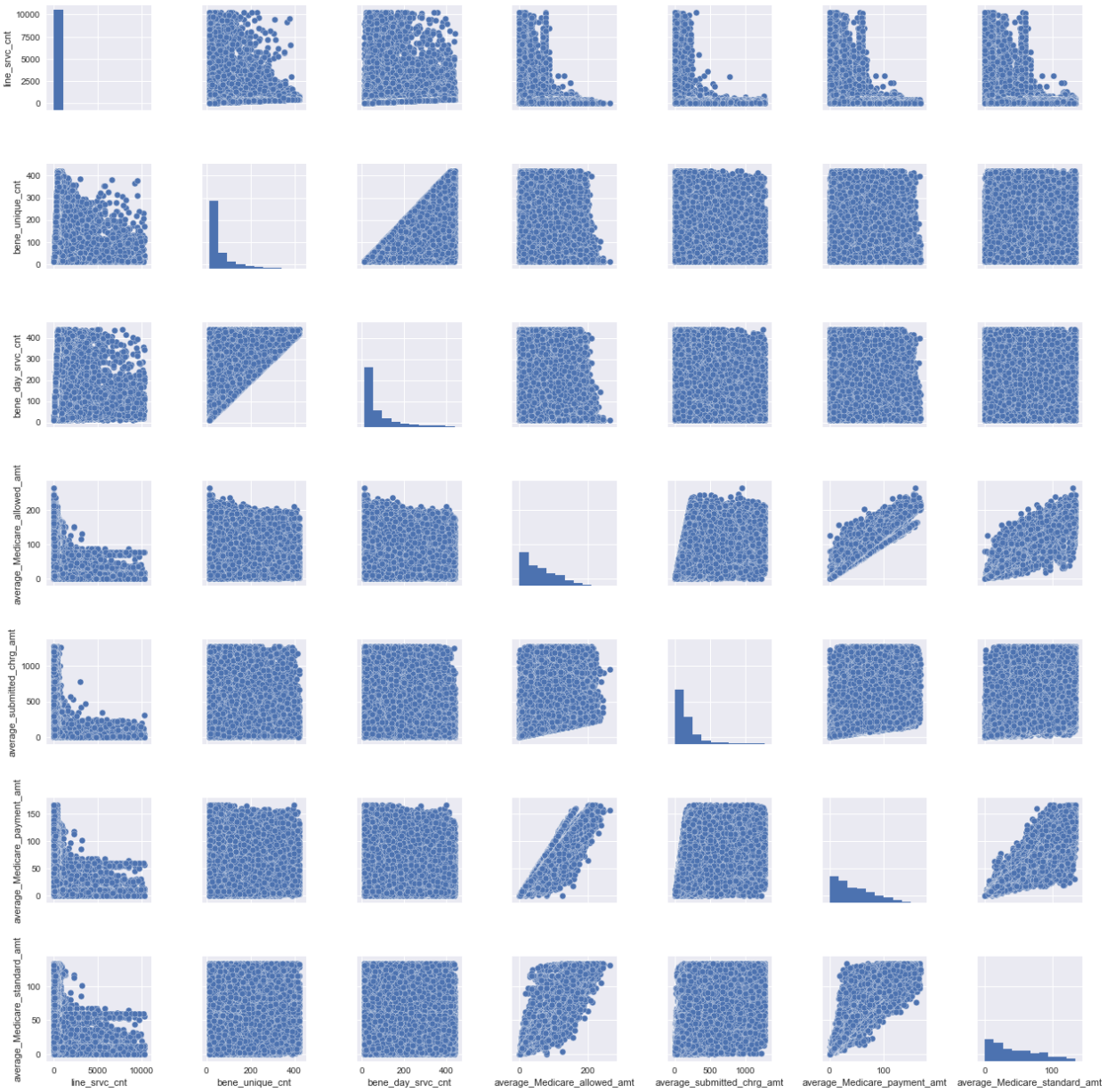
Next, features that were unnecessary, redundant, or simply too granular were removed. For example, nppes_provider_street1, was removed from the dataset because such data is too precise for this application, and is thus not useful. Additionally, categorical variables were one-hot encoded.

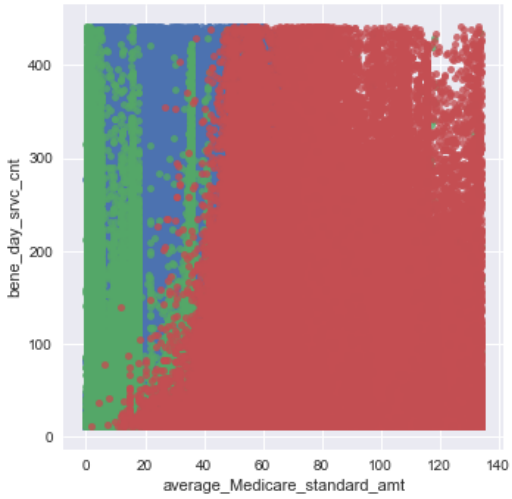
After cleaning the data to a manageable level, it was then possible to analyze it graphically, especially with the chart below showing every feature plotted against every other feature, as well as histograms for each feature.

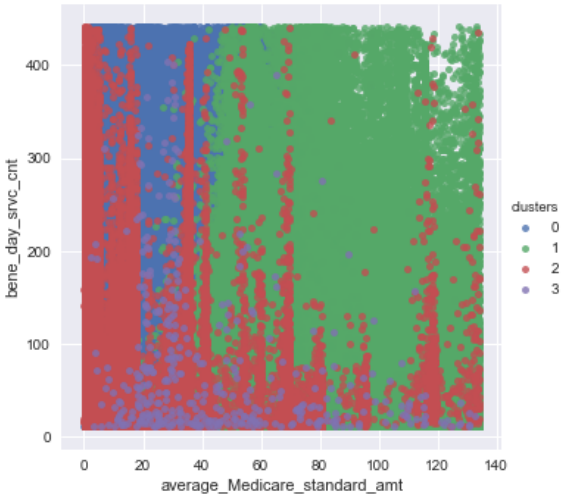
In order to facilitate clustering, a random sample of 1 million datapoints was drawn from the larger dataset, standardized, and passed through the k-means algorithm. Ideally, this random sampling would be repeated many times and the results averaged, to introduce an element of bootstrapping to the analysis.

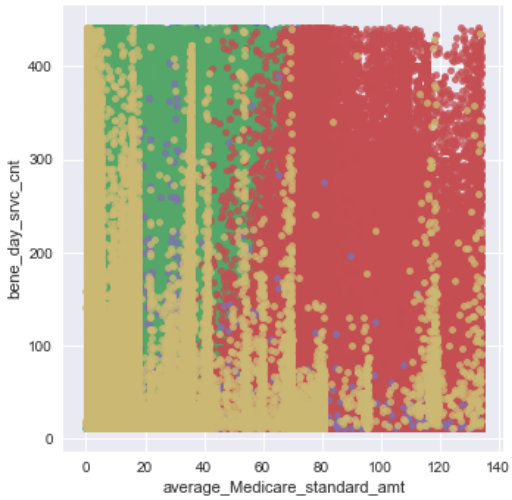
As these data relate to Medicare services and procedures, the business application explored was whether Medicare recipients could be categorized into various groups of service users, such as frequent/infrequent recipients, those who use certain types of services/procedures more than others, etc., so that data such as past Medicare data could be predictive of a patient's future behavior and needs.

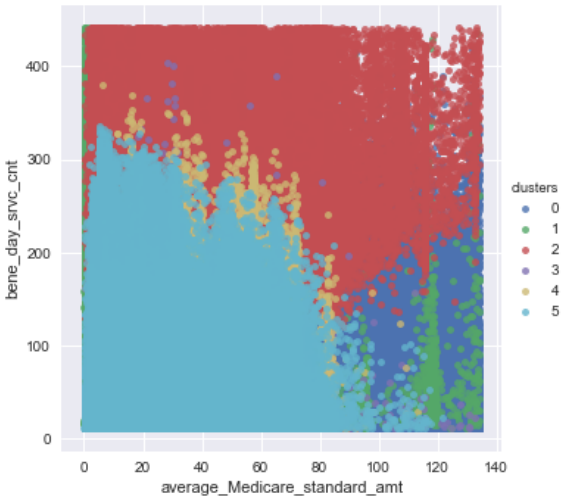
Based on qualitative and quantitative information, it appears that the optimal *k* number of clusters is 6.

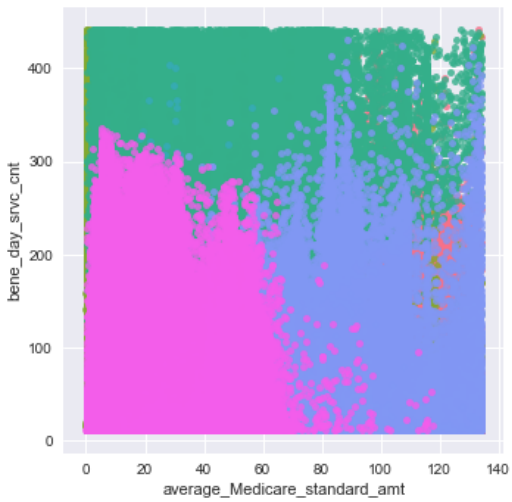

line_srvc_cnt
bene_unique_cnt
bene_day_srvc_cnt
average_Medicare_allowed_amt
average_submitted_chrg_amt
average_Medicare_payment_amt
average_Medicare_standard_amt

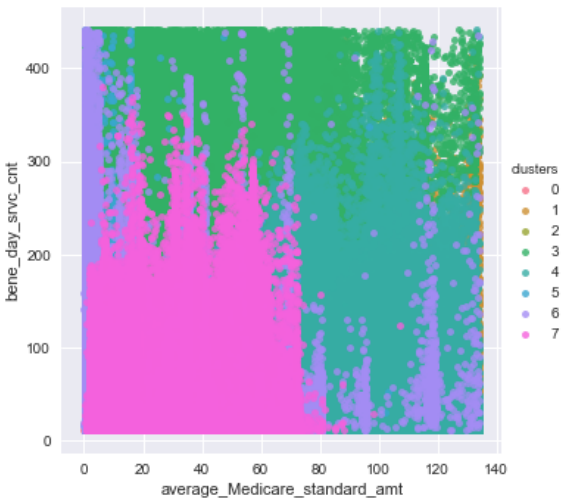

bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2

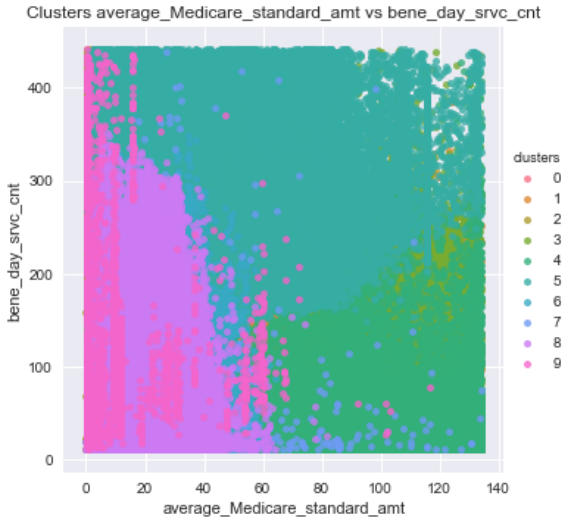

bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2
3

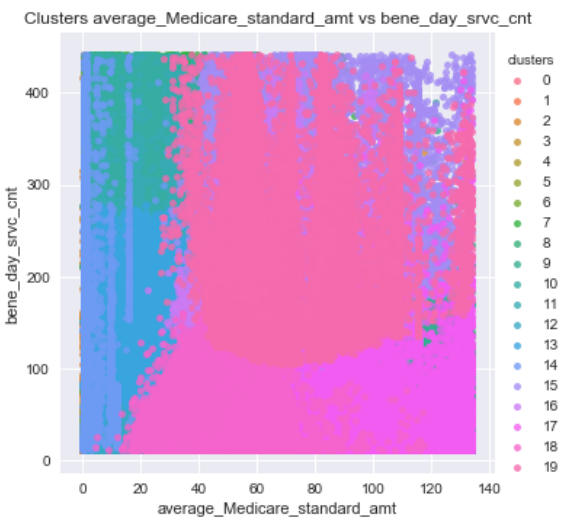

bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2
3
4


bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2
3
4
5


bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2
3
4
5
6


bene_day_srvc_cnt
average_Medicare_standard_amt
dusters
0
1
2
3
4
5
6
7


Clusters average_Medicare_standard_amt vs bene_day_srvc_cnt
bene_day_srvc_cnt
average_Medicare_standard_amt
clusters
0
1
2
3
4
5
6
7
8
9


Clusters average_Medicare_standard_amt vs bene_day_srvc_cnt
bene_day_srvc_cnt
average_Medicare_standard_amt
clusters
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19