Thomas Massion
IEMS 308 Winter 2020, HW#1: Clustering

A Look at Regional Differences in Medicare Coverage

Medicare PUF 2017 Dataset

Exploratory Data Analysis

The complete Medicare PUF dataset was imported into both Python and SQL for exploratory data analysis. SQL was used in part to do macro-level counting and sanity checks, but Python was the primary tool used for plotting distributions and exploring correlations (as well as clustering).

First to gauge the size of the data in SQL, using the command:

```
1    SELECT count(*) from medicarePUF;
```

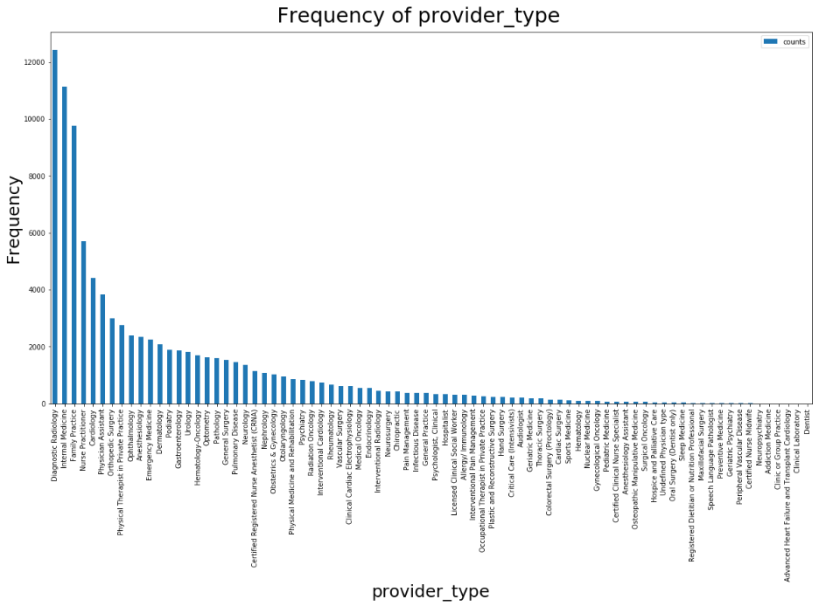
	count(*)
1	9847444

Seeing that the data has over 9.8 million records, I chose to initially randomly sample 1% of the data. This leaves us with around 98,000 records to perform EDA.

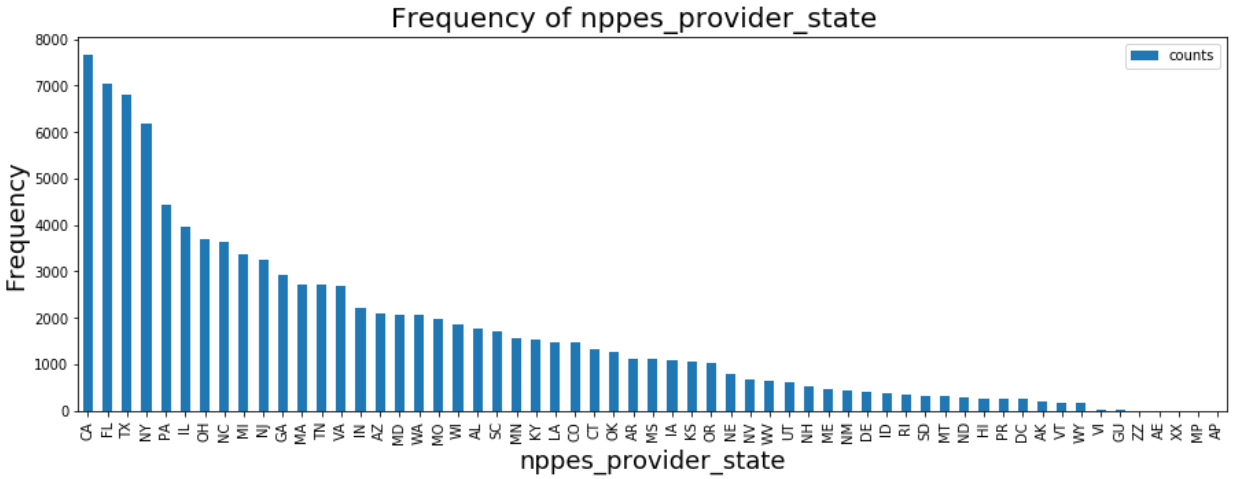
Distributions of features

EDA here begins with exploring the categorical variables relating to geographic and provider type differences in Medicare service.

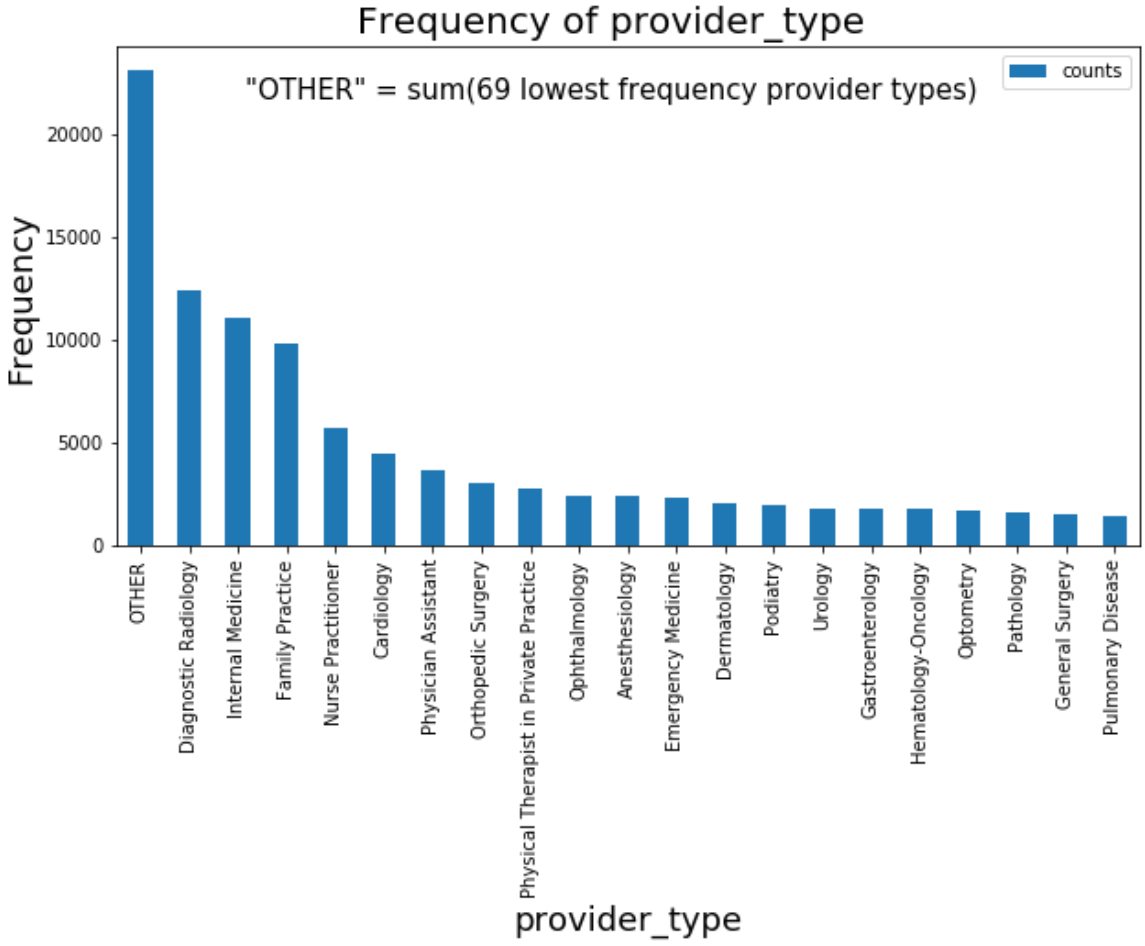
First looking at the distribution records across provider types:

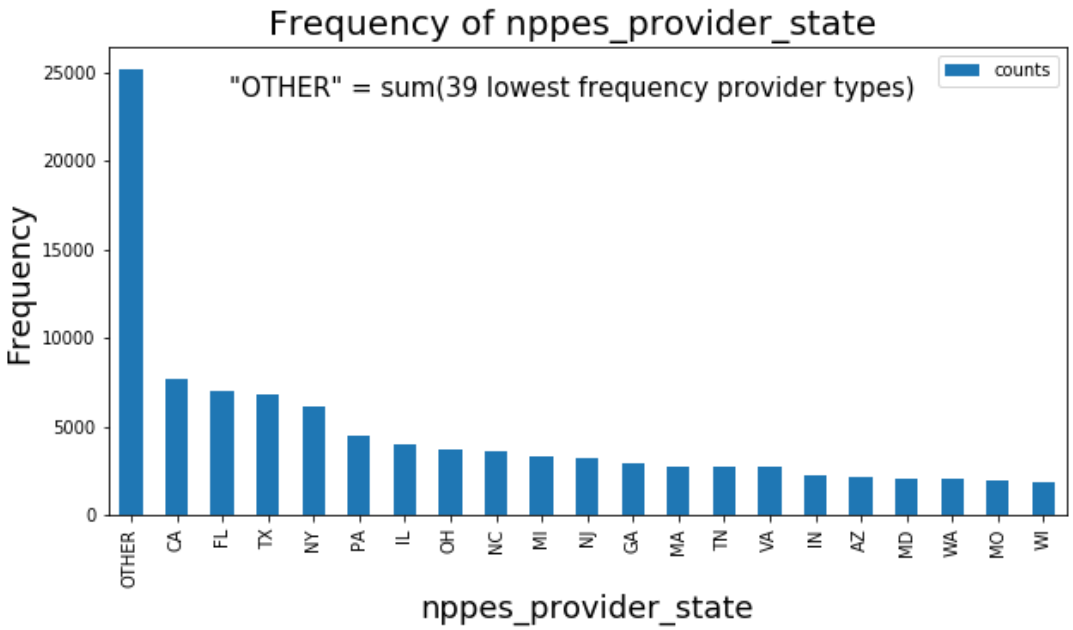


Then looking at the distribution records across states:



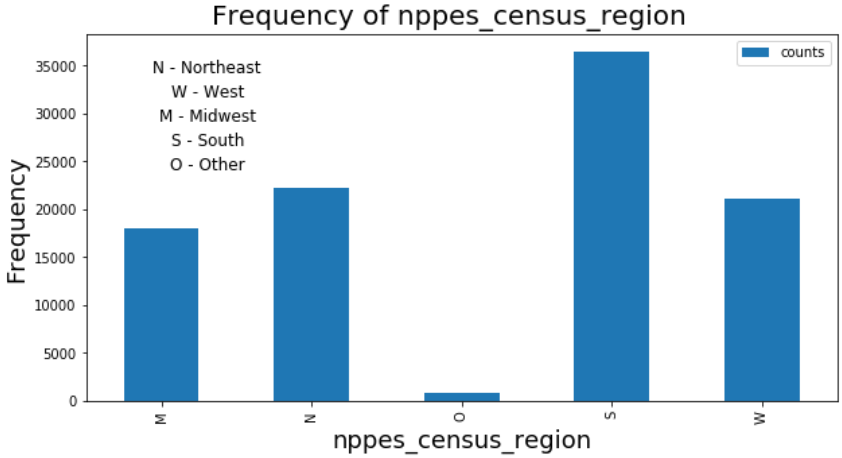
Seeing that for both providers and states, there are several leading provider types and states, followed by many provider types and states with much smaller frequencies. In hopes of decreasing the total number of bins so that these features could potentially be used for clustering, the effect of putting smaller frequencies into a collective "Other" bin was tested:





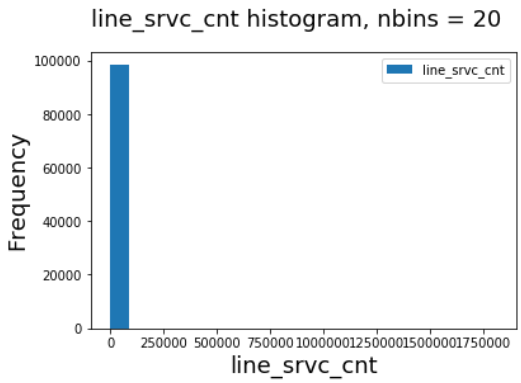
Here we see that the for both provider types and states, the high number of categories means that an "Other" bin will quickly accumulate in size with the addition of lowest frequency bins. Having condensed the categories into a total of 21 bins for both provider types and states already had the "Other" bin far surpassing other categories in terms of distribution.

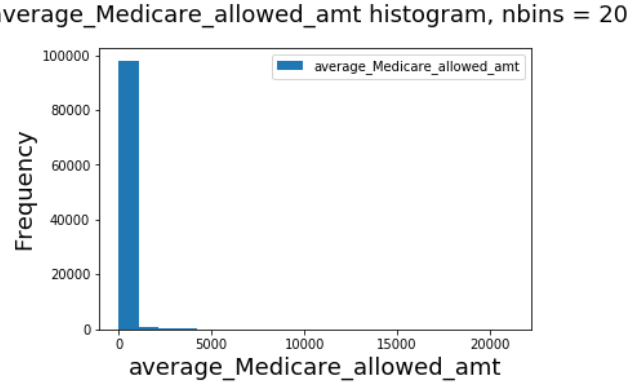
Although this "Other"-ing method didn't seem to work for provider type to be used for clustering, I decided to try to group the states into geographic census regions to have a geographic feature in my clustering.

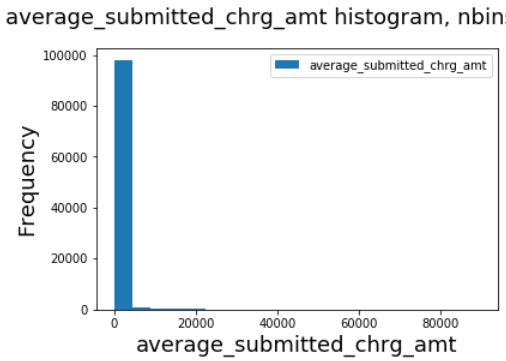


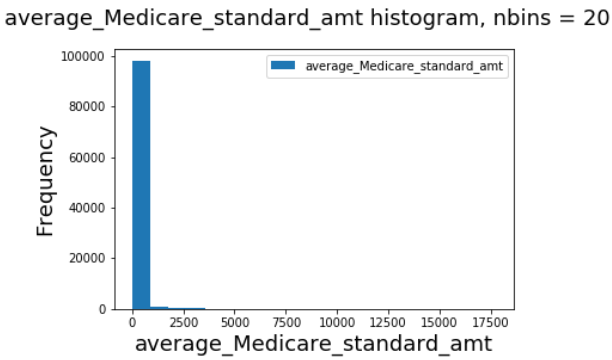
This classification is much more evenly distributed in terms of regions inside the US (Other includes non-mainland states and territories outside the US), and the small number of categories makes this much more suitable for clustering. Thus, we had a desirable regional feature for potential clustering.

Then the numerical features of the dataset were explored.









For these for 4 numeric features relating to volume of services (line_srvc_cnt), total charges (average_medicare_allowed_amt, average_submitted_chrg_amt), and medicare payments (average_medicare_standard_amt), we see that the data is skewed far to the right, suggesting some outliers in the data.

Upon having investigated this far left skew, the following observations and decisions were made:

- There are significantly more individual entity entries who *will* provide medicare
- The majority of medicare partificipation is a full yes for providers. The No's are less than 0.0005% of the data.

```
1  SELECT medicare_participation_indicator,count(medicare_participation_indicator)
2          from medicarePUF GROUP by medicare_participation_indicator ;
```

	medicare_participation_indicator	count(medicare_participation_indicator)
1	NULL	0
2	N	3328
3	Y	9844115

- Most providers are individuals and not organizations. Providers identified as organizations are less than ~5% of the data:

```
1   SELECT nppes_entity_code, count(nppes_entity_code)
2           from medicarePUF GROUP by nppes_entity_code ;
3
```

	nppes_entity_code	count(nppes_entity_code)
1	*NULL*	0
2	I	9416125
3	O	431318

- Organizations have drastically higher line service counts as opposed to individuals

Upon having contemplated this, **I chose to remove the non-Medicare participator** providers since I wanted to look at the amount of Medicare costs that were not covered. Thus, eliminating the No's would help from preventing more outliers down the road.

However, **I chose to keep the individual and organization providers**, since organizations have drastically higher line service counts as opposed to individuals, and thus, capture significantly more Medicare cost differences.

Side Note:
I found that services of specifically "Injection, daptomycin, 1 mg" were the source of many outliers in the data. SQL command: SELECT * from medicarePUF where hcpcs_code = 'J0878'). I chose to keep this data, as these records are for infectious disease vaccinations which are high service count while having very low cost, which agrees with the data.)

As for the far skew to the right, the majority of outliers could not be ruled out as being data input errors, so this further justifies needing to standardize and normalize before clustering.

<u>Introducing Variable for Uncovered Amount per Beneficiary</u>

I then chose to **introduce a new variable** which describes the yearly **average amount per beneficiary not covered by Medicare** for each nppes provider.

From the FAQ document provided for the dataset, CMS.gov reports that

"

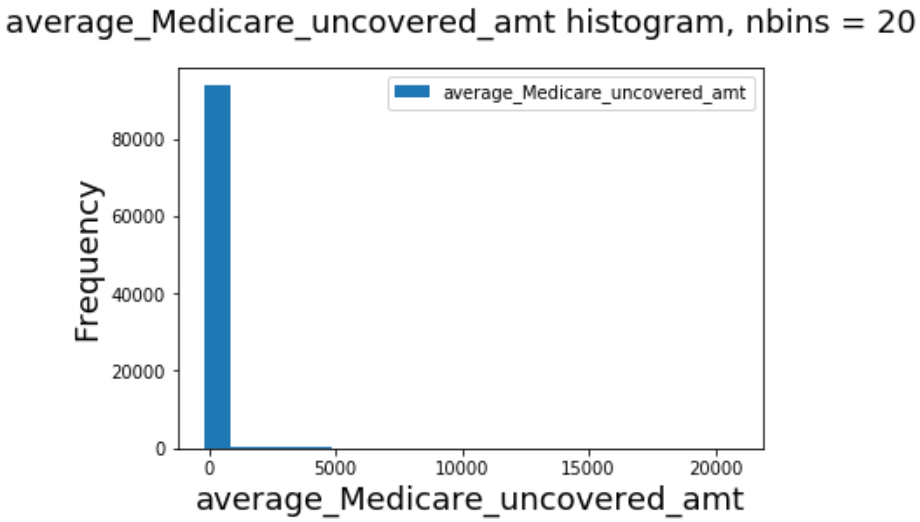
In the Physician and Other Supplier PUF, how are averages calculated for the average_Medicare_allowed_amt, average_submitted_chrg_amt, average_Medicare_payment_amt, and average_Medicare_standardized_amt variables? The average payment and charge variables reflect the total payments or charges for a given HCPCS code/place of service divided by the line_srvc_cnt (i.e., the number of services provided).
"

Source: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Physician_FAQ.pdf

Thus, the equation for average_Medicare_uncovered_amt is as follows:

$$\text{avg}_{\text{Uncovered}} = \text{lineSrvcCnt} * \left(\frac{\text{avg}_{\text{AllowedAmt}} - \text{avg}_{\text{StandardAmt}}}{\text{BeneUniqueCnt}} \right)$$

Observing the distribution for this new feature:



We see that this variable is also skewed to the right, thus we will need to standardize and normalize if used in clustering for this feature as well.

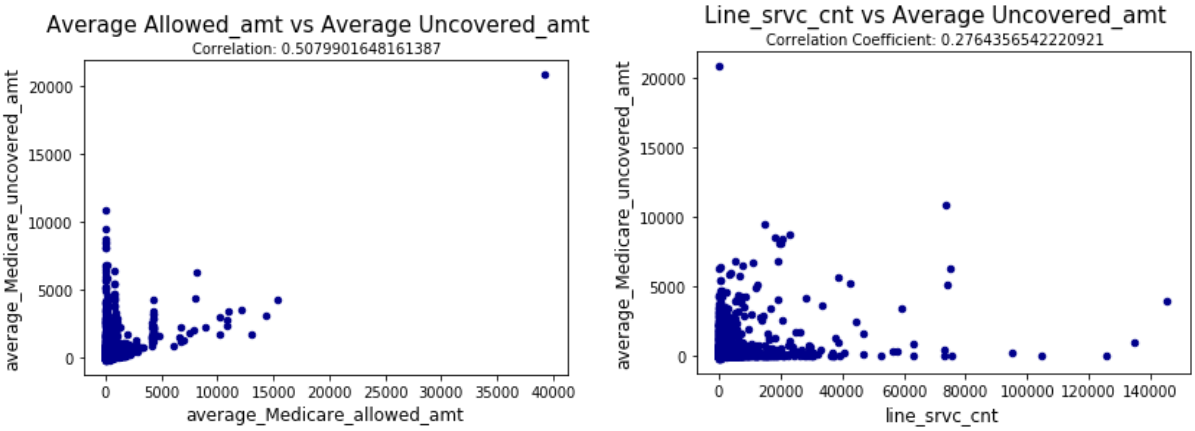
Correlations between features:

Shown here are the correlations between the numeric features of interest for clustering:

	line_srvc_cnt	average_Medicare_allowed_amt	average_Medicare_standard_amt	average_Medicare_uncovered_amt
line_srvc_cnt	1.000000	-0.023825	-0.023461	0.276436
average_Medicare_allowed_amt	-0.023825	1.000000	0.997131	0.507990
average_Medicare_standard_amt	-0.023461	0.997131	1.000000	0.499428
average_Medicare_uncovered_amt	0.276436	0.507990	0.499428	1.000000

Shown here are the correlations between the uncovered amts and the regional features:

	average_Medicare_uncovered_amt	nppes_census_region_M	nppes_census_region_N	nppes_census_region_S	nppes_census_region_W
_Medicare_uncovered_amt	1.000000	-0.012320	0.017436	-0.014692	0.010454
nppes_census_region_M	-0.012320	1.000000	-0.257675	-0.360514	-0.247589
nppes_census_region_N	0.017436	-0.257675	1.000000	-0.414765	-0.284846
nppes_census_region_S	-0.014692	-0.360514	-0.414765	1.000000	-0.398529
nppes_census_region_W	0.010454	-0.247589	-0.284846	-0.398529	1.000000



Here we see that the uncovered amount does not have a very strong correlation with any major other numeric variables related to cost, nor does it have any strong correlation to any particular region, thus prompting my central business question to be solved for clustering.

Business Question

Are there regional differences in Medicare coverage and how might that relate the provider place of service and provider gender?

Thus, I have chosen to cluster on the following variables:

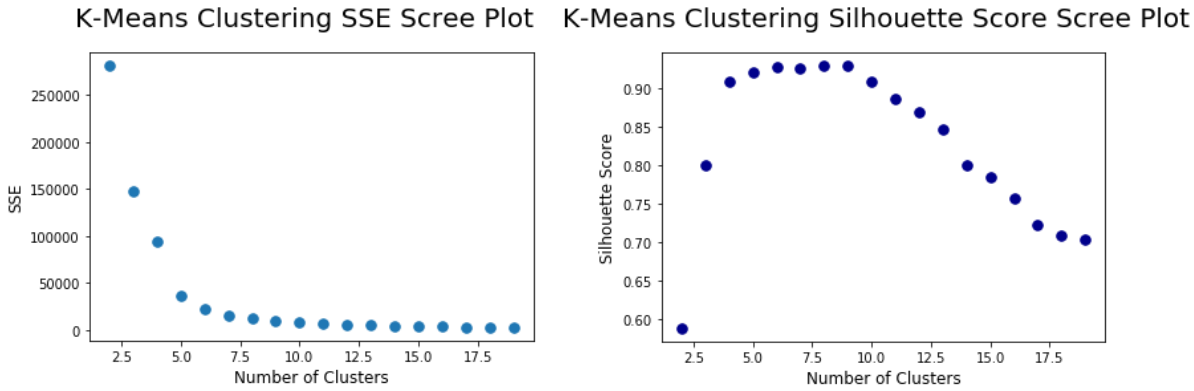
- average_Medicare_uncovered_amt
- place_of_service
- provider_gender

Hypothesis:

I believe there will be regional differences, in Medicare coverage, considering there are differences in state-level policies relating to mandatory insurance coverage and differences in poverty levels and elderly populations within regions. Thus, **I expect to see the clusters resulting from the features I have selected to align themselves with a census region or regions.**

K-Means Clustering:

Upon standardizing and normalizing the features selected for clustering, K-means clustering was performed starting with 2 clusters, then moving incrementally up to 20 clusters, with the SSE and silhouette score evaluated for each number of clusters along the way. The results are summarized in the scree plots below:



From this the optimal number of clusters based on Silhouette score was determined to be **9 clusters**.

Distribution of clusters:

```
dat['Cluster'].value_counts()
```

```
1    37445
0    27634
2    19299
3     9374
7      258
5       83
4       42
6        3
Name: Cluster, dtype: int64
```

The raw cluster results are shown below:

Cluster	average_Medicare_uncovered_amt	place_of_service_F	place_of_service_O	nppes_provider_gender_F	nppes_provider_gender_M
0	33.811369	1.000000	0.000000	0.000000	1.000000
1	31.026857	0.000000	1.000000	0.000000	1.000000
2	28.645668	0.000000	1.000000	1.000000	0.000000
3	30.190208	1.000000	0.000000	1.000000	0.000000
4	5483.346626	0.000000	1.000000	0.238095	0.761905
5	2540.776330	0.000000	1.000000	0.313253	0.686747
6	14432.842500	0.000000	1.000000	0.000000	1.000000
7	892.119065	0.069767	0.930233	0.217054	0.782946

Here we see that clusters with higher uncovered costs (clusters 4-7) were clustered **very highly with offices as a place of service** as indicated by the red rectangular region.

We also see that these high uncovered cost clusters **tend to lean towards having male providers**, while the high uncovered cost clusters (clusters 0-3) are split evenly between male and female.

Comparing these results with the distribution of clusters, we can also see that the **clusters of highest uncovered cost have the smallest populations**.

Now to officially examine the hypothesis for the business question of differing Medicare coverage within geographic census regions, the clusters are shown next to their average region attribution (just to clarify, the regions were **not** used as a feature for clustering):

Cluster	average_Medicare_uncovered_amt	nppes_census_region_M	nppes_census_region_N	nppes_census_region_S	nppes_census_region_W
0	33.811369	0.205146	0.211298	0.366903	0.210393
1	31.026857	0.165389	0.217839	0.389878	0.220109
2	28.645668	0.189440	0.237940	0.361210	0.204985
3	30.190208	0.206849	0.269362	0.320674	0.196821
4	5483.346626	0.095238	0.261905	0.452381	0.190476
5	2540.776330	0.132530	0.168675	0.445783	0.240964
6	14432.842500	0.333333	0.333333	0.333333	0.000000
7	892.119065	0.120155	0.259690	0.313953	0.298450

We can see from the indicated red regions that the highest uncovered costs clusters contain records most often geographically located in the **south of the US**.

Conclusion:

While acknowledging that clustering does not give way to formal statistical conclusions including tests of significance, we can conclude within the grounds of the tools used to answer the business question that there are indeed some regional differences in Medicare coverage. This is because we can see that highly uncovered clusters had samples mostly from the southern census region compared to other regions. Furthermore, these highly uncovered clusters often had male providers and took place in offices as opposed to facilities (facilities tend to be large-scale hospitals, clinics, etc.).

The aim of this business question is to look at disparities in coverage for Medicare to help give a starting point to understanding why Medicare might fall short in certain places. The average Medicare uncovered cost is the average yearly amount per beneficiary that Medicare does not cover for individuals to pay their medical fees. For most of the data within the subset of the data examined in this project, uncovered costs did not exceed $30, however what's more alarming is the highly uncovered clusters containing uncovered costs in the thousands of dollars (averages of around $892, $2540, $5483, and $14432). Considering Medicare is primarily for the elderly who can often be retired, and have little or no major sources of income, these high uncovered health costs can have serious impact on quality of life if uninsured by a third-party health insurance. Thus, these clustering results can point to further investigation in regional uncovered medical costs in terms of Medicare and the types of places where these uncovered costs occur.