IEMS 308
Morgan Thatcher

Homework 1: Clustering

Executive Summary

The publicly released Medicare Data includes information about health care providers, the services they provide, and the coverage of those services through Medicare. For the purposes of this analysis, the 2017 dataset was used. Exploratory data analysis in terms of variable definitions, variable types, and typical quantities was conducted to better understand sample features. In an effort to understand the importance of an MD provider credential, clustering was done for features of Medicare Submitted Cost and Unique Beneficiary columns. An initial hypothesis was that MD qualified providers would provide less common services at higher submitted cost prices. This hypothesis however was proven false through the use of cluster analysis. Whether a provider entity includes an MD does not relate to the popularity of the services provided or the pricing.

Findings

Description of Data

After obtaining the 2017 Medicare Data, an initial exploration of the variable definitions and types was necessary. Many features were categorical, describing both provider information and the service information. Numerical features included the number of beneficiaries of particular services as described through the variable names: line_srvc_cnt, bene_unique_cnt, and bene_day_srvc_cnt. Other numerical features included the Medicare amounts as given through variable names: average_Medicare_allowed_amt, average_submitted_chrg_amt, average_Medicare_payment_amt, and average_Medicare_standard_amt.

Data Exploration

In an effort to perform exploratory data analysis and better understand the trends in data, summary statistics and plots were generated on various features. Using all rows of the Medicare data and focusing primarily on continuous variables including Medicare Allowed Amount, Submitted Charge, Medicare Standard Amount, and Medicare Payment Amount, the following plots and tables were created. To understand how impactful outliers were on the factors, summary statistics were generated. In all cases, outliers skewed the descriptive statistics to greater values. For this portion of understanding the data, outliers were taken out according to the rule:

$$Q1 - 3IQR \geq DataPoint \geq 3IQR + Q3$$

Summary Statistics (With Outliers)

Variable	Min	Q1	Median	Mean	Q3	Max	IQR
Medicare Allowed Amount	0.0	24.26	65.00	101.60	113.50	56684.35	89.24
Submitted Charge Amount	0.0	58.0	146.0	351.6	300.0	100000.0	242
Medicare Standard Amount	0.0	20.09	47.76	78.22	85.25	44439.10	65.16

Medicare Payment Amount	0.0	19.28	49.96	77.53	85.19	44439.10	65.91

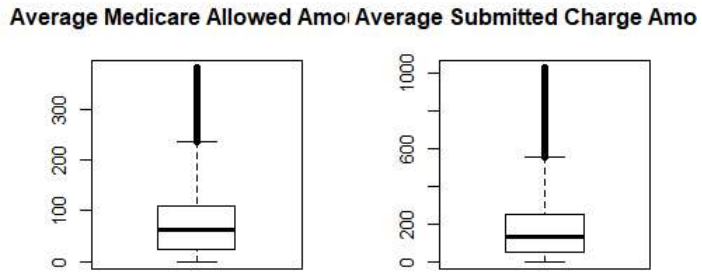
At this point, the outliers were removed from calculations according to the aforementioned rule using the interquartile range.

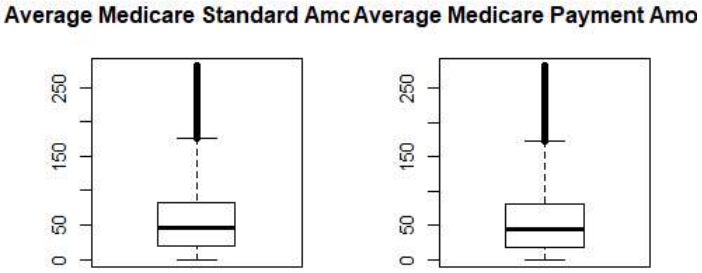
Summary Statistics (Without Outliers)

	Min	Q1	Median	Mean	Q3	Max	IQR
Medicare Allowed Amount	0.0	23.78	62.04	76.43	108.92	381.23	85.14
Submitted Charge Amount	0.0	53.0	132.0	193.03	252.0	1026.0	199
Medicare Standard Amount	0.0	19.41	46.26	58.01	82.22	280.73	62.81
Medicare Payment Amount	0.0	18.73	45.35	57.55	80.75	282.93	62.02

To obtain a more graphical representation of the statistical features of each variable, boxplots were generated, again without outliers. General findings from these plots and tables show that the Average Submitted Charge Amount skews larger than the Medicare Amounts granted, with Medicare Payment Amount having the lowest median and mean statistics. While the boxplots below might at first glance look similar, the y-axis demonstrates the large variation between values.

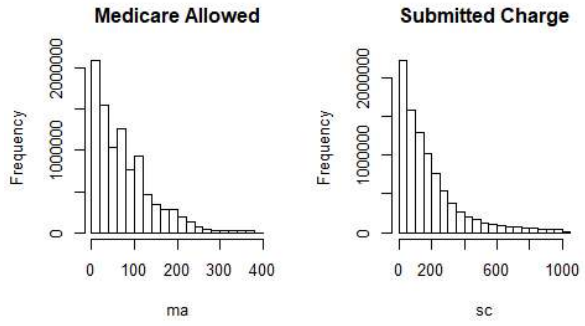
Box Plots

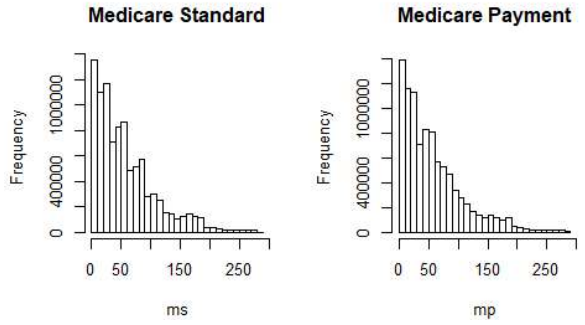




The same four variables were then plotted using histograms to understand the frequency of values. All variables have a right skew and further demonstrate the distribution of data points around the mean and median statistics.

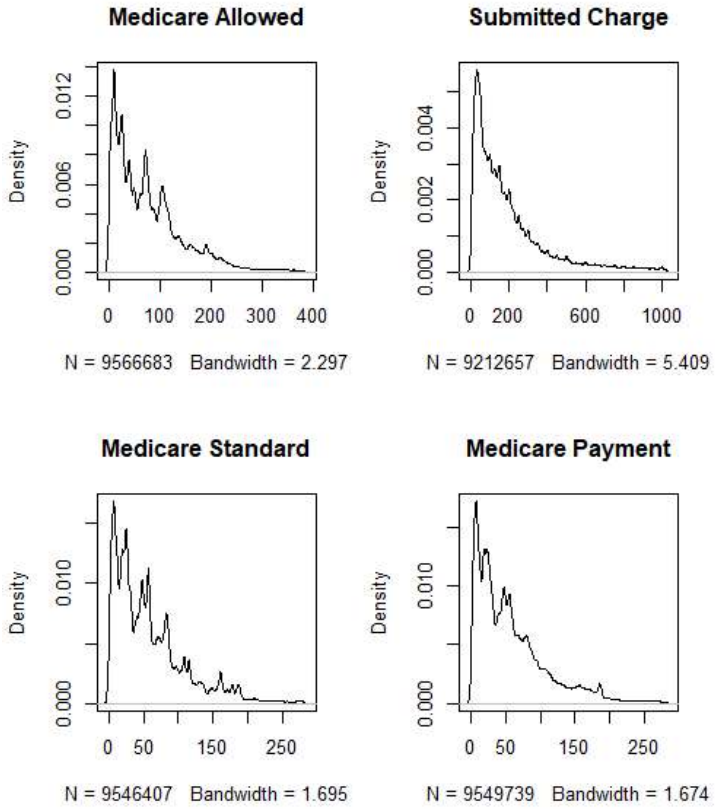
Histograms





To understand the distribution patterns that each feature followed, density plots were generated. As seen below, all variables have right skew. This graphic also shows the N-values for each variable which indicates, from the base amount of data points, how many outliers were calculated and removed.

Density Plots



Lastly, in exploratory data analysis, correlations were evaluated between the 4 variables previously used. From the matrix below, it's important to note that the lowest correlation, while still high, is between the submitted charge amount and the 3 Medicare amount variables.

Correlations

	Medicare Allowed	Submitted Charge	Medicare Standard	Medicare Payment
Medicare Allowed	1	0.7425943	0.9946656	0.9984357
Submitted Charge	0.7425943	1	0.7382782	0.7419106
Medicare Standard	0.9946656	0.7382782	1	0.9940894
Medicare Payment	0.9984357	0.7419106	0.9940894	1

The same types of exploratory statistics were also created for the other numerical variables including line_srvc_count, bene_unique_cnt, and bene_day_srvc_cnt. Using the same outlier IQR rule as aforementioned, summary statistics with and without outliers were obtained.

Summary Statistics (With Outliers)

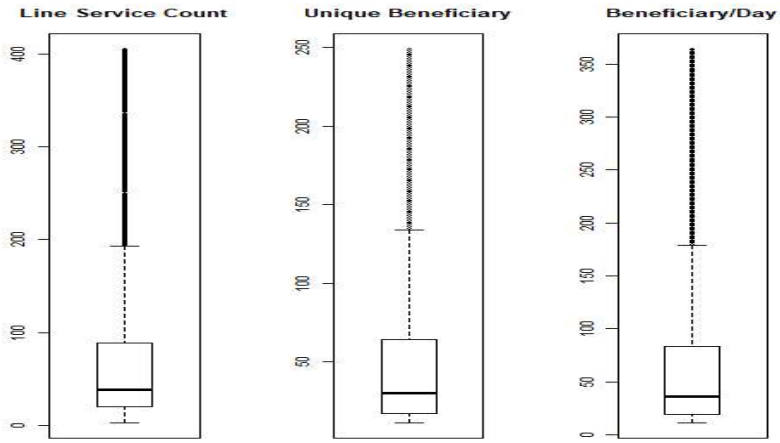
Variable	Min	Q1	Median	Mean	Q3	Max	IQR
Line Service Count	2	21	43	245	117	7195536	96
Beneficiary Unique Count	11	17	32	87.8	75	792873	58
Beneficiary/ Day Service Count	11	20	40	140.3	106.0	1504215	86

Summary Statistics (Without Outliers)

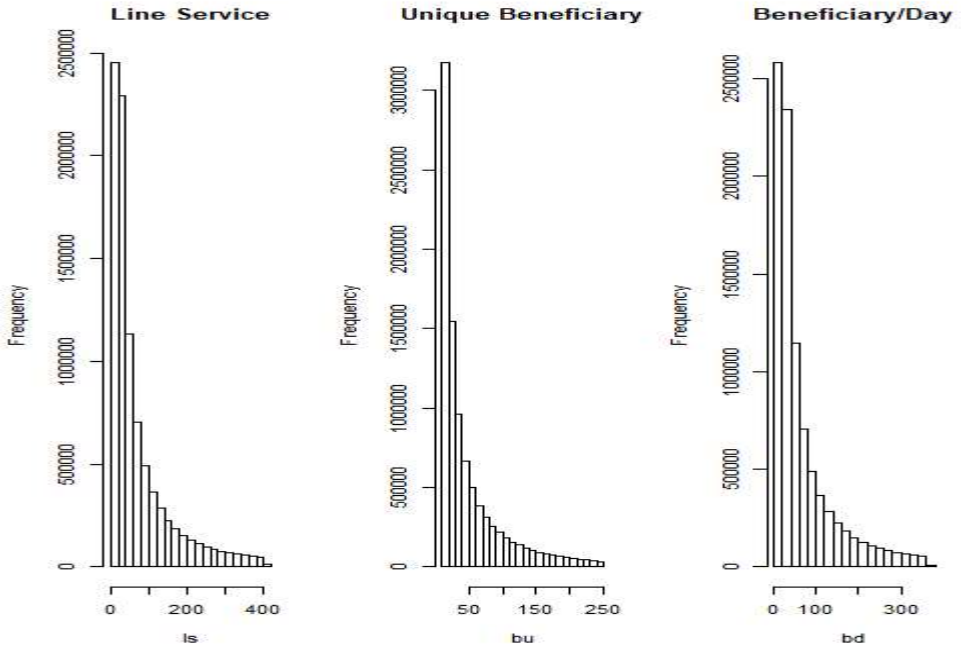
Variable	Min	Q1	Median	Mean	Q3	Max	IQR
Line Service Count	2.4	20	38	71.41	89	405	69
Beneficiary Unique	11	17	30	50.16	64	249	47

Count							
Beneficiary/ Day Service Count	11	19	36	66.38	83	364	64

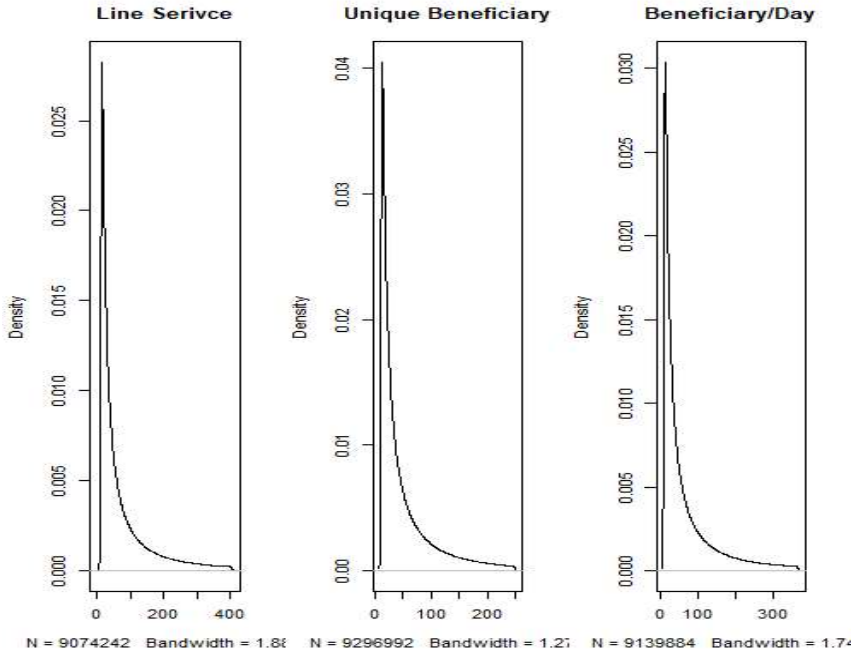
Box Plots



Histograms



Density Plots



Correlations

	Line Service	Unique Beneficiary	Beneficiary/Day
Line Service	1	0.4005589	0.4350068
Unique Beneficiary	0.4005589	1	0.9667719
Beneficiary/Day	0.4350068	0.9667719	1

Next, to understand the categorical variables, especially binary categorical variables, and the frequency of occurrence, a count and percent occurrence were created on each binary factor. For every binary variable, count of each type of response:

Variable	Count1	Percent1	Count2	Percent2
Hcpcs Drug indicator	Y: 616248	6.26	N: 9231195	93.74
Place of Service	F: 3769732	38.28	O: 6077711	61.72
Medicare Participation Indicator	Y: 9844115	99.97	N: 3328	0.03
Nppes Entity Code	I: 9416125	95.62	O: 431318	4.38
Nppes Provider Gender*	M: 6535107	64.40	F: 2881018	30.60

*nppes_provider_gender is not listed for 431319 of the samples which corresponds to the number of samples registered as organizations with nppes_entity_code(O)=431318.

After a primary data exploration, more analysis was necessary to understand which variables or factors affected others, and how that could be used to solve a business problem. In an effort to do this, SQL queries were performed on the data using a heuristic hypothesis approach. Several interesting findings appeared from this process.

In answering the hypothesized question: Do certain entitled providers submit higher amounts to Medicare for the same service? It was found that the MD title had a slightly higher overall average Submitted Amount, but when grouping by hcpcs code, MD entitled providers had a lower average. Since SQL alone could not provide a complete answer to the question, this business question was pursued using Clustering. The SQL results and query can be seen below.

SQL Query and Results

Results

Total: MD: 360.443179405997; Not MD: 351.588034210642

Group by HCPCS: MD: 1328.64959738638; Not MD: 1422.14188704305

Query

```
SELECT AVG(average_submitted_chrg_amt) FROM
(SELECT * FROM Medicare_Provider_Util_Payment_PUF_CY2017 WHERE
nppes_credentials LIKE "%MD%" OR
nppes_credentials LIKE "%M.D.%"
GROUP BY Medicare_Provider_Util_Payment_PUF_CY2017.hcpcs_code)
```

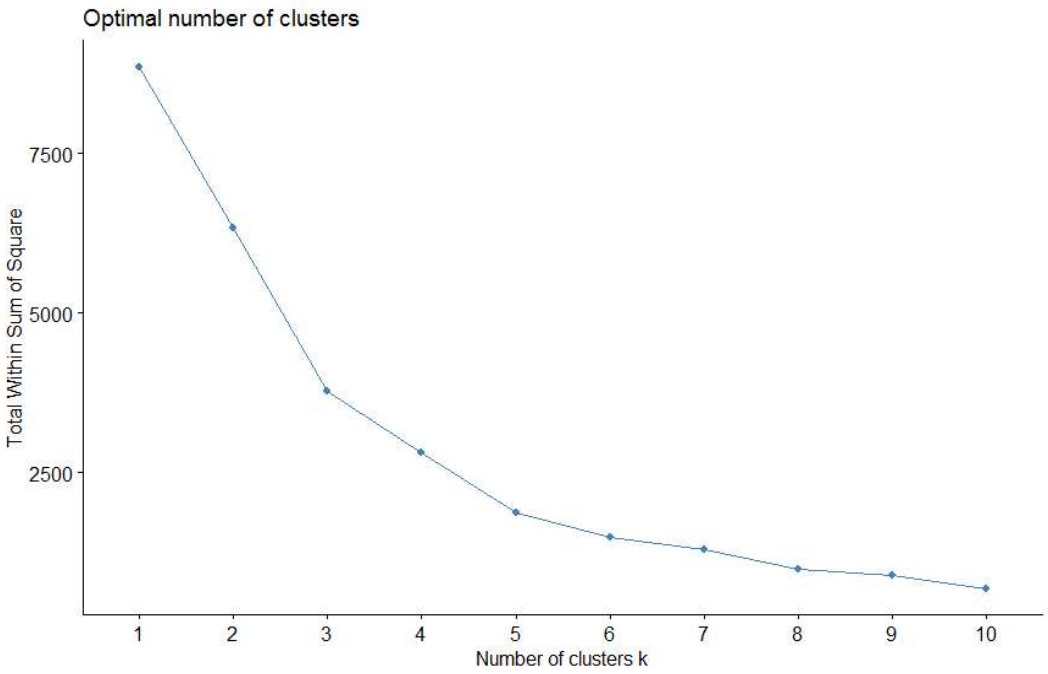
Business Question

The previous findings led to the creation of a business question to be solved through clustering. Does the inclusion of an MD title for a provider increase the amount submitted to Medicare for services and decrease the commonality of services provided? The business implications of this, if it were found to be true, would be for provider organizations to hire MD entitled individuals and for individual providers to either organize with MDs or obtain the credential in an effort to perform more interesting, less common, procedures for a higher price point.

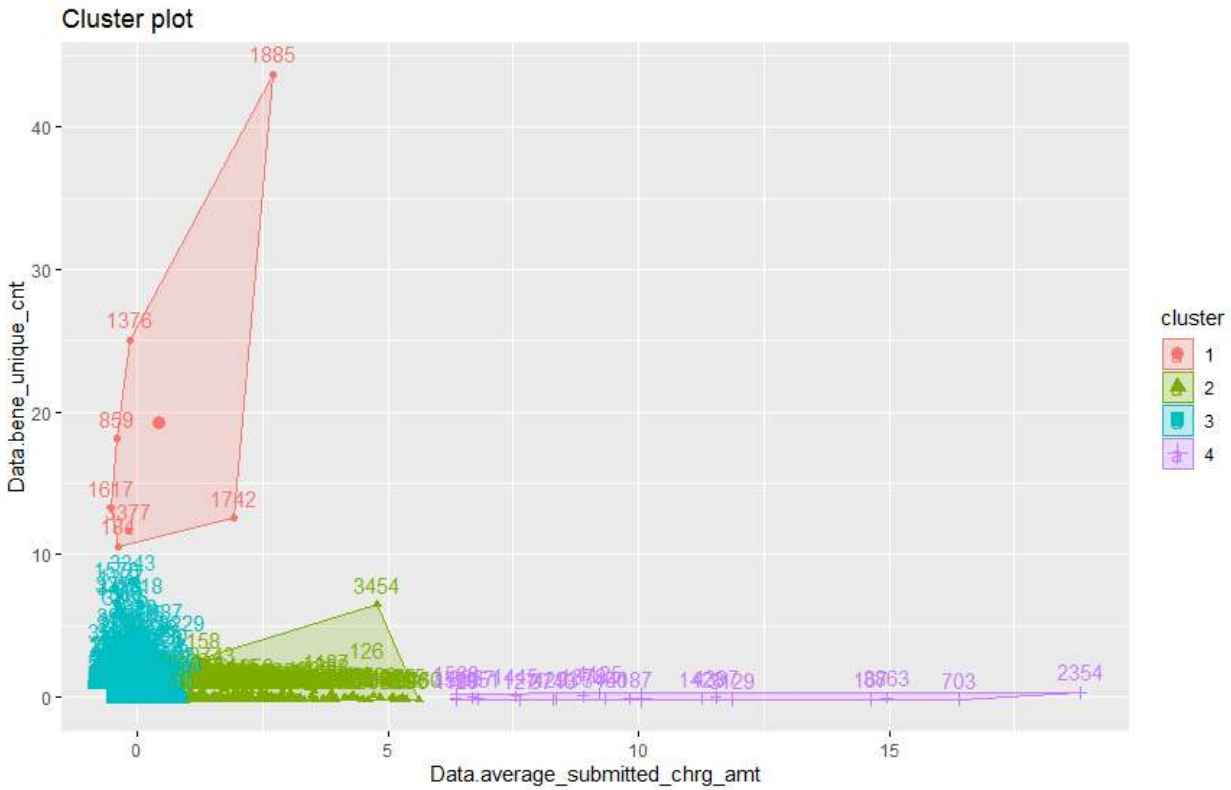
Clustering

To solve the aforementioned business question, clustering analysis was completed using features of submitted amount and beneficiary count as found in the variables: bene_unique_cnt and average_submitted_chrg_amt. To isolate part of the data set, and through the previous understanding of the number of outliers present in the data, an initial cleansing was completed in which, on both features, outliers were omitted. The data frame including the features described, with outliers and empty cells omitted, were scaled. Multiple numbers of clusters were fit to the features and plotted against each SSE to find the optimal number of clusters.

Optimal Cluster Plot



As seen in the optimal cluster plot, a kink in the curve is found at k=4. This number of clusters then was used to segment the data and can be seen in the plot below.



The average values for both factors, submitted amount and unique beneficiary count, as grouped by cluster were calculated and displayed against the average one-hot encoded value of MD entitlement. Cluster 4, with a larger than average submitted charge amount, had an average one-hot encoding of MD title of 0. This means that all services clustered in the high submitted amount were performed by providers without a MD credential. Cluster 1, with a high unique beneficiary count was similarly linked to an average of 0. Cluster 3, with the lowest submitted amount had a 0.579 average one-hot encoding. In regards to the business question, this solution seems to indicate that the MD credential is not associated with high submitted amount or low beneficiary/service.