# *Medicare 2018 Data Analysis*

*Bruno Zecchi*

## Executive Summary

There are countless healthcare services performed through Medicare every year. In this task we sought to find some insights on the healthcare industry by analyzing data provided by Medicare on their website. In particular, we chose to focus on prices charged for services that involved drugs, as well as on the amount that Medicare is allowed to cover for those services. We chose to perform this analysis by clustering and did so by using the K-Means algorithm. After performing initial data exploration and running models with different number of clusters, we determined that the data would be best split into three clusters. We then performed final analysis on those clusters and determined that the clusters represent services with different drugs. One is for Sipulecel-T, another for Pegfilgrastim, and the other for what we determine to be "miscellaneous". We thus see that the two drugs have unique characteristics that separate their services from all others. Although statistical analysis of the data might be finished, it is important to understand why the two drugs provide such clustering. This understanding might provide further knowledge on the strength of the model and provide more insights on the drug market.

## Problem Statement

In this task we analyze data provided by Medicare in 2018. The data is a record of services provided to Medicare patients throughout the year. Some features of the data include the price of the service and whether a drug was involved in the service. For this task we chose to understand whether drug services could be clustered into categories. Those categories could be related to the price charged for the service or on how much of the service Medicare chose to cover. This analysis would give us a better understanding of the pharmaceutical drug market and the interaction between providers, individuals, and Medicare.

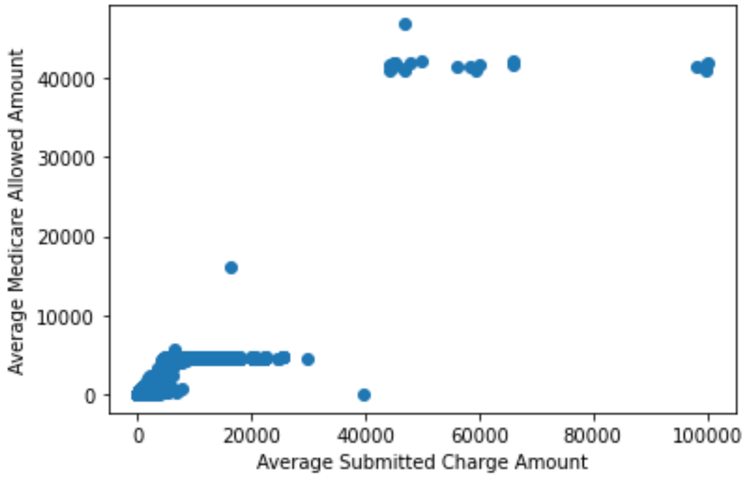
## Methodology

<u>Subset Selection:</u>

The data file is of significant size, so in order to make further processing of data smoother we chose to subset the data so that only relevant data is analyzed further. Since the business question tackles drug services prices, services that had the "drug indicator" feature marked as "Y" were stored in a separate data frame. Out of the initial 9 million rows, we now only work on six hundred thousand. All other rows are irrelevant for the task at hand.

Data Exploration:

Since we are interested in the price of drug services and on the Medicare allowed amount for that service, we chose to produce a scatter plot of that data, provided below.

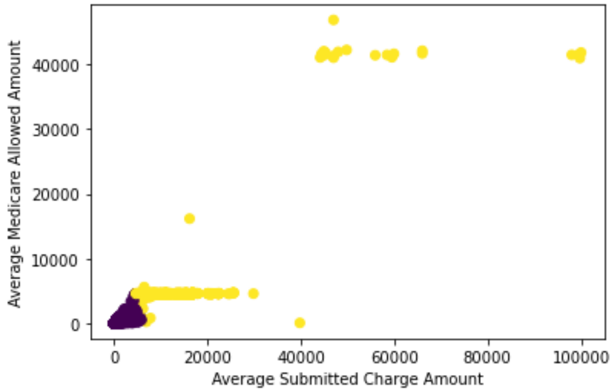


From the plot we realize that the data can be clustered based on the two features. It is worth noting that the data is not very linear, meaning that there are drug service prices that increase in charged amount, but that Medicare does not increase in its allowed amount.

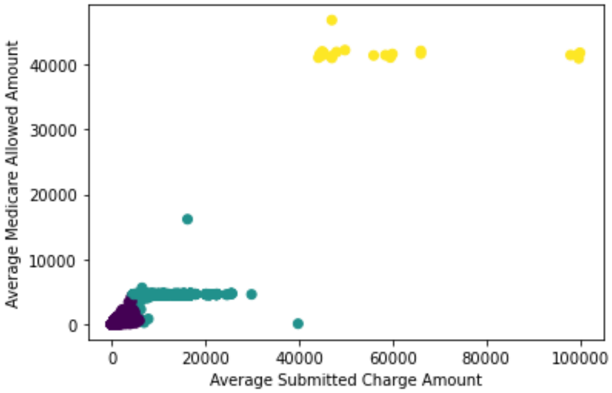
K-Means Clustering

We then set about clustering the data utilizing the K-Means algorithm. Using the K-Means function from *sklearn* in Python, we clustered the data with different numbers of clusters. We provide figures for the clusters and a table for the inertia values for each of the numbers.

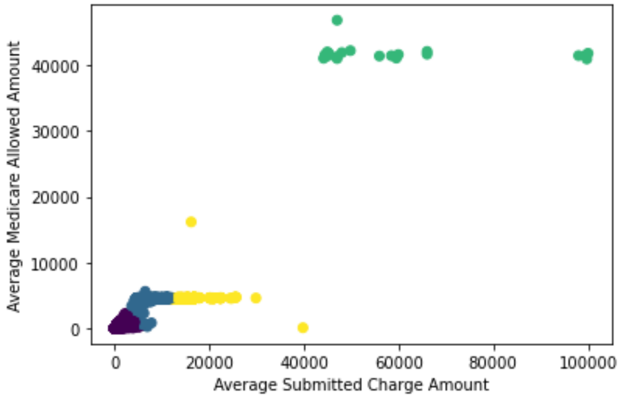
Number of clusters = 2



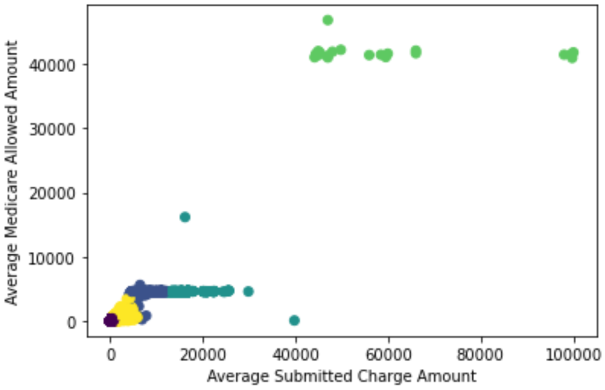
Number of clusters = 3

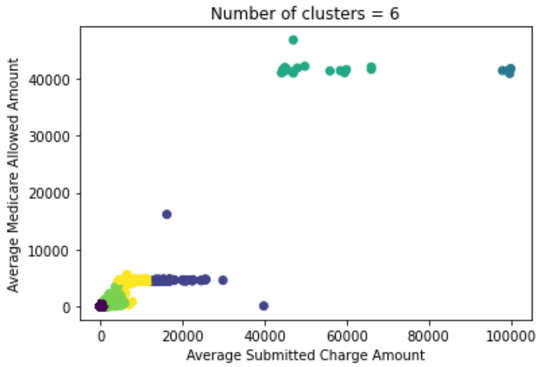


Number of clusters = 4



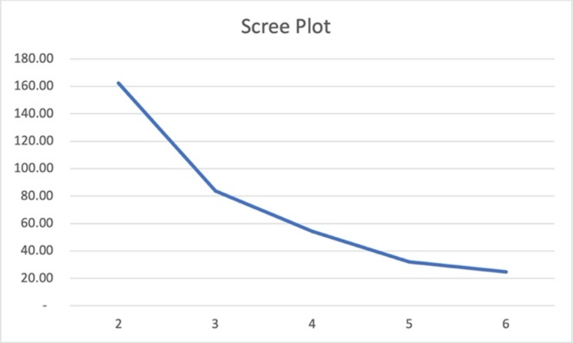
Number of clusters = 5





| Number of Clusters | Inertia |
|---|---|
| 2 | 162.5 million |
| 3 | 83.8 million |
| 4 | 54.3 million |
| 5 | 32.0 million |
| 6 | 24.8 million |

With the values from the table, we created the Scree Plot below:



By analyzing the clustering plots and the kink of the scree plot, we selected the optimal number of clusters to be three.

**Analysis**

We now set about understanding the reason for these clusters. By understanding what characteristic the model was trying to cluster, we can determine how accurate the model is. We first analyzed the cluster at the top right of the plot. We split off those data points into a separate dataframe, and then utilized Python's *describe()* function.

Through this analysis we saw that 95% of data points in that cluster represented drug services that were related to Sipulecel-T. This drug is not present in either of the other two clusters. Therefore, we can conclude that this cluster is related to services involving Sipulecel-T.

We next analyzed the middle cluster. We performed the same analysis as in the previous cluster and determined that this cluster is related to services involving Pegfilgrastim. 99.6% of data points in the cluster have that drug, and Pegfilgrastim is only present in 7 data points from other clusters. The bottom-left cluster can be considered the "other" or "miscellaneous" cluster, since it does not have any unique characteristics.