Abstract—Data science is evolving. In the minds of many business leaders, data science has evolved into a business necessity. Forward thinking leaders have formed data science departments which have proven to bring new business values across the entire organization. Data science departments are equipped with professionals who can work autonomously and create new data science applications. These data science applications are positioned to derive valuable insights from data. Effective data science teams have proven to create tremendous competitive advantages.

This paper explores the concept of data mining. Data mining is the process of extracting knowledge from data using different technologies, processes and algorithms. The paper looks into data mining from the perspective of an analysis conducted on data retrieved from a government website. The analysis considers a dataset which includes information comparing the charges for the 100 most common inpatient services and 30 common outpatient services. The analysis highlights the variation across the US in what medical providers charge for common services. The analysis grouped each cluster based upon medical diagnosis, medical provider city and state, total number of discharges, average covered charges and average total payment received from Medicare. The analysis focused on medical providers in New York, Florida, Pennsylvania, California, Maryland and New Jersey. Considering the identified clusters, further analysis can be conducted to determine if operating expenses or capital expenses should be evaluated or adjusted in effort to control costs and infuse equity in the payment system. Additional data can be collected to develop a predictive modeling application. The predictive model can be leveraged to predict future expenses and billings. Artificial neural networks and decision trees can also be a part of the predictive modeling strategy.

I. INTRODUCTION

Data are being released that show significant variation across the country and within communities in what medical providers charge for common hospital services. These data include information comparing the charges for the 100 most common inpatient services and 30 common outpatient services. The Department of Health And Human Services (DHHS) is looking to gain insight into the Inpatient Prospective Payment System (IPPS). IPPS is the Medicare payment system for hospital services; based on diagnosis related groups (DRG’s). The DHHS has requested an analysis which identifies variations in hospital service charges among common patient services. The analysis will provide detailed information on characteristics of the dataset. The insight will help DHHS determine if base operating
and capital rates paid to medical providers should be adjusted.

II. DATA MINING APPLICATIONS

The extensive amounts of data generated by healthcare transactions are both complex and voluminous. The complexity and size of the data are unlikely to be managed and processed by traditional data analysis methods. Data mining provides the methodology and technology to discover patterns, trends and relationships within complex healthcare data. Data mining can take this complex and voluminous healthcare data and identify insights which will improve decision making.

Several factors have increased the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse has led to multiple healthcare providers leveraging data mining tools to decrease their losses, reduce healthcare fraud and detect abuse. Another contribution factor for the increase in data mining applications in healthcare is healthcare organizations are under heightened pressure to make sound decisions based upon the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better positioned to meet their long term needs (Koh & Tan).

III. DATA MINING THEMES

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data (Tan, Kumar, Steinbach). Data mining is focused specifically on processes and algorithms that discover patterns from raw data. An efficient way to discover such patterns from raw data is by way of data mining themes. Data mining tasks can be classified into themes such as classification, regression, association rule learning, clustering, and anomaly detection, sequential pattern mining and artificial neural networks.

Classification analysis uses historical data patterns to predict behavior of new data. An example of such a prediction model is to identify customers who are likely to respond to a given offer. A regression analysis attempts to predict the numerical value of some variable. The predicted value of home prices on the East Coast during the next five years is an example of a regression analysis. Clustering attempts to group individuals in a population together by their similarity. Grouping individuals by political associations, gender and income is an example of a clustering analysis. Association rule learning identifies relationships between attributes within a dataset. A market analysis identifying commonly purchased items is a type of association rule learning. Anomaly detection helps identify hard to detect patterns in a large data set. Credit card companies and the IRS use such analysis to detect fraud. An artificial neural network is a connectionist computational system. It is a complex adaptive system. Artificial neural networks assist
doctors with their diagnosis by analyzing reported symptoms and or image data such as MRI’s or X-rays.

IV. DATA UNDERSTANDING

The analysis is based upon a dataset retrieved from data.cms.gov. The dataset represents a provider level summary of Inpatient Prospective Payment System (IPPS) discharges, average charges and average Medicare payments for the Top 100 Diagnosis-Related Groups (DRG). The dataset contains 163,065 instances and 11 attributes. The dataset contained a mixture of nominal and numeric values or discrete and continuous attributes. The attributes are defined as follows:

DRG Definition: Definition of diagnosis.

Provider ID: Provider Identifier billing for inpatient hospital services.

Provider Name: Name of provider.

Provider Street Address: Street address in which the provider is physically located.

Provider City: City in which the provider is physically located.

Provider State: State in which the provider is physically located.

Provider Zip code: Zip code in which the provider is physically located.

Provider Region: Region in which the provider is physically located.

Total Discharges: The number of discharges billed by the provider for inpatient hospital services.

Average Covered Payments: The provider's average charge for services covered by Medicare for all discharges in the DRG. These will vary from hospital to hospital because of differences in hospital charge structures.

Average Total Payments: The average of Medicare payments to the provider for the DRG including the DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included are co-payment and deductible amounts that the patient is responsible for.

V. DATA PREPARATION

Upon initial review of the dataset retrieved from data.cms.gov, there were no issues with data quality. The dataset was checked for missing values, duplicate data and outliers. There were no issues related to missing, duplicate or outliers within the data. Attributes such as Provider ID and attributes related to street address were excluded from the analysis.

VI. DATA MINING ALGORITHM

Cluster analysis groups data objects based on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar or related to one another and different from or unrelated to the objects in other groups. The greater the similarity or homogeneity within a group and the greater the difference between groups leads to a better or more distinct clustering. Clustering is useful in data concept construction, pattern detection and unsupervised learning processes.

The most popular centroid-based clustering algorithm is called k-means clustering. The k-means algorithm starts by creating k-initial cluster centers and assigns a centroid for each cluster. Each data point will be compared to a centroid and based upon the distance between the data point and centroid point, the data point is assigned to the closest centroid. The process iterates until the clusters remaining constant.

Considering the structure of the dataset and the objective of the analysis, clustering was selected as an appropriate data mining algorithm. I was in search of a richer IPPS data file which included discrete binary attributes or an IPPS data file for another time period. A more robust or additional dataset would have allowed for additional analysis. Completing a cluster analysis with a more robust data file using the SimpleKMeans algorithm in WEKA, would have resulted in a dataset that assigned each record to a cluster. This ancillary datasource would have been a qualifying dataset to be used for a classification model. If the dataset retrieved from data.cms.gov contained such data; I would have been able to apply the J48 algorithm to the resulting dataset in effort to develop a predictive model. So in essence, I would have used the original IPPS dataset to cluster data into similar/dissimilar groups. I would
have then used the resulting dataset that assigned each record to a cluster for the basis of a classification analysis. The classification analysis can be used to make a prediction of unknown future records. This approach would have been a combination of a supervised and unsupervised learning analytics. Considering the structure of the dataset retrieved from data.cms.gov, the classification algorithms were not available in WEKA. Thus a cluster analysis was applied to the IPPS dataset.

VII. RESULTS AND ANALYSIS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>DRG Definition</th>
<th>Provider State</th>
<th>Provider Region</th>
<th>Total Discharges</th>
<th>Average Covered Charges</th>
<th>Average Total Payments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>HEART FAILURE &amp; SHOCK W/ CC</td>
<td>NY</td>
<td>East Long Island</td>
<td>62</td>
<td>$29,810</td>
<td>$11,220</td>
</tr>
<tr>
<td>1</td>
<td>MISC DISORDERS OF NUTRITION, METABOLISM, FLUIDS/ELECTROLYTES W/O MCC</td>
<td>FL</td>
<td>Orlando</td>
<td>41</td>
<td>$21,244</td>
<td>$5,821</td>
</tr>
<tr>
<td>2</td>
<td>CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/ O/C/C/MCC</td>
<td>PA</td>
<td>Philadelphia</td>
<td>35</td>
<td>$32,253</td>
<td>$8,120</td>
</tr>
<tr>
<td>3</td>
<td>SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W/O MCC</td>
<td>CA</td>
<td>Los Angeles</td>
<td>36</td>
<td>$61,401</td>
<td>$11,656</td>
</tr>
<tr>
<td>4</td>
<td>MAJOR SMALL &amp; LARGE BOWEL PROCEDURES W/ MCC</td>
<td>MD</td>
<td>Baltimore</td>
<td>37</td>
<td>$77,372</td>
<td>$25,517</td>
</tr>
<tr>
<td>5</td>
<td>PERMANENT CARDIAC PACEMAKER IMPLANT W/ CC</td>
<td>NJ</td>
<td>Camden</td>
<td>48</td>
<td>$62,556</td>
<td>$14,575</td>
</tr>
</tbody>
</table>

**Cluster 0: Heart Failure & Shock**: Medical providers billing for inpatient hospital services tend to be in East Long Island NY. Medical providers in this region tend to bill for 62 inpatient hospital discharges. On average, the providers charge $29,810 for services covered by Medicare. Medical providers receive an average of $11,220 in Medicare payments.

**Cluster 1: Nutrition & Metabolism**: Medical providers billing for inpatient hospital services tend to be in Orlando FL. Medical providers in Orlando tend to bill for 41 inpatient hospital discharges. On average, the providers charge $21,244 for services covered by Medicare. Medical providers receive an average of $5,821 in Medicare payments.

**Cluster 2: Pulmonary Disease**: Medical providers billing for inpatient hospital services tend to be in Philadelphia PA. Medical providers located in Philadelphia tend to bill for 35 inpatient hospital discharges. On average, the providers charge $32,253 for services covered by Medicare. Medical providers receive an average of $8,120 in Medicare payments.

**Cluster 3: Severe Sepsis**: Medical providers billing for inpatient hospital services tend to be in Los Angeles CA. Medical providers in this region tend to bill for 36 inpatient hospital discharges. On average, the providers charge $61,401 for services covered by Medicare. Medical providers receive an average of $11,656 in Medicare payments.

**Cluster 4: Bowel Procedures**: Medical providers billing for inpatient hospital services tend to be in Baltimore MD. Medical providers in Baltimore tend to bill for 37 inpatient hospital discharges. On average, the providers charge $77,372 for services covered by Medicare. Medical providers receive an average of $25,517 in Medicare payments.

**Cluster 5: Cardiac Pacemaker**: Medical providers billing for inpatient hospital services tend to be in Camden NJ. Medical providers in Camden NJ tend to bill for 48 inpatient hospital discharges. On average, the providers charge $62,556 for services covered by Medicare. Medical providers receive an average of $14,575 in Medicare payments.
The characteristics of medical providers billing for inpatient hospital services vary. Specifically, the characteristics of providers with the highest average billing to Medicare tend to be located in Baltimore MD and also have one of the lower inpatient hospital discharges. Naturally, the medical providers with the highest average covered charges receive on average the highest IPPS payments. However, medical providers in Baltimore tend to receive 33% of what is billed to Medicare. Medical providers in East Long NY tend to receive a higher percentage of what is billed. Providers in East Long Island receive 38% of what is billed to Medicare or IPPS payments. Medical providers in Los Angeles tend to have one of the lower inpatient hospital discharges and also tend to receive the lowest % of IPPS payment. The percent of what medical providers in Los Angeles bill to Medicare vs. what they receive is 18%.

The DHHS can use the clustering data to assess the significant variation across the country and within communities in what medical providers charge for common services. The DHHS can further analyze the data to determine if the base operating and capital rates paid to hospitals should be adjusted in effort to control costs and reduce the variation of what medical providers charge for services. Based upon the analysis, cost adjustments can be made on the local, regional and national levels.

VIII. CONCLUSION

The analysis identified 6 clusters based upon attributes unique in terms of IPPS characteristics. The analysis clustered 6 medical procedures by region and provided metrics such as number of discharges and average Medicare payments to medical providers. Further analysis can be conducted to determine if operating expenses or capital expenses (significant variables that drive the IPPS payment structure) should be evaluated in effort to control costs and infuse equity within the payment system.

IX. FUTURE WORK

To continue work on this project, the next step will be to collect additional data which will allow for a predictive modeling application. The predictive model can be designed to predict future expenses and billings. An evolved phase of the analysis can be to also predict healthcare fraud. A model can also be developed to estimate or predict the length of time a patient will stay in a hospital. Artificial neural networks and decision trees can also be a part of the predictive modeling strategy.

REFERENCE

