Domain background

The US health system is mired with a high cost of medical care delivery and poor performance when compared with other developed countries. Post-operative complications significantly add a significant financial extra burden to the patients, the hospitals and the third-party payers, not to mention the physical and emotional damage inflicted on the patients.

I am a hospitalist physician and I care for such patients on a daily basis, I have worked for eight years now in non-profit hospitals where resources have not always been available.

Problem statement

There are different types of hospitals ownership in the US since it does not have a public health system like the rest of other western countries. I was curious to find out whether post-operative complications were more prevalent in certain type of hospitals of a given ownership compared to other types of hospitals with a different type of ownership. There are the state and federal owned hospitals, different types of non-profit hospitals and private or proprietary hospitals. Given that resources and workflow may vary between these different facilities, checking whether some type of hospitals have a better post-surgical outcome is worthy an investigation; since this will give us an insight on the ideal type of ownership of a hospital where the rate of post-operative adverse outcomes is less compared to others.

Datasets and inputs

Medicare compiled a set of different data tables on data.medicare.gov. These datasets are readily available to the public. I tried out many datasets for more two months to have a general feeling of what they entail, then I settled on two promising datasets. I downloaded the data as a csv format on my computer and then used pandas and numpy to analyze them in a Jupyter notebook, the two datasets I finally used were:

"Complications and Deaths Hospital" and *"Hospital_General_Information".*

I arranged the first dataset "Complications and Deaths Hospital" in a frame (after using "pivot_table" and other methods) where subsequently each row corresponded to the identity of one hospital and the columns/features corresponded to the score of four different post-operative complications. The features were:

1. Surgical wound dehiscence (meaning the wound doesn't properly close and it remains open several weeks after surgery).
2. Post Op Sepsis (meaning infectious complications following a surgery).
3. Joint Replacement complications.
4. Post Op Venous Thromboembolism (meaning blood clotting following surgery).

There were initially 4812 instances or hospitals. So, the initial data shape was (4812,4). I set the index for this dataset to "Provider ID" a unique identifier for each hospital.

The second dataset "Hospital_General_Information" had also 4812 rows but I only used two columns/features:

1. Provider ID – each hospital has a unique ID.
2. Hospital Ownership – this feature states who owns the hospital (government, non-profit, proprietary…)

Project design

I first selected the four features columns I mentioned above from the dataset "Complications_and_Deaths_Hospital", then I dropped all the rows with no values. I was left with a dataset of around 2043 rows. I then plotted all the dataset in a scatter matrix to learn about the distribution of each feature. All the features shared a distribution close to a Gaussian's distribution.

Then I removed the outliers in the top and bottom 2.5% of the data. Each row will be a single ID corresponding to the hospital ID in the dataset along the with four features mentioned above. The data shape after removing the outliers became (1941,4).

I chose to proceed with unsupervised learning to break up this dataset into different clusters anticipating that a cluster might stand out from the rest in terms of better performance (less post op complications).

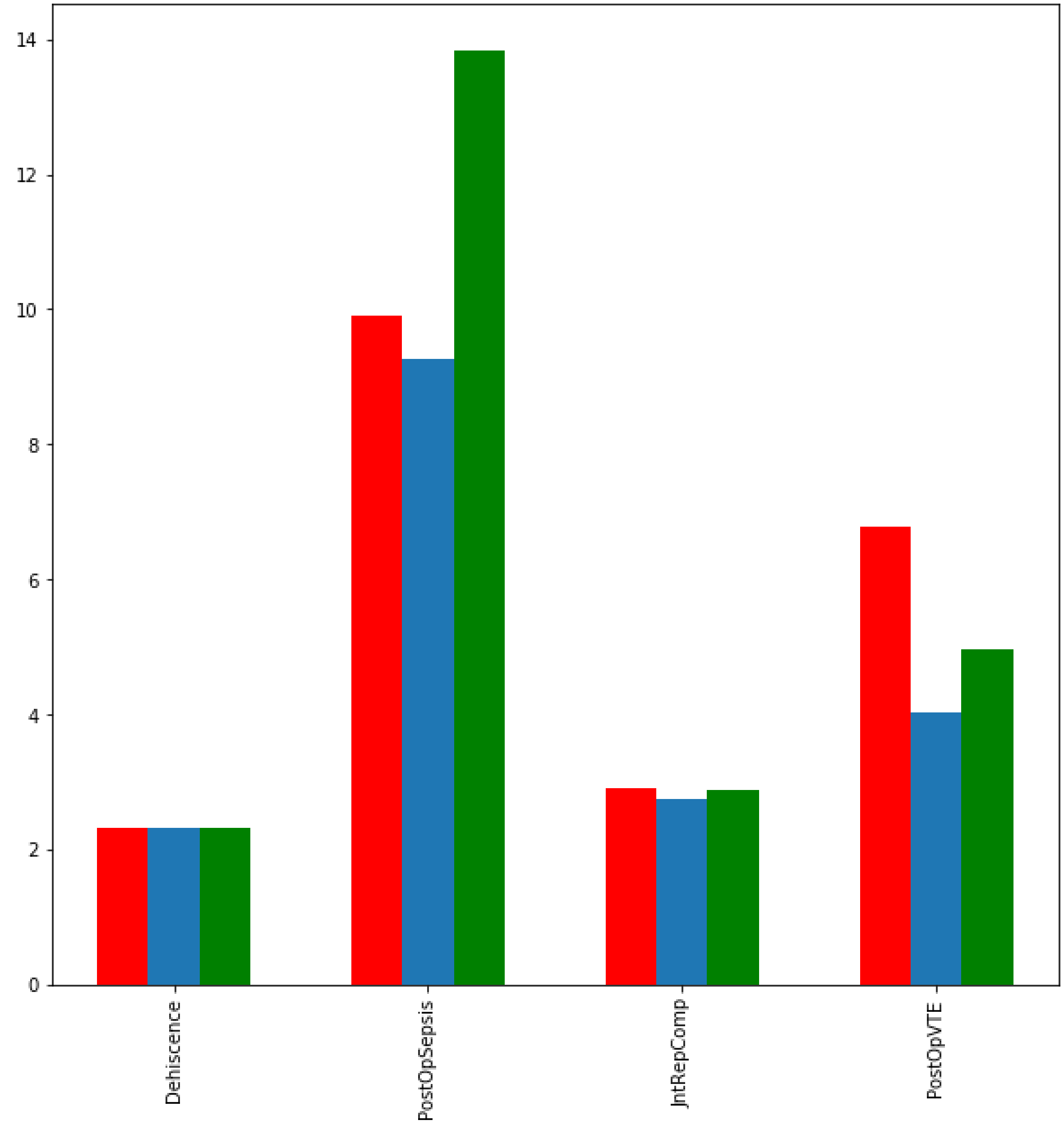
I used two clustering classes (MeanShift and KMeans) from the sklearn library.

MeanShift returned 4 clusters.

The silhouette scores of KMeans for n=2,3,4,5 were 0.39, 0.33, 0.31 and 0.29, respectively.

I tried KMeans with two, three and four clusters and plotted them as below, I ended up selecting three clusters for the project since it's between 3 and 4 and because I could discern a cluster that stands out from the rest in all of the features.

I calculated the mean (metric used) complication score for each feature in each cluster and then plotted them side by side like in the image below in order to discern which cluster performed the best.



Then, I used the second dataset to calculate how many hospitals in each category (government, non-profit and proprietary hospitals) exist in the best performing cluster (blue in the figure above) compared to the number of each category of those hospitals in the other two clusters (red and green grouped together). I could do the math because “Provider ID” is present in both tables and uniquely identified each hospital.

I ended up with the contingency table below:

	Gov Hospitals	Non-Profit	Proprietary
Best Cluster (blue)	135	702	204
Other Clusters	97	647	145

Ratio of Proprietary hospitals in good cluster 0.19

Ratio of Proprietary hospitals in other clusters 0.16

Ratio of Government hospitals in good cluster 0.13

Ratio of Government hospitals in other clusters 0.11

Ratio of Non-Profit hospitals in good cluster 0.67

Ratio of Non-Profit hospitals in other clusters 0.73

Solution statement

With the contingency table above, I could do a Chi square test for independence of variables between the two variables (Cluster's performance and Ownership of the hospital).

The chi test returned was 6.51 with a corresponding p-value of 0.038 (less than 5%). That indicated that these two variables were not independent. In other word, the presence of more Government and Proprietary hospitals and less non-profit hospitals in the good cluster may explain the lower rate of post-surgical complications.

The good cluster had a mean complication score of 22.96 compared to 18.32 in the other clusters. That is equivalent to 20.21% reduction rate in post-operative complications.

There was an article published in pubmed (a library of the National Institute of Health-NIH) by the title "Postoperative complications and implications on patient-centered outcome" which stated that Post-operative complications could reach a rate as high as 30% (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3637983).

Another article published in pubmed "The impact of complications on costs of major surgical procedures: a cost analysis of 1200 patients" (https://www.ncbi.nlm.nih.gov/pubmed/21562405) stated that the average cost per patient in uncomplicated post-operative course was $27,000 and could reach up to $159,000 when post-operative complications arose.

So in a model with more Government and Proprietary hospitals the complications rate could fall from that max of 30% to 24% when factoring the 20.21% reduction rate of the outperforming cluster and the cost per complication per patient would drop from $159,000 to $125,000.

Conclusion

It is difficult to explain the findings of the study above. From my personal experience, working as a physician in non-profit hospitals, I can speculate that a lot of non-profit hospitals are in unprivileged locations and are serving a patients' demographic of low income that lacks adequate insurance coverage. Therefore, the patients tend to be sicker and/or the hospitals may not have on board all the right specialists needed to perform certain procedures.

Proprietary hospitals tend to be located in middle and upper scale suburbs where most of the patients have adequate insurance coverage and tend to be less sick than patients in low income geographic areas and those facilities tend to have access to most of the specialists.

Government hospitals like the VA are usually staffed with all the specialists needed and their patients have VA benefits and access to outpatient medical care all the time.