Definition

Project overview

Every health system struggles to deliver a safe healthcare to its patients with an affordable cost and low rate of complications. I am personally a physician and I have always wondered how to create a safe environment and eliminate adverse side effects that occur in the hospital. There have been several recommendations made over the past several years by professional societies like the *World Health Organization* surgical safety checklist in 2007, the recommendation by the *Institute of Medicine* in its paper "Crossing the Quality Chasm" in 2005 to use information technology to store patient specific clinical information to exchange it among providers.

Problem statement

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, finding a model to reduce post-operative complications or finding what accounts for increased rate in complication have clearly health and financial benefits for the whole public. Since the complications scores were numeric with no well-defined categorical or numeric target, I chose to use *Unsupervised Learning* to tackle this problem and break up the post-op complications data into different clusters and check what characteristic set apart the clusters with low complications rate from others with higher complications rate.

Metric

In the dataset compiled by the "Department of health and human services", there is a score for post-surgical complications for each hospital in the country. I used

the *mean* score for some of those complications to break up the hospitals into ones with good performance and ones with poor performance. I chose the mean because most of those variables follow a *normal* distribution.

Then, I use the chi-square to see if there is a dependence between "performance" and type of "ownership" (Government hospital, Proprietary or Non-profit).

Data Exploration and Preprocessing

I utilized the dataset "Complications and Deaths Hospital" found at this link https://data.medicare.gov/Hospital-Compare/Complications-and-Deaths-Hospital/ynj2-r877, it is available to the public. I downloaded this dataset in early September 2017, so there might have been some updates since then.

*N.B: the datasets I used and uploaded to Github were the original datasets that I downloaded from data.medicare.gov in early September 2017. There were **no** missing rows; although looking at "Provider ID" of 10001 and then followed by the "Provider ID" of 10005 made it look like as if rows were missing.*

After some data wrangling, I was left with 4812 rows each corresponding to one hospital and 17 columns of medical and post-surgical complications scores. The scores were numerical values. Most of those features had normal distribution except for a few. I was only interested in the features related to post-surgical complications. These features were the following:

1. *"A wound that splits open after surgery on the abdomen or pelvis".*
2. *"Blood stream infection after surgery".*
3. *"Deaths among Patients with Serious Treatable Complications after Surgery".*
4. *Rate of complications for hip/knee replacement patients*
5. *Broken hip from a fall after surgery*
6. *Serious blood clots after surgery*

An analysis of the number of rows of NaN values for each feature:

Death rate for CABG: **3778**

Deaths among Patients with Serious Treatable Complications after Surgery: **2998**

Blood stream infection after surgery: **2513**

A wound that splits open after surgery on the abdomen or pelvis**: 2286**

Rate of complications for hip/knee replacement patients: **2052**

Broken hip from a fall after surgery: **1962**

Serious blood clots after surgery: **1846**

I renamed the features into the following (for the sake of programming and graphing):

Death rate for CABG: "**CABGmortality**".

Deaths among Patients with Serious Treatable Complications after Surgery: "**SerCOndPostOpMortality**"

Blood stream infection after surgery: "**PostOpSepsis**".

A wound that splits open after surgery on the abdomen or pelvis**:** "**Dehiscence**".

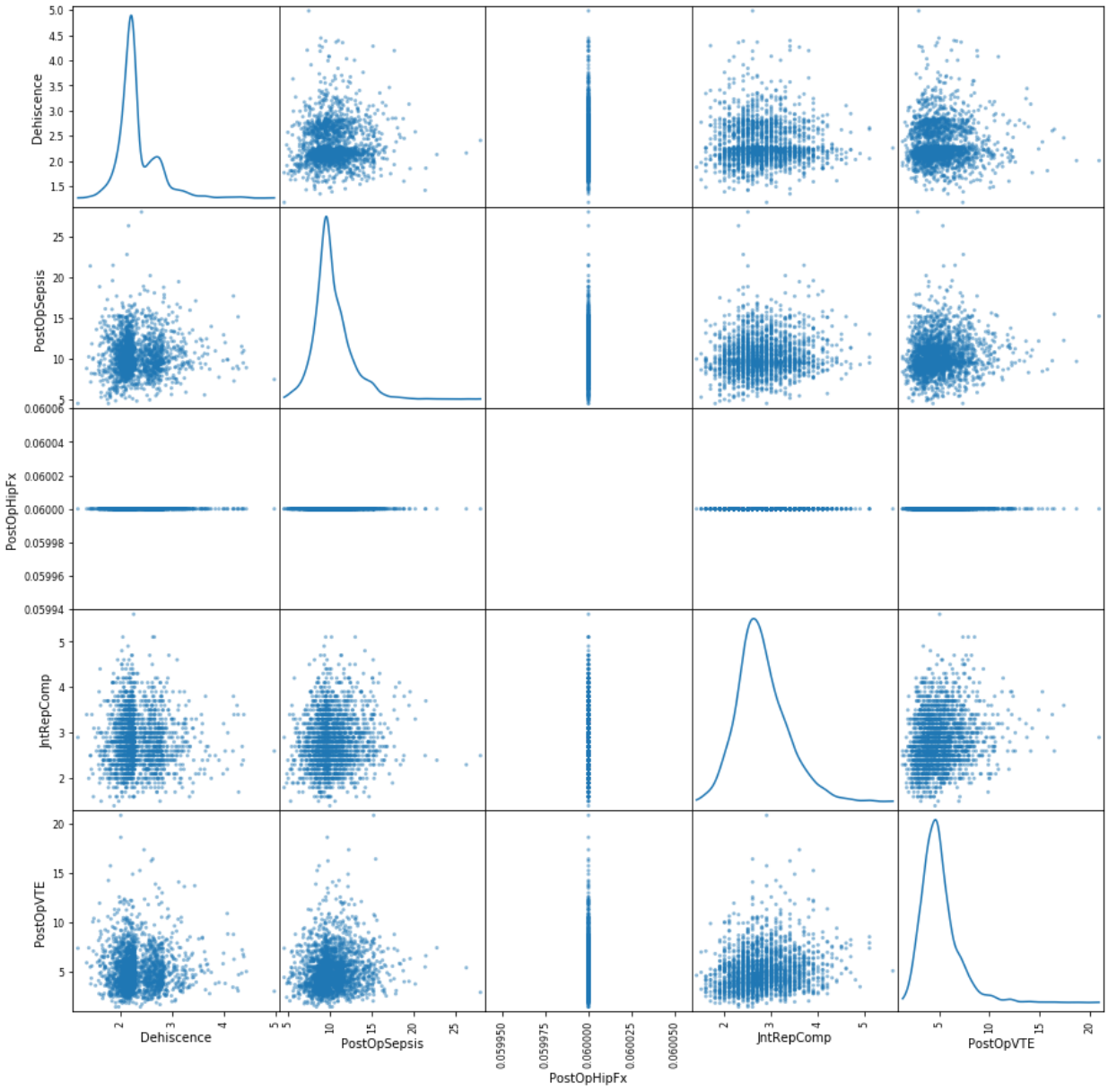
Rate of complications for hip/knee replacement patients: "**JntRepComp".**

Broken hip from a fall after surgery: "**PostOpHipFx".**

Serious blood clots after surgery: "**PostOpVTE**".

Given the high rate of missing values for "CABGmortality" (partly due to the fact that not all the hospitals perform this demanding surgery), I chose to remove it and not include it in the analysis. Same logic applied for "SerCondPostOpMortality" in addition to the fact that this complication is a little ambiguous in terms of what defines a serious complication and what makes it reversible, there was a large gray zone that stands in the way of a clear and precise definition. So, I dropped that feature too.

Below is shown the matrix of the different features distributions.



As shown above, most of the features had a distribution close to a Gaussian one except for "*Broken hip from a fall after surgery*" (PostOpHipFx) which was a uniform distribution of one value, therefore I removed this feature from the analysis. Now after removing "CABGmortality", and "PostOpHipFx" the dataset had 2043 rows with the four features ("Dehiscence", "PostOpSepsis", "JntRepComp", PostOpVTE").

Then, it was time to remove the outliers in the top and bottom 2.5%. That left the dataset with 1941 rows and the four features mentioned above or this final shape **(1941, 4).** A snapshot of the top five rows of the data is shown below:

Measure name	Dehiscence	PostOpSepsis	JntRepComp	PostOpVTE
Provider ID				
10001	2.72	5.72	4.2	3.32
10005	2.11	9.18	3.1	9.63
10011				
10016				
10024				

Algorithms and implementation

One reason why I chose **Unsupervised Learning** is the absence of an obvious target, but the most logical reason is that all the features are numerical and represent in a layman's language "Hospital complications after a surgery", so if I somehow find a target in a different set like say "Hospital safety", running a supervised machine learning algorithm with "Hospital safety" as a target and the above post-operative complications as the features WILL NOT give us any practical informative insight on what could be done to improve "safety". If our algorithm showed a strong negative correlation between "safety" and "JntRepComp"; that would not be a lot of information since anyways our goal is to reduce or find a process to reduce "ALL" post-op complications.

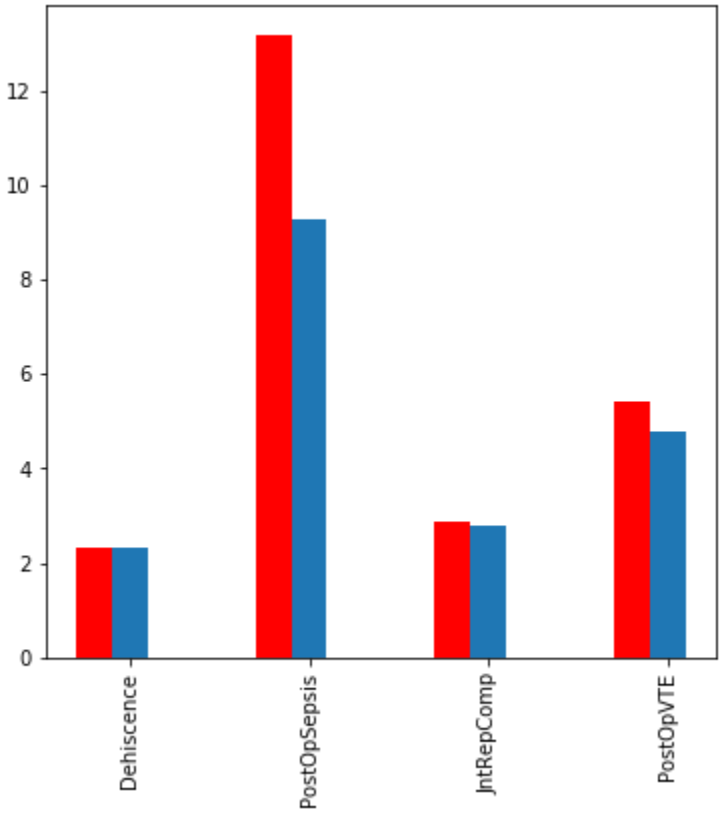
So, Unsupervised learning will break down the set into different clusters and one or several clusters will stand out from the rest by low post-op complications scores. Identifying "Good performing clusters" and see what sets them apart from the rest by associating them with another "characteristic/feature" from another dataset where this *"characteristic/feature"* is *more prevalent* in the *good* performing cluster compared to the *poor* performing clusters.

I will give a brief description of Unsupervised Learning and its underlying alorightms like KMeans and MeanShift. It is a process where linear algebra is used to calculate the "distances" between the different instances and group the instances with the "minimal distances" together in clusters. Using hierarchical (where the algorithm itself determines the number of clusters) versus non-hierarchical clustering (where we preset the number of clusters in the parameters

of the algorithm) will be the next challenge. So first, I tried hierarchical with MeanShift from the sklearn library. The optimal number of clusters returned was 4. When I analyzed this cluster's labels count, I found (**1189** instances with *label 0*, **50** instances with *label 1*, **1** instance with *label 2* and **1** instance with *label 3*). Obviously, MeanShift would not be the optimal algorithm for this problem.

So, I moved to KMeans (non-hierarchical clustering) where I calculated the silhouette score to find the optimal number of clusters for n=2, 3, 4 and 5. The silhouette scores were the following: (*n=2, Score=0.39*; *n=3, Score=0.33*; *n=4, Score=0.31*; *n=5, Score=0.29*). So, two clusters seemed to be the ideal way to proceed.

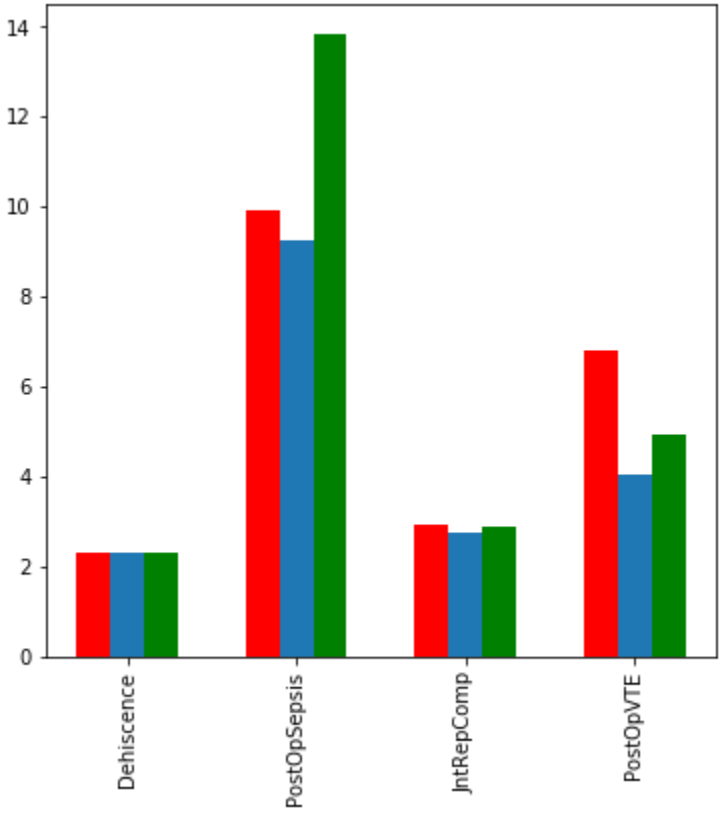
Now, I calculated the mean complication score for each feature in each cluster and graphed them sided by side like shown below:



The *red cluster (poor performer)* totaled 512 samples and the *blue cluster (good performer)* totaled 1429 samples.

Refinement

Since the blue cluster was three-fold the size of the red, I chose to increase the number of clusters from n=2 to n=3 so that the good cluster's size would be reduced and that way it would stand out better from the rest. Using KMeans again with n=3 produced the bar chart below:



Here, I obtained a blue *"good performing cluster"* of size *1045 samples* and another two *"poor performing clusters"* (red and green) of a cumulative size of *896 samples*. I felt that was a better clustering model compared to the two-clusters method.

Now, I moved to the second dataset "Hospital General Information" also publicly available on data.medicare.gov at this link https://data.medicare.gov/Hospital-Compare/Hospital-General-Information/xubh-q36u. There each row corresponded to a hospital and one of the columns specified the hospital's type of ownership (Government, private….). A snapshot of the first few rows showed:

Provider ID	**Hospital Ownership**	**Mortality national comparison**
10001	Government - Hospital District or Authority	Same as the national average
10005	Government - Hospital District or Authority	Below the national average
10006	Government - Hospital District or Authority	Below the national average
10007	Voluntary non-profit - Private	Same as the national average

Those were the types of ownership:

[*'Government - Federal', 'Government - Hospital District or Authority', 'Government - Local', 'Government - State', 'Physician', 'Proprietary', 'Tribal', 'Voluntary non-profit - Church', 'Voluntary non-profit - Other', 'Voluntary non-profit - Private'*]

After some data wrangling, I grouped the different subgroups of "government" hospitals together, I grouped the "non-profit" hospitals in another group and likewise for the "proprietary" hospitals. Using "Provider ID", I was able to calculate the number of each category of those hospitals in each cluster (the *good performer "blue"* on one side, the *poor performers "red" and "green"* on the other side).

	Government Hospitals	**Non-Profit Hospitals**	**Proprietary Hospitals**
Best Cluster (blue)	135	702	204

Other Clusters	97	647	145

The contingency table above showed two variables ("Cluster Performance" and "Type of ownership"). Now, we could analyze it with a Chi square test for independency of variables. Running Chi2 contingency test in python resulted in a **chi statistic of 6.51** and an **"alpha" value of 0.038**. That meant that those variables are likely dependent, and if we looked at the expected values below:

Ratio of Proprietary hospitals in *Blue cluster* **0.20** - Expected Ratio**: 0.18**

Ratio of Proprietary hospitals in *Other clusters* **0.16** - Expected Ratio: **0.18**

Ratio of Government hospitals in *Blue cluster* **0.13** - Expected Ratio: **0.12**

Ratio of Government hospitals in *Other clusters* **0.11** - Expected Ratio: **0.12**

Ratio of Non-Profit hospitals in *Blue cluster* **0.67** - Expected Ratio: **0.70**

Ratio of Non-Profit hospitals in *Other clusters* **0.73** - Expected Ratio: **0.70**

Based on the ratios above, we could say with a confidence of 96% that *Government* and *Proprietary Hospitals* are *more prevalent in the cluster with lower post-operative complications. Non-Profit Hospitals* are *more prevalent in the cluster with higher post-operative complications.*

Obstacles and pitfalls

Obviously, it was not a straightforward process to find some relation between hospitals' ownership and post-operative complications. It was a trial and error process where I tried many other variables. I tried to find a relationship between hospitals with low complications scores **and** hospitals' efficiency of use of electronic medical records and/or use of a preoperative checklist; hospitals with low complications scores **and** readmission rate (readmission rate has become recently a major concern since it increased cost of care and hospitals could be

financially penalized if their readmission rate was more than 10%); hospitals with low complications score **and** cost of care.

It was a tedious task and to my surprise and disappointment, I couldn't establish a statistically significant relationship until I tried hospitals' ownership. Now choosing the complications features was challenging since there was a total of 17 medical(non-surgical) and surgical complications together. Some complications scores were between 0 and 1 with sometimes a skewed distributions, so log scaling them resulted in negative numbers. With 17 complications it was impossible to discern on a bar chart which cluster performed the best. So, I worked my way down the features number and eventually settled on four surgical complications (which are very common complications in hospitals).

Benchmark model

Finding a benchmark model for validation and doing a head to head comparison was difficult given the nature of the data and project. The dataset "Hospital General Information" which contained the "Hospital Ownership" column also contained a column/feature *"Mortality national comparison"* where each hospital associated with a "Provider ID" had one of those four values for this latter feature:

Above the national average

Below the national average

Not Available

Same as the national average

Note that this dataset did NOT contain the post-operative complications features, that is why I used it as a benchmark model. So here again, I calculated the number of "Government", "Proprietary" and "Non-Profit" in each category of Mortality (Above, Below, Same). I dropped "Not Available". You could see below the resulting contingency table of the benchmark model.

Mortality national comparison	**Above the national average**	**Below the national average**	**Same as the national average**

Hospital Ownership			
Government	24	65	541
Non-Profit	298	211	1717
Proprietary	74	65	449

Ratio of **Government** with **mortality above** national average **0.038**

Ratio of **Government** with **mortality below** national average **0.103**

Ratio of **Non-Profit** with **mortality above** national average **0.13**

Ratio of **Non-Profit** with **mortality below** national average **0.095**

Ratio of **Proprietary** with **mortality above** national average **0.12**

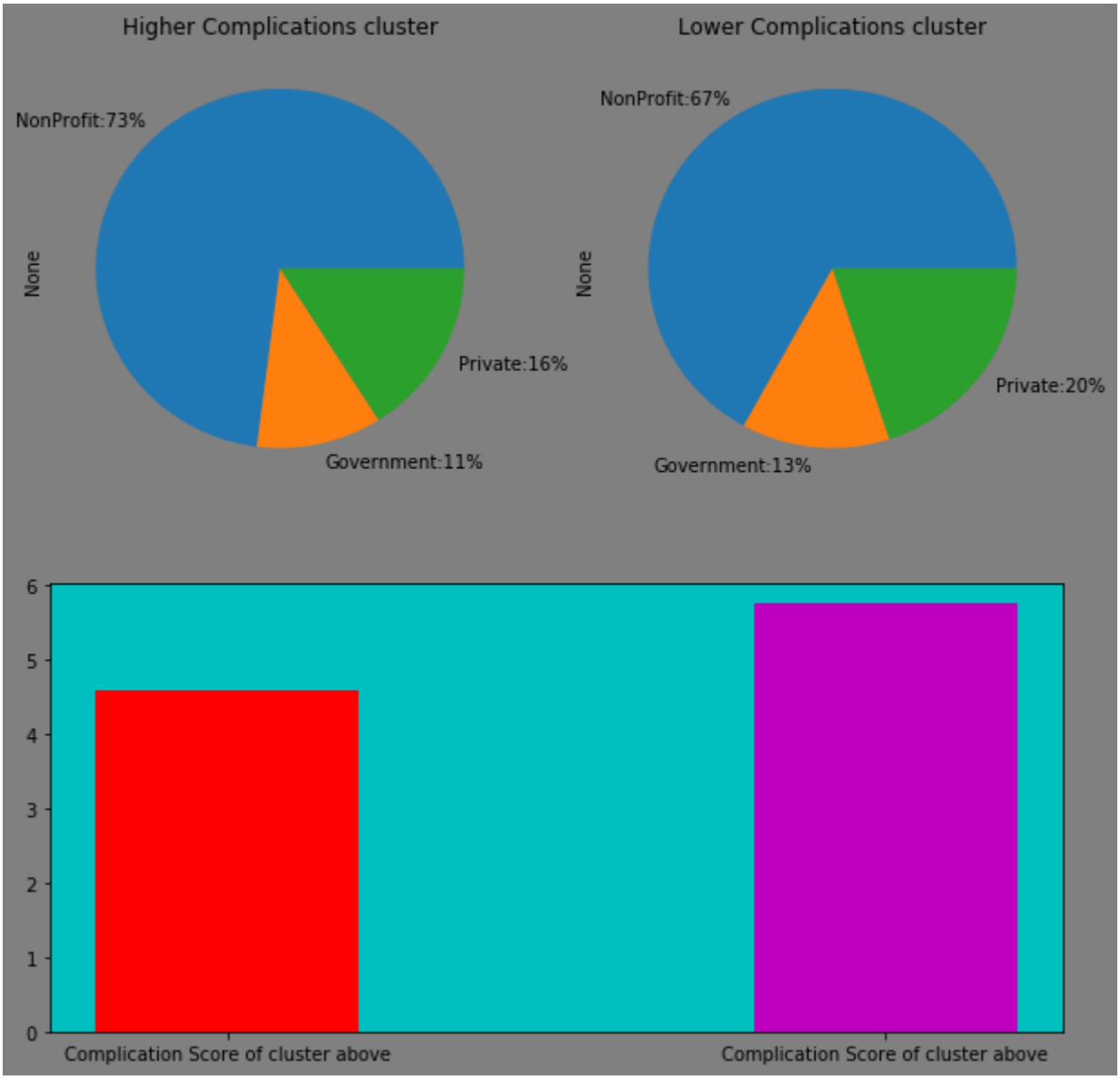
Ratio of **Proprietary** with **mortality below** national average **0.11**

Validation and results

Now it was time to validate this model of dependence between *"Hospitals' Ownernship" and "Post-operative complications"* against the **benchmark model** described in the paragraph above to evaluate the dependence between *"Hospitals' Ownership" and "Hospitals' mortality"*. Running the chi square contingency in python on the table above of the benchmark model yielded a **Chi statistic** of **46.42** and an **alpha value** of **2.00686e-09.** Those findings indicated a high likelihood of dependence between the two variables (type of Hospital Ownership on one side and Mortality on the other side).

Here again, "*Non-profit Hospitals*", showed a lower performance compared to Proprietary and Government hospitals, they had a higher proportion of hospitals with mortality rate above the national average and lower number of hospitals with mortality rate below the national average.

The visualization below should give us an idea about the findings above, where a small change in the ownership share was reflected with a substantial drop in the complication score.



Conclusion and reflections

Reflecting on what could be done to improve or expand on what was done above, I could think of using more complications features including non-surgical ones to include most of the complications (hospital's complications are not only surgical) so we could get a better understanding on what is the underlying factor behind adverse events in the hospitals. Probably more data extending over a longer period could be used to assess whether hospitals with lower complications have a lower cost of care or have better patient satisfaction.

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 throughout the year.