# ORIE 4741 Project Final Report

Kathy Byun, Raye Liu, Jody Zhu

December 13, 2020

## 1 Problem

The goal is to predict the length of stay and total cost of each cancer patient in New York State hospitals from demographics and health data using models. Furthermore, we want to explore whether clustering the patients into "homogeneous" groups can yield better predictions than in the general body. We will use linear regression with quadratic loss, huber loss, and quadratic regularization, low rank model, k-means clustering, and hierarchical clustering.

## 2 Background

Health data is very complicated but can provide beneficial information to both patients and healthcare providers when unlocked. Length of stay and total cost are themselves points of interest for planning ahead in terms of time and finances. Here, we will also rely on them when analyzing our clusters. Homogeneous clusters can provide intelligible insight as shown in research done at John Hopkins on Adjusted Clinical Group actuarial cells, or clusters. For instance, more consistent care can be delivered to patients with the same conditions. A greater number of early diagnoses can be made for individuals without known hereditary predispositions as well. On the hospital management side, well-segmented clusters can be used to help anticipate the high cost and long stay patients.

## 3 Dataset

The dataset used in this project is New York State's Statewide Planning and Research Cooperative System's (SPARCS) Hospital Inpatient Discharges. It contains information about patients discharged from hospitals in New York State in 2012. Some of the fields are race, age group, type of admission, diagnosis, severity index, length of stay, and total charges. To narrow down the scope we are starting with, we will only perform data analysis on cancer patients. The SPARCS_Cancer data has 35,804 rows/examples and 33 columns/features before preprocessing. There are 420 missing entries in Zip Code, 9,438 in Payment Typology 2, 21,636 in Payment Typology 3, and 0 otherwise. As we will see later in this report, these missing values will not impact our results.

## 4 Data Visualization

We generated multiple boxplots to get a better picture of our data. To gauge the feasibility of finding good clusters, we pick a simple grouping by facility ID which treats each facility separately. In Figure 1, the median length of stay varies slightly across hospitals in the Bronx, except for number 1175 (5-7 days vs. 13 days): approximately 75% of its values are greater than 50% of the values from the others. Facility

ID cannot be confirmed as an important feature yet because 1175 might be a special hospital. This is a decent start although our actual clusters are likely to be characterized by a combination of attributes.

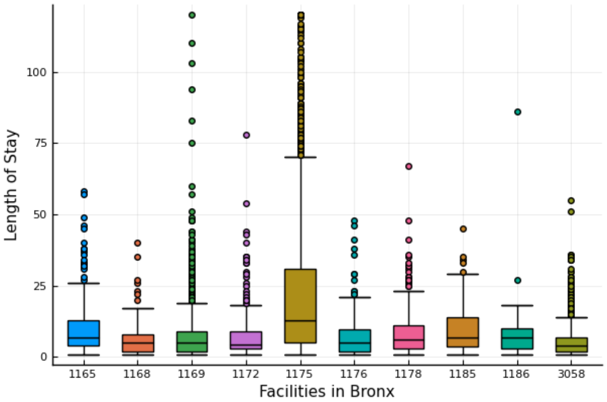


Figure 1: Length of Stay for each Facility in Bronx

Next, we look at APR Risk of Mortality. Prior knowledge tells us that more severe symptoms tend to have higher mortality rates and longer treatments. From Figure 2, this trend is evident in both length of stay and total charges. The differences between the boxplot at each level support the clustering method, and APR Risk of Mortality is likely to play a part as a significant feature to consider.

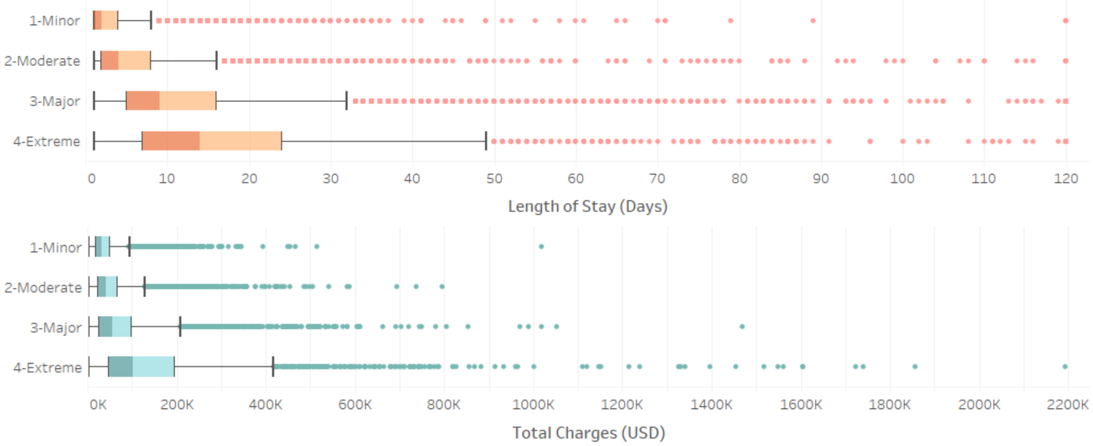


Figure 2: Length of Stay & Total Charges for each level of APR Risk of Mortality

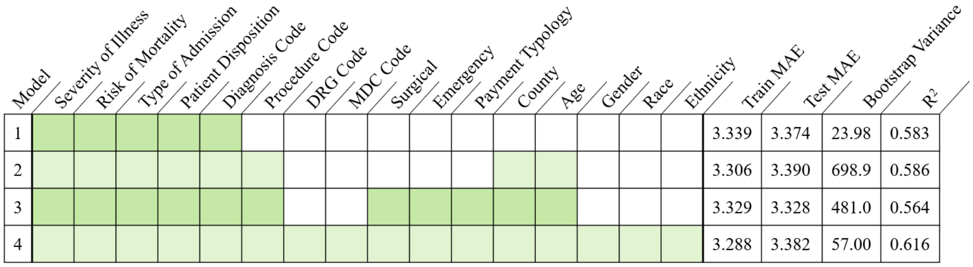
# 5 Data Preprocessing

Our first target feature Length of Stay is continuous. Its values are capped by 120+ which we replaced with 120. The other target feature Total Cost is continuous as well, ranging from 1,562.44 to 2,193,723.15, with a mean of 5,8879.05. APR Severity of Illness and Risk of Mortality are ordinal in nature and rewritten as integers 1 through 4 with 1 as Minor and 4 as Extreme. For Age Group, its 5 groups are translated into ordinal form in increasing age order. The remaining features are categorical, and the nominal values are converted to numbers using one-hot encoding which creates a column for each possible value and puts a 1 in the applicable column, 0 otherwise. This is simplified into a single column for binary categorical features.

# 6 Feature Selection

We use linear regression with quadratic loss to determine which features to include as well as to establish a baseline for later. Our dataset is split into training and testing sets (3:1) so that we can measure our models' accuracy. From the original 33 features, those that have the same value for every example are first removed. So are features that are replicates of each other (i.e. code and description) to avoid collinearity, which inflates variance and lowers model interpretability. The three columns with missing entries fall under this category.

Then, we perform 5-fold cross validation to find the combination of features with the best prediction ability and least likelihood of overfitting. A portion of the results is summarized in Figure 3. To examine how inliers did, mean absolute error is selected for its relative insensitivity to outliers. The second to last column is the expected variance from out-of-sample error calculated using bootstrap of 500 samples of size 5000. High variance is typically associated with overfitting. Lastly, a higher $R^2$ is better since it is a goodness-of-fit measure.



| Model | Severity of Illness | Risk of Mortality | Type of Admission | Patient Disposition | Diagnosis Code | Procedure Code | DRG Code | MDC Code | Surgical | Emergency | Payment Typology | County | Age | Gender | Race | Ethnicity | Train MAE | Test MAE | Bootstrap Variance | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | | | | | 3.339 | 3.374 | 23.98 | 0.583 |
| 2 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | 🟩 | 🟩 | | | | 3.306 | 3.390 | 698.9 | 0.586 |
| 3 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | 3.329 | 3.328 | 481.0 | 0.564 |
| 4 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 3.288 | 3.382 | 57.00 | 0.616 |

Figure 3: Cross-Validation Statistics

We suspected underfitting when taking fewer than 5 features, leading us to increase the number of features and ultimately deciding to include all the ones in the table. The MAE and bootstrap variance are decent. Most notably, its $R^2$ is a 3% improvement compared to the others. We will now transition to discuss models in further detail.

# 7 Models

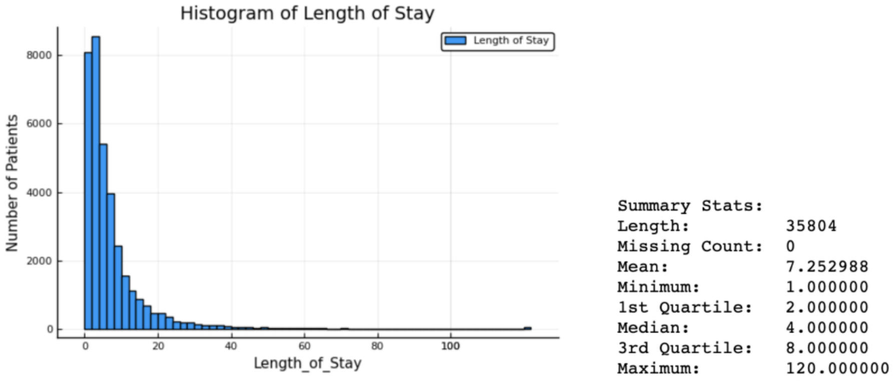
## 7.1 Linear Regression

First, we constructed a model with the quadratic loss function and a small quadratic regularizer, otherwise known as ridge regression. The target feature is length of stay. Although the dataset is right-skewed in Figure 5 for length of stay, it stays rather continuous through the maximum value. Quadratic loss keeps in consideration the right-most points that may or may not be outliers. The regularizer will make the solution unique.

In the case that the large values are outliers, we built a second model with Huber loss and a small quadratic regularizer. Huber loss is more robust by punishing outliers less (absolute value) than errors within a reasonable range (quadratic again). The regularizer serves the same purpose. We observe that the type of regularizer and lambda does not affect the results much if at all.

After 5-fold cross validation, the average test MAE is 3.382 for quadratic loss compared to a higher 3.464 for Huber loss. $R^2$ is 0.616 for quadratic loss versus a lower 0.467 for Huber loss. Furthermore, the mean length of stay is 7.19 for quadratic loss and 5.62 for Huber loss. The actual mean is 7.25 (Figure

4), so the predictions by our first model are overall more reliable.



```
Summary Stats:
Length:         35804
Missing Count:  0
Mean:           7.252988
Minimum:        1.000000
1st Quartile:   2.000000
Median:         4.000000
3rd Quartile:   8.000000
Maximum:        120.000000
```

Figure 4: Length of Stay Statistics

We will use linear regression with quadratic loss for cluster analysis. However, this model is not recommended for real world implementation. The optimal range of $R^2$ is between 0.7 and 0.9, which is not met. Linear regression is also not suitable and will have low accuracy if the data is not linear. There are likely still better models out there.

## 7.2 K-Means Clustering

The algorithm consists of two alternating steps: selecting $k$ numbers of centroids and assigning each data point to its closest centroid. Then, the means of the current clusters become the new centroids and so on until it converges. The parameter $k$ is the number of resulting clusters as well. We used the low rank models package and Python scikit-learn package to solve k-means clustering with the features from feature selection above.

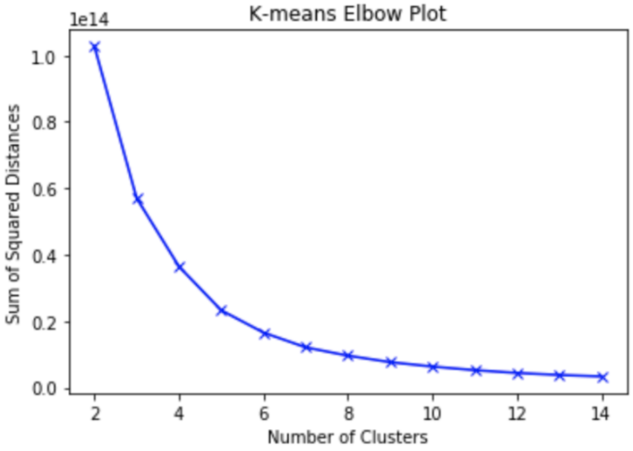
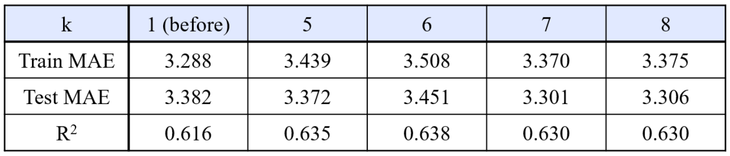


Figure 5: Sum of Squared Distance at each K

Figure 5 shows linear decrease approximately from when the number of clusters $k$ is 6. The value of $k$ where the plot starts to show linear relationship indicates the optimal number of clusters. We proceed to run k-means clustering for $5 \leq k \leq 8$. Our dataset is split 3:1 into training and testing sets. K-means clustering is applied to the training set. Using the model, each patient from the test set is then assigned to the closest training cluster. Finally, predictions are made by linear regression with quadratic loss.



| k | 1 (before) | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|
| Train MAE | 3.288 | 3.439 | 3.508 | 3.370 | 3.375 |
| Test MAE | 3.382 | 3.372 | 3.451 | 3.301 | 3.306 |
| $R^2$ | 0.616 | 0.635 | 0.638 | 0.630 | 0.630 |

Figure 6: Predicting on K Clusters Summary

For all clusters with $k > 1$ in Figure 6, their $R^2$ values are slightly greater than the linear model trained on original data with $k = 6$ boasting the highest. The 0.02 increase in $R^2$ when $k = 6$ indicates that 2% more of the variance in length of stay is explained by the features in the model. The test MAEs for linear models trained on clustered data are slightly smaller than that of the original as well. However, because the difference is small, it is not clear if the improvement is statistically significant.

## 7.3 Hierarchical Clustering

Hierarchical Clustering Analysis generally fall into two types: agglomerative (a ”bottom-up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy) and divisive (a ”top-down” approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy). In our dataset, we use the bottom-up agglomerative approach. Considering the uniqueness of each individual patient, we start with each of them as an individual cluster and then merge the clusters that share commonalities. We use all 35,804 observations and 395 columns (post-feature engineering and scaled). In this part, our goal is to find indicative clustering assignments of the patients and a potential cluster that represents the high-resource-consumption patients. To evaluate resource consumption, we will use Total Charges, Length of Stay, APR Mortality Risk, and APR Severity of Illness these four target variables as reference.

First, we select the optimal number of clusters $k$ using the R packages factoextra and NbClust. Figure 7 on the next page represent the evaluation results based on the Elbow method, the Silhouette method, and gap statistics. The Elbow method suggests an x-intercept of 4, meaning 4 is the optimal number of clusters. Silhouette suggests an optimal number of 2. Gap statistic suggests a cluster of one which is trivial. However, we want a higher level of interpretability, so we will re-evaluate by requiring one cluster of patients to have relatively small cardinality but high consumption of healthcare resources, which was noticed in data visualization. After training the models on 2, 4, 6, 9 and 16 (local maximum points in the Silhouette plots), none of them returned the characteristic cluster we are looking for. Thus, following the trend that the average silhouette width and gap statistics slowly increase after $k = 6$, we picked a large $k$ and trained a 100-cluster bottom-up model. 100 is not too excessive considering there are over 30,000 total observations. We will focus on the 100-cluster model which does return the particular high-resource-consuming cluster. Statistical analysis is preferred over validation for this type of unsupervised learning, so we will examine the quality of the clusters with a combination of standard deviation comparison and external manual interpretation (like for glass-box models).

The whole-set standard deviations of the potential target columns are: Total Charges: 75,993.88; APR Mortality Risk (ranged from 1 to 4): 0.9229; APR Severity of Illness (ranged from 1 to 4): 0.9227; Length of Stay: 11.0596. We calculate the standard deviations within each cluster and obtain the plots in Figure 8 on Page 7 (red line is the whole-set value). We can see a significant decrease in the standard deviation for most of the clusters. 76% of the clusters have a strictly lower standard deviation on Total Charges, 82% on APR Mortality Risk, 91% on APR Severity of Illness, 69% on Length of Stay. This indicates that the clustering method has done a decent job at correctly identifying the patients with less discrepancy and

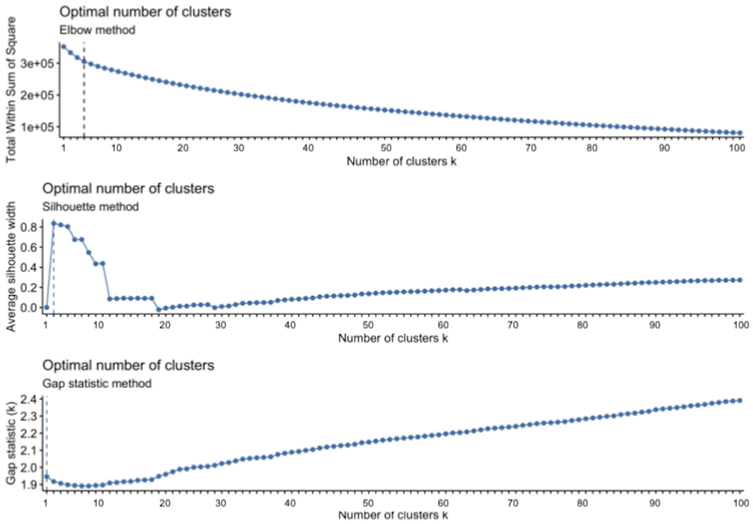
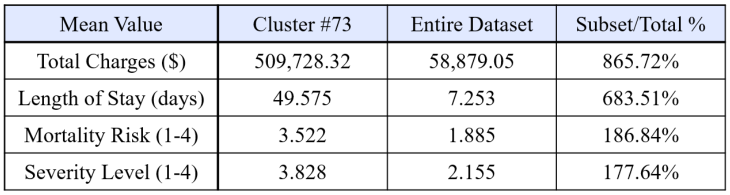


Figure 7: Optimal Cluster Number

assigning them to the same group. On the other hand, the outliers do exist, for instance, Cluster No.72 has a standard deviation drastically greater than the whole-set value. These outlier clusters are generated because their clustering is based on other features like diagnosis on certain APR-DRG, APR-MDC and APR-Procedure codes, and these codes are not directly or crucially influencing the target response variables we are examining. For instance, Cluster No.72 patients share in common that they have all been diagnosed with Mental Diseases and Disorders and over 90% percent of them has been diagnosed with Digestive Malignancy. These two disease categories have severity and mortality risk ranking all the way from the lowest to the highest based on the APR-DRG Weights System, thus it is hard to distinguish those with high mortality risk from those with relatively low mortality risk by only clustering from these two features. Even if most of these outlier clusters are not the "high resource users" we focus on in this project, we should still come up with solution to improve the clustering quality to lower down the standard deviation and eliminate these outliers.

Aiming to find the so-called "High Resource Users (HRU)" group, we calculate the mean value of the target response variables within each cluster and sort them in descending orders to find out those clusters that have high values. It is reasonable to view these four target columns (financial expense, risk of mortality, severity of disease, and length of hospitalization) as representations of the resource usage of a patient in the medical system. By definition, HRUs consist of a small proportion of the population that large proportion of health care spending is incurred by. We can clearly identify the cluster of HRUs, No.73, with a significantly peak in both Total Charges, Severity, Mortality, and Length of Stay.

There are 134 patients in Cluster No.73, 0.37% of the entire population. The statistics of this cluster compared with the whole population is listed in the table of Figure 9.



| Mean Value | Cluster #73 | Entire Dataset | Subset/Total % |
|---|---|---|---|
| Total Charges ($) | 509,728.32 | 58,879.05 | 865.72% |
| Length of Stay (days) | 49.575 | 7.253 | 683.51% |
| Mortality Risk (1-4) | 3.522 | 1.885 | 186.84% |
| Severity Level (1-4) | 3.828 | 2.155 | 177.64% |

Figure 9: Comparison of Resource Consumption between Cluster No.73 and the Whole Population

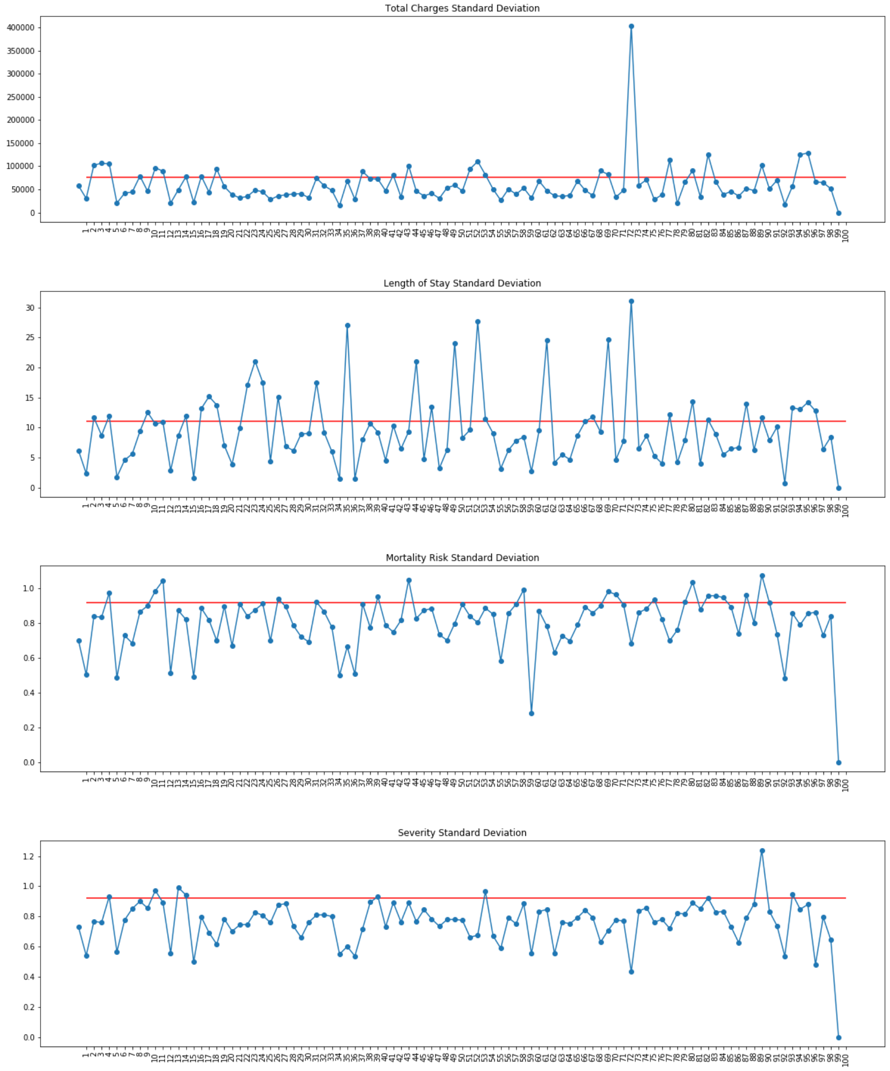
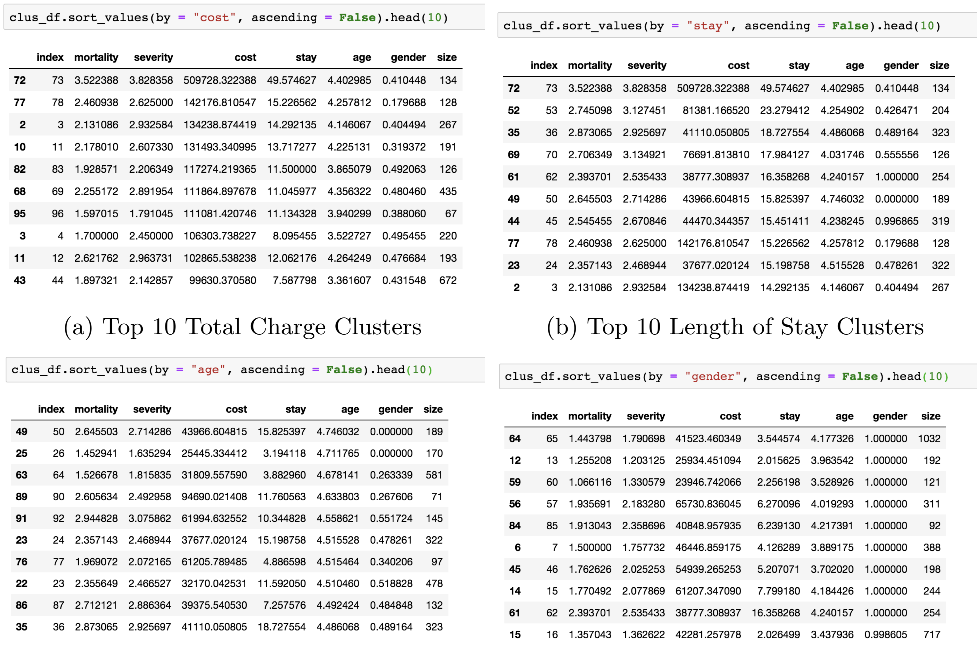


Figure 8: Standard Deviation Comparison

For understanding these statistics and finding out why patients in this clusters have an overwhelmingly larger expense and risk of death compared with other patients, we print out the columns data of this cluster. We find that the commonality these patients share is that they have all been diagnosed with APR-DRG code 4–Tracheostomy with MV 96+ hours with extensive procedure or ECMO, and they have all been through major surgeries (APR Surgical Description). Also, one common diagnosis is respiratory malignancy, which is presumably the cause of the using of ECMO (extracorporeal membrane oxygenation, used in open-heart surgery. It pumps and oxygenates a patient's blood outside the body, allowing the heart and lungs to rest). This lies in accordance for the long stay, high costs, and high mortality since ECMO is very costly to monitor and often it applies to critical patients.

More interesting and less obvious observations are made when we look at the top 10 clusters sorted in descending order (Figure 10) by the target columns as well as age and gender. During data preprocessing, gender was one-hot encoded as 1 for female and 0 for male, and a cluster consists of more males if the value in the gender column is less than 0.5. Amongst the high resource user clusters according to length of stay and total charges, many are majority male. For the longevial clusters, 5 out of 10 clusters (including the top 4) are predominantly female and the rest half is equal in gender distribution. This indicates that females tend to live longer than males.



```
clus_df.sort_values(by = "cost", ascending = False).head(10)
```

| | index | mortality | severity | cost | stay | age | gender | size |
|---|---|---|---|---|---|---|---|---|
| **72** | 73 | 3.522388 | 3.828358 | 509728.322388 | 49.574627 | 4.402985 | 0.410448 | 134 |
| **77** | 78 | 2.460938 | 2.625000 | 142176.810547 | 15.226562 | 4.257812 | 0.179688 | 128 |
| **2** | 3 | 2.131086 | 2.932584 | 134238.874419 | 14.292135 | 4.146067 | 0.404494 | 267 |
| **10** | 11 | 2.178010 | 2.607330 | 131493.340995 | 13.717277 | 4.225131 | 0.319372 | 191 |
| **82** | 83 | 1.928571 | 2.206349 | 117274.219365 | 11.500000 | 3.865079 | 0.492063 | 126 |
| **68** | 69 | 2.255172 | 2.891954 | 111864.897678 | 11.045977 | 4.356322 | 0.480460 | 435 |
| **95** | 96 | 1.597015 | 1.791045 | 111081.420746 | 11.134328 | 3.940299 | 0.388060 | 67 |
| **3** | 4 | 1.700000 | 2.450000 | 106303.738227 | 8.095455 | 3.522727 | 0.495455 | 220 |
| **11** | 12 | 2.621762 | 2.963731 | 102865.538238 | 12.062176 | 4.264249 | 0.476684 | 193 |
| **43** | 44 | 1.897321 | 2.142857 | 99630.370580 | 7.587798 | 3.361607 | 0.431548 | 672 |

(a) Top 10 Total Charge Clusters

```
clus_df.sort_values(by = "stay", ascending = False).head(10)
```

| | index | mortality | severity | cost | stay | age | gender | size |
|---|---|---|---|---|---|---|---|---|
| **72** | 73 | 3.522388 | 3.828358 | 509728.322388 | 49.574627 | 4.402985 | 0.410448 | 134 |
| **52** | 53 | 2.745098 | 3.127451 | 81381.166520 | 23.279412 | 4.254902 | 0.426471 | 204 |
| **35** | 36 | 2.873065 | 2.925697 | 41110.050805 | 18.727554 | 4.486068 | 0.489164 | 323 |
| **69** | 70 | 2.706349 | 3.134921 | 76691.813810 | 17.984127 | 4.031746 | 0.555556 | 126 |
| **61** | 62 | 2.393701 | 2.535433 | 38777.308937 | 16.358268 | 4.240157 | 1.000000 | 254 |
| **49** | 50 | 2.645503 | 2.714286 | 43966.604815 | 15.825397 | 4.746032 | 0.000000 | 189 |
| **44** | 45 | 2.545455 | 2.670846 | 44470.344357 | 15.451411 | 4.238245 | 0.996865 | 319 |
| **77** | 78 | 2.460938 | 2.625000 | 142176.810547 | 15.226562 | 4.257812 | 0.179688 | 128 |
| **23** | 24 | 2.357143 | 2.468944 | 37677.020124 | 15.198758 | 4.515528 | 0.478261 | 322 |
| **2** | 3 | 2.131086 | 2.932584 | 134238.874419 | 14.292135 | 4.146067 | 0.404494 | 267 |

(b) Top 10 Length of Stay Clusters

```
clus_df.sort_values(by = "age", ascending = False).head(10)
```

| | index | mortality | severity | cost | stay | age | gender | size |
|---|---|---|---|---|---|---|---|---|
| **49** | 50 | 2.645503 | 2.714286 | 43966.604815 | 15.825397 | 4.746032 | 0.000000 | 189 |
| **25** | 26 | 1.452941 | 1.635294 | 25445.334412 | 3.194118 | 4.711765 | 0.000000 | 170 |
| **63** | 64 | 1.526678 | 1.815835 | 31809.557590 | 3.882960 | 4.678141 | 0.263339 | 581 |
| **89** | 90 | 2.605634 | 2.492958 | 94690.021408 | 11.760563 | 4.633803 | 0.267606 | 71 |
| **91** | 92 | 2.944828 | 3.075862 | 61994.632552 | 10.344828 | 4.558621 | 0.551724 | 145 |
| **23** | 24 | 2.357143 | 2.468944 | 37677.020124 | 15.198758 | 4.515528 | 0.478261 | 322 |
| **76** | 77 | 1.969072 | 2.072165 | 61205.789485 | 4.886598 | 4.515464 | 0.340206 | 97 |
| **22** | 23 | 2.355649 | 2.466527 | 32170.042531 | 11.592050 | 4.510460 | 0.518828 | 478 |
| **86** | 87 | 2.712121 | 2.886364 | 39375.540530 | 7.257576 | 4.492424 | 0.484848 | 132 |
| **35** | 36 | 2.873065 | 2.925697 | 41110.050805 | 18.727554 | 4.486068 | 0.489164 | 323 |

(c) Top 10 Age Clusters

```
clus_df.sort_values(by = "gender", ascending = False).head(10)
```

| | index | mortality | severity | cost | stay | age | gender | size |
|---|---|---|---|---|---|---|---|---|
| **64** | 65 | 1.443798 | 1.790698 | 41523.460349 | 3.544574 | 4.177326 | 1.000000 | 1032 |
| **12** | 13 | 1.255208 | 1.203125 | 25934.451094 | 2.015625 | 3.963542 | 1.000000 | 192 |
| **59** | 60 | 1.066116 | 1.330579 | 23946.742066 | 2.256198 | 3.528926 | 1.000000 | 121 |
| **56** | 57 | 1.935691 | 2.183280 | 65730.836045 | 6.270096 | 4.019293 | 1.000000 | 311 |
| **84** | 85 | 1.913043 | 2.358696 | 40848.957935 | 6.239130 | 4.217391 | 1.000000 | 92 |
| **6** | 7 | 1.500000 | 1.757732 | 46446.859175 | 4.126289 | 3.889175 | 1.000000 | 388 |
| **45** | 46 | 1.762626 | 2.025253 | 54939.265253 | 5.207071 | 3.702020 | 1.000000 | 198 |
| **14** | 15 | 1.770492 | 2.077869 | 61207.347090 | 7.799180 | 4.184426 | 1.000000 | 244 |
| **61** | 62 | 2.393701 | 2.535433 | 38777.308937 | 16.358268 | 4.240157 | 1.000000 | 254 |
| **15** | 16 | 1.357043 | 1.362622 | 42281.257978 | 2.026499 | 3.437936 | 0.998605 | 717 |

(d) Top 10 Gender Clusters

Figure 10: Sorting Descending by the Target Columns

# 8 Conclusion and Future Work

The variations of linear regression returned decent results, but other models might be better suited given that the data is probably not linear. We were still able to use it to measure the performance of our k-means clusters, which saw slight advancement compared to without clustering. With the method of hierarchical clustering, we were able to find both more homogeneous and interpretable clusters. Standard deviation improved for approximately 70% of the clusters, and we were able to confirm the reliability of the clusters by cross-analyzing the characteristics of patients with external health information. Overall, we are fairly confident in this model and the predictions to be made using it.

Because health data pertains to life and death, there is the danger of our models becoming weapons of math destruction if used incorrectly. The same can be said about corrupting fairness. Doctors should not diagnose or treat patients solely based on our estimates. Hospitals should not shorten care for female patients because of the observations made from the clusters. This is not the purpose of the models in their current form. Clustering does support greater individual fairness by making similar predictions about similar patients, but adjustments are still needed to account for fairness in protected attributes. Until then, the models can aid patients in anticipating their stay and hospital management in staffing and purchase planning.

# 9 References

1. Na, Shi, et al. *Research on k-Means Clustering Algorithm: An Improved k-Means Clustering Algorithm.* 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, doi:10.1109/iitsi.2010.74.

2. Rosella, Laura. *Predicting High Health Care Resource Utilization in a Single-Payer Public Health Care System.* Medical Care, Oct. 2018.

3. Rouzbahman, Mahsa, et al. *Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU?* IEEE, 3 Feb. 2016, doi:10.1109/JBHI.2016.2525731.

4. S. Patel, S. Sihmar and A. Jatain, *A study of hierarchical clustering algorithms.* 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 537-541.

5. *The Johns Hopkins ACG System Technical Reference Guide.* The Johns Hopkins ACG System, Johns Hopkins School of Public Health, Dec. 2009, www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057914.pdf.

6. Udell, Madeleine, et al. *Generalized Low Rank Models.* 2016, doi:10.1561/9781680831412.