ORIE 4741 Final Project Report

Kathy JaYoung Byun, Raye Liu and Jody Zhu

December 13, 2020

# 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.

## 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.

# 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.

## 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.

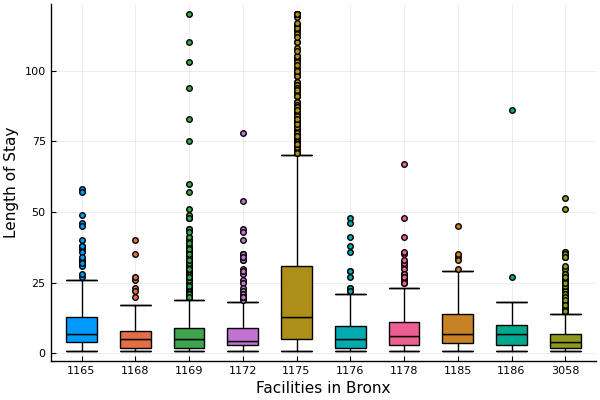


Figure 1: Length of Stay for each Facility in Bronx

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. This trend is evident both in Figure 2 with length of stay and in Figure 3 with total charges, our second target feature. 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.

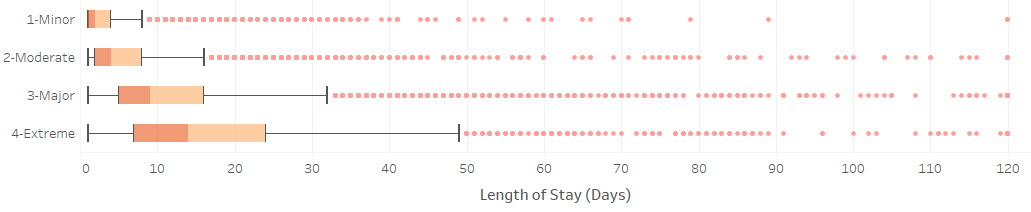


Figure 2: Length of Stay for each level of APR Risk of Mortality

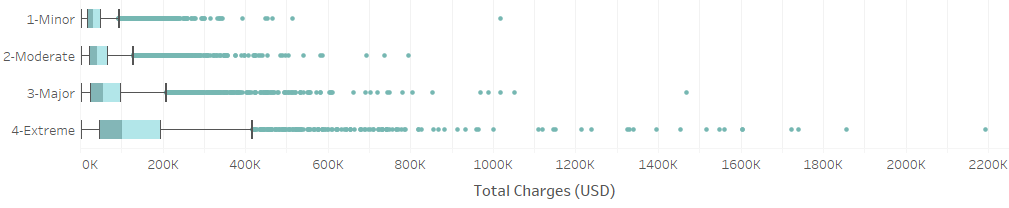


Figure 3: Total Charges for each level of APR Risk of Mortality

## 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.

# 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 4. 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 R2 is better since it is a goodness-of-fit measure.

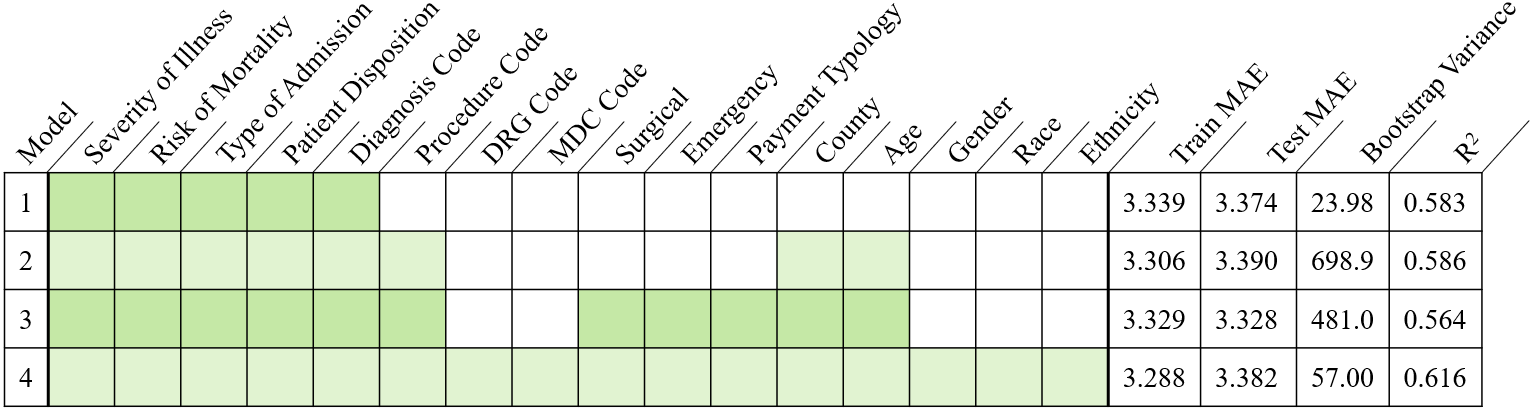


Figure 4: Cross-Validation Statistics

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 R2 is a 3% improvement compared to the others. We will now transition to discuss models in further detail.

# 

# Models

## 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. R2 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 5), so the predictions by our first model are overall more reliable.

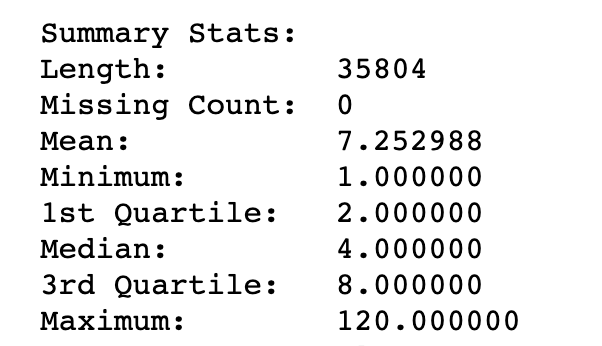
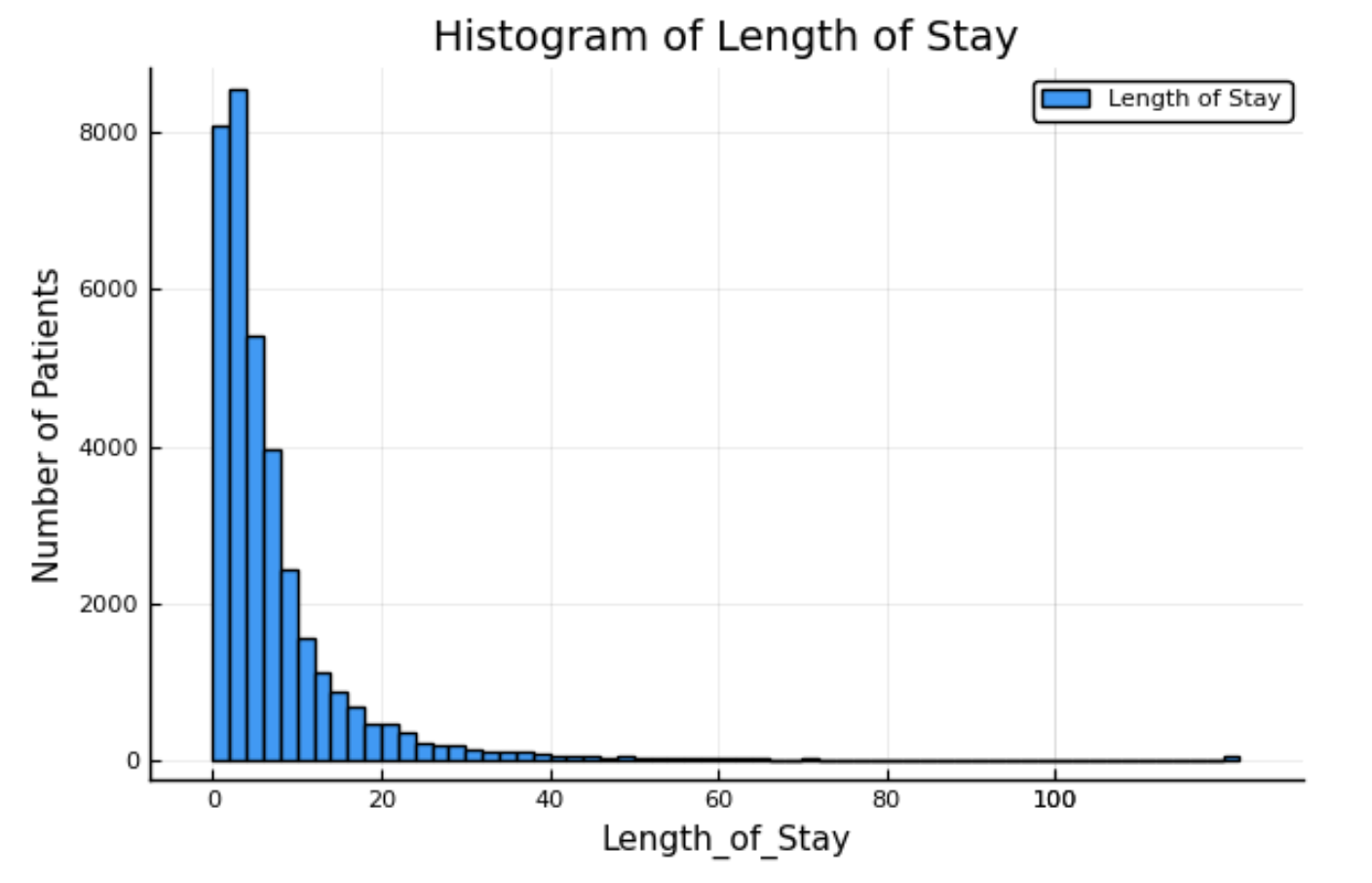


Figure 5: Length of Stay Statistics

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 R2 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.

## K-Means Clustering

The algorithm consists of two alternating steps: selecting ‘k’ 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.

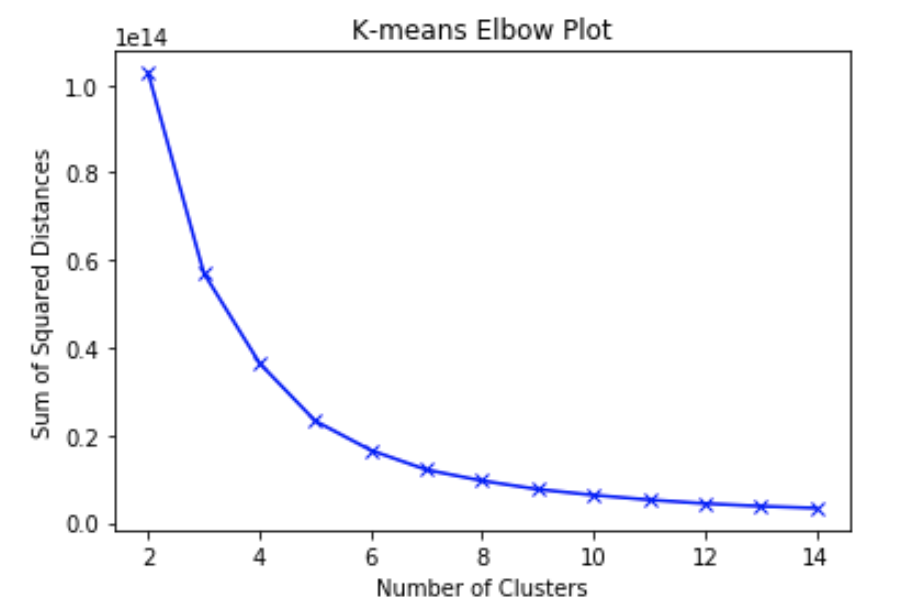


Figure 6: Sum of Squared Distance at each K

Figure 6 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 ≤ k ≤ 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.

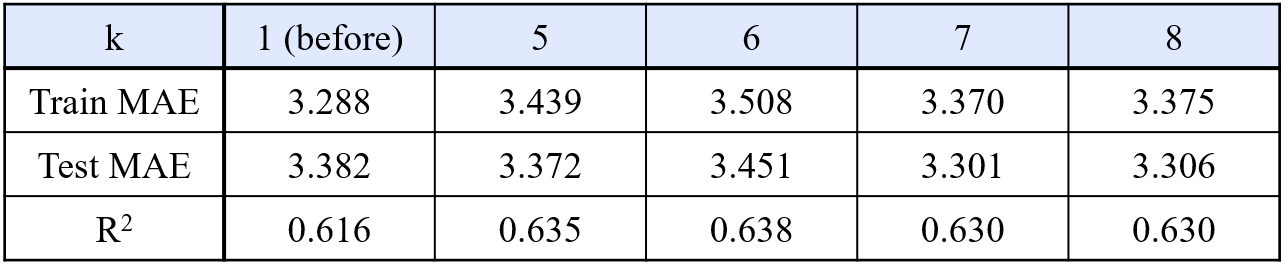


Figure 7: Predicting on K Clusters Summary

For all k > 1, their R2 values are slightly greater than the linear model trained on original data with k = 6 boasting the highest. The 0.02 increase in R2 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.

## Hierarchical Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis that seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: agglomerative and divisive. Agglomerative is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In general, the merges and splits are determined in a greedy manner.

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) that record the admission, diagnosis, expense, length of stay, and personal information of the patients in the dataset.

First, we select the optimal number of clusters used in training. This part is done in R with the *factoextra* and *NbClust* packages. The graphs below represent the evaluation results based on the Elbow method (Figure 8) and the Silhouette method (Figure 9). The Elbow method suggests an x-intercept of 4, meaning 4 is the optimal number of clusters. Silhouette suggests an optimal number of 2. 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 slowly increases after 20 clusters, we trained a 100-cluster bottom-up model. 100 is reasonable considering there are over 30,000 total observations.

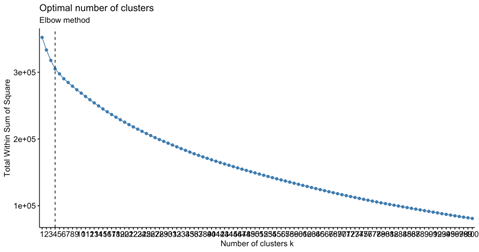


Figure 8: Caption

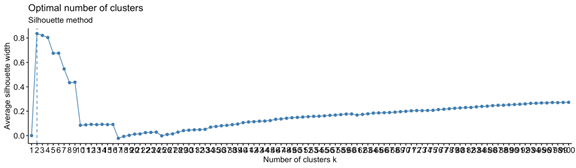


Figure 9: Caption

Due to page limitations, 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 interpretation (like for glass-box models).

The whole-set standard deviation is 11.06 for length of stay and 75,993.88 for total charges. We calculated the standard deviation for each cluster and plotted them against the whole-set value (Figure 10 and Figure 11). There is a significant decrease for most clusters: 69% of clusters have strictly lower standard deviation on length of stay and 76% on total charges. Therefore, the clustering method has done a decent job at assigning similar patients to the same group. While some outliers like cluster #72 do exist, a closer look reveals that they were generated based on other features like diagnosis which would see smaller variation on a different plot (i.e. all of cluster #72 is diagnosed with Mental Diseases and Disorders and 90% with Digestive Malignancy).

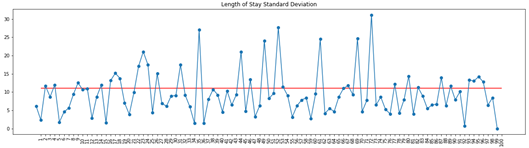


Figure 10: Length of Stay Cluster (blue) and Whole-Set (red) Standard Deviation

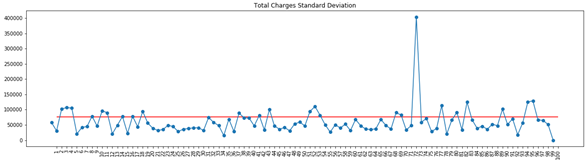


Figure 11: Total Charges Cluster (blue) and Whole-Set (red) Standard Deviation

This leads to a discussion on what information hospitals can learn from our clusters. Revisiting the small cardinality but high resource group, which we will call High Resource Users (HRU), cluster #73 only has 134 patients or 0.37% of the data population but 7 to 8 times longer stay and larger expenses as we see in Figure 12. To better understand these numbers, we find that they have all undergone tracheostomy, which is the insertion of a tube into the windpipe with 96+ hours of mechanical ventilation, or open-heart surgery with the use of ECMO, a blood pump that replaces the function of heart and lungs. This might seem obvious to medical and non-medical people alike as a reason for the long stay and high costs, but it is always good to check our results.

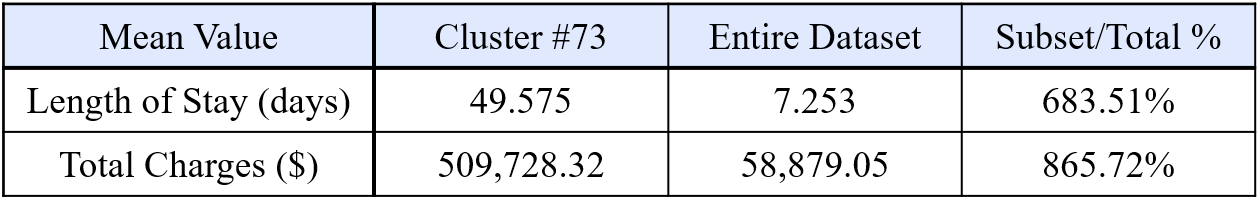


Figure 12: Mean Values of Cluster vs. Entire Dataset

More interesting and less obvious observations are made when we look at the top 10 clusters sorted in descending order (Figure 13). During data preprocessing, gender was one-hot encoded as 1 for female and 0 for male. This means a cluster consists of more males if the value in the gender column is less than 0.5. Out of the HRUs according to length of stay and total charges, more are majority male, and indeed females tend to live longer than males.

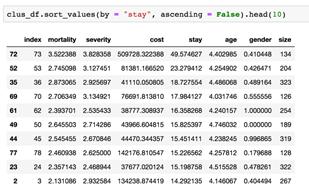
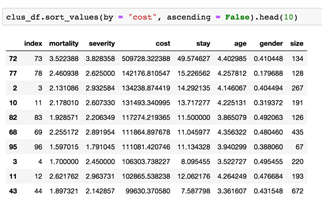
 

Figure 13: Caption

# 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.

# 

# Bibliography

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.

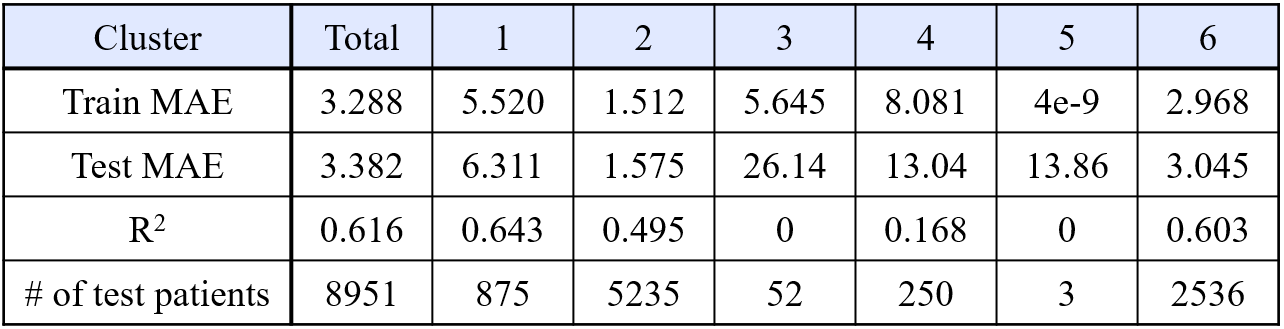


Figure 8: K=6, Prediction on each Individual Cluster

Next, we predicted the length of stay within each cluster for k = 6 (Figure 8). Linear models are overfitting because each cluster only has a small amount of data. In fact, for some clusters, the number of features exceeds the number of patients. We ran k-means clustering for 5 ≤ k ≤ 8. For all k, at least one cluster had a small amount of data. Better solutions would be to set a minimum size or alter the set-up to include a larger dataset like the entire New York State patient dataset.

Cluster 1 had a higher R2 value than the one calculated on the entire test set. This is what we hypothesized would occur. Our k-means clustering results can still be improved upon before being used to make decisions at hospitals.