Term Project

James Liang & David Huang

Cis3920

March 15, 2019

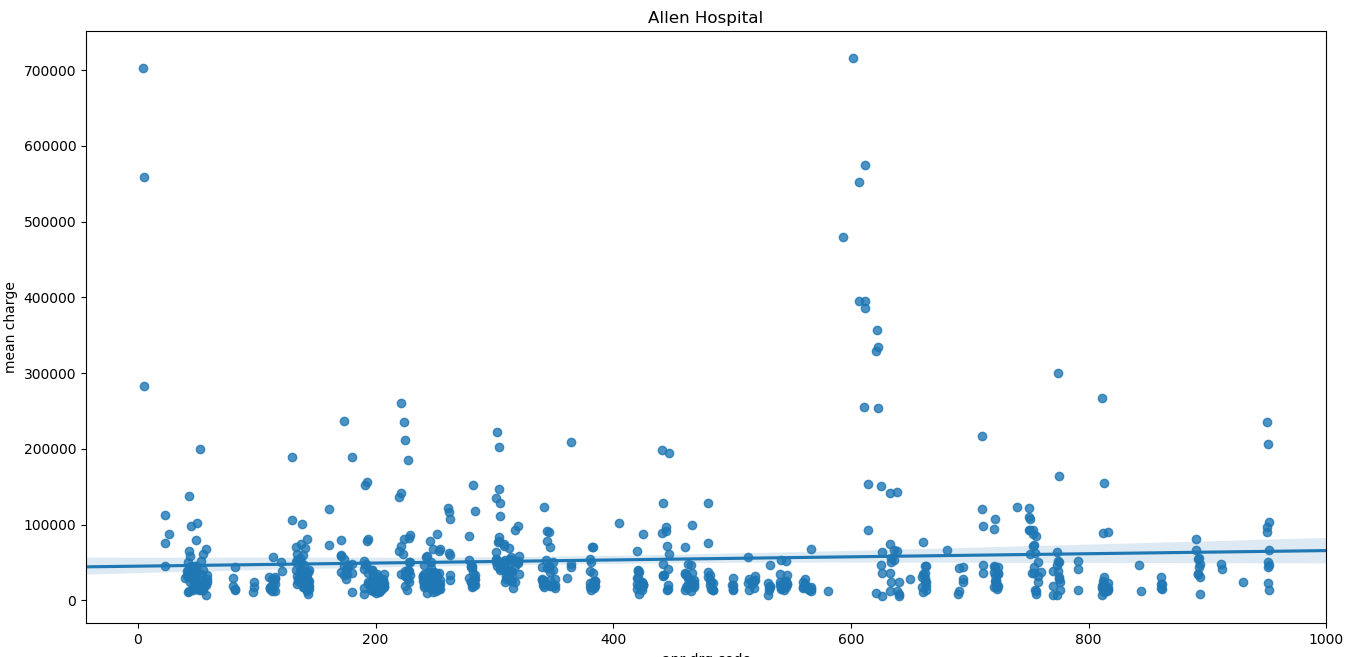
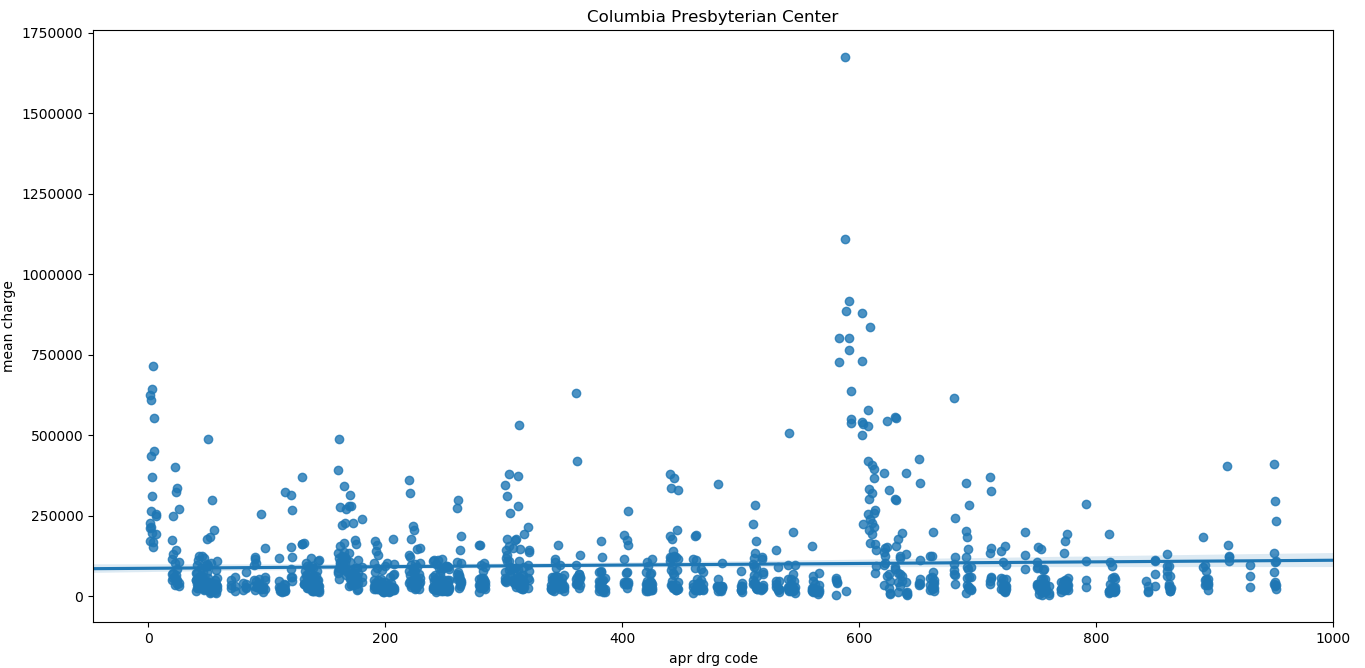
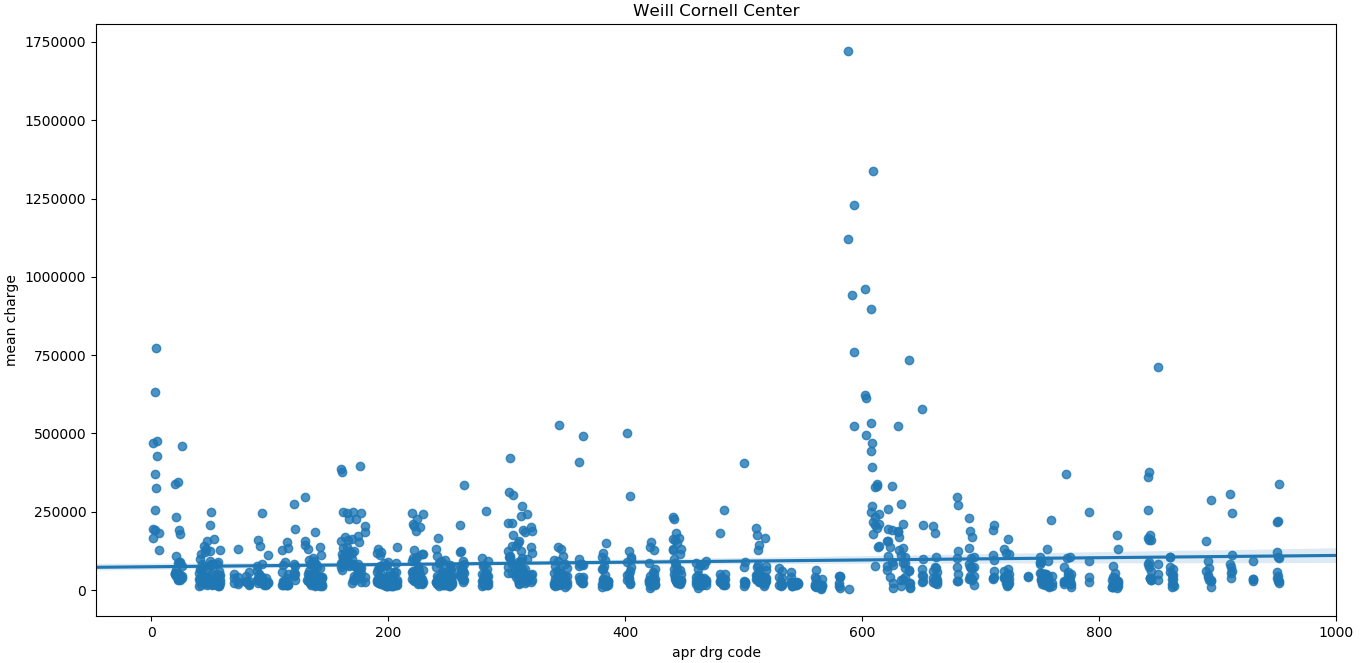
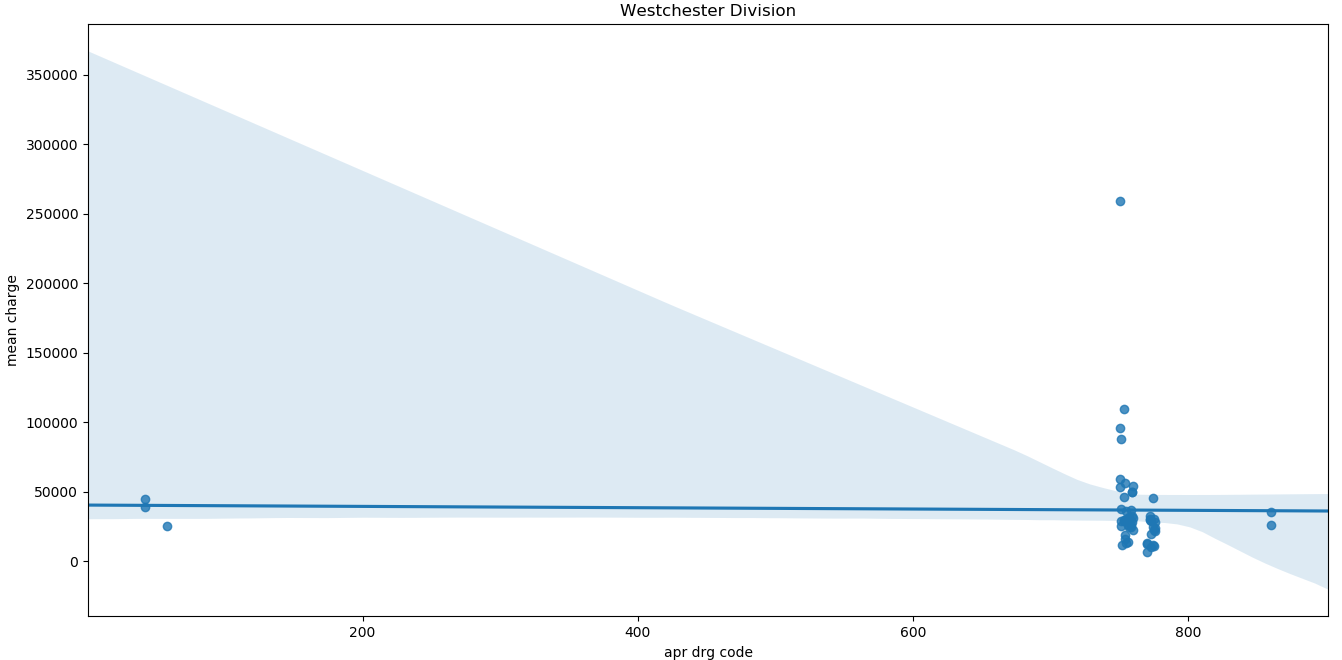
1. What is being trying to accomplish from this project?
   1. To identify which Hospital is the most profitable and which illness generates the most revenues therefore predicting whether to increasing equipment’s or specialist to improve treatment quality and patient satisfaction.
2. Project description/Abstract
   1. Due to the wide varieties of different services provided by these 4 hospital, we want to identify the illness that were the highest and the illness with the largest mode to better improve the efficiency and effectives in targeting those elements. We see that revenue is a great indicator to what illnesses that are the most prevalent and will require higher importance.
3. Introduction
   1. Due to some illnesses having higher capacity than others are creating overcrowded hospital. It has also decrease patient satisfaction with angry patients voicing their complaints against the hospital. We have gathered data from finance in order to complete our study on ways to remedy these issues.
4. Dataset being used and descriptions of the dataset
   1. The project will use data from 4 Hospital from New York Presbyterian Hospital. Using the data from New York Presbyterian Hospital - Westchester Division, New York Presbyterian Hospital - New York Weill Cornell Center, New York Presbyterian Hospital - Columbia Presbyterian Center and, New York Presbyterian Hospital - Allen Hospital. There are 14 attributes and they are listed below. There are 313 different illnesses that were services by these 4 hospital with the number of illness with the count of services during that year and we have the mean cost for each services.

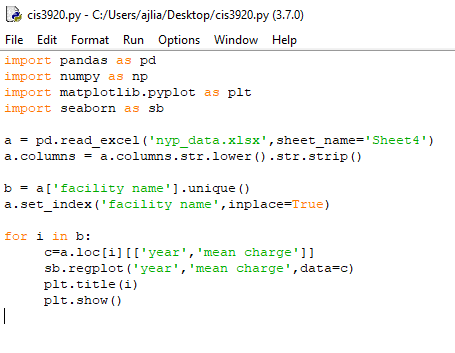
Year  Facility Id  Facility Name  APR DRG Code  APR Severity of Illness Code  APR DRG Description  APR Severity of Illness Description  APR Medical Surgical Code  APR Medical Surgical Description  Discharges  Mean Charge  Median Charge  Mean Cost  Median Cost

1. Literature Survey(existing work on this dataset)

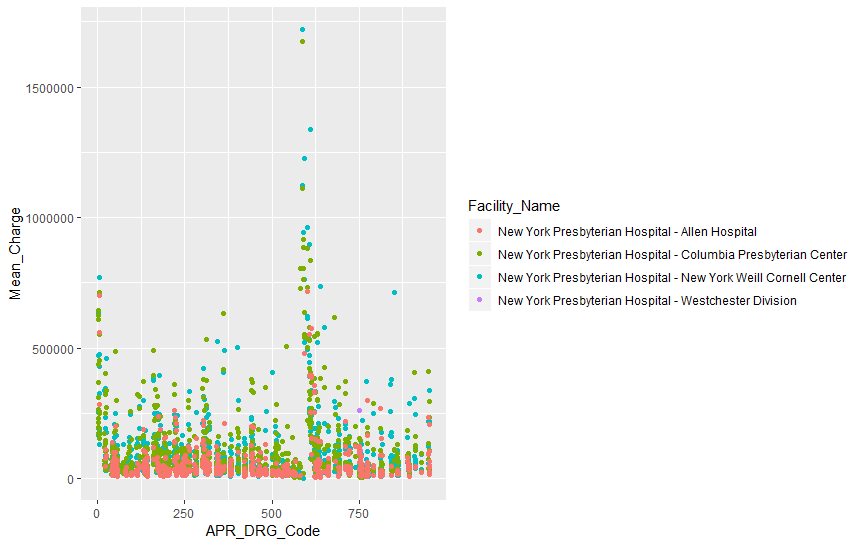


1. Project methodology - Steps being followed - refer to slide 55-56 Lecture#1 - "Intro to Data Mining.pdf"
   1. We will first integrate all four Hospitals data into one dataset. Then we will randomly sample 3000 records in total from all four hospitals, therefore reducing the data to a manageable training set. We will clean the data from any missing data created from the integration. We will test the range of mean cost to get an average cost of that treatment. In addition, we will search for outliers that may drastically change our plot. The data mining task is classification and prediction. We will use the Naïve Bayes, clustering, and decision tree as our data mining techniques.
2. Analysis & Results - Anticipated prediction algorithms being applied on the dataset  
   Linear Regression, Decision Tree, K-Nearest Neighbor
3. Appendix





**Linear Regressions Simple With R**:



A picture containing text

Description automatically generated

**Rcode:**

library(tidyverse)

view(nyp\_data)

nyp\_data

ggplot(data=nyp\_data) +

geom\_point (mapping = aes(x = APR\_DRG\_Code, y= Mean\_Charge, color = Facility\_Name))

ggplot(data=nyp\_data)+

geom\_point(mapping= aes(x=Mean\_Cost, y = Mean\_Charge, alpha = Discharges))

**Linear Regression More in Depth:**

**A close up of a mans face

Description automatically generated**

**A close up of a map

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**Rcode:**

library (ISLR)

library (MASS)

view(nyp\_data)

fix(nyp\_data)

names(nyp\_data)

attach(nyp\_data)

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code,data=nyp\_data)

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code)

lm.fit

summary(lm.fit)

coef(lm.fit)

confint(lm.fit)

plot (APR\_DRG\_Code, Mean\_Cost, col="blue")

abline (lm.fit, lwd=3)

par (mfrow=c (2,2))

plot(lm.fit)

plot(predict(lm.fit), residuals(lm.fit))

plot(predict(lm.fit), rstudent(lm.fit))

plot(hatvalues(lm.fit))

which.max(hatvalues(lm.fit))

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code+Discharges, data = nyp\_data)

lm.fit=lm (Mean\_Cost~., data=nyp\_data)

summary(lm.fit)

**Outputs for Linear Regression Complex:**

Call:

lm(formula = Mean\_Cost ~ APR\_DRG\_Code)

Coefficients:

(Intercept) APR\_DRG\_Code

31380.549 -5.712

Residuals:

Min 1Q Median 3Q Max

-28971 -20841 -14426 2737 412034

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31380.549 1373.508 22.847 <2e-16 \*\*\*

APR\_DRG\_Code -5.712 2.931 -1.949 0.0514 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40830 on 2986 degrees of freedom

Multiple R-squared: 0.00127, Adjusted R-squared: 0.0009359

F-statistic: 3.798 on 1 and 2986 DF, p-value: 0.0514

> coef(lm.fit)

(Intercept) APR\_DRG\_Code

31380.549031 -5.711536

confint(lm.fit)

2.5 % 97.5 %

(Intercept) 28687.4314 3.407367e+04

APR\_DRG\_Code -11.4579 3.482965e-02

Call:

lm(formula = Mean\_Cost ~ ., data = nyp\_data)

Residuals:

Min 1Q Median 3Q Max

-66635 -1239 -14 928 70514

Coefficients: (6 not defined because of singularities)

which.max(hatvalues(lm.fit))

220 #Most influential point. This point is Major Stomach, Esophageal & Duodenal Procedures which seems to make sense since stomach pain sounds like the most common illness. Sometimes minor, sometimes major.

**Decision tree:**

A close up of a device

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Rcode**:

library (ISLR)

library (MASS)

view(nyp\_data)

fix(nyp\_data)

names(nyp\_data)

attach(nyp\_data)

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code, data = nyp\_data)

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code)

lm.fit

summary(lm.fit)

coef(lm.fit)

confint(lm.fit)

plot (APR\_DRG\_Code, Mean\_Cost, col="blue")

abline(lm.fit, lwd = 3)

par(mfrow=c(2,2))

plot(lm.fit)

plot(predict(lm.fit), residuals(lm.fit))

plot(predict(lm.fit), rstudent(lm.fit))

plot(hatvalues(lm.fit))

which.max(hatvalues(lm.fit))

lm.fit=lm (Mean\_Cost~APR\_DRG\_Code+Discharges, data = nyp\_data)

lm.fit=lm (Mean\_Cost~., data=nyp\_data)

summary(lm.fit)

outputs for decision tree:

Classification tree:

tree(formula = High ~ . - Discharges, data = nyp\_data)

Variables actually used in tree construction:

[1] "Median\_Charge" "APR\_DRG\_Code" "Mean\_Charge"

[4] "Median\_Cost" "APR\_Severity\_of\_Illness\_Code" "Facility\_Id"

[7] "Mean\_Cost"

Number of terminal nodes: 18

Residual mean deviance: 0.02145 = 63.71 / 2970

Misclassification error rate: 0.003681 = 11 / 2988

table(tree.pred,High.test)

High.test

tree.pred No Yes

No 1149 5

Yes 2 3

> (1149+3)/1822

[1] 0.6322722

names(cv.nyp\_data)

[1] "size" "dev" "k" "method" #dev is cross-validation error rate

> cv.nyp\_data

$size

[1] 15 5 1

$dev

[1] 17 17 17 🡨 #we pick the smallest one but they’re all the same values of dev so we choose 5 because it cannot be 1 and 15 is too large, will cause more nodes less accuracy. Not sure if the data is supposed to be all the same for dev. This is cause for concern.

$k

[1] -Inf 0.000000 1.000000 1.142857

$method

[1] "misclass"

attr(,"class")

[1] "prune" "tree.sequence"

Pruned:

tree.pred No Yes

No 1152 6

Yes 2 2

> (1152+2)/1822

[1] 0.6333699

0.6333699-0.6322722 = +0.0010977

Increase of 0.0010977 for test observations correctly classified after pruning with a size of 5. More interpretable tree as well as higher classification accuracy albeit small.

**KNN**:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Rcode:**

library (ISLR)

library(class)

View(nyp\_data)

attach(nyp\_data)

summary(nyp\_data)

severity\_of\_illness1 = rep (1, length(`APR\_Severity\_of\_Illness\_Code`))

severity\_of\_illness1[`APR\_Severity\_of\_Illness\_Code`>median(`APR\_Severity\_of\_Illness\_Code`)]=4

boxplot (Mean\_Charge~severity\_of\_illness1, data=nyp\_data, main="Mean Charge vs Severity of Illness")

boxplot (Median\_Cost~severity\_of\_illness1, data=nyp\_data, main = "Median Cost vs Severity of Illness")

boxplot (Mean\_Cost~severity\_of\_illness1, data = nyp\_data, main = "Mean Cost vs Severity of Illness")

boxplot (Median\_Charge~severity\_of\_illness1, data = nyp\_data, main = "Median Charge vs Severity of Illness")

train = (Median\_Cost>Mean\_Cost)

test = !train

nyp\_data.train = nyp\_data[train, ]

nyp\_data.test = nyp\_data[test, ]

severity\_of\_illness1.test = severity\_of\_illness1[test]

lda.fit = lda(severity\_of\_illness1~Mean\_Charge + Median\_Cost + Mean\_Cost + Median\_Charge,data = nyp\_data,subset=train)

lda.pred=predict(lda.fit,nyp\_data.test)

mean(lda.pred$class != severity\_of\_illness1.test)

qda.fit = qda(severity\_of\_illness1~Mean\_Charge + Median\_Cost + Mean\_Cost + Median\_Charge,data = nyp\_data,subset=train)

qda.pred=predict(qda.fit,nyp\_data.test)

mean(qda.pred$class == severity\_of\_illness1.test)

train.B = cbind(Mean\_Charge, Median\_Cost, Mean\_Cost, Median\_Charge)[train,]

test.B = cbind(Mean\_Charge, Median\_Cost, Mean\_Cost, Median\_Charge)[test, ]

train.severity\_of\_illness1 = severity\_of\_illness1[train]

set.seed(1)

pred.knn = knn(train.B, test.B, train.severity\_of\_illness1, k =4)

table(pred.knn,severity\_of\_illness1.test)

mean(pred.knn == severity\_of\_illness1.test)

pred.knn = knn(train.B, test.B, train.severity\_of\_illness1, k =10)

table(pred.knn,severity\_of\_illness1.test)

mean(pred.knn == severity\_of\_illness1.test)

pred.knn = knn(train.B, test.B, train.severity\_of\_illness1, k =100)

table(pred.knn,severity\_of\_illness1.test)

mean(pred.knn == severity\_of\_illness1.test)

**Outputs for KNN:**

> mean(lda.pred$class == severity\_of\_illness1.test)

[1] 0.6537602 ~65% of LDA are correctly accurate.

> mean(qda.pred$class == severity\_of\_illness1.test)

[1] 0.7151899 ~71% of QDA are correctly accurate.

table(pred.knn,severity\_of\_illness1.test)

severity\_of\_illness1.test

pred.knn 1 4

1 1155 546

4 221 764

mean(pred.knn == severity\_of\_illness1.test)

[1] 0.7144453

> table(pred.knn,severity\_of\_illness1.test)

severity\_of\_illness1.test

pred.knn 1 4

1 1159 541

4 217 769

> mean(pred.knn == severity\_of\_illness1.test)

[1] 0.717796

> (1167+783)/2988

[1] 0.6526104 #At k = 4 ~65% correctly predicted for severity of illness. Not great but not bad.

> pred.knn = knn(train.B, test.B, train.severity\_of\_illness1, k =10)

table(pred.knn,severity\_of\_illness1.test)

severity\_of\_illness1.test

pred.knn 1 4

1 1178 503

4 198 807

> mean(pred.knn == severity\_of\_illness1.test)

[1] 0.7390171

> (1178+807)/2988

[1] 0.664324 #At k = 100, better at ~66% correctly predicted for severity of illness.

> pred.knn = knn(train.B, test.B, train.severity\_of\_illness1, k =100)

> table(pred.knn,severity\_of\_illness1.test)

severity\_of\_illness1.test

pred.knn 1 4

1 1253 675

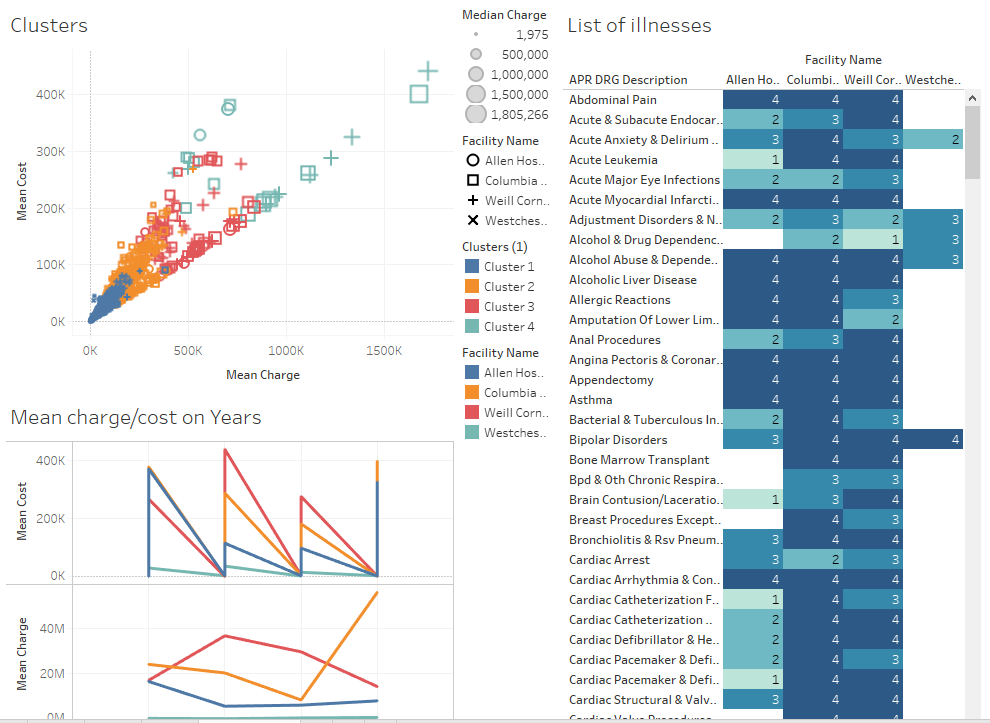
4 123 635

> mean(pred.knn == severity\_of\_illness1.test)

[1] 0.7029039

> (1253+635)/2988

[1] 0.6318608 #At k = 100, worst with ~63% accuracy. Conclusion: We can add more to k for higher percentage but there’s a limit to when it negatively affects accuracy rate.



**Inputs for Clustering**

|  |  |
| --- | --- |
| **Variables:** | Sum of Median Charge |
|  | Sum of Median Cost |
| **Level of Detail:** | Facility Name, Mean Charge, Mean Cost |
| **Scaling:** | Normalized |
|  |  |

**Summary Diagnostics**

|  |  |
| --- | --- |
| **Number of Clusters:** | 4 |
| **Number of Points:** | 2988 |
| **Between-group Sum of Squares:** | 28.523 |
| **Within-group Sum of Squares:** | 4.8336 |
| **Total Sum of Squares:** | 33.356 |
|  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | **Centers** | | | | | | | |  |
| **Clusters** |  | **Number of Items** |  | **Sum of Median Charge** | | | | **Sum of Median Cost** | | | |  |
| **Cluster 1** |  | 2469 |  | 36642.0 | | | | 12500.0 | | | |  |
| **Cluster 2** |  | 408 |  | 1.6314e+05 | | | | 59150.0 | | | |  |
| **Cluster 3** |  | 88 |  | 4.3696e+05 | | | | 1.3838e+05 | | | |  |
| **Cluster 4** |  | 23 |  | 8.6995e+05 | | | | 2.7504e+05 | | | |  |
| **Not Clustered** |  | 0 |  |  |  |  |  |  |  |  |  |  |

**Analysis of Variance:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | **Model** | |  | **Error** | |
| **Variable** |  | **F-statistic** |  | **p-value** |  | **Sum of Squares** | **DF** |  | **Sum of Squares** | **DF** |
| **Sum of Median Cost** |  | 872.2 |  | 0.0 |  | 18.42 | 3 |  | 21.01 | 2984 |
| **Sum of Median Charge** |  | 813.6 |  | 0.0 |  | 10.1 | 3 |  | 12.35 | 2984 |

1. How your contribution different from what is already existing (compare and contrast from Literature survey).  
     
   How our project contributes is that we used different statistical tools such as decision tree, linear regression as well as KNN to determine how to efficiently determine which hospitals use the most and least amount of money for different types of illnesses between severe and minor.
2. Struggles and Roadblocks  
   Was hard to figure out which variables in our data set to use in terms of best results for representation, the decision tree code where we had the same values for $dev which made it hard to determine which number of nodes to use for best accuracy.
3. Conclusion  
   To reduce cost we analyzed mean cost vs severity of illness, this will allow the hospitals in our dataset to realize which illnesses are most severe that require the most expensive cost vs the least severe and least expensive to determine how to properly allocate charges and generate more revenue to continue operating and to most efficiently handle patients in priority as well.
4. References

Cramer, Mary O. “Developing Performance Excellence at Developing Performance Excellence at NewYork NewYork -Presbyterian Hospital Presbyterian Hospital Driving Success and Lessons Learned Driving Success and Lessons Learned.” *cramer\_2aold*, 2006, [www.ehcca.com/presentations/qualitycolloquium4/cramer\_2aold.pdf](http://www.ehcca.com/presentations/qualitycolloquium4/cramer_2aold.pdf).  
<https://health.data.ny.gov/Health/NYP/gbkr-qq5u> #Dataset