Who is likely to pay higher Healthcare Cost for Teenagers in the State of Wisconsin?

Gil King and Kailyn Bates

Eco 520

November 24, 2020

Motivation

The year of 2020 has brought healthcare to the forefront of almost everyone's mind no matter where you are around the world. Healthcare is not equal around the world and Coronavirus has certainly shown where there is room for improvement especially here in the United States of America. The health care system is not a universally accessible system as it is not a right. It is a publicly and privately-funded patchwork of fragmented systems and programs. Here in the States we have a multi-payer system where you can have coverage in many forms. Insured Americans are covered by both public and private health insurance, with a majority of Amercians and their dependents are covered by private insurance plans through their employers. For those who are members of vulnerable population groups there are some government-funded programs, such as Medicaid and Medicare that provide health care coverage. Many children are covered under programs known as (CHIP), Children's Health Insurance Program. It is regulated by the federal government, but administered at the state level. Eligible children come from families at between 200 percent and 300 percent of the federal poverty level ($44,700 to $67,050 for a family of four). Under certain circumstances, pregnant women are also eligible for coverage in some states, and children of low income state (public) employees are covered. Average

hospital charges vary extremely across the country, with outward rhyme or reason with many treatments costing far more in some regions than others.

The goal of eliminating disparities in health care in the United States remains elusive especially on the state level. We have sourced a dataset from The US Agency for Healthcare; it conducted a nationwide survey of hospital costs that consists of hospital records of inpatient samples. Which led us to want to investigate what healthcare was like on a state level. For our project we will look at the state of Wisconsin. The state of Wisconsin ranks 33 out of 50 on hospital costs. Furthermore, health costs also differ greatly among hospitals within ethinic background. Our business question is who is more likely to pay more regarding the healthcare cost for children between the ages of 0-17 to reduce the scope of the project. To answer this question we look specifically in the The Kids' Inpatient Database (KID); a set of pediatric hospital inpatient databases included in the HCUP family and the The HCUP Cost-to-Charge Ratio Files; hospital-level files designed to supplement the data elements in the HCUP inpatient and emergency department databases.

These databases are created by AHRQ through a Federal-State-Industry partnership. We hypothesize that the results will show that the most vulnerable population incur the highest inpatient cost stays. To use a standard definition of vulnerable population we have relied on federal guidelines for statistical reporting and civil rights monitoring set by the Office of Management and Budget (OMB). The department has developed a minimum set of standardized categories for reporting on race and Hispanic ethnicity by federal agencies and recipients of federal funds (which almost all hospitals are).

Data and Empirical Methodology

The KID is a set of longitudinal hospital inpatient databases included in the HCUP family. These databases are created by AHRQ through a Federal-State-Industry partnership. This information is collected every year. The program consists of a series of independent annual surveys gathering health-related data on representative samples of state residents and communities. The information standards in this dataset has changed several times since 2015 and the latest adjustment is for 2018 as 2019 has not been made available. The long form purpose of this project is to make sure that health disparities are identified ahead of time, and that a framework for amassing this data is safe and reproducible. The key sampling periods are from 2018. Each KID sample is drawn from the sampling data frame consisting of discharge information submitted by HCUP Partners—statewide organizations that agree to participate in the KID.The majority of the data from key variables we will be using are integers. Not much factor conversion needed to take place. Age is represented as an integer from 0-17. Female is essentially a gender identifier with presence of women: 1 and non-women: 0. LOS is represented as an integer, Race represented as an integer scaling from 1-6, TOTCHG is an

integer, and APRDRG is also an integer. So the majority of the data we will be using is from 2018.

A Note On Data

AHRQ strongly advises researchers against using the KID to estimate State-specific statistics. Prior to 2012, State was available as a KID data element. However, these KID samples were not designed to display or share a representative sample of hospitals at the State level. AHRQ recommends that researchers employ the SID for State-level estimates.

Data from non-Partner States are missing completely from the sampling frame, and data from Partner States are sometimes incomplete because of different State reporting requirements, different State restrictions, or other data omissions. The KID is designed to represent hospitals and discharges nationally, including those outside the sampling frame.

To accomplish this, within each hospital sampling stratum the KID draws a sample of discharges from the sampling frame required to net a total of 10 percent of normal newborns and 80 percent of other pediatric discharges (younger than 21 years of age) nationally. The sampling strata are defined by census region (4 regions), hospital ownership (3 categories), urban-rural location, teaching status, and bed size (3 categories), with a separate category for children's hospitals. As a result, the proportion of KID discharges in a class that are from a given State is unlikely to equal the State's actual proportion of discharges in that class. Consequently, the sample of KID discharges is unlikely to be representative of discharges in the State, and the KID sample weights will not be appropriate at the State level either.

The level of this "misrepresentation" varies across the States in any given year of the KID, which further confounds State-to-State comparisons based on State-specific estimates from the KID. Moreover, for a given State the level of misrepresentation changes from year to year as States (and hospitals) enter and exit the sampling frame over time. This further confounds State-specific trends based on State-specific estimates from the KID. In summary, KID State-level estimates would be very imprecise at best and biased at worst.

Descriptive & Summary Statistics

```
      AGE              FEMALE           LOS              RACE           TOTCHG          APRDRG
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :  532   Min.   : 21.0
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1216   1st Qu.:640.0
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000   Median : 1536   Median :640.0
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078   Mean   : 2774   Mean   :616.4
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388   Max.   :952.0
```

This is the first summary statistic associated with the dataset. A background, the variables and their meaning used for this analysis include:

LOS	Length of Stay
TOTCHG	Total Hospital Discharge Costs
APRDRG	All patient refined diagnosis related groups
RACE	Ethnic Race of Origin
FEMALE (GENDER)	Gender Identity
AGE	Age

Currently, the average age of the children is 5 and the maximums have been predefined. The max age is 17 and the minimum age is 0, because the majority of children there are infants and below 1 year of age. The accompanying documentation also denotes that there are a large number of infants. The gender distribution is primarily female. The smallest charge according to TOTCHG is $532, and the highest total charged is $48,388. The maximum length of stay is 41 days and the

average amount of time people spend in Wisconsin hospitals is 2.8 or 3 days, coincidentally, the least amount of days someone could stay is 0. These are for the in and our patients, which could include transfers. The diagnosis group analysis (APRDRG) identified 63 groups total of which Group 44 or the Hematological & Immunological Diagnoses, showed the highest number of occurrences in Wisconsin hospitals.

.

The race distribution is categorized as:

1	White
2	Black
3	Hispanic
4	Asian or Pacific Islander
5	Native American
6	Other

The equations that we are using for this particular analysis is the classic linear regression analysis coupled with descriptive statistics of key variables answering certain questions we believe are important. There were about 14 factors across the dataset, but for the purpose of a more pointed analysis we decided to cut the variables down to 6 which we believe can help us determine if there is a relationship. The analysis of variance will help us determine if there is any one variable that matters the most.

```
cat <- aov(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)
cat <- lm(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)
```

The goal is to better understand which variables had a stronger effect on the bottom line. We are seeking to understand who frequents the hospital and who has the maximum expenditure. Is race related to the hospital utilization costs, Can length of stay be predicted from age, gender, and race? In the above code we wanted to see the relationships and aov to see if there was any one factor that stood out among the selected variable choices. The above seeks to find out if the length of stay can be predicted from the age, race or gender as well as a linear model to see if there could be a relationship at all from the total charges column.

```
aov(TOTCHG ~.,data=hosp_cost)
mod <- lm(TOTCHG ~ .,data=hosp_cost)
summary(mod)
```

Taking this approach will help us understand the variables and their behavior first. Understanding the shape of the data and these variables will give us an idea on how to approach the analysis because if we see high distributions in certain variables, we would want to know which ones they are, because they could mean a pattern, and furthermore how much impact exactly would they have.

To conduct our analysis we will be using the Kids' Inpatient Database (KID) and the attachable HCUP Cost-to-Charge Ratio Files to supplement the data elements in the HCUP

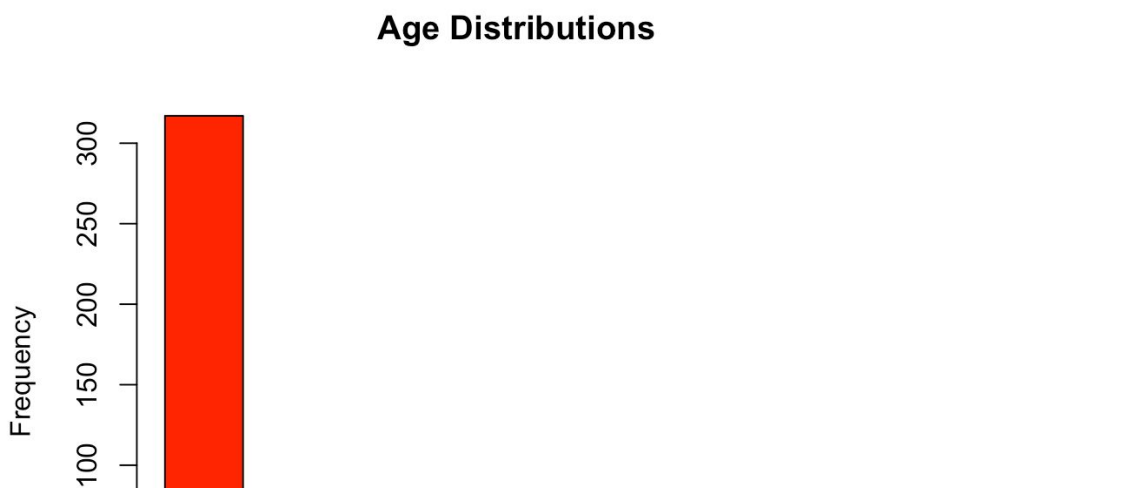
inpatient department database. We are specifically looking at the months of 2019 which are the most available and up to date. To reduce scope we filter it to look at just the State of Wisconsin, however we discovered a flaw in our methodology when trying to create categories. We found that our original plan to divide by ethnic groups is not as simple as it sounds especially in the healthcare system. Each hospital system has its own way of quantifying data and sometimes within systems there are nuances due to the merging of systems.

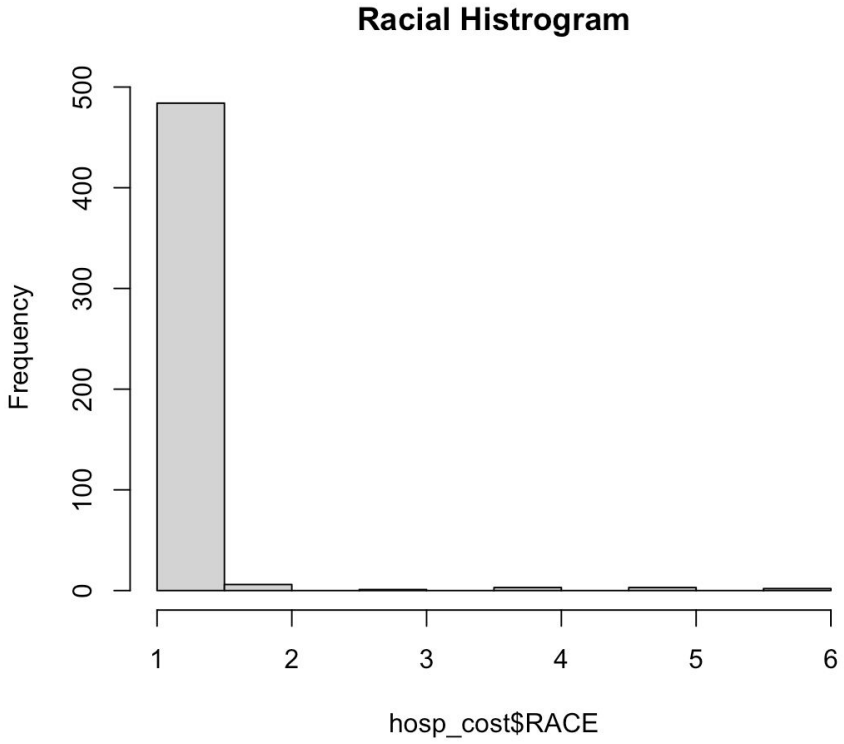
One of the positives is that each of the entities involved in the nation's health care system has some capability for the collection of race, ethnicity, and language data, some are better positioned than others to collect these data through self-report, the generally agreed-upon best way to define a person's racial and ethnic identity. In the future, information infrastructure may enable integrated data exchange so that all entities will not need to collect all data. For now, however, all health and health care entities have roles to play in collecting these data directly from individuals. Hospitals, community health centers, physician practices, health plans, and local, state, and federal agencies can all identify next steps toward improving or implementing direct data collection by under- standing the unique contexts in which they operate. Across all these entities, these data must be collected and stored responsibly.

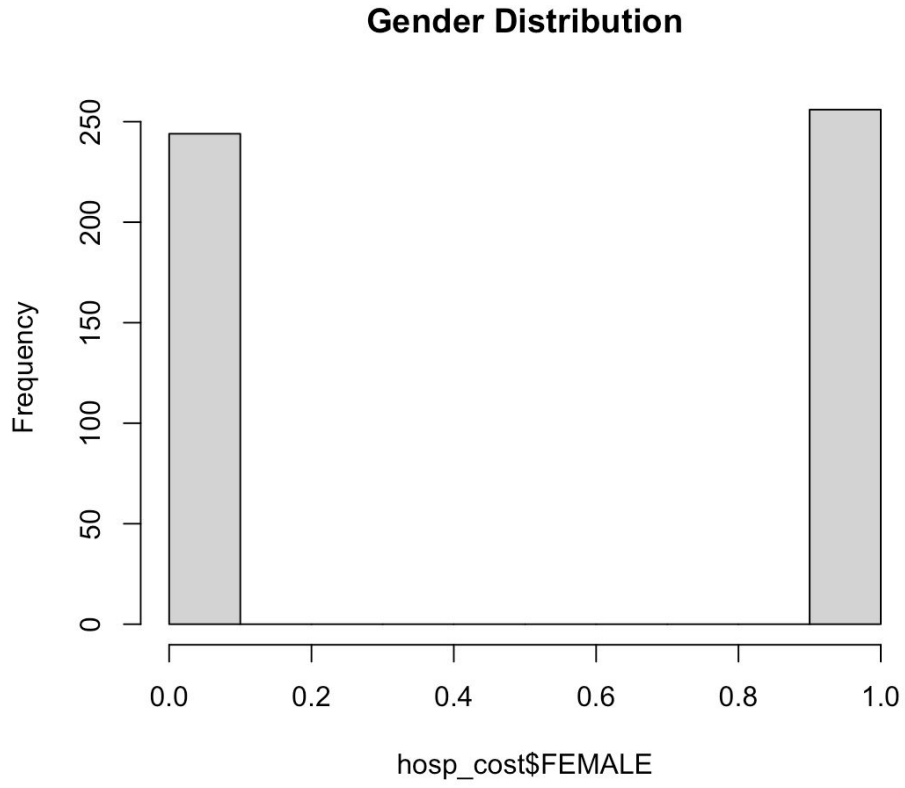
```
t <- table(hosp_cost$APRDRG)
d <- as.data.frame(t)
names(d)[1] = 'Diagnosis Group'
d
which.max(table(hosp_cost$APRDRG))
which.max(t)
which.max(d)
res <- aggregate(TOTCHG ~ APRDRG, data = hosp_cost, sum)
res
which.max(res$TOTCHG)
res[which.max(res$TOTCHG),]
```

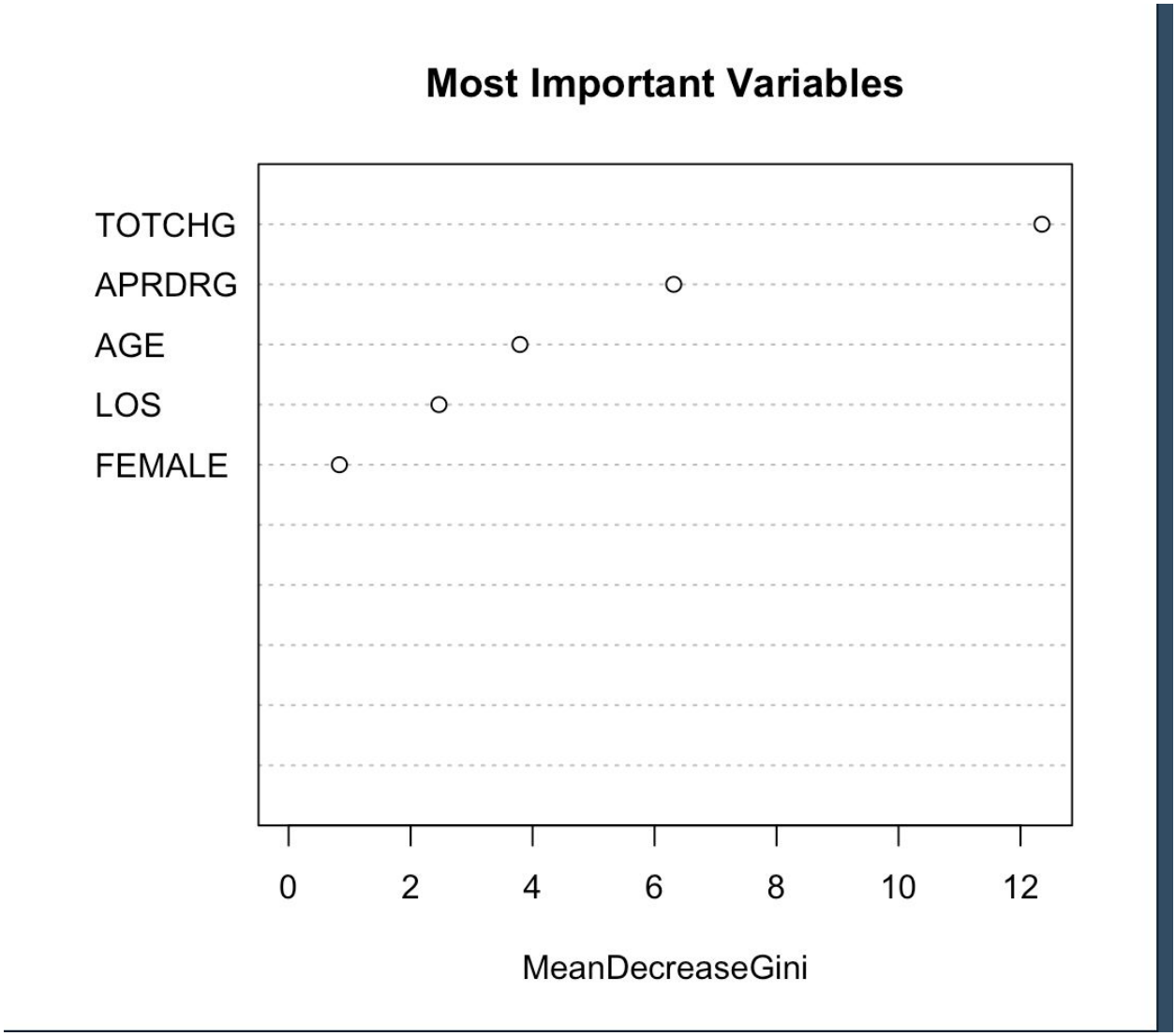
This bit of code tell us exactly which group is most likely to be charged the most and the answer really seemed to point to the obvious reasons hospitals exist, which is to treat people. The diagnosis groups were the best summary statistic we could have used because it told us which group was spending the most money during treatments and it was APRDRG group 44, of which the total charge of this group was $436822. Of all the numbered diagnosis groups, this one was charged the most in their hospital visits. The group is intracranial hemmorage group.

Visualizations




Racial Histrogram
Frequency
0
100
200
300
400
500
1
2
3
4
5
6
hosp_cost$RACE


Gender Distribution
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
hosp_cost$FEMALE


Most Important Variables
TOTCHG
APRDRG
AGE
LOS
FEMALE
0
2
4
6
8
10
12
MeanDecreaseGini

Predictive Results

Linear Regression(s)

lm(formula = TOTCHG ~ RACE, data = hosp_cost)				
Residuals:				
Min	1Q Median	1Q Median		3Q Max
-3049	-1551	-1223	-238	45615
Coefficients:				
Estimate Std. Error t value Pr(>|t|)				
(Intercept) 2772.7 177.6 15.615 <2e-16 ***				
RACE2 1429.5 1604.7 0.891 0.373				
RACE3 268.3 3910.5 0.069 0.945				
RACE4 -428.0 2262.4 -0.189 0.850				
RACE5 -746.0 2262.4 -0.330 0.742				
RACE6 -1423.7 2768.0 -0.514 0.607				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 3906 on 493 degrees of freedom				
(1 observation deleted due to missingness)				
Multiple R-squared: 0.002465,	Adjusted R-squared: -0.007652			
F-statistic: 0.2437 on 5 and 493 DF, p-value: 0.9429				

This formula was to determine if there was any relationship in higher charges based on race, as well as to possibly infer that malpractice might be present. In looking closer at this model, we can recognize that it is unbalanced based on the residuals. This could be attributed to the large amount of race category 1. In looking at the coefficients more closely, Race2 is the highest coefficient. Everything else is negative. The Standard error suggests there is a high variance in the results at 177.6, which isn't the best, but we still cannot reject the null hypothesis that race has a relationship, because the t value is so close to zero. The F-statistic is very close to 1, at .24, we can't say that there is a clear relationship as the data is imbalanced and the Fstat is not higher than 1.

Factors	Df	SumSq	MeanSq	F-value	Pr(>F)
Age	1	27	26.907	2.361	0.125
Female	1	17	16.51	1.449	0.229
Race	1	6	1.138	0.100	0.992
Residuals	491	5595	11.396		

Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race. The equation for this was

```
cat <- aov(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)
cat <- lm(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)
```

```
lm(formula = LOS ~ AGE + FEMALE + RACE, data = hosp_cost)

Residuals:
    Min      1Q  Median      3Q     Max
-3.211  -1.211  -0.857   0.143  37.789

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.85687    0.23160  12.335   <2e-16 ***
AGE         -0.03938    0.02258  -1.744   0.0818 .
FEMALE       0.35391    0.31292   1.131   0.2586
RACE2       -0.37501    1.39568  -0.269   0.7883
RACE3        0.78922    3.38581   0.233   0.8158
RACE4        0.59493    1.95716   0.304   0.7613
RACE5       -0.85687    1.96273  -0.437   0.6626
RACE6       -0.71879    2.39295  -0.300   0.7640
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.376 on 491 degrees of freedom
Multiple R-squar Adjusted R-squared:  -0.005433
F-statistic: 0.6156 on 7 and 491 DF,  p-value: 0.7432
```

Basically from the output we can see a higher F-statistic, which could show this model has a better fit to this data set for its size and shape. The gender information according to the coefficients don’t really show a relationship but a point of interest is age for length of stay. Age being a predictor for length of stay based on any number of the APRDRG categories. Age also generally makes sense with recovery times for certain illnesses. Race factor did not show that there was a relationship with length of stay either, which was surprising - but with an unbalanced dataset, we don’t want to speculate too much.

Random Forest Classification

	Confusion Matrix and Statistics						
	Reference						
	Prediction	1	2	3	4	5	6
	1	120	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	0	0	0	0	0
	4	0	0	0	1	0	0
	5	0	0	0	0	2	0
	6	1	0	0	0	0	1
	Overall Statistics						
		Accuracy :	0.992				
		95% CI :	(0.9562, 0.9998)				
		No Information Rate :	0.968				
		P-Value [Acc > NIR] :	0.08805				

		Kappa : 0.8862					
		Mcnemar's Test P-Value : NA					
		Statistics by Class:					
	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	
Sensitivity	0.9917	NA	NA	1	1	1	
Specificity	1	1	1	1	1	0.9919	
Pos Pred Value	0.968	NA	NA	1	1	0.5	
Neg Pred Value	0.968	NA	NA	1	1	1	
Prevalence	0.968	0	0	0.008	0.016	0.008	
Detection Rate	0.96	0	0	0.00.8	0.016	0.008	
Detection Prevalence	0.96	0	0	0.008	0.016	0.016	
Balanced Accuracy	0.9959	NA	NA	1	1	0.996	

Upon splitting and training the dataset we chose Random forest to see another model. The random forest model was used to test and see if the predicted value met the actual model value of the dataset. Reading the output it says that the model was .99 accurate but I'm unsure of its accuracy because of the smaller sample being used . For a confidence interval of 95%, the models of race against the predicted value turned out to be pretty high. Essentially, we were getting 99% on the test of all the variables in random forest, which turned out to actually be better than the confusion matrix method.

Performance

The best model in our opinion was the Random Forest classifier just because I saw more notes of accuracy as well as a dual model comparison of which this one outperformed the classical linear regression models. We really do think the imbalance of the dataset had some to do with the changing of the classifications and we were glad we were able to experiment with so many different methods. The linear model didn't do as well as our hunch, that race mattered, but class certainly would have been a predictor, again- a hunch. The FStats of the linear model weren't strong enough to say there was a real relationship, so even the comparative linear model did not yield results like we had thought. Although in it, there was a greater indicator of race being a factor.

The Random forest methods applied showed better accuracy and the confusion matrix kind of helped us see that the factors we were testing had more significance. A true limitation is a deeper understanding of how the shape of the data may mean something. Maybe the 6 variables were too low from 14. We wish we could learn how to iterate through the Aggregated diagnosis groups better to uncover more relationships, and in retrospect, with more skill we hope to be able to do this soon, so greater actionable items can be linked to more and more needs.

Summary of Project

Working through our business question, Who is more likely to pay more regarding the healthcare cost for children between the ages of 0-17 in the State of Wisconsin creates a window to look through the inner workings of a machine in constant use. Some pieces are overworked, some pieces are connected temporarily until new processes are made but most of all the system is doing its main goal; serving the customer. We have found that there were about 14 factors across the dataset that cut the variables down to 6 which we believe can help us determine if there is a relationship. These lead to our own conclusion of identifying which groups produced the highest spend amount. This did help in advising us that there wasn't a strong relationship between race, but more so regarding what the patient was being treated for. The highest group was autoimmune diseases an .

Like many projects our work is just a small snippet of an industry. In such a large industry, looking at small sections of data creates shortcomings. We must remember the long form purpose of this project is to make sure that health disparities are identified ahead of time and that a framework for amassing this data is safe and reproducible for the future. This was an interesting project and we look forward to growing our understanding of classical regression, linear probability modeling, and classification.

Bibliography

"Accompanying Ratio Files ." *Cost-to-Charge Ratio Files*, 12 Nov. 2020, www.hcup-us.ahrq.gov/db/ccr/costtocharge.jsp.

IOM (Institute of Medicine). 2009. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* Washington, DC: The National Academies Press.

Gay, James, et al. "Standards Documentation ." APR-DRGsV20, July 2003.

Utilization Project, Healthcare Cost. "Kids' Inpatient Database." *KID Database Documentation*, 2020, www.hcup-us.ahrq.gov/db/nation/kid/kiddbdocumentation.jsp.

Appendix: SAS or R Command and Data Files

```
library(ggplot2)
library(dplyr)
library(readr)
library(tidyr)
library(randomForest)



#READ IN DATA
hosp_cost <- read.csv("HospitalCosts.csv")

#DESCRIPTIVE STATS
str(hosp_cost)
head(hosp_cost)
tail(hosp_cost)
summary(hosp_cost)
table(hosp_cost)


head(hosp_cost$AGE)
summary(hosp_cost$AGE)
hist(hosp_cost$AGE, main="Age Distribution")
summary(as.factor(hosp_cost$AGE))
max(table(hosp_cost$AGE))
max(summary(as.factor(hosp_cost$AGE)))
which.max(table(hosp_cost$AGE))
age <- aggregate(TOTCHG ~ AGE, data = hosp_cost, sum)
max(age)

#whats the race distribution?
str(hosp_cost)
head(hosp_cost$RACE)
tail(hosp_cost$RACE)
summary(hosp_cost$RACE)
plot(hosp_cost$RACE, main="Race Distribution")
hist(hosp_cost$RACE, main="Racial Histrogram")
summary(as.factor(hosp_cost$RACE))
max(summary(as.factor(hosp_cost$RACE)))

#whats the gender distirbution?
```

```
str(hosp_cost$FEMALE)
head(hosp_cost$FEMALE)
summary(hosp_cost$FEMALE)
tail(hosp_cost$FEMALE)
hist(hosp_cost$FEMALE, main="Gender Distribution")


#Highest spending group
t <- table(hosp_cost$APRDRG)
d <- as.data.frame(t)
names(d)[1] = 'Diagnosis Group'
d
which.max(table(hosp_cost$APRDRG))
which.max(t)
which.max(d)
res <- aggregate(TOTCHG ~ APRDRG, data = hosp_cost, sum)
res
which.max(res$TOTCHG)
res[which.max(res$TOTCHG),]

#Race
table(hosp_cost$RACE)
hosp_cost$RACE <- as.factor(hosp_cost$RACE)
fit <- lm(TOTCHG ~ RACE,data=hosp_cost)
fit
summary(fit)
fit1 <- aov(TOTCHG ~ RACE,data=hosp_cost)
summary(fit1)
hosp_cost <- na.omit(hosp_cost)

#Age and Gender Analysis
table(hosp_cost$FEMALE)
a <- aov(TOTCHG ~ AGE+FEMALE,data=hosp_cost)
summary(a)
b <- lm(TOTCHG ~ AGE+FEMALE,data=hosp_cost)
summary(b)

#Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay
can be predicted from age, gender, and race.
table(hosp_cost$LOS)
cat <- aov(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)
cat <- lm(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
summary(cat)


#What's the variable that matters the most?
aov(TOTCHG ~.,data=hosp_cost)
mod <- lm(TOTCHG ~ .,data=hosp_cost)
```

```
summary(mod)



#ALTERNATIVE
str(hosp_cost)
summary(hosp_cost)
hist(hosp_cost$AGE, col="red", main="Age Distributions")
hist(hosp_cost$FEMALE, col="blue", main="Gender Distribution")
plot(hosp_cost$RACE, hosp_cost$TOTCHG, main="Cost Distribution by Race",
    xlab="Race", ylab="Highest Cost", pch=18)




#RANDOM FOREST

set.seed(1900)
train_ind <- sample(nrow(hosp_cost),round(0.75*nrow(hosp_cost)))
train     <- hosp_cost[train_ind,]
test      <- hosp_cost[-train_ind,]

str(hosp_cost)
rfModel <- randomForest(RACE ~ . , data = hosp_cost)
test$predicted <- predict(rfModel, test)
library(caret)
confusionMatrix(test$RACE, test$predicted)


library(MLmetrics)
F1_all <- F1_Score(test$RACE, test$predicted)
F1_all

options(repr.plot.width=5, repr.plot.height=4)
varImpPlot(rfModel,
          sort=T,
          n.var=10,
          main="Most Important Variables")
```