Week12

Ashley Hirt, Shreya Dubey, Rakib Kamal

11/18/2019

## Library Packages

library(data.table)  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ---------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::between() masks data.table::between()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::first() masks data.table::first()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::last() masks data.table::last()  
## x purrr::transpose() masks data.table::transpose()

library(viridis)

## Loading required package: viridisLite

theme\_set(theme\_minimal())  
  
# colors used for plotting  
mygreen <- "#22908C"  
mypurple <- "#450D54"  
  
mycolors <- c("Number of Admissions" = mygreen,   
 "Population" = mypurple)

## Read in Data

First, we read in our data and the columns. We will also get rid of Discharge.Year because all of the years were 2016.

ashley <- "G:/My Drive/DATA/Final Project/hospital.csv"  
rakib <- "C:\\Users\\Rakib Kamal\\Box Sync\\Fall 2019\\DS4001\\Projects\\Final\\hospital.csv"  
  
hosp <- fread(ashley,   
 check.names=T,  
 na.strings = c("", "NA")  
)  
  
hosp <- as.tibble(hosp)

## Warning: `as.tibble()` is deprecated, use `as\_tibble()` (but mind the new semantics).  
## This warning is displayed once per session.

# save original dataset before renaming  
hosp\_og <- hosp  
  
# rename columns  
colnames(hosp) <- c("Area", "County", "Certificate", "Facility.ID", "Facility.Name",  
 "Age", "Zip", "Gender", "Race", "Ethnicity", "Duration", "Type",  
 "Discharge.Status", "Discharge.Year", "CCS.Code", "CCS.Diagnosis",  
 "CCS.Procedure.Code", "CCS.Procedure", "DRG.Code", "DRG",  
 "MDC.Code", "MDC", "Severity.Code", "Severity", "Mortality.Risk",  
 "Surgical", "Payment1", "Payment2", "Payment3", "Birthweight",  
 "Abortion", "ED", "Charges", "Costs")  
  
unique(hosp$Discharge.Year)

## [1] 2016

# get rid of discharge.year   
hosp <- hosp[,-which(colnames(hosp)=="Discharge.Year")]

## Introduction

## Data Summary

## Examine Existing NAs

Then, we examine existing NAs. We see that, of our 2,343,429 rows, there are 1,725,198 incomplete cases. Please note that in any cases where we think there could be systematic missingness in categorical variables, we are keeping the variables as “unknown” or “not applicable” rather than recoding them as NAs. This is because randomForest removes any rows with NAs or forces you to impute them; instead, we’d like to keep them as a separate grouping. For continuous variables, we are recoding them as NAs to avoid introducing strings into a numeric variable.

sum(complete.cases(hosp) == F) # 1725198 incomplete cases

## [1] 1725198

# to look at the incomplete rows (assign data frame "incomplete"):  
incomplete <- hosp[!complete.cases(hosp),]  
  
# tally up the number of NAs in each column  
nas <- c()  
invisible(  
sapply(1:ncol(incomplete), function(x) {  
 nas[x] <<- sum(is.na(incomplete[,x]))  
})  
)  
names(nas) <- colnames(incomplete)  
nas # shows the number of NAs in each column

## Area County Certificate   
## 5325 5325 5325   
## Facility.ID Facility.Name Age   
## 5325 0 0   
## Zip Gender Race   
## 36215 0 0   
## Ethnicity Duration Type   
## 0 0 0   
## Discharge.Status CCS.Code CCS.Diagnosis   
## 0 0 0   
## CCS.Procedure.Code CCS.Procedure DRG.Code   
## 0 0 0   
## DRG MDC.Code MDC   
## 0 0 0   
## Severity.Code Severity Mortality.Risk   
## 0 67 67   
## Surgical Payment1 Payment2   
## 0 0 834420   
## Payment3 Birthweight Abortion   
## 1691076 0 0   
## ED Charges Costs   
## 0 0 0

Area, County, Certificate, and Facility.ID have 5325 instances of NAs. They all co-occur (an NA in one of these variables corresponds to an NA in the others as well). We also see that the Facility.Name variable is always “Redacted for Confidentiality” for these instances and Severity.Code is 0. Presumably, this is the reason that all the location information is NAs for these records. We will replace the corresponding Severity Codes with NAs and later create an indicator variable for these rows.

# look at facility name for these redacted rows  
unique(incomplete[is.na(incomplete$Area),]$Facility.Name)

## [1] "Redacted for Confidentiality"

# 0s don't make sense for Severity.Code and correspond to the redacted rows, replace with NAs  
hosp$Severity.Code <- ifelse(hosp$Severity.Code==0, NA, hosp$Severity.Code)

Zipcode has 36215 NAs. The data dictionary states that the patient zip code will be blank (NA) for the following confidentiality reasons: zip code population size less than 20,000 people; abortion records; less than 10 records with that zip code.

Severity AND Mortality.Risk each have 67 NAs, which also co-occur. Most of these records have no diganosis under the DRG column and they have a DRG.Code greater than 952. This doesn’t make sense, since the maximum DRG.Code used in New York State is 952. Thus, we assume these records had missing or invalid DRG Codes. Since Severity and mortality.risk are derived from the DRG coding system as well, it makes sense that they would also be NAs. Also, (inexplicably), for all of these records, the CCS.Diagnosis corresponds to some type of birth that was not an abortion.

# examine incomplete severity rows  
severity.incomplete <- incomplete[is.na(incomplete$Severity),]  
unique(severity.incomplete$DRG.Code)

## [1] 956 955

unique(severity.incomplete$DRG)

## [1] "Ungroupable"   
## [2] "Principal diagnosis invalid as discharge diagnosis"

# replace with NAs  
hosp$DRG.Code <- ifelse(hosp$DRG.Code > 952, NA, hosp$DRG.Code)

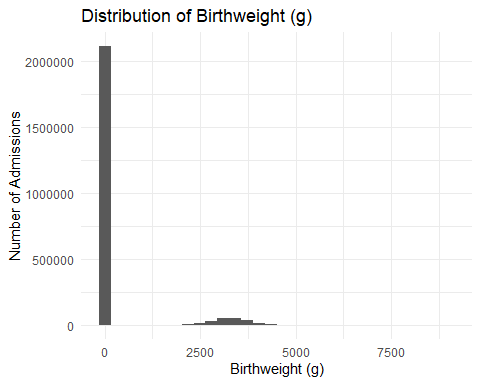
Payment2 has 834420 NAs. Payment3 has 1691076 NAs. These correspond to people that only had one or two payment types, respectively, so that makes sense and requires no further investigation.

## Change Invalid Values to NAs

To identify invalid values in each of our variables, we look at their distributions (either a histogram for quantitative variables or bar plots for categorical variables).

# examine birthweight plot for strange values  
ggplot(hosp, aes(x=Birthweight)) +  
 geom\_histogram() +   
 ggtitle("Distribution of Birthweight (g)") +  
 ylab("Number of Admissions") +  
 xlab("Birthweight (g)")

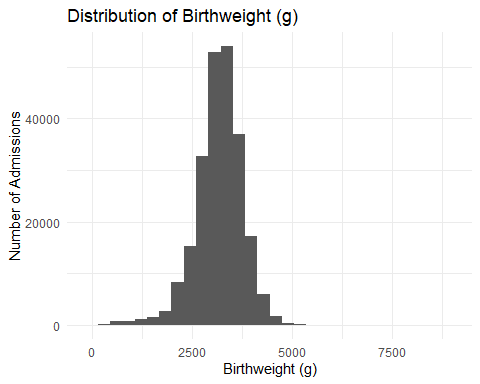
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 We see that there are a large number of 0s for birthweight, which indicates that the record does not correspond to a pregnancy/birth. We replace all 0s with “Not a birth”, reexamine the birthweight plot, and see that it appears normal (as expected)

hosp$Birthweight <- ifelse(hosp$Birthweight==0, NA, hosp$Birthweight)  
  
# reexamine birthweight plot  
ggplot(hosp, aes(x=Birthweight)) +  
 geom\_histogram() +   
 ggtitle("Distribution of Birthweight (g)") +  
 xlab("Birthweight (g)") +  
 ylab("Number of Admissions")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2110676 rows containing non-finite values (stat\_bin).

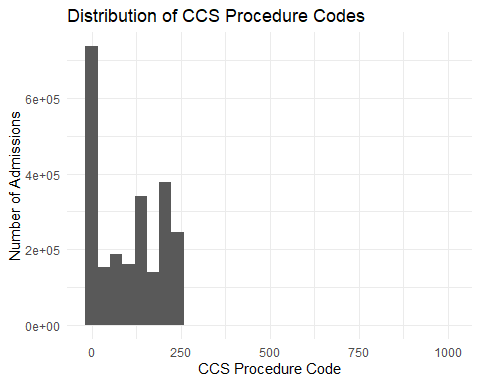


# none of the birthweights appear abnormal (they do follow a normal distribution)

Next, we look at the distribution of CCS Procedure Code. The highest CCS Procedure Code used is 231, but the highest code in our dataset is 999, which corresponds to “Ungroupable” procedures. We will make those procedures NAs. Also, a procedure code of 0 corresponds to “No procedure”, which we will later make an indicator variable for.

# examine distribution of CCS procedure code  
ggplot(hosp, aes(x=CCS.Procedure.Code)) +   
 geom\_histogram() +  
 ggtitle("Distribution of CCS Procedure Codes") +   
 xlab("CCS Procedure Code") +  
 ylab("Number of Admissions")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



unique(hosp[hosp$CCS.Procedure.Code>231, "CCS.Procedure.Code"])

## # A tibble: 1 x 1  
## CCS.Procedure.Code  
## <int>  
## 1 999

# the highest CCS Procedure code in our dataset is 999. We will make those NAs.   
hosp$CCS.Procedure.Code <- ifelse(hosp$CCS.Procedure.Code==999, NA, hosp$CCS.Procedure.Code)  
# show procedure code = 0  
hosp[hosp$CCS.Procedure.Code==0, c("CCS.Procedure.Code", "CCS.Procedure")][1,]

## # A tibble: 1 x 2  
## CCS.Procedure.Code CCS.Procedure  
## <int> <chr>   
## 1 0 NO PROC

We also see that there are 14 rows with an MDC code of 0, corresponding to the category “Pre-MDC or Ungroupable.” We replace these values with NAs.

nrow(hosp[hosp$MDC.Code==0,]) # 14 rows have MDC.code=0

## [1] 14

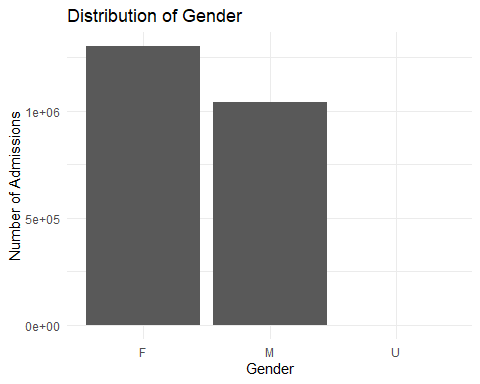
# replace with NAs  
hosp$MDC.Code <- ifelse(hosp$MDC.Code==0, NA, hosp$MDC.Code)

Also, we change the one invalid numeric value in Duration to a numeric value (120+ to 120).

hosp$Duration <- ifelse(hosp$Duration == "120 +", 120, hosp$Duration)  
hosp$Duration <- as.numeric(hosp$Duration)

Repeat for gender, and replace all 57 “U” genders with “Unknown”.

# examine gender plot for unusual values  
ggplot(hosp, aes(x=Gender)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Gender") +  
 xlab("Gender") +  
 ylab("Number of Admissions")



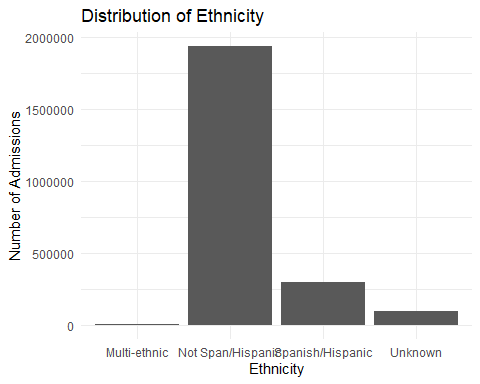
# replace all "U" genders with "Unknown"  
sum(hosp$Gender == "U") # only 57 records with unknown gender.

## [1] 57

hosp$Gender <- ifelse(hosp$Gender=="U", "Unknown", hosp$Gender)

Repeat for ethnicity, and keep all 97,000+ “unknown” ethnicities.

# examine ethnicity plot for unusual values  
ggplot(hosp, aes(x=Ethnicity)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Ethnicity") +  
 xlab("Ethnicity") +  
 ylab("Number of Admissions")

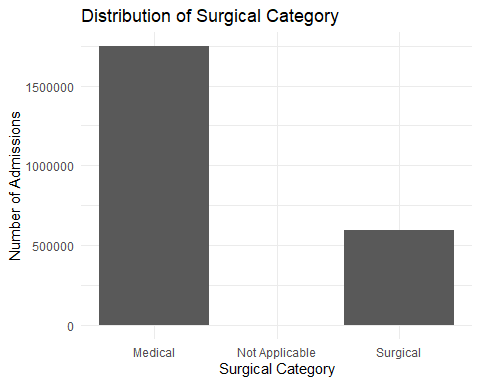


# look at number of unknown ethnicities  
sum(hosp$Ethnicity == "Unknown") # 97,180 records with Unknown ethnicities

## [1] 97180

Repeat for surgical, and replace all 67 “Not Applicable” entries with “Unknown”

# examine categories  
ggplot(hosp, aes(x=Surgical)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Surgical Category") +  
 xlab("Surgical Category") +  
 ylab("Number of Admissions")



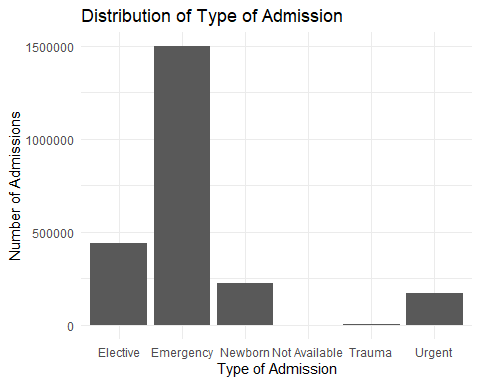
# replace surgical "Not Applicable" entries with "Unknown"  
sum(hosp$Surgical == "Not Applicable") # 67 Surgery "Not Applicable"s

## [1] 67

hosp$Surgical <- ifelse(hosp$Surgical=="Not Applicable", "Unknown", hosp$Surgical)

Repeat for Type of Admission, and replace any Unavailable values with “Unknown”

ggplot(hosp, aes(x=Type)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Type of Admission") +  
 xlab("Type of Admission") +  
 ylab("Number of Admissions")



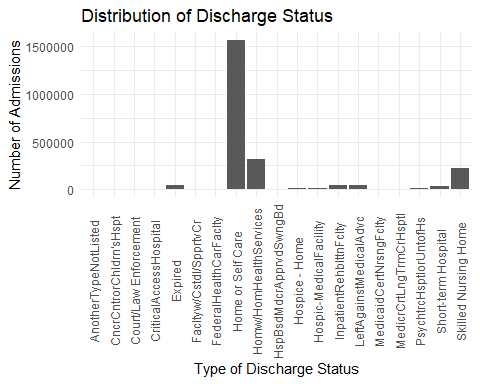
sum(hosp$Type=="Not Available") # 435 of these rows

## [1] 435

hosp$Type <- ifelse(hosp$Type == "Not Available", "Unknown", hosp$Type)

Repeat for discharge status, and replace unavailable values with “Unknown.”

ggplot(hosp, aes(x=Discharge.Status)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Discharge Status") +  
 xlab("Type of Discharge Status") +  
 ylab("Number of Admissions") +  
 theme(axis.text.x = element\_text(angle=90, vjust=.5)) +  
 scale\_x\_discrete(label=function(x) abbreviate(x, minlength=22))



sum(hosp$Discharge.Status=="Another Type Not Listed") #6976 "Another Type"s

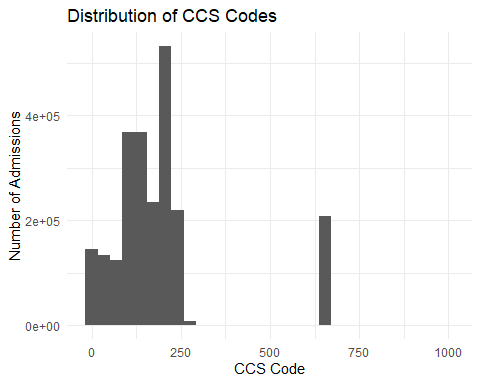
## [1] 6976

hosp$Discharge.Status <- ifelse(hosp$Discharge.Status=="Another Type Not Listed",   
 "Another Type", hosp$Discharge.Status)

Repeat for CCS code. From the distribution, we see that categories above 650-663 and 670 represent MHSA coding (mental health codes). 917 stands for unclassified E code and 999 is ungroupable, which we will recode as NAs

ggplot(hosp, aes(x=CCS.Code)) +   
 geom\_histogram() +  
 ggtitle("Distribution of CCS Codes") +  
 xlab("CCS Code") +  
 ylab("Number of Admissions")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#recode as NAs   
hosp$CCS.Code <- ifelse(hosp$CCS.Code==999, NA, hosp$CCS.Code)

Repeat for payment1. We will keep the 6,896 “Unknowns.” Note that we are not replacing payment2 or payment3 “Unknown” with NA because currently, NA stands for “there was not a second/third payment type” and we don’t want to combine these records. Perhaps the unknown category has predictive value and should be kept separate from the other NAs.

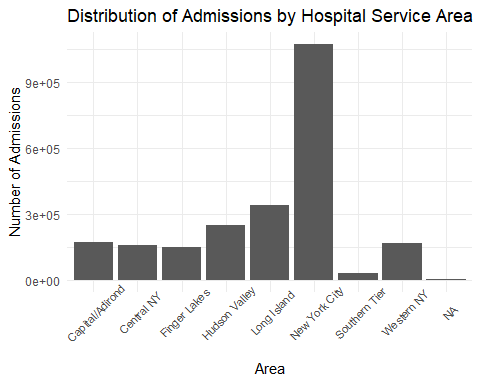
# replace payment1 "Unknown" with NA  
sum(hosp$Payment1 == "Unknown") #6896 Payment1 "Unknown"

## [1] 6896

## Examine Distributions of Variables

Next, we examine the distributions of our variables to get an idea of whether they appear appropriate.

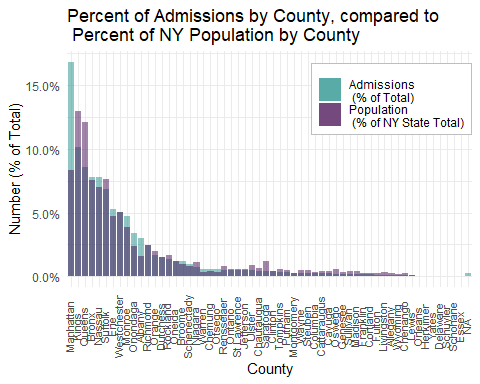
ggplot(hosp, aes(x=Area)) +   
 geom\_bar(stat="count") +   
 theme(axis.text.x = element\_text(angle = 45, size=8, vjust=.8)) +  
 ggtitle("Distribution of Admissions by Hospital Service Area") +  
 xlab("Area") +  
 ylab("Number of Admissions")

 This plot looks appropriate, and shows that most admissions come from New York City.

Next, we examine the distribution of county, adding in population data to see if the percent of total admissions for a county matches the percent of NY state population for each county. (We would expect that they should be close to eachother, if not equal.)

# get percent of admissions by county  
hosp %>%  
 count(County) %>%  
 mutate(perc = n/ nrow(hosp)) -> hosp2  
  
# load in new york population data  
ashley4 <- "C:/Users/Ashley Hirt/Documents/GitHub/FinalProject/county-pops.csv"  
countypops <- read.csv(ashley4, header=T)  
  
# left outer join of % admissions (hosp2) and population data (countypops) per county  
hosp3 <- merge(hosp2, countypops, by="County", all=T)  
  
# graph admissions by county with new york population data.  
mycolors <- c("Number of Admissions" = mygreen,   
 "Population" = mypurple)  
  
ggplot(hosp3, aes(x=reorder(County, -n), y=perc)) +   
 geom\_bar(aes(fill="Number of Admissions"), stat="identity", alpha=.5) +  
 geom\_bar(aes(y=prop.of.NY, fill="Population"),   
 stat="identity", alpha=.5) +  
 # geom\_point(aes(y=prop.of.NY, fill="Population", color="Population"),   
 # shape=21, stat="identity", alpha=.5) +  
 theme(axis.text.x = element\_text(angle = 90, size=8, vjust=0.2),  
 legend.position=c(.8,.8),  
 legend.background = element\_rect(fill="white", color="grey")) +  
 scale\_fill\_manual(name=NULL, values=mycolors,   
 labels=c("Admissions \n (% of Total)", "Population \n (% of NY State Total)"),  
 aesthetics=c("fill", "color")) +  
 ggtitle("Percent of Admissions by County, compared to \n Percent of NY Population by County") +  
 xlab("County") +  
 ylab("Number (% of Total)") +  
 scale\_y\_continuous(labels=scales::percent)

## Warning: Removed 8 rows containing missing values (position\_stack).

 From the above graph, we see that the green bar represents the percent of our dataset’s total admissions that come from each county. The light purple bar shows the population of each county, represented as a percent of NY state’s total population. Since our dataset ideally represents all of New York state, we would expect the two percentages to be equal (in the absence of any other influential factors). However, we see that some counties (notably, Manhattan, Nassau, Monroe, Onondaga, and Albany) disproportionately represent higher hospital admissions than their population numbers would predict. On the other hand, Kings, Queens, and Saratoga counties represent less of all hospital admissions than would be expected by their populations.

This could indicate disproportionate usage of the healthcare system by some counties over others. However, it could also indicate problems in the collection of our dataset (e.g., perhaps most hospitals sampled were in Manhattan and not evenly spread across all counties).

Next, we examine how cerficate compares to facility.name and facility.ID (are they redundant?).

name <- list()  
id <- list()  
  
# group by certificate, then get the unique facility name/ID with that certificate  
x <- hosp %>%  
 group\_by(Certificate) %>%  
 summarise(  
 name = list(unique(Facility.Name)),  
 id = list(unique(Facility.ID))  
 )  
x$name[c(1:10)]

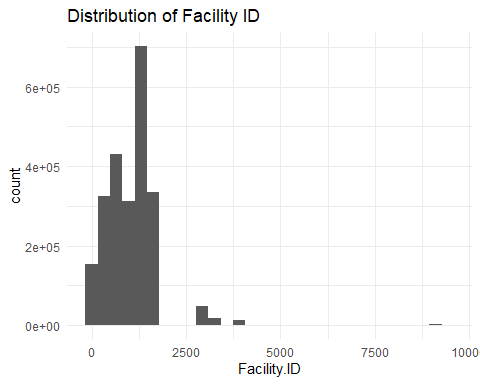
## [[1]]  
## [1] "Albany Medical Center Hospital"   
## [2] "Albany Medical Center - South Clinical Campus"  
##   
## [[2]]  
## [1] "Albany Memorial Hospital"  
##   
## [[3]]  
## [1] "St. Peter's Addiction Recovery Center"  
## [2] "St Peters Hospital"   
##   
## [[4]]  
## [1] "Cuba Memorial Hospital Inc"  
##   
## [[5]]  
## [1] "Memorial Hosp of Wm F & Gertrude F Jones A/K/A Jones Memorial Hosp"  
##   
## [[6]]  
## [1] "Our Lady of Lourdes Memorial Hospital Inc"  
##   
## [[7]]  
## [1] "United Health Services Hospitals Inc. - Binghamton General Hospital"  
## [2] "United Health Services Hospitals Inc. - Wilson Medical Center"   
##   
## [[8]]  
## [1] "Olean General Hospital"  
##   
## [[9]]  
## [1] "TLC Health Network Lake Shore Hospital"  
##   
## [[10]]  
## [1] "Auburn Community Hospital"

From this, we see that there are multiple hospital names for the same certificate number, but they are typically different campuses of the same main hospital (e.g. “Albany Medical Center Hospital” and “Albany Medical Center - South Clinical Campus”). The facility ID corresponds to these campus differences. So we assume that in our classification tasks, we will not use certificate or facility.name, we’ll just use facility.ID because it encapsulates the former two variables.

# examine distribution of facility ID  
ggplot(hosp, aes(x=Facility.ID)) +   
 geom\_histogram() +  
 ggtitle("Distribution of Facility ID")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 5325 rows containing non-finite values (stat\_bin).

 It appears unusual that some facility IDs are in the 9000s, but this is not mentioned in our data dictionary, so there is no way to know if they are valid or not.

high.id <- filter(hosp, Facility.ID>9000)  
unique(high.id$Area)

## [1] "Capital/Adirond" "Central NY"

unique(high.id$Facility.Name)

## [1] "St. Peter's Addiction Recovery Center"   
## [2] "Crouse Hospital - Commonwealth Division"  
## [3] "The Burdett Care Center"

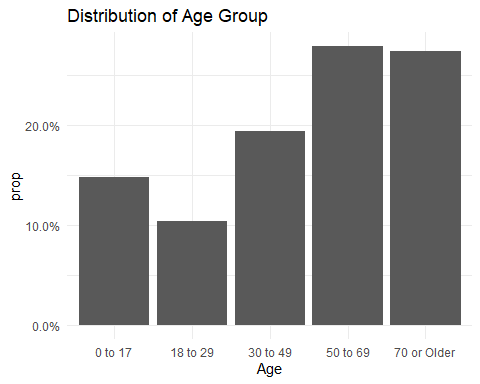
Examination of these rows indicates there are 3,151 rows with facility.ID>9000. They all come from one of 3 hospitals: St Peter’s Addiction Recovery Center, The Burdett Care Center, and Crouse Hospital - Commonwealth Division. Although their facility.ID values seem unusual, these records do appear valid, so we keep them in the dataset unchanged.

After performing this analysis, we drop the facility.name and certificate rows, but keep the Facility.ID variable.

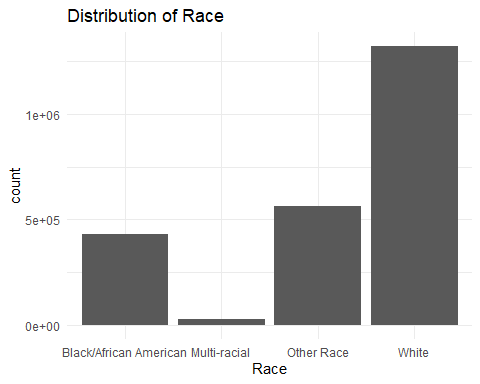
# drop columns Facility.Name and certificate  
hosp <- hosp[,-which(colnames(hosp) %in% c("Certificate", "Facility.Name"))]

Next, we examine a distribution of age groups.

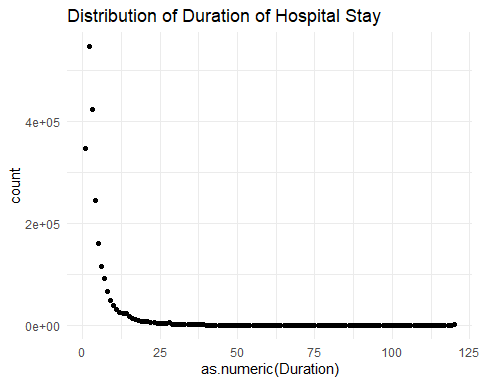
ggplot(hosp, aes(x=Age, y=..prop.., group=1)) +   
 geom\_bar() +  
 ggtitle("Distribution of Age Group") +  
 scale\_y\_continuous(labels=scales::percent)

 This appears normal, no NAs. Age groups 50-69, and 70+ have the highest amount of hospital discharge records, which is consistent with what is expected.

ggplot(hosp, aes(x=Race)) +   
 geom\_bar(stat="count") +  
 ggtitle("Distribution of Race")

 The distribution of race also appears normal, no NAs. Note that “other race” is explained in the data dictionary as coding for Asians, Pacific Islanders, and Native Americans.

y <- hosp %>% group\_by(Duration) %>% summarize(count = n())  
ggplot(y, aes(x=as.numeric(Duration), y=count)) +   
 geom\_point() +  
 ggtitle("Distribution of Duration of Hospital Stay")

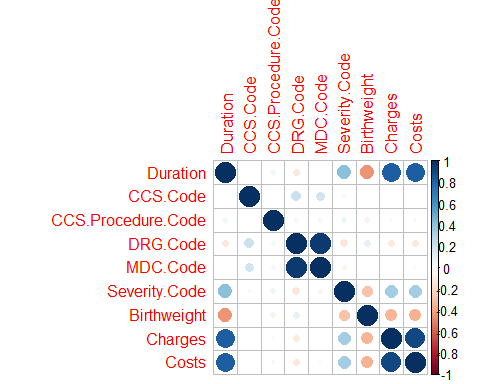
 Here we see that the vast majority of Hospital durations are less than 10 days. Also, the one row that was removed was “120+” bc it is not easily coerced to numeric. We know a priori from the data dictionary that this categorization of duration stands for any duration greater than 120 days.

Finally, we look at the correlations of our numeric variables with eachother.

library(corrplot)

## corrplot 0.84 loaded

hosp$Duration <- as.numeric(hosp$Duration)  
# get numeric columns  
num <- hosp[,c("Duration", "CCS.Code", "CCS.Procedure.Code", "DRG.Code", "MDC.Code", "Severity.Code", "Birthweight", "Charges", "Costs")]  
  
# get corrplot  
x <- cor(num[complete.cases(num)==T,]); corrplot(x)

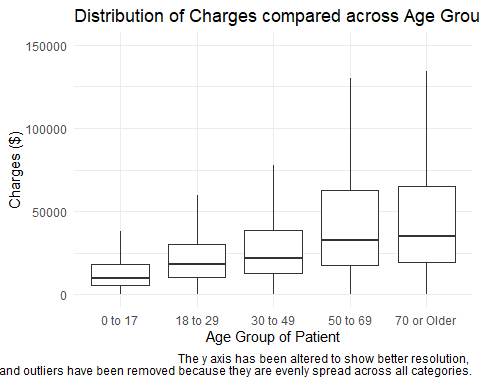
 From the corrplot, we see that DRG Code and MDC Code are highly correlated. This is surprising, considering that there are only 26 MDC codes which correspond to organ systems, while there are 300+ DRG Codes that represent specific diseases. We also see that there is a moderate positive correlation between severity code and duration, which makes sense since severity code is on a 1-4 scale with 4 being extremely severe. Also, there is a moderate negative correalation between birthweight and duration (again as expected, these likely represent premature babies that required more care). There is a strong positive correlation between charges and duration, as well as costs with duration and costs with charges. Finally, there is a weak negative correlation between birthweight and severity code (smaller babies have higher severity codes), a weak positive correlation of severity code with costs and severity code with charges (more severe cases cost the hospital more to treat, which is passed on to the patient), and a weak negative association between birthweight and charges/costs.

I am surprised to see the significance that birthweight has as a predictor for our other variables. Specifically, we can draw some preliminary conclusions: the lower the birthweight, the higher the duration, the higher the severity code, and the higher the charges/costs.

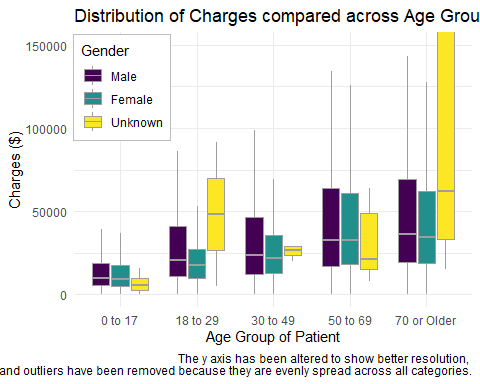
## Examine Potential Predictors of Charges

We have chosen to predict charges (cost charged to the patient) using the demographic and medical information in our dataset. To see how our predictors are correlated with this response variable (if at all), we examine graphs of boxplots for each predictor.

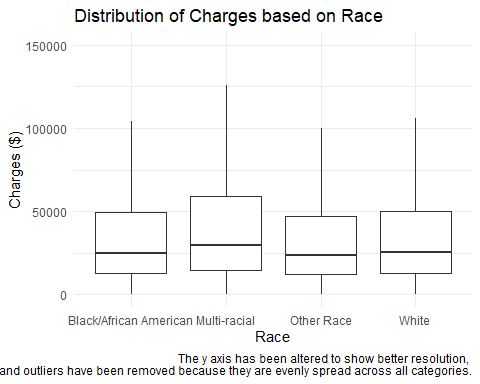
ggplot(hosp, aes(x=Age, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0, 150000)) +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.") +  
 ggtitle("Distribution of Charges compared across Age Groups") +  
 xlab("Age Group of Patient") +  
 ylab("Charges ($)")

 The above graph shows what one would generally expect: the older the patient, the higher the charges (end-of-life care is expensive and prolonged).

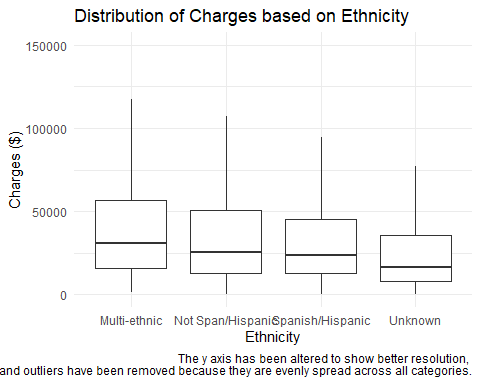
ggplot(hosp, aes(x=Age, y=Charges, fill=factor(Gender, levels=c("M", "F", "Unknown")))) +  
 geom\_boxplot(outlier.shape=NA, color="grey62") +  
 coord\_cartesian(ylim=c(0, 150000)) +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.", fill="Gender") +  
 ggtitle("Distribution of Charges compared across Age Groups") +  
 scale\_fill\_viridis\_d(labels=c("Male", "Female", "Unknown")) +  
 theme(legend.position=c(.12, .80),   
 legend.background=element\_rect(fill="white", color="grey")) +  
 xlab("Age Group of Patient") +  
 ylab("Charges ($)")

 Here, we see that charges don’t tend to change that much between males and females. The distribution is more spread for males (a higher upper whisker can be seen in all age groups), indicating more variability; however, the medians are not different from those of females. We do see that the “Unknown” gender distribution is different from that of males and females, with either markedly increased or decreased variability (as measured by the IQR). This is likely because there were only 57 Unknown gender records. Thus, we expect Gender (especially its Unknown level) to be somewhat predictive of Charges.

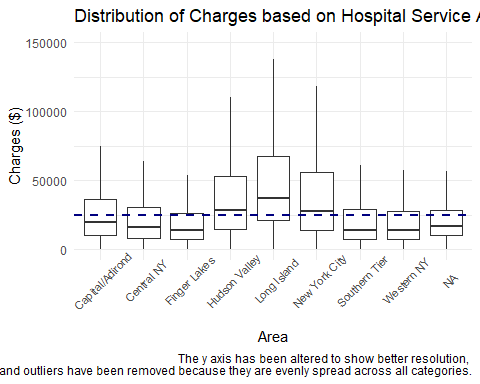
# charges vs. race  
ggplot(hosp, aes(x=Race, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,150000)) +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.") +  
 ggtitle("Distribution of Charges based on Race") +  
 xlab("Race") +  
 ylab("Charges ($)")



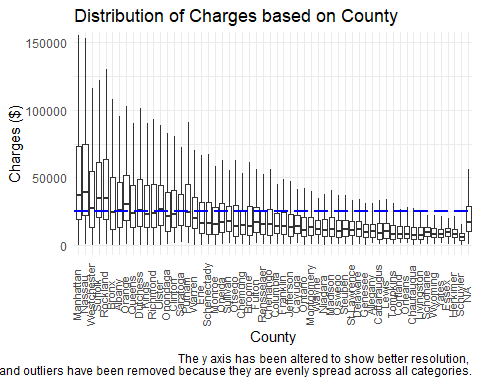
# charges vs. ethnicity  
ggplot(hosp, aes(x=Ethnicity, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,150000)) +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.") +  
 ggtitle("Distribution of Charges based on Ethnicity") +  
 xlab("Ethnicity") +  
 ylab("Charges ($)")

 Here, we see that the distribution of charges does not change much across different races or across different Ethnicities, so we would not expect Race or Ethnicity to be a good predictor of Charges.

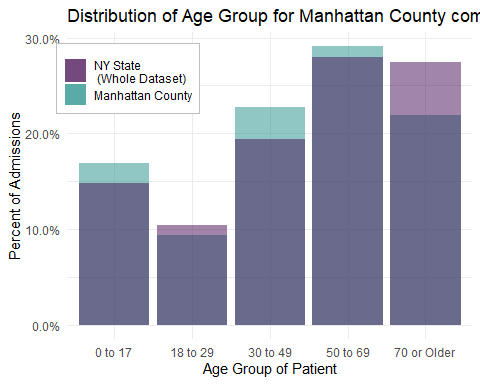
ggplot(hosp, aes(x=Area, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0, 150000)) +  
 ggtitle("Distribution of Charges based on Hospital Service Area") +  
 theme(axis.text.x = element\_text(angle = 45, size=8, vjust=.8)) +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.") +  
 xlab("Area") +  
 ylab("Charges ($)") +  
 geom\_hline(yintercept=median(hosp$Charges), color="navy blue", linetype="dashed", size=1)

 From this graph, we see that certain areas (Hudson Valley, Long Island, and NYC) have higher median charges (and overall distributions) than other areas. Thus, we would expect Area to be a good predictor of Charges.

ggplot(hosp, aes(x=reorder(County, -Charges), y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0, 150000)) +  
 ggtitle("Distribution of Charges based on County") +  
 labs(caption="The y axis has been altered to show better resolution, \n and outliers have been removed because they are evenly spread across all categories.") +  
 xlab("County") +  
 ylab("Charges ($)") +  
 theme(axis.text.x = element\_text(angle = 90, size=8, vjust=0.2)) +  
 geom\_hline(yintercept=median(hosp$Charges), size=1, color="blue", linetype="longdash")

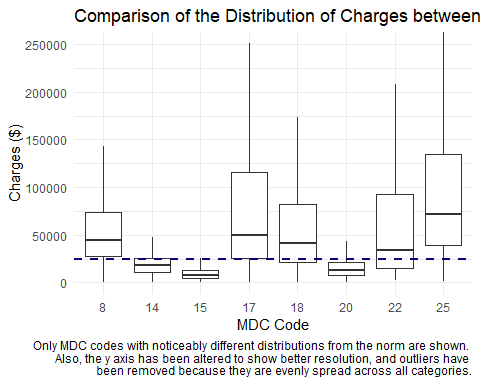
 Here, we see that Manhattan, Nassau, Westchester, Suffolk, and Rockland counties have the highest median charges. The first two of these counties also represent counties that were disproportionately represented in our dataset compared to their percent of the NY state population (see bar graph earlier in the report). I would hypothesize that hospitals in these counties are teaching hospitals or “the best”/highest funded in the state, which would both attract patients from other counties and attract rare/difficult cases that cost more to treat. Another explanation is that these counties treat older patients, which typically cost more to treat. We examine this explanation below for Manhattan County.

mycolors2 <- c("Manhattan" = mygreen,  
 "Dataset" = mypurple)  
ggplot(filter(hosp, County=="Manhattan"), aes(x=Age, y=..prop.., group=1)) +  
 geom\_bar(stat="count", aes(fill="Manhattan"), alpha=.5) +  
 scale\_y\_continuous(labels=scales::percent) +  
 geom\_bar(data=hosp, aes(x=Age, y=..prop.., group=1, fill="Dataset"), alpha=.5) +  
 scale\_fill\_manual(name=NULL, values=mycolors2, aesthetics=c("fill", "color"),   
 labels=c("NY State \n (Whole Dataset)", "Manhattan County")) +  
 theme(legend.position=c(.15,.85), legend.background=element\_rect(fill="white", color="grey")) +  
 ggtitle("Distribution of Age Group for Manhattan County compared to NY State") +  
 xlab("Age Group of Patient") +  
 ylab("Percent of Admissions")

 Here, we see that Manhattan County’s age distribution (shown in green) is actually shifted left compared to the age distribution of NY state (the whole dataset, shown in light purple). This means that Manhattan actually sees less patients that are on average younger than most other hospitals in the state. Thus, the hypothesis that Manhattan has higher median charges because it treats older patients seems untrue.

ggplot(hosp, aes(x=as.factor(MDC.Code), y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,250000)) +  
 xlab("MDC Code") +  
 ylab("Charges ($)") +  
 ggtitle("Comparison of the Distribution of Charges between MDC Codes") +  
 scale\_x\_discrete(limits=c("8", "14", "15", "17", "18", "20", "22", "25")) +  
 geom\_hline(yintercept=median(hosp$Charges), color="navy blue", linetype="dashed", size=1) +  
 labs(caption="Only MDC codes with noticeably different distributions from the norm are shown. \n Also, the y axis has been altered to show better resolution, and outliers have \n been removed because they are evenly spread across all categories.")

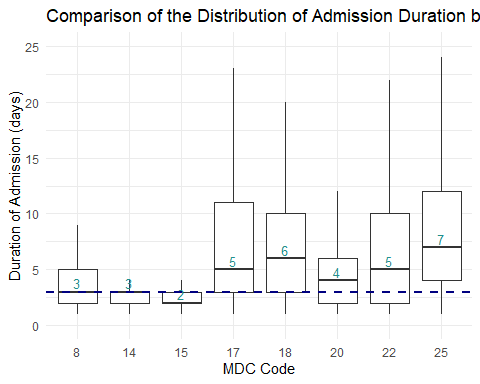
## Warning: Removed 1420735 rows containing missing values (stat\_boxplot).

 Here, we can see that diseases of certain body systems are associated with higher or lower median charges. MDC codes 8, 17, 18, 22, and 25 (respectively: Musculoskeletal, myeloproliferative, infectious disease, burns, and multiple significant trauma) have higher median charges, as expected. MDC codes 14, 15, and 20 (respectively: pregnancy/childbirth, newborn, and alcohol/drug use) have lower median charges and lower variations. We wanted to see if this difference could be explained by differences in admission durations and created the following plot.

# get median duration for each mdc.code (used for graphing)  
durs <- hosp %>%  
 group\_by(MDC.Code) %>%  
 summarise(  
 dur = median(Duration, na.rm=T),  
 charge = median(Charges)  
 ) %>%  
 arrange(desc(dur))  
  
# graph mdc code vs. duration  
ggplot(hosp, aes(x=as.factor(MDC.Code), y=Duration)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_text(data = durs, aes(x = as.factor(MDC.Code), y = dur, label = dur),   
 size = 3.5, vjust = -.3, color=mygreen) +  
 coord\_cartesian(ylim=c(0, 25)) +  
 scale\_x\_discrete(limits = c("8", "14", "15", "17", "18", "20", "22", "25")) +  
 xlab("MDC Code") +  
 ylab("Duration of Admission (days)") +  
 ggtitle("Comparison of the Distribution of Admission Duration between MDC Codes") +  
 geom\_hline(yintercept=median(hosp$Duration), size=1, linetype="dashed", color="navy blue")

## Warning: Removed 1420735 rows containing missing values (stat\_boxplot).

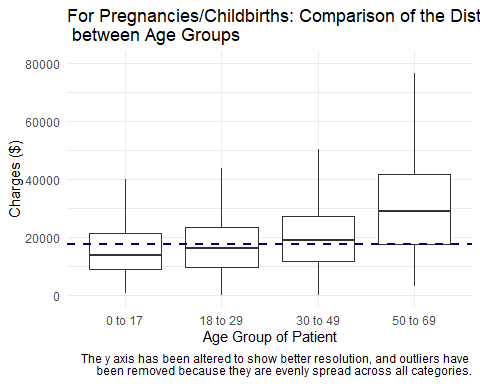
## Warning: Removed 18 rows containing missing values (geom\_text).



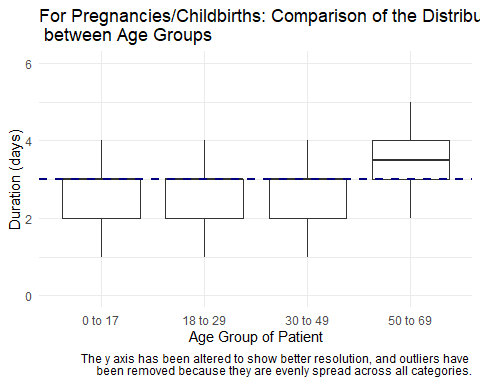
# only choosing to output MDC codes already identified as different from the median charges (previous section)

Here, we can see that some of the MDC codes’ that had noticeably different charge distributions have the same difference in their duration distributions (i.e. codes 17, 18, 22, and 25 had higher distributions of both variables, while codes 14 and 15 have lower distributions of both variables). However, we also see that codes 8 and 20 have distributions centering around the median of all records (shown as a navy blue dashed line), despite code 8 (Musculoskeletal) having a higher charge distribution and code 20 (Alcohol/Drug Use) having a lower charge distribution than normal. This shows that, although duration seems to encapsulate much of the variation in charges (and would obviously serve as a good predictor), it does not explain all of it.

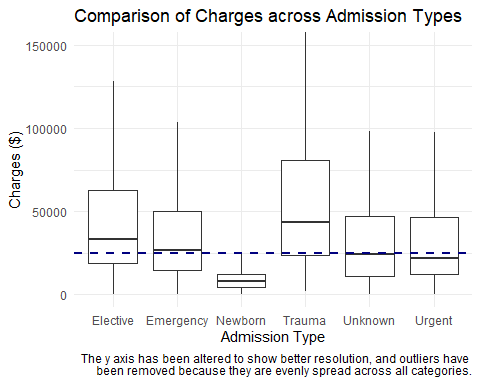
# for pregnancies: charges vs. age groups  
ggplot(filter(hosp, MDC.Code==14), aes(x=Age, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,80000)) +  
 xlab("Age Group of Patient") +  
 ylab("Charges ($)") +  
 ggtitle("For Pregnancies/Childbirths: Comparison of the Distribution of Charges \n between Age Groups") +  
 geom\_hline(yintercept=median(filter(hosp, MDC.Code==14)$Charges), color="navy blue", linetype="dashed", size=1) +  
 labs(caption="The y axis has been altered to show better resolution, and outliers have \n been removed because they are evenly spread across all categories.")



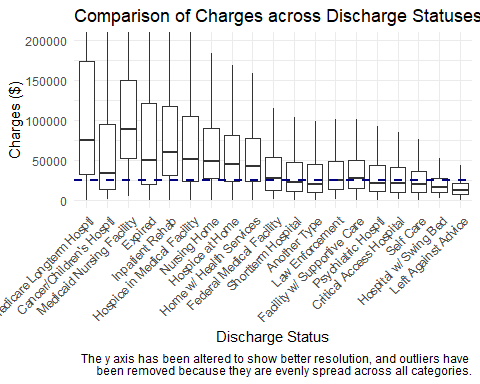
# for pregnancies: duration vs. age groups  
ggplot(filter(hosp, MDC.Code==14), aes(x=Age, y=as.numeric(Duration))) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,6)) +  
 xlab("Age Group of Patient") +  
 ylab("Duration (days)") +  
 ggtitle("For Pregnancies/Childbirths: Comparison of the Distribution of Duration \n between Age Groups") +  
 geom\_hline(yintercept=median(as.numeric(filter(hosp, MDC.Code==14)$Duration), na.rm=T), color="navy blue", linetype="dashed", size=1) +  
 labs(caption="The y axis has been altered to show better resolution, and outliers have \n been removed because they are evenly spread across all categories.")

 In the first graph, we see that older patients tend to have more expensive pregnancies/childbirths (as expected). The dark blue line represents the median charge of all pregnancies (MDC code=14). In the second graph, we see a possible explanation for this trend: older patients have longer admission durations, which is correlated with increased charges. The dark blue line represents the median duration of all pregnancies.

ggplot(hosp, aes(x=Type, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0, 150000)) +  
 ggtitle("Comparison of Charges across Admission Types") +  
 xlab("Admission Type") +  
 ylab("Charges ($)") +  
 labs(caption="The y axis has been altered to show better resolution, and outliers have \n been removed because they are evenly spread across all categories.") +  
 geom\_hline(yintercept=median(hosp$Charges), size=1, color="navy blue", linetype="dashed")

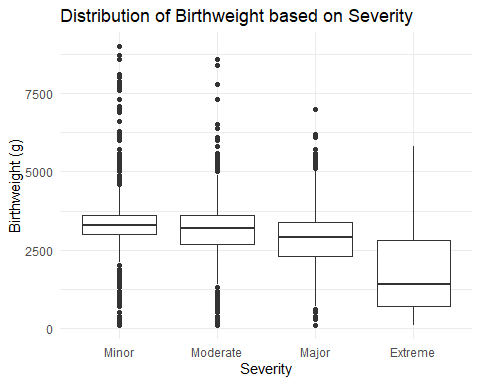
 This graph shows that Trauma admissions have higher charges, while Newborn admissions have lower charges (both as expected).

ggplot(hosp, aes(x=reorder(Discharge.Status, -Charges), y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0, 200000)) +  
 ggtitle("Comparison of Charges across Discharge Statuses") +  
 xlab("Discharge Status") +  
 ylab("Charges ($)") +  
 labs(caption="The y axis has been altered to show better resolution, and outliers have \n been removed because they are evenly spread across all categories.") +  
 geom\_hline(yintercept=median(hosp$Charges), size=1, color="navy blue", linetype="dashed") +  
 theme(axis.text.x = element\_text(angle=45, vjust=1, hjust=1)) +  
 scale\_x\_discrete(labels=c("Medicare Longterm Hosptl", "Cancer/Children's Hosptl",  
 "Medicaid Nursing Facility", "Expired", "Inpatient Rehab",  
 "Hospice in Medical Facility", "Nursing Home", "Hospice at Home",  
 "Home w/ Health Services", "Federal Medical Facility",  
 "Shortterm Hospital", "Another Type", "Law Enforcement",   
 "Facility w/ Supportive Care", "Psychiatric Hosptl",  
 "Critical Access Hospital", "Self Care", "Hospital w/ Swing Bed",  
 "Left Against Advice"))

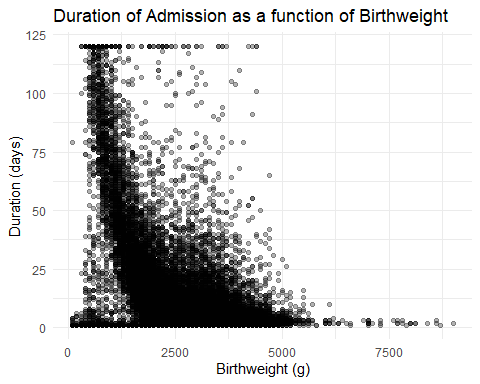
 Here, we see that Discharge Status does seem to have predictive value for Charges. Patients that were discharged to a longterm hospital, a nursing home, inpatient rehab, or hospice have higher median charges compared to those that were discharged to shortterm hospitals, supportive care, or self-care. This makes sense, as the former discharge locations would likely correspond to older patients or those with more serious conditions. Also, we note that “Expired” patients (those that died at the hospital) also have higher median charges.

Next, we move onto an analysis of birthweight.

# get df of only birth records  
births <- hosp[hosp$Birthweight>0 & !is.na(hosp$Birthweight),]  
  
# graph birthweight vs. severity  
ggplot(births[!is.na(births$Severity),], aes(  
 x=Severity, y=Birthweight)  
 ) +  
 geom\_boxplot() +  
 ggtitle("Distribution of Birthweight based on Severity") +  
 xlab("Severity") +  
 ylab("Birthweight (g)") +  
 scale\_x\_discrete(limits=c("Minor", "Moderate", "Major", "Extreme"))

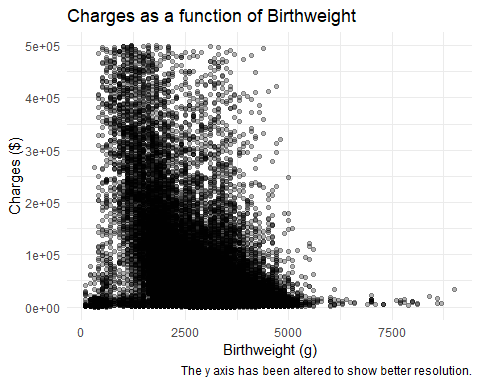


# graph birthweight vs. duration  
ggplot(births, aes(x=Birthweight, y=Duration)) +   
 geom\_point(alpha=.3) +  
 ggtitle("Duration of Admission as a function of Birthweight") +  
 xlab("Birthweight (g)") +  
 ylab("Duration (days)")



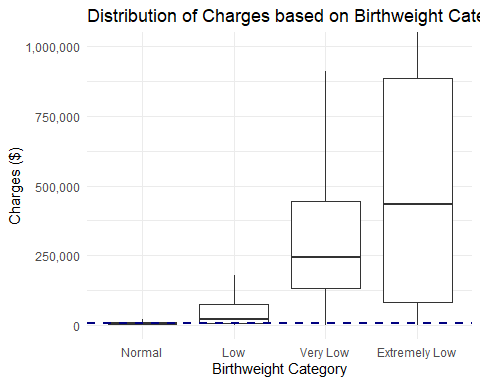
# graph birthweight vs. charges  
ggplot(births, aes(x=Birthweight, y=Charges)) +  
 geom\_point(alpha=.3) +  
 ggtitle("Charges as a function of Birthweight") +  
 xlab("Birthweight (g)") +  
 ylab("Charges ($)") +  
 ylim(0, 500000) + # to cut out outliers  
 labs(caption="The y axis has been altered to show better resolution.")

## Warning: Removed 1682 rows containing missing values (geom\_point).

 These graphs show many expected trends shown by our correlation plot earlier. As birthweight decreases, severity increases, the duration of the hospital admission increases, and charges increase.

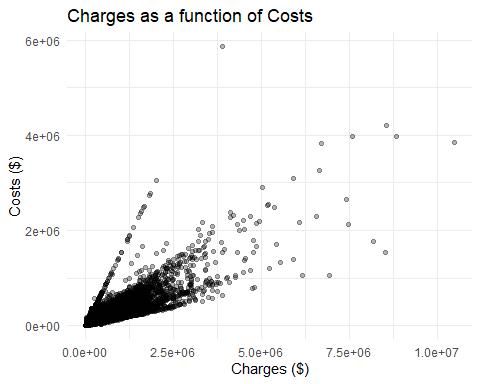
In terms of feature engineering, we decided to create a feature called birthweight.cat, which divided birthweights into the categories (based on medical literature) of “Normal”, “Low”, “Very Low”, and “Extremely Low.” Below, we see that a graph of charges distribution as a function of these categories shows clearer separation of charge distributions than birthweight or severity alone (the median birth charge is shown as a blue dashed line). Thus, we expect this feature to be a good predictor of charges.

# Engineered feature to create birthweight cutoffs (birthweight.cat)  
hosp <- hosp %>%  
 mutate(birthweight.cat = case\_when(  
 Birthweight < 2500 & Birthweight > 1500 ~ "Low",  
 Birthweight <= 1500 & Birthweight > 1000 ~ "Very Low",  
 Birthweight <= 1000 ~ "Extremely Low",  
 TRUE ~ "Normal"  
 ))  
# reinitialize births df  
births <- hosp[hosp$Birthweight>0 & !is.na(hosp$Birthweight),]   
  
# graph birthweight.cat vs. charges  
ggplot(births, aes(x=birthweight.cat, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 coord\_cartesian(ylim=c(0,1000000)) +  
 ggtitle("Distribution of Charges based on Birthweight Category") +  
 xlab("Birthweight Category") +  
 ylab("Charges ($)") +  
 scale\_x\_discrete(limits=c("Normal", "Low", "Very Low", "Extremely Low")) +  
 scale\_y\_continuous(labels=scales::comma) +  
 geom\_hline(yintercept=median(births$Charges), size=1, linetype="dashed", color="navy blue")



We also created a feature called charge ratio, defined as the total charges (amount billed to the consumer) divided by the costs (amount the hospital spent while taking care of you).

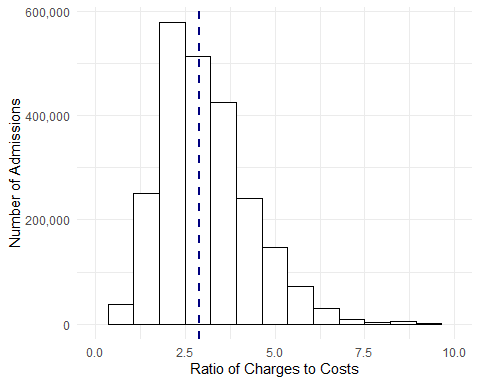
# create charge ratio variable  
hosp$chargeratio <- hosp$Charges / hosp$Costs  
  
ggplot(hosp, aes(x=Charges, y=Costs)) +   
 geom\_point(alpha=.3) +  
 ggtitle("Charges as a function of Costs") +  
 xlab("Charges ($)") +  
 ylab("Costs ($)")



ggplot(hosp, aes(x=chargeratio)) +  
 geom\_histogram(bins=15, fill="white", color="black") +  
 xlim(0,10) +  
 geom\_vline(xintercept=median(hosp$chargeratio), linetype="dashed", color="navy blue", size=1) +  
 xlab("Ratio of Charges to Costs") +  
 ylab("Number of Admissions") +  
 #ggtitle("Distribution of Charge Ratio") +  
 scale\_y\_continuous(labels=scales::comma)

## Warning: Removed 23815 rows containing non-finite values (stat\_bin).

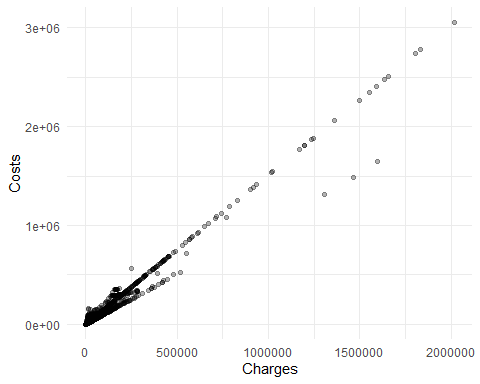
## Warning: Removed 2 rows containing missing values (geom\_bar).



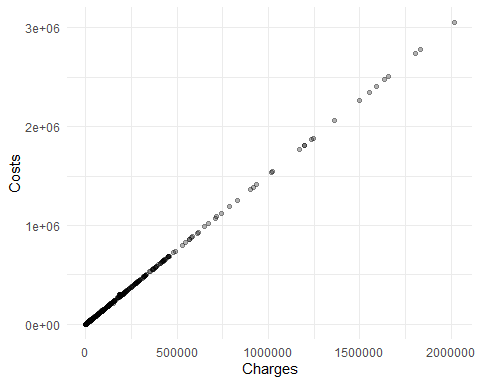
# This distribution is obviously right-skewed, with a median of around 2.   
# note removal of rows with 0 cost are being removed.

Investigating the weird magic slope

# in search of the magic slope  
temp <- hosp %>%  
 filter(Charges < 2500000) %>%  
 filter(chargeratio < 1)  
  
ggplot(temp, aes(x=Charges, y=Costs)) +  
 geom\_point(alpha=.3)

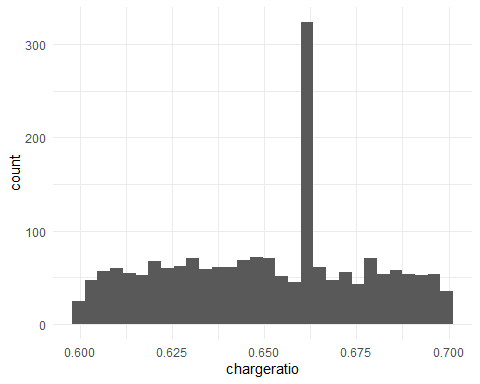


# magic slope looks (by visual inspection) to be 2/3  
# filter between .7 and .6 chargeratio  
temptemp <- temp %>%  
 filter(chargeratio < .7 & chargeratio > .6)  
ggplot(temptemp, aes(x=Charges, y=Costs)) +  
 geom\_point(alpha=.3)

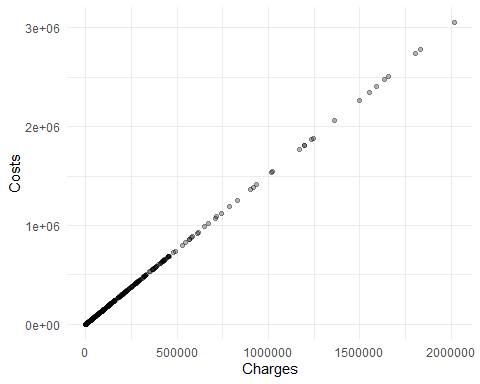


ggplot(temptemp, aes(x=chargeratio)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

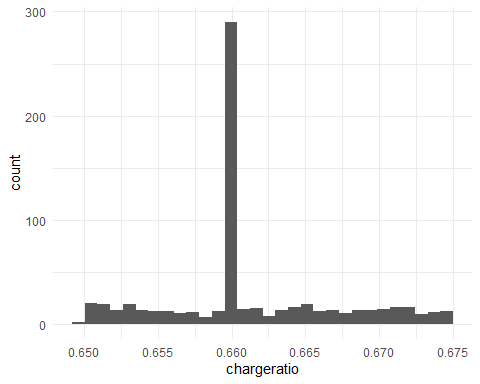


# filter between .675 and .65 chargeratio  
temptemptemp <- temptemp %>%  
 filter(chargeratio < .675 & chargeratio > .65)  
ggplot(temptemptemp, aes(x=Charges, y=Costs)) +  
 geom\_point(alpha=.3)

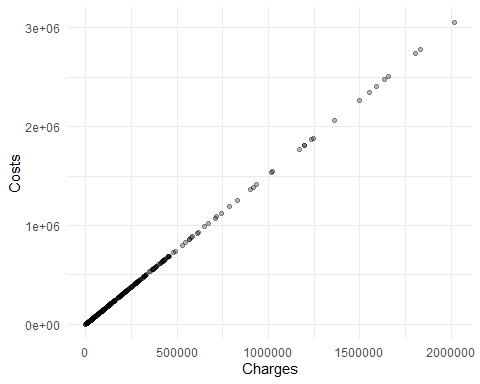


ggplot(temptemptemp, aes(x=chargeratio)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

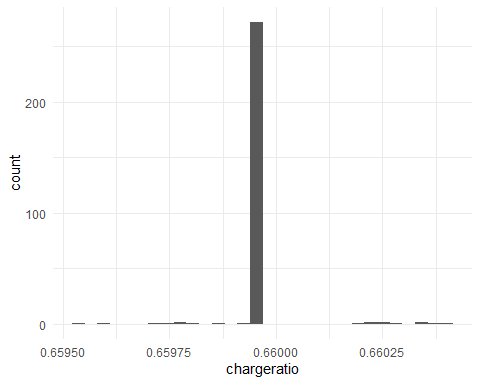


# filter between .6595 and .6605  
temptemptemptemp <- temptemptemp %>%  
 filter(chargeratio < .6605 & chargeratio > .6595)  
ggplot(temptemptemptemp, aes(x=Charges, y=Costs)) +  
 geom\_point(alpha=.3)



ggplot(temptemptemptemp, aes(x=chargeratio)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



table(round(temptemptemptemp$chargeratio, digits=5))

##   
## 0.65955 0.65959 0.65971 0.65976 0.65977 0.65979 0.65981 0.65986 0.65993   
## 1 1 1 1 1 1 1 1 1   
## 0.65994 0.65995 0.66018 0.66022 0.66026 0.66029 0.66033 0.66035 0.66038   
## 270 2 1 2 2 1 1 1 1   
## 0.66042   
## 1

# we see most observations fall right at .65994  
  
  
filtered <- filter(temptemptemptemp, round(chargeratio, digits=6)==.659944)  
table(filtered$CCS.Code) # most are code 131: respiratory failure

##   
## 14 78 82 95 99 106 108 118 127 131 146 148 155 159 162 199 201 238   
## 3 1 1 1 1 1 1 1 1 246 2 1 1 1 2 1 1 3

table(filtered$CCS.Procedure.Code) # most are code 216: resp intub/mech ventil

##   
## 0 41 175 177 193 196 197 216 231   
## 22 6 1 2 3 2 1 231 1

table(filtered$DRG.Code) # most are 130 or 133: resp system w/ ventilator support 96+ hours OR respiratory failure.

##   
## 40 49 52 130 133 140 194 197 201 240 244 248 254 380 463 468 862 890   
## 1 1 1 142 96 1 2 1 1 3 2 1 1 1 1 2 1 5   
## 892   
## 6

table(filtered$Surgical) # all medical

##   
## Medical   
## 269

table(filtered$Payment1) # almost all are Medicaid or Medicare

##   
## Blue Cross/Blue Shield Medicaid Medicare   
## 11 84 149   
## Miscellaneous/Other Private Health Insurance Self-Pay   
## 7 3 15

table(filtered$Facility.ID) # all Facility.ID=1486

##   
## 1486   
## 257

hosp\_og[hosp\_og$Facility.Id == 1486 & !is.na(hosp\_og$Facility.Id),c("Facility.Id", "Facility.Name")][1,] # this facility ID corresponds to the Henry J. Carter Specialty Hospital

## # A tibble: 1 x 2  
## Facility.Id Facility.Name   
## <int> <chr>   
## 1 1486 Henry J. Carter Specialty Hospital

## Create indicator variables

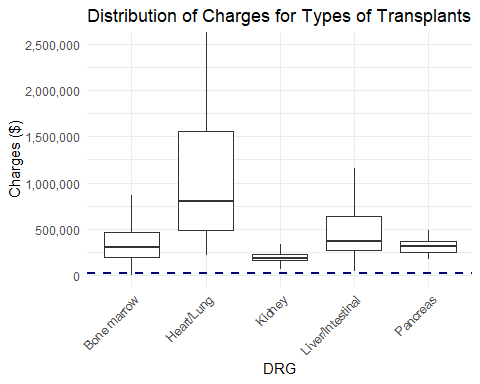
For all values replaced with NAs or other abnormalities noted above, we will create indicator variables in case there is systematic missingness.

hosp <- hosp %>%  
 mutate(redacted = ifelse(is.na(hosp$Area), 1, 0),  
 na.zip = ifelse(is.na(Zip), 1, 0),  
 invalid.drg = ifelse(is.na(DRG.Code), 1, 0),  
 na.payment1 = ifelse(Payment1=="Unknown", 1, 0),  
 no.payment2 = ifelse(is.na(Payment2), 1, 0),  
 unknown.payment2 = ifelse(Payment2=="Unknown", 1, 0),  
 no.payment3 = ifelse(is.na(Payment3), 1, 0),  
 unknown.payment3 = ifelse(Payment3 == "Unknown", 1, 0),  
 if.birth = ifelse(is.na(Birthweight), 0, 1),  
 na.gender = ifelse(is.na(Gender), 1, 0),  
 na.ethnicity = ifelse(is.na(Ethnicity), 1, 0),  
 na.surgical = ifelse(is.na(Surgical), 1, 0),  
 high.id = ifelse(Facility.ID > 9000, 1, 0),  
 na.type = ifelse(is.na(Type), 1, 0),  
 na.discharge = ifelse(is.na(Discharge.Status), 1, 0),  
 na.ccs = ifelse(is.na(CCS.Code), 1, 0),  
 na.ccs.procedure = ifelse(is.na(CCS.Procedure.Code), 1, 0),  
 na.mdc = ifelse(is.na(MDC.Code), 1, 0)  
 )

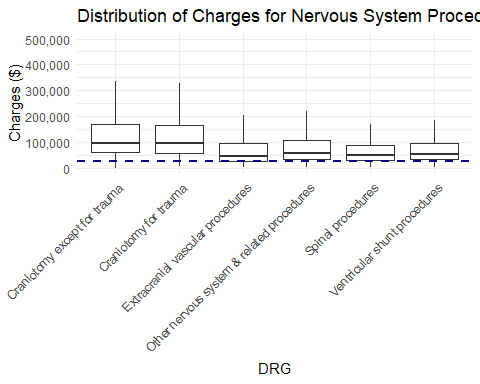
# Test Center

# feature engineering  
  
  
# median income for the diff zipcodes  
# read in income file  
rakib2 <- "C:\\Users\\Rakib Kamal\\Documents\\GitHub\\FinalProject\\zipcodes - Sheet1.csv"  
ashley2 <- "C:/Users/Ashley Hirt/Documents/GitHub/FinalProject/zipcodes - Sheet1.csv"  
incomes <- read.csv(file=ashley2)  
colnames(incomes)[1] <- "Zip" # rename zip.code column  
# perform left outer join on our hosp dataset and the income dataset, by Zipcode  
hosp <- merge(hosp, incomes, by="Zip", all=T)  
# rename the median income column  
colnames(hosp)[which(colnames(hosp)=="Median.Income..family.")] <- "Median.Income"  
  
  
# number of payment types  
hosp <- hosp %>%  
 mutate(numpay = case\_when(  
 is.na(Payment2) & is.na(Payment3) ~ 1,  
 !is.na(Payment2) & is.na(Payment3) ~ 2,  
 !is.na(Payment2) & !is.na(Payment3) ~ 3  
 ))  
  
# self-pay indicator  
hosp <- hosp %>%  
 mutate(selfpay = case\_when(  
 Payment1 == "Self-Pay" ~ 1,  
 Payment2 == "Self-Pay" ~ 1,  
 Payment3 == "self-Pay" ~ 1,  
 TRUE ~ 0  
 ))  
  
# overall grouping for each procedure (https://www.hcup-us.ahrq.gov/reports/statbriefs/sb154.jsp app  
procs <- hosp %>%  
 group\_by(CCS.Procedure.Code) %>%  
 summarise(  
 proc = unique(CCS.Procedure)  
 ) %>%  
 select(CCS.Procedure.Code, proc)  
  
hosp <- hosp %>%  
 mutate(CCS.Procedure.Group = case\_when(  
 CCS.Procedure.Code %in% c(1:9, 199, 188) ~ "Nervous System",  
 CCS.Procedure.Code %in% c(10:12) ~ "Endocrine System",  
 CCS.Procedure.Code %in% c(13:21, 220) ~ "Eyes",  
 CCS.Procedure.Code %in% c(22:33) ~ "ENT",  
 CCS.Procedure.Code %in% c(34:42, 216:217, 208) ~ "Respiratory System",  
 CCS.Procedure.Code %in% c(43:63, 201:203, 205, 225, 189:191, 193, 204) ~ "Cardiovascular System",  
 CCS.Procedure.Code %in% c(64:67, 222, 228) ~ "Hemic and Lymphatic System",  
 CCS.Procedure.Code %in% c(69:99, 194, 223, 184:186, 221) ~ "Digestive System",  
 CCS.Procedure.Code %in% c(100:112, 195, 187, 200) ~ "Urinary System",  
 CCS.Procedure.Code %in% c(113:118) ~ "Male Genital Organs",  
 CCS.Procedure.Code %in% c(119:123) ~ "Female Genital Organs",  
 CCS.Procedure.Code %in% c(124:141) ~ "Obstetrical Procedures",  
 CCS.Procedure.Code %in% c(142:164, 207, 212:215, 230) ~ "Musculoskeletal",  
 CCS.Procedure.Code %in% c(165:175, 182) ~ "Integumentary System and Breast",  
 CCS.Procedure.Code %in% c(176:181, 192, 183, 219, 218, 197, 198, 209, 210, 196, 211, 226, 227, 229, 231, 224) ~ "Misc or Diagnostic",  
 CCS.Procedure.Code == 0 ~ "Ungroupable",  
 is.na(CCS.Procedure.Code) ~ "No Procedure"  
 ))  
  
  
# median income for each county (used to extrapolate hospital tax funds in the area)  
# read in county data  
rakib3 <- "C:\\Users\\Rakib Kamal\\Documents\\GitHub\\FinalProject\\county.csv"  
ashley3 <- "C:/Users/Ashley Hirt/Documents/GitHub/FinalProject/county.csv"  
county <- read.csv(ashley3)  
# left outer join with county data  
hosp <- merge(hosp, county, by="County", all=T)

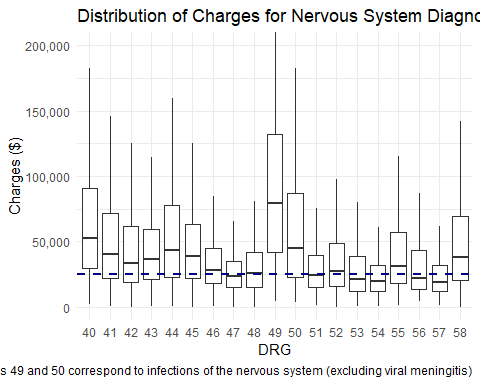
# graph transplant vs. charges  
ggplot(filter(hosp, DRG.Code < 4 | DRG.Code==440 | DRG.Code==6),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,2500000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 theme(axis.text.x = element\_text(angle=45, vjust=1, hjust=1)) +  
 scale\_x\_discrete(labels=c("Bone marrow", "Heart/Lung",   
 "Kidney", "Liver/Intestinal",  
 "Pancreas")) +  
 ggtitle("Distribution of Charges for Types of Transplants") +  
 xlab("DRG") +  
 ylab("Charges ($)")



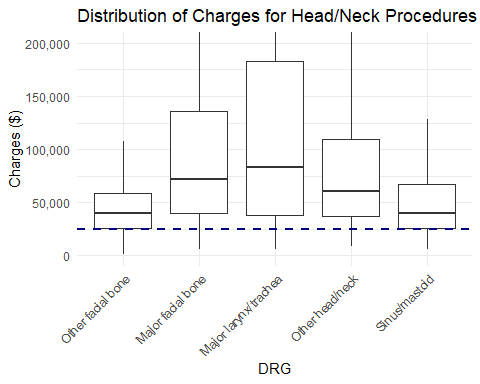
# graph nervous system procedures vs. charges  
ggplot(filter(hosp, DRG.Code <40 & DRG.Code >= 20),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,500000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 theme(axis.text.x = element\_text(angle=45, vjust=1, hjust=1)) +  
 ggtitle("Distribution of Charges for Nervous System Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



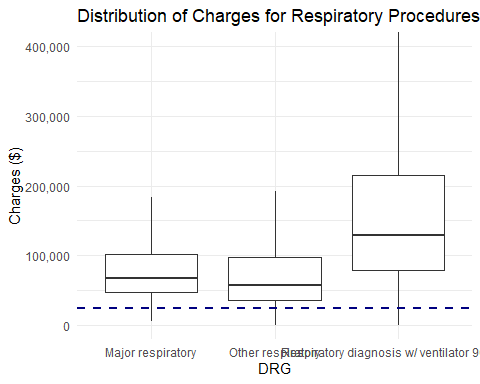
# graph nervous system diagnoses vs. charges  
ggplot(filter(hosp, DRG.Code <60 & DRG.Code >= 40),   
 aes(x=as.factor(DRG.Code), y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,200000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Nervous System Diagnoses") +  
 xlab("DRG") +  
 ylab("Charges ($)") +  
 labs(caption="Codes 49 and 50 correspond to infections of the nervous system (excluding viral meningitis)")



# graph head/neck procedures vs charges  
ggplot(filter(hosp, DRG.Code <=93 & DRG.Code >= 89),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,200000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 theme(axis.text.x=element\_text(angle=45, hjust=1, vjust=1)) +  
 scale\_x\_discrete(labels=c("Other facial bone", "Major facial bone",  
 "Major larynx/trachea", "Other head/neck",  
 "Sinus/mastoid")) +  
 ggtitle("Distribution of Charges for Head/Neck Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



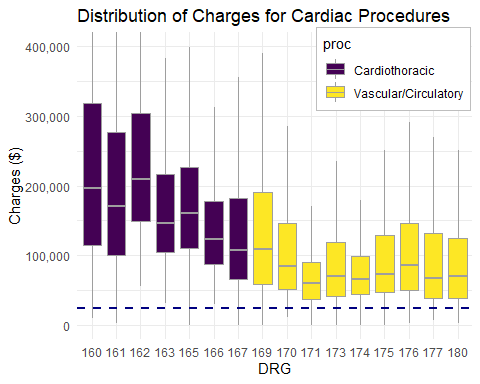
# graph respiratory procedures vs. charges  
ggplot(filter(hosp, DRG.Code <=130 & DRG.Code >= 120),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,400000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Major respiratory", "Other respiratory", "Respiratory diagnosis w/ ventilator 96+ hrs")) +  
 ggtitle("Distribution of Charges for Respiratory Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



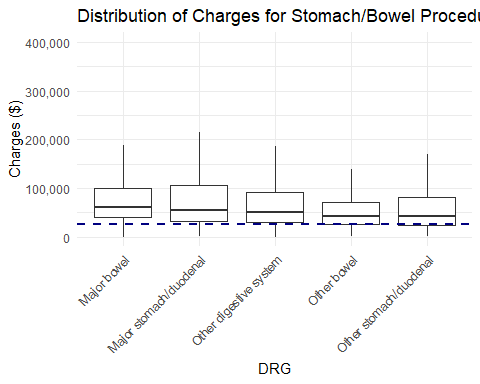
# create indicator var for cardiothoracic vs. vascular/circulatory procedures (purely for graphing)  
hosp\_temp <- hosp %>%  
 filter(DRG.Code <= 180 & DRG.Code >= 160) %>%  
 mutate(  
 proc = case\_when(  
 DRG.Code <= 167 ~ "Cardiothoracic",  
 DRG.Code > 167 ~ "Vascular/Circulatory"  
 )  
 )  
# graph cardiac procedures vs. charges  
ggplot(hosp\_temp, aes(x=as.factor(DRG.Code), y=Charges, fill=proc)) +  
 geom\_boxplot(outlier.shape=NA, color="grey62") +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,400000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_fill\_viridis\_d() +  
 ggtitle("Distribution of Charges for Cardiac Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)") +  
 theme(legend.position=c(.8,.88),   
 legend.background=element\_rect(fill="white", color="grey"),  
 legend.title=element\_text("Type of Cardiac Procedure"))

## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font  
## family not found in Windows font database

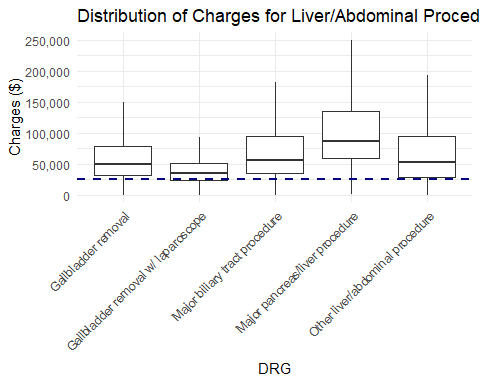
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :  
## font family not found in Windows font database



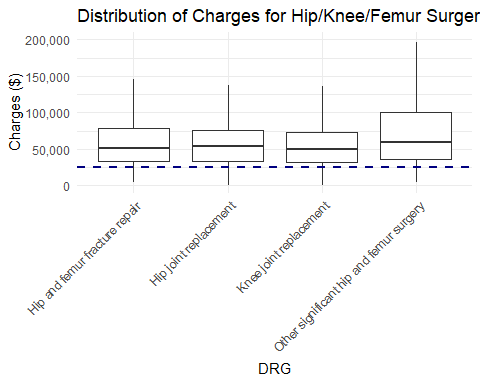
# graph stomach/bowel procedures vs. charges  
ggplot(filter(hosp, (DRG.Code <=223 & DRG.Code >= 220) | DRG.Code==229),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,400000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 theme(axis.text.x=element\_text(angle=45, vjust=1, hjust=1)) +  
 scale\_x\_discrete(labels=c("Major bowel", "Major stomach/duodenal",   
 "Other digestive system", "Other bowel", "Other stomach/duodenal")) +  
 ggtitle("Distribution of Charges for Stomach/Bowel Procedures (excl. hernia)") +  
 xlab("DRG") +  
 ylab("Charges ($)")



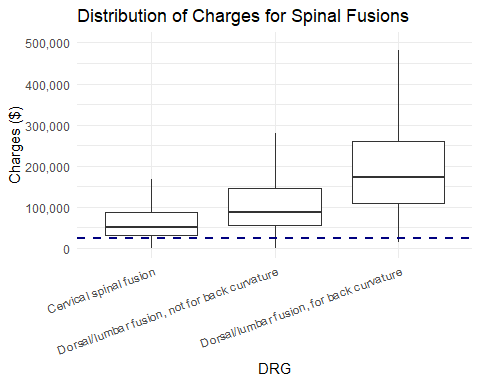
# graph liver/abdominal procedures vs. charges  
ggplot(filter(hosp, (DRG.Code <=264 & DRG.Code >= 260)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,250000)) +  
 theme(axis.text.x=element\_text(angle=45, hjust=1, vjust=1)) +  
 scale\_x\_discrete(labels=c("Gallbladder removal", "Gallbladder removal w/ laparoscope",   
 "Major biliary tract procedure", "Major pancreas/liver procedure",  
 "Other liver/abdominal procedure")) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Liver/Abdominal Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



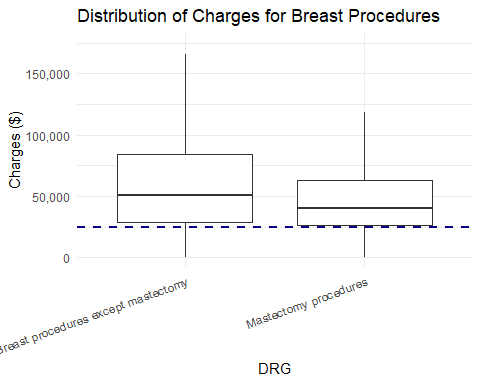
# graph hip/knee/femur surgery vs. charges  
ggplot(filter(hosp, DRG.Code %in% c(301, 302, 308, 309)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,200000)) +  
 theme(axis.text.x=element\_text(angle=45, hjust=1, vjust=1)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Hip/Knee/Femur Surgery") +  
 xlab("DRG") +  
 ylab("Charges ($)")



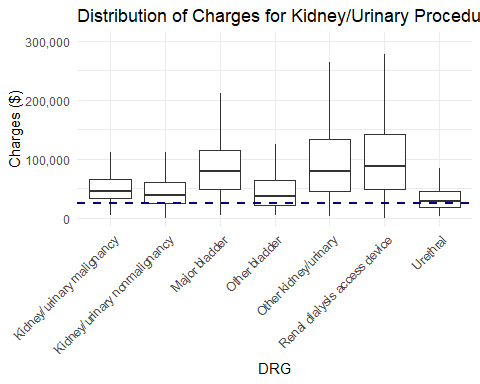
# graph spinal fusions vs. charges  
ggplot(filter(hosp, DRG.Code %in% c(303, 304, 321)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,500000)) +  
 scale\_x\_discrete(labels=c("Cervical spinal fusion",   
 "Dorsal/lumbar fusion, not for back curvature",   
 "Dorsal/lumbar fusion, for back curvature")) +  
 theme(axis.text.x=element\_text(angle=20, hjust=1, vjust=1)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Spinal Fusions") +  
 xlab("DRG") +  
 ylab("Charges ($)")



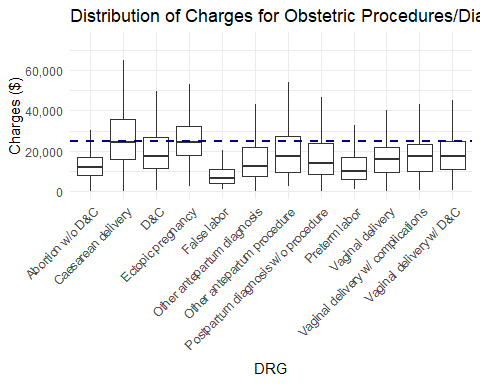
# graph breast procedures vs. charges  
ggplot(filter(hosp, (DRG.Code <=363 & DRG.Code >= 362)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,175000)) +  
 theme(axis.text.x=element\_text(angle=20, hjust=1, vjust=1)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Breast Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



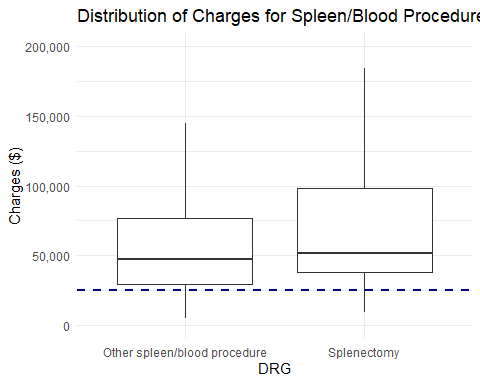
# graph kidney/urinary procedures vs. charges  
ggplot(filter(hosp, (DRG.Code <=447 & DRG.Code >= 441)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,300000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 theme(axis.text.x=element\_text(angle=45, hjust=1, vjust=1)) +  
 scale\_x\_discrete(labels=c("Kidney/urinary malignancy",   
 "Kidney/urinary nonmalignancy",  
 "Major bladder",  
 "Other bladder",  
 "Other kidney/urinary",  
 "Renal dialysis access device",  
 "Urethral")) +  
 ggtitle("Distribution of Charges for Kidney/Urinary Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



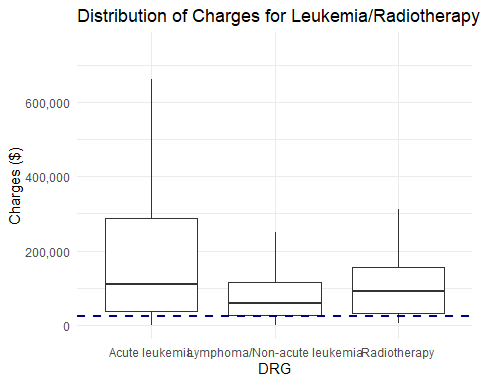
# graph ostetric procedures/diagnoses vs. charges  
ggplot(filter(hosp, (DRG.Code <=566 & DRG.Code >= 540)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,75000)) +  
 theme(axis.text.x=element\_text(angle=45, hjust=1, vjust=1)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Abortion w/o D&C",  
 "Caesarean delivery",  
 "D&C",  
 "Ectopic pregnancy",  
 "False labor",  
 "Other antepartum diagnosis",  
 "Other antepartum procedure",  
 "Postpartum diagnosis w/o procedure",  
 "Preterm labor",  
 "Vaginal delivery",  
 "Vaginal delivery w/ complications",  
 "Vaginal delivery w/ D&C")) +  
 ggtitle("Distribution of Charges for Obstetric Procedures/Diagnoses") +  
 xlab("DRG") +  
 ylab("Charges ($)")



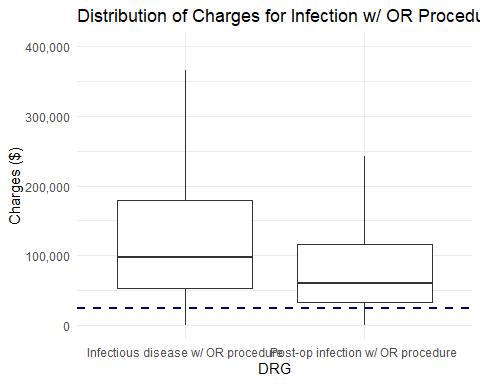
# graph spleen/blood procedures vs. charges  
ggplot(filter(hosp, (DRG.Code <=651 & DRG.Code >= 650)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,200000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Other spleen/blood procedure", "Splenectomy")) +  
 ggtitle("Distribution of Charges for Spleen/Blood Procedures") +  
 xlab("DRG") +  
 ylab("Charges ($)")



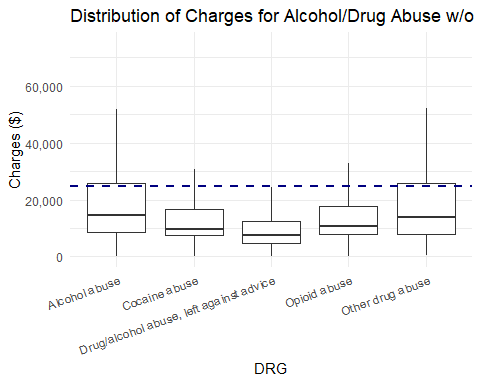
# graph leukemia/radiotherapy vs. charges  
ggplot(filter(hosp, (DRG.Code <=692 & DRG.Code >= 690)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,750000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Acute leukemia","Lymphoma/Non-acute leukemia",  
 "Radiotherapy")) +  
 ggtitle("Distribution of Charges for Leukemia/Radiotherapy") +  
 xlab("DRG") +  
 ylab("Charges ($)")



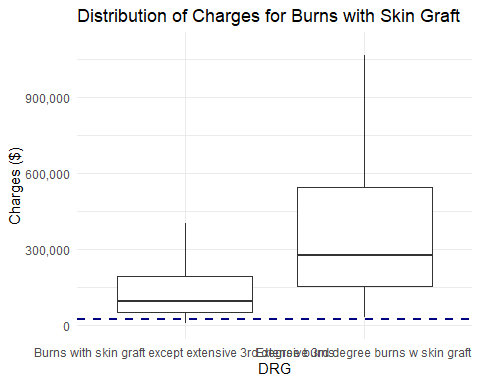
# graph infection w/ OR procedure vs. charges  
ggplot(filter(hosp, (DRG.Code <=711 & DRG.Code >= 710)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,400000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Infectious disease w/ OR procedure",   
 "Post-op infection w/ OR procedure")) +  
 ggtitle("Distribution of Charges for Infection w/ OR Procedure") +  
 xlab("DRG") +  
 ylab("Charges ($)")



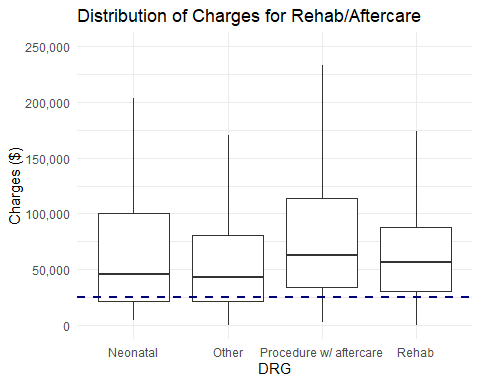
# graph alcohol/drug abuse w/o rehab vs charges  
ggplot(filter(hosp, (DRG.Code <=776 & DRG.Code >= 773) | DRG.Code == 770),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,75000)) +  
 theme(axis.text.x=element\_text(angle=20, hjust=1, vjust=1)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 scale\_x\_discrete(labels=c("Alcohol abuse", "Cocaine abuse",   
 "Drug/alcohol abuse, left against advice",   
 "Opioid abuse", "Other drug abuse")) +  
 ggtitle("Distribution of Charges for Alcohol/Drug Abuse w/o Rehab") +  
 xlab("DRG") +  
 ylab("Charges ($)")



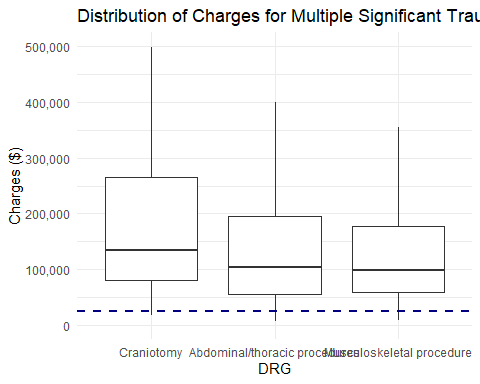
# graph burns w/ skin graft vs. charges  
ggplot(filter(hosp, (DRG.Code <=842 & DRG.Code >= 841)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,1100000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Burns with Skin Graft") +  
 xlab("DRG") +  
 ylab("Charges ($)")



# graph rehab/aftercare vs. charges  
ggplot(filter(hosp, (DRG.Code %in% c(850, 860, 862, 863))),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,250000)) +  
 scale\_x\_discrete(labels=c("Neonatal", "Other", "Procedure w/ aftercare", "Rehab")) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Rehab/Aftercare") +  
 xlab("DRG") +  
 ylab("Charges ($)")



# graph multiple significant trauma with procedure vs. charges  
ggplot(filter(hosp, (DRG.Code <=912 & DRG.Code >= 910)),   
 aes(x=DRG, y=Charges)) +  
 geom\_boxplot(outlier.shape=NA) +  
 geom\_hline(yintercept=median(hosp$Charges), linetype="dashed", size=1, color="navy blue") +  
 coord\_cartesian(ylim=c(0,500000)) +  
 scale\_y\_continuous(labels=scales::comma) +  
 ggtitle("Distribution of Charges for Multiple Significant Trauma with Procedure") +  
 scale\_x\_discrete(labels=c("Craniotomy", "Abdominal/thoracic procedures",   
 "Musculoskeletal procedure")) +  
 xlab("DRG") +  
 ylab("Charges ($)")



# create DRG vars based off preceding graphs  
hosp <- hosp %>%  
 mutate(drg.group = case\_when(  
 DRG.Code<4 | DRG.Code==6 | DRG.Code==440 ~ "Transplant",  
 DRG.Code < 40 & DRG.Code >=20 ~ "Nervous System Procedures",  
 DRG.Code == 49 | DRG.Code == 50 ~ "Nervous System Infection (excl. meningitis)",  
 DRG.Code >= 89 & DRG.Code <= 93 ~ "Head/Neck Procedures",  
 DRG.Code >= 120 & DRG.Code <= 130 ~ "Respiratory Procedures",  
 DRG.Code <= 223 & DRG.Code >= 220 ~ "Stomach/Bowel Procedures (excl. hernia)",  
 DRG.Code == 229 ~ "Stomach/Bowel Procedures (excl. hernia)",  
 DRG.Code <= 264 & DRG.Code >= 260 ~ "Liver/Abdominal Procedures",  
 DRG.Code %in% c(301, 302, 308, 309) ~ "Hip/Knee/Femur Surgery",  
 DRG.Code %in% c(303, 304, 321) ~ "Spinal Fusion",  
 DRG.Code %in% c(362, 363) ~ "Breast Procedures",  
 DRG.Code <= 447 & DRG.Code >= 441 ~ "Kidney/Urinary Procedures",  
 DRG.Code <=566 & DRG.Code >= 540 ~ "Obstetric Procedures/Diagnoses",  
 DRG.Code %in% c(650, 651) ~ "Spleen/Blood Procedures",  
 DRG.Code <=692 & DRG.Code >= 690 ~ "Leukemia/Radiotherapy",  
 DRG.Code <=711 & DRG.Code >= 710 ~ "Infection w/ OR Procedure",  
 (DRG.Code <=776 & DRG.Code >= 773) | DRG.Code == 770 ~ "Alcohol/Drug Abuse w/o Rehab",  
 DRG.Code <=842 & DRG.Code >= 841 ~ "Burns with Skin Graft",  
 DRG.Code %in% c(850, 860, 862, 863) ~ "Rehab/Aftercare",  
 DRG.Code <=912 & DRG.Code >= 910 ~ "Multiple Significant Trauma with Procedure",  
 TRUE ~ "Other"  
 ))

f <- hosp %>%  
 filter(complete.cases(.))  
myindex <- sample(1:nrow(f), 10000)  
ff <- f[myindex,]  
write.csv(f, file="sample for ml.csv", row.names=F)