MIS 500 Porfolio

Michael Pearson

2/3/2020

This is some stuff to get writing going

## Connect to the page and download the zipped files  
  
temp <- tempfile()  
  
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00296/dataset\_diabetes.zip", temp)  
  
post\_zipped <- unzip(temp)  
  
print(post\_zipped)

## [1] "./dataset\_diabetes/diabetic\_data.csv"  
## [2] "./dataset\_diabetes/IDs\_mapping.csv"

diabetes\_data <- read.csv("./dataset\_diabetes/diabetic\_data.csv", header = TRUE)  
  
print(head(diabetes\_data))

## encounter\_id patient\_nbr race gender age weight  
## 1 2278392 8222157 Caucasian Female [0-10) ?  
## 2 149190 55629189 Caucasian Female [10-20) ?  
## 3 64410 86047875 AfricanAmerican Female [20-30) ?  
## 4 500364 82442376 Caucasian Male [30-40) ?  
## 5 16680 42519267 Caucasian Male [40-50) ?  
## 6 35754 82637451 Caucasian Male [50-60) ?  
## admission\_type\_id discharge\_disposition\_id admission\_source\_id  
## 1 6 25 1  
## 2 1 1 7  
## 3 1 1 7  
## 4 1 1 7  
## 5 1 1 7  
## 6 2 1 2  
## time\_in\_hospital payer\_code medical\_specialty num\_lab\_procedures  
## 1 1 ? Pediatrics-Endocrinology 41  
## 2 3 ? ? 59  
## 3 2 ? ? 11  
## 4 2 ? ? 44  
## 5 1 ? ? 51  
## 6 3 ? ? 31  
## num\_procedures num\_medications number\_outpatient number\_emergency  
## 1 0 1 0 0  
## 2 0 18 0 0  
## 3 5 13 2 0  
## 4 1 16 0 0  
## 5 0 8 0 0  
## 6 6 16 0 0  
## number\_inpatient diag\_1 diag\_2 diag\_3 number\_diagnoses max\_glu\_serum  
## 1 0 250.83 ? ? 1 None  
## 2 0 276 250.01 255 9 None  
## 3 1 648 250 V27 6 None  
## 4 0 8 250.43 403 7 None  
## 5 0 197 157 250 5 None  
## 6 0 414 411 250 9 None  
## A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride  
## 1 None No No No No No  
## 2 None No No No No No  
## 3 None No No No No No  
## 4 None No No No No No  
## 5 None No No No No No  
## 6 None No No No No No  
## acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone  
## 1 No No No No No No  
## 2 No No No No No No  
## 3 No Steady No No No No  
## 4 No No No No No No  
## 5 No Steady No No No No  
## 6 No No No No No No  
## acarbose miglitol troglitazone tolazamide examide citoglipton insulin  
## 1 No No No No No No No  
## 2 No No No No No No Up  
## 3 No No No No No No No  
## 4 No No No No No No Up  
## 5 No No No No No No Steady  
## 6 No No No No No No Steady  
## glyburide.metformin glipizide.metformin glimepiride.pioglitazone  
## 1 No No No  
## 2 No No No  
## 3 No No No  
## 4 No No No  
## 5 No No No  
## 6 No No No  
## metformin.rosiglitazone metformin.pioglitazone change diabetesMed  
## 1 No No No No  
## 2 No No Ch Yes  
## 3 No No No Yes  
## 4 No No Ch Yes  
## 5 No No Ch Yes  
## 6 No No No Yes  
## readmitted  
## 1 NO  
## 2 >30  
## 3 NO  
## 4 NO  
## 5 NO  
## 6 >30

## Convert to a data table for easier subsetting  
  
## Now convert to a data table so that the genders can be broken out into separate tables for analysis  
  
library(data.table)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following objects are masked from 'package:data.table':  
##   
## dcast, melt

library(dplyr)

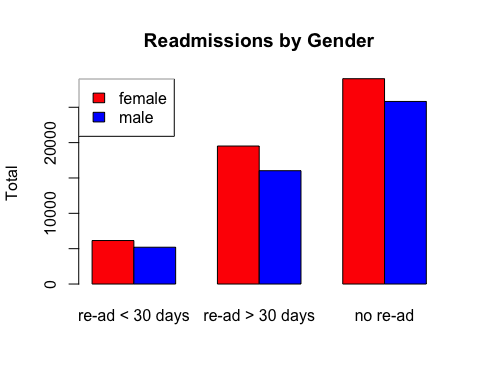
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

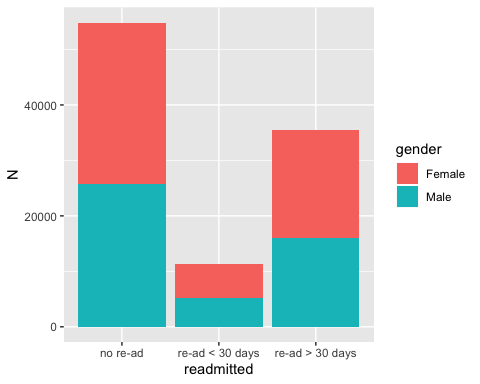
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

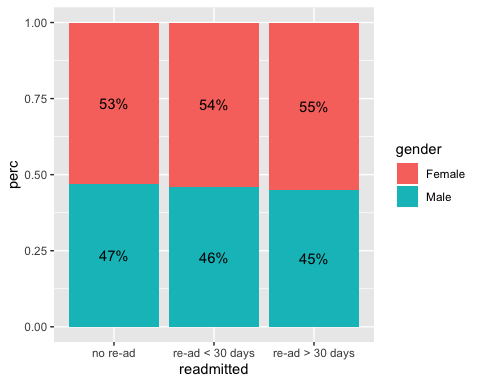
library(ggplot2)  
##diabetes\_data <- data.table(diabetes\_data)  
##setkey(diabetes\_data, gender)  
##males <- diabetes\_data["Male"]  
##females <- diabetes\_data["Female"]  
##setkey(diabetes\_data, readmitted)  
##more\_than\_30 <- diabetes\_data[">30"]  
##less\_than\_30 <- diabetes\_data["<30"]  
  
##nope <- diabetes\_data["NO"]  
##setkey(nope, gender)  
##male\_nope <- nope["Male"]  
##female\_nope <- nope["Female"]  
##setkey(more\_than\_30, gender)  
##male\_more\_than\_30 <- more\_than\_30["Male"]  
##female\_more\_than\_30 <- more\_than\_30["Female"]  
##setkey(less\_than\_30, gender)  
##male\_less\_than\_30 <- less\_than\_30["Male"]  
##female\_less\_than\_30 <- less\_than\_30["Female"]  
  
## For plotting  
gender\_readmit <- diabetes\_data[, c(4,50)]  
  
try <- table(gender\_readmit)  
gender\_readmit\_table <- try[1:2,]  
colnames(gender\_readmit\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
colors <- c("red", "blue")  
barplot(gender\_readmit\_table, main = "Readmissions by Gender", ylab = "Total", beside = TRUE, col = colors)  
legend("topleft", c("female", "male"), fill = colors)



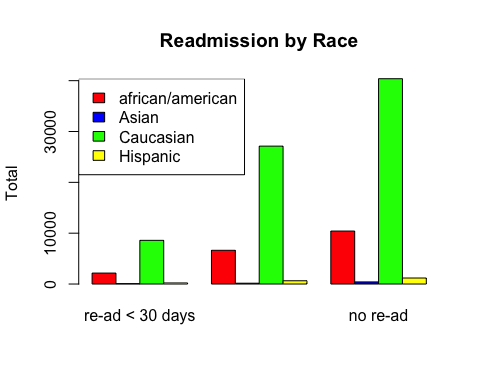
try1 <- data.table(gender\_readmit\_table)  
  
ggplot(try1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = gender, cumulative = TRUE)) + geom\_col()



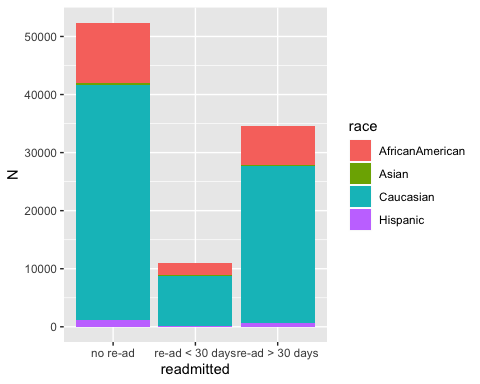
ggplot(try1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = gender, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



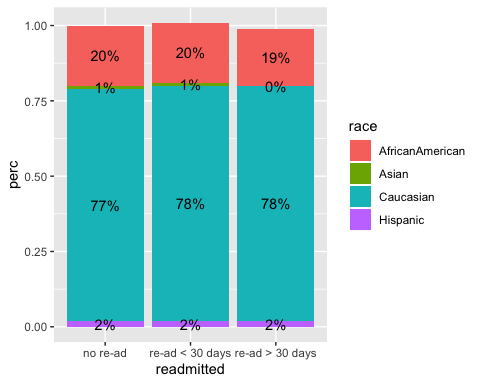
race\_readmit <- diabetes\_data[, c(3,50)]  
  
race\_readmit\_table <- table(race\_readmit)  
race\_readmit\_table <- race\_readmit\_table[2:5,]  
colnames(race\_readmit\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
rcolors <- c("red", "blue", "green", "yellow")  
barplot(race\_readmit\_table, main = "Readmission by Race", ylab = "Total", beside = TRUE, col = rcolors)  
legend("topleft", legend = c("african/american", "Asian", "Caucasian", "Hispanic"), fill = rcolors)



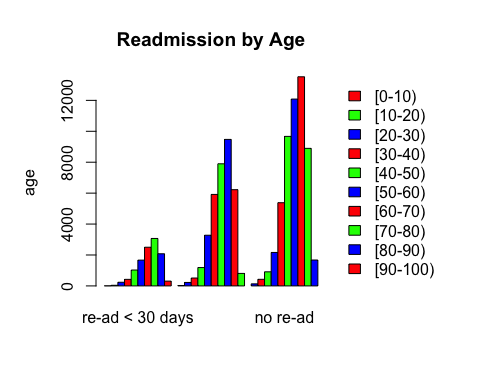
race\_readmit\_table\_1 <- data.table(race\_readmit\_table)  
  
ggplot(race\_readmit\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = race, cumulative = TRUE)) + geom\_col()



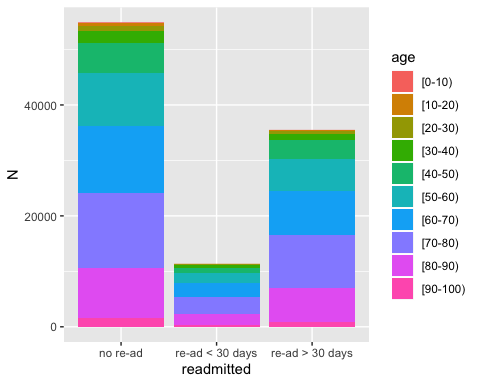
ggplot(race\_readmit\_table\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = race, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



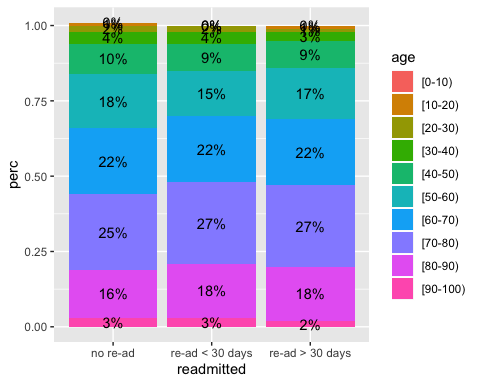
age\_readmit <- diabetes\_data[, c(5,50)]  
age\_readmit\_table <- table(age\_readmit)  
colnames(age\_readmit\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
par(mfrow=c(1, 1), mar=c(5, 5, 4, 8))  
barplot(age\_readmit\_table, main = "Readmission by Age", ylab="age", legend.text = TRUE, beside = TRUE, col = rainbow(3),args.legend = list(x = "topright", bty = "n", inset=c(-0.5, 0)))



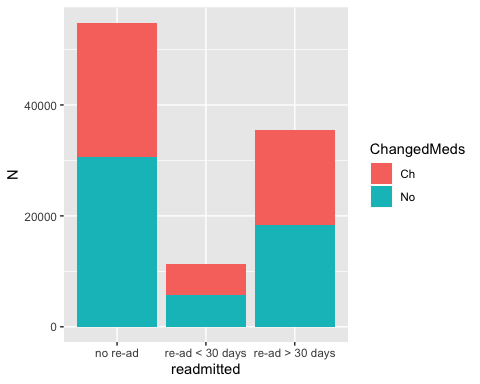
age\_readmit\_table\_1 <- data.table(age\_readmit\_table)  
ggplot(age\_readmit\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = age, cumulative = TRUE)) + geom\_col()



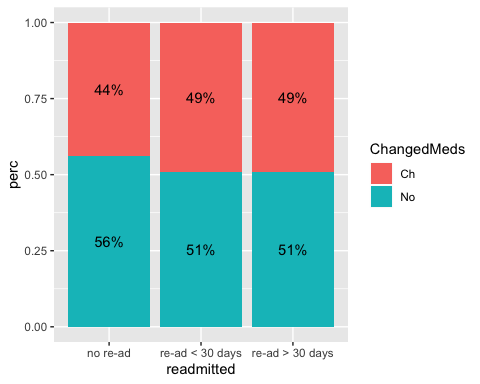
ggplot(age\_readmit\_table\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = age, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



changes <- table(diabetes\_data$change, diabetes\_data$readmitted)  
colnames(changes) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
changes\_1 <- data.table(changes)  
colnames(changes\_1) <- c("ChangedMeds", "readmitted", "N")  
ggplot(changes\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = ChangedMeds, cumulative = TRUE)) + geom\_col()



ggplot(changes\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = ChangedMeds, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



## Or another way  
## For all data, Chi-Square of readamission versus gender  
  
gender\_test <- table(diabetes\_data$gender, diabetes\_data$readmitted)  
  
gender\_test <- gender\_test[1:2,]  
  
all\_data\_gender <- chisq.test(gender\_test)  
  
print(all\_data\_gender)

##   
## Pearson's Chi-squared test  
##   
## data: gender\_test  
## X-squared = 34.896, df = 2, p-value = 2.645e-08

changed\_meds\_test <- chisq.test(changes)  
  
print(changed\_meds\_test)

##   
## Pearson's Chi-squared test  
##   
## data: changes  
## X-squared = 215.83, df = 2, p-value < 2.2e-16

gender\_changes <- table(diabetes\_data$gender, diabetes\_data$change)  
  
print(changed\_meds\_test)

##   
## Pearson's Chi-squared test  
##   
## data: changes  
## X-squared = 215.83, df = 2, p-value < 2.2e-16

print(chisq.test(gender\_changes))

## Warning in chisq.test(gender\_changes): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: gender\_changes  
## X-squared = 23.007, df = 2, p-value = 1.009e-05

## now, summarize the data  
  
weight\_all = summary(diabetes\_data$weight)  
print(weight\_all)

## ? [0-25) [100-125) [125-150) [150-175) [175-200) [25-50)   
## 98569 48 625 145 35 11 97   
## [50-75) [75-100) >200   
## 897 1336 3

print(chisq.test(weight\_all))

##   
## Chi-squared test for given probabilities  
##   
## data: weight\_all  
## X-squared = 853255, df = 9, p-value < 2.2e-16

yup <- table(diabetes\_data$patient\_nbr)  
  
print(quantile(yup))

## 0% 25% 50% 75% 100%   
## 1 1 1 1 40

## For all data, Chi-Square for race versus admission data  
  
  
race\_test <- table(diabetes\_data$race, diabetes\_data$readmitted)  
  
race\_test <- race\_test[1:5,]  
  
all\_data\_race <- chisq.test(race\_test)  
  
print(all\_data\_race)

##   
## Pearson's Chi-squared test  
##   
## data: race\_test  
## X-squared = 253.65, df = 8, p-value < 2.2e-16

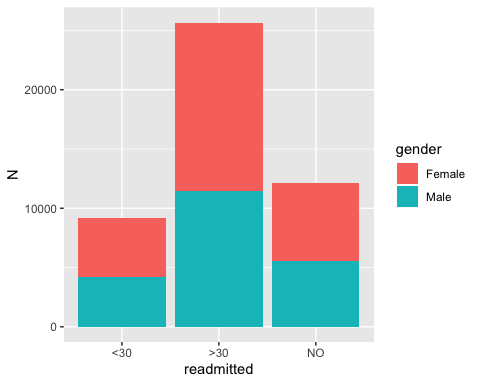
## For all data, Chi-square for age versus readmission data  
  
  
age\_test <- table(diabetes\_data$age, diabetes\_data$readmitted)  
  
age\_test <- age\_test[2:10,]  
  
all\_data\_age <- chisq.test(age\_test)  
  
print(all\_data\_age)

##   
## Pearson's Chi-squared test  
##   
## data: age\_test  
## X-squared = 260.71, df = 16, p-value < 2.2e-16

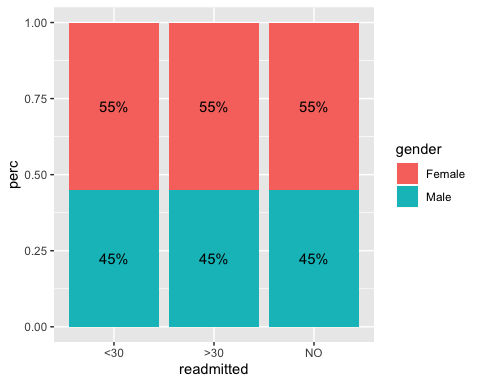
## Evaluate only patients with multiple encounter IDs  
  
  
Dupe\_diabetes\_data <- diabetes\_data[diabetes\_data$patient\_nbr %in% names(which(table(diabetes\_data$patient\_nbr)>1)),]  
  
## Duplicate patient ID, chi-square of gender versus readmission  
  
  
dupes\_gender\_table <- table(Dupe\_diabetes\_data$gender, Dupe\_diabetes\_data$readmitted)  
  
dupes\_gender\_table <- dupes\_gender\_table[1:2,]  
  
dupes\_gender\_test <- chisq.test(dupes\_gender\_table)  
  
print(dupes\_gender\_test)

##   
## Pearson's Chi-squared test  
##   
## data: dupes\_gender\_table  
## X-squared = 3.4571, df = 2, p-value = 0.1775

dupes\_gender\_test\_1 <- data.table(dupes\_gender\_table)  
  
colnames(dupes\_gender\_test\_1) <- c("gender", "readmitted", "N")  
ggplot(dupes\_gender\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = gender, cumulative = TRUE)) + geom\_col()



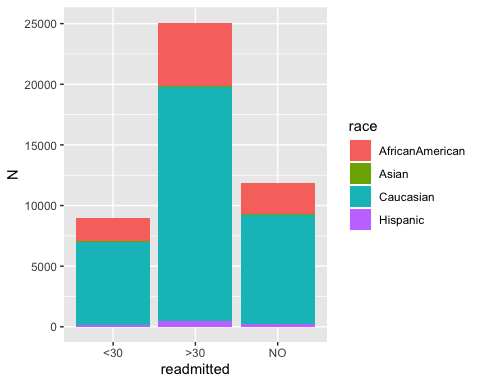
ggplot(dupes\_gender\_test\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = gender, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



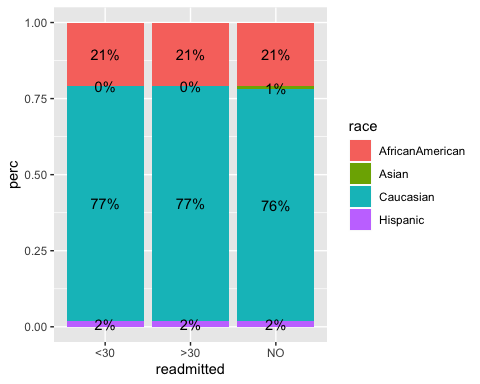
## Duplicate patient ID, chi-square of race versus readmission  
  
dupes\_race\_table <- table(Dupe\_diabetes\_data$race, Dupe\_diabetes\_data$readmitted)  
  
## Remove the unknown races  
  
dupes\_race\_table <- dupes\_race\_table[2:5,]  
  
dupes\_race\_test <- chisq.test(dupes\_race\_table)  
  
print(dupes\_race\_test)

##   
## Pearson's Chi-squared test  
##   
## data: dupes\_race\_table  
## X-squared = 7.1093, df = 6, p-value = 0.3109

dupes\_race\_test\_1 <- data.table(dupes\_race\_table)  
  
colnames(dupes\_race\_test\_1) <- c("race", "readmitted", "N")  
ggplot(dupes\_race\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = race, cumulative = TRUE)) + geom\_col()



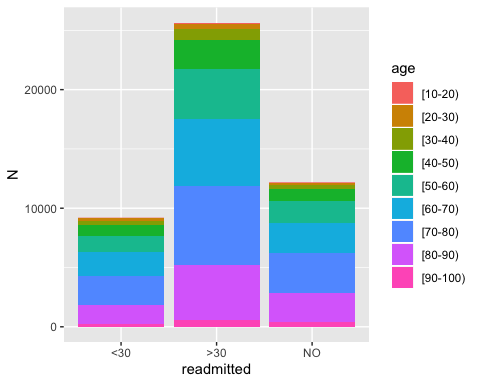
ggplot(dupes\_race\_test\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = race, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



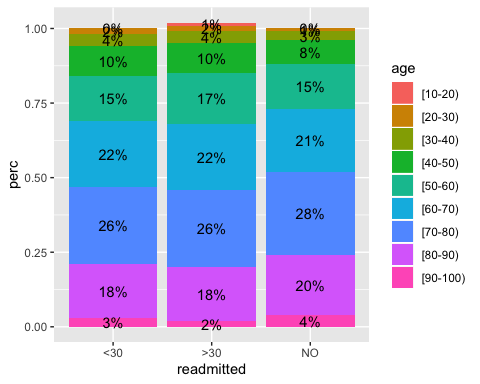
## Duplicate patient ID, chi-square of age versus readmission  
  
dupe\_age\_table <- table(Dupe\_diabetes\_data$age, Dupe\_diabetes\_data$readmitted)  
  
dupe\_age\_table <- dupe\_age\_table[2:10,]  
  
dupe\_age\_test <- chisq.test(dupe\_age\_table)  
  
print(dupe\_age\_test)

##   
## Pearson's Chi-squared test  
##   
## data: dupe\_age\_table  
## X-squared = 189.61, df = 16, p-value < 2.2e-16

dupe\_age\_test\_1 <- data.table(dupe\_age\_table)  
colnames(dupe\_age\_test\_1) <- c("age", "readmitted", "N")  
ggplot(dupe\_age\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = age, cumulative = TRUE)) + geom\_col()



ggplot(dupe\_age\_test\_1 %>% group\_by(readmitted) %>% mutate(perc = round(N/sum(N),2)), aes(x = readmitted, y = perc, fill = age, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.5))



## Patients numbers only once (may have been readmitted)  
  
No\_duplicated\_data <- diabetes\_data[diabetes\_data$patient\_nbr %in% names(which(table(diabetes\_data$patient\_nbr)==1)),]  
  
print(head(No\_duplicated\_data))

## encounter\_id patient\_nbr race gender age weight  
## 1 2278392 8222157 Caucasian Female [0-10) ?  
## 2 149190 55629189 Caucasian Female [10-20) ?  
## 3 64410 86047875 AfricanAmerican Female [20-30) ?  
## 4 500364 82442376 Caucasian Male [30-40) ?  
## 5 16680 42519267 Caucasian Male [40-50) ?  
## 6 35754 82637451 Caucasian Male [50-60) ?  
## admission\_type\_id discharge\_disposition\_id admission\_source\_id  
## 1 6 25 1  
## 2 1 1 7  
## 3 1 1 7  
## 4 1 1 7  
## 5 1 1 7  
## 6 2 1 2  
## time\_in\_hospital payer\_code medical\_specialty num\_lab\_procedures  
## 1 1 ? Pediatrics-Endocrinology 41  
## 2 3 ? ? 59  
## 3 2 ? ? 11  
## 4 2 ? ? 44  
## 5 1 ? ? 51  
## 6 3 ? ? 31  
## num\_procedures num\_medications number\_outpatient number\_emergency  
## 1 0 1 0 0  
## 2 0 18 0 0  
## 3 5 13 2 0  
## 4 1 16 0 0  
## 5 0 8 0 0  
## 6 6 16 0 0  
## number\_inpatient diag\_1 diag\_2 diag\_3 number\_diagnoses max\_glu\_serum  
## 1 0 250.83 ? ? 1 None  
## 2 0 276 250.01 255 9 None  
## 3 1 648 250 V27 6 None  
## 4 0 8 250.43 403 7 None  
## 5 0 197 157 250 5 None  
## 6 0 414 411 250 9 None  
## A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride  
## 1 None No No No No No  
## 2 None No No No No No  
## 3 None No No No No No  
## 4 None No No No No No  
## 5 None No No No No No  
## 6 None No No No No No  
## acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone  
## 1 No No No No No No  
## 2 No No No No No No  
## 3 No Steady No No No No  
## 4 No No No No No No  
## 5 No Steady No No No No  
## 6 No No No No No No  
## acarbose miglitol troglitazone tolazamide examide citoglipton insulin  
## 1 No No No No No No No  
## 2 No No No No No No Up  
## 3 No No No No No No No  
## 4 No No No No No No Up  
## 5 No No No No No No Steady  
## 6 No No No No No No Steady  
## glyburide.metformin glipizide.metformin glimepiride.pioglitazone  
## 1 No No No  
## 2 No No No  
## 3 No No No  
## 4 No No No  
## 5 No No No  
## 6 No No No  
## metformin.rosiglitazone metformin.pioglitazone change diabetesMed  
## 1 No No No No  
## 2 No No Ch Yes  
## 3 No No No Yes  
## 4 No No Ch Yes  
## 5 No No Ch Yes  
## 6 No No No Yes  
## readmitted  
## 1 NO  
## 2 >30  
## 3 NO  
## 4 NO  
## 5 NO  
## 6 >30

gender\_nodupes\_table <- table(No\_duplicated\_data$gender, No\_duplicated\_data$readmitted)  
  
gender\_nodupes\_table <- gender\_nodupes\_table[1:2,]  
  
gender\_nodupes\_test <- chisq.test(gender\_nodupes\_table)  
  
print(gender\_nodupes\_test)

##   
## Pearson's Chi-squared test  
##   
## data: gender\_nodupes\_table  
## X-squared = 4.0943, df = 2, p-value = 0.1291

## Now the race data  
  
race\_nodupes\_table <- table(No\_duplicated\_data$race, No\_duplicated\_data$readmitted)  
  
race\_nodupes\_table <- race\_nodupes\_table[2:5,]  
  
race\_nodupes\_test <- chisq.test(race\_nodupes\_table)  
  
print(race\_nodupes\_test)

##   
## Pearson's Chi-squared test  
##   
## data: race\_nodupes\_table  
## X-squared = 171.53, df = 6, p-value < 2.2e-16

## Now for the age data  
  
age\_nodupes\_table <- table(No\_duplicated\_data$age, No\_duplicated\_data$readmitted)  
  
age\_nodupes\_table <- age\_nodupes\_table[2:10,]  
  
age\_nodupes\_test <- chisq.test(age\_nodupes\_table)  
  
print(age\_nodupes\_test)

##   
## Pearson's Chi-squared test  
##   
## data: age\_nodupes\_table  
## X-squared = 365.76, df = 16, p-value < 2.2e-16

## Some plots  
  
  
  
## remove ? for race