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

## Loading required package: grid

library(ggpubr)

## Loading required package: magrittr

##print(summary(diabetes\_data))

Now for some tablature ## For plotting

## Some stats about the data

table(diabetes\_data$gender)

##   
## Female Male Unknown/Invalid   
## 54708 47055 3

table(diabetes\_data$race)

##   
## ? AfricanAmerican Asian Caucasian   
## 2273 19210 641 76099   
## Hispanic Other   
## 2037 1506

table(diabetes\_data$weight)

##   
## ? [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

table(diabetes\_data$time\_in\_hospital)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 14208 17224 17756 13924 9966 7539 5859 4391 3002 2342 1855 1448   
## 13 14   
## 1210 1042

table(diabetes\_data$num\_lab\_procedures)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 3208 1101 668 378 286 282 323 366 933 838 689 506 404 336 461   
## 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30   
## 544 743 730 977 763 676 694 836 744 1048 1094 831 1048 1340 1354   
## 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45   
## 1311 1336 1297 1677 1907 1962 2079 2213 2101 2201 2117 2113 2804 2496 2376   
## 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60   
## 2189 2106 2058 2066 1924 1925 1838 1802 1888 1836 1839 1747 1708 1624 1610   
## 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75   
## 1638 1444 1450 1324 1238 1212 1093 1043 991 826 798 749 661 612 526   
## 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90   
## 465 421 357 334 293 283 225 230 160 156 128 91 102 73 65   
## 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105   
## 61 48 56 45 46 28 31 26 9 13 13 8 6 3 6   
## 106 107 108 109 111 113 114 118 120 121 126 129 132   
## 5 1 4 4 3 3 2 1 1 1 1 1 1

table(diabetes\_data$num\_procedures)

##   
## 0 1 2 3 4 5 6   
## 46652 20742 12717 9443 4180 3078 4954

table(diabetes\_data$readmitted)

##   
## <30 >30 NO   
## 11357 35545 54864

table(diabetes\_data$diabetesMed)

##   
## No Yes   
## 23403 78363

table(diabetes\_data$change)

##   
## Ch No   
## 47011 54755

table(diabetes\_data$insulin)

##   
## Down No Steady Up   
## 12218 47383 30849 11316

table(diabetes\_data$insulin)

##   
## Down No Steady Up   
## 12218 47383 30849 11316

table(diabetes\_data$examide)

##   
## No   
## 101766

table(diabetes\_data$num\_medications)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 262 470 900 1417 2017 2699 3484 4353 4913 5346 5795 6004 6086 5707 5792   
## 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30   
## 5430 4919 4523 4078 3691 3230 2868 2426 2109 1888 1608 1432 1233 1000 849   
## 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45   
## 712 623 504 453 388 287 293 228 209 179 146 128 127 100 88   
## 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60   
## 92 74 60 61 55 43 54 40 33 32 37 26 25 20 23   
## 61 62 63 64 65 66 67 68 69 70 72 74 75 79 81   
## 14 15 14 8 12 5 7 7 5 2 3 1 2 1 1

table(diabetes\_data$number\_outpatient)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11   
## 85027 8547 3594 2042 1099 533 303 155 98 83 57 42   
## 12 13 14 15 16 17 18 19 20 21 22 23   
## 30 31 28 20 15 8 5 3 7 7 5 2   
## 24 25 26 27 28 29 33 34 35 36 37 38   
## 3 2 2 3 1 2 2 1 2 2 1 1   
## 39 40 42   
## 1 1 1

table(diabetes\_data$number\_emergency)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11   
## 90383 7677 2042 725 374 192 94 73 50 33 34 23   
## 12 13 14 15 16 18 19 20 21 22 24 25   
## 10 12 3 3 5 5 4 4 2 6 1 2   
## 28 29 37 42 46 54 63 64 76   
## 1 1 1 1 1 1 1 1 1

table(diabetes\_data$max\_glu\_serum)

##   
## >200 >300 None Norm   
## 1485 1264 96420 2597

table(diabetes\_data$A1Cresult)

##   
## >7 >8 None Norm   
## 3812 8216 84748 4990

table(diabetes\_data$number\_inpatient)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11   
## 67630 19521 7566 3411 1622 812 480 268 151 111 61 49   
## 12 13 14 15 16 17 18 19 21   
## 34 20 10 9 6 1 1 2 1

table(diabetes\_data$number\_diagnoses)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 219 1023 2835 5537 11393 10161 10393 10616 49474 17 11 9   
## 13 14 15 16   
## 16 7 10 45

table(diabetes\_data$age)

##   
## [0-10) [10-20) [20-30) [30-40) [40-50) [50-60) [60-70) [70-80)   
## 161 691 1657 3775 9685 17256 22483 26068   
## [80-90) [90-100)   
## 17197 2793

length(unique(diabetes\_data$patient\_nbr))

## [1] 71518

This section is where we will check out the full data set

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")  
  
print(gender\_readmit\_table)

## readmitted  
## gender re-ad < 30 days re-ad > 30 days no re-ad  
## Female 6152 19518 29038  
## Male 5205 16027 25823

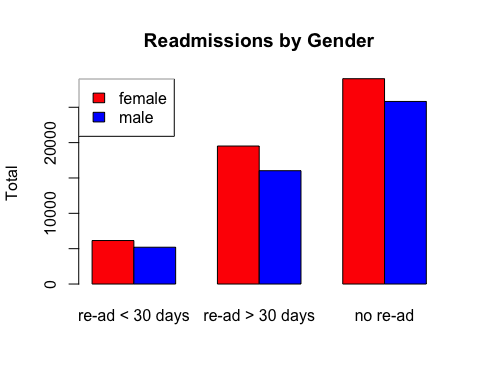
## CHI SQUARE GENDER FOR FULL TABLE  
  
  
chisq.test(gender\_readmit\_table)

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

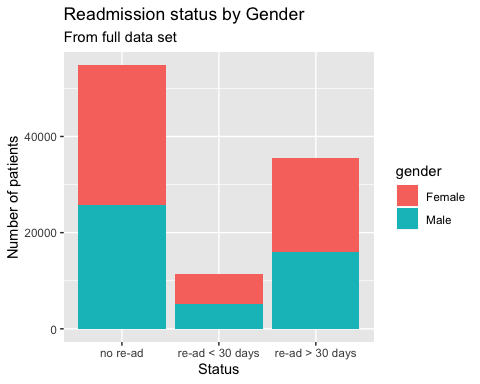
assocstats(gender\_readmit\_table)

## X^2 df P(> X^2)  
## Likelihood Ratio 34.908 2 2.6289e-08  
## Pearson 34.896 2 2.6447e-08  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.019   
## Cramer's V : 0.019

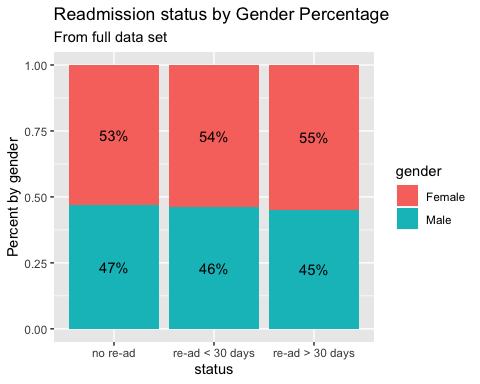
colors <- c("red", "blue")  
barplot(gender\_readmit\_table, main = "Readmissions by Gender", ylab = "Total", beside = TRUE, col = colors)  
legend("topleft", c("female", "male"), fill = colors)



try1 <- data.table(gender\_readmit\_table)  
  
p <- ggplot(try1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = gender, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Gender", subtitle = "From full data set")  
p1 <- p + ylab("Number of patients") + xlab("Status")  
  
p1



q <- 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))  
q <- q + labs(title = "Readmission status by Gender Percentage", subtitle = "From full data set")  
q1 <- q + ylab("Percent by gender") + xlab("status")  
q1



##ggarrange(p, q, ncol = 2, nrow = 1)  
  
##ggarrange(p, q, ncol = 1, row = 2)  
  
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")  
  
print(race\_readmit\_table)

## readmitted  
## race re-ad < 30 days re-ad > 30 days no re-ad  
## AfricanAmerican 2155 6634 10421  
## Asian 65 161 415  
## Caucasian 8592 27124 40383  
## Hispanic 212 642 1183

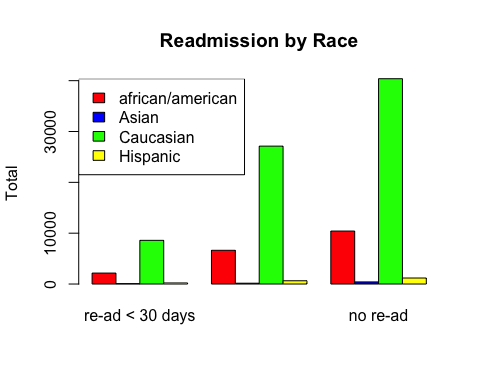
## CHI SQUARE FOR RACE FULL DATA  
  
chisq.test(race\_readmit\_table)

##   
## Pearson's Chi-squared test  
##   
## data: race\_readmit\_table  
## X-squared = 62.736, df = 6, p-value = 1.249e-11

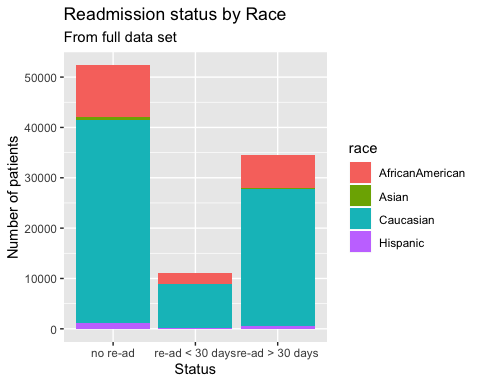
assocstats(race\_readmit\_table)

## X^2 df P(> X^2)  
## Likelihood Ratio 64.029 6 6.8092e-12  
## Pearson 62.736 6 1.2493e-11  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.025   
## Cramer's V : 0.018

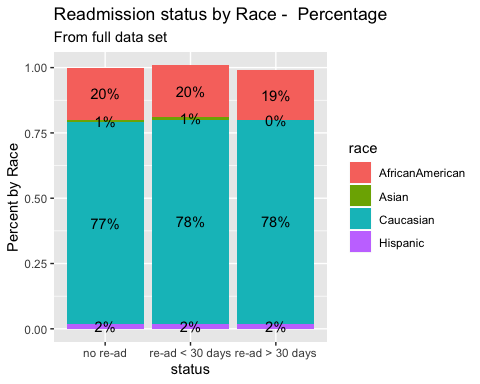
barplot(race\_readmit\_table, main = "Readmission by Race", ylab = "Total", beside = TRUE, col = rcolors)  
legend("topleft", legend = c("african/american", "Asian", "Caucasian", "Hispanic"), fill = rcolors)



race\_readmit\_table\_1 <- data.table(race\_readmit\_table)  
  
p <- ggplot(race\_readmit\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = race, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Race", subtitle = "From full data set")  
p2 <- p + ylab("Number of patients") + xlab("Status")  
p2



q <- 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))  
q <- q + labs(title = "Readmission status by Race - Percentage", subtitle = "From full data set")  
q2 <- q + ylab("Percent by Race") + xlab("status")  
q2



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")  
  
print(age\_readmit\_table)

## readmitted  
## age re-ad < 30 days re-ad > 30 days no re-ad  
## [0-10) 3 26 132  
## [10-20) 40 224 427  
## [20-30) 236 510 911  
## [30-40) 424 1187 2164  
## [40-50) 1027 3278 5380  
## [50-60) 1668 5917 9671  
## [60-70) 2502 7897 12084  
## [70-80) 3069 9475 13524  
## [80-90) 2078 6223 8896  
## [90-100) 310 808 1675

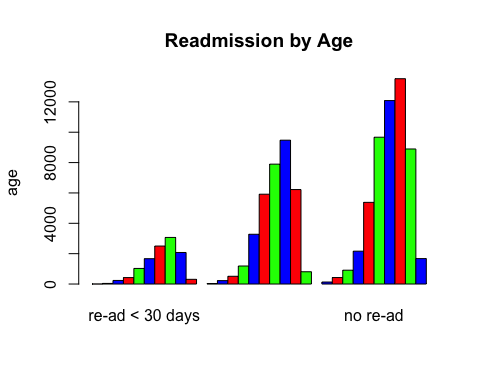
chisq.test(age\_readmit\_table)

##   
## Pearson's Chi-squared test  
##   
## data: age\_readmit\_table  
## X-squared = 313.17, df = 18, p-value < 2.2e-16

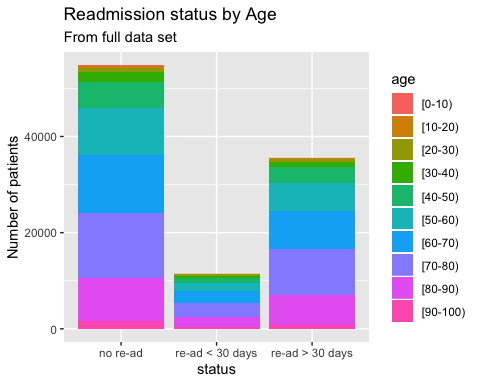
assocstats(age\_readmit\_table)

## X^2 df P(> X^2)  
## Likelihood Ratio 326.04 18 0  
## Pearson 313.17 18 0  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.055   
## Cramer's V : 0.039

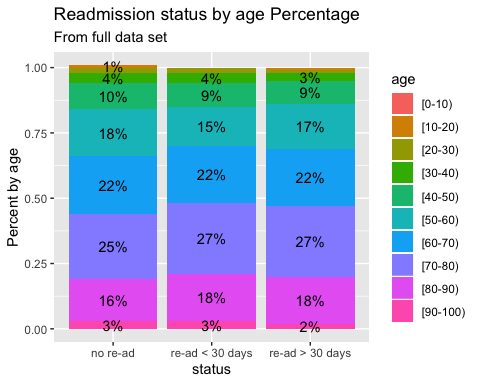
barplot(age\_readmit\_table, main = "Readmission by Age", ylab="age", legend.text = TRUE, beside = TRUE, col = rainbow(3),args.legend = list(x = "topright", bty = "n", inset=c(-0.5, 0)))



age\_readmit\_table\_1 <- data.table(age\_readmit\_table)  
p <- ggplot(age\_readmit\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = age, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Age", subtitle = "From full data set")  
p3 <- p + ylab("Number of patients") + xlab("status")  
p3



q <- 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), check\_overlap = TRUE)  
q <- q + labs(title = "Readmission status by age Percentage", subtitle = "From full data set")  
q3 <- q + ylab("Percent by age") + xlab("status")  
q3



## ggarrange(p1, q1, p2, q2, p3, q3, heights = c(2,2,2,2,2,2), nrow = 3, ncol = 2)

Test Changes

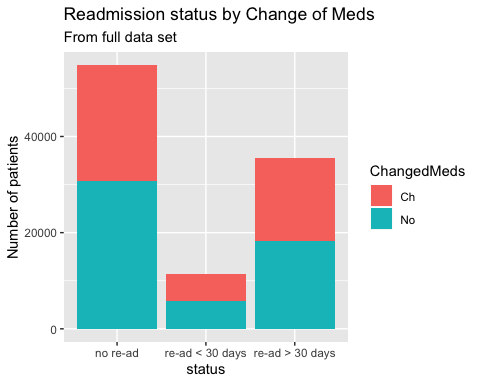
changes <- table(diabetes\_data$change, diabetes\_data$readmitted)  
colnames(changes) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
chisq.test(changes)

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

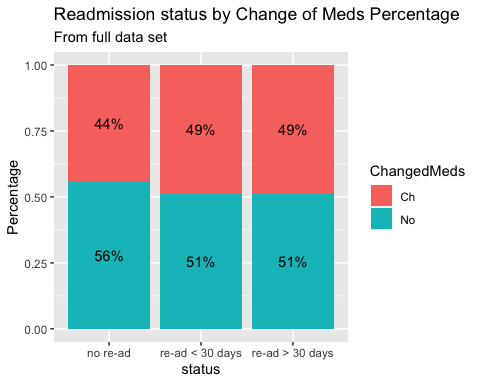
assocstats(changes)

## X^2 df P(> X^2)  
## Likelihood Ratio 215.82 2 0  
## Pearson 215.83 2 0  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.046   
## Cramer's V : 0.046

changes\_1 <- data.table(changes)  
colnames(changes\_1) <- c("ChangedMeds", "readmitted", "N")  
p <- ggplot(changes\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = ChangedMeds, cumulative = TRUE)) + geom\_col()  
  
p <- p + labs(title = "Readmission status by Change of Meds", subtitle = "From full data set")  
p <- p + ylab("Number of patients") + xlab("status")  
p



q <- 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),,check\_overlap = TRUE)  
q <- q + labs(title = "Readmission status by Change of Meds Percentage", subtitle = "From full data set")  
q <- q + ylab("Percentage") + xlab("status")  
q



changy <- diabetes\_data[diabetes\_data$change == "Ch",]  
table

## function (..., exclude = if (useNA == "no") c(NA, NaN), useNA = c("no",   
## "ifany", "always"), dnn = list.names(...), deparse.level = 1)   
## {  
## list.names <- function(...) {  
## l <- as.list(substitute(list(...)))[-1L]  
## nm <- names(l)  
## fixup <- if (is.null(nm))   
## seq\_along(l)  
## else nm == ""  
## dep <- vapply(l[fixup], function(x) switch(deparse.level +   
## 1, "", if (is.symbol(x)) as.character(x) else "",   
## deparse(x, nlines = 1)[1L]), "")  
## if (is.null(nm))   
## dep  
## else {  
## nm[fixup] <- dep  
## nm  
## }  
## }  
## miss.use <- missing(useNA)  
## miss.exc <- missing(exclude)  
## useNA <- if (miss.use && !miss.exc && !match(NA, exclude,   
## nomatch = 0L))   
## "ifany"  
## else match.arg(useNA)  
## doNA <- useNA != "no"  
## if (!miss.use && !miss.exc && doNA && match(NA, exclude,   
## nomatch = 0L))   
## warning("'exclude' containing NA and 'useNA' != \"no\"' are a bit contradicting")  
## args <- list(...)  
## if (!length(args))   
## stop("nothing to tabulate")  
## if (length(args) == 1L && is.list(args[[1L]])) {  
## args <- args[[1L]]  
## if (length(dnn) != length(args))   
## dnn <- if (!is.null(argn <- names(args)))   
## argn  
## else paste(dnn[1L], seq\_along(args), sep = ".")  
## }  
## bin <- 0L  
## lens <- NULL  
## dims <- integer()  
## pd <- 1L  
## dn <- NULL  
## for (a in args) {  
## if (is.null(lens))   
## lens <- length(a)  
## else if (length(a) != lens)   
## stop("all arguments must have the same length")  
## fact.a <- is.factor(a)  
## if (doNA)   
## aNA <- anyNA(a)  
## if (!fact.a) {  
## a0 <- a  
## a <- factor(a, exclude = exclude)  
## }  
## add.na <- doNA  
## if (add.na) {  
## ifany <- (useNA == "ifany")  
## anNAc <- anyNA(a)  
## add.na <- if (!ifany || anNAc) {  
## ll <- levels(a)  
## if (add.ll <- !anyNA(ll)) {  
## ll <- c(ll, NA)  
## TRUE  
## }  
## else if (!ifany && !anNAc)   
## FALSE  
## else TRUE  
## }  
## else FALSE  
## }  
## if (add.na)   
## a <- factor(a, levels = ll, exclude = NULL)  
## else ll <- levels(a)  
## a <- as.integer(a)  
## if (fact.a && !miss.exc) {  
## ll <- ll[keep <- which(match(ll, exclude, nomatch = 0L) ==   
## 0L)]  
## a <- match(a, keep)  
## }  
## else if (!fact.a && add.na) {  
## if (ifany && !aNA && add.ll) {  
## ll <- ll[!is.na(ll)]  
## is.na(a) <- match(a0, c(exclude, NA), nomatch = 0L) >   
## 0L  
## }  
## else {  
## is.na(a) <- match(a0, exclude, nomatch = 0L) >   
## 0L  
## }  
## }  
## nl <- length(ll)  
## dims <- c(dims, nl)  
## if (prod(dims) > .Machine$integer.max)   
## stop("attempt to make a table with >= 2^31 elements")  
## dn <- c(dn, list(ll))  
## bin <- bin + pd \* (a - 1L)  
## pd <- pd \* nl  
## }  
## names(dn) <- dnn  
## bin <- bin[!is.na(bin)]  
## if (length(bin))   
## bin <- bin + 1L  
## y <- array(tabulate(bin, pd), dims, dimnames = dn)  
## class(y) <- "table"  
## y  
## }  
## <bytecode: 0x7fb7c006b4e0>  
## <environment: namespace:base>

## Or another way

## For all data, Chi-Square of readamission versus gender

## 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,]  
colnames(dupes\_gender\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
print(dupes\_gender\_table)

##   
## re-ad < 30 days re-ad > 30 days no re-ad  
## Female 5027 14229 6630  
## Male 4164 11438 5533

dupes\_gender\_test <- assocstats(dupes\_gender\_table)  
  
chisq.test(dupes\_gender\_table)

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

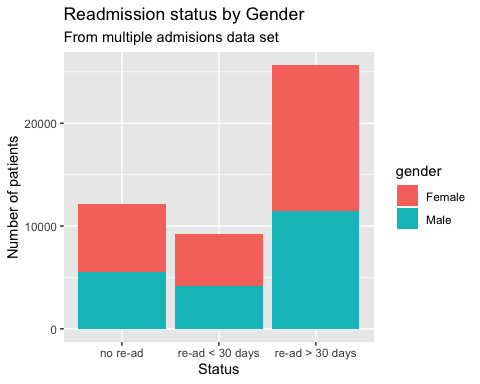
assocstats(dupes\_gender\_table)

## X^2 df P(> X^2)  
## Likelihood Ratio 3.4567 2 0.17758  
## Pearson 3.4571 2 0.17754  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.009   
## Cramer's V : 0.009

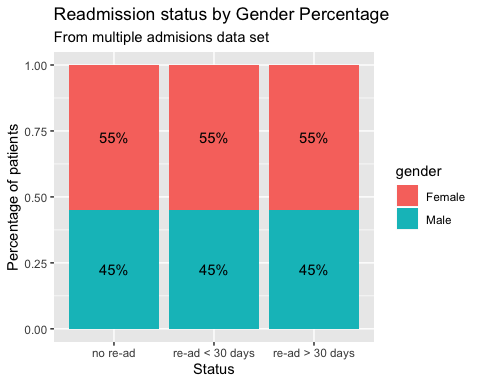
print(dupes\_gender\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 3.4567 2 0.17758  
## Pearson 3.4571 2 0.17754  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.009   
## Cramer's V : 0.009

dupes\_gender\_test\_1 <- data.table(dupes\_gender\_table)  
  
colnames(dupes\_gender\_test\_1) <- c("gender", "readmitted", "N")  
p <- ggplot(dupes\_gender\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = gender, cumulative = TRUE)) + geom\_col()  
  
p <- p + labs(title = "Readmission status by Gender", subtitle = "From multiple admisions data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
p



p <- 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))  
  
p <- p + labs(title = "Readmission status by Gender Percentage", subtitle = "From multiple admisions data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
p



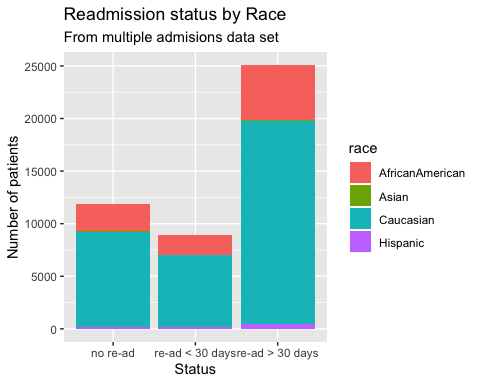
## Duplicate patient ID, chi-square of race versus readmission  
  
dupes\_race\_table <- table(Dupe\_diabetes\_data$race, Dupe\_diabetes\_data$readmitted)  
  
  
## Remove the unknown races  
  
dupes\_race\_table <- dupes\_race\_table[2:5,]  
colnames(dupes\_race\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
print(dupes\_race\_table)

##   
## re-ad < 30 days re-ad > 30 days no re-ad  
## AfricanAmerican 1892 5227 2534  
## Asian 43 103 69  
## Caucasian 6839 19316 9031  
## Hispanic 162 444 209

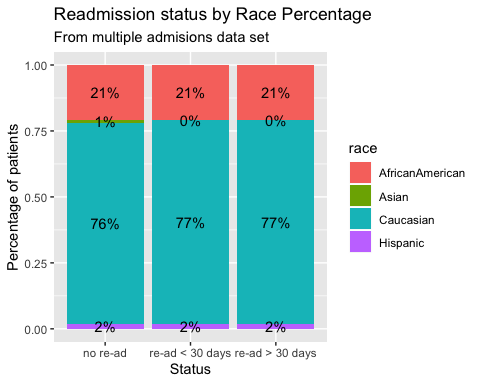
dupes\_race\_test <- assocstats(dupes\_race\_table)  
  
print(dupes\_race\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 6.9496 6 0.32553  
## Pearson 7.1093 6 0.31085  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.012   
## Cramer's V : 0.009

dupes\_race\_test\_1 <- data.table(dupes\_race\_table)  
  
colnames(dupes\_race\_test\_1) <- c("race", "readmitted", "N")  
  
p <- ggplot(dupes\_race\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = race, cumulative = TRUE)) + geom\_col()  
  
p <- p + labs(title = "Readmission status by Race", subtitle = "From multiple admisions data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
p



p <- 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))  
p <- p + labs(title = "Readmission status by Race Percentage", subtitle = "From multiple admisions data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
p



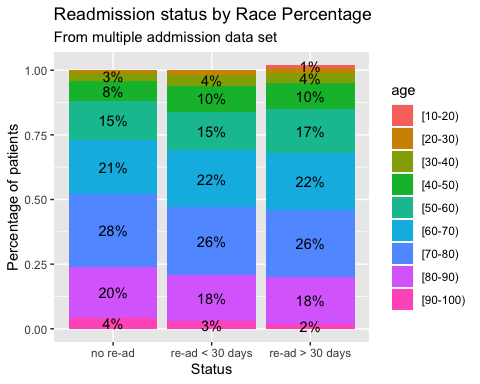
## Duplicate patient ID, chi-square of age versus readmission  
  
dupe\_age\_table <- table(Dupe\_diabetes\_data$age, Dupe\_diabetes\_data$readmitted)  
  
  
print(dupe\_age\_table)

##   
## <30 >30 NO  
## [0-10) 2 8 5  
## [10-20) 29 157 60  
## [20-30) 209 402 147  
## [30-40) 374 911 348  
## [40-50) 896 2448 973  
## [50-60) 1396 4259 1855  
## [60-70) 1999 5609 2562  
## [70-80) 2412 6691 3351  
## [80-90) 1617 4581 2428  
## [90-100) 257 601 434

dupe\_age\_table <- dupe\_age\_table[2:10,]  
  
colnames(dupe\_age\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
dupe\_age\_test <- assocstats(dupe\_age\_table)  
  
print(dupe\_age\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 187.89 16 0  
## Pearson 189.61 16 0  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.063   
## Cramer's V : 0.045

dupe\_age\_test\_1 <- data.table(dupe\_age\_table)  
colnames(dupe\_age\_test\_1) <- c("age", "readmitted", "N")  
p <- ggplot(dupe\_age\_test\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = age, cumulative = TRUE)) + geom\_col()  
  
p <- p + labs(title = "Readmission status by Age", subtitle = "From multiple addmission data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
  
p <- 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),,check\_overlap = TRUE)  
  
p <- p + labs(title = "Readmission status by Race Percentage", subtitle = "From multiple addmission data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
  
p

 ## Patients numbers only once (may have been readmitted)

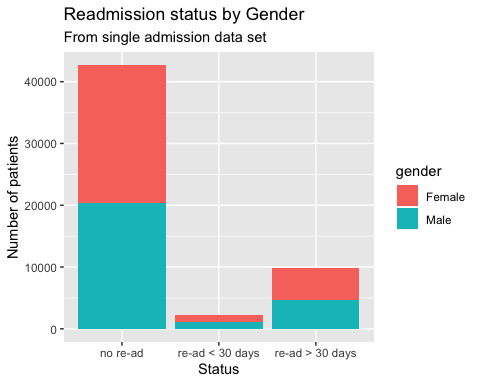
No\_duplicated\_data <- diabetes\_data[diabetes\_data$patient\_nbr %in% names(which(table(diabetes\_data$patient\_nbr)==1)),]  
  
##print(head(No\_duplicated\_data))  
  
gender\_nodupes\_table <- table(No\_duplicated\_data$gender, No\_duplicated\_data$readmitted)  
  
colnames(gender\_nodupes\_table) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
  
gender\_nodupes\_table <- gender\_nodupes\_table[1:2,]  
  
  
print(gender\_nodupes\_table)

##   
## re-ad < 30 days re-ad > 30 days no re-ad  
## Female 1125 5289 22408  
## Male 1041 4589 20290

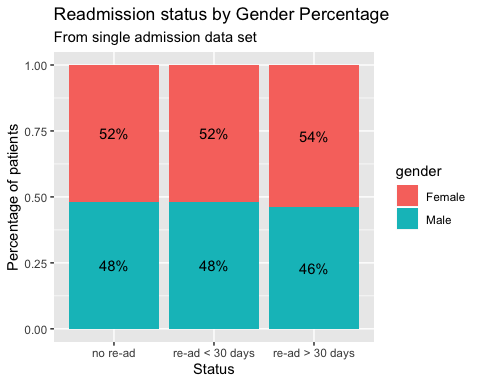
gender\_nodupes\_test <- assocstats(gender\_nodupes\_table)  
  
print(gender\_nodupes\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 4.0962 2 0.12898  
## Pearson 4.0943 2 0.12910  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.009   
## Cramer's V : 0.009

gender\_nodupes\_table\_1 <- data.table(gender\_nodupes\_table)  
colnames(gender\_nodupes\_table\_1) <- c("gender", "readmitted", "N")  
p <- ggplot(gender\_nodupes\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = gender, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Gender", subtitle = "From single admission data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
p



p <- ggplot(gender\_nodupes\_table\_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))  
p <- p + labs(title = "Readmission status by Gender Percentage", subtitle = "From single admission data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
  
p



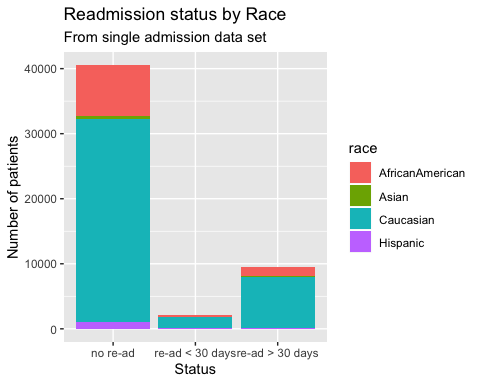
## Now the race data  
  
race\_nodupes\_table <- table(No\_duplicated\_data$race, No\_duplicated\_data$readmitted)  
  
colnames(race\_nodupes\_table ) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
  
race\_nodupes\_table <- race\_nodupes\_table[2:5,]  
  
print(race\_nodupes\_table)

##   
## re-ad < 30 days re-ad > 30 days no re-ad  
## AfricanAmerican 263 1407 7887  
## Asian 22 58 346  
## Caucasian 1753 7808 31352  
## Hispanic 50 198 974

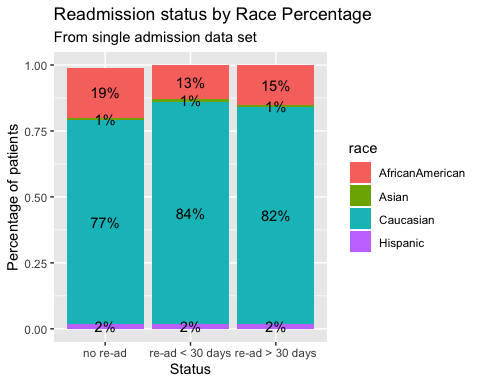
race\_nodupes\_test <- assocstats(race\_nodupes\_table)  
  
print(race\_nodupes\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 180.05 6 0  
## Pearson 171.53 6 0  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.057   
## Cramer's V : 0.041

race\_nodupes\_table\_1 <- data.table(race\_nodupes\_table)  
colnames(race\_nodupes\_table\_1) <- c("race", "readmitted", "N")  
p <- ggplot(race\_nodupes\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = race, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Race", subtitle = "From single admission data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
p



p <- ggplot(race\_nodupes\_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))  
p <- p + labs(title = "Readmission status by Race Percentage", subtitle = "From single admission data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
  
p



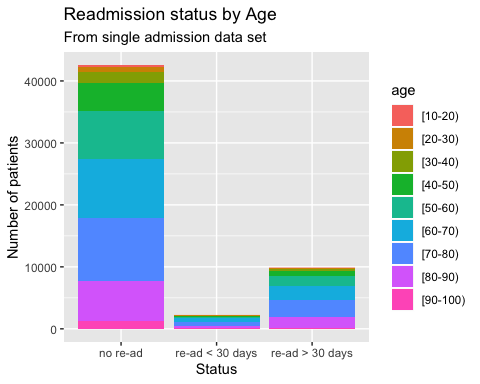
## Now for the age data  
  
age\_nodupes\_table <- table(No\_duplicated\_data$age, No\_duplicated\_data$readmitted)  
  
colnames(age\_nodupes\_table ) <- c("re-ad < 30 days", "re-ad > 30 days", "no re-ad")  
  
  
  
age\_nodupes\_table <- age\_nodupes\_table[2:10,]  
  
print(age\_nodupes\_table)

##   
## re-ad < 30 days re-ad > 30 days no re-ad  
## [10-20) 11 67 367  
## [20-30) 27 108 764  
## [30-40) 50 276 1816  
## [40-50) 131 830 4407  
## [50-60) 272 1658 7816  
## [60-70) 503 2288 9522  
## [70-80) 657 2784 10173  
## [80-90) 461 1642 6468  
## [90-100) 53 207 1241

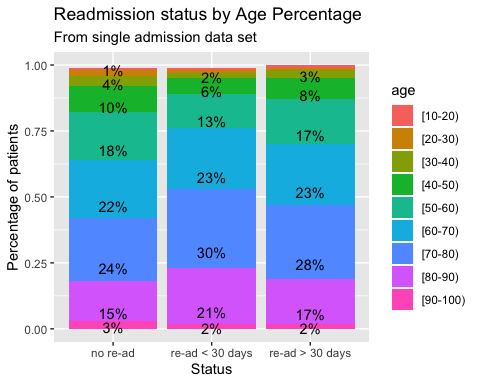
age\_nodupes\_test <- assocstats(age\_nodupes\_table)  
  
print(age\_nodupes\_test)

## X^2 df P(> X^2)  
## Likelihood Ratio 378.06 16 0  
## Pearson 365.76 16 0  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.082   
## Cramer's V : 0.058

age\_nodupes\_table\_1 <- data.table(age\_nodupes\_table)  
colnames(age\_nodupes\_table\_1) <- c("age", "readmitted", "N")  
p <- ggplot(age\_nodupes\_table\_1 %>% group\_by(readmitted) , aes(x = readmitted, y = N, fill = age, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Readmission status by Age", subtitle = "From single admission data set")  
p <- p + ylab("Number of patients") + xlab("Status")  
p



p <- ggplot(age\_nodupes\_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 = 0.2), check\_overlap = TRUE)  
p <- p + labs(title = "Readmission status by Age Percentage", subtitle = "From single admission data set")  
p <- p + ylab("Percentage of patients") + xlab("Status")  
  
p

 ## Other things to predict - number of procesdures

## remove ? for race  
  
## For only with diabetes in diagnoses 1-3  
  
dog <- diabetes\_data  
dog$diag\_1 <- as.numeric(dog$diag\_1)  
dog$diag\_2 <- as.numeric(dog$diag\_2)  
dog$diag\_3 <- as.numeric(dog$diag\_3)  
diabetes\_data$x <- dog$diag\_1 >= 249 & dog$diag\_1 < 251  
diabetes\_data$y <- dog$diag\_2 >= 249 & dog$diag\_2 < 251  
diabetes\_data$z <- dog$diag\_3 >= 249 & dog$diag\_3 < 251  
diabetes\_data$xyz <- diabetes\_data$x == "TRUE" | diabetes\_data$y == "TRUE" | diabetes\_data$z == "TRUE"  
  
coded <- diabetes\_data[diabetes\_data$xyz == "TRUE",]  
dim(coded)

## [1] 3862 54

gender\_code <- table(coded$gender, coded$readmitted)  
gender\_code <- gender\_code[1:2,]  
dim(gender\_code)

## [1] 2 3

print("The gender data is")

## [1] "The gender data is"

print(assocstats(gender\_code))

## X^2 df P(> X^2)  
## Likelihood Ratio 1.1160 2 0.57236  
## Pearson 1.1159 2 0.57237  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.017   
## Cramer's V : 0.017

age\_code <- table(coded$age, coded$readmitted)  
print("The age data is ")

## [1] "The age data is "

print(assocstats(age\_code))

## X^2 df P(> X^2)  
## Likelihood Ratio 16.529 18 0.55567  
## Pearson NaN 18 NaN  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: NaN   
## Cramer's V : NaN

race\_code <- table(coded$race, coded$readmitted)  
race\_code <- race\_code[2:5,]  
  
print("The race data is ")

## [1] "The race data is "

print(assocstats(race\_code))

## X^2 df P(> X^2)  
## Likelihood Ratio 3.3801 6 0.75985  
## Pearson 3.1912 6 0.78449  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.029   
## Cramer's V : 0.021

coded\_first <- diabetes\_data[diabetes\_data$x == "TRUE",]  
dim(coded\_first)

## [1] 12 54

race\_code\_first <- table(coded\_first$race, coded\_first$readmitted)  
race\_code\_first <- race\_code\_first[2:5,]  
print("The first diagnosis data is ")

## [1] "The first diagnosis data is "

print(assocstats(race\_code\_first))

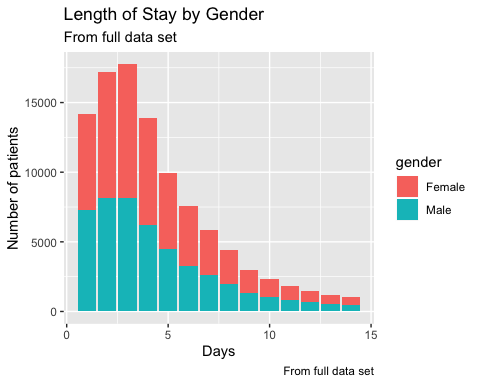
## X^2 df P(> X^2)  
## Likelihood Ratio 5.5617 6 0.47403  
## Pearson NaN 6 NaN  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: NaN   
## Cramer's V : NaN

## A t-test for the data - do males and females have the same mean lenght of stay in the hospital

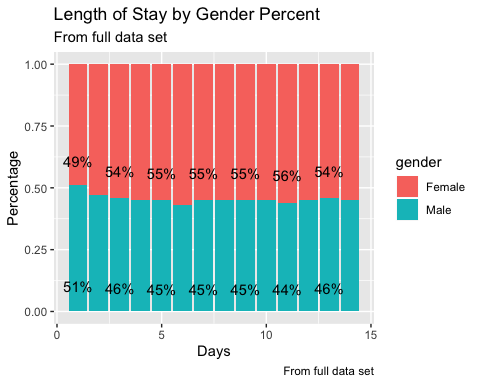
male\_stay <- diabetes\_data[diabetes\_data$gender =="Male",]  
female\_stay <- diabetes\_data[diabetes\_data$gender == "Female",]  
t.test(male\_stay$time\_in\_hospital, female\_stay$time\_in\_hospital)

##   
## Welch Two Sample t-test  
##   
## data: male\_stay$time\_in\_hospital and female\_stay$time\_in\_hospital  
## t = -9.5409, df = 99425, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2157958 -0.1422436  
## sample estimates:  
## mean of x mean of y   
## 4.299777 4.478797

tablos <- table(diabetes\_data$gender, diabetes\_data$time\_in\_hospital)  
tablos <- tablos[1:2,]  
tablos\_1 <- data.table(tablos)  
colnames(tablos\_1) <- c("gender", "length\_of\_stay", "total")  
tablos\_1$length\_of\_stay <- as.numeric(tablos\_1$length\_of\_stay)  
p <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) , aes(x = length\_of\_stay, y = total, fill = gender, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Length of Stay by Gender", subtitle = "From full data set", caption = "From full data set")  
p <- p + ylab("Number of patients") + xlab("Days")  
p



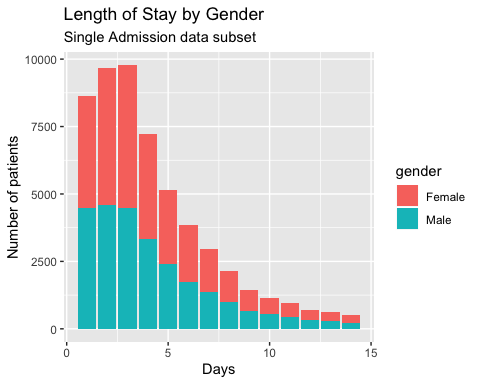
q <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) %>% mutate(perc = round(total/sum(total),2)), aes(x = length\_of\_stay, y = perc, fill = gender, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.2), check\_overlap = TRUE)  
q <- q + labs(title = "Length of Stay by Gender Percent", subtitle = "From full data set", caption = "From full data set")  
q <- q + ylab("Percentage") + xlab("Days")  
q



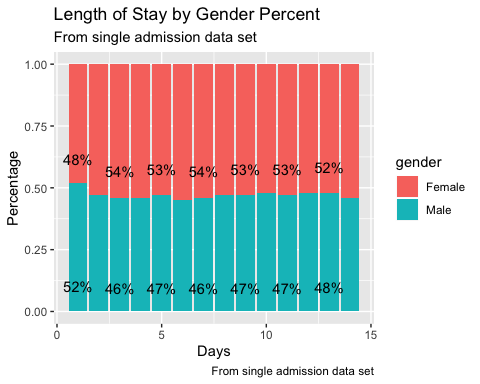
male\_nodupe\_stay <- No\_duplicated\_data[No\_duplicated\_data$gender == "Male",]  
female\_nodupe\_stay <- No\_duplicated\_data[No\_duplicated\_data$gender == "Female",]   
t.test(male\_nodupe\_stay$time\_in\_hospital, female\_nodupe\_stay$time\_in\_hospital)

##   
## Welch Two Sample t-test  
##   
## data: male\_nodupe\_stay$time\_in\_hospital and female\_nodupe\_stay$time\_in\_hospital  
## t = -4.5067, df = 53980, p-value = 6.597e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.16263957 -0.06405043  
## sample estimates:  
## mean of x mean of y   
## 4.159259 4.272604

tablos <- table(No\_duplicated\_data$gender, No\_duplicated\_data$time\_in\_hospital)  
tablos <- tablos[1:2,]  
tablos\_1 <- data.table(tablos)  
colnames(tablos\_1) <- c("gender", "length\_of\_stay", "total")  
tablos\_1$length\_of\_stay <- as.numeric(tablos\_1$length\_of\_stay)  
p <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) , aes(x = length\_of\_stay, y = total, fill = gender, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Length of Stay by Gender", subtitle = "Single Admission data subset")  
p <- p + ylab("Number of patients") + xlab("Days")  
p



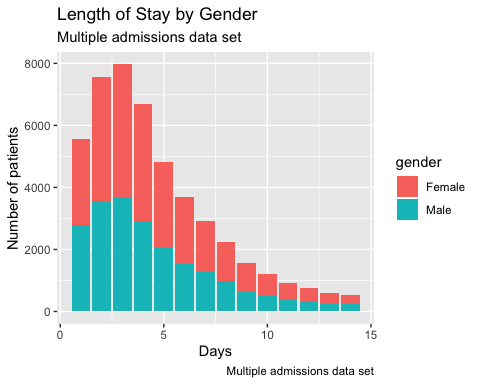
q <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) %>% mutate(perc = round(total/sum(total),2)), aes(x = length\_of\_stay, y = perc, fill = gender, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.2), check\_overlap = TRUE)  
q <- q + labs(title = "Length of Stay by Gender Percent", subtitle = "From single admission data set", caption = "From single admission data set")  
q <- q + ylab("Percentage") + xlab("Days")  
q



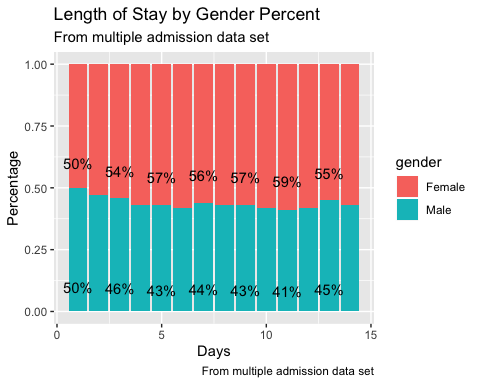
male\_dupe\_stay <- Dupe\_diabetes\_data[Dupe\_diabetes\_data$gender == "Male",]  
female\_dupe\_stay <-Dupe\_diabetes\_data[Dupe\_diabetes\_data$gender == "Female",]   
t.test(male\_dupe\_stay$time\_in\_hospital, female\_dupe\_stay$time\_in\_hospital)

##   
## Welch Two Sample t-test  
##   
## data: male\_dupe\_stay$time\_in\_hospital and female\_dupe\_stay$time\_in\_hospital  
## t = -8.4278, df = 45232, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2912149 -0.1813197  
## sample estimates:  
## mean of x mean of y   
## 4.472108 4.708375

tablos <- table(Dupe\_diabetes\_data$gender, Dupe\_diabetes\_data$time\_in\_hospital)  
tablos <- tablos[1:2,]  
tablos\_1 <- data.table(tablos)  
colnames(tablos\_1) <- c("gender", "length\_of\_stay", "total")  
tablos\_1$length\_of\_stay <- as.numeric(tablos\_1$length\_of\_stay)  
p <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) , aes(x = length\_of\_stay, y = total, fill = gender, cumulative = TRUE)) + geom\_col()  
p <- p + labs(title = "Length of Stay by Gender", subtitle = "Multiple admissions data set", caption = "Multiple admissions data set")  
p <- p + ylab("Number of patients") + xlab("Days")  
p



q <- ggplot(tablos\_1 %>% group\_by(length\_of\_stay) %>% mutate(perc = round(total/sum(total),2)), aes(x = length\_of\_stay, y = perc, fill = gender, cumulative = TRUE)) + geom\_col() +  
 geom\_text(aes(label = paste0(perc\*100,"%")),   
 position = position\_stack(vjust = 0.2), check\_overlap = TRUE)  
q <- q + labs(title = "Length of Stay by Gender Percent", subtitle = "From multiple admission data set", caption = "From multiple admission data set")  
q <- q + ylab("Percentage") + xlab("Days")  
q



## Now to test to see if Medicare patients have differing results

medi <- diabetes\_data[diabetes\_data$payer\_code == "MC",]  
no\_medi <- diabetes\_data[diabetes\_data$payer\_code != "MC",]  
total\_no\_perc <- sum(diabetes\_data$readmitted == "NO")/dim(diabetes\_data)[1]  
total\_lt\_30 <- sum(diabetes\_data$readmitted == "<30")/dim(diabetes\_data)[1]  
total\_mt\_30 <- sum(diabetes\_data$readmitted == ">30")/dim(diabetes\_data)[1]  
total\_no\_medi <- sum(medi$readmitted == "NO")/dim(medi)[1]  
total\_lt30\_medi <- sum(medi$readmitted == "<30")/dim(medi)[1]  
total\_mt30\_medi <- sum(medi$readmitted == ">30")/dim(medi)[1]  
total\_no\_medi\_no <- sum(no\_medi$readmitted == "NO")/dim(no\_medi)[1]  
total\_no\_medi\_lt30 <- sum(no\_medi$readmitted == "<30")/dim(no\_medi)[1]  
total\_no\_medi\_mt30 <- sum(no\_medi$readmitted == ">30")/dim(no\_medi)[1]  
frame\_it <- data.frame(total\_no\_perc, total\_lt\_30, total\_mt\_30)  
frame\_jt <- data.frame(total\_no\_medi, total\_lt30\_medi, total\_mt30\_medi)  
frame\_kt <- data.frame(total\_no\_medi\_no, total\_no\_medi\_lt30, total\_no\_medi\_mt30)  
colnames(frame\_it) <- c("no\_re-ad", "re-ad\_lt\_30", "re-ad\_mt\_30")  
colnames(frame\_jt) <- c("no\_re-ad", "re-ad\_lt\_30", "re-ad\_mt\_30")  
colnames(frame\_kt) <- c("no\_re-ad", "re-ad\_lt\_30", "re-ad\_mt\_30")  
framez <- rbind(frame\_it, frame\_jt, frame\_kt)  
framez <- 1000\* framez  
chisq.test(framez)

##   
## Pearson's Chi-squared test  
##   
## data: framez  
## X-squared = 3.0438, df = 4, p-value = 0.5505