set.seed(1680) # for reproducibility

library(ggplot2)

library(cluster) # for gower similarity and pam

library(Rtsne) # for t-SNE plot

library(ggplot2) # for visualization

df=df[-1,]

data = df[df$provider\_type=='Obstetrics/Gynecology' && df$nppes\_provider\_country=='US',]

data1= sqldf('select hcpcs\_code, sum (line\_srvc\_cnt) as ct from data group by hcpcs\_code')

data1 = sqldf('select \* from data1 order by ct desc')

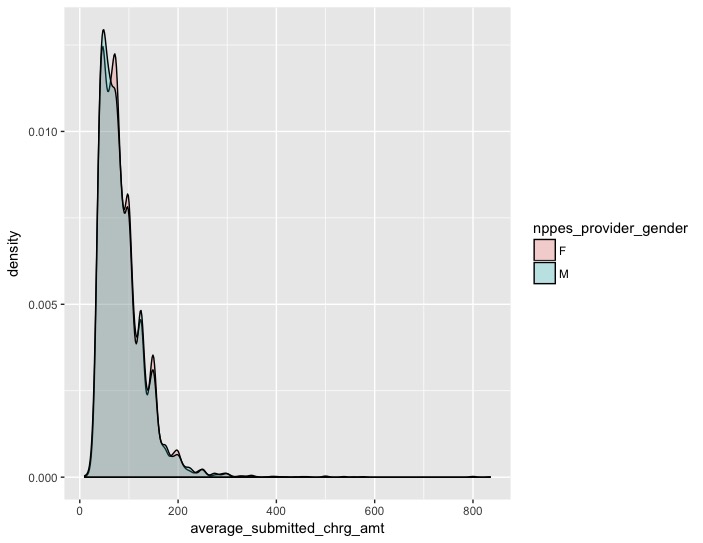
#pick top 20 most frequently used hcpcs: 99213 G0101 99214 J0897 Q0091 J1071 81002 99212 G0328 J1644 76830 J9267 99204 G0202

# J0585 99203 77052 81003 J1050 82270

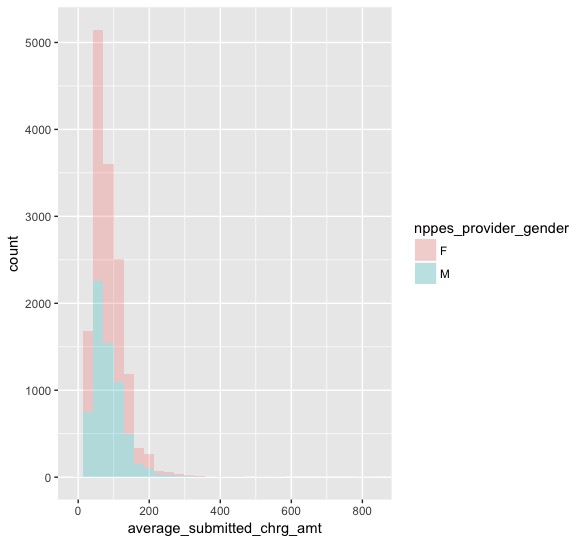
data3=data[data$hcpcs\_code=='G0101',]

data4=data3[data3$nppes\_provider\_gender!='',]

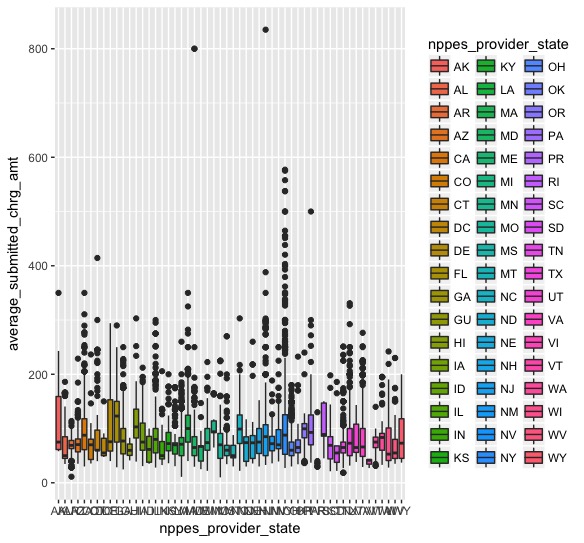
ggplot(data4,aes(x=average\_submitted\_chrg\_amt, fill=nppes\_provider\_gender)) + geom\_density(alpha=0.25)



ggplot(data4,aes(x=average\_submitted\_chrg\_amt, fill=nppes\_provider\_gender)) + geom\_histogram(alpha=0.25)

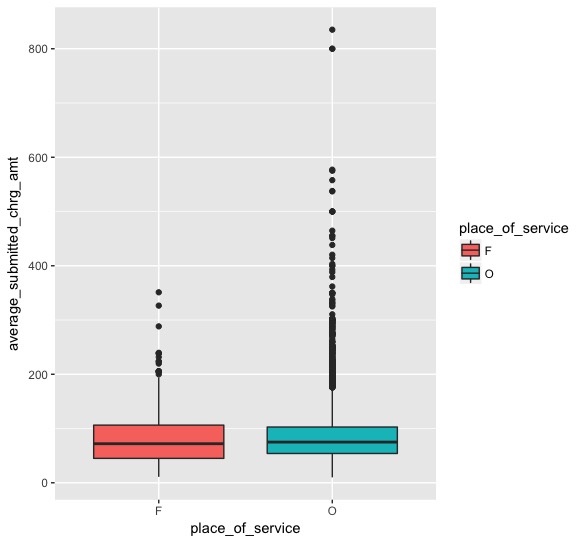


ggplot(data4, aes(x=nppes\_provider\_state, y = average\_submitted\_chrg\_amt, fill =nppes\_provider\_state )) + geom\_boxplot()



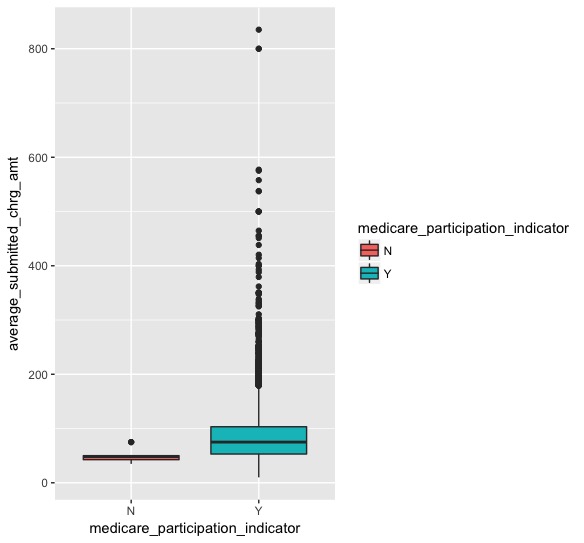
#different place of services show slightly different average submitted charge amt

ggplot(data4, aes(x=place\_of\_service, y = average\_submitted\_chrg\_amt, fill =place\_of\_service )) + geom\_boxplot()



#medicare participation or not significantly affect average submmited chr\_amt for the same service

ggplot(data4, aes(x=medicare\_participation\_indicator, y = average\_submitted\_chrg\_amt, fill =medicare\_participation\_indicator )) + geom\_boxplot()



data\_cluster = subset(data4,select=c(npi,nppes\_provider\_gender,nppes\_provider\_state,medicare\_participation\_indicator,

line\_srvc\_cnt,average\_submitted\_chrg\_amt))

#get gower\_dist which provides a way to calculate distance between discrete variables

gower\_dist <- daisy(data\_cluster[, 1:4],

metric = "gower",

type = list(logratio = 3))

summary(gower\_dist)

301476 dissimilarities, summarized :

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0018601 0.1034400 0.2358700 0.2314500 0.3271400 0.7773500

Metric : mixed ; Types = I, I, I, I, N, N

Number of objects : 777

# Calculate silhouette width for many k using PAM

sil\_width <- c(NA)

for(i in 2:10){

pam\_fit <- pam(gower\_dist,

diss = TRUE,

k = i)

sil\_width[i] <- pam\_fit$silinfo$avg.width

}

# Plot sihouette width (higher is better)

plot(1:10, sil\_width,

xlab = "Number of clusters",

ylab = "Silhouette Width")

lines(1:10, sil\_width)



#choose K = and apply pam clustering

pam\_fit <- pam(gower\_dist, diss = TRUE, k = 6)

#result interpretation

data\_cluster[pam\_fit$medoids, ]

703912 F NY

791147 F TX

163284 F PA

8820125 F CA

1995749 M NY

1126226 M TX

medicare\_participation\_indicator line\_srvc\_cnt

703912 Y 31

791147 Y 80

163284 Y 48

8820125 Y 17

1995749 Y 78

1126226 Y 25

average\_submitted\_chrg\_amt

703912 40.0000

791147 91.6875

163284 71.0000

8820125 100.0000

1995749 134.4533

1126226 86.2500

pam\_results <- data\_cluster %>%

mutate(cluster = pam\_fit$clustering) %>%

group\_by(cluster) %>%

do(the\_summary = summary(.))

[[1]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 NY : 621 : 0

F:2086 OH : 121 N: 3

M: 0 MI : 94 Y:2083

FL : 83

IL : 81

NC : 76

(Other):1010

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 18.49 Min. :1

1st Qu.: 26.00 1st Qu.: 53.00 1st Qu.:1

Median : 31.00 Median : 75.00 Median :1

Mean : 40.15 Mean : 89.06 Mean :1

3rd Qu.: 37.00 3rd Qu.:105.00 3rd Qu.:1

Max. :428.00 Max. :577.10 Max. :1

[[2]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 TX :600 : 0

F:1939 FL :158 N: 2

M: 0 NJ :156 Y:1937

VA : 98

MD : 85

CT : 77

(Other):765

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 29.54 Min. :2

1st Qu.: 55.00 1st Qu.: 50.00 1st Qu.:2

Median : 80.00 Median : 71.77 Median :2

Mean : 94.56 Mean : 80.92 Mean :2

3rd Qu.:118.00 3rd Qu.:100.00 3rd Qu.:2

Max. :563.00 Max. :350.00 Max. :2

[[3]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 PA :650 : 0

F:1582 FL :101 N: 2

M: 121 IL : 75 Y:1701

NJ : 65

OH : 61

NC : 54

(Other):697

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 18.49 Min. :3

1st Qu.: 41.00 1st Qu.: 59.00 1st Qu.:3

Median : 48.00 Median : 77.59 Median :3

Mean : 52.79 Mean : 86.94 Mean :3

3rd Qu.: 57.00 3rd Qu.:107.24 3rd Qu.:3

Max. :542.00 Max. :350.00 Max. :3

[[4]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 CA : 685 : 0

F:2830 OH : 167 N: 11

M: 112 IL : 157 Y:2931

MI : 141

FL : 125

MA : 112

(Other):1555

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 11.18 Min. :4

1st Qu.: 13.00 1st Qu.: 58.00 1st Qu.:4

Median : 17.00 Median : 78.00 Median :4

Mean : 23.07 Mean : 87.58 Mean :4

3rd Qu.: 21.00 3rd Qu.:108.00 3rd Qu.:4

Max. :599.00 Max. :350.00 Max. :4

[[5]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 NY : 460 : 0

F: 0 NJ : 185 N: 0

M:2295 FL : 152 Y:2295

CA : 123

PA : 119

VA : 99

(Other):1157

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 11.00 Min. :5

1st Qu.: 59.00 1st Qu.: 50.00 1st Qu.:5

Median : 78.00 Median : 73.00 Median :5

Mean : 92.28 Mean : 86.61 Mean :5

3rd Qu.:112.00 3rd Qu.:100.00 3rd Qu.:5

Max. :782.00 Max. :835.14 Max. :5

[[6]]

nppes\_provider\_gender nppes\_provider\_state medicare\_participation\_indicator

: 0 TX : 460 : 0

F: 0 FL : 295 N: 3

M:3981 OH : 227 Y:3978

MI : 194

NC : 188

PA : 177

(Other):2440

line\_srvc\_cnt average\_submitted\_chrg\_amt cluster

Min. : 11.00 Min. : 10.00 Min. :6

1st Qu.: 17.00 1st Qu.: 53.46 1st Qu.:6

Median : 25.00 Median : 75.00 Median :6

Mean : 28.14 Mean : 84.02 Mean :6

3rd Qu.: 35.00 3rd Qu.:102.00 3rd Qu.:6

Max. :398.00 Max. :500.00 Max. :6

tsne\_obj <- Rtsne(gower\_dist, is\_distance = TRUE)

tsne\_data <- tsne\_obj$Y %>%

data.frame() %>%

setNames(c("X", "Y")) %>%

mutate(cluster = factor(pam\_fit$clustering))

ggplot(aes(x = X, y = Y), data = tsne\_data) +

geom\_point(aes(color = cluster))