EDA of Human Cardiac Tissue-specific Proteome (Feb 23 2021) CaseOLAP Scores

Ashlyn Jew

Load libaries

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2)
```

Load Data

```r
# Human cardiac tissue-specific proteome - Feb 23 2021
human <- read_csv("https://raw.githubusercontent.com/asjew/heart_caseolap_EDA/main/Data/Human%20cardiac
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   protein = col_character(),
##   IHD = col_double(),
##   CM = col_double(),
##   ARR = col_double(),
##   VD = col_double(),
##   CHD = col_double(),
##   CCD = col_double(),
##   VOO = col_double(),
##   OTH = col_double()
## )
```

```
head(human)
```

```
## # A tibble: 6 x 9
##   protein    IHD      CM      ARR      VD     CHD     CCD     VOO     OTH
##   <chr>    <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 q8n4c6  0      0       0.0179  0.0105 0       0.0138 0      0
## 2 o60902  0.0554 0.0396  0.0508  0.0327 0.0410  0.0259 0.0292 0.0350
## 3 q18pe1  0      0.00622 0       0      0       0      0      0
## 4 p12821  0.128  0.165   0.0718  0.0538 0.0318  0.0380 0.0376 0.128
## 5 q9hbx9  0.0116 0.0251  0.00637 0      0.00659 0      0      0.00631
## 6 q8izh2  0      0       0       0      0       0      0      0.0163
```

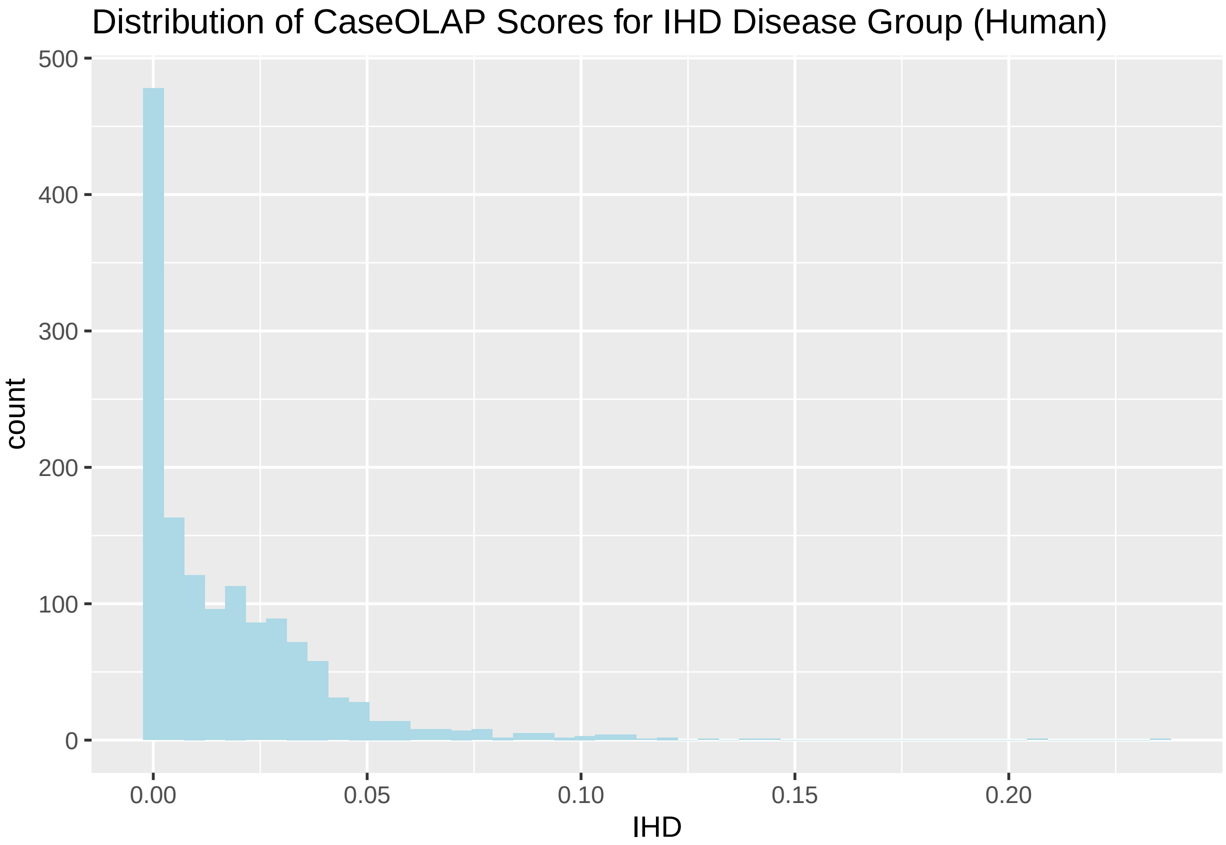
```
dim(human)
```

```
## [1] 1427    9
```

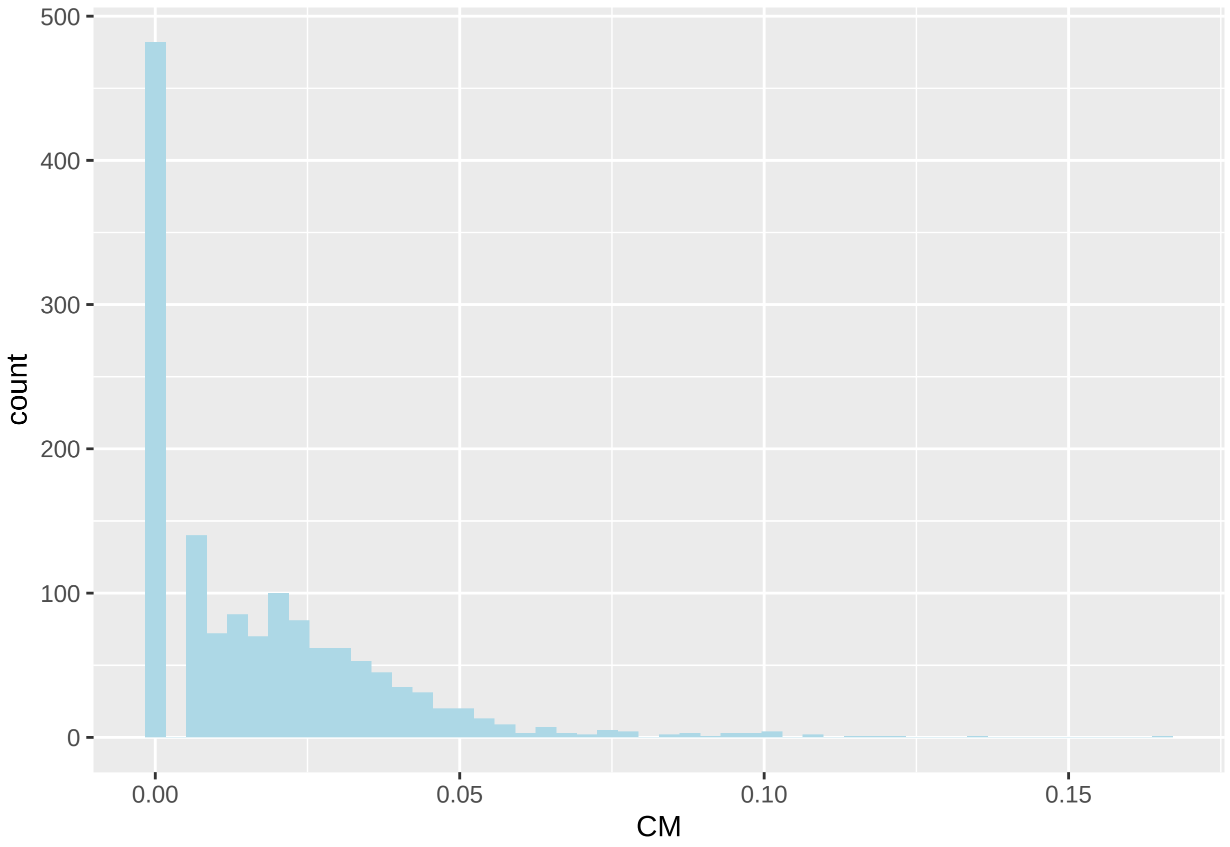
Exploratory Data Analysis

Histogram for each group

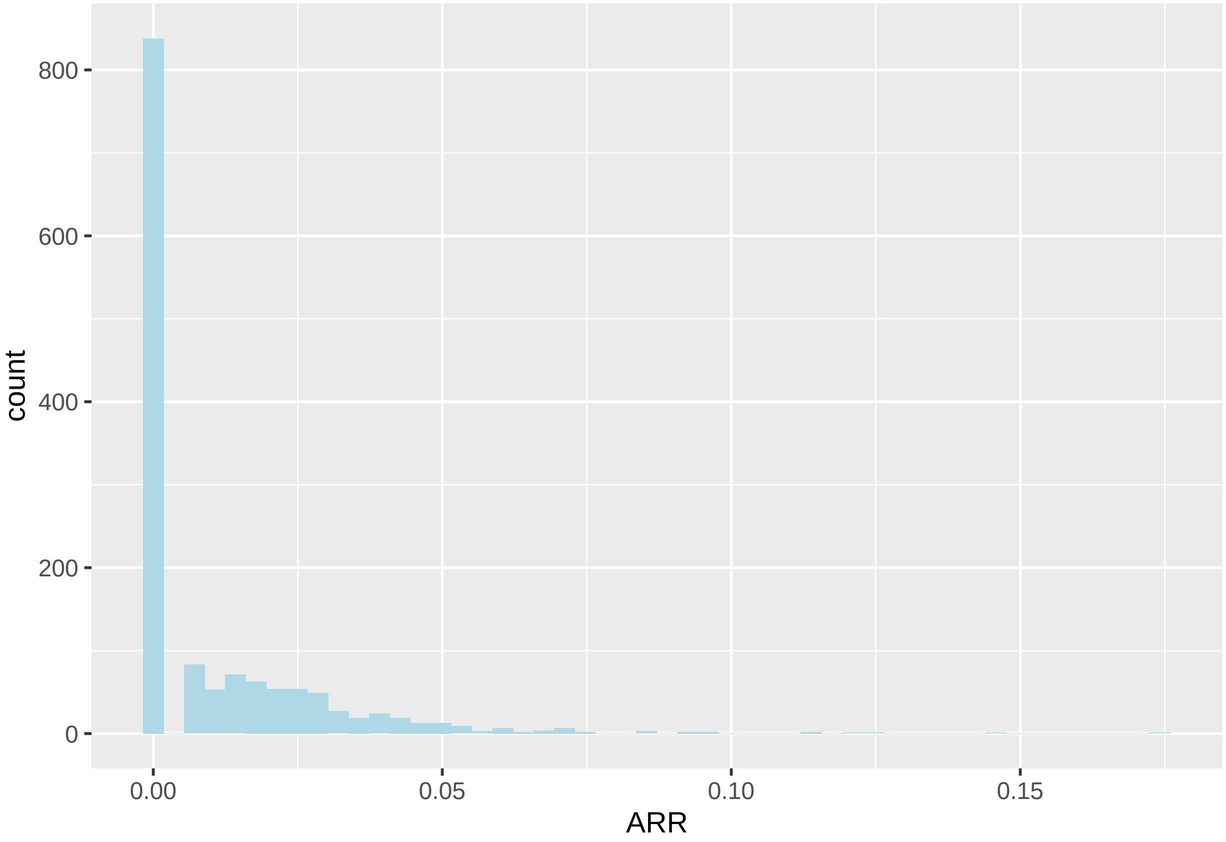
```
ggplot(human, aes(x = IHD)) + geom_histogram(fill = "lightblue", bins = 50) + ggtitle("Distribution of C
```



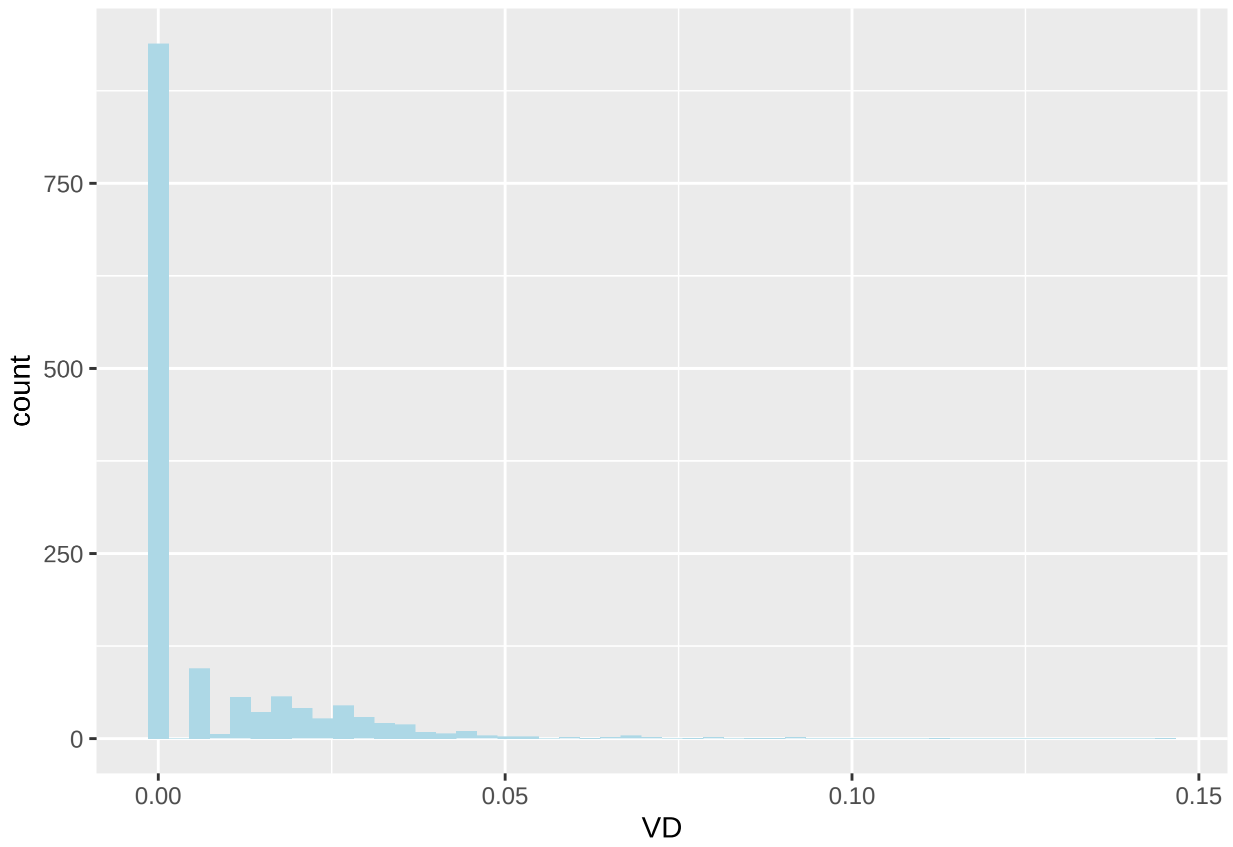
```r
ggplot(human, aes(x = CM)) + geom_histogram(fill = "lightblue", bins = 50)
```



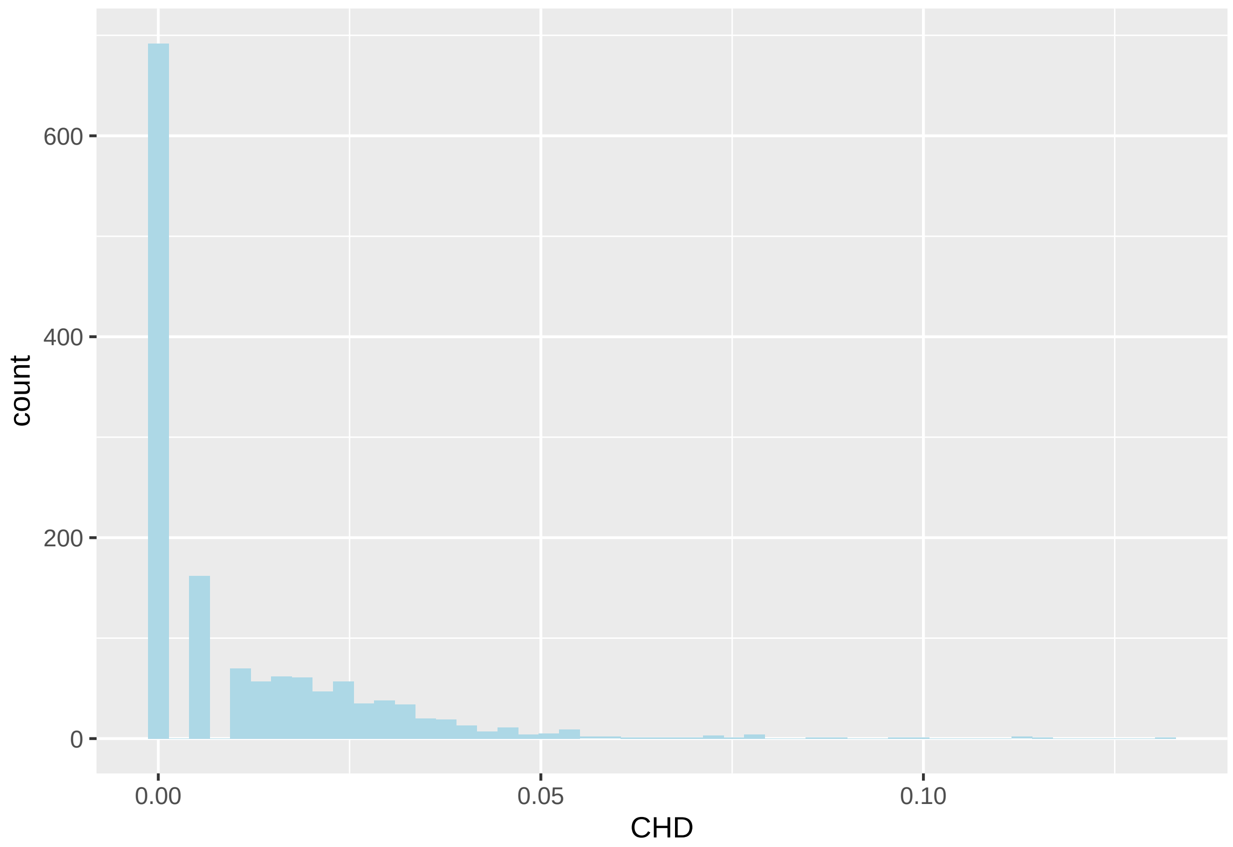
```r
ggplot(human, aes(x = ARR)) + geom_histogram(fill = "lightblue", bins = 50)
```



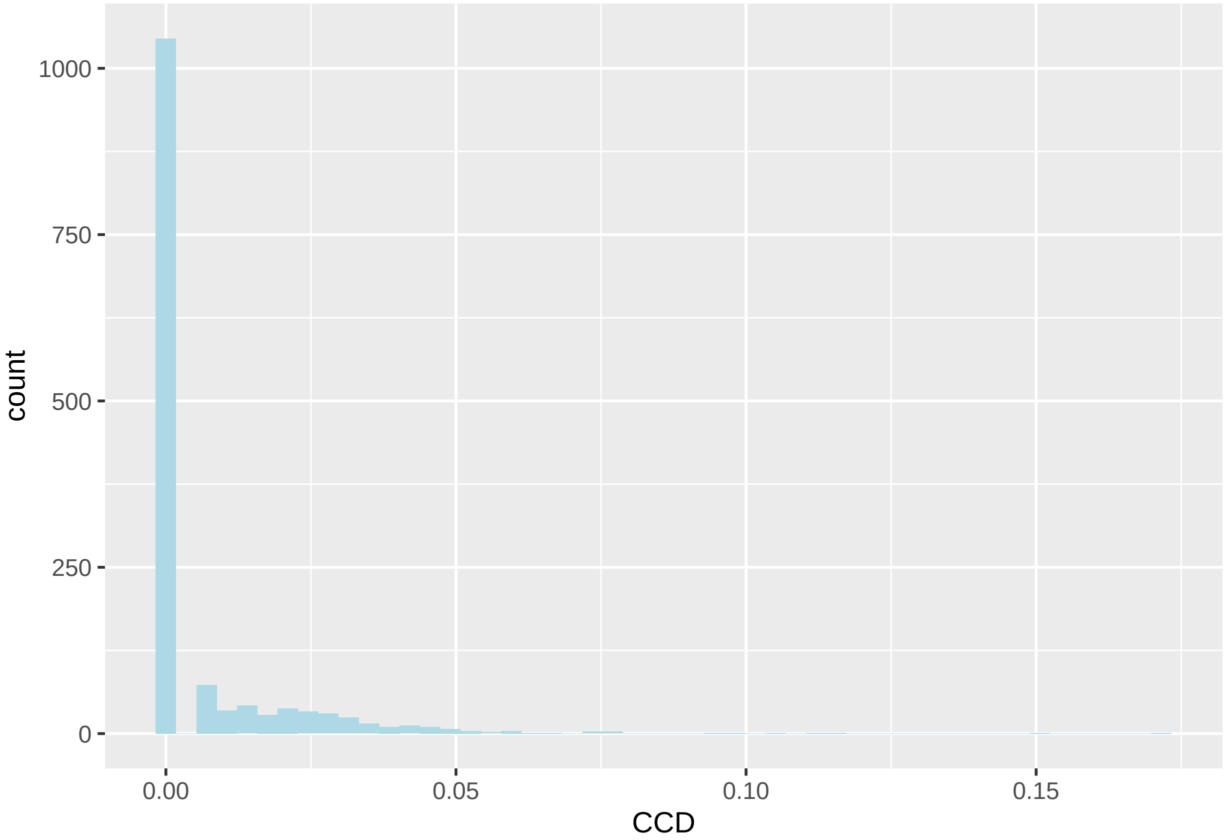
```r
ggplot(human, aes(x = VD)) + geom_histogram(fill = "lightblue", bins = 50)
```



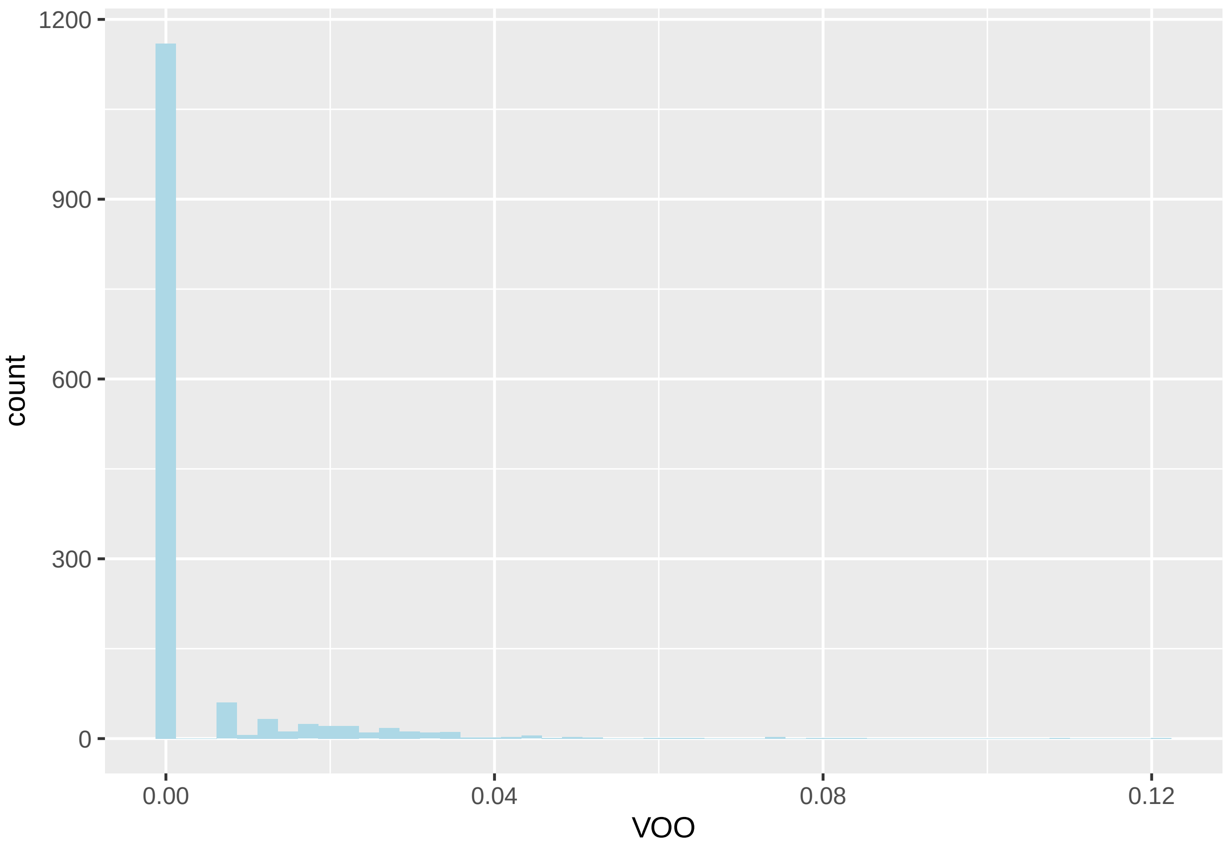
```r
ggplot(human, aes(x = CHD)) + geom_histogram(fill = "lightblue", bins = 50)
```



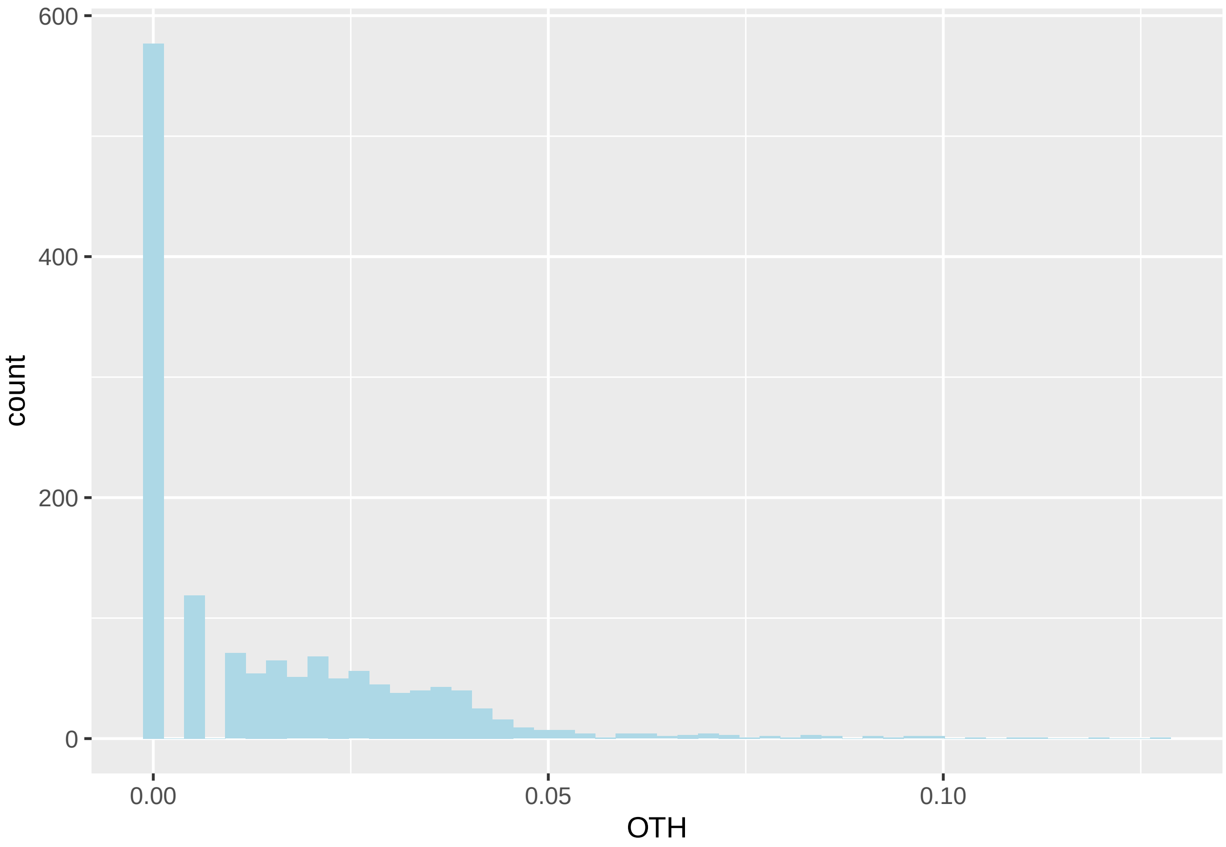
```
ggplot(human, aes(x = CCD)) + geom_histogram(fill = "lightblue", bins = 50)
```



```r
ggplot(human, aes(x = VOO)) + geom_histogram(fill = "lightblue", bins = 50)
```

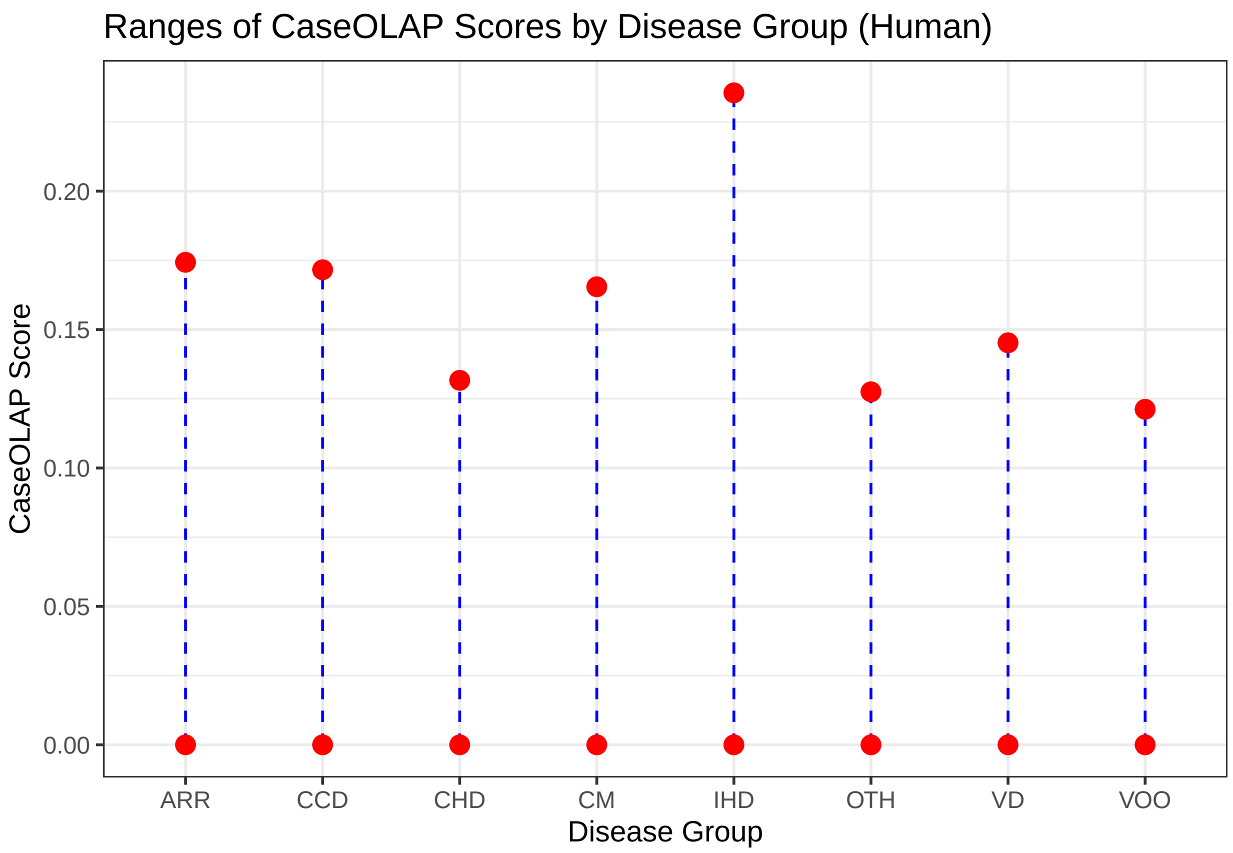


```r
ggplot(human, aes(x = OTH)) + geom_histogram(fill = "lightblue", bins = 50)
```



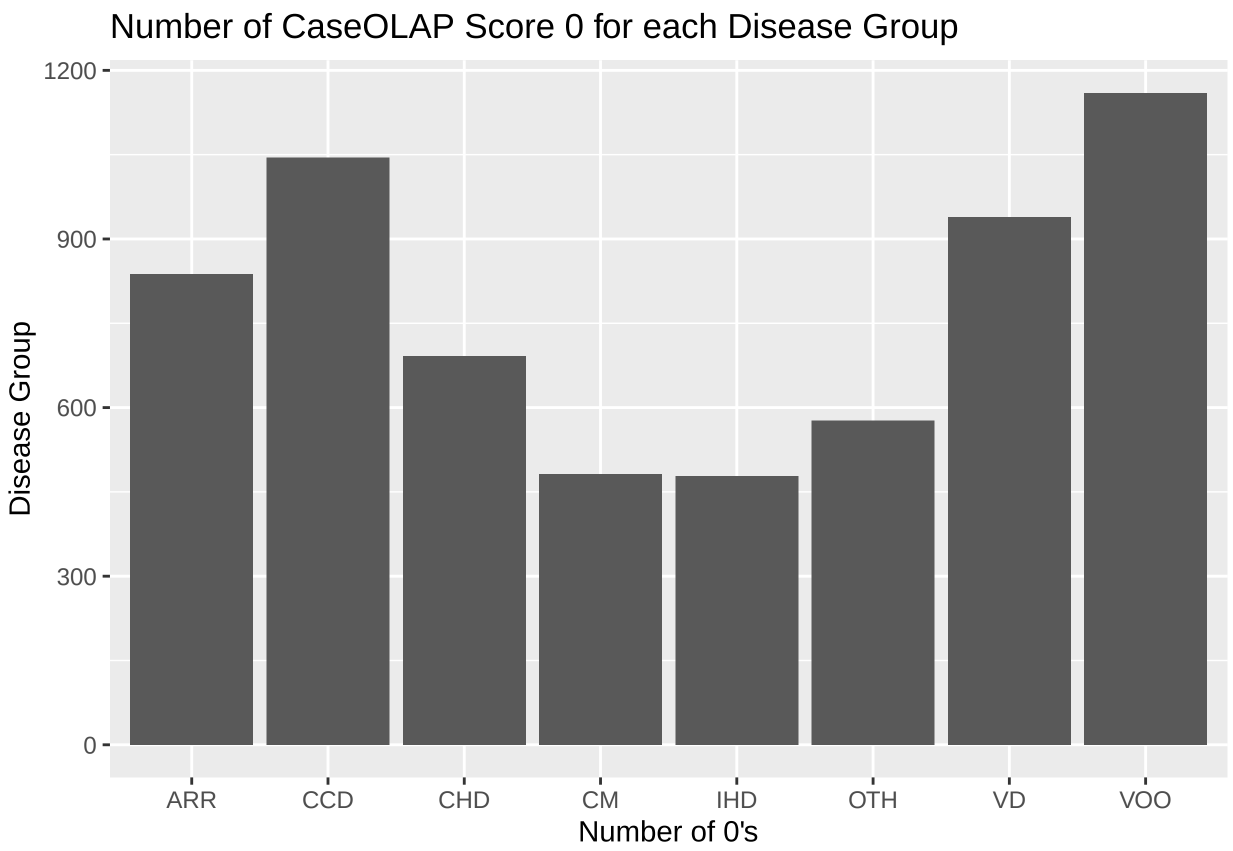
Ranges of CaseOLAP scores by group

```
human_min <- sapply(human[2:9], min)
human_max <- sapply(human[2:9], max)
human_ranges <- data.frame(id=c("IHD", "CM", "ARR","VD", "CHD", "CCD", "VOO", "OTH"),
                           min=human_min, max=human_max)

ggplot(human_ranges, aes(x=id))+
  geom_linerange(aes(ymin=min,ymax=max),linetype=2,color="blue")+
  geom_point(aes(y=min),size=3,color="red")+
  geom_point(aes(y=max),size=3,color="red")+
  theme_bw() + ggtitle("Ranges of CaseOLAP Scores by Disease Group (Human)") +
  xlab("Disease Group") + ylab("CaseOLAP Score")
```



Number of zeroes in each group

```
human_num_zero <- sapply(human[2:9], function(x) sum(x == 0))
human_zero <- data.frame(id=c("IHD", "CM", "ARR","VD", "CHD", "CCD", "VOO", "OTH"),
                         num_zero = human_num_zero)

ggplot(data = human_zero, aes(x=id, y=num_zero)) + geom_bar(stat="identity") +
  ggtitle("Number of CaseOLAP Score 0 for each Disease Group") +
  xlab("Number of 0's") +ylab("Disease Group")
```



Top 20 Analysis

```
# Summary Statistics
summary(human[2:9])
```

```
##       IHD                 CM                ARR                 VD          
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000  
##  Median :0.01160   Median :0.01244   Median :0.00000   Median :0.000000  
##  Mean   :0.01755   Mean   :0.01735   Mean   :0.01060   Mean   :0.007668  
##  3rd Qu.:0.02686   3rd Qu.:0.02706   3rd Qu.:0.01773   3rd Qu.:0.010497  
##  Max.   :0.23552   Max.   :0.16546   Max.   :0.17430   Max.   :0.145220  
##       CHD                 CCD                 VOO                 OTH          
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000  
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000  
##  Median :0.006545   Median :0.000000   Median :0.000000   Median :0.009995  
##  Mean   :0.011114   Mean   :0.006543   Mean   :0.004095   Mean   :0.015072  
##  3rd Qu.:0.018652   3rd Qu.:0.006889   3rd Qu.:0.000000   3rd Qu.:0.025302  
##  Max.   :0.131668   Max.   :0.171596   Max.   :0.121190   Max.   :0.127521
```

```
# Get top 20 proteins for each group
human_IHD <- human %>% arrange(desc(IHD))
```

```r
t20_humanIHD <- human_IHD[1:20, ]$protein

human_CM <- human %>% arrange(desc(CM))
t20_humanCM <- human_CM[1:20, ]$protein

human_ARR <- human %>% arrange(desc(ARR))
t20_humanARR <- human_ARR[1:20, ]$protein

human_VD <- human %>% arrange(desc(VD))
t20_humanVD <- human_VD[1:20, ]$protein

human_CHD <- human %>% arrange(desc(CHD))
t20_humanCHD <- human_CHD[1:20, ]$protein

human_CCD <- human %>% arrange(desc(CCD))
t20_humanCCD <- human_CCD[1:20, ]$protein

human_VOO <- human %>% arrange(desc(VOO))
t20_humanVOO <- human_VOO[1:20, ]$protein

human_OTH <- human %>% arrange(desc(OTH))
t20_humanOTH <- human_OTH[1:20, ]$protein

# Find the proteins that appear in more than one top 20 list
Reduce(intersect, list(t20_humanIHD, t20_humanCM, t20_humanARR,
                       t20_humanVD, t20_humanCHD, t20_humanCCD, t20_humanVOO, t20_humanOTH))
```

```
## [1] "o15534" "q9nzs2" "q16836" "o95461" "p26358"
```

```r
# Combine top 20 lists into a dataframe
t20_human <- data.frame(t20_humanIHD, t20_humanCM, t20_humanARR,
                        t20_humanVD, t20_humanCHD, t20_humanCCD, t20_humanVOO, t20_humanOTH)

# Count the number of times each protein appears in the dataframe
sort(table(c(t20_humanIHD, t20_humanCM, t20_humanARR, t20_humanVD,
             t20_humanCHD, t20_humanCCD, t20_humanVOO, t20_humanOTH)))
```

```
## 
## o00300 o00400 o14788 o15266 o43612 p12643 p19474 p26678 p27169 p29323 p40763 
##      1      1      1      1      1      1      1      1      1      1      1 
## p45379 p48357 p52952 p53602 p63252 q03135 q07869 q11206 q13936 q16635 q53gg5 
##      1      1      1      1      1      1      1      1      1      1      1 
## q8nfu7 q8tct9 q92574 q96pn6 q96q15 q96qv1 q99959 q9h4e5 q9uhl9 q9ulz3 q9y6j6 
##      1      1      1      1      1      1      1      1      1      1      1 
## q9y6m7 o75369 p11532 p12081 p34949 p42574 p50402 q06124 q15004 q92736 q9y3q4 
##      1      2      2      2      2      2      2      2      2      2      2 
## o14649 o94925 o95433 p05154 p15382 p51787 q12809 q96q05 q9nv58 q9ui32 p12821 
##      3      3      3      3      3      3      3      3      3      3      4 
## q01638 q92688 o43557 p61244 q05682 p10275 o15534 o95461 p26358 q16836 q9nzs2 
##      4      4      5      6      6      7      8      8      8      8      8
```

```r
# Proteins that appeared only once
t20_once <- names(which(sort(table(c(t20_humanIHD, t20_humanCM, t20_humanARR,
                                     t20_humanVD, t20_humanCHD, t20_humanCCD,
                                     t20_humanVOO, t20_humanOTH))) == 1))

# Create dataframe with caseolap scores of proteins that appeared only once
t20_once_caseolap <- data.frame(matrix(0, nrow = 32, ncol = 9))
unlist(human[which(human$protein == t20_once[1]), 2:9])
```

```
##        IHD         CM        ARR         VD        CHD        CCD        VOO
## 0.08474036 0.05258247 0.03292985 0.05430054 0.02818079 0.00000000 0.06293712
##        OTH
## 0.04411905
```

```r
for(i in 1:32)
{
  t20_once_caseolap[i, ] <- c(t20_once[i],
                              unlist(human[which(human$protein == t20_once[i]), 2:9]))
}
colnames(t20_once_caseolap) <- c("protein", "IHD", "CM", "ARR","VD", "CHD", "CCD", "VOO", "OTH")

# Columns with the highest caseolap score
colnames(t20_once_caseolap[2:9])[apply(t20_once_caseolap[2:9],1,which.max)]
```

```
##  [1] "IHD" "OTH" "VOO" "CHD" "IHD" "VOO" "CCD" "CM"  "IHD" "OTH" "CM"  "VD"
## [13] "CM"  "CHD" "IHD" "CCD" "IHD" "CM"  "CM"  "CCD" "CHD" "VOO" "IHD" "OTH"
## [25] "OTH" "CHD" "CHD" "IHD" "CHD" "CCD" "VD"  "IHD"
```

```r
# IHD
t20_once_caseolap[which(colnames(t20_once_caseolap[2:9])[apply(t20_once_caseolap[2:9],
                                                             1,which.max)] == "IHD"), 1]
```

```
## [1] "o00300" "o43612" "p27169" "p53602" "q03135" "q8nfu7" "q96qv1" "q9ulz3"
```

```r
#CM
t20_once_caseolap[which(colnames(t20_once_caseolap[2:9])[apply(t20_once_caseolap[2:9],
                                                             1,which.max)] == "CM"), 1]
```

```
## [1] "p26678" "p40763" "p48357" "q07869" "q11206"
```

```r
#ARR
t20_once_caseolap[which(colnames(t20_once_caseolap[2:9])[apply(t20_once_caseolap[2:9],
                                                             1,which.max)] == "ARR"), 1]
```

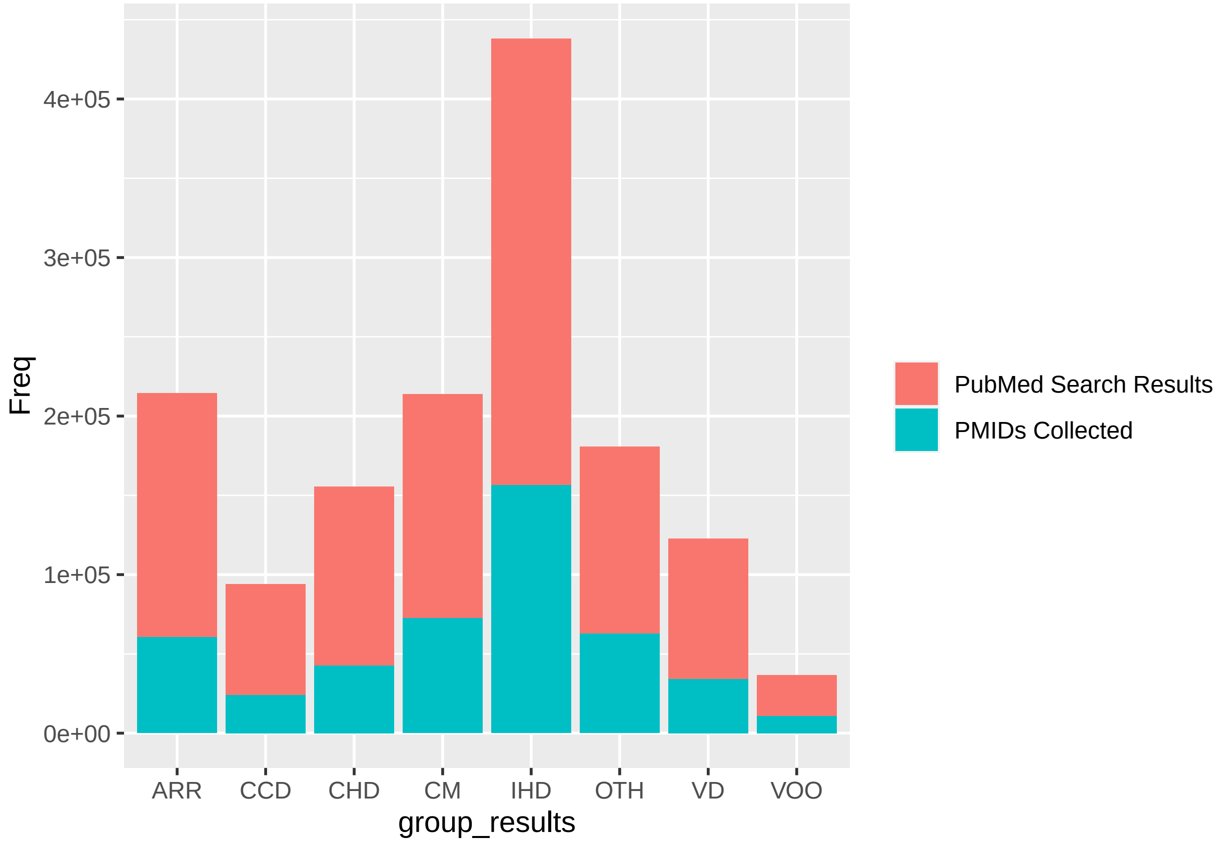
```
## character(0)
```

```r
# NONE

#VD
t20_once_caseolap[which(colnames(t20_once_caseolap[2:9])[apply(t20_once_caseolap[2:9],
                                                             1,which.max)] == "VD"), 1]
```

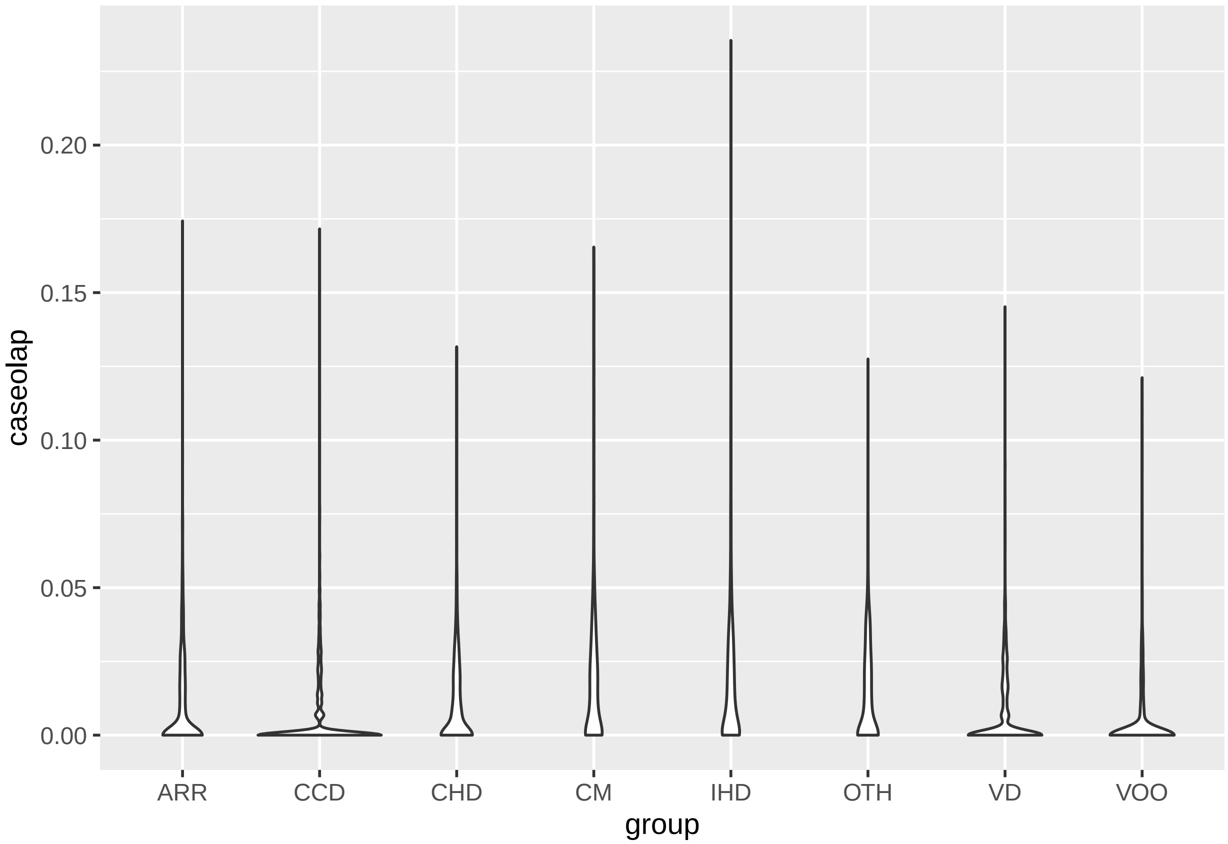
```
## [1] "p45379" "q9uhl9"
```

PMIDs

```r
# Number of pmid collected in dataset
IHD_pmid <- 156402
CM_pmid <- 72527
ARR_pmid <- 60535
VD_pmid <- 34242
CHD_pmid <- 42621
CCD_pmid <- 24119
VOO_pmid <- 10850
OTH_pmid <- 62708

# Myocardial Ischemia[MeSH Terms]
IHD_results <- 438259
# (Heart Failure[MeSH Terms]) OR (Cardiomyopathies[MeSH Terms])
CM_results <- 214025
# Arrhythmias, Cardiac[MeSH Terms]
ARR_results <- 214459
# Heart Valve Diseases[MeSH Terms]
VD_results <- 122701
# Heart Defects, Congenital[MeSH Terms]
CHD_results <- 155415
# Cardiac Conduction System Disease[MeSH Terms]
CCD_results <- 93934
# Ventricular Outflow Obstruction[MeSH Terms]
VOO_results <- 36612
# (((((Cardiomegaly[MeSH Terms]) OR (Endocarditis[MeSH Terms])) OR
# (Heart Arrest[MeSH Terms])) OR (Heart Rupture[MeSH Terms])) OR
# (Ventricular Dysfunction[MeSH Terms])) OR (Pericarditis[MeSH Terms])
OTH_results <- 180768


group_results <- c(rep("IHD", IHD_results), rep("CM", CM_results),
                   rep("ARR", ARR_results), rep("VD" , VD_results),
                   rep("CHD", CHD_results), rep("CCD", CCD_results),
                   rep("VOO" , VOO_results), rep("OTH" , OTH_results))
group_pmid <- c(rep(1, IHD_pmid), rep(0, IHD_results - IHD_pmid),
                rep(1, CM_pmid), rep(0, CM_results - CM_pmid),
                rep(1, ARR_pmid), rep(0, ARR_results - ARR_pmid),
                rep(1, VD_pmid), rep(0, VD_results - VD_pmid),
                rep(1, CHD_pmid), rep(0, CHD_results - CHD_pmid),
                rep(1, CCD_pmid), rep(0, CCD_results - CCD_pmid),
                rep(1, VOO_pmid), rep(0, VOO_results - VOO_pmid),
                rep(1, OTH_pmid), rep(0, OTH_results - OTH_pmid))
group_counts <- table(group_pmid, group_results)

ggplot(as.data.frame(group_counts), aes(group_results, Freq, fill=group_pmid)) +
  geom_bar(stat="identity") + scale_fill_discrete(name = "", labels = c("PubMed Search Results", "PMIDs
```

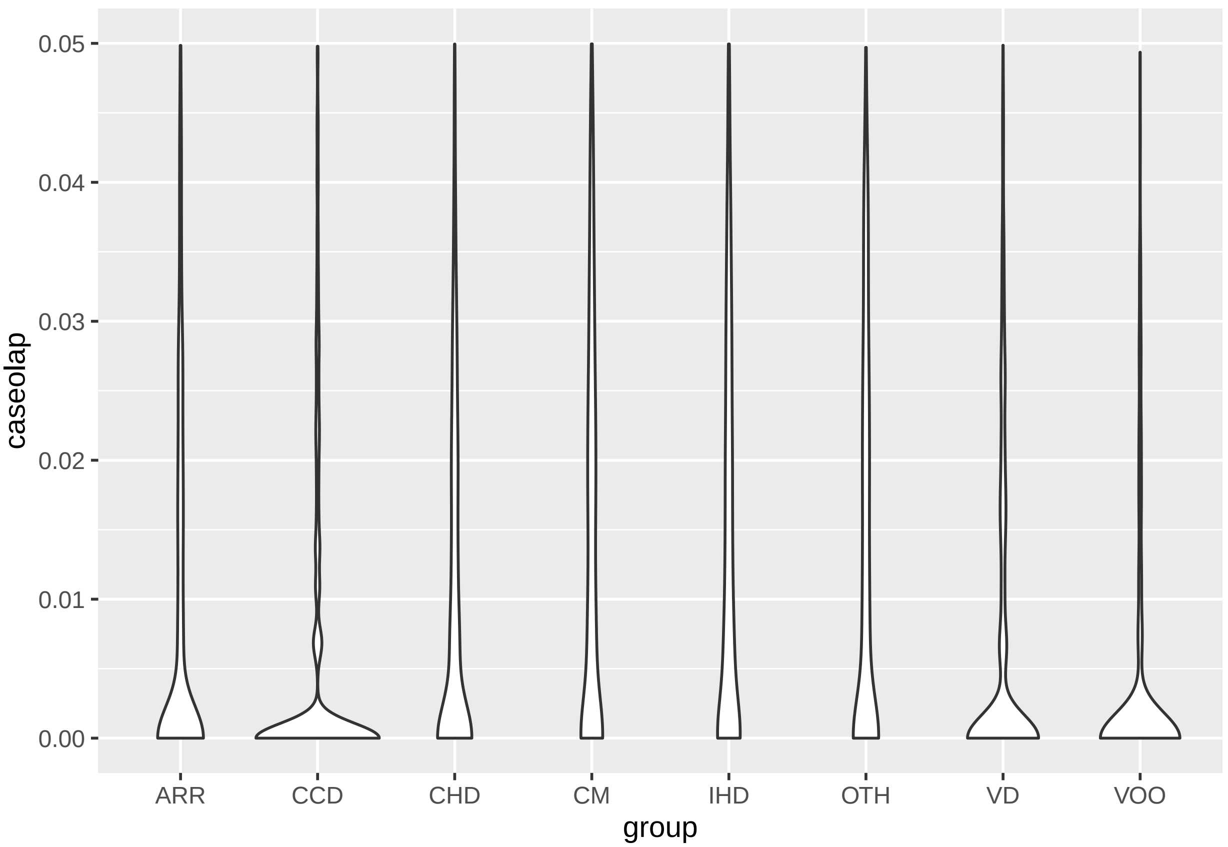


```r
# Violin plot for each group
caseolap_IHD <- cbind.data.frame("caseolap" = as.numeric(human$IHD),
                                 "group" = rep("IHD", length(human$IHD)))
caseolap_CM <- cbind.data.frame("caseolap" = as.numeric(human$CM),
                                "group" = rep("CM", length(human$CM)))
caseolap_ARR <- cbind.data.frame("caseolap" = as.numeric(human$ARR),
                                 "group" = rep("ARR", length(human$ARR)))
caseolap_VD <- cbind.data.frame("caseolap" = as.numeric(human$VD),
                                "group" = rep("VD", length(human$VD)))
caseolap_CHD <- cbind.data.frame("caseolap" = as.numeric(human$CHD),
                                 "group" = rep("CHD", length(human$CHD)))
caseolap_CCD <- cbind.data.frame("caseolap" = as.numeric(human$CCD),
                                 "group" = rep("CCD", length(human$CCD)))
caseolap_VOO <- cbind.data.frame("caseolap" = as.numeric(human$VOO),
                                 "group" = rep("VOO", length(human$VOO)))
caseolap_OTH <- cbind.data.frame("caseolap" = as.numeric(human$OTH),
                                 "group" = rep("OTH", length(human$OTH)))

caseolap <- rbind(caseolap_IHD, caseolap_CM, caseolap_ARR, caseolap_VD,
                  caseolap_CHD, caseolap_CCD, caseolap_VOO, caseolap_OTH)
ggplot(caseolap, aes(x=group, y=caseolap)) +
  geom_violin()
```



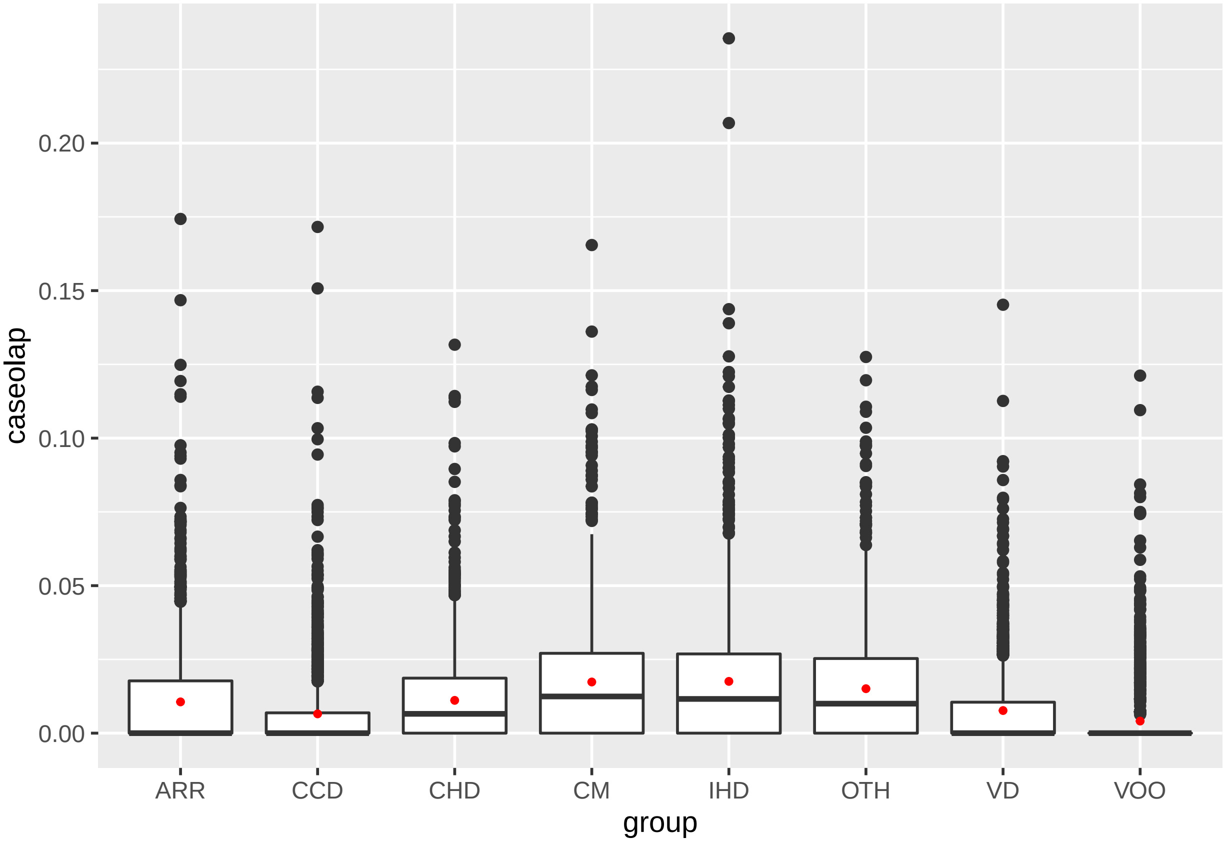
```
ggplot(caseolap, aes(x=group, y=caseolap)) +
  geom_violin() + scale_y_continuous(limits = c(0, 0.05))
```

```
## Warning: Removed 383 rows containing non-finite values (stat_ydensity).
```



```
# Boxplot for each group
caseolap <- rbind(caseolap_IHD, caseolap_CM, caseolap_ARR, caseolap_VD,
                  caseolap_CHD, caseolap_CCD, caseolap_VOO, caseolap_OTH)
ggplot(caseolap, aes(x=group, y=caseolap)) + geom_boxplot() +
    stat_summary(fun.y=mean, geom="point", shape=20, color="red", fill="red")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```


0.20
0.15
0.10
0.05
0.00
caseolap
ARR
CCD
CHD
CM
IHD
OTH
VD
VOO
group