EDA of Mouse & Rat Cardiac Tissue-specific Proteome (Feb 23 2021) CaseOLAP Scores

Ashlyn Jew

Load libaries

```
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()
```

```
library(ggplot2)
```

Load Data

```
# Mouse+rat cardiac tissue-specific proteome - Feb 23 2021
mouserat <- read_csv("https://raw.githubusercontent.com/asjew/heart_caseolap_EDA/main/Data/Mouse%2Brat%
```

```
##
## -- 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()
## )
```

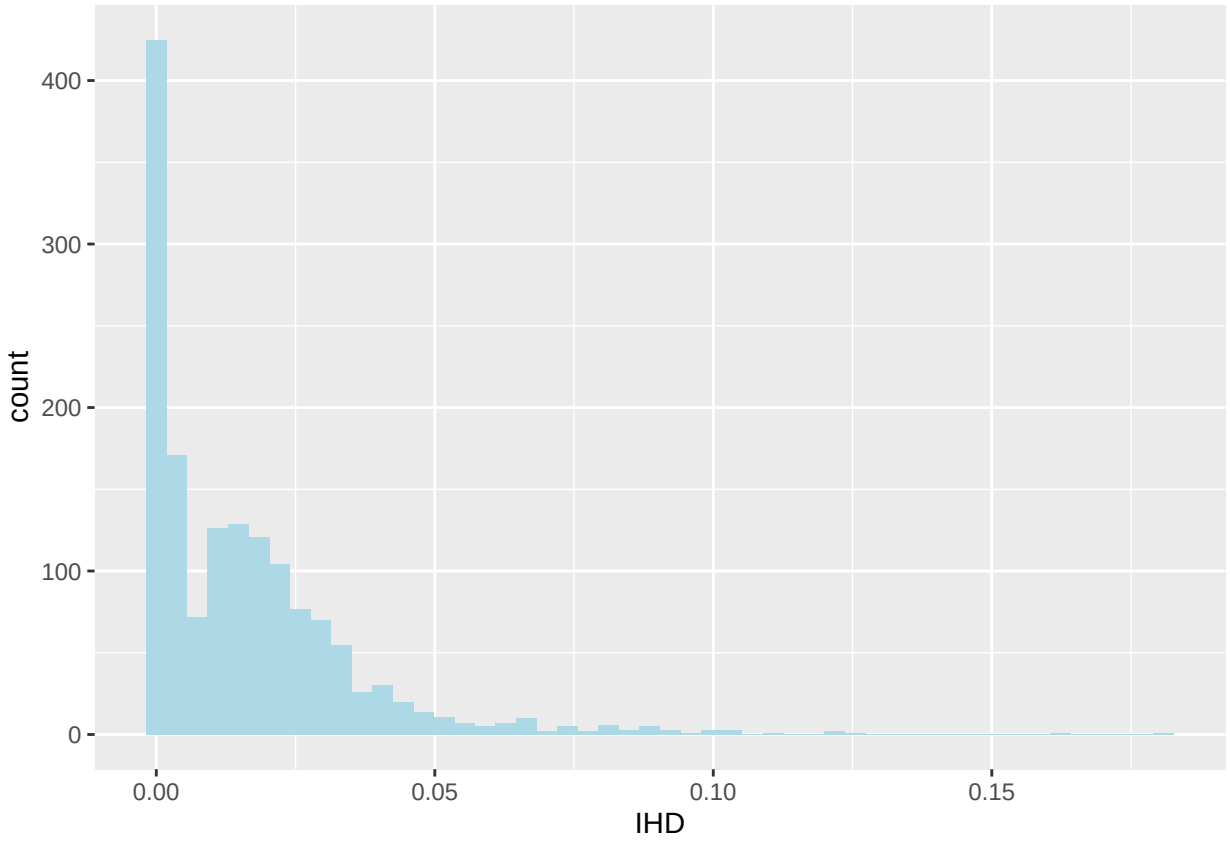
```
head(mouserat)
```

```
## # A tibble: 6 x 9
##   protein     IHD      CM    ARR     VD     CHD     CCD     VOO     OTH
##   <chr>     <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 p28076  0.0101  0.0106  0      0      0       0       0       0
## 2 o08573  0.00502 0.0203  0      0      0       0       0       0.00537
## 3 q63488  0.0151  0       0      0.0112 0.00898 0       0.00609 0
## 4 p56677  0       0       0      0      0.0113  0       0       0
## 5 p26645  0.0382  0.0243  0.0150 0.0229 0.00886 0.00906 0.0201  0.00530
## 6 p49586  0       0.00531 0      0      0.00567 0       0       0
```

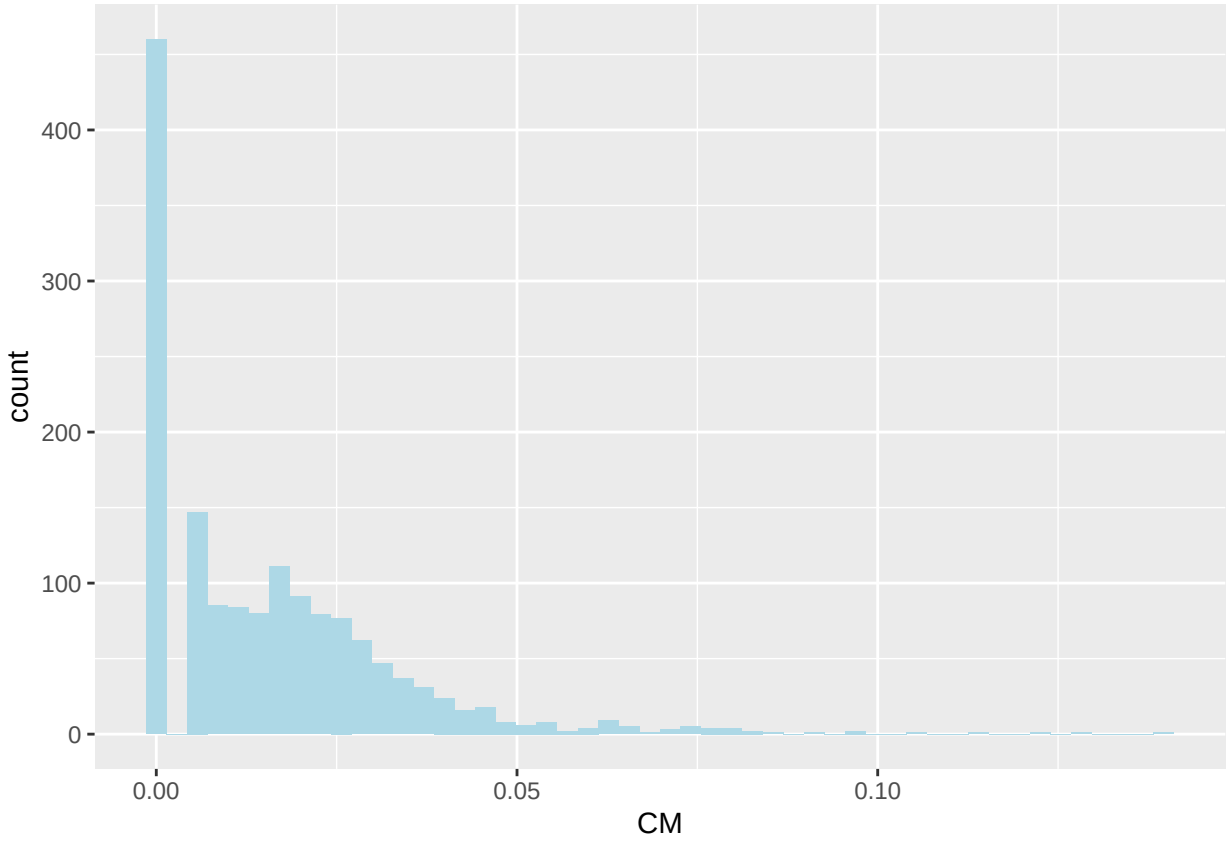
Exploratory Data Analysis

Histogram for each group

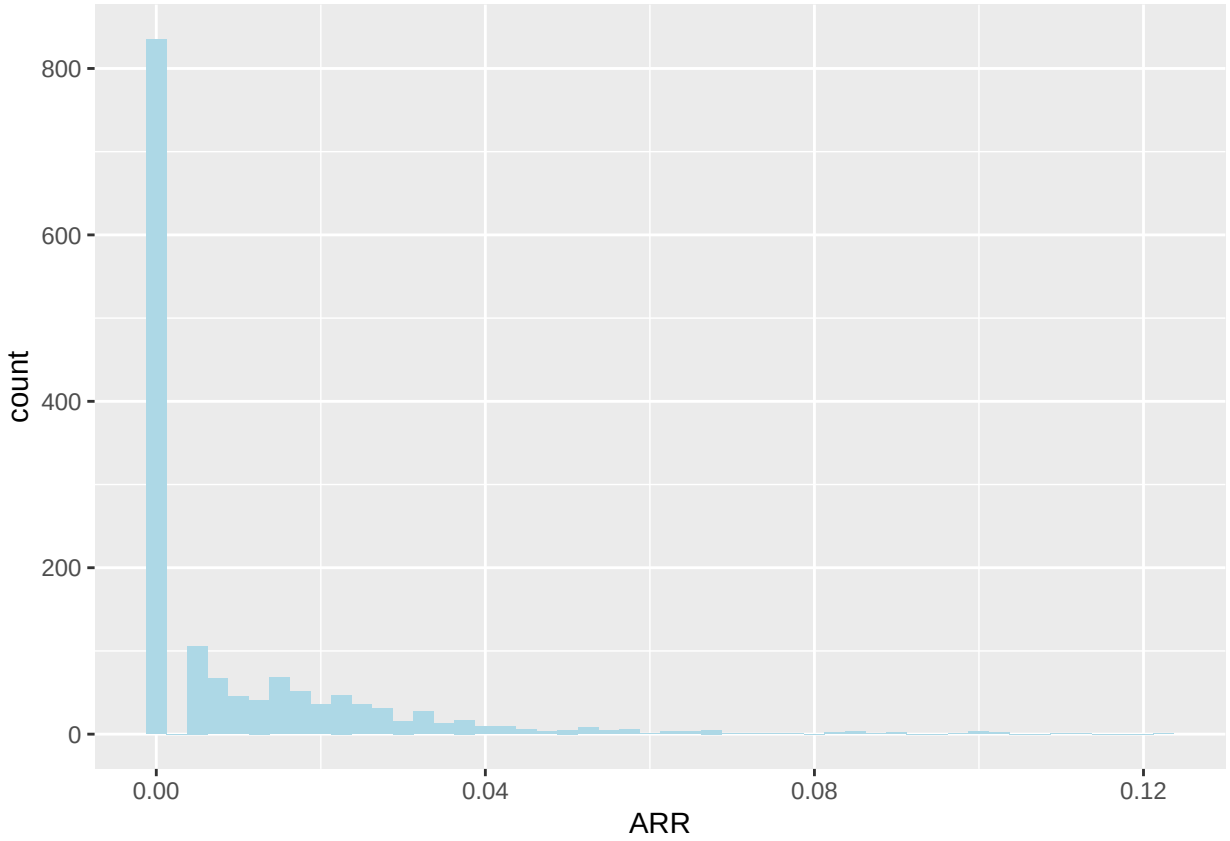
```
ggplot(mouserat, aes(x = IHD)) + geom_histogram(fill = "lightblue", bins = 50)
```



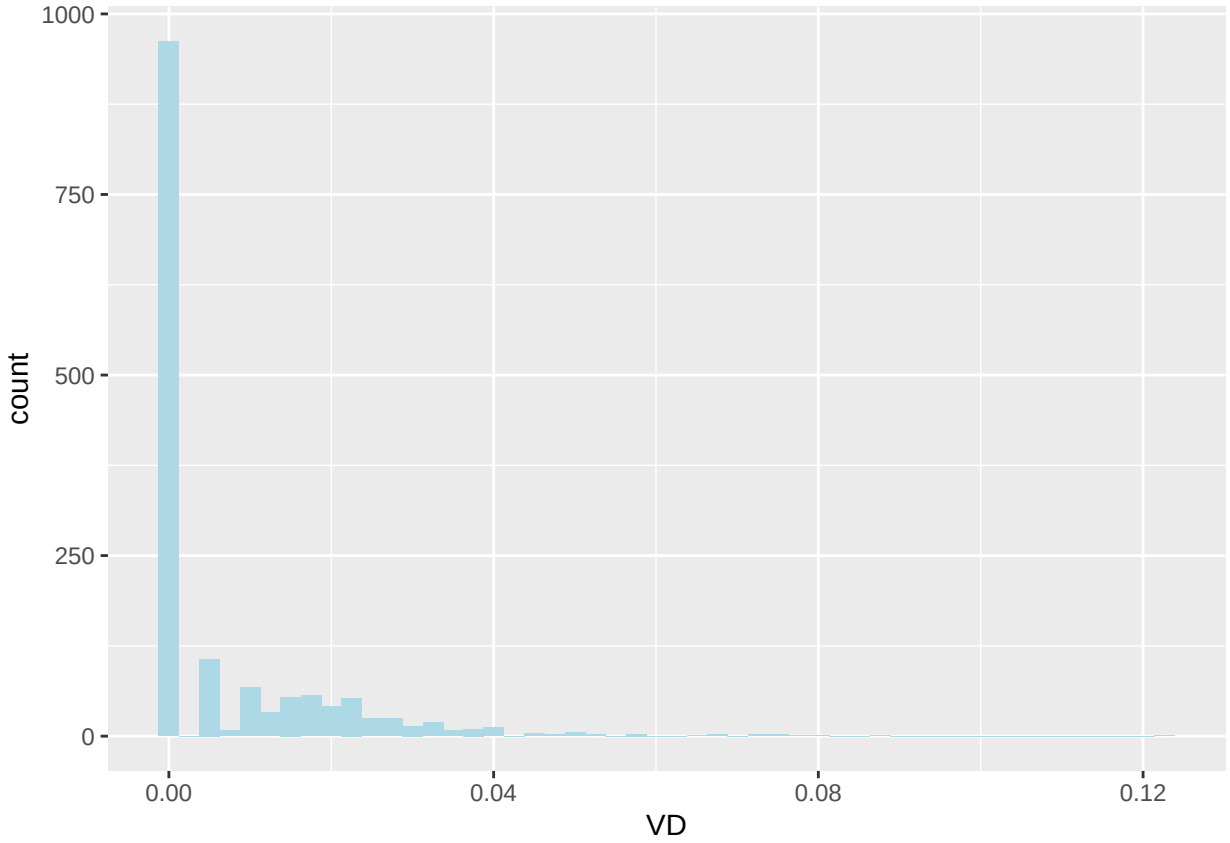

```
ggplot(mouserat, aes(x = CM)) + geom_histogram(fill = "lightblue", bins = 50)
```



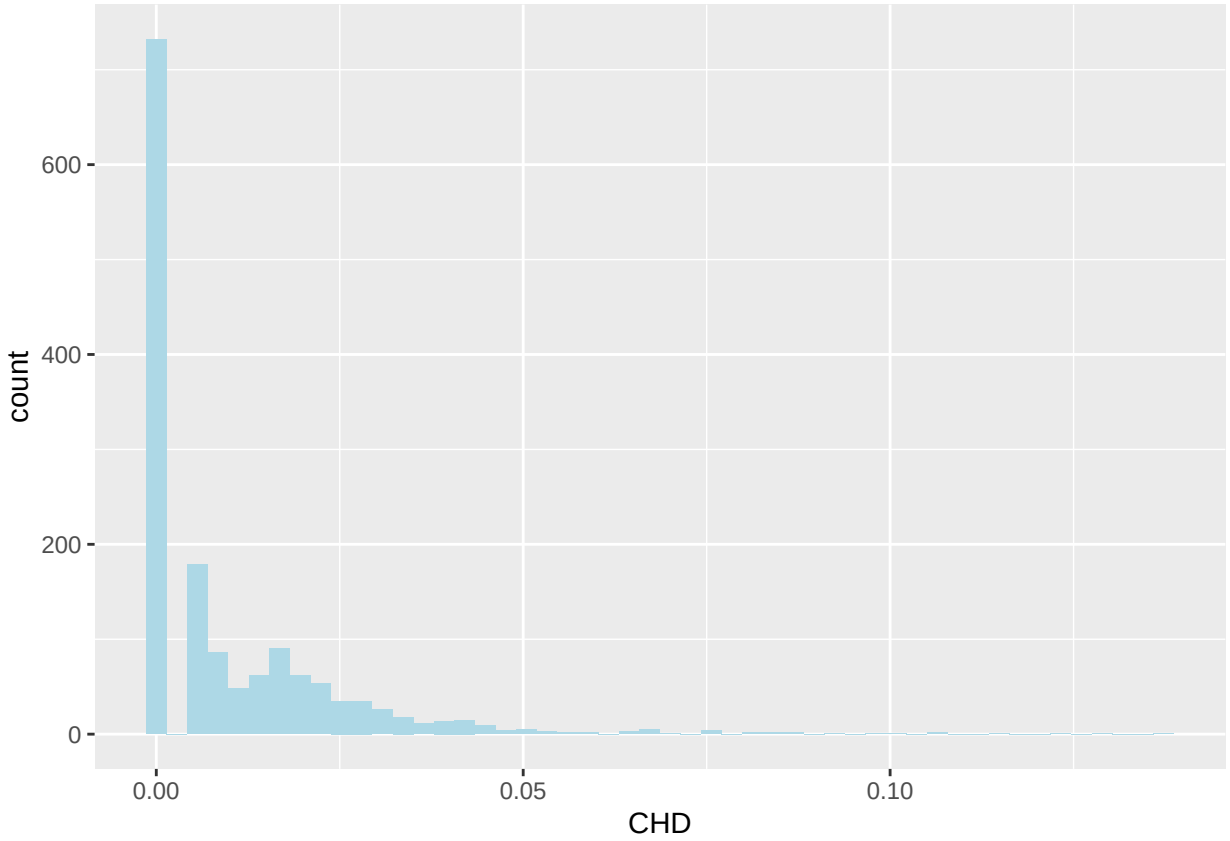

```
ggplot(mouserat, aes(x = ARR)) + geom_histogram(fill = "lightblue", bins = 50)
```



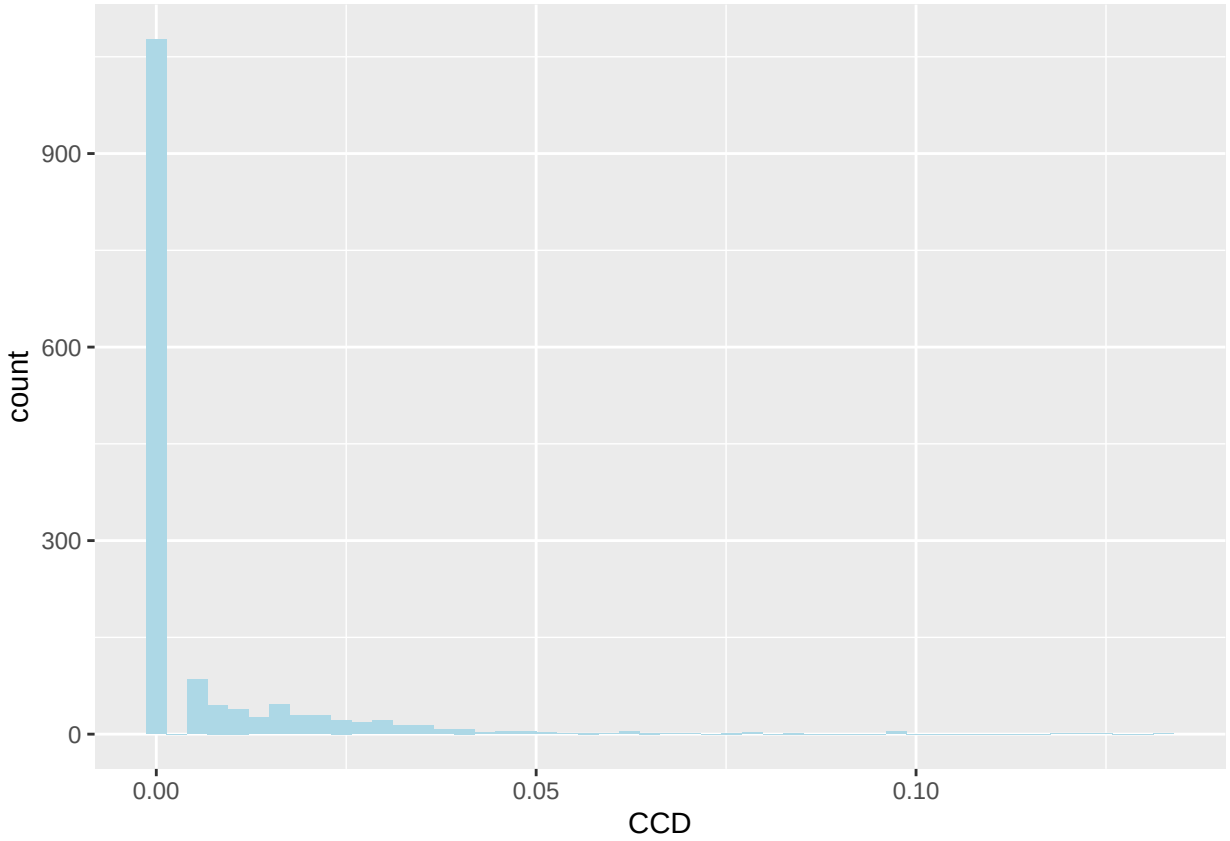

```
ggplot(mouserat, aes(x = VD)) + geom_histogram(fill = "lightblue", bins = 50)
```



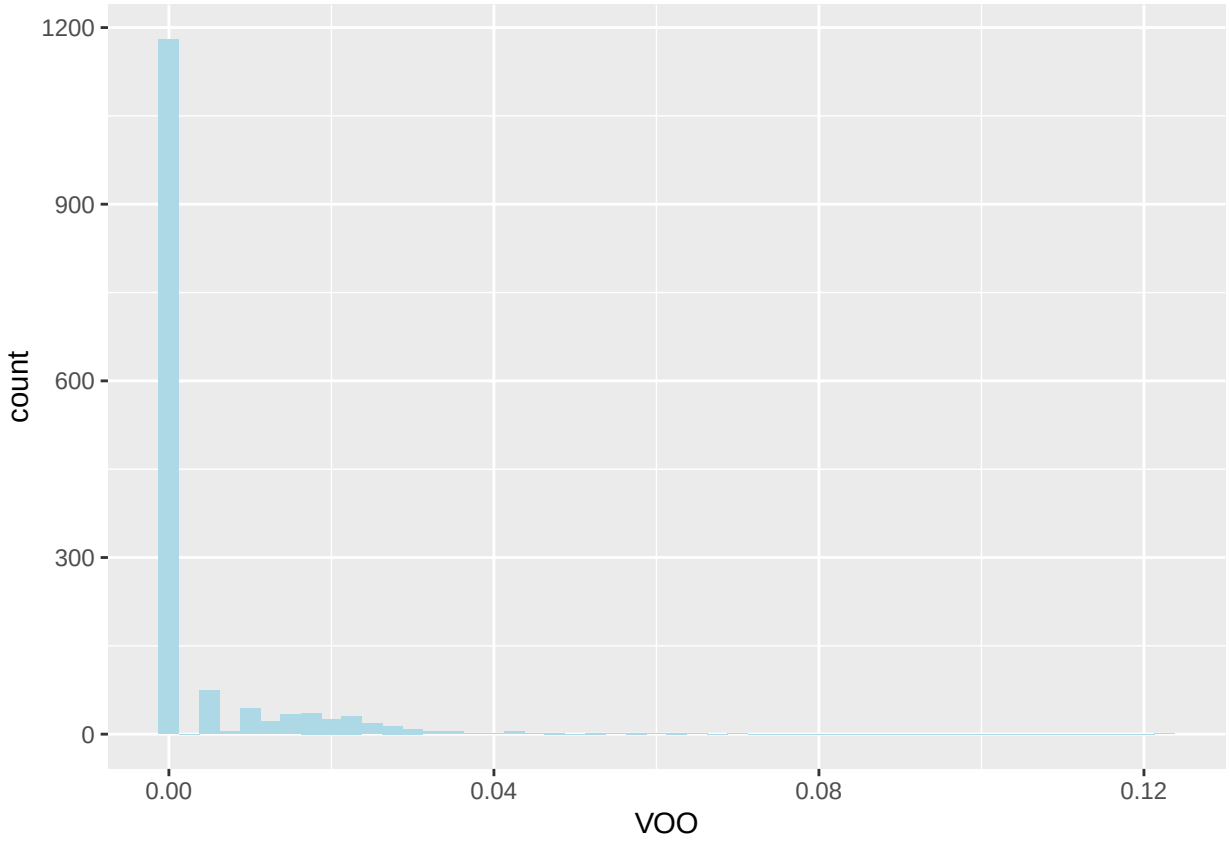

```
ggplot(mouserat, aes(x = CHD)) + geom_histogram(fill = "lightblue", bins = 50)
```



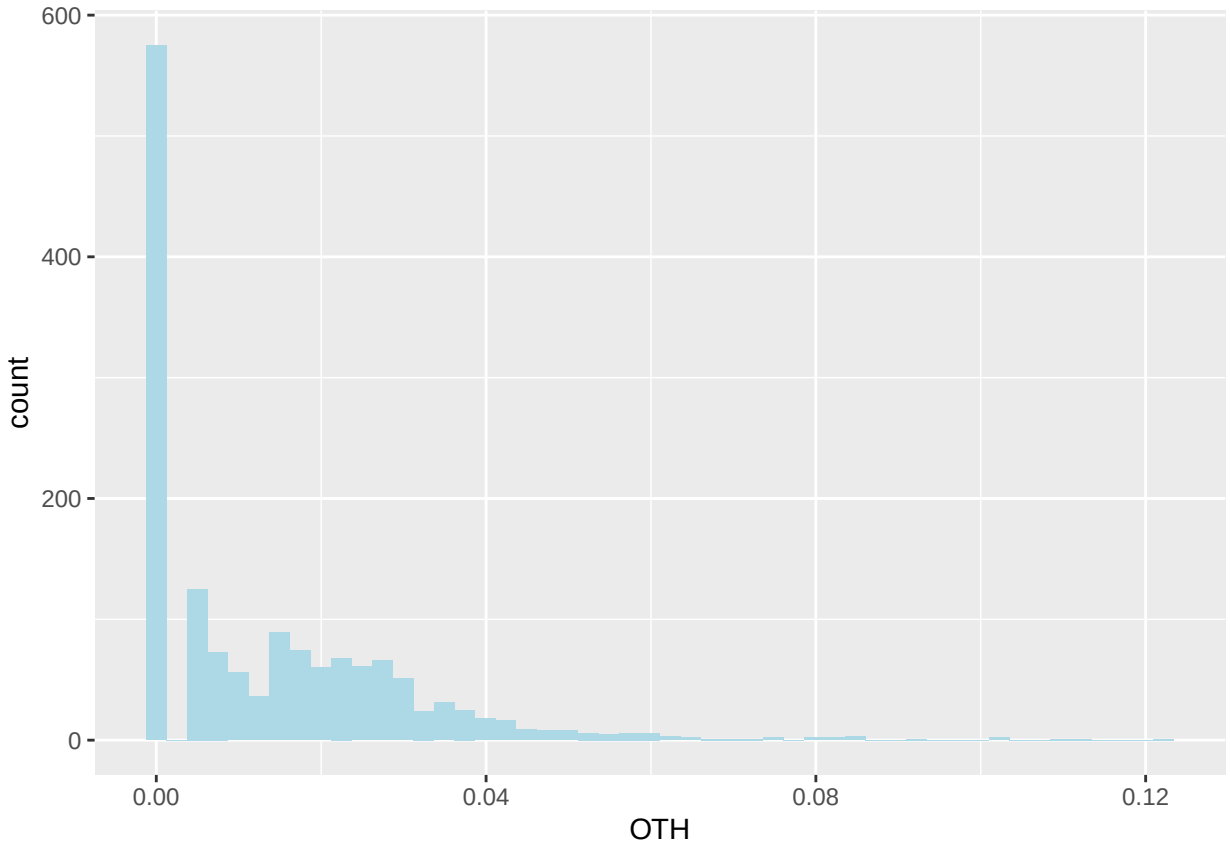

```
ggplot(mouserat, aes(x = CCD)) + geom_histogram(fill = "lightblue", bins = 50)
```




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

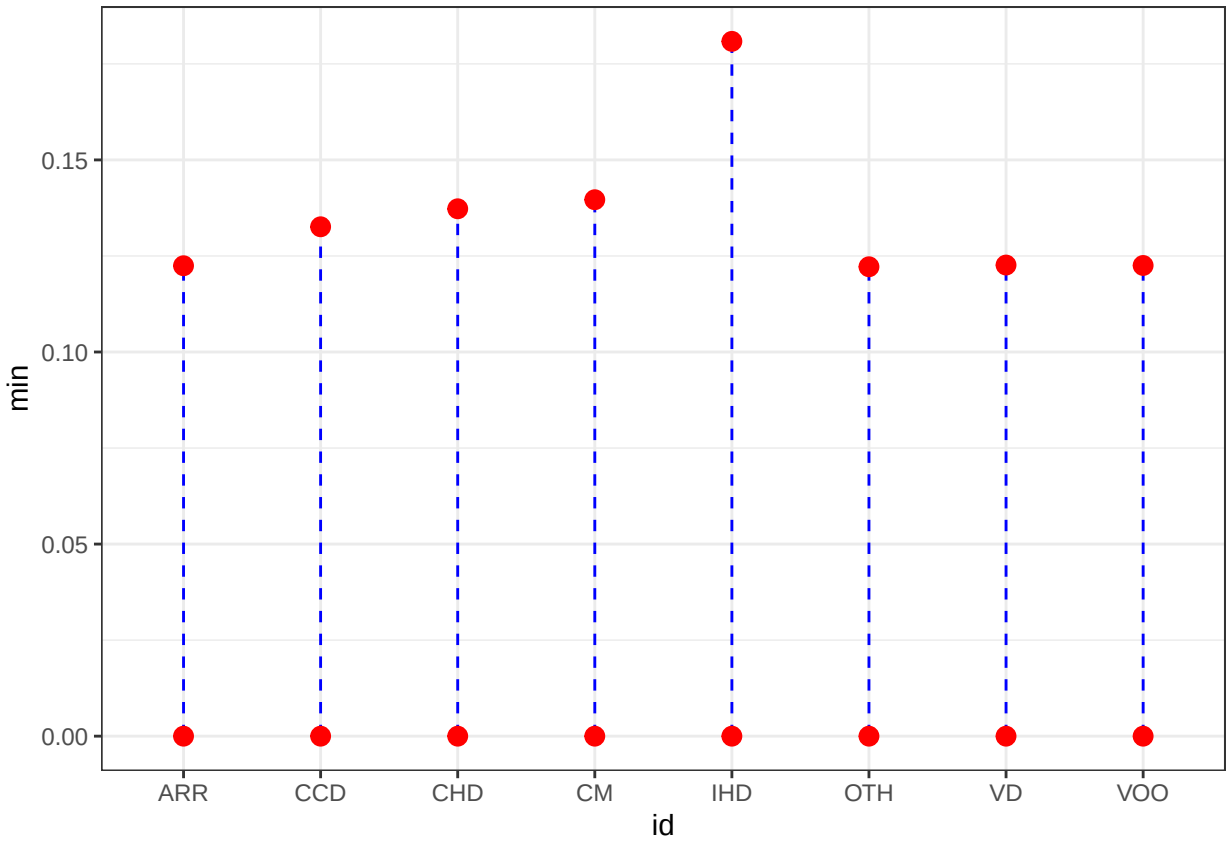



```
ggplot(mouserat, aes(x = OTH)) + geom_histogram(fill = "lightblue", bins = 50)
```



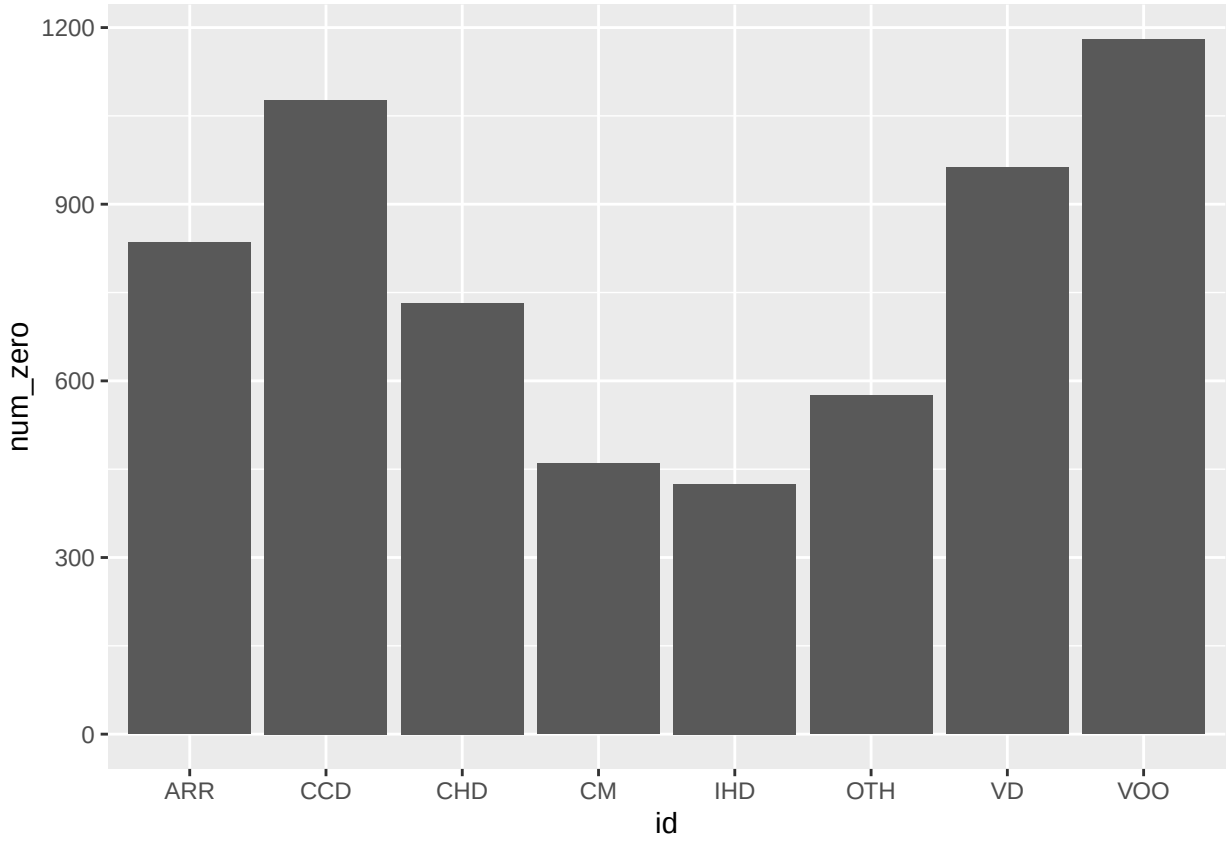
Ranges of CaseOLAP scores by group

```
mouserat_min <- sapply(mouserat[2:9], min)
mouserat_max <- sapply(mouserat[2:9], max)
mouserat_ranges <- data.frame(id=c("IHD", "CM", "ARR","VD", "CHD", "CCD", "VOO", "OTH"),
                              min=mouserat_min, max=mouserat_max)

ggplot(mouserat_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()
```



Number of zeroes in each group

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

ggplot(data = mouserat_zero, aes(x=id, y=num_zero)) + geom_bar(stat="identity")
```



Top 20 Analysis

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

```
##       IHD                CM                ARR                 VD
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.01167   Median :0.01234   Median :0.000000   Median :0.000000
##  Mean   :0.01630   Mean   :0.01578   Mean   :0.009965   Mean   :0.006979
##  3rd Qu.:0.02367   3rd Qu.:0.02429   3rd Qu.:0.015130   3rd Qu.:0.011235
##  Max.   :0.18087   Max.   :0.13964   Max.   :0.122457   Max.   :0.122633
##       CHD                CCD                VOO                OTH
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.005644   Median :0.000000   Median :0.000000   Median :0.00852
##  Mean   :0.010258   Mean   :0.006299   Mean   :0.003941   Mean   :0.01391
##  3rd Qu.:0.015946   3rd Qu.:0.005789   3rd Qu.:0.000000   3rd Qu.:0.02293
##  Max.   :0.137284   Max.   :0.132598   Max.   :0.122521   Max.   :0.12219
```

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

```
t20_mouseratIHD <- mouserat_IHD[1:20, ]$protein

mouserat_CM <- mouserat %>% arrange(desc(CM))
t20_mouseratCM <- mouserat_CM[1:20, ]$protein

mouserat_ARR <- mouserat %>% arrange(desc(ARR))
t20_mouseratARR <- mouserat_ARR[1:20, ]$protein

mouserat_VD <- mouserat %>% arrange(desc(VD))
t20_mouseratVD <- mouserat_VD[1:20, ]$protein

mouserat_CHD <- mouserat %>% arrange(desc(CHD))
t20_mouseratCHD <- mouserat_CHD[1:20, ]$protein

mouserat_CCD <- mouserat %>% arrange(desc(CCD))
t20_mouseratCCD <- mouserat_CCD[1:20, ]$protein

mouserat_VOO <- mouserat %>% arrange(desc(VOO))
t20_mouseratVOO <- mouserat_VOO[1:20, ]$protein

mouserat_OTH <- mouserat %>% arrange(desc(OTH))
t20_mouseratOTH <- mouserat_OTH[1:20, ]$protein

# Find the proteins that appear in more than one top 20 list
Reduce(intersect, list(t20_mouseratIHD, t20_mouseratCM, t20_mouseratARR, t20_mouseratVD,
                       t20_mouseratCHD, t20_mouseratCCD, t20_mouseratVOO, t20_mouseratOTH))
```

```
## [1] "q62052" "o35973" "q9z1m7"
```

```
# Combine top 20 lists into a dataframe
t20_mouserat <- data.frame(t20_mouseratIHD, t20_mouseratCM, t20_mouseratARR, t20_mouseratVD,
                           t20_mouseratCHD, t20_mouseratCCD, t20_mouseratVOO, t20_mouseratOTH)

# Count the number of times each protein appears in the dataframe
sort(table(c(t20_mouseratIHD, t20_mouseratCM, t20_mouseratARR, t20_mouseratVD,
             t20_mouseratCHD, t20_mouseratCCD, t20_mouseratVOO, t20_mouseratOTH)))
```

```
##
## o09161 o35111 o54912 o88775 p11152 p25446 p35561 p35859 p37200 p42859 p49817
##      1      1      1      1      1      1      1      1      1      1      1
## p51111 p51437 p52430 p52631 p70490 p98106 q08369 q63945 q66hs7 q6a051 q8vi04
##      1      1      1      1      1      1      1      1      1      1      1
## q8vii8 q91xj0 q9epb4 q9jlj0 q9quk6 q9z2z6 b0lpn4 e9q401 o54990 p11531 p15383
##      1      1      1      1      1      1      2      2      2      2      2
## p22387 p35235 p55213 p63086 p70677 q01705 q09137 q62230 q91vb4 q92015 q9dbd0
##      2      2      2      2      2      2      2      2      2      2      2
## q9z0u5 o08962 o35219 p15389 p41971 p47820 p97414 p97523 q07969 q3unx5 q61140
##      2      3      3      3      3      3      3      3      3      3      3
## q91zz5 o54754 p32507 q8vhj4 q8bsd5 q9r0c0 q08874 q8vig1 q9jlr5 o35973 q62052
##      3      4      4      4      5      5      6      6      7      8      8
## q9z1m7
##      8
```