Exploring Data

R markdown disadvantages

Back to HTRU2

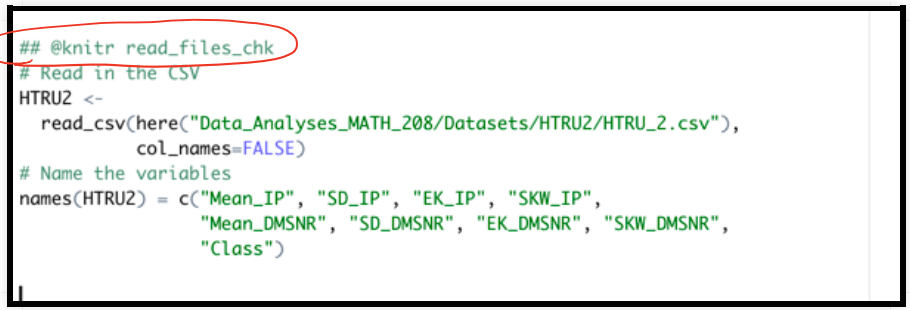
R markdown

```
read_chunk(here("Data_Analyses_MATH_208/Scripts/HTRU2.R"))
```

```
# Read in the CSV
HTRU2 <-
  read_csv(here("Data_Analyses_MATH_208/Datasets/HTRU2/HTRU_2.csv"),
           col_names=FALSE)
# Name the variables
names(HTRU2) = c("Mean_IP", "SD_IP", "EK_IP", "SKW_IP",
                 "Mean_DMSNR", "SD_DMSNR", "EK_DMSNR", "SKW_DMSNR",
                 "Class")
```

The script file

```
include_graphics(here("Documents/knitr_chunk_graphic.png"))
```



```
## @knitr read_files_chk
# Read in the CSV
HTRU2 <-
  read_csv(here("Data_Analyses_MATH_208/Datasets/HTRU2/HTRU_2.csv"),
           col_names=FALSE)
# Name the variables
names(HTRU2) = c("Mean_IP", "SD_IP", "EK_IP", "SKW_IP",
                 "Mean_DMSNR", "SD_DMSNR", "EK_DMSNR", "SKW_DMSNR",
                 "Class")
```

The R markdown file

```
include_graphics(here("Documents/knitr_chunk_graphic2.png"))
```


```
HTRU2 <- HTRU2 %>%
  mutate(Class=factor(ifelse(Class==0,"Negative","Positive")))
```

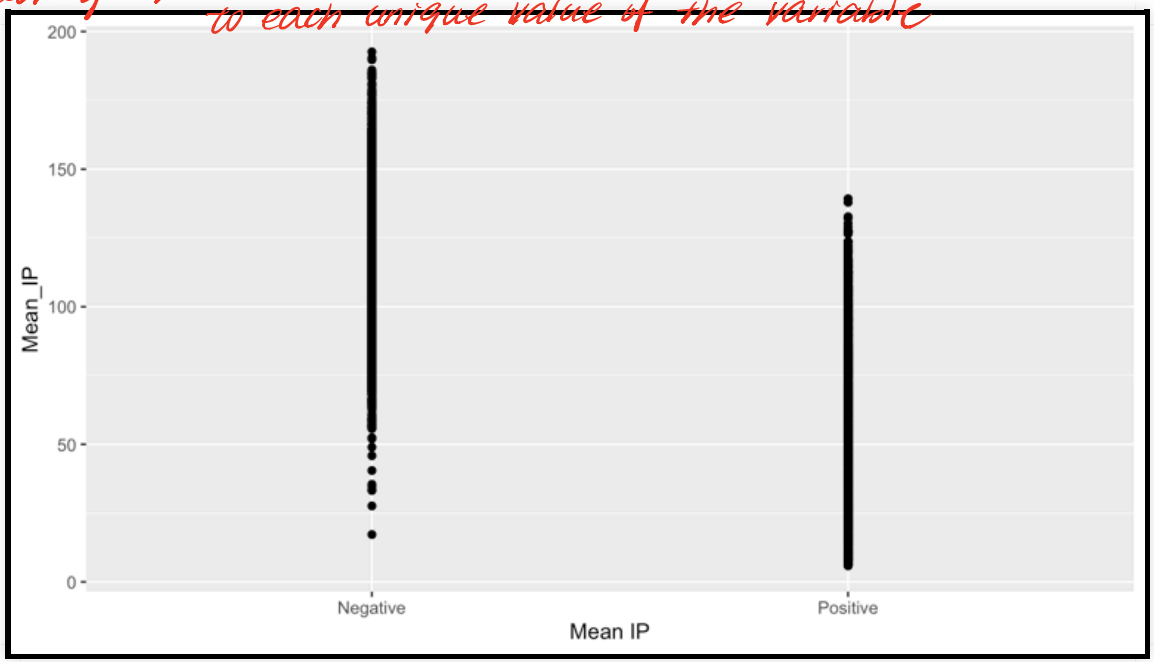
ggplot grammar (drawing with words)

First (bad attempt) Mean IP

aesthetic (a visual property of the objects in your plot)

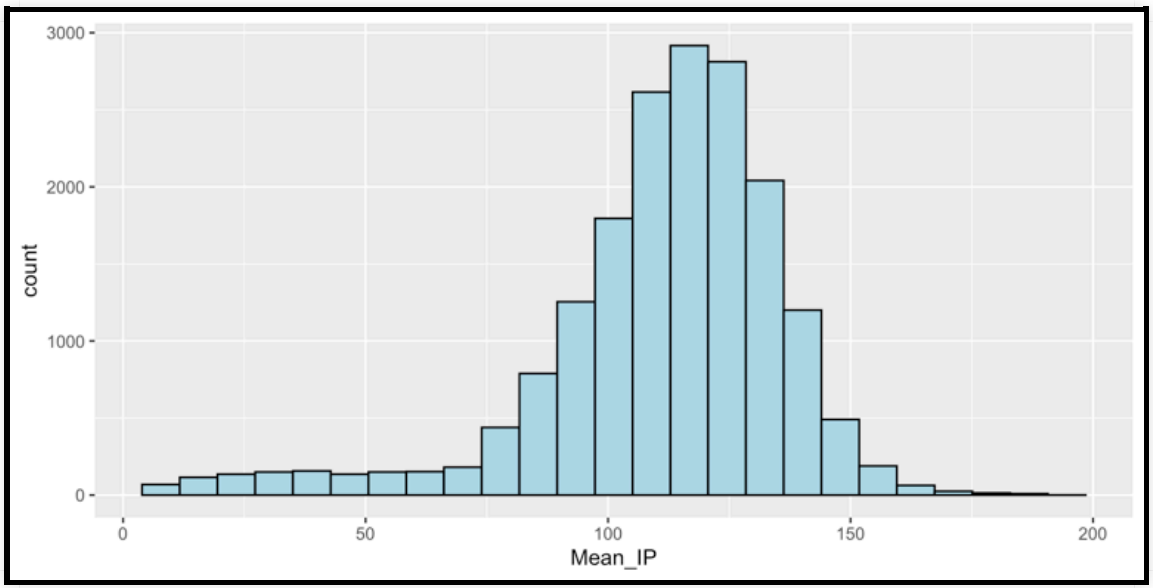
```
ggplot(HTRU2,aes(x=Class,y=Mean_IP)) + geom_point() +
  xlab("Mean IP")
```

scaling => ggplot2 will automatically assign a unique level of the aesthetic to each unique value of the variable



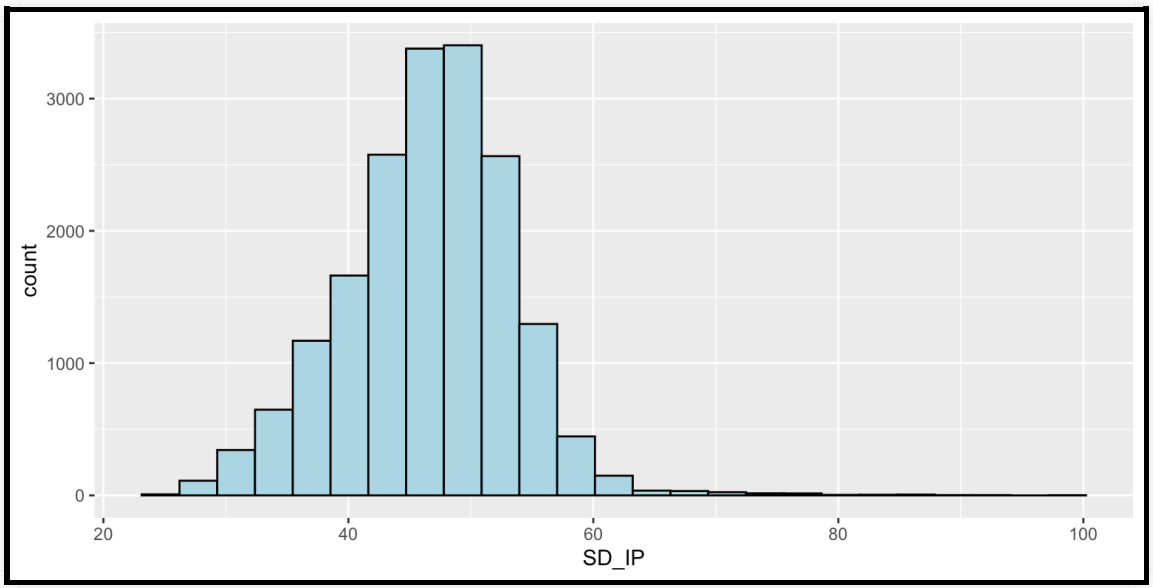
Back to HTRU2: Mean IP

```
ggplot(HTRU2, aes(x=Mean_IP)) +
  geom_histogram(bins=25,col="black",fill="lightblue")
```



Back to HTRU2: SD IP

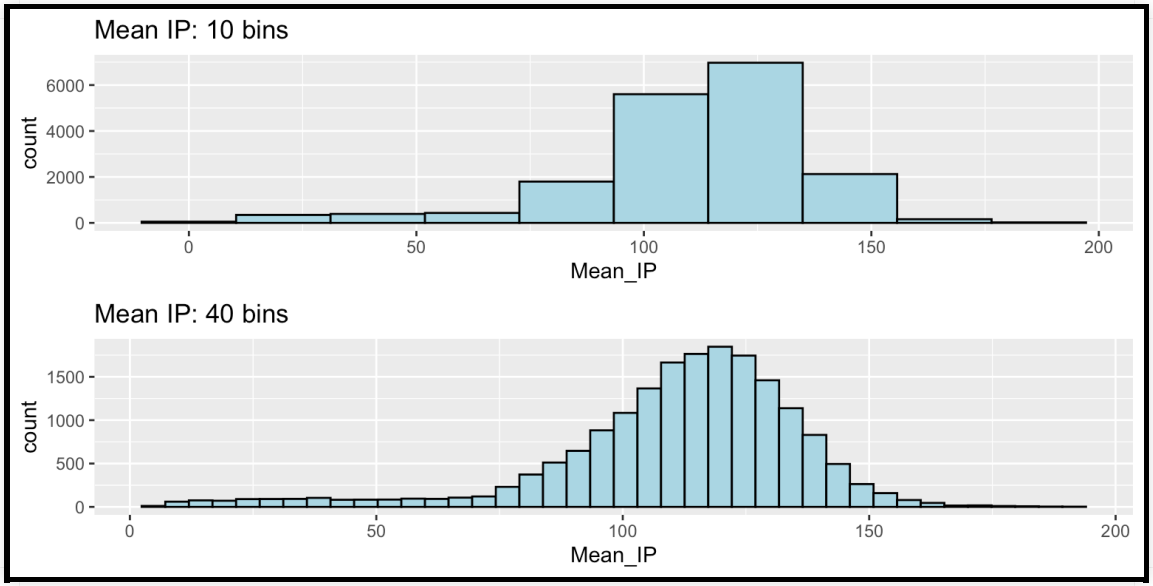
```
ggplot(HTRU2, aes(x=SD_IP)) +
  geom_histogram(bins=25,col="black",fill="lightblue")
```



Back to HTRU2: Changing Bins

```
p1 = ggplot(HTRU2, aes(x=Mean_IP)) +
  geom_histogram(bins=10,col="black",fill="lightblue") +
  ggtitle("Mean IP: 10 bins")
p2 = ggplot(HTRU2, aes(x=Mean_IP)) +
  geom_histogram(bins=40,col="black",fill="lightblue") +
  ggtitle("Mean IP: 40 bins")
grid.arrange(p1,p2)
```

More bins, more bars.


Mean IP: 10 bins
count
6000
4000
2000
0
0
50
100
150
200
Mean_IP
Mean IP: 40 bins
count
1500
1000
500
0
0
50
100
150
200
Mean_IP

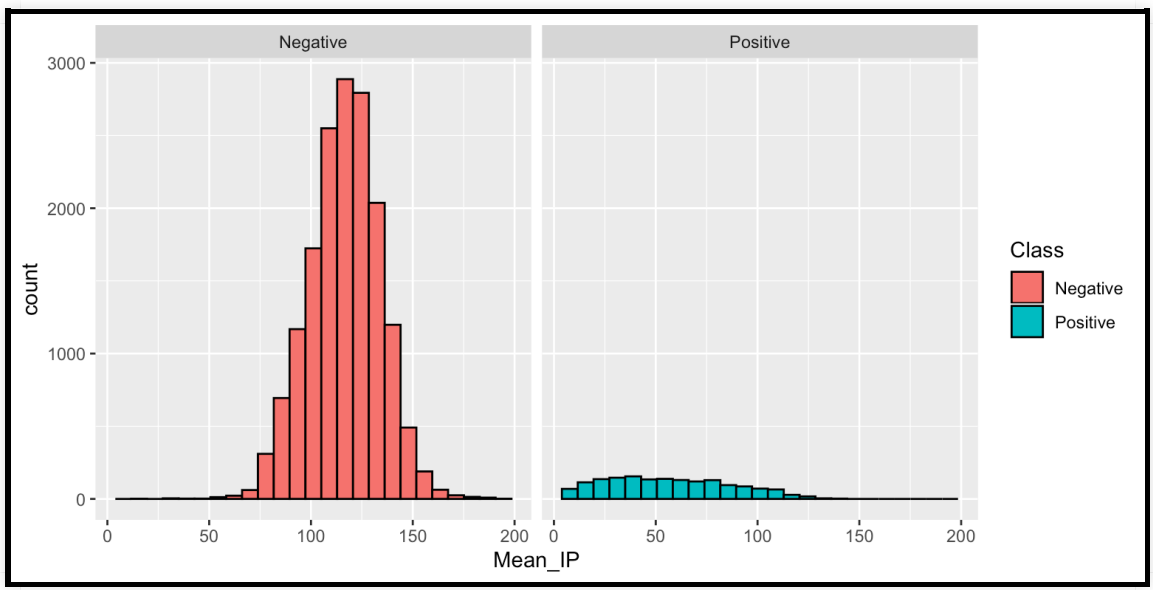
Adding additional information by faceting

split your plot into facets

subplots that each display one subset of the data.

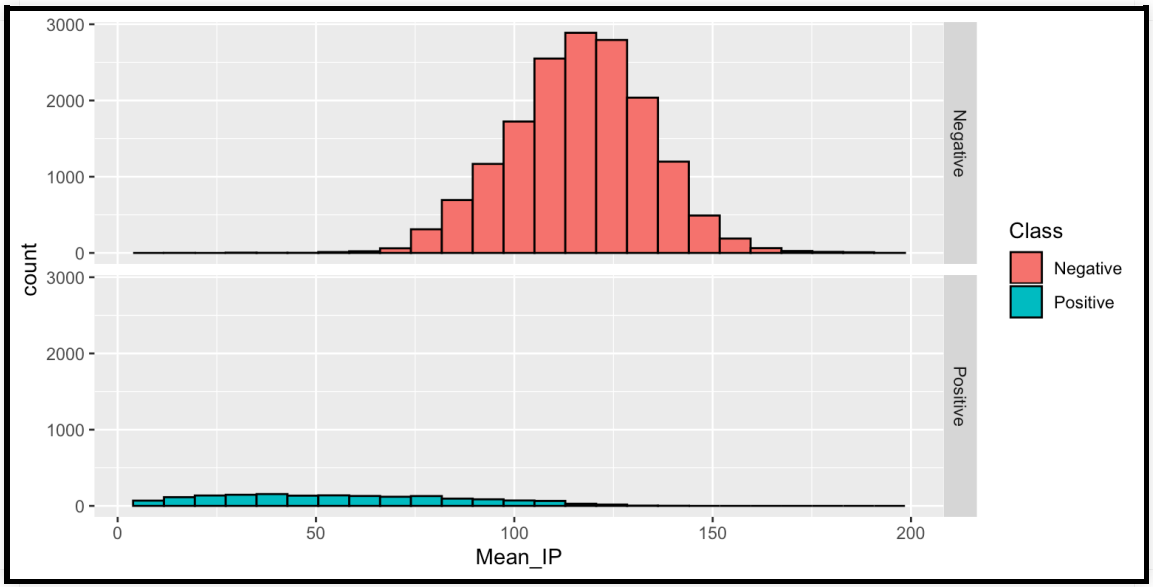
```
ggplot(HTRU2, aes(x=Mean_IP,group=Class,fill=Class)) +
  geom_histogram(bins=25,col="black") +
    facet_wrap(~Class)
```

variable name (discrete)



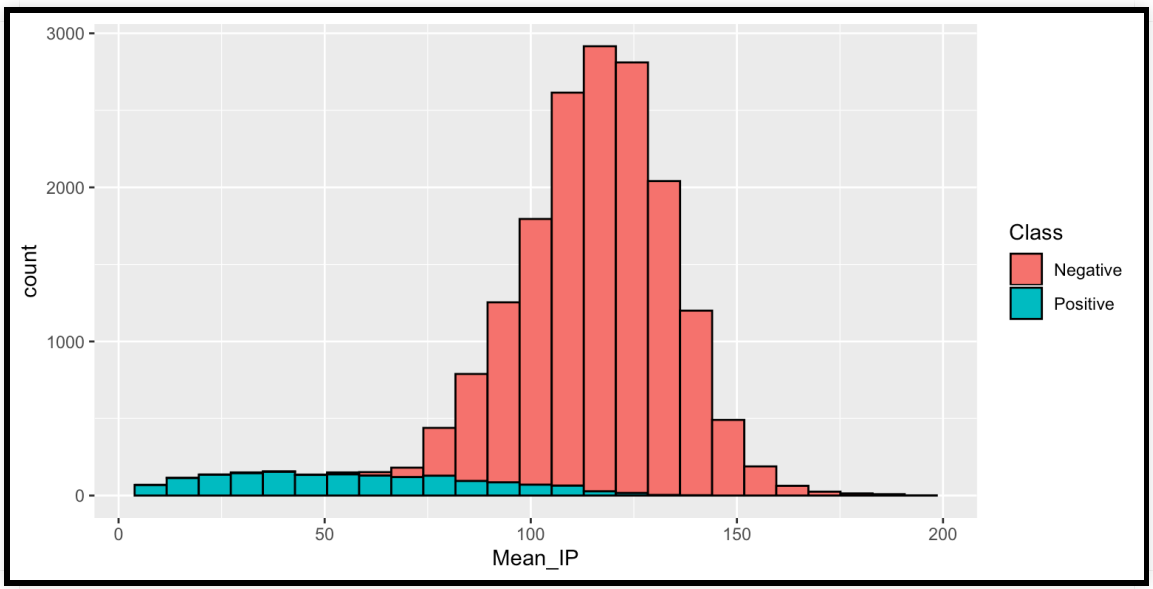
Adding additional information by faceting

```
ggplot(HTRU2, aes(x=Mean_IP,group=Class,fill=Class)) +
  geom_histogram(bins=25,col="black") +
  facet_grid(rows=vars(Class))
```



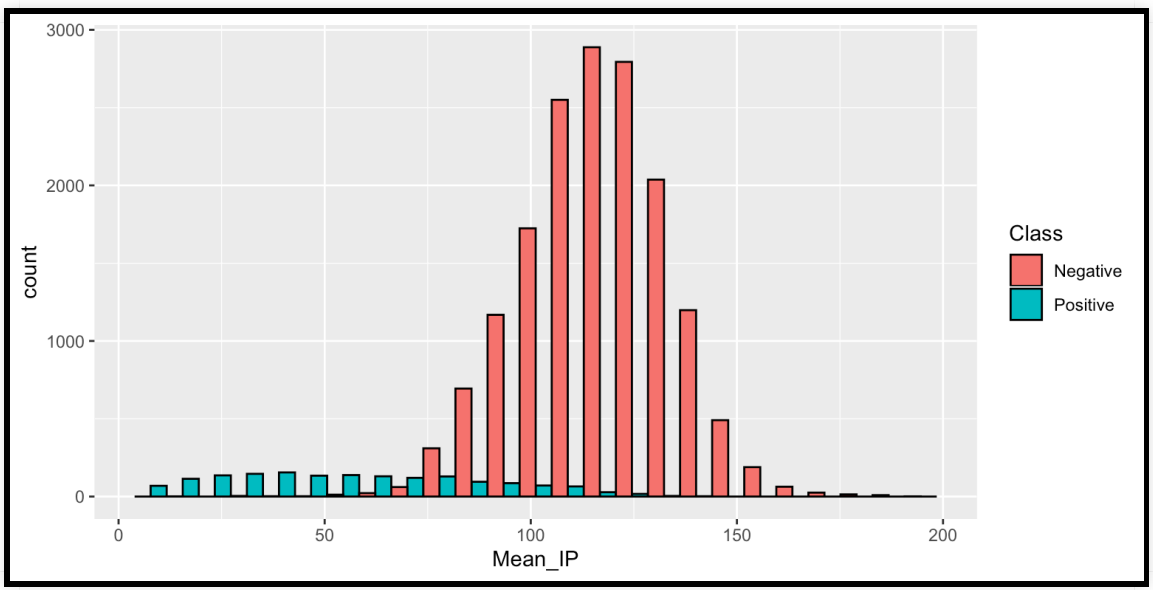
Adding additional information within the same plot

```
ggplot(HTRU2, aes(x=Mean_IP,group=Class,fill=Class)) +
  geom_histogram(bins=25,col="black")
```



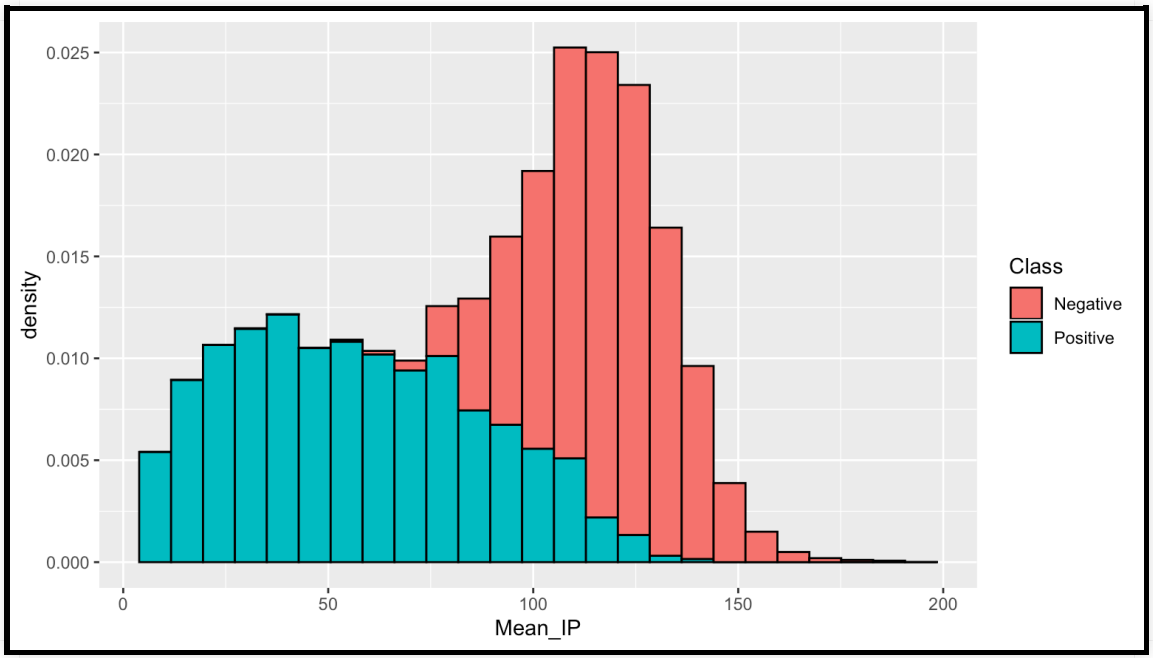
Adding additional information within the same plot

```
ggplot(HTRU2, aes(x=Mean_IP,group=Class,fill=Class)) +
  geom_histogram(bins=25,col="black",position="dodge")
```



Probability histograms

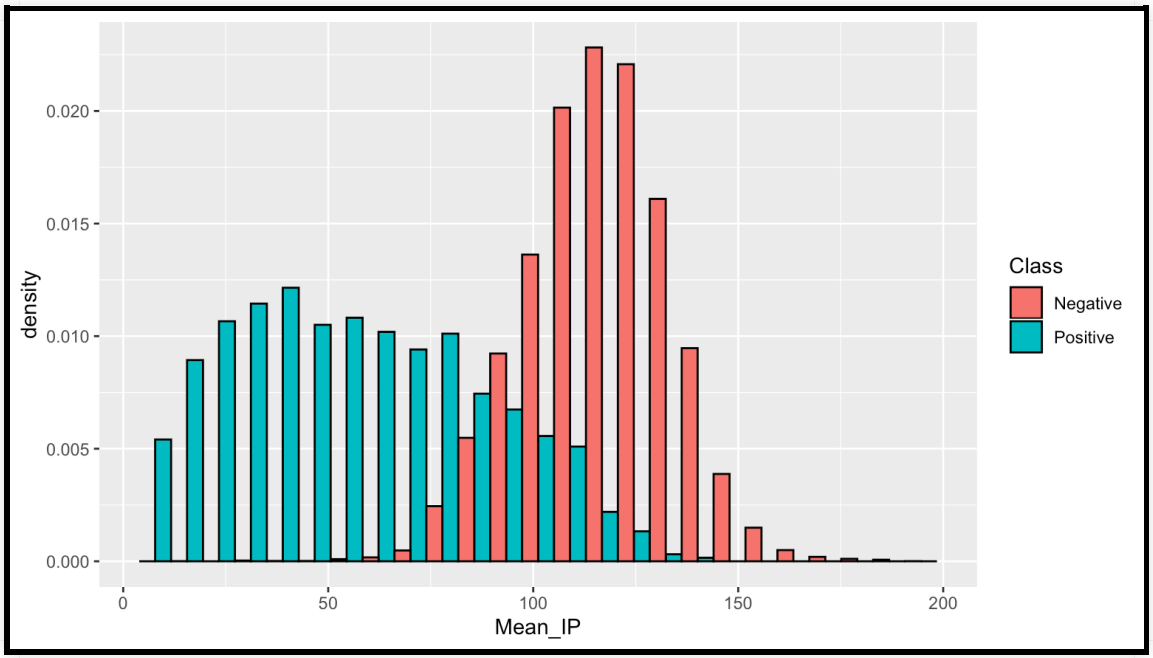
```
ggplot(HTRU2,aes(x=Mean_IP,group=Class,fill=Class)) +
  geom_histogram(aes(y=..density..),bins=25,col="black")
```



Probability histograms

```
ggplot(HTRU2,aes(x=Mean_IP,group=Class,fill=Class)) +

        geom_histogram(aes(y=..density..),bins=25,col="black",position="
```

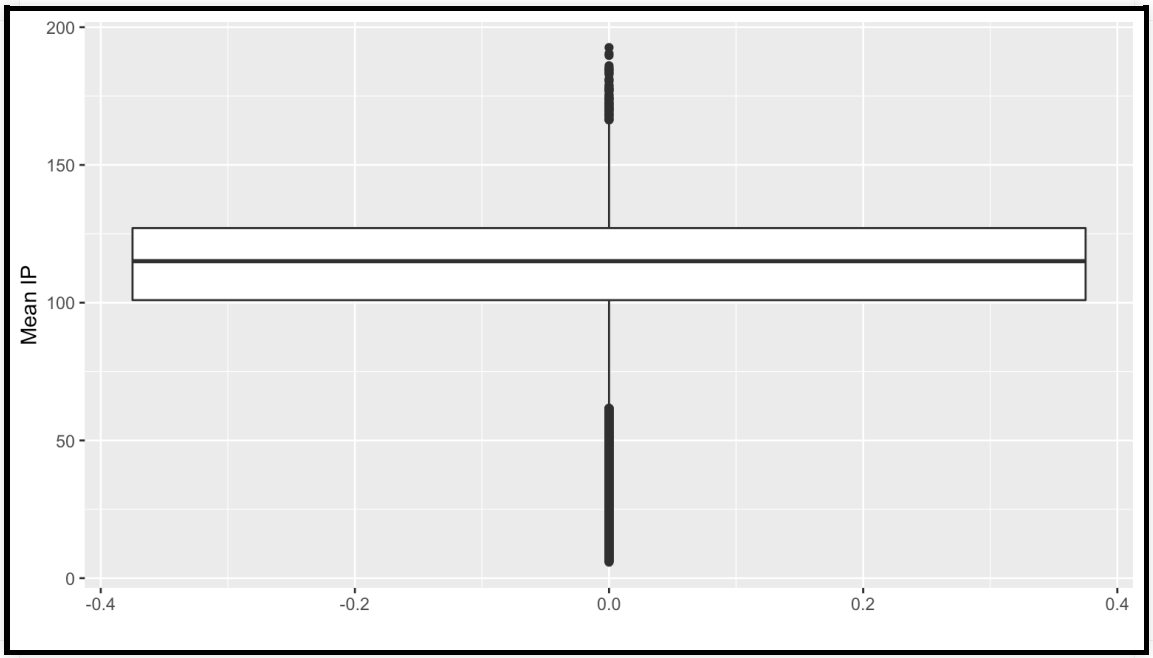


Boxplots

Boxplots: Overview

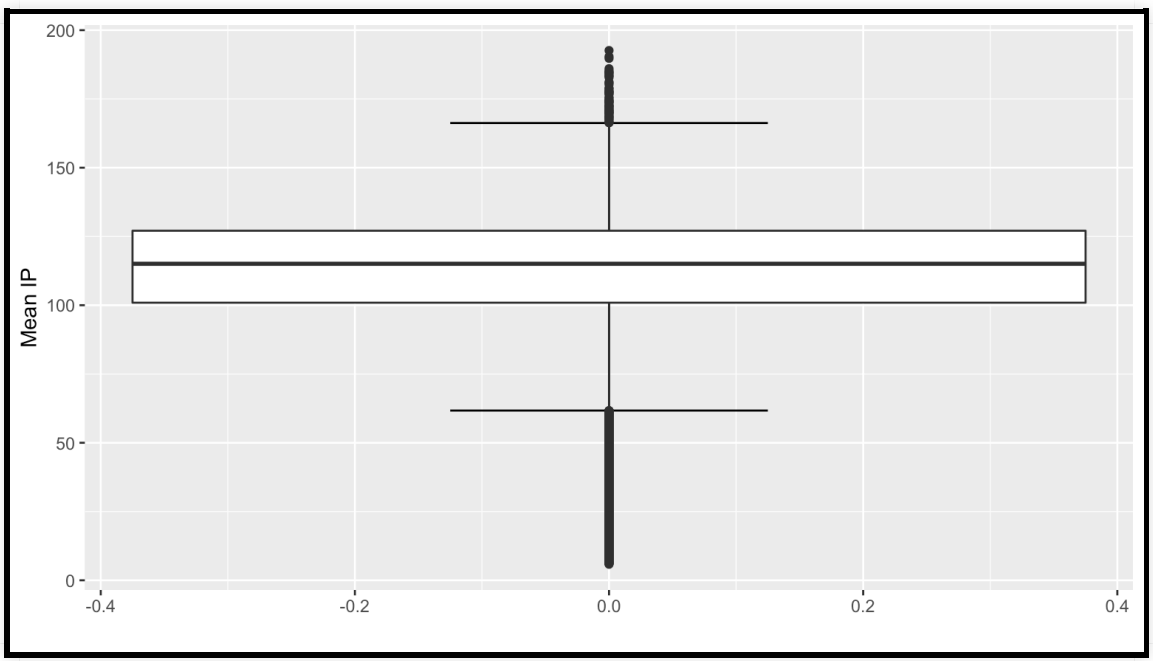
Boxplots

```
ggplot(HTRU2,aes(x=NULL,y=Mean_IP)) + geom_boxplot() +
  xlab("") + ylab("Mean IP")
```



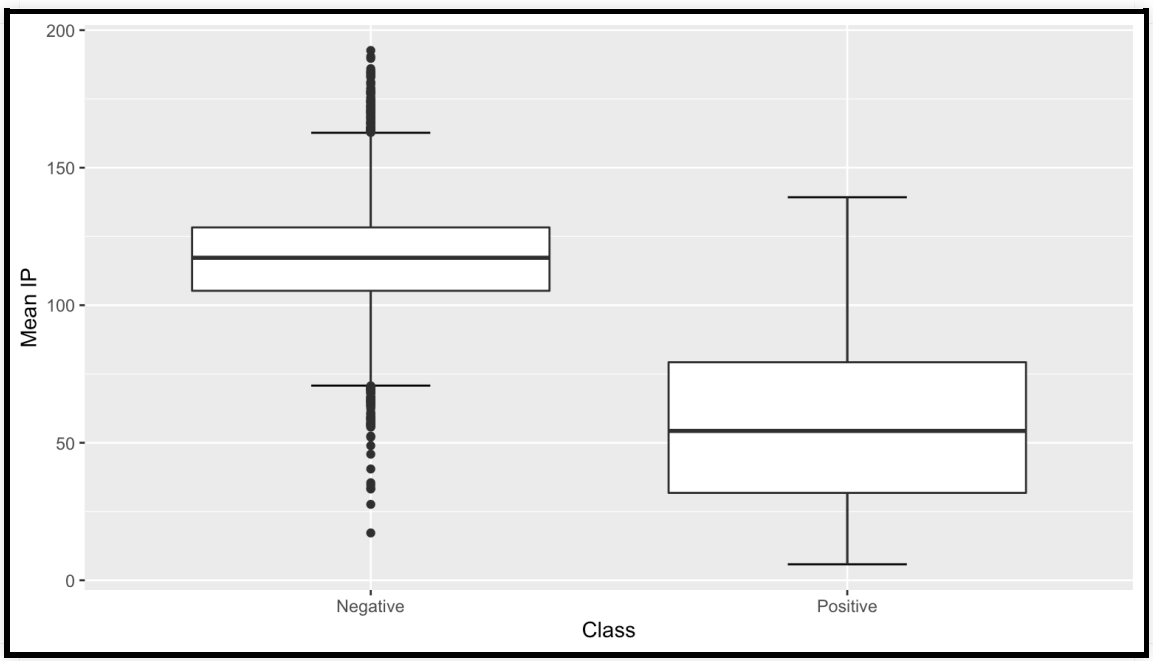
Boxplots

```
ggplot(HTRU2,aes(x=NULL,y=Mean_IP)) +
  stat_boxplot(geom="errorbar",width=0.25) + geom_boxplot() +
  xlab("") + ylab("Mean IP")
```



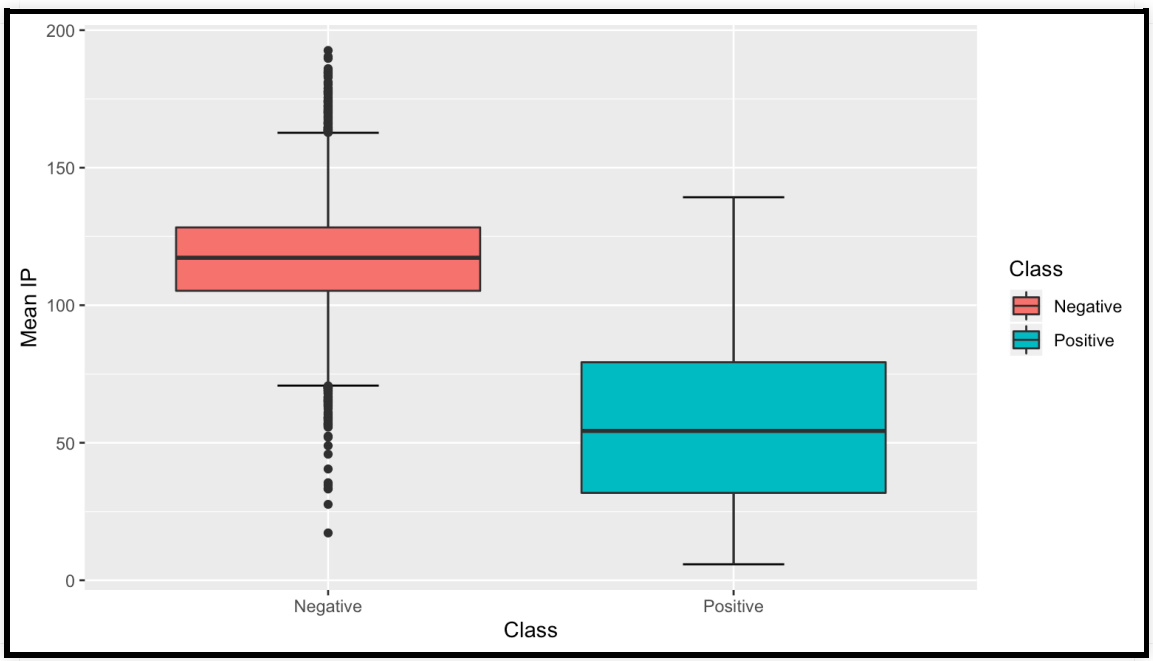
Boxplots

```
ggplot(HTRU2,aes(x=Class,y=Mean_IP)) +
  stat_boxplot(geom="errorbar",width=0.25) + geom_boxplot() +
  ylab("Mean IP")
```



Boxplots

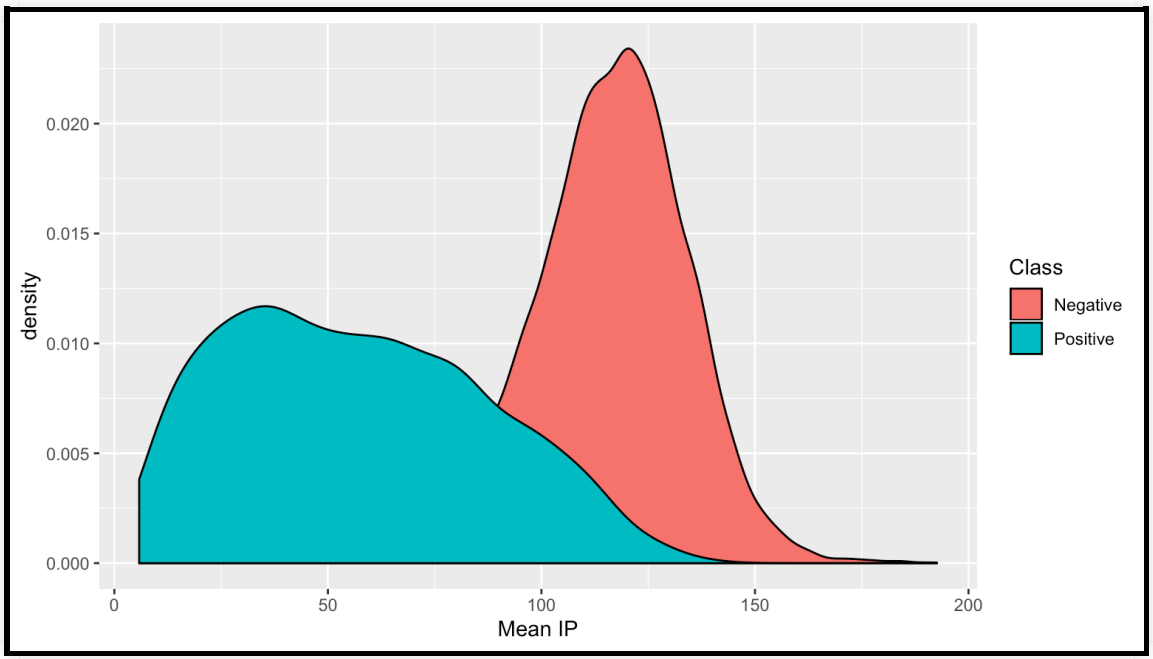
```
ggplot(HTRU2,aes(x=Class,y=Mean_IP,fill=Class)) +
  stat_boxplot(geom="errorbar",width=0.25) + geom_boxplot() +
  ylab("Mean IP")
```



Other plots

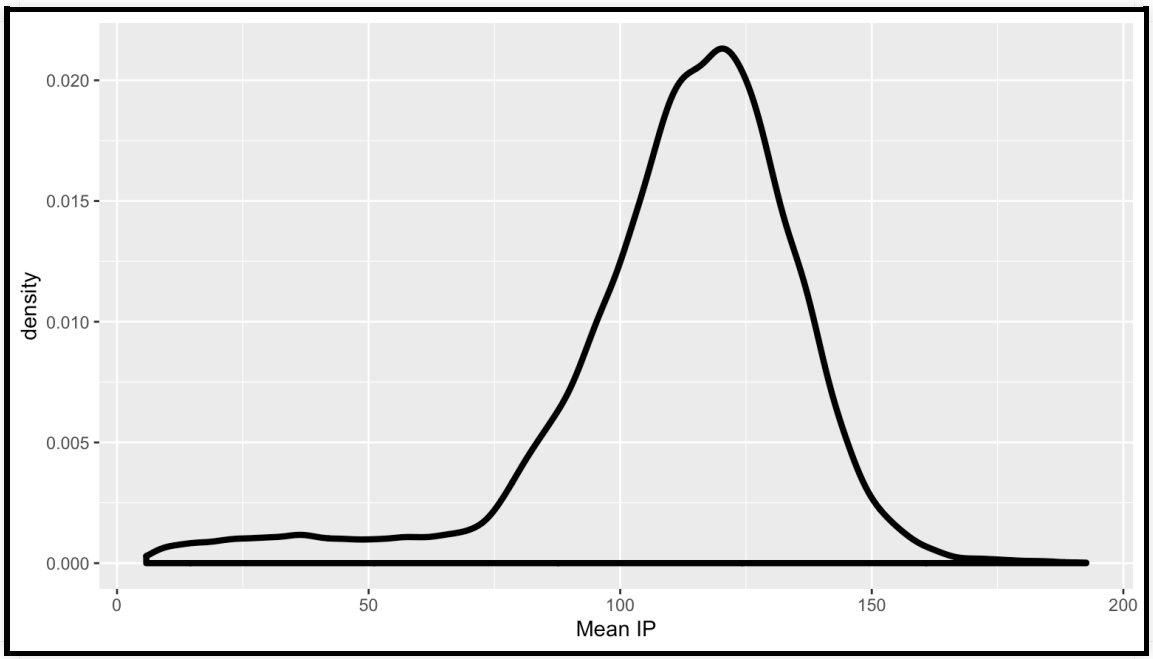
Density plots

```
ggplot(HTRU2,aes(x=Mean_IP,fill=Class)) +
  geom_density()+ xlab("Mean IP")
```



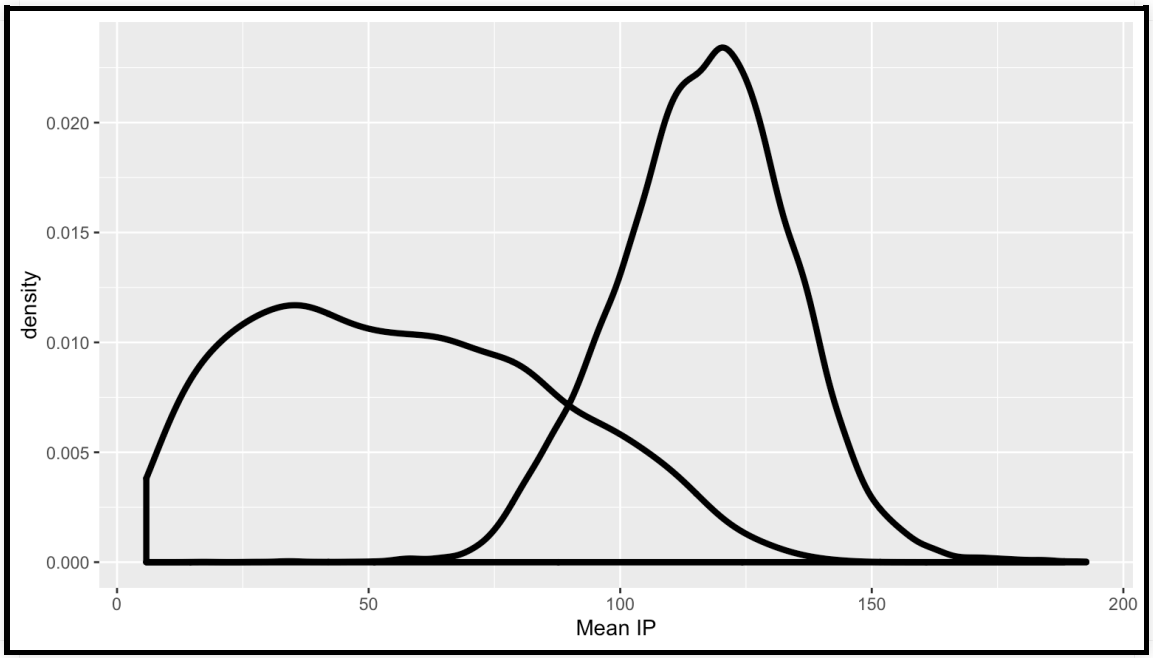
Density plots

```
ggplot(HTRU2,aes(x=Mean_IP)) +
  geom_density(size=1.5) + xlab("Mean IP")
```



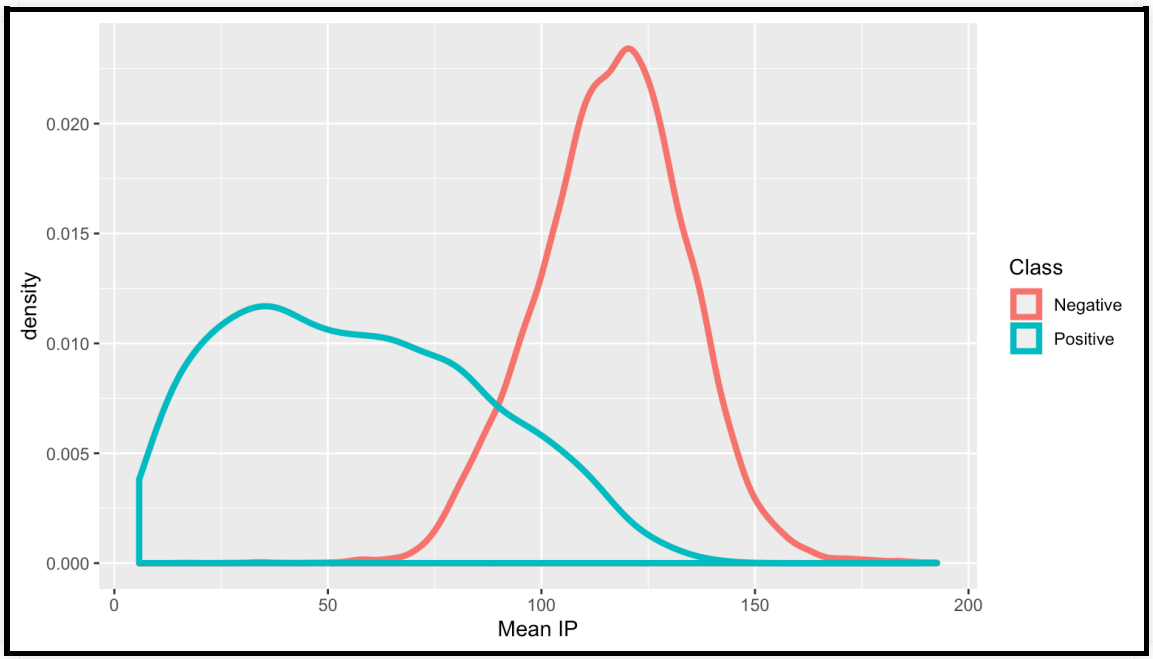
Density plots

```
ggplot(HTRU2,aes(x=Mean_IP,group=Class)) +
  geom_density(size=1.5) + xlab("Mean IP")
```



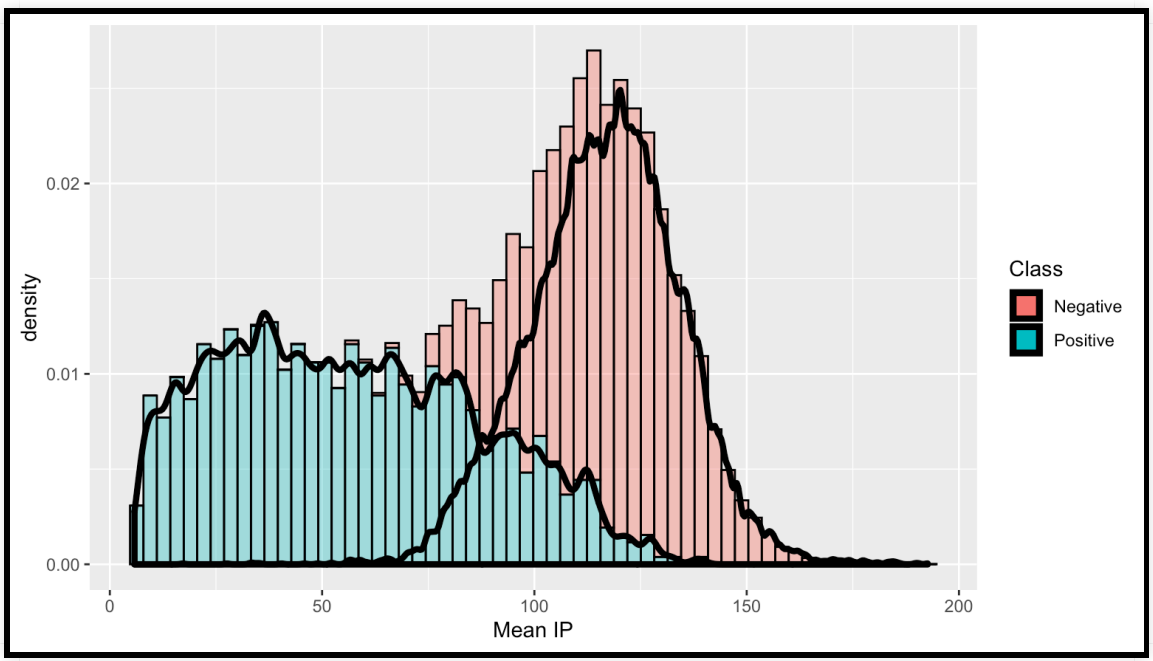
Density plots

```
ggplot(HTRU2,aes(x=Mean_IP,col=Class)) +
  geom_density(size=1.5) + xlab("Mean IP")
```



Density plots

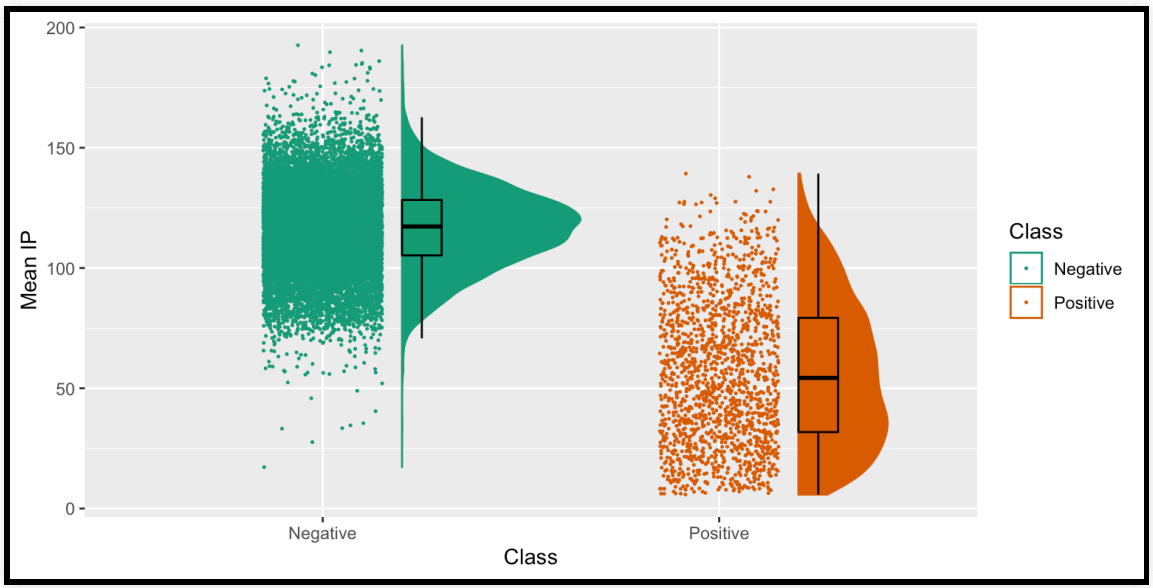
```
ggplot(HTRU2,aes(x=Mean_IP,group=Class)) +
  geom_histogram(aes(y=..density..,fill=Class),
                 col="black",alpha=0.4,bins=60)+
  geom_density(size=1.5,adjust=0.25) + xlab("Mean IP")
```



Raincloud plots

```
library(cowplot)
source(here("Documents/rain_cloud.R"))
ggplot(HTRU2,aes(x=Class,y=Mean_IP,fill=Class,col=Class)) +
  geom_flat_violin(position=position_nudge(x=0.2,y=0),adjust=1) +
  #note that here we need to set the x-variable to a numeric variable
  #and  bump it to get the boxplots to line up with the rainclouds.
  geom_boxplot(aes(x = as.numeric(Class)+0.25, y = Mean_IP),
               outlier.shape = NA, alpha = 0.3, width = .1,
               colour = "BLACK") +
  geom_point(position=position_jitter(width=0.15),size=0.25) +
  xlab("Class") + ylab("Mean IP") +
  guides(fill = FALSE)+
  scale_colour_brewer(palette = "Dark2")+
  scale_fill_brewer(palette = "Dark2")
```

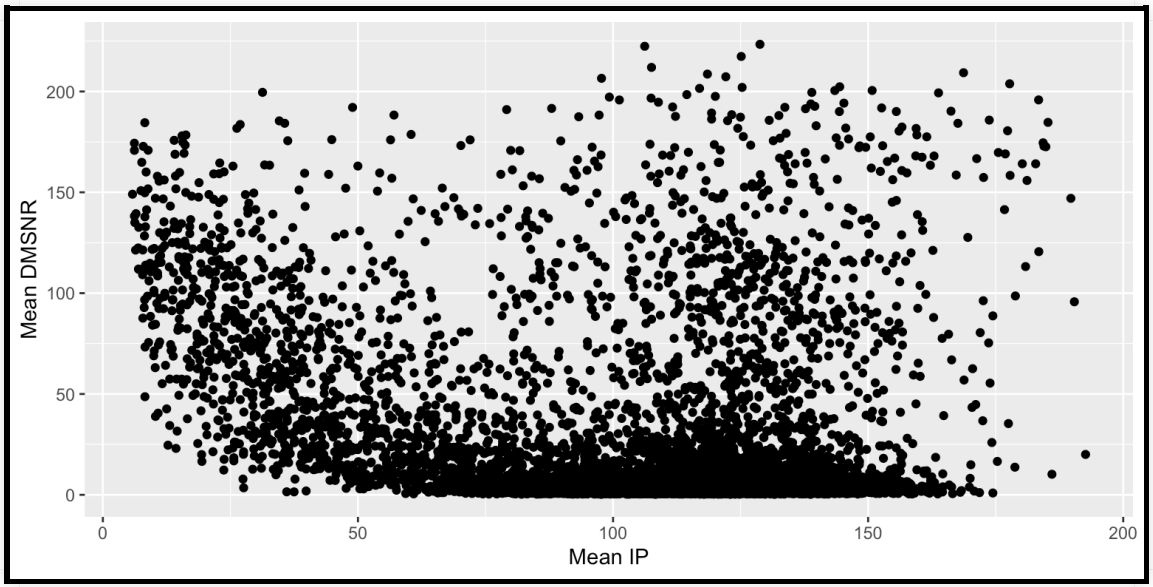
Raincloud plots



Bivariate quantitative measures

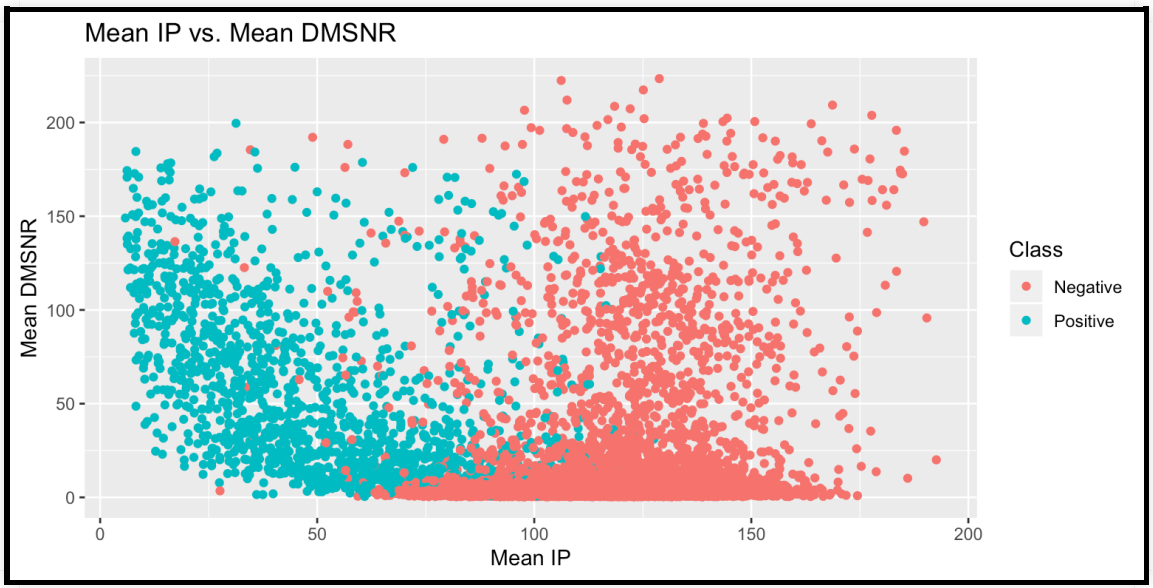
Scatterplots

```
ggplot(HTRU2,aes(x=Mean_IP,y=Mean_DMSNR)) + geom_point() +
  xlab("Mean IP") + ylab("Mean DMSNR")
```



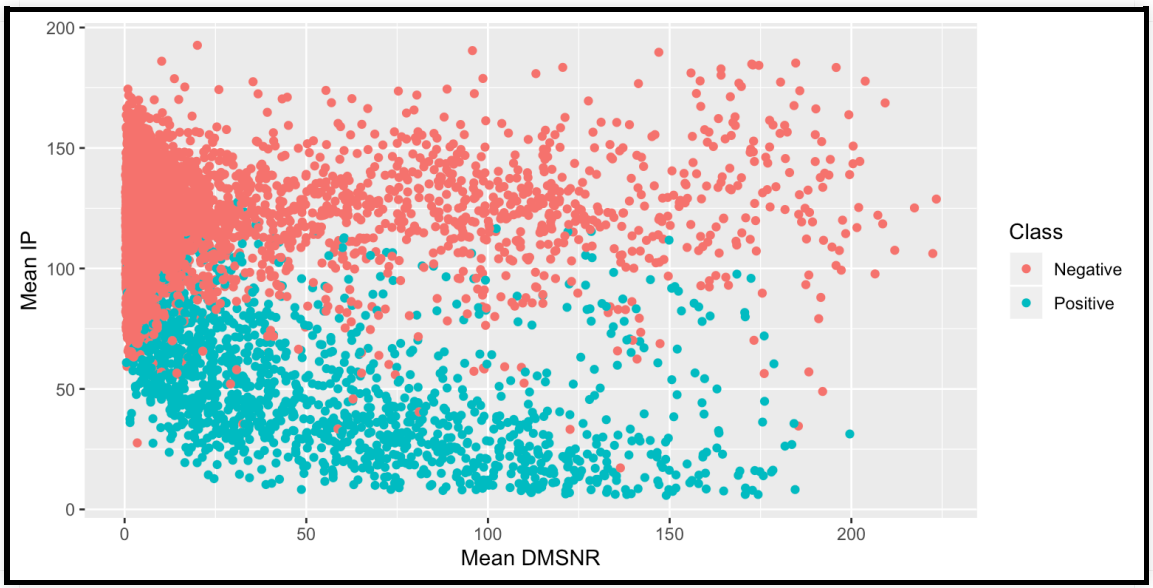
Scatterplots with class info

```
ggplot(HTRU2,aes(x=Mean_IP,y=Mean_DMSNR,col=Class)) + geom_point() +
  labs(x="Mean IP", y="Mean DMSNR", title="Mean IP vs. Mean DMSNR")
```



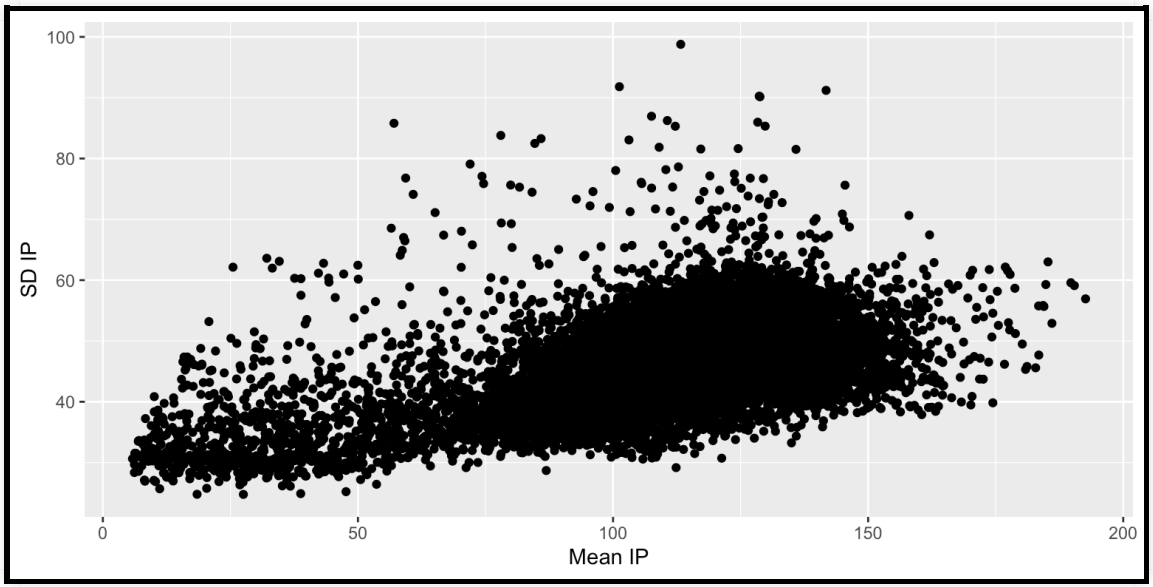
Scatterplots with class info

```
ggplot(HTRU2,aes(x=Mean_IP,y=Mean_DMSNR,col=Class)) + geom_point() +
  labs(x="Mean IP",y="Mean DMSNR")+    coord_flip()
```



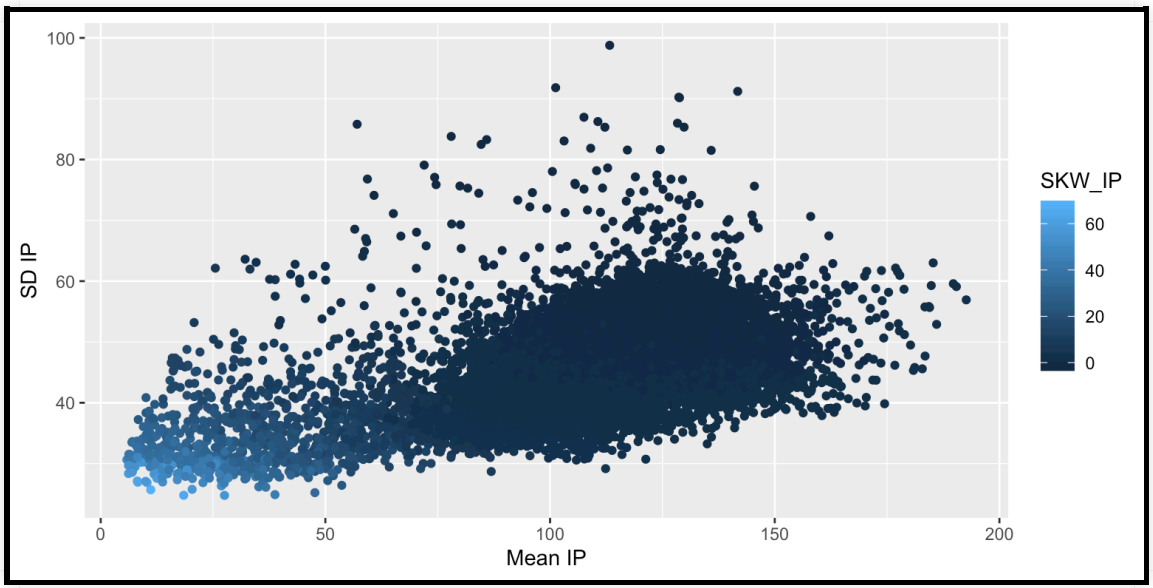
Scatterplots with SD info

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) + geom_point() +
  labs(x="Mean IP",y="SD IP")
```



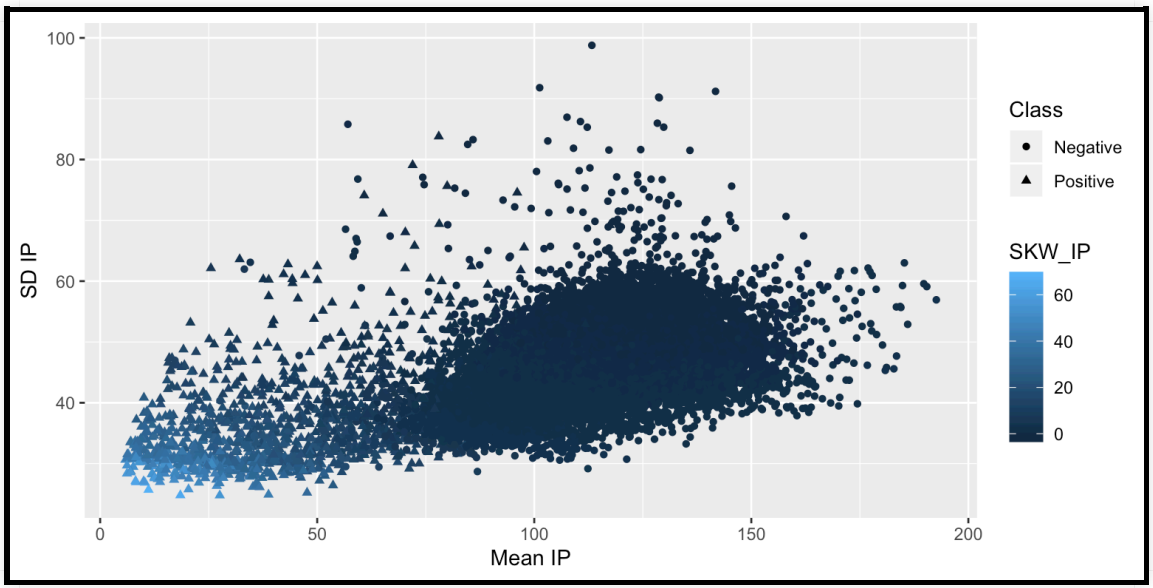
Scatterplots with SD and Skew info

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP,col=SKW_IP)) + geom_point() +
  labs(x="Mean IP",y="SD IP")
```



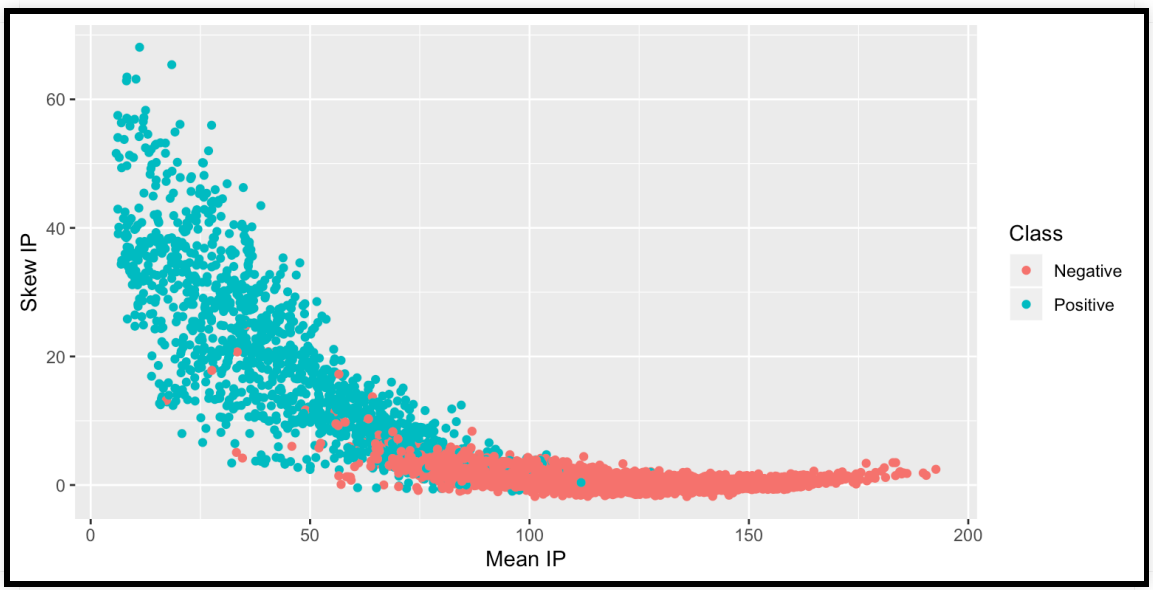
Scatterplots with SD and Skew AND class info

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP,col=SKW_IP, shape=Class)) +
  geom_point() + labs(x="Mean IP",y="SD IP")
```



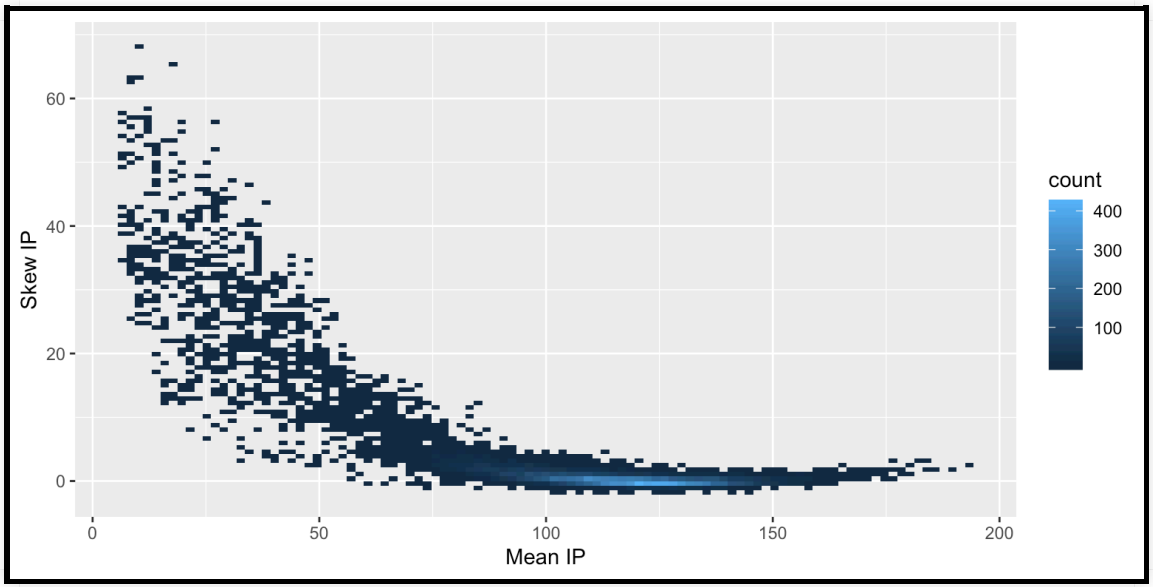
Scatterplots with Mean and Skew AND class info

```
ggplot(HTRU2,aes(x=Mean_IP,y=SKW_IP, col=Class)) +
  geom_point() + labs(x="Mean IP",y="Skew IP")
```



Two-dimensional histograms

```
ggplot(HTRU2,aes(x=Mean_IP,y=SKW_IP)) + geom_bin2d(bins=100) +
        labs(x="Mean IP",y="Skew IP")
```



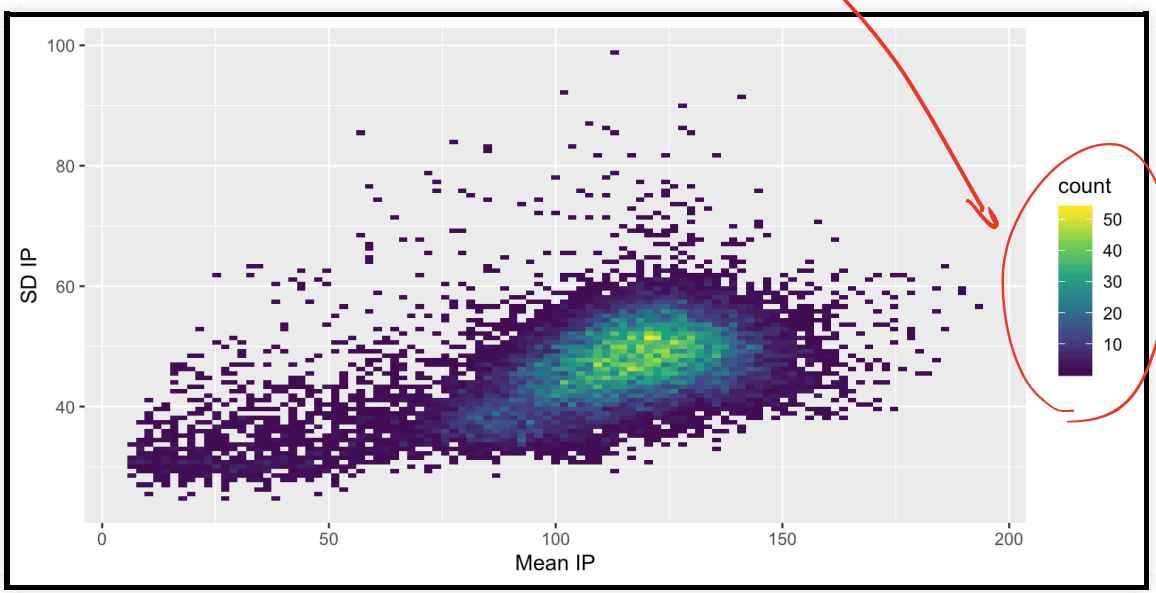
Two-dimensional histograms

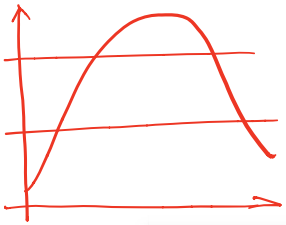
aesthetic

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) + geom_bin2d(bins=100) +
  scale_fill_continuous(type = "viridis") +
  labs(x="Mean IP",y="SD IP")
```

change state

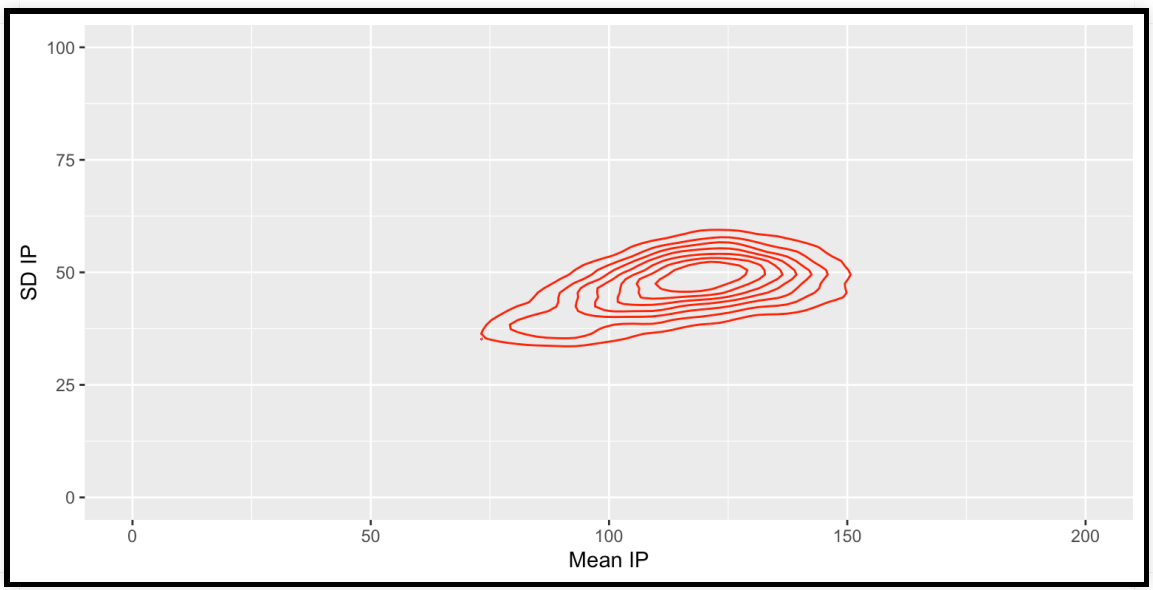
numeric





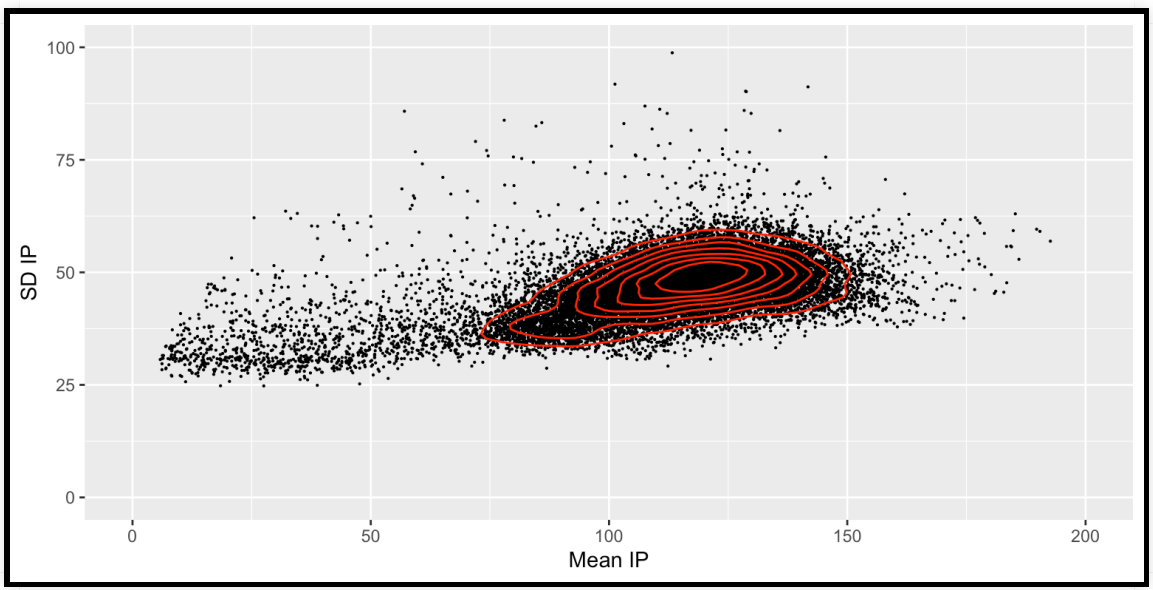
Two-dimensional contour plots

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) +
  geom_density_2d(col="red") + labs(x="Mean IP",y="SD IP") +
  ylim(c(0,100)) +
  xlim(c(0,200))
```



Two-dimensional contour plots

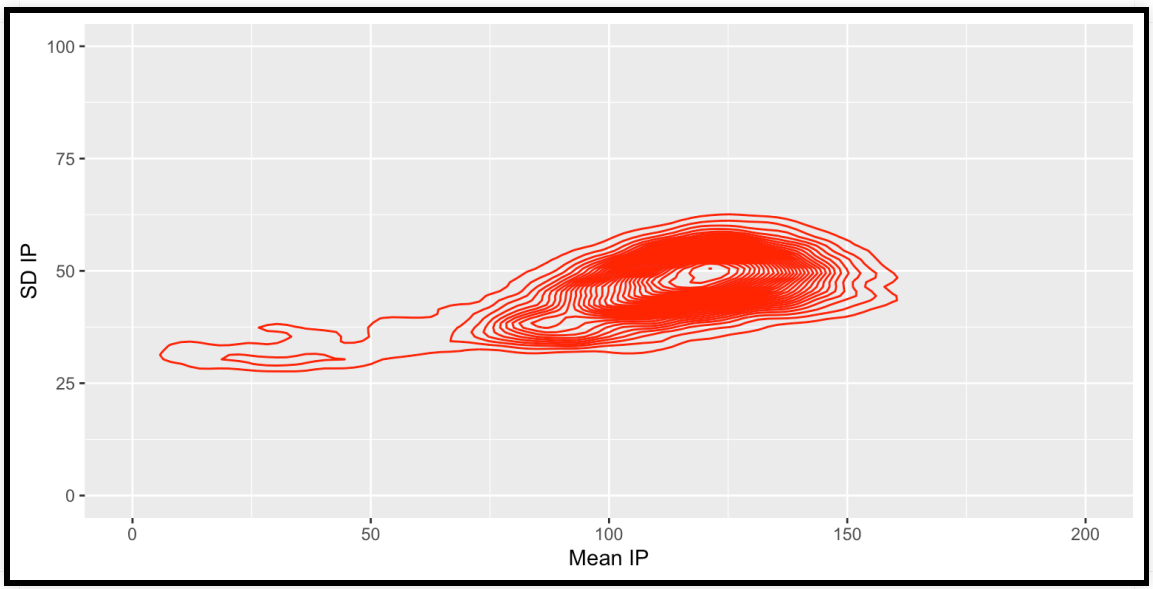
```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) + geom_point(size=0.1) +
  geom_density_2d(col="red") + labs(x="Mean IP",y="SD IP") +
  ylim(c(0,100)) +
  xlim(c(0,200))
```



Two-dimensional contour plots

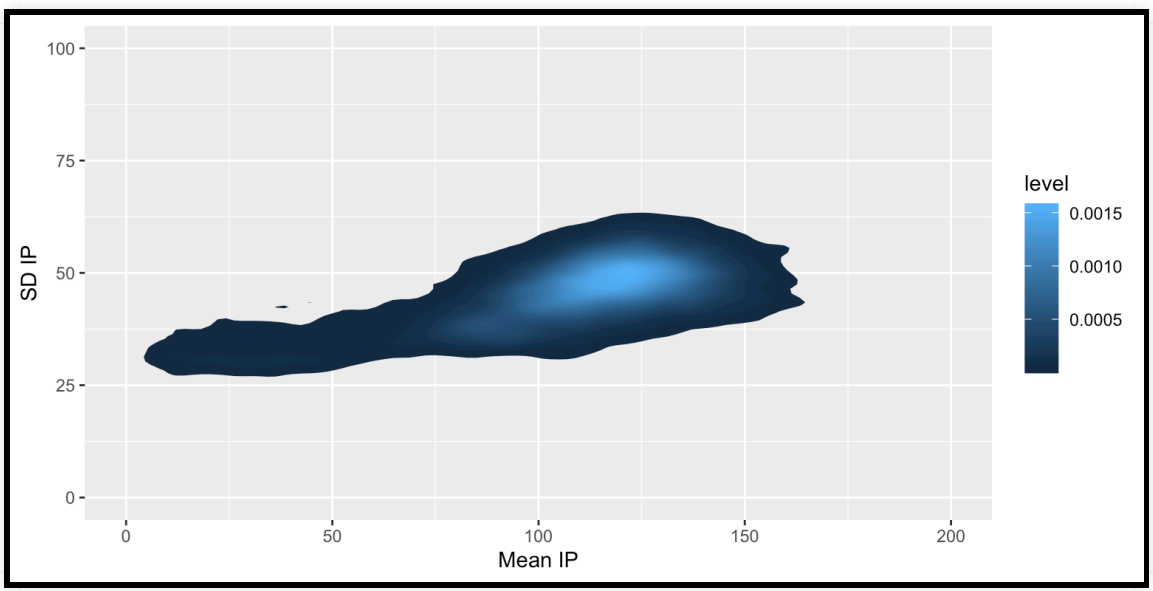
```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) +
  geom_density_2d(col="red",bins=30) +
  labs(x="Mean IP",y="SD IP")+
  ylim(c(0,100)) +
  xlim(c(0,200))
```

increase contours we use



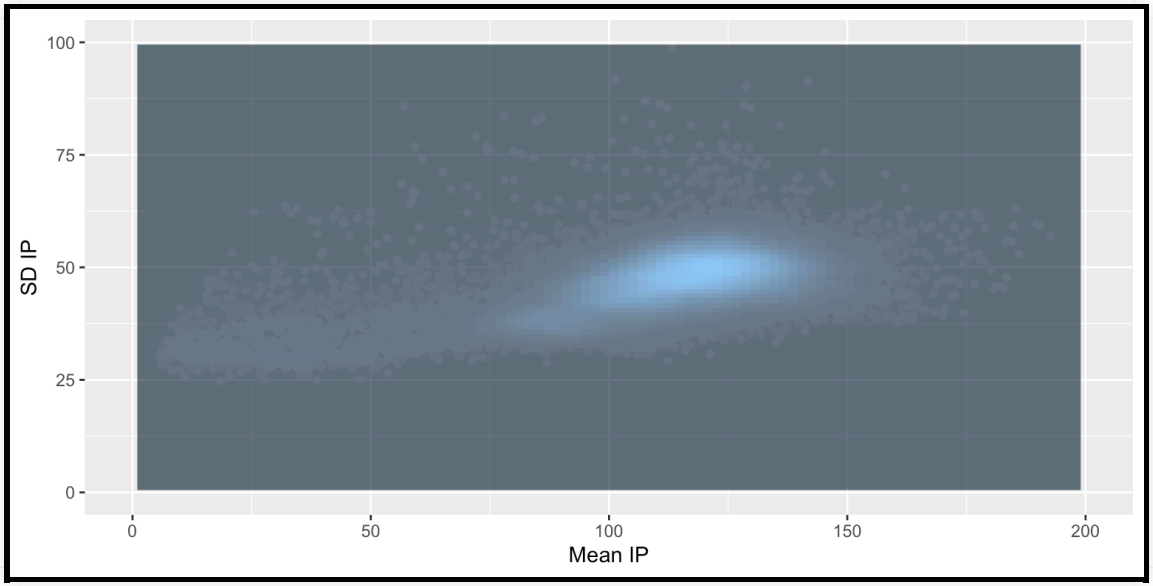
Two-dimensional contour plots

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) +
  stat_density_2d(bins=50,aes(fill = ..level..),
                  geom = "polygon") +
  labs(x="Mean IP",y="SD IP")+
  ylim(c(0,100)) +
  xlim(c(0,200))
```



Two-dimensional contour plots

```
ggplot(HTRU2,aes(x=Mean_IP,y=SD_IP)) + geom_point(col="white")+
  stat_density_2d(bins=50,alpha=0.7,aes(fill = ..density..),
                  geom = "raster", contour = FALSE) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0))+
  theme(legend.position='none')+
  labs(x="Mean IP",y="SD IP")+
  ylim(c(0,100)) +
  xlim(c(0,200))
```

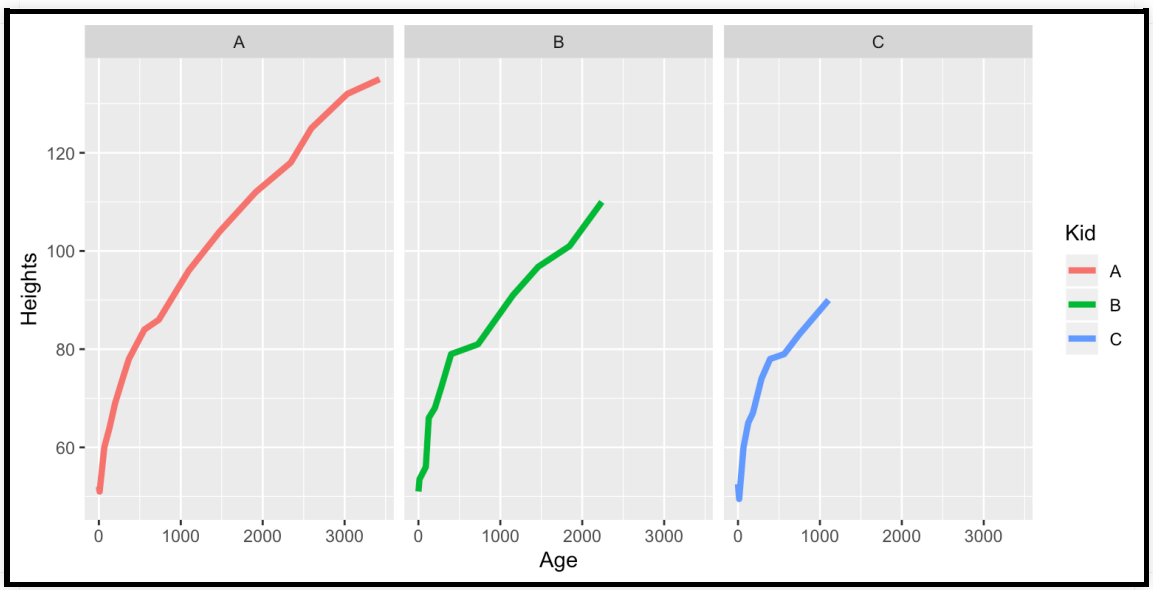

100
75
50
25
0
SD IP
0
50
100
150
200
Mean IP

Plotting time series data

Back to the kids data

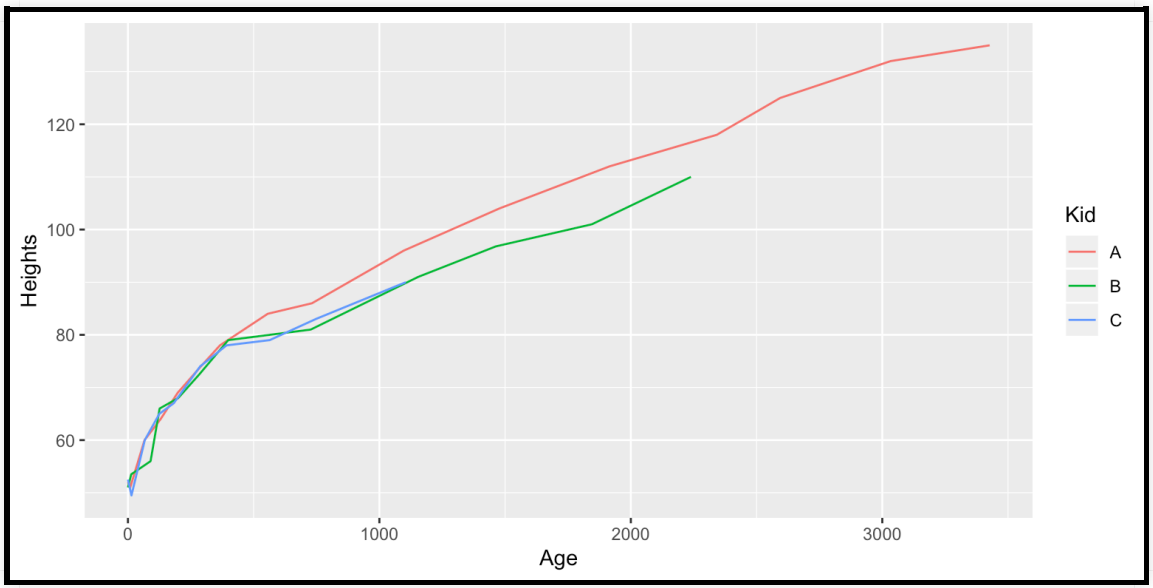
Using faceting

```
ggplot(kids_heights_tbl_grp,aes(x=Age,y=Heights,group=Kid,col=Kid)) +
  geom_line(size=1.5) +
  facet_wrap(~Kid)
```



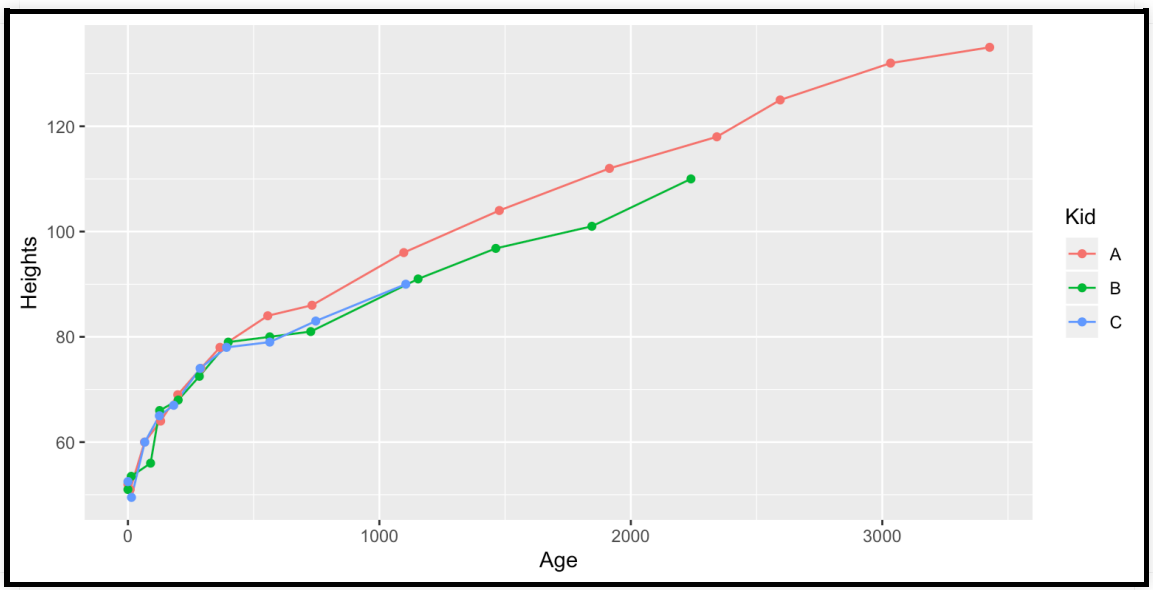
All together

```
ggplot(kids_heights_tbl_grp,aes(x=Age,Heights,group=Kid,col=Kid)) +
  geom_line()
```



Adding points

```
ggplot(kids_heights_tbl_grp,aes(x=Age,Heights,group=Kid,col=Kid)) +
  geom_line() + geom_point()
```



Using the tstibble packgae

```
library(tsbox) # Need to install, allows conversion
kids_heights_ts_tbl = kids_heights_tbl_grp %>% ts_tbl()
ggplot(kids_heights_ts_tbl,aes(x=Age,y=Heights,col=Kid)) + geom_line()
```