Pitch Arsenal Classification

Prompt: Given a sample of Trackman data for a certain pitcher within a season, classify the pitches in the sample.

Data import and prep

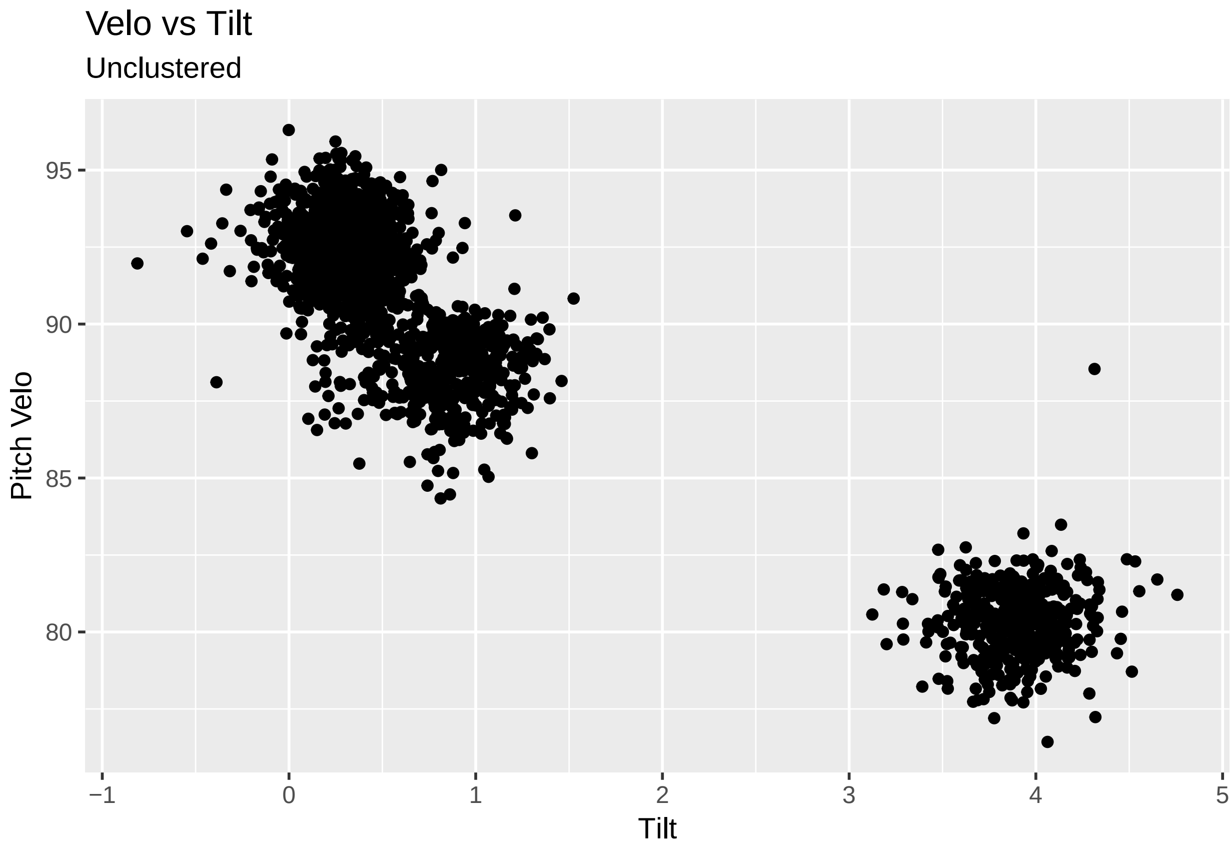
```r
#read in the data and convert to a data.table
all = data.table(read.csv("./trackman_data.csv",na.strings = "NULL"))

#remove the 6 rows without any Trackman data
tmd = all[complete.cases(all[,list(rel_speed,horz_break
                                   ,induced_vert_break,spin_rate)]),]
```

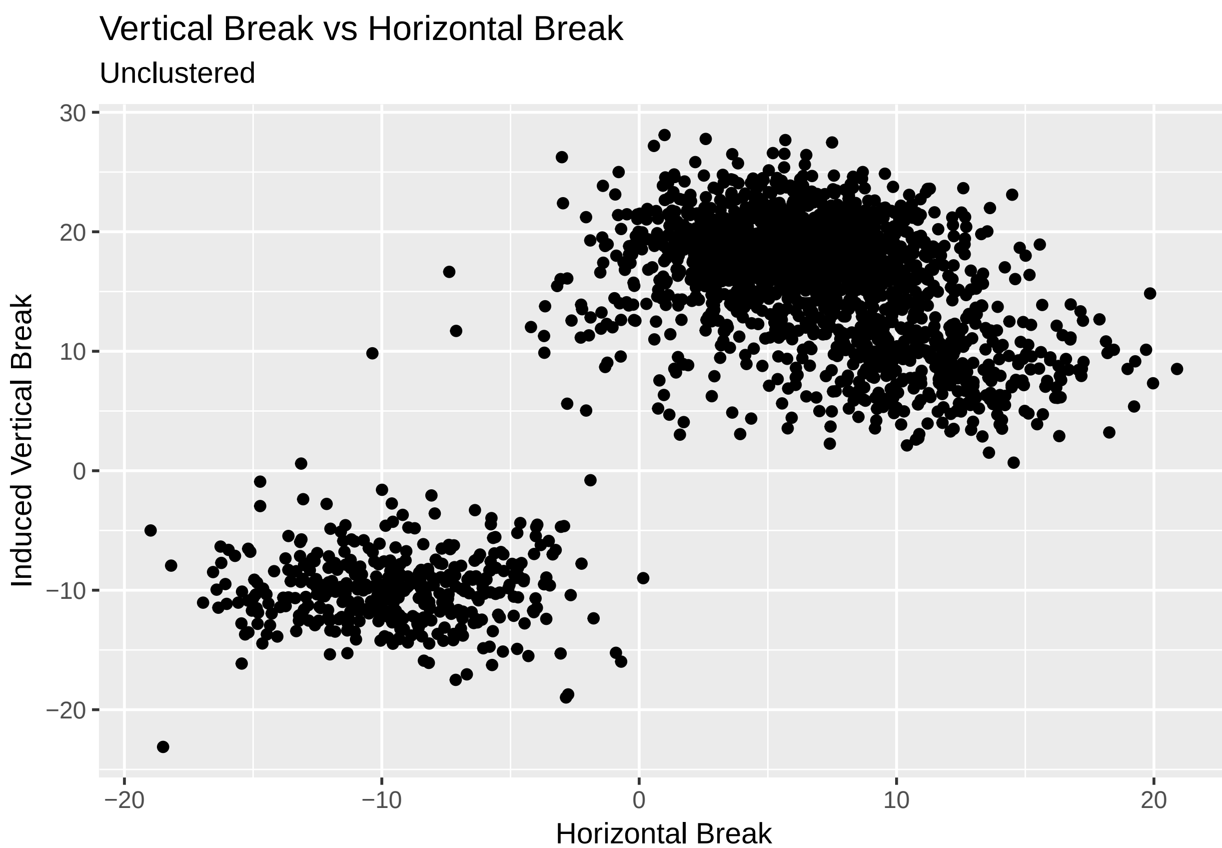
```r
#use break numbers to make a "tilt" variable (or spin direction)
tmd[,tilt := atan(horz_break/induced_vert_break)]
tmd[induced_vert_break<0,tilt:= pi+tilt]

#the atan cutoff separated one point from its apparent cluster so cheat
#a little here to get it back with the rest of its similar points
tmd[tilt < -1.5,tilt:=tilt+2*pi]
```

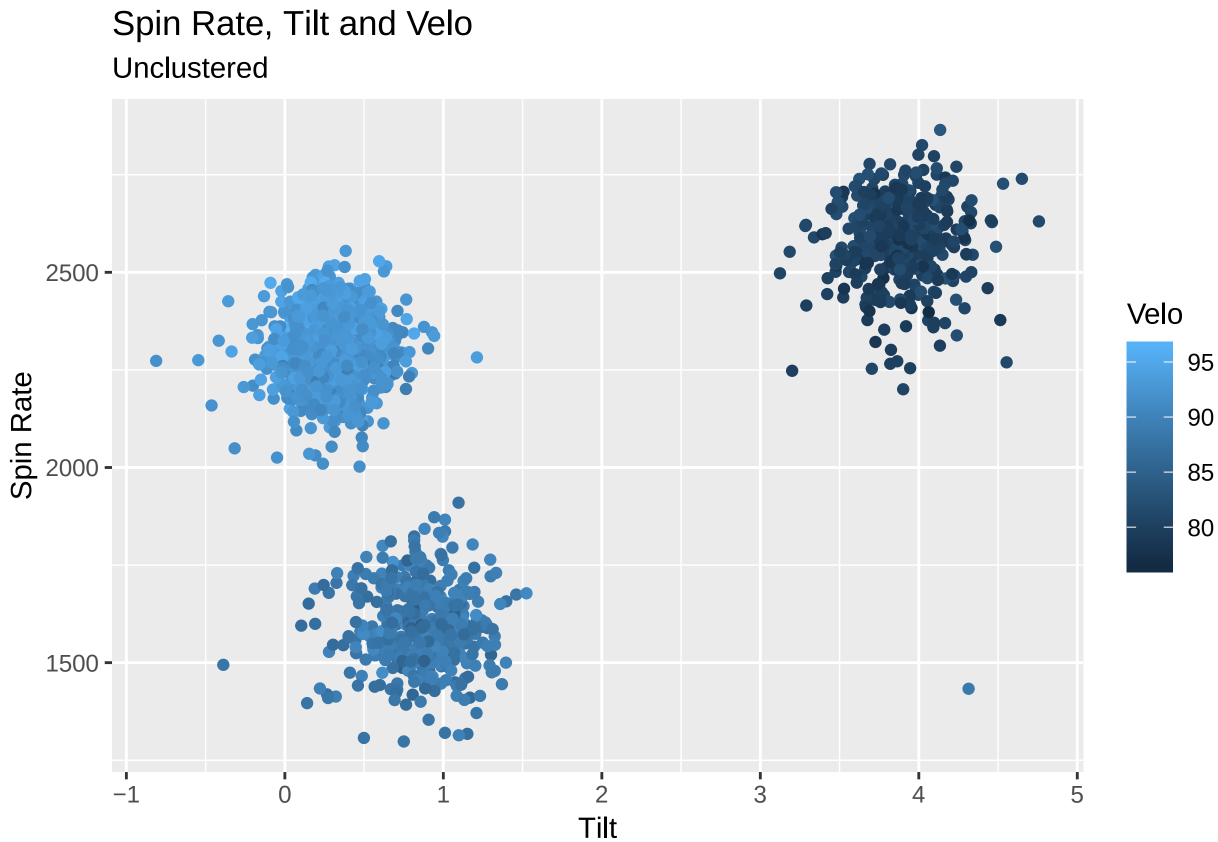
```r
#create several scatterplots to look at the data
ggplot(tmd,aes(x=tilt,y=rel_speed))+geom_point()+
  ggtitle("Velo vs Tilt",subtitle = "Unclustered")+
  xlab("Tilt")+ylab("Pitch Velo")
```




```r
ggplot(tmd,aes(x=horz_break,y=induced_vert_break))+geom_point()+
  ggtitle("Vertical Break vs Horizontal Break",subtitle = "Unclustered")+
  xlab("Horizontal Break")+ylab("Induced Vertical Break")
```




```r
ggplot(tmd,aes(x=tilt,y=spin_rate))+geom_point(aes(color = rel_speed))+
  labs(title = "Spin Rate, Tilt and Velo",subtitle = "Unclustered"
       ,color = "Velo")+
  xlab("Tilt")+ylab("Spin Rate")
```


Spin Rate, Tilt and Velo
Unclustered
Spin Rate
2500
2000
1500
−1
0
1
2
3
4
5
Tilt
Velo
95
90
85
80

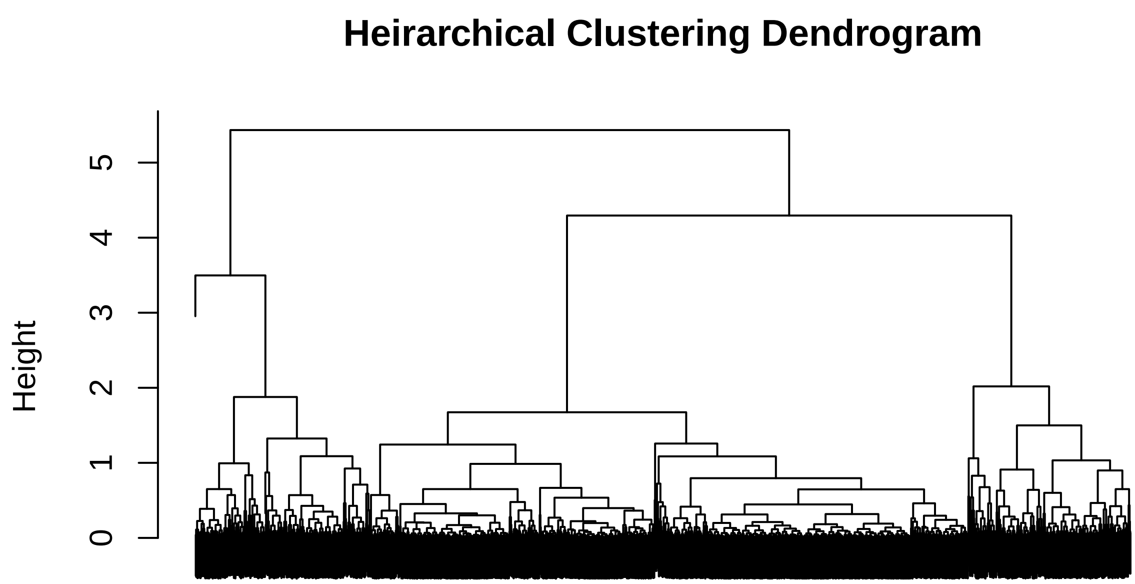
Clustering

Based on these charts Spin Rate and Tilt seem to break into 3 very clear clusters so we'll cluster across those 2 variables.

```
#filter out only the columns for the clustering analysis and scale
cld = scale(tmd[,list(spin_rate,tilt)])

#Perform Heirarchical clustering analysis on cld just to verify 3 clusters
hc = hclust(dist(cld),method="complete")

#plot the dendrogram
plot(hc, main="Heirarchical Clustering Dendrogram"
     ,xlab = "",sub = "",labels = FALSE)
```

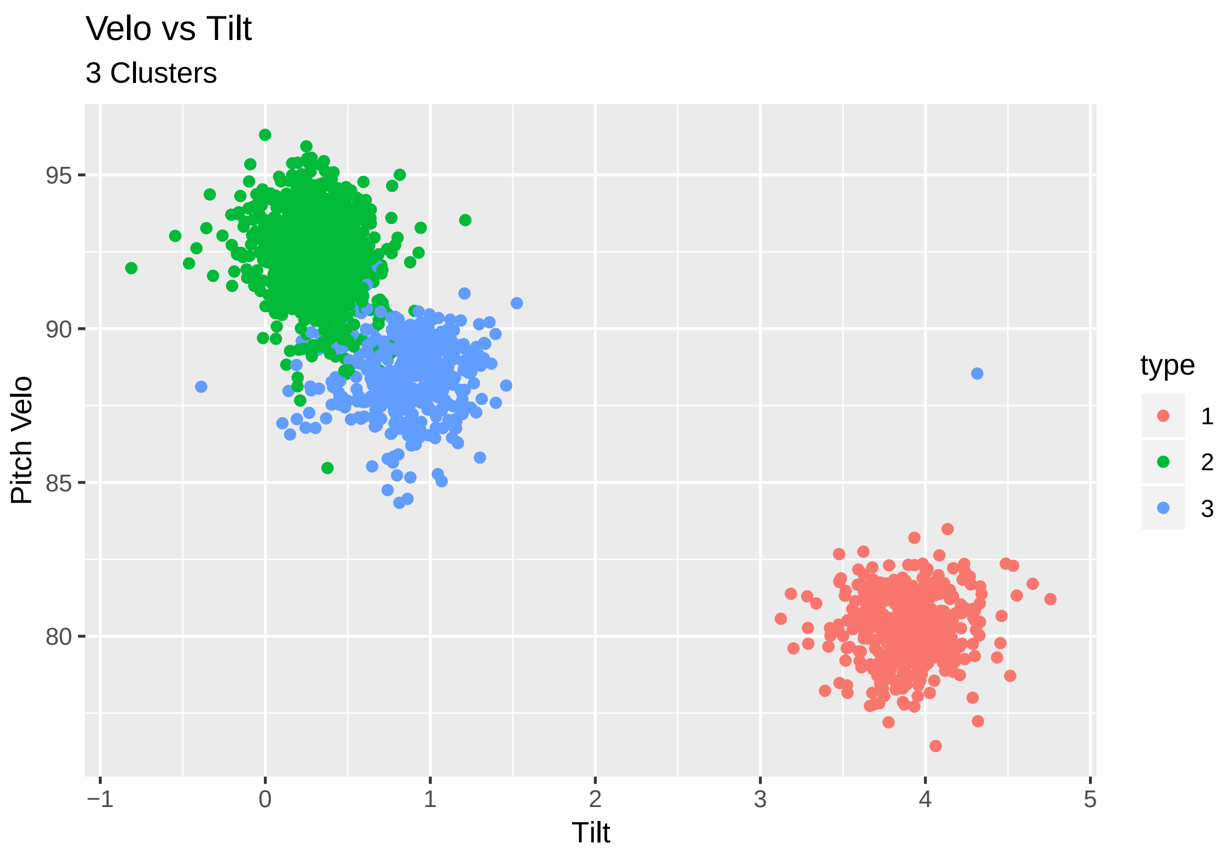


Based on the dendrogram, 3 clusters still appears to be the right amount.

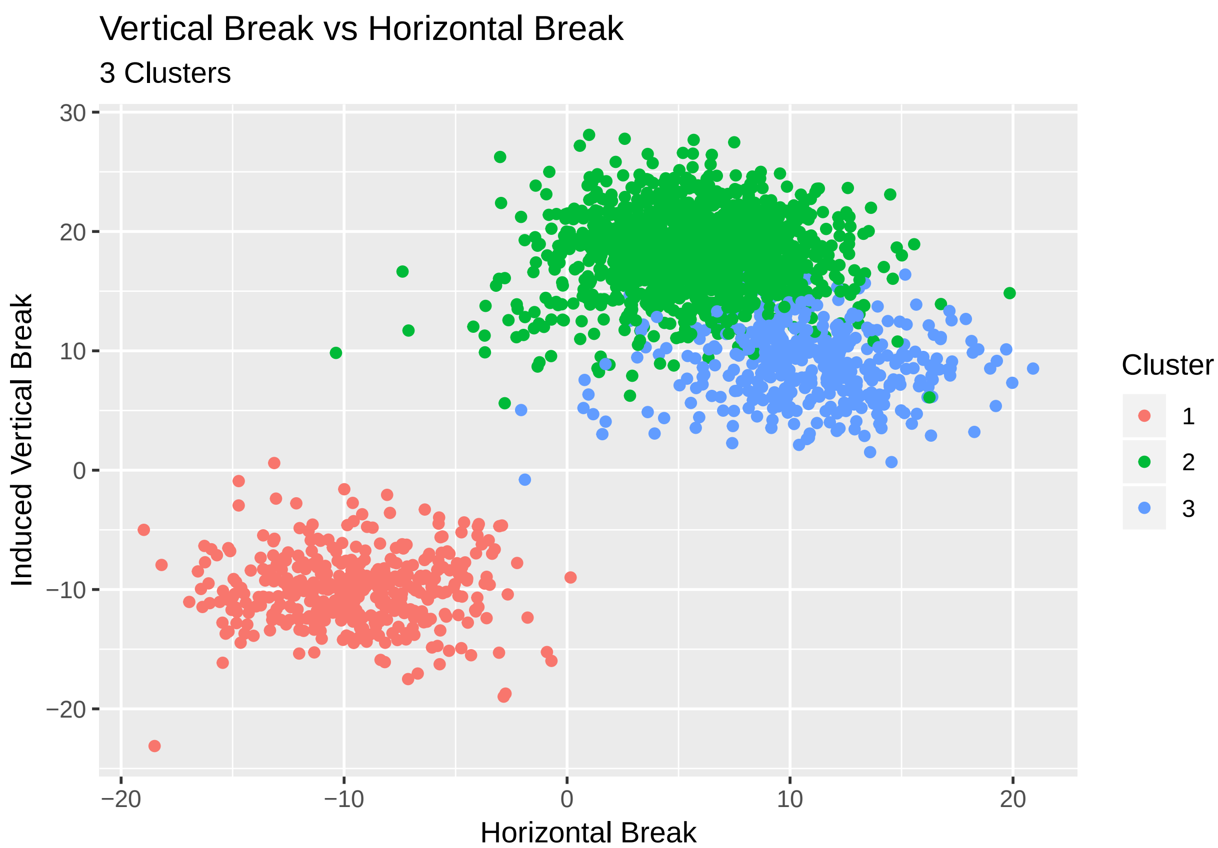
```
#label the pitches in tmd by cluster
cl = kmeans(cld,3)
tmd[,type := as.factor(cl$cluster)]
```

Recreate the same charts as above but colored by cluster to evaluate the accuracy of my clusters.

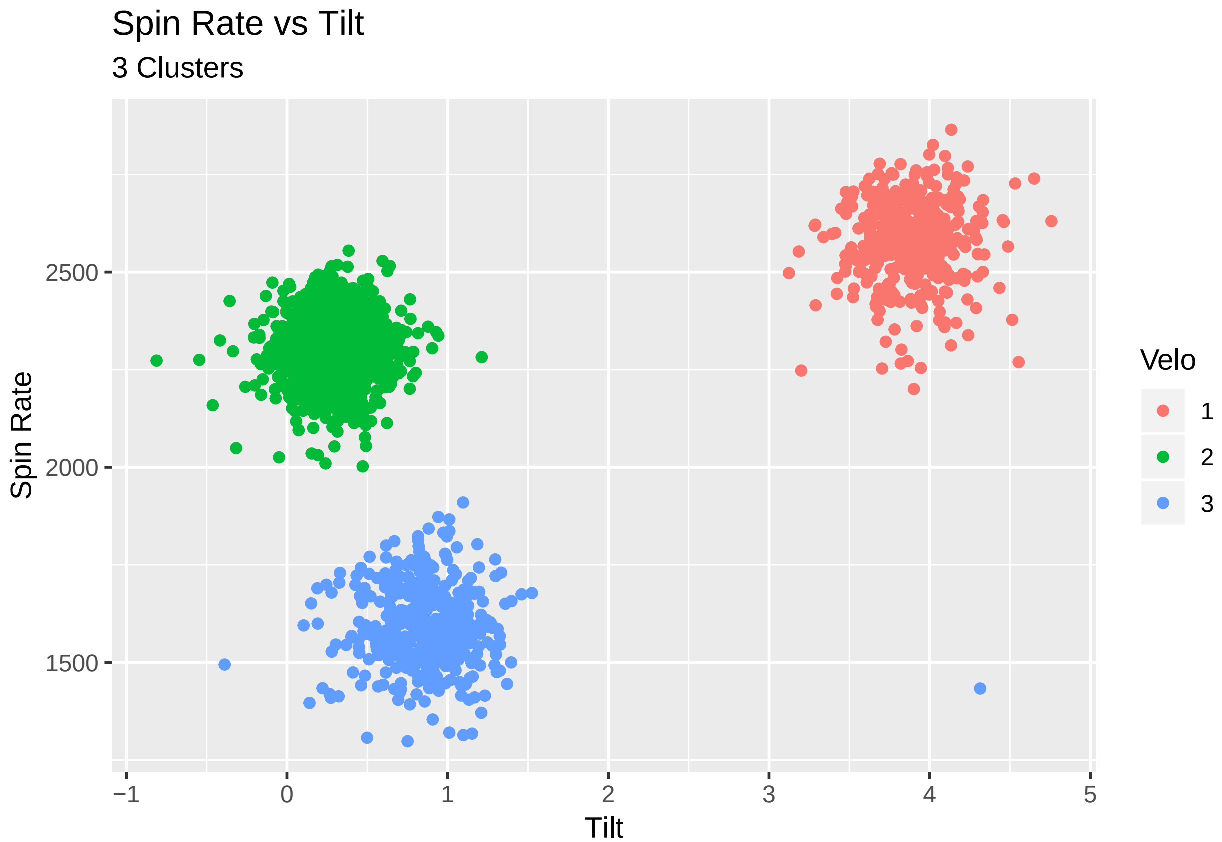
```
#plot the pitches by cluster to determine which is better
ggplot(tmd,aes(x=tilt,y=rel_speed))+
  geom_point(aes(color = type))+
  ggtitle("Velo vs Tilt",subtitle = "3 Clusters")+
  xlab("Tilt")+ylab("Pitch Velo")
```




```
ggplot(tmd,aes(x=horz_break,y=induced_vert_break))+
  geom_point(aes(color = type))+
  labs(title="Vertical Break vs Horizontal Break",subtitle = "3 Clusters"
       ,color = "Cluster")+
  xlab("Horizontal Break")+ylab("Induced Vertical Break")
```




```
ggplot(tmd,aes(x=tilt,y=spin_rate))+geom_point(aes(color = type))+
  labs(title = "Spin Rate vs Tilt",subtitle = "3 Clusters"
       ,color = "Velo")+
  xlab("Tilt")+ylab("Spin Rate")
```



Velocity and movement can run together between 2 of the clusters, but the velo vs tilt graph does support that these are the correct clusters, so move forward by manually defining each cluster.

```
#get average spin rate for each cluster in a table
sm = tmd[,list(spin = mean(spin_rate)),by = type]

#Assign each cluster its proper pitch type name using some basic logic
levels(tmd$type)[sm[spin == min(sm$spin),type]] = "Changeup"
levels(tmd$type)[sm[spin == max(sm$spin),type]] = "Curveball"
levels(tmd$type)[levels(tmd$type) %in% 1:3] = "Fastball"
```

View statistics by cluster

```r
#add binary columns of in/out of zone, swing, and strike.
tmd[,inzone := ifelse(abs(plate_loc_side) < .75 &
                        plate_loc_height > 1.5 &
                        plate_loc_height < 3.75,1,0)]
tmd[,swing := ifelse(pitch_call %in% c("FoulBall","InPlay"
                                        ,"StrikeSwinging")
                     ,1,0)]
tmd[,strike := ifelse(pitch_call %in% c("BallCalled"
                                         ,"BallIntentional"
                                         ,"HitByPitch")
                      ,0,1)]

#print a table of pitch-level metrics to analyze the pitch types
tmd[,list(velo=mean(rel_speed),spin=mean(spin_rate)
          ,h_break = mean(horz_break),v_break = mean(induced_vert_break)
          ,zone_pct = mean(inzone)
          ,strike_pct = mean(strike)
          ,chase = sum(swing*(1-inzone))/sum(1-inzone)
          ,whiff = sum(pitch_call == "StrikeSwinging")/sum(swing)
          ,exit_velo = mean(exit_speed,na.rm = TRUE)
          ,angle = mean(angle,na.rm = TRUE)
          ,p = length(exit_speed),HR = sum(play_result == "HomeRun"))
    ,by = type]
```

```
##         type     velo     spin   h_break   v_break  zone_pct strike_pct
## 1:  Fastball 92.47173 2307.577  5.813665 18.200301 0.5017135  0.6244003
## 2:  Changeup 88.46511 1593.990 10.757623  9.118594 0.4219114  0.6596737
## 3: Curveball 80.29920 2580.327 -9.448372 -9.920399 0.3627204  0.5012594
##        chase     whiff exit_velo      angle    p HR
## 1: 0.2132050 0.1785174  83.09793 23.9472968 1459  6
## 2: 0.4274194 0.3466135  84.05011 -0.9400901  429  0
## 3: 0.2687747 0.3602941  84.77678  6.1180341  397  0
```

Output the classifications as a CSV

```r
#add the classifications back to the original data
setkey(tmd,pitch_id)
setkey(all,pitch_id)
classified = tmd[all][,list(pitch_id,type)][order(pitch_id)]

#set the few records with no Trackman data as "undefined"
classified[is.na(type),type := "Undefined"]

#write back to a csv
write.csv(classified
          ,"./Classified_pitches.csv",row.names = FALSE)
```