Hockey Big Data Competition

Joshua

2/4/2021

```
data <- read.csv("https://raw.githubusercontent.com/bigdatacup/Big-Data-Cup-2021/main/hackathon_scoutin

View(data)
```

```
Wdata <- read.csv("https://raw.githubusercontent.com/bigdatacup/Big-Data-Cup-2021/main/hackathon_womens

View(Wdata)
```

```
dim(data) #75873 rows  21 columns
```

```
## [1] 75873    21
```

```
dim(Wdata) #24002 rows   21 columns
```

```
## [1] 24002    21
```

```
#The only difference is the gender.
```

Testing an assumption

```
#commented out for now, as it causes some knitting problems
#u <-unique(data["Home.Team"])
#v <- unique(data["Away.Team"])
#length(u)==length(v)
#length(setdiff(u,v))
#length(setdiff(v,u))
#length(intersect(u,v))==length(u) False.
```

ˆFinding:Not all the teams in Home.Team are in Away.Team.

Ideas after examining the datasets: Q.How do players' performances vary?

```
length(unique(data$Player))
```

```
## [1] 408
```

#so not a good idea.

Q.Making clusters and identifying successful clusters and clusters that need improvements? A:No, the datas are already classified so no need to use clustering algorithms.

Analyze shot: Shot attempts that are unsuccessful (block, miss or save) Analyze goal: Shot attempts that are successful (goal) Analyze play: Pass attempts that are successful

Which skills get affected the most and least between these two types of matches(Home and away)?

How does it vary between two genders?

The rest above sound good. I will proceed on. I need to simplify/clean the dataset first.

Simplifying dataset: #2nd x-coord, 2nd y-coord. Many of them are NA

```
#Collecting 2020 data only
teams <- group_by(data,Home.Team)

shots <- teams %>% filter(Event=="Shot")
shots$date <- as.Date(shots$game_date,format= "%Y-%m-%d")
shots <- subset(shots,game_date>"2020-01-01" & game_date<"2021-01-01")
shots
```

```
## # A tibble: 2,303 x 22
## # Groups:   Home.Team [11]
##    game_date Home.Team Away.Team Period Clock Home.Team.Skate~ Away.Team.Skate~
##    <chr>     <chr>     <chr>      <int> <chr>            <int>            <int>
##  1 2020-01-~ Kingston~ Erie Ott~      1 18:53                5                5
##  2 2020-01-~ Kingston~ Erie Ott~      1 18:40                5                5
##  3 2020-01-~ Kingston~ Erie Ott~      1 18:28                5                5
##  4 2020-01-~ Kingston~ Erie Ott~      1 18:23                5                5
##  5 2020-01-~ Kingston~ Erie Ott~      1 17:51                5                5
##  6 2020-01-~ Kingston~ Erie Ott~      1 17:05                5                5
##  7 2020-01-~ Kingston~ Erie Ott~      1 16:24                5                5
##  8 2020-01-~ Kingston~ Erie Ott~      1 14:06                5                5
##  9 2020-01-~ Kingston~ Erie Ott~      1 13:43                5                5
## 10 2020-01-~ Kingston~ Erie Ott~      1 13:07                5                5
## # ... with 2,293 more rows, and 15 more variables: Home.Team.Goals <int>,
## #   Away.Team.Goals <int>, Team <chr>, Player <chr>, Event <chr>,
## #   X.Coordinate <int>, Y.Coordinate <int>, Detail.1 <chr>, Detail.2 <chr>,
## #   Detail.3 <chr>, Detail.4 <chr>, Player.2 <chr>, X.Coordinate.2 <int>,
## #   Y.Coordinate.2 <int>, date <date>
```

```
nrow(shots) #2303
```

```
## [1] 2303
```

```
goal <- teams %>% filter(Event=="Goal")
goal$date <- as.Date(goal$game_date,format= "%Y-%m-%d")
goal <- subset(goal,game_date>"2020-01-01" & game_date<"2021-01-01")
goal
```

```
## # A tibble: 132 x 22
## # Groups:   Home.Team [11]
##    game_date Home.Team Away.Team Period Clock Home.Team.Skate~ Away.Team.Skate~
##    <chr>     <chr>     <chr>      <int> <chr>            <int>            <int>
##  1 2020-01-~ Kingston~ Erie Ott~      1 0:56                  5                5
##  2 2020-01-~ Kingston~ Erie Ott~      2 16:55                 5                5
##  3 2020-01-~ Kingston~ Erie Ott~      2 15:34                 5                5
##  4 2020-01-~ Kingston~ Erie Ott~      2 14:21                 5                5
##  5 2020-01-~ Kingston~ Erie Ott~      2 10:03                 5                5
##  6 2020-01-~ Kingston~ Erie Ott~      2 1:36                  5                5
##  7 2020-01-~ Kingston~ Erie Ott~      3 16:20                 5                5
##  8 2020-01-~ Kingston~ Erie Ott~      3 15:22                 5                5
##  9 2020-01-~ Kingston~ Erie Ott~      3 8:06                  5                5
## 10 2020-01-~ Ottawa 6~ Erie Ott~      1 11:19                 5                5
## # ... with 122 more rows, and 15 more variables: Home.Team.Goals <int>,
## #   Away.Team.Goals <int>, Team <chr>, Player <chr>, Event <chr>,
## #   X.Coordinate <int>, Y.Coordinate <int>, Detail.1 <chr>, Detail.2 <chr>,
## #   Detail.3 <chr>, Detail.4 <chr>, Player.2 <chr>, X.Coordinate.2 <int>,
## #   Y.Coordinate.2 <int>, date <date>
```

```
nrow(goal) #132
```

```
## [1] 132
```

```
pass <- teams %>% filter(Event=="Play")
pass$date<- as.Date(pass$game_date,format= "%Y-%m-%d")
pass <- subset(pass,game_date>"2020-01-01" & game_date<"2021-01-01")
pass
```

```
## # A tibble: 11,303 x 22
## # Groups:   Home.Team [11]
##    game_date Home.Team Away.Team Period Clock Home.Team.Skate~ Away.Team.Skate~
##    <chr>     <chr>     <chr>      <int> <chr>            <int>            <int>
##  1 2020-01-~ Kingston~ Erie Ott~      1 19:56                5                5
##  2 2020-01-~ Kingston~ Erie Ott~      1 19:51                5                5
##  3 2020-01-~ Kingston~ Erie Ott~      1 19:45                5                5
##  4 2020-01-~ Kingston~ Erie Ott~      1 19:43                5                5
##  5 2020-01-~ Kingston~ Erie Ott~      1 19:36                5                5
##  6 2020-01-~ Kingston~ Erie Ott~      1 19:25                5                5
##  7 2020-01-~ Kingston~ Erie Ott~      1 19:21                5                5
##  8 2020-01-~ Kingston~ Erie Ott~      1 19:15                5                5
##  9 2020-01-~ Kingston~ Erie Ott~      1 19:13                5                5
## 10 2020-01-~ Kingston~ Erie Ott~      1 19:01                5                5
## # ... with 11,293 more rows, and 15 more variables: Home.Team.Goals <int>,
## #   Away.Team.Goals <int>, Team <chr>, Player <chr>, Event <chr>,
## #   X.Coordinate <int>, Y.Coordinate <int>, Detail.1 <chr>, Detail.2 <chr>,
## #   Detail.3 <chr>, Detail.4 <chr>, Player.2 <chr>, X.Coordinate.2 <int>,
## #   Y.Coordinate.2 <int>, date <date>
```

```
nrow(pass) #11303
```

```
## [1] 11303
```

-Analyzing shots made in 2020-

```
table(shots$Detail.2)
```

```
##
## Blocked  Missed  On Net
##     556     674    1073
```

```
shots1 <- shots %>% filter(Period=="1")
shots2 <- shots %>% filter(Period=="2")
shots3 <- shots %>% filter(Period=="3")

table(shots1$Detail.2)
```

```
##
## Blocked  Missed  On Net
##     184     218     334
```

```
table(shots2$Detail.2)
```

```
##
## Blocked  Missed  On Net
##     182     232     363
```

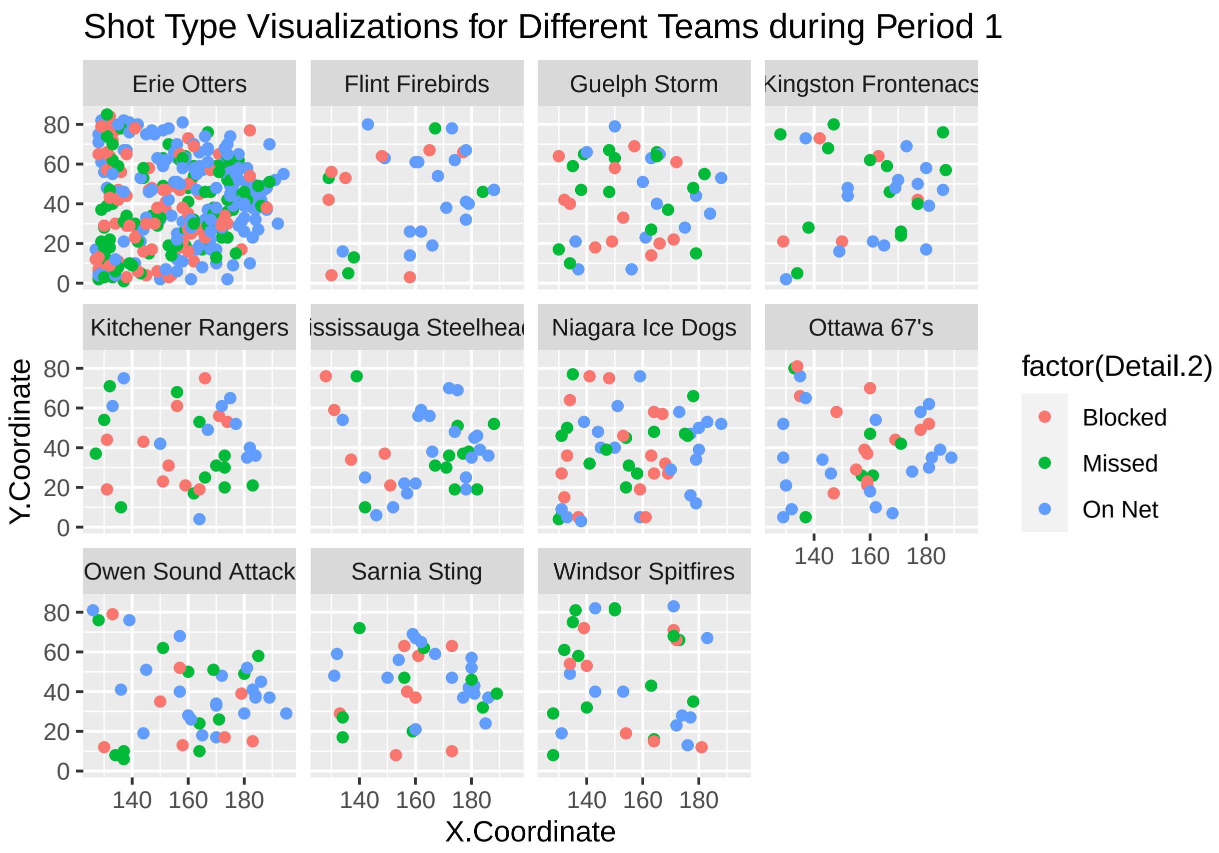
```
table(shots3$Detail.2)
```

```
##
## Blocked  Missed  On Net
##     188     220     370
```

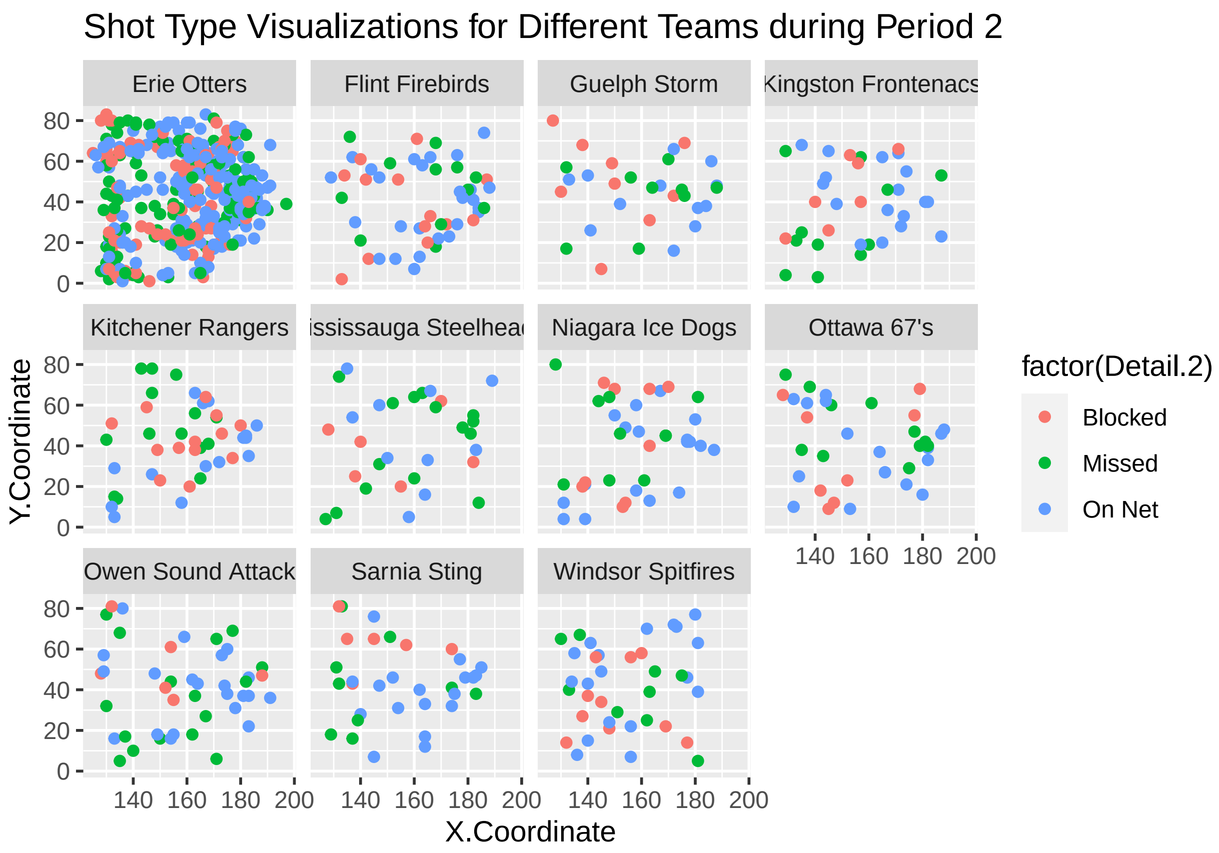
Observation: 1)The count of On Net generally increases over the period 2) The count of missed spiked during 2nd period but not by significant margin 3) The count of blocked is lowest in 2nd period and highest in 3rd period, but the difference is not significant. 4)On Net is the shot type with the highest count; blocked is the shot type with the lowest count.

Suggestion: 1) Work on the precision training in order to reduce the shot on Net.

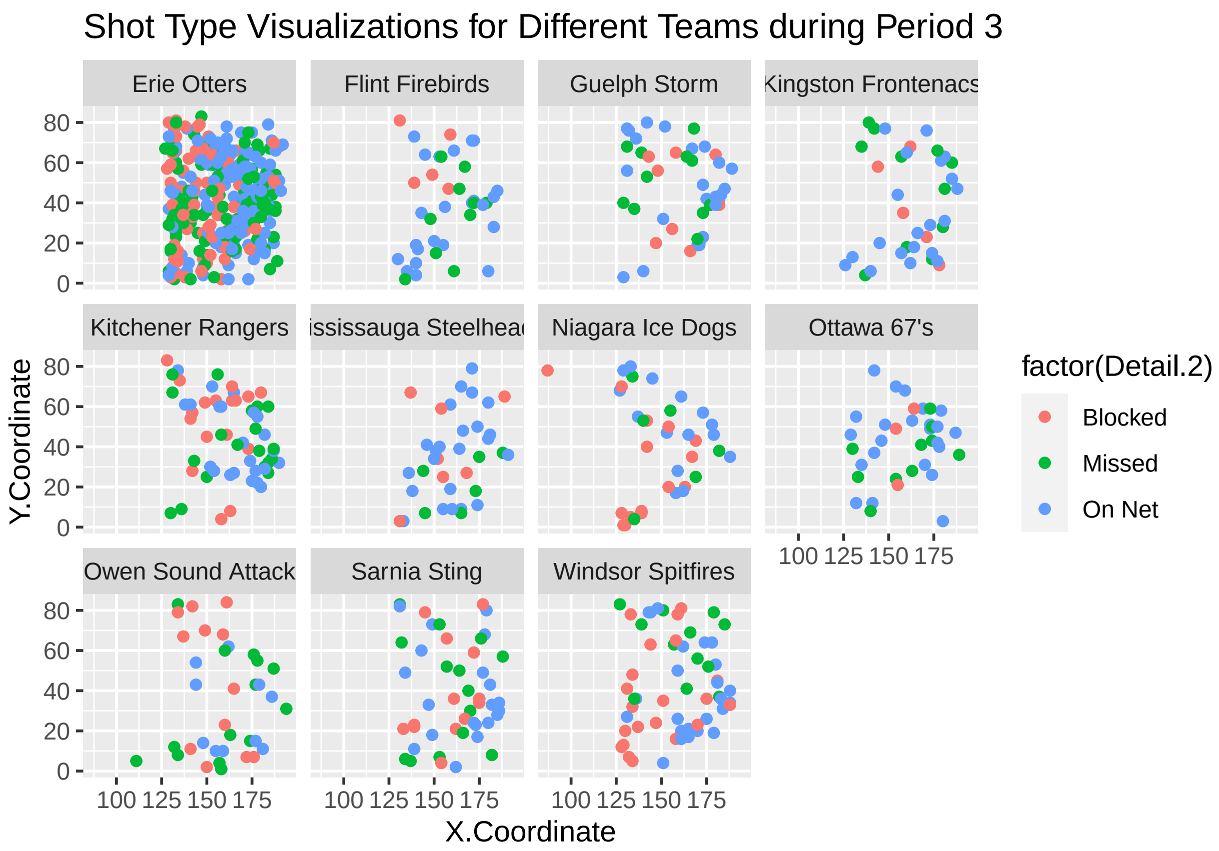
```
shots1 <- shots %>% filter(Period=="1")
ggplot(shots1,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.2)))+ggtitle("Shot Type Visualizations
```



```
shots2 <- shots %>% filter(Period=="2")
ggplot(shots2,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.2)))+ggtitle("Shot Type Visualizations
```



```
shots3 <- shots %>% filter(Period=="3")
ggplot(shots3,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.2)))+ggtitle("Shot Type Visualizations
```



-Analyzing goals made in 2020-

```
table(goal$Detail.1)
```

```
##
##  Deflection     Slapshot     Snapshot Wrap Around    Wristshot
##           9            3           65           2           53
```

```
goal1 <- goal %>% filter(Period=="1")
goal2 <- goal %>% filter(Period=="2")
goal3 <- goal %>% filter(Period=="3")

table(goal1$Detail.1)
```

```
##
## Deflection    Slapshot    Snapshot   Wristshot
##          3           2          15          15
```

```
table(goal2$Detail.1)
```

```
##
## Deflection    Snapshot   Wristshot
##          3          27          20
```

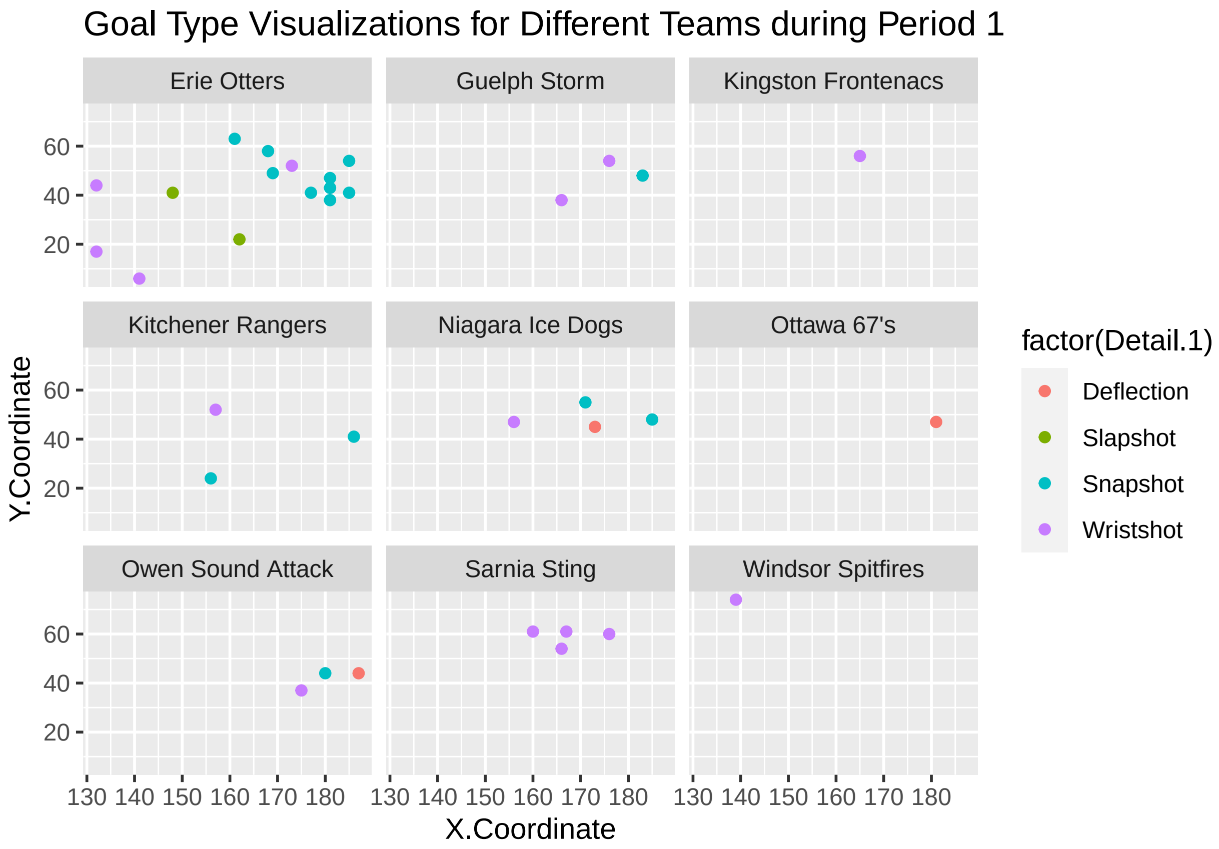
```
table(goal3$Detail.1)
```

```
##
##   Deflection     Slapshot     Snapshot Wrap Around     Wristshot
##            3            1           21            2           18
```

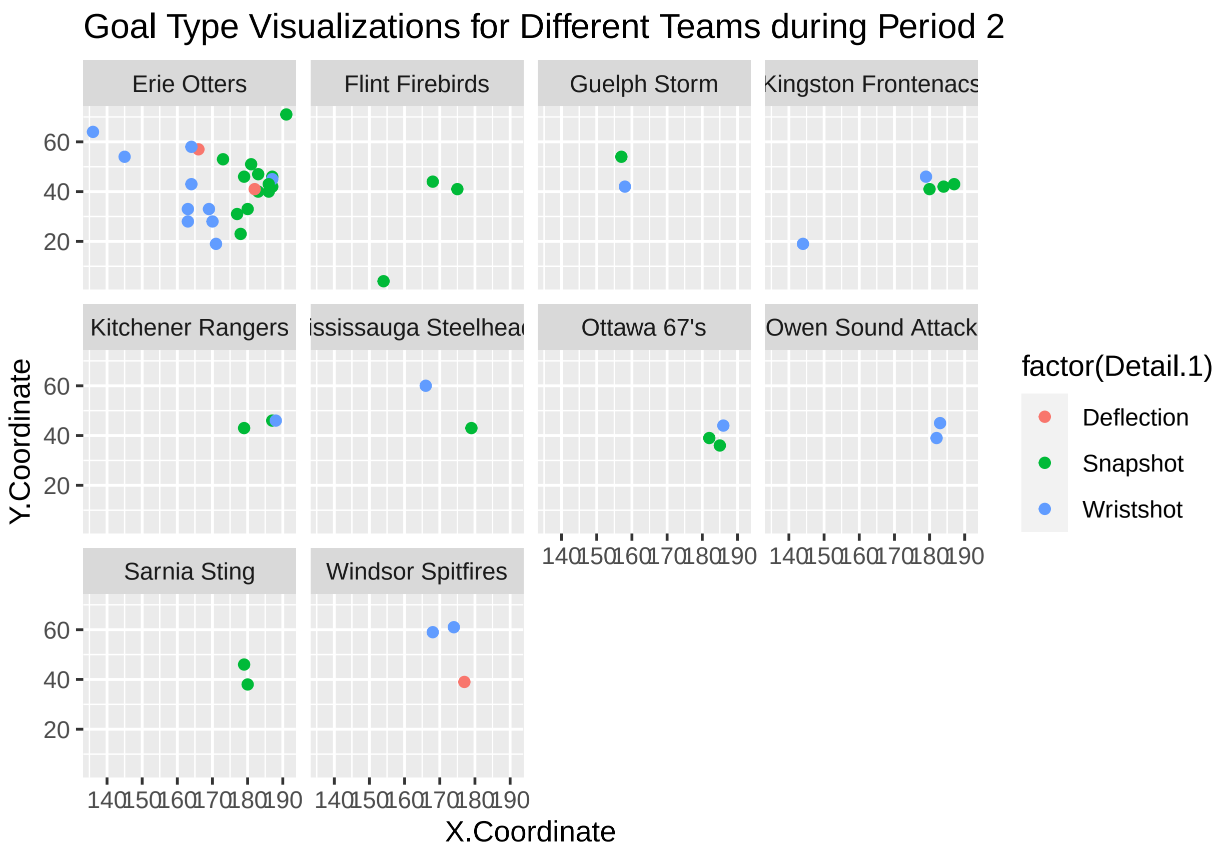
Observation: 1)Snapshot takes approximately 49% of the overall goals made in 2020. 2) Snapshot is the goal type that was of highest count, whereas wrap around is the goal type that is of the lowest count.

Suggestion: 1) Work on developing the goal skills of the 3 lowest count: wrap around, slapshot, and deflection.

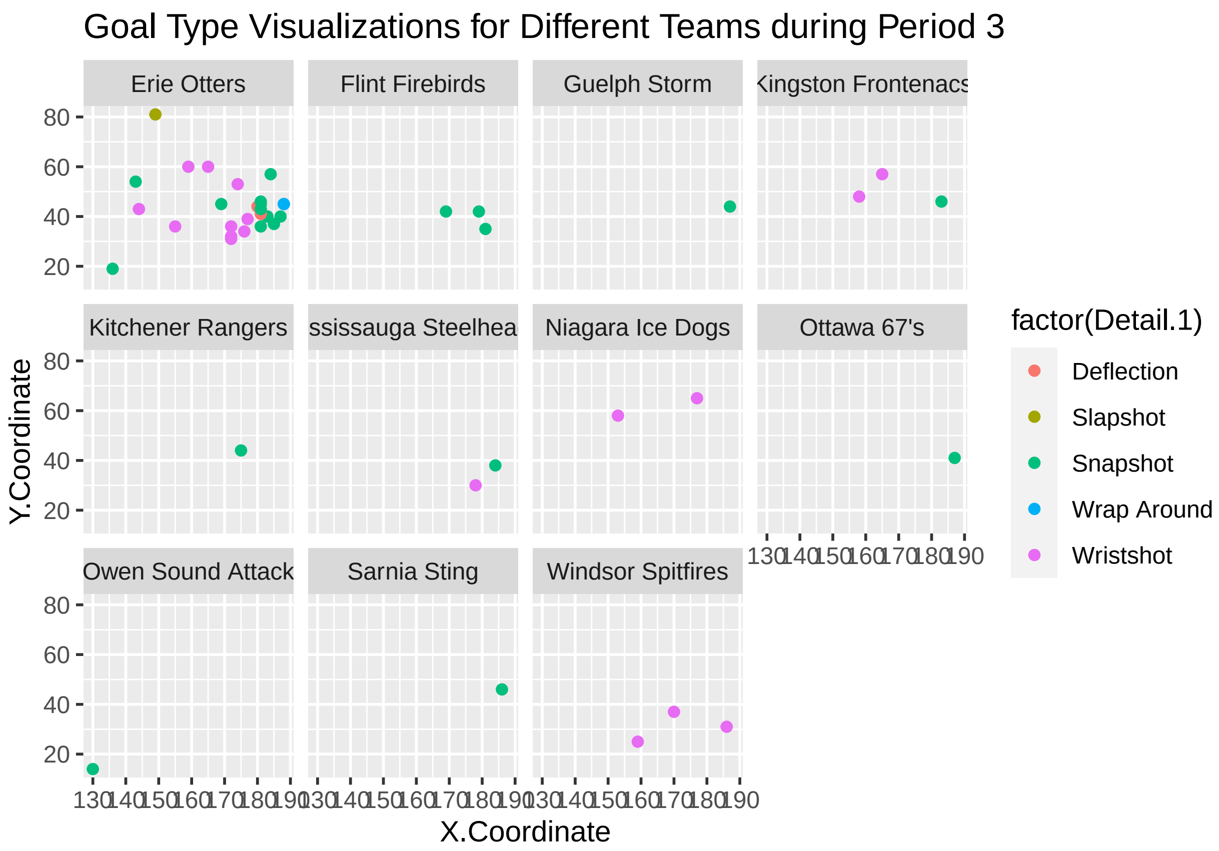
```
goal1 <- goal %>% filter(Period=="1")
ggplot(goal1,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Goal Type Visualizations
```



```
goal2 <- goal %>% filter(Period=="2")
ggplot(goal2,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Goal Type Visualizations
```



```
goal3 <- goal %>% filter(Period=="3")
ggplot(goal3,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Goal Type Visualizations
```



-Analyzing pass made in 2020-

```
table(pass$Detail.1)
```

```
##
##   Direct Indirect
##     7959     3344
```

```
pass1 <- pass %>% filter(Period=="1")
pass2 <- pass %>% filter(Period=="2")
pass3 <- pass %>% filter(Period=="3")

table(pass1$Detail.1)
```

```
##
##   Direct Indirect
##     2793     1104
```

```
table(pass2$Detail.1)
```

```
##
##   Direct Indirect
##     2619     1077
```

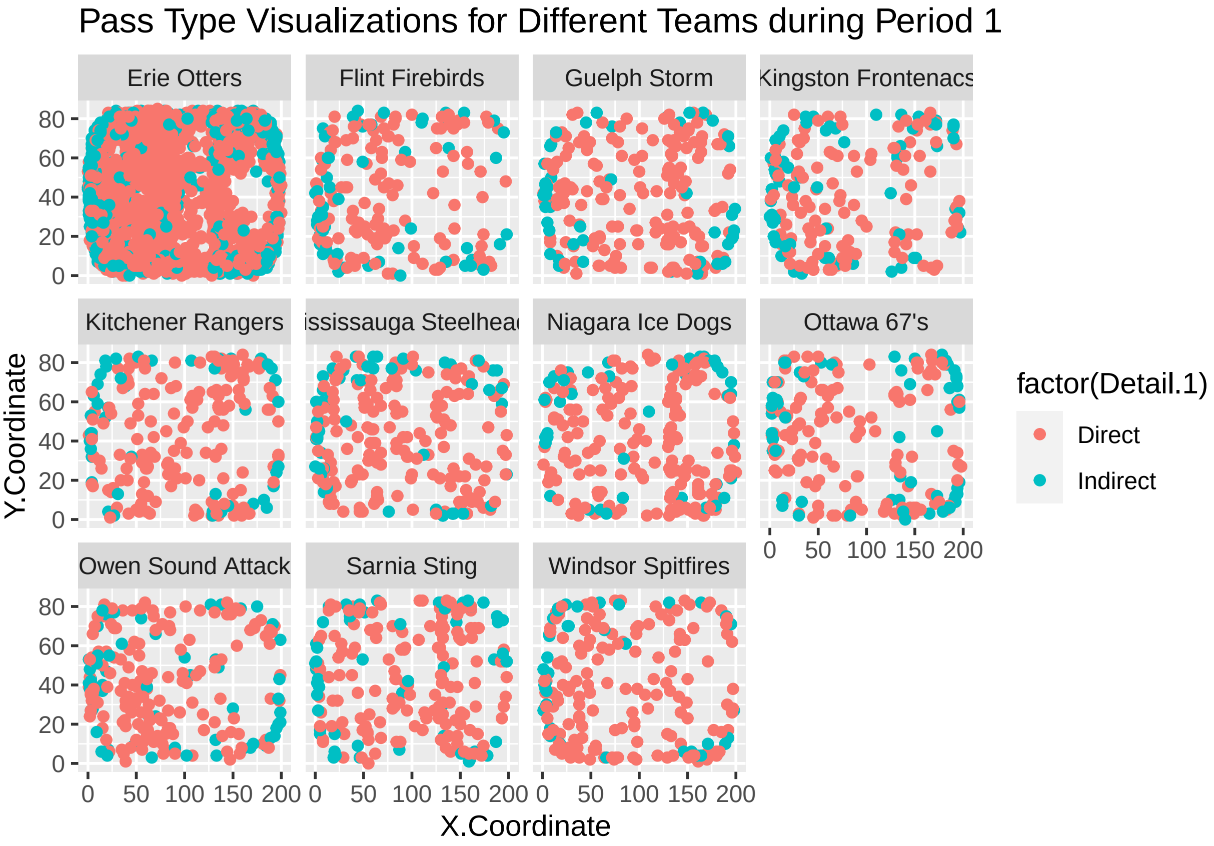
```
table(pass3$Detail.1)
```

```
##
##   Direct Indirect
##     2487     1137
```

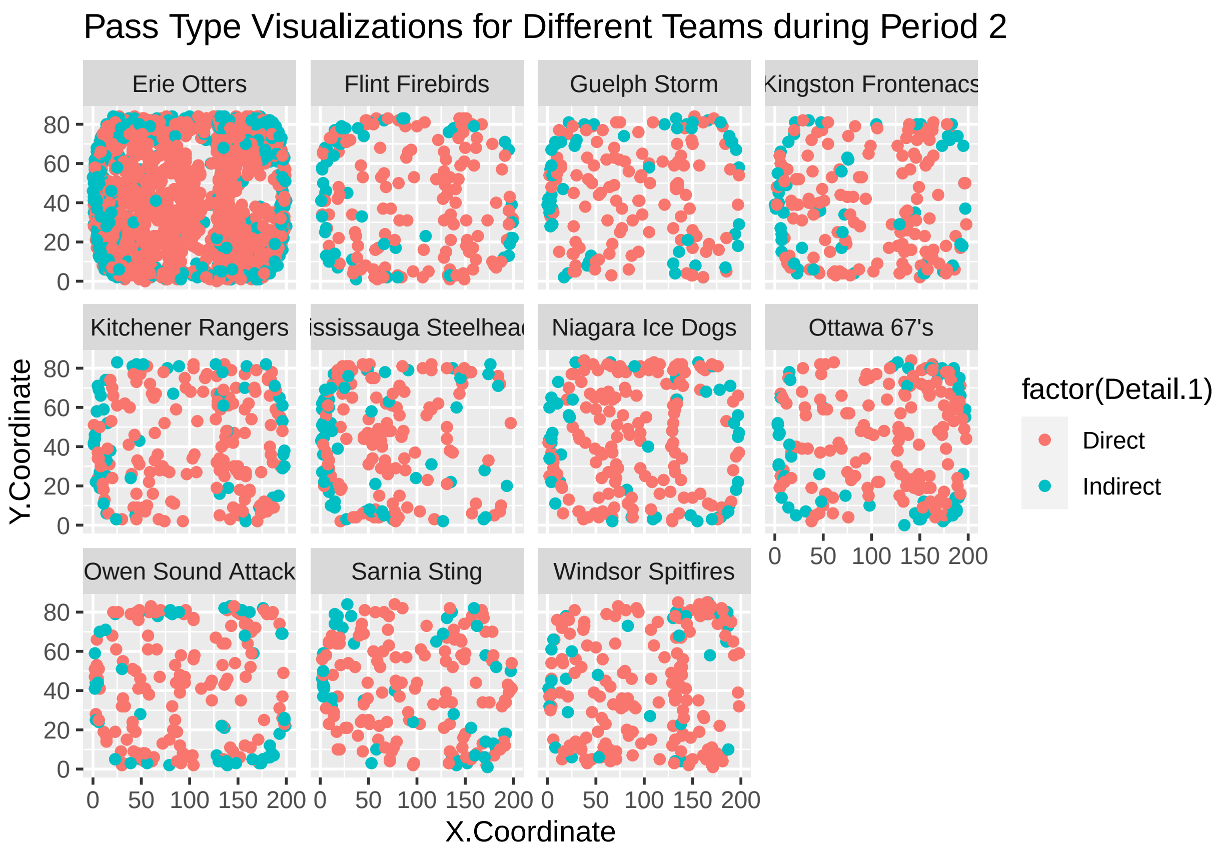
Observation: 1)The count of Direct passes are higher than that of indirect passes in all 3 periods.

Suggestion: 1) Dedicating more time into practicing either direct passes, which are performed more during matches, or into indirect passes to have more options when it comes to playing styles

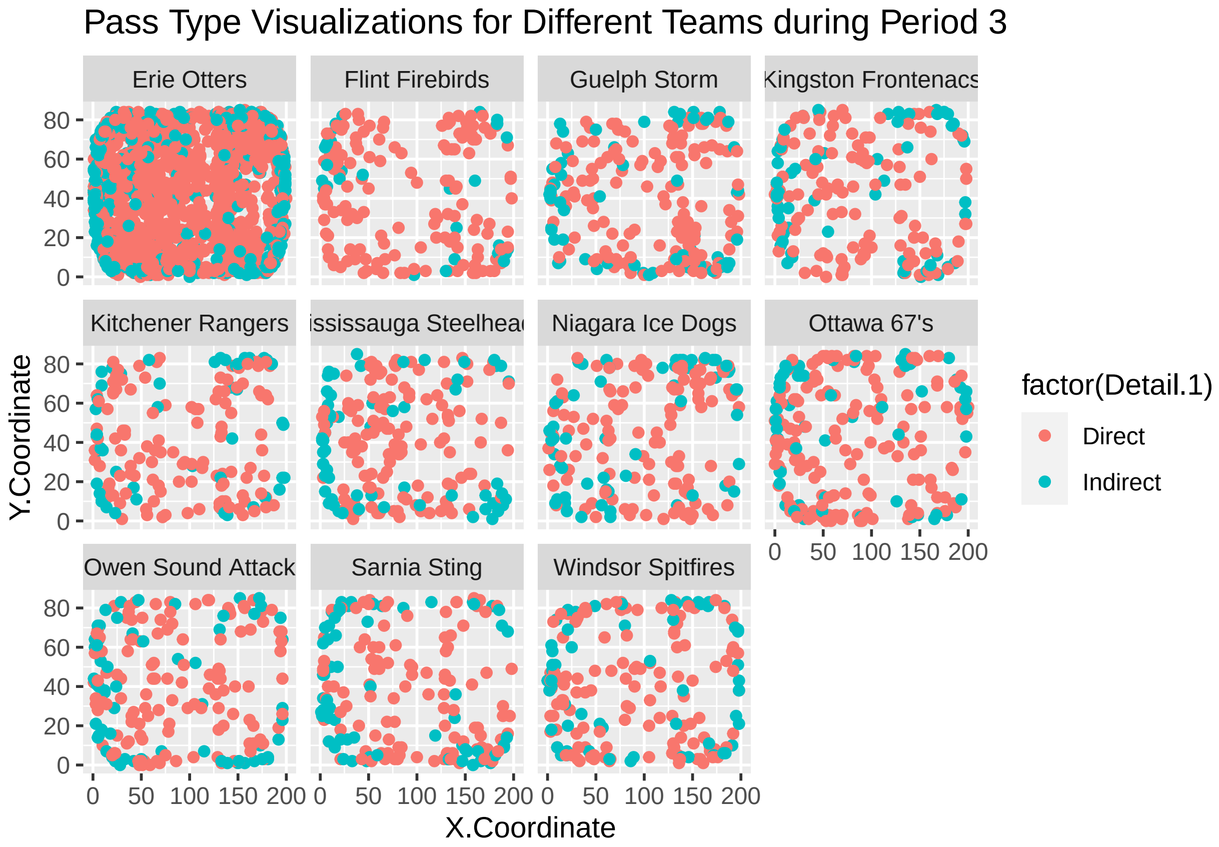
```
pass1 <- pass %>% filter(Period=="1")
ggplot(pass1,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Pass Type Visualizations
```



```
pass2 <- pass %>% filter(Period=="2")
ggplot(pass2,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Pass Type Visualizations
```



```
pass3 <- pass %>% filter(Period=="3")
ggplot(pass3,aes(x=X.Coordinate,y=Y.Coordinate,col=factor(Detail.1)))+ggtitle("Pass Type Visualizations
```



Observation: 1)Whereas direct passes are made from everywhere throughout all 3 periods, indirect passes generally seem to occur near the ends of the rings rather than the middle body of the ring in the 1st period. Then, over the 2nd and 3rd periods, more indirect passes begin to occur slightly more near th middle body of the ring.

Suggestion: In terms of strategy, the players should watch out more for indirect passes near the middle part of the ring, particularly during period 2. Especially if the teams are ssissasuga steelhead and Guelph storm.