Foul Ball Data

Introduction

Baseball, America's favorite pastime, draws hundreds of thousands of spectators to stadiums nationwide each season. In recent years, a conversation has risen concerning the dangers foul balls can pose to unsuspecting spectators. While many attend games in hopes Many say the safety nets should be extended to protect a larger portion of the sideline stands. In this analysis we are trying to determine which section(s) would require more protection. The data was collected by Baseball Savant from the games that the most foul balls were hit, from the top ten stadiums in fouls balls hit across the league. The range of these games were from the opening day in March, all the way up to June 5, 2019. This is an observational study because there was no interface in the play to adjust for the produced foul balls. This sample should be considered a biased sample because the sample only comes from 10 different stadiums which could have different features that produce foul balls more frequently than the rest of the league. For instance the grandstand behind home plate could cast a shadow on the field that causes batters to undercut the ball more resulting in more pop ups behind home plate. The data is interesting to us because it shows the type of foul balls hit in each location of the park. This data could be of use to the class because it could optimize their seat selection to catch a foul ball or to be protected from a line drive. We also would like to analyze the difference between the predicted zone from the foul ball and the actual camera zone which it ended up landing in. This will help educate us on how accurate the model used for predicting the landing of foul balls actually is.

Data Cleaning

```
FoulBalls2 <- FoulBalls %>% filter(!is.na(type_of_hit))
FoulBalls3 <- FoulBalls2 %>% filter(!is.na(predicted_zone))
FoulBalls2 <- FoulBalls2 %>% filter(!is.na(camera_zone))
```

We used functions in R to remove any empty values from the following variables: type_of_hit, predicited_zone, camera_zone. This ensured our data wasr accurate and represented all variables that we wanted to look at for the purposes of this project.

Simple bar graph

```
f <- FoulBalls %>% complete.cases(FoulBalls) %>%
  mutate(exitvcategory = ifelse(exit_velocity > 90, "high risk", NA),
         exitvcategory = ifelse(exit_velocity >= 60 & exit_velocity <= 90, "medium risk",
exitvcategory),
         exitvcategory = ifelse(exit_velocity < 60, "low risk", exitvcategory))


ggplot(data=f, mapping=aes(x=used_zone, fill = used_zone)) +
  geom_bar(fill = "blue")
```

This bar graph shows the foul ball measurements for each zone. We can see that most foul balls land in zone 1, and there are no foul balls in zone two. Few foul balls land in zones three, six, and seven. Zones four and five have a considerable number of foul balls, and these zones are not protected by nets like zone one is.

Velocity graph

This graph shows the risk a foul ball has depending on its exit velocity. Foul balls with a velocity of above ninety miles per hour are considered “high risk.” Foul balls with an exit velocity greater than sixty miles per hour are considered “medium risk.” “Low risk” foul balls are considered so with an exit velocity less than sixty miles per hour. We see that the most risky zones are three through seven, as these areas are unprotected by nets and are used zones for “medium” and “high” risk foul balls.

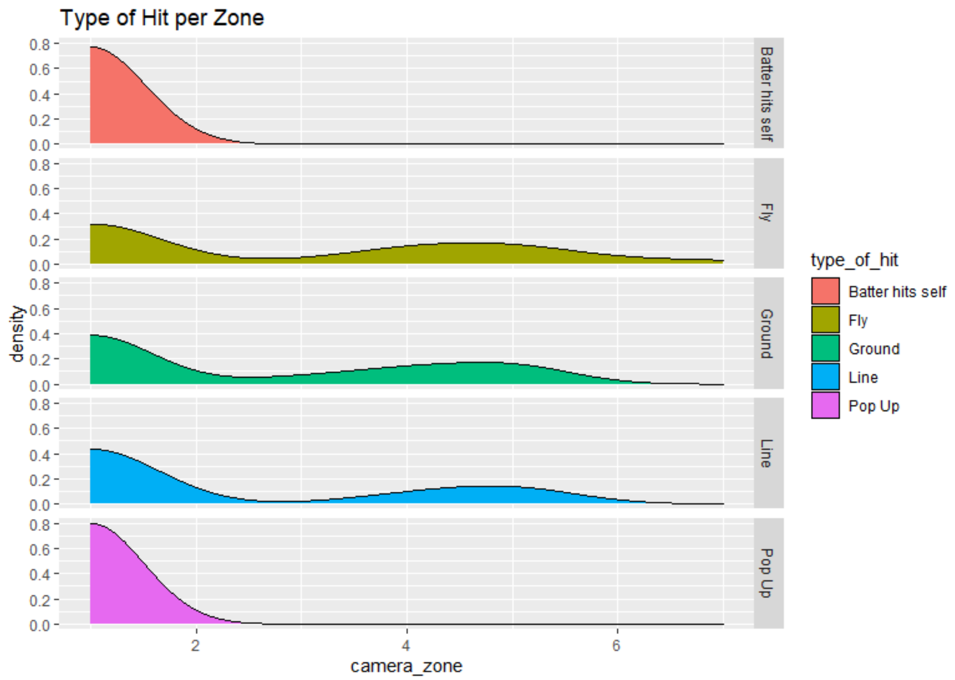
```
f <- FoulBalls %>% complete.cases(FoulBalls) %>%
  mutate(exitvcategory = ifelse(exit_velocity > 90, "high risk", NA),
         exitvcategory = ifelse(exit_velocity >= 60 & exit_velocity <= 90, "medium risk",
exitvcategory),
         exitvcategory = ifelse(exit_velocity < 60, "low risk", exitvcategory))

ggplot(data=f, mapping=aes(x=used_zone, fill = exitvcategory)) +
geom_bar(position='fill')
```

Data Analysis

Camera Zone Density Plot:

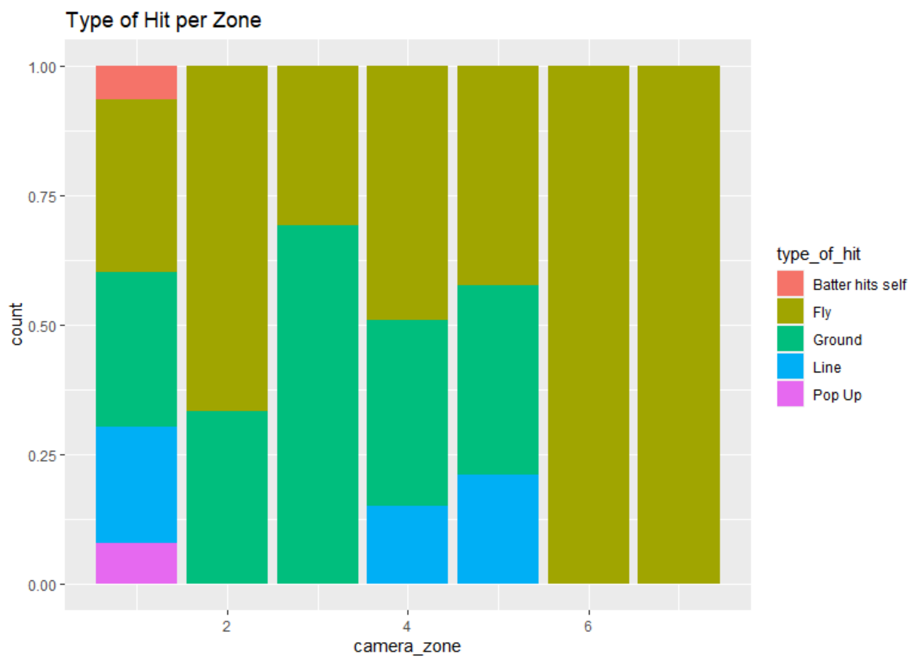
```
ggplot(data=FoulBalls2, mapping=aes(x=camera_zone,
fill=type_of_hit))
     + geom_density() +
     facet_grid(row=vars(type_of_hit)) +
     labs(title="Type of Hit per Zone")
```



This graph shows the density of each type of foul ball hit for the seven different camera zones. “Batter hits self” and “Pop Up” hits were most likely to land in zone 1 or zone 2. These types of hit are low risk for spectators, as those zones are the most protected from nets, and the least likely to go into the stands. “Fly,” “Ground,” and “Line” hits are less predictable, and are more likely to land in unprotected zones. “Fly” and “Line” hits are more likely to land in the stands than “Ground” hits, making them the more dangerous types of hit.

Camera Zone Bar Graph:

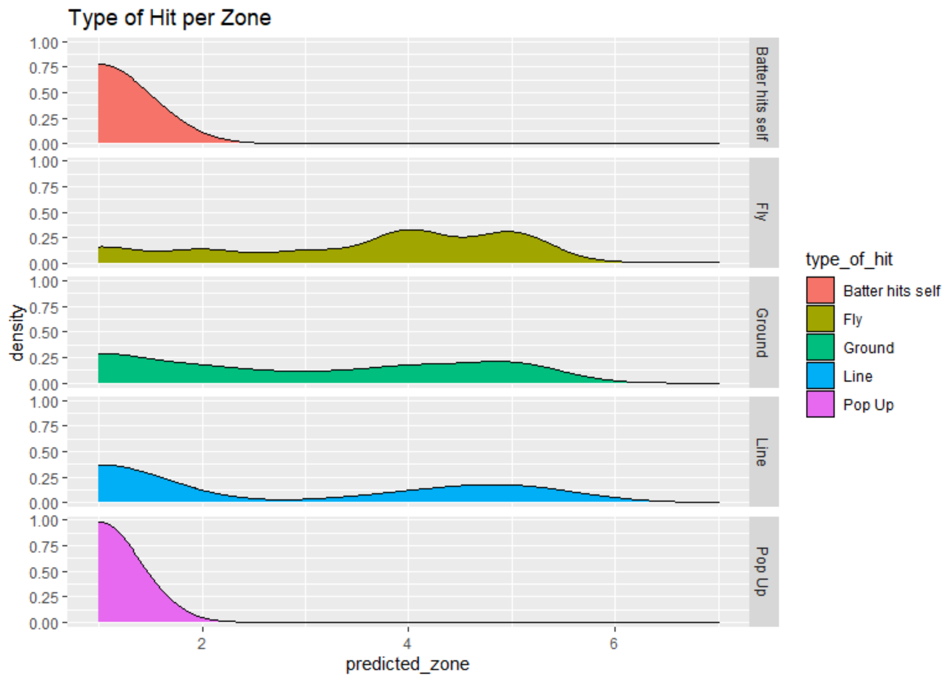
```
ggplot(data=FoulBalls2, mapping=aes(x=camera_zone,
fill=type_of_hit))
     + geom_bar(position='fill') +
     labs(title="Type of Hit per Zone")
```



This bar graph compares the type of hit for each camera zone. The graph shows that “Fly” hits are common in all seven zones, and the “Batter hits self” and “Pop Up” hits are exclusively located in zone 1. Zone 1 remains the safest place to spectate the game, as it is the most protected area. Zones 6 and 7 are populated exclusively by Fly balls as no other foul ball has the power to reach such lengths. For the most part, the further away from home plate a zone is the less diverse the type of hits become.

Predicted Zone Density Plot:

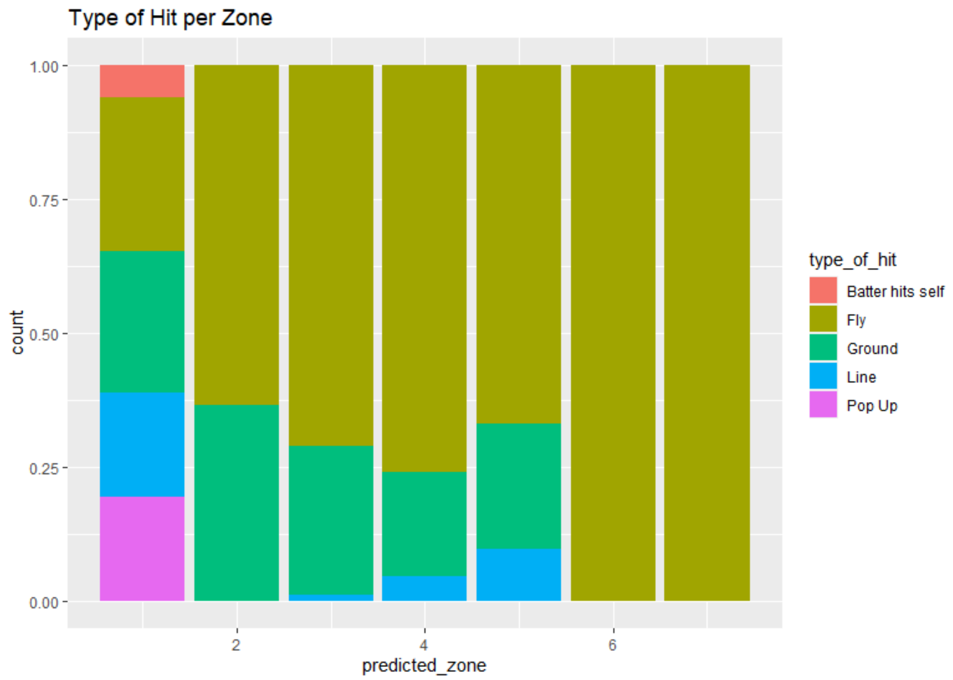
```
ggplot(data=FoulBalls3, mapping=aes(x=predicted_zone,
fill=type_of_hit)) +
     geom_density() +
     facet_grid(row=vars(type_of_hit))+labs(title="Type of Hit
per Zone")
```



This graph shows the density of each type of foul ball hit for the seven different predicted landing zones. “Batter hits self” and “Pop Up” hits were most likely to land in zone 1 or zone 2. These types of hit are low risk for spectators, as those zones are the most protected from nets, and the least likely to go into the stands. “Fly,” “Ground,” and “Line” hits are less predictable, and are more likely to land in unprotected zones. “Fly” and “Line” hits are more likely to land in the stands than “Ground” hits, making them the more dangerous types of hit.

Bar Graph:

```
ggplot(data=FoulBalls3, mapping=aes(x=predicted_one,
fill=type_of_hit)) +
     geom_bar(position='fill') +
     labs(title="Type of Hit per Zone")
```



This bar graph compares the type of hit for each predicted landing zone. The graph shows that "Fly" hits are common in all seven zones, and the "Batter hits self" and "Pop Up" hits are exclusively located in zone 1. Zone 1 remains the safest place to spectate the game, as it is the most protected area. Zones 6 and 7 are populated exclusively by Fly balls as no other foul ball has the power to reach such lengths. For the most part, the further away from home plate a zone is the less diverse the type of hits become.

Conclusions

Based on the concentration of fly balls and lines, it would seem that the zones most likely to end up with dangerous foul balls are zones 1-5. Because zone 1 is adjacent to home plate, it receives the most protection out of any zone for the spectators with netting. This makes zone 1 one of the safest zones from foul balls because the fans stay protected behind the barrier. That makes the most interesting and most dangerous zones include zones 2-5. If you want to catch a foul ball we suggest a seat in zone 2-5, but if you want to avoid an unwanted bruise zones 1 and 7 seem far safer. The camera zone graphs and predicted zone graphs look extremely similar to each other with very deviation. This means the model used to predict the landing zone is very accurate and that it is very uncommon for the model to misjudge where the foul ball will land.

Limitations and Recommendations

Our data is limited by the sample size of the foul balls. The data was taken from the most foul ball heavy games in the season which probably means there was a reason for the most foul balls such as the weather or field adaptations. For instance if it rained one of the games the heavier balls could cause more batters to foul them off instead of hitting them in play. Our recommendation for conducting this experiment in the future is to include weather in the data set and include multiple other games per stadium despite how rigorous the data might be. This will give us a better sample size and weed out variables such as weather and field differences.

```
ggplot(data=f, mapping=aes(x=used_zone, fill =
exitvcategory))+geom_bar(position='fill')

f <- FoulBalls %>%
  mutate( exitvcategory = ifelse(exit_velocity > 90, "high risk", NA),
          exitvcategory = ifelse(exit_velocity >= 60 & exit_velocity <= 90, "medium risk",
exitvcategory),
          exitvcategory = ifelse(exit_velocity < 60, "low risk", exitvcategory))
```