Significance Tests & Visualizations

Luke Welsh

08-03-2022

Motivation: Many people have claimed that the MLB changes their baseballs for nationally televised games in order to produce more home runs, which makes games more entertaining and get more viewers. This idea has been referred to as ‘juicing baseballs.’ In this document, I will use MLB pitch-by-pitch data along with the channel that each game was broadcasted on in order to take a statistical stance on this heated topic.

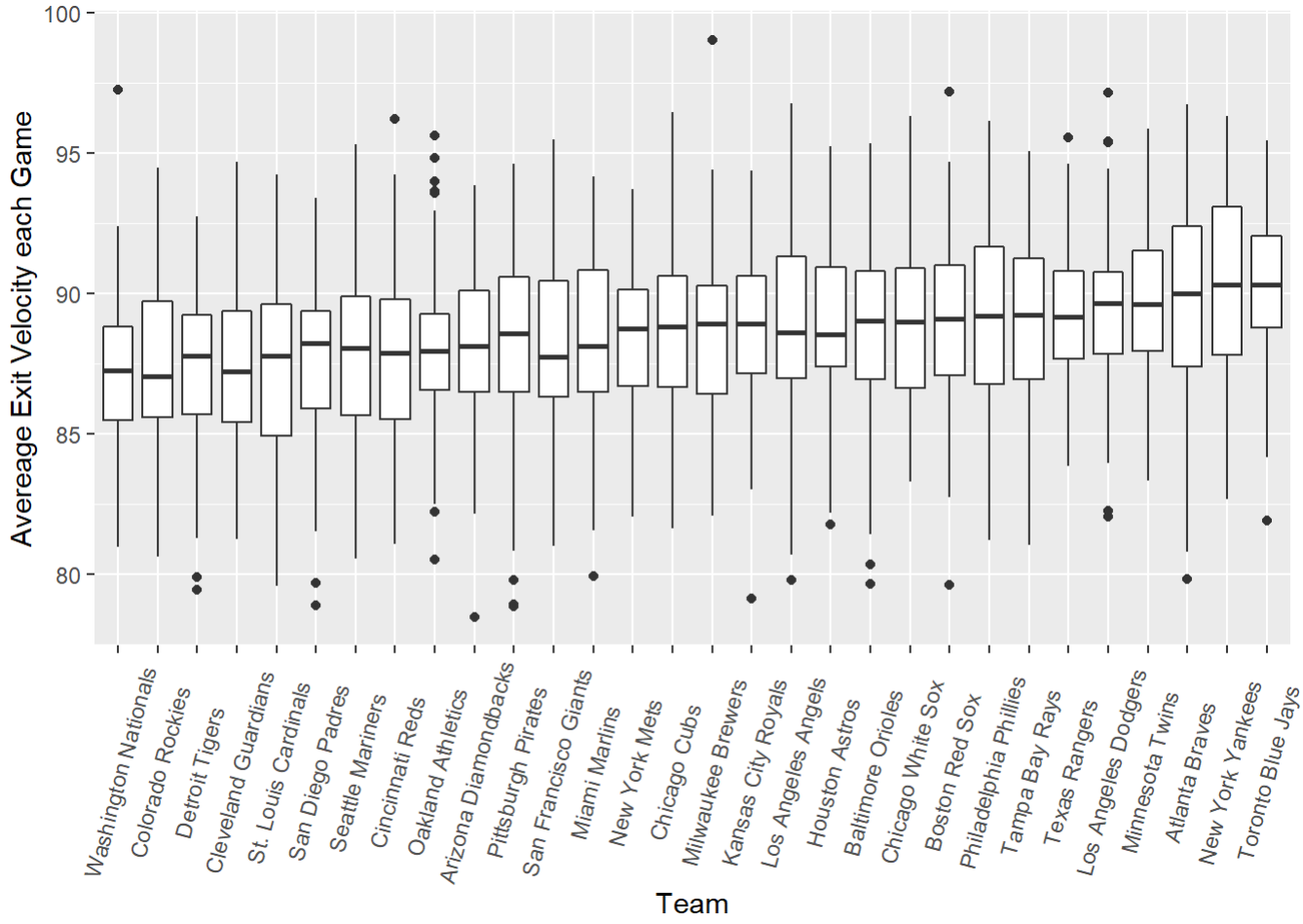
Background: The secret has been out that baseball is losing viewers and money due to the slow pace of play. There are more and more strikeouts every year which means less balls in play. Starting in 2019, it was heavily speculated that baseballs were vastly different than ever before as uncanny amounts of home runs were hit. In 2019, four separate teams broke the previous record for home runs in a season, the Minnesota Twins being the first to break it with a full month of baseball play. It was identified later that, league wide, there were new baseballs being used. Later, in 2020, two different brands of balls had to be used due to supply chain issues resultant of COVID-19. It has been debated and explained that these balls performed very differently. The following year, the idea that nationally broadcasted games had balls that went further began; highlighted by the Field of Dreams game. This game was the first of its kind as it brought the Yankees and White Sox to Iowa for a heavily advertised game in which the league paid homage to the all-time classic movie: *Field of Dreams*. The game ended with a score of 9-8 with 8 home runs including one to walk it off for the White Sox. This brings us to the 2022 season. Follow along below to find out if there is a statistical difference among the average exit velocities between games that are locally and nationally televised.

Disclaimer: This is not polished due to running out of time. Please forgive any poor graph labels

```
library(ggplot2)
all_games = read.csv("all_game_sum.csv")
all_games_tv = read.csv("all_game_tv.csv")
```

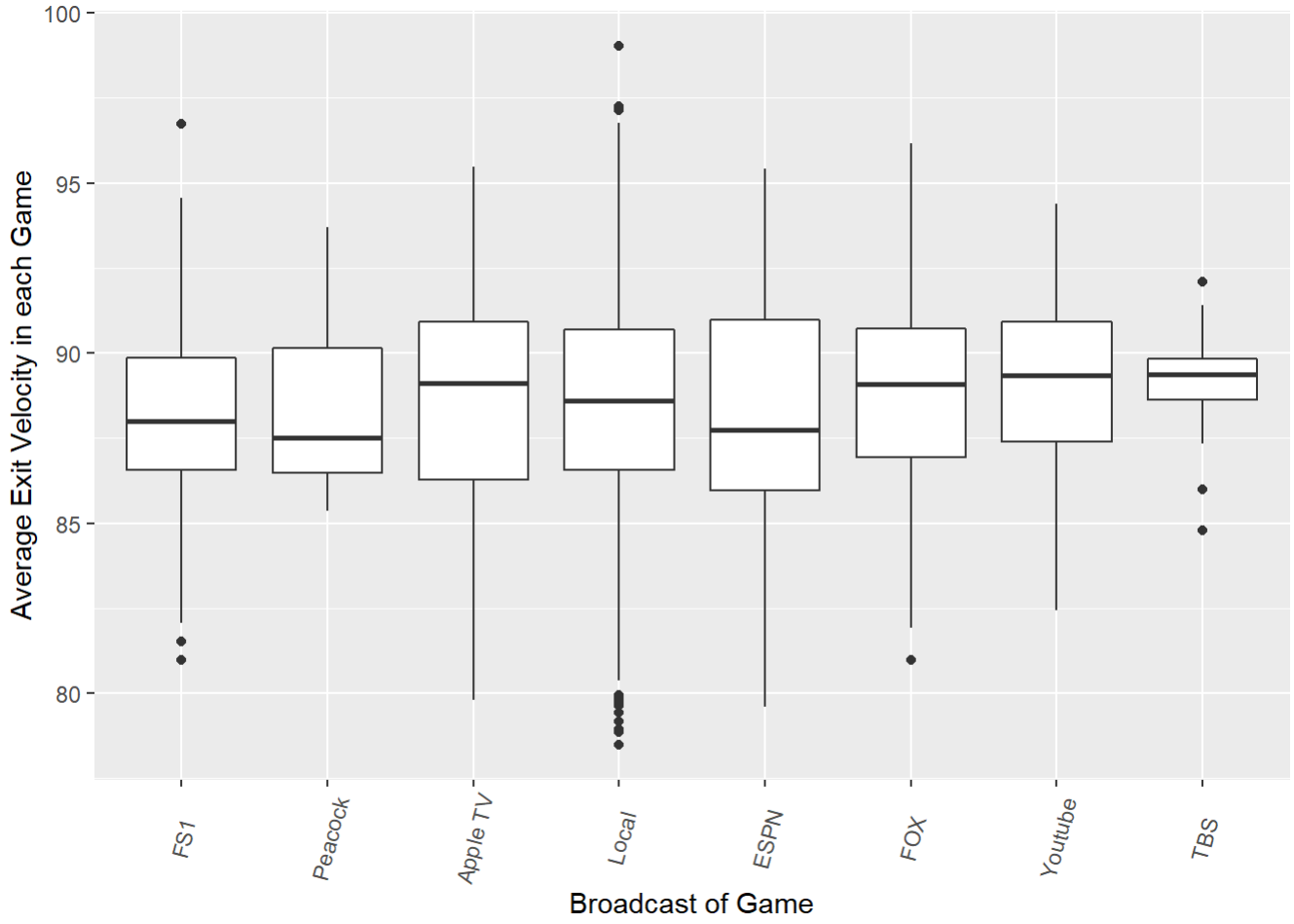
Showing variation in average exit velocity between different teams. Ascending mean from left to right.

```
ggplot(data=all_games, mapping = aes(x=reorder(batting_team, velo), y=velo)) +
  geom_boxplot()+
  theme(axis.text.x = element_text(angle=75, vjust = 0.5))+
  labs(x = 'Team', y = 'Avereage Exit Velocity each Game')
```



Showing variation in average exit velocity by network. Ascending mean from left to right.

```
ggplot(data=all_games_tv, mapping = aes(x=reorder(tv_provider, velo), y=velo)) +
  geom_boxplot()+
  theme(axis.text.x = element_text(angle=75, vjust = 0.5))+
  labs(x='Broadcast of Game', y = 'Average Exit Velocity in each Game')
```



Visualizing distribution of the statistic game_diff as season goes on.

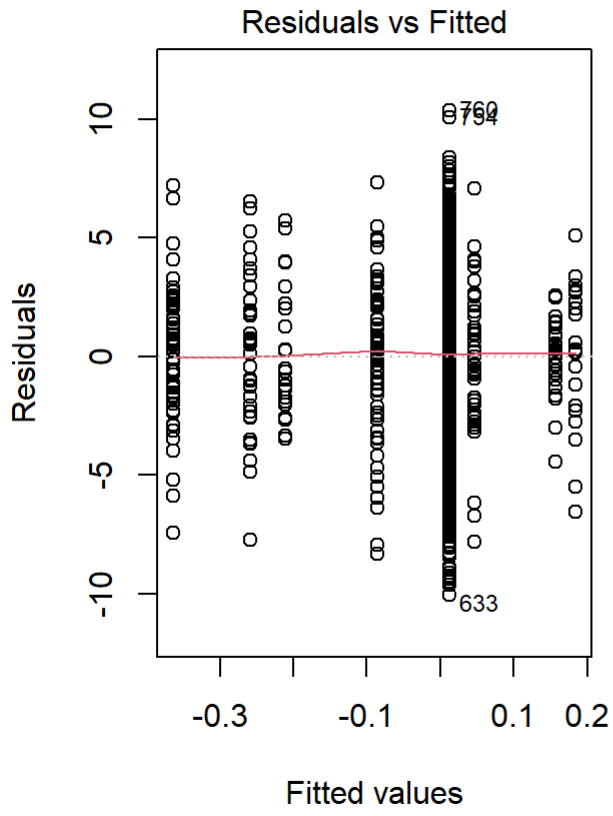
```
ggplot(all_games_tv)+
geom_point(aes(x = X/2, y = game_diff, color = tv_provider), alpha = 0.5)+
labs(x='Game Number in Season', y="Exit Velocity difference from team's season average")
```

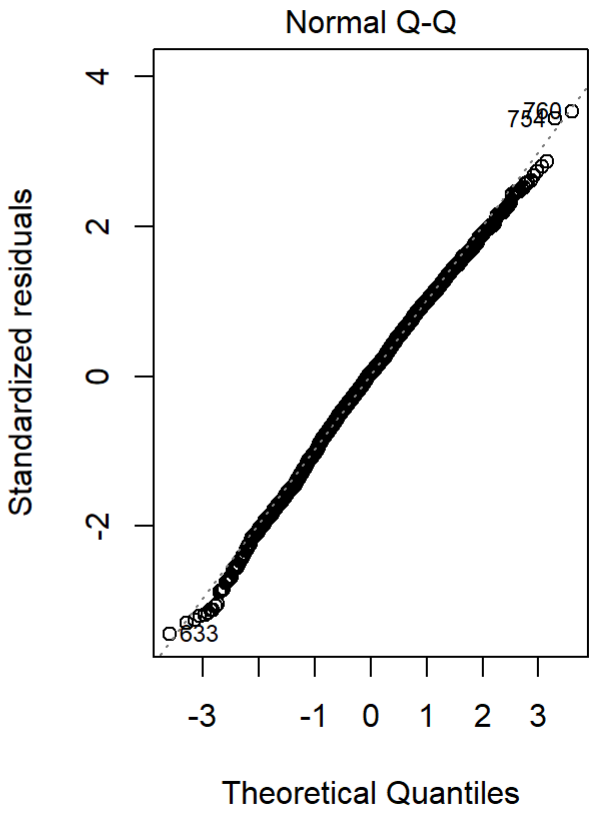
Here, in the ANOVA test, we will test with significance level alpha = 0.05

```
aov.tv = aov(game_diff ~ tv_provider, data = all_games_tv)
summary(aov.tv)
```

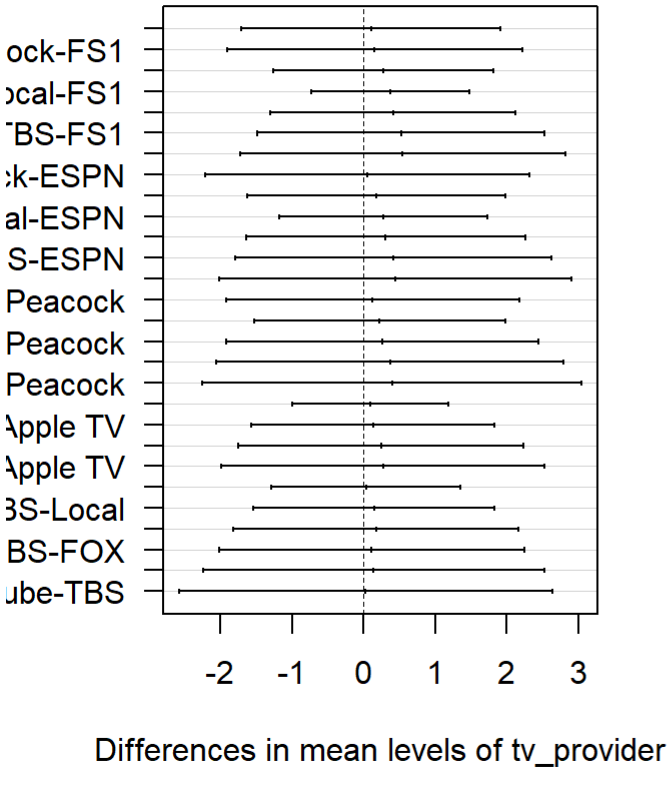
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## tv_provider    7     15   2.117   0.246  0.973
## Residuals   3102  26638   8.588
```

```
par(mfrow=c(1,2))
plot(aov.tv,which=1:2,ask=F)
```






```
plot(TukeyHSD(aov.tv, ordered = TRUE, conf.level = 0.95), las=1)
```



In our ANOVA summary, we can see that the p value is 0.973, which is greater than the significance level. So we fail to reject the null hypothesis that exit velocities remain constant between games broadcast on different channels. Further, a visualization of the comparisons between each channel is shown, and all ranges contain the value 0, meaning that all differences can be attributed to random chance.

Conclusion: Further steps that could be made in this investigation would be to (1) look more at the 2021 season which is where more public speculation is and (2) create an Exit Velocity Over Expected Machine Learning Algorithm to get a better prediction, pitch-by-pitch, of which balls are getting hit harder than expected.

Thank You for Reading and I hope that you enjoyed!