Fan Ratings

Steven Miller

April 6, 2019

## Fan Ratings Dataset

ratings <- read.csv('data/fan\_ratings.csv')

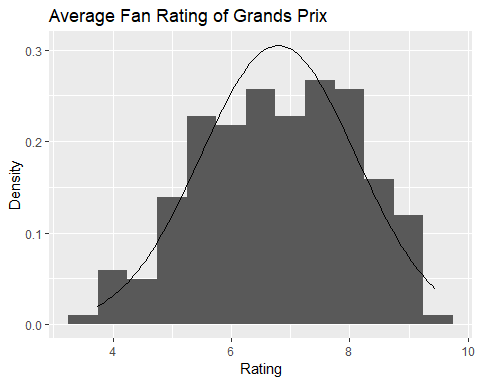
Summary of Fan Ratings:

summary(ratings)

## Y R GPNAME P1   
## Min. :2008 Min. : 1.000 Australian GP: 11 Hamilton:62   
## 1st Qu.:2010 1st Qu.: 5.000 Belgian GP : 11 Vettel :52   
## Median :2013 Median :10.000 British GP : 11 Rosberg :22   
## Mean :2013 Mean : 9.822 Chinese GP : 11 Button :14   
## 3rd Qu.:2016 3rd Qu.:14.000 Hungarian GP : 11 Alonso :13   
## Max. :2018 Max. :21.000 Monaco GP : 11 Webber : 9   
## (Other) :136 (Other) :30   
## P2 P3 RATING   
## Hamilton :30 Raikkonen:26 Min. :3.740   
## Vettel :29 Vettel :26 1st Qu.:5.777   
## Rosberg :25 Hamilton :21 Median :6.812   
## Alonso :19 Alonso :16 Mean :6.793   
## Raikkonen:19 Ricciardo:15 3rd Qu.:7.801   
## Webber :16 Webber :15 Max. :9.449   
## (Other) :64 (Other) :83

Y - Year of Race  
R - Round of Race in Season  
GPNAME - Name of Race  
P1 - Race Winner  
P2 - Runner-Up  
P3 - Third Place  
RATING - Average fan rating on a scale from 1 to 10

ggplot(ratings, aes(RATING)) + geom\_histogram(aes(y=..density..), binwidth = 0.5) +   
 stat\_function(fun = dnorm, args = list(mean = mean(ratings$RATING), sd = sd(ratings$RATING))) +   
 labs(title="Average Fan Rating of Grands Prix", y="Density", x="Rating")



Is it possible that certain drivers winning a race would have an impact on the fan rating? It seems like a possibility to me. A fan of a certain driver might be more likely to give a race a higher rating if they enjoyed the race. Looking at the summary of the ratings table, it appears that the five winningest drivers in the data set were Hamilton, Vettel, Rosberg, Button, and Alonso. We also know that the mean rating is 6.793. Let’s see what the average rating is for races where these drivers won.

agg <- aggregate(ratings$RATING, by = list(ratings$P1), FUN = mean)  
agg <- agg[order(-agg$x),]  
names(agg) <- c("Driver","Rating")  
print(agg)

## Driver Rating  
## 8 Maldonado 8.274000  
## 11 Ricciardo 8.108571  
## 7 Kubica 7.809000  
## 13 Verstappen 7.732500  
## 4 Button 7.307286  
## 10 Raikkonen 7.045400  
## 5 Hamilton 6.945645  
## 1 Alonso 6.799077  
## 14 Vettel 6.569173  
## 15 Webber 6.517000  
## 12 Rosberg 6.449364  
## 2 Barichello 6.202000  
## 6 Kovalainen 6.202000  
## 9 Massa 6.104333  
## 3 Bottas 4.836667

Recalling that our mean rating was 6.793 it appears that the five winningest drivers are all pretty close to this mean. From this, one could hypothesize that fans find it more or less exciting when a driver who doesn’t typically win races wins.

## Ergast F1 Dataset

The first variable we’ll take a look at is the duration of pit stops. Throughout many seasons, the rules regarding pit stops have changed. Over time, various seasons of racing will have had various average durations for pit stops. This can be a matter of seconds but leads to big changes in strategic decision making that can impact the way a race plays out.

pitstops <- read.csv('data/pit\_stops.csv', header=FALSE, col.names = c("raceId","driverId","stop","lap","time","duration","milliseconds"))  
summary(pitstops)

## raceId driverId stop lap   
## Min. : 841.0 Min. : 1.0 Min. :1.000 Min. : 1.00   
## 1st Qu.: 871.0 1st Qu.: 16.0 1st Qu.:1.000 1st Qu.:13.00   
## Median : 909.0 Median :807.0 Median :2.000 Median :24.00   
## Mean : 915.9 Mean :447.4 Mean :1.793 Mean :24.83   
## 3rd Qu.: 959.0 3rd Qu.:822.0 3rd Qu.:2.000 3rd Qu.:35.00   
## Max. :1011.0 Max. :848.0 Max. :6.000 Max. :74.00   
##   
## time duration milliseconds   
## 14:56:46: 5 22.303 : 6 Min. : 12897   
## 14:05:02: 4 22.838 : 6 1st Qu.: 21811   
## 14:18:36: 4 21.012 : 5 Median : 23375   
## 14:19:03: 4 21.900 : 5 Mean : 45738   
## 14:20:17: 4 22.105 : 5 3rd Qu.: 25567   
## 14:20:51: 4 22.273 : 5 Max. :2011266   
## (Other) :6824 (Other):6817

head(pitstops)

## raceId driverId stop lap time duration milliseconds  
## 1 841 153 1 1 17:05:23 26.898 26898  
## 2 841 30 1 1 17:05:52 25.021 25021  
## 3 841 17 1 11 17:20:48 23.426 23426  
## 4 841 4 1 12 17:22:34 23.251 23251  
## 5 841 13 1 13 17:24:10 23.842 23842  
## 6 841 22 1 13 17:24:29 23.643 23643

raceId - the database id of the race the pit stop took place in driverId - the database id of the driver making a pit stop stop - the stop number out of all stops a driver made in a particular race lap - the lap number the pit stop was conducted on time - the time of day the pit stop took place duration - how long the pit stop took milliseconds - duration, but in milliseconds

We can see from the summary that there are some vast outliers in the milliseconds column. We have a max value of almost forty minutes and a minimum value of 12.897 seconds. The mean is 45.738 and a median of 23.375. Based on these values, it’s clear that the outliers are skewing the data. With a little bit of domain knowledge, I can safely say that any pit stops taking that long were not done by vehicles who were actually competitive during the race and can pretty safely be excluded from further analysis.

stopsd <- sd(pitstops$milliseconds)  
print(stopsd)

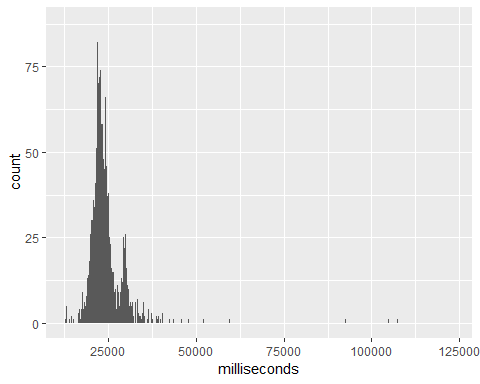
## [1] 171321.9

Looking at the standard deviation of pit stop duration, it appears to be 171 milliseconds. I feel comfortable excluding any values greater than one-half standard deviation over the mean from further analysis, as this would indicate pit stops over two minutes in length.

pitstops <- pitstops[pitstops$milliseconds<=mean(pitstops$milliseconds)+stopsd/2,]

With those values removed from the data, we can now further analyze the duration.

ggplot(pitstops, aes(milliseconds)) + geom\_histogram(binwidth = 60)



We can now see the peak around 25 seconds as well as a second, smaller spike around 30 seconds. This is likely due to rule changes at some point that either sped up or slowed down the average pit stop duration. It would be interesting, I think, to see how the average time has changed over the years. If we combine some tables together we can look into this further.

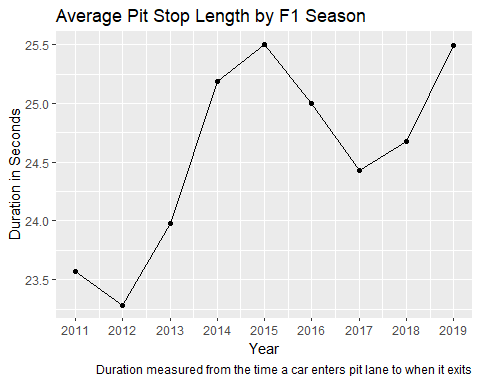
## F1 races table

races <- read.csv('data/races.csv', header=FALSE, col.names=c("raceId","year","round","circuitId","name","date","time","url"))  
head(races)

## raceId year round circuitId name date time  
## 1 1 2009 1 1 Australian Grand Prix 2009-03-29 06:00:00  
## 2 2 2009 2 2 Malaysian Grand Prix 2009-04-05 09:00:00  
## 3 3 2009 3 17 Chinese Grand Prix 2009-04-19 07:00:00  
## 4 4 2009 4 3 Bahrain Grand Prix 2009-04-26 12:00:00  
## 5 5 2009 5 4 Spanish Grand Prix 2009-05-10 12:00:00  
## 6 6 2009 6 6 Monaco Grand Prix 2009-05-24 12:00:00  
## url  
## 1 http://en.wikipedia.org/wiki/2009\_Australian\_Grand\_Prix  
## 2 http://en.wikipedia.org/wiki/2009\_Malaysian\_Grand\_Prix  
## 3 http://en.wikipedia.org/wiki/2009\_Chinese\_Grand\_Prix  
## 4 http://en.wikipedia.org/wiki/2009\_Bahrain\_Grand\_Prix  
## 5 http://en.wikipedia.org/wiki/2009\_Spanish\_Grand\_Prix  
## 6 http://en.wikipedia.org/wiki/2009\_Monaco\_Grand\_Prix

By joining the year data from the races table to the pit stops table by matching the raceId of the pit stop, we can now summarize the pit stop data to a mean value by year.

pitstops <- merge(x=pitstops, y=races, by="raceId", all.x = TRUE)  
agg <- aggregate(pitstops$milliseconds, by = list(pitstops$year), FUN = mean)  
names(agg) <- c("Year","Milliseconds")  
ggplot(agg, aes(x=Year, y=Milliseconds/1000)) + geom\_line() + geom\_point() + labs(y="Duration in Seconds", title="Average Pit Stop Length by F1 Season", caption="Duration measured from the time a car enters pit lane to when it exits") + scale\_x\_continuous(breaks=seq(2011,2019,1))



From this graph, it appears that the variance from season to season is minimal and unlikely to be a reason for changes in fan ratings.