Section 3

Steven Miller

April 17, 2019

It's hard to display my final dataset in a single table, as it's more of a database spread across multiple tables. What I'll be doing instead is getting my data into a form that's useful for analysis of some of the questions I wanted to investigate.

The first question I want to look at was in regards to how the length of a race track impacts the fan rating of races held at the circuit. At the very minimum, I will need the circuits table from my primary dataset, and the fan ratings table.

```
circuits <- read.csv('data/circuits.csv')
fan_ratings <- read.csv('data/fan_ratings.csv')

head(circuits)
```

```
##   circuitId  circuitRef                          name     location
## 1         1 albert_park Albert Park Grand Prix Circuit    Melbourne
## 2         2      sepang   Sepang International Circuit Kuala Lumpur
## 3         3     bahrain  Bahrain International Circuit       Sakhir
## 4         4   catalunya Circuit de Barcelona-Catalunya    MontmelÌ_
## 5         5    istanbul                 Istanbul Park     Istanbul
## 6         6      monaco              Circuit de Monaco  Monte-Carlo
##     country       lat       lng alt
## 1 Australia -37.84970 144.96800  10
## 2  Malaysia   2.76083 101.73800  NA
## 3   Bahrain  26.03250  50.51060  NA
## 4     Spain  41.57000   2.26111  NA
## 5    Turkey  40.95170  29.40500  NA
## 6    Monaco  43.73470   7.42056  NA
##                                                      url
## 1   http://en.wikipedia.org/wiki/Melbourne_Grand_Prix_Circuit
## 2   http://en.wikipedia.org/wiki/Sepang_International_Circuit
## 3  http://en.wikipedia.org/wiki/Bahrain_International_Circuit
## 4 http://en.wikipedia.org/wiki/Circuit_de_Barcelona-Catalunya
## 5                 http://en.wikipedia.org/wiki/Istanbul_Park
## 6              http://en.wikipedia.org/wiki/Circuit_de_Monaco
```

```
head(fan_ratings)
```

```
##      Y  R        GPNAME         P1         P2        P3 RATING
## 1 2008  1 Australian GP   Hamilton   Heidfeld   Rosberg  7.609
## 2 2008 10     German GP   Hamilton     Piquet     Massa  7.180
## 3 2008 11  Hungarian GP Kovalainen      Glock Raikkonen  6.202
## 4 2008 12   European GP      Massa   Hamilton    Kubica  3.977
## 5 2008 13    Belgian GP      Massa   Heidfeld  Hamilton  7.736
## 6 2008 14    Italian GP     Vettel Kovalainen    Kubica  8.153
```

Taking a look at each table, I notice an immediate issue. There is no column that will easily join the data from one table to another. We will need additional data.

```
races = read.csv('data/races.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

```
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
```

The races table contains the year and round number of the event, as does the fan rating column. Using this data along with the circuit_id value should be enough to get us fan scores broken down by each circuit. To keep our merge function simple, I'll create a new column in each table that contains the year and round number concatentated. This is the value I'll use to join the two tables together.

```
races$yr <- paste(races$year,races$round)
fan_ratings$yr <- paste(fan_ratings$Y, fan_ratings$R)
new_frame <- merge(x = fan_ratings, y = races, by="yr", all.x = TRUE)
new_frame <- new_frame[,c("year","round","circuitId","RATING")]
head(new_frame)
```

```
##   year round circuitId RATING
## 1 2008     1         1  7.609
## 2 2008    10        10  7.180
## 3 2008    11        11  6.202
## 4 2008    12        12  3.977
## 5 2008    13        13  7.736
## 6 2008    14        14  8.153
```

Now that all of the data has been joined into a single, useful table, I can aggregate the data to get an average race rating based on the circuit.

```
rating_by_circuit <- new_frame %>% group_by(circuitId) %>% summarize(mean_rating = mean(RATING))
rating_by_circuit <- merge(x = rating_by_circuit, y = circuits, by="circuitId", all.x = TRUE)
truncated_rating_bc <- rating_by_circuit[,c("name","mean_rating")]
truncated_rating_bc <- truncated_rating_bc[order(-truncated_rating_bc$mean_rating),]
print(truncated_rating_bc)
```

```
##                             name mean_rating
## 19                   NÌ_rburgring    7.723000
## 26        Circuit of the Americas    7.398000
## 9             Silverstone Circuit    7.363091
## 29              Baku City Circuit    7.360000
## 7       Circuit Gilles Villeneuve    7.330800
## 17 Shanghai International Circuit    7.263273
## 18   AutÌ_dromo JosÌ© Carlos Pace    7.241200
## 13   Circuit de Spa-Francorchamps    7.162091
```

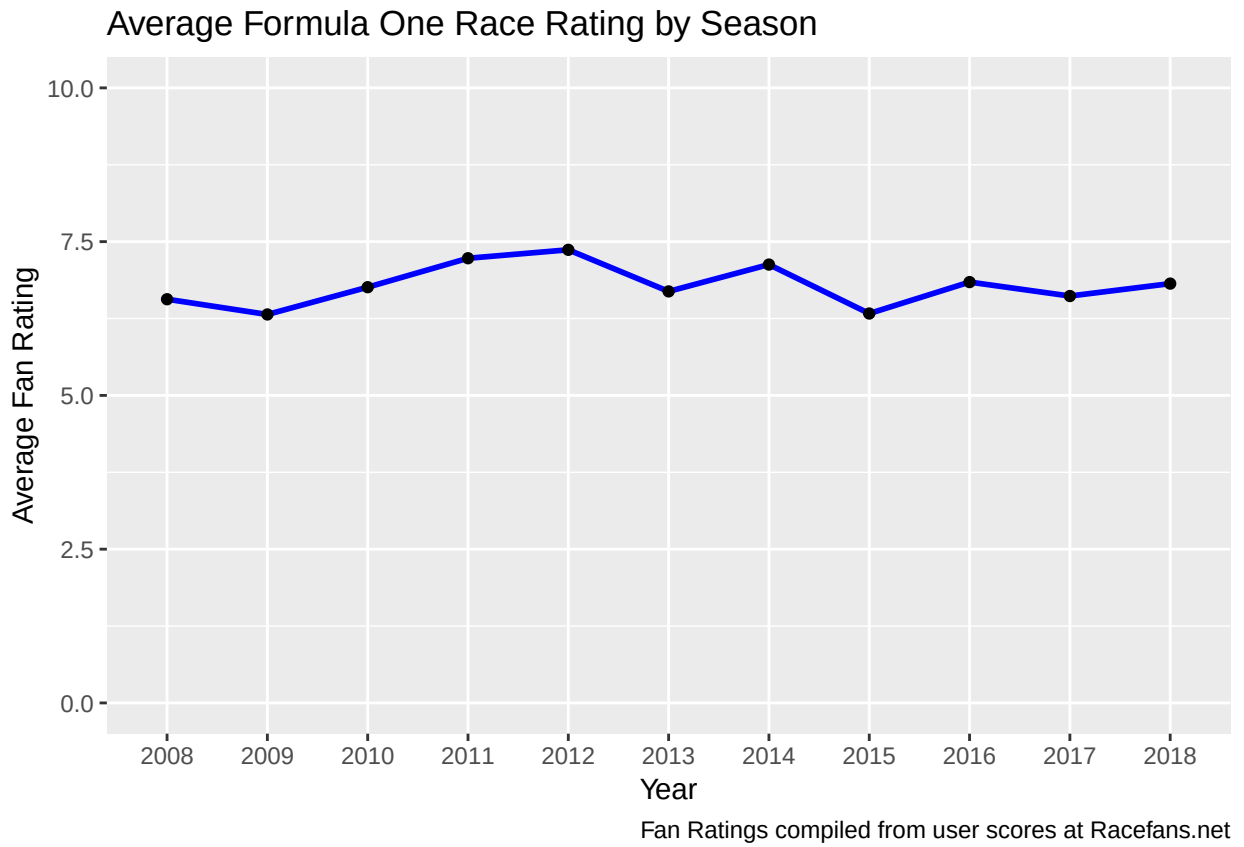
```
## 3   Bahrain International Circuit    7.120400
## 1  Albert Park Grand Prix Circuit    7.114727
## 2    Sepang International Circuit    7.047900
## 11                    Hungaroring    7.002727
## 5                   Istanbul Park    6.845500
## 27                  Red Bull Ring    6.836000
## 24   Korean International Circuit    6.740000
## 14   Autodromo Nazionale di Monza    6.688500
## 16                  Fuji Speedway    6.660000
## 10                Hockenheimring    6.642500
## 23            Circuit Paul Ricard    6.470000
## 20                 Suzuka Circuit    6.403000
## 15      Marina Bay Street Circuit    6.374300
## 4  Circuit de Barcelona-Catalunya    6.354000
## 6               Circuit de Monaco    6.344545
## 21             Yas Marina Circuit    6.166000
## 22 AutÌ_dromo Hermanos RodrÌ_guez    6.050000
## 25    Buddh International Circuit    5.750333
## 12        Valencia Street Circuit    5.488200
## 28                 Sochi Autodrom    5.310000
## 8   Circuit de Nevers Magny-Cours    3.977000
```

I now have the final table for the analysis of fan ratings by circuit. It appears that the Nurburgring has the highest average rating, while Magny-Cours has the loweset. At this point, I have realized that while my initial question was going to examine the impact the circuit length had on scores, I do not presently have that information available. I do believe that I can retrieve it, along with some information about weather, by scraping Wikipedia.

The next question I'll need to prepare data for is "How have rule changes in recent years impacted the quality of races?". I don't currently have data on rule changes, but these typically take place in between seasons. A summary of the average ratings of races by season will be sufficient for an initial analysis.

```
rating_by_season <- fan_ratings %>% group_by(Y) %>% summarize(mean_rating = mean(RATING))

ggplot(rating_by_season, aes(Y,mean_rating)) + geom_line(color="blue", size=1) + geom_point() + scale_x
```




```
print(rating_by_season)
```

```
## # A tibble: 11 x 2
##        Y mean_rating
##    <int>       <dbl>
##  1  2008        6.56
##  2  2009        6.32
##  3  2010        6.76
##  4  2011        7.23
##  5  2012        7.37
##  6  2013        6.69
##  7  2014        7.13
##  8  2015        6.33
##  9  2016        6.84
## 10  2017        6.62
## 11  2018        6.82
```

The average ratings actually look pretty consistent from year to year. Based on these, we probably don't need to do further analysis on how the number of winners in a season impacts season ratings. The variance is too small for a meaningful impact.

The next thing to look at is how different teams and drivers impact the fan ratings.

```
rating_by_winner <- fan_ratings %>% group_by(P1) %>% summarize(mean_rating = mean(RATING))
print(rating_by_winner[order(rating_by_winner$mean_rating, decreasing=TRUE),])
```

```
## # A tibble: 15 x 2
##    P1         mean_rating
```

```
##    <fct>            <dbl>
##  1 Maldonado         8.27
##  2 Ricciardo         8.11
##  3 Kubica            7.81
##  4 Verstappen        7.73
##  5 Button            7.31
##  6 Raikkonen         7.05
##  7 Hamilton          6.95
##  8 Alonso            6.80
##  9 Vettel            6.57
## 10 Webber            6.52
## 11 Rosberg           6.45
## 12 Barichello        6.20
## 13 Kovalainen        6.20
## 14 Massa             6.10
## 15 Bottas            4.84
```

It appears that there's a pretty significant variation in rating based on the winner. What about specific teams?

```
results <- read.csv('data/results.csv')
results <- results[results$positionOrder==1,]
constructors <- read.csv('data/constructors.csv')
results <- merge(x = results, y = constructors, by="constructorId", all.x = TRUE)
results <- merge(x = results, y = races, by="raceId", all.x = TRUE)
results <- merge(x = results, y = fan_ratings, by="yr", all.x = TRUE)
results <- results[complete.cases(results$RATING),]
rating_by_const <- results %>% group_by(name.x) %>% summarize(mean_rating = mean(RATING))
rating_by_const <- rating_by_const[complete.cases(rating_by_const),]
print(rating_by_const[order(rating_by_const$mean_rating , decreasing=TRUE),])
```

```
## # A tibble: 10 x 2
##    name.x     mean_rating
##    <fct>            <dbl>
##  1 Lotus F1          8.28
##  2 Williams          8.27
##  3 Toro Rosso        8.15
##  4 BMW Sauber        7.81
##  5 McLaren           7.69
##  6 Red Bull          6.71
##  7 Ferrari           6.67
##  8 Mercedes          6.55
##  9 Renault           6.48
## 10 Brawn             6.02
```

We can see a pretty big discrepancy in ratings between winning constructors as well.

```
model <- lm(formula = RATING ~ name.x + P1 + GPNAME, data=results)
summary(model)
```

```
##
## Call:
## lm(formula = RATING ~ name.x + P1 + GPNAME, data = results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.2175 -0.6889  0.0781  0.6243  3.1678
##
## Coefficients: (3 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.79012    1.28178   5.297 4.34e-07 ***
## name.xBrawn             -2.44873    1.46043  -1.677   0.0958 .
## name.xFerrari           -0.44108    1.27345  -0.346   0.7296
## name.xLotus F1           1.94629    1.68035   1.158   0.2487
## name.xMcLaren           -0.16906    1.28622  -0.131   0.8956
## name.xMercedes          -0.85126    1.24762  -0.682   0.4961
## name.xRed Bull          -0.77534    1.36144  -0.569   0.5699
## name.xRenault           -0.49622    1.48910  -0.333   0.7394
## name.xToro Rosso         1.05924    1.81653   0.583   0.5607
## name.xWilliams           1.23251    1.71849   0.717   0.4744
## P1Barichello             2.06111    1.16525   1.769   0.0791 .
## P1Bottas                -1.26425    0.76266  -1.658   0.0996 .
## P1Button                 0.94589    0.59685   1.585   0.1152
## P1Hamilton               0.16098    0.32104   0.501   0.6169
## P1Kovalainen            -1.00858    1.27745  -0.790   0.4311
## P1Kubica                     NA         NA      NA       NA
## P1Maldonado                  NA         NA      NA       NA
## P1Massa                 -0.46421    0.65490  -0.709   0.4796
## P1Raikkonen             -0.79714    0.78008  -1.022   0.3086
## P1Ricciardo              1.67614    0.87173   1.923   0.0565 .
## P1Rosberg                    NA         NA      NA       NA
## P1Verstappen             1.43050    0.94505   1.514   0.1323
## P1Vettel                -0.02825    0.56392  -0.050   0.9601
## P1Webber                -0.26569    0.74490  -0.357   0.7219
## GPNAMEAustralian GP      0.67347    0.54580   1.234   0.2193
## GPNAMEAustrian GP        0.87946    0.71264   1.234   0.2192
## GPNAMEAzerbaijan GP      1.10908    1.35613   0.818   0.4148
## GPNAMEBahrain GP         0.98783    0.56525   1.748   0.0827 .
## GPNAMEBelgian GP         0.99048    0.55431   1.787   0.0761 .
## GPNAMEBrazilian GP       1.11907    0.55603   2.013   0.0460 *
## GPNAMEBritish GP         1.13501    0.55139   2.058   0.0414 *
## GPNAMECanadian GP        1.01888    0.58269   1.749   0.0825 .
## GPNAMEChinese GP         0.76622    0.54602   1.403   0.1627
## GPNAMEEuropean GP       -0.73288    0.64746  -1.132   0.2596
## GPNAMEFrench GP         -1.90783    1.33282  -1.431   0.1545
## GPNAMEGerman GP          0.60844    0.58336   1.043   0.2987
## GPNAMEHungarian GP       0.58953    0.57786   1.020   0.3094
## GPNAMEIndian GP         -0.23619    0.78764  -0.300   0.7647
## GPNAMEItalian GP         0.33189    0.56453   0.588   0.5575
## GPNAMEJapanese GP        0.18430    0.58571   0.315   0.7535
## GPNAMEKorean GP          0.66284    0.71398   0.928   0.3548
## GPNAMEMalaysian GP       0.77744    0.56403   1.378   0.1702
## GPNAMEMexican GP        -0.44465    0.81584  -0.545   0.5866
## GPNAMEMonaco GP          0.47688    0.55614   0.857   0.3926
## GPNAMERussian GP        -0.39328    0.70791  -0.556   0.5794
## GPNAMESingapore GP       0.18991    0.55038   0.345   0.7306
## GPNAMESpanish GP         0.25137    0.57613   0.436   0.6633
## GPNAMETurkish GP         0.86033    0.72763   1.182   0.2390
## GPNAMEUnited States GP   1.20335    0.62291   1.932   0.0554 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.142 on 143 degrees of freedom
## Multiple R-squared:  0.4025, Adjusted R-squared:  0.2145
## F-statistic: 2.141 on 45 and 143 DF,  p-value: 0.0003772
```

Looking at our summary of the model, we see only a few variables with significant p-values. Based on their coefficients, it appears that Brawn is a consistent detractor to ratings (Good thing they no longer exist!), Rubens Barichello was a solid boost to ratings (unfortunately he is retired), Valteri Bottas is bad for ratings, and Daniel Ricciardo is good for ratings. When it comes to grands prix the events with significant p-values all have positive coefficients. We can see this with Bahrain, Belgium, Brazil, the United Kingdom, Canada, and the United States.