Analysis of Formula One Fan Reviews

Name: Steven Miller

Date: 2019-04-06

Section 1

- Explain what your interests are in the data sets identified.
    - As a lifelong fan of auto racing, I was curious to take a look at data generated from auto races. I found a few data sets focused on the results of Formula One races as well as a data set of fan ratings of those races. I thought it would be interesting to take a look at the race-related data and see if any of the values there can be identified as drivers of fan ratings.
- What is the target audience for this research?
    - I believe other fans of auto racing would find the research interesting. I also believe that decision makers within auto racing could use the information to make changes and take action in the future.
- Identify the Packages that are needed for your project.
    - ggplot2 will be useful for data visualization
    - dplyr will be helpful for working with tables of information
    - pastecs will be useful for generating statistical information on data and analyzing distributions
    - rmarkdown will be critical for reporting my findings
    - sqldf will be useful for working with my primary dataset. My dataset currently contains various tables of data that will need to be joined to make it useful for analysis.
- Original source where the data was obtained is cited and, if possible, hyperlinked.
    - Two datasets were found on Kaggle, one data set is from Ergast.com.
        - http://ergast.com/mrd/db/#csv
        - https://www.kaggle.com/codingminds/formula-1-race-fan-ratings/version/1#
        - https://www.kaggle.com/cjgdev/formula-1-race-data-19502017
        - I will be collecting additional data through the scraping of Wikipedia articles to fill in gaps that exist.
- Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
    - The data was all originally collected by other fans who wanted to gain further insight into the sport. To the best of my knowledge, the data is contained in its original format will all variables intact. A quick look through the data sets did not indicate any null values at all. All values appear to be complete and intact with no imputations to fill in null values.

Section 2

- Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
    - By analyzing the results and other data from Formula One races and comparing this data to fan ratings of each race, I hope to better identify what factors contribute to a "good" race.
- Provide a concise explanation of how you plan to address this problem statement.
    - A large portion of the project will be focused on hypothesizing on factors that might be contributors to the quality of a race and seeing how those factors for each race correlate with the fan ratings.
- Discuss how your proposed approach will address (fully or partially) this problem.
    - By analyzing the correlations of these factors to the fan ratings, we can get a better understanding of which factors do and do not contribute to the quality of a race.
- List at least 6 research questions you aim to answer.
    - How have rule changes in recent years impacted the quality of races?
    - Are fans more engaged when more drivers are winning throughout a season or when a few drivers dominate?
    - Does weather play a factor in the quality of a race?
    - Are fans of a specific driver or team leading to biases in the data? Do we tend to find a higher score when specific drivers or teams win?
    - How does the circuit play a factor into the quality of a race? Are longer circuits more exciting?
    - Are races with more pit stops more or less exciting?
- Explain how your analysis may help the consumer of your research findings (recall you target audience from Section 1).
    - This analysis could help decision makers in Formula One adjust moving forward to provide a more engaging product for fans. For example, if a positive correlation is found between wet weather races and a good rating, adjustments to the calendar could be made to schedule races at times of the year when the weather is more likely to be rainy. If it is found that pit stops increase the excitement of a race, decision makers could partner with tire manufacturer Pirelli to develop softer tire compounds leading to higher degradation and more pit stops during an event.
- What types of plots and tables will help you to illustrate the findings to your research questions?
    - Histograms will be needed initially to get a better understanding of the initial data. In some cases, we will be looking at the impact a categorical value has on a quantitative value (for example, comparing the average rating of a race won by a specific rate team). In these situations, tables will be useful. Much of the analysis will focus on correlations. Correlation matrices will be helpful here, but so will scatterplots to allow the visualization of data.
- What do you not know how to do right now that you need to learn to answer your research questions?
    - The main area I need to learn more about is working with a relational database in R. The dataset I am using is spread across multiple tables and the data needed to fully present an analysis is not always present within the same table.

- I will also need to get a better understanding of data transformations through the dplyr package. I expect dplyr to be sufficient for my needs, but if I find myself needing something different I will have to learn more about the packages available for R.

Section 3

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 regard 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 concatenated. 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)
```

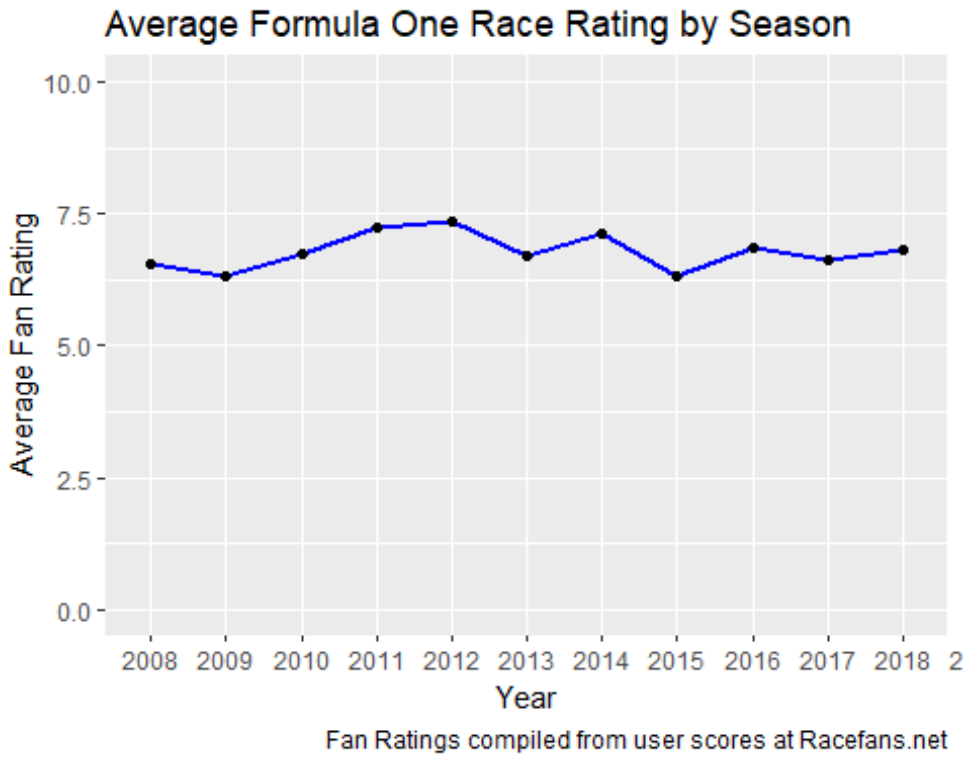
```
##                              name mean_rating
## 19                    NÌ_rburgring    7.723000
## 26         Circuit of the Americas    7.398000
## 9              Silverstone Circuit    7.363091
## 29               Baku City Circuit    7.360000
## 7        Circuit Gilles Villeneuve    7.330800
## 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 lowest. 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_discrete(name="Year",limits=c(2008:2019)) +
scale_y_continuous(name="Average Fan Rating", limits=c(0,10)) +
ggtitle("Average Formula One Race Rating by Season") + labs(caption = "Fan
Ratings compiled from user scores at Racefans.net")
```




```
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, but further analysis will still be necessary.

I feel I learned a good bit by working through this first research question. I gained a much better understanding of merging dataframes and summarizing the data with an aggregate function. Moving forward I will need to find some more data involving both circuit length and weather on race day. I believe this information is attainable through Wikipedia and will be working on finding a way to collect that data. While researching R functions, I came across a package called sqldf. It essentially allows you to work with R data frames as if they were database tables. Given the nature of my dataset, this sounds very useful and I would like to try working with it moving forward.

Section 4

By performing exploratory data analysis and analyzing relationships between variables in my data both visually and statistically, I hope to uncover new information about my data. I will be looking at the data with graphs including histograms, scatterplots and also line graphs as much of the data is time-oriented.

Each of the research questions I'm looking at will require the data to be structured in a specific way. My dataset most closely resembles the structure of a relational database. As a result, a lot of joining of separate data frames will be involved in order to create summary information for final analysis. New variables will also need to be created in some instances. For example, one of the questions I hope to answer is how the weather in a race impacts the fan rating of the event. My dataset does not have information on race weather, and I was unable to find a suitable source, however I feel the question is important and want to get an answer. To do this, I will be scraping parts of the Wikipedia entry for each event and looking at the words used to determine the weather for each event and creating new variables for analysis.

One question I was interested in was what the relationship is between the number of drivers who win in a single year and the average rating of the races in that year. My hypothesis is that more winners leads to a greater amount of unpredictability and potentially more interest. To summarize this data, I would first need to get the count of unique race winners grouped by season. From there I would also need to look at the fan ratings table and find the mean rating by season. I can then join these two tables together based on the year column and perform further analysis.

While much of the data I'll be analyzing at the final step will be easily represented in table form, scatterplots will be the most beneficial in providing a visual analysis of the relationships I am interested in gaining further insight into for most of the six research questions. Comparing the number of unique race winners in a season versus the average fan rating would be best represented with a scatterplot. Looking at the number of pit stops in a race versus fan rating would also be best accomplished with a scatterplot. Looking at rule changes over recent years would be best accomplished with a line graph, just visualizing how the average rating has changed over years. Analyzing weather would likely be best accomplished with a box plot where we can visually compare the differences between two categorical variables. This would apply for analyzing the ratings based on winning drivers as well. A good chunk of data will not be used in general simply because I don't have the fan rating data for the races represented. My data for all races dates back to 1950, however I only have data on fan ratings for 2008 and on. As a result, all data related to races from the year 2007 or before will be excluded.

Simple and multiple regression will be part of my analysis. I'll want to see how each variable individually relates to the fan rating but will also want to create a model using everything that is found to be statistically significant.

Section 5 Summary

By looking at the data I was able to identify which events, drivers and teams make an impact on the fan response to a Formula One race. By running a regression I was able to identify the variables that have a statistically significant impact on the fan response to the race.

I saw one constructor, Brawn GP, led to a significant negative impact. This makes sense, as their short tenure in the sport was quite dominating and led to races where you knew who the winner would be before the race had even started.

Looking at drivers we found that Rubens Barrichello was a significant boost to ratings, while Daniel Ricciardo offered a slightly smaller boost. Valteri Bottas, however, brings down ratings.

I also found several events that are consistently good for ratings. These include the grands prix in Belgium, the United Kingdom, Brazil, the United States, and Canada. This is likely the most actionable piece of information as we can do further analysis into the characteristics of these courses that lend themselves to good races.

This analysis was limited by the data I have access to. While I can obtain quite a lot of data surrounding an event, the fan scores used are user-submitted ratings from one fan website. I was not given the number of scores submitted for each race so the sample size is not known. We could be dealing with a very small set of passionate fans, which does not necessarily extrapolate out to a larger viewing audience.