**DSC 520 Final Project**

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/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.

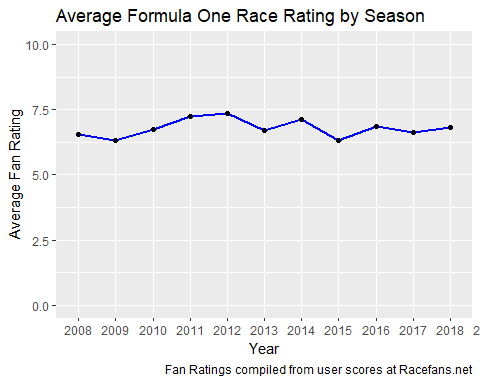
rating\_by\_circuit <- new\_frame %>% group\_by(circuitId) %>% summarize(mean\_rating = mean(RATING))  
rating\_by\_circuit <- merge(x = rating\_by\_circuit, y = circuits, by="circuitId", all.x = TRUE)  
truncated\_rating\_bc <- rating\_by\_circuit[,c("name","mean\_rating")]  
truncated\_rating\_bc <- 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**

* Discuss how you plan to uncover new information in the data that is not self-evident.
* What are different ways you could look at this data to answer the questions you want to answer?
* Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
* How could you summarize your data to answer key questions?
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
* What do you not know how to do right now that you need to learn to answer your questions?
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Suggestion from the course professor: Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you ﬁlter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efﬁcient?

**Section 5 Summary**

* Overall, write a coherent narrative that tells a story with the data as you complete this section.
* Summarize the problem statement you addressed.
* Summarize how you addressed this problem statement (the data used and the methodology employed).
* Summarize the interesting insights that your analysis provided.
* Summarize the implications to the consumer (target audience) of your analysis.
* Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
* In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.