Analysis of Formula 1 Drivers
Aurora Gardner

Formula One is a sport of intense competition and innovation. However, recent years have sparked debate over whether a select few drivers dominate the sport throughout seasons. Additionally, the ever-evolving nature of the sport allows insights to be drawn about the relationship between driver performance and innovation.

The dataset F1DriversDataset by user DD on Kaggle offers information about each driver in Formula One from the dawn of the sport in 1950 until the most recent completed season in 2023. It includes variables such as measures of driver performance, rates of performance, and the seasons and decade in which each driver competed.

F1DriversDataset is a well maintained dataset that required minimal cleaning. There are few missing values, and the missing values that exist are entered as "NA" and are logical in the context of the variables. Some variable names did not match the others, so columns such as "Championship Years" were changed to "Championship_Years" to match the other column names that contained underscores in place of spaces. No variables were removed, as I explored a variety of ideas that involved many different variables.

The debate over whether a few drivers dominate Formula One races has increased throughout recent years. This notion leads to a decline in perceived competitiveness in the races, which is a primary appeal of the sport. However, Formula One has experienced a recent drastic rise in popularity, causing increased media coverage of drivers. My first goal is to evaluate if the increased perception that Formula One is dominated by only a few drivers is actually based on an objective decrease in race winners.

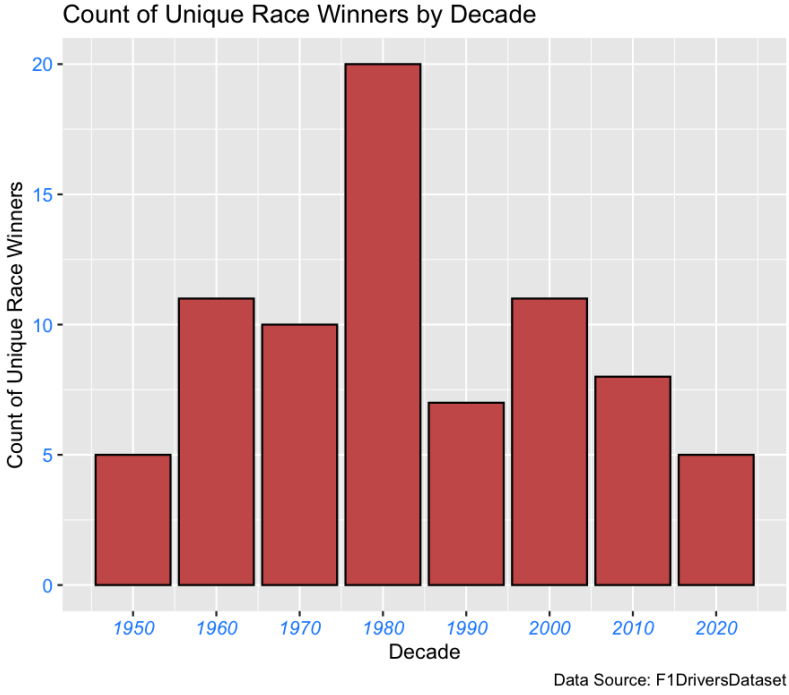


Figure 1 displays the count of unique winners on the Y-axis and each decade on the X-axis. A bar chart provides a clear comparison between counts of drivers for each decade. Code available in Appendix A.

The highest count of race winners exists in the 1980's. However, this value, 20 drivers, is far greater than the values of any other decade. The lowest counts of race winners exist in the 1950's, during the infancy of the sport, and the 2020's, which have only recently begun.

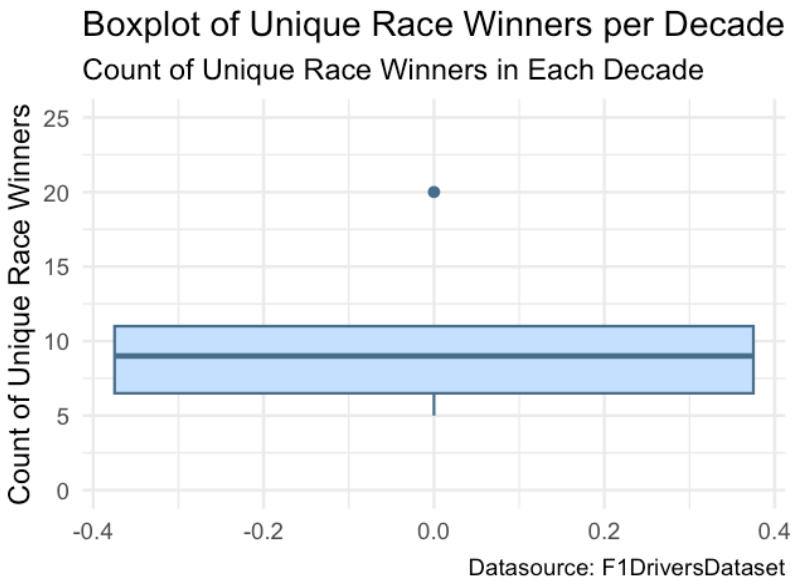


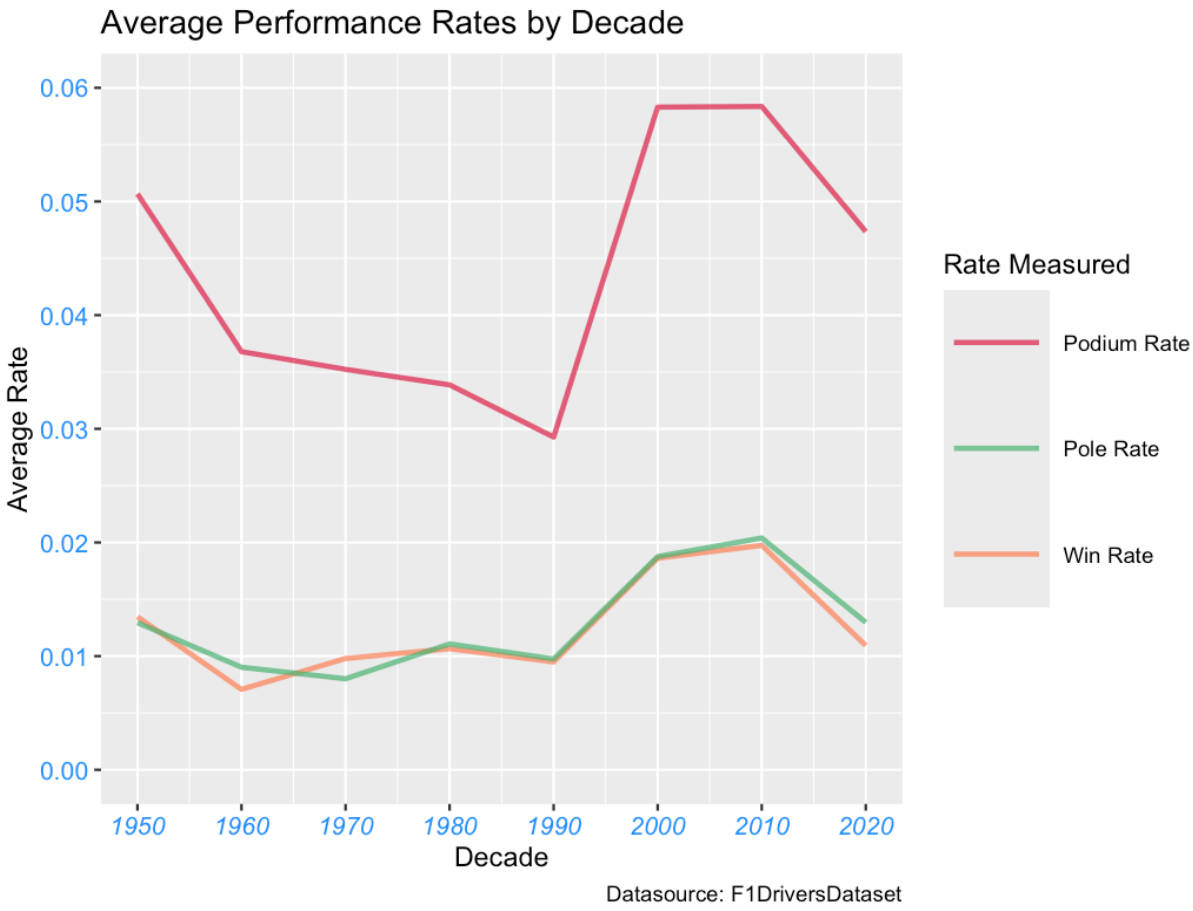
Figure 2 is a boxplot for the count of unique race winners in each decade. It shows the distribution of the number of unique race winners of each decade, including the median, quartiles, and an outlier of 20 drivers. Code in Appendix A.

Figure 2 shows that the count of unique race winners in the 1980's, 20 drivers, is an outlier. To assess the counts of drivers, I set aside the outlier in the 1980's and the skewed or incomplete data in the 1950's and 2020's. The range of the remaining counts is seven to eleven drivers, with the lowest value in the 1990's. The count of unique race winners over the decades has remained relatively stable, so the notion that the sport is increasingly dominated by few drivers is not based on tangible data.

Constant innovation in Formula One allows insights to be drawn regarding the relationship between driver performance and advancements in the sport over time.

To visualize the trends in driver performance, I grouped drivers by decade and calculated the mean of the performance rates in each decade. Pole rate and win rate were very closely related. Podium rate followed similar trends with greater values, because it includes the number of times each driver achieved first, second, or third place instead of exclusively first place as measured in the win rate. The most notable spike in all measures of performance occurred from the 1990's to the 2000's. A possible factor in this spike is innovation in the engineering of the cars. Advancements such as anti-lock braking and traction control were introduced at this time, but the most notable innovation is the kinetic energy recovery system (KERS). KERS allows the cars to conserve energy while they are braking.

Figure 3:
Figure 3 is a line graph displaying the trends in the average rates of three measures of performance, podium rate, pole rate, and win rate, over time. Code in Appendix B.



Grouping the data, isolating variables, and manipulating data in groups proved to be particularly challenging and meticulous. Additionally, choosing the most substantial findings was difficult, as the F1DriversDataset offers a variety of variables that could be used in many different approaches. For example, I experimented with exploring the relationship between years of experience and driver performance.

Despite the notion that Formula One seasons are increasingly dominated by a select few drivers, this perception is not supported by data. This belief may be caused by increased media coverage of select drivers in the sport. Additionally, the trends in driver performance over time reflect major innovations in Formula One. The collaboration between human performance and machines create a unique opportunity for growth in F1, in which advancements in both engineering and driver efficiency are required to edge out the competition.

References

DD. (October 2023). F1DriversDataset, Version 3. Retrieved October 13, 2024 from https://www.kaggle.com/datasets/dubradave/formula-1-drivers-dataset

F1—The official home of formula 1® racing. (n.d.). Formula 1® - The Official F1® Website. Retrieved October 7, 2024, from https://www.formula1.com/en.html

How long is a Formula 1 car? F1 Car length explained! - Las Motorsport. (2023, October 5). https://las-motorsport.com/f1/news/how-long-is-a-formula-1-car-f1-car-length-explained/5048/

Lap around the track of formula 1 technological innovations. (2019, April 2). Jobs.Ca. https://www.jobs.ca/lap-around-track-formula-1-technological-innovations/

Wendorf, M. (n.d.). *Formula 1: How much has changed since 1950?* Interesting Engineering. Retrieved October 7, 2024, from https://interestingengineering.com/culture/formula-1-a-lot-has-changed-since-1950

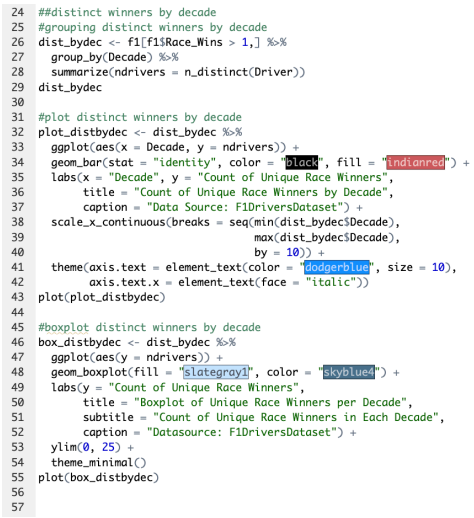
Why do title-winning F1 teams dominate for so much longer now? (2023, August 15). *RACER*. https://racer.com/2023/08/15/why-do-title-winning-f1-teams-dominate-for-so-much-longer-now/

Appendix A

```
##R code for figures 1 and 2
#grouping distinct winners by decade
dist_bydec <- f1[f1$Race_Wins > 1,] %>%
  group_by(Decade) %>%
  summarize(ndrivers = n_distinct(Driver))
dist_bydec

#plot distinct winners by decade
plot_distbydec <- dist_bydec %>%
  ggplot(aes(x = Decade, y = ndrivers)) +
  geom_bar(stat = "identity", color = "black", fill = "indianred") +
  labs(x = "Decade", y = "Count of Unique Race Winners",
       title = "Count of Unique Race Winners by Decade",
       caption = "Data Source: F1DriversDataset") +
  scale_x_continuous(breaks = seq(min(dist_bydec$Decade),
                                  max(dist_bydec$Decade),
                                  by = 10)) +
  theme(axis.text = element_text(color = "dodgerblue", size = 10),
        axis.text.x = element_text(face = "italic"))
plot(plot_distbydec)

#boxplot distinct winners by decade
box_distbydec <- dist_bydec %>%
  ggplot(aes(y = ndrivers)) +
  geom_boxplot(fill = "slategray1", color = "skyblue4") +
  labs(y = "Count of Unique Race Winners",
       title = "Boxplot of Unique Race Winners per Decade",
       subtitle = "Count of Unique Race Winners in Each Decade",
       caption = "Datasource: F1DriversDataset") +
  ylim(0, 25) +
  theme_minimal()
plot(box_distbydec)
```



Appendix B

```
##R code for figure 3
#average rates by decade
avg_bydec <- f1 %>%
  group_by(Decade) %>%
  summarise(
    avg_win_rate = mean(Win_Rate),
    avg_pole_rate = mean(Pole_Rate),
    avg_pod_rate = mean(Podium_Rate)
  )
avg_bydec

#plot average rates by decade
plot_avgbydec <- ggplot(avg_bydec, aes(x = Decade)) +
  geom_line(aes(y = avg_win_rate, color = "Win Rate"),
            linewidth = 1, alpha = 0.7) +
  geom_line(aes(y = avg_pole_rate, color = "Pole Rate"),
            linewidth = 1, alpha = 0.7) +
  geom_line(aes(y = avg_pod_rate, color = "Podium Rate"),
            linewidth = 1, alpha = 0.7) +
  labs(title = "Average Performance Rates by Decade",
       x = "Decade",
       y = "Average Rate",
       color = "Rate Measured",
       caption = "Datasource: F1DriversDataset") +
  scale_color_manual(values = c("Win Rate" = "coral",
                                "Pole Rate" = "mediumseagreen",
                                "Podium Rate" = "#DC143C")) +
  scale_x_continuous(breaks = seq(min(avg_bydec$Decade),
                                  max(avg_bydec$Decade),
                                  by = 10)) +
  scale_y_continuous(breaks = seq(0, 0.06, by = 0.01),
                     limits = c(0, 0.06)) +
  theme(axis.text = element_text(color = "dodgerblue", size = 10),
        axis.text.x = element_text(face = "italic"),
        legend.key.size = unit(1.5, "cm"))
plot(plot_avgbydec)
```