Formula 1 Project Report: Data Wrangling
Charlie Rolecek

1. Introduction and Background

The addition of a third Grand Prix to take place in The United States, along with the hit Netflix series "Drive to Survive" have created a growing American fan base for Formula 1. This new American fan base presents the opportunity to set more accurate, and more interesting sports gambling lines. According to Statista, in 2021 total revenue from legal sports gambling in the United States reached over 4.33 billion dollars[1]. The third grand prix in America will take place on the Las Vegas Strip, one of the most notorious gambling cities in the world. Currently FanDuel creates and sets gambling lines and odds for Formula 1 races, and prop bets such as: who will have the fastest lap, team of the winner, top 6 finish, and more.

In this project, I will use Formula 1 race data along with the 2021 Formula 1 World Drivers Championship Standings to examine the relation between the duration of driver's pit stops and the driver's final placement within the standings to better understand how minimal mistakes can shape the result of the season.

2. Data

The two primary sources of data I used in this project were from Kaggle and Wikipedia. Kaggle's open data set on Formula 1 contained data from 1950 to 2022 containing 14 files, which I only used 3 of: pit_stops.csv, drivers.csv, and races.csv[2]. My other data source came from the 2021 Formula 1 World Championship Wikipedia page[3].

2.1 2021 Formula 1 World Championship

To collect the final driver standings of the 2021 racing season, I used the main page of the 2021 Formula 1 World Championship Wikipedia page. This page contained a table of driver standing, all driver names, outcomes of the 2021 Grand Prix's, and total points collected throughout the season. From here, I wrote a scraping script to pull the World Drivers Championship standings table, convert it into a data frame that contained 23 observations of 25 variables. All scraping code can be examined in the R script "csrolecek_project_scraping.R".

2.2 Formula 1 World Championship (1950-2022)

The Kaggle open data set is a collection of 14 files with a combined 120 columns with varying observations for each. All files were published to Kaggle under a .csv format, and variables of interest were spread across 3 files: pit_stops.csv, drivers.csv, and races.csv. After reading the csv files into RStudio I horizontally merged pit_stops.csv and drivers.csv by the shared variable of driverId. I combined two variables of forename and surname to driver in order to match the scraped table from Wikipedia. Following this, another horizontal merge was needed to integrate the 3rd csv file with the first 2 using raceId. The final merged data frame was then cleaned by, removing irrelevant and erroneous columns, renaming columns, and reordering columns. After cleaning the merged data frame, I subset a new data frame to capture only races that were in 2021. Next, a final horizontal merge was needed to integrate the merged data set

and the scraped Wikipedia table. Finally, I wrote this final merged data frame to a new csv. All Integration code can be examined in the R script "csrolecek_project_integration.R".

Table 1 Data Dictionary

Column	Data Type	Description
driver	text	Name of the driver
race_Id	numeric	Unique ID given for each race
race_name	text	Name of the race (ex. Monaco Grand Prix)
driver_id	numeric	Unique ID given to each driver
pitstop_seconds	numeric	total seconds driver spent in the pitstop
pitstop_milliseconds	numeric	total milliseconds driver spent in the pitstop
stop_number	numeric	Number of amount of times driver has used pitstop
lap_number	numeric	Lap number that driver entered pitstop
year_of_race	numeric	Year that the race took place
race_number	numeric	Denoting the order of the races 1-22
driver_number	numeric	Number that the driver has on his helmet and car
date_of_birth	Timestamp	Date of driver's birthday
nationality	text	Nationality of the driver
circuit_Id	num	Unique ID given for each circuit
driver_final_standing	num	Final ranking of the World Drivers Championship
season_total_points	num	Total amount of points earned during the 2021 racing season

3. Analysis

The goal of this project is to examine the role pit stop time plays in Formula 1. Especially if a lower pitstop time leads to more points earned during the racing seasons.

3.1

Which driver had the lowest pitstop time throughout the entire season and which driver had the lowest average pitstop time? In order to answer this question, I created a dplyr[4] summary table that shows the driver_Id along with their minimum, maximum and average pitstop times. Table 2 will show these results.

Table 2 Driver Summary

driver_Id	Min.pitstop_seconds	Max.pitstop_seconds	Avg.pitstops_seconds
9	19.708	24.501	22.10450
1	15.432	40.266	23.99371
830	17.807	35.245	24.01173
844	15.092	32.648	24.07041
815	15.277	44.608	24.13912
849	18.153	39.081	24.49160
846	15.866	38.267	24.50775
822	17.837	36.341	24.52388
817	14.994	38.128	24.69093
4	15.432	57.601	24.94618
840	17.935	43.124	25.02838
8	15.064	36.457	25.02978
852	14.943	40.740	25.13128
20	14.945	52.043	25.30447
847	18.179	46.315	25.37972
853	15.054	49.729	25.45196
841	14.881	54.673	25.76118
842	15.085	56.083	25.96675
854	15.058	51.222	26.52511
839	17.818	56.733	26.99377

The table is able to show that the driver with the lowest average time spent in the pitstop during the 2021 racing season was Robert Kubica however, Kubica only raced in 2 Grand Prix's in 2021. Taking this into account shows that driver_Id 1, Lewis Hamilton, has the lowest average time spent in the pitstop during the 2021 racing season.

3.2

Is there a relationship between lower average pitstop times and a higher final standing? Using linear regression, I used the average pitstop time for each driver against the drivers final standing. To do so, I used a ggplot[5] scatterplot showing drivers final standing at the end of the season and average pitstop time in seconds (Figure 1).

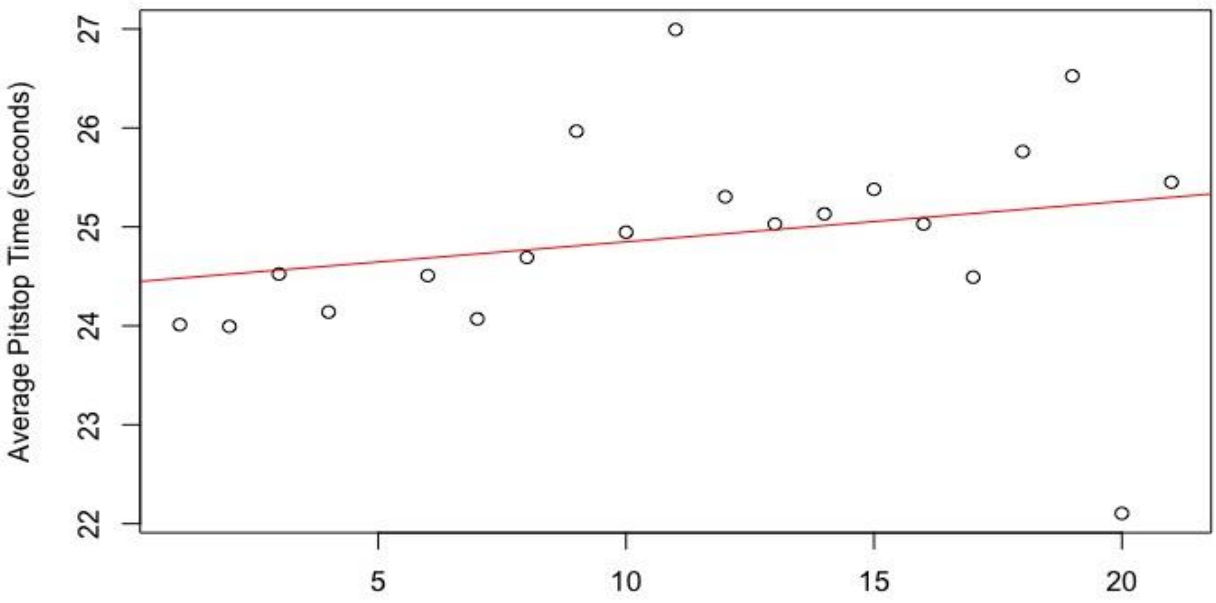

Figure 1 Scatterplot

This scatterplot does show that a positive, yet weak, relationship between high average pitstop times and high final standings. Next, I checked the normality with a Q-Q plot (Figure 2) and found that since marks are close to the reference line, the pitstop times in the data are very close to normal distribution.

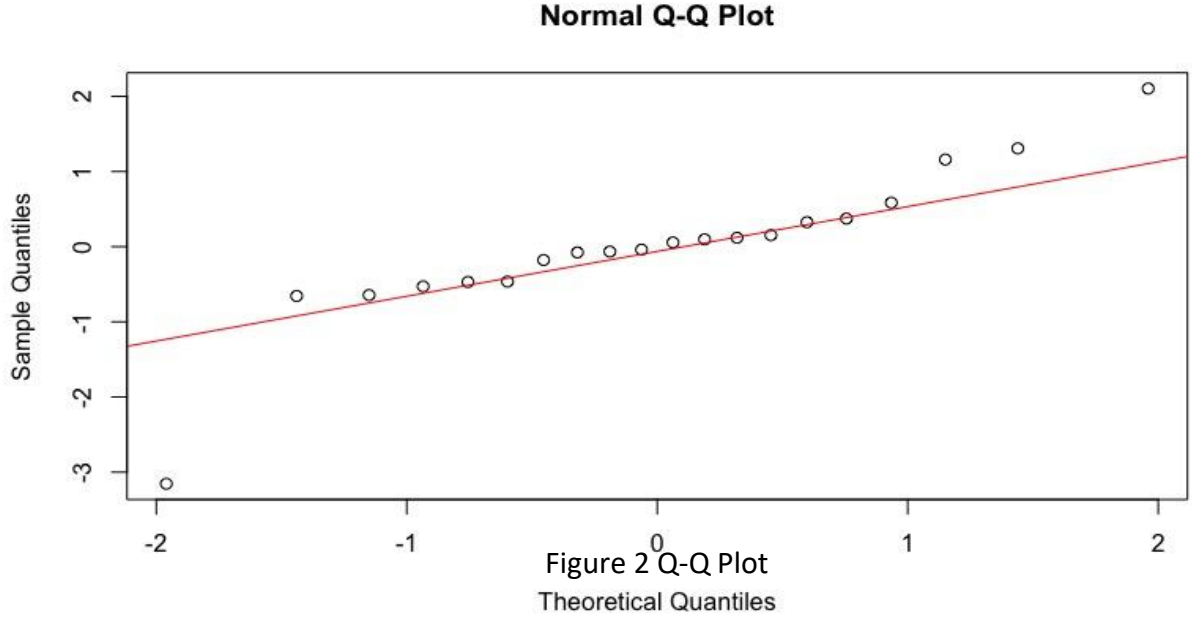

Figure 2 Q-Q Plot

3.3

Does the nationality of the driver have any impact on pitstop time? To evaluate this question, I create a boxplot showing pitstop times in seconds and the nationality of the drivers (Figure 3). While all of these nationalities seem very close to each other there are a few that stick out. The polish driver had very little outliers making me believe that this is Robert Kubica, who I spoke about earlier. On top of this Dutch and British drivers have shorter pitstop times and less outliers when compared to the rest.

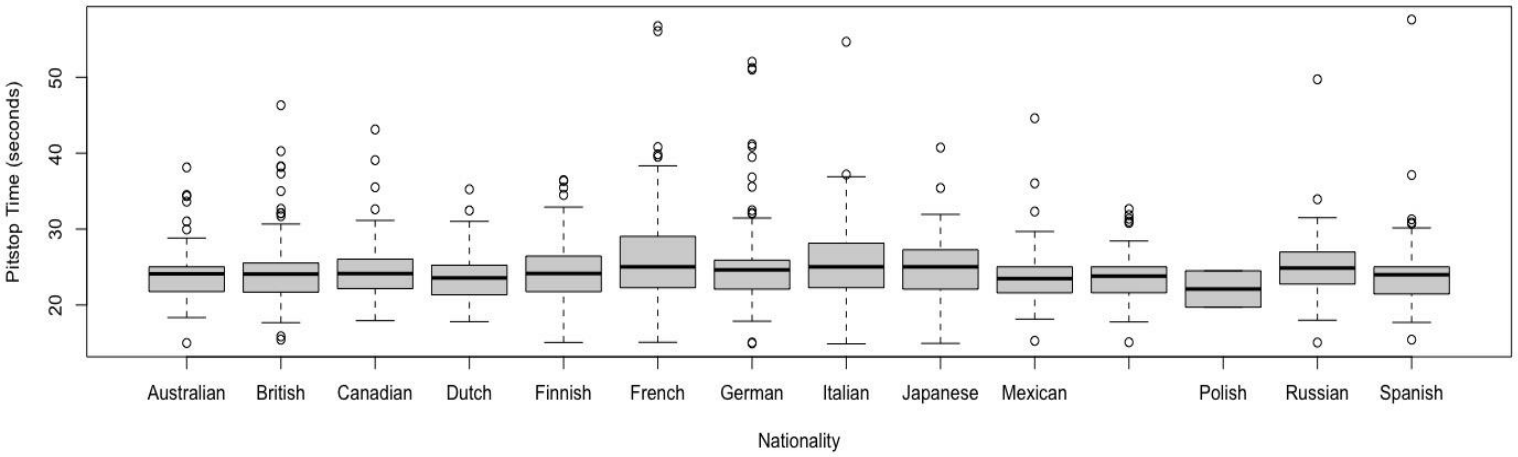

Figure 3 Boxplot

All code for descriptive analysis and visualization can be examined in the R script "csrolecek_projet_descriptive_analysis.R"

4. Conclusion and Limitations

The use of the Formula 1 Championship dataset from Kaggle that contained race, driver, and pitstop data combined with 2021 world driver championship standings scraped from Wikipedia, helped me explore the relationship between the duration of pitstops to driver standings, driver nationality, and simply which driver, or which driver's team, is the most consistent crew for pitstops. The use of summary statistics, visualizations of data, linear regression, and the bartlett test, helped show that there is a positive relationship of higher duration pitstops and higher ranking in standings. This means that the less amount of time a driver takes in their pitstop their final standing will be closer to first. Furthermore, there are multiple limitations to this project. First, the data obtained from Kaggle had missing values for pitstop_seconds, and I had to impute the mean leading to different results than real data. On top of this, this project does not account for engine strength or volume which could lead to lower pit times. Next this project does not take driver skill into account. In formula 1 some drivers are better than others and it is as simple as that, skill can decrease amount of time in the pit or raise total points earned throughout a season. Future work on this topic could be very interesting, it could examine the possibility of certain circuits having pits that are longer/shorter than others, thus changing the impact of the pitstop time.

References

1. https://www.statista.com/topics/1368/gambling/
2. https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020
3. https://en.wikipedia.org/wiki/2021_Formula_One_World_Championship
4. https://cran.r-project.org/web/packages/dplyr/
5. https://cran.r-project.org/web/packages/ggplot2/