DMWR Final Project: An Exploration of F1 Data

Avi Kodali

2022-05-05

Part 1: Ergast Data

One of my biggest passions is Formula 1 racing. In fact, I have watched each and every race religiously since the start of the Covid-19 pandemic. Therefore, I thought it would be fun to work with Formula 1 data, and see what I could find through analyzing it. The first part of my final project is using the historical F1 data from https://ergast.com/mrd/. This is an API that has access to F1 data from its inception in 1950 until now. This includes things like race results, pit stop times, driver data, season data.

Pitstop Data

I decided to first look at pit stop data from this API, which has data specifically from 2011 until now. My first step was to attempt to use the API to get the pit stop data. However, this proved to be very difficult/inefficient because each query to the API for the pit stop data only allows access to the pit stop data of a specific year and round number of an F1 season. This would mean that I would have to have to combine the pit stop dataframes of roughly 200 races at some point, as well as combine that data with other data from Ergast such as constructor names, race locations, etc. Therefore, I decided to explore the site more to see if I had any other options, and I discovered that I could download the data from the Ergast database from https://ergast.com/mrd/db/. This made obtaining the dataframe I wanted more obtainable. So using the CSV database tables from Ergast, I cleaned and combined the different tables to gain my final pit stop dataframe, which part of can be seen below.

raceName	raceDate	circuitName	circuitAltitude	dateTime	year	round	driverName	driverCode	constructorName	stop	lap	duration	milliseconds	seconds
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:05:23	2011	1	Jaime Alguersuari	ALG	Toro Rosso	1	1	26.898	26898	26.898
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:05:52	2011	1	Michael Schumacher	MSC	Mercedes	1	1	25.021	25021	25.021
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:20:48	2011	1	Mark Webber	WEB	Red Bull	1	11	23.426	23426	23.426
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:22:34	2011	1	Fernando Alonso	ALO	Ferrari	1	12	23.251	23251	23.251
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:24:10	2011	1	Felipe Massa	MAS	Ferrari	1	13	23.842	23842	23.842
Australian Grand Prix	2011-03-27	Albert Park Grand Prix Circuit	10	2011-03-27 17:24:29	2011	1	Rubens Barrichello	BAR	Williams	1	13	23.643	23643	23.643

Now I can save this clean pitstop dataframe into a csv file so that anyone can use it to start their own analysis.

```
write.csv(pitstops, file = "pitstops_full.csv")
```

Now with this cleaned data, I can do some analysis to see what I can find. First I want to see which constructor have the fastest average pit stop time.

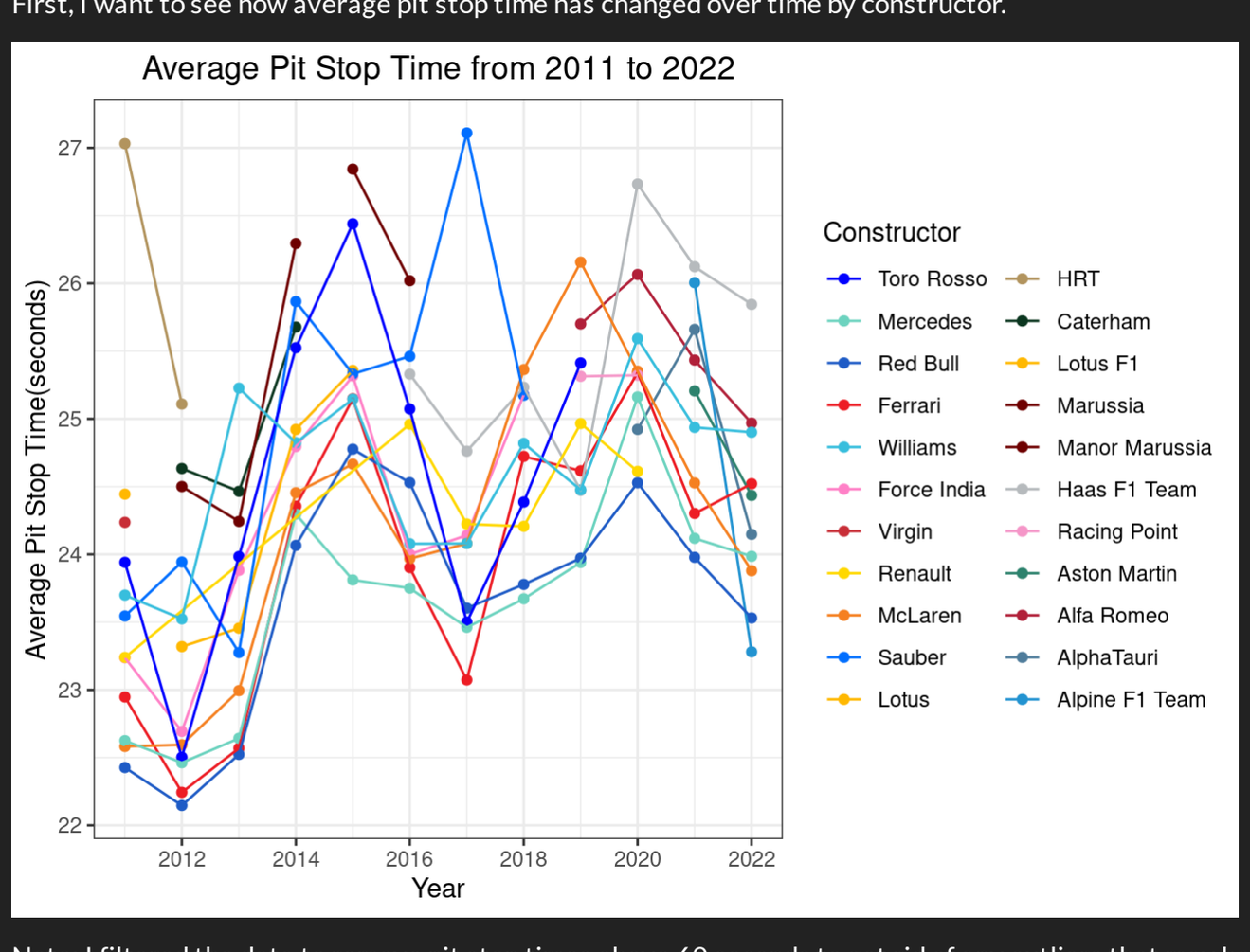
constructorName	avgPitStopTime	numberOfRaces
Virgin	24.23619	77
Lotus	24.44421	85
Lotus F1	32.46375	285
HRT	32.67777	150
Caterham	33.92462	241
Marussia	34.39684	233
Toro Rosso	44.48698	683
Sauber	44.54653	621
Force India	50.84577	609
Renault	63.26183	403

From this table, it seems that Virgin is the constructor with the fastest average pit stop time. However, a big part of this is probably because the Virgin only raced for two years(2010, 2011). I think I should also point out that there are two Lotus teams because they are technically two different teams(Lotus turned into Lotus F1 in the 2012 F1 season). Another interesting insight from this table is that most of these fastest average pit stop times are from teams that do not currently exist right now, i.e. teams that raced in the early 2010s. I think this might be because average pit stop times overall have increased from 2011 to now. However, I think it is important to note that the summary statistics for this table include outliers that could have been caused by pit stop problems, retirements, etc. Therefore, I decided to filter out these outliers and see how the results change.

constructorName	avgPitStopTime	numberOfRaces
Mercedes	23.56088	845
Red Bull	23.57966	853
Ferrari	23.81331	839
McLaren	24.05513	828
Force India	24.05836	593
Virgin	24.23619	77
Lotus F1	24.23822	281
Renault	24.29716	390
Lotus	24.44421	85
Toro Rosso	24.49106	671

This table is interesting because the top three constructors with the fastest average pit stop times(Mercedes, Red Bull, Ferrari) are the top three constructors in the past decade, i.e they are the teams that mainly fought for the F1 World Championship. Furthermore, a lot of the constructors in this table are still older teams that don't race right now. With the insights from these tables, I can try plotting the data.

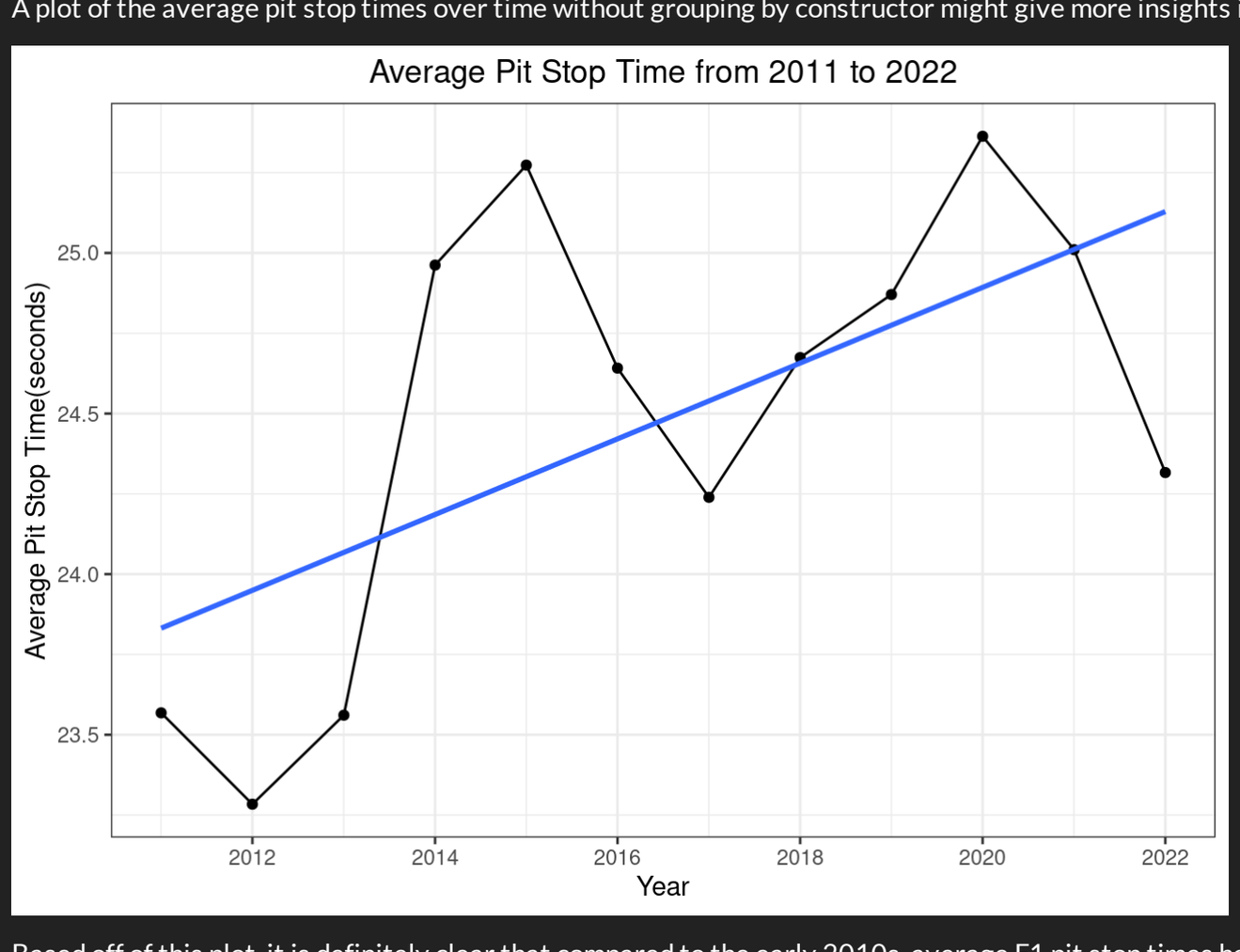
First, I want to see how average pit stop time has changed over time by constructor.



Note: I filtered the data to remove pit stop times above 60 seconds to get rid of any outliers that may be caused by pit stop problems, retirements, etc.

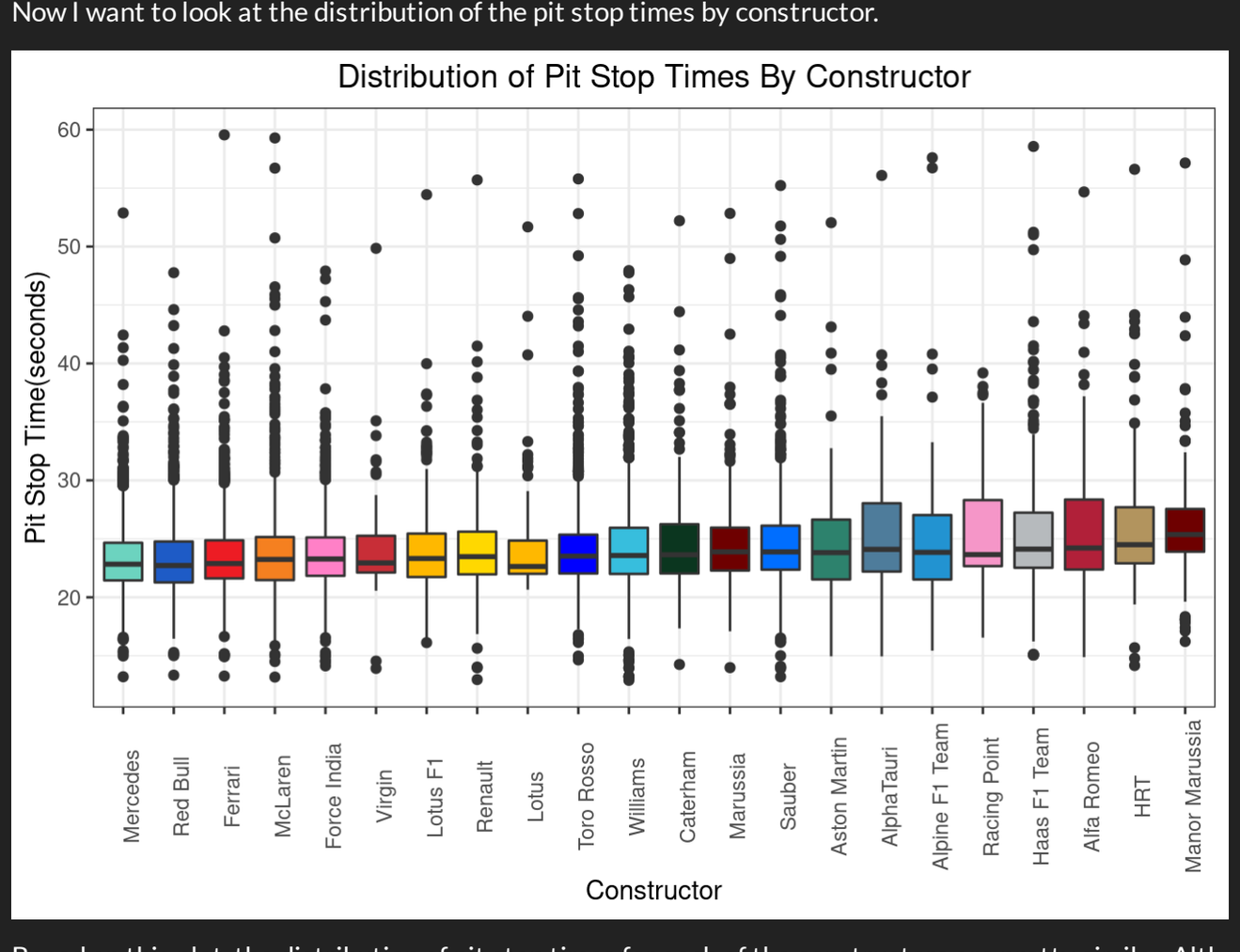
Although, there isn't a clearly visible pattern in average pit stop times over the years, overall, I would say that compared to the early 2010s, average pit stop times have slightly increased.

A plot of the average pit stop times over time without grouping by constructor might give more insights into this.



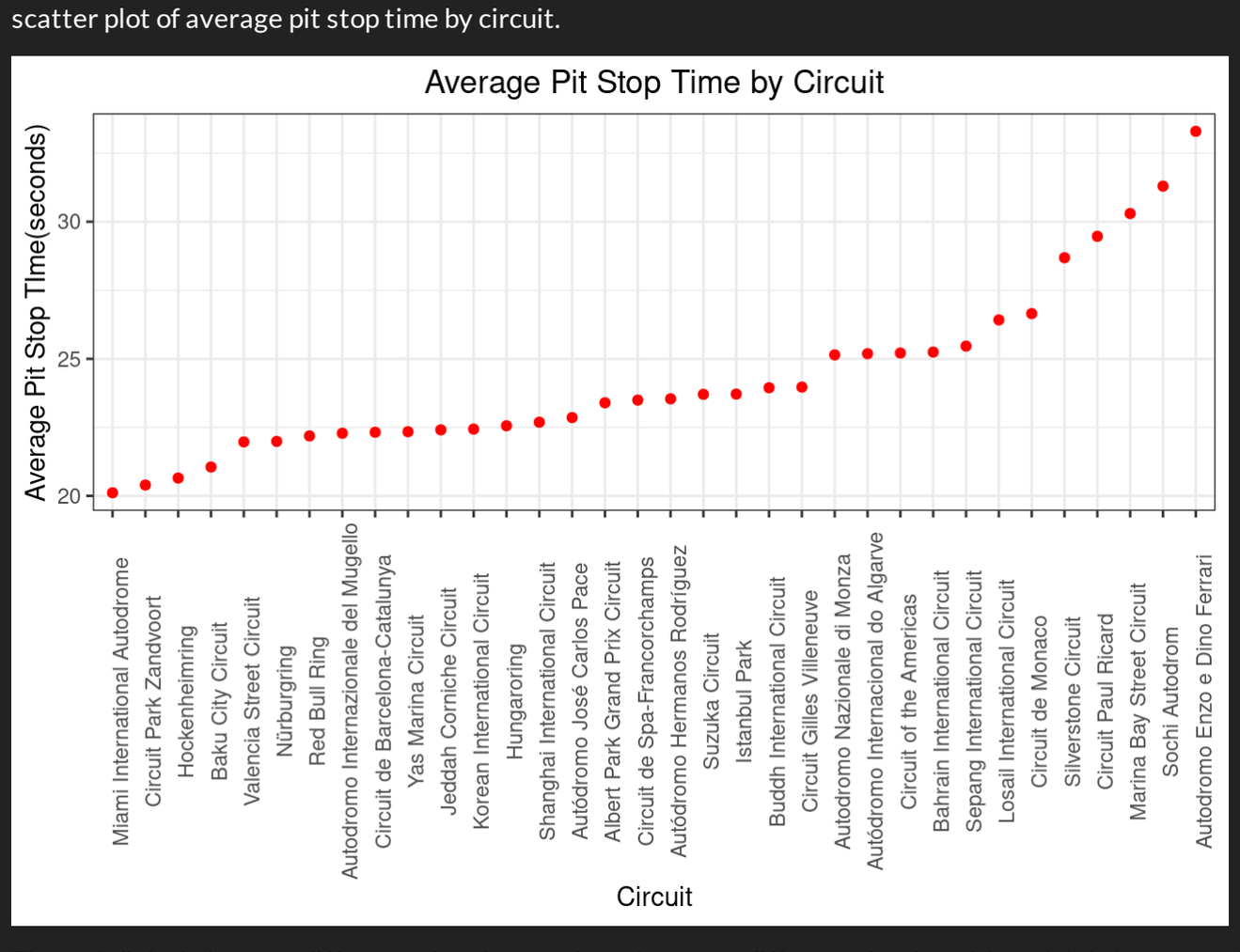
Based off of this plot, it is definitely clear that compared to the early 2010s, average F1 pit stop times have increased over time.

Now I want to look at the distribution of the pit stop times by constructor.



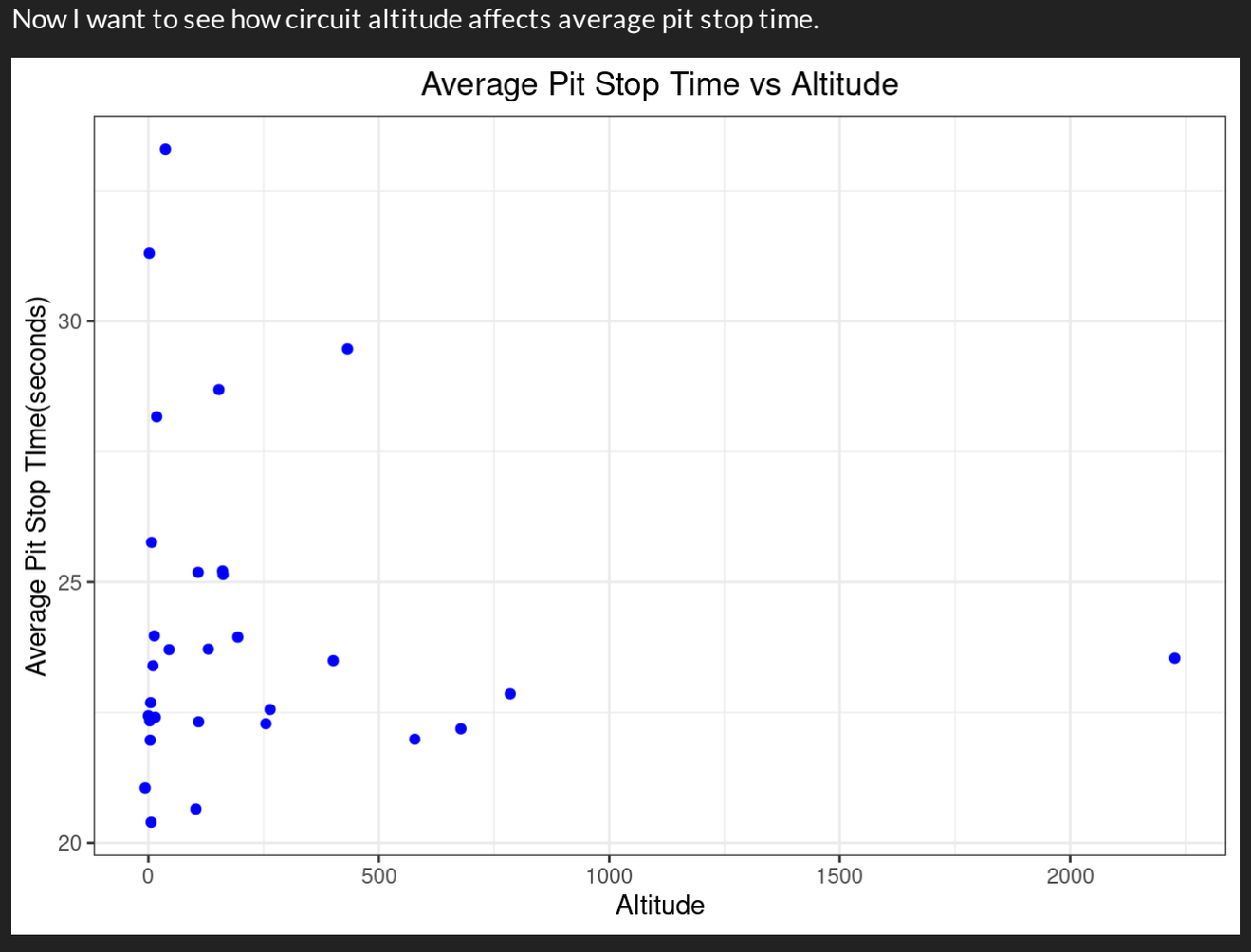
Based on this plot, the distribution of pit stop times for each of the constructors are pretty similar. Although there are some that are a little less consistent(bigger IQR) like the Alpine F1 Team, the IQR of all the teams is between 20 and 30 seconds.

After looking at pit stop data by constructor, I want to see if the circuit at which the race occurs affects pit stop time. To do this, I can look a scatter plot of average pit stop time by circuit.



There definitely is some difference for pit stop times between different circuits, although I do have to point out that Circui Park Zandvoort has the fastest average pit stop time because it is a very new circuit that only had one race in 2021. This is a pretty interesting insight. These differences could be due to the length of the pit lane at each circuit or perhaps the overall environment at each circuit(like temperature, humidity, etc.).

Now I want to see how circuit altitude affects average pit stop time.



Based on the plot, it does not seem like there is a clear relationship between altitude and average pit stop time.

2021 F1 Season Data

Having satisfied my curiosity about this pit stop data, I want to now make some plots to look at the 2021 season of F1 specifically. This was one of the closest and most controversial seasons of F1 in the turbo-hybrid era, resulting in Max Verstappen winning the Driver's World Championship after Lewis Hamilton dominated the previous four seasons. First I have to make a couple of complete, clean, dataframes to contain all of the needed data.

First, I made the driver standings dataframe for the 2021 F1 season. (Preview below)

city	country	circuitName	raceName	raceDate	year	round	driverName	driverCode	points	position	wins
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Lewis Hamilton	HAM	25	1	1
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Max Verstappen	VER	18	2	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Valtteri Bottas	BOT	16	3	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Lando Norris	NOR	12	4	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Sergio Pérez	PER	10	5	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Charles Leclerc	LEC	8	6	0

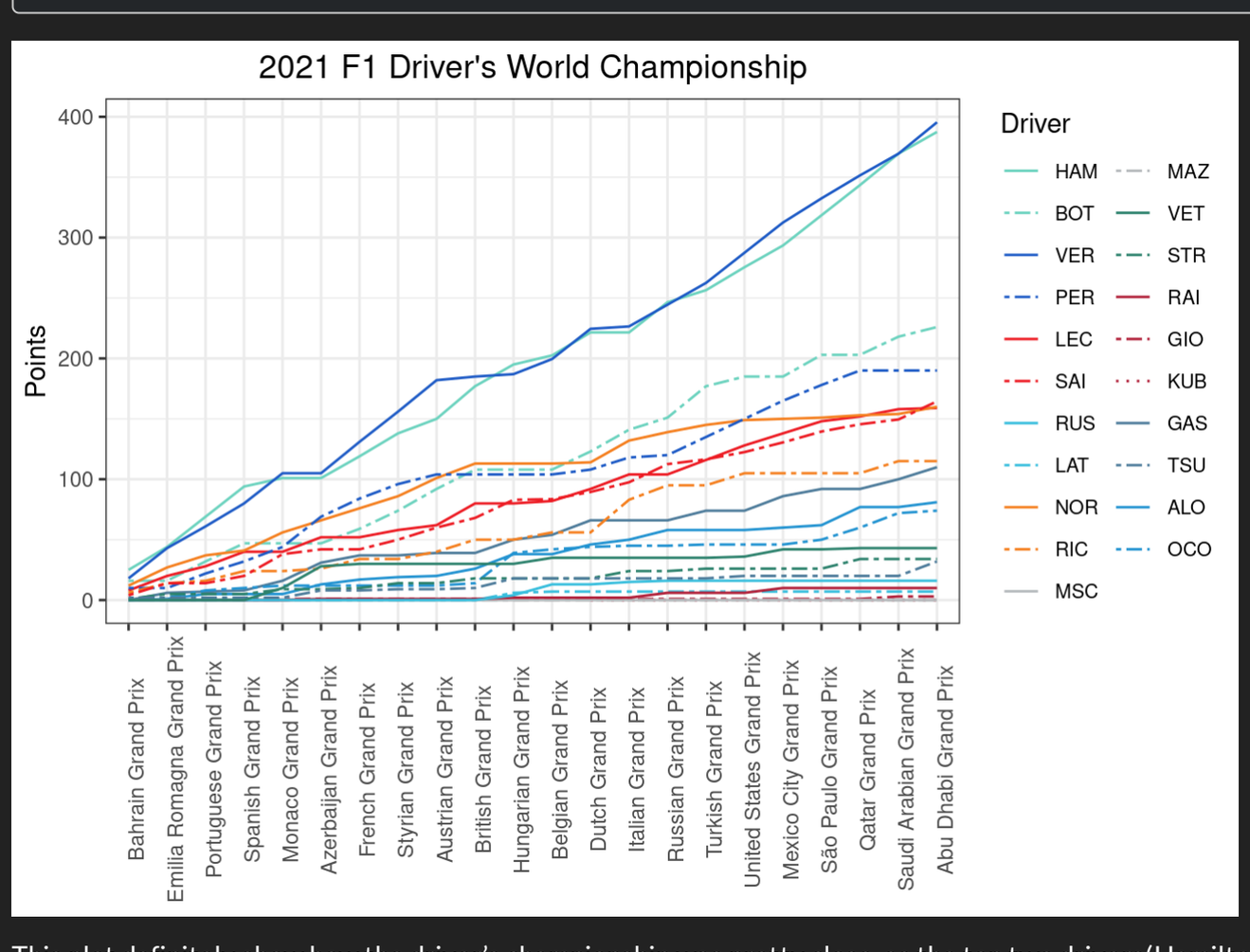
Next, I made the constructor standings dataframe for the 2021 F1 season. (Preview below)

city	country	circuitName	raceName	raceDate	year	round	constructorName	points	position	wins
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Mercedes	41	1	1
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Red Bull	28	2	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	McLaren	18	3	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Ferrari	12	4	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	AlphaTauri	2	5	0
Sakhir	Bahrain	Bahrain International Circuit	Bahrain Grand Prix	2021-03-28	2021	1	Aston Martin	1	6	0

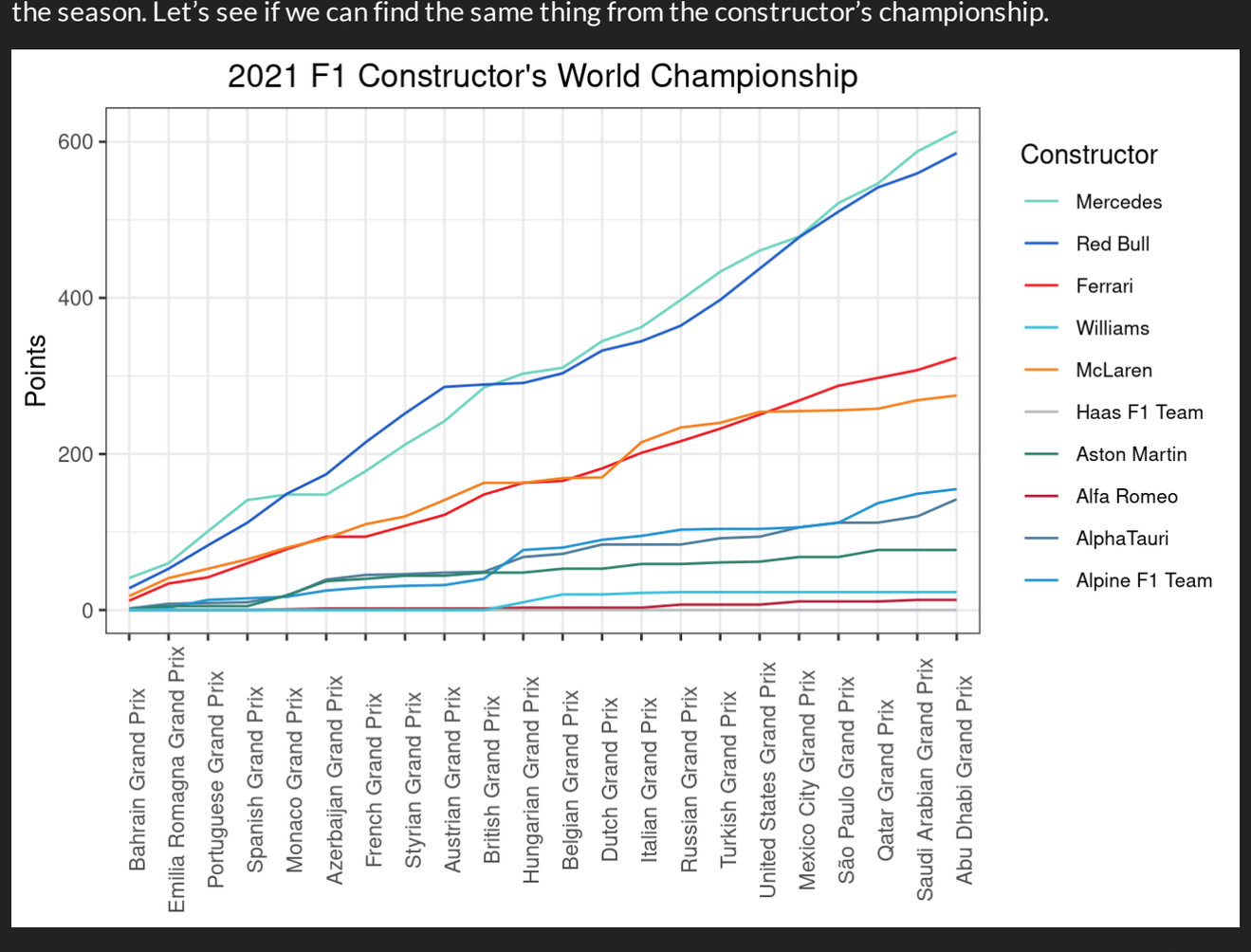
Now I can get started on working on a couple of plots.

I have included the code for the F1 World Driver's championship plot because it took me a lot of time to make, and personally I am proud of it. It looks a lot like the graphs the official F1 Instagram account posted.

```
driver_standings_2021 %>%
  ggplot(aes(x = fct_reorder(raceName, raceDate), y = points,
             color = driverCode, linetype = driverCode, group = driverCode)) +
  geom_line() +
  scale_color_manual(values = drivers_2021_colors) +
  scale_linetype_manual(values = drivers_2021_lines) +
  theme_bw() +
  theme(axis.text.x.bottom = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        axis.title.x = element_blank(),
        legend.text = element_text(size = 8)) +
  labs(title = "2021 F1 Driver's World Championship",
       y = "Points",
       color = "Driver",
       linetype = "Driver")
```



This plot definitely show how the driver's championship was pretty close, as the top two drivers(Hamilton and Verstappen) exchanged places in terms of points total. And that was not the only close battle, as there are several driver who were close to each other throughout the season. Let's see if we can find the same thing from the constructor's championship.



The constructor's championship was also pretty close. From the plot, there are three battles that for most of the season: Mercedes vs. Redbull, Ferrari vs. Mclaren, and Alphatauri vs. Alpine F1 Team. Although making and fine-tuning these plots was somewhat thrilling itself, watching these battles unfold live was more thrilling.

F1 Twitter Sentiment Analysis

Next I want to try and do some sentiment analysis on F1 Twitter posts. Specifically, I want to look at tweets around the time of the 2021 Abu Dhabi Grand Prix. This is because this race made the 2021 F1 season very controversial. The 2021 season of F1 was controversial because of the way the race ended, and the World Champion was decided. The last race of this season was the Abu Dhabi Grand Prix, and in this race, Hamilton was well on his way to winning the race and the championship, but in the last couple of laps, the controversial decision by the race director in regards to the safety car allowed Verstappen to catch up and overtake Hamilton rather easily. At first, I wanted to do this by scraping Twitter using the Twitter API, but this proved to be difficult because the developer account I had did not have the functionality needed to do this, and any packages in R that helped utilize the API were not optimized for Twitter's new v2 API. Therefore, I had to find another way, and luckily I found a dataset on https://www.kaggle.com/datasets/kaushiksuresh147/formula-1-trending-tweets, which has trending F1 tweets from Twitter starting from August 2021. If I filtered this dataset to include only tweets from December 4 to December 17, this would give me a good amount of tweets around when the Abu Dhabi Grand Prix occurred(as well as for the previous grand prix that was also pretty controversial).

Here, I filtered out the appropriate dates, and selected variables I thought might be useful(other than the text itself).

dateTime	tweet_date	text
2021-12-04 17:47:49	2021-12-04	Like circuit #SaudiArabianGP #F1 👍🏻
2021-12-04 17:48:23	2021-12-04	We've got a hacker in the open lobby 😂
#F1 #SaudiArabianGP #Quali https://t.co/x2qkapKBc2		
2021-12-04 17:48:24	2021-12-04	Dangerous track. Many collisions avoided only for lucky #SaudiArabiaGP #F1 #SaudiArabianGP
2021-12-04 17:48:28	2021-12-04	Leclerc and the two Mercedes are waiting at the end of the pitlane for the green light for Q3 to start. Eager beavers. #F1
2021-12-04 17:48:30	2021-12-04	This could be interesting #Q3 #F1
2021-12-04 17:48:31	2021-12-04	#Q3 start #F1 https://t.co/SMZ98ba3Ln

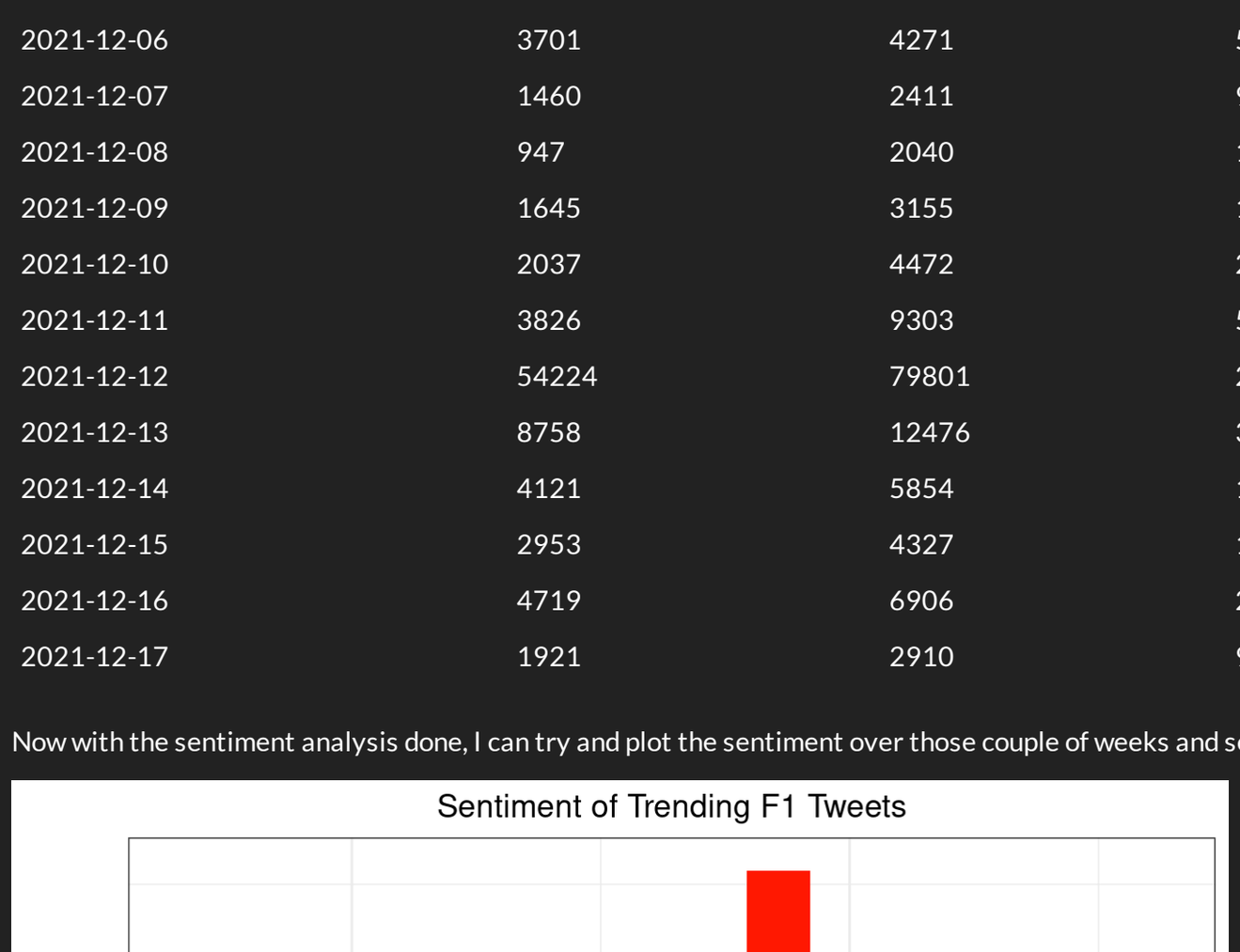
Now I have to make a dataframe that follows the tidy text format.

tweet_date	word
2021-12-04	like
2021-12-04	circuit
2021-12-04	saudiarabiangp
2021-12-04	f1
2021-12-04	we've
2021-12-04	got

With this tidy data, I can try and perform sentiment analysis.

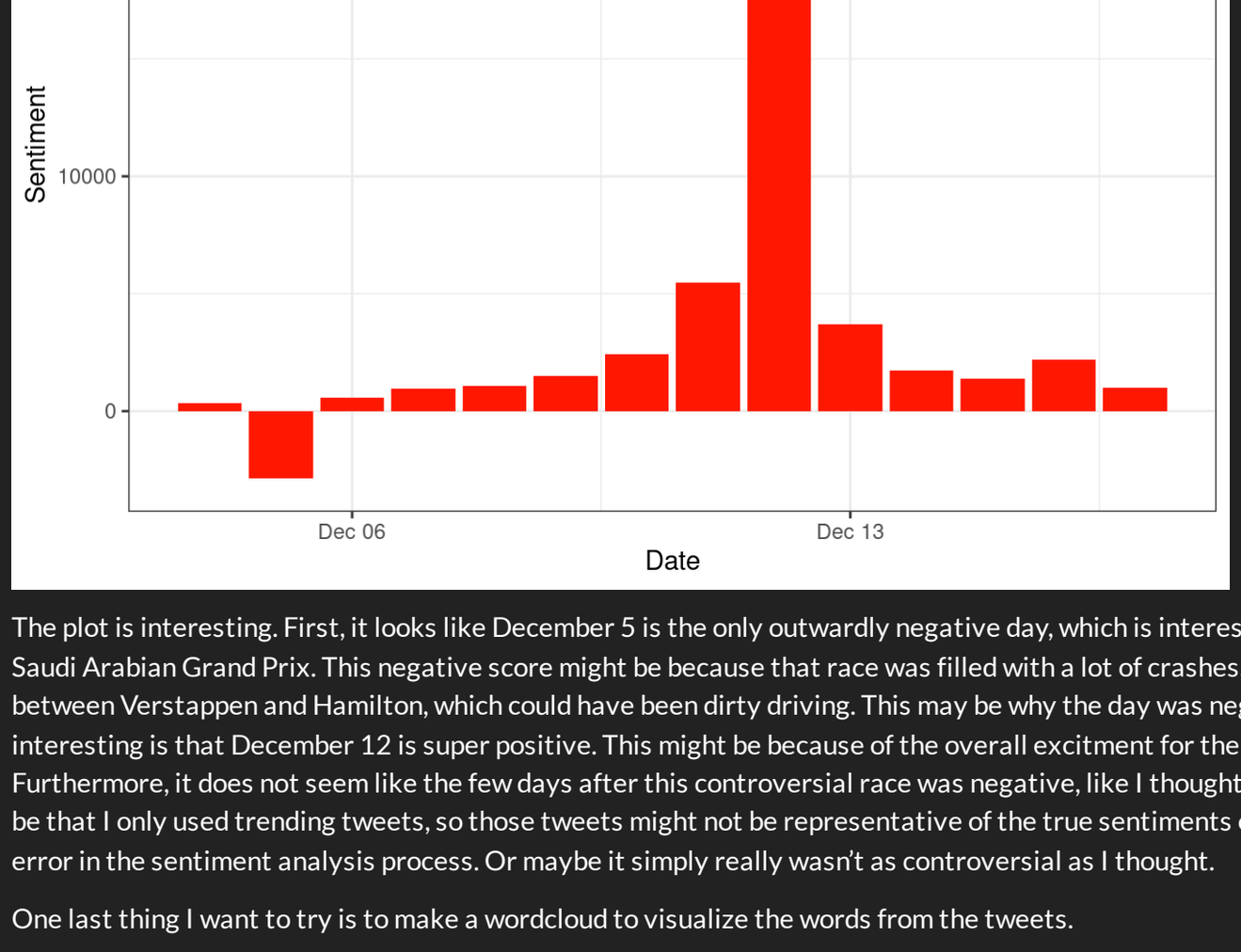
tweet_date	negative	positive	sentiment
2021-12-04	2912	3242	330
2021-12-05	28953	26089	-2864
2021-12-06	3701	4271	570
2021-12-07	1460	2411	951
2021-12-08	947	2040	1093
2021-12-09	1645	3155	1510
2021-12-10	2037	4472	2435
2021-12-11	3826	9303	5477
2021-12-12	54224	79801	25577
2021-12-13	8758	12476	3718
2021-12-14	4121	5854	1733
2021-12-15	2953	4327	1374
2021-12-16	4719	6906	2187
2021-12-17	1921	2910	989

Now with the sentiment analysis done, I can try and plot the sentiment over those couple of weeks and see if I can find anything interesting.



The plot is interesting. First, it looks like December 5 is the only outwardly negative day, which is interesting because it was the day of the Saudi Arabian Grand Prix. This negative score might be because that race was filled with a lot of crashes, and highlighting it all was contact between Verstappen and Hamilton, which could have been dirty driving. This may be why the day was negative. Another thing that is interesting is that December 12 is super positive. This might be because of the overall excitment for the last, championship deciding race. Furthermore, it does not seem like the few days after this controversial race was negative, like I thought it would be. A factor in this might be that I only used trending tweets, so those tweets might not be representative of the true sentiments on Twitter. Or maybe I made an error in the sentiment analysis process. Or maybe it simply really wasn't as controversial as I thought.

One last thing I want to try is to make a wordcloud to visualize the words from the tweets.



From this wordcloud, one thought I had was that the context of the word might be important. Clearly, words related to winning or wins are used a lot, but in the context of that tweet, Verstappen winning might have been negative. Also, December 12 might be so positive because there was an influx of tweets proclaiming Verstappen as race winner and World Champion, so that might be why it was so positive. Furthermore, although there are a lot of positive words, there are also plenty of negative words, especially curse words. Since the occurrence of those positive words is higher, that might be why we do not have a lot of days that are clearly negative.

With this, my final project covering the exploration of F1 data is done. I found a lot of insights through my analysis of various sets of F1 data, and it was pretty enjoyable. I hope you found what I did interesting as well.