- In this document columns of the data files and their sources are described.

List of data files

1. 2014-07-06_BritishGP-ExampleTwitterTweets.csv
2. t0WHNqwDjXc-ExampleYoutubeComments.csv
3. youtube_sentiments.csv
4. twitter_sentiments.csv
5. aggregated_basedOnFanRatings.csv

2014-07-06_BritishGP-ExampleTwitterTweets.csv - This is an example data file which consists of the tweets for the British Grand Prix.

t0WHNqwDjXc-ExampleYoutubeComments.csv - This is an example data file which consists of the comments extracted from youtube. "t0WHNqwDjXc" is the video id of the youtube video. This is the highlights video of 2017 British Grand Prix.

youtube_sentiments.csv - This contains data about 84 Grand Prix races conducted during 2017 – 2021. There is a corresponding race highlight video posted by the official Formula 1 youtube channel https://www.youtube.com/c/F1/about .

Race Id – A unique id to each race which is collected from the egast developer F1 race dataset. http://ergast.com/mrd/

Grand Prix – Name of the grand prix.

Year – year in which the grand prix is conducted.

View count – The view count of the highlights video from youtube.

Like count – The like count of the highlights video from youtube.

Dislike count – The dislike count of the highlights video from youtube.

Likecount.viewcount - Ratio of like count and view count.

Dislikecount.viewcount - Ratio of dislike count and view count.

Likecount.dislikecount - Sum of like and dislike counts.

X.likecount.dislikecount..viewcount - Ratio of sum of like and dislike counts and view count.

Commentcount – Comment count of the highlights video from youtube.

Number.of.extracted.comments - number of comments of the highlights video that could be extracted from the video. Some of the comments could not extracted in case if the user restricts its access.

Numberofusefulcomments – The comments from which at least one sentence with a positive or negative sentiment could be inferred.

Totalpos – Total number of positive sentiments from the extracted comments for highlights video of a particular race.

Totalneg - Total number of negative sentiments from the extracted comments for highlights video of a particular race.

Excitement_index – Ratio of the sum of totalpos and totalneg to the numberofusefulcomments.

Wetweather – Boolean to indicate if a particular race is declared as wet weather race or not. This is extracted from reddit website.
https://www.reddit.com/r/formula1/comments/g0plkr/list_of_wet_weather_races_and_wins_by_driver/

Unfin – Number of drivers who did not finish the race. I.e they exited the race due to various technical reasons or crashes. This info is aggregated from the results info from the ergast dataset http://ergast.com/mrd/

Novertakes – Number of overtakes that occurred in a race. This is extracted from reddit.
https://www.reddit.com/r/formula1/comments/nf4jkq/f1_overtaking_database_19942020/

Totaldrivers – Total number of drivers in the race. This is also extracted from the results info from the ergast dataset http://ergast.com/mrd/

constructorWon - The constructor which won the corresponding race.

Note – The statistics like count, dislike count, view count and comment count for the individual race highlight videos cannot be accurate to present count as they can change overtime.

twitter_sentiments.csv - This file contains data about 116 Grand Prix races conducted during 2014 – 2021. There are twitter handles for every grand prix, tweets are collected using twitter api and using these twitter handles which are unique to each race.

Race Id – A unique id to each race which is collected from the egast developer F1 race dataset. http://ergast.com/mrd/

Year – year in which the grand prix is conducted.

Grand Prix – Name of the grand prix.

Date – The date on which the particular grand prix is conducted.

Total_tweets – The number of tweets that could be queried and extracted using the twitter api for each individual race.

Useful_tweets - The number of tweets from which at least one sentence with a positive or negative sentiment could be inferred.

Positive_tweets – Total number of positive sentiments from the extracted tweets of a particular race.

Negative_tweets - Total number of negative sentiments from the extracted tweets of a particular race.

Excitement_index – Ratio of the sum of totalpos and totalneg to the useful_tweets.

Novertakes – Number of overtakes that occurred in a race. This is extracted from reddit.
https://www.reddit.com/r/formula1/comments/nf4jkq/f1_overtaking_database_19942020/

Wetweather – Boolean to indicate if a particular race is declared as wet weather race or not. This is extracted from reddit website.
https://www.reddit.com/r/formula1/comments/g0plkr/list_of_wet_weather_races_and_wins_by_driver/

Unfin – Number of drivers who did not finish the race. I.e they exited the race due to various technical reasons or crashes. This info is aggregated from the results info

from the ergast dataset http://ergast.com/mrd/

Totaldrivers – Total number of drivers in the race. This is also extracted from the results info from the ergast dataset http://ergast.com/mrd/

constructorWon - The constructor which won the corresponding race.

aggregated_basedOnFanRatings.csv - This file consists of data about 202 Grand Prix races conducted during 2008 – 2018.

Race Id – A unique id to each race which is collected from the egast developer F1 race dataset. http://ergast.com/mrd/

Year – year in which the grand prix is conducted.

GPNAME – Name of the grand prix.

Novertakes – Number of overtakes that occurred in a race. This is extracted from reddit.
https://www.reddit.com/r/formula1/comments/nf4jkq/f1_overtaking_database_19942020/

Wetweather – Boolean to indicate if a particular race is declared as wet weather race or not. This is extracted from reddit website.
https://www.reddit.com/r/formula1/comments/g0plkr/list_of_wet_weather_races_and_wins_by_driver/

Unfin – Number of drivers who did not finish the race. I.e they exited the race due to various technical reasons or crashes. This info is aggregated from the results info from the ergast dataset http://ergast.com/mrd/

Totaldrivers – Total number of drivers in the race. This is also extracted from the results info from the ergast dataset http://ergast.com/mrd/

constructorWon - The constructor which won the corresponding race.

RATING – The rating for each individual race given by the fans. This data is collected from kaggle (https://www.kaggle.com/codingminds/formula-1-race-fan-ratings) And is provided by racefans.net.