Analysis Report

Table of Contents

Data Cleaning

There are a total of 11 csv files provided as source data. In those 11 files, 2 files are for driver standing information. Analysis on the 2 files indicated that ‘driver_standings.csv’ and ‘'driverStandings.csv’ has the same number of columns, with the former having around 1200 more rows, and rows that exist in both files have the same data. Hence, the file with more data ‘driver_standings.csv’ was used.

The file with the parent table ‘results.csv’ also had 2 places cleaned. One is fixing incorrect values for column ‘postionOrder’. And the other fix was filling in missing values for the column ‘fastestLapTime’ from file ‘lapTimes.csv’, which recorded lap time details for each lap that each driver had done for each race.

Assumptions:

Below is a consolidated list of all assumptions made and throughout cleaning and analysis of the data sets.

- For the 2 driver standing files, use the sheet with more data: driver_standings
- For the mismatching values from in results.csv, given both position and positionText have the same value, assume the error is with positionOrder
- For the missing raceId in results.csv, the extra raceIds in races.csv are records of races for the 2019, 2020 season which happened after the date 2018-11-25, the last race for 2018 season.
- For the data error in results.csv, it is not feasible to go through each row and research whether the data are accurate. Hence for this task, will have to stick with what is provided by the files.

Tasks

1. Constructors’ success based on championship season wins

Constructor championships did not happen until 1958, hence this report will consider the wins from 1958 and onwards. The bar chart below shows the no.of wins for all the constructors who have at least won 1 championship. From the bar chart, Ferrari is the constructor with the most wins, and has far more wins compared to the next constructor McLaren. In the analysis script and output dataset, there is a separate graph and table which includes data from 1950.

Do note that the numbers below do not exactly match official f1 records, as explored and discussed further in the analysis script (Tasks Analysis.ipynb’) and mentioned in section assumptions, this was mainly due to data issues with the source data.

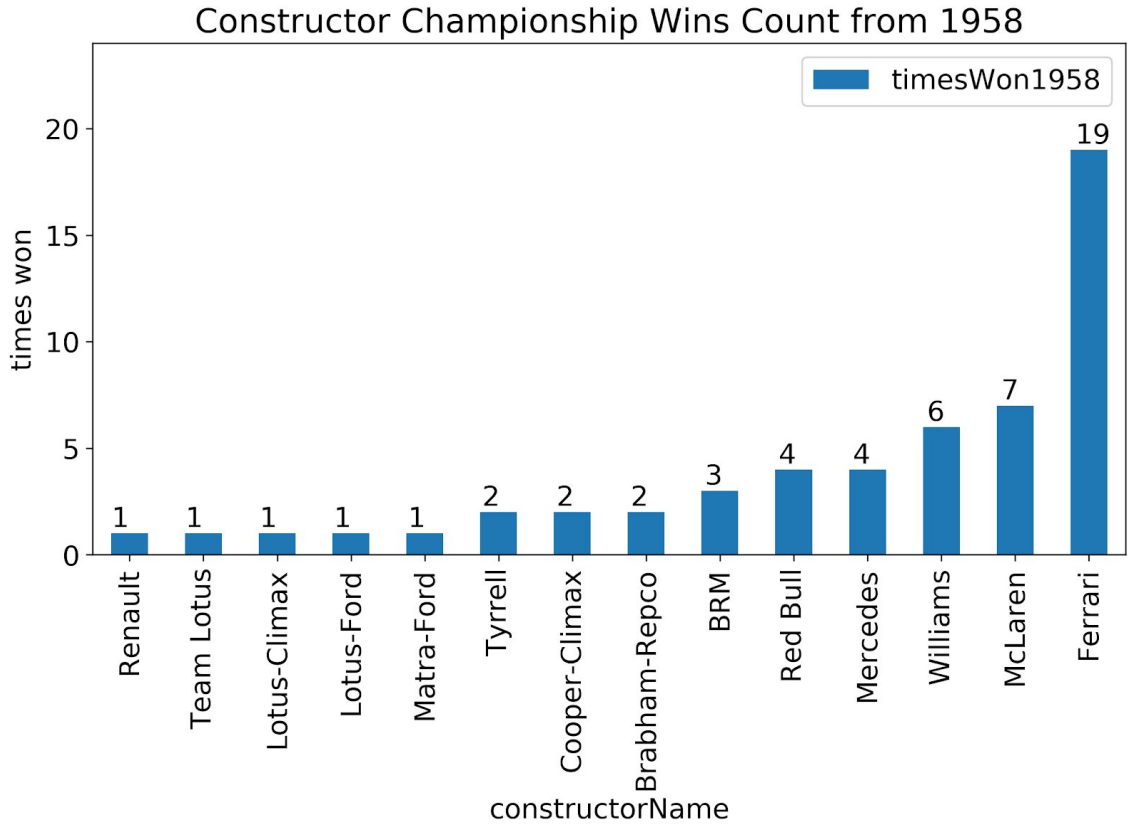

figure 1, constructor no.of championship wins

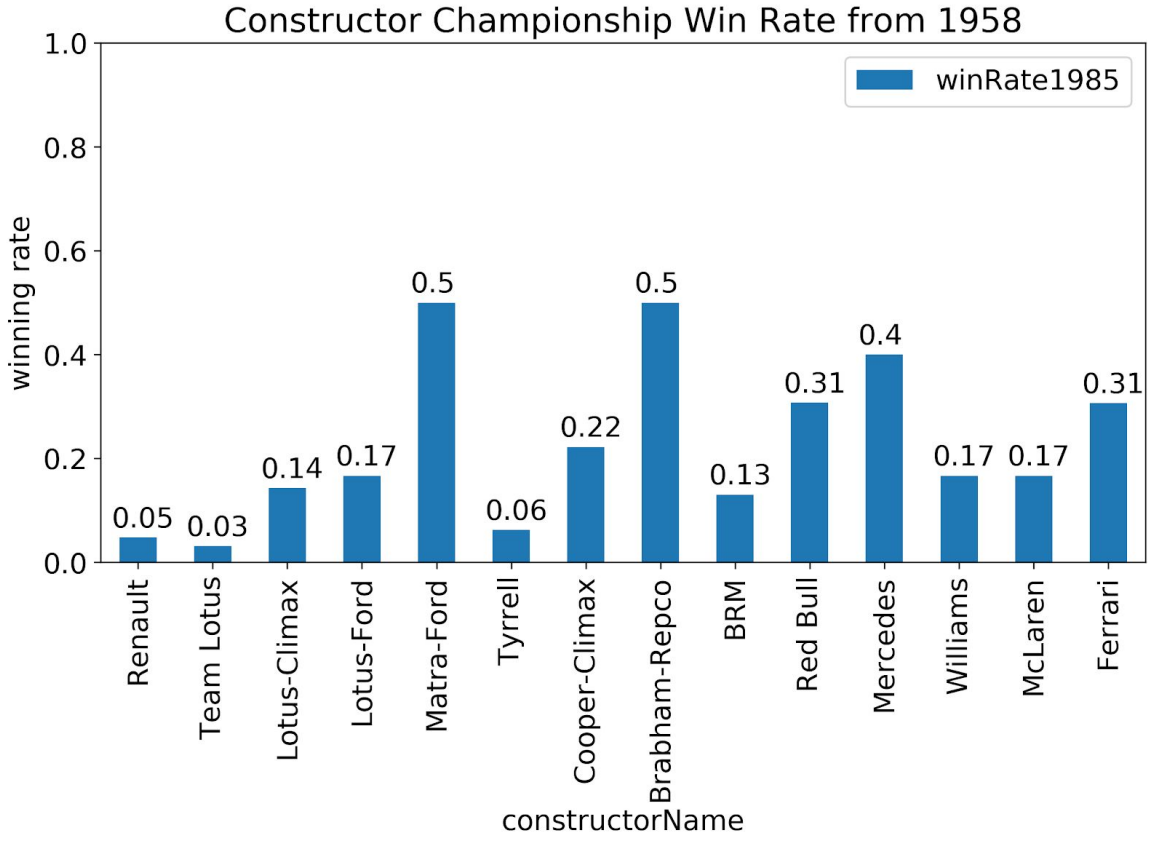

figure 1, constructor championship win rates

Upon checking the win rate, calculated based on the no.of wins over no.of seasons participated. It can be seen that Ferrari is sitting at 0.31, with the highest being Martra-Ford and Brabham-Repco at 0.5. However, Martra-Ford has only participated in 2 championships, and 4 for Brabham-Repco.

2. Drivers' success based on race wins

Figure 3 below shows the top 10 drivers based on no.of race wins, as well as their stats on winning second and third place. One interesting thing is that 9 out of these drivers are from Europe, with only 1 driver from Japan (refer to table 1 below).

Looking at the win rates graph for drivers (figure 4 below), the highest win rate is close to 0.5 by driver Lee Wallard, who did not appear in the top 10 graph for counting the total times they have won the races. Lewis Hamilton, who had the most wins, only ranked on the 6th position for win rates.

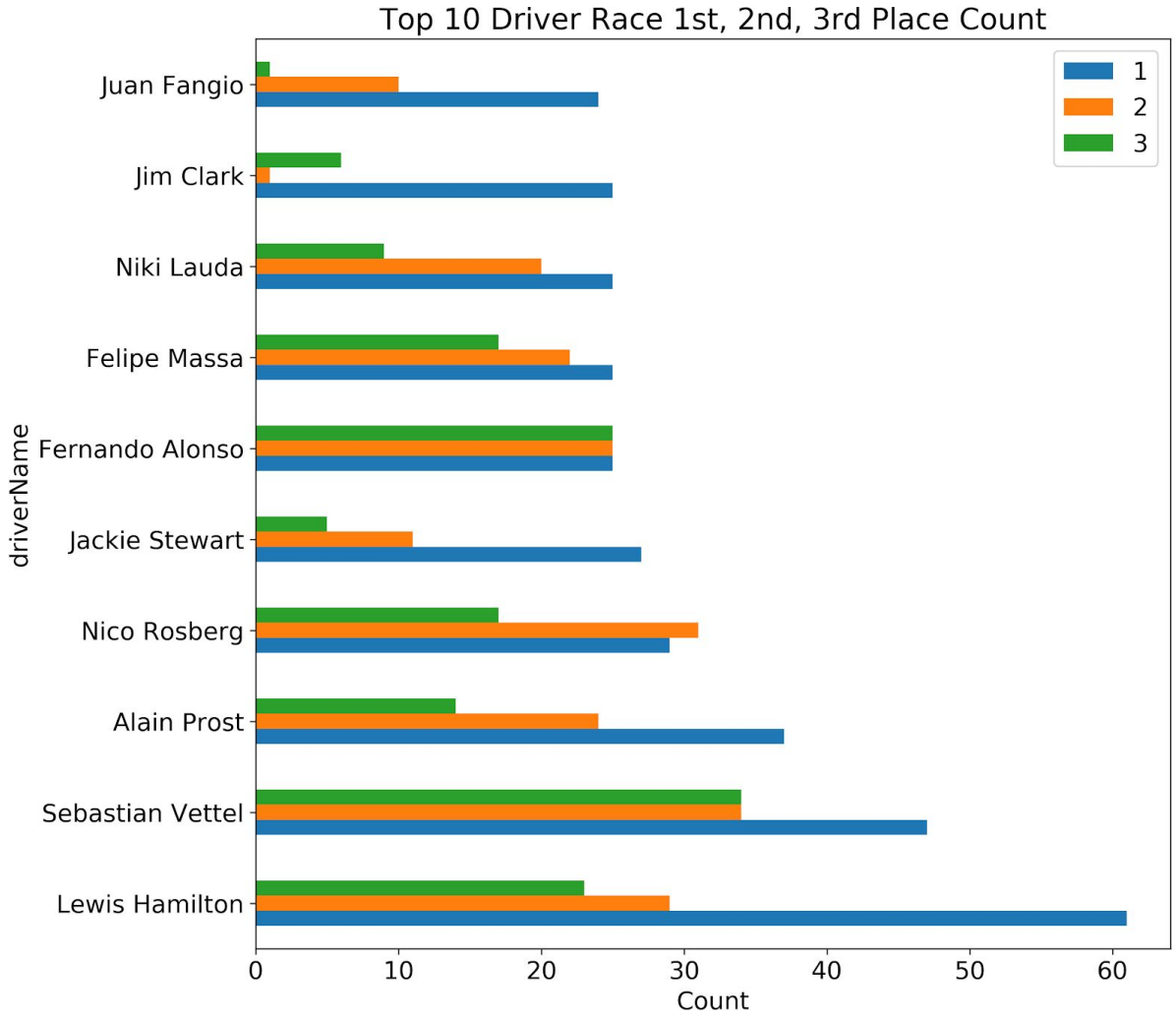

figure 3, top 10 drivers by race wins

driverRef	number	code	dob	nationality	appearances	winRate	driverName
hamilton	44.0	HAM	07/01/1985	British	207	0.294686	Lewis Hamilton
heidfeld	NaN	HEI	10/05/1977	German	177	0.028249	Nick Heidfeld
rosberg	6.0	ROS	27/06/1985	German	203	0.142857	Nico Rosberg
alonso	14.0	ALO	29/07/1981	Spanish	289	0.086505	Fernando Alonso
kovalainen	NaN	KOV	19/10/1981	Finnish	110	0.045455	Heikki Kovalainen
nakajima	NaN	NAK	11/01/1985	Japanese	36	0.027778	Kazuki Nakajima
bourdais	NaN	BOU	28/02/1979	French	25	0.080000	SÌ©bastien Bourdais
raikkonen	7.0	RAI	17/10/1979	Finnish	266	0.041353	Kimi RÌ_ikkÌ¦nen
kubica	NaN	KUB	07/12/1984	Polish	76	0.026316	Robert Kubica
glock	NaN	GLO	18/03/1982	German	94	0.042553	Timo Glock

table 1, top 10 driver's details by race wins

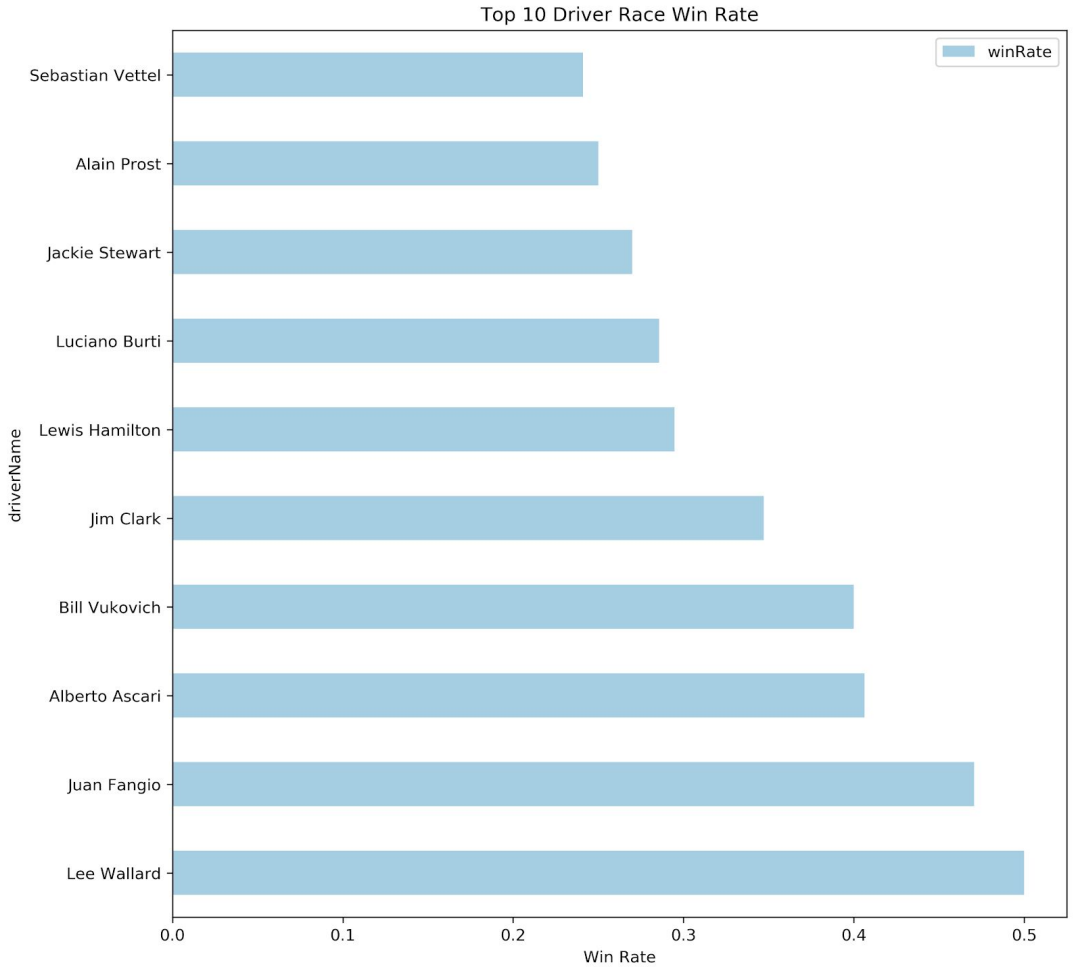

figure 4, driver race win rates

3. Risks associated with race tracks based on accidents and collisions

This was analysed based on the records in the status column in status.csv, which consisted of words such as collision, accident, as well as other status which could have been caused by accidents/collisions. During the analysis, two tables were generated, one was only based on status specifically mentioned collision or accidents. And the other table based on all the status potentially due to accident/collision as well. Table below lists them out as tier 1 and tier 2 status.

Status	**Tier**
Fatal accident, Accident, Collision, Collision damage	1
Injured, Injury, Eye injury, Broken wing, Front wing, Heat shield fire, Oil leak, Fuel leak, Engine fire, Spun off, Puncture, Chassis, Safety, Safety concerns, Spun off	2

table 3 status tier

The following treemap graphs are produced by the 2 different datasets, based on the top 20 most dangerous tracks. It can be observed that no matter the tier 1 or both tier 1 and tier 2 status were included, the most dangerous tracks never changes, it is monaco, which is Circuit de Monaco in Monaco. For the rest of the tracks, they are mostly the same tracks as well despite which tier of status is being looked at, just their position on the most dangerous scale would change.

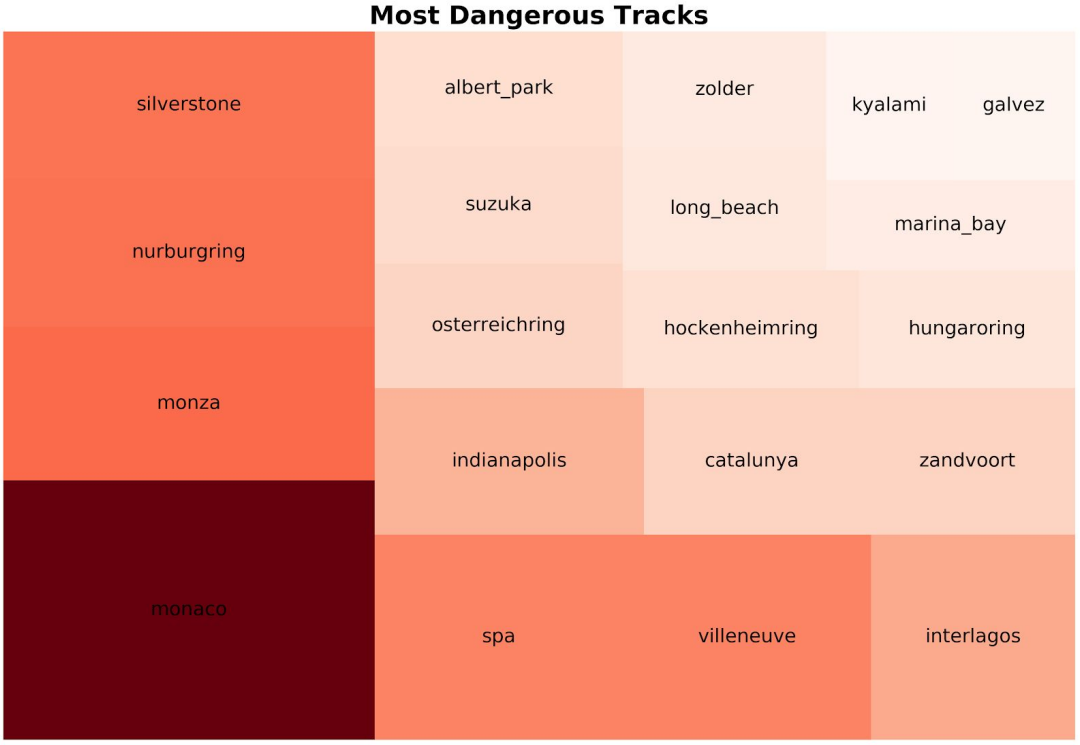

figure 5, dangerous track tier1

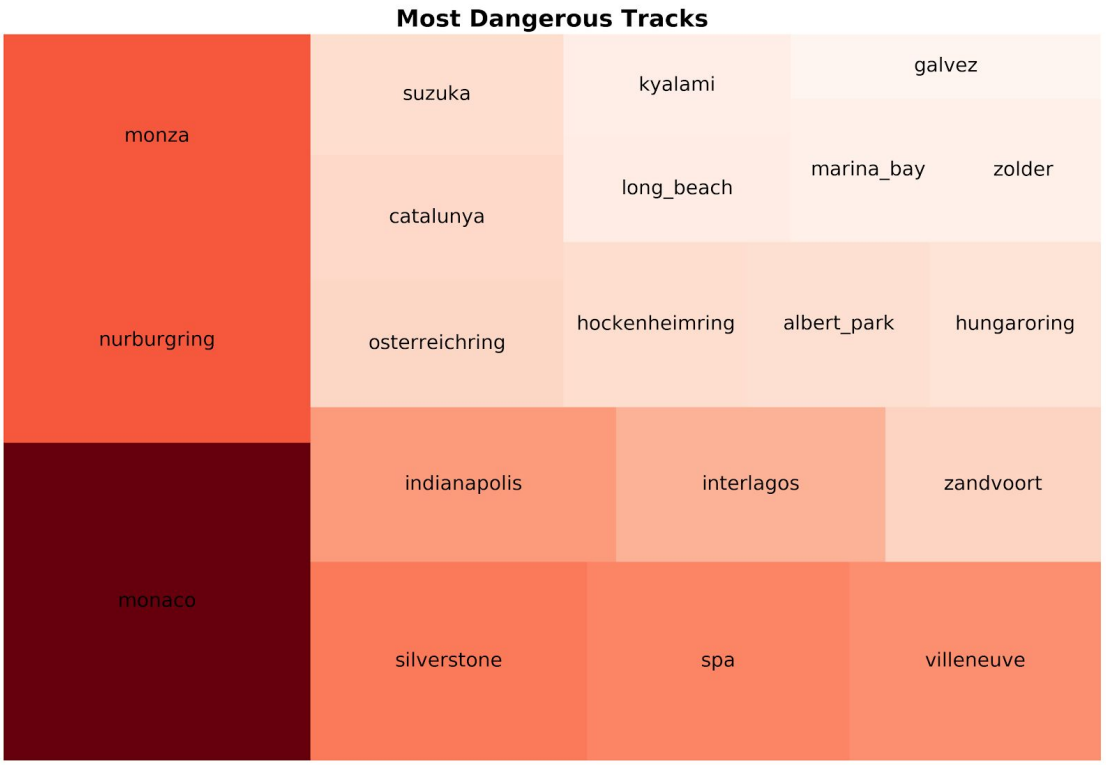

figure 6, dangerous track tier2

4. Fastest race tracks based on lap times

Figure 7 below shows the lap times for each of the tracks. It shows the quickest track is Red Bull Ring, with the quickest single lap time of 67.4 seconds , which can be seen from table 4.

	circuitId	circuitName	fastestLapTimeSeconds
31	70	Red Bull Ring	67.400
22	23	A1-Ring	68.337
18	19	Indianapolis Motor Speedway	70.399
17	18	AutÌ_dromo JosÌ© Carlos Pace	71.000
6	7	Circuit Gilles Villeneuve	73.622
9	10	Hockenheimring	73.780
5	6	Circuit de Monaco	74.439
7	8	Circuit de Nevers Magny-Cours	75.045
3	4	Circuit de Barcelona-Catalunya	75.641
10	11	Hungaroring	76.207
19	20	NÌ_rburgring	78.354

table 4, quickest tracks

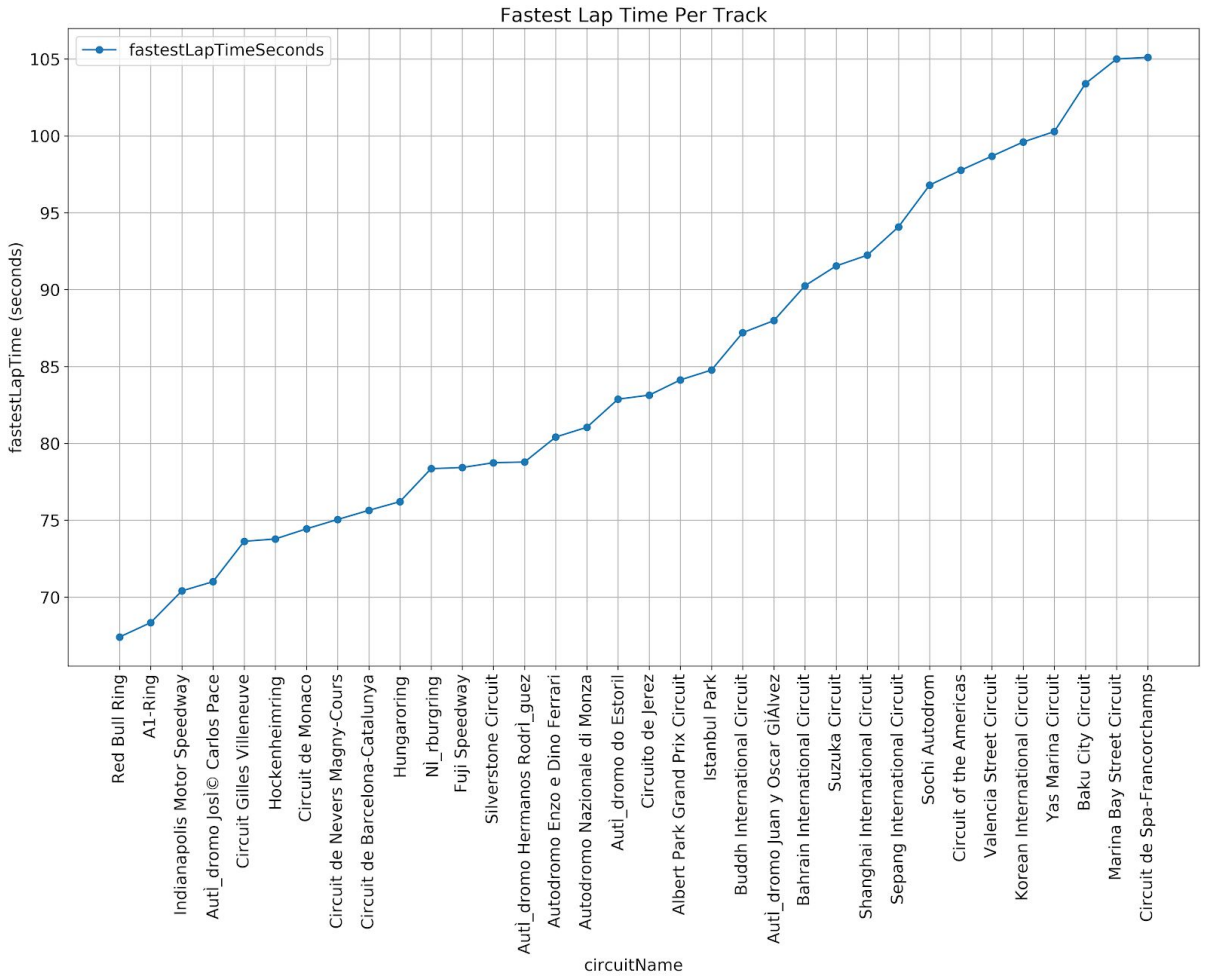

figure 7, quickest tracks line chart

5. Recommend engineered features

To win championships, teams and players need to win as many races as possible, which is the point they get. The below correlation matrix indicates that grid has a negative 0.36 correlation with points. Making grid a potential significant feature to assist the machine learning model in predicting the correct winner.

	resultId	raceId	driverId	constructorId	number	grid	position	positionOrder	points	laps	milliseconds	fastestLap
resultId	1.000000	0.697219	0.468243	0.475907	0.211245	-0.000071	0.091598	-0.027998	0.175539	0.082383	0.312763	0.269450
raceId	0.697219	1.000000	0.618638	0.327582	0.215549	-0.009704	0.065449	0.061486	0.094530	0.015263	0.245213	0.222743
driverId	0.468243	0.618638	1.000000	0.248030	0.254334	0.061253	0.167014	0.074550	-0.097159	0.046384	0.294030	0.092797
constructorId	0.475907	0.327582	0.248030	1.000000	0.208789	0.120045	0.185173	0.101200	-0.074732	0.010727	0.330859	0.063550
number	0.211245	0.215549	0.254334	0.208789	1.000000	0.189036	0.254152	0.241480	-0.148713	-0.036386	0.212313	0.027622
grid	-0.000071	-0.009704	0.061253	0.120045	0.189036	1.000000	0.651252	0.131765	-0.356515	0.084996	0.109112	-0.009618
position	0.091598	0.065449	0.167014	0.185173	0.254152	0.651252	1.000000	0.999978	-0.645920	-0.056088	-0.028064	0.047228
positionOrder	-0.027998	0.061486	0.074550	0.101200	0.241480	0.131765	0.999978	1.000000	-0.564764	-0.658559	-0.029167	-0.296632
points	0.175539	0.094530	-0.097159	-0.074732	-0.148713	-0.356515	-0.645920	-0.564764	1.000000	0.258257	-0.058453	0.200755
laps	0.082383	0.015263	0.046384	0.010727	-0.036386	0.084996	-0.056088	-0.658559	0.258257	1.000000	0.599665	0.673261
milliseconds	0.312763	0.245213	0.294030	0.330859	0.212313	0.109112	-0.028064	-0.029167	-0.058453	0.599665	1.000000	0.225515
fastestLap	0.269450	0.222743	0.092797	0.063550	0.027622	-0.009618	0.047228	-0.296632	0.200755	0.673261	0.225515	1.000000
rank	0.012649	-0.015812	0.160892	0.255656	0.198565	0.608204	0.741391	0.648907	-0.540120	-0.217628	0.076088	-0.263967
statusId	0.036760	0.091815	0.082242	0.132550	0.185504	-0.138471	0.371253	0.528214	-0.268009	-0.353389	-0.009488	-0.196327

From results in task1 about constructor wins, it was also observed that certain constructors had won much championship than the rest, particularly Ferrari. Even though the correlation is very low with points, it does have significant impact directly linked to the result. Hence, constructor could also assist with the model.

In addition, the data set could be expanded and further analysis could be conducted to assist. For example, weather could have an important impact on the result, particularly it increases the chance of accidents.

Overall, recommended engineered features are grid and constructor.

Appendix

1. Source Data Files

File Name	Columns
circuits.csv	circuitId, circuitRef, name, location, country, lat, lng, alt, url
constructors.csv	constructorId, constructorRef, name, nationality, url
constructor_standings.csv	constructorStandingsId, raceId, constructorId, points, position, positionText, wins
drivers.csv	driverId, driverRef, number, code, forename, surname, dob, nationality, url
'driverStandings.csv	driverStandingsId, raceId, driverId, points, position, positionText, wins
driver_standings.csv	driverStandingsId, raceId, driverId, points, position, positionText, wins
lapTimes.csv	raceId, driverId, lap, position, time, milliseconds
pitStops.csv	raceId, driverId, stop, lap, time, duration, milliseconds
races.csv	raceId, year, round, circuitId, name, date, time, url
results.csv	resultId, raceId, driverId, constructorId, number, grid, position, positionText, positionOrder, points, laps, time, milliseconds, fastestLap, rank, fastestLapTime, fastestLapSpeed, statusId
status.csv	statusId, status

2. Analysis Scripts

File Name	Description
‘Explore & Cleaning.ipynb’	Cleaning/wrangling on the source data.
‘Tasks Analysis.ipynb’	Analysis script

3. Cleaned Data

File Name	Description
results_wrangled.csv	Cleaned results.csv

4. Analysis Result Data

File Name	Description
constructor_win_1958plus.csv	constructor championship win result for 1958 and later
constructor_win_alltime.csv	constructor championship win result counting from 1950
dangerous_tracks_tier1.csv	dangerous tracks based on status including ‘accident’ or ‘collision’
dangerous_tracks_tier2.csv	dangerous tracks based on status including ‘accident’ or ‘collision’ plus other status potentially caused by accident or collision
driver_wins.csv	win result for drivers (includes all the position counts)
driver_wins_slimed.csv	win result for drivers (only 1st place count)
fastest_tracks.csv	tracks with their fastest lap times