SUMMARY

Data

Description of the data set

The data in the F1-data set was gathered over a 70 years period in early 1950 to 2021. It contains 13 files within includes: circuits, constructor_results, constructor_standings, constructors, driver_standings, drivers, lap_times, pit_stops, qualifying, races, results, seasons and status.

The dataset consists of all information on the formula 1 races, driver, constructors, qualifying, circuits, lap times, pit stops, championships from 1950 till the latest 2021 season.

We can use the results file to calculate the circuit distance.

2.1 Initial steps

The very first step is always to check if the data needs cleaning by looking for duplicate rows, zero values or Nans where they shouldn't be. Since the data sets files are many and too big, we can inspect 2 data sets visually. The head of a dataset looks like this:

Circuit.csv

	circuitId	circuitRef	name	location	country	lat	lng	alt	url
0	1	albert_park	Albert Park Grand Prix Circuit	Melbourne	Australia	-37.84970	144.96800	10	http://en.wikipedia.org/wiki/Melbourne_Grand_P...
1	2	sepang	Sepang International Circuit	Kuala Lumpur	Malaysia	2.76083	101.73800	18	http://en.wikipedia.org/wiki/Sepang_Internatio...
2	3	bahrain	Bahrain International Circuit	Sakhir	Bahrain	26.03250	50.51060	7	http://en.wikipedia.org/wiki/Bahrain_Internati...
3	4	catalunya	Circuit de Barcelona-Catalunya	Montmeló	Spain	41.57000	2.26111	109	http://en.wikipedia.org/wiki/Circuit_de_Barcel...
4	5	istanbul	Istanbul Park	Istanbul	Turkey	40.95170	29.40500	130	http://en.wikipedia.org/wiki/Istanbul_Park
...	...	...	...	...	...	...	...	...	...
72	73	BAK	Baku City Circuit	Baku	Azerbaijan	40.37250	49.85330	-7	http://en.wikipedia.org/wiki/Baku_City_Circuit
73	74	hanoi	Hanoi Street Circuit	Hanoi	Vietnam	21.01660	105.76600	9	http://en.wikipedia.org/wiki/Hanoi_Street_Circuit
74	75	portimao	Autódromo Internacional do Algarve	Portimão	Portugal	37.22700	-8.62670	108	http://en.wikipedia.org/wiki/Algarve_Internati...
75	76	mugello	Autodromo Internazionale del Mugello	Mugello	Italy	43.99750	11.37190	255	http://en.wikipedia.org/wiki/Mugello_Circuit
76	77	jeddah	Jeddah Street Circuit	Jeddah	Saudi Arabia	21.54330	39.17280	15	http://en.wikipedia.org/wiki/Jeddah_Street_Cir...

77 rows × 9 columns

results.csv

ructorId	number	grid	position	positionText	positionOrder	points	laps	time	milliseconds	fastestLap	rank	fastestLapTime	fastestLapSpeed	statusId
1	22	1	1	1	1	10.0	58	1:34:50.616	5690616	39	2	1:27.452	218.300	1
2	3	5	2	2	2	8.0	58	+5.478	5696094	41	3	1:27.739	217.586	1
3	7	7	3	3	3	6.0	58	+8.163	5698779	41	5	1:28.090	216.719	1
4	5	11	4	4	4	5.0	58	+17.181	5707797	58	7	1:28.603	215.464	1
1	23	3	5	5	5	4.0	58	+18.014	5708630	43	1	1:27.418	218.385	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	6	18	16	16	16	0.0	70	\N	\N	62	11	1:08.520	226.865	11
117	5	11	17	17	17	0.0	69	\N	\N	51	8	1:08.420	227.196	4
210	47	19	18	18	18	0.0	69	\N	\N	56	18	1:09.394	224.007	12
210	9	20	19	19	19	0.0	69	\N	\N	49	19	1:09.757	222.842	12
214	31	17	\N	R	20	0.0	0	\N	\N	\N	0	\N	\N	4

The first data set files look fine, counting the number of valid entries confirms this meanwhile, the second data set contains zero values or NaNs in the variables.

Description of the data set

Pandas **describe ()** can provide a quick summary of the data set as outlined in the notebook. However, without looking at the data in more detail, we cannot yet state what we think about the distance the drivers drove in one lap. In the notebook I calculated the distance since we require the Lap speed and time to get the distance of the circuit. We can analyze the circuit distance by doing the following analyses and method below.

Methods & Analysis

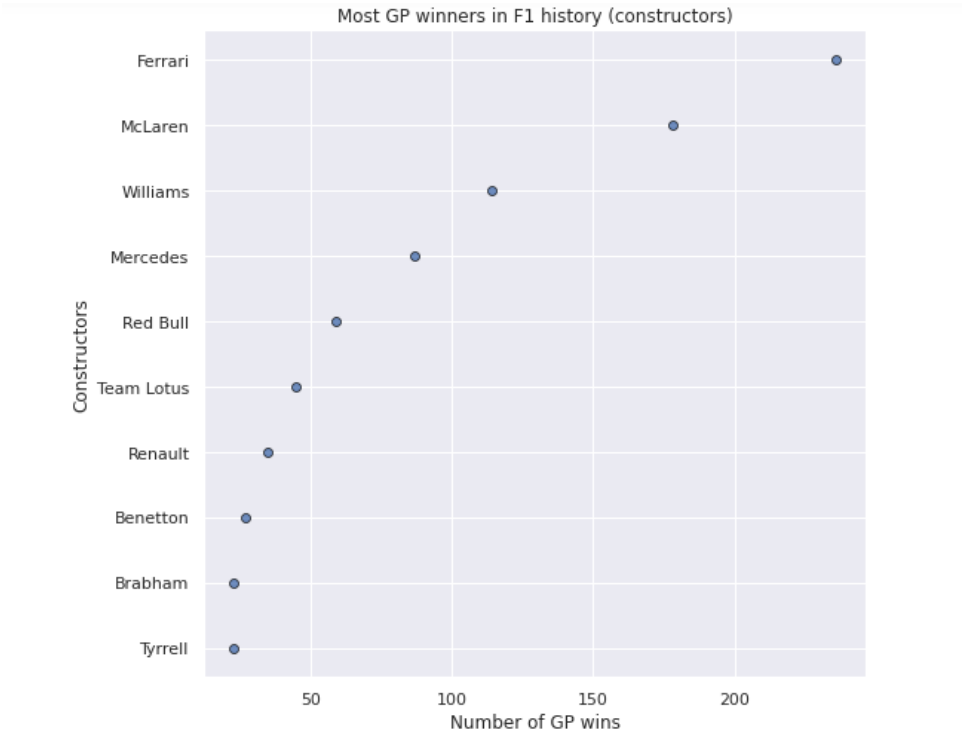
A model can be created using Linear Regression, a machine learning approach that allows us to map numeric inputs to numeric outputs, or models the relationship between one or more variables.

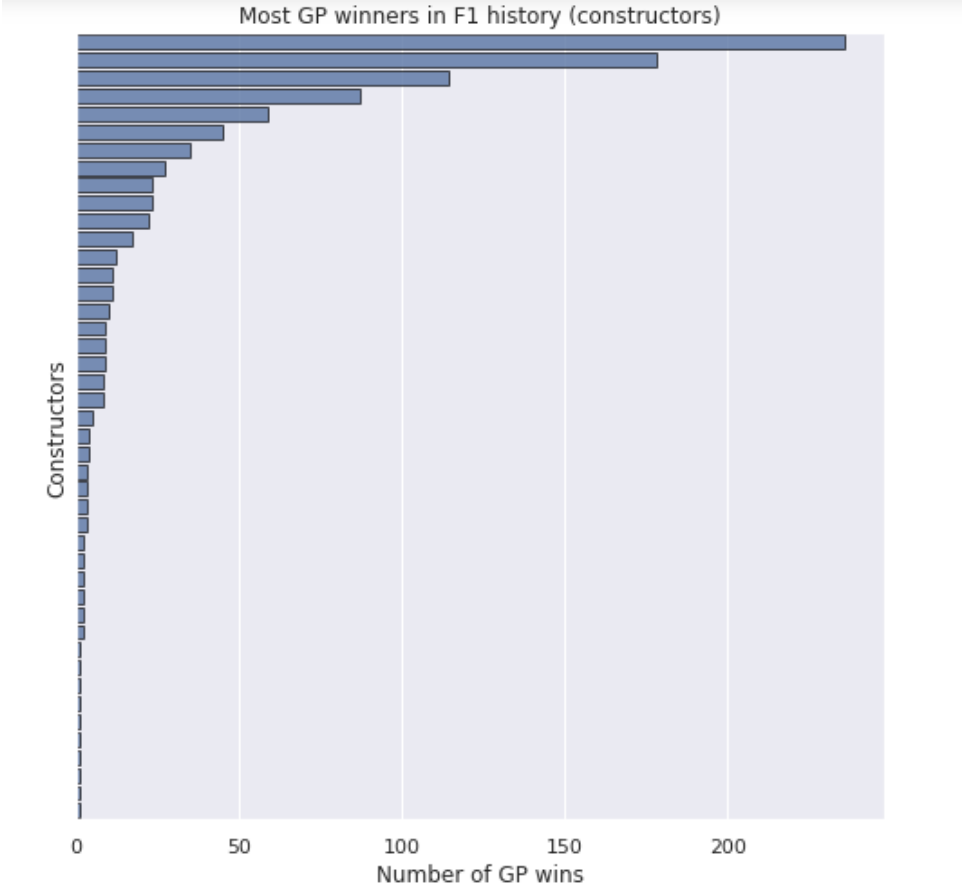
In Linear Regression, there are two sorts of variables: dependent variables and independent variables

Though there are different ways we can get the predicted circuit distance output by analyzing the datasets in Results.csv. Another method of analysis is Gaussian Process (GP). Gaussian Process is an addition to standard scikit-learn estimator API. Gaussian Process Regressor allows prediction, without prior fitting hence provides an additional method sample_y(x), which evaluates samples drawn from the GPR at given inputs. GP is a distribution over functions that takes two parameters, namely the mean (m) and the kernel function K to ensure smoothness.

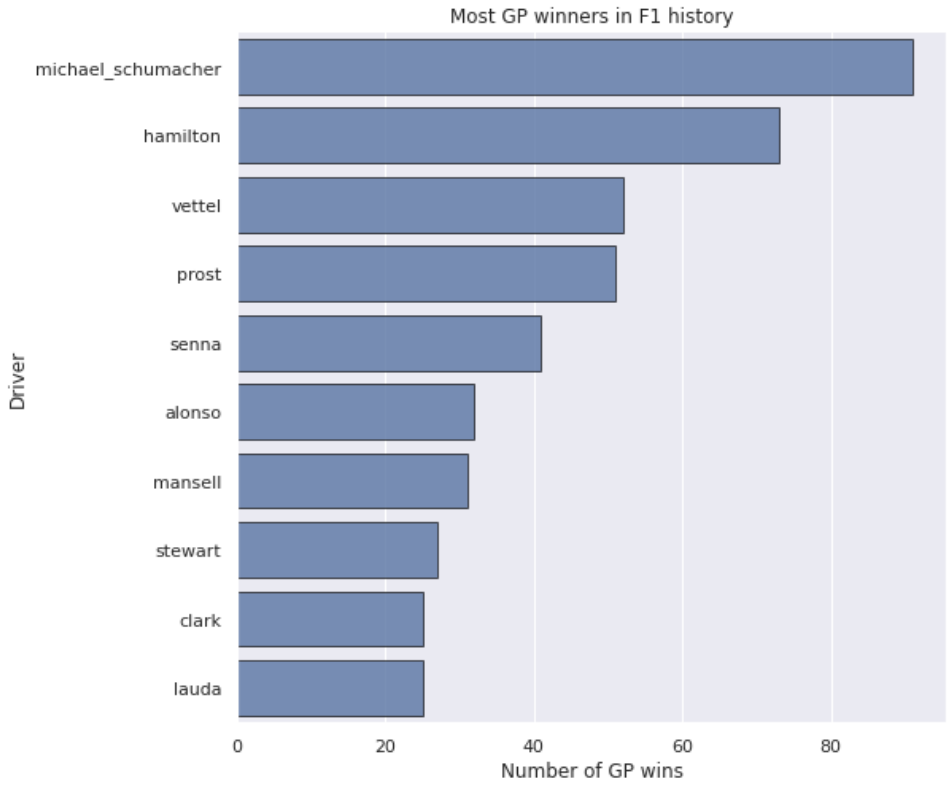
For my case, Linear progression was much better compared to Gaussian Process since it cleans data before finding the data prediction.

Linear regression can also be seen in the relationship between the constructors and the Number of GP wins in the most GP winners, The drivers and Most GP winners in F1 history as show below





These are the most successful Formula One drivers in terms of total GP victories. Michael Schumacher is the most successful driver, with 91 victories. Lewis Hamilton and Sebastian Vettel have 73 and 52 victories in Formula One, respectively. These two drivers are the only ones on the list who aren't retired, thus they have a chance to challenge Schumacher's dominance.

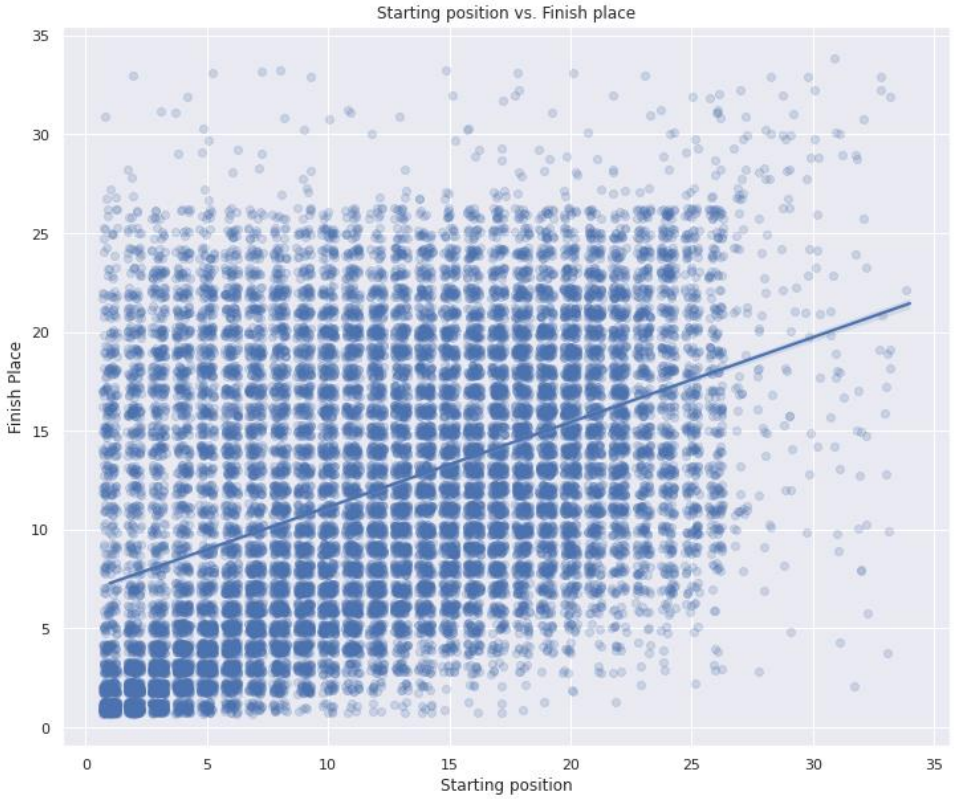


In linear regression, the number of GP wins made affects the list of the highest constructor the relationship between the two variables.

Bivariate Exploration

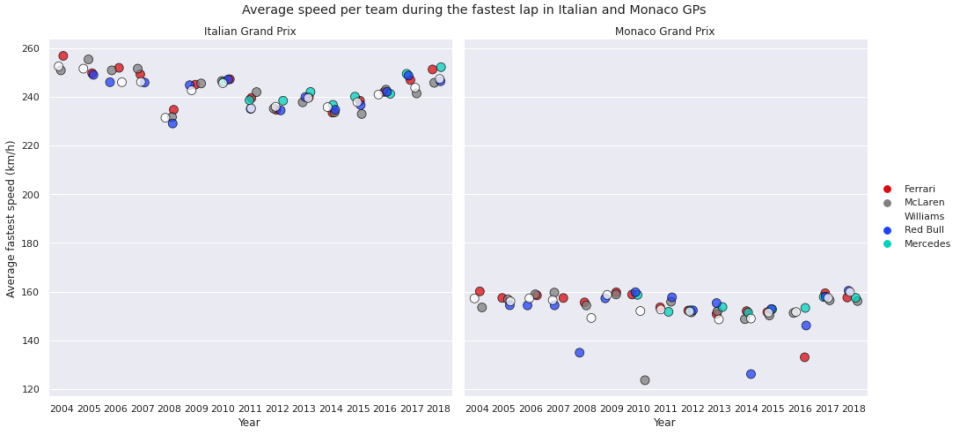
I'll look into the relationship between beginning and finishing positions in this part. In addition, I'll investigate how speed has evolved over time at various tracks.

The diagram below we can see a linear link between the starting and finishing positions on this graph, which is exactly what we expected. We can also see that the majority of races had up to 25 drivers, with some having even more than 30.



Multivariate Exploration

In this section I will analyze the difference in speeds by teams in Mozna (Italian GP) and Monaco. I will try to see if some teams are better than others on different tracks.



Conclusion

In conclusion, the driver needs the circuit distance to calculate the speed required to win the championships race in a particular circuit.

The main findings of this analysis are:

- Top 10 GP driver winners in F1 history
- Top 10 GP constructor winners in F1 history
- Highest time taken for each laps
- Highest speed taken for each lap
- Distance for each circuit
- Drivers with the most points in F1 history

Reference

1. *Higham, Peter (2003). The Complete Book of Formula One. Motorbooks International.*
2. *Hayhoe, David & Holland, David (2006). Grand Prix Data Book (4th edition). Haynes, Sparkford, UK.*
3. *Sandipan Dey (2020). Gaussian Process Regression with Python. Simply Data Science.*
4. *Jordan, (2007). Privateer era is over. ITV-F1.com.*
5. *FIA Rules & Regulations Sporting Regulations, (2006). Formula One.*
6. *Smith and Luke, (2020). All 10 Formula 1 teams sign up for a new Concorde Agreement. Autosport.com.*
7. *Benson Andrew (2 June, 2020). Formula 1 season to start with races in Europe. BBC Sport.*
8. *Baldwin, Alan (17 February 2001). F1 Plans Return of Traction Control. The Independent. Newspaper Publishing.*
9. *Blunsden, John (20 December 1986). Filling Balestre's shoes is no job for a back-seat driver. Financial Times.*