Capstone Project
Background and Objectives

SPORTS ANALYTICS

DATA SCIENCE
INSTITUTE

Background

1. Background

Sports analytics involves using data and statistical analysis to gain insights into various aspects of sports performance, strategy, and driver behavior, ultimately aiding decision-making for teams and athletes.

This Capstone project is all about using sports analytics to understand and improve performance in Formula 1 (F1) races. By digging into the numbers, we want to figure out how teams and drivers have been doing over the years.

This project brings together sports and data to uncover interesting insights in Formula 1.

Background

2. Primary Objectives

1. **To analyze team performance over the years (Team Performance Analytics)**

1. **To analyze** Driver **performance over the years**
2. Consider additional anomalies throughout the seasons
   - e.g. Covid influence and regulation changes
3. To predict the qualifying and race outcomes for the 2023 season
   - Predict the top 3 finishers for each race for the 2023 season.

Background

3. Data

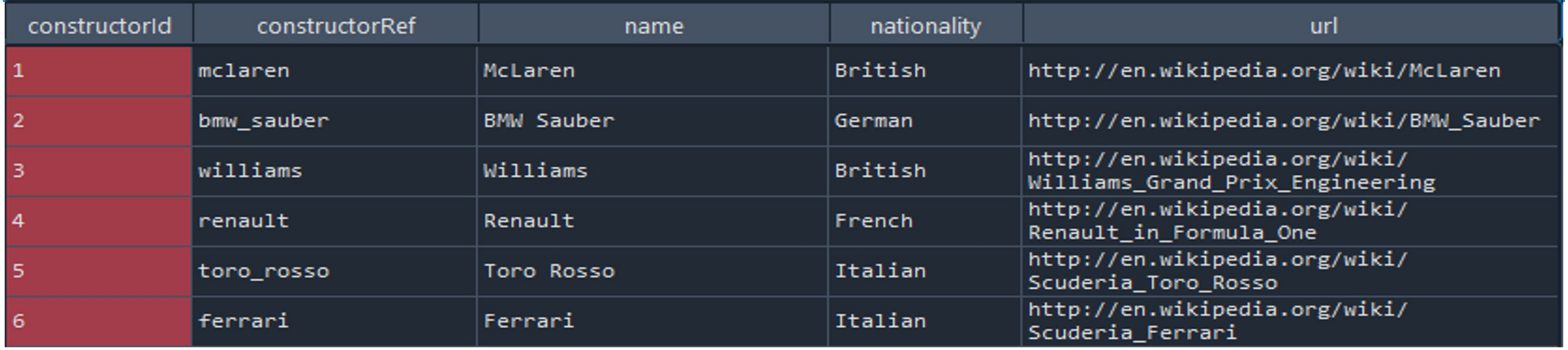
The following datasets are available:

1. constructors
2. drivers
3. qualifying
4. races
5. results

Data: constructors

Content

constructors **data contains identif**ying **data for each** team (or Constructor) **for the entire** history **from** 1950 **to 202**3**.**



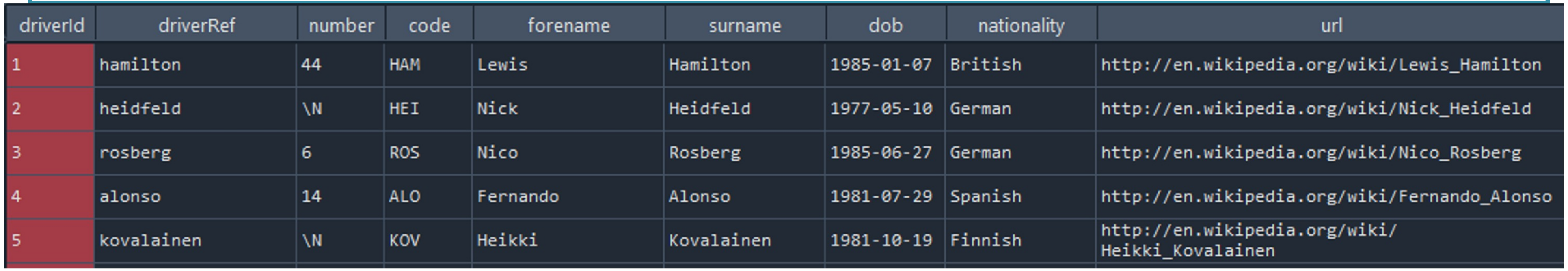
constructorId	constructorRef	name	nationality	url
1	mclaren	McLaren	British	http://en.wikipedia.org/wiki/McLaren
2	bmw_sauber	BMW Sauber	German	http://en.wikipedia.org/wiki/BMW_Sauber
3	williams	Williams	British	http://en.wikipedia.org/wiki/Williams_Grand_Prix_Engineering
4	renault	Renault	French	http://en.wikipedia.org/wiki/Renault_in_Formula_One
5	toro_rosso	Toro Rosso	Italian	http://en.wikipedia.org/wiki/Scuderia_Toro_Rosso
6	ferrari	Ferrari	Italian	http://en.wikipedia.org/wiki/Scuderia_Ferrari

Columns	Description	Type	Possible values
constructorId	Identifying Number of team	Integer	
constructorRef	A username type reference	Character	
name	Team Name	Character	
nationality	Nationality of Team	Character	
url	Link to Wiki page of Team	Character	

Data: drivers

Content

drivers **data contains identif**ying **data for each** driver **for the entire** history **from** 1950 **to 202**3**.**



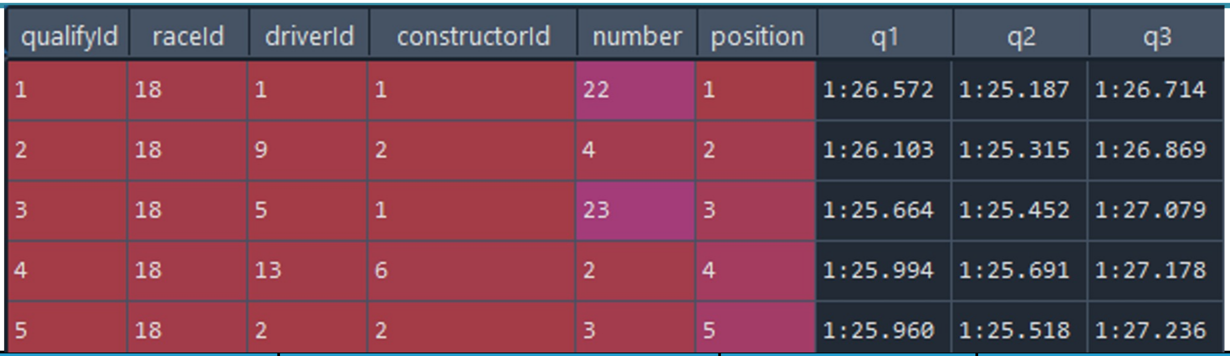
driverId	driverRef	number	code	forename	surname	dob	nationality	url
1	hamilton	44	HAM	Lewis	Hamilton	1985-01-07	British	http://en.wikipedia.org/wiki/Lewis_Hamilton
2	heidfeld	\N	HEI	Nick	Heidfeld	1977-05-10	German	http://en.wikipedia.org/wiki/Nick_Heidfeld
3	rosberg	6	ROS	Nico	Rosberg	1985-06-27	German	http://en.wikipedia.org/wiki/Nico_Rosberg
4	alonso	14	ALO	Fernando	Alonso	1981-07-29	Spanish	http://en.wikipedia.org/wiki/Fernando_Alonso
5	kovalainen	\N	KOV	Heikki	Kovalainen	1981-10-19	Finnish	http://en.wikipedia.org/wiki/Heikki_Kovalainen

Columns	Description	Type	Possible values
driverId	Identifying Number of Driver	Integer	
driverRef	A username type reference	Character	
number	Driver number used in races	Character	
code	Driver screen reference	Character	
forename	Driver first name	Character	
surname	Driver surname	Character	
dob	Driver date of birth	Date	
nationality	Nationality of driver	Character	
url	Link to Wiki page of driver	Character	

Data: qualifying

Content

qualifying **data contains data for each** qualifying session **for the entire** history **from** 1950 **to 202**3**.**



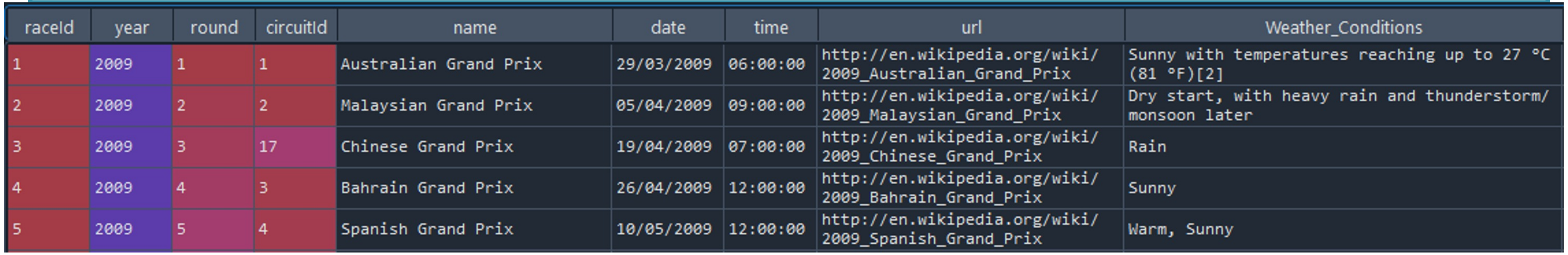
qualifyId	raceId	driverId	constructorId	number	position	q1	q2	q3
1	18	1	1	22	1	1:26.572	1:25.187	1:26.714
2	18	9	2	4	2	1:26.103	1:25.315	1:26.869
3	18	5	1	23	3	1:25.664	1:25.452	1:27.079
4	18	13	6	2	4	1:25.994	1:25.691	1:27.178
5	18	2	2	3	5	1:25.960	1:25.518	1:27.236

Columns	Description	Type	Possible values
qualifyId	Identifying Number of quali session	Integer	
raceId	Identifying Number of the race	Integer	
driverId	Identifying Number of the driver	Integer	
constructorId	Identifying Number of the team	Integer	
number	Car number for the session	Integer	
position	Qualifying position for the race	Integer	
q1	Round 1 Qualifying Time	Character	
q2	Round 2 Qualifying Time	Character	
q3	Round 3 Qualifying Time	Character	

Data: races

Content

races **data contains data for each** race **for the entire** history **from** 1950 **to 202**3**.**



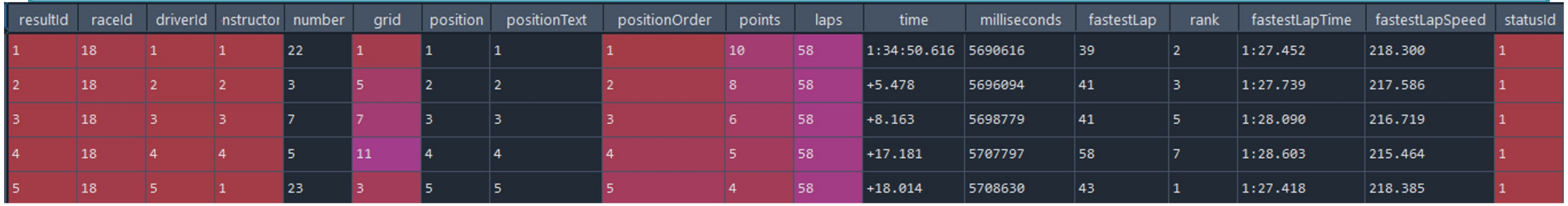
raceId	year	round	circuitId	name	date	time	url	Weather_Conditions
1	2009	1	1	Australian Grand Prix	29/03/2009	06:00:00	http://en.wikipedia.org/wiki/2009_Australian_Grand_Prix	Sunny with temperatures reaching up to 27 °C (81 °F)[2]
2	2009	2	2	Malaysian Grand Prix	05/04/2009	09:00:00	http://en.wikipedia.org/wiki/2009_Malaysian_Grand_Prix	Dry start, with heavy rain and thunderstorm/monsoon later
3	2009	3	17	Chinese Grand Prix	19/04/2009	07:00:00	http://en.wikipedia.org/wiki/2009_Chinese_Grand_Prix	Rain
4	2009	4	3	Bahrain Grand Prix	26/04/2009	12:00:00	http://en.wikipedia.org/wiki/2009_Bahrain_Grand_Prix	Sunny
5	2009	5	4	Spanish Grand Prix	10/05/2009	12:00:00	http://en.wikipedia.org/wiki/2009_Spanish_Grand_Prix	Warm, Sunny

Columns	Description	Type	Possible values
raceId	Identifying Number of race	Integer	
year	Year of race	Integer	
round	Which round the race was held	Integer	
circuitId	ID of the circuit the race was held	Integer	
name	Name of the Grand Prix	Character	
date	Date of the race	Character	
time	Time of the race	Character	
url	Link to Wiki page of race	Character	
Weather_Conditions	Weather conditions for the race	Character	

Data: results

Content

results **data contains** the results **for each** race **for the entire** history **from** 1950 **to 202**3**.**



resultId	raceId	driverId	nstructor	number	grid	position	positionText	positionOrder	points	laps	time	milliseconds	fastestLap	rank	fastestLapTime	fastestLapSpeed	statusId
1	18	1	1	22	1	1	1	1	10	58	1:34:50.616	5690616	39	2	1:27.452	218.300	1
2	18	2	2	3	5	2	2	2	8	58	+5.478	5696094	41	3	1:27.739	217.586	1
3	18	3	3	7	7	3	3	3	6	58	+8.163	5698779	41	5	1:28.090	216.719	1
4	18	4	4	5	11	4	4	4	5	58	+17.181	5707797	58	7	1:28.603	215.464	1
5	18	5	1	23	3	5	5	5	4	58	+18.014	5708630	43	1	1:27.418	218.385	1

Columns	Description	Type
resultId	Identifying Number of the result	Integer
raceId	Identifying Number of the race	Integer
driverId	Identifying Number of the driver	Integer
constructorId	Identifying Number of the team	Integer
number	Car number for the session	Integer
grid	Starting position for the race	Integer
position	Finishing position for the race	Integer
positionText	Finishing position for the race in Text for retirements	Character
positionOrder	Finishing order classification for the race	Integer

Columns	Description	Type
points	Points gained for finishing position	Integer
laps	Number of laps completed	Integer
time	Time taken to complete race	Character
milliseconds	Time taken to complete race in milliseconds	Character
fastestLap	Lap number in which the driver completed their fastest lap	Character
rank	Total Rank	Character
fastestLapTime	Time taken for fastest lap	Character
fastestLapSpeed	Avg. speed for fasktest lap in kmh	Character
StatusId	ID for final status of Driver	Character

Next steps

Data management

- Compile 2008-2022 datasets using Driver/ Team/Race ID
- Data cleaning , Handling missing values and completing Basic Data checks
- Check if any variables needed to be feature coded i.e made into groups or want to be left as continuous variables
- Complete EDA of combined dataset to identify significant Statistics and Visualization.

Descriptive Statistics & Data visualization

- Explore data for team performance analytics and player performance analytics.
- How can this data be presented better visually ?
- Once again post Data visualization check if any variable needs to be feature coded

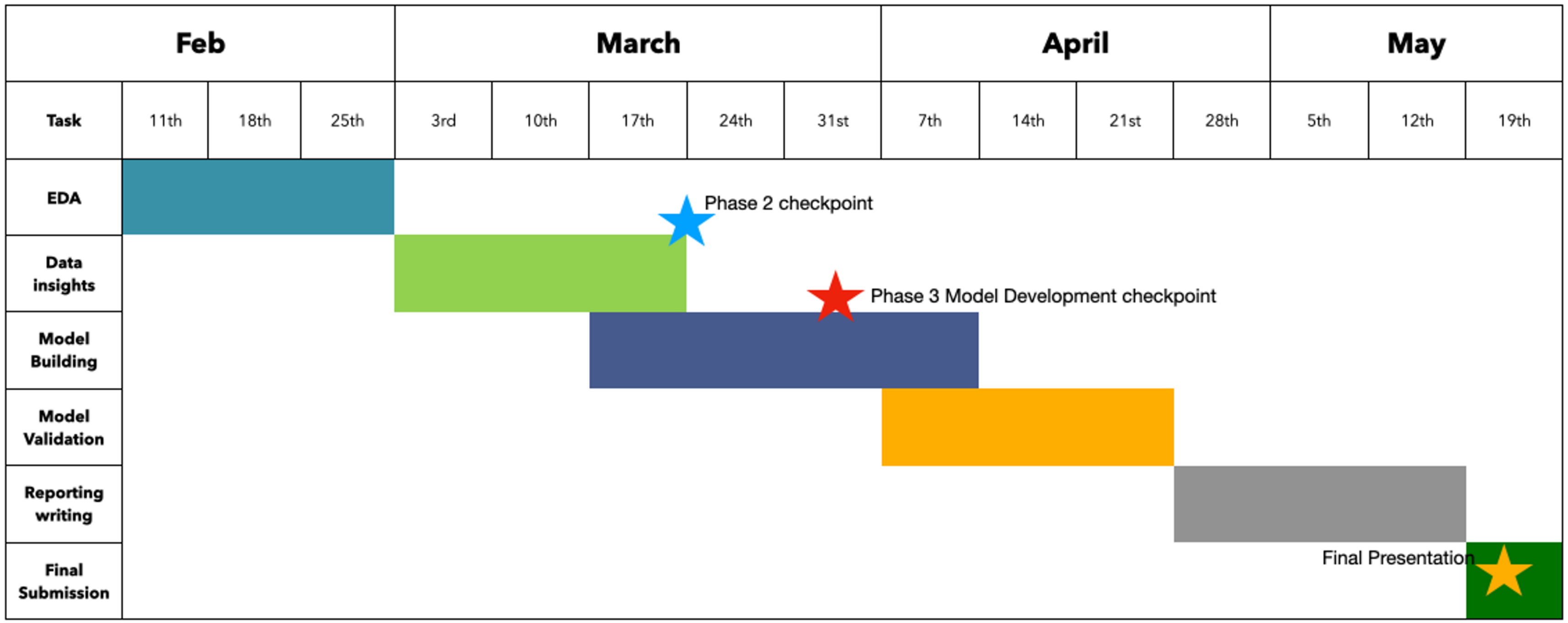
Predictive modelling

- Develop a model to predict the top 3 finishers for each race
- Using different Predictive model techniques to find Significant variables
- Ensure you follow all steps like Train and test data , checking for Multicollinearity
- Check if any other ML technique fits better

Final Output

- Deliver UI for User to interact with Model and allow for insights

Proposed Project Timeline



	Feb			March					April				May		
Task	11th	18th	25th	3rd	10th	17th	24th	31st	7th	14th	21st	28th	5th	12th	19th
EDA	🟦	🟦	🟦				★ Phase 2 checkpoint								
Data insights				🟩	🟩	🟩		★ Phase 3 Model Development checkpoint							
Model Building						🟦	🟦	🟦	🟦						
Model Validation									🟧	🟧	🟧				
Reporting writing												■	■	■	
Final Submission													Final Presentation		🟩 ★