SC1015 Mini-Project

Formula One

Lab Group 3
Dang Huy Phuong
Clara Heng Yih Xian

01

About our Project

Formula One

- “Who will win?”
- Complexity and unpredictability of success

***Motivation**: To better understand which factors contribute most to a driver’s success + predict who will win*

PROBLEM STATEMENT

Which driver will finish in the top position in the Driver’s Championship at the end of the season and which of the new drivers have the potential to become a top F1 driver?

Dataset

Kaggle
Formula 1 World Championship (1950 - 2023)[1]

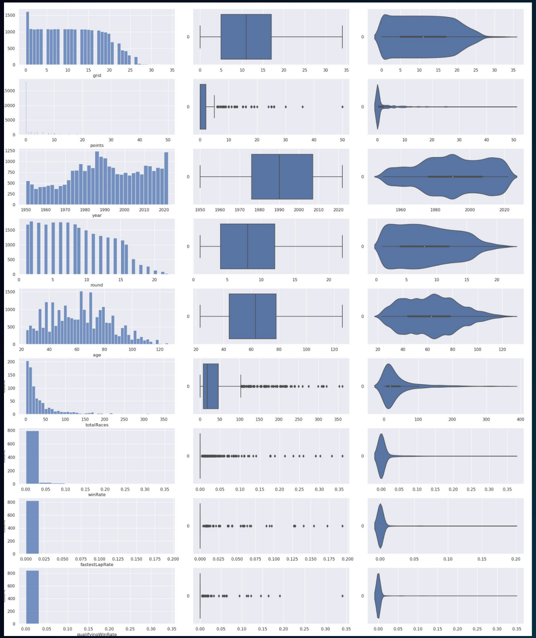
Kept up-to-date, many useful information

Data exploration and cleaning

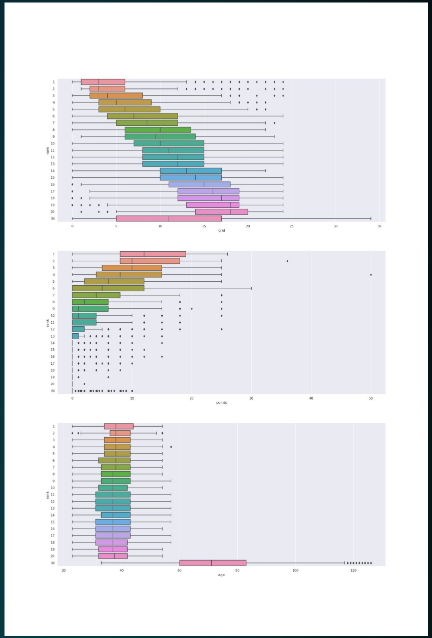
02

Breakdown of Variable

Numerical Variable



Categorical Variable



Feature Engineering

	winRate	fastestLapRate	qualifyingWinRate	age
0	0.331190	0.192926	0.340836	38
1	0.000000	0.010309	0.005155	46
2	0.111650	0.097087	0.145631	38
3	0.088643	0.063712	0.063712	42
4	0.009009	0.018018	0.009009	42

- We create new variables to capture important information in dataset
  - winRate = (number of winning) / (total races)
  - fastestLapRate = (number of fastest laps) / (total laps)
  - qualifyingWinRate = (number of winning qualify) / (total qualify races)

Correlation Matrix

- Plot the correlation matrix for points against all other numeric variables
- Interesting Findings:
    - Points and Age (-0.25)
    - Points and Grid (-0.37)

Core Analysis 03

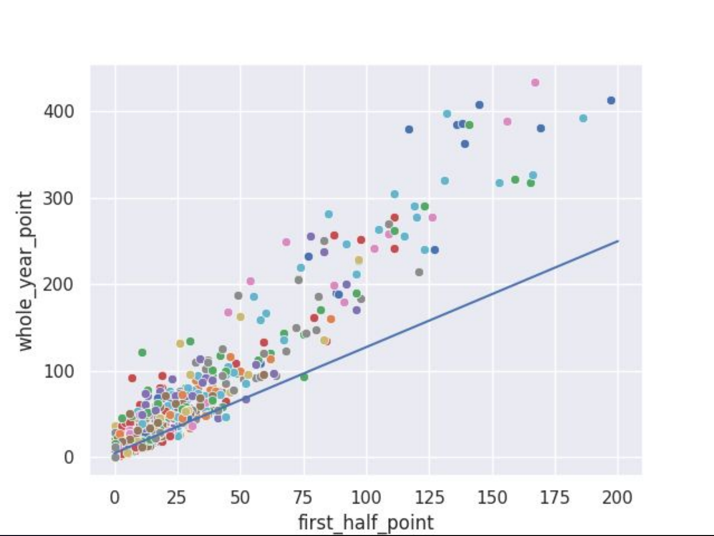
Linear Regression

Train (80%):

Explained Variance (R^2) : 0.79

Test (20%):

Explained Variance (R^2) : 0.87



Polynomial Regression

Train (80%):

Explained Variance (R^2) : 0.81

Test (20%):

Explained Variance (R^2) : 0.85

Polynomials	Form	Degree	Examples
Linear Polynomial	$p(x): ax+b, a \neq 0$	Polynomial with Degree 1	$x + 8$
Quadratic Polynomial	$p(x): ax^2+b+c, a \neq 0$	Polynomial with Degree 2	$3x^2-4x+7$
Cubic Polynomial	$p(x): ax^3+bx^2+cx, a \neq 0$	Polynomial with Degree 3	$2x^3+3x^2+4x+6$

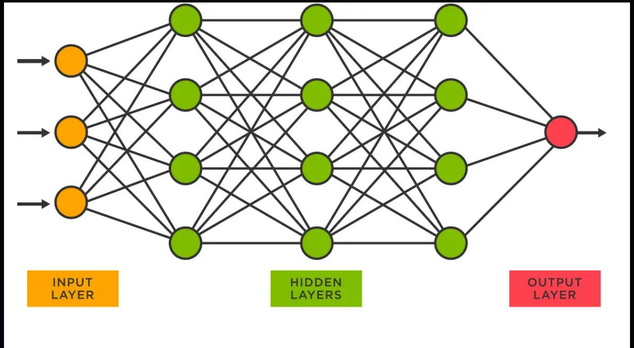
Neural Network

Train (80%):

Accuracy score : 0.86

Test (20%):

Accuracy score : 0.83



Which driver will finish in the top position in the Driver's Championship at the end of the season and **which of the new drivers have the potential to become a top F1 driver**?

Unsupervised Learning

- K-means
- DBSCAN

Lewis Hamilton, Mercedes, 1st position, celebrates on arrival in Parc Ferme
Photo by: Jerry Andre / Motorsport Images

1. Lewis Hamilton - 103 wins

- First race: 2007 Australian Grand Prix
- World Championships: 7 (2008, 2014-15, 2017-20)
- Number of races: 310
- Number of wins: 103
- Number of pole positions: 103
- Career points: 4415.5

	driverId	driverRef	totalRaces	winRate	fastestLapRate	qualifyingWinRate	age
0	1	hamilton	311.0	0.331190	0.192926	0.340836	38
1	2	heidfeld	194.0	0.000000	0.010309	0.005155	46
2	3	rosberg	206.0	0.111650	0.097087	0.145631	38
3	4	alonso	361.0	0.088643	0.063712	0.063712	42
4	5	kovalainen	111.0	0.009009	0.018018	0.009009	42

CLUSTERING PROCESS

“Newer drivers”

Data normalisation

Train model

Plot clusters

Elbow plot

Repeat using DBSCAN

Predictions

Defining "newer drivers"

	driverId	driverRef
845	847	russell
850	852	tsunoda
853	855	zhou
854	856	de_vries
855	857	piastri

CLUSTERING PROCESS

“Newer drivers”

Data normalisation

Train model

Plot clusters

Elbow plot

Repeat using DBSCAN

Predictions

CLUSTERING PROCESS

"Newer drivers"

Data normalisation

Train model

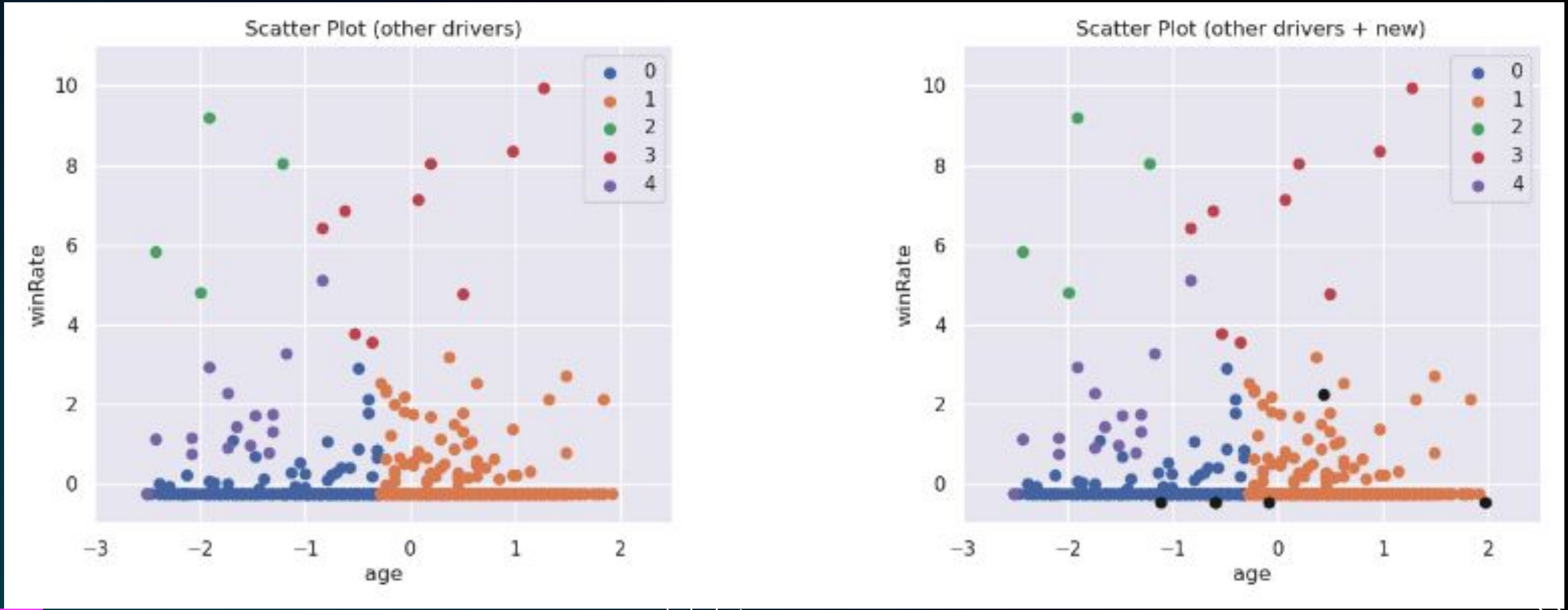
Plot clusters

Elbow plot

Repeat using DBSCAN

Predictions

K-means



Scatter Plot (other drivers)
Scatter Plot (other drivers + new)
winRate
age
0
1
2
3
4

CLUSTERING PROCESS

“Newer drivers”

Data normalisation

Train model

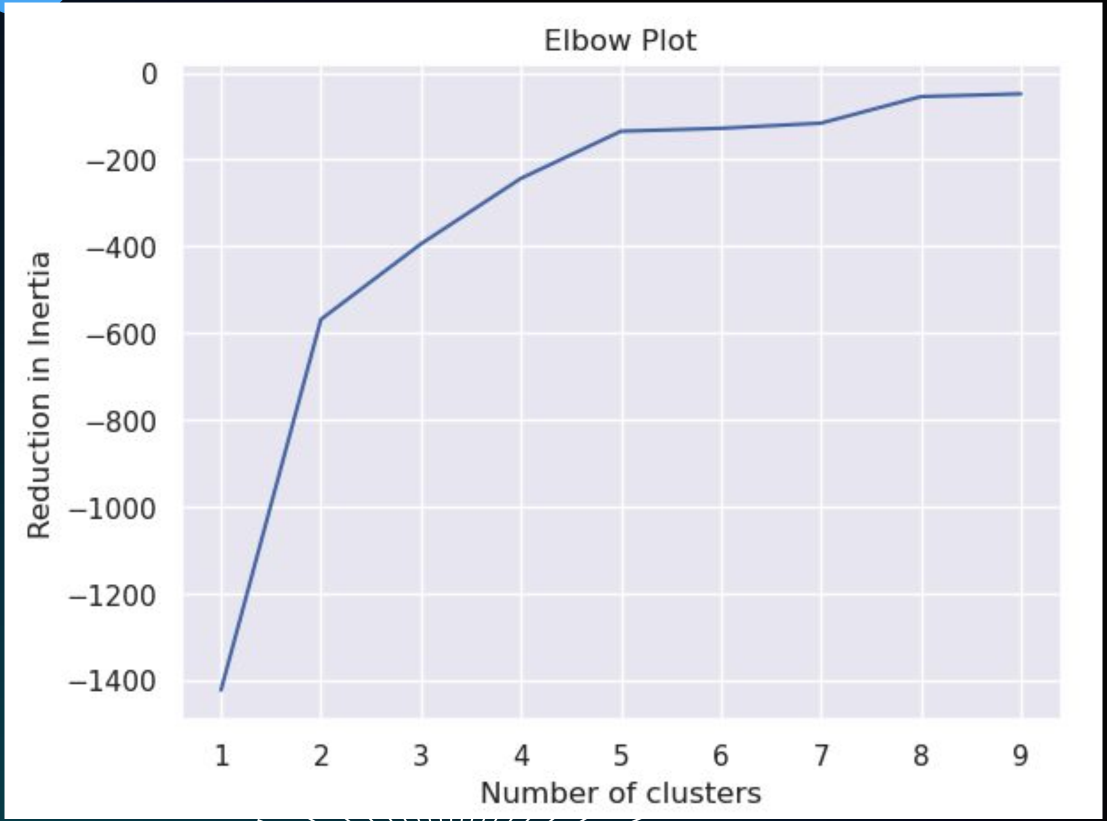
Plot clusters

Elbow plot

Repeat using DBSCAN

Predictions

Elbow plot



Elbow Plot
0
−200
−400
−600
−800
−1000
−1200
−1400
Reduction in Inertia
1
2
3
4
5
6
7
8
9
Number of clusters

CLUSTERING PROCESS

"Newer drivers"

Data normalisation

Train model

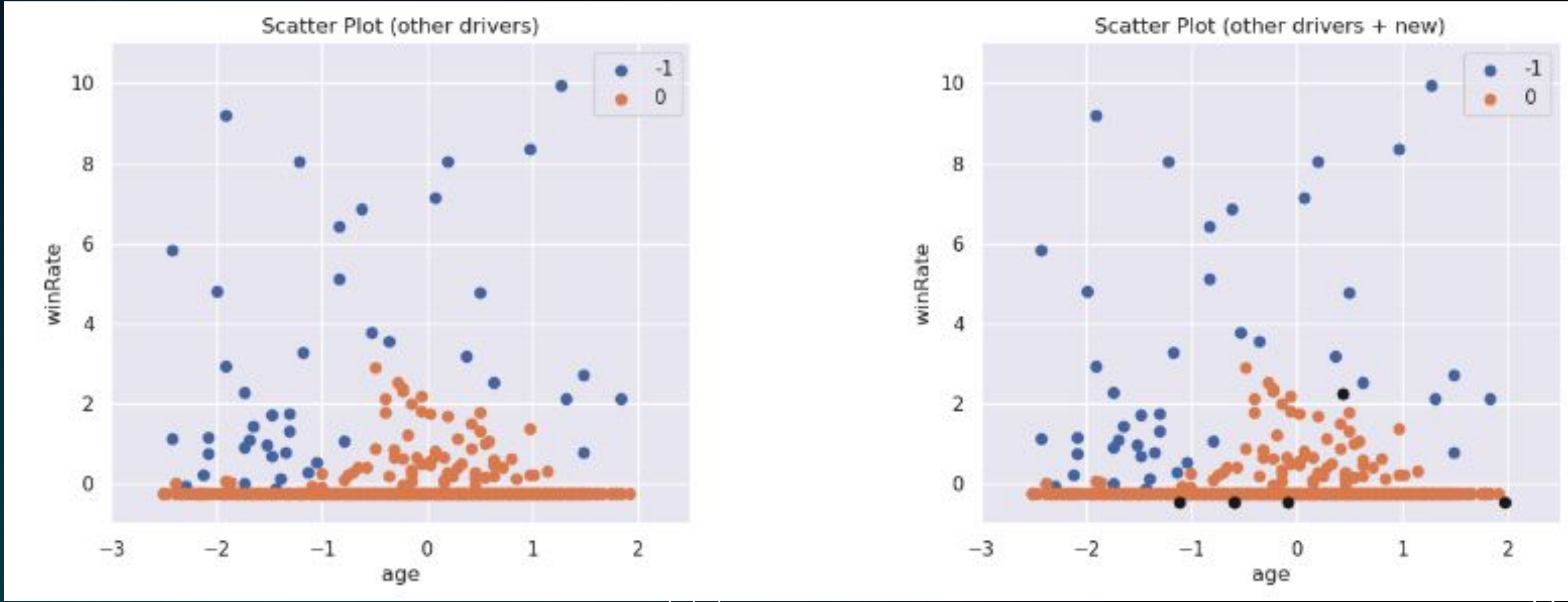
Plot clusters

Elbow plot

Repeat using DBSCAN

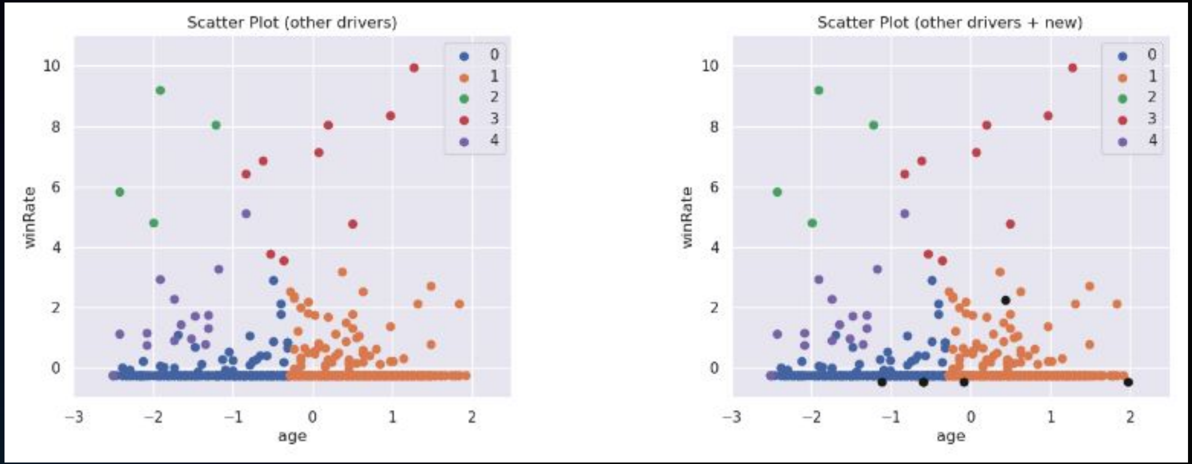
Predictions

DBSCAN

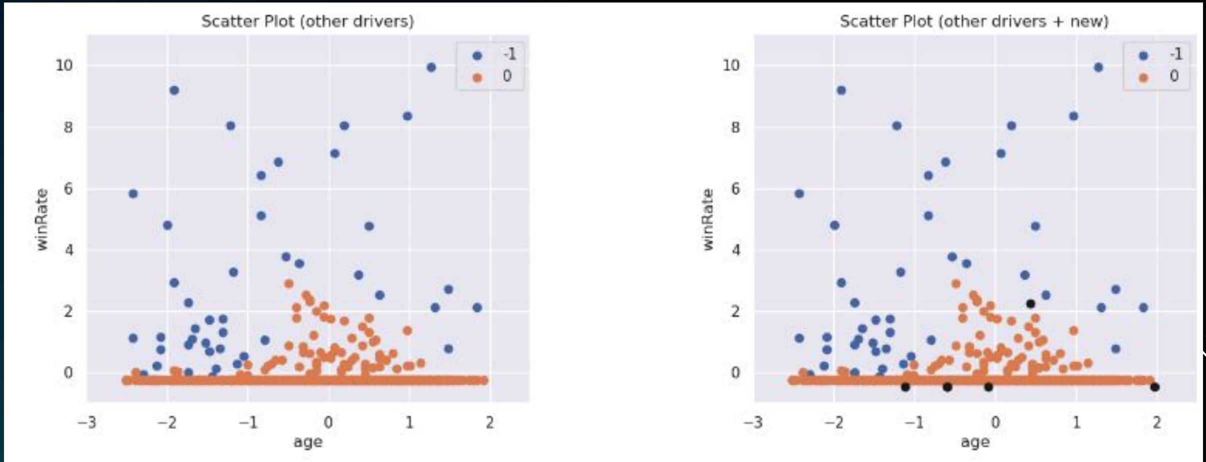


Scatter Plot (other drivers)
Scatter Plot (other drivers + new)
winRate
age
-1
0
10
8
6
4
2
0
−3
−2
−1
0
1
2

K-means



DBSCAN



K-means

driverId	driverRef	totalRaces	winRate	fastestLapRate	qualifyingWinRate	age	cluster

> From here, we can see that none of the newer drivers share the same characteristics as the top drivers as they do not fall into the same clusters as them.

DBSCAN

	driverId	driverRef	totalRaces	winRate	fastestLapRate	qualifyingWinRate	age	cluster	cluster_dbscan
845	847	russell	83.0	0.012048	0.060241	0.012048	25	7	-1
850	852	tsunoda	45.0	0.000000	0.000000	0.000000	23	0	-1
853	855	zhou	23.0	0.000000	0.043478	0.000000	24	0	-1
854	856	de_vries	8.0	0.000000	0.000000	0.000000	28	5	-1
855	857	piastri	1.0	0.000000	0.000000	0.000000	22	6	-1
856	858	sargeant	1.0	0.000000	0.000000	0.000000	23	0	-1

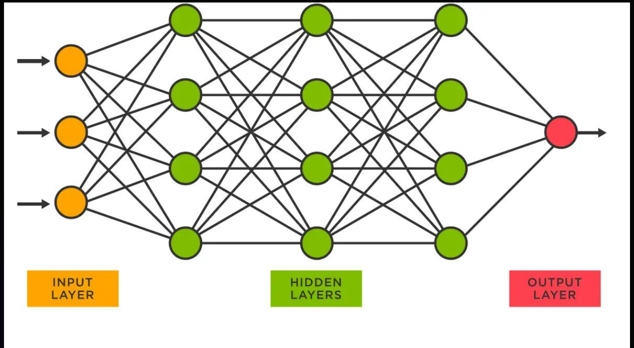
> From here, we can see that according to DBSCAN, all of the newer drivers share the same characteristics as the top drivers.

04

Outcomes and Insights

Supervised Learning

- High accuracy
- Help teams forecast driver performance



Unsupervised Learning

K-means

driverId	driverRef	totalRaces	winRate	fastestLapRate	qualifyingWinRate	age	cluster

> From here, we can see that none of the newer drivers share the same characteristics as the top drivers as they do not fall into the same clusters as them.

DBSCAN

	driverId	driverRef	totalRaces	winRate	fastestLapRate	qualifyingWinRate	age	cluster	cluster_dbscan
845	847	russell	83.0	0.012048	0.060241	0.012048	25	7	-1
850	852	tsunoda	45.0	0.000000	0.000000	0.000000	23	0	-1
853	855	zhou	23.0	0.000000	0.043478	0.000000	24	0	-1
854	856	de_vries	8.0	0.000000	0.000000	0.000000	28	5	-1
855	857	piastri	1.0	0.000000	0.000000	0.000000	22	6	-1
856	858	sargeant	1.0	0.000000	0.000000	0.000000	23	0	-1

> From here, we can see that according to DBSCAN, all of the newer drivers share the same characteristics as the top drivers.

Unsupervised Learning

- Different results
- DBSCAN may not be the most accurate

Data Driven Insights

- Correlation matrix: Points obtained by the driver are highly correlated to the age and grid of the driver.
- DBSCAN: Younger drivers more likely to win more often
- Highlight to teams: focus on the potential of younger drivers and placing well in qualifying races.

	DBSCAN_cluster	driverId	winRate	fastestLapRate	qualifyingWinRate	age
0	-1	324.854167	0.102665	0.044837	0.039205	58.979167
1	0	432.109589	0.002954	0.000149	0.000048	83.845579

Main learning points

- Polynomial regression could result in better fit on the training set, but it risks worse performance on the validation set
- It is very important to prepare data for clustering
- Learnt more about the different types of clustering models and their algorithms

Recommendations

- More complex models can be integrated to improve performance of prediction such as Recurrent Neural Network to better capture time series data.
- More types of clustering algorithms used - hierarchical clustering
- More data fed into clustering algorithm to improve accuracy, when appropriate

Thank You

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**