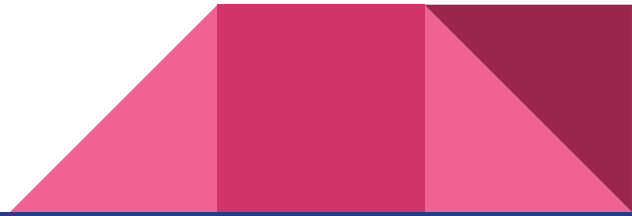
Football Player Recommender System

A General Assembly Capstone Project

Tan Jun Yu, SG-DSI-11

Outline

1) Introduction and Problem Statement
2) Extendability
    a) Business Impact
3) Project Details
    a) Dataset(s)
    b) Cleaning and Preprocessing
    c) Modelling
    d) Feature Engineering
    e) Finetuning and Evaluation
4) Summary of Results
    a) Answering the Problem Statement
5) Learning Points
6) Future Work/Research



Introduction

Rewind to Ballon d'Or 2019..

- An annual awards ceremony that recognises the best-performing footballers for that year
- Lionel Messi, arguably the best player in the world, just won his SIXTH Ballon d'Or

- His speech: *"I'm aware of how old I am, and I enjoy these moments so much because I know that the moment of my retirement is approaching. It's difficult. I know I have some years left. Time flies and everything goes so quickly. I hope to keep enjoying football, my family and everything else in my life.*

Problem Statement

Replacing Lionel Messi - who are the players most similar to him?

- Aim: to build a content-based recommender system to identify the players that are most similar to Lionel Messi.. and beyond!
  - Not just Messi, but this recommender system should be able to recommend similar players for any given player.

Extendability

- **HR/Recruitment Analytics**
  - Lionel Messi = elite top performer/employee
  - Finding similar employees/candidates for a vacant role (or about-to-be-vacant)
  - Hopefully, reduce the effect of politics and bootlicking when it comes to promotions!

Business Impact

According to SHRM (Society for Human Resource Management), the average monetary impact of a typical hiring process was $4,425 in 2017.

Business Impact

- Better fit of candidates -> less time and resources spent on training new person
- Hiring/promoting internally > externally, in terms of costs

Hence, everyone benefits! The company, the hiring department/team, the candidate.

Project Details - The Dataset(s)

- Kaggle: FIFA 2020 game dataset
  - 18278 rows (players) x 104 columns (features)
- Initially planned to merge with Football Manager 2017 dataset..
  - 150k rows (players) x 89 columns (features)

Project Details - Dataset Merging Issues

- Tried merging FIFA17 with FM17 dataset.
- Many issues faced, but the dealbreaker was **dealing with multiple duplicated names**.
  - FM17 had too many duplicated names
  - FM17 had no ‘club’ feature, which can be used to identify and separate duplicated players

Project Details - Dataset Merging Issues

Suppose the players can be identified in the end:

- If merge FIFA17 (18k players) onto FM17 (150k players) on player names:
  - Many FM17 players absent in FIFA17 will have a lot of null values.
  - Can attempt missing data imputation but too tedious
    - Mean/median/mode imputation not correct as quality of player matters.
- Conversely, merging FM17 onto FIFA17 was also too tedious:
  - Just too many duplicated names in FM17.

Ultimately, settle with FIFA20 dataset.

Project Details - Cleaning and Preprocessing

- De-accent names using unidecode library
  - “Mesut Özil” -> “Mesut Ozil”
- Removing “booster scores” using regex

lwb	ldm	cdm	rdm	rwb	lb	lcb	cb	rcb	rb
68+2	66+2	66+2	66+2	68+2	63+2	52+2	52+2	52+2	63+2
65+3	61+3	61+3	61+3	65+3	61+3	53+3	53+3	53+3	61+3
66+3	61+3	61+3	61+3	66+3	61+3	46+3	46+3	46+3	61+3
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
66+3	63+3	63+3	63+3	66+3	61+3	49+3	49+3	49+3	61+3

Project Details - Cleaning and Preprocessing

- **Goalkeepers:**
  1) No ‘gk’ positional score, but goalkeeping attributes/scores present
  2) Null values for all other positional scores

```
['short_name', 'long_name', 'age', 'nationality', 'club', 'overall',
 'potential', 'value_eur', 'wage_eur', 'player_positions',
 'preferred_foot', 'international_reputation', 'weak_foot',
 'skill_moves', 'work_rate', 'body_type', 'real_face',
 'release_clause_eur', 'player_tags', 'team_position',
 'team_jersey_number', 'loaned_from', 'joined', 'contract_valid_until',
 'nation_position', 'nation_jersey_number', 'pace', 'shooting',
 'passing', 'dribbling', 'defending', 'physic', 'gk_diving',
 'gk_handling', 'gk_kicking', 'gk_reflexes', 'gk_speed',
 'gk_positioning', 'player_traits', 'attacking_crossing',
 'attacking_finishing', 'attacking_heading_accuracy',
 'attacking_short_passing', 'attacking_volleys', 'skill_dribbling',
 'skill_curve', 'skill_fk_accuracy', 'skill_long_passing',
 'skill_ball_control', 'movement_acceleration', 'movement_sprint_speed',
 'movement_agility', 'movement_reactions', 'movement_balance',
 'power_shot_power', 'power_jumping', 'power_stamina', 'power_strength',
 'power_long_shots', 'mentality_aggression', 'mentality_interceptions',
 'mentality_positioning', 'mentality_vision', 'mentality_penalties',
 'mentality_composure', 'defending_marking', 'defending_standing_tackle'
```

```
fifa_pos_scores = ['lw', 'lf', 'cf', 'rf', 'rw', 'lam', 'cam', 'ram', 'lm', 'lcm', 'cm',
       'rcm', 'rm', 'lwb', 'ldm', 'cdm', 'rdm', 'rwb', 'lb', 'lcb', 'cb',
       'rcb', 'rb', 'ls', 'st', 'rs']
```

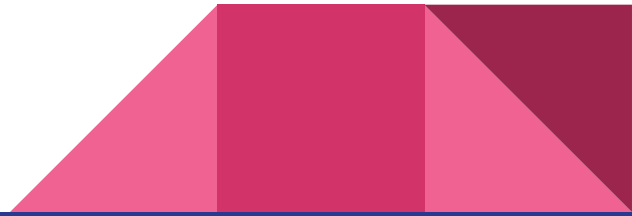
Project Details - Cleaning and Preprocessing

Therefore:

- Feature engineer ‘gk’ score by taking mean of all goalkeeping attributes
- Give a value of ‘0’ for all non-goalkeeping positions

Likewise, for non-goalkeepers:

- Value of ‘0’ for ‘gk’ score



Project Details - Cleaning and Preprocessing

- **Duplicated names**
  - 1700 duplicated short names
- Concatenate ‘short_name’ with ‘club’
  - E.g. ‘L. Messi’ -> ‘L. Messi, FC Barcelona’
  - Still had 16 duplicates

	...	...	...	...	...
15847	Zhang Lu	Zhang Lu	31	China PR	Shanghai Greenland Shenhua FC
15782	Zhang Wei	Zhang Wei	26	China PR	Shanghai SIPG FC
18275	Zhang Wei	Zhang Wei	19	China PR	Hebei China Fortune FC
15845	Zhang Yuan	Yuan Zhang	29	China PR	Shenzhen FC
17918	Zhang Yuan	Zhang Yuan	22	China PR	Tianjin Quanjian FC

1700 rows × 92 columns

Project Details - Cleaning and Preprocessing

- For these remaining 16, concatenate ‘long_name’ with ‘club’

15086	M. Langer, FC Schalke 04	M. Langer	Marcel Langer	22	Germany	FC Schalke 04
10669	M. Langer, FC Schalke 04	M. Langer	Michael Langer	34	Austria	FC Schalke 04
7220	M. Pedersen, Tromsø IL	M. Pedersen	Morten Gamst Pedersen	37	Norway	Tromsø IL
15441	M. Pedersen, Tromsø IL	M. Pedersen	Marcus Holmgren Pedersen	19	Norway	Tromsø IL
17953	S. Ohlsson, IFK Göteborg	S. Ohlsson	Samuel Ohlsson	18	Sweden	IFK Göteborg
10483	S. Ohlsson, IFK Göteborg	S. Ohlsson	Sebastian Ohlsson	26	Sweden	IFK Göteborg
14427	Y. Takahashi, Sagan Tosu	Y. Takahashi	Gao Qiao You Zhi	26	Japan	Sagan Tosu
9852	Y. Takahashi, Sagan Tosu	Y. Takahashi	Gao Qiao Yi Xi	34	Japan	Sagan Tosu

16 rows × 93 columns

Project Details - Modelling

- Quick and Dirty RecSys (RS_basic)
  - Cosine Similarity transformation (from notes) of 72 features
- Results (mehhh):

```
get_sim_players('L. Messi', RS_basic)
```

```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The higher it is, the more similar the player is to L. Messi.
```

	L. Messi, FC Barcelona
name_club	
J. Sildero, Uruguay	0.957726
K. Coman, FC Bayern München	0.948896
A. Januzaj, Real Sociedad	0.943269
P. Dybala, Juventus	0.941714
J. Brandt, Borussia Dortmund	0.937238

Project Details - Feature Engineering

1) Players' Playable Positions
   - One-hot encoding

	name_club	player_positions	team_position	all_positions
0	L. Messi, FC Barcelona	RW, CF, ST	RW	RW, CF, ST, RW
1	Cristiano Ronaldo, Juventus	ST, LW	LW	ST, LW, LW
2	Neymar Jr, Paris Saint-Germain	LW, CAM	CAM	LW, CAM, CAM
3	J. Oblak, Atlético Madrid	GK	GK	GK, GK
4	E. Hazard, Real Madrid	LW, CF	LW	LW, CF, LW

Project Details - Feature Engineering

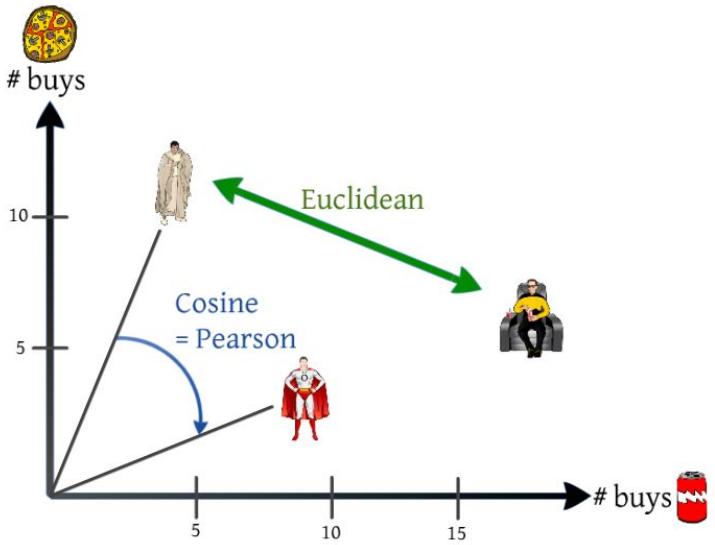
1) Players' Styles/Characteristics
   - Concatenation, regex, followed by one-hot encoding

	name_club	player_tags	player_traits
0	L. Messi, FC Barcelona	#Dribbler, #Distance Shooter, #Crosser, #FK Sp...	Beat Offside Trap, Argues with Officials, Earl...
1	Cristiano Ronaldo, Juventus	#Speedster, #Dribbler, #Distance Shooter, #Acr...	Long Throw-in, Selfish, Argues with Officials,...
2	Neymar Jr, Paris Saint-Germain	#Speedster, #Dribbler, #Playmaker , #Crosser,...	Power Free-Kick, Injury Free, Selfish, Early C...
3	J. Oblak, Atlético Madrid	NaN	Flair, Acrobatic Clearance
4	E. Hazard, Real Madrid	#Speedster, #Dribbler, #Acrobat	Beat Offside Trap, Selfish, Finesse Shot, Spee...

Project Details - Modelling and Finetuning

1) Combine all new features into dataframe
   - 72 features -> 138 features
2) Use **Euclidean Distance**, instead of Cosine Similarity, to transform
   - A player of skill level (10, 10, 10) should NOT be close to another of skill level (100, 100, 100).

   - **Magnitude matters!**



Project Details - Modelling and Finetuning

Results from new RecSys (RS_new):

- Looks better!

```
get_sim_players('L. Messi', RS_new)
```

```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The lower it is, the more similar the player is to L. Messi.
```

	L. Messi, FC Barcelona
name_club	
M. Salah, Liverpool	5.453304
P. Dybala, Juventus	5.602830
K. Mbappe, Paris Saint-Germain	5.744785
A. Januzaj, Real Sociedad	5.934160
D. Mertens, Napoli	5.987330

```
get_sim_players('L. Messi', RS_basic)
```

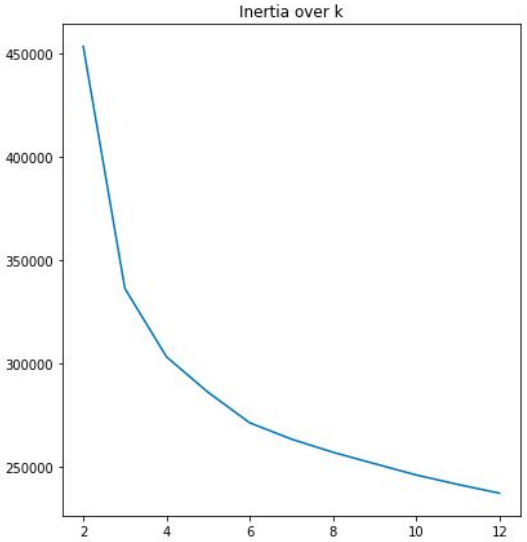
```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The higher it is, the more similar the player is to L. Messi.
```

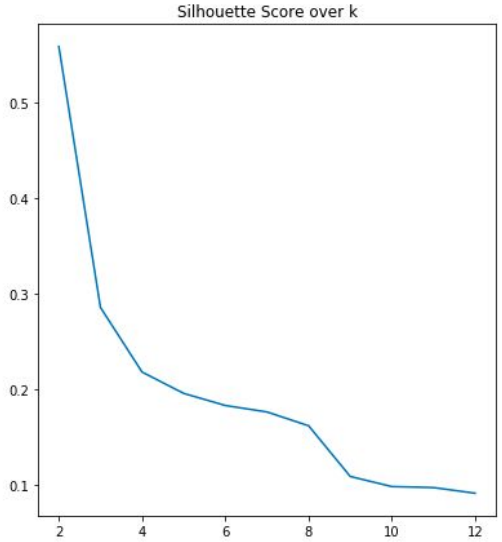
	L. Messi, FC Barcelona
name_club	
J. Sildero, Uruguay	0.957726
K. Coman, FC Bayern München	0.948896
A. Januzaj, Real Sociedad	0.943269
P. Dybala, Juventus	0.941714
J. Brandt, Borussia Dortmund	0.937238

Project Details - Finetuning and Evaluation

- **KMeans Clustering**

BUT curse of high dimensionality (138 features)!



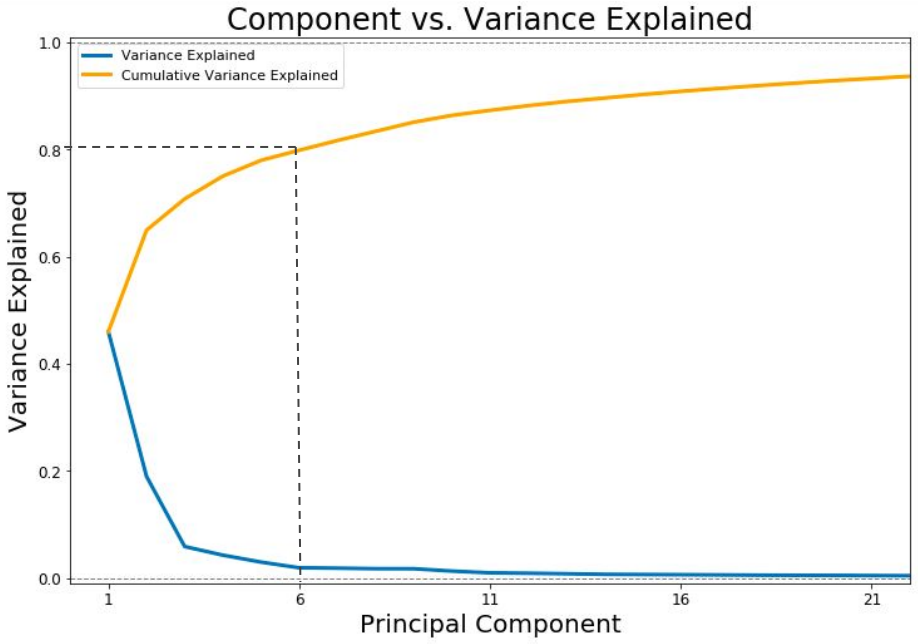


Project Details - Finetuning and Evaluation

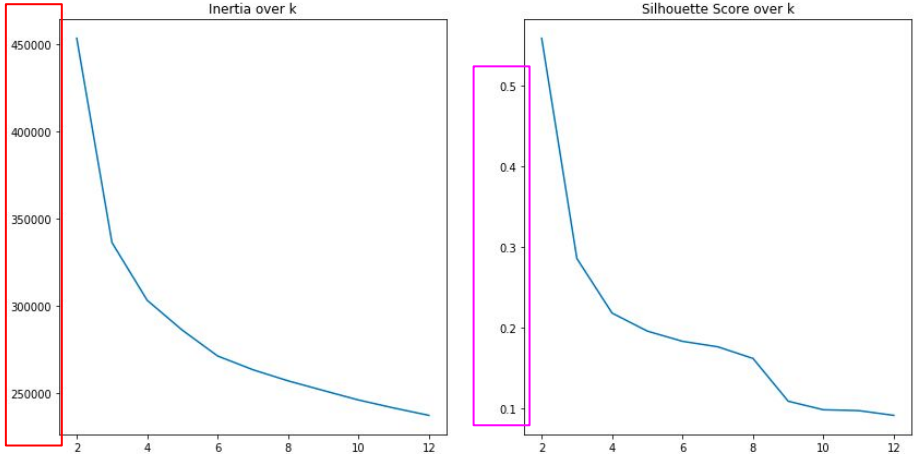
Feature Reduction via

Principal Component Analysis (PCA)

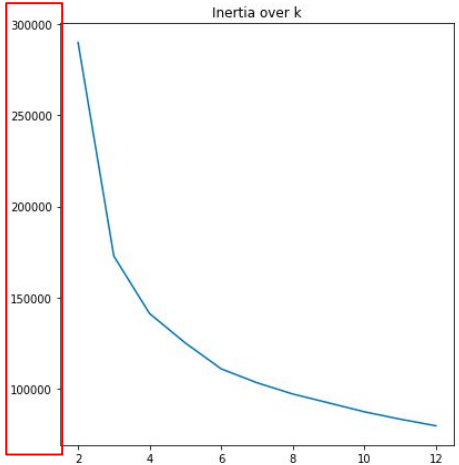
- 6 PCA features chosen to explain 80% of variance

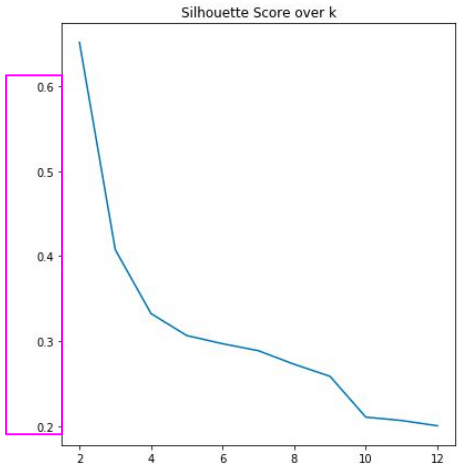


More importantly, for clustering, at every k, inertia and silhouette scores are lower.



- No-PCA clustering





- With-PCA clustering

Project Details - Finetuning and Evaluation

As such, settle on post-PCA DataFrame and transform into RecSys model.

name_club	PCA_feature0	PCA_feature1	PCA_feature2	PCA_feature3	PCA_feature4	PCA_feature5
L. Messi, FC Barcelona	-4.282541	-7.000847	4.181654	-2.690114	3.424324	1.010420
Cristiano Ronaldo, Juventus	-3.478443	-6.530892	4.712754	0.662317	4.401527	0.367672
Neymar Jr, Paris Saint-Germain	-3.599457	-7.437620	2.731434	-2.909759	3.022723	2.329960
J. Oblak, Atlético Madrid	16.187793	-1.246609	2.864576	-1.875956	5.027067	0.708078
E. Hazard, Real Madrid	-3.627067	-6.650310	2.626116	-2.402977	3.052530	0.928096

Project Details - Finetuning and Evaluation

Post-PCA RecSys (RS_PCA) results:

- Much much better!

```
get_sim_players('L. Messi', RS_PCA)
```

```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The lower it is, the more similar the player is to L. Messi.
```

	L. Messi, FC Barcelona
name_club	
E. Hazard, Real Madrid	1.788764
Neymar Jr, Paris Saint-Germain	2.170535
M. Reus, Borussia Dortmund	2.630180
P. Dybala, Juventus	2.695661
D. Mertens, Napoli	2.895664

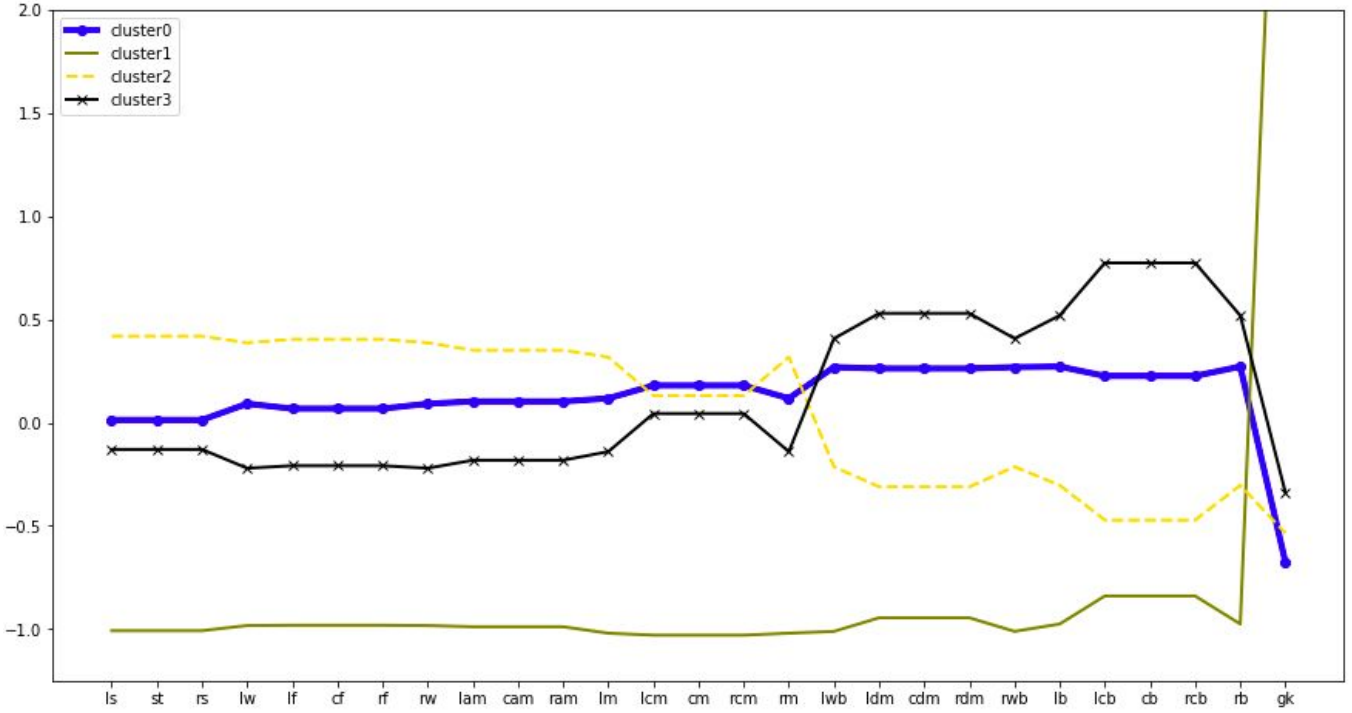
```
get_sim_players('L. Messi', RS_new)
```

```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The lower it is, the more similar the player is to L. Messi.
```

	L. Messi, FC Barcelona
name_club	
M. Salah, Liverpool	5.453304
P. Dybala, Juventus	5.602830
K. Mbappe, Paris Saint-Germain	5.744785
A. Januzaj, Real Sociedad	5.934160
D. Mertens, Napoli	5.987330

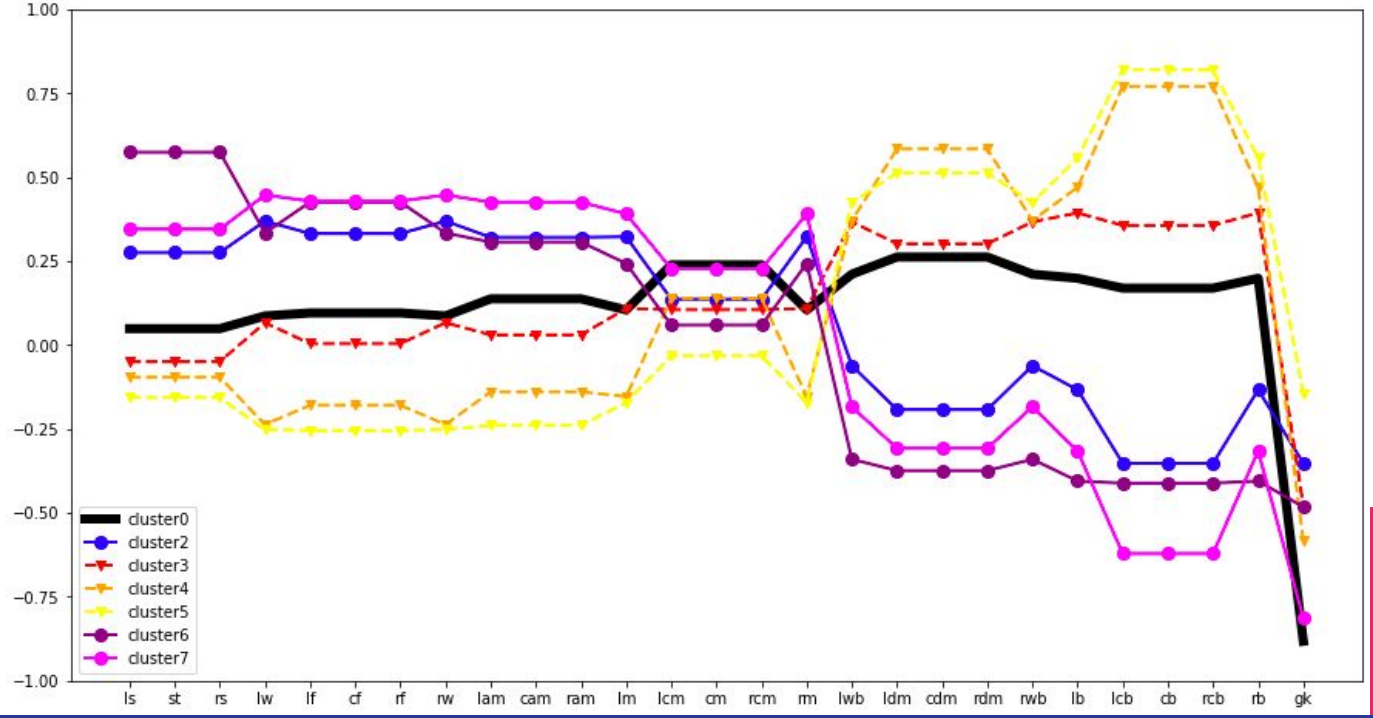
Project Details - Finetuning and Evaluation

K = 4 - simple, but can be better



Project Details - Finetuning and Evaluation

Chosen K = 8



Summary of Results

- Final Model: Post-PCA Dataset of **6 principal components/features**
- Distance metric transformation: **Euclidean distance**
- Chosen clusters of **k = 8; inertia = 97259; silhouette score = 0.27**
- Cluster characteristics:
  - 0: Central midfielders, ‘engines’ of the team
  - 1: Goalkeepers
  - 2: Average-low quality forwards and attacking midfielders
  - 3: Wingbacks
  - 4: High quality and technical central defenders
  - 5: Average-low quality central defenders
  - 6: ‘Pure’ forwards, out-and-out strikers, target men
  - 7: High quality forwards and attacking midfielders

Answering the Problem Statement

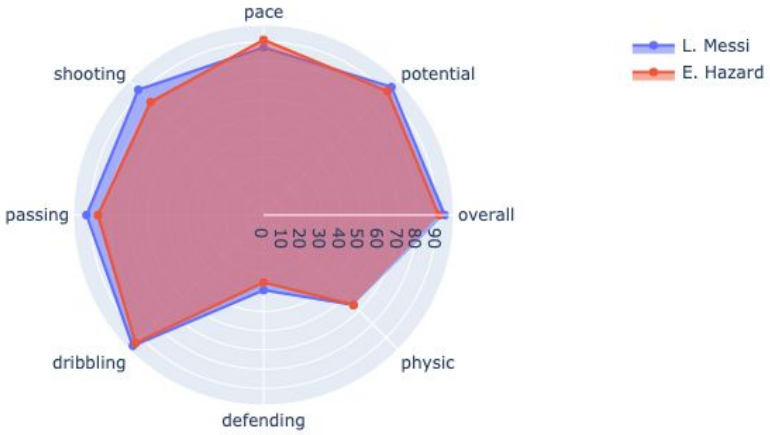
```
get_sim_players('L. Messi', RS_PCA)
```

```
Here are the top 5 players that are similar to L. Messi, FC Barcelona.
The distances below represent the similarity.
The lower it is, the more similar the player is to L. Messi.
```

	L. Messi, FC Barcelona
name_club	
E. Hazard, Real Madrid	1.788764
Neymar Jr, Paris Saint-Germain	2.170535
M. Reus, Borussia Dortmund	2.630180
P. Dybala, Juventus	2.695661
D. Mertens, Napoli	2.895664

Answering the Problem Statement

- Visual Spiderplots for Comparison



Learning Points

- First attempt at **unsupervised learning**
  - Multiple iterations of Similarity Transformation
  - Multiple iterations of clustering
  - Tedious!
- Understanding and making sense of data/results
  - Evaluating unsupervised models is tough
- Learning about all the different distance-based metrics
  - Knowing how and when to apply which metric

Future Work/Research

- Deployment phase (WIP)
- Mapping **synergy** of a potential player (employee) in the team (company/department)
  - E.g. Liverpool FC's midfielders and strikers tend to have a high Attack Work Rate and high Defense Work Rate.
    - Take the mean of the team's Attack/Defense Work Rate and see whether a particular player is above/below that mean
  - In reality (for businesses and companies), might be hard to collect data on such attributes.
- 'Dithering' to introduce some noise into the recommended players
  - Possibly recommend employees from other related teams