USL League One Player Analysis

Report for Capstone Project: Final

In Collaboration with the Oakland Roots Soccer Club
Root Team's Final Report —

1. Executive Summary

Problem

Oakland Roots needs help recruiting talented soccer players to their team. In particular, they are looking to find players with skill sets lacking in the Roots team and at an affordable price.

Solution

With data sourced and scraped from soccer statistics websites, we have provided a comprehensive analysis/plan of attack for Oakland Roots. We delve into player statistics, team dynamics, field layouts, and feature/skillset importance to make better-informed recruitment decisions.

Highlights

Using various analytic techniques, we uncovered resources & statistics that provide the Oakland Roots team strategies to recruit talent. These include 3 main categories (1) *Talent-based recruitments* (based on specific features of desired players), (2) *Similarity-based recruitment* (comparing to other players), and (3) *Team-based recommendations & blindspots* (based on what team dynamics lead to winning championships). More specifically, this breaks down into these loosely-grouped analytic categories:

[1] **Undervalued Players (Market Value Model)**
[2] **Similar Players (Cosine Similarity)**
[3] **Important Features of Valuable Players (Random Forest)**
[4] **Important Features Per *Position***
[5] **Team-by-Team Comparisons**
[6] **Statistics by Features**
[7] **Player Country of Origin (Choropleth)**
[8] **Radar Plots Comparing Most Similar Players**

2. Datasets & Data Retrieval

Databasing

Wyscout (2020 & 2021)

The **Wyscout** dataset is retrieved from the Wyscout data platform as provided by the client (Jordan). Wyscout is the largest soccer data database on the internet. It compares players in the USL League One — ranking players by shots, crosses, & successful tackles in the USL Championship Games. Moreover, they have team statistics — correlation stats of *formation play* & *winning rate of each team* for multiple or singular games. Statistics range from *broad* (team/games level) to *narrow* (players/minute-by-minute level). We acquired 2020 & 2021 datasets on all the players in USL League One — extracting player positions, features, & statistics.

American Soccer Analysis Dataset

The **American Soccer Analysis** (ASA) dataset is downloaded to visualize the data. The dataset contains three separate datasets of *xGoals*, *xPass*, and *Goals Added* (g+). The metric *xGoals* (xG) assesses the probability (%) of any shot scoring (a goal). Specifically, it quantifies the difficulty of a scoring shot using various predictive metrics — *players' goals*, *key passes*, & *assists*. Likewise, the metric *xPass* (xP) assesses the probability (%) of a pass being successful (making it to a teammate). Other columns include — passing percentage (%), number of passes (#), passing distance (m), & player's touch percentage of the ball (%). The Goals Added (g+) dataset measures a player's *total contribution* in *attack* & *defense* during gameplay. Each column in the dataset are factors — dribbling, fouling, passing, & receiving — can affect a team's chance to score or overturn possession.

Webscraping

TransferMarkt Webscrape

Transfermarkt is a website where anyone can look up a soccer team with player statistics, game stats, & a player's market value ($). The market value on Transfermarkt is determined by the *administrator* in charge of each regional league & a *data scout*. Aforesaid data scout considers factors like (1) the *level* of the league the player is in, (2) the player's position, (3) age, & (4) recent trading bids from other teams (higher bids increase market value). The market value is helpful for sports managers of a team to get an idea of a player's skill level. We webscraped every player's statistics—including market value—from the website. This dataset was primarily combined with Wyscout's and ASA datasets (matching by player names) for modeling and visualization purposes. This data was then used in a Moneyball-like attempt to *identify undervalued players* according to their skill level & market value.

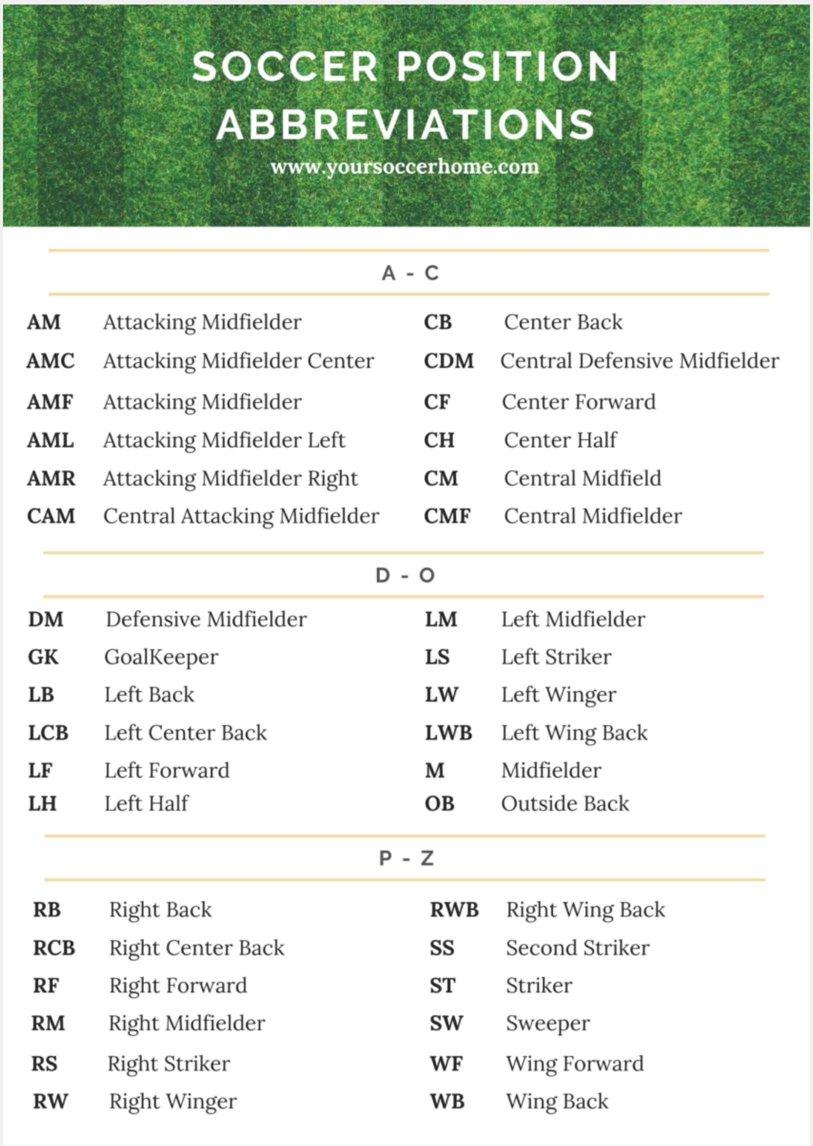
Sofa Score Dataset Webscrape (2020 & 2021)

We found *ratings* of each player in the USL League One on a website called **SofaScore**. This website built an algorithm that quantifies a player's performance by aggregating player statistics into one “rating”. The SofaScore ratings for 2020 and 2021 were scraped from the site. This dataset was merged with Wyscout to construct feature selection, ratings, and a salary model.

3. Data Processing

Position Categorization

Owing to the complexity of the datasets in this project, biases afflict the descriptions of players' positions from different sources. Moreover, stats on positions will also differ due to the opportunities different positions are presented with. Offensive players, for instance, will have more goal-scoring opportunities whilst defensive players rank higher in deep passes. So — categorizing by position is inherently valuable. Nonetheless, it makes sense to *aggregate players in similar positions* so that the categories are more meaningfully predictive when we conduct machine learning.



SOCCER POSITION ABBREVIATIONS

www.yoursoccerhome.com

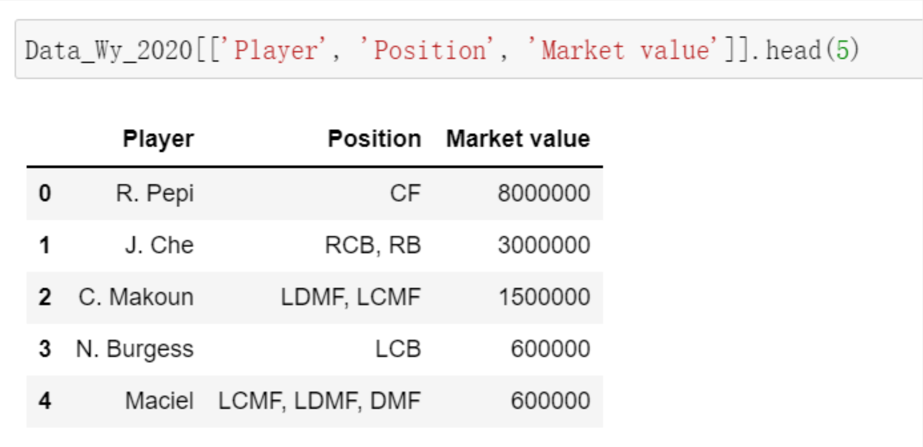
	A - C		
AM	Attacking Midfielder	CB	Center Back
AMC	Attacking Midfielder Center	CDM	Central Defensive Midfielder
AMF	Attacking Midfielder	CF	Center Forward
AML	Attacking Midfielder Left	CH	Center Half
AMR	Attacking Midfielder Right	CM	Central Midfield
CAM	Central Attacking Midfielder	CMF	Central Midfielder
	D - O		
DM	Defensive Midfielder	LM	Left Midfielder
GK	GoalKeeper	LS	Left Striker
LB	Left Back	LW	Left Winger
LCB	Left Center Back	LWB	Left Wing Back
LF	Left Forward	M	Midfielder
LH	Left Half	OB	Outside Back
	P - Z		
RB	Right Back	RWB	Right Wing Back
RCB	Right Center Back	SS	Second Striker
RF	Right Forward	ST	Striker
RM	Right Midfielder	SW	Sweeper
RS	Right Striker	WF	Wing Forward
RW	Right Winger	WB	Wing Back

```
FB = ['RB', 'RWB', 'LB', 'LWB']
CB = ['LCB', 'RCB', 'CH', 'LH', 'RH', 'CB']
CM = ['CDM', 'CM', 'M', 'DMF', 'RDMF', 'LDMF', 'DM', 'RCMF']
AM = ['AMC', 'AMF', 'LAMF', 'RAMF', 'AML', 'AMR', 'CAM', 'AM']
 W = ['WF','LWF', 'RWF', 'RM', 'RW', 'LM', 'LW', 'RS', 'LS']
CF = ['CF', 'LF', 'RF', 'ST', 'SS', 'S', '']
GK = ['GK','G']
```

Shown above are the position aggregations/consolidations we decided on. Some positions have twin labels (GK & G) so these are grouped to avoid redundancy. Rightsided and leftsided variants of the same position are grouped together, as are other variants (WF & WB grouped with W). Center variants of positions received their own groupings *separate* from left- or rightsided counterparts.

Wyscout Data Cleaning

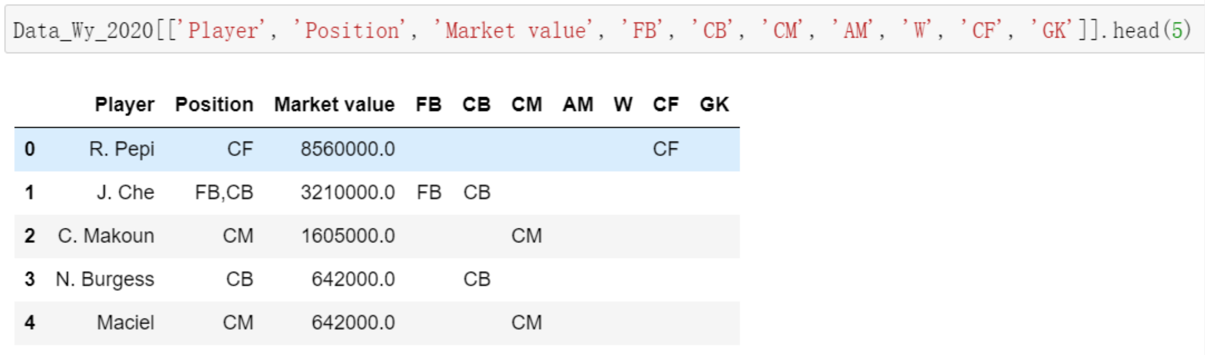
In the source data from Wyscout, the player's *position* is based on the description (often including multiple positions), and the player's *market value* is based on the euro (€).



```
Data_Wy_2020[['Player', 'Position', 'Market value']].head(5)
```

	Player	Position	Market value
0	R. Pepi	CF	8000000
1	J. Che	RCB, RB	3000000
2	C. Makoun	LDMF, LCMF	1500000
3	N. Burgess	LCB	600000
4	Maciel	LCMF, LDMF, DMF	600000

When we cleaned up the data, we first converted the *market value* of the players from **Euros** (€) to **USD** ($). Secondly, we split the original position column by each position and converted each to the description standard we developed (above). We also added 7 new columns to the data according to the current position standard and replaced the boolean value returned according to the player's position with the corresponding position character.



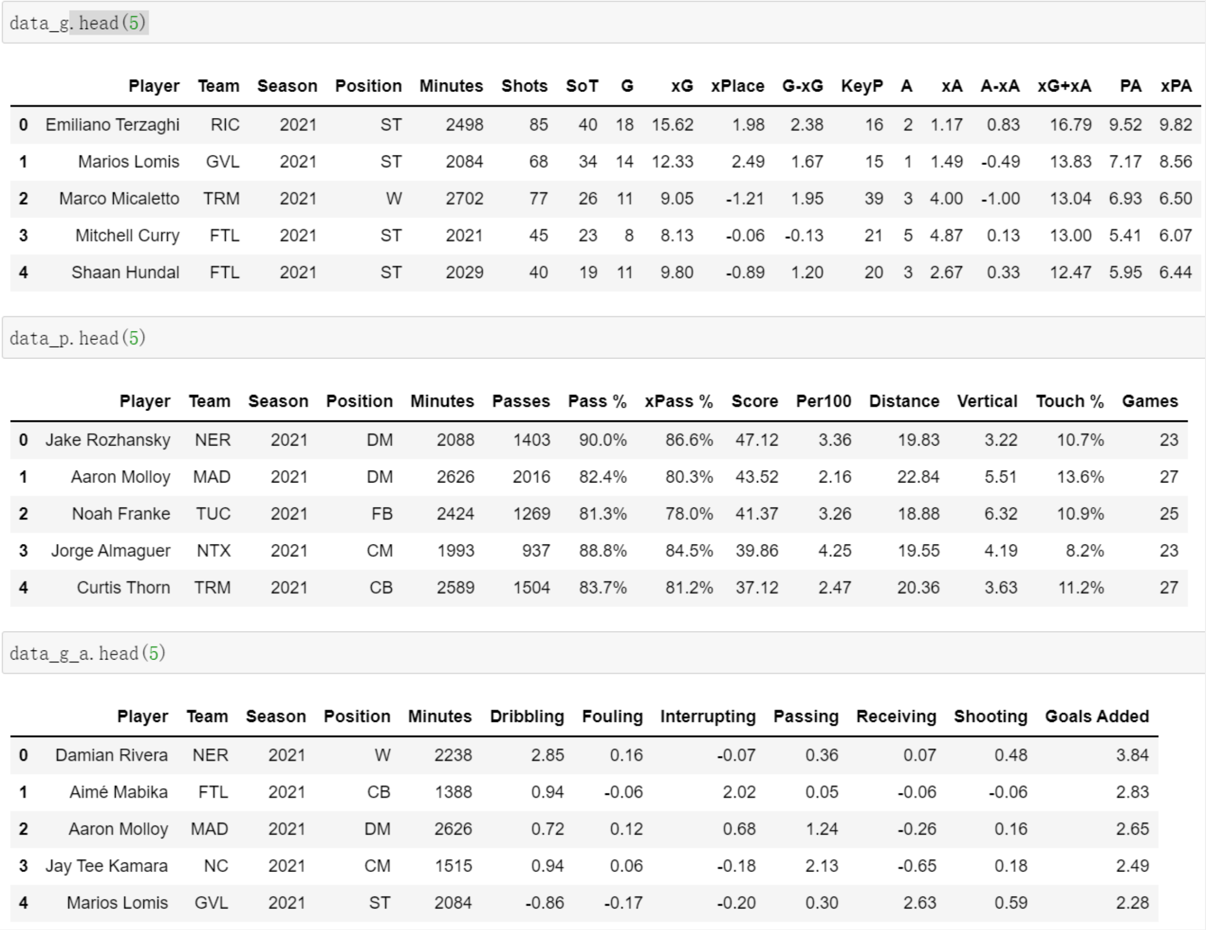
```
Data_Wy_2020[['Player', 'Position', 'Market value', 'FB', 'CB', 'CM', 'AM', 'W', 'CF', 'GK']].head(5)
```

	Player	Position	Market value	FB	CB	CM	AM	W	CF	GK
0	R. Pepi	CF	8560000.0						CF	
1	J. Che	FB,CB	3210000.0	FB	CB					
2	C. Makoun	CM	1605000.0			CM				
3	N. Burgess	CB	642000.0		CB					
4	Maciel	CM	642000.0			CM				

The final data is stored in two CSV files 'datasets/new/Data_Wy_2020_new.csv' and 'datasets/new/Data_Wy_2021_new.csv'.

ASA Dataset

The ASA source data has three CSV files with columns of team player features in the three files, and the player names are arranged in a jumbled/random order.



```
data_g.head(5)
```

	Player	Team	Season	Position	Minutes	Shots	SoT	G	xG	xPlace	G-xG	KeyP	A	xA	A-xA	xG+xA	PA	xPA
0	Emiliano Terzaghi	RIC	2021	ST	2498	85	40	18	15.62	1.98	2.38	16	2	1.17	0.83	16.79	9.52	9.82
1	Marios Lomis	GVL	2021	ST	2084	68	34	14	12.33	2.49	1.67	15	1	1.49	-0.49	13.83	7.17	8.56
2	Marco Micaletto	TRM	2021	W	2702	77	26	11	9.05	-1.21	1.95	39	3	4.00	-1.00	13.04	6.93	6.50
3	Mitchell Curry	FTL	2021	ST	2021	45	23	8	8.13	-0.06	-0.13	21	5	4.87	0.13	13.00	5.41	6.07
4	Shaan Hundal	FTL	2021	ST	2029	40	19	11	9.80	-0.89	1.20	20	3	2.67	0.33	12.47	5.95	6.44

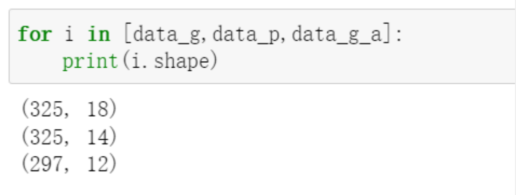
```
data_p.head(5)
```

	Player	Team	Season	Position	Minutes	Passes	Pass %	xPass %	Score	Per100	Distance	Vertical	Touch %	Games
0	Jake Rozhansky	NER	2021	DM	2088	1403	90.0%	86.6%	47.12	3.36	19.83	3.22	10.7%	23
1	Aaron Molloy	MAD	2021	DM	2626	2016	82.4%	80.3%	43.52	2.16	22.84	5.51	13.6%	27
2	Noah Franke	TUC	2021	FB	2424	1269	81.3%	78.0%	41.37	3.26	18.88	6.32	10.9%	25
3	Jorge Almaguer	NTX	2021	CM	1993	937	88.8%	84.5%	39.86	4.25	19.55	4.19	8.2%	23
4	Curtis Thorn	TRM	2021	CB	2589	1504	83.7%	81.2%	37.12	2.47	20.36	3.63	11.2%	27

```
data_g_a.head(5)
```

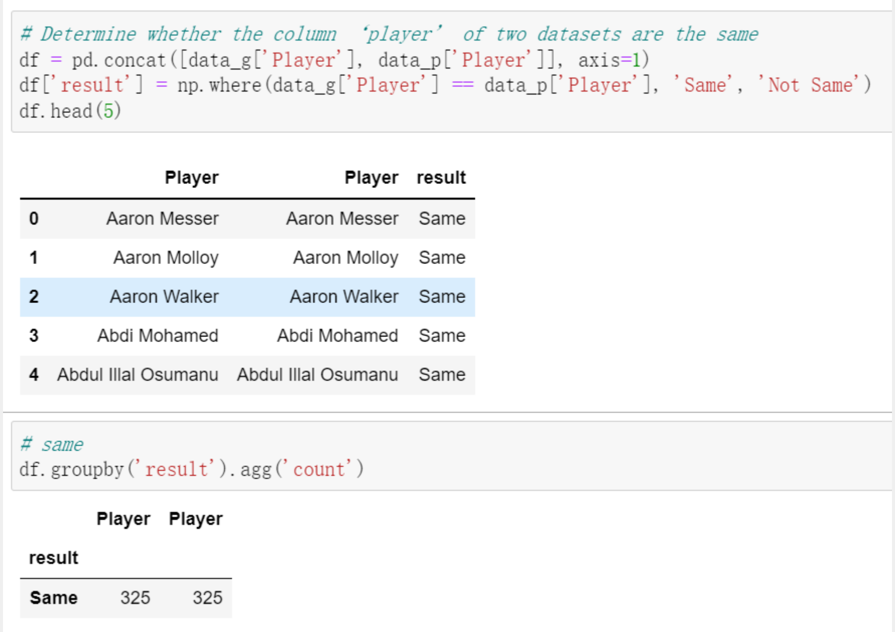
	Player	Team	Season	Position	Minutes	Dribbling	Fouling	Interrupting	Passing	Receiving	Shooting	Goals Added
0	Damian Rivera	NER	2021	W	2238	2.85	0.16	-0.07	0.36	0.07	0.48	3.84
1	Aimé Mabika	FTL	2021	CB	1388	0.94	-0.06	2.02	0.05	-0.06	-0.06	2.83
2	Aaron Molloy	MAD	2021	DM	2626	0.72	0.12	0.68	1.24	-0.26	0.16	2.65
3	Jay Tee Kamara	NC	2021	CM	1515	0.94	0.06	-0.18	2.13	-0.65	0.18	2.49
4	Marios Lomis	GVL	2021	ST	2084	-0.86	-0.17	-0.20	0.30	2.63	0.59	2.28

To successfully merge the three CSV files into one, we first rearranged the data by player initials and reset the indexes. After that, we found that the players in the three data frames were not consistently the same when we reviewed the shape of the three data frames. We first compared the players in the first two data frames and found that the players all matched and were then merged:



```
for i in [data_g, data_p, data_g_a]:
    print(i.shape)
```

```
(325, 18)
(325, 14)
(297, 12)
```



```
# Determine whether the column ‘player’ of two datasets are the same
df = pd.concat([data_g['Player'], data_p['Player']], axis=1)
df['result'] = np.where(data_g['Player'] == data_p['Player'], 'Same', 'Not Same')
df.head(5)
```

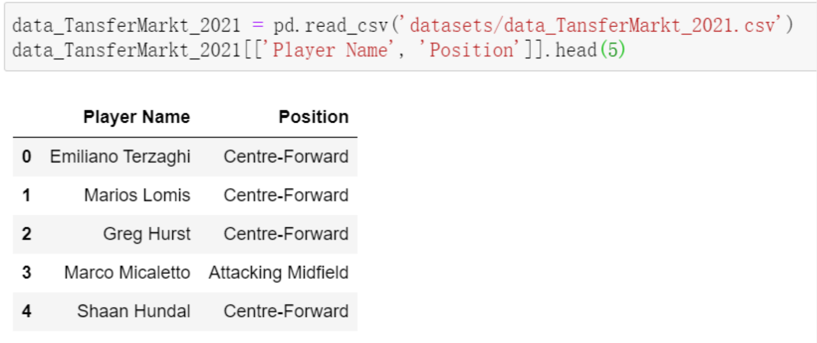
	Player	Player	result
0	Aaron Messer	Aaron Messer	Same
1	Aaron Molloy	Aaron Molloy	Same
2	Aaron Walker	Aaron Walker	Same
3	Abdi Mohamed	Abdi Mohamed	Same
4	Abdul Illal Osumanu	Abdul Illal Osumanu	Same

```
# same
df.groupby('result').agg('count')
```

	Player	Player
result		
Same	325	325

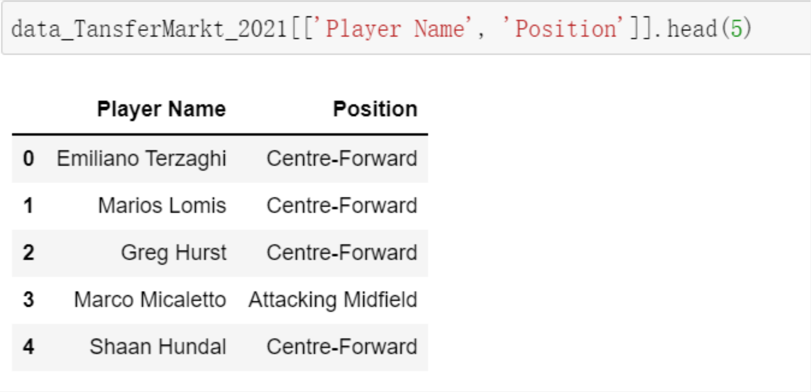
The merged data frame is then compared with the third data frame to find the list of players missing from the third data frame. And merge the data, the missing player data as Null value. The final data is stored as 'datasets/data_ASA.csv'.

TransferMarkt Data



```
data_TansferMarkt_2021 = pd.read_csv('datasets/data_TansferMarkt_2021.csv')
data_TansferMarkt_2021[['Player Name', 'Position']].head(5)
```

	Player Name	Position
0	Emiliano Terzaghi	Centre-Forward
1	Marios Lomis	Centre-Forward
2	Greg Hurst	Centre-Forward
3	Marco Micaletto	Attacking Midfield
4	Shaan Hundal	Centre-Forward



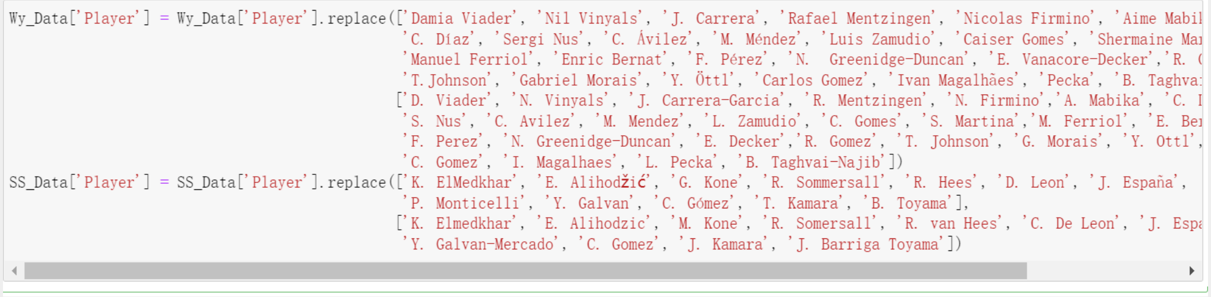
```
data_TansferMarkt_2021[['Player Name', 'Position']].head(5)
```

	Player Name	Position
0	Emiliano Terzaghi	Centre-Forward
1	Marios Lomis	Centre-Forward
2	Greg Hurst	Centre-Forward
3	Marco Micaletto	Attacking Midfield
4	Shaan Hundal	Centre-Forward

TransferMarkt has only one source for this data, but it describes the player's position as a full name and not an abbreviation, again not meeting current standards. We implemented data cleaning by constructing regular expressions to extract the initial capital letters from the position descriptions and overwrite them. The final data is stored in 'datasets/new/data_TansferMarkt_2021_new.csv'.

Data Consolidation (Wyscout Dataset ＆ Sofa Score Dataset)

In order to make the data better serve the subsequent progress, we decided to merge the Rating column of the data from the Wyscout source with the data from the SofaScore source via the Player column. In order to make the data better serve the subsequent progress, we decided to merge the Rating column of the data from the Wyscout source with the data from the SofaScore source via the Player column. However, since the player names of the SofaScore source data are fully spelled, they do not match the Wyscout data. We match by constructing a regular expression that keeps the player's last name and replaces the player's first name with an acronym. Player names with middle names or special characters that are not alphabetic are matched using manual substitution.



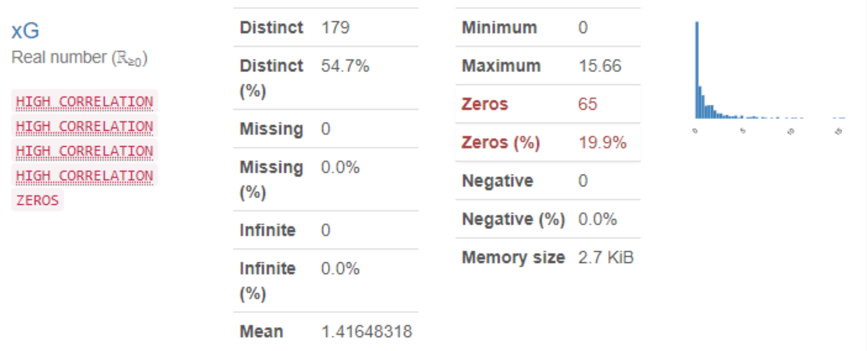
```
Wy_Data['Player'] = Wy_Data['Player'].replace(['Damia Viader', 'Nil Vinyals', 'J. Carrera', 'Rafael Mentzingen', 'Nicolas Firmino', 'Aime Mabil
                                               'C. Díaz', 'Sergi Nus', 'C. Ávilez', 'M. Méndez', 'Luis Zamudio', 'Caiser Gomes', 'Shermaine Mar
                                               'Manuel Ferriol', 'Enric Bernat', 'F. Pérez', 'N.  Greenidge-Duncan', 'E. Vanacore-Decker','R. (
                                               'T.Johnson', 'Gabriel Morais', 'Y. Öttl', 'Carlos Gomez', 'Ivan Magalhães', 'Pecka', 'B. Taghva
                                              ['D. Viader', 'N. Vinyals', 'J. Carrera-Garcia', 'R. Mentzingen', 'N. Firmino','A. Mabika', 'C. I
                                               'S. Nus', 'C. Avilez', 'M. Mendez', 'L. Zamudio', 'C. Gomes', 'S. Martina','M. Ferriol', 'E. Be
                                               'F. Perez', 'N. Greenidge-Duncan', 'E. Decker','R. Gomez', 'T. Johnson', 'G. Morais', 'Y. Ottl',
                                               'C. Gomez', 'I. Magalhaes', 'L. Pecka', 'B. Taghvai-Najib'])
SS_Data['Player'] = SS_Data['Player'].replace(['K. ElMedkhar', 'E. Alihodžić', 'G. Kone', 'R. Sommersall', 'R. Hees', 'D. Leon', 'J. España',
                                               'P. Monticelli', 'Y. Galvan', 'C. Gómez', 'T. Kamara', 'B. Toyama'],
                                              ['K. Elmedkhar', 'E. Alihodzic', 'M. Kone', 'R. Somersall', 'R. van Hees', 'C. De Leon', 'J. Espa
                                               'Y. Galvan-Mercado', 'C. Gomez', 'J. Kamara', 'J. Barriga Toyama'])
```

The final data is saved in two CSV files 'datasets/data_WyScout_Rating_2021.csv' and 'datasets/data_WyScout_Rating_2020.csv'.

5. Data analysis

Wyscout 2021 EDA

xG



xG
Real number ($\mathbb{R}_{\geq 0}$)

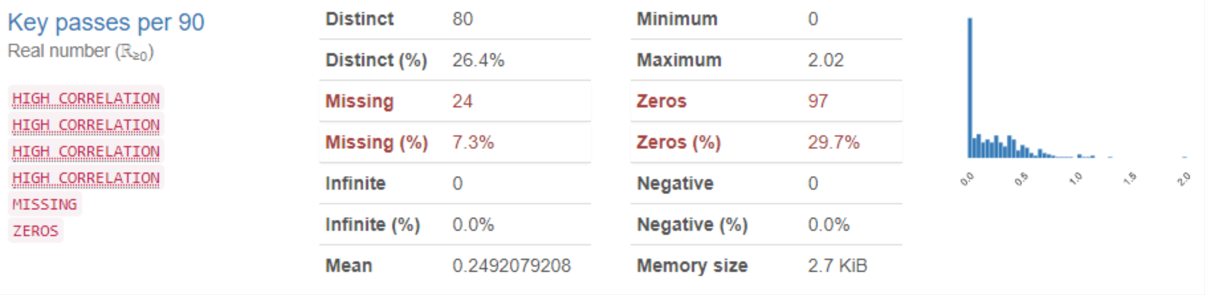
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	179
Distinct (%)	54.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1.41648318

Minimum	0
Maximum	15.66
Zeros	65
Zeros (%)	19.9%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

xG is the metric of the probability that a shot will result in a goal based on the characteristics and the previous actions of that shot. The average score is xG in Wyscout 2021 dataset is 1.4. We would expect the average player in the league to score at least 1 goal in all their games. The maximum xG is 18, so we would expect that player to achieve nearly an average of 18 goals in the 2021 season.

xA



Key passes per 90

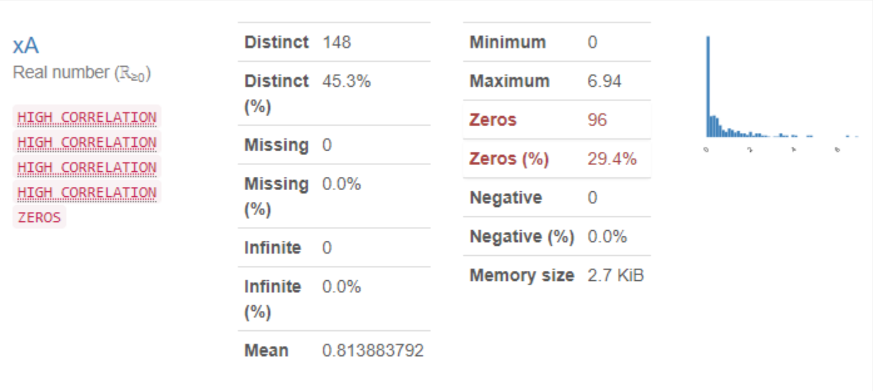
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	80	Minimum	0
Distinct (%)	26.4%	Maximum	2.02
Missing	24	Zeros	97
Missing (%)	7.3%	Zeros (%)	29.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.2492079208	Memory size	2.7 KiB

xA is the value of a Shot that's assisted by this pass, and this metric is essential for all forward positions. The average of xA is 0.81, which means players can expect to have one assist in a pass to score a goal. The maximum assist is 6.94. This player represents an effective forward position player that can pass a critical on-ball action to contribute to a goal.

Keyasses



xA

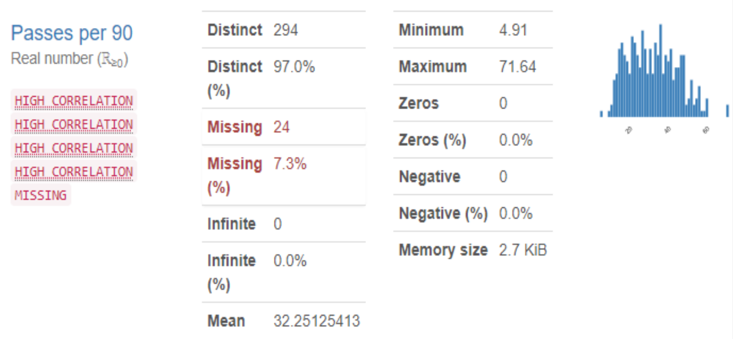
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	148	Minimum	0
Distinct (%)	45.3%	Maximum	6.94
Missing	0	Zeros	96
Missing (%)	0.0%	Zeros (%)	29.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.813883792	Memory size	2.7 KiB

A **keypass** is a pass that immediately creates a clear goal-scoring opportunity for a teammate. Keypasses for players in a 90 minutes game are relatively low with an average of 0.2 number of passes per 90 minutes. The maximum of key passes for a player is 2.02.

Passes Per 90



Passes per 90

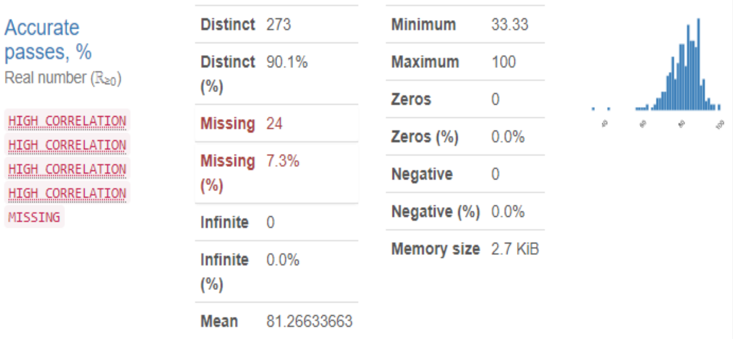
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	294	Minimum	4.91
Distinct (%)	97.0%	Maximum	71.64
Missing	24	Zeros	0
Missing (%)	7.3%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	32.25125413	Memory size	2.7 KiB

Passes Per 90 is the total passes in a 90-minute game for all games in the 2021 season is 294 passes. The players pass the ball 32 times per game, with a minimum pass of 4 and maximum passes of 71. Having a higher volume of passes maximizes the chance of scoring.

Accurate Passes (%)



Accurate passes, %
Real number (ℝ≥0)

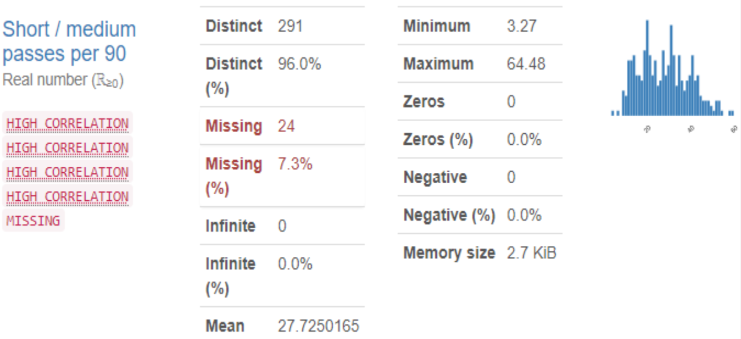
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	273
Distinct (%)	90.1%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	81.26633663

Minimum	33.33
Maximum	100
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

The **passing accuracy** is calculated by the actual number of passes / completed passes. With the accurate passes percentage, the average of accurate passes is 81%, which indicates that players' passes are 81% successful during the games.

Short/Medium Passes Per 90



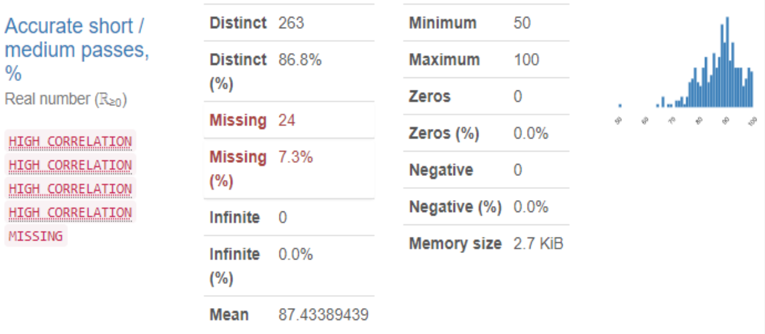
Short / medium passes per 90
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	291
Distinct (%)	96.0%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	27.7250165

Minimum	3.27
Maximum	64.48
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

Accurate Short/Medium Passes (%)



Accurate short / medium passes, %
Real number (ℝ≥0)

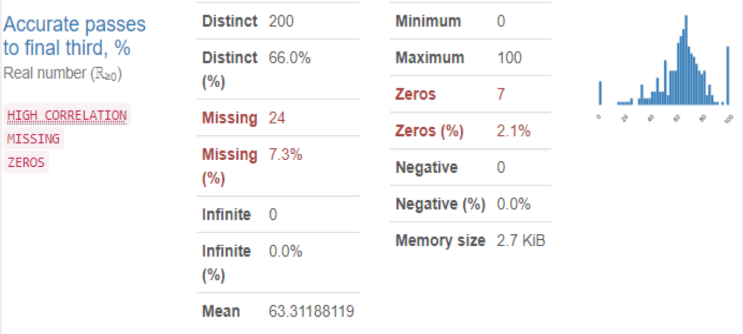
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	263
Distinct (%)	86.8%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	87.43389439

Minimum	50
Maximum	100
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

The Oakland Roots is a passing team; our team can focus on finding players with high passing numbers and percentages. Short/medium passes are a pass of fewer than 40 meters long. Players average about 27 short/medium passes per 90, and the maximum passes of one player are 64 times. The average percentage of accurate short/medium passes is 87%, with a minimum of 50% accurate passes from all the players.

Accurate passes to final third (%)

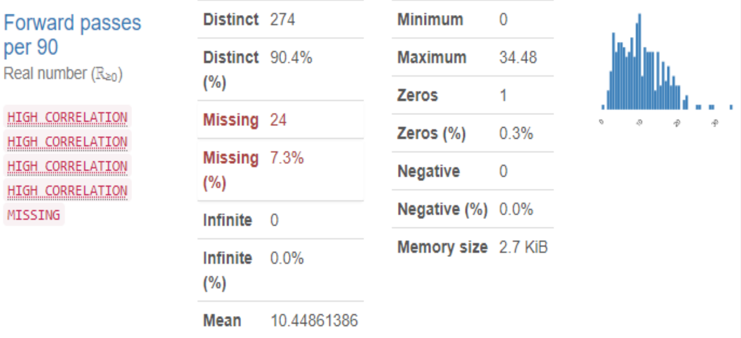


Accurate passes to final third, %
Real number (ℝ≥0)

HIGH CORRELATION
MISSING
ZEROS

Distinct	200	Minimum	0
Distinct (%)	66.0%	Maximum	100
Missing	24	Zeros	7
Missing (%)	7.3%	Zeros (%)	2.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	63.31188119	Memory size	2.7 KiB

Forward passes per 90



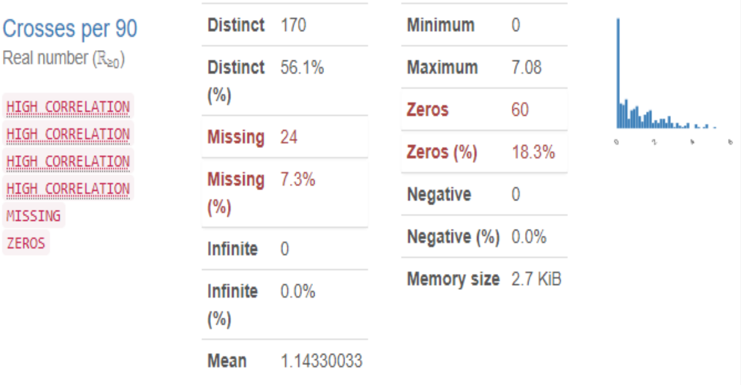
Forward passes per 90
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	274	Minimum	0
Distinct (%)	90.4%	Maximum	34.48
Missing	24	Zeros	1
Missing (%)	7.3%	Zeros (%)	0.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	10.44861386	Memory size	2.7 KiB

These two metrics are essential for all defensive positions. Accurate passes into the final third can create key passes to score a goal. The average percentage of successful passes into the final third is 63%. There are about ten forward passes from players passed forward per 90.

Crosses per 90

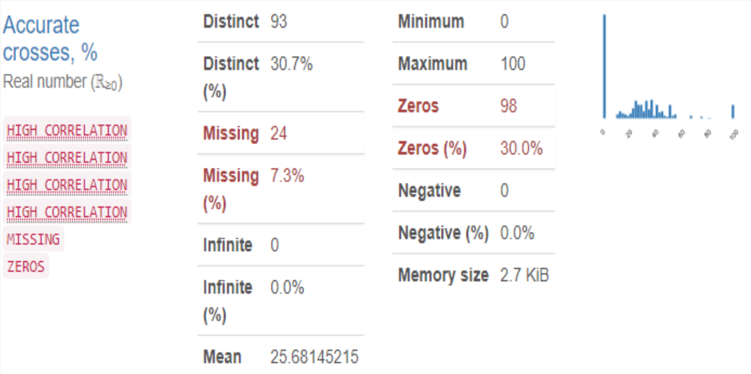


Crosses per 90
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	170	Minimum	0
Distinct (%)	56.1%	Maximum	7.08
Missing	24	Zeros	60
Missing (%)	7.3%	Zeros (%)	18.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1.14330033	Memory size	2.7 KiB

Accurate crosses (%)



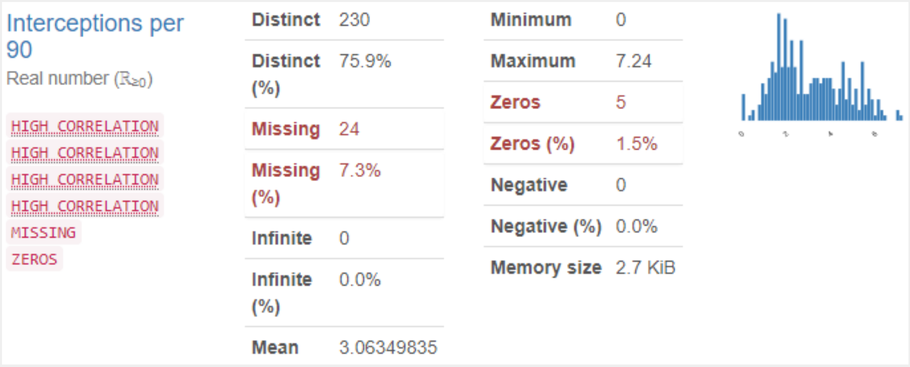
Accurate crosses, %
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	93	Minimum	0
Distinct (%)	30.7%	Maximum	100
Missing	24	Zeros	98
Missing (%)	7.3%	Zeros (%)	30.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	25.68145215	Memory size	2.7 KiB

A cross is a ball played from the offensive flanks aimed towards a teammate in the area in front of the opponent's goal. These statistics are important for outside back players, wingers, and forward players. The average number of crosses per 90 is 1.14, with a maximum of 7, which means at least one cross attempt will happen per 90 minutes. The average of accurate crosses percent is 25 %, saying the average of successful crosses is 25%.

Interception per 90



Interceptions per 90
Real number ($\mathbb{R}_{\geq 0}$)

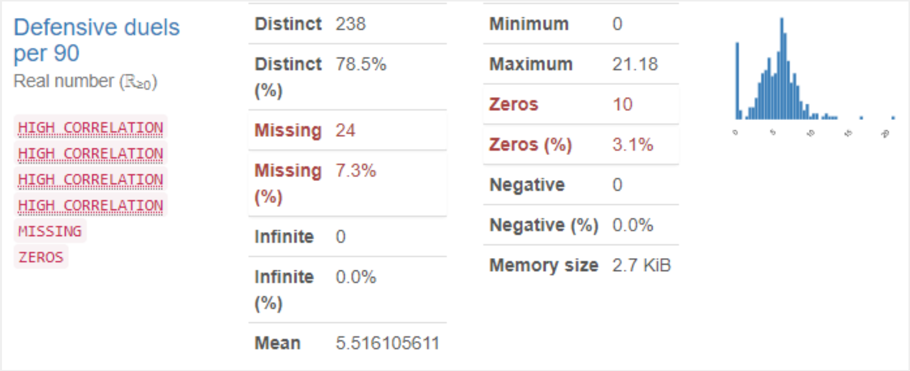
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	230
Distinct (%)	75.9%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	3.06349835

Minimum	0
Maximum	7.24
Zeros	5
Zeros (%)	1.5%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

An interception is an act of a player actively intercepting the ball by anticipating its movement when the opponent is shooting, passing, or crossing. The average interception per 90 is three times with a maximum interception of 7 times per 90.

Defensive Duels Per 90, Defensive Duels won (%)



Defensive duels per 90
Real number ($\mathbb{R}_{\geq 0}$)

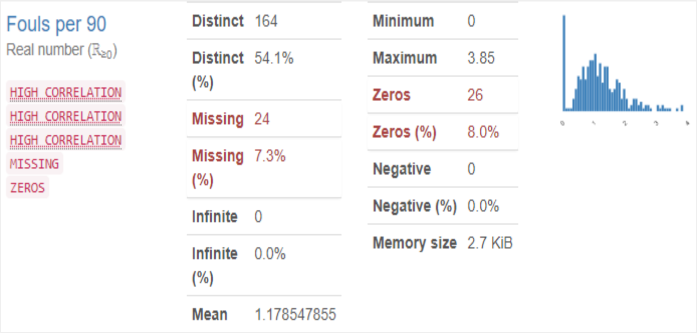
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	238
Distinct (%)	78.5%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	5.516105611

Minimum	0
Maximum	21.18
Zeros	10
Zeros (%)	3.1%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

A defensive duel is when a player attempts to dispossess an opposition player to stop an attack from progressing. This metric is critical for midfielders to prevent other players from passing the ball to forward positions. Players have a mean of 5.51, which means they attempt to duel for the ball in the defensive position at least 5.51 times per game. The average percentage of defensive duels won is 56%.

Fouls per 90



Fouls per 90

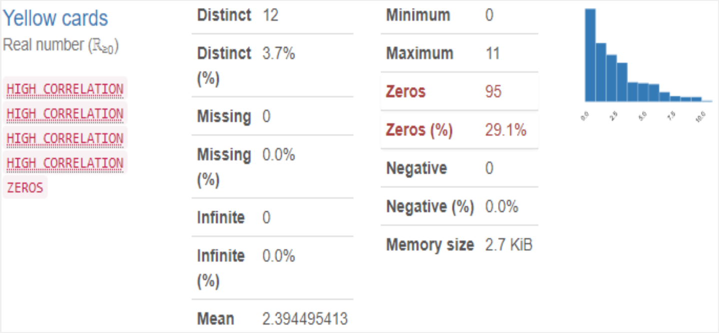
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING
ZEROS

Distinct	164
Distinct (%)	54.1%
Missing	24
Missing (%)	7.3%
Infinite	0
Infinite (%)	0.0%
Mean	1.178547855

Minimum	0
Maximum	3.85
Zeros	26
Zeros (%)	8.0%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

Yellow Cards



Yellow cards

Real number ($\mathbb{R}_{\geq 0}$)

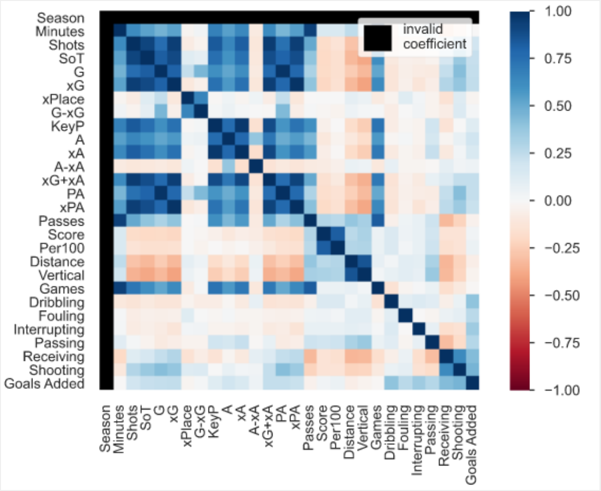
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	12
Distinct (%)	3.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.394495413

Minimum	0
Maximum	11
Zeros	95
Zeros (%)	29.1%
Negative	0
Negative (%)	0.0%
Memory size	2.7 KiB

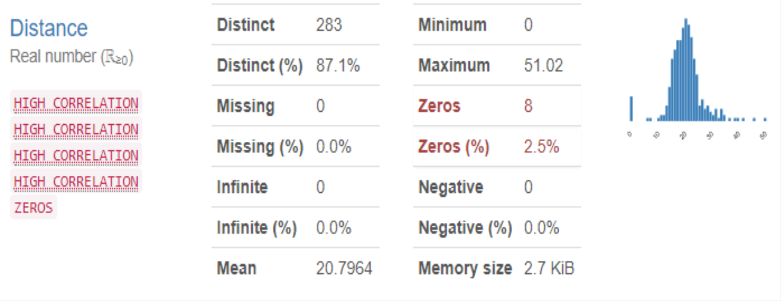
Players that have a lower number of fouls and yellow cards mean players will be more available on the field. Players foul at least once per game, and players may receive two yellow cards per season.

ASA Dataset



The ASA dataset looks relatively clean but with few missing values in dribbling, fouling, interrupting, passing, receiving, and shooting columns. These columns require precise measurement and recording of professional types of equipment. The American Soccer Analysis only retrieves player statistics data from other resources to calculate the xGoals, Xpass, and Goals Added. The columns that don't have null values have high correlations.

Distance



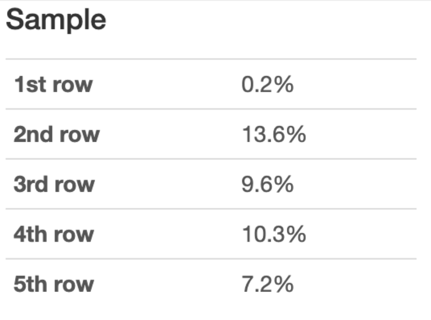
Distance
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	283	Minimum	0
Distinct (%)	87.1%	Maximum	51.02
Missing	0	Zeros	8
Missing (%)	0.0%	Zeros (%)	2.5%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	20.7964	Memory size	2.7 KiB

Distance is the average distance of a pass, and the center players should have a short pass. The average distance of players on the field is 20, and the maximum distance is 51.

Touch%

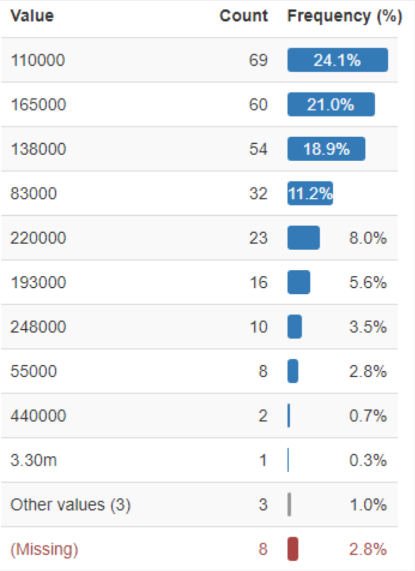


Sample

1st row	0.2%
2nd row	13.6%
3rd row	9.6%
4th row	10.3%
5th row	7.2%

A touch or missed touch of the ball is when the player is not doing a Pass or clearly identifiable action. For fullback, attacking, and central midfielders, touch % is crucial for the touch percentage being in the double digits matters.

Transfermarkt Dataset



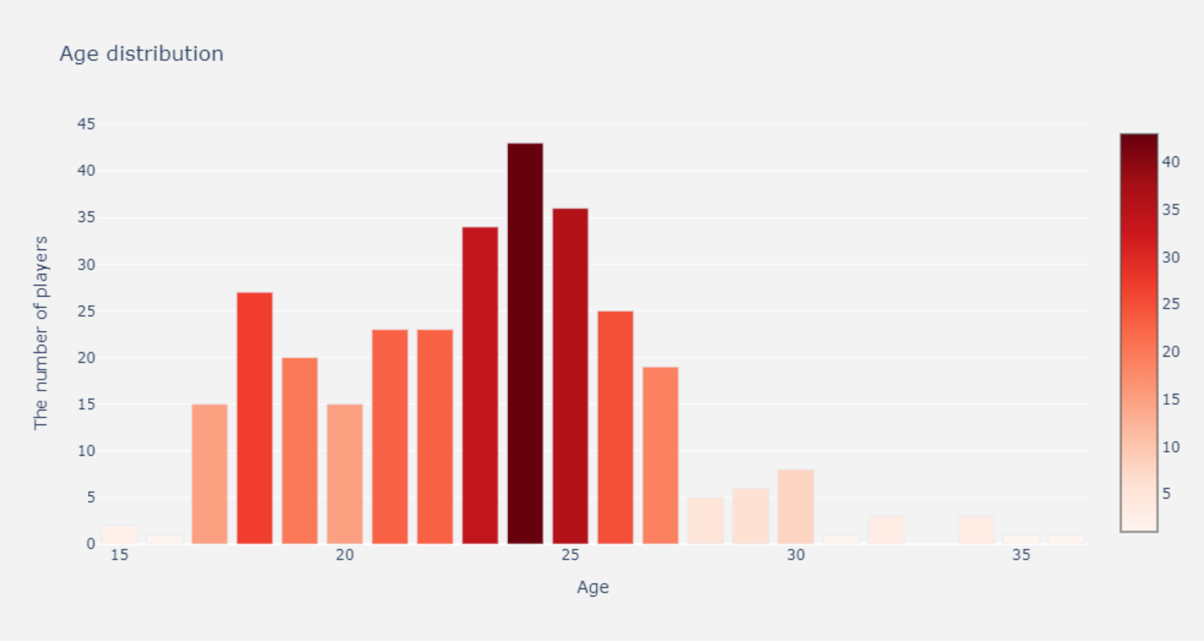
Value	Count	Frequency (%)
110000	69	24.1%
165000	60	21.0%
138000	54	18.9%
83000	32	11.2%
220000	23	8.0%
193000	16	5.6%
248000	10	3.5%
55000	8	2.8%
440000	2	0.7%
3.30m	1	0.3%
Other values (3)	3	1.0%
(Missing)	8	2.8%

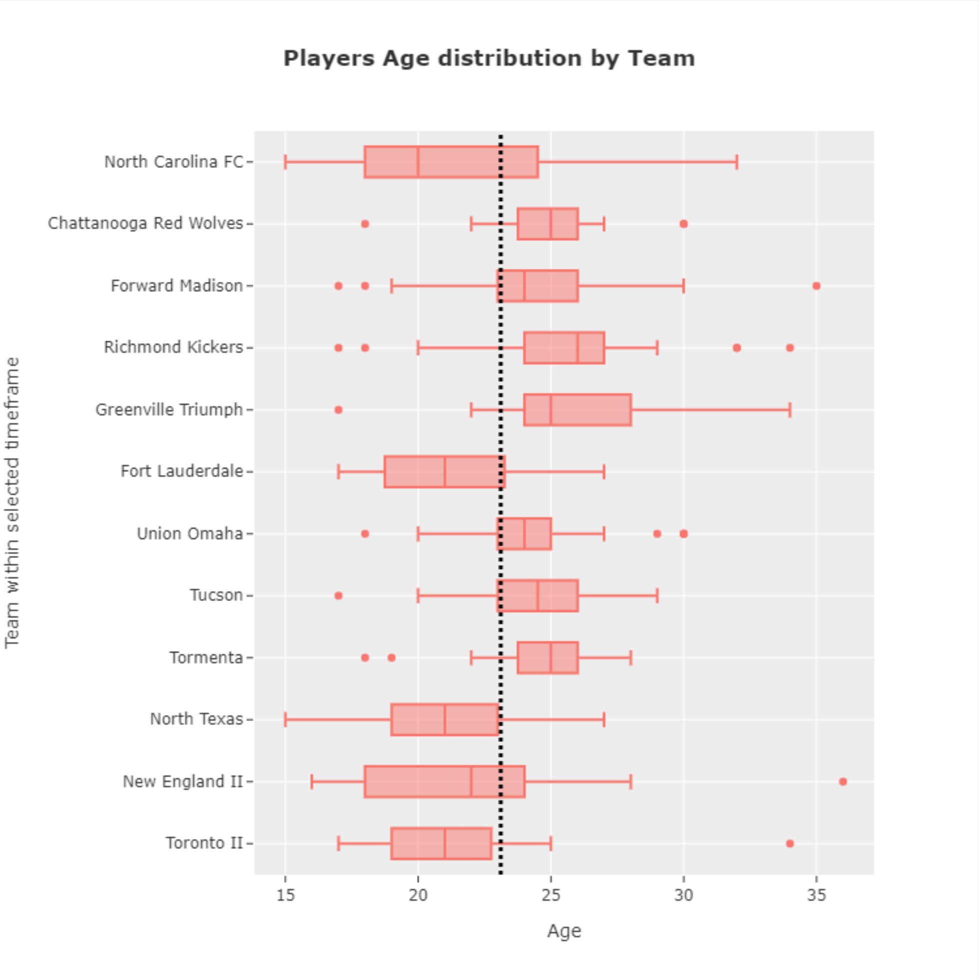
The column we focus on for the Transfermarkt dataset is on salary. The high frequency of the salary value is $110,000 with a frequency of 24.1%. There are eight missing values in this column. These might be new players; the data scouts have no current salary information or do not have enough to generate a salary.

6. Data visualization

Age Distributions (WyScout)

Plotting the age of soccer players reveals that *more than half* the players are younger than 25, peaking at 24-year-olds. 18 features a peak since recruitment efforts ramp up once the player enters adulthood. After 25, however, there is a sharp decline. Some skilled players play well into their 30s, while most end their soccer careers early. This is likely more indicative of career & lifestyle options available to such players than the decrepit age. However, past the mid-30s it becomes impossible to maintain peak physical condition.





The Box Plots of player ages by team reveal some peculiar patterns. We observe **North Carolina FC** & **North Texas** have perfectly continuous player ages and all other teams have outliers. As a whole, **North Carolina FC**, **Fort Lauderdale**, **North Texas**, **New England Ⅱ**, & **Toronto Ⅱ** have generally below average player ages & youthful teams. In contrast, the rest of the teams generally have above-average player ages. The following is a table of the age stats per team —

Team	Observations
North Carolina FC:	① This team has the largest age span ② The median age is lower than the average, and more than half of the players are younger than the average. This team is the youngest team among all teams ③ The team has the youngest players
Chattanooga Red Wolves	① This team has the highest age concentration, with most players between the ages of 24 and 26 ② The median age is higher than the average, and more than three-quarters of the players are older than the average
Forward Madison	① The median age is higher than the average, and three-quarters of the players are older than the average

Richmond Kickers	① This team has the largest age gap, with four outliers ② Median age is above average and more than three-quarters of the players are above average
Greenville Triumph	① This team is the *oldest of all teams* ② The median age is above the mean & more than 3/4 of the players are older than the average
Fort Lauderdale	① The median age is below the average, and three-quarters of the players are below the average
Union Omaha	① Most of the players are between 23-25 years old ② Median age is higher than average, & approx 3/4 of the players are older than the average
Tuscon	① The median age is higher than the average and three-quarters of the players are older than the average
Tormenta	① Most of the players are between 24-26 years old ② Median age is higher than the average & more than 3/4 of the players are older than the average
North Texas	①The median age is below the average and more than three-quarters of the players are below the average ② The team has the youngest player
New England II	① The team has the oldest player. ② The median age is below the average, and more than one-half of the players are below the average.
Toronto II	① The median age is lower than the average, and more than three-quarters of the players are below the average.

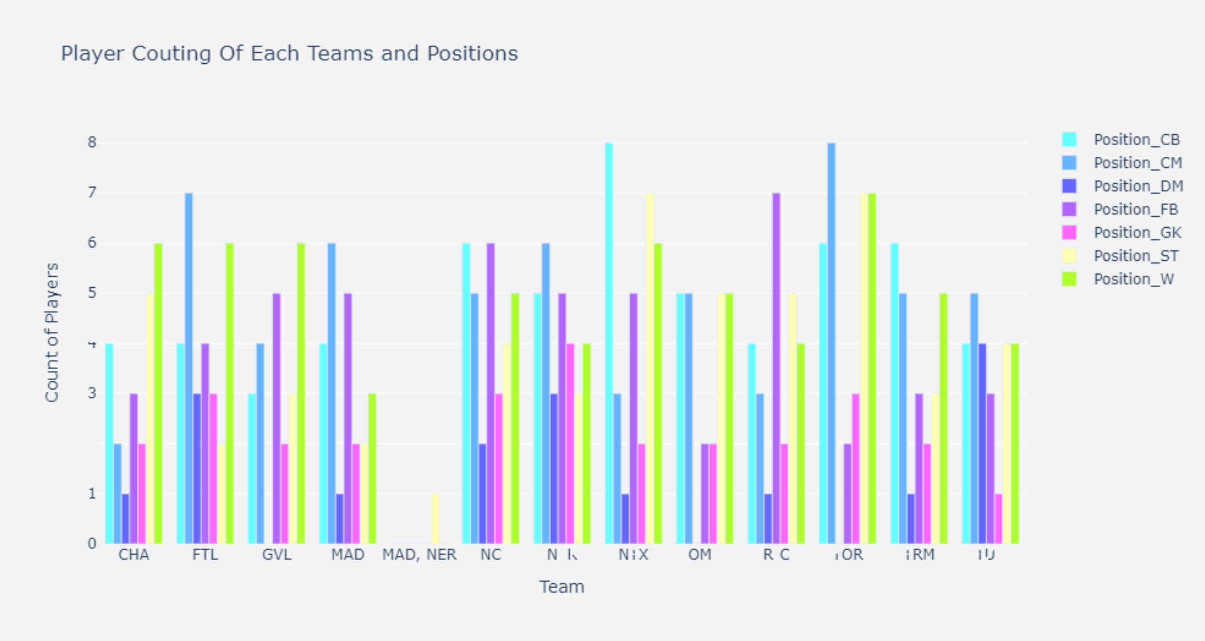
Counts by Team & Position (ASA)

We plotted the different teams' player types against the data from the American Soccer Analysis source. We found an interesting statistic that **MAD** and **NER** share the same player with the position **ST**. In general, the most popular position is **CM**, the more popular positions are **CB**, **FB**, and **W**, and the least popular position is **DM**. The following is a detailed analysis of each team.

Team	**Observations**
North Carolina FC:	① The team focuses on developing players in positions **CB, CM, FB, ST, & W**

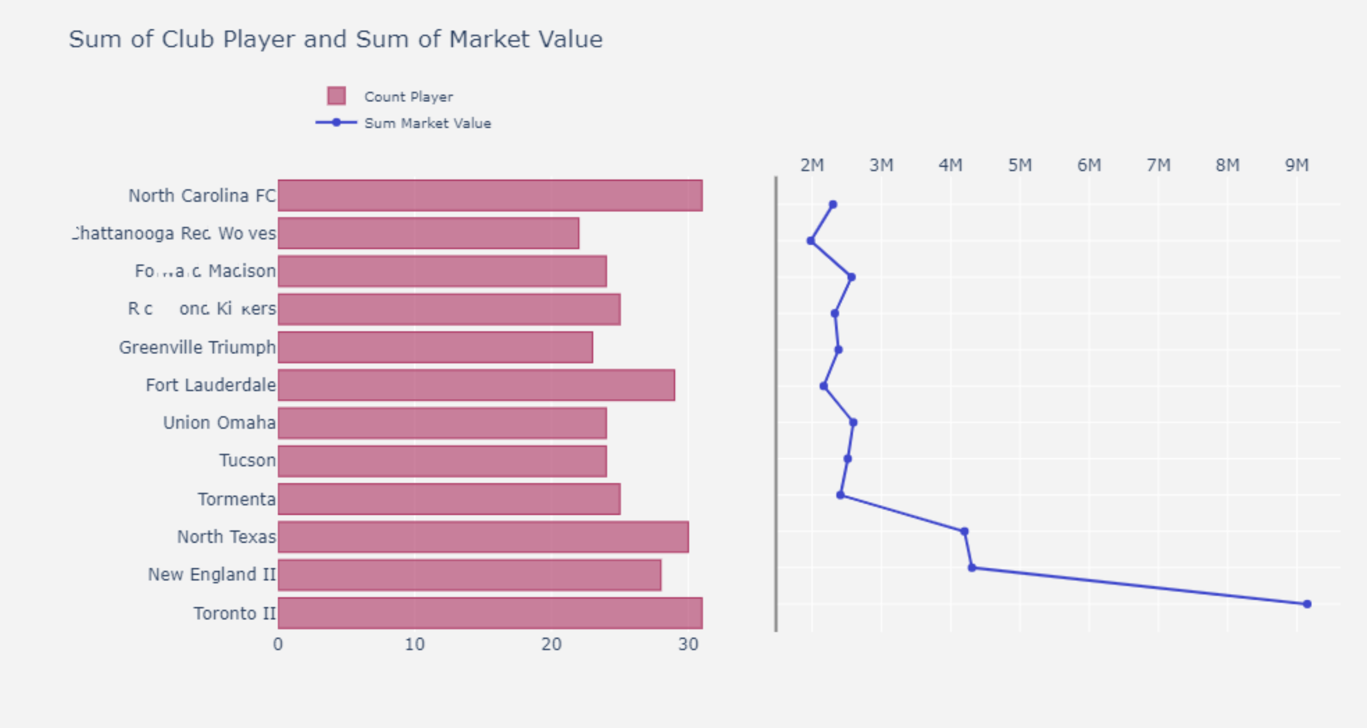
	② The team has the most players in positions **CB & FB**
Chattanooga Red Wolves	① The team focuses on developing players in positions **CB, ST, & W** ② The team has the most players in the position **W**
Forward Madison	① The team focuses on developing players in positions **CB, CM, & FB** ② The team has the most players in the position **CM**
Richmond Kickers	① The team focuses on developing players in positions **CB, FB, ST, & W** ② The team has the most players in the position **FB**
Greenville Triumph	① The team focuses on developing players in positions **CM, FB, & W** ② The team has the most players in the position **W** ③ The team has *no players* in the **DM** position.
Fort Lauderdale	① The team focuses on developing players in positions **CB, CM, FB, & W** ② The team has the most players in the position **CM**.
Union Omaha	① The team focuses on developing players in positions **CB, CM, ST, & W** ② The team has the most players in **CB, CM, ST, & W** ③ The team's player positions are *flatter* (than others) ④ The team has *no players* in the **DM** position
Tuscon	① The team focuses on developing players in positions **CB, CM, DM, ST, & W** ② The team has the most players in positions **CM & DM**
Tormenta	① The team focuses on developing players in positions **CB, CM, & W** ②The team has the most players in the position **CB**
North Texas	① The team focuses on developing players mostly in **CB, FB, ST, & W** ② The team has the most players in positions **CB & ST**
New England II	① The team focuses on developing players in positions **CB, CM, FB, GK, & W** ② The team has the most players in positions **CM & GK**
Toronto II	① The team focuses on developing players in positions **CB, CM, ST, & W**

	② The team has the most players in **CM, ST, & W** ③ The team has *no player* with a **DM** position



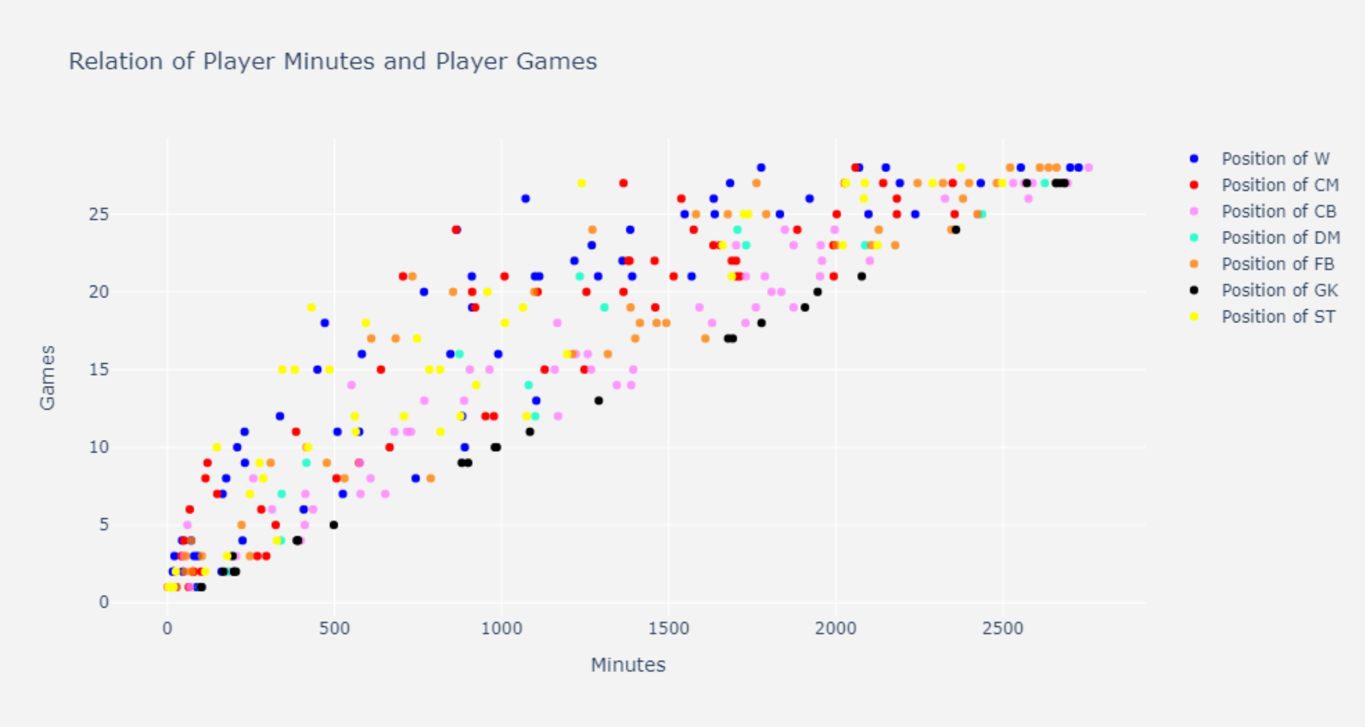
Sum of Club Players & Sum of Market Value (WyScout)

By mapping the number of players and market value of different teams with data from Wyscout sources. We can find that North Carolina FC and Toronto II are the two teams with the most players, with 31 players. Although the number of players in the same, we can find that the total value of Toronto II's players is more than four times the total value of North Carolina FC's players by the market value line chart. North Texas and New England II are in the second tier, followed by Toronto II, which is the most powerful and well-financed of all teams. The remaining teams are in the third tier, and the Chattanooga Red Wolves have the lowest stats in comparison.



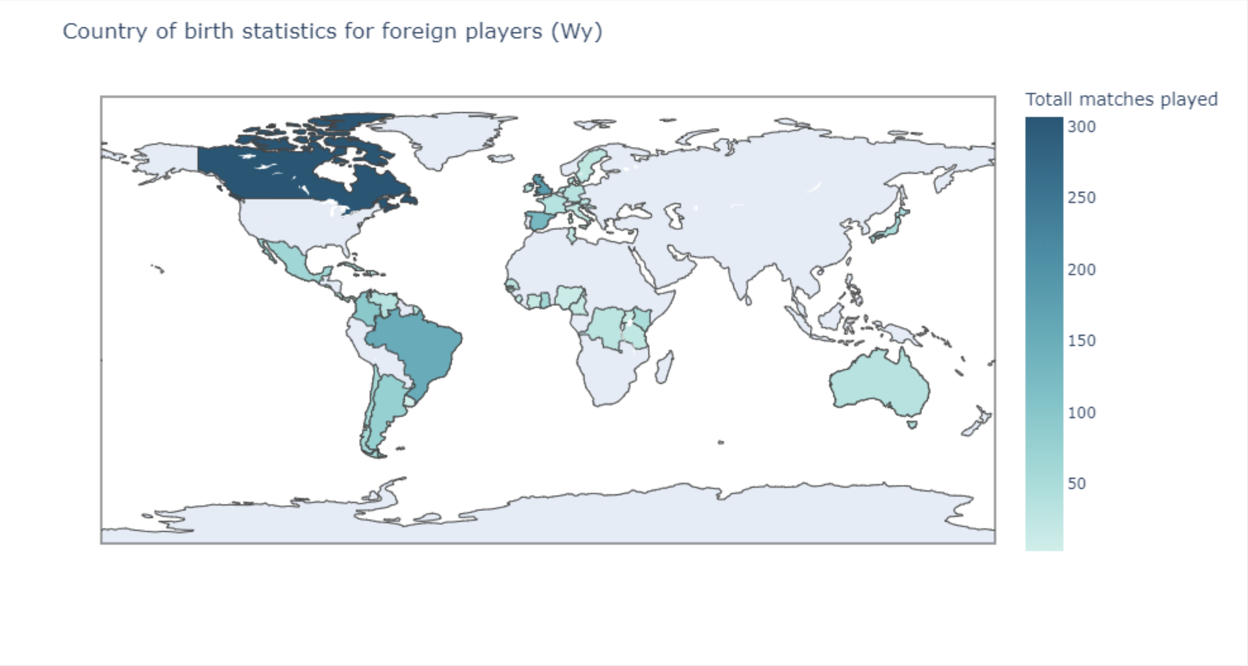
Relation of all player minutes and player games (ASA)

We plot the number of games played and minutes played for different positions of players from the American Soccer Analysis source. We find the black dots representing the goalkeepers form a *baseline* since the goalkeepers are *rarely replaced during a game*. The closer the data point is to this line, the more time the position represented by that data point appears in a game, which also reflects the importance of that position. As a whole, CB and DM are closer to the baseline overall. The remaining W, CM, FB, and ST are more spread out. So CB and DM have a higher average presence time and are more important.



Country of Origin by Player (WyScout)

We plot the birthplaces of the players in the data from WyScout sources. Since the data is recorded for USL league one players, the highest number of players are from the US mainland. *In fact, ~61% of players are from the United States*. Not surprising. To show what foreign countries American teams recruit from, we exclude players whose origin is the United States —



We find that a plurality of foreign players hail from Canada, followed closely by England and Spain, then Brazil, Argentina, and Columbia. Outside of North America, players on American teams are mostly from Europe and South America.

7. Features selection (Wyscout Dataset)

In our models, we took the counsel of our advisor (Jordan) as to what features mattered most by position. We compiled it into the table below. Given these groupings, we used each feature's *perceived importance* as weighted values for predictive models of skill set.

- **The most important feature of the position**
- Important features of the position
- Universally important features
- *Features with little reference value*

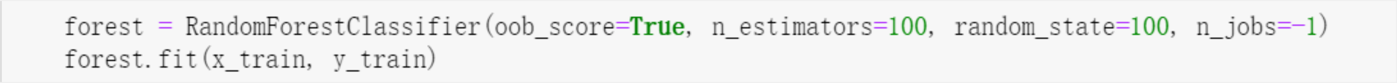
⚽	Position-Specific Feature Selection
FB	**Successful defensive actions per 90** Passes per 90, Accurate passes %, Short/medium passes per 90, Accurate short/medium passes % *Goals, Non-penalty goals, Assists, Key passes per 90, Duels per 90, Duels won %, Shots blocked per 90, Interceptions per 90, Crosses per 90, Accurate crosses %, Crosses to goalie box per 90, Received passes per 90, Accurate long passes %, Shot assists per 90, Passes to final third per 90, Accurate passes to final third %*
CB	**Successful defensive actions per 90, Duels per 90, Duels won %, Accurate long passes %, Through passes per 90, Accurate through passes %** Passes to final third per 90, Accurate passes to final third %, Dribbles per 90, Successful dribbles Passes per 90, Accurate passes %, Short/medium passes per 90, Accurate short/medium passes % *Goals, Non-penalty goals, Assists, Shots blocked per 90, Interceptions per 90, Received passes per 90, Key passes per 90*
CM	**Aerial duels per 90, Touches in box per 90, Average pass length m** Shot assists per 90, Passes to final third per 90, Accurate passes to final third %, Smart Pass per 90, Accurate smart passes % Passes per 90, Accurate passes %, Short/medium passes per 90, Accurate short/medium passes % *Goals, Non-penalty goals, Assists, Duels per 90, Duels won %, Sliding tackles per 90, Interceptions per 90, Received passes per 90, PAdj Interceptions*
AM	**Touches in box per 90, Forward passes per 90, Accurate forward passes %, Average pass length m, Second assists per 90** Successful attacking actions per 90, Offensive duels won %, Shot assists per 90, Third assists per 90, Passes to final third per 90, Accurate passes to final third %, Passes to penalty area per 90, Accurate passes to penalty area % Passes per 90, Accurate passes %, Short/medium passes per 90, Accurate short/medium passes % *Goals, Non-penalty goals, Assists, Duels per 90, Duels won %, Key passes per 90, Received passes per 90, Key passes per 90*

W	**Goals, Non-penalty goals, Assists, Crosses per 90, Accurate crosses %, Crosses to goalie box per 9, Dribbles per 90, Successful dribbles %** Successful attacking actions per 90, Offensive duels won %, Shot assists per 90, Key passes per 90 Passes per 90, Accurate passes %, Short / medium passes per 90, Accurate short / medium passes % *Key passes per 90, Duels per 90, Duels won %, Shots on target %, Goal conversion%, Received passes per 90, Passes to penalty area per 90, Accurate passes to penalty area %*
CF	**Goals, Non-penalty goals, Shots on target %, Goal conversion %, Key passes per 90** Aerial duels per 90, Aerial duels won %, Successful attacking actions per 90, Offensive duels won %, Shot assists per 90 Passes per 90, Accurate passes %, Short / medium passes per 90, Accurate short / medium passes % *Assists, Pass%, Key passes per 90, Duels per 90, Duels won %, Touches in box per 90, Received passes per 90*
GK	**Accurate short / medium passes %, Save rate %** Short / medium passes per 90, Shots against *Long passes per 90, Accurate long passes %, Conceded goals per 90, Clean sheet*

After discussing with Oakland Roots, we decide to select different features of each position.

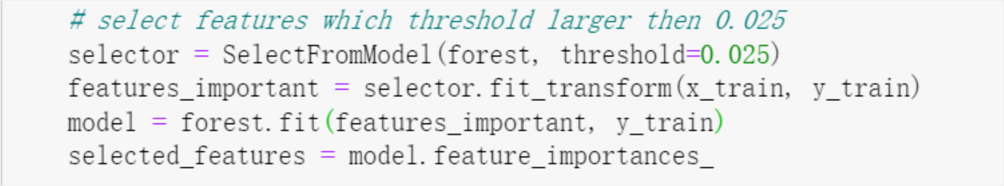
Deploy Random Forest Algorithm

To be able to further select the important features from the features of each position above, we decided to build a random forest model with the features of each position above as the independent variable and 'Rating' as the dependent variable.



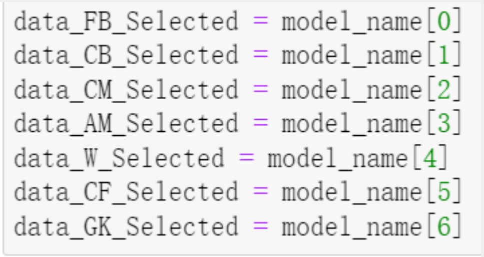
```
forest = RandomForestClassifier(oob_score=True, n_estimators=100, random_state=100, n_jobs=-1)
forest.fit(x_train, y_train)
```

Based on the model built, we exported a table of feature importance to the outside (we did not keep the table because it was only a temporary reference standard for our subsequent process.). With the importance table, we decided to control the number of features per position to about 20 features after observing the importance values of most of the features, so we further tuned the random forest model by setting the threshold of feature selection at 0.025.



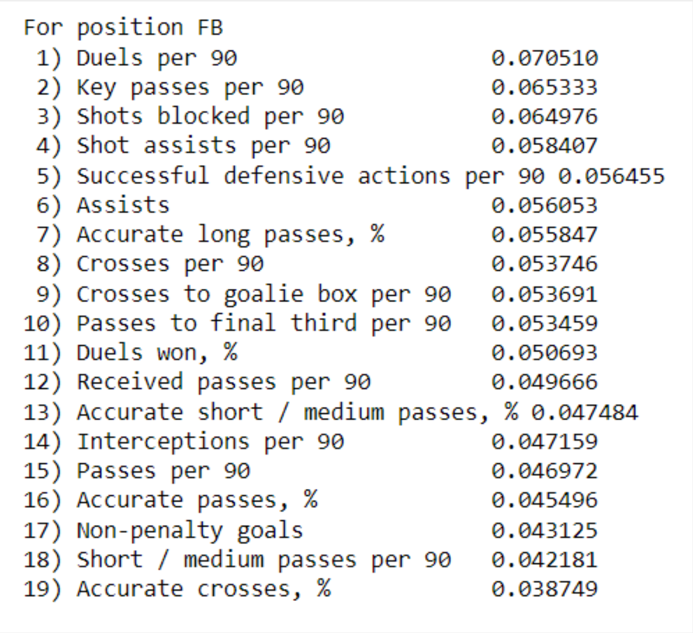
```
# select features which threshold larger then 0.025
selector = SelectFromModel(forest, threshold=0.025)
features_important = selector.fit_transform(x_train, y_train)
model = forest.fit(features_important, y_train)
selected_features = model.feature_importances_
```

Finally, we downscale and subdivide the overall data into seven subsets based on the features selected by the random forest model.

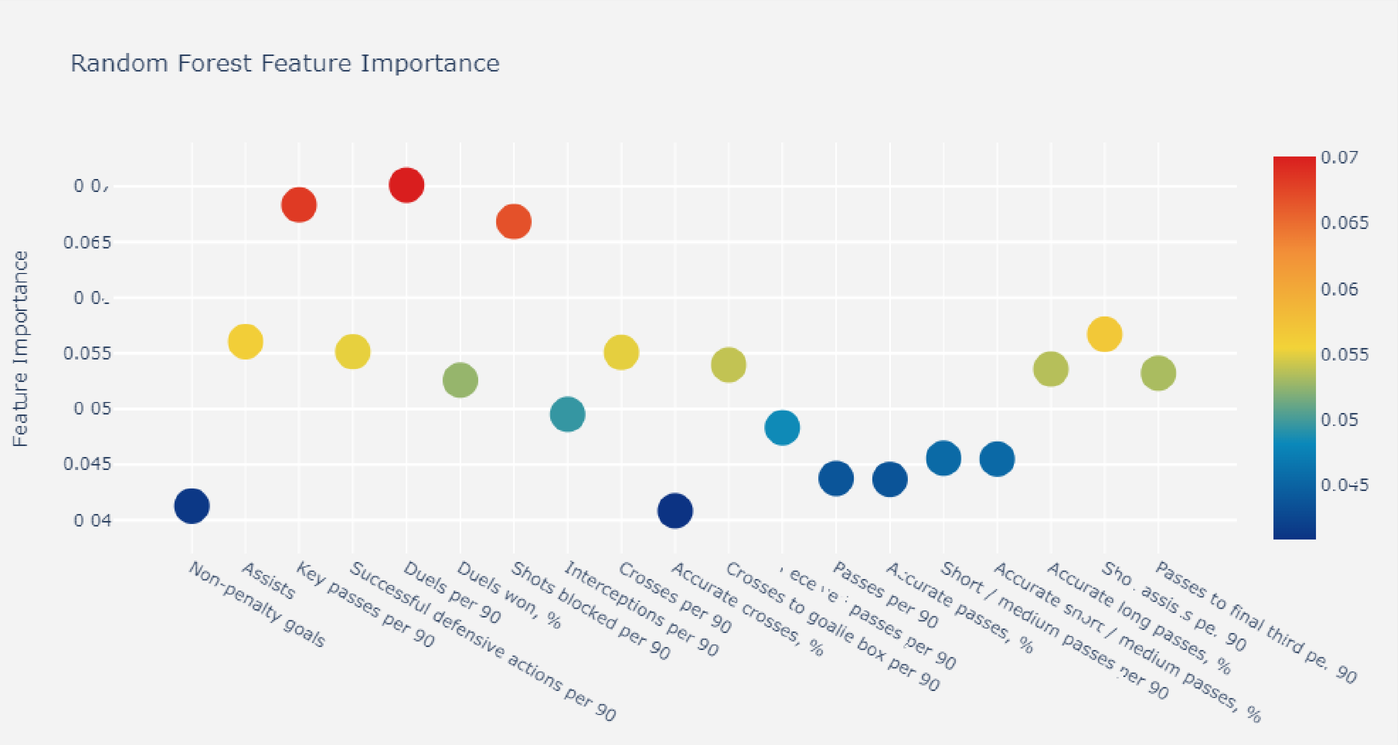


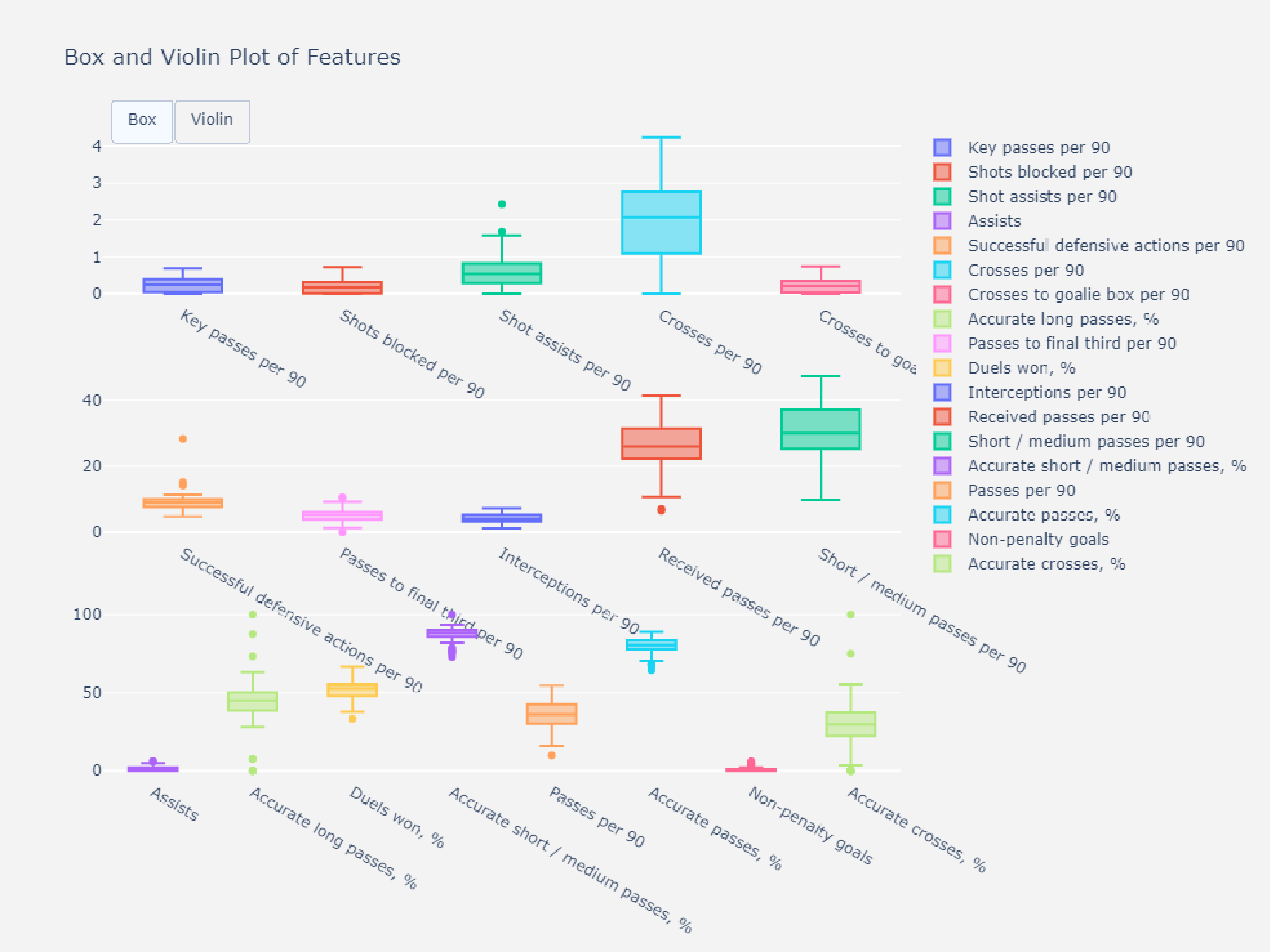
```
data_FB_Selected = model_name[0]
data_CB_Selected = model_name[1]
data_CM_Selected = model_name[2]
data_AM_Selected = model_name[3]
data_W_Selected = model_name[4]
data_CF_Selected = model_name[5]
data_GK_Selected = model_name[6]
```

Finally, we output the feature selection results for each position and plot the corresponding features in scatter plots, box plots, and violin plots. (Below is the result of our FB example.)



```
For position FB
 1) Duels per 90                    0.070510
 2) Key passes per 90               0.065333
 3) Shots blocked per 90            0.064976
 4) Shot assists per 90             0.058407
 5) Successful defensive actions per 90 0.056455
 6) Assists                         0.056053
 7) Accurate long passes, %         0.055847
 8) Crosses per 90                  0.053746
 9) Crosses to goalie box per 90    0.053691
10) Passes to final third per 90    0.053459
11) Duels won, %                    0.050693
12) Received passes per 90          0.049666
13) Accurate short / medium passes, % 0.047484
14) Interceptions per 90            0.047159
15) Passes per 90                   0.046972
16) Accurate passes, %              0.045496
17) Non-penalty goals               0.043125
18) Short / medium passes per 90    0.042181
19) Accurate crosses, %             0.038749
```





The two plots above are for the feature selection on the **Fullback** position. **Fullbacks** are responsible for defending both sides of the field. Their task is to prevent the opponent from passing the ball from the sideline into the penalty area, and they also help make passes for attacks. In the scatter plot, the important features of a **FB** are key passes per 90, duels per 90, and shots blocked per 90. These are all essential features for **FB** to prevent opponents from passing the ball, creating a chance for goals. There are also box and violin plots of the features to show the min, median, max, 0.25, 0.75, upper max, and outliers for each of the features.

8. Cosine distance

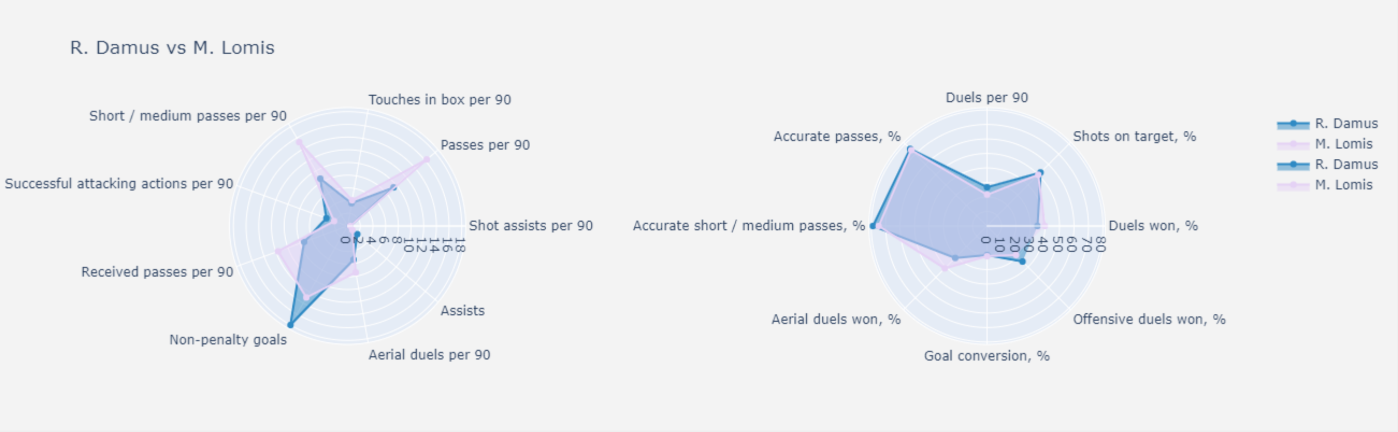
Explanation

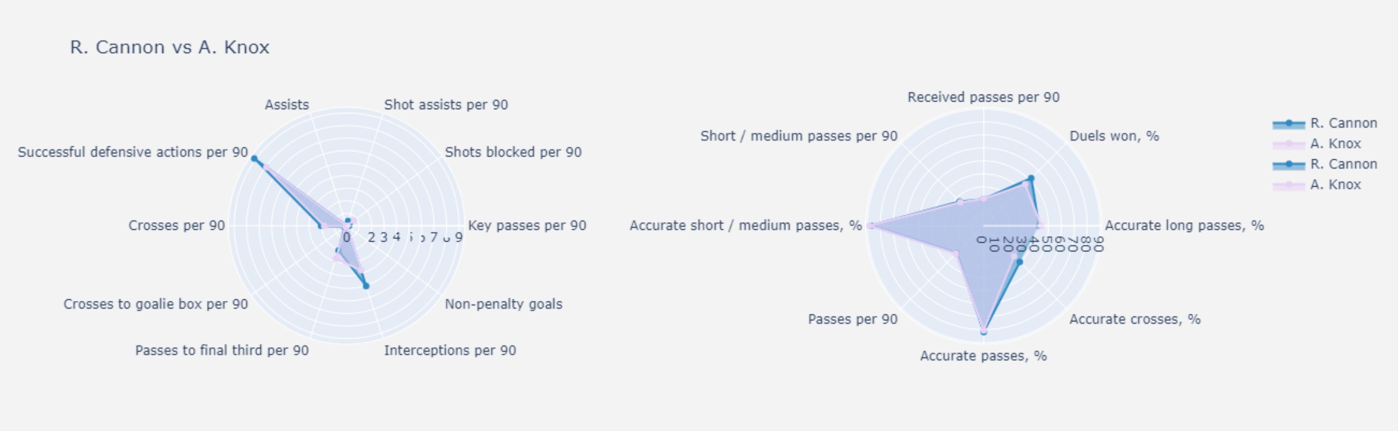
The client provided us with four ideal players (Reggie Cannon, James Sands, Ronaldo Damus, and Cole Bassett). When planning a player acquisition, these players have the statistics the client thought would best match his preference for these four positions (LB/RB, CB, FWD, and AM). These four positions were categorized into broader categories in previous steps (CB, CF, and FB). To locate other players similar to these five players, we calculated the cosine distance between the ideal players and similar players.

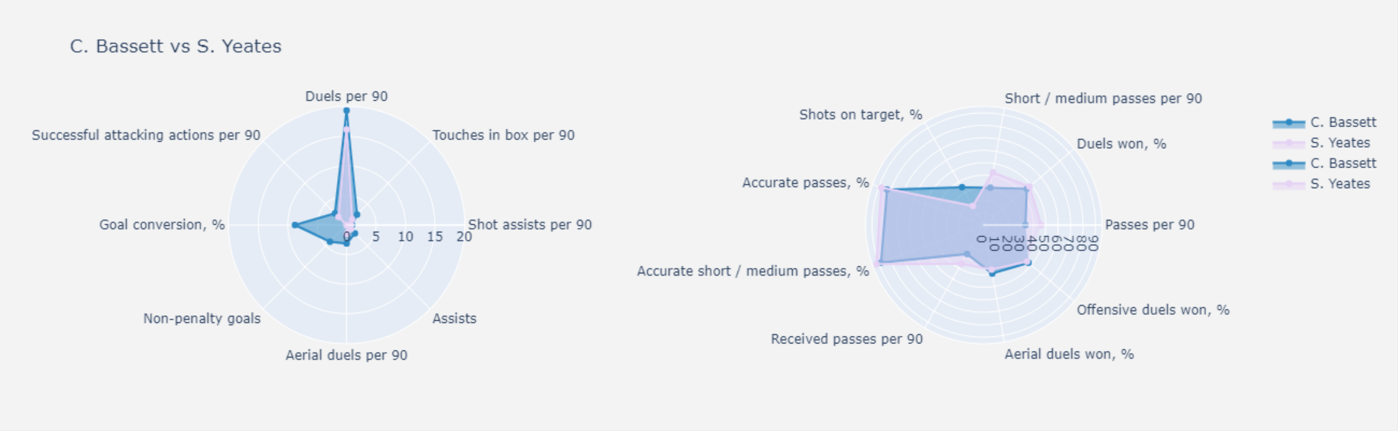
By extracting the model player's statistics, and the essential features from the positions data generated by the RF model. The dataset was normalized and separated the data into x, all the players' features, and y, the player's name. Next, we calculated the cosine similarity between the players. The angle between the players will give the percent similarity between the ideal player and comparable players. We then use cosine distance (1-cosine similarity) to determine the distance between the two players. If the players' cosine distance is closer to 0, the two players are alike. Players with more significant cosine similarity and the least cosine distance are the top similar players.

Radar Plot

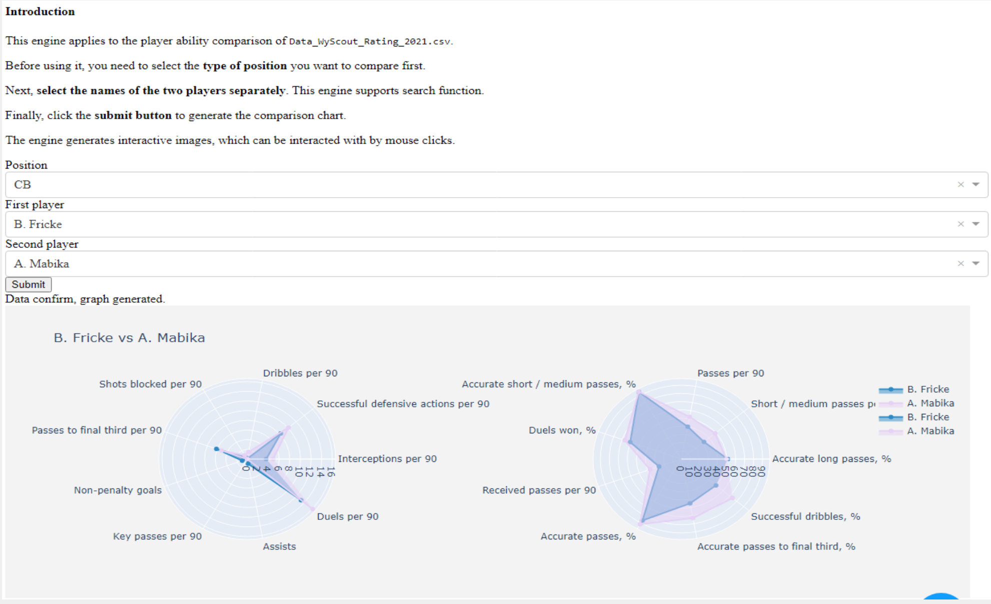
Based on the sample of exemplary players provided to us by Root Team, we selected a series of samples similar to the exemplary players by calculating the Cosine distance, and compared the players with the shortest Cosine distance with the exemplary sample and plotted the radar chart. Below are the radar plots of the model players and the sample that is most similar to them.







To enable our clients to compare all players, we have implemented this feature in **Dash**. The client can select the position of the player he wants to search for and then select the two players he wants to compare to get an idea of the player. This **Dashboard** can be viewed in a **Jupyter notebook**.



Introduction

This engine applies to the player ability comparison of `Data_WyScout_Rating_2021.csv`.

Before using it, you need to select the **type of position** you want to compare first.

Next, **select the names of the two players separately**. This engine supports search function.

Finally, click the **submit button** to generate the comparison chart.

The engine generates interactive images, which can be interacted with by mouse clicks.

Position: CB

First player: B. Fricke

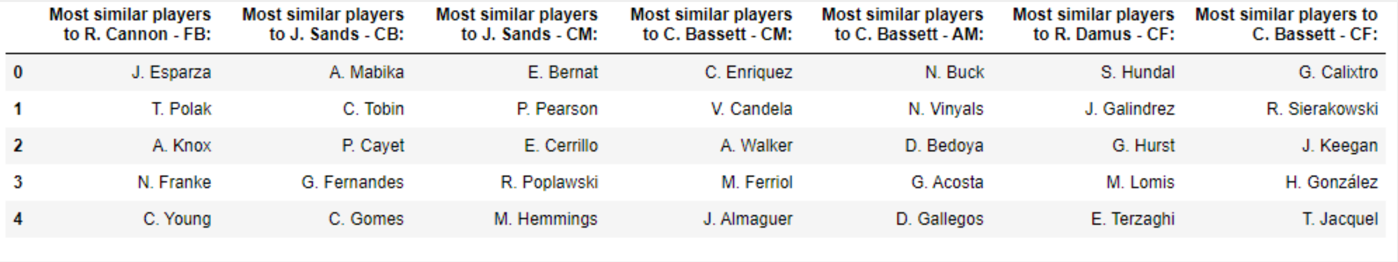
Second player: A. Mabika

Submit

Data confirm, graph generated.

Another Way Cosine Distance

For another way to find cosine similarity, we compiled the features on each position from the data_wy_rating_2021 to the dictionary. Then we implemented the cosine similarity to find the top 5 players similar to the ideal players. Lastly, we combined the two datasets with Transfermarkt's dataset to show the market value for similar players.



	Most similar players to R. Cannon - FB:	Most similar players to J. Sands - CB:	Most similar players to J. Sands - CM:	Most similar players to C. Bassett - CM:	Most similar players to C. Bassett - AM:	Most similar players to R. Damus - CF:	Most similar players to C. Bassett - CF:
0	J. Esparza	A. Mabika	E. Bernat	C. Enriquez	N. Buck	S. Hundal	G. Calixtro
1	T. Polak	C. Tobin	P. Pearson	V. Candela	N. Vinyals	J. Galindrez	R. Sierakowski
2	A. Knox	P. Cayet	E. Cerrillo	A. Walker	D. Bedoya	G. Hurst	J. Keegan
3	N. Franke	G. Fernandes	R. Poplawski	M. Ferriol	G. Acosta	M. Lomis	H. González
4	C. Young	C. Gomes	M. Hemmings	J. Almaguer	D. Gallegos	E. Terzaghi	T. Jacquel

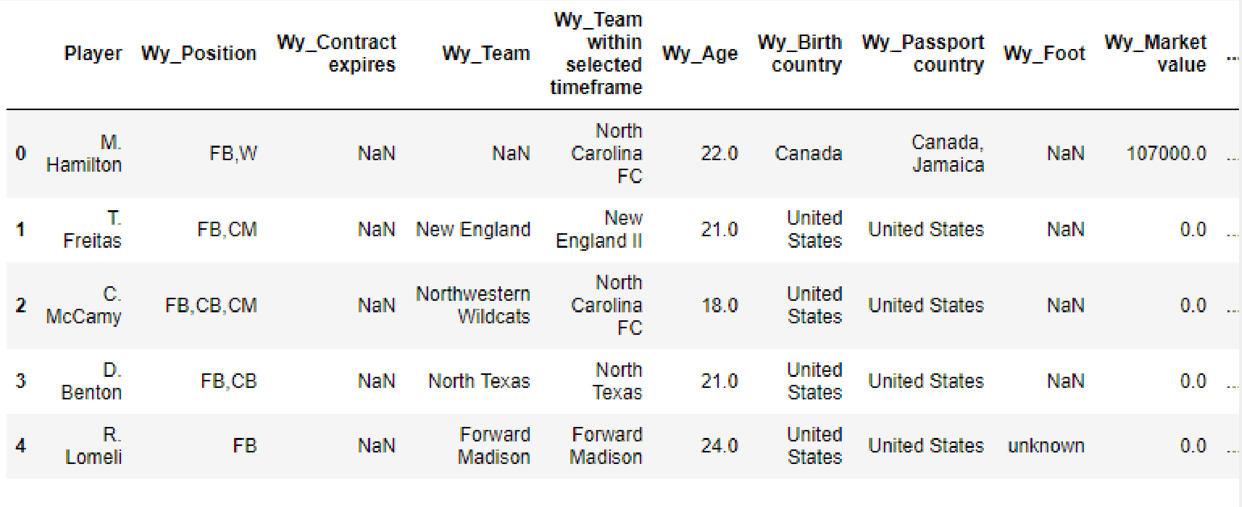
Results

The main functionality of the cosine similarity model is to help the client acquire the perfect player under different requirements. First, the client can use this model to find suitable players that can match the skill set and level he targets for each position. After locating the potential players, the radar plots will display different metrics for the client to explore in detail whether those players can match his skills preferences. Furthermore, the client can review player stats with the players' details extracted from Transfermarkt's dataset. The output will include the player's contract date and market value, so the client can assess different aspects if he is considering trading players in the future. For example, if an ideal player he chose has a contract that will soon expire, the client can assume a smooth transaction when trading players. If the contract still has time left, the client might need to pay fees for the player during trading due to breach of contract. The results generated by this model serve as a supplement opinion. It is only suggesting potential undervalued players the Oakland Roots should consider. The club should select players for acquisitions based on the actual situation before considering the selection of players.

8. Rating model

By obtaining the rating of the players' performance characteristics from the SofaScore website and merging it with the Wy data set using data processing techniques, a complete data set was generated for analysis and modeling. The rating score is the target value(y), and each position's different essential performance metrics are used as input variables for modeling and prediction. The rating prediction model is built using the player performance data of the 2020 season combined with the random forest algorithm, and the player performance data of the 2021 season is input to the model as the independent variable required for prediction. The rating value of the 2021 season is predicted and fetched through the model.

At the same time, we compare the predicted Ratings with the real values. We considered the players with a low real Rating value and high predicted Rating value as potential players. This group of players is considered the players who can reveal their potential through strenuous practical training to perform better in the future. Then, all models screen the candidates and obtain the player's personal information (e.g., left and right foot habits, contract information, salary information, etc.) from the Wyscout and Transfermarkt datasets.



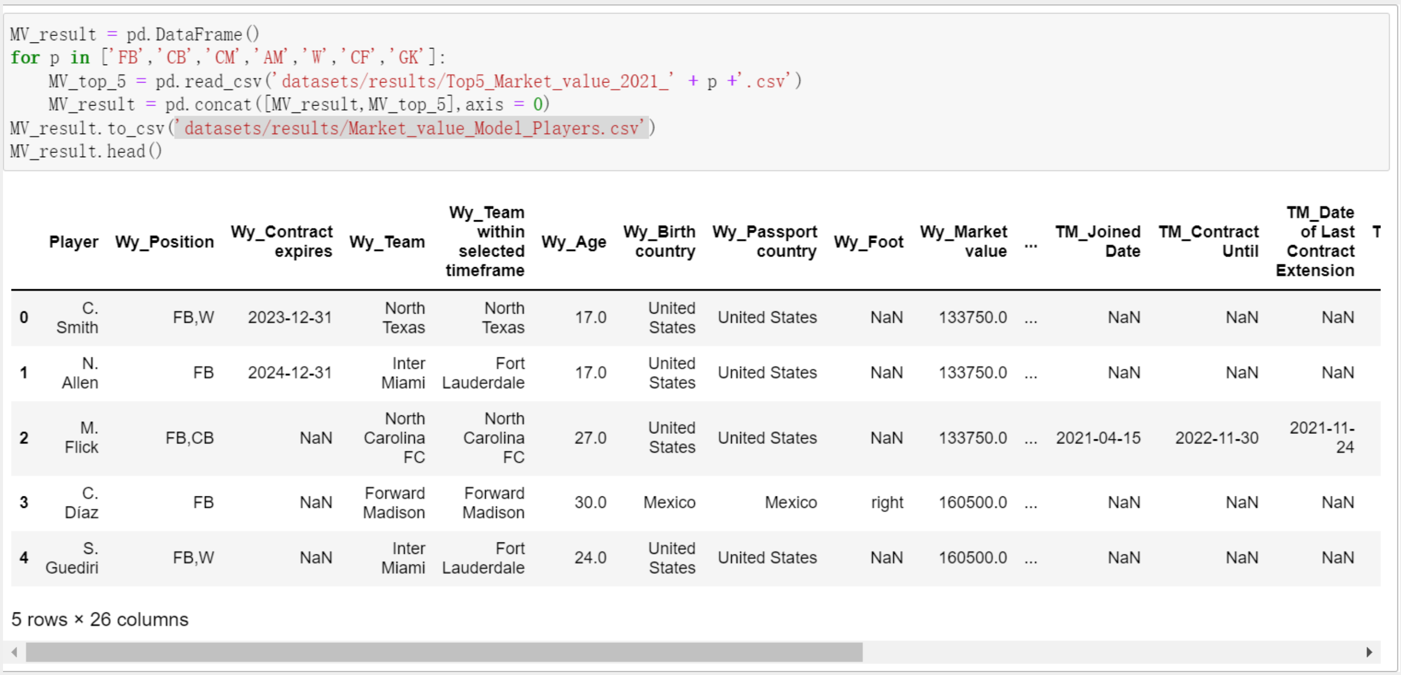
	Player	Wy_Position	Wy_Contract expires	Wy_Team	Wy_Team within selected timeframe	Wy_Age	Wy_Birth country	Wy_Passport country	Wy_Foot	Wy_Market value	...
0	M. Hamilton	FB,W	NaN	NaN	North Carolina FC	22.0	Canada	Canada, Jamaica	NaN	107000.0	...
1	T. Freitas	FB,CM	NaN	New England	New England II	21.0	United States	United States	NaN	0.0	...
2	C. McCamy	FB,CB,CM	NaN	Northwestern Wildcats	North Carolina FC	18.0	United States	United States	NaN	0.0	...
3	D. Benton	FB,CB	NaN	North Texas	North Texas	21.0	United States	United States	NaN	0.0	...
4	R. Lomeli	FB	NaN	Forward Madison	Forward Madison	24.0	United States	United States	unknown	0.0	...

9. Market value model

Market Value Forecast

We modeled and predicted by using salary data from Wy dataset and different characteristics of different positions. We use the important data of players in 2020 season combined with Random Forest algorithm to build the salary prediction model, and input the performance data of players in 2021 season as the variables needed for prediction into the model, and get the salary value of players in 2021 season through the model. At the same time, we compare the projected salary value with the real salary value, and we define the players whose real salary value is lower than the projected salary value as valuable players with potential, and use them as the player selection evaluation criteria. Through the random forest model, we exported the top 5 players of each position market value and stored the data in 'datasets/results/Top5_Market_value_2021_' + p + '.csv' in order by position.

To enable a unified view, we then read and merge these seven CSV files into a single CSV file stored in 'datasets/results/Market_value_Model_Players.csv'.



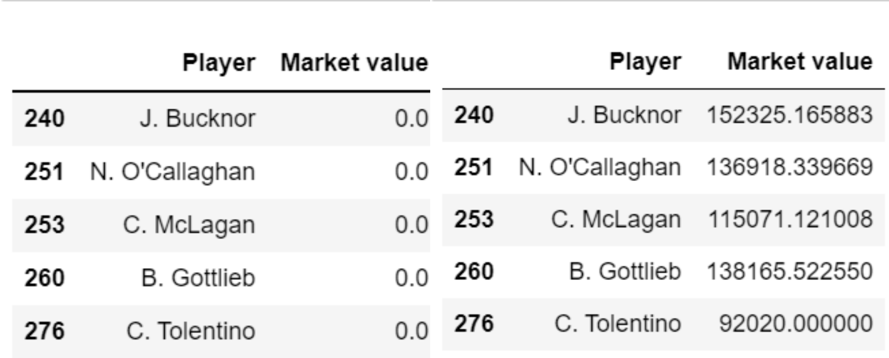
```
MV_result = pd.DataFrame()
for p in ['FB','CB','CM','AM','W','CF','GK']:
    MV_top_5 = pd.read_csv('datasets/results/Top5_Market_value_2021_' + p +'.csv')
    MV_result = pd.concat([MV_result,MV_top_5],axis = 0)
MV_result.to_csv('datasets/results/Market_value_Model_Players.csv')
MV_result.head()
```

	Player	Wy_Position	Wy_Contract expires	Wy_Team	Wy_Team within selected timeframe	Wy_Age	Wy_Birth country	Wy_Passport country	Wy_Foot	Wy_Market value	...	TM_Joined Date	TM_Contract Until	TM_Date of Last Contract Extension	T
0	C. Smith	FB,W	2023-12-31	North Texas	North Texas	17.0	United States	United States	NaN	133750.0	...	NaN	NaN	NaN	
1	N. Allen	FB	2024-12-31	Inter Miami	Fort Lauderdale	17.0	United States	United States	NaN	133750.0	...	NaN	NaN	NaN	
2	M. Flick	FB,CB	NaN	North Carolina FC	North Carolina FC	27.0	United States	United States	NaN	133750.0	...	2021-04-15	2022-11-30	2021-11-24	
3	C. Díaz	FB	NaN	Forward Madison	Forward Madison	30.0	Mexico	Mexico	right	160500.0	...	NaN	NaN	NaN	
4	S. Guediri	FB,W	NaN	Inter Miami	Fort Lauderdale	24.0	United States	United States	NaN	160500.0	...	NaN	NaN	NaN	

5 rows × 26 columns

Market Value supplement

For players with zero market value in the 2020 data, we construct a model set consisting of random forest, decision tree, and KNN to predict the due market value of these players. We first divided the 2020 data into a training set and a test set based on whether the market value is zero or not. (A sample of players with zero market value was used as the test set, and a sample of players with non-zero market value was used as the training set.) When training the model, we use market value as the dependent variable and other characteristics as the independent variables. Finally, we bring the test set into the model and take the average of the three results output by the model set as the final output and replace the zeros.



	Player	Market value
240	J. Bucknor	0.0
251	N. O'Callaghan	0.0
253	C. McLagan	0.0
260	B. Gottlieb	0.0
276	C. Tolentino	0.0

	Player	Market value
240	J. Bucknor	152325.165883
251	N. O'Callaghan	136918.339669
253	C. McLagan	115071.121008
260	B. Gottlieb	138165.522550
276	C. Tolentino	92020.000000

Our original plan was to add continuity to the data by re-submitting the completed data to the previous model, but we eventually abandoned the idea because the sample size was too small to guarantee the accuracy of the model.

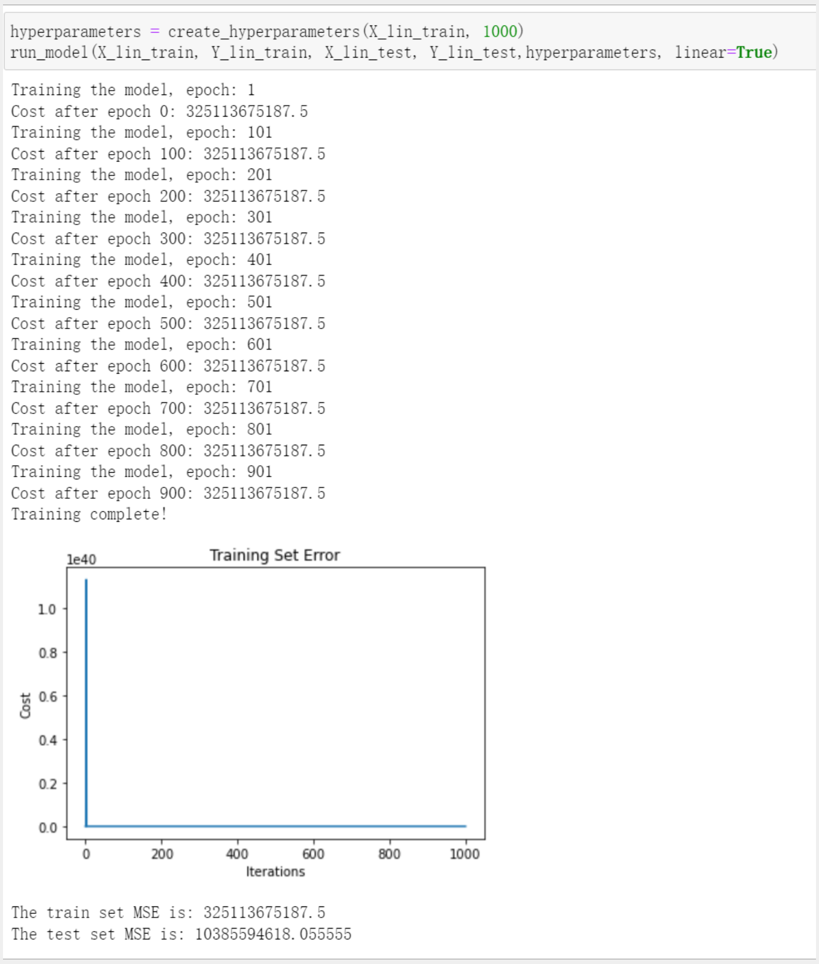
10. Abandon cases

A neural network model for market value prediction

The intention of using a neural network in the market value section is to predict the market value of players in 2021, using players' market value in the Wyscout 2020 dataset as training data. To use 2020's dataset, we have to fill in the null market values for some players. The first step is to split the 2020 dataset by positions, choose the players with market values as the training dataset, and select the player with no market values as the test dataset. In the second step, we used a random forest classifier to choose essential features relevant to market value. After data processing, we create the feedforward neural network.

In the neural network, we initialized parameters, including the weights and biases for each layer; the weights are randomly initialized from Gaussian distribution to ensure the weights are not too large. We used the relu function in the hidden layers in the activation function. We then created forward propagation and lost function to minimize loss during training. The backpropagation function is used to calculate the partial derivative of the cost function. We have set up a mean squared error to evaluate if the model is learning. In the train_model function, we combined the functions and trained them using gradient descent. Lastly, we plot the training set error over the number of iterations and the final evaluation metric output for training and test sets.

The performance of the neural network wasn't that great. The cost after the epoch is very high, and the MSE for both train and test set shows this is a terrible neural network model. This model's problems can be underfitting because the MSE on the test set is smaller than the train set. Another thing can be that the functions are miswritten, or we might have to explore different options for gradient descent.



```
hyperparameters = create_hyperparameters(X_lin_train, 1000)
run_model(X_lin_train, Y_lin_train, X_lin_test, Y_lin_test,hyperparameters, linear=True)
```

```
Training the model, epoch: 1
Cost after epoch 0: 325113675187.5
Training the model, epoch: 101
Cost after epoch 100: 325113675187.5
Training the model, epoch: 201
Cost after epoch 200: 325113675187.5
Training the model, epoch: 301
Cost after epoch 300: 325113675187.5
Training the model, epoch: 401
Cost after epoch 400: 325113675187.5
Training the model, epoch: 501
Cost after epoch 500: 325113675187.5
Training the model, epoch: 601
Cost after epoch 600: 325113675187.5
Training the model, epoch: 701
Cost after epoch 700: 325113675187.5
Training the model, epoch: 801
Cost after epoch 800: 325113675187.5
Training the model, epoch: 901
Cost after epoch 900: 325113675187.5
Training complete!
```

```
The train set MSE is: 325113675187.5
The test set MSE is: 10385594618.055555
```

11. Next Steps

Our team has enjoyed working on this project, and there are some objectives we think would furthermore benefit the Oakland Roots. First, implementing a better predictive algorithm for the undervalued players in TransferMarkt Data will go long. TransferMarkt's website has many aspects that would benefit the team for player acquisition. Having a robust algorithm to predict undervalued players is crucial. Second, during our meeting with the client, we pitched the idea of building a model to detect players whose skill sets are better suited to other field positions. If this money ball algorithm is successful, the player could potentially help play multiple positions; this would benefit the players by adding more value to their skills and will also help the team to discover more potential ways to plan tactics.

Another idea for the next capstone team is to build an interactive dashboard for the Oakland Roots to use. With players and team information constantly updating, the Oakland Roots can use this dashboard to monitor and select players or teams to compare and command. The next team would have to develop a GitHub workflow and recurrent web scraping for the client to utilize because the information on different websites is constantly changing.

12. References

1. Wyscout Dataset
https://footballdata.wyscout.com/

2. ASA Dataset
https://app.americansocceranalysis.com/#!/usl1

3. TransferMarkt Dataset Webscrape
https://www.transfermarkt.us/usl-league-one/torschuetzenliste/wettbewerb/USC3/ajax/yw1/saison_id/2020/altersklasse/alle/detailpos//plus/1/page/{}?ajax=yw1
https://www.transfermarkt.us/usl-league-one/gastarbeiter/wettbewerb/USC3/saison_id/2020

4. Sofa Score Dataset Webscrape (2020&2021)
https://www.sofascore.com/tournament/football/usa/usl-league-one/13362