Mubashir Ali
Muhammad Abdullah

BUSINESS UNDERSTANDING

BACKGROUND:

This dataset is taken for the footballers around the world who performs under FIFA. This dataset shows the different type of attributes to show the details about the football players around the world.

BUSINESS OBJECTIVES:

The objective is to enhance the correlations among the attributes of the data.

DATA UNDERSTANDING

COLLECT INITIAL DATA:

The data provided is consists of:

- data.xlsx: contains information about the player's id, name, age, height, salary, performance, nationality etc as a total of 17 attributes and 299 rows.

DESCRIPTION OF THE DATA:

The file containing the data required for data mining is data.xlsx. Below is the description of each of its fields and their data type.

VARIABLE NAME	DESCRIPTION	TYPE
ID	Id of the player	Integer
Name	Name of the player	Nominal
Age	Age of the player	Integer
Nationality	Nationality of the player	Nominal
Overall Performance	Overall performance of the player in his career	Integer
Potential	His potential to deal within the match	Integer
Club	Domestic club in which the player played	Nominal
Assets in €M	Total assets of the player in his entire career in € million	Real

Salary in €K	His salary in € thousands	Integer
Preferred Foot	The foot he used to preferred during the match	Nominal
International Reputation	International reputation of the player	Integer
Weak Foot	The weak foots' counts which the player uses in the match	Integer
Skill Moves	Skill moves for the player	Integer
Height	Height in feet inches of the player	Nominal
Weight in lbs	Weight of the player in pounds	Integer
Stamina	Stamina of the player between the matches	Integer
Strength	Strength of the player in the match	Integer

DATA PREPARATION:

The data should be free from null or bogus values as it is not a good practice to have such values.

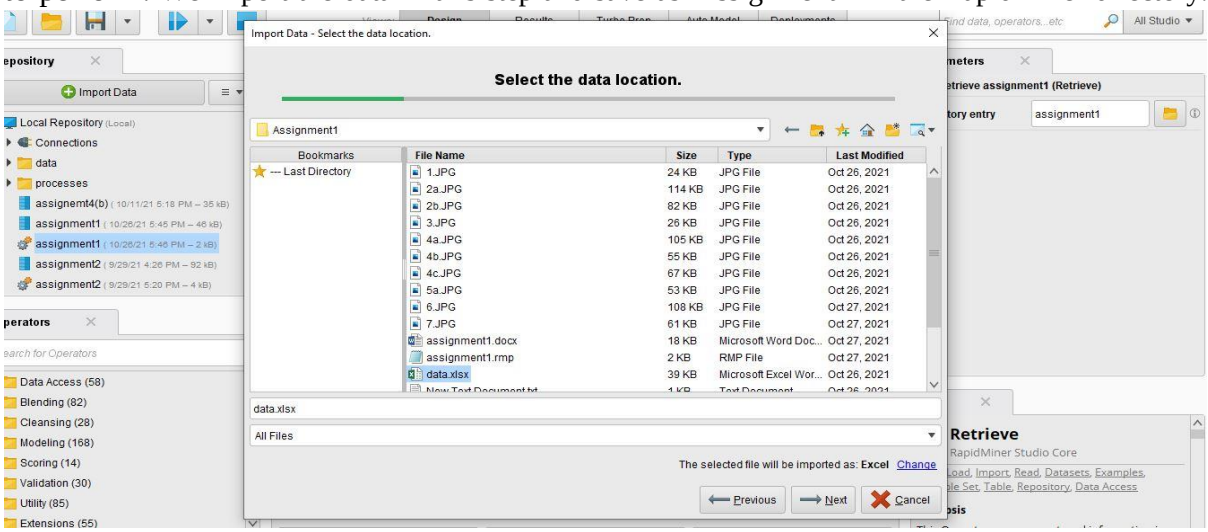
CLEAN DATA:

This process was performed using the RapidMiner tool.

The dataset doesn't contain any null value in it so it is already in a clean shape. So after this, we again have the same data as it is shown in the above table.

FORMAT DATA:

The data was cleaned and in this process it was added to RapidMiner directory for the procedure to perform. We import the data in this step and save as 'Assignment1' in the RapidMiner directory.



MODELING:

In this model we add first the dataset and after that we look into the results by running it. (See below figures)

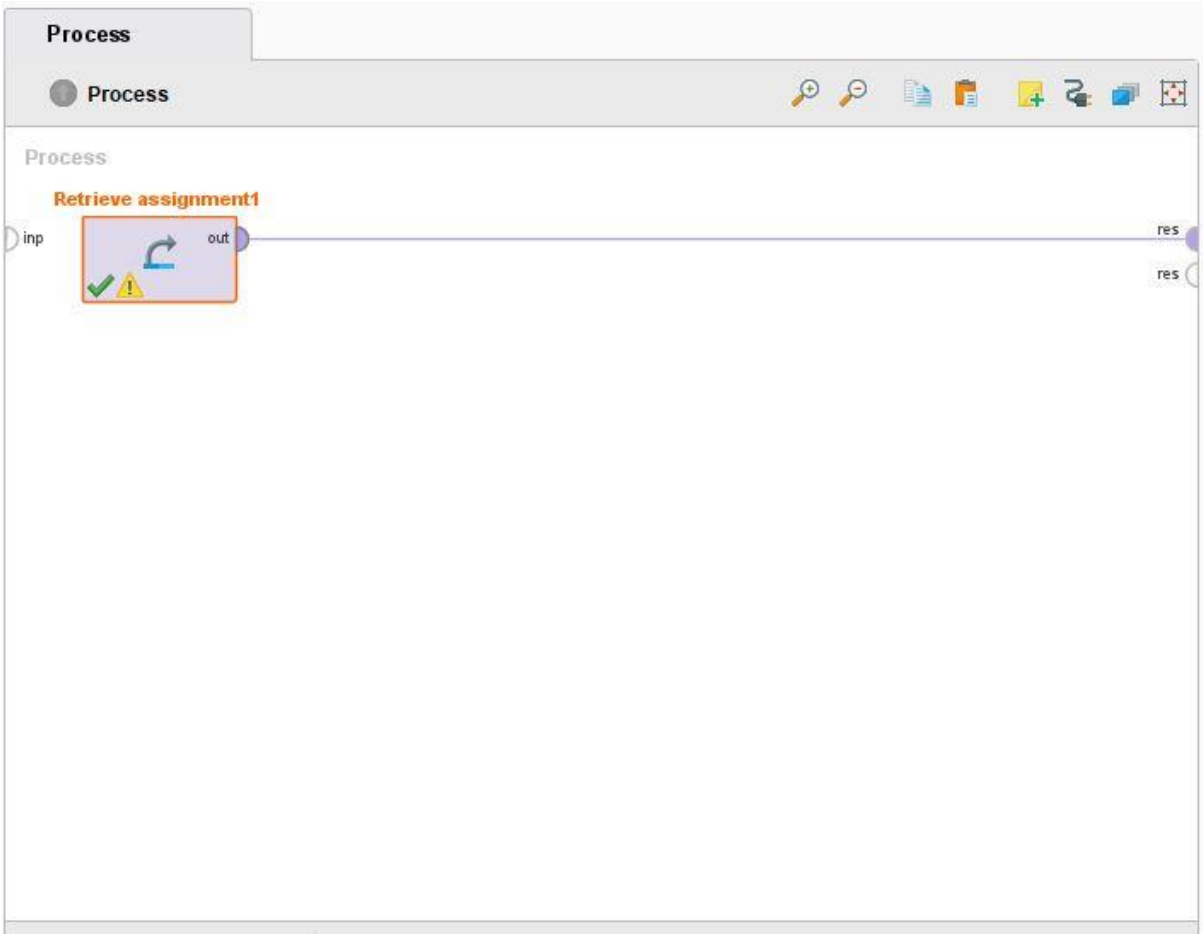


Figure 1: This fig shows the dataset while adding after in the process.

After running the process we get the below 2 figures (Figure2 & Figure3):

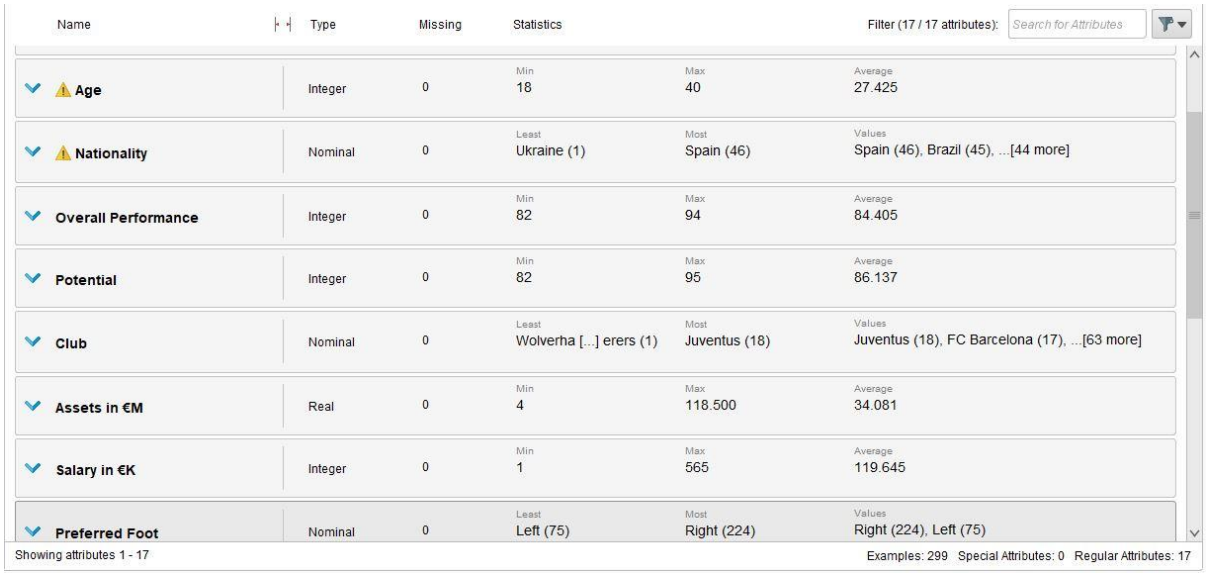

Open in Turbo Prep Auto Model

Filter (299 / 299 examples): all

Row No.	ID	Name	Age	Nationality	Overall Perf...	Potential	Club	Assets in €M	Salary in €K	Preferred Fo...	Internationa
1	158023	L. Messi	31	Argentina	94	94	FC Barcelona	110.500	565	Left	5
2	20801	Cristiano Ro...	33	Portugal	94	94	Juventus	77	405	Right	5
3	190871	Neymar Jr	26	Brazil	92	93	Paris Saint-G...	118.500	290	Right	5
4	193080	De Gea	27	Spain	91	93	Manchester ...	72	260	Right	4
5	192985	K. De Bruyne	27	Belgium	91	92	Manchester ...	102	355	Right	4
6	183277	E. Hazard	27	Belgium	91	91	Chelsea	93	340	Right	4
7	177003	L. Modri?	32	Croatia	91	91	Real Madrid	67	420	Right	4
8	176580	L. Su?rez	31	Uruguay	91	91	FC Barcelona	80	455	Right	5
9	155862	Sergio Ramos	32	Spain	91	91	Real Madrid	51	380	Right	4
10	200389	J. Oblak	25	Slovenia	90	93	Atlético Madrid	68	94	Right	3
11	188545	R. Lewandow...	29	Poland	90	90	FC Bayern M...	77	205	Right	4
12	182521	T. Kroos	28	Germany	90	90	Real Madrid	76.500	355	Right	4
13	182493	D. God?n	32	Uruguay	90	90	Atlético Madrid	44	125	Right	3
14	168542	David Silva	32	Spain	90	90	Manchester ...	60	285	Left	4

ExampleSet (299 examples, 0 special attributes, 17 regular attributes)

Figure 2: This fig shows the results of 'Data' section



Name	Type	Missing	Statistics		
Age	Integer	0	Min 18	Max 40	Average 27.425
Nationality	Nominal	0	Least Ukraine (1)	Most Spain (46)	Values Spain (46), Brazil (45), ...[44 more]
Overall Performance	Integer	0	Min 82	Max 94	Average 84.405
Potential	Integer	0	Min 82	Max 95	Average 86.137
Club	Nominal	0	Least Wolverha [...] erers (1)	Most Juventus (18)	Values Juventus (18), FC Barcelona (17), ...[63 more]
Assets in €M	Real	0	Min 4	Max 118.500	Average 34.081
Salary in €K	Integer	0	Min 1	Max 565	Average 119.645
Preferred Foot	Nominal	0	Least Left (75)	Most Right (224)	Values Right (224), Left (75)

Showing attributes 1 - 17 Examples: 299 Special Attributes: 0 Regular Attributes: 17

Figure 3: This fig shows the 'Statistics' section after running the dataset

After running the process while just entering the dataset we add the operator **'Correlation Matrix**' with parameters as *'random seed'* to '*2001*' and *'send email'* as *'never'*.

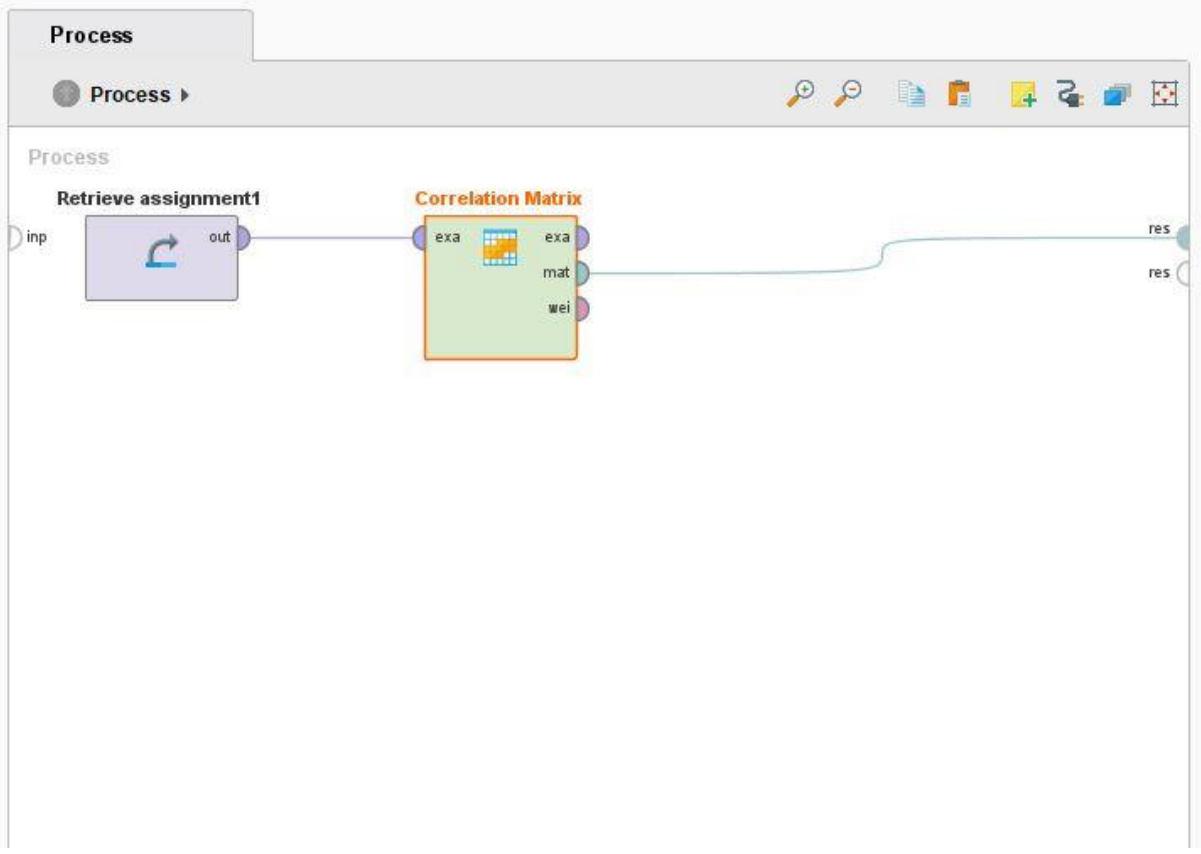


Figure 4: This fig shows the real process of the assignment i.e. 'Correlation Matrix' operator added to the process.

Later on, we run the process and the following results of the correlation operator.

Attributes	ID	Name	Age	National...	Overall Performance	Potential	Club	Assets i...	Salary i...	Preferr...	Internat...	Weak F...	Skill Mo.
ID	1	?	-0.717	?	-0.267	0.139	?	0.146	-0.117	-0.082	-0.483	0.001	0.039
Name	?	1	?	?	?	?	?	?	?	?	?	?	?
Age	-0.717	?	1	?	0.230	-0.429	?	-0.275	0.093	0.127	0.425	-0.052	-0.171
Nationality	?	?	?	1	?	?	?	?	?	?	?	?	?
Overall Performance	-0.267	?	0.230	?	1	0.668	?	0.802	0.772	0.007	0.648	0.100	0.022
Potential	0.139	?	-0.429	?	0.668	1	?	0.771	0.542	-0.105	0.259	0.078	0.021
Club	?	?	?	?	?	?	1	?	?	?	?	?	?
Assets in €M	0.146	?	-0.275	?	0.802	0.771	?	1	0.721	-0.055	0.414	0.228	0.291
Salary in €K	-0.117	?	0.093	?	0.772	0.542	?	0.721	1	-0.046	0.546	0.132	0.165
Preferred Foot	-0.082	?	0.127	?	0.007	-0.105	?	-0.055	-0.046	1	0.094	0.280	-0.098
International Reputation	-0.483	?	0.425	?	0.648	0.259	?	0.414	0.546	0.094	1	0.104	0.121
Weak Foot	0.001	?	-0.052	?	0.100	0.078	?	0.228	0.132	0.280	0.104	1	0.336
Skill Moves	0.039	?	-0.171	?	0.022	0.021	?	0.291	0.165	-0.098	0.121	0.336	1
Height	?	?	?	?	?	?	?	?	?	?	?	?	?
Weight in lbs	-0.157	?	0.162	?	0.062	0.037	?	-0.148	-0.059	0.084	0.080	-0.184	-0.583

Figure 5: The 'Data' tab results are shown in this fig

First Attribute	Second Attribute	Correlation
Nationality	Height	?
Nationality	Weight in lbs	?
Nationality	Stamina	?
Nationality	Strength	?
Overall Performance	Potential	0.668
Overall Performance	Club	?
Overall Performance	Assets in €M	0.802
Overall Performance	Salary in €K	0.772
Overall Performance	Preferred Foot	0.007
Overall Performance	International Reputation	0.648
Overall Performance	Weak Foot	0.100
Overall Performance	Skill Moves	0.022
Overall Performance	Height	?
Overall Performance	Weight in lbs	0.062
Overall Performance	Stamina	0.044

Figure 6: This fig shows the correlation among the attributes

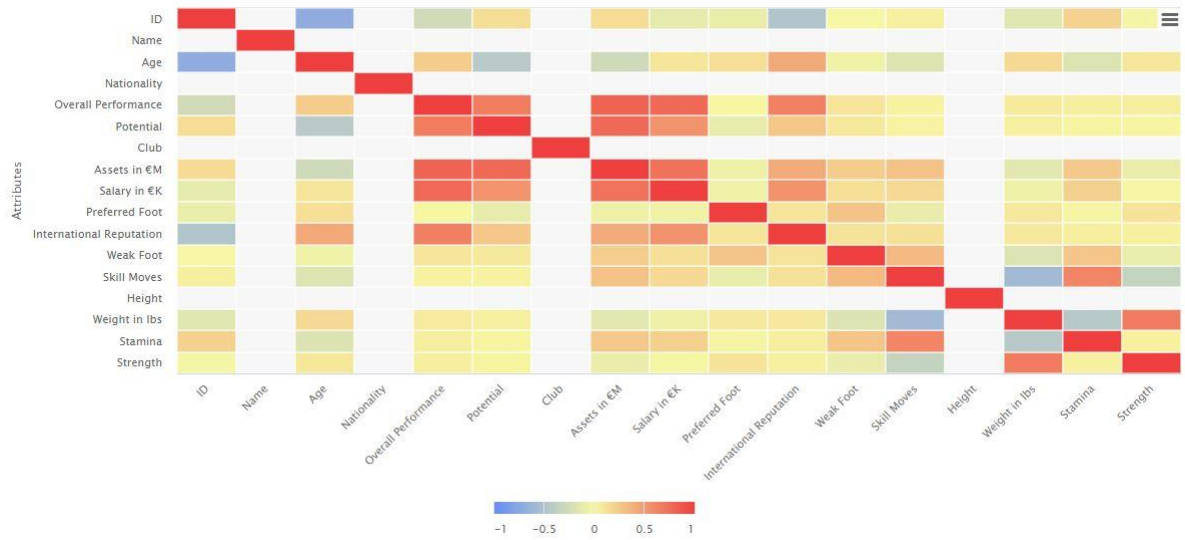


Figure 7: This fig shows the 'Matrix Visualization' tab

First Attribute	Second Attribute	Correlation ↑
Age	International Reputation	0.425
Potential	Salary in €K	0.542
Salary in €K	International Reputation	0.546
Skill Moves	Stamina	0.618
Overall Performance	International Reputation	0.648
Overall Performance	Potential	0.668
Weight in lbs	Strength	0.671
Assets in €M	Salary in €K	0.721
Potential	Assets in €M	0.771
Overall Performance	Salary in €K	0.772
Overall Performance	Assets in €M	0.802
ID	Name	?
ID	Nationality	?
ID	Club	?
ID	Height	?

Figure 8: This fig shows the 'Overall Performance' and 'Salary in €K' also with 'Assets in €M' and their correlation among them. These correlation values are among the highest positive correlation values that we get.

REPORT:

What correlations exist?

The Correlation exists among the 13 attributes as we found out through the process explained earlier.

How strong are they?

They (attributes) make a strong relationship among them as can be shown from the results as well. There exist negative as well as positive linear relationships between them.

Are they surprising to you? If so, why?

No, they aren't surprising to me as well because I choice the dataset which is suitable to our requirement.

What other attributes would you like to add?

Nothing other attributes because we already added 17 attributes which are more than the enough for us.

Are there any you would like to eliminate now? Why?

Yes, we'll eliminate the 'Nationality' and 'Club' attributes because these attributes are not good to have a correlation as they are not integer type.

REFERENCES:

Dataset

https://www.kaggle.com/karangadiya/fifa19