PCA/Cluster Analysis/ Regression

Team Name: Neo

EBAC 4 (Part Time), Institute of Systems Science, NUS

[Anusuya Manickavasagam, Muni Ranjan, Pradeep Kumar, Kesavan Sridharan]

[e0146864, e0146946, e0147017, e0146771] @u.nus.edu

Abstract

This work is part of assignment to find a suitable dataset to perform the PCA a popular dimension reduction technique, perform clustering technique followed by the regression analysis on reduced orthogonal variables and summarize prediction quality.

# Executive Summary

Our business problem is to look for a suitable dataset for PCA or Factor Analysis techniques. We will apply all the learnt techniques of data exploration and sanitization followed by running the PCA analysis to reduce the dimensions. Perform clustering and then regression techniques to predict some independent variable.

Our key findings are:

* On a high level, our dataset is about the soccer players data and their wages/overall value based on various features like Acceleration, Balance, Agility, dribbling, finishing, diving, kicking and many others.
* We found that 64 variables may be reduced as correlation matrix showed us features very similar to each other. Hence, we applied PCA and observed that dimension reduced from 64 variables to 6 variables which explains the 90% variance of the data.
* Overall principle components were profiled in to 6 components, namely: **Attacker, Defender, PlayerFieldSkills, StrengthAge, Valuation, Athletic.**
* We applied hierarchical clustering technique and found 4 clusters. Different numbers of clusters were created and test. Cluster with least overlapping using dendogram was selected. Their profiles were like PCA grouping. Groups were defined as below:
  1. Cluster 1: Attacker
  2. Cluster 2: Valued Players based on Age & Strength
  3. Cluster 3: Valued Athlete based on Age & Strength
  4. Cluster 4: Field Skilled Defender

This grouping is quite important from business perspective. For example: If there is a business group with some budget in hand and want to form a Soccer team, they could make use of the clusters above to choose players and form a high performing team.

* Our linear regression output showed a stable model with R-square of **0.81** and

RMSE of about 3.

**Please note:**

Although not mandatory, but we ran the 5-fold cross validation using stepwise regression in Jmp to test our model. We used the stepwise regression with backward elimination approach to find the best performing model.

# Data Description

## Data source

Data was collected from Kaggle: <https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset/data> which originated from a website in China named: <https://sofifa.com/>

## Data Summary

It contained **75 variables** and approx. **17,981 observations**. Approximately half of the columns had values represented as string, though underlying structure of those columns were continuous. For example: a value of 93 was represented as ‘90+3’. We sanitized this data and perform arithmetic to this string expression using R and generated the final numerical value.

After filtering out some nominal columns, we considered only **65** variables for analysis. The details of those variables are as below.

**Table 1:** Variable Description

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S. No. | Name | Type | Description | Data Type |
| **1** | **Overall** | **Response** | **Overall Point** | **Numerical** |
| 2 | Value | Explanatory | Self-Explanatory | Numerical |
| 3 | Wage | Explanatory | Self-Explanatory | Numerical |
| 4 | Special | Explanatory | Self-Explanatory | Numerical |
| 5 | Acceleration | Explanatory | Self-Explanatory | Numerical |
| 6 | Aggression | Explanatory | Self-Explanatory | Numerical |
| 7 | Agility | Explanatory | Self-Explanatory | Numerical |
| 8 | Balance | Explanatory | Self-Explanatory | Numerical |
| 9 | Ball.control | Explanatory | Self-Explanatory | Numerical |
| 10 | Composure | Explanatory | Self-Explanatory | Numerical |
| 11 | Crossing | Explanatory | Self-Explanatory | Numerical |
| 12 | Curve | Explanatory | Self-Explanatory | Numerical |
| 13 | Dribbling | Explanatory | Self-Explanatory | Numerical |
| 14 | Finishing | Explanatory | Self-Explanatory | Numerical |
| 15 | Free.kick.accuracy | Explanatory | Self-Explanatory | Numerical |
| 16 | GK.diving | Explanatory | Self-Explanatory | Numerical |
| 17 | GK.handling | Explanatory | Self-Explanatory | Numerical |
| 18 | GK.kicking | Explanatory | Self-Explanatory | Numerical |
| 19 | GK.positioning | Explanatory | Self-Explanatory | Numerical |
| 20 | GK.reflexes | Explanatory | Self-Explanatory | Numerical |
| 21 | Heading.accuracy | Explanatory | Self-Explanatory | Numerical |
| 22 | Interceptions | Explanatory | Self-Explanatory | Numerical |
| 23 | Jumping | Explanatory | Self-Explanatory | Numerical |
| 24 | Long.passing | Explanatory | Self-Explanatory | Numerical |
| 25 | Long.shots | Explanatory | Self-Explanatory | Numerical |
| 26 | Marking | Explanatory | Self-Explanatory | Numerical |
| 27 | Penalties | Explanatory | Self-Explanatory | Numerical |
| 28 | Positioning | Explanatory | Self-Explanatory | Numerical |
| 29 | Reactions | Explanatory | Self-Explanatory | Numerical |
| 30 | Short.passing | Explanatory | Self-Explanatory | Numerical |
| 31 | Shot.power | Explanatory | Self-Explanatory | Numerical |
| 32 | Sliding.tackle | Explanatory | Self-Explanatory | Numerical |
| 33 | Sprint.speed | Explanatory | Self-Explanatory | Numerical |
| 34 | Stamina | Explanatory | Self-Explanatory | Numerical |
| 35 | Standing.tackle | Explanatory | Self-Explanatory | Numerical |
| 36 | Strength | Explanatory | Self-Explanatory | Numerical |
| 37 | Vision | Explanatory | Self-Explanatory | Numerical |
| 38 | Volleys | Explanatory | Self-Explanatory | Numerical |
| 39 | CAM | Explanatory | Center Attacking Midfielder | Numerical |
| 40 | CB | Explanatory | Center Back | Numerical |
| 41 | CDM | Explanatory | Center Defensive Midfielder | Numerical |
| 42 | CF | Explanatory | Center Forward | Numerical |
| 43 | CM | Explanatory | Center Midfielder | Numerical |
| 45 | LAM | Explanatory | Left Attacking Midfielder | Numerical |
| 46 | LB | Explanatory | Left Back | Numerical |
| 47 | LCB | Explanatory | Left Center Back | Numerical |
| 48 | LCM | Explanatory | Left Center Midfielder | Numerical |
| 49 | LDM | Explanatory | Left Defensive Midfielder | Numerical |
| 50 | LF | Explanatory | Left Forward | Numerical |
| 51 | LM | Explanatory | Left Midfielder | Numerical |
| 52 | LS | Explanatory | Left Stricker | Numerical |
| 53 | LW | Explanatory | Left Wing | Numerical |
| 54 | LWB | Explanatory | Left Wing Back | Numerical |
| 55 | RAM | Explanatory | Right Attacking Midfielder | Numerical |
| 56 | RB | Explanatory | Right Back | Numerical |
| 57 | RCB | Explanatory | Right Center Back | Numerical |
| 58 | RCM | Explanatory | Right Center Midfielder | Numerical |
| 59 | RDM | Explanatory | Right Defensive Midfielder | Numerical |
| 60 | RF | Explanatory | Right Forward | Numerical |
| 61 | RM | Explanatory | Right Midfielder | Numerical |
| 62 | RS | Explanatory | Right Stricker | Numerical |
| 63 | RW | Explanatory | Right Wing | Numerical |
| 64 | RWB | Explanatory | Right Wing Back | Numerical |
| 65 | ST | Explanatory | Striker | Numerical |

## Business Statement

Based on the dataset and combining the assignment objectives, a hypothetical business statement is formed, as below:

*“A new Football franchisee would like to form a team. They have data for approx. 18k players worldwide. Overall, 64 features are available like age, strength, playing style etc. Franchisee would like to define 5-6 features which incorporates all the variables. Also, they would like to apply cluster techniques to create a pool of homogenous player together to form a diverse team.*

*At the end, they would like to create a regression model which could predict overall points for a new player given the available attributes.”*

# Data Selection and Preprocessing

We followed the following process to analyze the dataset:

**Figure 1**: Process to Explore the data

## Tools Used

* **R** – Used for Data exploration, imputation using kNN
* **JMP** – PCA, Regression and Clustering
* **Tableau –** For cluster profiling

## Dealing with missing values

* 1. **Filling up missing Values in Numerical variables with kNN imputation in R**: To proceed further, all the missing values were replaced with imputed values using k-Nearest Neighbour Imputation based on a variation of the Gower Distance for numerical, categorical, ordered and semi-continuous variables. It generates multiple imputations for incomplete multivariate data. A small piece of code has been written in R to implement imputation on numerical data.

## 3.3 Correlation Check Of Numerical Values

The second step is to analyze the **multi-collinearity effects** in between the numerical variables. Data analysis tool in Excel has been used to compute the correlation matrix using **Pearson’s**

**coefficient** and for better interpretation, the results have been compiled in a correlation chart. Please find below the correlation matrix among all numeric variables:



**Table 2:** Correlation Matrix

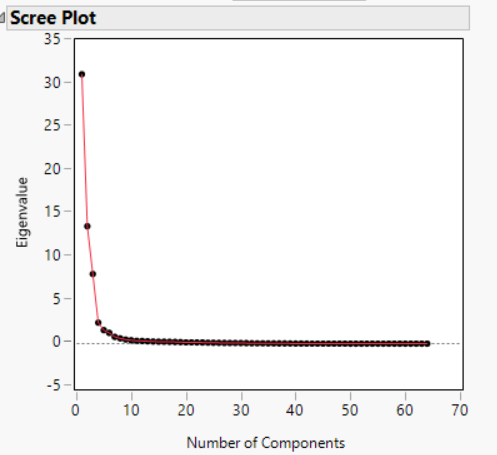
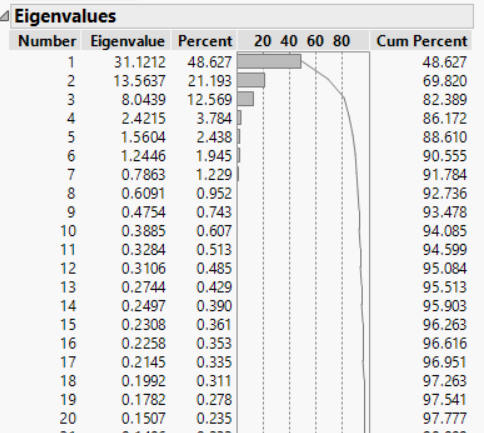
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Value | | Wage | Special | Acceleration | Aggression | Agility | Balance | Ball.control |
| Value | | 1.00 | 0.85 | 0.38 | 0.18 | 0.19 | 0.20 | 0.11 | 0.31 |
| Wage | | 0.85 | 1.00 | 0.37 | 0.15 | 0.21 | 0.17 | 0.09 | 0.29 |
| Special | | 0.38 | 0.37 | 1.00 | 0.65 | 0.67 | 0.69 | 0.58 | 0.91 |
| Acceleration | | 0.18 | 0.15 | 0.65 | 1.00 | 0.25 | 0.80 | 0.70 | 0.67 |
| Aggression | | 0.19 | 0.21 | 0.67 | 0.25 | 1.00 | 0.24 | 0.18 | 0.54 |
| Agility | | 0.20 | 0.17 | 0.69 | 0.80 | 0.24 | 1.00 | 0.77 | 0.70 |
| Balance | | 0.11 | 0.09 | 0.58 | 0.70 | 0.18 | 0.77 | 1.00 | 0.60 |
| Ball.control | | 0.31 | 0.29 | 0.91 | 0.67 | 0.54 | 0.70 | 0.60 | 1.00 |
| Composure | | 0.40 | 0.39 | 0.80 | 0.44 | 0.58 | 0.49 | 0.37 | 0.76 |
| Crossing | | 0.25 | 0.24 | 0.86 | 0.66 | 0.47 | 0.69 | 0.62 | 0.84 |
| Curve | | 0.29 | 0.27 | 0.85 | 0.60 | 0.39 | 0.68 | 0.58 | 0.83 |
| Dribbling | | 0.27 | 0.25 | 0.86 | 0.74 | 0.42 | 0.76 | 0.66 | 0.93 |
| Finishing | | 0.26 | 0.23 | 0.71 | 0.60 | 0.23 | 0.63 | 0.51 | 0.79 |
| Free.kick.accuracy | | 0.27 | 0.25 | 0.81 | 0.49 | 0.40 | 0.58 | 0.51 | 0.77 |
| GK.diving | | -0.03 | -0.04 | -0.67 | -0.59 | -0.57 | -0.52 | -0.50 | -0.78 |
| GK.handling | | -0.03 | -0.03 | -0.67 | -0.59 | -0.57 | -0.52 | -0.51 | -0.78 |
| GK.kicking | | -0.03 | -0.04 | -0.67 | -0.58 | -0.57 | -0.52 | -0.50 | -0.78 |
| GK.positioning | | -0.03 | -0.04 | -0.67 | -0.58 | -0.56 | -0.52 | -0.50 | -0.78 |
| GK.reflexes | | -0.03 | -0.04 | -0.67 | -0.58 | -0.57 | -0.52 | -0.50 | -0.78 |
| Heading.accuracy | | 0.19 | 0.21 | 0.65 | 0.33 | 0.69 | 0.26 | 0.17 | 0.65 |
| Interceptions | | 0.14 | 0.16 | 0.57 | 0.15 | 0.74 | 0.13 | 0.15 | 0.40 |
| Jumping | | 0.14 | 0.15 | 0.31 | 0.21 | 0.36 | 0.21 | 0.18 | 0.18 |
| Long.passing | | 0.30 | 0.29 | 0.85 | 0.43 | 0.58 | 0.52 | 0.46 | 0.78 |
| Long.shots | | 0.28 | 0.26 | 0.83 | 0.57 | 0.39 | 0.64 | 0.52 | 0.83 |
| Marking | | 0.08 | 0.11 | 0.51 | 0.14 | 0.72 | 0.09 | 0.12 | 0.36 |
| Penalties | | 0.24 | 0.24 | 0.73 | 0.53 | 0.33 | 0.56 | 0.48 | 0.77 |
| Positioning | | 0.26 | 0.24 | 0.81 | 0.67 | 0.38 | 0.70 | 0.59 | 0.86 |
| Reactions | | 0.53 | 0.50 | 0.59 | 0.19 | 0.40 | 0.28 | 0.14 | 0.43 |
| Short.passing | | 0.32 | 0.31 | 0.90 | 0.56 | 0.60 | 0.61 | 0.54 | 0.91 |
| Shot.power | | 0.28 | 0.27 | 0.83 | 0.54 | 0.49 | 0.57 | 0.45 | 0.83 |
| Sliding.tackle | | 0.08 | 0.12 | 0.51 | 0.16 | 0.71 | 0.11 | 0.15 | 0.37 |
| Sprint.speed | | 0.18 | 0.15 | 0.65 | 0.92 | 0.28 | 0.75 | 0.64 | 0.66 |
| Stamina | | 0.21 | 0.20 | 0.79 | 0.61 | 0.64 | 0.56 | 0.47 | 0.72 |
| Standing.tackle | | 0.10 | 0.13 | 0.54 | 0.15 | 0.73 | 0.11 | 0.14 | 0.40 |
| Strength | | 0.14 | 0.17 | 0.19 | -0.16 | 0.46 | -0.24 | -0.40 | 0.08 |
| Vision | | 0.35 | 0.32 | 0.75 | 0.46 | 0.30 | 0.59 | 0.49 | 0.72 |
| Volleys | | 0.29 | 0.27 | 0.76 | 0.57 | 0.32 | 0.62 | 0.51 | 0.79 |
| CAM | | 0.45 | 0.41 | 0.73 | 0.49 | 0.18 | 0.63 | 0.48 | 0.69 |
| CB | | 0.19 | 0.22 | 0.13 | -0.32 | 0.44 | -0.30 | -0.31 | -0.15 |
| CDM | | 0.31 | 0.32 | 0.34 | -0.17 | 0.46 | -0.08 | -0.12 | 0.07 |
| CF | | 0.44 | 0.40 | 0.75 | 0.55 | 0.20 | 0.65 | 0.50 | 0.73 |
| CM | | 0.49 | 0.46 | 0.75 | 0.31 | 0.35 | 0.47 | 0.34 | 0.61 |
| ID | | -0.14 | -0.21 | -0.23 | 0.12 | -0.23 | -0.02 | 0.05 | -0.11 |
| LAM | | 0.44 | 0.41 | 0.77 | 0.53 | 0.22 | 0.66 | 0.51 | 0.73 |
| LB | | 0.29 | 0.30 | 0.37 | -0.03 | 0.45 | -0.01 | -0.03 | 0.08 |
| LCB | | 0.19 | 0.22 | 0.13 | -0.31 | 0.45 | -0.30 | -0.30 | -0.14 |
| LCM | | 0.48 | 0.46 | 0.77 | 0.33 | 0.37 | 0.48 | 0.36 | 0.63 |
| LDM | | 0.31 | 0.32 | 0.37 | -0.14 | 0.48 | -0.06 | -0.10 | 0.10 |
| LF | | 0.43 | 0.40 | 0.78 | 0.58 | 0.23 | 0.67 | 0.52 | 0.76 |
| LM | | 0.44 | 0.40 | 0.80 | 0.61 | 0.25 | 0.70 | 0.55 | 0.75 |
| LS | | 0.44 | 0.41 | 0.79 | 0.54 | 0.29 | 0.61 | 0.43 | 0.76 |
| LW | | 0.41 | 0.38 | 0.79 | 0.64 | 0.22 | 0.73 | 0.58 | 0.78 |
| LWB | | 0.35 | 0.35 | 0.52 | 0.10 | 0.48 | 0.15 | 0.10 | 0.24 |
| RAM | | 0.44 | 0.40 | 0.80 | 0.56 | 0.25 | 0.68 | 0.53 | 0.77 |
| RB | | 0.29 | 0.30 | 0.39 | -0.01 | 0.47 | 0.01 | -0.02 | 0.11 |
| RCB | | 0.19 | 0.22 | 0.14 | -0.31 | 0.46 | -0.29 | -0.30 | -0.14 |
| RCM | | 0.48 | 0.46 | 0.79 | 0.35 | 0.39 | 0.50 | 0.38 | 0.66 |
| RDM | | 0.31 | 0.32 | 0.38 | -0.13 | 0.49 | -0.05 | -0.09 | 0.11 |
| RF | | 0.43 | 0.39 | 0.80 | 0.60 | 0.25 | 0.69 | 0.53 | 0.78 |
| RM | | 0.43 | 0.40 | 0.82 | 0.63 | 0.27 | 0.71 | 0.57 | 0.77 |
| RS | | 0.43 | 0.41 | 0.80 | 0.55 | 0.30 | 0.62 | 0.44 | 0.77 |
| RW | | 0.41 | 0.37 | 0.81 | 0.65 | 0.24 | 0.74 | 0.59 | 0.79 |
| RWB | | 0.35 | 0.35 | 0.53 | 0.12 | 0.49 | 0.16 | 0.11 | 0.26 |
| ST | | 0.43 | 0.41 | 0.80 | 0.56 | 0.30 | 0.62 | 0.44 | 0.78 |

The results of **correlation matrix** show that **most of the variables are correlated with each other**. That is why out of 65 variables, less than 10 variables explain more than 90% of variance.

# Dimension Reduction Using PCA

JMP was used to perform PCA. After the data was processed and run for the dimension reduction, scree plot was generated as below in Figure (2). Based on this, **6 factors appear** good to be extracted.

**Figure 2:** Scree Plot **Figure 3:** Eigenvalues Cum Percent

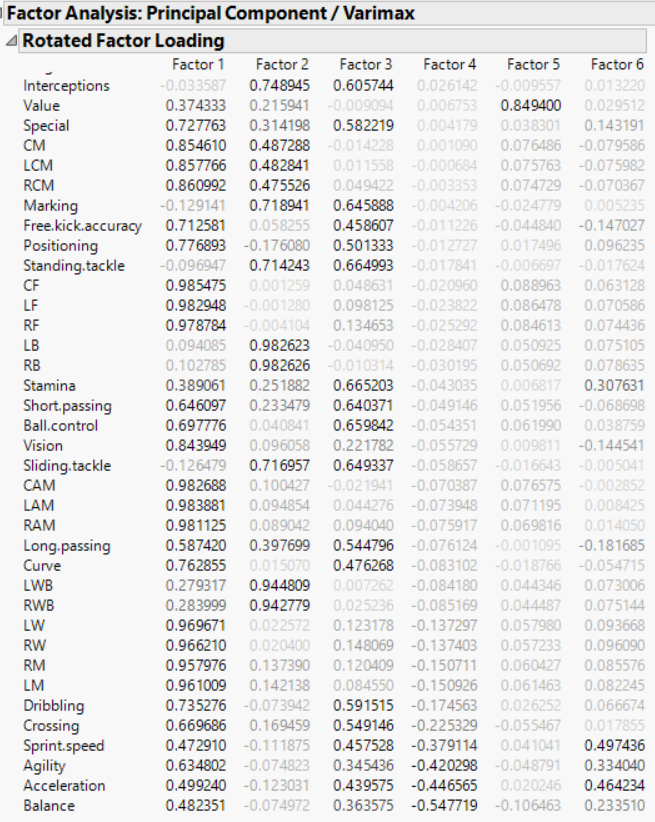
Furthermore, the cumulative variance was observed using Eignvalues table shown in Figure (3) above. **90%** of the variance is explained by these variables.

Upon analyzing the Varimax Rotated factor loading as shown in Figure (4) below, we extracted following profiles:

**Table (3):** Profiling based on Principal Components

|  |  |
| --- | --- |
| Profiles |  |
| 1 | Attacker |
| 2 | Defender |
| 3 | PlayerFieldSkills |
| 4 | StrengthAge |
| 5 | Valuation |
| 6 | Athletic |

**Figure 4:** Rotated Factor Loading



# Clusters

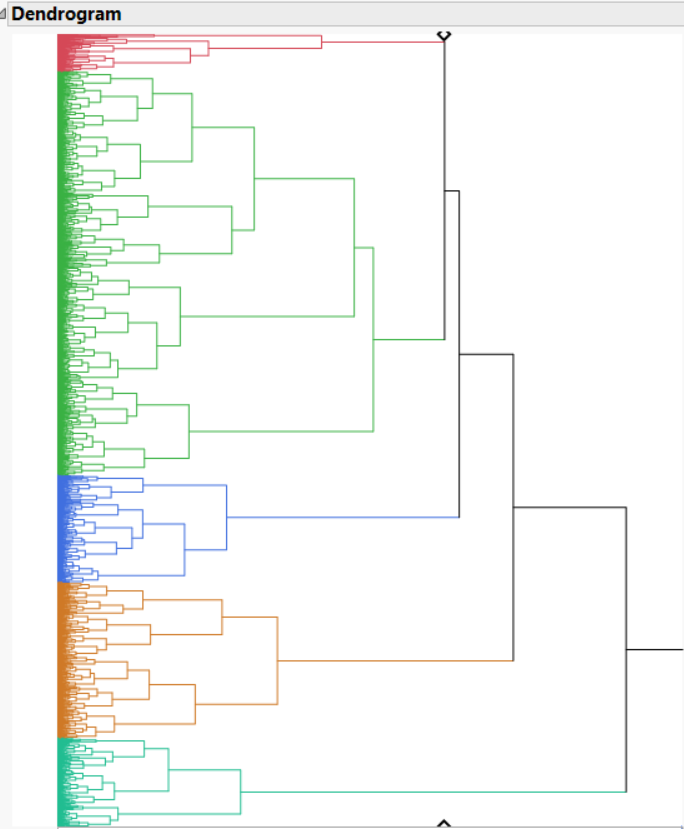
We used the Hierarchical clustering technique on the reduced dimensions identified in section (4) above. Upon analyzing the Dendogram diagram, we found that 4 clusters were found to be suitable for the dataset.

**Table (4): Clusters based on the hierarchical clustering technique**

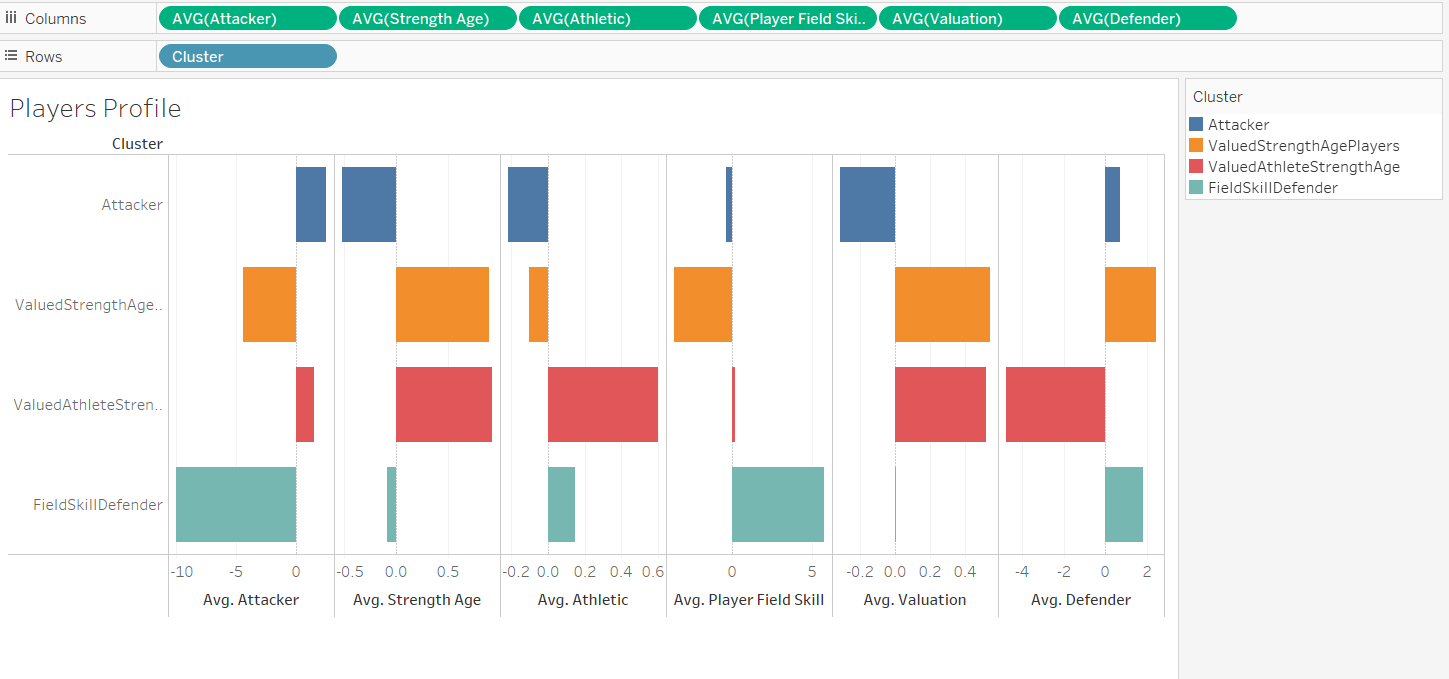
|  |  |
| --- | --- |
| Clusters |  |
| 1 | Attacker |
| 2 | Valued Strength Age Players |
| 3 | Valued Athlete Strength Age |
| 4 | Field Skill Defender |

We then pushed that data to **Tableau** to map those clusters to the different principal components we identified. Figure (5) and (6) shows the dendogram and clusters extracted. The cluster extraction method helps define the franchisee a pool of players based on different skills. They could further analyze the clusters to choose the player.

**Figure (5):** Dendogram



**Figure (6):** Clusters Vs Principal components

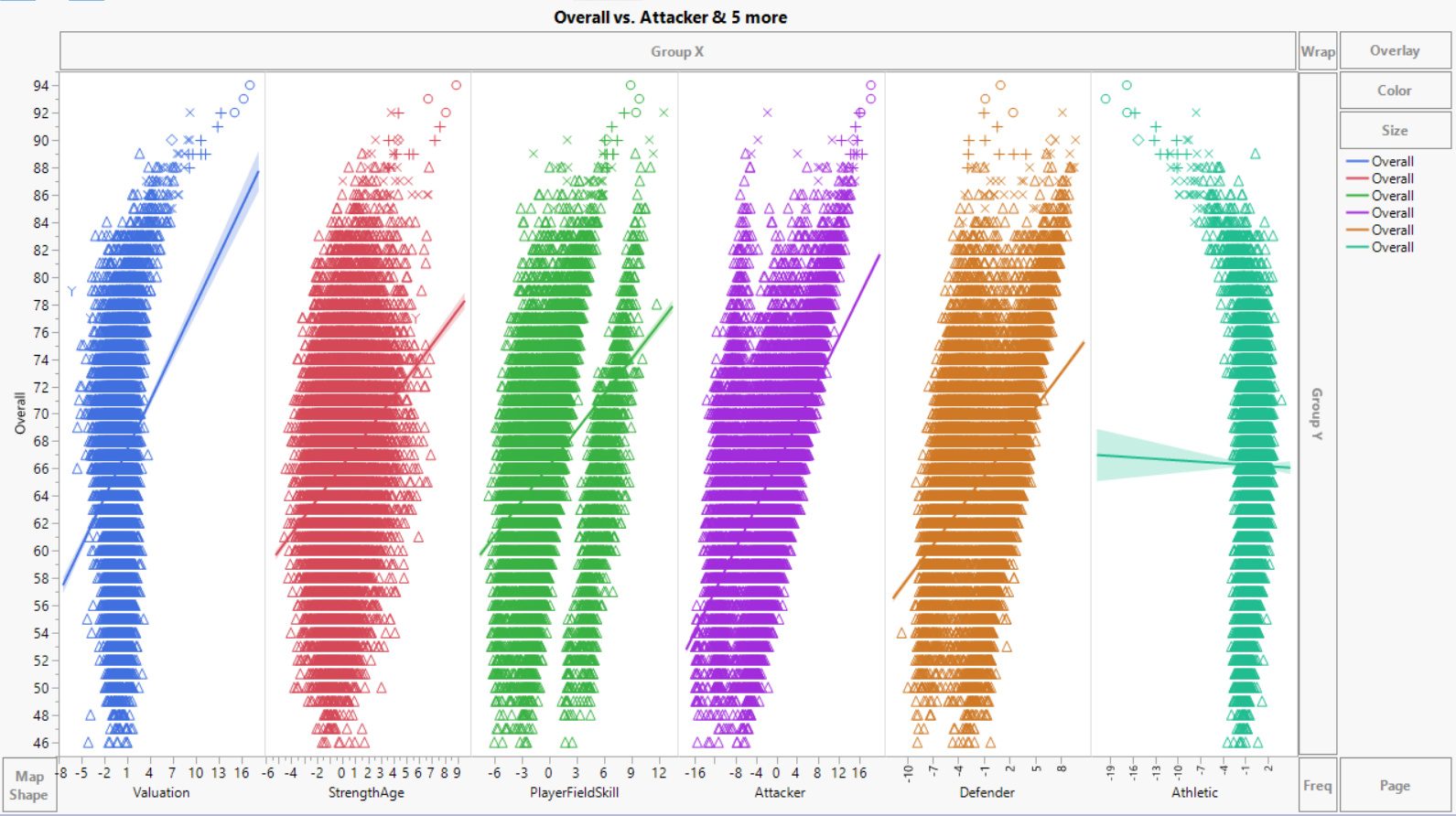


# Regression Model Analysis

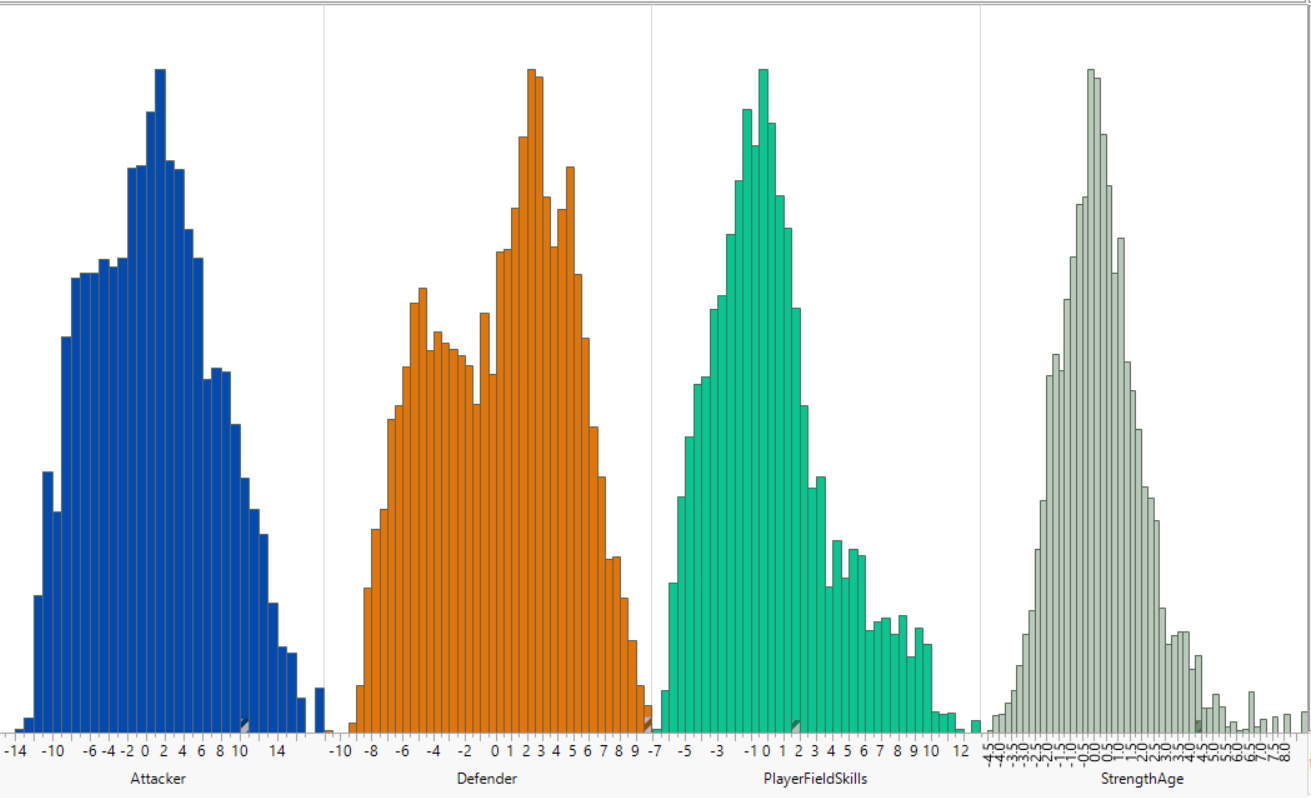
Before running the regression model, we checked for major assumptions for a linear regression model:

* Scatter-plot analysis showed little linear relation between dependent variables with independent variable as shown in Figure (7) below.
* Outlier presence was not detected
* Histogram showed that variables to be multivariate normal as shown in Figure (8) below.
* Residual scatter plot found to be homoscedastic as shown in Figure (9) below.

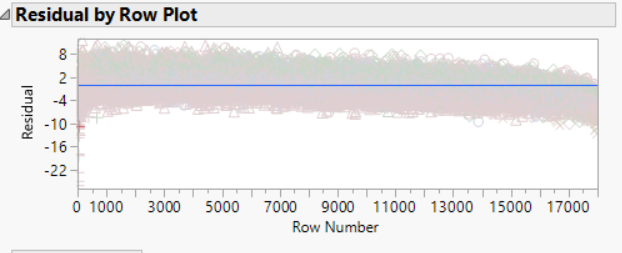
**Figure (7)**: Scatter-plot Y v/s X’s



**Figure (8):** Normal distribution for Multivariate



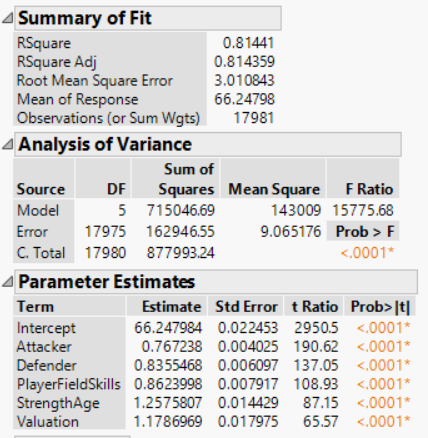
**Figure (9):** Residual plot



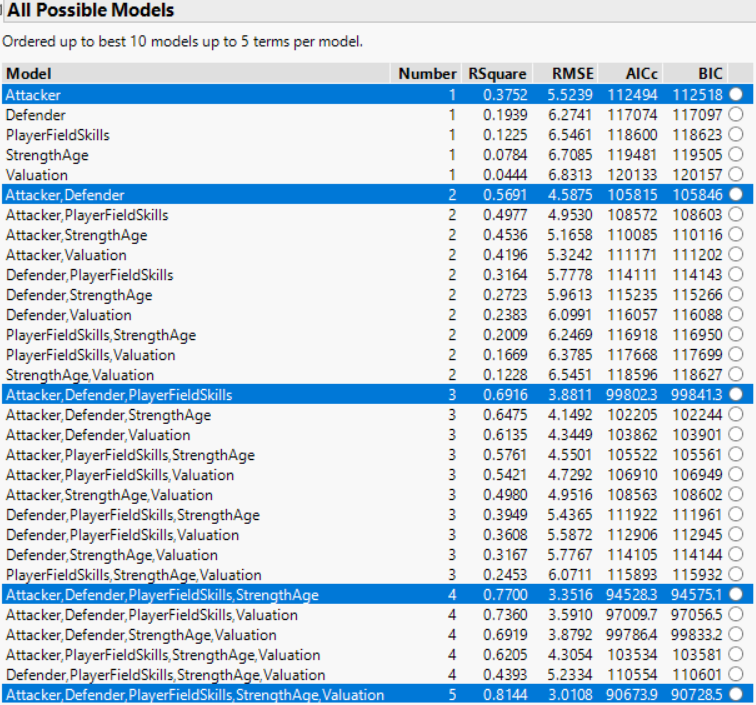
## Model Fit & Parameter Estimates

Figure (10) below shows the summary of fit and parameter estimates. Using the stepwise regression and 5-fold cross validation, we created and validated the model. Overall, though RSquare is showing a stable model prediction but overall Prediction error is high. This is expected as the data was less linear with the response variables. Figure (11) below shows the possible backward elimination performed.

**Figure (10):** Summary of Fit



**Figure (11):** All Possible Models using Stepwise



# References

* Lecture Notes
* Online Learning resources

# Files Description:

|  |  |
| --- | --- |
| File Name | Description |
| PCA\_FIFA.R | Feature Processing and PCA (1) |
| Principal Components.jrp | PCA in Jmp (used for report) |
| Hierarchical Cluster.jrp | Hierarchical clusters output in Jmp |
| Fit Model.jrp | Regression output in Jmp |
| FIFA\_NUM\_imputed.Rda | Imputed data for R for caching |
| FIFA\_CompleteDataset.csv | Dataset csv format |
| CompleteDataset\_Cluster.xlsx | Dataset xls format with cluster saved |
| CompleteDataset.xlsx | Complete Dataset xls format |
| CompleteDataset.jmp | Jmp loading dataset |
| Cluster\_Tableau.twb | Tableau for the cluster charts |