Soccer_AssociationRule

Henry

The Business Problem and Approach

Library needed packages

```
library(dplyr)
library(magrittr)
library(tidyr)
library(arules)
library(ggplot2)
```

Load in the needed table

```
setwd("C:\\Users\\fengy\\Desktop\\Fall\\3. EDA\\HW 2\\HW 2")
match <- read.csv('match.csv')
player = read.csv('player.csv')
p_attributes = read.csv('player_attribute.csv')
team = read.csv('team.csv')
t_attributes = read.csv('team_attributes.csv')
```

Player Analysis

Description and Rationale for the chosen analysis

For team Roma, the biggest asset is the player. And player is also one of the big sucess factor for the winning game.

In this section of analyses, we will first plot the players frequency of participating the winning games, both home match and away match. From this discriptive analyses, we might give the coach an overview of specifc player with winning games.

Further, in order to locate the players who associate with the winning game. We perform association rules to find the association relationship between player and winning/losing game. We also segment the data into home match and away match. Our assumption here is that the association rule between players and players with higher lift might lead to more possibility of success if the coach arrange certain player together.

Execution and result

Munging the data at first We try to munge the data to prepare for the association rule matrix. The dataframe in the end contains columns match_id, player, player_api_id and player name.

```
# Locate the roma info
long_team_name <- 'Roma'
roma_record <- team %>%
  collect() %>%
  filter(grepl(long_team_name, team_long_name))

# Get the Roma_home_team_matches
home_matches <- filter(match, home_team_api_id == roma_record$team_api_id)

# match id & goal diff
match_outcomes_per_match <- match %>%
  mutate(goal_diff =home_team_goal - away_team_goal) %>%
  select(id, goal_diff)

colnames(match_outcomes_per_match) <- c('match_id', 'goal_diff')

# match/ player position/ player_api_id
roma_players_per_match <- select(home_matches, id, matches("home_player_[[:digit:]]")) %>%
  collect() %>%
  gather(player, player_api_id, -id)

colnames(roma_players_per_match) <- c('match_id', 'player','player_api_id')

# create table : match_id, player, player_api_id & player name
roma_player_id <- roma_players_per_match %>% distinct(player_api_id)
roma_player_info <- merge(roma_player_id, player, by = 'player_api_id')
roma_player_info2 <- roma_player_info %>% select(player_api_id, player_name)
roma_player_per_match2 <- left_join(roma_players_per_match, roma_player_info2, by = 'player_a
pi_id')
head(roma_player_per_match2)
```

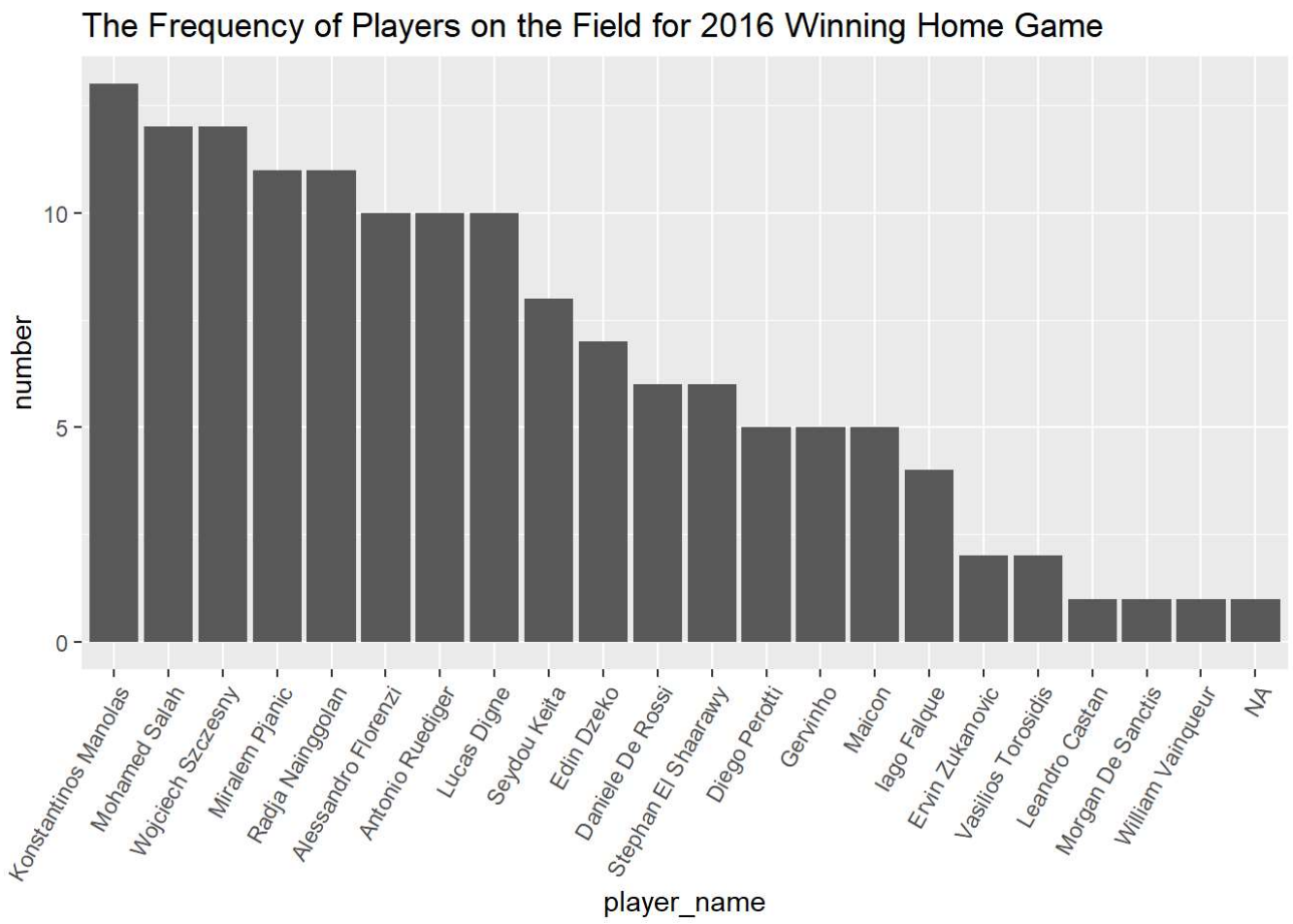
```
##   match_id        player player_api_id player_name
## 1    10263 home_player_1         39351        Doni
## 2    10288 home_player_1         39351        Doni
## 3    10312 home_player_1         39351        Doni
## 4    10331 home_player_1         19344       Artur
## 5    10349 home_player_1         39351        Doni
## 6    10387 home_player_1         39351        Doni
```

Use simple EDA to plot the distribution of 2015/2016 season winning and away game player

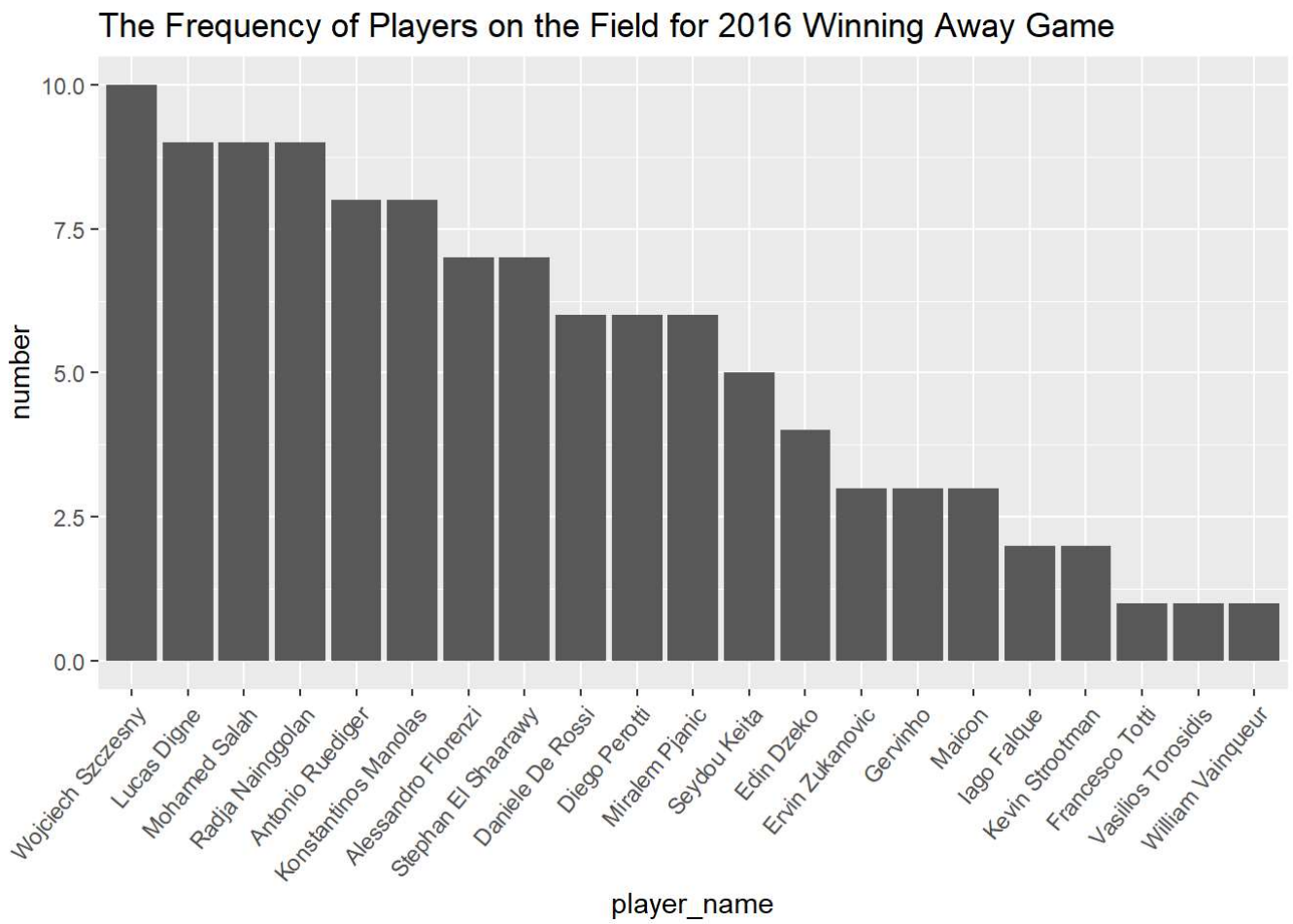
```
win16_player_count <- win_16_playerlist %>%
  group_by(player_name) %>%
  summarise(number = n()) %>%
  arrange(desc(number))

win16_player_count <- transform(win16_player_count, player_name = reorder(player_name, -numbe
r))

ggplot(win16_player_count, aes(x=player_name, y =number)) +
  geom_bar(stat='identity')+
  ggtitle('The Frequency of Players on the Field for 2016 Winning Home Game')+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
ggplot(win16aw_player_count, aes(x=player_name, y =number)) +
  geom_bar(stat='identity')+
  ggtitle('The Frequency of Players on the Field for 2016 Winning Away Game')+
  theme(axis.text.x = element_text(angle = 50, hjust = 1))
```



Interpretation From the two graphs, we are able to find the slight difference of players who are assigned to home and away match. In the winning home match, Manolas (center back) leads the most frequency. And in the distribution of away game, Szczensny (goal keeper) had the most appearance. And Manolas became the player with 6th highest frequency.

We can have the evidence that the formation of team players in home and away game is slightly different. Goal keeper was in the top three list, but players like Digne (Left back) (away match 2nd higher, home match 8th higher) and Pjanic (midfielder) (home match 4th highest and away match 12th highest).

Furthermore, based on the findings, we would continue with the division of home and away team. We will further use association rule on these two segmentation.

Association Rule for Player of winning/losing game at home matches

```
win_rule = sort(win_rule, by = 'lift', decreasing = TRUE)
inspect(head(win_rule,20))
```

```
##      lhs                          rhs                        support confidence     lift count
## [1]  {Wojciech Szczesny}       => {Mohamed Salah}          0.1060606  0.8750000 7.218750    14
## [2]  {Mohamed Salah}           => {Wojciech Szczesny}      0.1060606  0.8750000 7.218750    14
## [3]  {Konstantinos Manolas,
##       Wojciech Szczesny}       => {Mohamed Salah}          0.1060606  0.8750000 7.218750    14
## [4]  {Konstantinos Manolas,
##       Mohamed Salah}           => {Wojciech Szczesny}      0.1060606  0.8750000 7.218750    14
## [5]  {Alessandro Florenzi,
##       Konstantinos Manolas}    => {Wojciech Szczesny}      0.1060606  0.6666667 5.500000    14
## [6]  {Maarten Stekelenburg}    => {Erik Lamela}            0.1287879  0.8947368 5.135011    17
## [7]  {Erik Lamela}             => {Maarten Stekelenburg}   0.1287879  0.7391304 5.135011    17
## [8]  {Erik Lamela,
##       Francesco Totti}         => {Maarten Stekelenburg}   0.1060606  0.7368421 5.119114    14
## [9]  {Francesco Totti,
##       Maarten Stekelenburg}    => {Erik Lamela}            0.1060606  0.8750000 5.021739    14
## [10] {Erik Lamela}             => {Pablo Osvaldo}          0.1060606  0.6086957 4.228833    14
## [11] {Pablo Osvaldo}           => {Erik Lamela}            0.1060606  0.7368421 4.228833    14
## [12] {Konstantinos Manolas,
##       Miralem Pjanic}          => {Wojciech Szczesny}      0.1060606  0.5000000 4.125000    14
## [13] {Konstantinos Manolas,
##       Miralem Pjanic}          => {Mohamed Salah}          0.1060606  0.5000000 4.125000    14
## [14] {Antonio Ruediger}        => {Konstantinos Manolas}   0.1060606  1.0000000 4.000000    14
## [15] {Lucas Digne}             => {Konstantinos Manolas}   0.1060606  1.0000000 4.000000    14
## [16] {Wojciech Szczesny}       => {Konstantinos Manolas}   0.1212121  1.0000000 4.000000    16
## [17] {Mohamed Salah}           => {Konstantinos Manolas}   0.1212121  1.0000000 4.000000    16
## [18] {Seydou Keita}            => {Konstantinos Manolas}   0.1363636  1.0000000 4.000000    18
## [19] {Konstantinos Manolas}    => {Seydou Keita}           0.1363636  0.5454545 4.000000    18
## [20] {Mohamed Salah,
##       Wojciech Szczesny}       => {Konstantinos Manolas}   0.1060606  1.0000000 4.000000    14
```

Interpretation From the association rule, we collated all the data from the dataset, and subset it with roma home winning match. We found Mohamed Salah (forward) and Wojciech Szcaesny (goal keeper) were a good match in the previous matches record. And also, Konstantinos Manolas, Mohamed Salah, and Wojciech Szaesny were another good combination.

```
lose_rule = sort(lose_rule, by = 'lift', decreasing = TRUE)
inspect(head(lose_rule,20))
```

```
##      lhs                       rhs                     support confidence lift count
## [1]  {Gervinho}             => {Vasilios Torosidis} 0.1052632          1  9.5     2
## [2]  {Vasilios Torosidis}   => {Gervinho}           0.1052632          1  9.5     2
## [3]  {Juan Manuel Iturbe}   => {Morgan De Sanctis}  0.1052632          1  9.5     2
## [4]  {Morgan De Sanctis}    => {Juan Manuel Iturbe} 0.1052632          1  9.5     2
## [5]  {Gervinho,
##       Miralem Pjanic}       => {Vasilios Torosidis} 0.1052632          1  9.5     2
## [6]  {Miralem Pjanic,
##       Vasilios Torosidis}   => {Gervinho}           0.1052632          1  9.5     2
## [7]  {Francesco Totti,
##       Gervinho}             => {Vasilios Torosidis} 0.1052632          1  9.5     2
## [8]  {Francesco Totti,
##       Vasilios Torosidis}   => {Gervinho}           0.1052632          1  9.5     2
## [9]  {Alessandro Florenzi,
##       Juan Manuel Iturbe}   => {Morgan De Sanctis}  0.1052632          1  9.5     2
## [10] {Alessandro Florenzi,
##       Morgan De Sanctis}    => {Juan Manuel Iturbe} 0.1052632          1  9.5     2
## [11] {Juan Manuel Iturbe,
##       Miralem Pjanic}       => {Morgan De Sanctis}  0.1052632          1  9.5     2
## [12] {Miralem Pjanic,
##       Morgan De Sanctis}    => {Juan Manuel Iturbe} 0.1052632          1  9.5     2
## [13] {Francesco Totti,
##       Jose Angel}           => {Gabriel Heinze}     0.1052632          1  9.5     2
## [14] {Francesco Totti,
##       Gervinho,
##       Miralem Pjanic}       => {Vasilios Torosidis} 0.1052632          1  9.5     2
## [15] {Francesco Totti,
##       Miralem Pjanic,
##       Vasilios Torosidis}   => {Gervinho}           0.1052632          1  9.5     2
## [16] {Alessandro Florenzi,
##       Juan Manuel Iturbe,
##       Miralem Pjanic}       => {Morgan De Sanctis}  0.1052632          1  9.5     2
## [17] {Alessandro Florenzi,
##       Miralem Pjanic,
##       Morgan De Sanctis}    => {Juan Manuel Iturbe} 0.1052632          1  9.5     2
## [18] {Francesco Totti,
##       Jose Angel,
##       Pablo Osvaldo}        => {Gabriel Heinze}     0.1052632          1  9.5     2
## [19] {Francesco Totti,
##       Jose Angel,
##       Miralem Pjanic}       => {Gabriel Heinze}     0.1052632          1  9.5     2
## [20] {Daniele De Rossi,
##       Francesco Totti,
##       Jose Angel}           => {Gabriel Heinze}     0.1052632          1  9.5     2
```

Interpretation For the losing rule, the result showed that the combination of frequent players were are in the losing matches. We can't tell the significance from the table since the support, confidence, lift and count are almost equal with the column. It might provide the info that these twenty combination might not be a strong pairs for a winning game.

Association Rule for Player of winning/losing game at home matches

```
win_rule_aw = sort(win_rule_aw, by = 'lift', decreasing = TRUE)
inspect(head(win_rule_aw,20))
```

```
##      lhs                       rhs                    support confidence     lift count
## [1]  {Antonio Ruediger,
##       Konstantinos Manolas} => {Lucas Digne}       0.1067961  1.0000000 7.357143    11
## [2]  {Antonio Ruediger,
##       Konstantinos Manolas,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1067961  1.0000000 7.357143    11
## [3]  {Konstantinos Manolas,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1262136  0.9285714 6.831633    13
## [4]  {Antonio Ruediger}     => {Lucas Digne}       0.1165049  0.9230769 6.791209    12
## [5]  {Antonio Ruediger,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1165049  0.9230769 6.791209    12
## [6]  {Alessandro Florenzi,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1165049  0.9230769 6.791209    12
## [7]  {Alessandro Florenzi,
##       Konstantinos Manolas,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1165049  0.9230769 6.791209    12
## [8]  {Lucas Digne}          => {Antonio Ruediger}  0.1165049  0.8571429 6.791209    12
## [9]  {Lucas Digne,
##       Wojciech Szczesny}    => {Antonio Ruediger}  0.1165049  0.8571429 6.791209    12
## [10] {Radja Nainggolan,
##       Wojciech Szczesny}    => {Mohamed Salah}     0.1165049  0.8571429 6.791209    12
## [11] {Konstantinos Manolas,
##       Radja Nainggolan,
##       Wojciech Szczesny}    => {Lucas Digne}       0.1067961  0.9166667 6.744048    11
## [12] {Konstantinos Manolas,
##       Lucas Digne}          => {Antonio Ruediger}  0.1067961  0.8461538 6.704142    11
## [13] {Konstantinos Manolas,
##       Lucas Digne,
##       Wojciech Szczesny}    => {Antonio Ruediger}  0.1067961  0.8461538 6.704142    11
## [14] {Antonio Ruediger}     => {Wojciech Szczesny} 0.1262136  1.0000000 6.437500    13
## [15] {Wojciech Szczesny}    => {Antonio Ruediger}  0.1262136  0.8125000 6.437500    13
## [16] {Mohamed Salah}        => {Wojciech Szczesny} 0.1262136  1.0000000 6.437500    13
## [17] {Wojciech Szczesny}    => {Mohamed Salah}     0.1262136  0.8125000 6.437500    13
## [18] {Lucas Digne}          => {Wojciech Szczesny} 0.1359223  1.0000000 6.437500    14
## [19] {Wojciech Szczesny}    => {Lucas Digne}       0.1359223  0.8750000 6.437500    14
## [20] {Antonio Ruediger,
##       Lucas Digne}          => {Wojciech Szczesny} 0.1165049  1.0000000 6.437500    12
```

Interpretation

From the away winning matches, we can observe that Lucas Digne is frequently appearing in the winning matches with high lift, which indicated that most of the winning away matches, he might play a influencial role. Lucas Dinge was a loaned player for Roma in the season 2015/2016. We recommend that Roma should search for the player equipped with the similar player attribute with Lucas Dinge. The cooperation between the potential player and the original players might improve the winning rate of away matches.

```
lose_rule_aw = sort(lose_rule_aw, by = 'lift', decreasing = TRUE)
inspect(head(lose_rule_aw,20))
```

```
##      lhs                       rhs                 support confidence     lift count
## [1]  {Marco Borriello,
##       Simone Perrotta}      => {Julio Sergio}    0.1020408  1.0000000 8.166667     5
## [2]  {Daniele De Rossi,
##       Fernando Gago,
##       Pablo Osvaldo}        => {Simon Kjaer}     0.1020408  1.0000000 8.166667     5
## [3]  {Daniele De Rossi,
##       Fernando Gago,
##       Maarten Stekelenburg,
##       Pablo Osvaldo}        => {Simon Kjaer}     0.1020408  1.0000000 8.166667     5
## [4]  {Michael Bradley}      => {Ivan Piris}      0.1020408  0.8333333 6.805556     5
## [5]  {Ivan Piris}           => {Michael Bradley} 0.1020408  0.8333333 6.805556     5
## [6]  {Christian Panucci,
##       John Arne Riise}      => {Julio Baptista}  0.1020408  0.8333333 6.805556     5
## [7]  {Jeremy Menez,
##       Marco Borriello}      => {Julio Sergio}    0.1020408  0.8333333 6.805556     5
## [8]  {Fernando Gago,
##       Jose Angel}           => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [9]  {Aleandro Rosi,
##       Fernando Gago}        => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [10] {Fernando Gago,
##       Pablo Osvaldo}        => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [11] {Christian Panucci,
##       Daniele De Rossi,
##       John Arne Riise}      => {Julio Baptista}  0.1020408  0.8333333 6.805556     5
## [12] {Jeremy Menez,
##       Marco Borriello,
##       Marco Cassetti}       => {Julio Sergio}    0.1020408  0.8333333 6.805556     5
## [13] {Fernando Gago,
##       Jose Angel,
##       Maarten Stekelenburg} => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [14] {Aleandro Rosi,
##       Fernando Gago,
##       Maarten Stekelenburg} => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [15] {Fernando Gago,
##       Maarten Stekelenburg,
##       Pablo Osvaldo}        => {Simon Kjaer}     0.1020408  0.8333333 6.805556     5
## [16] {Julio Sergio}         => {Marco Borriello} 0.1224490  1.0000000 6.125000     6
## [17] {Marco Borriello}      => {Julio Sergio}    0.1224490  0.7500000 6.125000     6
## [18] {Jeremy Menez,
##       Julio Sergio}         => {Marco Borriello} 0.1020408  1.0000000 6.125000     5
## [19] {Julio Sergio,
##       Marco Cassetti}       => {Marco Borriello} 0.1020408  1.0000000 6.125000     5
## [20] {Julio Sergio,
##       Simone Perrotta}      => {Marco Borriello} 0.1020408  1.0000000 6.125000     5
```

Interpretation

From the assoication rule of losing away matches, the combinations of Marco Borriello (foward), Simon Perrotta (Central) and Julio Sergio (goal keeper) and Rossi (midfielder), Gago (midfielder) and Osvaldo(striker) have high lift. We would suggest the coach that he shouldn't arrange these players to play in the same formation. Substituting one or two of them for renewing the player formation is recommended.