Exploratory Data Analytics and Visualization- HW1

Team 12

September 30, 2019

Contents

The Business Problem and Our Approach

Our Task

AS Roma, a prestigious club in Italian football has turned to data to help them gain an edge over rivals in the race to win top spot of the Serie A league. To that end, we have been hired by AS Roma to be their resident data scientist group. The coach of the team wants us to find patterns that he can exploit to increase success on the field and decrease failure. While allowed to use any tools and techniques that can bring AS Roma valuable insights into how better they may compete, the coach is keen that association rules play a part of our analysis.

Defining Success

Roma is an ambitious club that has set its sight on the top 3 league spots and nothing less. The club strictly defines success as consisting of purely wins. Draws are viewed unfavorably, almost as bad as a loss.

Key Question

The key question of our analysis is what strategies Roma can take to defeat AC Milan. Across all the years of match data (2008/09-2012/13), AC Milan is the team we find the next hardest to win against, only after Juventus. Our reason for choosing to focus on AC Milan rather than Juventus is that Roma has a very low match win rate (only around 18%) against Juventus.The next four to five teams that Roma finds hard to beat lose between 30-40% of the matches played against Roma, AC Milan being at 31%. Thus, it is fairly evident that Juventus could be a lot harder to beat than AC Milan. Irrespective of the team Roma wins against, the same 3 points are up for grabs. Hence it makes more sense to target AC Milan.

Since this data science approach to supplementing game strategy is new to Roma, they would like to see the results against one opponent, before placing all their faith in the process blindly. If the results are positive, such an analysis can be extended to the other teams Roma wants to focus next on defeating.

An assumption we hold all through the analysis is that Roma has currently just begun a domestic season and transferring of players in-and-out of the club is not a possibility. With this assumption we limit the focus of our analysis to strategies AS Roma can adopt to maximize wins using only their current squad.

Our Approach

We first looked at players who have played against AC Milan. More specifically, we first classified players (excluding goalkeepers) into three types-defenders, attackers and midfielders. Then using clustering, we subgrouped these three types of players. In order to devise strategies for Roma, a team weaker than AC Milan, we looked at matches where AC Milan has lost to its weaker teams. We not only looked at what types of players were involved in those matches, but their formations. By doing so, we attempted to find out nonobvious patterns of players and their formations that have defeated AC Milan. Lastly, based on Roma's own situation, we offered recommendations.

Data Cleaning

Load Dependencies

```
library(lubridate)
library(data.table)
library(sqldf)
library(dplyr)
library(magrittr)
library(tidyverse)
library(ggplot2)
library(readxl)
library(dplyr)
library(tidyr)
library(stringr)
library(dbplyr)
library(xml2)
library(naniar)
library(hash)
```

```
library(plyr)
library(purrr)
library(arules)
library(arulesViz)
library(cluster)
library(fmsb)
library(RColorBrewer)
library(scales)
library(clue)
library(reshape2)
library(gridExtra)
```

Load datasets

```
con <- src_sqlite("euro_soccer.sqlite")

long_team_name <- 'Roma'
#each of the followint tables are just dplyr connections to the database tables
#if or when I need to bring the table to local memory I need to run table <- collect(table)
country_tbl <- tbl(con, "country")
league_tbl <- tbl(con, "league")
match_tbl <- tbl(con, "match")
player_tbl <- tbl(con, "player")
player_atts_tbl <- tbl(con, "player_attributes")
team_tbl <- tbl(con, "team")
team_atts_tbl <- tbl(con, "team_attributes")

country <- country_tbl %>% collect()
league <- league_tbl %>% collect()
match <- match_tbl %>% collect()
player <- player_tbl %>% collect()
player_atts <- player_atts_tbl %>% collect()
team <- team_tbl %>% collect()
team_atts <- team_atts_tbl %>% collect()


roma_record <- team_tbl %>%
                collect() %>%
                filter(grepl(long_team_name, team_long_name))
```

Missing values

There are a total of seven tables in the database. We first look at how many missing values are there in each table.

```
head(miss_var_summary(country))
```

```
## # A tibble: 2 x 3
##   variable n_miss pct_miss
##   <chr>     <int>    <dbl>
## 1 id            0        0
## 2 name          0        0
```

```r
head(miss_var_summary(league))
```

```
## # A tibble: 3 x 3
##   variable   n_miss pct_miss
##   <chr>       <int>    <dbl>
## 1 id              0        0
## 2 country_id      0        0
## 3 name            0        0
```

```r
head(miss_var_summary(player))
```

```
## # A tibble: 6 x 3
##   variable           n_miss pct_miss
##   <chr>               <int>    <dbl>
## 1 id                      0        0
## 2 player_api_id           0        0
## 3 player_name             0        0
## 4 player_fifa_api_id      0        0
## 5 birthday                0        0
## 6 height                  0        0
```

```r
head(miss_var_summary(team))
```

```
## # A tibble: 5 x 3
##   variable         n_miss pct_miss
##   <chr>             <int>    <dbl>
## 1 team_fifa_api_id     11     3.68
## 2 id                    0     0
## 3 team_api_id           0     0
## 4 team_long_name        0     0
## 5 team_short_name       0     0
```

```r
head(miss_var_summary(team_atts))
```

```
## # A tibble: 6 x 3
##   variable             n_miss pct_miss
##   <chr>                 <int>    <dbl>
## 1 buildUpPlayDribbling    969     66.5
## 2 id                        0      0
## 3 team_fifa_api_id          0      0
## 4 team_api_id               0      0
## 5 date                      0      0
## 6 buildUpPlaySpeed          0      0
```

For country, league and player tables, there are no missing values; And for the team table, there are a small number of missing values in team_fifa_api_id. But in this analysis, team_fifa_api_id is not used, so we just keep the table as it is; And for the team attributes table, there are a large number of missing values in buildUpPlayDribbling.

```
head(miss_var_summary(match))
```

```
## # A tibble: 6 x 3
##   variable n_miss pct_miss
##   <chr>     <int>    <dbl>
## 1 PSH       14811     57.0
## 2 PSD       14811     57.0
## 3 PSA       14811     57.0
## 4 BSH       11818     45.5
## 5 BSD       11818     45.5
## 6 BSA       11818     45.5
```

```
head(miss_var_summary(player_atts))
```

```
## # A tibble: 6 x 3
##   variable            n_miss pct_miss
##   <chr>                <int>    <dbl>
## 1 attacking_work_rate   3230     1.76
## 2 volleys               2713     1.47
## 3 curve                 2713     1.47
## 4 agility               2713     1.47
## 5 balance               2713     1.47
## 6 jumping               2713     1.47
```

For match table, there are a large number of missing values across the table; For player table, each column has a small amount of missing data.

Data Exploration

The manager's goal is to effectively increase AS Roma's win rate against AC Milan. For this analysis, we assume that AS Roma is just as competent, if not more, as the other teams featured below in the league's rankings and in terms of the win-rate analysis.

The factors we chose to look at are the team attributes of the weaker team on the day of the match ranging from the strength of the defense to other variables such as buildUpPlayPassing. We added a variable to check if AC Milan is playing home or away, and finally added the positions that AC Milan and the less competent team played in every match. The formations are denoted in a format similar to '422' or '4321', and are encoded as multiple levels for both the weak team (weakerItems variable) and AC Milan's formation (acmItems). To identify our target feature,we determine the formation and match result for both home and away teams for every match played in Italy Serie A.

Define our target matches

As it is only meaningful for Roma to form special strategies against its stronger teams. Here we tried to find out a list of teams that are stronger than Roma in Italian Serie A league. We calculated the win rates of Roma against other teams in the Italian Serie A league. We defined stronger teams as the ones whom Roma has win rate of over 50%.

We discovered that there are nine stronger teams. And the win rate of Roma against AC Milan is around 30%, making AC Milan a reasonable target for our analysis.

Define target features

Description and Rationale for Chosen Analysis

The goal here is to find out the factors that matter and limit our scope to such factors and leverage insight out of the data to guide AS Roma's strategy. We performed a regression analysis to partial out the impact of each of these factors holding the other variables fixed. This gives us a robust sense of what factors truly matter when defeating AC Milan in a match. The response variable here is a binary outcome of whether AC Milan lost (encoded as 1, because that's a favorable scenario for us, and 0). We've also accounted for the interaction effects of the home and away position combinations.

Execution and Result

```
summary(lr)
```

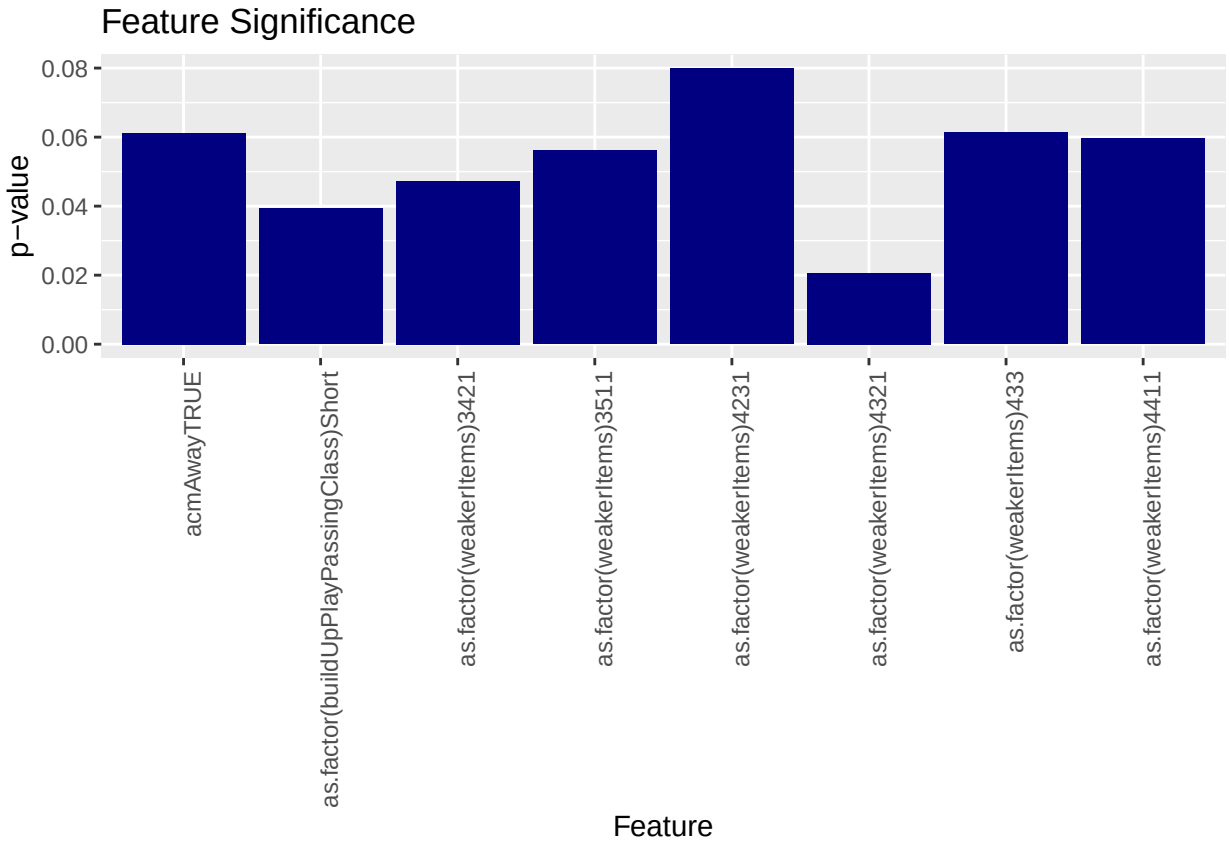
```
##
## Call:
## lm(formula = acmLoss ~ as.factor(weakerItems) + as.factor(acmItems) +
##     as.factor(weakerItems) * as.factor(acmItems) + acmAway +
##     defenceTeamWidth + as.factor(buildUpPlayPassingClass) + as.factor(defencePressureClass) +
##     as.factor(buildUpPlaySpeedClass) + as.factor(buildUpPlayDribblingClass) +
##     as.factor(chanceCreationPassingClass) + as.factor(chanceCreationCrossingClass) +
##     as.factor(chanceCreationShootingClass) + as.factor(defenceAggressionClass) +
##     defenceTeamWidth + +as.factor(defenceDefenderLineClass),
##     data = acmm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8476 -0.2354 -0.0298  0.1206  0.8389
##
## Coefficients: (22 not defined because of singularities)
##                                                  Estimate Std. Error
## (Intercept)                                      1.566431   0.985677
## as.factor(weakerItems)352                       -0.712540   0.530899
## as.factor(weakerItems)433                       -0.937907   0.492584
## as.factor(weakerItems)442                        0.008005   0.635760
## as.factor(weakerItems)3412                      -0.740735   0.572595
## as.factor(weakerItems)3421                      -1.357651   0.671762
## as.factor(weakerItems)3511                      -1.069083   0.549754
## as.factor(weakerItems)4231                      -0.929850   0.522700
## as.factor(weakerItems)4312                      -0.789460   0.500597
## as.factor(weakerItems)4321                      -1.367876   0.576786
## as.factor(weakerItems)4411                      -1.113955   0.580966
## as.factor(weakerItems)41212                     -1.022731   0.864243
## as.factor(acmItems)442                          -0.204754   0.238968
## as.factor(acmItems)4231                         -0.112692   0.478275
## as.factor(acmItems)4312                         -0.138486   0.536873
## acmAwayTRUE                                      0.192997   0.101320
## defenceTeamWidth                                -0.005822   0.006867
## as.factor(buildUpPlayPassingClass)Mixed          0.150328   0.182272
## as.factor(buildUpPlayPassingClass)Short          0.625997   0.297667
## as.factor(defencePressureClass)Medium           -0.181624   0.199556
## as.factor(buildUpPlaySpeedClass)Fast            -0.188715   0.203074
## as.factor(buildUpPlaySpeedClass)Slow            -0.289378   0.487607
## as.factor(buildUpPlayDribblingClass)Lots        -0.084718   0.263154
```

```
## as.factor(buildUpPlayDribblingClass)Normal          -0.255691   0.243695
## as.factor(chanceCreationPassingClass)Risky           0.182079   0.171774
## as.factor(chanceCreationPassingClass)Safe            0.231739   0.262445
## as.factor(chanceCreationCrossingClass)Normal         0.201838   0.180104
## as.factor(chanceCreationShootingClass)Lots           0.039058   0.302631
## as.factor(chanceCreationShootingClass)Normal         0.109023   0.261083
## as.factor(defenceAggressionClass)Press              -0.313012   0.613587
## as.factor(defenceDefenderLineClass)Offside Trap      0.010053   0.164321
## as.factor(weakerItems)352:as.factor(acmItems)442     0.108170   0.328741
## as.factor(weakerItems)433:as.factor(acmItems)442     0.142019   0.298213
## as.factor(weakerItems)442:as.factor(acmItems)442    -0.999888   0.644268
## as.factor(weakerItems)3412:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)3421:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)3511:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)4231:as.factor(acmItems)442    0.546207   0.393404
## as.factor(weakerItems)4312:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)4321:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)4411:as.factor(acmItems)442          NA         NA
## as.factor(weakerItems)41212:as.factor(acmItems)442         NA         NA
## as.factor(weakerItems)352:as.factor(acmItems)4231          NA         NA
## as.factor(weakerItems)433:as.factor(acmItems)4231          NA         NA
## as.factor(weakerItems)442:as.factor(acmItems)4231   -0.725334   0.769045
## as.factor(weakerItems)3412:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)3421:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)3511:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)4231:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)4312:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)4321:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)4411:as.factor(acmItems)4231         NA         NA
## as.factor(weakerItems)41212:as.factor(acmItems)4231        NA         NA
## as.factor(weakerItems)352:as.factor(acmItems)4312   -0.164398   0.648711
## as.factor(weakerItems)433:as.factor(acmItems)4312    0.572532   0.583247
## as.factor(weakerItems)442:as.factor(acmItems)4312   -0.497092   0.722905
## as.factor(weakerItems)3412:as.factor(acmItems)4312         NA         NA
## as.factor(weakerItems)3421:as.factor(acmItems)4312         NA         NA
## as.factor(weakerItems)3511:as.factor(acmItems)4312         NA         NA
## as.factor(weakerItems)4231:as.factor(acmItems)4312  -0.247094   0.660150
## as.factor(weakerItems)4312:as.factor(acmItems)4312   0.260108   0.614574
## as.factor(weakerItems)4321:as.factor(acmItems)4312   0.768427   0.748870
## as.factor(weakerItems)4411:as.factor(acmItems)4312         NA         NA
## as.factor(weakerItems)41212:as.factor(acmItems)4312        NA         NA
##                                                    t value Pr(>|t|)
## (Intercept)                                          1.589   0.1168
## as.factor(weakerItems)352                           -1.342   0.1841
## as.factor(weakerItems)433                           -1.904   0.0613 .
## as.factor(weakerItems)442                            0.013   0.9900
## as.factor(weakerItems)3412                          -1.294   0.2003
## as.factor(weakerItems)3421                          -2.021   0.0473 *
## as.factor(weakerItems)3511                          -1.945   0.0561 .
## as.factor(weakerItems)4231                          -1.779   0.0799 .
## as.factor(weakerItems)4312                          -1.577   0.1196
## as.factor(weakerItems)4321                          -2.372   0.0206 *
## as.factor(weakerItems)4411                          -1.917   0.0595 .
## as.factor(weakerItems)41212                         -1.183   0.2409
```

```
## as.factor(acmItems)442                                    -0.857   0.3946
## as.factor(acmItems)4231                                   -0.236   0.8145
## as.factor(acmItems)4312                                   -0.258   0.7972
## acmAwayTRUE                                                1.905   0.0612 .
## defenceTeamWidth                                          -0.848   0.3995
## as.factor(buildUpPlayPassingClass)Mixed                    0.825   0.4125
## as.factor(buildUpPlayPassingClass)Short                    2.103   0.0393 *
## as.factor(defencePressureClass)Medium                     -0.910   0.3661
## as.factor(buildUpPlaySpeedClass)Fast                      -0.929   0.3561
## as.factor(buildUpPlaySpeedClass)Slow                      -0.593   0.5549
## as.factor(buildUpPlayDribblingClass)Lots                  -0.322   0.7485
## as.factor(buildUpPlayDribblingClass)Normal                -1.049   0.2979
## as.factor(chanceCreationPassingClass)Risky                 1.060   0.2930
## as.factor(chanceCreationPassingClass)Safe                  0.883   0.3804
## as.factor(chanceCreationCrossingClass)Normal               1.121   0.2665
## as.factor(chanceCreationShootingClass)Lots                 0.129   0.8977
## as.factor(chanceCreationShootingClass)Normal               0.418   0.6776
## as.factor(defenceAggressionClass)Press                    -0.510   0.6117
## as.factor(defenceDefenderLineClass)Offside Trap            0.061   0.9514
## as.factor(weakerItems)352:as.factor(acmItems)442           0.329   0.7432
## as.factor(weakerItems)433:as.factor(acmItems)442           0.476   0.6355
## as.factor(weakerItems)442:as.factor(acmItems)442          -1.552   0.1255
## as.factor(weakerItems)3412:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)3421:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)3511:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)4231:as.factor(acmItems)442          1.388   0.1697
## as.factor(weakerItems)4312:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)4321:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)4411:as.factor(acmItems)442             NA       NA
## as.factor(weakerItems)41212:as.factor(acmItems)442            NA       NA
## as.factor(weakerItems)352:as.factor(acmItems)4231             NA       NA
## as.factor(weakerItems)433:as.factor(acmItems)4231             NA       NA
## as.factor(weakerItems)442:as.factor(acmItems)4231         -0.943   0.3490
## as.factor(weakerItems)3412:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)3421:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)3511:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)4231:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)4312:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)4321:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)4411:as.factor(acmItems)4231            NA       NA
## as.factor(weakerItems)41212:as.factor(acmItems)4231           NA       NA
## as.factor(weakerItems)352:as.factor(acmItems)4312         -0.253   0.8007
## as.factor(weakerItems)433:as.factor(acmItems)4312          0.982   0.3299
## as.factor(weakerItems)442:as.factor(acmItems)4312         -0.688   0.4941
## as.factor(weakerItems)3412:as.factor(acmItems)4312            NA       NA
## as.factor(weakerItems)3421:as.factor(acmItems)4312            NA       NA
## as.factor(weakerItems)3511:as.factor(acmItems)4312            NA       NA
## as.factor(weakerItems)4231:as.factor(acmItems)4312        -0.374   0.7094
## as.factor(weakerItems)4312:as.factor(acmItems)4312         0.423   0.6735
## as.factor(weakerItems)4321:as.factor(acmItems)4312         1.026   0.3086
## as.factor(weakerItems)4411:as.factor(acmItems)4312            NA       NA
## as.factor(weakerItems)41212:as.factor(acmItems)4312           NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4171 on 66 degrees of freedom
## Multiple R-squared:  0.4989, Adjusted R-squared:  0.1876
## F-statistic: 1.603 on 41 and 66 DF,  p-value: 0.04323
```

```
pval = data.frame(summary(lr)[4][1])
pval$feature = row.names(data.frame(summary(lr)[4][1]))

pval = pval %>% filter(coefficients.Pr...t.. < 0.1) %>% arrange(coefficients.Pr...t..)
ggplot(data = pval, aes(x= feature, y = coefficients.Pr...t..)) +
   geom_bar(stat="identity", fill="naVyblue") +
     labs(title="Feature Significance",
        x ="Feature", y = "p-value") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Interpretation and Conclusion

The results gives us a clear picture of what variables matter (significance). The formation of the weaker Team is highly significant in determining the outcome of a match against AC Milan, whereas AC Milan's formations did not turn out to be signicant. We can note that certain positions that the weaker team plays significantly reduce the odds of winning against AC Milan. Moreover, variables such as whether AC Milan is playing home/away (acmAway) and buildUpPlayPassingClass explain the match outcome significantly. There is evidence to suspect that the formation of the weaker team is a significant explainer of the team's strategy in defeating AC Milan, and hence we form the team's formation as a basis in our Association Rules analysis to answer the question - Which team formations increase the odds of AS Roma winning against AC Milan?

Data preparation for Analysis

We first subsetted the matches where the stronger teams have lost domestically. Within those matches, we found the ones that are associated with AC Milan as our target matches for clustering. The output for this step is the attributes information of all players and player formation in the targeted matches.

Find the matches where the stronger teams have lost domestically

Total: 519 matches

Find out matches associated with AC Milan

Total: 74 matches

Find out the player attributes of opponent teams of AC Milan in the matches:

Total: 3531 player records

Get player roles of the winning teams in the matches where AC Milan has lost domestically:

Total: 410 unique players

Final dataframe for clustering and divide dataframe for clustering::

K-mean Clustering Modeling on Player Attributes

To understand the types of players Roma and the other teams have, we have decided to adopt a clustering approach, more specifically k-means clustering.

Description and Rationale for the Chosen Analysis The goal of the clustering is to recognize non-obvious player attribute combinations in each position (attack, mid-field, and defense). The algorithm we chose is K-mean clustering, since it is able to aggregate collections of players based on their similarities. The dataset we used includes matches where different teams played against AC Milan. We are interested in players that showed up in those matches and classify them based on 35 attributes (e.g.: agility, shot-power and passing) they have.

Train the model

Using the elbow curve as a heuristic to decide on the appropriate number of clusters to have within each of the three roles (attackers, midfielders, and defenders), we arrive at 5 clusters within attackers, 8 clusters within midfielders, and 2 clusters within defenders. AS Roma has players belonging only to a few of these clusters, and given that our focus is on Roma's current season campaign, we assume that the overall squad cannot be changed drastically. We thus limit our focus to clusters that Roma already has players from, and identifying what clusters work best against AC Milan so that the coach may make selections from these recommended clusters and maximize Roma's chance of winning.

Defenders:

Based on the elbow chart, we clustered defenders into 2 groups.

Interpretation:

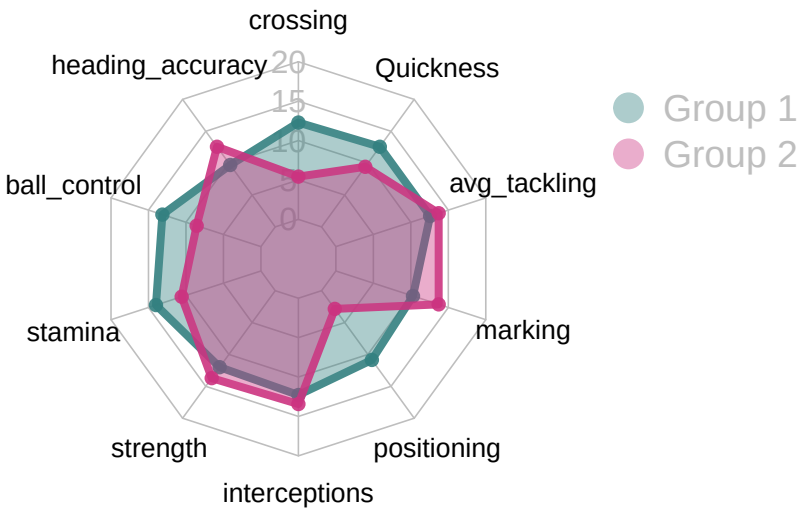
Following were identified as the relevant clusters for defenders:

Centrebacks (group 2 from defender radar chart): The primary responsibility of these players is to prevent opposition from scoring goals. They are tall, strong individuals who are very good at marking, intercepting, tackling, and heading the ball. However, their skillset is generally limited to these defensive attributes, and they are generally not very quick or creative when in possession of the football. Example of Roma players:

Fullbacks/Wingbacks (group 1 from the defender radar chart): These players play a supporting role to centre backs while also providing width to attacks through overlapping runs. They are wide defenders, and are very good at crossing the ball. They also have a solid workrate, having to run up the pitch to support attacking moves and also falling back to support their defense. Example of Roma players:

visualization

```
def_char_data = data.frame("crossing"=c(100,30,73,49),
"heading_accuracy"=c(100,30,64,74),
"ball_control"=c(100,30,76,60),
"stamina"=c(100,30,79,67),
"strength"=c(100,30,72,78),
"interceptions"=c(100,30,73,77),
"positioning"=c(100,30,68,40),
"marking"=c(100,30,66,78),
"avg_tackling"=c(100,30,74,78),
"Quickness"=c(100,30,74,63))
rownames(def_char_data) = c("1","2","Group 1","Group 2")


colors_border=c( rgb(0.2,0.5,0.5,0.9), rgb(0.8,0.2,0.5,0.9))
colors_in=c( rgb(0.2,0.5,0.5,0.4), rgb(0.8,0.2,0.5,0.4))


# plot with default options:
radarchart(def_char_data , axistype=1 ,
           #custom polygon
           pcol=colors_border , pfcol=colors_in , plwd=4 , plty=1,
           #custom the grid
           cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5), cglwd=0.8,
           #custom labels
           vlcex=0.8,
           title = "Defender Attributes by Cluster Group")

# Add a legend
legend(x=1.5, y=1, legend = rownames(def_char_data[-c(1,2),]),
       bty = "n", pch=20 , col=colors_in , text.col = "grey", cex=1.2, pt.cex=3)
```



To illustrate our interpretation, we perform a radar chart for these two types of defenders. For Group1 defenders (recognized as Fullbacks/Wingbacks), they have a better skill set in terms of quickness, crossing and ball control, which enables them to break through opponents and create oppotunities for crossing. In adiition, their good stamina and positioning helps them to attack forwards and defend backwards on the side. On the contrary, group 1 defenders (recognized as Centrebacks) receives a better score in strength, interception and marking. This allows them to win the ball back and make clearance kick at the last moment.

Mid-fielders:

Based on the elbow graph, we clustered midfielders into 8 groups.

Interpretation:

Following were identified as the relevant clusters for mid-fielders:

Enforcers (group 2 in midfielder radar chart): These players specialize in winning the ball back from the opposition early in midfield. They tend not to be as technically gifted as players from the other clusters, but make it up in their ability to intercept and tackle. They are strong, aggressive players that shield their team's defensive line, and turn the ball to box-to-box midfielders or other attacking players further up the football pitch.

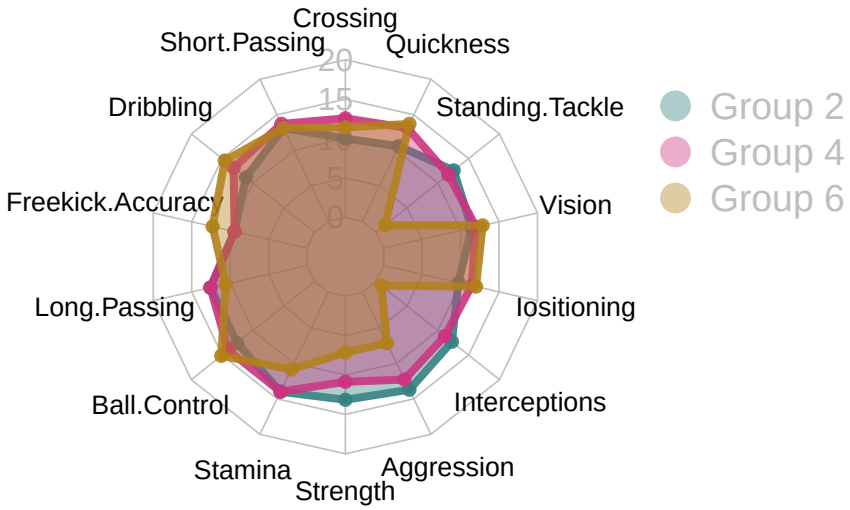
Box-to-box midfielders (group 4 in midfield radar chart): These players are versatile midfielders possessing a very balanced skillset. They are the engine of a team, having a lot of stamina and constantly running. They are good at dribbling and have good ball control. Their role in the team is to facilitate speedy transition from defence to attack, while also supporting 'enforcers' in winning the ball back when the ball is not in their team's possession.

Attacking midfielders (group 6 in midfield radar chart): These players have a very special skill set, and are not as versatile as box-to-box midfielders. These players play behind the target man or second striker, and are largely responsible for creating opportunities. They have the best ball-control, vision and

dribbling amongst other types of midfielders, and are quick too. They are generally more accurate with freekicks, and have high spatial awareness. Thus, they are very intelligent in terms of the positions they take up, exploiting little pockets of space from where they can play in the final pass to the strikers.

Visualization:

```
## Midfielders
mid_char_data = data.frame("Crossing"=c(100,30,65,74,70),
                           "Short Passing"=c(100,30,76,78,76),
                           "Dribbling"=c(100,30,69,76,81),
                           "Freekick Accuracy"=c(100,30,63,63,73),
                           "Long Passing"=c(100,30,74,74,67),
                           "Ball Control"=c(100,30,74,79,83),
                           "Stamina"=c(100,30,79,79,68),
                           "Strength"=c(100,30,76,68,55),
                           "Aggression"=c(100,30,78,73,55),
                           "Interceptions"=c(100,30,73,69,33),
                           "Iositioning"=c(100,30,64,70,72),
                           "Vision"=c(100,30,70,72,75),
                           "Standing Tackle"=c(100,30,74,71,35),
                           "Quickness"=c(100,30,67,76,78))
rownames(mid_char_data) = c("1","2","Group 2","Group 4","Group 6")


colors_border=c( rgb(0.2,0.5,0.5,0.9), rgb(0.8,0.2,0.5,0.9) , rgb(0.7,0.5,0.1,0.9))
colors_in=c( rgb(0.2,0.5,0.5,0.4), rgb(0.8,0.2,0.5,0.4) , rgb(0.7,0.5,0.1,0.4))


# plot with default options:
radarchart(mid_char_data , axistype=1 ,
           #custom polygon
           pcol=colors_border , pfcol=colors_in , plwd=4 , plty=1,
           #custom the grid
           cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5), cglwd=0.8,
           #custom labels
           vlcex=0.8,
           title = "Midfielder Attributes by Cluster Group")

# Add a legend
legend(x=1.5, y=1, legend = rownames(mid_char_data[-c(1,2),]),
       bty = "n", pch=20 , col=colors_in , text.col = "grey", cex=1.2, pt.cex=3)
```



The interprretation is slightly harder in mid-fielder than defenders, as the role and positions are more diverse. However, based on the radar chart, it is apprant that group 6 midfielders (recognized as attacking-midfielders) receive a low score in tackles and interceptions, but they are good at passing, ball control and freekick accuracy which enables them to make penetrate pass for goal oppotunities.

Compared to group 6 midfielders, group 2 (recognized as enforcers) midfielders does not have great ball techniques, but instead, they have great strength, interceptions and tackle skills. Lastly, group 4 midfielders (recognized as Box-to-Box midfielders) tend to have a more balanced skill set, which enables them to win the ball back and make smooth transition from defense to attack.

Attackers:

Based on elbow chart, we clustered attackers into 5 groups.

Interpretation:

Following were identified as the relevant clusters for attackers:

Wide forwards (group 1 in attacker radar chart): These players are lightning-quick, great dribblers and have the best balance among attackers belonging to other clusters. They add speed and width to a team's attack, which stretches out opposition defences giving more space to strikers. Many of these forwards also possess great crossing ability.

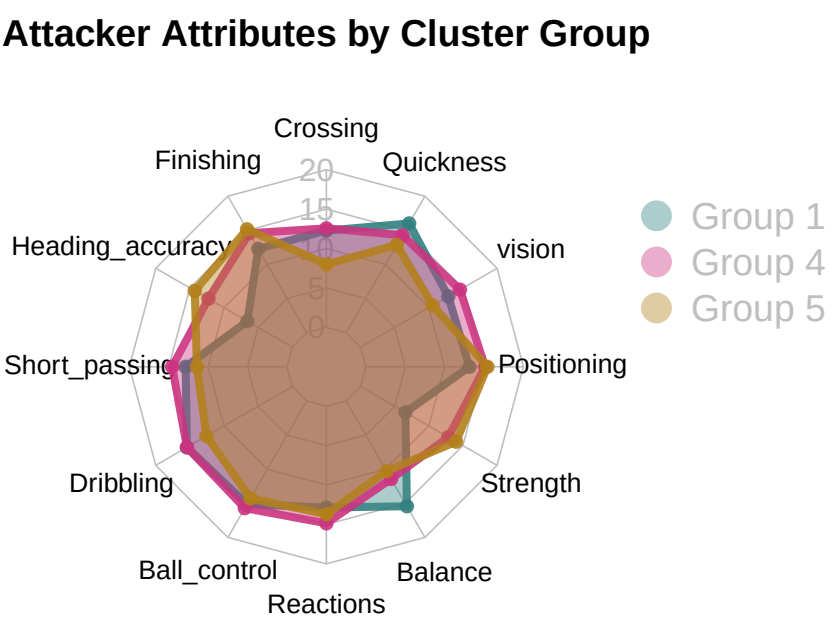
Second-strikers (group 4 in attacker radar chart): Unlike target-men, these individuals are mobile forwards with good technical ability (dribbling skills and ball control) and link-up play that only support out-and-out strikers in scoring goals, but can also create opportunities for other players, especially the wide forwards. Their excellent reactions, ball-control, vision (eye for a pass) and short-passing ability allows them to quickly pick-out passes thus creating opportunities for other teammates.

Target-man/out-and-out striker (group 5 in attacker radar chart): These players are prolific goalscorers, known to be lethal in front of goal. What they lack in speed, they make up in strength and po-

sitioning. They are well-built, both in terms of weight and height, challenging defenders physically. Among all attacking clusters, they have the best heading accuracy, and are among the best at finishing.

Visualization:

```
## Attackers
att_char_data = data.frame("Crossing" = c(100,30,73,74,58),
                           "Finishing"=c(100,30,73,81,83),
                           "Heading_accuracy"=c(100,30,53,73,80),
                           "Short_passing"=c(100,30,75,81,70),
                           "Dribbling"=c(100,30,84,84,74),
                           "Ball_control"=c(100,30,83,85,80),
                           "Reactions"=c(100,30,75,82,78),
                           "Balance"=c(100,30,84,70,66),
                           "Strength"=c(100,30,53,75,79),
                           "Positioning"=c(100,30,76,83,84),
                           "vision"=c(100,30,75,81,67),
                           "Quickness"=c(100,30,86,80,75))
rownames(att_char_data) = c("1","2","Group 1","Group 4","Group 5")


colors_border=c( rgb(0.2,0.5,0.5,0.9), rgb(0.8,0.2,0.5,0.9) , rgb(0.7,0.5,0.1,0.9) )
colors_in=c( rgb(0.2,0.5,0.5,0.4), rgb(0.8,0.2,0.5,0.4) , rgb(0.7,0.5,0.1,0.4) )


# plot with default options:
radarchart(att_char_data , axistype=1 ,
            #custom polygon
            pcol=colors_border , pfcol=colors_in , plwd=4 , plty=1,
            #custom the grid
            cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5), cglwd=0.8,
            #custom labels
            vlcex=0.8,
           title = "Attacker Attributes by Cluster Group")


# Add a legend
legend(x=1.5, y=1, legend = rownames(att_char_data[-c(1,2),]),
       bty = "n", pch=20 , col=colors_in , text.col = "grey", cex=1.2, pt.cex=3)
```



As shown on the radar chart, group 1 attackers (recognized as wide forwards) have a quick spped and great balance. This enables them to break through opponent's defenders and make squares pass. On the contrary, Group 5 attackers (recognized as target-man/out-and-out striker) performs better in terms of positioning, strength, finishing and heading accuracy. This allows them to be in the position timely and make the final shot. Lastly, group 4 attackers (recognized as second strikers) have a great vision and passing skills. They can make killer pass to the target-man, and their good positionings and reactions allow them to follow up opportunities that target-man may miss.

Descriptive Analysis of AS Roma and AC Milan Team Formations

Problem Description

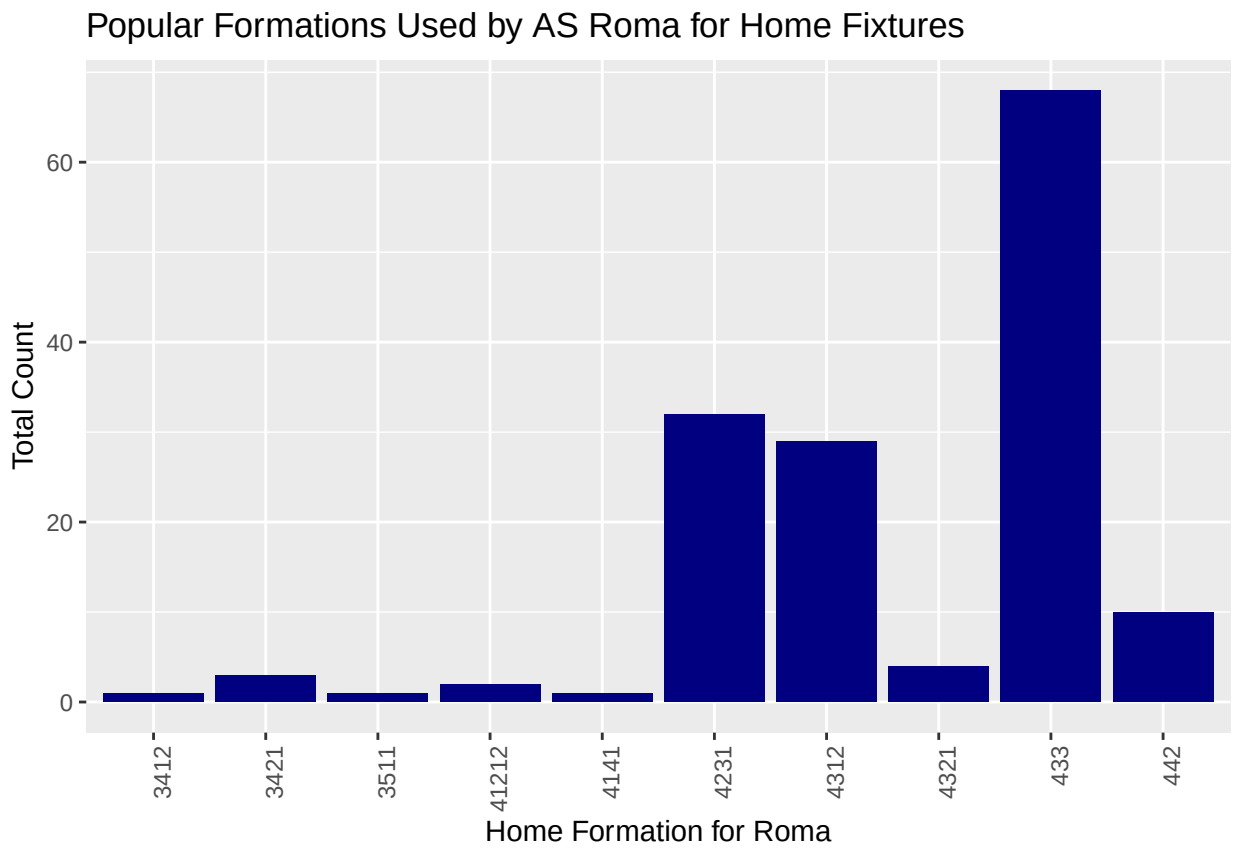
Now that we have clustered players who have played against AC Milan into relatively homogeneous segments, we need to establish how to set our team up in terms of field positions. This is called the formation of the team, and it dictates the general structure of the team for a particular game.

Description and Rationale for the Chosen Analysis

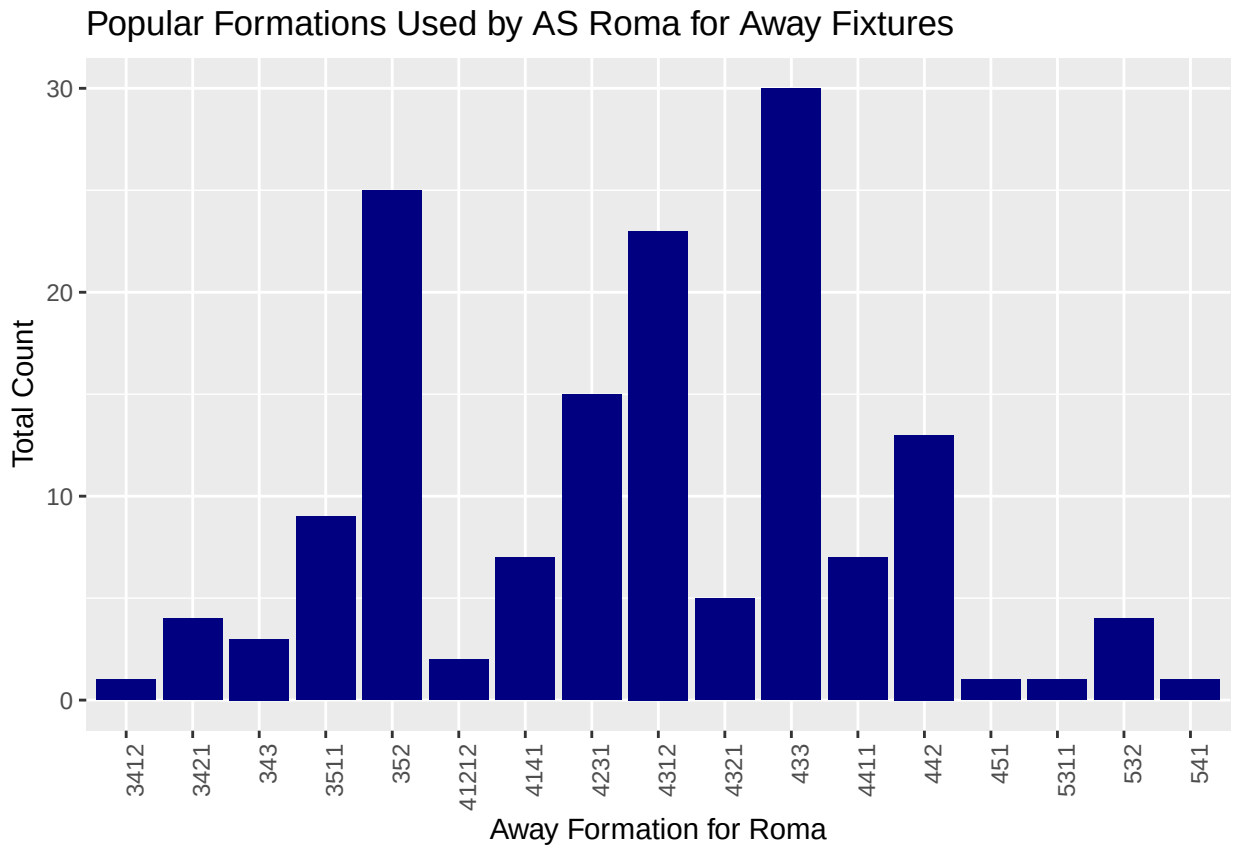
To study the most common formations used by Roma and AC Milan, we used simple descriptive analysis and clear visualizations. We believe these to be more than sufficient in bringing out trends regarding current Roma and AC Milan team formation strategies. We determine the formation excluding the position of the goalkeeper as he has a default position for all teams across leagues.

Execution and Results

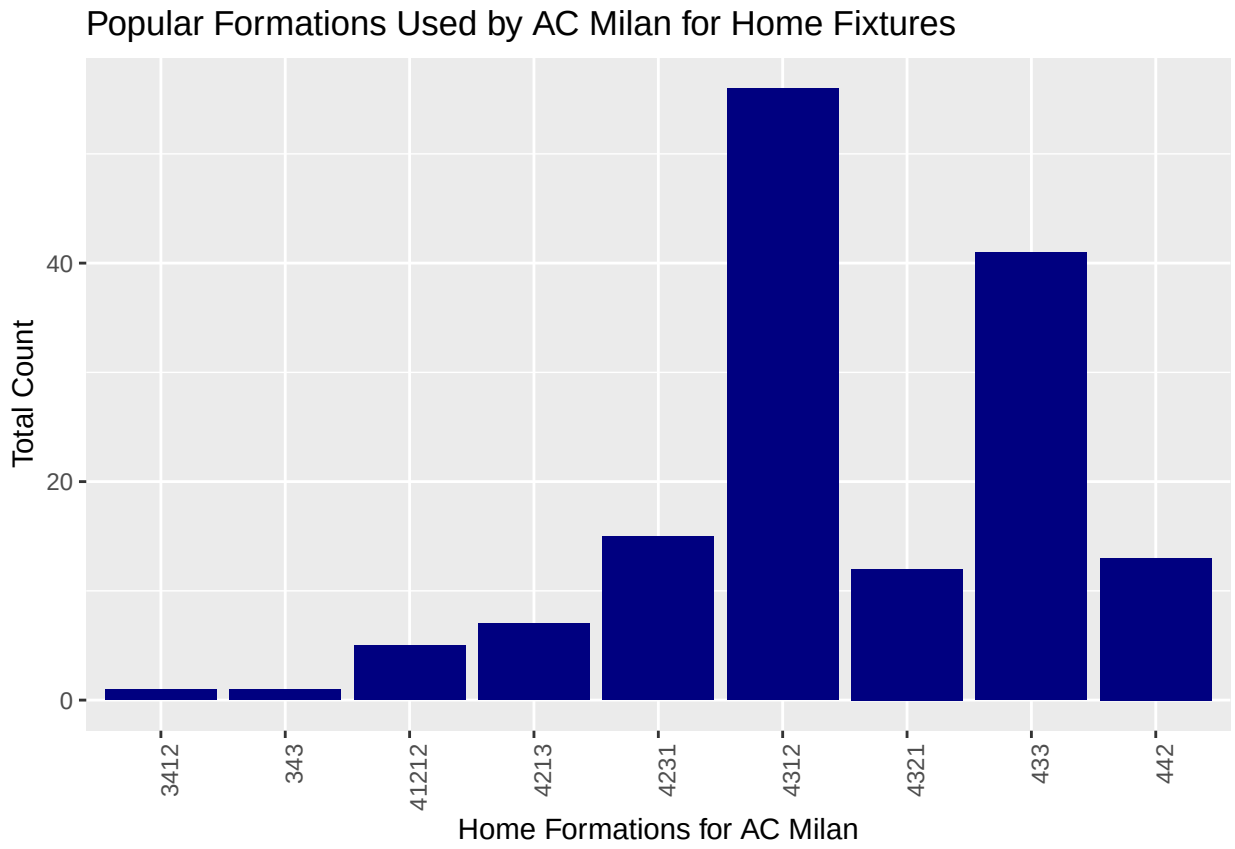
```
#Home team : Roma
# Formation : 433 (70 matches), 4231(30 matches), 4312(30 matches)
match_table %>% filter(home_team_api_id==roma_record$team_api_id)%>%
  ggplot(., aes(x=home_formation, fill = "blue") ) +
  geom_histogram(stat="count",fill = "naVyblue") +
  theme(axis.text.x = element_text(angle=90, hjust=1),legend.position = "none") +
  labs(x = "Home Formation for Roma", y = "Total Count",fill = "League Names")+
  scale_fill_brewer(palette="Paired")+
  ggtitle("Popular Formations Used by AS Roma for Home Fixtures")
```



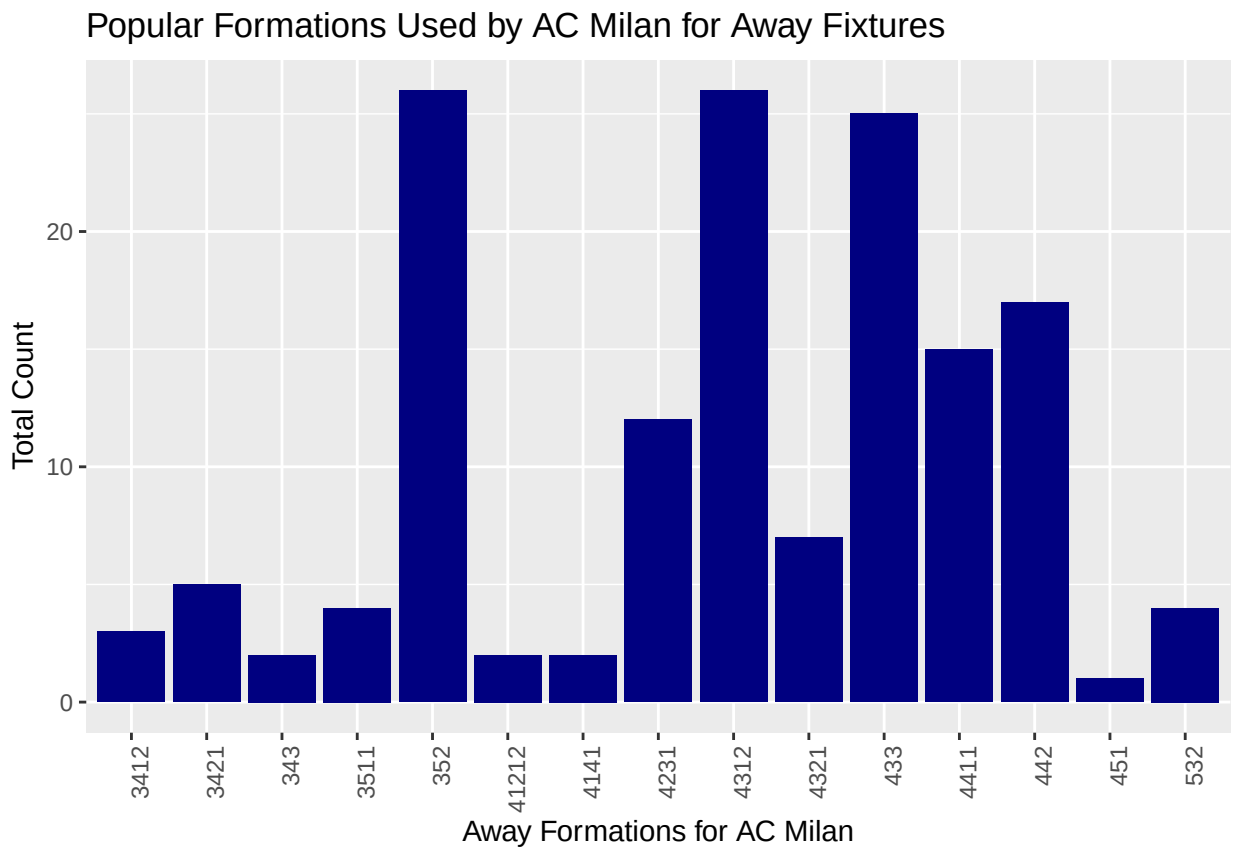

```
#Away team : Roma
# Formation : 433 (30 matches), 352(25 matches), 4312 (23 matches)
match_table %>% filter(home_team_api_id==roma_record$team_api_id)%>%
  ggplot(., aes(x=away_formation, fill = "blue") ) +
  geom_histogram(stat="count",fill = "naVyblue") +
  theme(axis.text.x = element_text(angle=90, hjust=1),legend.position = "none")+
  labs(x = "Away Formation for Roma", y = "Total Count",fill = "League Names")+
  scale_fill_brewer(palette="Paired")+
  ggtitle("Popular Formations Used by AS Roma for Away Fixtures")
```




```
#Home team : ACM
# Formation : 4312 (50 matches), 433 (40 matches)
match_table %>% filter(home_team_api_id==acm_str_id)%>%
  ggplot(., aes(x=home_formation,fill = "blue") ) +
  geom_histogram(stat="count",fill = "naVyblue")+
  theme(axis.text.x = element_text(angle=90, hjust=1),legend.position = "none")+
  labs(x = "Home Formations for AC Milan", y = "Total Count",fill = "League Names")+
  scale_fill_brewer(palette="Paired")+
  ggtitle("Popular Formations Used by AC Milan for Home Fixtures")
```




```
#Away team : ACM
# Formation : 352 and 4312(25 matches), 433(25 matches), 442 (17 matches)
match_table %>% filter(home_team_api_id==acm_str_id)%>%
  ggplot(., aes(x=away_formation,fill = "blue") ) +
  geom_histogram(stat="count",fill="naVyblue")+
  theme(axis.text.x = element_text(angle=90, hjust=1),legend.position = "none")+
  labs(x = "Away Formations for AC Milan", y = "Total Count",fill = "League Names")+
  scale_fill_brewer(palette="Paired")+
  ggtitle("Popular Formations Used by AC Milan for Away Fixtures")
```



Interpretation and Conclusion

Based on our analysis on formations, we observe that Roma's frequent play formation is 4-3-3 and 4-3-1-2. Irrespective of whether the game is played in the home ground or away. We also observe that AC Milan has a tendency to use 4-3-3 and 4-3-1-2 in both home and away matches too.

However, to be more confident about formation and team strategy, we believe that we should expand the scope of analysis to also look at formations of other teams that managed to beat AC Milan.

Association Rule Analysis on the formation and player types

Our key question for this step is: What can we learn from these teams who are weaker than AC Milan but won the matches? It is highly likely that AC Milan players, overall, have a better skill set than their opponents, what are other factors that can compensate for the disadvantage of players' abilities?

Description and Rationale for the Chosen Analysis

We believe that the formation of the team revealed something about the strategy of the team, such as whether the team is pressuring mid-field or playing defensive counterback. In addition, after a team determines its playstyle, which player Roma should pick, or more generally, what types of players to pick. Association Rules algorithm is a good tool focusing on finding co-occurring associations in dataset. In this case, we believe that using association rule analysis would help us define non-obvious patterns of formation and playertypes combinations. Therefore, for every match, we put the formation of teams and player types of all winning team's players, besides goalkeeper as our dataset for association rules mining.

Data preparation for Association Rules

We first matched every player to the playertypes(clusters created) they belong to. Then we subsetted the matches where AC Milan has lost(total: 26 matches). In order to find out the non-obvious patterns on the combination of team formation and playertypes, we added the formations of both teams used in those matches to the playertypes infomation. The output of this step is a dataset ready for performing Association Rules where the each row contains information of the formations of two teams and the playertypes assigned by the winning team.

Perform Associastion Rules

Get datasets ready for running Association Rules.

```
#Load matches as "transactions"
matches = read.transactions("ACM_26info_for_AR.csv", format = "basket",
                            sep = ",", rm.duplicates = TRUE)

#matches_formation <- acm_info_for_AR[,1:2]
# matches_postype <- acm_info_for_AR[,3:12]
# write.csv(matches_formation, "ACM_26matches_formation.csv", row.names = FALSE)
# write.csv(matches_postype, "ACM_26matches_postype.csv", row.names = FALSE)

matches_formation = read.transactions("ACM_26matches_formation.csv", format = "basket",
                            sep = ",", rm.duplicates = TRUE)
matches_postype = read.transactions("ACM_26matches_postype.csv", format = "basket",
                            sep = ",", rm.duplicates = TRUE)
```

Above we loaded three datasets ready for running Association Rules:

"matches_formation" dataset contains only the formations of both teams for each match record;

"matches_postype" dataset contains only the playertypes of tne teams who won AC Milan for each match record;

"matches" dataset contains formations of both teams and playertypes of tne teams who won AC Milan for each match record.

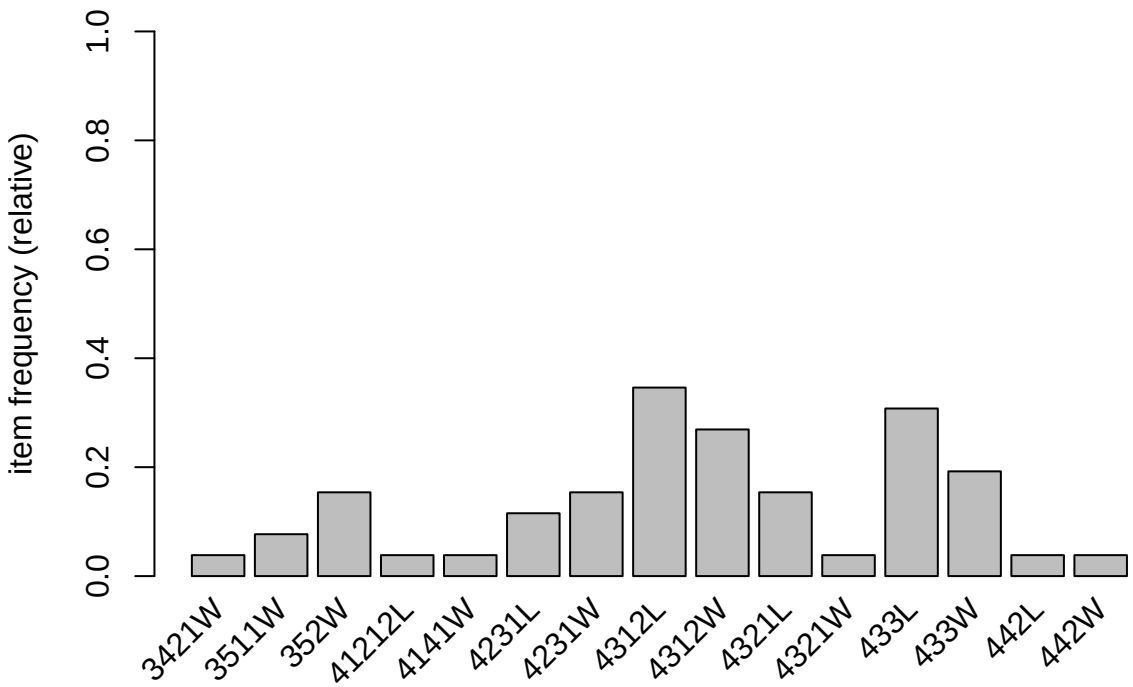
Apply Association Rules to Formation table

In this section, we applied Association rules on the "match_formation" table to find out the most frequently used formations that the winning team used to win AC Milan.

A. Frequent formation used to win ACM

We first explore the frequencies of the formations that were frequently used by the winning teams and AC Milan in those target matches.

```
# Find the most frequently used formation
itemFrequencyPlot(matches_formation, ylim = c(0, 1))
```




```
# The abslute support count for each formation used
itemFrequency(matches_formation, type = "absolute")
```

Interpretation and Conclusion:

We noticed that most teams that are weaker than AC Milan used 4312, 433, and 4231 formations to win AC Milan. And AC Milan lost mostly in those matches when using 4312 and 433 formations. Although AC Milan will use different formations for different matches, 4312 and 433 are two most common formations we suggest to focus on when designing the optimal winning strategies against AC Milan.

B. Which formation Roma should use when facing specific formation of AC Milan

We then apply Association rules on the "match_formation" table to find out which formation those winning teams use frequently when facing specific formation of AC Milan. Since we found out that 4312 and 433 are two most common formations in those matches AC Milan lost, we will use Association Rules to target these specific two formations respectively.

B-1: When AC Milan is playing **4312 formation**

```
rules <- apriori(matches_formation, parameter = list(supp = 0.01, conf = 0.1 ))
formation_res <- inspect(subset(rules, subset = lhs %in% "4312L")) %>% arrange(desc(count))
```

```
formation_res
```

```
##       lhs        rhs    support confidence      lift count
## 1 {4312L} => {4312W} 0.11538462  0.3333333 1.2380952     3
## 2 {4312L} => {4231W} 0.07692308  0.2222222 1.4444444     2
```

```
## 3 {4312L} =>  {433W} 0.07692308  0.2222222 1.1555556     2
## 4 {4312L} => {4321W} 0.03846154  0.1111111 2.8888889     1
## 5 {4312L} =>  {352W} 0.03846154  0.1111111 0.7222222     1
```

Interpretation and Conclusion:

From the result of Association rules, we found that the probability of winning teams using 4312 formation is 33% when AC Milan used 4312 formation lost, the probability of winning teams using 4231 formation is 22% when AC Milan used 4312 formation lost, and the probability of winning teams using 4333 formation is 22% when AC Milan used 4312 formation lost. Therefore, 4312, 4231, and 433 formation can be considered when playing against ACM's 4312 formation.

B-2: When AC Milan is playing **433 formation**

```
rules <- apriori(matches_formation, parameter = list(supp = 0.01, conf = 0.1 ))
formation_res <- inspect(subset(rules, subset = lhs %in% "433L")) %>% arrange(desc(count))
```

```
formation_res
```

```
##      lhs        rhs    support confidence     lift count
## 1 {433L} => {4312W} 0.11538462      0.375 1.392857     3
## 2 {433L} =>  {433W} 0.07692308      0.250 1.300000     2
## 3 {433L} => {3511W} 0.03846154      0.125 1.625000     1
## 4 {433L} => {4231W} 0.03846154      0.125 0.812500     1
## 5 {433L} =>  {352W} 0.03846154      0.125 0.812500     1
```

Interpretation:

From the result of Association rules, we found that the probability of winning teams using 4312 formation is 37% when AC Milan used 433 formation lost, and the probability of winning teams using 433 formation is 25% when AC Milan used 433 formation lost. Therefore, 4312, 433 can be considered when playing against ACM's 433 formation.

Conclusion

From the results of the Association Rules on the targeted 2 formations AC Milan used respectively, we found that 4312, 433, and 4231 are three formations we could suggest Roma to consider using to win these 2 common AC Milan's formations. Since we found out that 4312 and 433 are two most common formations in those matches AC Milan lost, we will use Association Rules to target these specific two formations respectively. From previews analysis on Roma's frequently used formations, 4312 and 433 are within the most familiar formations for Roma. Therefore, we will next fucus on which types of players to suggest for Roma when Roma playing these 2 formations against AC Milan.

Apply Association Rules to both Formation and Playertypes

From the previews analysis, we found that 4312 and 433 are the 2 target formations we suggest Roma to focus on when playing against AC Milan. In order to form a better strategy on the field, we also need to consider which types of players Roma should assign when playing those formations respectively. Hence, we next apply Association rules on the "matches" table to find out which types of players co-occurred frequently with the formations we specified.

A. When considering playing **4312 formation** against AC Milan:

```
rules <- apriori(matches, parameter = list(supp = 0.1, conf = 1))
both_res <- inspect(subset(rules, subset = lhs %in% "4312W")) %>% arrange(desc(count))
names(both_res) <- c('lhs','arrow','rhs', 'support', 'confidence', 'lift','count')
```

```
both_res %>% filter('count' >= 4)
```

```
##                     lhs arrow   rhs   support confidence     lift count
## 1          {4312W,D 1}    => {D 2} 0.2307692          1 1.040000     6
## 2          {4312W,D 2}    => {D 1} 0.2307692          1 1.529412     6
## 3          {4312W,M 6}    => {D 1} 0.1923077          1 1.529412     5
## 4          {4312W,M 6}    => {D 2} 0.1923077          1 1.040000     5
## 5      {4312W,D 1,M 6}    => {D 2} 0.1923077          1 1.040000     5
## 6      {4312W,D 2,M 6}    => {D 1} 0.1923077          1 1.529412     5
## 7          {4312W,A 1}    => {M 6} 0.1538462          1 1.857143     4
## 8          {4312W,A 1}    => {D 1} 0.1538462          1 1.529412     4
## 9          {4312W,A 1}    => {D 2} 0.1538462          1 1.040000     4
## 10     {4312W,A 1,M 6}    => {D 1} 0.1538462          1 1.529412     4
## 11     {4312W,A 1,D 1}    => {M 6} 0.1538462          1 1.857143     4
## 12     {4312W,A 1,M 6}    => {D 2} 0.1538462          1 1.040000     4
## 13     {4312W,A 1,D 2}    => {M 6} 0.1538462          1 1.857143     4
## 14     {4312W,A 1,D 1}    => {D 2} 0.1538462          1 1.040000     4
## 15     {4312W,A 1,D 2}    => {D 1} 0.1538462          1 1.529412     4
## 16 {4312W,A 1,D 1,M 6}    => {D 2} 0.1538462          1 1.040000     4
## 17 {4312W,A 1,D 2,M 6}    => {D 1} 0.1538462          1 1.529412     4
## 18 {4312W,A 1,D 1,D 2}    => {M 6} 0.1538462          1 1.857143     4
## 19        {4312W,433L}    => {D 1} 0.1153846          1 1.529412     3
## 20        {4312W,433L}    => {D 2} 0.1153846          1 1.040000     3
## 21         {4312W,A 5}    => {D 1} 0.1153846          1 1.529412     3
## 22         {4312W,A 5}    => {D 2} 0.1153846          1 1.040000     3
## 23    {4312W,433L,D 1}    => {D 2} 0.1153846          1 1.040000     3
## 24    {4312W,433L,D 2}    => {D 1} 0.1153846          1 1.529412     3
## 25     {4312W,A 5,D 1}    => {D 2} 0.1153846          1 1.040000     3
## 26     {4312W,A 5,D 2}    => {D 1} 0.1153846          1 1.529412     3
```

Interpretation and Conclusion:

From the result of Association rules, we found that D1 and D2 almost co-occurred in all the matches when winning teams play 4312 formation, which is reasonable, since it requires 4 defenders in the 4312 formations and it usually needs two types of defenders to cooperate in the match. Given this, we would suggest Roma to use these two types of players(D1,D2) when playing 4312 formation. In terms of attackers and Mid-fielders, we found that Mid-fielder type 6 (M6), Attacker type 1 (A1), and Attacker type 5 (A5) co-occurred 100% when D1 or D2 is being used in the match. Therefore, we suggest Roma to consider using players who are identified as these 3 types of players when playing 4312 formation against AC Milan.

B. When considering playing **433 formation** against AC Milan:

```
rules <- apriori(matches, parameter = list(supp = 0.1, conf = 1))
both_res <- inspect(subset(rules, subset = lhs %in% "433W")) %>% arrange(desc(count))
names(both_res) <- c('lhs','arrow','rhs', 'support', 'confidence', 'lift','count')
```

```
both_res %>% filter(count >= 4)
```

```
##               lhs arrow   rhs   support confidence     lift count
## 1          {433W}    => {D 1} 0.1923077          1 1.529412     5
## 2          {433W}    => {D 2} 0.1923077          1 1.040000     5
## 3      {433W,D 1}    => {D 2} 0.1923077          1 1.040000     5
## 4      {433W,D 2}    => {D 1} 0.1923077          1 1.529412     5
```

```
## 5      {433W,M 2}    => {D 1} 0.1538462          1 1.529412     4
## 6      {433W,M 2}    => {D 2} 0.1538462          1 1.040000     4
## 7      {433W,A 5}    => {D 1} 0.1538462          1 1.529412     4
## 8      {433W,A 5}    => {D 2} 0.1538462          1 1.040000     4
## 9  {433W,D 1,M 2}    => {D 2} 0.1538462          1 1.040000     4
## 10 {433W,D 2,M 2}    => {D 1} 0.1538462          1 1.529412     4
## 11 {433W,A 5,D 1}    => {D 2} 0.1538462          1 1.040000     4
## 12 {433W,A 5,D 2}    => {D 1} 0.1538462          1 1.529412     4
```

Interpretation and Conclusion:

From the result of Association rules, we also found that D1 and D2 almost co-occurred in all the matches when winning teams play 433 formation, which is reasonable as the same reason mentioned in the previews case. Given this, we would suggest Roma to use these two types of players(D1,D2) when playing 433 formation. In terms of attackers and Mid-fielders, we found that Mid-fielder type 2 (M2), Mid-fielder type 5 (M5), and Attacker type 5 (A5) co-occurred 100% when D1 or D2 is being assigned in the match. Therefore, we suggest Roma to consider using players who are identified as these 3 types of players when playing 433 formation against AC Milan.

Final Takeaways for Roma

What have we found?

Data preparation for roma player cluster prediction:

In order to give a more specific recommendation to AS Roma on which players to assign, we use the clustering model built to predict which clusters Roma's current players belong to (based on each position: Defender, mid-fielder, attacker). Considering Roma is having a different playerbase, we searched and gathered all the players' information who are active in the most recent years (2016 and 2017) to offer a more realistic player assignment suggestion.(Since one player can have multiple roles during years of matches, we counted the most frequently played role for each player). This dataset is stored as a csv file call "Roma_16_17_players".

Summarzing strategies of each formation for Roma:

1. When Roma is considering using 4312 formation: M6,A1,A5

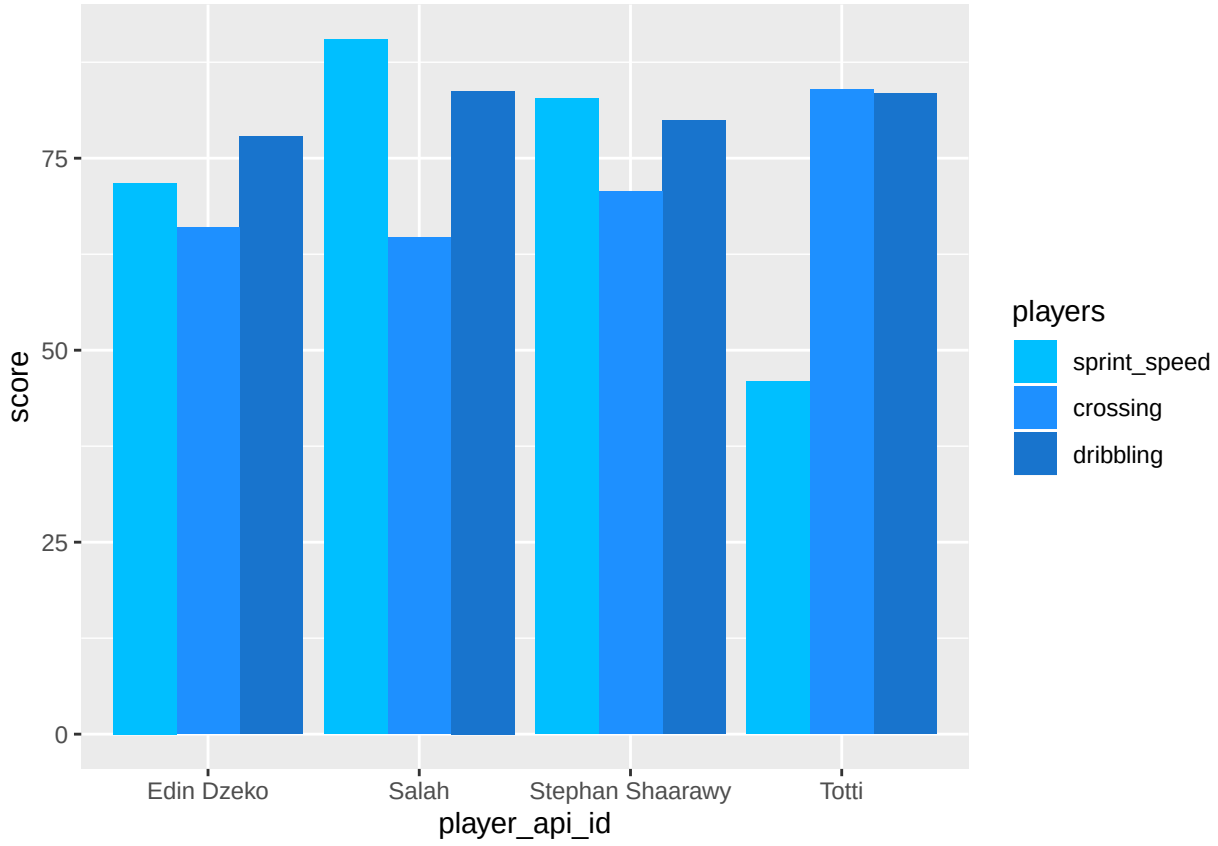
```
# Find all the players in Roma that's suggested for 433 and 4312 formations
roma_16_17_player <- read.csv('Roma_16_17_players.csv')
Roma_players_4312 <- roma_16_17_player  %>%
  mutate(Matches_played = Freq) %>%
  select(-c('X','X.1', 'player_fifa_api_id','id','Freq', 'birthday','Matches_played')) %>%
  filter(player_type %in% c('A-1','A-5','M-6'))

Roma_players_4312
```

```
##   player_id player_type         player_name height weight
## 1     15403         A-1          Edin Dzeko 193.04    185
## 2     30682         M-6    Daniele De Rossi 185.42    183
## 3     30714         A-1     Francesco Totti 180.34    181
## 4    161328         A-1 Stephan El Shaarawy 177.80    159
## 5    245572         A-5  Juan Manuel Iturbe 170.18    137
## 6    264842         A-1 Alessandro Florenzi 172.72    148
## 7    292462         A-1       Mohamed Salah 175.26    159
```

Inerpretation:

According to Roma's current squad, Daniele De Rossi and Juan Manuel Iturbe would be the recommended players if Roma is playing 4-3-1-2 against AC Milan based on their player types (group 6 in midfield and group 5 in attack). In addition, a group 1 attack player is needed, but there are a few players who are all recognized as this type of players. Based on the interpretation from the radar chart, these players are wide forwards who are lightning-quick and great dribblers with great crossing ability. Therefore, these skills are compared among group1 attack players in Roma (we exclude Alessandro Florenzi since he normally plays midfielder or fullback), and we are looking for players that have a relatively high and balanced skill set as the recommended candidate.

```
roma_a1_id = c(15403,30714,161328,292462)

roma_a1_player = player_atts %>%
  filter(player_api_id %in% roma_a1_id)

roma_a1_player_atts = roma_a1_player %>%
  group_by(player_api_id) %>%
  summarise_at(vars("sprint_speed","crossing","dribbling"),mean)

roma_a1_player_atts$player_api_id = as.factor(roma_a1_player_atts$player_api_id)
roma_a1_player_atts$player_api_id = c("Edin Dzeko","Totti","Stephan Shaarawy","Salah")

attributes <- c("sprint_speed","crossing","dribbling")

mydata <- melt(roma_a1_player_atts,id.vars="player_api_id",variable.name="players",value.name="score")

ggplot(mydata,aes(player_api_id,score,fill=players))+
  geom_bar(stat="identity",position="dodge") +
  scale_fill_manual(values = c( "deepskyblue","dodgerblue", "dodgerblue3"))
```



Inerpretation:

From the plot above, notice that Totti has a high score in crossing and dribbling, but his speed could be an disadvantage as a wide forward. Also, Salah receives the highest score in speed and dribbling, but his crossing score is the lowest among these players. Then, comparing the remaining two players (EdinDzeko and Stephan Shaarwy) who have a more balanced skillset, Stephan outscored Edin in all three skills. From a perspective of a more balanced and more skilled wide forward, Stephan Shaarawy would be the recommended player.

2. When Roma is considering using 433 formation:

```
# Find all the players in Roma that's suggested for 433 and 4312 formations
player = c(1,2,4,5,7,8)
Roma_players_433 <- roma_16_17_player  %>%
  mutate(Matches_played = Freq) %>%
  select(-c('X','X.1', 'player_fifa_api_id','id','Freq', 'birthday','Matches_played')) %>%
  filter(player_type %in% c('A-5','M-2','M-4'))
Roma_players_433 = Roma_players_433[player,]

Roma_players_433
```

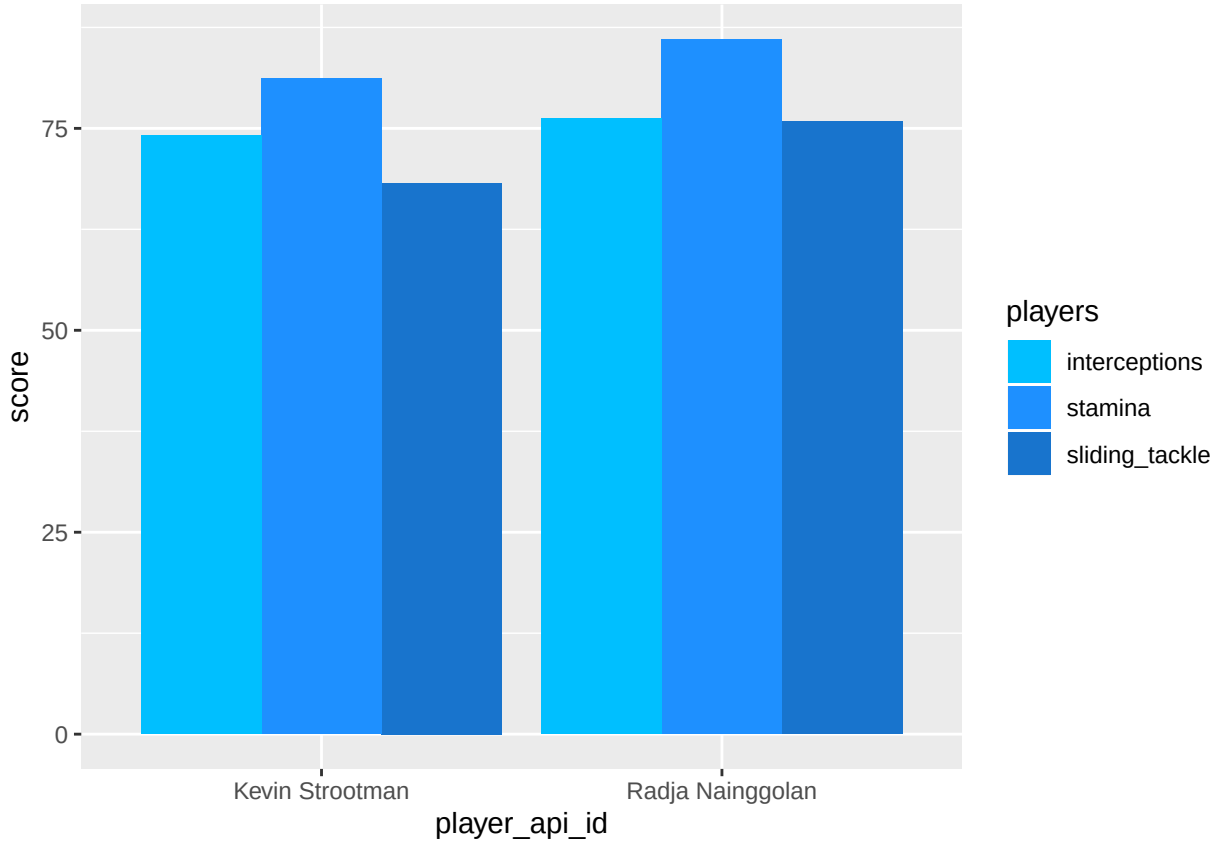
```
##   player_id player_type            player_name height weight
## 1     30714         M-4         Francesco Totti 180.34    181
## 2     41433         M-2        Radja Nainggolan 175.26    143
## 4    114558         M-2         Kevin Strootman 185.42    172
## 5    237606         M-4 Leandro Daniel Paredes 180.34    165
```

```
## 7    245572         A-5     Juan Manuel Iturbe 170.18    137
## 8    264842         M-4    Alessandro Florenzi 172.72    148
```

Inerpretation:

Juan Manuel Iturbe (group 5 in attack) would be the recommended player when Roma is playing 4-3-3 against AC Milan. Radja Nainggolan and Kevin Stroot man are classified as group 2 in midfield, which is recognized as enforcers. They play a role in winning the ball back from the opposition early in midfield. A strong enforcer is someone who plays aggressively with great skills in interception, tackle and stamina to shield team's defensive line.

```
roma_m2_id = c(41433,114558)

roma_m2_player = player_atts %>%
  filter(player_api_id %in% roma_m2_id)

roma_m2_player_atts = roma_m2_player %>%
  group_by(player_api_id) %>%
  summarise_at(vars("interceptions","stamina","sliding_tackle"),mean)

roma_m2_player_atts$player_api_id = as.factor(roma_m2_player_atts$player_api_id)
roma_m2_player_atts$player_api_id = c("Radja Nainggolan","Kevin Strootman")

attributes <- c("interceptions","stamina","tackles")


mydata <- melt(roma_m2_player_atts,id.vars="player_api_id",
               variable.name="players",value.name="score")

ggplot(mydata,aes(player_api_id,score,fill=players))+
  geom_bar(stat="identity",position="dodge") +
  scale_fill_manual(values = c( "deepskyblue","dodgerblue", "dodgerblue3"))
```



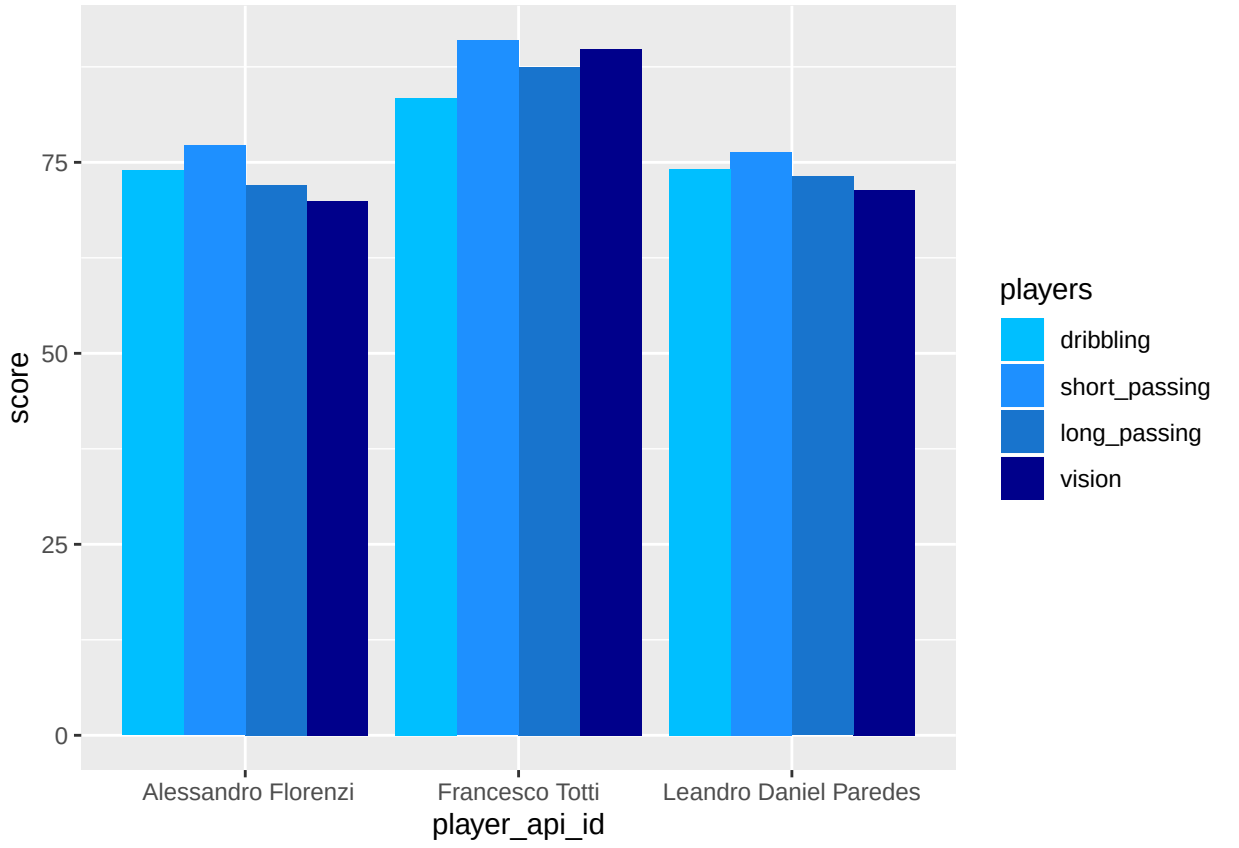
Inerpretation:

Although both players have a relatively balanced skill set in interception, stamina and tackles, Radja Nainggolan is slightly preferred since he outscores Kevin Strootman in all three attributes that we are interested in.

Lastly, the box-to-box midfielders (group 4 in midfield). These players are the engine of the team who are good at dribbling, passing with great vision on the field. Three players (Francesco Totti, Leandro Daniel Paredes and Alessandro Florenzi) show up in this position, and their skill sets are compared below.

```
roma_m4_id = c(30714,237606,264842)

roma_m4_player = player_atts %>%
  filter(player_api_id %in% roma_m4_id)

roma_m4_player_atts = roma_m4_player %>%
  group_by(player_api_id) %>%
  summarise_at(vars("dribbling","short_passing","long_passing","vision"),mean)

roma_m4_player_atts$player_api_id = as.factor(roma_m4_player_atts$player_api_id)
roma_m4_player_atts$player_api_id = c("Francesco Totti", "Leandro Daniel Paredes"," Alessandro Florenzi

attributes <- c("dribbling","short_passing","long_passing","vision")


mydata <- melt(roma_m4_player_atts,id.vars="player_api_id",
               variable.name="players",value.name="score")
```

```
ggplot(mydata,aes(player_api_id,score,fill=players))+
  geom_bar(stat="identity",position="dodge") +
  scale_fill_manual(values = c( "deepskyblue","dodgerblue", "dodgerblue3","dark blue"))
```



Inerpretation:

According to the plot, it is apparent that Francesco Totti is the desired player for the box-to-box midfielder. He outscored the other two players in every attribute that we compare. In addition, he joined in Roma in 1992 and already stayed for 25 years. He would be very familiar with the play-style that Roma typically play.

Final Recommendation

By performing association rules, we recognize a few formations AC Milan used to play, and we recommend a few formations that commonly beat AC Milan. Two of the most common winning formations (4-2-3-1 and 4-3-3) happen to be the common formation Roma is currently playing. Then, for each formation, we recommend some players based on their position and types from the cluster. In addition, for players with similar play style, we evaluate their attributes in different matches and pick the player with a higher and balanced skill set.

To be specific, if Roma decides to play 4-3-1-2 against AC Milan's 4-3-3, we would recommend De Rossi as attack midfielder and Juan Manuel Iturbe as the target man. In addition, for the wide forward, Stephan Shaarawy is the preferred player than Edin Dzeko and Salah, as it has a more balanced skill set and is good at crossing. Although this is not the recommendation for every player in Roma, this could be used as a suggestion for key positions and key players for a particular formation.

Finally, due to AC Milan's overall good performance in the Italian Series A, we do not find a lot of matches (in total of 26 matches) where they are beaten by a weak team. In order to get more information and make more informed recommendations about our opponent, we will need to collect more matches not only from the domestic league, but from the continential league, like the European championship as well.