Do skills collectively explain players in different positions? We have 34 skill features for every player. It's hard to look how players of different positions do in all 34 skills. We decided to use principal component analysis (PCA) without scaling to visualize them in reduced dimension.

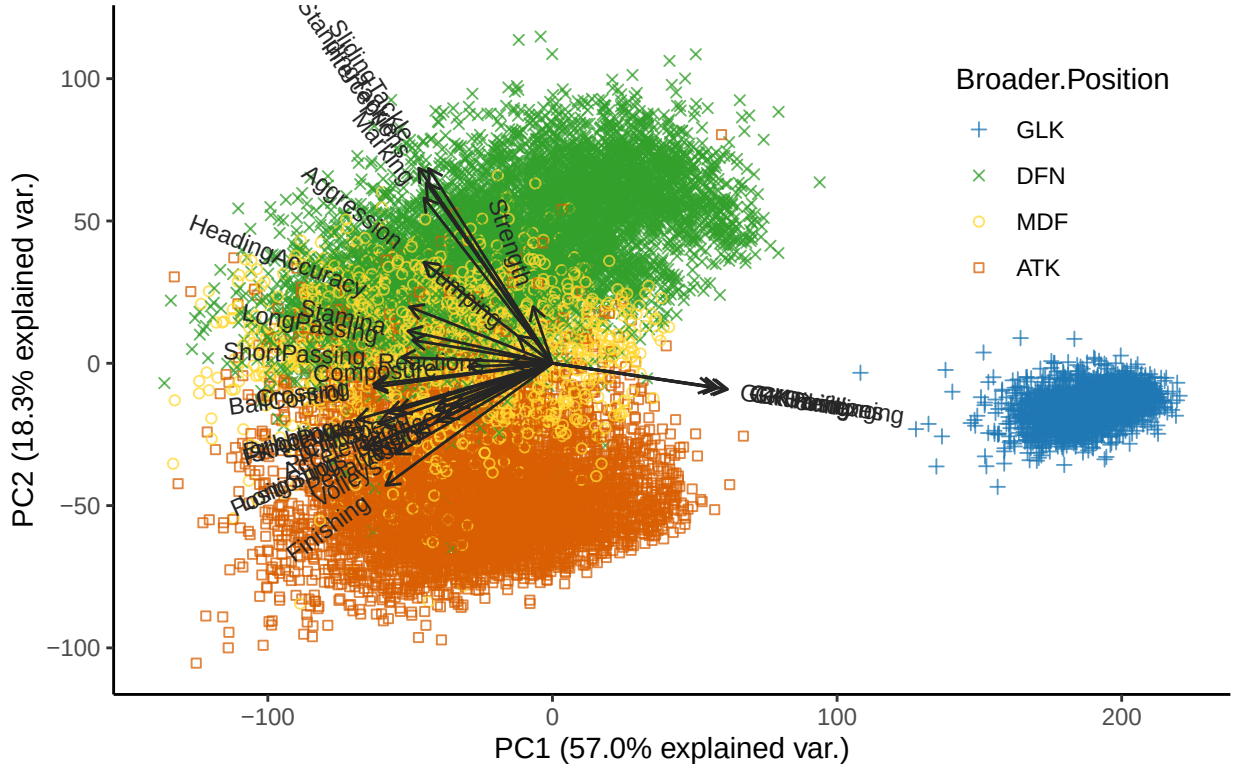
We also grouped all the players into 4 broader categories of positions based on correlations of their positional scores.

Biplot We try to look into the biplot with all the player's scores in first 2 PCs and the loadings of skills. The loadings show how strongly each skill influences the principal components.

```
setwd("D:/msc-ds/course-resource/data-visualization/project")
rm(list=ls())

library(RColorBrewer)
library(ggbiplot)

soccer.preprocessed <- read.csv(
  "soccer-preprocessed.csv",
  encoding = "UTF-8"
)

# Broader.Position
GLK <- c(
  "GK"
)
DFN <- c(
  "LB",
  "LCB",
  "CB",
  "RCB",
  "RB",
  "LWB",
  "LDM",
  "CDM",
  "RDM",
  "RWB"
)
MDF <- c(
  "LCM",
  "CM",
  "RCM"
)
ATK <- c(
  "LM",
  "RM",
  "LAM",
  "CAM",
  "RAM",
  "LW",
  "LF",
  "CF",
  "RF",
  "RW",
  "LS",
```

```
  "ST",
  "RS"
)

soccer.preprocessed$Broader.Position <- ifelse(
  soccer.preprocessed$Position %in% GLK, "GLK", ifelse(
    soccer.preprocessed$Position %in% DFN, "DFN", ifelse(
      soccer.preprocessed$Position %in% MDF, "MDF", "ATK"
    )
  )
)

soccer.preprocessed$Broader.Position <- factor(
  soccer.preprocessed$Broader.Position,
  levels=c("GLK","DFN","MDF", "ATK")
)
soccer.pca <- prcomp(soccer.preprocessed[, 46:79])

pallete.set2 <- brewer.pal(n=8, name="Set2")
pallete.paired <- brewer.pal(n=12, name="Paired")
pallete.dark2 <- brewer.pal(n=8, name="Dark2")
pallete.greys <- brewer.pal(n=9, name="Greys")

colors <- c(
  pallete.paired[2],
  pallete.paired[4],
  pallete.set2[6],
  pallete.dark2[2]
)
shapes <- c(3, 4, 21, 22)

soccer.pca.biplot <- ggbiplot(
  soccer.pca,
  obs.scale=1,
  var.scale=1,
  pc.biplot=T,
  col=c(pallete.greys[3], pallete.dark2[4]),
  groups=soccer.preprocessed$Broader.Position,
  alpha=0
) +
scale_color_manual(
  name="Broader.Position",
  values=colors
) +
scale_shape_manual(
  name="Broader.Position",
  values=shapes
) +
geom_point(
  aes(
    colour=soccer.preprocessed$Broader.Position,
    shape=soccer.preprocessed$Broader.Position
  ),
```

```
  size=1.5,
  alpha=0.75
) +
theme_classic() +
theme(
  legend.direction='vertical',
  legend.position=c(0.85, 0.75),
)

layer_arrows <- soccer.pca.biplot$layers[[1]]
layer_texts <- soccer.pca.biplot$layers[[3]]
layer_points <- soccer.pca.biplot$layers[[4]]

layer_arrows$aes_params$colour <- pallete.greys[8]
layer_texts$aes_params$colour <- pallete.greys[8]

soccer.pca.biplot$layers <- c(
  layer_points,
  layer_texts,
  layer_arrows
)
soccer.pca.biplot
```



Observations:

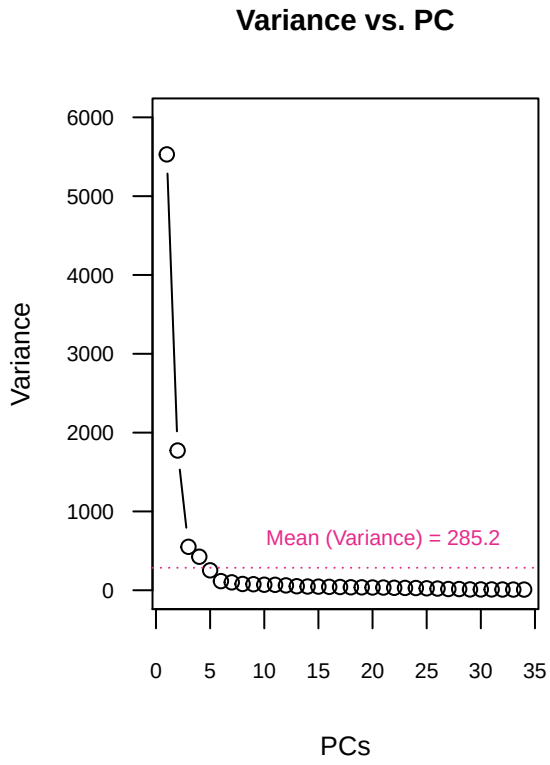
1. The first 2 PC axes explains 75% of total variance of the player's skills.

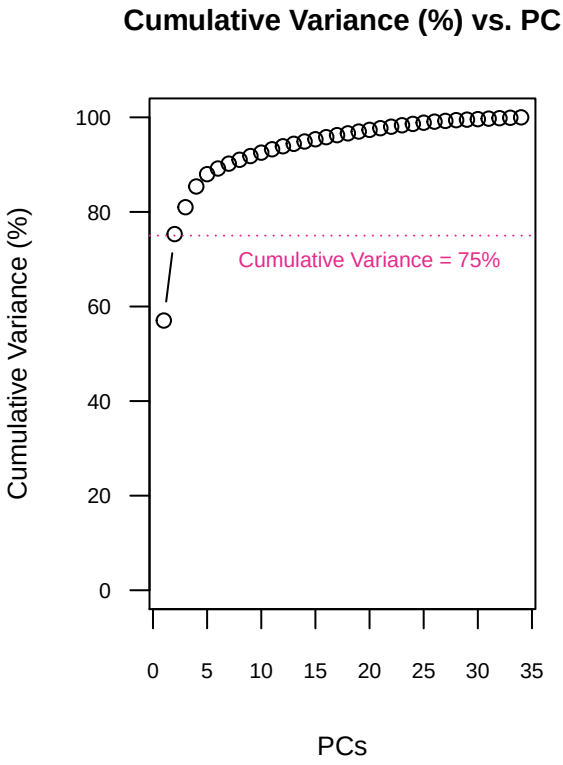
2. PC1 is strongly influenced by the goalkeeper specific skills, e.g. GKDiving, GKKicking, splitting the players into two major clusters goalkeepers and non goalkeepers.
3. PC2 summarises players playing in attacking, midfield and defending positions. Players in attacking positions have low scores in PC2 while defensive players have high scores on that axis. Midfielders scores in between these two positions.
4. We have some outliers but the average trend supports point 1 and 2.
5. Having almost zero angles between all goalkeeper skills indicates they are perfectly positively correlated and collectively explains goalkeepers space.
6. Skills like Standing Tackle, Sliding Tackle, Marking, Interception, Aggression, Strength with a small angles in between them explains their high positive correlations and also they are the skills that defenders share.
7. Finishing, Penalties, Volleys are some other skills with positive correlations that are most prevalent in players from attacking positions.

People with knowledge about soccer may find it easy to spot on the skills of players in different positions. But from principal component analysis of the soccer data and with biplot anyone can now relate these aspects.

```
setwd("D:/msc-ds/course-resource/data-visualization/project")
rm(list=ls())

library(RColorBrewer)

soccer.preprocessed <- read.csv(
  "soccer-preprocessed.csv",
  encoding="UTF-8"
)
soccer.ev <- eigen(
  cov(
    soccer.preprocessed[, 46:79]
  )
)$values

pallete.dark2 <- brewer.pal(n=8, name="Dark2")

mm <- rbind(c(1,2))
ww <- c(1,1)
hh <- c(10)
layout(mat=mm, widths=ww, heights=hh)

plot(
  soccer.ev,
  type="b",
  xlab="PCs",
  ylab="Variance",
  ylim=c(0,6000),
  main="Variance vs. PC",
  cex.main=0.90,
  cex.lab=0.80,
```

```
  cex.axis=0.70,
  las=1,
  xpd=T
)
abline(
  h=mean(soccer.ev),
  lty="dotted",
  col=pallete.dark2[4]
)
text(
  x=21,
  y=650,
  labels=paste0("Mean (Variance) = ", round(mean(soccer.ev), 1)),
  col=pallete.dark2[4],
  cex=0.70
)


plot(
  100*cumsum(soccer.ev)/sum(soccer.ev),
  type="b",
  xlab="PCs",
  ylab="Cumulative Variance (%)",
  ylim=c(0,100),
  main="Cumulative Variance (%) vs. PC",
  cex.main=0.90,
  cex.lab=0.80,
  cex.axis=0.70,
  las=1,
  xpd=T
)
abline(
  h=75,
  lty="dotted",
  col=pallete.dark2[4]
)
text(
  x=20,
  y=70,
  labels="Cumulative Variance = 75%",
  col=pallete.dark2[4],
  cex=0.70
)
```





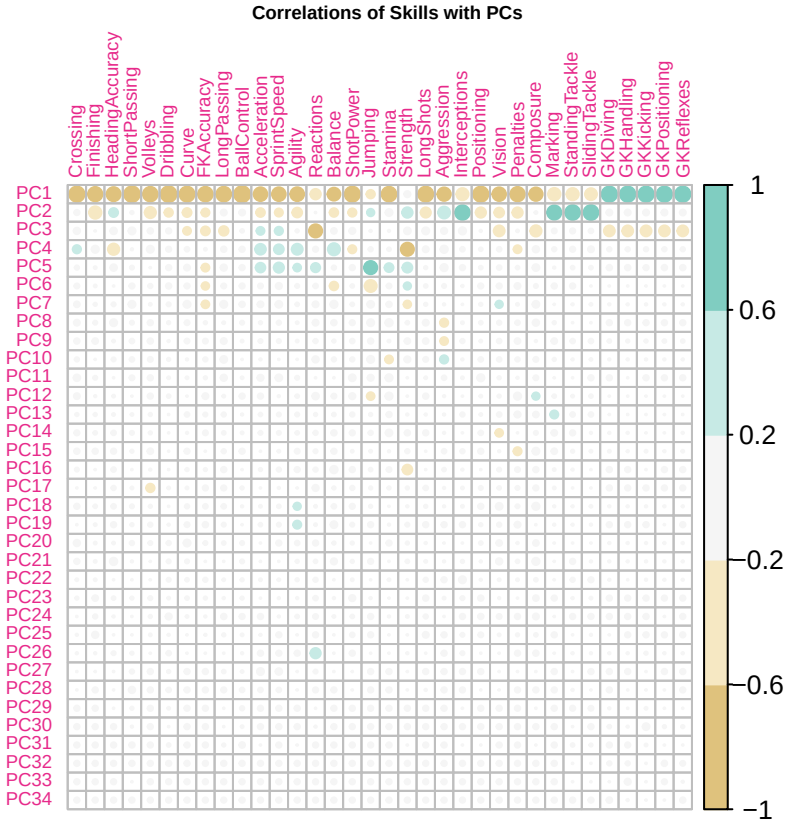
Scree Plots

Observations:

1. Mean Variance line from the variance scree plot suggests 5 PCs would be a sensible choice if we would need to do some further machine learning modeling like clustering.
2. Cumulative variance line from the R.H.S. plot suggests if we were to visualize the variance with simple 2-dimensional scatter plot we would still have 75% variance.

```
setwd("D:/msc-ds/course-resource/data-visualization/project")
rm(list=ls())

library(RColorBrewer)
library(corrplot)

soccer.preprocessed <- read.csv(
  "soccer-preprocessed.csv",
  encoding = "UTF-8"
)
soccer.pca <- prcomp(soccer.preprocessed[, 46:79])

pallete.dark2 <- brewer.pal(n=8, name="Dark2")
pallete.brbg <- brewer.pal(n=11, name="BrBG")

par(
```

```
  cex.main=0.60
)
corrplot(
  round(cor(soccer.pca$x[, 1:34], soccer.preprocessed[, 46:79]), 2),
  col=pallete.brbg[4:8],
  tl.cex=0.6,
  tl.col=pallete.dark2[4],
  mar=c(1, 1, 1, 1),
  title="Correlations of Skills with PCs"
)
```



Correlation Plots

Observations:

1. Apart from the visualisation of how skills are correlated with PCs, this correlation plot also reaffirms the fact that significant amount of variance of skill information can be explained only by a few PCs.