Load the Data, Explore Variables

We begin by loading in the data and taking a quick look at the variables that we'll be using in our PCA for this exercise. Pay attention to the packages in the library() statements, as they will be required for other elements of the code and you may need to install them!

```
> library(reshape2) #melt correlation matrix into 3 columns
> library(ggplot2) #correlation heatmap
> library(ggfortify) #autoplot bi-plot
> library(viridis) # magma palette
> library(plotrix) # color.legend
> load('fifa.RData')
> head(fifa)
```

```
              Name Age                                          Photo
1 Cristiano Ronaldo  32  https://cdn.sofifa.org/48/18/players/20801.png
2          L. Messi  30 https://cdn.sofifa.org/48/18/players/158023.png
3            Neymar  25 https://cdn.sofifa.org/48/18/players/190871.png
4         L. Suárez  30 https://cdn.sofifa.org/48/18/players/176580.png
5          M. Neuer  31 https://cdn.sofifa.org/48/18/players/167495.png
6    R. Lewandowski  28 https://cdn.sofifa.org/48/18/players/188545.png
  Nationality                                Flag Overall Potential
1    Portugal https://cdn.sofifa.org/flags/38.png      94        94
2   Argentina https://cdn.sofifa.org/flags/52.png      93        93
3      Brazil https://cdn.sofifa.org/flags/54.png      92        94
4     Uruguay https://cdn.sofifa.org/flags/60.png      92        92
5     Germany https://cdn.sofifa.org/flags/21.png      92        92
6      Poland https://cdn.sofifa.org/flags/37.png      91        91
                Club                                  Club.Logo  Value  Wage
1      Real Madrid CF https://cdn.sofifa.org/24/18/teams/243.png €95.5M €565K
2        FC Barcelona https://cdn.sofifa.org/24/18/teams/241.png  €105M €565K
3 Paris Saint-Germain  https://cdn.sofifa.org/24/18/teams/73.png  €123M €280K
4        FC Barcelona https://cdn.sofifa.org/24/18/teams/241.png   €97M €510K
5    FC Bayern Munich  https://cdn.sofifa.org/24/18/teams/21.png   €61M €230K
6    FC Bayern Munich  https://cdn.sofifa.org/24/18/teams/21.png   €92M €355K
  Special Acceleration Aggression Agility Balance Ball.control Composure
1    2228           89         63      89      63           93        95
2    2154           92         48      90      95           95        96
3    2100           94         56      96      82           95        92
4    2291           88         78      86      60           91        83
5    1493           58         29      52      35           48        70
6    2143           79         80      78      80           89        87
  Crossing Curve Dribbling Finishing Free.kick.accuracy GK.diving GK.handling
1       85    81        91        94                76         7          11
2       77    89        97        95                90         6          11
3       75    81        96        89                84         9           9
```

```
4       77    86        86        94                84        27          25
5       15    14        30        13                11        91          90
6       62    77        85        91                84        15           6
  GK.kicking GK.positioning GK.reflexes Heading.accuracy Interceptions Jumping
1         15              14          11               88            29      95
2         15              14           8               71            22      68
3         15              15          11               62            36      61
4         31              33          37               77            41      69
5         95              91          89               25            30      78
6         12               8          10               85            39      84
  Long.passing Long.shots Marking Penalties Positioning Reactions Short.passing
1           77         92      22        85          95        96            83
2           87         88      13        74          93        95            88
3           75         77      21        81          90        88            81
4           64         86      30        85          92        93            83
5           59         16      10        47          12        85            55
6           65         83      25        81          91        91            83
  Shot.power Sliding.tackle Sprint.speed Stamina Standing.tackle Strength
1         94             23           91      92              31       80
2         85             26           87      73              28       59
3         80             33           90      78              24       53
4         87             38           77      89              45       80
5         25             11           61      44              10       83
6         88             19           83      79              42       84
  Vision Volleys position
1     85      88        1
2     90      85        1
3     80      83        1
4     84      88        1
5     70      11        4
6     78      87        1
```

```
> summary(fifa[,13:46])
```

```
  Acceleration     Aggression       Agility         Balance       Ball.control
 Min.   :11.00   Min.   :11.00   Min.   :14.00   Min.   :11.00   Min.   : 8
 1st Qu.:56.00   1st Qu.:43.00   1st Qu.:55.00   1st Qu.:56.00   1st Qu.:53
 Median :67.00   Median :58.00   Median :65.00   Median :66.00   Median :62
 Mean   :64.48   Mean   :55.74   Mean   :63.25   Mean   :63.76   Mean   :58
 3rd Qu.:75.00   3rd Qu.:69.00   3rd Qu.:74.00   3rd Qu.:74.00   3rd Qu.:69
 Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :95
   Composure        Crossing        Curve        Dribbling       Finishing
 Min.   : 5.00   Min.   : 5.0   Min.   : 6.0   Min.   : 2.00   Min.   : 2.00
 1st Qu.:51.00   1st Qu.:37.0   1st Qu.:34.0   1st Qu.:48.00   1st Qu.:29.00
 Median :60.00   Median :54.0   Median :48.0   Median :60.00   Median :48.00
 Mean   :57.82   Mean   :49.7   Mean   :47.2   Mean   :54.94   Mean   :45.18
 3rd Qu.:67.00   3rd Qu.:64.0   3rd Qu.:62.0   3rd Qu.:68.00   3rd Qu.:61.00
 Max.   :96.00   Max.   :91.0   Max.   :92.0   Max.   :97.00   Max.   :95.00
 Free.kick.accuracy   GK.diving      GK.handling      GK.kicking
 Min.   : 4.00      Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
 1st Qu.:31.00      1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00
 Median :42.00      Median :11.00   Median :11.00   Median :11.00
 Mean   :43.08      Mean   :16.78   Mean   :16.55   Mean   :16.42
 3rd Qu.:57.00      3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
 Max.   :93.00      Max.   :91.00   Max.   :91.00   Max.   :95.00
 GK.positioning   GK.reflexes    Heading.accuracy Interceptions
 Min.   : 1.00   Min.   : 1.00   Min.   : 4.00    Min.   : 4.00
 1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.:44.00    1st Qu.:26.00
 Median :11.00   Median :11.00   Median :55.00    Median :52.00
 Mean   :16.54   Mean   :16.91   Mean   :52.26    Mean   :46.53
 3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:64.00    3rd Qu.:64.00
 Max.   :91.00   Max.   :90.00   Max.   :94.00    Max.   :92.00
    Jumping       Long.passing     Long.shots       Marking
 Min.   :15.00   Min.   : 7.00   Min.   : 3.00   Min.   : 4.00
```

```
1st Qu.:58.00   1st Qu.:42.00   1st Qu.:32.00   1st Qu.:22.00
Median :66.00   Median :56.00   Median :51.00   Median :48.00
Mean   :64.84   Mean   :52.37   Mean   :47.11   Mean   :44.09
3rd Qu.:73.00   3rd Qu.:64.00   3rd Qu.:62.00   3rd Qu.:63.00
Max.   :95.00   Max.   :93.00   Max.   :92.00   Max.   :92.00
  Penalties      Positioning      Reactions     Short.passing
Min.   : 5.00   Min.   : 2.00   Min.   :28.00   Min.   :10.00
1st Qu.:39.00   1st Qu.:38.00   1st Qu.:55.00   1st Qu.:53.00
Median :50.00   Median :54.00   Median :62.00   Median :62.00
Mean   :48.92   Mean   :49.53   Mean   :61.85   Mean   :58.22
3rd Qu.:61.00   3rd Qu.:64.00   3rd Qu.:68.00   3rd Qu.:68.00
Max.   :92.00   Max.   :95.00   Max.   :96.00   Max.   :92.00
  Shot.power    Sliding.tackle   Sprint.speed      Stamina
Min.   : 3.00   Min.   : 4.00   Min.   :11.00   Min.   :12.00
1st Qu.:46.00   1st Qu.:24.00   1st Qu.:57.00   1st Qu.:56.00
Median :59.00   Median :52.00   Median :67.00   Median :66.00
Mean   :55.57   Mean   :45.56   Mean   :64.72   Mean   :63.13
3rd Qu.:68.00   3rd Qu.:64.00   3rd Qu.:75.00   3rd Qu.:74.00
Max.   :94.00   Max.   :91.00   Max.   :96.00   Max.   :95.00
Standing.tackle    Strength         Vision         Volleys
Min.   : 4.00   Min.   :20.00   Min.   :10.00   Min.   : 4.00
1st Qu.:26.00   1st Qu.:58.00   1st Qu.:43.00   1st Qu.:30.00
Median :54.00   Median :66.00   Median :54.00   Median :44.00
Mean   :47.41   Mean   :65.24   Mean   :52.93   Mean   :43.13
3rd Qu.:66.00   3rd Qu.:74.00   3rd Qu.:64.00   3rd Qu.:57.00
Max.   :92.00   Max.   :98.00   Max.   :94.00   Max.   :91.00
```

These variables are scores on the scale of [0,100] that measure 34 key abilities of soccer players. No player has ever earned a score of 100 on any of these attributes - no player is *perfect*!

It would be natural to assume some correlation between these variables and indeed, we see lots of it: What jumps out right away are the "GK" (Goal Keeping) abilities - these attributes have *very* strong positive correlation with one another and negative correlation with the other abilities. After all, goal keepers are not traditionally well known for their dribbling, passing, and finishing abilities!

Outside of that, we see a lot of red in this correlation matrix – many attributes share a lot of information. This is the type of situation where PCA shines.

```
> cor.matrix = cor(fifa[,13:46])
> cor.matrix = melt(cor.matrix)
> ggplot(data = cor.matrix, aes(x=Var1, y=Var2, fill=value)) +
+   geom_tile(color = "white")+
+   scale_fill_gradient2(low = "blue", high = "red", mid = "white",
+    midpoint = 0, limit = c(-1,1), space = "Lab",
+     name="Correlation") +  theme_minimal()+
+       theme(axis.title.x = element_blank(),axis.title.y = element_blank(),
+             axis.text.x = element_text(angle = 45, vjust = 1,
+         size = 9, hjust = 1))+coord_fixed()
```

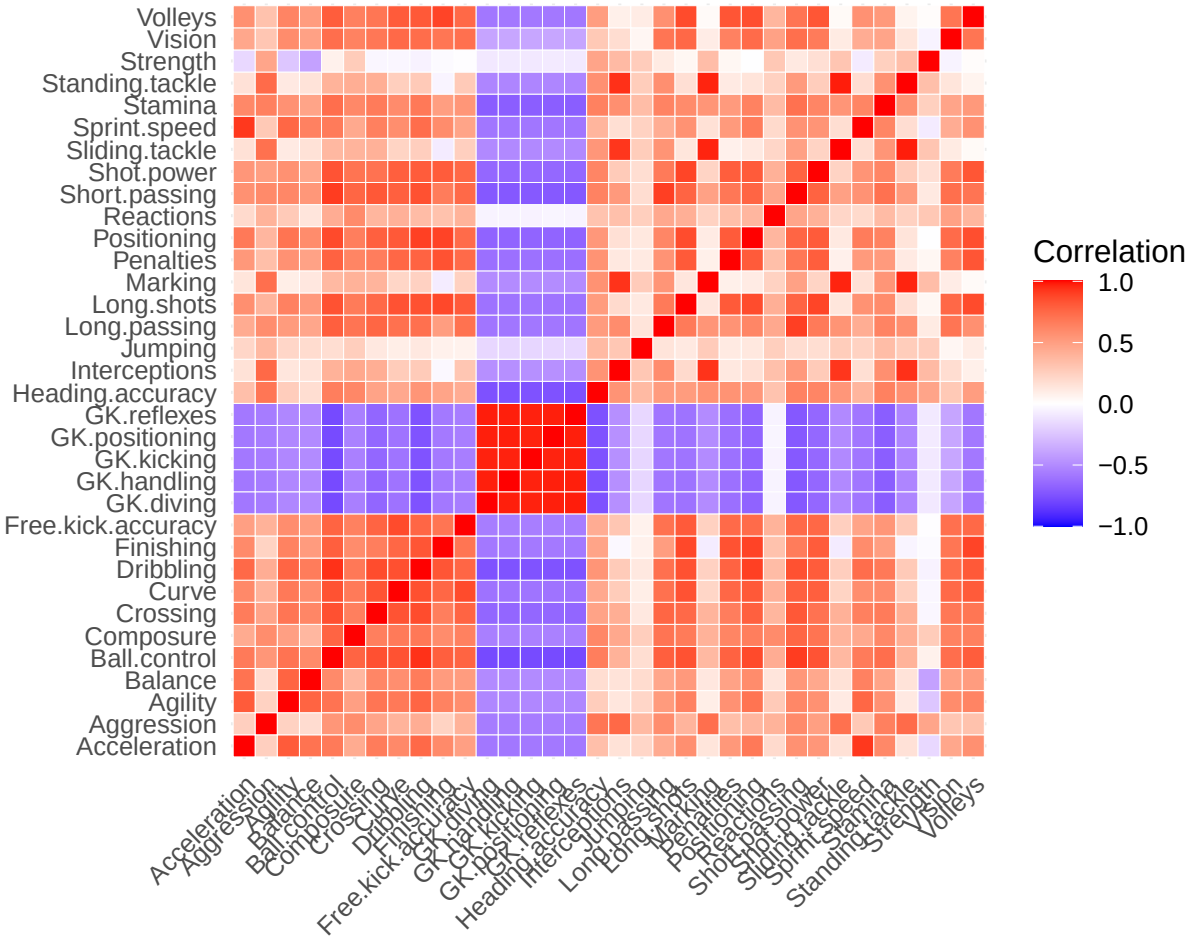


Figure 1: Heatmap of correlation matrix for 34 variables of interest

Principal Components Analysis

Let's take a look at the principal components analysis. Since the variables are on the same scale, I'll start with **covariance PCA** (the default in R's prcomp() function.

```
> fifa.pca = prcomp(fifa[,13:46] )
> summary(fifa.pca)
```

```
Importance of components:
                           PC1     PC2      PC3      PC4      PC5      PC6
Standard deviation     74.8371 43.5787 23.28767 20.58146 16.12477 10.71539
Proportion of Variance  0.5647  0.1915  0.05468  0.04271  0.02621  0.01158
Cumulative Proportion   0.5647  0.7561  0.81081  0.85352  0.87973  0.89131
                           PC7     PC8     PC9   PC10    PC11    PC12    PC13
Standard deviation     10.17785 9.11852 8.98065 8.5082 8.41550 7.93741 7.15935
Proportion of Variance  0.01044 0.00838 0.00813 0.0073 0.00714 0.00635 0.00517
Cumulative Proportion   0.90175 0.91013 0.91827 0.9256 0.93270 0.93906 0.94422
                          PC14    PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     7.06502 6.68497 6.56406 6.50459 6.22369 6.08812 6.00578
Proportion of Variance 0.00503 0.00451 0.00434 0.00427 0.00391 0.00374 0.00364
Cumulative Proportion  0.94926 0.95376 0.95811 0.96237 0.96628 0.97001 0.97365
                          PC21    PC22    PC23    PC24    PC25    PC26   PC27
Standard deviation     5.91320 5.66946 5.45018 5.15051 4.86761 4.34786 4.1098
Proportion of Variance 0.00353 0.00324 0.00299 0.00267 0.00239 0.00191 0.0017
Cumulative Proportion  0.97718 0.98042 0.98341 0.98609 0.98848 0.99038 0.9921
                          PC28    PC29    PC30    PC31   PC32    PC33    PC34
Standard deviation     4.05716 3.46035 3.37936 3.31179 3.1429 3.01667 2.95098
Proportion of Variance 0.00166 0.00121 0.00115 0.00111 0.0010 0.00092 0.00088
Cumulative Proportion  0.99374 0.99495 0.99610 0.99721 0.9982 0.99912 1.00000
```

```
> # Loadings on first 3 components:
> fifa.pca$rotation[,1:3]
```

```
                            PC1           PC2          PC3
Acceleration       -0.13674335  0.0944478107 -0.141193842
Aggression         -0.15322857 -0.2030537953  0.105372978
Agility            -0.13598896  0.1196301737 -0.017763073
Balance            -0.11474980  0.0865672989 -0.072629834
Ball.control       -0.21256812  0.0585990154  0.038243802
Composure          -0.13288575 -0.0005635262  0.163887637
Crossing           -0.21347202  0.0458210228  0.124741235
Curve              -0.20656129  0.1254947094  0.180634730
Dribbling          -0.23090613  0.1259819707 -0.002905379
Finishing          -0.19431248  0.2534086437  0.006524693
Free.kick.accuracy -0.18528508  0.0960404650  0.219976709
GK.diving           0.20757999  0.0480952942  0.326161934
GK.handling         0.19811125  0.0464542553  0.314165622
GK.kicking          0.19261876  0.0456942190  0.304722126
GK.positioning      0.19889113  0.0456384196  0.317850121
GK.reflexes         0.21081755  0.0489895700  0.332751195
Heading.accuracy   -0.17218607 -0.1115416097 -0.125135161
Interceptions      -0.15038835 -0.3669025376  0.162064432
Jumping            -0.03805419 -0.0579221746  0.012263523
Long.passing       -0.16849827 -0.0435009943  0.224584171
Long.shots         -0.21415526  0.1677851237  0.157466462
Marking            -0.14863254 -0.4076616902  0.078298039
Penalties          -0.16328049  0.1407803994  0.024403976
Positioning        -0.22053959  0.1797895382  0.020734699
Reactions          -0.04780774  0.0001844959  0.250247098
Short.passing      -0.18176636 -0.0033124240  0.118611543
Shot.power         -0.19592137  0.0989340925  0.101707386
```

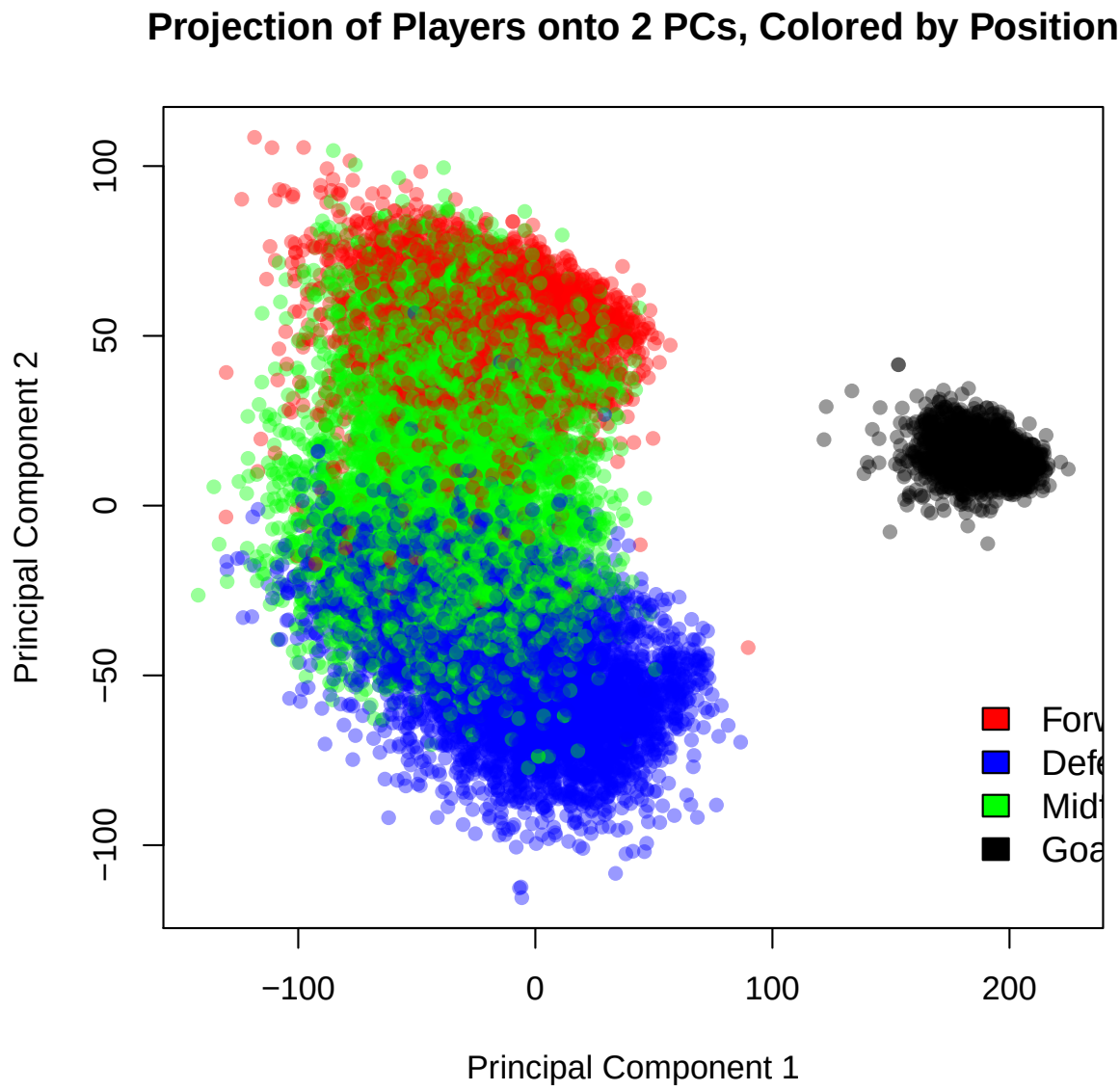
```
Sliding.tackle     -0.14977558 -0.4024030355  0.069945935
Sprint.speed       -0.13387287  0.0804847541 -0.146049405
Stamina            -0.17231648 -0.0634639786 -0.016509650
Standing.tackle    -0.15992073 -0.4039763876  0.086418583
Strength           -0.02186264 -0.1151018222  0.096053864
Vision             -0.13027169  0.1152237536  0.260985686
Volleys            -0.18465028  0.1888480712  0.076974579
```

It's clear we can capture a large amount of the variance in this data with just a few components. In fact just **2 components yield 76% of the variance!**

Now let's look at some projections of the players onto those 2 principal components. The scores are located in the fifa.pca$x matrix.

The plot easily separates the field players from the goal keepers, and the forwards from the defenders. As one might

```
> plot(fifa.pca$x[,1],fifa.pca$x[,2], col=alpha(c('red','blue','green','black')[as.factor(fifa$position)],0.4), pch=16, xl
> legend(175,-50, c('Forward','Defense','Midfield','GoalKeeper'), c('red','blue','green','black'), bty = 'n', cex=1.1)
```



expect, midfielders are sandwiched by the forwards and defenders, as they play both roles on the field. The labelling of player position was imperfect and done using a list of the players' preferred positions, and it's likely we are seeing that in some of the players labeled as midfielders that appear above the cloud of red points.

The BiPlot

BiPlots can be tricky when we have so much data and so many variables. As you will see, the default image leaves much to be desired, and will motivate our move to the ggfortify library to use the autoplot() function. The image takes too long to render and is practically unreadable with the whole dataset, so I demonstrate the default biplot() function with a sample of the observations.

```
> biplot(fifa.pca$x[sample(1:16501,2000),],fifa.pca$rotation[,1:2], cex=0.5, arrow.len = 0.1)
```

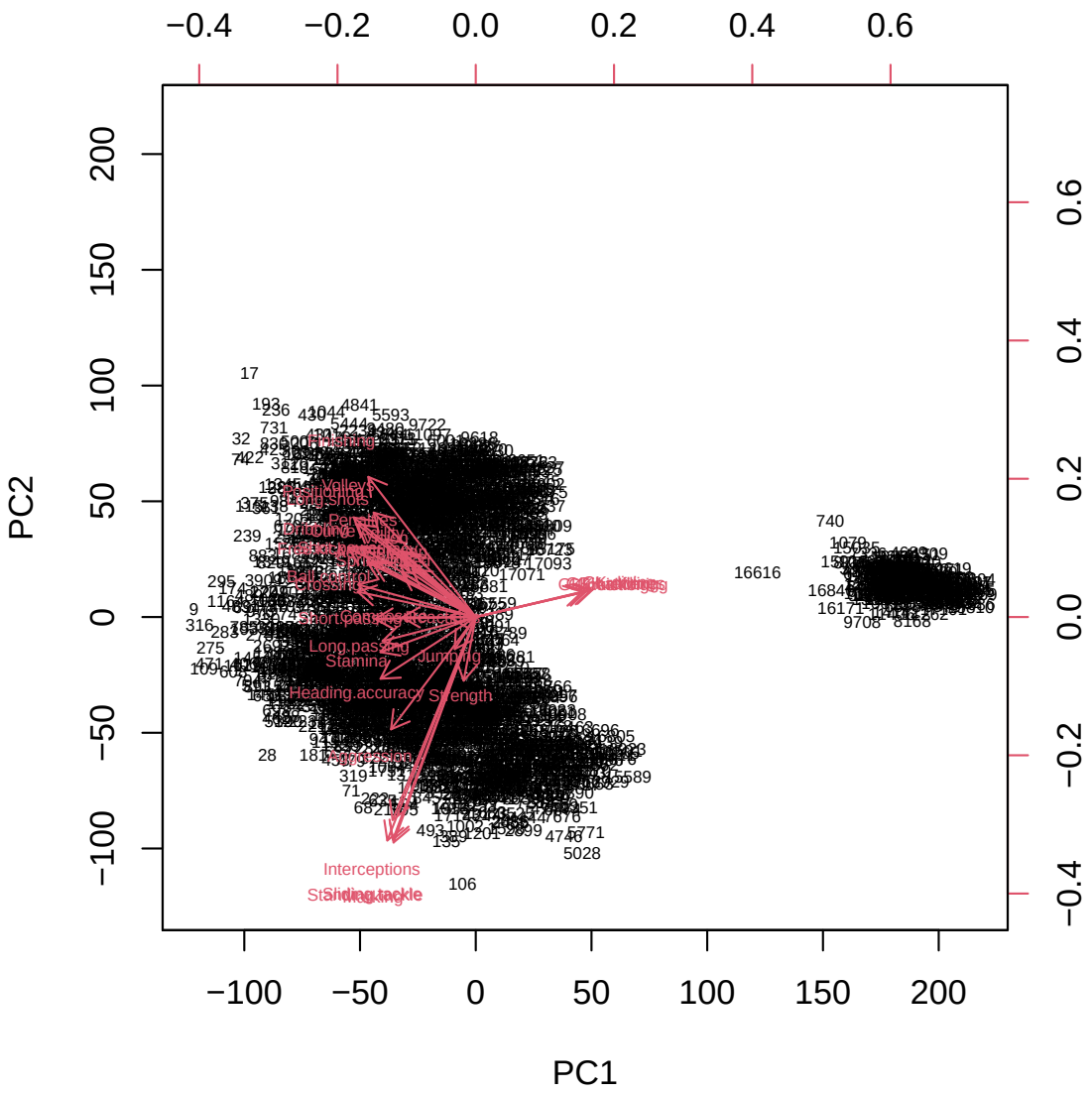


Figure 2: Default biplot function is not great here

The autoplot function uses the ggplot2 package and is superior when we have more data.

Many expected conclusions can be drawn from this biplot. The defenders tend to have stronger skills of *interception*, *slide tackling*, *standing tackling*, and *marking*, while forwards are generally stronger when it comes to *finishing*, *long.shots*, *volleys*, *agility* etc. Midfielders are likely to be stronger with *crossing*, *passing*, *ball.control*, and *stamina*.

```
> autoplot(fifa.pca, data = fifa,
+          colour = alpha(c('red','blue','green','orange')[as.factor(fifa$pos)],0.4),
+          loadings = TRUE, loadings.colour = 'black',
+          loadings.label = TRUE, loadings.label.size = 3.5, loadings.label.alpha = 1,
+          loadings.label.fontface='bold',
+          loadings.label.colour = 'black',
+          loadings.label.repel=T)
```

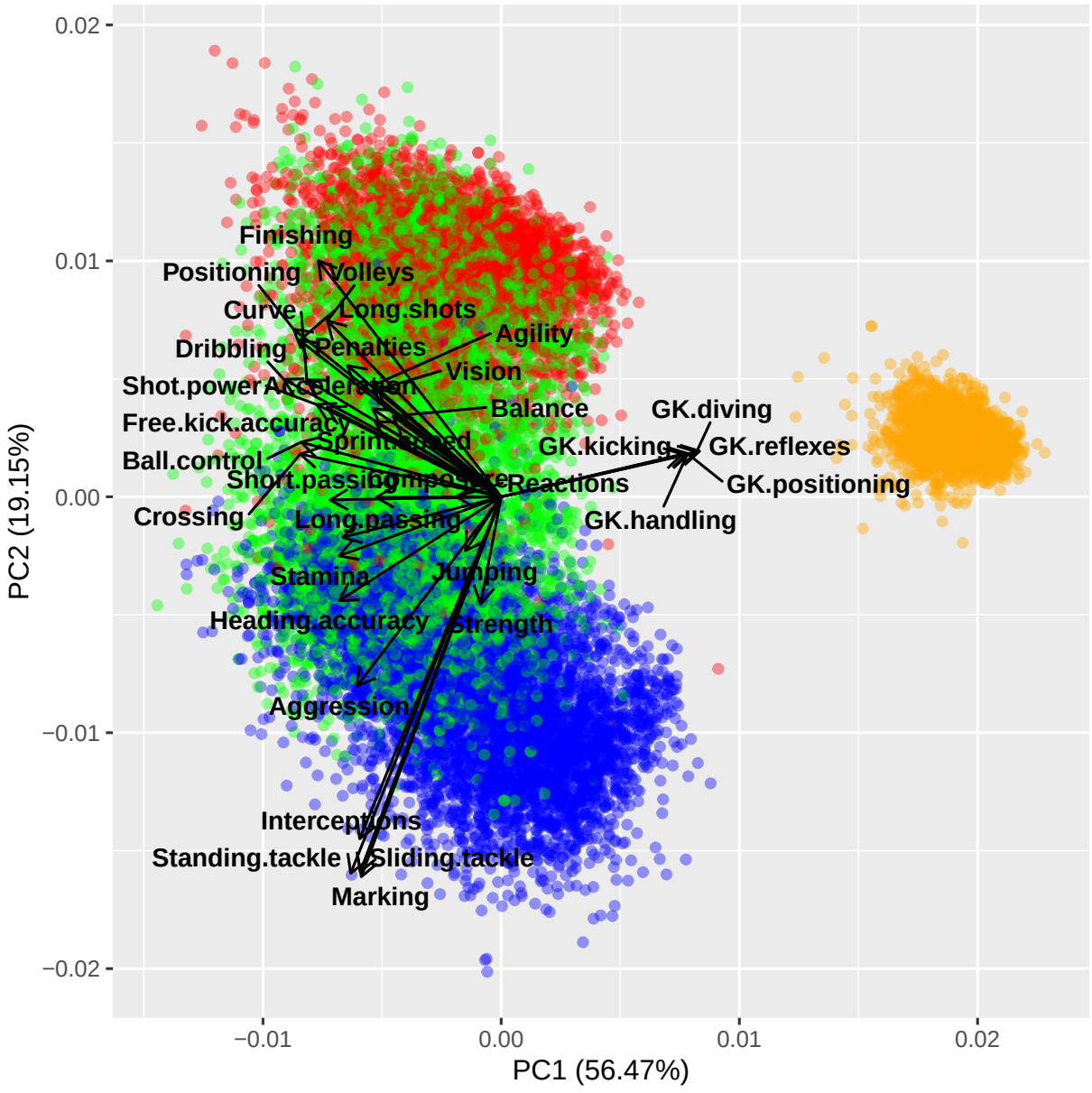
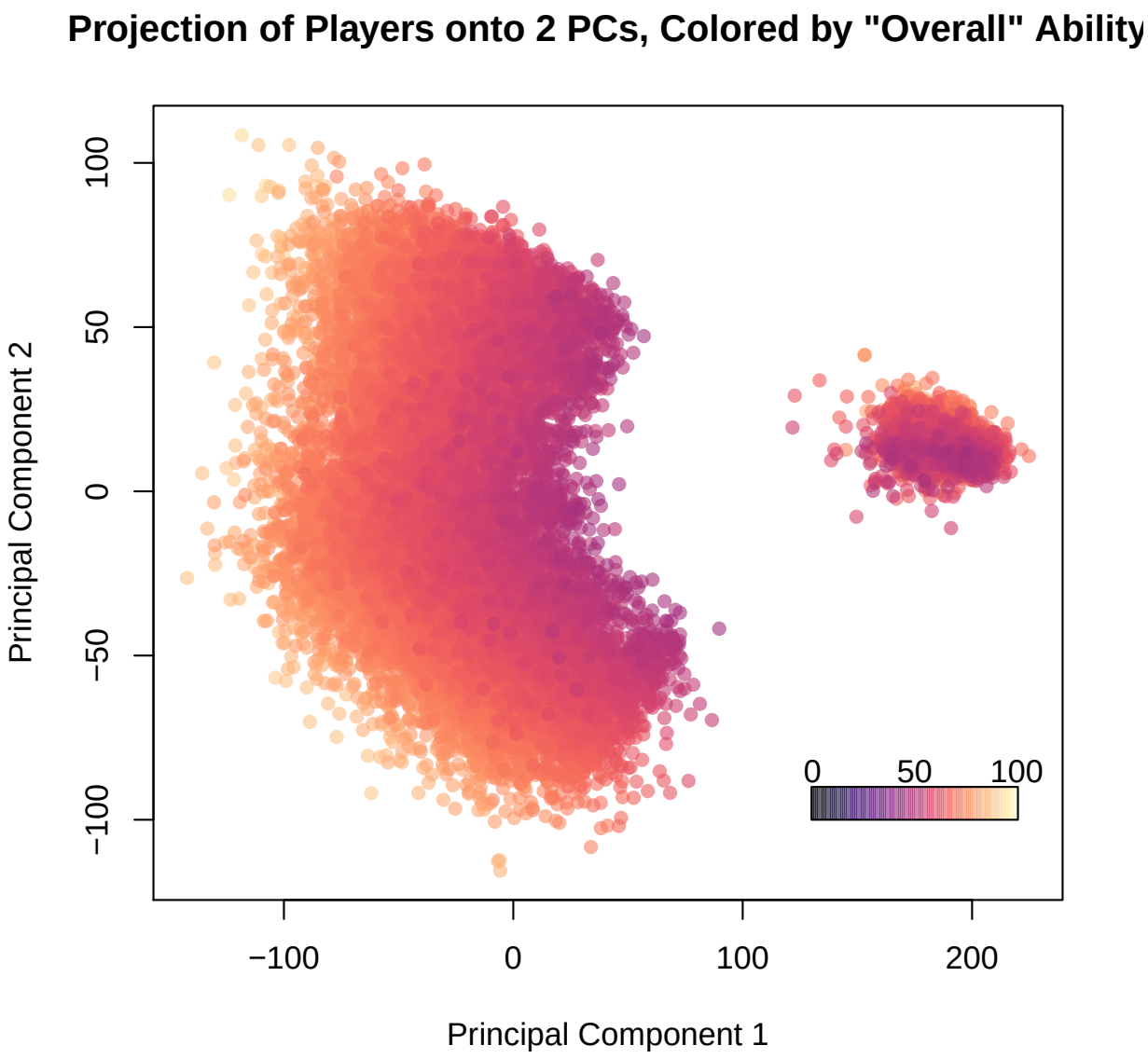


Figure 3: autoplot() biplot functionality has many more options for readability.

Further Exploration

Let's see what happens if we color by the variable 'overall' which is designed to rank a player's overall quality of play.

```
> palette(alpha(magma(100),0.6))
> plot(fifa.pca$x[,1],fifa.pca$x[,2], col=fifa$Overall,pch=16, xlab = 'Principal Component 1', ylab='Principal Component 2
> color.legend(130,-100,220,-90,seq(0,100,50),alpha(magma(100),0.6),gradient="x")
```



We can attempt to label some of the outliers, too. First, we'll look at the 0.001 and 0.999 quantiles to get a sense of what coordinates we want to highlight. Then we'll label any players outside of those bounds and surely find some familiar names.

What about by wage? First we need to convert their salary, denominated in Euros, to a numeric variable.

```
> # This first chunk is identical to the chunk above. I have to reproduce the plot to label it.
> palette(alpha(magma(100),0.6))
> plot(fifa.pca$x[,1], fifa.pca$x[,2], col=fifa$Overall,pch=16, xlab = 'Principal Component 1', ylab='Principal Component
+      xlim=c(-175,250), ylim = c(-150,150))
> color.legend(130,-100,220,-90,seq(0,100,50),alpha(magma(100),0.6),gradient="x")
> # Identify quantiles (high/low) for each PC
> (quant1h = quantile(fifa.pca$x[,1],0.9997))
```

```
  99.97%
215.4003
```

```
> (quant1l = quantile(fifa.pca$x[,1],0.0003))
```

```
    0.03%
-130.1493
```

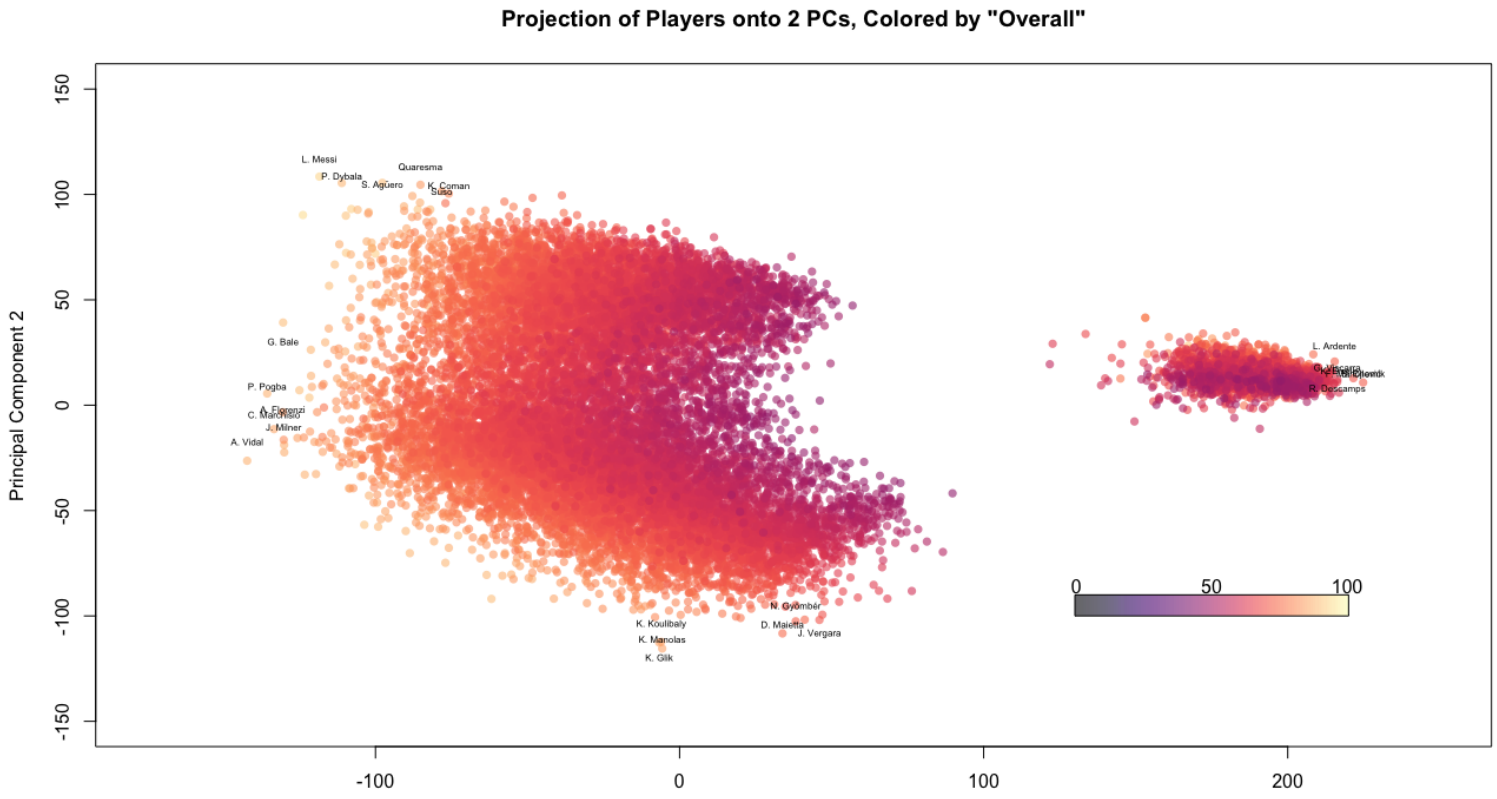
```
> (quant2h = quantile(fifa.pca$x[,2],0.9997))
```

```
 99.97%
100.208
```

```
> (quant2l = quantile(fifa.pca$x[,2],0.0003))
```

```
    0.03%
-101.8846
```

```
> # Next I create a logical vector which identifies the outliers
> # (i.e. TRUE = outlier, FALSE = not outlier)
> outliers = fifa.pca$x[,1] > quant1h | fifa.pca$x[,1] < quant1l |
+                   fifa.pca$x[,2] > quant2h | fifa.pca$x[,2] < quant2l
> # Here I label them by name, jittering the coordinates of the text so it's more readable
> text(jitter(fifa.pca$x[outliers,1],factor=1), jitter(fifa.pca$x[outliers,2],factor=600), fifa$Name[outliers], cex=0.7)
```



```
> # First, observe the problem with the Wage column as it stands
> head(fifa$Wage)
```

```
[1] "€565K" "€565K" "€280K" "€510K" "€230K" "€355K"
```

```
> # Use regular expressions to remove the Euro sign and K from the wage column
> # then covert to numeric
> fifa$Wage = as.numeric(gsub('[€K]', '', fifa$Wage))
> # new data:
> head(fifa$Wage)
```

```
[1] 565 565 280 510 230 355
```

```
> palette(alpha(magma(100),0.6))
> plot(fifa.pca$x[,1], fifa.pca$x[,2], col=fifa$Wage,pch=16, xlab = 'Principal Component 1', ylab='Principal Component 2',
> color.legend(130,-100,220,-90,c(min(fifa$Wage),max(fifa$Wage)),alpha(magma(100),0.6),gradient="x")
```