Assessing Video Game Mechanics

Danya Zhang

2023-08-14

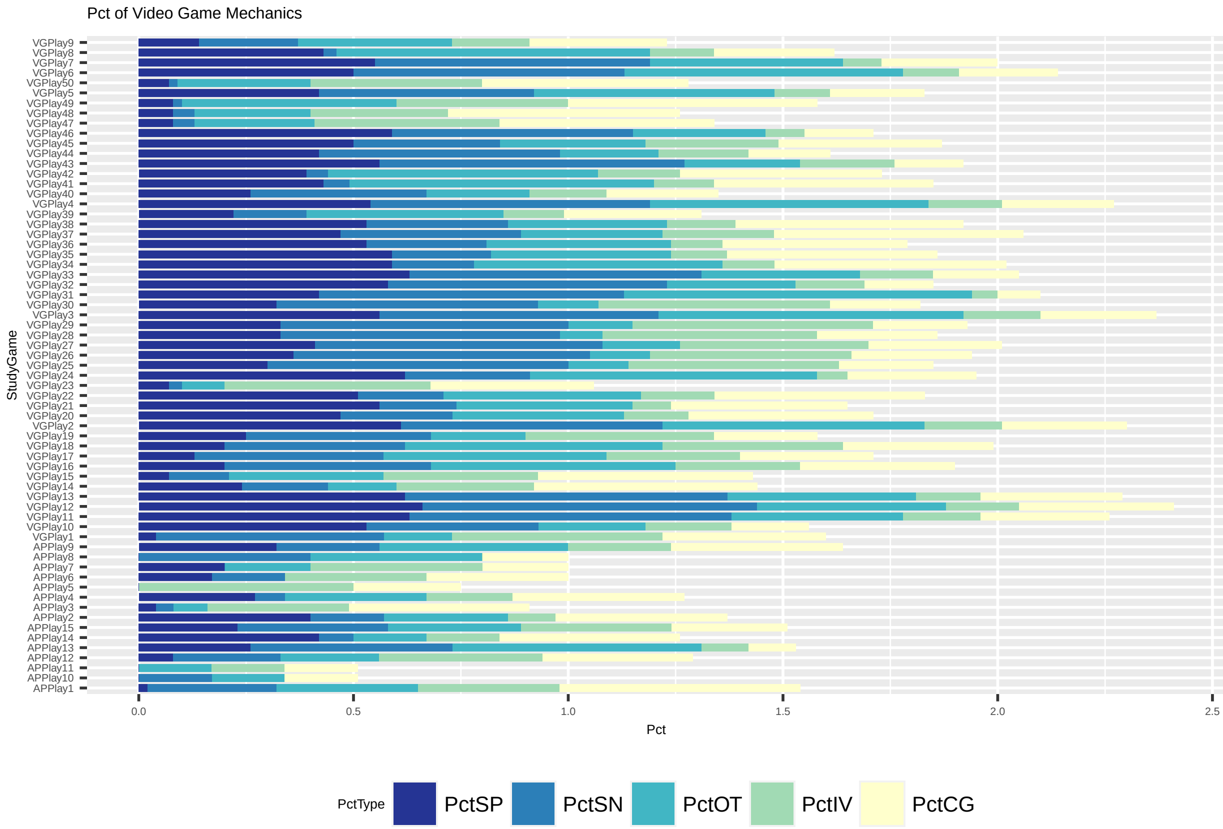
Visualizations

```r
# read in the data
game_df <- read_excel("TopSelling_MSSP_July28_GameData.xlsx")
participant_df <- read_excel("TopSelling_MSSP_July28_ParticipantData.xlsx")

# data cleaning: change unlist all columns in game_df
game_df <- as.data.frame(game_df)
participant_df <- as.data.frame(participant_df)
```

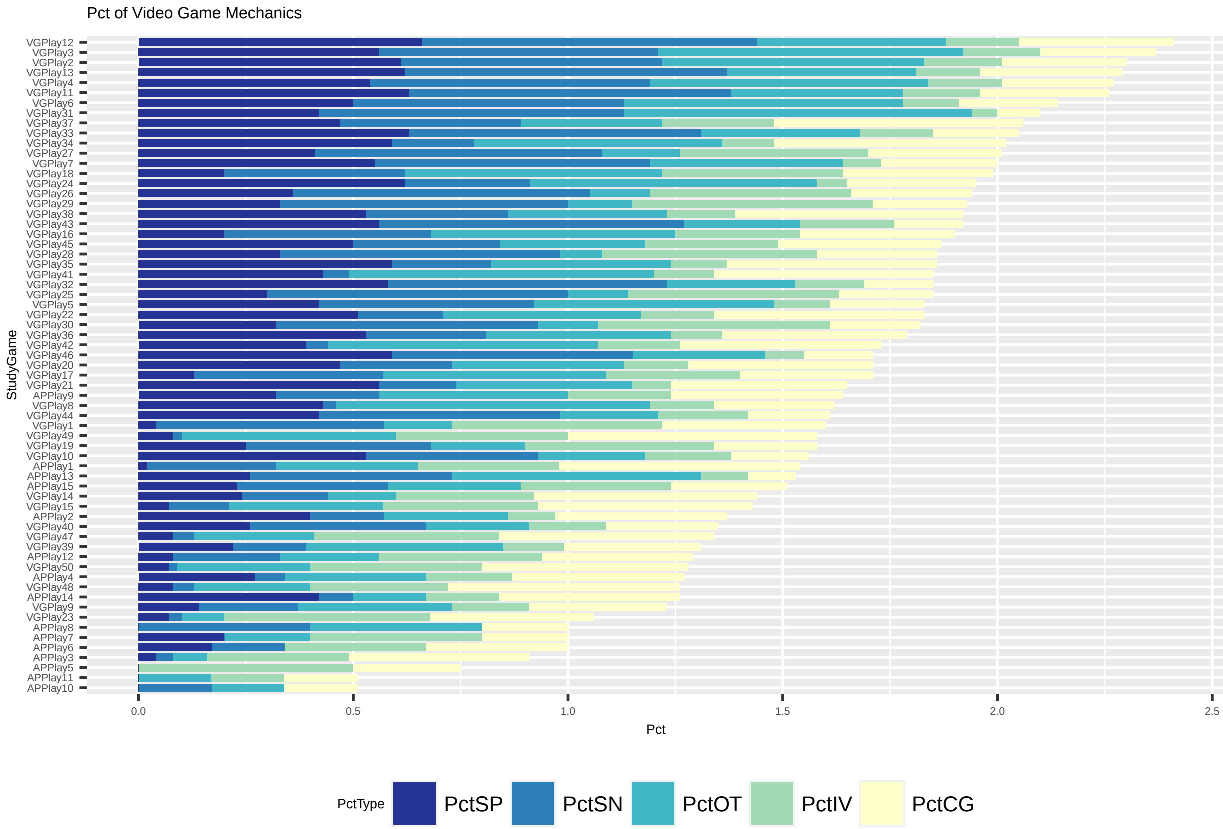
```r
# create a subset of the game_df with just Num variables
game_df_pct_subset_longer <- game_df %>%
  dplyr::select(StudyGame,GameName,NumFam,PctOT,PctIV,PctSN,PctSP,PctCG) %>%
  pivot_longer(cols = c(PctOT,PctIV,PctSN,PctSP,PctCG),
               names_to = "PctType",
               values_to = "Pct")
```

```r
# stacked bar chart for the 5 game mechanics
ggplot(data = game_df_pct_subset_longer,
       aes(x = StudyGame, y = Pct, fill = PctType)) +
  geom_bar(position="stack", stat="identity", width = 0.75) +
  coord_flip() +
  scale_fill_brewer(palette = "YlGnBu") +
  theme(text = element_text(size=5),
        legend.position = "bottom", legend.box = "horizontal",
        legend.text = element_text(size = 8)) +
  xlab("StudyGame") +
  labs(title = "Pct of Video Game Mechanics") +
  guides(fill = guide_legend(reverse = TRUE))
```



The graph above shows the percent of each mechanic that was selected for each game.

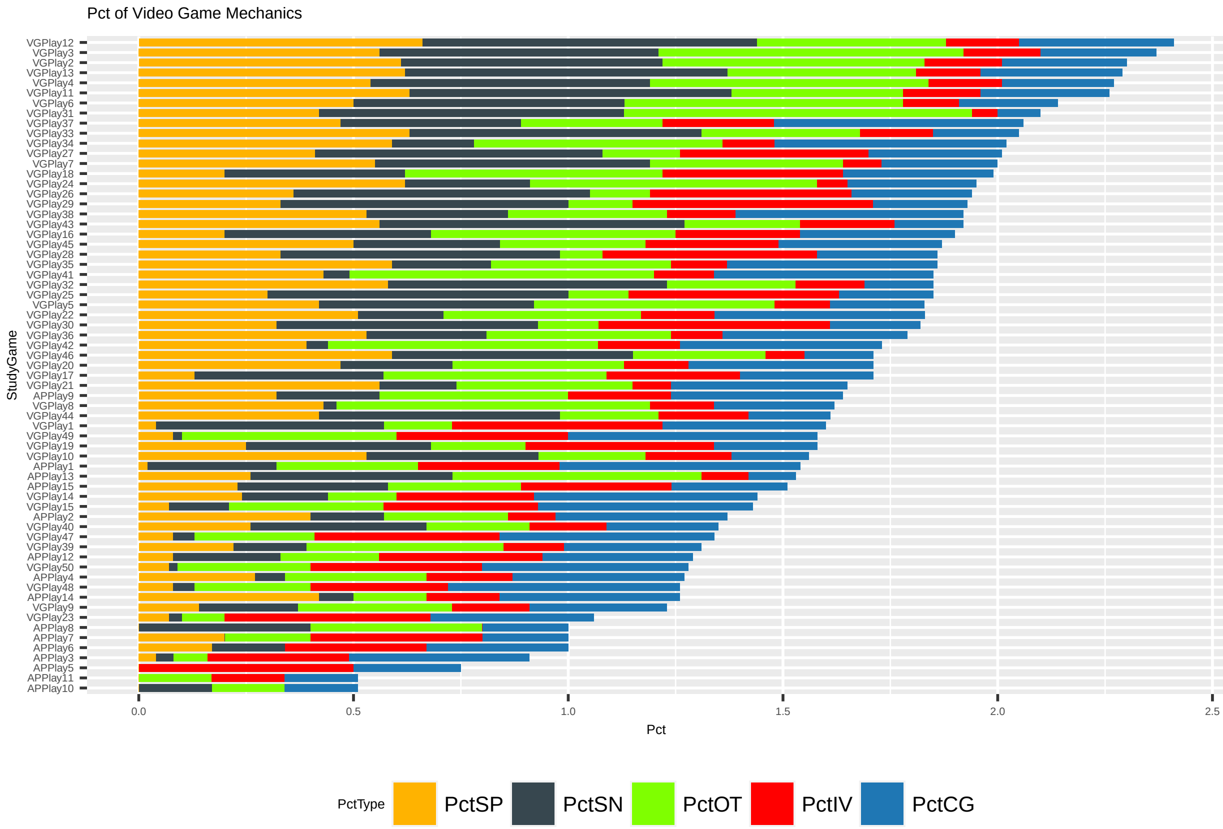
```r
# order by length of bars
ggplot(data = game_df_pct_subset_longer,
      aes(x = reorder(StudyGame, Pct, sum), y = Pct, fill = PctType)) +
  geom_bar(position="stack", stat="identity", width = 0.75) +
  coord_flip() +
  scale_fill_brewer(palette = "YlGnBu") +
  theme(text = element_text(size=5),
        legend.position = "bottom", legend.box = "horizontal",
        legend.text = element_text(size = 8)) +
  xlab("StudyGame") +
  labs(title = "Pct of Video Game Mechanics") +
  guides(fill = guide_legend(reverse = TRUE))
```



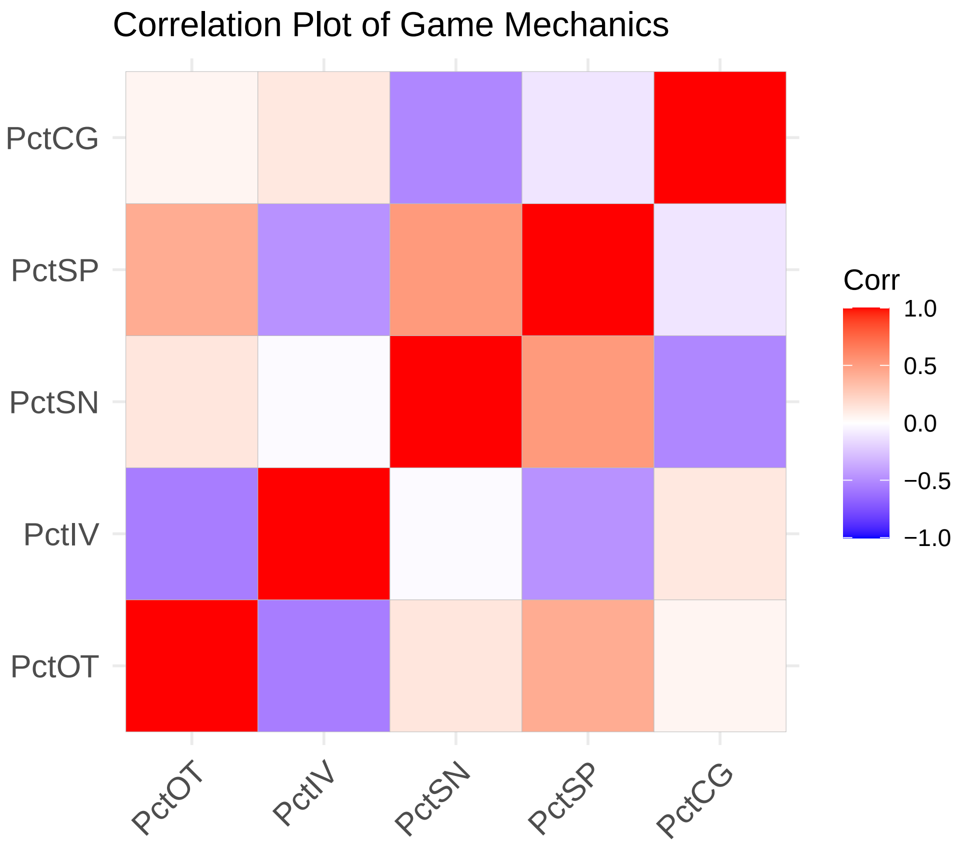
We order the bars by length for easy reading.

```
# manually changing the color scheme
# you can use hexcodes OR
# R recognized color names
custom_colors <- c("#1f77b4", "red", "chartreuse1", "#37474f", "#ffb300")

ggplot(data = game_df_pct_subset_longer,
       aes(x = reorder(StudyGame, Pct, sum), y = Pct, fill = PctType)) +
  geom_bar(position = "stack", stat = "identity", width = 0.75) +
  coord_flip() +
  scale_fill_manual(values = custom_colors) +  # Use the custom color palette
  theme(text = element_text(size = 5),
        legend.position = "bottom", legend.box = "horizontal",
        legend.text = element_text(size = 8)) +
  xlab("StudyGame") +
  labs(title = "Pct of Video Game Mechanics") +
  guides(fill = guide_legend(reverse = TRUE))
```




```r
# let's make correlation visualization for the 5 game mechanics
game_df_pct <- game_df %>%
  dplyr::select(starts_with("Pct"))
ggcorrplot(cor(game_df_pct)) +
  labs(title = "Correlation Plot of Game Mechanics")
```



To interpret this plot, first we need to understand what a correlation coefficient is. If you take a look at the legend, the values of the grid can be associated with a value between -1 and 1. If the square is red, meaning closer to 1, that means the two `Pct` variables at the cross section are *positively* correlated. Vice verse, if the square is blue, meaning closer to -1, then the two `Pct` variables at the cross section are *negatively* correlated.

As an example, let's look at the square at the bottom row, fourth column; it is a salmon color. This means that the variable `PctSP` and `PctOT` are slightly positively correlated. In other words, games that exhibit the `PctSP` mechanic are likely to also exhibit the `PctOT` mechanic. For blue squares, the two variables at the cross section are less likely to exhibit in a single game.

```
# run a principal components analysis (PCA) to see which games are similar
pca_results <- prcomp(as.matrix(game_df_pct), scale = FALSE)
pander(summary(pca_results))
```

Table 1: Principal Components Analysis

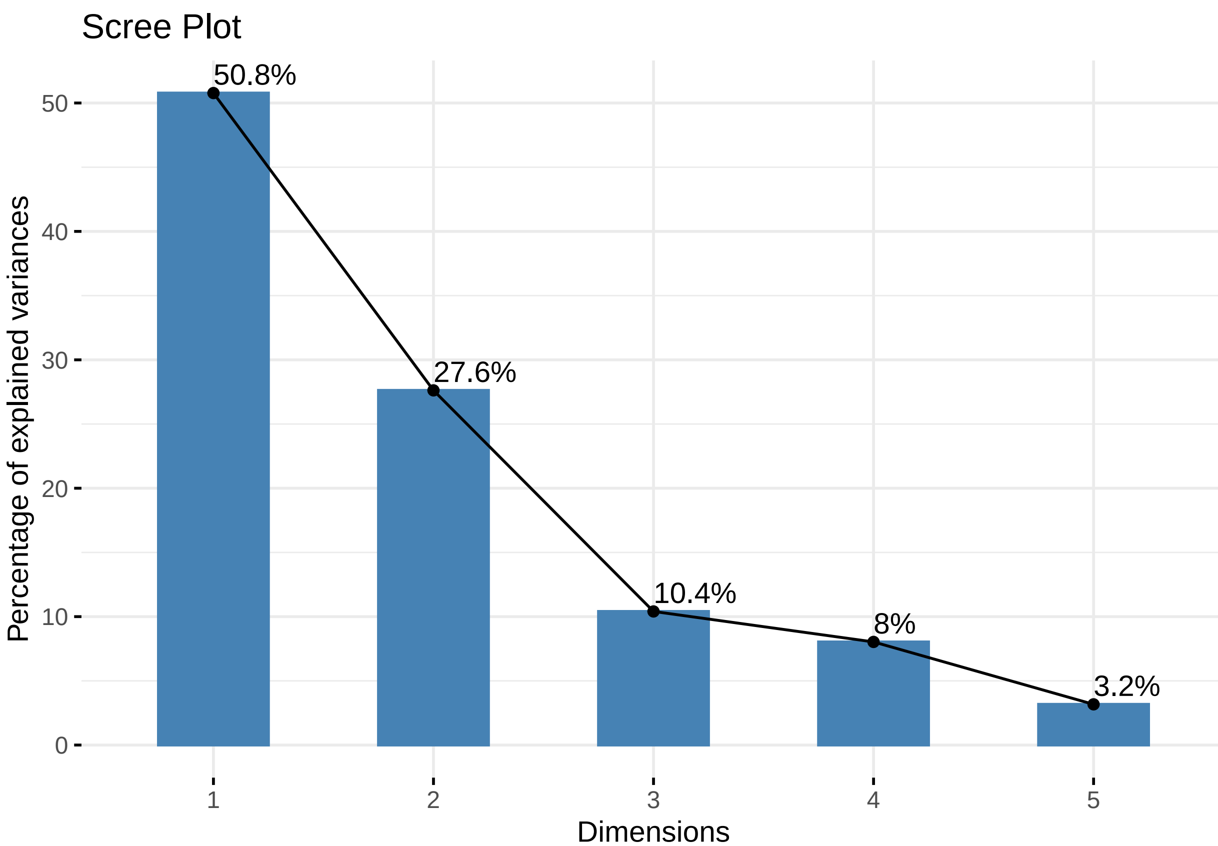
	PC1	PC2	PC3	PC4	PC5
PctOT	0.3215	0.6316	-0.5615	0.3705	0.2125
PctIV	-0.2029	-0.4321	0.04801	0.6428	0.5972
PctSN	0.7014	-0.5471	-0.2174	0.1958	-0.3508
PctSP	0.5761	0.2365	0.7079	-0.02107	0.3326
PctCG	-0.1777	0.2432	0.3661	0.6409	-0.6037

	PC1	PC2	PC3	PC4	PC5
Standard deviation	0.3001	0.2213	0.1359	0.1194	0.075
Proportion of Variance	0.5077	0.2762	0.1041	0.08031	0.03171

	PC1	PC2	PC3	PC4	PC5
Cumulative Proportion	0.5077	0.7839	0.888	0.9683	1

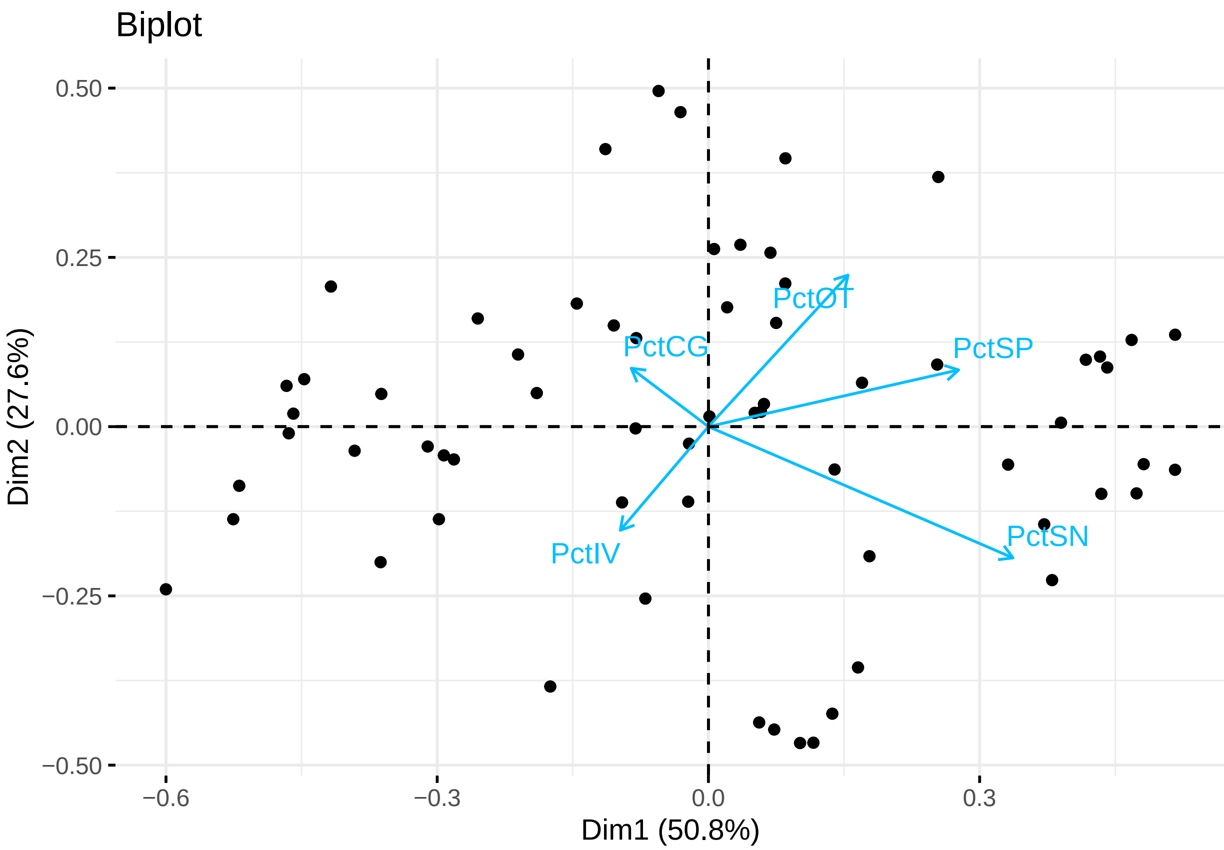
The Proportion of Variance row shows that what percentage of total variance each principal component accounts for. PC1 accounts for 50.77% of the total variance. In simpler timers, you could say that 50.77% of the data can be explained by PC1. PC2 accounts for 27.62% of the total variance, and so on and so forth.

```
fviz_eig(pca_results,
         title = "Scree Plot", addlabels = TRUE, bar_width=0.5)
```



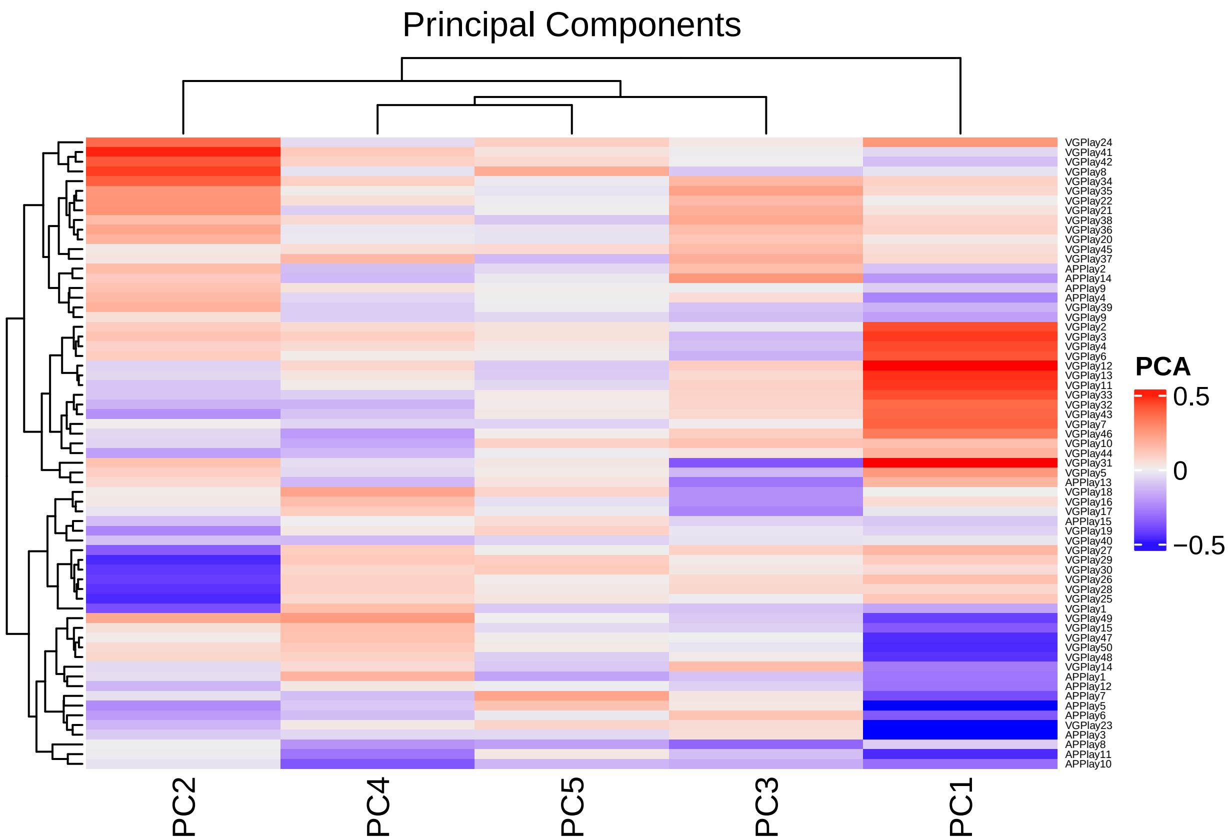
Alternatively, you can visualize this with a scree plot.

```
# we'd like to visualize the PCA results
fviz_pca_biplot(pca_results,
                repel = TRUE,
                col.var = "deepskyblue",
                title = "Biplot", geom="point")
```




```
# Install and load required packages
pca_scores <- pca_results$x
rownames(pca_scores) <- game_df$StudyGame

Heatmap(pca_scores,
        name = "PCA",
        column_title = "Principal Components",
        show_column_names = TRUE,
        cluster_columns = TRUE,
        show_row_dend = TRUE,
        show_row_names = TRUE,
        row_names_gp = gpar(fontsize = 4),)
```



Each PC on the x-axis represents a linearly independent measurement of similarity across games. These linearly independent measurements of similarity are in decreasing order of explained variance. So the type of similarity that PC1 explains, explains the most variance in the data. Whereas, the type of similarity that PC5 explains, explains the 5th most variance in the data.

With that in mind, notice that there are groups of rows that are shaded similarly for each PC. Those groups are similar according to that PC metric. For example, there are a group of games in the second quarter (0.25-0.50 range) of rows of PC1 that are red. They all have high measurements of the type of similarity described by PC1. The bottom quarter of PC1 (0.75-1.0 rows), are also grouped and have the least amount of the kind of similarity described by PC1. The same can also be seen for the 1st and 3rd quarters of rows of PC2, and the middle rows of PC3. The games in these groupings should be analyzed to uncover the latent similarity that the PC is describing.