Jared De La Serna

Dr. Chii Dean Lin

Stat 520

10 May 2024

2022 FIFA World Cup Principal Component Analysis Report

Data Introduction and Summary

For our data topic, we decided to focus on analyzing the 2022 World Cup. This data set includes the player statistics Gls (Goals), Ast (Assists), Shots, Successful Dribbles, Crosses, Age, Tackles Made, Blocks, Interceptions, and Clearance. Additionally, the data set looked at a total of 680 player entries (across 32 total countries that qualified for the World Cup). Focusing on the background, the World Cup is held every four years and decides who the best soccer nation is in the world. The winner of the 2022 World Cup was Argentina, the runner-up was France, and the third place was Croatia. Moreover, our data set came from Kaggle and a link to the data set will be provided in the references section.

Methods

When looking into our data set, we noticed a PCA analysis might be most appropriate. PCA allows us to condense our data set down while also maintaining that most of the variance is accounted for. Individual principal components create overall associations between the variables, usually looking at the first few PCs (these account for the most variance). Plots correlated with the PCA method highlight visual assumptions we can make about items scattered throughout the visualizations. Furthermore, visuals help make correlations between PC examples like PC1 vs PC2, PC2 vs PC3, and so on.

PC Categorization

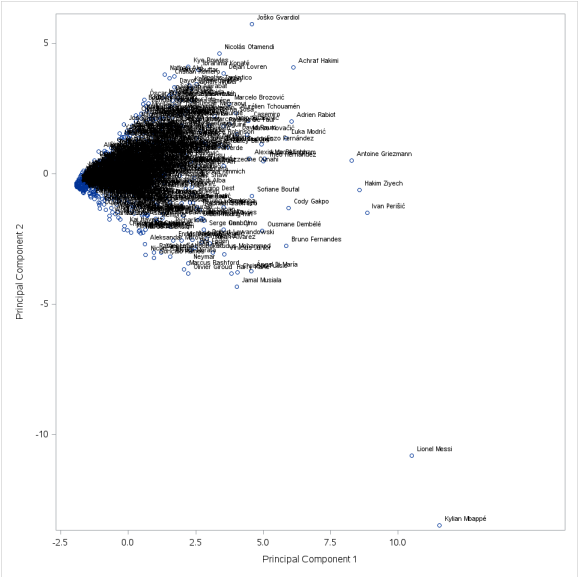
	Prin1	Prin2	Prin3
Gls	0.270142	-.402143	0.113846
Ast	0.321391	-.189834	0.144400
Shots	0.366982	-.378729	0.056503
Successful_dribbles	0.347927	-.296664	-.138654
Crosses	0.347910	-.142465	-.017343
Age	-.003990	0.031939	0.939967
Tackles_made	0.372440	0.321602	-.116618
Blocks	0.389655	0.277939	-.103234
Interceptions	0.326922	0.407826	0.005417
Clearence	0.222538	0.449981	0.188798

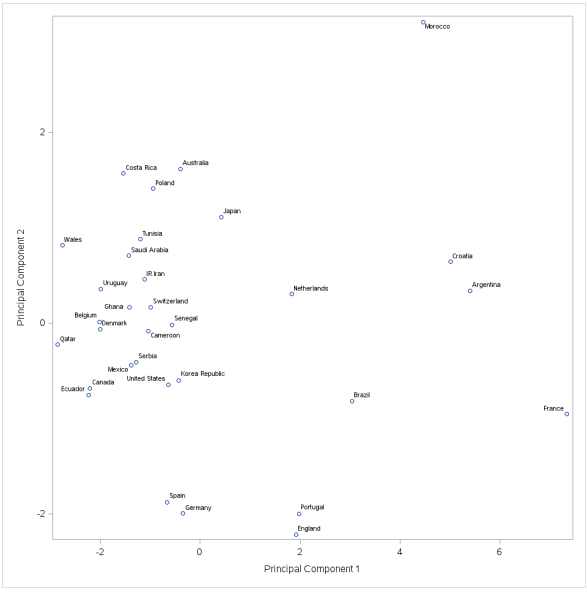
In the above output, three PC components are retained because the variance explained starts to tailor off after this point. PC1 appears to be representing the average of the statistics of every player. We can see similar values between almost all variables excluding Age. Additionally, PC2 appears to be representing a comparison between defense vs offense metrics. Similar values appear for each side of the ball (offense is negative, defense is positive), but again here we are excluding Age. Therefore, Prin3 appears to be based on age, as this variable's principal component three value of .939967 stands out compared to the others. We can interpret this PC as dealing with a player's experience. Note that the above output is for player names only, but the PCA analysis structured by the team yielded similar results.

PC1 vs PC2 Result

In the below plots, a comparison between PC1 vs PC2 was made between individual players, along with individual countries. First, looking at the players, it is no surprise to see Messi and Mbappe stand out. These two are undeniably in the conversation of being GOATs(Greatest Of All Time), and our below output displays it. Both have high PC1 and low PC2 values, meaning they have overall high statistics, with most of these statistics being geared towards offense. Shifting to the teams, the top three countries shine. It is interesting to see the differences between these top countries. For example, France has a lower PC2 value when compared to Argentina and

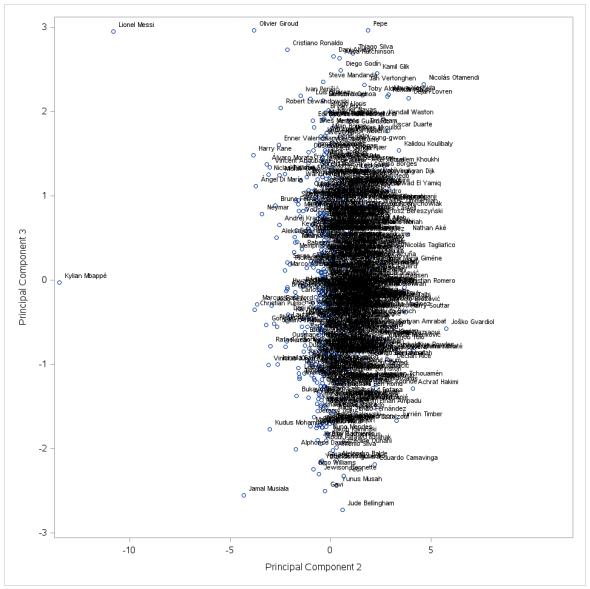
Croatia. Therefore, we can say that France displays a more offensive approach while the other two have a more balanced approach being near 0 for PC2.

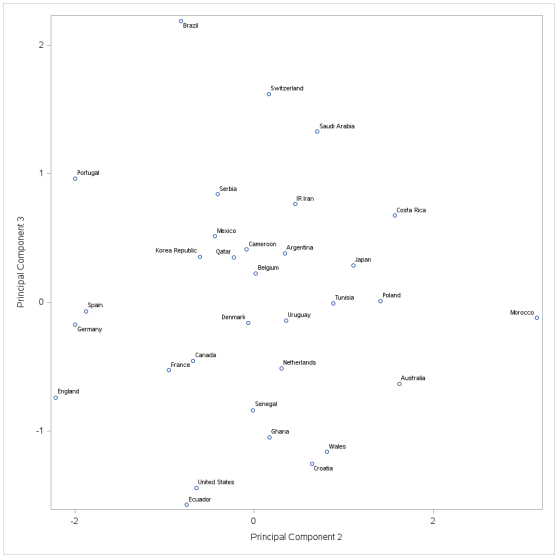




PC2 vs PC3 Result

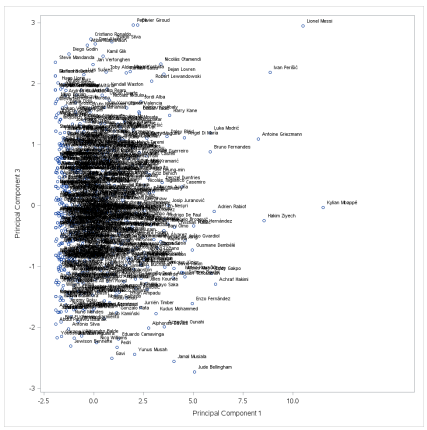
Next, a comparison between PC2 vs PC3 was made below. Here again, Messi and Mbappe stand out. However, this time we can clearly see that Messi is a bit older than Mbappe. Interestingly enough, young teenagers started to appear with the introduction of PC3. For example, near the bottom of the player output names like Jamal Musiala and Jude Bellingham are shown. Musiala and Bellingham were both 19 years old at the time of the World Cup. Additionally, Musialia's negative PC2 value means he is an offensive player while Bellingham is more of a balanced player being near 0 for his PC2 value. For the teams, it is easy to spot the oldest vs youngest country. Based on the below plot, we can say that Ecuador was the youngest team(lowest PC3 value) and Brazil was the oldest team (highest PC3 value). Additionally, a team like Morocco had an average age (close to 0 for PC3). Although Morocco is average in terms of age, we can conclude they are the best defensive team (highest PC2 value).

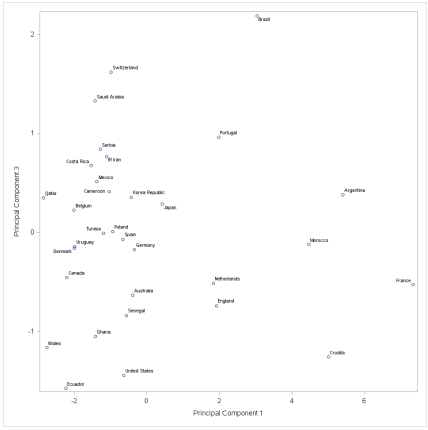




PC1 vs PC3 Results

Finally, the last comparison involved looking at PC1 vs PC3. Starting off with the players, the below plot can give us insight about players who are young but inexperienced/immature. An example of this includes Gavi, who was 17 years old at the time of the 2022 World Cup. However, Gavi's PC1 value is not that high, meaning he does not statistically perform the same as the likes of Mbappe and Messi. In terms of the teams, we can dive into the age differences between the top teams. Although each of the top teams have high PC1 values, their PC3 values differ from one another. For example, Croatia has a low PC3 value while Argentina and France are closer to 0. Therefore, Croatia is younger on average compared to the others.





References

Tittobobby. "FIFA World Cup 2022 Player Stats." Kaggle, 2022,

www.kaggle.com/datasets/tittobobby/fifa-world-cup-2022-player-stats.