Overview

This project is a detailed analysis of player's performance in international and club soccer games. Generally, players are most often criticized by general people for not fulfilling their expectations in terms of their performance in international games. In this project we are going to analyze the same by comparing their performance between club and international games.

Hypothesis Summary

Since this project is an evaluation of player's performance between club and international games, the hypotheses statements will be as below:

Null Hypothesis: There is no significant difference in the player's performance between the club and international games.

Alternative Hypothesis: There is a significant difference in the player's performance between the club and international games.

Since z-score is our evaluation tool, to be more precise on hypotheses, z-score more than 1 will be considered as a significant difference in performance and z-score less than or equal to 1 will be considered as not significant.

Dataset Summary

The main dataset of this project consists of two csv files. One contains club data which includes the player's performance in club games. The second csv file contains data of the same year which includes the player's performance in international games. These data have been collected from Wikipedia's official website i.e., www.wikipedia.org. The csv file containing data of international games has twelve players in rows with 8 columns showing their country and

appearances as well as goals in three years (i.e., 2006, 2010 and 2014). Similarly, the second csv file containing club data has twelve players in rows and 8 columns showing their respective clubs, their appearances as well as goals during the same three years (i.e., 2006, 2010 and 2014). Data analysis has been done based on player's appearances on those games and how many goals they scored; hence, those are important columns of the dataset.

Analysis Methods

Data analysis is the crucial part of hypothesis testing. In general, a player's performance is evaluated primarily by how many goals he/she scored. Hence, we have chosen the same approach to measure their performance. Their performance comparison between both games by looking only at their goals will not be an ideal approach since their appearances between those games might be different. Hence, during analysis, we have calculated the goal per appearance to have more accuracy in our conclusion. Goal per appearance is calculated by dividing total goals scores of an individual player by the total number of games they appeared in. In other words, it is an average goal scored per their appearance i.e., lambda (λ). Lambda has been calculated by using below formula.

$$\lambda = \frac{Total\ Goals}{Total\ no.of\ Apperances}$$

Under Poisson distribution of statistics, Lambda is the mean (average) value as well as the variance. Hence, goal per appearances i.e., Lambda (λ) of each individual player for both club and international games has been calculated. In order to compare the calculated lambda values between two games, z-scores for individual players has been calculated using below formula.

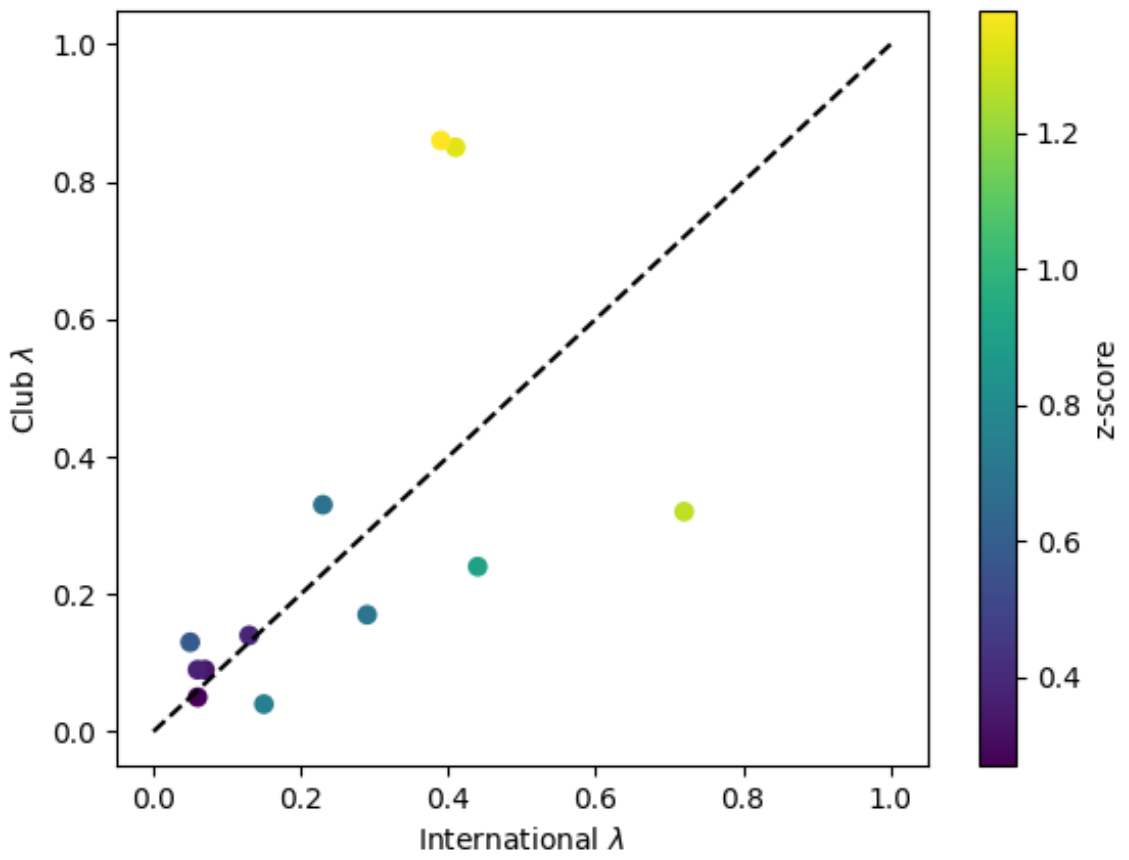
$$z - score = \frac{Max\ (Lambda_{club},\ Lambda_{Intl})}{Min\ (Lambda_{club},\ Lambda_{Intl})}$$

Inference

We have used the same criteria that we had specified during hypothesis formulation to evaluate the z-score to make a conclusion. When we run the code in python, below output has been generated:

Player	Lambda (Club)	Lambda (Intl)	Z-Score	Remarks
Cristiano Ronaldo	0.85	0.41	1.33	>1
Lionel Messi	0.86	0.39	1.38	>1
Sergio Ramos	0.13	0.05	0.58	<1
Rafael Márquez	0.04	0.15	0.75	<1
Xavi	0.09	0.07	0.34	<1
Andrés Iniesta	0.14	0.13	0.39	<1
Andrés Guardado	0.09	0.06	0.37	<1
Wesley Sneijder	0.33	0.23	0.69	<1
Tim Cahill	0.24	0.44	0.90	<1
Miroslav Klose	0.32	0.72	1.27	>1
Philipp Lahm	0.05	0.06	0.27	<1
Bastian Schweinsteiger	0.17	0.29	0.70	<1

Looking at the above z-scores, we can conclude that only 3 players out of 12 (Cristiano Ronaldo, Lionel Messi & Miroslav Klose) have significantly different performances (based on respective z-scores: 1.33, 1.38 & 1.27 which is >1). The scatter plot below is also an output from python after plotting the data where x-axis shows Lambda_Intl and y-axis shows Lambda_Club. The color of dots represents the z-scores. There are only 3 outliers in the scatter plot as well. Since 9 out of 12 player's z-scores are below 1, the null hypothesis has been accepted and a conclusion has been made that there is no significant difference between the player's performance in club and international games.



Conclusion

From the above analysis, we have already concluded that there is no significant difference between a player's performance in international and club games. This analysis is useful for many stakeholders. The countries participating in world cup games can evaluate their player's club performances and manage their team accordingly. Also, soccer clubs can evaluate a player's performance in this world cup or previous world cups and make decisions to hire new players or manage a team to improve their club performances. People involved in sports betting can make better predictions based on this analysis. Sports commentators can analyze the players performance in both games to notice players and mark them while commenting during games. In future, if data

collection from a reliable source is possible, researchers can do more in depth research with collecting data of a player's assists, fouls, shots on target and more. Researchers can also do research on goalkeeper's and defender's performance based on how many saves and fouls they made.