A Comparative Analysis of Match Outcomes for India and Pakistan in Asia Cup History

1. Problem Definition

Historically, the rivalry between the Indian and the Pakistani cricket teams have always been prevalent. This competition is also highlighted throughout the Asia Cup. Asia cup is a tournament that includes several teams that represent different countries in the continent of Asia. One of the most important actions that a captain of a cricket team makes is the coin toss. The coin toss can help set up the framework for the match. If won, the captain is able to make the decision of Batting or Bowling first. This decision is strategically important based on the strengths and weaknesses of the two teams. Usually the tension between the two teams starts here.

This project aims to determine: *To what extent does winning the toss influence the final match outcome for the two historically rival teams: India and Pakistan?* Furthermore, I would like to explore if the decision made (voluntarily or involuntarily) has a visible impact on the final outcome of the match.

This relationship between the decision made after coin toss and the result of the match overall provides the fans and sports analysts with data to back up claims. These claims can range from whether the coin toss, associated with luck, is a determining factor in the outcome of the specific match, and how significant the coin toss is to the outcome.

Because India and Pakistan are historically rivals when it comes to Cricket, it is interesting to understand if something as trivial as a coin flipping in the air can set up the team for success or failure. The analysis of historical data can also set up the basis for claims that fans make about the performance of a team.

2. Data Description

This dataset from Kaggle has data about each individual match that was played in the Asia Cup between the years of 1984 and 2022.

Each row of the dataset represents an individual match. Each team gets their own row which means that there are duplicates in the matches. But the difference between the two rows is the perspective through which the data is taken. For example, if there is a game between India and Pakistan, there are two rows that represent the same game. One of the rows will have all the stats about the Indian team with the opponent being Pakistan, while the other row will have all the statistics about the Pakistani team with the opponent listed as India.

The key columns and features that this research will be focused on are “Team”, “Toss”, “Selection”, and “Result”. These columns are necessary for the research question because it deals with the relationship between the coin toss affecting the selection and the outcome of the match for the particular team.

The dataset is 250 columns and 20 rows which can be interpreted as 250 matches with 20 data points related to the match, four of which are mentioned earlier. This is a relatively large dataset with a lot of information. Because of this, I decided to only conduct the analysis on two teams: India and Pakistan. When filtered to just India, there were 57 rows (matches). When filtered to Pakistan, there were 54 rows (matches). Since the two numbers are relatively close in proximity, there will be no biases regarding the outcome of the analysis.

Although the dataset has a lot of detailed information about each of the matches that were played in the Asia Cup, some details that could make the analysis even more accurate could be the time of day the match was played, team rankings, and even dew factor since it can determine the performance of the cricket ball itself which can affect the performance of teams. Another factor that is missing is the choice of toss that each of the team chose. The data jumps directly to whether the team won the toss or not. Although it is not important to the performance, there could be interesting trends.

The dataset makes the fair weather assumption which means that all of the teams played their matches in the same weather conditions. Each match is different. The weather could affect the outcome of the selection that comes after the toss. Also by grouping the matches together, the dataset is assuming that the rules and regulations stayed consistent across the years and the different matches.

3. Data Cleaning & Preparation

I made notable cleaning decisions that ensured that the data was easy to read and the dataset was prepared for the analysis.

I made sure to standardize the columns “Result” and “Toss” to make them both lowercase because of the conflict of having duplicate values because of the lack of consistent capitalization. I then converted each string value from these two columns (“win” and “lose”) into 1 and 0 respectively. Converting them into integers will set me up for the analysis where I will be able to calculate the mean of the data.

The value of the mean found by the function .mean() determines winning/losing percentage outcomes, or the winning rate (if decimal is greater than 0.5, we can assume majority wins in the particular scenario).

I also made sure to filter out the “no result” and the null rows because it does not contribute appropriately to the question. No result cannot be deemed the same as winning/losing because the outcome was not determined. These rows did not add to the value to the question because the research is only concerned with the relationship between selection after toss and match results.

Another decision that I made that was crucial to the research question was to make two different variables that held the data of two different teams: India and Pakistan. This made it easier to analyze data because the variables already had the data filtered out. Analyzing a smaller dataset is easier because there is less room for error and confusion.

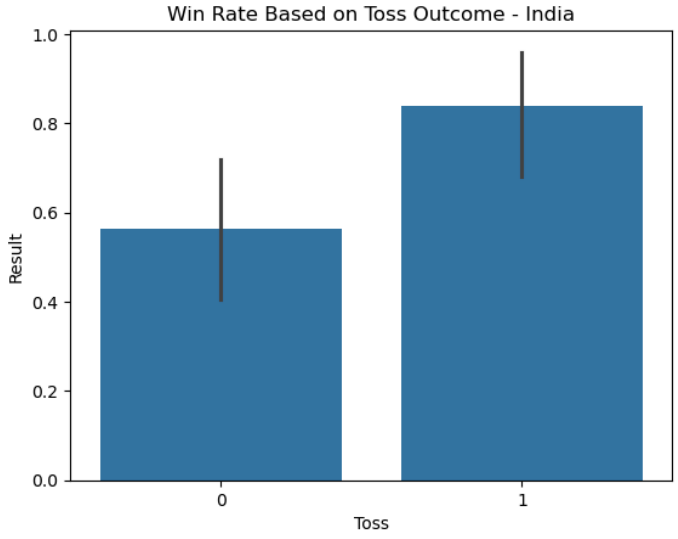
During this part, I grouped together the results of win and win d/l, and respectively with loss, because I made the assumption that the weightage of both of the results would be the same.

Some tradeoffs of this assumption was that the outcome of the analysis assumes that the winning/losing happened under similar circumstances, rather than them being different.

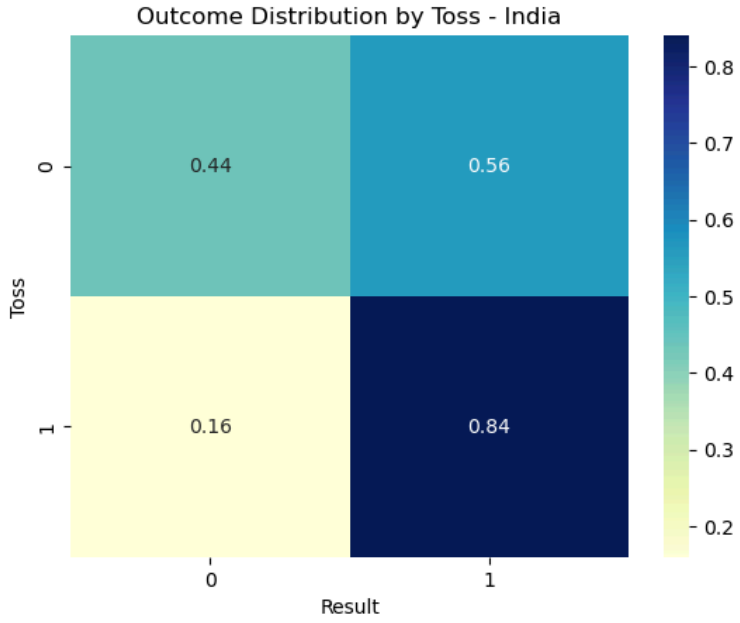
4. Data Understanding & Visualization

I made sure to analyze the two countries separately so that I can compare the results at the end.

India

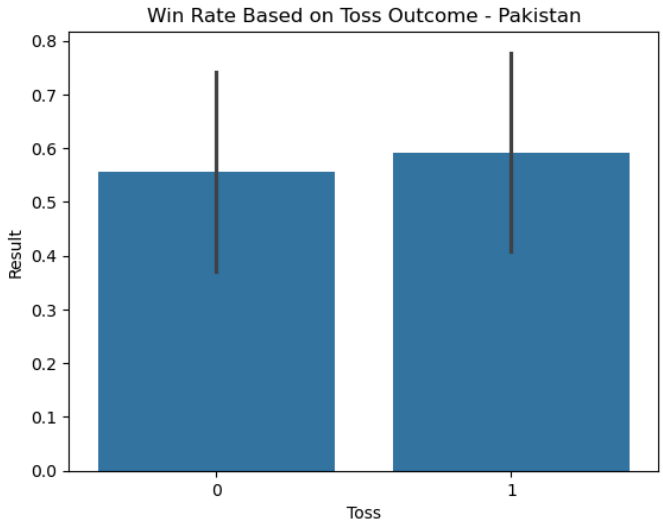


Toss	count	mean
0	32	0.5625
1	25	0.8400

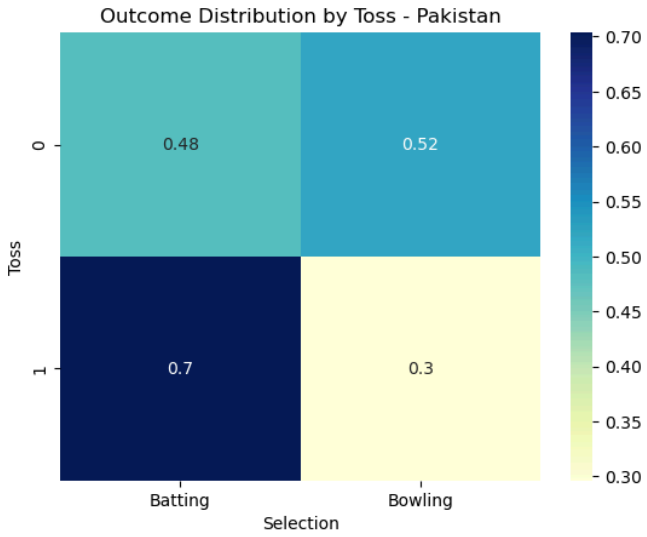


Result	0	1
Toss		
0	0.4375	0.5625
1	0.1600	0.8400

The first visualization that I made was this barplot which shows the count and percentage of wins for India based on whether they won or lost the toss. If the mean, which is the average of the "Result" category, is greater than 0.5, then the rate of winning is high for the particular team. In this case, team India had a higher percentage of wins when the coin toss was in their favor.



Toss	count	mean
0	27	0.555556
1	27	0.592593



Toss	count	mean
0	27	0.555556
1	27	0.592593

On the other hand, the analysis for Pakistan's team shows that the outcome of the toss does not affect the team's overall result, but winning is slightly more favorable (by 0.4).

Toss	Selection	count	mean
0	Batting	12	0.416667
	Bowling	20	0.650000
1	Batting	7	0.714286
	Bowling	18	0.888889

India

		count	mean
Toss	Selection		
0	Batting	13	0.384615
	Bowling	14	0.714286
1	Batting	19	0.578947
	Bowling	8	0.625000

Pakistan

The above tables show the relationship between winning the coin toss, selection and the winning percentage. For team India, whenever the coin toss is in their favor, there is a clear increase in the winning chances. On the other hand, Pakistan does not have a clear indicator similar to India. Bowling has higher winning percentages in both categories than batting does selection wise.

These visualizations help me understand the data because of the neat format they are displayed in. In the heatmaps, there are color indicators that help see which combinations were most and least favorable for different teams. The bar graph's length of the bar helps to see what outcome was more and less favorable to the respective team.

5. Storytelling & Interpretation

The data tells an intriguing tale of the interplay between strategy and luck in international cricket. My research indicates that though winning the toss significantly raises India's chances of winning, it is a much less accurate predictor for Pakistan. India's winning percentage rises dramatically when they win the toss, indicating that they make good use of the tactical advantage of batting or bowling first. The trends show that India is more dependable at turning this early advantage into a win, even though both teams prefer bowling first, which is linked to higher winning rates. With only a 0.4 difference in results regardless of the toss, Pakistan's data, however, suggests that individual skill or on-field performance likely have a greater influence on their match results than the initial tactical decision. It's important to avoid assuming that winning the toss equates to victory. Even though the data indicates a correlation, it disregards the opponent's skill level and the specifics of the match. Gaining an advantage by winning the toss does not guarantee success.

6. Limitations, Ethics, & Reflection

After reflecting on this exploration, there could be several factors outside the scope of the data that could have likely influenced the results.

The dataset lacks critical environmental metadata, such as dew factor and pitch type. In Asian conditions, dew often makes bowling second much harder, which might be the real reason behind the high win rates for bowling first, rather than the toss itself.

The data does not include team rankings or player availability. We do not know if India won several tosses against weaker opponents. This could make the toss advantage artificially inflated in the statistics and may not paint the full picture. This data also only includes the

matches from the Asia cup which excludes a majority of the countries that are popular for cricket. This could be harmful because the analysis is not applicable to countries that are not in the dataset since we do not have that information.

I was surprised to see how much more sensitive India's win rate was to the toss compared to Pakistan's. This could showcase the strategies that the different teams use in preparation for the match. Team India might use the winning of a Toss to their advantage.

If I had more time, I would like to merge this with weather data to see if the advantage of the coin toss being on your side fluctuates based on what time of day the match is held. Ethical considerations also arise when the result of this data analysis is used for sports betting or predicting the outcome of a future match. This is because this undermines the complexity of human performance and oversimplifies the entire game down to a coin flip.

7. **Code**

   https://github.com/rchivate/DTSC_2301_2302/tree/main/Cricket%20Asia%20Cup

8. **References & AI Use Transparency**

   https://www.kaggle.com/datasets/hasibalmuzdadid/asia-cup-cricket-1984-to-2022

Generative AI was used to assist in formatting and generating the html from a google document that Riya hand typed (bold, italics, etc.).