

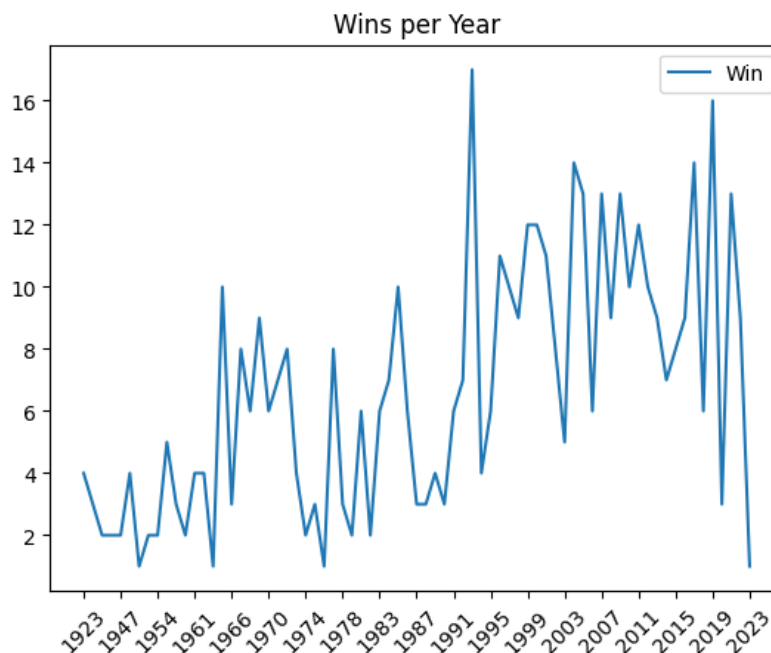
Introduction

Growing up I always had a passion for soccer and would play it often with family at events such as birthday parties or just at a simple get together. I was always small so I usually got hurt in those games but that did not stop me from playing what I love. I also had always loved playing FIFA on Xbox and I played from Fifa12 to Fifa17. Playing the game, I would usually end up playing as Mexico as that is in my blood and we treasure the sport of soccer and hold it as a traditional sport in our hearts. With this project I wanted to explore some data with Mexico and their history with the sport of soccer. I was able to find data on this topic via [Kaggle](#).

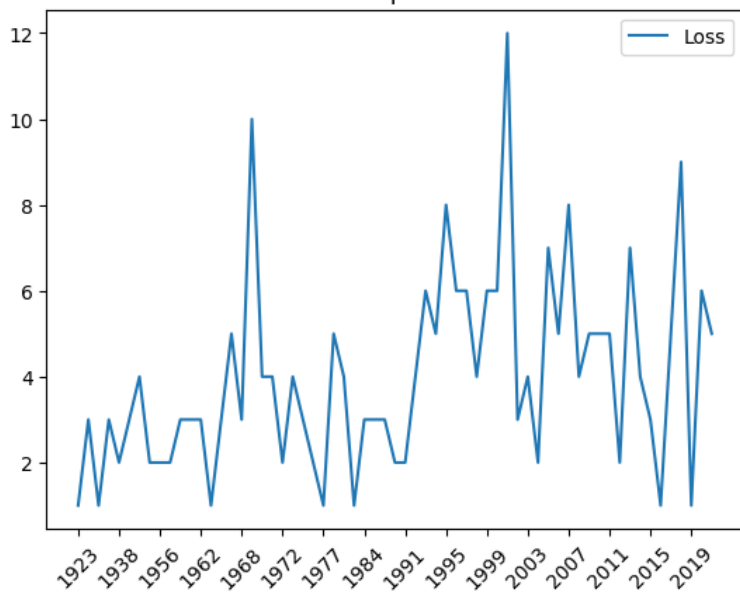
Methods

I decided to split my coding in four parts to better organize the process. The parts consisted of data formatting, data results, further digging, and final data results. The first part was simply just 'cleaning' the data so that I was only focusing on observations that involved Mexico. I then had to make a new column to indicate whether the game resulted in a win, loss, or a tie for Mexico. I also had to consider the penalty data that I had and factor in whether that particular game was a win or a loss. The second part I tried to do some visualization by plotting points of the number of wins, losses, and ties, per year. I also included a graph that had all three graphs together. I then calculated the overall percentage of wins, losses, and ties for Mexico. For part three, I was wondering if Simpson's paradox had a role in part two's results so I decided to group results based on the tournament played as I figured maybe there could be a difference between a friendly match and a world cup match. Part four showed the results of part three.

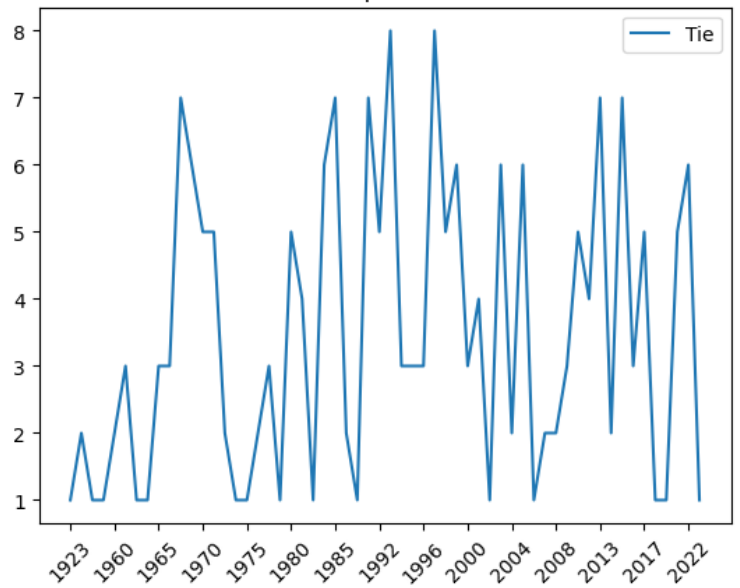
Results:



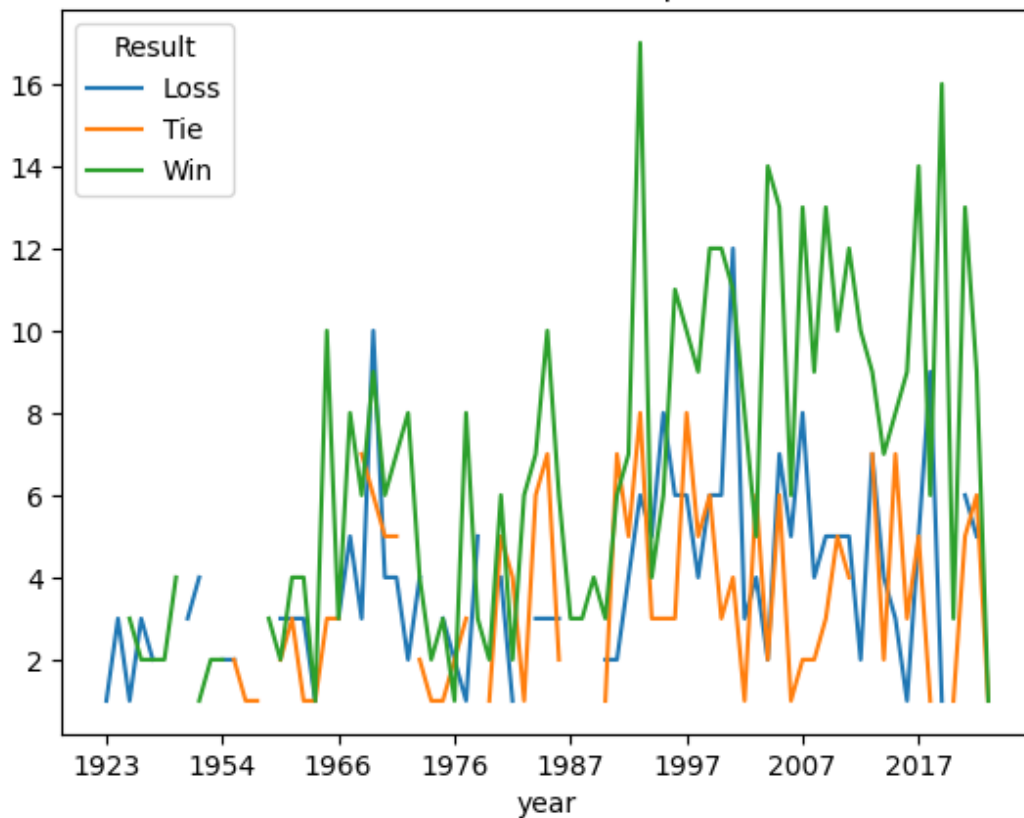
Losses per Year



Ties per Year



Wins, Losses and Ties per Year



DATA SCIENCE DISCOVERY - 107

		Overall		
		Win %	Loss %	Tie %
		51.65	26.8	21.6
Tournament Grouping				
	Win %	Loss %	Tie %	Greater_Percentage
CONCACAF Championship	58.974	15.385	25.641	Win
CONCACAF Championship qualification	56.25	18.75	25.0	Win
CONCACAF Nations League	70.0	10.0	20.0	Win
Confederations Cup	44.444	44.444	11.111	Win/Loss
Copa América	43.75	35.417	20.833	Win
FIFA World Cup	28.333	50.0	21.667	Loss
FIFA World Cup qualification	65.625	14.375	20.0	Win
Friendly	47.149	29.167	23.684	Win
Gold Cup	74.684	13.924	11.392	Win
Kirin Cup	0.0	50.0	50.0	Loss/Tie
Korea Cup	50.0	25.0	25.0	Win
Lunar New Year Cup	66.667	33.333	0.0	Win
NAFC Championship	100.0	0.0	0.0	Win
NAFU Championship	33.333	33.333	33.333	Equal Percentage
Pan American Championship	7.692	69.231	23.077	Loss
USA Cup	46.154	15.385	38.462	Win

Conclusion

This project was very fun for me as this was my first solo project with no guidance other than google and a little bit of chatgpt with the plot graphs as I had trouble understanding how to use the functions. Looking at the graphs, there seems to be an interesting story to tell but looking at the data as a whole there doesn't seem to be a powerful story. With part four's results, I expected there to be more losses per tournament than wins but what I saw was that there were more wins per tournament than losses or ties. I guess this kind of follows the overall results as win percentage is higher than losses and ties. However, I still think that Simpson's Paradox still has a role in this. Perhaps grouping by tournaments wasn't the right move but perhaps grouping by intervals of recent and old matches. Maybe even consider the average career of a player and split up the intervals like that. There is a lot more I would like to do with this data and that I will do but I wasn't able to do due to the time for the submission of this project. I strongly believe that with further analysis there could be a great story to tell with this data.