In professional tennis, there are four tiers of events, Grand Slams, Masters 1000, ATP 500 and ATP 250. Players receive the most points in their Association of Tennis Professionals (ATP) rankings for winning a Grand Slam, and the least from winning an ATP 250 tournament. There are four Grand Slams throughout the year, the Australian Open (Hard surface), Roland Garros (Clay surface), Wimbledon (Grass surface), and the U.S. Open (Hard surface). This leads to more competitive players wanting to play in the Grand Slams. In the ATP, Grand Slam tournaments are played in a best of five format, and non-Grand Slam tournaments in a best of three.

In Tennis, there are three different types of surfaces that are played on. The options are Grass, Hard, and Clay. The surfaces are important to keep track of as the speed of tennis changes, e.g., clay generally slows the ball down whereas grass speeds it up. Individual players perform better on certain surfaces and may favor playing on their preferred surface.

This dataset contains information for each player on each surface. To be included in the data set, players must have played a minimum of 10 matches overall or 5 matches on a particular surface. This data was filtered so only players who have recorded data on all three surfaces are present. The original data is available in the file atp\_2023.csv and summarized in the files grand\_slam\_summarized.csv and court\_surface\_summarized.csv

In this worksheet, we will investigate win percentages by looking at distributions, shapes, and differences between Grand Slams and Non-Grand Slams. Additionally, we will compare between the three surfaces. There will be questions about each of these, some of them being more open ended than others.

1. Below is a bee swarm plot showing the distribution of win percentage in Non-Grand Slam Tournaments and Grand Slam Tournaments. Comment on the difference between the two.



The distribution on the Non-Grand Slam side is much less spread out and there appear to be more observations. The range is smaller as well. On the Grand Slam side, the range goes all the way from 1 to 0, and there are many observations at 0 dragging down the median and mean.

1. Typically, a longer series favors the better player as they have more opportunity to win. Why does this plot support that theory and if so why?

This plot supports that theory because in the best of 3 tournaments, the variability is generally smaller (i.e., less predictable). For example, there is not a single player with a zero-win percentage. In the best of 5 tournaments, there appear to be many players with a zero-win percentage. This suggests that the increase in the series length means better players are very likely to win, and worse players are very likely to lose.

1. What is the shape, center, and spread of the distribution of win percentages displayed below? Is there a big difference between the win three surfaces?



All three distributions look similar and normally distributed. The range goes from 0 to 1 on all three. The center differs, it is around .5 for clay and hard, but closer to .6 for grass.

1. As you can see above, the center of win percentage distribution for grass is higher than the others. Throughout the year, there are far fewer grass tournaments played, and Wimbledon a Grand Slam is one of these tournaments. Discuss why this might contribute to grass having a higher win percentage distribution.

As a Grand Slam, Wimbledon attracts better players that will have higher win percentages. In Question 1, we found that Grand Slam matches are more predictable, so better players usually win. Additionally, Wimbledon has a higher impact on the distribution in Grass than the other Grand Slams do on their respective surfaces due to the smaller number of Grass tournaments leading to an uptick in Win Percentage.