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. This data is available in the file atp\_2023.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.

A graph of a diagram

Description automatically generated with medium confidence

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?
2. 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?

A diagram of a graph

Description automatically generated with medium confidenceA graph of different types of plants

Description automatically generated with medium confidence

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.