In professional tennis the rankings are typically based on set parameters. Throughout the year there are tournaments that are worth different numbers of points. There are four tiers of events, Grand Slams, Masters 1000, ATP 500 and ATP 250. You receive the most points in your Association of Tennis Professionals (ATP) rankings for winning a Grand Slam, and the least from winning an ATP 250 tournament. In the ATP, Grand Slam tournaments are played in a best of 5 format, and non Grand Slam tournaments in a best of 3. In the WTA, all tournaments are played in a best of 3 format.

In Tennis, there are also 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. Certain players perform better on certain surfaces. This dataset contains information for each player on each surface.

In this worksheet, we will look at distributions, shapes, differences between ATP and WTA, and differences between Grand Slams and Non-Grand Slams. There will be questions about each of these, some of them being more open ended than others.

(In order 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)

1. A graph of different types of plants

   Description automatically generated with medium confidenceWhat is the shape, center, and spread of each distribution? Is there a big difference between the 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. A diagram of a graph

   Description automatically generated with medium confidenceCalculate the IQR and range of the distribution for each surface.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Surface | Min | Q1 | Med | Q3 | Max |
| Clay | 0.000 | 0.367 | 0.500 | 0.594 | 0.893 |
| Grass | 0.167 | 0.429 | 0.583 | 0.683 | 1.000 |
| Hard | 0.000 | 0.373 | 0.462 | 0.573 | 0.927 |

Clay: IQR = 0.594 – 0.367, Range = 0.893 – 0.000

Grass: IQR = 0.683 – 0.429, Range = 1.000 – 0.167

Hard: IQR = 0.573 – 0.373, Range = 0.927 – 0.000

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

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, 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.