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 data that we will be using, the provider has calculated an Elo ranking. This ranking considers what the tournaments are, as well as who each player is playing. So, playing a "harder" opponent in a Grand Slam counts for more than playing an "easy" opponent. For instance, if number 9 Elo ranked Casper Ruud plays number 1 Elo ranked Jannik Sinner and wins, it will be worth more than if Ruud played number 111 Elo ranked Stan Wawrinka.

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 the data including this Elo ranking and look at distributions, shapes, and a multiple linear regression model. 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 sizes of bars

   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 chart with different colored squares

   Description automatically generatedCalculate the IQR and range of the distribution for each surface.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Surface | Min | Q1 | Med | Q3 | Max |
| Clay | 0.000 | 0.433 | 0.500 | 0.667 | 0.833 |
| Grass | 0.200 | 0.421 | 0.575 | 0.714 | 1.000 |
| Hard | 0.125 | 0.394 | 0.510 | 0.619 | 0.893 |

Clay: IQR = 0.667 – 0.433, Range = 0.833 – 0.000

Grass: IQR = 0.714 – 0.421, Range = 1.000 – 0.200

Hard: IQR = 0.619 – 0.394, Range = 0.893 – 0.000

1. A chart with different colored squares

   Description automatically generatedBelow is the same information as questions 1 and 2. Go through and do the same steps as 1 and 2.

A graph of a bar graph

Description automatically generated with medium confidence

1. Discuss if there are any differences between the ATP and the WTA. Are there any problems with this data and how it was collected.

Interpret the coefficients for return points won percentage and grass and clay surfaces.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** |
| Constant | -0.1898 | 0.0994 | -1.91 | 0.057 |
| ReturnPointsWonPercentage | 2.079 | 0.253 | 8.23 | 0.000 |
| EloRank | -0.001332 | 0.000120 | -11.07 | 0.000 |
| Surface |  |  |  |  |
| Grass | 0.1016 | 0.0209 | 4.85 | 0.000 |
| Hard | 0.0269 | 0.0166 | 1.62 | 0.106 |

Return Points: A 1 unit increase in the return points won percentage is associated with a 2.079 increase in win percentage holding all else constant.

Grass: Compared to playing on Clay courts, we would expect the win percentage for playing on grass courts to be 0.1016 higher, on average, holding all else constant.

1. Interpret the R-Squared value and decide if you think that this is a good model.

|  |  |  |
| --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** |
| 0.125993 | 48.20% | 47.57% |

About 47.57% of the variability of win percentage is explained by the multiple linear regression model.

This isn’t a great R-Squared value, so we should look at different models or transformations to our model that might be able to explain the data better.

1. Based off of the following plots, do you believe that this model meets the assumptions of a multiple linear regression model?

A graph with blue dots

Description automatically generatedA graph with a line

Description automatically generated

As we can see in the Q-Q plot on the left, this looks normally distributed as a vast majority of the points follow the line. On our plot of fitted values vs. residuals, the variance seems consistent with a slight narrowness towards the lower values.