ATP Tennis Results 2015: Lab 5

For this lab, I chose six attributes to focus on from the original dataset. They are: **w_1stIn**, **l_1stIn**, **w_2ndWon**, **l_2ndWon**, **w_bpFaced**, and **w_bpSaved**. These variables represent various statistics of a specific tennis match. The **w_** denotes that the variable is attributed to the winner of the match, while the **l_** denotes that the variable is attributed to the loser of the match. In respective order, the statistics are: the number of first serves in, the number of second service points won, number of break points faces, and number of break points saved.

Before explaining the insights gained from the brushing effects, it is beneficial to explain the statistics and what they mean. In a tennis match, for every point played, the player serving the ball has two chances to get the ball across the net. The number of first serves in (**1stIn**) denotes how many of the first chances were successful. The number of second service points won (**2ndWon**), denotes how many of the points were won by the player when the point was played on the player's second serve. A break point occurs when a player is serving, but his opponent is one point away from winning the game (a set in tennis is played until a player reaches 6-7 games, depending on the circumstances). The number of break points faced (**bpFaced**) denotes the number of break points the opponent achieves against the player. The number of break points saved (**bpSaved**) denotes how many break points the player manages to recover. This means that the opponent is no longer one point away from winning the game.

At the beginning of the video, before I brush on the parallel coordinates, the bar chart for the winner's number of break points faced is shown. As we can see, among the matches in the dataset, the most frequent number of break points faced by the winning player is zero. This means that the winner, more often than not, faces zero break points in a match. This makes sense because if the winner does not face break points, then the opponent is not on the verge of winning a game on the winner's serve. In order to win sets, the opponent must "break" the player's serve. If the player does not face break points, then the opponent has minimal chances of winning the match. This is one of the ways a player becomes the winner of a match.

Data story 1:

However, tennis matches are not always that straightforward, as the visualizations will suggest. When the high values for the **w_bpFaced** variable are brushed on the parallel coordinates, we see the paths of the data observations that fall into that category. When the winner faces a large amount of break points, the **w_bpSaved** (number of break points saved) is also high. This shows that the winning players, even when confronted with a large amount of break points, are still able to deny that point to their opponent and save the break points. If the **w_bpSaved** was low compared to the **w_bpFaced**, then the winner would likely not be the winner. This is why we see on the bar chart that the frequency of high amounts of **w_bpFaced** is not high.

Data story 2:

When **w_bpFaced** is high, we are made aware of another interesting insight. The **l_1stIn** variable is usually higher than the **w_1stIn** variable. This insight is not very trivial and requires a deeper analysis. When a player has a high number of break points faced, this usually means that the player is not serving very well. This can mean that the player is serving the ball at slow speeds, which makes it easier for the opponent to dominate the point. It can also mean that the player is failing to get the ball in play (failing two out of the two serve chances results in a loss of the point). Therefore, it would make sense for the **w_1stIn** to be lower when the **w_bpFaced** is high. In this case, perhaps the player is missing many of their first serve chances and must take

their second serve chance. This would result in either a loss of the point after gameplay or a loss of the point due to a miss of the second serve. When the **w_bpFaced** is high, this also usually means that the opponent is playing well, since they are continuously threatening the other player's serve. This makes it likely for the player to have a high number of first serves in. However, if the player has a high number of break points faced, how do they end up being the winner? It seems counterintuitive, but another insight will help explain how this is possible.

Data story 3:

The brushing on the **w_bpFaced** on the parallel coordinates also reveals that the **w_2ndWon** is higher than the **l_2ndWon** when the value is high. The **l_2ndWon** is significantly lower than the **w_2ndWon**. This statistic is imperative to explaining the eventual success of a winner despite the amount of break points faced. This is because second service points capture more fully the entirety of the match. That is, second service points are played throughout the match (every time the player misses their first serve). On the other hand, break points can be very scattered. It is possible for break points to occur all in one game of a match (one game out of a minimum of 12 games if it is a three-set match). Therefore, break points may not be the most surefire way of predicting the winner of a match. However, second service points are much more reliable. That means that if a player's number of second service points won are high, they are not losing many points on their own serve throughout the match. In this way, second service points describe a player's consistency throughout the match. The more points the player wins on their own serve, the better. Ultimately, the opponent loses because they are less consistent in their own serve. This cancels out the effects from imposing break points on the winning player.

Data story 4:

Taking into account the variables that have been discussed thus far, the MDS plot also reflects an interesting insight. Without the brushing, we see that there are two clusters of variables. One consists of four variables, and the other consists of two. When we brush on the **w_bpFaced** and **w_bpSaved**, we see that these two variables make up the cluster of two variables. Therefore, the discussion above about how break points may not be a great predictor of who wins a tennis match is correct.

Data story 5:

The last observation we can make from brushing comes from analyzing the scatterplot matrix. When high amounts of **l_2ndWon** are brushed, we see a positive correlation between **w_1stIn** and **l_1stIn**. However, the number for each of these variables are scattered. That is, the **w_1stIn** and **l_1stIn** variables are not high or low. Instead, they are considerably representative of the original data. This may mean that the number of second service points won are independent from the number of first serves in. This is an interesting observation because it would be easy to assume that if **l_2ndWon** is high, then the **l_stIn** should be high as well. This suggests that a player's second serve outcome cannot be predicted by their first serve outcome.