Assignment #1: Data Exploration and Static Visualization Design

For this exploratory analysis I played with data and transformed it to propose hypothesizes and visualized data to find my story.

**Data Transformation:** Given data has 10 rows for 10 matches and each row contain data variables for both players. I pivot down this data and created new row for each player. I created two new derived data columns – Victory , Won Set1. Reason for transforming data is the ease of reading and understanding data on player level. It is easier to compare player level performance as level as match level information if we have data in pivot down form.

Victory – 0 or 1. If player won (1) or lost that match (1). Derived from *‘Winning Player’* Column.

Won Set1 – 0 or 1. If player won set 1 or lost the 1st set. Derived from ‘*Results’* Column.

**Assumption1:** More number of matches played := Higher Probability of winning match.

More number of matches played := More number of matches won

This is a **player level exploration** of data. So two parameters I want to see w.r.t to total matches played are :

* Number of matches won by player
* Probability of Player winning = \_\_\_\_\_\_No. of matches won by the player\_\_\_\_\_\_\_

Total No. of final matches played by the Player

**Visualization of data :**

* **Use of bubble plots**: Bubble plots are best to depict the size difference. Larger the size of bubble, higher the number of matches played by that player. Bubbles also have large area which can be filled with colors to represent another measure.
* **Use of Sequential Color**: Sequential Color are best to depict the continuous range of values, here I wanted to show if probability (color) will increase with no. of matches played(size). Shades of blue are chosen as it is easy to spot the difference in its sequential shades and also it is subtle color.
* **Use of Text:** I used text labels to mark Players and no. of matches they won.

Hence, I used size, color and text successfully to highlight the purpose of this hypothesis.

**Conclusion:** Assumption1 is wrong as we can clearly see in plot neither no. of wins always increase with size of

bubble nor color of bubble always gets darker with size of bubble.

**Assumption 2:**  Win set1 := Win the match

For this hypothesis I will be using two derived columns – Victory, Won Set1.

This a **match level as well as player level exploration** of data.

**Visualization of data:**

* **Axis:** x-axis represents players and y-axis represents Year of match. This classic use of 2-D plot allows us to see all 10 matches and who were the players playing that match and against whom.
* **Triangular Shapes:** Upward triangle depicts winning of Set1 and Downward triangle depicts loss of set. I used these shapes as they are mostly used depict win and loss in daily life like share market, hence they are easy to spot and interpret their meaning.
* **Color of Triangles:** Color of triangle depicts if player won the match or not. I used red to depict loss and green to depict green because they are very natural colors which are used at many places (like traffic lights, share market) to show gain and loss and easy to interpret.

Hence I used axis, shape and color to represent 4- dimensional data which give information both at math level and player level.

**Conclusion:** From the plot it can be seen that if a player won first set he also won the match. Only exception is Novak Djokovic who lost first set in two matches but still managed to win both matches hence only two matches are outliers.

**My Story:** If a player has played a greater number of final matches then he may or may not win a greater number of matches. Concluding that just reaching at finals every time does not increase player’s chance of winning. But if a player manages to win his 1st set then it is highly probable that he will win the match. It may be due to increase in his confidence after winning his first set.