Workflow Methodology for Big Ant Studios Data Test

Step 1:

Combine all ATP and Challenger tabs into a new tab (All\_Athletes) so all athletes are together.

Step 2: Wrangle data into clean formatting

* Add a new column referencing the competition name for each athlete.
* Reformat the Date\_Of\_Birth column into DD/MM/YY format and use “ **=DATEDIF(cell\_number, TODAY(), "Y")** “ to fill in missing age values.
* Use **Ctrl + H** to find values and replace them:
* *Gender* = convert all data in either “Male” or “Female”.
* *Handedness* = convert all data into either “Right” or “Left”.
* *Backhand* = convert all data into either 1 or 2 depending on how many hands they use on a backhand shot.
* Convert heights into centimetres (cm):
* If in metres (*i.e. 1.78m*) -> **multiply by 100** ( *= 178cm*)
* If in feet and inches (*i.e. 6ft 2in*) -> **multiply feet by 12 and add inches** ( *= 6\*12 + 2 = 74 inches*) -> **multiply inches by 2.56** ( *= 74 inches \* 2.56 = 189cm*).
* Convert weight into kilogram (kg):
* To convert pounds (lb) into kg, use formula **=ROUND(cell\_number \* 0.45359237, 1).**
* Cells highlighted in RED represent data that was deemed unfillable, as it related to sensitive information (e.g., birthdates and ages) that could not be correctly generated for fictional athletes.

Step 3: Standardise attribute data layout.

* Standardise the format of the attribute data onto the 0-100 scale.
* Some inputted data was in the form:
* “X / 10” -> multiply data points by 10 and remove “/ 10” using **Ctrl + H**.
* “X / 100” -> remove the “/ 100” using **Ctrl + H**.
* Data in decimal format Eg. 6.4 -> multiply data by 10 -> 64.

Step 4: Fill in ATP\_Ranking, Ranking\_Points and Challenger\_Ranking data.

* One of the columns was completed using the other two as relative markers. This is denoted where cells are PURPLE. For example, *Kaden Simmons*’ ATP\_Ranking was inputted as they had the same number of Ranking\_Points as *Milo Hepner*.

Step 5:

Create new dataframes labelled “Male” and “Female” and copy the respective athlete information and attribute data into each.

* I thought it was important to separate athletes based on gender for more accuracy of player skill averages and comparisons against each other.

Step 6: For both “Male” and “Female” datasets:

* Calculate the mean and median values for each attribute.
* Calculate the mean attribute rating for each player.
* Sort players by ATP\_Ranking.
* Where athletes have a small amount of missing data (1-3 points), use their average attribute rating to fill in empty cells. This is denoted where cells are GREEN. This is a suitable method as it is less likely to distort their profile than the attribute mean and stays consistent with their overall skill level.
* If an athlete has a large amount of data missing, fill in using the average of the 2 athletes above them and the 2 athletes below them in the ATP rankings. This is denoted where cells are BLUE. This is a suitable method of completing data as it preserves realism and reflects quality based on ATP\_Ranking.
* *Sonia Munoz* had a large amount of missing data and was the highest ranked female player in the dataset by some margin. As a result, I couldn't use the averages of players higher ranked so instead assigned attribute ratings based on how much higher the other stats were compared to the players below. These are denoted in PURPLE.

Step 7: Create two new tabs titled “Male\_Player\_Overalls” and “Female\_Player\_Overalls”

* Copy and paste player names and complete attribute stats into the relative tabs.

Step 8: Generating attribute overalls and player overalls

* Each attribute (Serve, Forehand, Backhand and Attack) is generated an overall which is then used to generate the overall for the player.
* Equation: Attribute OVERALL = **SUM((Attribute\_Rating \* Weighting) + … + …)**
* Attribute weightings: Each component of the attribute is weighted according to my opinion on the importance of it during match play.

| **Attribute** | **Percentage** | **Weighting** |
| --- | --- | --- |
| **Serve\_Flat** | **40** | **0.4** |
| **Serve\_Top\_Spin** | **35** | **0.35** |
| **Serve\_Slice** | **25** | **0.25** |
| **Forehand\_Flat** | **25** | **0.25** |
| **Forehand\_Top\_Spin** | **20** | **0.2** |
| **Forehand\_Slice** | **20** | **0.2** |
| **Forehand\_Volley** | **15** | **0.15** |
| **Forehand\_Drop\_Shot** | **15** | **0.15** |
| **Forehand\_Lob** | **5** | **0.05** |
| **Backhand\_Flat** | **25** | **0.25** |
| **Backhand\_Top\_Spin** | **20** | **0.2** |
| **Backhand\_Slice** | **20** | **0.2** |
| **Backhand\_Volley** | **15** | **0.15** |
| **Backhand\_Drop\_Shot** | **15** | **0.15** |
| **Backhand\_Lob** | **5** | **0.05** |
| **Power** | **30** | **0.3** |
| **Reflexes** | **20** | **0.2** |
| **Serve\_Speed** | **15** | **0.15** |
| **Stamina** | **15** | **0.15** |
| **Running\_Speed** | **20** | **0.2** |

* Generate player overalls using:

=**SUM((Serve\_Attribute\*Weighting)+(Forehand\_Attribute\*Weighting)+(Backhand\_Attribute\*Weighting)+(Attack\_Attribute\*Weighting))**

Weightings are based on my opinions of ranking of importance during match play.

| **Attribute** | % | Weighting |
| --- | --- | --- |
| **Serve** | 25 | 0.25 |
| **Forehand** | 25 | 0.25 |
| **Backhand** | 15 | 0.15 |
| **Attack** | 35 | 0.35 |

All players have generated an overall rating 0-100 and can be ranked from best to worst. It is important to note that the ratings may not fully align with the ATP\_Ranking as ATP Rankings are based on tournament points (performance & consistency), while the overall ranking is built from attributes (skills). So a “high skill” player in the dataset might not match perfectly with a “high ranking” player.

Time used:

*Data Cleaning (Steps 1-4)*: 2 hours

*Data Inputting (Steps 5 and 6)*: 1.5 hours

*Weightings and Overall Generation*: (Steps 7 and 8): 1.5 hours

*Explanation*: 30 minutes

**TOTAL**: 5.5 hours