Mini Project 1 - Tennis Dataset

Data Source:

Input file name : “TennisData.csv”

All data is in csv format, ready for use within standard spreadsheet applications. The dataset used is Tennis data for ATP (Men) and WTA (Women) player’s data from 2013 to 2016.

This dataset is a subset of the original dataset that is available on Kaggle and http://www.tennis-data.co.uk/data.php .

Data Exploration and Cleaning Steps:

a. While combining the data from Kaggle and tennis-data.co.uk, a few columns which were relevant to betting were dropped. The focus of the analysis was more on the players, their wins, losses, tournaments, and different types of surfaces. Other columns were identified as irrelevant and dropped while preparing the source file.
b. On initial review, few columns had null values. Although the ideal practice is to replace these with median or mean values, for this analysis, the null values are replaced with 0 after creating data frames. This is done using info(), fillna(0).

Target Analysis:

The dataset is used to answer the following questions:

a. Does expertise on one tennis surface influence win rate on other type of tennis surfaces?

Unit of analysis: Win, Loss, Surface

Comparison: For each player, compute the correlation coefficient of wins across different Court surfaces

Output: Should be in a file (Surface_Influence.csv) with player wins across different surfaces

b. **Which Tennis Player is better suited for all surfaces?**

   Unit of analysis: Win, Loss, Surface

   Comparison: For each player, compute the average percentage win rate irrespective of the surface and find the most adaptable player.

   Output: Should be in a file (Overall_Top3.csv) with top 3 player with highest win rate across different surfaces

The program reads the TennisData.csv and answers both questions using pandas and numpy. Before answering the questions, the program checks the dataset by getting pre-requisite information like shape, information, description of the dataframes.

This provides with the dataframe schema, statistical information on quantitative columns and information of Null Value count. This information is used to clean the data by filtering and updating Null Values with 0.

Every question has individual dataframe copies. Even if their purpose is same and can be reused, for the purpose of demonstration , different copies of dataframe with same values are created.

The output files are relatively simple in structure as compared to the input files.

1. Question 1 output is coefficient correlation of different surfaces to help identify what surface expertise is beneficial to play and win on other surfaces
2. Question 2 output is average of win percentages across different surfaces for top 3 players. This is used to infer the player with highest combined average across all the 3 surfaces and statistically identify the best allrounder.