MVIZ5303 Data Choice Submission

Evan O’Neil
2/11/25

Data Source

The dataset was obtained from Kaggle's "ATP Tennis Rankings, Results & Stats (1968-2023)" collection, specifically focusing on the 2023 match results.

Source URL:
https://www.kaggle.com/datasets/warcoder/atp-tennis-rankings-results-and-stats1968-2023/data

Data Access Methods

The data can be accessed in two ways:

1. Direct Download:
   - Visit the Kaggle Source URL
   - Click the "Download" button in the upper right corner, you may need to sign up for a Kaggle account
   - Extract the ZIP file and locate "atp_matches_2023.csv"

2. In the R Studio environment using kaggler:

```
 library(kaggler)

# Set up Kaggle credentials (needs to be done once)
# kgl_auth(creds_file = "kaggle.json")

# Download the dataset
 kgl_download_dataset("warcoder/atp-tennis-rankings-results-and-stats1968-2023")

# Unzip and read the 2023 matches file
matches_2023 <- read_csv("atp_matches_2023.csv")
```

Why I Chose This Dataset

I chose this dataset because it offers comprehensive professional tennis match statistics from the 2023 ATP Tour season. The data presents multiple opportunities for interesting visualizations and analysis, including:

- Performance patterns across different playing surfaces
- Tournament-level analysis of player statistics
- Service and return game effectiveness metrics
- Match duration and competitive balance indicators

On a personal note, I work with data representing bleak issues most of the time and I chose to take a break from that and focus on a personal interest of mine. However, I will return to climate, disasters, and the fraying social fabric of our society shortly.

Dataset Description

The file "atp_matches_2023.csv" contains:

- Number of rows: 2,242 (representing individual matches)
- Number of columns: 49
- Format: CSV
- File size: 470kb

Key data points include:

- Match outcomes and scores
- Player statistics (serves, returns, break points)
- Tournament information (surface, level, draw size)
- Player rankings and points
- Match duration and game statistics

The data is well-structured and requires minimal cleaning, though some columns contain floating-point numbers that may need special handling for visualization purposes.

There is a possibility to include an extra data set, I have two ideas at this point though I haven't identified a specific dataset to use yet.

1) Match prize money to tournaments
2) Match racquet make and model to players