Gary Kelly – C16380531

TECHNOLOGICAL UNIVERSITY DUBLIN

Visualising Data Assignment 2020 Dataset exploration

Table of Contents

[Table of Tables 1](#_Toc39077254)

[Dataset Exploration 2](#_Toc39077255)

[Kaggle Datasets: 2](#_Toc39077256)

[Scraped Datasets: 5](#_Toc39077257)

[Tournament Index 5](#_Toc39077258)

[Location Dataset 6](#_Toc39077259)

[Final Dataset 7](#_Toc39077260)

[Appendix 8](#_Toc39077261)

[WebScraper.py 8](#_Toc39077262)

[Api.py 9](#_Toc39077263)

[TournamentIndex.py 9](#_Toc39077264)

# Table of Tables

[Table 1 Kaggle Datasets Used Columns 3](#_Toc39077269)

[Table 2 Scraped Dataset Used Columns 5](#_Toc39077270)

[Table 3 Tournament Index Dataset Used Columns 6](#_Toc39077271)

[Table 4 Location Dataset Used Columns 6](#_Toc39077272)

[Table 5 Final Dataset 7](#_Toc39077273)

# Dataset Exploration

In this section, the datasets used for this project will be discussed. The following image shows an overview of the datasets used and where they were accessed from. This image will be expanded on in this document, going into further detail on the columns taken from each dataset.

A picture containing parking, meter, side, screen

Description automatically generated

Figure Overview of Used Datasets

## Kaggle Datasets:

**URL of Datasets**: [https://www.kaggle.com/hakeem/atp-and-wta-tennis-data/data#](https://www.kaggle.com/hakeem/atp-and-wta-tennis-data/data)

**Context of Datasets:**

The link above contains two datasets – df\_atp.csv and df\_wta.csv. These datasets store the ATP and WTA matches results and betting odds from year 2000 until 2019. The ATP dataset contains match results dating back to January 2000 and betting odds dating back to 2001. The WTA dataset contains match results and betting odds from 2007 onwards.

**Description of Datasets:**

Both of these datasets share a similar column structure and are comprised of two parts – results data and betting data. For this project none of the betting data columns were of use and so they were dropped. The following table shows the columns taken from the results data of this dataset, a description of what they are and the column in which they were used for in the final dataset:

|  |  |  |
| --- | --- | --- |
| Column | Description | Final Use |
| ATP | A unique tournament identifier for the male tournaments which start at 1 and end at 66  (Continuous Numeric) | Tournament.ID |
| WTA | A unique tournament identifier for the female tournaments which start at 1 and end at 58.  (Continuous Numeric) | Tournament.ID |
| Location | Venue of tournament  (Location) | Location |
| Tournament | Name of tournament  (Text) | Tournament |
| Series | Name of ATP tennis series  (Categorical) | Tier |
| Tier | Tier of WTP tennis series  (Categorical) | Tier |
| Court | Type of court (outdoors or indoors)  (Categorical) | Court |
| Surface | Type of surface (clay, hard, carpet or grass)  (Categorical) | Surface |
| Round | Round of match  (Categorical) | NumRounds |
| Best of | Maximum number of sets playable in match  (Categorical) | Best of |
| WRank | ATP Entry ranking of the match winner as of the start of the tournament  (Continuous Numeric) | AvgRank, MinRank |
| LRank | ATP Entry ranking of the match loser as of the start of the tournament  (Continuous Numeric) | AvgRank, MinRank |
| WPts | ATP Entry points of the match winner as of the start of the tournament  (Continuous Numeric) | AvgPts, MinPts |
| LPts | ATP Entry points of the match loser as of the start of the tournament  (Continuous Numeric) | AvgPts, MinPts |

Table Kaggle Datasets Used Columns

**Calculated Columns:**

As can be seen in the above tables, some of the columns used from the above datasets were useful in deriving a new column for the final dataset. In this section I will discuss how the values of these columns were used in order to calculate the values for the resulting final dataset.

* **Tournament.ID:** The tournament id column is the index column for the final dataset. It’s composed using the ATP and WTA columns from the original datasets. Each of the ATP tournament ids remain the same as in their dataset and each of the WTA tournament ids are increased by the max value of the ATP tournament id column (66) in order to remain unique.
* **NumRounds:** The NumRounds column is used to store the number of rounds for each tournament in the final dataset. It’s calculated by counting the number of unique values in the Round column of the Kaggle datasets for each competition.
* **AvgRank:** The AvgRank column of the final dataset shows the average ATP/WTA ranking of the entrants to each competition last year. It’s calculated by adding the mean of the WRank and LRank columns of each of the rows for the first round of each competition and dividing by 2. By doing so you get the mean ranking of a player in the first round of the tournament.
* **MinRank:** The MinRank column of the final dataset shows the lowest ATP/WTA ranking of the entrants to each competition last year. It’s calculated by getting the minimum value of the WRank and LRank columns for the first round of each competition and seeing which value is lower. By doing so you get the lowest ranked player in the first round of the tournament.
* **AvgPoints:** The AvgPoints column of the final dataset shows the average ATP/WTA ranking points of the entrants to each competition last year. It’s calculated similarly to the AvgRank column except uses the WPts and LPts columns instead.
* **MinPoints:** The MinPoints column of the final dataset shows the lowest ATP/WTA ranking points of the entrants to each competition last year. It’s calculated similarly to the MinRank column except uses the WPts and LPts columns instead.

**Acknowledgements:**

The Data is scraped from http://www.tennis-data.co.uk.

Tennis-Data would like to acknowledge the following sources which are currently utilised in the compilation of Tennis-Data's results and odds files.

* Results:
  + Xscores - http://www.xscores.com/
  + ATPtennis.com - http://www.atptennis.com/
  + ATP Tour Rankings and Results Page - http://www.stevegtennis.com/
  + Livescore - http://www.livescore.net/
* Rankings:
  + ATPtennis.com - http://www.atptennis.com/
  + ATP Tour Rankings and Results Page - http://www.stevegtennis.com/
  + WTA Tour Rankings - http://www.sonyericssonwtatour.com

Betting odds for matches generally represent the most recent before play starts, as reported by oddsportal.com and the individual bookmakers.

## Scraped Datasets:

**URL of Datasets**: Data was scraped from the ATP and WTA tournament websites <https://www.atptour.com/en/tournaments> and <https://www.wtatennis.com/tournaments>.

**Context of Datasets:**

The links above are the sources for the information scraped into the prize\_money.csv file by the webScraper.py and api.py files. These files are shown in the appendix. This information contains information regarding each tournament such as the dates and the prize money.

**Description of Datasets:**

Both of these datasets share a similar column structure with 5 shared columns. The WTA scraped dataset has some additional information like location data and court data which is already contained in my Kaggle dataset so will be dropped. This dataset is joined using a left merge with the Kaggle datasets described above using the Tournament and Type columns. The following table shows the columns taken from both of the scraped datasets and their name in the resulting final dataset:

|  |  |  |
| --- | --- | --- |
| Column | Description | Final Use |
| Title | Tournament Name  (Text) | Tournament  (Used for merge) |
| Type | Tournament Type  (Categorical) | Type  (Used for merge) |
| startDate | Start date for tournament  (Temporal) | startDate |
| endDate | End date for tournament  (Temporal) | endDate |
| prizeMoney | Total Financial Commitment of tournament  (Continuous Numeric) | prizeMoney |
| prizeMoneyCurrency | Currency of tournament prize money  (Categorical) | prizeMoneyCurrency |

Table Scraped Dataset Used Columns

## Tournament Index

**Context of Dataset:**

In order to merge the Kaggle datasets with the scraped datasets, another intermediary dataset had to be created in order to serve as an index for the Tournament names in both datasets. This dataset was created in python by reading each of the above datasets into dataframes and using the str.contains function to find the tournament values that were the same and storing these names in a new dataset. The code for this program can be seen in the appendix under TournamentIndex.py.

**Description of Dataset:**

The dataset contains three columns which maps the tournament names in the Kaggle dataset to the related tournament name in the Scraped dataset.

|  |  |  |
| --- | --- | --- |
| Column | Description | Final Use |
| Tournament | Tournament name  (Kaggle Dataset) | Tournament  (Used for merge) |
| Title | Tournament name  (Scraped dataset) | Tournament |
| Type | Tournament type | Type |

Table Tournament Index Dataset Used Columns

## Location Dataset

**URL of Datasets**: <https://simplemaps.com/data/world-cities>

**Context of Dataset:**

This dataset was used to gain further geographic data on the locations of each tournament for mapping. The dataset contains a list of about 15 thousand prominent cities.

**Description of Dataset:**

The original dataset contains 11 columns but only 5 of these were relevant for the final dataset. These columns can be seen below:

|  |  |  |
| --- | --- | --- |
| Column | Description | Final Use |
| City Ascii | City  (Ascii Text) | City  (Used for merge) |
| Country | Country of city  (Text) | Country |
| Lat | Latitude of city  (Temporal) | Lat |
| Lng | Longitude of city  (Temporal) | Lng |
| Population | Population of city  (Continuous Numeric) | Population |

Table Location Dataset Used Columns

## Final Dataset

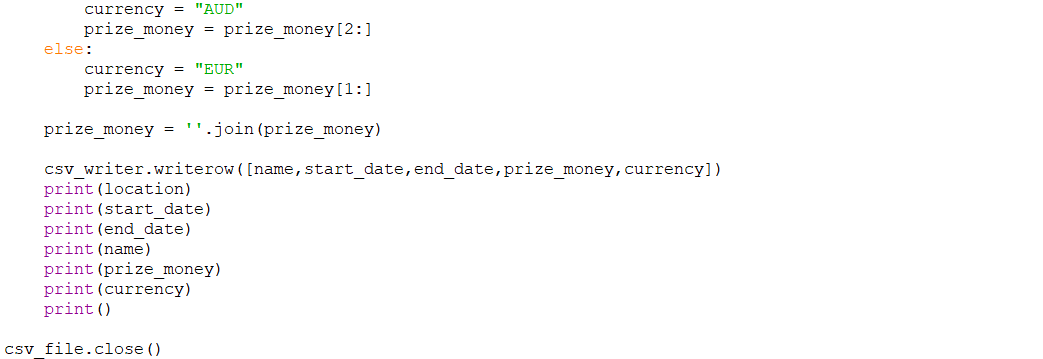
The resulting final dataset can be seen below:

|  |  |  |
| --- | --- | --- |
| Column | Description | Values |
| Tournament.ID | Unique identifier for tournament | Identifier |
| Tournament | Tournament name | Text |
| Type | Tournament type | Categorical  (ATP / WTA) |
| Location | Location of tournament | Text |
| Country | Country of tournament | Text |
| Lat | Latitude of location | Co-ordinates |
| Lng | Longitude of location | Co-ordinates |
| Best.of | Maximum number of sets playable in match | Categorical  (3 / 5) |
| Court | Type of court  (Categorical) | Categorical  (outdoors / indoors) |
| Surface | Type of surface  (Categorical) | Categorical  (clay, hard, carpet or grass) |
| Tier | Tier of tournament | Categorical  (ATP250, ATP500, Grand Slam, International, Masters 1000, Masters Cup, Premier and Tour Championships) |
| StartDate | Starting date for tournament | Date |
| EndDate | End date for tournament | Date |
| AvgPoints | Average points of entrants in last year’s tournament | Continuous Numeric |
| MinPoints | Minimum points of entrants in last year’s tournament | Continuous Numeric |
| AvgRank | Average rank of entrants in last year’s tournament | Continuous Numeric |
| MinRank | Minimum rank of entrants in last year’s tournament | Continuous Numeric |
| PrizeMoney | Total Financial Commitment of tournament | Continuous Numeric |
| PrizeMoneyCurrency | Currency of prize money for tournament | Categorical  (EUR, USD, AUD) |

Table Final Dataset

# Appendix

## WebScraper.py



## Api.py



## TournamentIndex.py

