# 1b. ATP Initial Exploration Part 2

This notebook is a continuation of "1a. ATP Initial Exploration Part 1". The first notebook focused on cleaning and wrangling steps. This second notebook will create subsets to use for analysis.

# Table of Contents

## List of All New Subsets

- New Main Dataframe: "df_matchstats"
- By Number of Sets: “df_BO3” and “df_BO5”
- Hard Courts (3 total): "df_hard", “df_hard_BO3” and “df_hard_BO5”
- Clay Courts (3 total): "df_clay", “df_clay_BO3” and “df_clay_BO5”
- Grass Courts (3 total): "df_grass", “df_grass_BO3” and “df_grass_BO5”

## 1. Importing Libraries and PKL File

```
#Import Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import scipy
import matplotlib

#Set Path
path = r'/Users/tristansavella/Desktop/Important Things/Data
Analytics/CareerFoundry/Data Immersion/Achievement 6/Master Folder
ATP/02 Data'
```

```
#Import df_post2000
df_post2000 = pd.read_pickle(os.path.join(path, 'Prepared
Data','df_post2000.pkl'))

#Show all columns
pd.set_option('display.max_columns', None)

#Show all rows
pd.set_option('display.max_rows', None)

#Check Head
df_post2000.head()
```

```
       tourney_id tourney_name surface  draw_size tourney_level
tourney_date  \
119317   2000-301     Auckland    Hard         32             A
20000110
119318   2000-301     Auckland    Hard         32             A
20000110
119319   2000-301     Auckland    Hard         32             A
20000110
119320   2000-301     Auckland    Hard         32             A
20000110
119321   2000-301     Auckland    Hard         32             A
20000110

        match_num winner_id  winner_seed winner_entry
winner_name  \
119317          1    103163          1.0          NaN           Tommy
Haas
119318          2    102607          NaN            Q        Juan
Balcells
119319          3    103252          NaN          NaN       Alberto
Martin
119320          4    103507          7.0          NaN  Juan Carlos
Ferrero
119321          5    102103          NaN            Q         Michael
Sell

       winner_hand  winner_ht winner_ioc  winner_age loser_id
loser_seed  \
119317           R      188.0        GER        21.7   101543
NaN
119318           R      190.0        ESP        24.5   102644
NaN
119319           R      175.0        ESP        21.3   102238
NaN
119320           R      183.0        ESP        19.9   103819
NaN
119321           R      180.0        USA        27.3   102765
```

```
4.0

       loser_entry           loser_name loser_hand  loser_ht loser_ioc
\
119317         NaN         Jeff Tarango          L     180.0       USA

119318         NaN     Franco Squillari          L     183.0       ARG

119319         NaN  Alberto Berasategui          R     173.0       ESP

119320         NaN        Roger Federer          R     185.0       SUI

119321         NaN       Nicolas Escude          R     185.0       FRA


        loser_age           score best_of round  minutes  w_#aces
w_#dfs  \
119317       31.1     7-5 4-6 7-5       3   R32    108.0     18.0
4.0
119318       24.3         7-5 7-5       3   R32     85.0      5.0
3.0
119319       26.5         6-3 6-1       3   R32     56.0      0.0
0.0
119320       18.4         6-4 6-4       3   R32     68.0      5.0
1.0
119321       23.7  0-6 7-6(7) 6-1       3   R32    115.0      1.0
2.0

        w_#ServePoints  w_#1stServesIn  w_#1stWon  w_#2ndWon
w_#ServeGames  \
119317            96.0            49.0       39.0       28.0
17.0
119318            76.0            52.0       39.0       13.0
12.0
119319            55.0            35.0       25.0       12.0
8.0
119320            53.0            28.0       26.0       15.0
10.0
119321            98.0            66.0       39.0       14.0
13.0

        w_bpSaved  w_#bpFaced  l_#aces  l_#dfs  l_#ServePoints  \
119317        3.0         5.0      7.0     8.0           106.0
119318        5.0         6.0      5.0    10.0            74.0
119319        1.0         1.0      0.0     6.0            56.0
119320        0.0         0.0     11.0     2.0            70.0
119321        6.0        11.0      8.0     8.0            92.0

        l_#1stServesIn  l_#1stWon  l_#2ndWon  l_#ServeGames  l_bpSaved
\
119317            55.0        39.0        29.0             17.0         4.0

119318            32.0        25.0        18.0             12.0         3.0

119319            33.0        20.0         7.0              8.0         7.0

119320            43.0        29.0        14.0             10.0         6.0

119321            46.0        34.0        18.0             12.0         5.0


        l_#bpFaced  winner_rank  winner_rank_points  loser_rank  \
119317         7.0         11.0              1612.0        63.0
119318         6.0        211.0               157.0        49.0
119319        11.0         48.0               726.0        59.0
119320         8.0         45.0               768.0        61.0
119321         9.0        167.0               219.0        34.0

        loser_rank_points  w_#2ndServePoints  w_%1stServesIn  w_
%1stWon  \
119317              595.0               47.0              51
79
119318              723.0               24.0              68
75
119319              649.0               20.0              63
71
119320              616.0               25.0              52
92
119321              873.0               32.0              67
59

        w_%2ndWon  l_#2ndServePoints  l_%1stServesIn  l_%1stWon  l_
%2ndWon  \
119317         59               51.0              51         70
56
119318         54               42.0              43         78
42
119319         60               23.0              58         60
30
119320         60               27.0              61         67
51
119321         43               46.0              50         73
39

        Year
119317  2000
119318  2000
119319  2000
119320  2000
119321  2000
```

# 2. Creating Subsets

2.1 Create main dataframe, 'df_matchstats', consisting only of match statistics

2.2 Divide 'df_matchstats' into two groups; best of 3 and best of 5

2.3 Create Subsets for Matches Played By Surface (Hard, Clay and Grass)

2.4 Create Subsets for Matches Played by Big 3 (Participate, Win and Lose)

# Final List of Subsets

- Main Dataframe: "df_matchstats"
- By Number of Sets: “df_BO3” and “df_BO5”
- Hard Courts (3 total): "df_hard", “df_hard_BO3” and “df_hard_BO5”
- Clay Courts (3 total): "df_clay", “df_clay_BO3” and “df_clay_BO5”
- Grass Courts (3 total): "df_grass", “df_grass_BO3” and “df_grass_BO5”
- Big 3 Wins (3 total): "df_big3_win", “df_big3_win_BO3” and “df_big3_win_BO5”
- Big 3 Loses (3 total): "df_big3_lose", “df_big3_lose_BO3” and “df_big3_lose_BO5

```
#drop irrelevant columns

df_matchstats = df_post2000.drop(columns = ['tourney_date',
                                            'draw_size',
                                            'match_num',
                                            'score',
                                            'winner_entry',
                                            'winner_hand',
                                            'loser_entry',
                                            'loser_hand',
                                            'winner_rank_points',
                                            'loser_rank_points',
                                            'winner_seed',
                                            'loser_seed',
                                            ])

#Reorder Columns

df_matchstats = df_matchstats[['tourney_id',
                               'Year',
                                       'tourney_name',
                                       'surface',
                                       'tourney_level',
                                       'winner_id',
                               'winner_ioc',
                                       'winner_name',
                                       'winner_age',
                                       'winner_rank',
                               'winner_ht',
                                       'loser_id',
                               'loser_ioc',
                                       'loser_name',
                                       'loser_rank',
                               'loser_ht',
                                       'loser_age',
                                       'best_of',
                                       'round',
                                       'minutes',
                                       'w_#ServeGames',
                                       'w_#aces',
                                       'w_#dfs',
                                       'w_#ServePoints',
                                       'w_#1stServesIn',
                                       'w_#2ndServePoints',
                                       'w_%1stServesIn',
                                       'w_#1stWon',
                                       'w_%1stWon',
                                       'w_#2ndWon',
                                       'w_%2ndWon',
                                       'w_bpSaved',
                                       'w_#bpFaced',
                                       'l_#ServeGames',
                                       'l_#aces',
                                       'l_#dfs',
                                       'l_#ServePoints',
                                       'l_#1stServesIn',
                                       'l_#2ndServePoints',
                                       'l_%1stServesIn',
                                       'l_#1stWon',
                                       'l_%1stWon',
                                       'l_#2ndWon',
                                       'l_%2ndWon',
                                       'l_bpSaved',
                                       'l_#bpFaced']]

#check head
df_matchstats.head()
```

```
       tourney_id  Year tourney_name surface tourney_level
winner_id  \
119317   2000-301  2000     Auckland    Hard             A    103163

119318   2000-301  2000     Auckland    Hard             A    102607

119319   2000-301  2000     Auckland    Hard             A    103252

119320   2000-301  2000     Auckland    Hard             A    103507

119321   2000-301  2000     Auckland    Hard             A    102103
```

```
       winner_ioc          winner_name  winner_age  winner_rank  
winner_ht  \
119317        GER           Tommy Haas        21.7         11.0   
188.0
119318        ESP        Juan Balcells        24.5        211.0   
190.0
119319        ESP       Alberto Martin        21.3         48.0   
175.0
119320        ESP  Juan Carlos Ferrero        19.9         45.0   
183.0
119321        USA         Michael Sell        27.3        167.0   
180.0

       loser_id loser_ioc           loser_name  loser_rank  
loser_ht  \
119317   101543       USA         Jeff Tarango        63.0     180.0   

119318   102644       ARG     Franco Squillari        49.0     183.0   

119319   102238       ESP  Alberto Berasategui        59.0     173.0   

119320   103819       SUI        Roger Federer        61.0     185.0   

119321   102765       FRA       Nicolas Escude        34.0     185.0   


        loser_age best_of round  minutes  w_#ServeGames  w_#aces  
w_#dfs  \
119317       31.1       3   R32    108.0           17.0     18.0   
4.0
119318       24.3       3   R32     85.0           12.0      5.0   
3.0
119319       26.5       3   R32     56.0            8.0      0.0   
0.0
119320       18.4       3   R32     68.0           10.0      5.0   
1.0
119321       23.7       3   R32    115.0           13.0      1.0   
2.0

        w_#ServePoints  w_#1stServesIn  w_#2ndServePoints  w_
%1stServesIn  \
119317            96.0            49.0               47.0   
51
119318            76.0            52.0               24.0   
68
119319            55.0            35.0               20.0   
63
119320            53.0            28.0               25.0   
```

```
52
119321            98.0            66.0               32.0
67

        w_#1stWon  w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved
w_#bpFaced  \
119317       39.0         79       28.0         59        3.0
5.0
119318       39.0         75       13.0         54        5.0
6.0
119319       25.0         71       12.0         60        1.0
1.0
119320       26.0         92       15.0         60        0.0
0.0
119321       39.0         59       14.0         43        6.0
11.0

        l_#ServeGames  l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
\
119317           17.0      7.0     8.0           106.0            55.0

119318           12.0      5.0    10.0            74.0            32.0

119319            8.0      0.0     6.0            56.0            33.0

119320           10.0     11.0     2.0            70.0            43.0

119321           12.0      8.0     8.0            92.0            46.0


        l_#2ndServePoints  l_%1stServesIn  l_#1stWon  l_%1stWon
l_#2ndWon  \
119317               51.0              51       39.0         70
29.0
119318               42.0              43       25.0         78
18.0
119319               23.0              58       20.0         60
7.0
119320               27.0              61       29.0         67
14.0
119321               46.0              50       34.0         73
18.0

        l_%2ndWon  l_bpSaved  l_#bpFaced
119317         56        4.0         7.0
119318         42        3.0         6.0
119319         30        7.0        11.0
119320         51        6.0         8.0
119321         39        5.0         9.0
```

```
#describe
df_matchstats.describe()

         winner_age   winner_rank     winner_ht    loser_rank
loser_ht  \
count  62530.000000  62441.000000  62332.000000  62234.000000
61801.000000
mean      26.407667     62.159607    186.182603     95.891731
185.611317
std        3.928157     86.099346      6.826430    132.046004
6.776734
min       14.900000      1.000000    163.000000      1.000000
163.000000
25%       23.500000     17.000000    183.000000     35.000000
181.000000
50%       26.200000     42.000000    185.000000     65.000000
185.000000
75%       29.100000     79.000000    190.000000    106.000000
190.000000
max       42.300000   1890.000000    211.000000   2159.000000
211.000000

          loser_age       minutes  w_#ServeGames       w_#aces
w_#dfs  \
count  62528.000000  61074.000000   62530.000000  62530.000000
62530.000000
mean      26.540982    106.384141      12.507484      6.915832
2.652167
std        4.002836     40.944761       4.221022      5.547041
2.291411
min       14.500000      3.000000       0.000000      0.000000
0.000000
25%       23.600000     76.000000       9.000000      3.000000
1.000000
50%       26.400000     99.000000      11.000000      6.000000
2.000000
75%       29.300000    129.000000      15.000000      9.000000
4.000000
max       46.000000   1146.000000      90.000000    113.000000
26.000000

       w_#ServePoints  w_#1stServesIn  w_#2ndServePoints  w_
%1stServesIn  \
count    62530.000000    62530.000000        62530.00000
62530.000000
mean        77.922469       47.881449           30.04102
61.061427
std         29.176144       18.902521           13.02635
8.180105
min          3.000000        1.000000            1.00000
```

```
12.000000
25%          56.000000       34.000000          21.00000
56.000000
50%          73.000000       45.000000          28.00000
61.000000
75%          94.000000       58.000000          37.00000
67.000000
max         491.000000      361.000000         130.00000
98.000000

          w_#1stWon     w_%1stWon     w_#2ndWon     w_%2ndWon
w_bpSaved  \
count  62530.000000  62530.000000  62530.000000  62530.000000
62530.000000
mean      36.233664     76.289637     16.654630     56.223413
3.464753
std       13.541931      8.141773      6.983291     10.496364
3.078031
min        0.000000      0.000000      0.000000      0.000000
0.000000
25%       27.000000     71.000000     12.000000     50.000000
1.000000
50%       34.000000     76.000000     16.000000     56.000000
3.000000
75%       43.000000     82.000000     20.000000     63.000000
5.000000
max      292.000000    100.000000     82.000000    129.000000
24.000000

         w_#bpFaced  l_#ServeGames       l_#aces        l_#dfs  \
count  62530.000000   62530.000000  62530.000000  62530.000000
mean       5.038893      12.300112      5.108460      3.378682
std        4.034177       4.222271      4.882427      2.530127
min        0.000000       0.000000      0.000000      0.000000
25%        2.000000       9.000000      2.000000      2.000000
50%        4.000000      11.000000      4.000000      3.000000
75%        7.000000      15.000000      7.000000      5.000000
max       30.000000      91.000000    103.000000     26.000000

       l_#ServePoints  l_#1stServesIn  l_#2ndServePoints  l_
%1stServesIn  \
count    62530.000000    62530.000000       62530.000000
62530.000000
mean        80.973885       48.472013          32.501871
59.158452
std         29.154055       19.195435          12.949551
8.272118
min          3.000000        1.000000           1.000000
12.000000
25%         60.000000       35.000000          23.000000
```

```
54.000000
50%          76.000000        45.000000          30.000000
59.000000
75%          97.000000        59.000000          40.000000
65.000000
max         489.000000       328.000000         161.000000
97.000000

          l_#1stWon     l_%1stWon     l_#2ndWon     l_%2ndWon
l_bpSaved  \
count  62530.000000  62530.000000  62530.000000  62530.000000
62530.000000
mean      32.332768     65.442875     14.972477     44.859332
4.778330
std       14.349978      9.668002      7.210501     10.276116
3.270382
min        0.000000      0.000000      0.000000      0.000000
0.000000
25%       22.000000     60.000000     10.000000     38.000000
2.000000
50%       30.000000     66.000000     14.000000     45.000000
4.000000
75%       40.000000     72.000000     19.000000     52.000000
7.000000
max      284.000000    100.000000    101.000000    100.000000
27.000000

         l_#bpFaced
count  62530.000000
mean       8.630977
std        4.142344
min        0.000000
25%        6.000000
50%        8.000000
75%       11.000000
max       38.000000
```

## 2.2 Best of 3 or 5

```
#Checking for matches played best of 3 and matches played best of 5
df_matchstats['best_of'].value_counts(dropna = False)

best_of
3    50549
5    11981
Name: count, dtype: int64

#Subset: Matches Played Best of 3 Sets
df_BO3 = df_matchstats[df_matchstats['best_of']==3]
```

```
#checking shape
df_BO3.shape
#50549 was the number I was looking for

(50549, 46)

#Subset: Matches Played Best of 5 Sets
df_BO5 = df_matchstats[df_matchstats['best_of']==5]

#checking shape
df_BO5.shape
#11981 was the number I was looking for

(11981, 46)
```

## 2.3: Create Subsets based on Surface:

```
#checking for matches by surface
df_matchstats['surface'].value_counts(dropna = False)

surface
Hard      34182
Clay      20073
Grass      6836
Carpet     1439
Name: count, dtype: int64
```

### Hard Courts

- "df_hard"
- “df_hard_BO3”
- “df_hard_BO5”

```
#creating main df for hard courts
df_hard = df_matchstats[df_matchstats['surface']=='Hard']

#checking shape
df_hard.shape
#correct number of matches

(34182, 46)

#creating 'df_hard_BO3'
df_hard_BO3 = df_BO3[df_BO3['surface']=='Hard']

#checking head
df_hard_BO3.head()

       tourney_id  Year tourney_name surface tourney_level  
winner_id  \
119317   2000-301  2000     Auckland    Hard             A    103163
```

```
119318   2000-301  2000     Auckland    Hard             A    102607

119319   2000-301  2000     Auckland    Hard             A    103252

119320   2000-301  2000     Auckland    Hard             A    103507

119321   2000-301  2000     Auckland    Hard             A    102103


       winner_ioc          winner_name  winner_age  winner_rank  
winner_ht  \
119317        GER           Tommy Haas        21.7         11.0   
188.0
119318        ESP        Juan Balcells        24.5        211.0   
190.0
119319        ESP       Alberto Martin        21.3         48.0   
175.0
119320        ESP  Juan Carlos Ferrero        19.9         45.0   
183.0
119321        USA         Michael Sell        27.3        167.0   
180.0

       loser_id loser_ioc           loser_name  loser_rank  
loser_ht  \
119317   101543       USA         Jeff Tarango        63.0     180.0

119318   102644       ARG     Franco Squillari        49.0     183.0

119319   102238       ESP  Alberto Berasategui        59.0     173.0

119320   103819       SUI        Roger Federer        61.0     185.0

119321   102765       FRA       Nicolas Escude        34.0     185.0


        loser_age best_of round  minutes  w_#ServeGames  w_#aces  
w_#dfs  \
119317       31.1       3   R32    108.0           17.0     18.0   
4.0
119318       24.3       3   R32     85.0           12.0      5.0   
3.0
119319       26.5       3   R32     56.0            8.0      0.0   
0.0
119320       18.4       3   R32     68.0           10.0      5.0   
1.0
119321       23.7       3   R32    115.0           13.0      1.0   
2.0

        w_#ServePoints  w_#1stServesIn  w_#2ndServePoints  w_
%1stServesIn  \
119317            96.0            49.0                  47.0
51
119318            76.0            52.0                  24.0
68
119319            55.0            35.0                  20.0
63
119320            53.0            28.0                  25.0
52
119321            98.0            66.0                  32.0
67

        w_#1stWon  w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved
w_#bpFaced  \
119317       39.0         79       28.0         59        3.0
5.0
119318       39.0         75       13.0         54        5.0
6.0
119319       25.0         71       12.0         60        1.0
1.0
119320       26.0         92       15.0         60        0.0
0.0
119321       39.0         59       14.0         43        6.0
11.0

        l_#ServeGames  l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
\
119317           17.0      7.0     8.0           106.0            55.0

119318           12.0      5.0    10.0            74.0            32.0

119319            8.0      0.0     6.0            56.0            33.0

119320           10.0     11.0     2.0            70.0            43.0

119321           12.0      8.0     8.0            92.0            46.0


        l_#2ndServePoints  l_%1stServesIn  l_#1stWon  l_%1stWon
l_#2ndWon  \
119317               51.0              51       39.0         70
29.0
119318               42.0              43       25.0         78
18.0
119319               23.0              58       20.0         60
7.0
119320               27.0              61       29.0         67
14.0
119321               46.0              50       34.0         73
18.0
```

```
        l_%2ndWon  l_bpSaved  l_#bpFaced
119317         56        4.0         7.0
119318         42        3.0         6.0
119319         30        7.0        11.0
119320         51        6.0         8.0
119321         39        5.0         9.0
```

```
#checking shape
df_hard_BO3.shape
```

```
(28096, 46)
```

```
#creating 'df_hard_BO5'
df_hard_BO5 = df_BO5[df_BO5['surface']=='Hard']
```

```
#checking head
df_hard_BO5.head()
```

```
       tourney_id  Year          tourney_name surface tourney_level
winner_id  \
119876   2000-337  2000                Vienna    Hard             A
102450
120032   2000-357  2000     Stuttgart Masters    Hard             M
101965
120267   2000-403  2000         Miami Masters    Hard             M
101948
120330   2000-404  2000  Indian Wells Masters    Hard             M
102374
120996   2000-429  2000             Stockholm    Hard             A
102563

       winner_ioc       winner_name  winner_age  winner_rank
winner_ht  \
119876        GBR        Tim Henman        26.0         10.0
185.0
120032        RSA    Wayne Ferreira        29.1         19.0
185.0
120267        USA      Pete Sampras        28.6          2.0
185.0
120330        ESP     Alex Corretja        25.9         26.0
180.0
120996        SWE  Thomas Johansson        25.6         57.0
180.0

       loser_id loser_ioc          loser_name  loser_rank  loser_ht  \
119876   103163       GER         Tommy Haas        28.0     188.0
120032   103720       AUS     Lleyton Hewitt         8.0     180.0
120267   102856       BRA    Gustavo Kuerten         6.0     190.0
120330   102358       SWE     Thomas Enqvist        10.0     190.0
120996   102338       RUS  Yevgeny Kafelnikov        5.0     190.0
```

```
        loser_age best_of round  minutes  w_#ServeGames  w_#aces
w_#dfs  \
119876       22.5       5     F    124.0           15.0     10.0
1.0
120032       19.6       5     F    251.0           27.0     18.0
5.0
120267       23.5       5     F    198.0           22.0     20.0
9.0
120330       26.0       5     F    120.0           14.0     10.0
1.0
120996       26.7       5     F     95.0           14.0      7.0
4.0

        w_#ServePoints  w_#1stServesIn  w_#2ndServePoints  w_
%1stServesIn  \
119876            84.0            56.0               28.0
66
120032           194.0           100.0               94.0
51
120267           148.0            94.0               54.0
63
120330            77.0            38.0               39.0
49
120996            76.0            36.0               40.0
47

        w_#1stWon  w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved
w_#bpFaced  \
119876       46.0         82       18.0         64        3.0
3.0
120032       69.0         69       52.0         55       14.0
20.0
120267       75.0         79       30.0         55        5.0
6.0
120330       30.0         78       28.0         71        6.0
7.0
120996       29.0         80       21.0         52        2.0
5.0

        l_#ServeGames  l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
\
119876           15.0      5.0     3.0            99.0            51.0

120032           26.0      8.0     8.0           170.0            84.0

120267           21.0     16.0     4.0           159.0            90.0

120330           15.0      8.0     7.0           100.0            54.0

120996           14.0      3.0    10.0            77.0            36.0
```

```
        l_#2ndServePoints  l_%1stServesIn  l_#1stWon  l_%1stWon
l_#2ndWon  \
119876               48.0              51       38.0         74
25.0
120032               86.0              49       64.0         76
40.0
120267               69.0              56       61.0         67
43.0
120330               46.0              54       34.0         62
25.0
120996               41.0              46       25.0         69
13.0

        l_%2ndWon  l_bpSaved  l_#bpFaced
119876         52        6.0         9.0
120032         46        8.0        14.0
120267         62       11.0        14.0
120330         54        8.0        13.0
120996         31        2.0         9.0
```

```
#checking shape
df_hard_BO5.shape
#6,086 + 28096 = 34182 --> checks out
```

```
(6086, 46)
```

## Clay Courts

- "df_clay"
- “df_clay_BO3”
- “df_clay_BO5”

```
#creating main df for clay courts
df_clay = df_matchstats[df_matchstats['surface']=='Clay']

#checking shape
df_clay.shape
#correct number of matches
```

```
(20073, 46)
```

```
#creating 'df_clay_BO3'
df_clay_BO3 = df_BO3[df_BO3['surface']=='Clay']

#checking head
df_clay_BO3.head()
```

```
       tourney_id  Year tourney_name surface tourney_level
winner_id  \
119348   2000-306  2000  St. Poelten    Clay             A    102247

119349   2000-306  2000  St. Poelten    Clay             A    102287

119350   2000-306  2000  St. Poelten    Clay             A    102869

119351   2000-306  2000  St. Poelten    Clay             A    103082

119352   2000-306  2000  St. Poelten    Clay             A    102446


       winner_ioc      winner_name  winner_age  winner_rank  winner_ht
\
119348        ITA  Andrea Gaudenzi        26.8         74.0      183.0

119349        ESP    Albert Portas        26.5         71.0      188.0

119350        ESP      Galo Blanco        23.6         70.0      173.0

119351        GER  Markus Hantschk        22.5         94.0      188.0

119352        UKR  Andrei Medvedev        25.7         21.0      193.0


       loser_id loser_ioc       loser_name  loser_rank  loser_ht
loser_age  \
119348   103017       GER   Nicolas Kiefer         8.0     183.0
22.8
119349   103242       ESP       Juan Giner       399.0     178.0
21.8
119350   102987       BRA         Andre Sa        86.0     185.0
23.0
119351   103819       SUI    Roger Federer        54.0     185.0
18.7
119352   102795       USA  Scott Humphries       720.0     185.0
23.9

       best_of round  minutes  w_#ServeGames  w_#aces  w_#dfs
w_#ServePoints  \
119348       3   R32     76.0            8.0      1.0     3.0
50.0
119349       3   R32     87.0           13.0      4.0     6.0
74.0
119350       3   R32     36.0            5.0      1.0     0.0
21.0
119351       3   R32     54.0            7.0      0.0     0.0
36.0
119352       3   R32    122.0           15.0      5.0     5.0
108.0

        w_#1stServesIn  w_#2ndServePoints  w_%1stServesIn
```

```
w_#1stWon  \
119348            35.0               15.0              70       19.0

119349            35.0               39.0              47       26.0

119350            12.0                9.0              57        7.0

119351            17.0               19.0              47       14.0

119352            69.0               39.0              63       46.0


        w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved  w_#bpFaced
l_#ServeGames  \
119348         54        8.0         53        4.0         8.0
8.0
119349         74       22.0         56        2.0         5.0
12.0
119350         58        7.0         77        0.0         1.0
4.0
119351         82       15.0         78        0.0         0.0
8.0
119352         66       18.0         46        7.0        12.0
16.0

        l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
l_#2ndServePoints  \
119348      0.0     4.0            53.0            24.0
29.0
119349      0.0     2.0            71.0            46.0
25.0
119350      0.0     3.0            25.0            11.0
14.0
119351      2.0     2.0            64.0            36.0
28.0
119352      4.0     4.0           104.0            67.0
37.0

        l_%1stServesIn  l_#1stWon  l_%1stWon  l_#2ndWon  l_%2ndWon
l_bpSaved  \
119348              45       11.0         45       12.0         41
5.0
119349              64       22.0         47       13.0         52
3.0
119350              44        4.0         36        4.0         28
0.0
119351              56       19.0         52       10.0         35
6.0
119352              64       45.0         67       12.0         32
4.0
```

```
        l_#bpFaced
119348        10.0
119349         9.0
119350         4.0
119351        11.0
119352        10.0
```

```
#checking shape
df_clay_BO3.shape
```

```
(16996, 46)
```

```
#creating 'df_clay_BO5'
df_clay_BO5 = df_BO5[df_BO5['surface']=='Clay']
```

```
#checking head
df_clay_BO5.head()
```

```
       tourney_id  Year          tourney_name surface tourney_level
winner_id  \
119588   2000-317  2000             Amsterdam    Clay             A
101320
119666   2000-321  2000     Stuttgart Outdoor    Clay             A
102644
120486   2000-410  2000   Monte Carlo Masters    Clay             M
101611
120549   2000-414  2000       Hamburg Masters    Clay             M
102856
120612   2000-416  2000          Rome Masters    Clay             M
102796

       winner_ioc        winner_name  winner_age  winner_rank
winner_ht  \
119588        SWE  Magnus Gustafsson        33.5         63.0
185.0
119666        ARG   Franco Squillari        24.9         20.0
183.0
120486        FRA     Cedric Pioline        30.8         12.0
188.0
120549        BRA    Gustavo Kuerten        23.6          7.0
190.0
120612        SWE      Magnus Norman        23.9          4.0
188.0

       loser_id loser_ioc       loser_name  loser_rank  loser_ht
loser_age  \
119588   103171       NED   Raemon Sluiter       142.0     185.0
22.2
119666   103292       ARG    Gaston Gaudio        41.0     175.0
21.6
```

```
120486   103103       SVK   Dominik Hrbaty        24.0     183.0
22.2
120549   103498       RUS      Marat Safin        14.0     193.0
20.2
120612   102856       BRA  Gustavo Kuerten         6.0     190.0
23.6

       best_of round  minutes  w_#ServeGames  w_#aces  w_#dfs
w_#ServePoints  \
119588       5     F    155.0           21.0      6.0     4.0
118.0
119666       5     F    228.0           22.0      8.0     2.0
130.0
120486       5     F    160.0           17.0     12.0     5.0
127.0
120549       5     F    232.0           28.0      7.0     1.0
165.0
120612       5     F    185.0           20.0     14.0     1.0
121.0

        w_#1stServesIn  w_#2ndServePoints  w_%1stServesIn
w_#1stWon  \
119588            82.0               36.0              69       71.0

119666            63.0               67.0              48       45.0

120486            61.0               66.0              48       48.0

120549            86.0               79.0              52       64.0

120612            81.0               40.0              66       57.0


        w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved  w_#bpFaced
l_#ServeGames  \
119588         86       19.0         52        0.0         0.0
19.0
119666         71       38.0         56        4.0         9.0
23.0
120486         78       33.0         50        7.0        10.0
17.0
120549         74       41.0         51        4.0        11.0
28.0
120612         70       21.0         52        5.0         9.0
19.0

        l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
l_#2ndServePoints  \
119588      4.0     2.0           121.0            70.0
51.0
```

```
119666       2.0     5.0           168.0           110.0
58.0
120486       7.0     7.0           113.0            57.0
56.0
120549      13.0     2.0           177.0            94.0
83.0
120612      15.0     5.0           142.0            68.0
74.0

        l_%1stServesIn  l_#1stWon  l_%1stWon  l_#2ndWon  l_%2ndWon
l_bpSaved  \
119588              57       46.0         65       33.0         64
3.0
119666              65       63.0         57       32.0         55
10.0
120486              50       42.0         73       22.0         39
4.0
120549              53       66.0         70       38.0         45
4.0
120612              47       48.0         70       34.0         45
15.0

        l_#bpFaced
119588         6.0
119666        18.0
120486         8.0
120549        11.0
120612        21.0

#checking shape
df_clay_BO5.shape
#16996 + 3077 = 20073 --> checks out

(3077, 46)
```

## Grass Courts

- "df_grass"
- “df_grass_BO3”
- “df_grass_BO5”

```
#creating main df for grass courts
df_grass = df_matchstats[df_matchstats['surface']=='Grass']

#checking shape
df_grass.shape
#correct number of matches

(6836, 46)
```

```
#creating 'df_grass_BO3'
df_grass_BO3 = df_BO3[df_BO3['surface']=='Grass']

#checking head
df_grass_BO3.head()
```

```
       tourney_id  Year  tourney_name surface tourney_level winner_id
\
119410   2000-311  2000  Queen's Club   Grass             A    102179

119411   2000-311  2000  Queen's Club   Grass             A    101150

119412   2000-311  2000  Queen's Club   Grass             A    101086

119413   2000-311  2000  Queen's Club   Grass             A    102257

119414   2000-311  2000  Queen's Club   Grass             A    102533


       winner_ioc       winner_name  winner_age  winner_rank
winner_ht  \
119410        FRA     Antony Dupuis        27.2        105.0
185.0
119411        ITA    Gianluca Pozzi        34.9         76.0
180.0
119412        USA     Ronald Agenor        35.5         97.0
180.0
119413        GBR     Greg Rusedski        26.7         20.0
193.0
119414        GER  Jens Knippschild        25.3        114.0
190.0

       loser_id loser_ioc         loser_name  loser_rank  loser_ht
loser_age  \
119410   101320       SWE  Magnus Gustafsson        77.0     185.0
33.4
119411   101733       NED      Jan Siemerink       125.0     183.0
30.1
119412   101820       SUI        Marc Rosset        34.0     201.0
29.5
119413   101965       RSA     Wayne Ferreira        42.0     185.0
28.7
119414   102755       USA          Alex Witt       498.0       NaN
24.2

       best_of round  minutes  w_#ServeGames  w_#aces  w_#dfs
w_#ServePoints  \
119410       3   R64    113.0           15.0      5.0     0.0
102.0
119411       3   R64    112.0           13.0      4.0     5.0
```

```
92.0
119412       3   R64     42.0           6.0      2.0     0.0
31.0
119413       3   R64     67.0          11.0     14.0     5.0
61.0
119414       3   R64     96.0          14.0      7.0     3.0
70.0

        w_#1stServesIn  w_#2ndServePoints  w_%1stServesIn
w_#1stWon  \
119410            63.0               39.0              61       45.0

119411            47.0               45.0              51       35.0

119412            22.0                9.0              70       20.0

119413            34.0               27.0              55       28.0

119414            34.0               36.0              48       29.0


        w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved  w_#bpFaced
l_#ServeGames  \
119410         71       22.0         56       10.0        12.0
14.0
119411         74       25.0         55        9.0        11.0
13.0
119412         90        4.0         44        0.0         0.0
5.0
119413         82       15.0         55        3.0         4.0
11.0
119414         85       22.0         61        2.0         4.0
14.0

        l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
l_#2ndServePoints  \
119410      6.0     7.0            78.0            44.0
34.0
119411      6.0     7.0            97.0            54.0
43.0
119412      3.0     5.0            36.0            16.0
20.0
119413      7.0     7.0            68.0            33.0
35.0
119414      8.0     4.0            97.0            50.0
47.0

        l_%1stServesIn  l_#1stWon  l_%1stWon  l_#2ndWon  l_%2ndWon
l_bpSaved  \
119410              56       34.0         77       17.0         50
```

```
4.0
119411              55       33.0         61       20.0         46
11.0
119412              44       12.0         75        7.0         35
2.0
119413              48       22.0         66       19.0         54
5.0
119414              51       32.0         64       26.0         55
5.0

        l_#bpFaced
119410         7.0
119411        16.0
119412         4.0
119413         8.0
119414         8.0
```

```
#checking shape
df_grass_BO3.shape
```

```
(4044, 46)
```

```
#creating 'df_grass_BO5'
df_grass_BO5 = df_BO5[df_BO5['surface']=='Grass']
```

```
#checking head
df_grass_BO5.head()
```

```
       tourney_id  Year tourney_name surface tourney_level
winner_id  \
121574   2000-540  2000    Wimbledon   Grass             G    101948

121575   2000-540  2000    Wimbledon   Grass             G    102344

121576   2000-540  2000    Wimbledon   Grass             G    103566

121577   2000-540  2000    Wimbledon   Grass             G    102925

121578   2000-540  2000    Wimbledon   Grass             G    103252


       winner_ioc       winner_name  winner_age  winner_rank
winner_ht  \
121574        USA      Pete Sampras        28.8          3.0
185.0
121575        SVK      Karol Kucera        26.3         44.0
188.0
121576        FRA    Michael Llodra        20.1        158.0
190.0
121577        USA  Justin Gimelstob        23.4         99.0
196.0
```

```
121578        ESP    Alberto Martin        21.8         67.0
175.0

       loser_id loser_ioc       loser_name  loser_rank  loser_ht
loser_age  \
121574   103181       CZE       Jiri Vanek        80.0     185.0
22.1
121575   102286       ZIM      Wayne Black       166.0     170.0
26.6
121576   102223       MAR      Karim Alami        29.0     185.0
27.0
121577   102443       GBR      Barry Cowan       172.0     188.0
25.8
121578   102381       AUT  Werner Eschauer       176.0     188.0
26.1

       best_of round  minutes  w_#ServeGames  w_#aces  w_#dfs
w_#ServePoints  \
121574       5  R128     83.0           14.0     10.0     4.0
73.0
121575       5  R128     88.0           13.0     13.0     9.0
68.0
121576       5  R128     78.0           13.0      4.0     4.0
72.0
121577       5  R128    159.0           21.0     28.0     7.0
131.0
121578       5  R128    128.0           18.0      3.0     1.0
91.0

        w_#1stServesIn  w_#2ndServePoints  w_%1stServesIn
w_#1stWon  \
121574            48.0               25.0              65       43.0

121575            36.0               32.0              52       33.0

121576            44.0               28.0              61       33.0

121577            81.0               50.0              61       64.0

121578            74.0               17.0              81       57.0


        w_%1stWon  w_#2ndWon  w_%2ndWon  w_bpSaved  w_#bpFaced
l_#ServeGames  \
121574         89       15.0         60        2.0         2.0
14.0
121575         91       18.0         56        2.0         3.0
13.0
121576         75       17.0         60        2.0         4.0
12.0
```

```
121577         79       27.0         54        4.0         6.0
19.0
121578         77       11.0         64        5.0         7.0
19.0

        l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn
l_#2ndServePoints  \
121574      3.0     3.0            94.0            59.0
35.0
121575      3.0     5.0            87.0            43.0
44.0
121576      3.0     7.0            80.0            50.0
30.0
121577     10.0     7.0           141.0            94.0
47.0
121578      2.0     4.0           124.0            91.0
33.0

        l_%1stServesIn  l_#1stWon  l_%1stWon  l_#2ndWon  l_%2ndWon
l_bpSaved  \
121574              62       40.0         67       16.0         45
2.0
121575              49       22.0         51       25.0         56
6.0
121576              62       30.0         60        9.0         30
8.0
121577              66       67.0         71       17.0         36
12.0
121578              73       55.0         60       18.0         54
8.0

        l_#bpFaced
121574         6.0
121575        12.0
121576        15.0
121577        17.0
121578        13.0

#checking shape
df_grass_BO5.shape
#4,044 + 2,792 = 6,836 --> Checks Out

(2792, 46)
```

# 3. Exporting Subsets

List of New Main Dataframe and Subsets

- Main Dataframe: "df_matchstats"
- By Number of Sets: “df_BO3” and “df_BO5”

- Hard Courts (3 total): "df_hard", “df_hard_BO3” and “df_hard_BO5”
- Clay Courts (3 total): "df_clay", “df_clay_BO3” and “df_clay_BO5”
- Grass Courts (3 total): "df_grass", “df_grass_BO3” and “df_grass_BO5”

```
#df_matchstats (main dataframe)
df_matchstats.to_pickle(os.path.join(path, 'Prepared Data',
'df_matchstats.pkl'))

#save df_matchstats as CSV
df_matchstats.to_csv(os.path.join(path,'Prepared
Data','df_matchstats.csv'))

#df_BO3
df_BO3.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_BO3.pkl'))

#df_BO5
df_BO5.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_BO5.pkl'))

#df_hard
df_hard.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_hard.pkl'))

#df_hard_BO3
df_hard_BO3.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_hard_BO3.pkl'))

#df_hard_BO5
df_hard_BO5.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_hard_BO5.pkl'))

#df_clay
df_clay.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_clay.pkl'))

#df_clay_BO3
df_clay_BO3.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_clay_BO3.pkl'))

#df_clay_BO5
df_clay_BO5.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_clay_BO5.pkl'))

#df_grass
df_grass.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_grass.pkl'))

#df_grass_BO5
df_grass_BO5.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_grass_BO5.pkl'))
```

```
#df_grass_BO3
df_grass_BO3.to_pickle(os.path.join(path, 'Prepared Data',
'Subsets','df_grass_BO3.pkl'))
```