1. ATP Initial Exploration

This is the first of two note books conducting an initial exploration of the ATP Dataset. This notebook focuses on exploring the raw data and steps to cleaning/wrangling. The second notebook will take the prepared data and create subsets for further analysis.

Table of Contents

- There were no match statistics before 1991

3D. Changing Data Types for Certain Variables

3E. Final Cleaning: Converting Derived Percentage Variables to Integers

4. Export PKLs

1. Import

```
#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 Datasets
df_matches = pd.read_csv(os.path.join(path, 'Original Data',
'matches.csv'), index_col = False)
df_players = pd.read_csv(os.path.join(path, 'Original Data',
'players.csv'), index_col = False)
df_rankings = pd.read_csv(os.path.join(path, 'Original Data',
'rankings.csv'), index_col = False)

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

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

2. Initial Exploration

2A. Matches

```
df_matches.head()
```

```
  tourney_id tourney_name surface  draw_size tourney_level
tourney_date  \
0  1968-2029       Dublin   Grass         32             A
19680708
1  1968-2029       Dublin   Grass         32             A
```

```
19680708
2  1968-2029       Dublin   Grass         32             A
19680708
3  1968-2029       Dublin   Grass         32             A
19680708
4  1968-2029       Dublin   Grass         32             A
19680708

   match_num  winner_id  winner_seed winner_entry      winner_name  \
0        270     112411          NaN          NaN    Douglas Smith
1        271     126914          NaN          NaN  Louis Pretorius
2        272     209523          NaN          NaN     Cecil Pedlow
3        273     100084          NaN          NaN        Tom Okker
4        274     100132          NaN          NaN  Armistead Neely

  winner_hand  winner_ht winner_ioc  winner_age  loser_id  loser_seed
\
0           U        NaN        AUS         NaN    110196         NaN

1           R        NaN        RSA         NaN    209536         NaN

2           U        NaN        IRL         NaN    209535         NaN

3           R      178.0        NED        24.3    209534         NaN

4           R        NaN        USA        21.3    209533         NaN


  loser_entry       loser_name loser_hand  loser_ht loser_ioc
loser_age  \
0         NaN  Peter Ledbetter          U       NaN       UNK
24.0
1         NaN  Maurice Pollock          U       NaN       IRL
NaN
2         NaN      John Mulvey          U       NaN       IRL
NaN
3         NaN  Unknown Fearmon          U       NaN       NaN
NaN
4         NaN   Harry Sheridan          U       NaN       IRL
NaN

     score  best_of round  minutes  w_ace  w_df  w_svpt  w_1stIn
w_1stWon  \
0  6-1 7-5        3   R32      NaN    NaN   NaN     NaN      NaN
NaN
1  6-1 6-1        3   R32      NaN    NaN   NaN     NaN      NaN
NaN
2  6-2 6-2        3   R32      NaN    NaN   NaN     NaN      NaN
NaN
3  6-1 6-1        3   R32      NaN    NaN   NaN     NaN      NaN
```

```
NaN
4  6-2 6-4         3   R32      NaN    NaN   NaN     NaN      NaN
NaN

   w_2ndWon  w_SvGms  w_bpSaved  w_bpFaced  l_ace  l_df  l_svpt
l_1stIn  \
0       NaN      NaN        NaN        NaN    NaN   NaN     NaN
NaN
1       NaN      NaN        NaN        NaN    NaN   NaN     NaN
NaN
2       NaN      NaN        NaN        NaN    NaN   NaN     NaN
NaN
3       NaN      NaN        NaN        NaN    NaN   NaN     NaN
NaN
4       NaN      NaN        NaN        NaN    NaN   NaN     NaN
NaN

   l_1stWon  l_2ndWon  l_SvGms  l_bpSaved  l_bpFaced  winner_rank  \
0       NaN       NaN      NaN        NaN        NaN          NaN
1       NaN       NaN      NaN        NaN        NaN          NaN
2       NaN       NaN      NaN        NaN        NaN          NaN
3       NaN       NaN      NaN        NaN        NaN          NaN
4       NaN       NaN      NaN        NaN        NaN          NaN

   winner_rank_points  loser_rank  loser_rank_points
0                 NaN         NaN                NaN
1                 NaN         NaN                NaN
2                 NaN         NaN                NaN
3                 NaN         NaN                NaN
4                 NaN         NaN                NaN

#Shape
df_matches.shape

(188161, 49)

#Columns

df_matches.columns

#I will need to rename some of these variables

Index(['tourney_id', 'tourney_name', 'surface', 'draw_size',
'tourney_level',
       'tourney_date', 'match_num', 'winner_id', 'winner_seed',
'winner_entry',
       'winner_name', 'winner_hand', 'winner_ht', 'winner_ioc',
'winner_age',
       'loser_id', 'loser_seed', 'loser_entry', 'loser_name',
'loser_hand',
       'loser_ht', 'loser_ioc', 'loser_age', 'score', 'best_of',
```

```
'round',
       'minutes', 'w_ace', 'w_df', 'w_svpt', 'w_1stIn', 'w_1stWon',
'w_2ndWon',
       'w_SvGms', 'w_bpSaved', 'w_bpFaced', 'l_ace', 'l_df', 'l_svpt',
       'l_1stIn', 'l_1stWon', 'l_2ndWon', 'l_SvGms', 'l_bpSaved',
'l_bpFaced',
       'winner_rank', 'winner_rank_points', 'loser_rank',
'loser_rank_points'],
      dtype='object')

#Checking for missing values
df_matches.isnull().sum()

#Seed; there is no seeding for team tournaments, such as Davis Cup;
also, not all players are seeded
#the fact that more
#For match statistics: similar amount of missing data for these

tourney_id                 0
tourney_name               0
surface                 2317
draw_size                  0
tourney_level              0
tourney_date               0
match_num                  0
winner_id                  0
winner_seed           118467
winner_entry          171891
winner_name                0
winner_hand               17
winner_ht              16237
winner_ioc                10
winner_age              1335
loser_id                   0
loser_seed            152824
loser_entry           160432
loser_name                 0
loser_hand                64
loser_ht               28698
loser_ioc                 69
loser_age               4825
score                      8
best_of                    0
round                      0
minutes                98650
w_ace                  95941
w_df                   95942
w_svpt                 95942
w_1stIn                95942
w_1stWon               95942
```

```
w_2ndWon                95942
w_SvGms                 95941
w_bpSaved               95942
w_bpFaced               95942
l_ace                   95942
l_df                    95941
l_svpt                  95942
l_1stIn                 95942
l_1stWon                95942
l_2ndWon                95942
l_SvGms                 95941
l_bpSaved               95942
l_bpFaced               95942
winner_rank             34964
winner_rank_points      82188
loser_rank              43327
loser_rank_points       83807
dtype: int64
```

```
#duplicates check

df_matches_dups = df_matches[df_matches.duplicated()]
df_matches_dups.shape

#no duplicates
```

```
(0, 49)
```

```
df_matches.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 188161 entries, 0 to 188160
Data columns (total 49 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   tourney_id          188161 non-null  object 
 1   tourney_name        188161 non-null  object 
 2   surface             185844 non-null  object 
 3   draw_size           188161 non-null  int64  
 4   tourney_level       188161 non-null  object 
 5   tourney_date        188161 non-null  int64  
 6   match_num           188161 non-null  int64  
 7   winner_id           188161 non-null  int64  
 8   winner_seed         69694 non-null   float64
 9   winner_entry        16270 non-null   object 
 10  winner_name         188161 non-null  object 
 11  winner_hand         188144 non-null  object 
 12  winner_ht           171924 non-null  float64
 13  winner_ioc          188151 non-null  object 
 14  winner_age          186826 non-null  float64
```

```
 15  loser_id            188161 non-null  int64
 16  loser_seed          35337 non-null   float64
 17  loser_entry         27729 non-null   object
 18  loser_name          188161 non-null  object
 19  loser_hand          188097 non-null  object
 20  loser_ht            159463 non-null  float64
 21  loser_ioc           188092 non-null  object
 22  loser_age           183336 non-null  float64
 23  score               188153 non-null  object
 24  best_of             188161 non-null  int64
 25  round               188161 non-null  object
 26  minutes             89511 non-null   float64
 27  w_ace               92220 non-null   float64
 28  w_df                92219 non-null   float64
 29  w_svpt              92219 non-null   float64
 30  w_1stIn             92219 non-null   float64
 31  w_1stWon            92219 non-null   float64
 32  w_2ndWon            92219 non-null   float64
 33  w_SvGms             92220 non-null   float64
 34  w_bpSaved           92219 non-null   float64
 35  w_bpFaced           92219 non-null   float64
 36  l_ace               92219 non-null   float64
 37  l_df                92220 non-null   float64
 38  l_svpt              92219 non-null   float64
 39  l_1stIn             92219 non-null   float64
 40  l_1stWon            92219 non-null   float64
 41  l_2ndWon            92219 non-null   float64
 42  l_SvGms             92220 non-null   float64
 43  l_bpSaved           92219 non-null   float64
 44  l_bpFaced           92219 non-null   float64
 45  winner_rank         153197 non-null  float64
 46  winner_rank_points  105973 non-null  float64
 47  loser_rank          144834 non-null  float64
 48  loser_rank_points   104354 non-null  float64
dtypes: float64(29), int64(6), object(14)
memory usage: 70.3+ MB
```

```
df_matches.describe()

#the following columns should be turned into strings:
'tourney_date','match_num','winner_id','loser_id','best_of
```

```
           draw_size  tourney_date      match_num      winner_id  \
count  188161.000000  1.881610e+05  188161.000000  188161.000000
mean       52.926292  1.993350e+07      76.618598  103820.251673
std        36.446303  1.544445e+05     110.714957   11470.048991
min         2.000000  1.968011e+07       1.000000  100001.000000
25%        32.000000  1.980041e+07      10.000000  100402.000000
50%        32.000000  1.993030e+07      25.000000  101686.000000
75%        64.000000  2.006072e+07      80.000000  103898.000000
```

```
max       128.000000  2.022113e+07    1701.000000  211468.000000

        winner_seed      winner_ht     winner_age       loser_id  \
count  69694.000000  171924.000000  186826.000000  188161.000000
mean       6.280225     184.449187      25.658362  104967.986995
std        5.509548       6.667033       4.045128   14866.251405
min        1.000000     160.000000      14.300000  100001.000000
25%        2.000000     180.000000      22.700000  100502.000000
50%        5.000000     185.000000      25.200000  101843.000000
75%        8.000000     188.000000      28.200000  104252.000000
max       35.000000     211.000000      58.700000  211805.000000

         loser_seed       loser_ht      loser_age        best_of  \
count  35337.000000  159463.000000  183336.000000  188161.000000
mean       7.667402     184.226592      25.709391       3.441218
std        5.995551       6.655036       4.190362       0.830009
min        1.000000     160.000000      14.400000       1.000000
25%        4.000000     180.000000      22.700000       3.000000
50%        6.000000     185.000000      25.300000       3.000000
75%       10.000000     188.000000      28.300000       3.000000
max       35.000000     211.000000      63.600000       5.000000

            minutes         w_ace          w_df        w_svpt
w_1stIn  \
count  89511.000000  92220.000000  92219.000000  92219.000000
92219.000000
mean     103.497403      6.517708      2.736258     78.068142
47.581724
std       39.365772      5.341289      2.367377     29.523162
19.216689
min        0.000000      0.000000      0.000000      0.000000
0.000000
25%       75.000000      3.000000      1.000000     56.000000
34.000000
50%       96.000000      5.000000      2.000000     73.000000
44.000000
75%      125.000000      9.000000      4.000000     94.000000
58.000000
max     1146.000000    113.000000     26.000000    491.000000
361.000000

          w_1stWon      w_2ndWon       w_SvGms     w_bpSaved
w_bpFaced  \
count  92219.000000  92219.000000  92220.000000  92219.000000
92219.000000
mean      35.873746     16.733883     12.396758      3.527549
5.167406
std       13.836610      6.989782      4.120292      3.086390
4.063021
min        0.000000      0.000000      0.000000      0.000000
```

```
0.000000
25%       26.000000     12.000000      9.000000      1.000000
2.000000
50%       33.000000     16.000000     11.000000      3.000000
4.000000
75%       43.000000     21.000000     15.000000      5.000000
7.000000
max      292.000000     82.000000     90.000000     24.000000
34.000000

              l_ace          l_df        l_svpt       l_1stIn
l_1stWon  \
count  92219.000000  92220.000000  92219.000000  92219.000000
92219.000000
mean       4.828745      3.488224     80.908284     48.011549
31.894892
std        4.670710      2.618833     29.458713     19.390121
14.449465
min        0.000000      0.000000      0.000000      0.000000
0.000000
25%        2.000000      2.000000     59.000000     34.000000
22.000000
50%        4.000000      3.000000     76.000000     45.000000
29.000000
75%        7.000000      5.000000     97.000000     58.000000
40.000000
max      103.000000     26.000000    489.000000    328.000000
284.000000

          l_2ndWon       l_SvGms     l_bpSaved     l_bpFaced
winner_rank  \
count  92219.000000  92220.000000  92219.000000  92219.000000
153197.000000
mean      14.985263     12.197387      4.812002      8.742884
75.255716
std        7.220377      4.129834      3.275387      4.131839
121.053512
min        0.000000      0.000000     -6.000000      0.000000
1.000000
25%       10.000000      9.000000      2.000000      6.000000
17.000000
50%       14.000000     11.000000      4.000000      8.000000
44.000000
75%       19.000000     15.000000      7.000000     11.000000
86.000000
max      101.000000     91.000000     28.000000     38.000000
2101.000000

       winner_rank_points     loser_rank  loser_rank_points
count       105973.000000  144834.000000      104354.000000
```

```
mean          1366.471611     112.884150          859.219896
std           1726.089469     162.191701          987.192154
min              0.000000       1.000000            0.000000
25%            489.000000      37.000000          361.000000
50%            846.000000      70.000000          630.000000
75%           1532.000000     118.000000         1013.000000
max          16950.000000    2159.000000        16950.000000
```

Initial Findings:

- Lots of missing statistics in many matches: figure out why data is missing here. Were these tournaments low profile? Too old?

2B. Players

```
df_players.shape
```

```
(58687, 8)
```

```
df_players.head()
```

```
   player_id name_first name_last hand         dob  ioc  height
wikidata_id
0     100001    Gardnar    Mulloy    R  19131122.0  USA   185.0
Q54544
1     100002     Pancho    Segura    R  19210620.0  ECU   168.0
Q54581
2     100003      Frank   Sedgman    R  19271002.0  AUS   180.0
Q962049
3     100004   Giuseppe     Merlo    R  19271011.0  ITA     NaN
Q1258752
4     100005    Richard  Gonzalez    R  19280509.0  USA   188.0
Q53554
```

```
df_players.columns
```

```
Index(['player_id', 'name_first', 'name_last', 'hand', 'dob', 'ioc',
'height',
       'wikidata_id'],
      dtype='object')
```

```
#duplicates check

df_players_dups = df_players[df_players.duplicated()]
df_players_dups.shape

#no duplicates
```

```
(0, 8)
```

```
df_players.isnull().sum()
```

```
player_id          0
name_first       355
name_last         41
hand             240
dob            13547
ioc              101
height         55899
wikidata_id    53793
dtype: int64
```

2C. Rankings --> I will most likely not use this CSV

3. Wrangling and Cleaning Steps

Cleaning and Wrangling Steps:

A. Rename following columns

- w_ace --> w_#aces
- w_df --> w_#dfs
- w_svpt --> w_#servepoints
- w_1stin --> w_#1stservesin
- w_1stWon --> w_#1stWon
- w_2ndWon --> w_#2ndWon
- w_SvGms --> w_#SvGms
- w_ bpSaved --> w_#bpSaved
- w_bpFaced --> w_#bpFaced
- l_ace --> l_#aces
- l_df --> l_#dfs
- l_svpt --> l_#servepoints
- l_1stin --> l_#1stservesin
- l_1stWon --> l_#1stWon
- l_2ndWon --> l_#2ndWon
- l_SvGms --> l_#SvGms
- l_bpSaved --> l_#bpSaved
- l_bpFaced --> l_#bpFaced

B. Create/derive following variables

Serve Statistics

- "w_#2ndservepoints" = "w_#servepoints" - "w_#1stservesin"
- "w_%1stservesin" = "w_#1stservesin"/"w_#servepoints"
- "w_%1stWon" = "w_#1stWon"/"w_#1stservesin"
- "w_%2ndWon" = "w_#2ndWon"/"w_#2ndservepoints"
- "l_#2ndservepoints" = "l_#servepoints" - "l_#1stservesin"

- "l_%1stservesin" = "l_#1stservesin"/"l_#servepoints"
- "l_%1stWon" = "l_#1stWon"/"l_#1stservesin"
- "l_%2ndWon" = "l_#2ndWon"/"l_#2ndservepoints"

Other

- Year (first four digits of "tourney_date"

C. New Main Dataframe

- New Main DF: Remove entries prior to 1991 AND with missing match statistics: "df_matchstats"

D. Changing Data Types

- Change the following variables' data types from integers to strings:

3A. Renaming Columns

```
#Renaming Columns

df_matches.rename(columns =
                  {'w_ace' : 'w_#aces',
                   'w_df' : 'w_#dfs',
                   'w_svpt' : 'w_#ServePoints',
                   'w_1stIn' : 'w_#1stServesIn',
                   'w_1stWon' : 'w_#1stWon',
                   'w_2ndWon' : 'w_#2ndWon',
                   'w_SvGms' : 'w_#ServeGames',
                   'w_ bpSaved' : 'w_#bpSaved',
                   'w_bpFaced' : 'w_#bpFaced',
                    'l_ace' : 'l_#aces',
                   'l_df' : 'l_#dfs',
                   'l_svpt' : 'l_#ServePoints',
                   'l_1stIn' : 'l_#1stServesIn',
                   'l_1stWon' : 'l_#1stWon',
                   'l_2ndWon' : 'l_#2ndWon',
                   'l_SvGms' : 'l_#ServeGames',
                   'l_ bpSaved' : 'l_#bpSaved',
                   'l_bpFaced' : 'l_#bpFaced'}, inplace = True)

df_matches.head()
```

```
  tourney_id tourney_name surface  draw_size tourney_level
tourney_date  \
0  1968-2029       Dublin   Grass         32             A
19680708
1  1968-2029       Dublin   Grass         32             A
19680708
2  1968-2029       Dublin   Grass         32             A
19680708
3  1968-2029       Dublin   Grass         32             A
19680708
```

```
4  1968-2029       Dublin   Grass         32             A
19680708

   match_num  winner_id  winner_seed winner_entry      winner_name  \
0        270     112411          NaN          NaN    Douglas Smith
1        271     126914          NaN          NaN  Louis Pretorius
2        272     209523          NaN          NaN     Cecil Pedlow
3        273     100084          NaN          NaN        Tom Okker
4        274     100132          NaN          NaN  Armistead Neely

  winner_hand  winner_ht winner_ioc  winner_age  loser_id  loser_seed
\
0           U        NaN        AUS         NaN    110196         NaN

1           R        NaN        RSA         NaN    209536         NaN

2           U        NaN        IRL         NaN    209535         NaN

3           R      178.0        NED        24.3    209534         NaN

4           R        NaN        USA        21.3    209533         NaN


  loser_entry       loser_name loser_hand  loser_ht loser_ioc
loser_age  \
0         NaN  Peter Ledbetter          U       NaN       UNK
24.0
1         NaN  Maurice Pollock          U       NaN       IRL
NaN
2         NaN      John Mulvey          U       NaN       IRL
NaN
3         NaN  Unknown Fearmon          U       NaN       NaN
NaN
4         NaN   Harry Sheridan          U       NaN       IRL
NaN

     score  best_of round  minutes  w_#aces  w_#dfs  w_#ServePoints  \
0  6-1 7-5        3   R32      NaN      NaN     NaN             NaN
1  6-1 6-1        3   R32      NaN      NaN     NaN             NaN
2  6-2 6-2        3   R32      NaN      NaN     NaN             NaN
3  6-1 6-1        3   R32      NaN      NaN     NaN             NaN
4  6-2 6-4        3   R32      NaN      NaN     NaN             NaN

   w_#1stServesIn  w_#1stWon  w_#2ndWon  w_#ServeGames  w_bpSaved
w_#bpFaced  \
0             NaN        NaN        NaN            NaN        NaN
NaN
1             NaN        NaN        NaN            NaN        NaN
NaN
2             NaN        NaN        NaN            NaN        NaN
```

```
NaN
3              NaN         NaN         NaN              NaN         NaN
NaN
4              NaN         NaN         NaN              NaN         NaN
NaN

   l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn  l_#1stWon
l_#2ndWon  \
0      NaN     NaN              NaN              NaN         NaN
NaN
1      NaN     NaN              NaN              NaN         NaN
NaN
2      NaN     NaN              NaN              NaN         NaN
NaN
3      NaN     NaN              NaN              NaN         NaN
NaN
4      NaN     NaN              NaN              NaN         NaN
NaN

   l_#ServeGames  l_bpSaved  l_#bpFaced  winner_rank
winner_rank_points  \
0             NaN         NaN          NaN           NaN
NaN
1             NaN         NaN          NaN           NaN
NaN
2             NaN         NaN          NaN           NaN
NaN
3             NaN         NaN          NaN           NaN
NaN
4             NaN         NaN          NaN           NaN
NaN

   loser_rank  loser_rank_points
0         NaN                NaN
1         NaN                NaN
2         NaN                NaN
3         NaN                NaN
4         NaN                NaN
```

3B. Deriving New Variables

```
#winner's 2nd serve points played

df_matches['w_#2ndServePoints'] = df_matches['w_#ServePoints'] -
df_matches['w_#1stServesIn']

#winner's First Serve Percentage

df_matches['w_%1stServesIn'] =
df_matches['w_#1stServesIn']/df_matches['w_#ServePoints']
```

```
#winner's Percentage of First Serve Points Won

df_matches['w_%1stWon'] =
df_matches['w_#1stWon']/df_matches['w_#1stServesIn']

#winner's Percentage of Second Serve Points Won

df_matches['w_%2ndWon'] =
df_matches['w_#2ndWon']/df_matches['w_#2ndServePoints']

#loser's 2nd serve points played

df_matches['l_#2ndServePoints'] = df_matches['l_#ServePoints'] -
df_matches['l_#1stServesIn']

#loser's First Serve Percentage

df_matches['l_%1stServesIn'] =
df_matches['l_#1stServesIn']/df_matches['l_#ServePoints']

#loser's Percentage of First Serve Points Won

df_matches['l_%1stWon'] =
df_matches['l_#1stWon']/df_matches['l_#1stServesIn']

#loser's Percentage of Second Serve Points Won

df_matches['l_%2ndWon'] =
df_matches['l_#2ndWon']/df_matches['l_#2ndServePoints']
```

Derive "Year"

```
df_matches['Year'] = df_matches['tourney_date']

#convert 'year' to string

df_matches['Year'] = df_matches['Year'].astype('str')

df_matches['Year'] = df_matches['Year'].str[:4]

df_matches.head()
```

```
  tourney_id tourney_name surface  draw_size tourney_level
tourney_date  \
0  1968-2029       Dublin   Grass         32             A
19680708
1  1968-2029       Dublin   Grass         32             A
19680708
2  1968-2029       Dublin   Grass         32             A
19680708
3  1968-2029       Dublin   Grass         32             A
19680708
```

```
4  1968-2029       Dublin   Grass         32             A
19680708

   match_num  winner_id  winner_seed winner_entry      winner_name  \
0        270     112411          NaN          NaN    Douglas Smith
1        271     126914          NaN          NaN  Louis Pretorius
2        272     209523          NaN          NaN     Cecil Pedlow
3        273     100084          NaN          NaN        Tom Okker
4        274     100132          NaN          NaN  Armistead Neely

  winner_hand  winner_ht winner_ioc  winner_age  loser_id  loser_seed
\
0           U        NaN        AUS         NaN    110196         NaN

1           R        NaN        RSA         NaN    209536         NaN

2           U        NaN        IRL         NaN    209535         NaN

3           R      178.0        NED        24.3    209534         NaN

4           R        NaN        USA        21.3    209533         NaN


  loser_entry       loser_name loser_hand  loser_ht loser_ioc
loser_age  \
0         NaN  Peter Ledbetter          U       NaN       UNK
24.0
1         NaN  Maurice Pollock          U       NaN       IRL
NaN
2         NaN      John Mulvey          U       NaN       IRL
NaN
3         NaN  Unknown Fearmon          U       NaN       NaN
NaN
4         NaN   Harry Sheridan          U       NaN       IRL
NaN

     score  best_of round  minutes  w_#aces  w_#dfs  w_#ServePoints  \
0  6-1 7-5        3   R32      NaN      NaN     NaN             NaN
1  6-1 6-1        3   R32      NaN      NaN     NaN             NaN
2  6-2 6-2        3   R32      NaN      NaN     NaN             NaN
3  6-1 6-1        3   R32      NaN      NaN     NaN             NaN
4  6-2 6-4        3   R32      NaN      NaN     NaN             NaN

   w_#1stServesIn  w_#1stWon  w_#2ndWon  w_#ServeGames  w_bpSaved
w_#bpFaced  \
0             NaN        NaN        NaN            NaN        NaN
NaN
1             NaN        NaN        NaN            NaN        NaN
NaN
2             NaN        NaN        NaN            NaN        NaN
```

```
NaN
3             NaN         NaN         NaN             NaN         NaN
NaN
4             NaN         NaN         NaN             NaN         NaN
NaN

   l_#aces  l_#dfs  l_#ServePoints  l_#1stServesIn  l_#1stWon
l_#2ndWon  \
0      NaN     NaN             NaN             NaN         NaN
NaN
1      NaN     NaN             NaN             NaN         NaN
NaN
2      NaN     NaN             NaN             NaN         NaN
NaN
3      NaN     NaN             NaN             NaN         NaN
NaN
4      NaN     NaN             NaN             NaN         NaN
NaN

   l_#ServeGames  l_bpSaved  l_#bpFaced  winner_rank
winner_rank_points  \
0            NaN         NaN         NaN           NaN
NaN
1            NaN         NaN         NaN           NaN
NaN
2            NaN         NaN         NaN           NaN
NaN
3            NaN         NaN         NaN           NaN
NaN
4            NaN         NaN         NaN           NaN
NaN

   loser_rank  loser_rank_points  w_#2ndServePoints  w_%1stServesIn  \
0         NaN                NaN                NaN             NaN
1         NaN                NaN                NaN             NaN
2         NaN                NaN                NaN             NaN
3         NaN                NaN                NaN             NaN
4         NaN                NaN                NaN             NaN

   w_%1stWon  w_%2ndWon  l_#2ndServePoints  l_%1stServesIn  l_%1stWon
\
0        NaN        NaN                NaN             NaN        NaN

1        NaN        NaN                NaN             NaN        NaN

2        NaN        NaN                NaN             NaN        NaN

3        NaN        NaN                NaN             NaN        NaN

4        NaN        NaN                NaN             NaN        NaN
```

```
   l_%2ndWon  Year
0        NaN  1968
1        NaN  1968
2        NaN  1968
3        NaN  1968
4        NaN  1968
```

3C. Creating New Main Dataframe

New Main Dataframe

New Main DF: Remove entries prior to 2000 AND with missing match statistics: "df_matchstats"

```
df_post2000 = df_matches[df_matches['Year']>= '2000']

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        0.510417
0.795918
119318              723.0               24.0        0.684211
0.750000
119319              649.0               20.0        0.636364
0.714286
119320              616.0               25.0        0.528302
0.928571
119321              873.0               32.0        0.673469
0.590909

        w_%2ndWon  l_#2ndServePoints  l_%1stServesIn  l_%1stWon  l_
%2ndWon  \
119317   0.595745               51.0        0.518868   0.709091
0.568627
119318   0.541667               42.0        0.432432   0.781250
0.428571
119319   0.600000               23.0        0.589286   0.606061
0.304348
119320   0.600000               27.0        0.614286   0.674419
0.518519
119321   0.437500               46.0        0.500000   0.739130
0.391304
```

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

```
df_post2000.shape
```

```
(68844, 58)
```

```
#Checking for missing values

df_post2000.isnull().sum()

# Unimportant missing values: winner_seed, winner_entry, winner_hand,
loser_seed, loser_entry, loser_hand, minutes
# Remove data where serve statistics are missing
# to deal with later: winner_ht, loser_ht
```

```
tourney_id                0
tourney_name              0
surface                   0
draw_size                 0
tourney_level             0
tourney_date              0
match_num                 0
winner_id                 0
winner_seed           40268
winner_entry          60288
winner_name               0
winner_hand              11
winner_ht              1499
winner_ioc                0
winner_age                5
loser_id                  0
loser_seed            53132
loser_entry           54835
loser_name                0
loser_hand               44
loser_ht               3039
loser_ioc                 0
loser_age                 3
score                     0
best_of                   0
round                     0
minutes                7706
w_#aces                6289
w_#dfs                 6289
```

```
w_#ServePoints          6289
w_#1stServesIn          6289
w_#1stWon               6289
w_#2ndWon               6289
w_#ServeGames           6288
w_bpSaved               6289
w_#bpFaced              6289
l_#aces                 6289
l_#dfs                  6289
l_#ServePoints          6289
l_#1stServesIn          6289
l_#1stWon               6289
l_#2ndWon               6289
l_#ServeGames           6288
l_bpSaved               6289
l_#bpFaced              6289
winner_rank              536
winner_rank_points       536
loser_rank              1398
loser_rank_points       1398
w_#2ndServePoints       6289
w_%1stServesIn          6294
w_%1stWon               6294
w_%2ndWon               6305
l_#2ndServePoints       6289
l_%1stServesIn          6293
l_%1stWon               6295
l_%2ndWon               6304
Year                       0
dtype: int64
```

I will primarily be looking at serve statistics and their impact on matches/match outcomes, therefore, I will remove entries in which serve stats are missing

```
#New Main Dataframe

df_post2000 = df_post2000[df_post2000['w_#1stServesIn'].notna()]

df_post2000.shape
```

```
(62555, 58)
```

```
df_post2000.isnull().sum()
#remove remaining missing serve statistics
```

```
tourney_id              0
tourney_name            0
surface                 0
draw_size               0
tourney_level           0
tourney_date            0
```

```
match_num                0
winner_id                0
winner_seed          34357
winner_entry         54062
winner_name              0
winner_hand              5
winner_ht              198
winner_ioc               0
winner_age               0
loser_id                 0
loser_seed           47113
loser_entry          48648
loser_name               0
loser_hand              24
loser_ht               730
loser_ioc                0
loser_age                2
score                    0
best_of                  0
round                    0
minutes               1463
w_#aces                  0
w_#dfs                   0
w_#ServePoints           0
w_#1stServesIn           0
w_#1stWon                0
w_#2ndWon                0
w_#ServeGames            0
w_bpSaved                0
w_#bpFaced               0
l_#aces                  0
l_#dfs                   0
l_#ServePoints           0
l_#1stServesIn           0
l_#1stWon                0
l_#2ndWon                0
l_#ServeGames            0
l_bpSaved                0
l_#bpFaced               0
winner_rank             89
winner_rank_points      89
loser_rank             296
loser_rank_points      296
w_#2ndServePoints        0
w_%1stServesIn           5
w_%1stWon                5
w_%2ndWon               16
l_#2ndServePoints        0
l_%1stServesIn           4
```

```
l_%1stWon                  6
l_%2ndWon                 15
Year                       0
dtype: int64
```

Remove remaining missing columns

- w_%1stWon 5
- w_%2ndWon 16
- l_#2ndServePoints 0
- l_%1stServesIn 4
- l_%1stWon 6
- l_%2ndWon 15

```
df_post2000 = df_post2000[df_post2000['w_%1stWon'].isnull()== False]
df_post2000 = df_post2000[df_post2000['w_%2ndWon'].isnull()== False]
df_post2000 = df_post2000[df_post2000['l_#2ndServePoints'].isnull()==
False]
df_post2000 = df_post2000[df_post2000['l_%1stServesIn'].isnull()==
False]
df_post2000 = df_post2000[df_post2000['l_%1stWon'].isnull()== False]
df_post2000 = df_post2000[df_post2000['l_%2ndWon'].isnull()== False]

df_post2000.isnull().sum()
#No more missing serve statistics
```

```
tourney_id                 0
tourney_name               0
surface                    0
draw_size                  0
tourney_level              0
tourney_date               0
match_num                  0
winner_id                  0
winner_seed            34337
winner_entry           54042
winner_name                0
winner_hand                5
winner_ht                198
winner_ioc                 0
winner_age                 0
loser_id                   0
loser_seed             47097
loser_entry            48629
loser_name                 0
loser_hand                24
loser_ht                 729
loser_ioc                  0
loser_age                  2
score                      0
```

```
best_of                   0
round                     0
minutes                1456
w_#aces                   0
w_#dfs                    0
w_#ServePoints            0
w_#1stServesIn            0
w_#1stWon                 0
w_#2ndWon                 0
w_#ServeGames             0
w_bpSaved                 0
w_#bpFaced                0
l_#aces                   0
l_#dfs                    0
l_#ServePoints            0
l_#1stServesIn            0
l_#1stWon                 0
l_#2ndWon                 0
l_#ServeGames             0
l_bpSaved                 0
l_#bpFaced                0
winner_rank              89
winner_rank_points       89
loser_rank              296
loser_rank_points       296
w_#2ndServePoints         0
w_%1stServesIn            0
w_%1stWon                 0
w_%2ndWon                 0
l_#2ndServePoints         0
l_%1stServesIn            0
l_%1stWon                 0
l_%2ndWon                 0
Year                      0
dtype: int64
```

3D. Changing Data Type

The following variables should be changed into object:

'winner_id', 'loser_id', 'best_of'

```
#Checking for Mixed Data Types
for col in df_post2000.columns.tolist():
  weird = (df_post2000[[col]].applymap(type) !=
df_post2000[[col]].iloc[0].apply(type)).any(axis = 1)
  if len (df_post2000[weird]) > 0:
    print (col)
```

```
#The following variables have mixed data types, but will likely not be
used for analysis. They should all be strings

winner_entry
winner_hand
loser_entry
loser_hand

#checking data type for each variable

df_post2000.info()

<class 'pandas.core.frame.DataFrame'>
Index: 62530 entries, 119317 to 188160
Data columns (total 58 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tourney_id          62530 non-null  object
 1   tourney_name        62530 non-null  object
 2   surface             62530 non-null  object
 3   draw_size           62530 non-null  int64
 4   tourney_level       62530 non-null  object
 5   tourney_date        62530 non-null  int64
 6   match_num           62530 non-null  int64
 7   winner_id           62530 non-null  int64
 8   winner_seed         28193 non-null  float64
 9   winner_entry        8488 non-null   object
 10  winner_name         62530 non-null  object
 11  winner_hand         62525 non-null  object
 12  winner_ht           62332 non-null  float64
 13  winner_ioc          62530 non-null  object
 14  winner_age          62530 non-null  float64
 15  loser_id            62530 non-null  int64
 16  loser_seed          15433 non-null  float64
 17  loser_entry         13901 non-null  object
 18  loser_name          62530 non-null  object
 19  loser_hand          62506 non-null  object
 20  loser_ht            61801 non-null  float64
 21  loser_ioc           62530 non-null  object
 22  loser_age           62528 non-null  float64
 23  score               62530 non-null  object
 24  best_of             62530 non-null  int64
 25  round               62530 non-null  object
 26  minutes             61074 non-null  float64
 27  w_#aces             62530 non-null  float64
 28  w_#dfs              62530 non-null  float64
 29  w_#ServePoints      62530 non-null  float64
 30  w_#1stServesIn      62530 non-null  float64
 31  w_#1stWon           62530 non-null  float64
 32  w_#2ndWon           62530 non-null  float64
```

```
 33  w_#ServeGames        62530 non-null  float64
 34  w_bpSaved            62530 non-null  float64
 35  w_#bpFaced           62530 non-null  float64
 36  l_#aces              62530 non-null  float64
 37  l_#dfs               62530 non-null  float64
 38  l_#ServePoints       62530 non-null  float64
 39  l_#1stServesIn       62530 non-null  float64
 40  l_#1stWon            62530 non-null  float64
 41  l_#2ndWon            62530 non-null  float64
 42  l_#ServeGames        62530 non-null  float64
 43  l_bpSaved            62530 non-null  float64
 44  l_#bpFaced           62530 non-null  float64
 45  winner_rank          62441 non-null  float64
 46  winner_rank_points   62441 non-null  float64
 47  loser_rank           62234 non-null  float64
 48  loser_rank_points    62234 non-null  float64
 49  w_#2ndServePoints    62530 non-null  float64
 50  w_%1stServesIn       62530 non-null  float64
 51  w_%1stWon            62530 non-null  float64
 52  w_%2ndWon            62530 non-null  float64
 53  l_#2ndServePoints    62530 non-null  float64
 54  l_%1stServesIn       62530 non-null  float64
 55  l_%1stWon            62530 non-null  float64
 56  l_%2ndWon            62530 non-null  float64
 57  Year                 62530 non-null  object
dtypes: float64(37), int64(6), object(15)
memory usage: 28.1+ MB
```

Change the following to objects: 'winner_id', 'loser_id', 'best_of'

```
#changing winner_id

df_post2000['winner_id'] = df_post2000['winner_id'].astype('object')

#changing loser_id

df_post2000['loser_id'] = df_post2000['loser_id'].astype('object')

#changing best_of

df_post2000['best_of'] = df_post2000['best_of'].astype('object')
```

Convert all percentages into integers

This will be useful later when making categorical plots

```
#Convert all Percentage Variables to Integer by multiplying by 100
(for categorical plot bins)
df_post2000['w_%1stServesIn'] = df_post2000['w_
%1stServesIn'].apply(lambda x: int(x * 100))
```

```
df_post2000['w_%1stWon'] = df_post2000['w_%1stWon'].apply(lambda x:
int(x * 100))
df_post2000['w_%2ndWon'] = df_post2000['w_%2ndWon'].apply(lambda x:
int(x * 100))
df_post2000['l_%1stServesIn'] = df_post2000['l_
%1stServesIn'].apply(lambda x: int(x * 100))
df_post2000['l_%1stWon'] = df_post2000['l_%1stWon'].apply(lambda x:
int(x * 100))
df_post2000['l_%2ndWon'] = df_post2000['l_%2ndWon'].apply(lambda x:
int(x * 100))

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
```

4. Export New Main Dataframe

```
df_post2000.to_pickle(os.path.join(path, 'Prepared Data',
'df_post2000.pkl'))
```

On Script 1b. ATP Initial Exploration Part

1. Importing Libraries and PKL File

2. Creating Subsets

- df_matchstats
- df_carpet
- df_hard
- df_grass
- df_clay
- df_big3_win
- df_big3_lose
- df_big3

```

```