teams_pre

March 9, 2020

[4]:
```
import math
import numpy as np
import pandas as pd
```

[5]:
```
df = pd.read_csv('../data/lahman/mlb_data/Teams.csv')
```

[6]:
```
df.columns
```

[6]:
```
Index(['yearID', 'lgID', 'teamID', 'franchID', 'divID', 'Rank', 'G', 'Ghome',
       'W', 'L', 'DivWin', 'WCWin', 'LgWin', 'WSWin', 'R', 'AB', 'H', '2B',
       '3B', 'HR', 'BB', 'SO', 'SB', 'CS', 'HBP', 'SF', 'RA', 'ER', 'ERA',
       'CG', 'SHO', 'SV', 'IPouts', 'HA', 'HRA', 'BBA', 'SOA', 'E', 'DP', 'FP',
       'name', 'park', 'attendance', 'BPF', 'PPF', 'teamIDBR',
       'teamIDlahman45', 'teamIDretro'],
      dtype='object')
```

[7]:
```
df.head()
```

[7]:

	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	...	DP	\
0	1919	AL	BOS	BOS	NaN	6	138	66	66	71	...	118	
1	1919	NL	BRO	LAD	NaN	5	141	70	69	71	...	84	
2	1919	NL	BSN	ATL	NaN	6	140	68	57	82	...	111	
3	1919	AL	CHA	CHW	NaN	1	140	70	88	52	...	116	
4	1919	NL	CHN	CHC	NaN	3	140	71	75	65	...	87	

	FP	name	park	attendance	BPF	PPF	teamIDBR	\
0	0.975	Boston Red Sox	Fenway Park I	417291	94	94	BOS	
1	0.963	Brooklyn Robins	Ebbets Field	360721	103	103	BRO	
2	0.966	Boston Braves	Braves Field	167401	95	98	BSN	
3	0.969	Chicago White Sox	Comiskey Park	627186	100	99	CHW	
4	0.969	Chicago Cubs	Wrigley Field	424430	100	99	CHC	

	teamIDlahman45	teamIDretro
0	BOS	BOS
1	BRO	BRO
2	BSN	BSN
3	CHA	CHA

```
4              CHN          CHN

[5 rows x 48 columns]
```

[8]:
```
df = df.drop(columns=['teamIDlahman45', 'teamIDBR'])
```

The first step is to ensure we're only using one ID per team. It would be best to just use Retrosheet's values, so our first step is to see where teamID differs from teamIDretro. Once we come up with a way to fix these differences, we'll want to write it as a script that we can use elsewhere - for example, in the batting table where we're using the regular teamID values.

[9]:
```
df[(df['teamID'] != df['teamIDretro'])][['yearID', 'teamID', 'teamIDretro',␣
↪'name']]
```

[9]:

	yearID	teamID	teamIDretro	name
551	1953	ML1	MLN	Milwaukee Braves
568	1954	ML1	MLN	Milwaukee Braves
585	1955	ML1	MLN	Milwaukee Braves
601	1956	ML1	MLN	Milwaukee Braves
617	1957	ML1	MLN	Milwaukee Braves
633	1958	ML1	MLN	Milwaukee Braves
649	1959	ML1	MLN	Milwaukee Braves
665	1960	ML1	MLN	Milwaukee Braves
683	1961	ML1	MLN	Milwaukee Braves
702	1962	ML1	MLN	Milwaukee Braves
722	1963	ML1	MLN	Milwaukee Braves
742	1964	ML1	MLN	Milwaukee Braves
762	1965	ML1	MLN	Milwaukee Braves
867	1970	ML4	MIL	Milwaukee Brewers
891	1971	ML4	MIL	Milwaukee Brewers
915	1972	ML4	MIL	Milwaukee Brewers
939	1973	ML4	MIL	Milwaukee Brewers
963	1974	ML4	MIL	Milwaukee Brewers
987	1975	ML4	MIL	Milwaukee Brewers
1011	1976	ML4	MIL	Milwaukee Brewers
1035	1977	ML4	MIL	Milwaukee Brewers
1061	1978	ML4	MIL	Milwaukee Brewers
1087	1979	ML4	MIL	Milwaukee Brewers
1113	1980	ML4	MIL	Milwaukee Brewers
1139	1981	ML4	MIL	Milwaukee Brewers
1165	1982	ML4	MIL	Milwaukee Brewers
1191	1983	ML4	MIL	Milwaukee Brewers
1217	1984	ML4	MIL	Milwaukee Brewers
1243	1985	ML4	MIL	Milwaukee Brewers
1269	1986	ML4	MIL	Milwaukee Brewers
1295	1987	ML4	MIL	Milwaukee Brewers
1321	1988	ML4	MIL	Milwaukee Brewers

```
1347    1989    ML4         MIL              Milwaukee Brewers
1373    1990    ML4         MIL              Milwaukee Brewers
1399    1991    ML4         MIL              Milwaukee Brewers
1425    1992    ML4         MIL              Milwaukee Brewers
1453    1993    ML4         MIL              Milwaukee Brewers
1481    1994    ML4         MIL              Milwaukee Brewers
1509    1995    ML4         MIL              Milwaukee Brewers
1537    1996    ML4         MIL              Milwaukee Brewers
1565    1997    ML4         MIL              Milwaukee Brewers
1801    2005    LAA         ANA  Los Angeles Angels of Anaheim
1831    2006    LAA         ANA  Los Angeles Angels of Anaheim
1861    2007    LAA         ANA  Los Angeles Angels of Anaheim
1891    2008    LAA         ANA  Los Angeles Angels of Anaheim
1921    2009    LAA         ANA  Los Angeles Angels of Anaheim
1951    2010    LAA         ANA  Los Angeles Angels of Anaheim
1981    2011    LAA         ANA  Los Angeles Angels of Anaheim
2010    2012    LAA         ANA  Los Angeles Angels of Anaheim
2040    2013    LAA         ANA  Los Angeles Angels of Anaheim
2070    2014    LAA         ANA  Los Angeles Angels of Anaheim
2100    2015    LAA         ANA  Los Angeles Angels of Anaheim
2130    2016    LAA         ANA  Los Angeles Angels of Anaheim
2160    2017    LAA         ANA  Los Angeles Angels of Anaheim
2190    2018    LAA         ANA  Los Angeles Angels of Anaheim
```

So clearly we have three teams where the IDs differ. We need to ask a few questions though:

Do they differ on those teams every time? We can't just take that for granted.

[10]:
```
df[df['franchID'] == 'ANA']['teamID'].value_counts()
```

[10]:
```
CAL    32
LAA    18
ANA     8
Name: teamID, dtype: int64
```

[11]:
```
df[(df['teamID'] != df['teamIDretro'])][['teamID', 'teamIDretro', 'name']].
 ↪shape[0]
```

[11]:
```
55
```

[12]:
```
df[(df['teamID'] == 'ML1')].shape[0] + df[(df['teamID'] == 'ML4')].shape[0] +␣
 ↪df[(df['teamID'] == 'LAA')].shape[0]
```

[12]:
```
59
```

Unfortunately we have a disparity of 4, so we need to find out where that is.

[13]:
```
df[(df['teamID'] == 'ML1') & (df['teamID'] == df['teamIDretro'])]
```

[13]:
```
Empty DataFrame
Columns: [yearID, lgID, teamID, franchID, divID, Rank, G, Ghome, W, L, DivWin,
WCWin, LgWin, WSWin, R, AB, H, 2B, 3B, HR, BB, SO, SB, CS, HBP, SF, RA, ER, ERA,
CG, SHO, SV, IPouts, HA, HRA, BBA, SOA, E, DP, FP, name, park, attendance, BPF,
PPF, teamIDretro]
Index: []

[0 rows x 46 columns]
```

[14]:
```
df[(df['teamID'] == 'ML4') & (df['teamID'] == df['teamIDretro'])]
```

[14]:
```
Empty DataFrame
Columns: [yearID, lgID, teamID, franchID, divID, Rank, G, Ghome, W, L, DivWin,
WCWin, LgWin, WSWin, R, AB, H, 2B, 3B, HR, BB, SO, SB, CS, HBP, SF, RA, ER, ERA,
CG, SHO, SV, IPouts, HA, HRA, BBA, SOA, E, DP, FP, name, park, attendance, BPF,
PPF, teamIDretro]
Index: []

[0 rows x 46 columns]
```

[15]:
```
df[(df['teamID'] == 'LAA') & (df['teamID'] == df['teamIDretro'])]
```

[15]:
```
     yearID lgID teamID franchID divID  Rank    G  Ghome   W   L  ...  SOA  \
680    1961   AL    LAA      ANA   NaN     8  162     82  70  91  ...  973
699    1962   AL    LAA      ANA   NaN     3  162     81  86  76  ...  858
719    1963   AL    LAA      ANA   NaN     9  161     81  70  91  ...  889
739    1964   AL    LAA      ANA   NaN     5  162     81  82  80  ...  965

       E   DP     FP                name                park  attendance  BPF  \
680  192  154  0.969  Los Angeles Angels  Wrigley Field (LA)      603510  111
699  175  153  0.972  Los Angeles Angels      Dodger Stadium     1144063   97
719  163  155  0.974  Los Angeles Angels      Dodger Stadium      821015   94
739  138  168  0.978  Los Angeles Angels      Dodger Stadium      760439   90

     PPF  teamIDretro
680  112          LAA
699   97          LAA
719   94          LAA
739   90          LAA

[4 rows x 46 columns]
```

[16]:
```
df[(df['teamID'] == 'LAA') & (df['teamID'] != df['teamIDretro'])]
```

[16]:
```
      yearID lgID teamID franchID divID  Rank    G  Ghome    W   L  ...   SOA  \
1801    2005   AL    LAA      ANA     W     1  162     81   95  67  ...  1126
1831    2006   AL    LAA      ANA     W     2  162     81   89  73  ...  1164
```

```
1861    2007   AL    LAA      ANA     W     1  162     81   94  68  ...  1156
1891    2008   AL    LAA      ANA     W     1  162     81  100  62  ...  1106
1921    2009   AL    LAA      ANA     W     1  162     81   97  65  ...  1062
1951    2010   AL    LAA      ANA     W     3  162     81   80  82  ...  1130
1981    2011   AL    LAA      ANA     W     2  162     81   86  76  ...  1058
2010    2012   AL    LAA      ANA     W     3  162     81   89  73  ...  1157
2040    2013   AL    LAA      ANA     W     3  162     81   78  84  ...  1200
2070    2014   AL    LAA      ANA     W     1  162     81   98  64  ...  1342
2100    2015   AL    LAA      ANA     W     3  162     81   85  77  ...  1221
2130    2016   AL    LAA      ANA     W     4  162     81   74  88  ...  1136
2160    2017   AL    LAA      ANA     W     2  162     81   80  82  ...  1312
2190    2018   AL    LAA      ANA     W     4  162     81   80  82  ...  1386

        E   DP     FP                          name  \
1801   87  139  0.986  Los Angeles Angels of Anaheim
1831  124  154  0.979  Los Angeles Angels of Anaheim
1861  101  154  0.983  Los Angeles Angels of Anaheim
1891   91  159  0.985  Los Angeles Angels of Anaheim
1921   85  174  0.986  Los Angeles Angels of Anaheim
1951  113  116  0.981  Los Angeles Angels of Anaheim
1981   93  157  0.985  Los Angeles Angels of Anaheim
2010   98  141  0.984  Los Angeles Angels of Anaheim
2040  112  135  0.981  Los Angeles Angels of Anaheim
2070   83  127  0.986  Los Angeles Angels of Anaheim
2100   93  108  0.984  Los Angeles Angels of Anaheim
2130   97  148  0.983  Los Angeles Angels of Anaheim
2160   80  135  0.986  Los Angeles Angels of Anaheim
2190   76  173  0.987  Los Angeles Angels of Anaheim

                          park  attendance  BPF  PPF  teamIDretro
1801             Angel Stadium     3404686   98   97          ANA
1831             Angel Stadium     3406790  100  100          ANA
1861             Angel Stadium     3365632  101  100          ANA
1891             Angel Stadium     3336747  103  102          ANA
1921             Angel Stadium     3240386   99   98          ANA
1951             Angel Stadium     3250816   98   98          ANA
1981             Angel Stadium     3166321   93   93          ANA
2010  Angel Stadium of Anaheim     3061770   92   92          ANA
2040  Angel Stadium of Anaheim     3019505   94   94          ANA
2070  Angel Stadium of Anaheim     3095935   96   95          ANA
2100  Angel Stadium of Anaheim     3012765   94   95          ANA
2130  Angel Stadium of Anaheim     3016142   95   95          ANA
2160  Angel Stadium of Anaheim     3019585   96   96          ANA
2190  Angel Stadium of Anaheim     3020216   97   97          ANA

[14 rows x 46 columns]
```

[17]:
```
df['franchID'].unique()
```

[17]:
```
array(['BOS', 'LAD', 'ATL', 'CHW', 'CHC', 'CIN', 'CLE', 'DET', 'SFG',
       'NYY', 'OAK', 'PHI', 'PIT', 'BAL', 'STL', 'MIN', 'ANA', 'TEX',
       'HOU', 'NYM', 'KCR', 'WSN', 'SDP', 'MIL', 'SEA', 'TOR', 'COL',
       'FLA', 'ARI', 'TBD'], dtype=object)
```

[18]:
```
df[(df['franchID'].isnull())]
```

[18]:
```
Empty DataFrame
Columns: [yearID, lgID, teamID, franchID, divID, Rank, G, Ghome, W, L, DivWin,
WCWin, LgWin, WSWin, R, AB, H, 2B, 3B, HR, BB, SO, SB, CS, HBP, SF, RA, ER, ERA,
CG, SHO, SV, IPouts, HA, HRA, BBA, SOA, E, DP, FP, name, park, attendance, BPF,
PPF, teamIDretro]
Index: []

[0 rows x 46 columns]
```

[19]:
```
df['franchID'].nunique()
```

[19]:
```
30
```

It looks like it will be easiest to just use the franchise ID - they stay consistent throughout and there are only ever 30 max. We'll need a way to map to these values from an external script so we can use it in other files.