teams_pre

February 27, 2020

[18]:
```
import math
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[19]:
```
df = pd.read_csv('./data/lahman/mlb_data/Teams.csv')
```

[20]:
```
df.columns
```

[20]:
```
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')
```

[21]:
```
df.head()
```

[21]:

	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	...	DP	FP	name	park	attendance	BPF	PPF	teamIDBR	teamIDlahman45	teamIDretro
0	1919	AL	BOS	BOS	NaN	6	138	66	66	71	...	118	0.975	Boston Red Sox	Fenway Park I	417291	94	94	BOS	BOS	BOS
1	1919	NL	BRO	LAD	NaN	5	141	70	69	71	...	84	0.963	Brooklyn Robins	Ebbets Field	360721	103	103	BRO	BRO	BRO
2	1919	NL	BSN	ATL	NaN	6	140	68	57	82	...	111	0.966	Boston Braves	Braves Field	167401	95	98	BSN	BSN	BSN
3	1919	AL	CHA	CHW	NaN	1	140	70	88	52	...	116	0.969	Chicago White Sox	Comiskey Park	627186	100	99	CHW		
4	1919	NL	CHN	CHC	NaN	3	140	71	75	65	...	87	0.969	Chicago Cubs	Wrigley Field	424430	100	99	CHC		

```
3             CHA          CHA
4             CHN          CHN

[5 rows x 48 columns]
```

[22]:
```
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.

[23]:
```
df[(df['teamID'] != df['teamIDretro'])][['yearID', 'teamID', 'teamIDretro',␣
↪'name']]
```

[23]:

	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.

[30]:
```
df[df['franchID'] == 'ANA']['teamID'].value_counts()
```

[30]:
```
CAL    32
LAA    18
ANA     8
Name: teamID, dtype: int64
```

[18]:
```
df[(df['teamID'] != df['teamIDretro'])][['teamID', 'teamIDretro', 'name']].
 ↪shape[0]
```

[18]:
```
55
```

[16]:
```
df[(df['teamID'] == 'ML1')].shape[0] + df[(df['teamID'] == 'ML4')].shape[0] +␣
 ↪df[(df['teamID'] == 'LAA')].shape[0]
```

[16]:
```
59
```

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

[24]:
```
df[(df['teamID'] == 'ML1') & (df['teamID'] == df['teamIDretro'])]
```

[24]:
```
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, teamIDBR, teamIDlahman45, teamIDretro]
Index: []

[0 rows x 48 columns]
```

[25]:
```
df[(df['teamID'] == 'ML4') & (df['teamID'] == df['teamIDretro'])]
```

[25]:
```
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, teamIDBR, teamIDlahman45, teamIDretro]
Index: []

[0 rows x 48 columns]
```

[37]:
```
df[(df['teamID'] == 'LAA') & (df['teamID'] == df['teamIDretro'])]
```

[37]:

	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	...	DP	\
680	1961	AL	LAA	ANA	NaN	8	162	82	70	91	...	154	
699	1962	AL	LAA	ANA	NaN	3	162	81	86	76	...	153	
719	1963	AL	LAA	ANA	NaN	9	161	81	70	91	...	155	
739	1964	AL	LAA	ANA	NaN	5	162	81	82	80	...	168	

	FP	name	park	attendance	BPF	PPF	\
680	0.969	Los Angeles Angels	Wrigley Field (LA)	603510	111	112	
699	0.972	Los Angeles Angels	Dodger Stadium	1144063	97	97	
719	0.974	Los Angeles Angels	Dodger Stadium	821015	94	94	
739	0.978	Los Angeles Angels	Dodger Stadium	760439	90	90	

	teamIDBR	teamIDlahman45	teamIDretro
680	LAA	LAA	LAA
699	LAA	LAA	LAA
719	LAA	LAA	LAA
739	LAA	LAA	LAA

[4 rows x 48 columns]

[36]:
```
df[(df['teamID'] == 'LAA') & (df['teamID'] != df['teamIDretro'])]
```

[36]:

	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	...	DP	\
1801	2005	AL	LAA	ANA	W	1	162	81	95	67	...	139	
1831	2006	AL	LAA	ANA	W	2	162	81	89	73	...	154	
1861	2007	AL	LAA	ANA	W	1	162	81	94	68	...	154	
1891	2008	AL	LAA	ANA	W	1	162	81	100	62	...	159	
1921	2009	AL	LAA	ANA	W	1	162	81	97	65	...	174	
1951	2010	AL	LAA	ANA	W	3	162	81	80	82	...	116	
1981	2011	AL	LAA	ANA	W	2	162	81	86	76	...	157	
2010	2012	AL	LAA	ANA	W	3	162	81	89	73	...	141	
2040	2013	AL	LAA	ANA	W	3	162	81	78	84	...	135	
2070	2014	AL	LAA	ANA	W	1	162	81	98	64	...	127	
2100	2015	AL	LAA	ANA	W	3	162	81	85	77	...	108	
2130	2016	AL	LAA	ANA	W	4	162	81	74	88	...	148	
2160	2017	AL	LAA	ANA	W	2	162	81	80	82	...	135	
2190	2018	AL	LAA	ANA	W	4	162	81	80	82	...	173	

	FP	name	park	\
1801	0.986	Los Angeles Angels of Anaheim	Angel Stadium	
1831	0.979	Los Angeles Angels of Anaheim	Angel Stadium	
1861	0.983	Los Angeles Angels of Anaheim	Angel Stadium	
1891	0.985	Los Angeles Angels of Anaheim	Angel Stadium	
1921	0.986	Los Angeles Angels of Anaheim	Angel Stadium	
1951	0.981	Los Angeles Angels of Anaheim	Angel Stadium	
1981	0.985	Los Angeles Angels of Anaheim	Angel Stadium	
2010	0.984	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2040	0.981	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2070	0.986	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2100	0.984	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2130	0.983	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2160	0.986	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	
2190	0.987	Los Angeles Angels of Anaheim	Angel Stadium of Anaheim	

	attendance	BPF	PPF	teamIDBR	teamIDlahman45	teamIDretro
1801	3404686	98	97	LAA	ANA	ANA
1831	3406790	100	100	LAA	ANA	ANA
1861	3365632	101	100	LAA	ANA	ANA
1891	3336747	103	102	LAA	ANA	ANA
1921	3240386	99	98	LAA	ANA	ANA
1951	3250816	98	98	LAA	ANA	ANA
1981	3166321	93	93	LAA	ANA	ANA
2010	3061770	92	92	LAA	ANA	ANA
2040	3019505	94	94	LAA	ANA	ANA
2070	3095935	96	95	LAA	ANA	ANA
2100	3012765	94	95	LAA	ANA	ANA
2130	3016142	95	95	LAA	ANA	ANA
2160	3019585	96	96	LAA	ANA	ANA
2190	3020216	97	97	LAA	ANA	ANA

```
[14 rows x 48 columns]
```

[29]:
```
df['franchID'].unique()
```

[29]:
```
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)
```

[40]:
```
df[(df['franchID'].isnull())]
```

[40]:
```
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, teamIDBR, teamIDlahman45, teamIDretro]
Index: []

[0 rows x 48 columns]
```

[31]:
```
df['franchID'].nunique()
```

[31]:
```
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.

[ ]:
```

```