convert_gamelogs

May 7, 2020

Processing Game Logs

The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at “www.retrosheet.org”.

We’re going to use Retrosheet game logs as input data to our predictive model. The first thing we need to do is process them to fit our needs.

[135]:

```
import pandas as pd
import os
```

[156]:

```
df = pd.read_csv('../core/data/lahman/mlb_data/Teams.csv')
df = df[['teamID', 'franchID']]
team_dict = df.set_index('teamID').to_dict()['franchID']
team_dict['MLN'] = 'ATL'


def get_team(team):
    return team_dict[team] if team_dict[team] is not None else team
```

These are the columns of the Retrosheet game logs. This metadata was obtained here: https://www.retrosheet.org/gamelogs/glfields.txt

[137]:

```
columns = [
    'date',
    'game_number',
    'day_of_week',
    'visit_team',
    'visit_league',
    'visit_game_number',
    'home_team',
    'home_league',
    'home_game_number',
    'visit_score',
    'home_score',
    'game_length_outs',
    'day_night',
    'completion_info',
    'forfeit_info',
    'protest_info',
    'park_id',
    'attendance',
    'time_minutes',
    'visit_line_score',
    'home_line_score',
    'visit_ab',
    'visit_h',
    'visit_2b',
    'visit_3b',
    'visit_hr',
    'visit_rbi',
    'visit_sh',
    'visit_sf',
    'visit_hbp',
    'visit_bb',
    'visit_ibb',
    'visit_k',
    'visit_sb',
    'visit_cs',
    'visit_gidp',
    'visit_ci',
    'visit_lob',
    'visit_pitchers_used',
    'visit_individual_er',
    'visit_team_er',
    'visit_wp',
    'visit_bk',
    'visit_po',
    'visit_assists',
    'visit_e',
    'visit_passed_balls',
    'visit_double_plays',
    'visit_triple_plays',
    'home_ab',
    'home_h',
    'home_2b',
    'home_3b',
    'home_hr',
    'home_rbi',
    'home_sh',
    'home_sf',
    'home_hbp',
    'home_bb',
    'home_ibb',
    'home_k',
    'home_sb',
    'home_cs',
    'home_gidp',
    'home_ci',
    'home_lob',
    'home_pitchers_used',
    'home_individual_er',
    'home_team_er',
    'home_wp',
    'home_bk',
    'home_po',
    'home_assists',
    'home_e',
    'home_passed_balls',
    'home_double_plays',
    'home_triple_plays',
    'hp_ump_id',
    'hp_ump_name',
    '1b_ump_id',
    '1b_ump_name',
    '2b_ump_id',
    '2b_ump_name',
    '3b_ump_id',
    '3b_ump_name',
    'lf_ump_id',
    'lf_ump_name',
    'rf_ump_id',
    'rf_ump_name',
    'visit_manager_id',
    'visit_manager_name',
    'home_manager_id',
    'home_manager_name',
    'winning_pitcher_id',
    'winning_pitcher_name',
    'losing_pitcher_id',
    'losing_pitcher_name',
    'saving_pitcher_id',
    'saving_pitcher_name',
    'winning_rbi_batter_id',
    'winning_rbi_batter_name',
    'visit_sp_id',
    'visit_sp_name',
    'home_sp_id',
    'home_sp_name',
    'visit_player_1_id',
    'visit_player_1_name',
    'visit_player_1_pos',
    'visit_player_2_id',
    'visit_player_2_name',
    'visit_player_2_pos',
    'visit_player_3_id',
    'visit_player_3_name',
    'visit_player_3_pos',
    'visit_player_4_id',
    'visit_player_4_name',
    'visit_player_4_pos',
    'visit_player_5_id',
    'visit_player_5_name',
    'visit_player_5_pos',
    'visit_player_6_id',
    'visit_player_6_name',
    'visit_player_6_pos',
    'visit_player_7_id',
    'visit_player_7_name',
    'visit_player_7_pos',
    'visit_player_8_id',
    'visit_player_8_name',
    'visit_player_8_pos',
    'visit_player_9_id',
    'visit_player_9_name',
    'visit_player_9_pos',
    'home_player_1_id',
    'home_player_1_name',
    'home_player_1_pos',
    'home_player_2_id',
    'home_player_2_name',
    'home_player_2_pos',
    'home_player_3_id',
    'home_player_3_name',
    'home_player_3_pos',
    'home_player_4_id',
    'home_player_4_name',
    'home_player_4_pos',
    'home_player_5_id',
    'home_player_5_name',
    'home_player_5_pos',
    'home_player_6_id',
    'home_player_6_name',
    'home_player_6_pos',
    'home_player_7_id',
    'home_player_7_name',
    'home_player_7_pos',
    'home_player_8_id',
    'home_player_8_name',
    'home_player_8_pos',
    'home_player_9_id',
    'home_player_9_name',
    'home_player_9_pos',
    'additional_info',
    'acquisition_info'
]
```

The script is broken up here, then I later explore what I need to do to process the data. At the end I combine that all into one loop.

```
[12]: for year in range(1919, 2020):
          file_path = '../core/data/retrosheet/gamelogs/GL{}'.format(year)
          df = pd.read_csv(file_path + '.TXT', delimiter = ',', header = 0, names =␣
       ↪columns)
          if os.path.exists(file_path + '.TXT'):
              os.remove(file_path + '.TXT')
          df.to_csv(file_path + '.csv')
```

```
[109]: df = pd.read_csv('../core/data/retrosheet/gamelogs/GL2015.csv')
```

We don't want every column, so we'll specify exactly which ones to use

```
[110]: df = df[[
               'date',
               'visit_team',
               'home_team',
               'visit_score',
               'home_score',
               'game_length_outs',
               'day_night',
               'park_id',
               'visit_manager_id',
               'home_manager_id',
               'visit_sp_id',
               'home_sp_id',
               'visit_player_1_id',
               'visit_player_2_id',
               'visit_player_3_id',
               'visit_player_4_id',
               'visit_player_5_id',
               'visit_player_6_id',
               'visit_player_7_id',
               'visit_player_8_id',
               'visit_player_9_id',
               'home_player_1_id',
               'home_player_2_id',
               'home_player_3_id',
               'home_player_4_id',
               'home_player_5_id',
               'home_player_6_id',
               'home_player_7_id',
               'home_player_8_id',
               'home_player_9_id'
           ]]
```

[111]:

```
df
```

[111]:

```
          date visit_team home_team  visit_score  home_score  \
0     20150406        MIN       DET            0           4
1     20150406        CLE       HOU            0           2
2     20150406        CHA       KCA            1          10
3     20150406        TOR       NYA            6           1
4     20150406        TEX       OAK            0           8
...        ...        ...       ...          ...         ...
2423  20151004        CHN       MIL            3           1
2424  20151004        WAS       NYN            0           1
2425  20151004        MIA       PHI            2           7
2426  20151004        CIN       PIT            0           4
2427  20151004        COL       SFN            7           3

      game_length_outs day_night park_id visit_manager_id home_manager_id  \
0                   51         D   DET05         molip001        ausmb001
1                   51         N   HOU03         frant001        hinca001
2                   51         D   KAN06         ventr001        yoste001
3                   54         D   NYC21         gibbj001        giraj001
4                   51         N   OAK01         banij001        melvb001
...                ...       ...     ...              ...             ...
2423                54         D   MIL06         maddj801        counc001
2424                51         D   NYC20         willm003        collt801
2425                51         D   PHI13         jennd801        mackp101
2426                51         D   PIT08         pricb801        hurdc001
2427                54         D   SFO03         weisw001        bochb002

      ... visit_player_9_id home_player_1_id home_player_2_id  \
0     ...          schaj002         davir003         kinsi001
1     ...          ramij003         altuj001         sprig001
2     ...          johnm006         escoa003         mousm001
3     ...          travd001         ellsj001         gardb001
4     ...          odorr001         gentc001         fulds001
...   ...               ...              ...              ...
2423  ...          hared001         genns001         petes002
2424  ...          roart001         granc001         wrigd002
2425  ...          conla001         galvf001         altha001
2426  ...          smitj004         polag001         harrj002
```

```
2427  ...          bergc001          pagaa001          tomlk001

     home_player_3_id home_player_4_id home_player_5_id home_player_6_id  \
0            cabrm001         martv001         martj006         cespy001
1            valbl001         gatte001         cartc002         castj006
2            cainl001         hosme001         morak001         gorda001
3            beltc001         teixm001         mccab002         headc001
4            zobrb001         butlb003         davii001         lawrb002
...               ...              ...              ...              ...
2423         linda001         davik003         santd002         pereh001
2424         murpd006         cespy001         dudal001         darnt001
2425         franm004         ruf-d001         franj004         blana001
2426         mccua001         walkn001         marts002         alvap001
2427         duffm002         poseb001         parkj002         willm008

     home_player_7_id home_player_8_id home_player_9_id
0            castn001         avila001         iglej001
1            lowrj001         rasmc001         marij002
2            riosa002         peres002         infao001
3            rodra001         drews001         gregd001
4            vogts001         semim001         sogae001
...               ...              ...              ...
2423         seguj002         maldm001         lopej004
2424         confm001         tejar001         degrj001
2425         krate001         ruppc001         buchd001
2426         cervf001         mercj002         happj001
2427         noonn001         willj005         cainm001

[2428 rows x 33 columns]
```

```
[112]: df['date'] = df['date'].astype(str)
```

'date' isn't very useful, so we'll export it to three separate columns.

```
[113]: df['year'] = df['date'].str[0:4].astype(int)
       df['month'] = df['date'].str[4:6].astype(int)
       df['day'] = df['date'].str[6:8].astype(int)
```

```
[114]: df
```

```
[114]:          date visit_team home_team  visit_score  home_score  \
       0     20150406        MIN       DET            0           4
       1     20150406        CLE       HOU            0           2
       2     20150406        CHA       KCA            1          10
       3     20150406        TOR       NYA            6           1
       4     20150406        TEX       OAK            0           8
       ...        ...        ...       ...          ...         ...
```

```
2423  20151004        CHN       MIL            3           1
2424  20151004        WAS       NYN            0           1
2425  20151004        MIA       PHI            2           7
2426  20151004        CIN       PIT            0           4
2427  20151004        COL       SFN            7           3

      game_length_outs day_night park_id visit_manager_id home_manager_id  \
0                   51         D   DET05         molip001        ausmb001
1                   51         N   HOU03         frant001        hinca001
2                   51         D   KAN06         ventr001        yoste001
3                   54         D   NYC21         gibbj001        giraj001
4                   51         N   OAK01         banij001        melvb001
...                ...       ...     ...              ...             ...
2423                54         D   MIL06         maddj801        counc001
2424                51         D   NYC20         willm003        collt801
2425                51         D   PHI13         jennd801        mackp101
2426                51         D   PIT08         pricb801        hurdc001
2427                54         D   SFO03         weisw001        bochb002

      ... home_player_3_id home_player_4_id home_player_5_id home_player_6_id  \
0     ...         cabrm001         martv001         martj006         cespy001
1     ...         valbl001         gatte001         cartc002         castj006
2     ...         cainl001         hosme001         morak001         gorda001
3     ...         beltc001         teixm001         mccab002         headc001
4     ...         zobrb001         butlb003         davii001         lawrb002
...   ...              ...              ...              ...              ...
2423  ...         linda001         davik003         santd002         pereh001
2424  ...         murpd006         cespy001         dudal001         darnt001
2425  ...         franm004         ruf-d001         franj004         blana001
2426  ...         mccua001         walkn001         marts002         alvap001
2427  ...         duffm002         poseb001         parkj002         willm008

     home_player_7_id home_player_8_id home_player_9_id  year month day
0            castn001         avila001         iglej001  2015     4   6
1            lowrj001         rasmc001         marij002  2015     4   6
2            riosa002         peres002         infao001  2015     4   6
3            rodra001         drews001         gregd001  2015     4   6
4            vogts001         semim001         sogae001  2015     4   6
...               ...              ...              ...   ...   ...  ..
2423         seguj002         maldm001         lopej004  2015    10   4
2424         confm001         tejar001         degrj001  2015    10   4
2425         krate001         ruppc001         buchd001  2015    10   4
2426         cervf001         mercj002         happj001  2015    10   4
2427         noonn001         willj005         cainm001  2015    10   4

[2428 rows x 36 columns]
```

We aren't going to use every column in the final model, but we want to make sure that the ones we will are in the proper format.

```
[115]: night_game = pd.get_dummies(df['day_night'], drop_first=True)
```

```
[116]: night_game
```

```
[116]:       N
        0     0
        1     1
        2     0
        3     0
        4     1
        ...  ..
        2423  0
        2424  0
        2425  0
        2426  0
        2427  0

        [2428 rows x 1 columns]
```

```
[117]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2428 entries, 0 to 2427
Data columns (total 36 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   date                2428 non-null   object
 1   visit_team          2428 non-null   object
 2   home_team           2428 non-null   object
 3   visit_score         2428 non-null   int64
 4   home_score          2428 non-null   int64
 5   game_length_outs    2428 non-null   int64
 6   day_night           2428 non-null   object
 7   park_id             2428 non-null   object
 8   visit_manager_id    2428 non-null   object
 9   home_manager_id     2428 non-null   object
 10  winning_pitcher_id  2428 non-null   object
 11  losing_pitcher_id   2428 non-null   object
 12  saving_pitcher_id   1291 non-null   object
 13  visit_sp_id         2428 non-null   object
 14  home_sp_id          2428 non-null   object
 15  visit_player_1_id   2428 non-null   object
 16  visit_player_2_id   2428 non-null   object
 17  visit_player_3_id   2428 non-null   object
 18  visit_player_4_id   2428 non-null   object
 19  visit_player_5_id   2428 non-null   object
 20  visit_player_6_id   2428 non-null   object
 21  visit_player_7_id   2428 non-null   object
 22  visit_player_8_id   2428 non-null   object
 23  visit_player_9_id   2428 non-null   object
 24  home_player_1_id    2428 non-null   object
 25  home_player_2_id    2428 non-null   object
 26  home_player_3_id    2428 non-null   object
 27  home_player_4_id    2428 non-null   object
 28  home_player_5_id    2428 non-null   object
 29  home_player_6_id    2428 non-null   object
 30  home_player_7_id    2428 non-null   object
 31  home_player_8_id    2428 non-null   object
 32  home_player_9_id    2428 non-null   object
 33  year                2428 non-null   int64
 34  month               2428 non-null   int64
 35  day                 2428 non-null   int64
dtypes: int64(6), object(30)
memory usage: 683.0+ KB
```

[118]:

```
df.insert(loc=6, column='night_game', value=night_game)
```

[119]:

```
df
```

[119]:

```
          date visit_team home_team  visit_score  home_score  \
0     20150406        MIN       DET            0           4
1     20150406        CLE       HOU            0           2
2     20150406        CHA       KCA            1          10
3     20150406        TOR       NYA            6           1
4     20150406        TEX       OAK            0           8
...        ...        ...       ...          ...         ...
2423  20151004        CHN       MIL            3           1
2424  20151004        WAS       NYN            0           1
2425  20151004        MIA       PHI            2           7
2426  20151004        CIN       PIT            0           4
2427  20151004        COL       SFN            7           3

      game_length_outs  night_game day_night park_id visit_manager_id  ...  \
0                   51           0         D   DET05         molip001  ...
1                   51           1         N   HOU03         frant001  ...
2                   51           0         D   KAN06         ventr001  ...
3                   54           0         D   NYC21         gibbj001  ...
4                   51           1         N   OAK01         banij001  ...
...                ...         ...       ...     ...              ...  ...
2423                54           0         D   MIL06         maddj801  ...
2424                51           0         D   NYC20         willm003  ...
2425                51           0         D   PHI13         jennd801  ...
```

```
2426                 51            0         D   PIT08          pricb801  ...
2427                 54            0         D   SF003          weisw001  ...

     home_player_3_id home_player_4_id home_player_5_id home_player_6_id  \
0            cabrm001         martv001         martj006         cespy001
1            valbl001         gatte001         cartc002         castj006
2            cainl001         hosme001         morak001         gorda001
3            beltc001         teixm001         mccab002         headc001
4            zobrb001         butlb003         davii001         lawrb002
...               ...              ...              ...              ...
2423         linda001         davik003         santd002         pereh001
2424         murpd006         cespy001         dudal001         darnt001
2425         franm004         ruf-d001         franj004         blana001
2426         mccua001         walkn001         marts002         alvap001
2427         duffm002         poseb001         parkj002         willm008

     home_player_7_id home_player_8_id home_player_9_id  year month day
0            castn001         avila001         iglej001  2015     4   6
1            lowrj001         rasmc001         marij002  2015     4   6
2            riosa002         peres002         infao001  2015     4   6
3            rodra001         drews001         gregd001  2015     4   6
4            vogts001         semim001         sogae001  2015     4   6
...               ...              ...              ...   ...   ...  ..
2423         seguj002         maldm001         lopej004  2015    10   4
2424         confm001         tejar001         degrj001  2015    10   4
2425         krate001         ruppc001         buchd001  2015    10   4
2426         cervf001         mercj002         happj001  2015    10   4
2427         noonn001         willj005         cainm001  2015    10   4

[2428 rows x 37 columns]
```

[120]:
```
to_drop = ['date', 'day_night']
```

[121]:
```
df = df.drop(columns=to_drop)
```

[122]:
```
df
```

[122]:
```
     visit_team home_team  visit_score  home_score  game_length_outs  \
0           MIN       DET            0           4                51
1           CLE       HOU            0           2                51
2           CHA       KCA            1          10                51
3           TOR       NYA            6           1                54
4           TEX       OAK            0           8                51
...         ...       ...          ...         ...               ...
2423        CHN       MIL            3           1                54
2424        WAS       NYN            0           1                51
2425        MIA       PHI            2           7                51
```

```
2426        CIN       PIT            0           4                51
2427        COL       SFN            7           3                54

      night_game park_id visit_manager_id home_manager_id winning_pitcher_id  \
0              0   DET05         molip001        ausmb001           pricd001
1              1   HOU03         frant001        hinca001           keucd001
2              0   KAN06         ventr001        yoste001           venty001
3              0   NYC21         gibbj001        giraj001           hutcd001
4              1   OAK01         banij001        melvb001           grays001
...          ...     ...              ...             ...                ...
2423           0   MIL06         maddj801        counc001           hared001
2424           0   NYC20         willm003        collt801           clipt001
2425           0   PHI13         jennd801        mackp101           garcl005
2426           0   PIT08         pricb801        hurdc001           happj001
2427           0   SFO03         weisw001        bochb002           brotr001

      ... home_player_3_id home_player_4_id home_player_5_id home_player_6_id  \
0     ...         cabrm001         martv001         martj006         cespy001
1     ...         valbl001         gatte001         cartc002         castj006
2     ...         cainl001         hosme001         morak001         gorda001
3     ...         beltc001         teixm001         mccab002         headc001
4     ...         zobrb001         butlb003         davii001         lawrb002
...   ...              ...              ...              ...              ...
2423  ...         linda001         davik003         santd002         pereh001
2424  ...         murpd006         cespy001         dudal001         darnt001
2425  ...         franm004         ruf-d001         franj004         blana001
2426  ...         mccua001         walkn001         marts002         alvap001
2427  ...         duffm002         poseb001         parkj002         willm008

     home_player_7_id home_player_8_id home_player_9_id  year month day
0            castn001         avila001         iglej001  2015     4   6
1            lowrj001         rasmc001         marij002  2015     4   6
2            riosa002         peres002         infao001  2015     4   6
3            rodra001         drews001         gregd001  2015     4   6
4            vogts001         semim001         sogae001  2015     4   6
...               ...              ...              ...   ...   ...  ..
2423         seguj002         maldm001         lopej004  2015    10   4
2424         confm001         tejar001         degrj001  2015    10   4
2425         krate001         ruppc001         buchd001  2015    10   4
2426         cervf001         mercj002         happj001  2015    10   4
2427         noonn001         willj005         cainm001  2015    10   4

[2428 rows x 35 columns]
```

[123]:
```
df['visit_team'] = df['visit_team'].apply(get_team)
```

[124]:
```
df['home_team'] = df['home_team'].apply(get_team)
```

Early games only took place during the day, so we need to handle the effects of using one-hot encoding when dropping first with those.

```
[147]: file_path = '../core/data/retrosheet/gamelogs/GL{}'.format(1919)
       df = pd.read_csv(file_path + '.TXT', delimiter = ',', header = 0, names =␣
        ↪columns)
```

```
[148]: df['day_night'].nunique()
```

```
[148]: 1
```

```
[149]: pd.get_dummies(df['day_night'])
```

```
[149]:       D
       0     1
       1     1
       2     1
       3     1
       4     1
       ...  ..
       1112  1
       1113  1
       1114  1
       1115  1
       1116  1

       [1117 rows x 1 columns]
```

Final Script

```
[172]: for year in range(1919, 2020):
           file_path = '../core/data/retrosheet/gamelogs/GL{}'.format(year)
           df = pd.read_csv(file_path + '.TXT', delimiter = ',', header = 0, names =␣
        ↪columns)
           df = df[[
               'date',
               'visit_team',
               'home_team',
               'visit_score',
               'home_score',
               'game_length_outs',
               'day_night',
               'park_id',
               'visit_manager_id',
               'home_manager_id',
               'winning_pitcher_id',
               'losing_pitcher_id',
               'saving_pitcher_id',
               'visit_sp_id',
               'home_sp_id',
               'visit_player_1_id',
               'visit_player_2_id',
               'visit_player_3_id',
               'visit_player_4_id',
               'visit_player_5_id',
               'visit_player_6_id',
               'visit_player_7_id',
               'visit_player_8_id',
               'visit_player_9_id',
               'home_player_1_id',
               'home_player_2_id',
               'home_player_3_id',
               'home_player_4_id',
               'home_player_5_id',
               'home_player_6_id',
               'home_player_7_id',
               'home_player_8_id',
               'home_player_9_id',
           ]]
           df['date'] = df['date'].astype(str)
           df['year'] = df['date'].str[0:4].astype(int)
           df['month'] = df['date'].str[4:6].astype(int)
           df['day'] = df['date'].str[6:8].astype(int)
           night_game = pd.get_dummies(df['day_night'], drop_first=(df['day_night'].
        ↪nunique() > 1))
           df.insert(loc=6, column='night_game', value=night_game)
           df = df.drop(columns=['date', 'day_night'])
           df['visit_team'] = df['visit_team'].apply(get_team)
           df['home_team'] = df['home_team'].apply(get_team)
           if os.path.exists(file_path + '.TXT'):
               os.remove(file_path + '.TXT')
           df.to_csv(file_path + '.csv', index=False)
```

[170]:
```
df = pd.read_csv('../core/data/retrosheet/gamelogs/GL2015.csv')
```

[173]:
```
df
```

[173]:
```
     visit_team home_team  visit_score  home_score  game_length_outs  \
0           MIN       DET            0           4                51
1           CLE       HOU            0           2                51
2           CHW       KCR            1          10                51
3           TOR       NYY            6           1                54
4           TEX       OAK            0           8                51
...         ...       ...          ...         ...               ...
2423        CHC       MIL            3           1                54
```

```
2424        WSN       NYM            0           1                51
2425        FLA       PHI            2           7                51
2426        CIN       PIT            0           4                51
2427        COL       SFG            7           3                54

      night_game park_id visit_manager_id home_manager_id winning_pitcher_id  \
0              0   DET05         molip001        ausmb001           pricd001
1              1   HOU03         frant001        hinca001           keucd001
2              0   KAN06         ventr001        yoste001           venty001
3              0   NYC21         gibbj001        giraj001           hutcd001
4              1   OAK01         banij001        melvb001           grays001
...          ...     ...              ...             ...                ...
2423           0   MIL06         maddj801        counc001           hared001
2424           0   NYC20         willm003        collt801           clipt001
2425           0   PHI13         jennd801        mackp101           garcl005
2426           0   PIT08         pricb801        hurdc001           happj001
2427           0   SFO03         weisw001        bochb002           brotr001

      ... home_player_3_id home_player_4_id home_player_5_id home_player_6_id  \
0     ...         cabrm001         martv001         martj006         cespy001
1     ...         valbl001         gatte001         cartc002         castj006
2     ...         cainl001         hosme001         morak001         gorda001
3     ...         beltc001         teixm001         mccab002         headc001
4     ...         zobrb001         butlb003         davii001         lawrb002
...   ...              ...              ...              ...              ...
2423  ...         linda001         davik003         santd002         pereh001
2424  ...         murpd006         cespy001         dudal001         darnt001
2425  ...         franm004         ruf-d001         franj004         blana001
2426  ...         mccua001         walkn001         marts002         alvap001
2427  ...         duffm002         poseb001         parkj002         willm008

     home_player_7_id home_player_8_id home_player_9_id  year month day
0            castn001         avila001         iglej001  2015     4   6
1            lowrj001         rasmc001         marij002  2015     4   6
2            riosa002         peres002         infao001  2015     4   6
3            rodra001         drews001         gregd001  2015     4   6
4            vogts001         semim001         sogae001  2015     4   6
...               ...              ...              ...   ...   ...  ..
2423         seguj002         maldm001         lopej004  2015    10   4
2424         confm001         tejar001         degrj001  2015    10   4
2425         krate001         ruppc001         buchd001  2015    10   4
2426         cervf001         mercj002         happj001  2015    10   4
2427         noonn001         willj005         cainm001  2015    10   4

[2428 rows x 35 columns]
```

When we do the actual script, we don’t want the column names hardcoded into it. So I’ve pasted those to .csv files but I need to process them a bit.

```
[252]: gla = pd.read_csv('../core/data/retrosheet/rs_gl_cols_all.csv', header=None)
       gl = pd.read_csv('../core/data/retrosheet/rs_gl_cols.csv', header=None)
```

```
[254]: gl
```

```
[254]:                                  0
       0                         'date',
       1                   'visit_team',
       2                    'home_team',
       3                  'visit_score',
       4                   'home_score',
       5             'game_length_outs',
       6                    'day_night',
       7                      'park_id',
       8             'visit_manager_id',
       9              'home_manager_id',
       10          'winning_pitcher_id',
       11           'losing_pitcher_id',
       12           'saving_pitcher_id',
       13                'visit_sp_id',
       14                 'home_sp_id',
       15           'visit_player_1_id',
       16           'visit_player_2_id',
       17           'visit_player_3_id',
       18           'visit_player_4_id',
       19           'visit_player_5_id',
       20           'visit_player_6_id',
       21           'visit_player_7_id',
       22           'visit_player_8_id',
       23           'visit_player_9_id',
       24            'home_player_1_id',
       25            'home_player_2_id',
       26            'home_player_3_id',
       27            'home_player_4_id',
       28            'home_player_5_id',
       29            'home_player_6_id',
       30            'home_player_7_id',
       31            'home_player_8_id',
       32             'home_player_9_id'
```

We need to get rid of whitespace, commas and quotation marks.

```
[189]: gla.iloc[156][0].replace(',', '')
```

```
[189]: "   'home_player_9_id'"
```

[190]:
```
from functools import reduce
```

[241]:
```
def trim_cell(cell):
    replacements = {' ': '', "'": '', ',': ''}
    string = cell[0]
    return reduce(lambda a, kv: a.replace(*kv), replacements.items(), string)
```

[200]:
```
print(trim_cell(gla.iloc[156]))
```

```
home_player_9_id
```

[255]:
```
gla = gla.apply(trim_cell, axis=1)
```

[256]:
```
type(gla)
```

[256]:
```
pandas.core.series.Series
```

[257]:
```
gl = gl.apply(trim_cell, axis=1)
```

[258]:
```
gla.to_csv('../core/data/retrosheet/rs_gl_cols_all.csv', header=None)
gl.to_csv('../core/data/retrosheet/rs_gl_cols.csv', header=None)
```

[259]:
```
gla = pd.read_csv('../core/data/retrosheet/rs_gl_cols_all.csv', header=None)
gl = pd.read_csv('../core/data/retrosheet/rs_gl_cols.csv', header=None)
```

[260]:
```
gl
```

[260]:
```
      0                  1
0     0               date
1     1        visit_team
2     2         home_team
3     3       visit_score
4     4        home_score
5     5   game_length_outs
6     6         day_night
7     7           park_id
8     8   visit_manager_id
9     9    home_manager_id
10   10  winning_pitcher_id
11   11   losing_pitcher_id
12   12   saving_pitcher_id
13   13        visit_sp_id
14   14         home_sp_id
15   15   visit_player_1_id
16   16   visit_player_2_id
17   17   visit_player_3_id
18   18   visit_player_4_id
```

```
19  19   visit_player_5_id
20  20   visit_player_6_id
21  21   visit_player_7_id
22  22   visit_player_8_id
23  23   visit_player_9_id
24  24    home_player_1_id
25  25    home_player_2_id
26  26    home_player_3_id
27  27    home_player_4_id
28  28    home_player_5_id
29  29    home_player_6_id
30  30    home_player_7_id
31  31    home_player_8_id
32  32    home_player_9_id
```

[261]:

```
gla[1].tolist()
```

[261]:

```
['date',
 'game_number',
 'day_of_week',
 'visit_team',
 'visit_league',
 'visit_game_number',
 'home_team',
 'home_league',
 'home_game_number',
 'visit_score',
 'home_score',
 'game_length_outs',
 'day_night',
 'completion_info',
 'forfeit_info',
 'protest_info',
 'park_id',
 'attendance',
 'time_minutes',
 'visit_line_score',
 'home_line_score',
 'visit_ab',
 'visit_h',
 'visit_2b',
 'visit_3b',
 'visit_hr',
 'visit_rbi',
 'visit_sh',
 'visit_sf',
 'visit_hbp',
'visit_bb',
'visit_ibb',
'visit_k',
'visit_sb',
'visit_cs',
'visit_gidp',
'visit_ci',
'visit_lob',
'visit_pitchers_used',
'visit_individual_er',
'visit_team_er',
'visit_wp',
'visit_bk',
'visit_po',
'visit_assists',
'visit_e',
'visit_passed_balls',
'visit_double_plays',
'visit_triple_plays',
'home_ab',
'home_h',
'home_2b',
'home_3b',
'home_hr',
'home_rbi',
'home_sh',
'home_sf',
'home_hbp',
'home_bb',
'home_ibb',
'home_k',
'home_sb',
'home_cs',
'home_gidp',
'home_ci',
'home_lob',
'home_pitchers_used',
'home_individual_er',
'home_team_er',
'home_wp',
'home_bk',
'home_po',
'home_assists',
'home_e',
'home_passed_balls',
'home_double_plays',
'home_triple_plays',
'hp_ump_id',
'hp_ump_name',
'1b_ump_id',
'1b_ump_name',
'2b_ump_id',
'2b_ump_name',
'3b_ump_id',
'3b_ump_name',
'lf_ump_id',
'lf_ump_name',
'rf_ump_id',
'rf_ump_name',
'visit_manager_id',
'visit_manager_name',
'home_manager_id',
'home_manager_name',
'winning_pitcher_id',
'winning_pitcher_name',
'losing_pitcher_id',
'losing_pitcher_name',
'saving_pitcher_id',
'saving_pitcher_name',
'winning_rbi_batter_id',
'winning_rbi_batter_name',
'visit_sp_id',
'visit_sp_name',
'home_sp_id',
'home_sp_name',
'visit_player_1_id',
'visit_player_1_name',
'visit_player_1_pos',
'visit_player_2_id',
'visit_player_2_name',
'visit_player_2_pos',
'visit_player_3_id',
'visit_player_3_name',
'visit_player_3_pos',
'visit_player_4_id',
'visit_player_4_name',
'visit_player_4_pos',
'visit_player_5_id',
'visit_player_5_name',
'visit_player_5_pos',
'visit_player_6_id',
'visit_player_6_name',
'visit_player_6_pos',
'visit_player_7_id',
 'visit_player_7_name',
 'visit_player_7_pos',
 'visit_player_8_id',
 'visit_player_8_name',
 'visit_player_8_pos',
 'visit_player_9_id',
 'visit_player_9_name',
 'visit_player_9_pos',
 'home_player_1_id',
 'home_player_1_name',
 'home_player_1_pos',
 'home_player_2_id',
 'home_player_2_name',
 'home_player_2_pos',
 'home_player_3_id',
 'home_player_3_name',
 'home_player_3_pos',
 'home_player_4_id',
 'home_player_4_name',
 'home_player_4_pos',
 'home_player_5_id',
 'home_player_5_name',
 'home_player_5_pos',
 'home_player_6_id',
 'home_player_6_name',
 'home_player_6_pos',
 'home_player_7_id',
 'home_player_7_name',
 'home_player_7_pos',
 'home_player_8_id',
 'home_player_8_name',
 'home_player_8_pos',
 'home_player_9_id',
 'home_player_9_name',
 'home_player_9_pos',
 'additional_info',
 'acquisition_info']
```

[247]:
```
gl
```

[247]:
```
                      0
0            visit_team
1             home_team
2           visit_score
3            home_score
4      game_length_outs
5             day_night
```

```
6               park_id
7      visit_manager_id
8       home_manager_id
9    winning_pitcher_id
10    losing_pitcher_id
11    saving_pitcher_id
12          visit_sp_id
13           home_sp_id
14    visit_player_1_id
15    visit_player_2_id
16    visit_player_3_id
17    visit_player_4_id
18    visit_player_5_id
19    visit_player_6_id
20    visit_player_7_id
21    visit_player_8_id
22    visit_player_9_id
23     home_player_1_id
24     home_player_2_id
25     home_player_3_id
26     home_player_4_id
27     home_player_5_id
28     home_player_6_id
29     home_player_7_id
30     home_player_8_id
31     home_player_9_id
```

[ ]: