preliminary_clean

December 14, 2022

This is the first jupyter notebook where I'd be making preliminary cleaning of the datasets so as to use for analysis in a later notebook. After the cleaning is done here, I'll then export it to MySQL to make use of `JOINS` to merge different datasets and columns to another.

0.1 Match

[160]:

```
import pandas as pd
match = pd.read_csv('Match.csv')
match.head()
```

[160]:

```
   id  country_id  league_id     season  stage                 date  \
0   1           1          1  2008/2009      1  2008-08-17 00:00:00
1   2           1          1  2008/2009      1  2008-08-16 00:00:00
2   3           1          1  2008/2009      1  2008-08-16 00:00:00
3   4           1          1  2008/2009      1  2008-08-17 00:00:00
4   5           1          1  2008/2009      1  2008-08-16 00:00:00

   match_api_id  home_team_api_id  away_team_api_id  home_team_goal  …  \
0        492473              9987              9993               1  …
1        492474             10000              9994               0  …
2        492475              9984              8635               0  …
3        492476              9991              9998               5  …
4        492477              7947              9985               1  …

    SJA   VCH   VCD   VCA   GBH   GBD   GBA   BSH   BSD   BSA
0  4.00  1.65  3.40  4.50  1.78  3.25  4.00  1.73  3.40  4.20
1  3.80  2.00  3.25  3.25  1.85  3.25  3.75  1.91  3.25  3.60
2  2.50  2.35  3.25  2.65  2.50  3.20  2.50  2.30  3.20  2.75
3  7.50  1.45  3.75  6.50  1.50  3.75  5.50  1.44  3.75  6.50
4  1.73  4.50  3.40  1.65  4.50  3.50  1.65  4.75  3.30  1.67

[5 rows x 115 columns]
```

[161]:

```
match.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25979 entries, 0 to 25978
Columns: 115 entries, id to BSA
```

```
dtypes: float64(96), int64(9), object(10)
memory usage: 22.8+ MB
```

```
[162]: match.dtypes
```

```
[162]: id                int64
       country_id        int64
       league_id         int64
       season           object
       stage             int64
                         …
       GBD             float64
       GBA             float64
       BSH             float64
       BSD             float64
       BSA             float64
       Length: 115, dtype: object
```

Change date datatype

```
[163]: match['date'] = pd.to_datetime(match['date']).dt.date
```

```
[164]: match['date'] = pd.to_datetime(match['date'])
```

```
[165]: match.columns
```

```
[165]: Index(['id', 'country_id', 'league_id', 'season', 'stage', 'date',
              'match_api_id', 'home_team_api_id', 'away_team_api_id',
              'home_team_goal',
              …
              'SJA', 'VCH', 'VCD', 'VCA', 'GBH', 'GBD', 'GBA', 'BSH', 'BSD', 'BSA'],
             dtype='object', length=115)
```

```
[166]: print(match.columns.tolist())
```

```
['id', 'country_id', 'league_id', 'season', 'stage', 'date', 'match_api_id',
'home_team_api_id', 'away_team_api_id', 'home_team_goal', 'away_team_goal',
'home_player_X1', 'home_player_X2', 'home_player_X3', 'home_player_X4',
'home_player_X5', 'home_player_X6', 'home_player_X7', 'home_player_X8',
'home_player_X9', 'home_player_X10', 'home_player_X11', 'away_player_X1',
'away_player_X2', 'away_player_X3', 'away_player_X4', 'away_player_X5',
'away_player_X6', 'away_player_X7', 'away_player_X8', 'away_player_X9',
'away_player_X10', 'away_player_X11', 'home_player_Y1', 'home_player_Y2',
'home_player_Y3', 'home_player_Y4', 'home_player_Y5', 'home_player_Y6',
'home_player_Y7', 'home_player_Y8', 'home_player_Y9', 'home_player_Y10',
'home_player_Y11', 'away_player_Y1', 'away_player_Y2', 'away_player_Y3',
'away_player_Y4', 'away_player_Y5', 'away_player_Y6', 'away_player_Y7',
'away_player_Y8', 'away_player_Y9', 'away_player_Y10', 'away_player_Y11',
'home_player_1', 'home_player_2', 'home_player_3', 'home_player_4',
'home_player_5', 'home_player_6', 'home_player_7', 'home_player_8',
'home_player_9', 'home_player_10', 'home_player_11', 'away_player_1',
'away_player_2', 'away_player_3', 'away_player_4', 'away_player_5',
'away_player_6', 'away_player_7', 'away_player_8', 'away_player_9',
'away_player_10', 'away_player_11', 'goal', 'shoton', 'shotoff', 'foulcommit',
'card', 'cross', 'corner', 'possession', 'B365H', 'B365D', 'B365A', 'BWH',
'BWD', 'BWA', 'IWH', 'IWD', 'IWA', 'LBH', 'LBD', 'LBA', 'PSH', 'PSD', 'PSA',
'WHH', 'WHD', 'WHA', 'SJH', 'SJD', 'SJA', 'VCH', 'VCD', 'VCA', 'GBH', 'GBD',
'GBA', 'BSH', 'BSD', 'BSA']
```

from the `match` dataset, the data below is what I'd be needing for this analysis.

```
[167]: match = match[['id', 'country_id', 'league_id', 'date', 'season', 'stage',␣
        ↪'home_team_api_id', 'away_team_api_id', 'home_team_goal', 'away_team_goal']]
```

```
[168]: match.isnull().sum()
```

```
[168]: id                  0
       country_id          0
       league_id           0
       date                0
       season              0
       stage               0
       home_team_api_id    0
       away_team_api_id    0
       home_team_goal      0
       away_team_goal      0
       dtype: int64
```

```
[171]: match.duplicated().sum()
```

```
[171]: 0
```

```
[172]: match.shape
```

```
[172]: (25979, 10)
```

```
[173]: # drop columns whose name contains specific string from dataframe
       # match[match.columns.drop(list(match.filter(regex='away_')))]
```

```
[174]: match.to_csv('Match.csv', index=False)
```

0.2 Player attributes

```
[70]: play_att = pd.read_csv('Player_Attributes.csv')
      play_att.head()
```

[70]:

```
   id  player_fifa_api_id  player_api_id                 date  overall_rating  \
0   1              218353         505942  2016-02-18 00:00:00            67.0   
1   2              218353         505942  2015-11-19 00:00:00            67.0   
2   3              218353         505942  2015-09-21 00:00:00            62.0   
3   4              218353         505942  2015-03-20 00:00:00            61.0   
4   5              218353         505942  2007-02-22 00:00:00            61.0   

   potential preferred_foot attacking_work_rate defensive_work_rate  crossing  \
0       71.0          right              medium              medium      49.0   
1       71.0          right              medium              medium      49.0   
2       66.0          right              medium              medium      49.0   
3       65.0          right              medium              medium      48.0   
4       65.0          right              medium              medium      48.0   

   …  vision  penalties  marking  standing_tackle  sliding_tackle  \
0  …    54.0       48.0     65.0             69.0            69.0   
1  …    54.0       48.0     65.0             69.0            69.0   
2  …    54.0       48.0     65.0             66.0            69.0   
3  …    53.0       47.0     62.0             63.0            66.0   
4  …    53.0       47.0     62.0             63.0            66.0   

   gk_diving  gk_handling  gk_kicking  gk_positioning  gk_reflexes  
0        6.0         11.0        10.0             8.0          8.0  
1        6.0         11.0        10.0             8.0          8.0  
2        6.0         11.0        10.0             8.0          8.0  
3        5.0         10.0         9.0             7.0          7.0  
4        5.0         10.0         9.0             7.0          7.0  

[5 rows x 42 columns]
```

[71]:

```
play_att.shape
```

[71]:

```
(183978, 42)
```

[72]:

```
play_att.isnull().sum().sum()
```

[72]:

```
47301
```

[73]:

```
play_att.duplicated().sum()
```

[73]:

```
0
```

[74]:

```
play_att.dropna(inplace=True)
```

[76]:

```
play_att.isnull().sum().sum()
```

[76]:

```
0
```

[77]:
```
play_att.shape
```

[77]:
```
(180354, 42)
```

[78]:
```
play_att.columns
```

[78]:
```
Index(['id', 'player_fifa_api_id', 'player_api_id', 'date', 'overall_rating',
       'potential', 'preferred_foot', 'attacking_work_rate',
       'defensive_work_rate', 'crossing', 'finishing', 'heading_accuracy',
       'short_passing', 'volleys', 'dribbling', 'curve', 'free_kick_accuracy',
       'long_passing', 'ball_control', 'acceleration', 'sprint_speed',
       'agility', 'reactions', 'balance', 'shot_power', 'jumping', 'stamina',
       'strength', 'long_shots', 'aggression', 'interceptions', 'positioning',
       'vision', 'penalties', 'marking', 'standing_tackle', 'sliding_tackle',
       'gk_diving', 'gk_handling', 'gk_kicking', 'gk_positioning',
       'gk_reflexes'],
      dtype='object')
```

[79]:
```
play_att.drop(['volleys', 'curve', 'vision', 'standing_tackle', 'gk_reflexes'],␣
 ↪axis=1, inplace=True)
```

Convert date datatype

[81]:
```
play_att['date'] = pd.to_datetime(play_att['date']).dt.date
```

[82]:
```
play_att['date']= pd.to_datetime(play_att['date'])
```

[84]:
```
print(play_att.dtypes)
```

```
id                              int64
player_fifa_api_id              int64
player_api_id                   int64
date                   datetime64[ns]
overall_rating                float64
potential                     float64
preferred_foot                 object
attacking_work_rate            object
defensive_work_rate            object
crossing                      float64
finishing                     float64
heading_accuracy              float64
short_passing                 float64
dribbling                     float64
free_kick_accuracy            float64
long_passing                  float64
ball_control                  float64
acceleration                  float64
sprint_speed                  float64
agility                       float64
```

```
reactions                     float64
balance                       float64
shot_power                    float64
jumping                       float64
stamina                       float64
strength                      float64
long_shots                    float64
aggression                    float64
interceptions                 float64
positioning                   float64
penalties                     float64
marking                       float64
sliding_tackle                float64
gk_diving                     float64
gk_handling                   float64
gk_kicking                    float64
gk_positioning                float64
dtype: object
```

```
[85]: play_att.to_csv('Player_Attributes.csv', index=False)
```

0.3 Player

```
[57]: play = pd.read_csv('Player.csv')
      play.head()
```

```
[57]:    id  player_api_id         player_name  player_fifa_api_id  \
      0   1         505942  Aaron Appindangoye              218353
      1   2         155782     Aaron Cresswell              189615
      2   3         162549         Aaron Doran              186170
      3   4          30572       Aaron Galindo              140161
      4   5          23780        Aaron Hughes               17725

                    birthday  height  weight
      0  1992-02-29 00:00:00  182.88     187
      1  1989-12-15 00:00:00  170.18     146
      2  1991-05-13 00:00:00  170.18     163
      3  1982-05-08 00:00:00  182.88     198
      4  1979-11-08 00:00:00  182.88     154
```

Convert birthday from datetime to date

```
[59]: play['birthday'] = pd.to_datetime(play['birthday']).dt.date
```

```
[62]: play['birthday'] = pd.to_datetime(play['birthday'])
```

```
[63]: play.dtypes
```

```
[63]: id                              int64
      player_api_id                   int64
      player_name                    object
      player_fifa_api_id              int64
      birthday               datetime64[ns]
      height                        float64
      weight                          int64
      dtype: object
```

```
[65]: play.isnull().sum()
```

```
[65]: id                    0
      player_api_id         0
      player_name           0
      player_fifa_api_id    0
      birthday              0
      height                0
      weight                0
      dtype: int64
```

```
[67]: play.duplicated().sum()
```

```
[67]: 0
```

```
[68]: play.shape
```

```
[68]: (11060, 7)
```

```
[69]: play.to_csv('Player.csv', index=False)
```

0.4 Team Attributes

```
[86]: team_att = pd.read_csv('Team_Attributes.csv')
      team_att.head()
```

```
[86]:    id  team_fifa_api_id  team_api_id                 date  buildUpPlaySpeed  \
      0   1               434         9930  2010-02-22 00:00:00                60
      1   2               434         9930  2014-09-19 00:00:00                52
      2   3               434         9930  2015-09-10 00:00:00                47
      3   4                77         8485  2010-02-22 00:00:00                70
      4   5                77         8485  2011-02-22 00:00:00                47

        buildUpPlaySpeedClass  buildUpPlayDribbling buildUpPlayDribblingClass  \
      0              Balanced                   NaN                    Little
      1              Balanced                  48.0                    Normal
      2              Balanced                  41.0                    Normal
      3                  Fast                   NaN                    Little
      4              Balanced                   NaN                    Little
```

```
   buildUpPlayPassing buildUpPlayPassingClass  … chanceCreationShooting  \
0                  50                   Mixed  …                     55
1                  56                   Mixed  …                     64
2                  54                   Mixed  …                     64
3                  70                    Long  …                     70
4                  52                   Mixed  …                     52

   chanceCreationShootingClass chanceCreationPositioningClass  \
0                       Normal                      Organised
1                       Normal                      Organised
2                       Normal                      Organised
3                         Lots                      Organised
4                       Normal                      Organised

   defencePressure defencePressureClass  defenceAggression  \
0               50               Medium                 55
1               47               Medium                 44
2               47               Medium                 44
3               60               Medium                 70
4               47               Medium                 47

  defenceAggressionClass defenceTeamWidth  defenceTeamWidthClass  \
0                  Press               45                 Normal
1                  Press               54                 Normal
2                  Press               54                 Normal
3                 Double               70                   Wide
4                  Press               52                 Normal

  defenceDefenderLineClass
0                    Cover
1                    Cover
2                    Cover
3                    Cover
4                    Cover

[5 rows x 25 columns]
```

Change date datatype

[87]:
```
team_att['date'] = pd.to_datetime(team_att['date']).dt.date
```

[88]:
```
team_att['date'] = pd.to_datetime(team_att['date'])
```

[89]:
```
team_att.dtypes
```

[89]:
```
id                                         int64
team_fifa_api_id                           int64
```

```
team_api_id                                int64
date                              datetime64[ns]
buildUpPlaySpeed                           int64
buildUpPlaySpeedClass                     object
buildUpPlayDribbling                     float64
buildUpPlayDribblingClass                 object
buildUpPlayPassing                         int64
buildUpPlayPassingClass                   object
buildUpPlayPositioningClass               object
chanceCreationPassing                      int64
chanceCreationPassingClass                object
chanceCreationCrossing                     int64
chanceCreationCrossingClass               object
chanceCreationShooting                     int64
chanceCreationShootingClass               object
chanceCreationPositioningClass            object
defencePressure                            int64
defencePressureClass                      object
defenceAggression                          int64
defenceAggressionClass                    object
defenceTeamWidth                           int64
defenceTeamWidthClass                     object
defenceDefenderLineClass                  object
dtype: object
```

[90]:
```
team_att.isnull().sum().sum()
```

[90]:
```
969
```

[91]:
```
team_att.duplicated().sum()
```

[91]:
```
0
```

[92]:
```
team_att.dropna(inplace=True)
```

[93]:
```
team_att.shape
```

[93]:
```
(489, 25)
```

[94]:
```
team_att.to_csv('Team_Attributes.csv', index=False)
```

0.5 Team

[95]:
```
team = pd.read_csv('Team.csv')
team.head()
```

[95]:
```
   id  team_api_id  team_fifa_api_id     team_long_name team_short_name
0   1         9987             673.0          KRC Genk             GEN
```

```
1   2         9993            675.0        Beerschot AC               BAC
2   3        10000          15005.0    SV Zulte-Waregem               ZUL
3   4         9994           2007.0    Sporting Lokeren               LOK
4   5         9984           1750.0  KSV Cercle Brugge               CEB
```

```
[96]: team.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                299 non-null    int64
 1   team_api_id       299 non-null    int64
 2   team_fifa_api_id  288 non-null    float64
 3   team_long_name    299 non-null    object
 4   team_short_name   299 non-null    object
dtypes: float64(1), int64(2), object(2)
memory usage: 11.8+ KB
```

```
[97]: team.isnull().sum().sum()
```

```
[97]: 11
```

```
[98]: team.duplicated().sum()
```

```
[98]: 0
```

```
[99]: team.dropna(inplace=True)
```

```
[100]: team.to_csv('Team.csv', index=False)
```

```
[ ]:
```