```
import pandas as pd
import numpy as np

data20 = pd.read_csv("2020.csv")

data = pd.read_csv("2021.csv")
```

```
players21 = set(data.name)
players20 = set(data20.name)
players20 = players20.intersection(players21)
data20 = data20[data20['name'].isin(players20)]
data = data[data['name'].isin(players20)]
```

```
data20['total_points_20'] = data20['total_points']
data20['GW_20'] = data20['GW']
data20['name_20'] = data20['name']
data20 = data20[['name_20','GW_20','total_points_20']]
cols = ['name','GW']
cols20 = ['name_20','GW_20']
data20.set_index(cols20)
data.set_index(cols)
data = data.join(data20)
data = data[(data['name']==data['name_20']) & (data['GW']==data['GW_20'])]
```

```
team_names = sorted(list(set(list((data.team)))))
team_ids = np.arange(1,21)
data['opponent_team'] = data['opponent_team'].map(dict(zip(team_ids, team_names)))
```

```
#in previous year
goals_conceded_by_team = {'Arsenal': 39,'Aston Villa': 46,'Brighton': 46,'Burnley': 55,
                          'Chelsea': 36,'Crystal Palace': 66,'Everton': 48,'Fulham': 55,
                          'Leeds': 54,'Leicester': 50,'Liverpool': 42,'Man City': 32,
                          'Man Utd': 44,'Newcastle': 62,'Sheffield Utd': 63,'Southampton': 68,
                          'Spurs': 45,'West Brom': 76,'West Ham': 47,'Wolves': 52}

goals_scored= {'Man City':83,'Man Utd':73,'Leicester':68,
               'Liverpool':68,'Spurs':68,'Leeds':62,'West Ham':62,
               'Chelsea':58,'Arsenal':55,'Aston Villa':55,'Southampton':48,
               'Everton':47,'Newcastle':46,'Brighton':41,'Crystal Palace':41,
               'Wolves':36,'West Brom':35,'Burnley':33,'Fulham':27,'Sheffield Utd':20}

team_wins = {'Man City': 27,'Man Utd': 21, 'Leicester': 20,'Liverpool': 20,
             'Chelsea': 19,'West Ham': 19,'Arsenal': 18,'Leeds': 18,'Spurs': 18,
             'Everton': 17,'Aston Villa': 16,'Crystal Palace': 12,'Newcastle': 12,
             'Southampton': 12,'Wolves': 12,'Burnley': 10,'Brighton': 9,
             'Sheffield Utd': 7,'Fulham': 5,'West Brom': 5}
```

```
#we measure goals_conceded as an inverse indicator of defense strength
data['opp_defense_rank'] = data['opponent_team'].map(goals_conceded_by_team)
data.sort_values(by =['opp_defense_rank'], inplace = True)
data['opp_defense_rank']= pd.qcut(data['opp_defense_rank'],q = 4, labels = False)

#Goals scored are an attribute of team strength
data['opp_attack_rank'] = data['opponent_team'].map(goals_scored)
data.sort_values(by =['opp_attack_rank'], inplace = True)
data['opp_attack_rank']= pd.qcut(data['opp_attack_rank'],q = 4, labels = False)

#Final rankings naturally give an idea of the overall team strength/quality

data['team_cluster_rank'] = data['opponent_team'].map(goals_scored)
data.sort_values(by =['team_cluster_rank'], inplace = True)
data['team_cluster_rank']= pd.qcut(data['team_cluster_rank'],q = 4, labels = False)

data['opp_cluster_rank'] = data['opponent_team'].map(goals_scored)
data.sort_values(by =['opp_cluster_rank'], inplace = True)
data['opp_cluster_rank']= pd.qcut(data['opp_cluster_rank'],q = 4, labels = False)
```

```
data['opponent_team'] = data['opponent_team'].map(dict(zip(team_names, team_ids)))
data['team'] = data['team'].map(dict(zip(team_names, team_ids)))
```

```
initval = data[data['GW']==1][["name","value"]]
data['initval'] = data['name'].map(dict(zip(initval.name, initval.value)))
```

```
data['pos_id'] = data['position']
data = pd.get_dummies(data,columns=['position'])
pos_ids = np.array([k for k in data['pos_id'].unique()])
data['pos_id']=data['pos_id'].apply(lambda x : np.where(x == pos_ids)[0][0])
```

```
data['game_avg_7']= data.groupby(['name']
                                 )['total_points_20'].rolling(7).mean().reset_index(0,drop=True)

data.sort_values(by =['game_avg_7'], inplace = True)
data['rank']= pd.qcut(data['game_avg_7'],q = 4, labels = False)
data['diff_from_avg'] = data['total_points'] - data['game_avg_7']
```

```
pts=data.groupby(['name']).sum()
pts = pts[pts['total_points']>50]
chosen_players = pts.index
data = data[data['name'].isin(chosen_players)]
```

```
data['was_homee']=data['was_home']
data = pd.get_dummies(data,columns=['was_home'])
data['was_home']=data['was_homee']
data = data.dropna()
```

```
data = data[['name', 'total_points','position_DEF','GW','team',
             'opponent_team','position_FWD', 'position_GK',
             'position_MID', 'game_avg_7' ,'pos_id',
             'initval','rank','diff_from_avg',
            'opp_defense_rank','opp_attack_rank','team_cluster_rank',
            'opp_cluster_rank','was_home','total_points_20']]
data.sort_values(by=['GW'],inplace=True)
data.to_csv('data.csv',index=False)
```