Clustering-Analysis-in-Soccer-teams-by-playing-style

June 6, 2018

In [1]:
```
# Original Source: https://www.kaggle.com/teamaker/clustering-teams-based-on-style-of-pl
# Dataset https://www.kaggle.com/teamaker/clustering-teams-based-on-style-of-play/data
import pandas as pd
import numpy as np
import sqlite3
```

In [2]:
```
con = sqlite3.connect("./database.sqlite")
team=pd.read_sql_query('select * from Team', con)
team_attr=pd.read_sql_query('select * from Team_Attributes', con)
con.close()

df = pd.merge(team, team_attr, how='inner', left_on='team_api_id', right_on='team_api_id
print (df.shape)
```

```
(1458, 29)
```

In [3]:
```
df.sample(n=10)
```

Out[3]:

	id_x	team_api_id	team_fifa_api_id_x	team_long_name	\
1073	35290	10264	1896.0	SC Braga	
1212	43035	10267	461.0	Valencia CF	
1051	35285	9768	237.0	Sporting CP	
496	15617	9823	21.0	FC Bayern Munich	
576	15631	9788	23.0	Borussia Mönchengladbach	
1176	39395	9925	78.0	Celtic	
202	3474	9879	144.0	Fulham	
472	12595	10242	294.0	ES Troyes AC	
1302	43050	10281	462.0	Real Valladolid	
1113	35774	2033	111540.0	S.C. Olhanense	

	team_short_name	id_y	team_fifa_api_id_y	date	\
1073	BRA	233	1896	2012-02-22 00:00:00	
1212	VAL	1311	461	2013-09-20 00:00:00	
1051	SCP	1197	237	2014-09-19 00:00:00	
496	BMU	147	21	2014-09-19 00:00:00	
576	GLA	220	23	2014-09-19 00:00:00	
1176	CEL	291	78	2012-02-22 00:00:00	

```
202              FUL   482                 144  2013-09-20 00:00:00
472              TRO  1291                 294  2014-09-19 00:00:00
1302             VAL  1324                 462  2014-09-19 00:00:00
1113             OLH   928              111540  2010-02-22 00:00:00

      buildUpPlaySpeed buildUpPlaySpeedClass            ...             \
1073                53             Balanced            ...
1212                20                 Slow            ...
1051                55             Balanced            ...
496                 37             Balanced            ...
576                 31                 Slow            ...
1176                70                 Fast            ...
202                 52             Balanced            ...
472                 49             Balanced            ...
1302                55             Balanced            ...
1113                45             Balanced            ...

      chanceCreationShooting chanceCreationShootingClass  \
1073                      52                      Normal
1212                      55                      Normal
1051                      55                      Normal
496                       40                      Normal
576                       49                      Normal
1176                      60                      Normal
202                       59                      Normal
472                       46                      Normal
1302                      41                      Normal
1113                      45                      Normal

      chanceCreationPositioningClass defencePressure defencePressureClass  \
1073                      Organised              23                  Deep
1212                      Organised              51                Medium
1051                      Organised              60                Medium
496                       Free Form              61                Medium
576                       Free Form              43                Medium
1176                      Organised              50                Medium
202                       Organised              39                Medium
472                       Organised              43                Medium
1302                      Organised              51                Medium
1113                      Free Form              50                Medium

      defenceAggression defenceAggressionClass  defenceTeamWidth  \
1073                 48                  Press                39
1212                 38                  Press                59
1051                 60                  Press                54
496                  59                  Press                40
576                  50                  Press                41
1176                 50                  Press                60
```

```
202                  39                  Press                  57
472                  42                  Press                  52
1302                 50                  Press                  60
1113                 45                  Press                  60

     defenceTeamWidthClass  defenceDefenderLineClass
1073                Normal                     Cover
1212                Normal                     Cover
1051                Normal                     Cover
496                 Normal                     Cover
576                 Normal                     Cover
1176                Normal                     Cover
202                 Normal                     Cover
472                 Normal                     Cover
1302                Normal                     Cover
1113                Normal                     Cover

[10 rows x 29 columns]
```

```
In [4]: cols_to_keep = ['date', 'team_long_name', u'buildUpPlaySpeed', u'buildUpPlayDribbling',
                u'buildUpPlayPassing', u'chanceCreationPassing', u'chanceCreationCrossing',
              u'chanceCreationShooting', u'defencePressure', u'defenceAggression', u'defenceTea
        df = df[cols_to_keep]
```

```
In [5]: old_df = df.copy(deep=True)
```

```
In [6]: aggs = df.groupby('team_long_name')['date'].max().to_frame()
        df.drop('date', axis=1, inplace=True)
        df.drop_duplicates(subset='team_long_name', keep='last', inplace=True)
        df = df.merge(right=aggs, right_index=True, left_on='team_long_name', how='right')
        df = df.dropna()
        df.set_index('team_long_name', inplace=True)
        df.drop('date', axis=1, inplace=True)
        print (df.shape)
```

```
(260, 9)
```

```
In [7]: index = old_df.team_long_name[old_df.team_long_name.str.contains(pat = 'Manches')].index
        for i in index:
            print(old_df.iloc[[i]])
```

```
                  date     team_long_name  buildUpPlaySpeed  \
97  2010-02-22 00:00:00  Manchester United                70

    buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
97                   NaN                  45                     45

    chanceCreationCrossing  chanceCreationShooting  defencePressure  \
```

```
97                    70                      65              40

    defenceAggression  defenceTeamWidth
97                 50               40
                   date     team_long_name  buildUpPlaySpeed  \
98  2011-02-22 00:00:00  Manchester United                65

    buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
98                   NaN                  40                     65

    chanceCreationCrossing  chanceCreationShooting  defencePressure  \
98                      65                      70               45

    defenceAggression  defenceTeamWidth
98                 45               65
                   date     team_long_name  buildUpPlaySpeed  \
99  2012-02-22 00:00:00  Manchester United                46

    buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
99                   NaN                  54                     46

    chanceCreationCrossing  chanceCreationShooting  defencePressure  \
99                      60                      55               40

    defenceAggression  defenceTeamWidth
99                 50               56
                    date     team_long_name  buildUpPlaySpeed  \
100  2013-09-20 00:00:00  Manchester United                46

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
100                   NaN                  38                     46

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
100                      68                      37               49

     defenceAggression  defenceTeamWidth
100                49               56
                    date     team_long_name  buildUpPlaySpeed  \
101  2014-09-19 00:00:00  Manchester United                46

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
101                  34.0                  54                     49

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
101                      72                      56               42

     defenceAggression  defenceTeamWidth
101                41               56
```

```
                  date     team_long_name  buildUpPlaySpeed  \
102  2015-09-10 00:00:00  Manchester United                38

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
102                  42.0                  44                     49

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
102                      44                      40               54

     defenceAggression  defenceTeamWidth
102                 53                56
                  date   team_long_name  buildUpPlaySpeed  \
151  2010-02-22 00:00:00  Manchester City                70

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
151                   NaN                  60                     55

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
151                      70                      70               45

     defenceAggression  defenceTeamWidth
151                 55                45
                  date   team_long_name  buildUpPlaySpeed  \
152  2011-02-22 00:00:00  Manchester City                70

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
152                   NaN                  60                     70

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
152                      65                      75               65

     defenceAggression  defenceTeamWidth
152                 65                50
                  date   team_long_name  buildUpPlaySpeed  \
153  2012-02-22 00:00:00  Manchester City                45

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
153                   NaN                  55                     40

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
153                      38                      48               50

     defenceAggression  defenceTeamWidth
153                 60                54
                  date   team_long_name  buildUpPlaySpeed  \
154  2013-09-20 00:00:00  Manchester City                36

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
```

```
154                   NaN                   34                       39

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
154                      44                      38               42

     defenceAggression  defenceTeamWidth
154                 44                54
                    date   team_long_name  buildUpPlaySpeed  \
155  2014-09-19 00:00:00  Manchester City                59

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
155                  35.0                  29                     32

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
155                      49                      62               42

     defenceAggression  defenceTeamWidth
155                 44                54
                    date   team_long_name  buildUpPlaySpeed  \
156  2015-09-10 00:00:00  Manchester City                59

     buildUpPlayDribbling  buildUpPlayPassing  chanceCreationPassing  \
156                  35.0                  29                     38

     chanceCreationCrossing  chanceCreationShooting  defencePressure  \
156                      36                      24               48

     defenceAggression  defenceTeamWidth
156                 47                54
```

0.1 PCA

```
In [8]: from sklearn.preprocessing import StandardScaler
        from sklearn.decomposition import PCA

        x = df.copy()
        x_normalized = StandardScaler().fit(x).transform(x)

        pca = PCA(n_components = 3).fit(x_normalized)
        print (pca.explained_variance_ratio_)
```

```
[0.19677852 0.1868464  0.12810027]
```

```
In [9]: np.sum([pca.explained_variance_ratio_])
```

```
Out[9]: 0.5117251758763541
```

We can improve this!

0.2 Multi Dimensional Scaling (MDS).

```
In [10]: from sklearn.manifold import MDS

         mds = MDS(n_components = 2, n_init = 10)
         mds_2 = MDS(n_components = 3, n_init = 10)
         x_mds = mds.fit_transform(x_normalized)
         x_mds_2 = mds_2.fit_transform(x_normalized)

In [11]: team_names = ['Liverpool', 'ES Troyes AC', 'Real Madrid CF', 'Arsenal', 'Swansea City',
         team_maps = {}
         for team in team_names:
             print (team, df.index.get_loc(team))
             team_maps[team] = df.index.get_loc(team)

Liverpool 23
ES Troyes AC 79
Real Madrid CF 222
Arsenal 20
Swansea City 42
Manchester United 18


In [12]: def diff(a,b):
             sum = 0
             for i in range(len(a)):
                 sum += (a[i]-b[i])*(a[i]-b[i])
             return sum

In [13]: print ("Before MDS: ")
         i=-1
         for team in team_names:
             i+=1
             for j in range(i+1,4):
                 print ("{} and {} diff : {}".format(team, team_names[j],
                                                     diff(df.iloc[team_maps[team]].tolist(),
                                                     df.iloc[team_maps[team_names[j]]].tolist())

Before MDS:
Liverpool and ES Troyes AC diff : 1263.0
Liverpool and Real Madrid CF diff : 1402.0
Liverpool and Arsenal diff : 636.0
ES Troyes AC and Real Madrid CF diff : 1067.0
ES Troyes AC and Arsenal diff : 1219.0
Real Madrid CF and Arsenal diff : 2138.0


In [14]: print ("After MDS (keeping 3 components) : ")
         i=-1
```

```
        for team in team_names:
            i+=1
            for j in range(i+1,4):
                print ("{} and {} diff : {}".format(team, team_names[j],
                                                    diff(x_mds_2[team_maps[team]], x_mds_2[team_
```

```
After MDS (keeping 3 components) :
Liverpool and ES Troyes AC diff : 13.68006081989818
Liverpool and Real Madrid CF diff : 12.116927780573683
Liverpool and Arsenal diff : 3.369952164783861
ES Troyes AC and Real Madrid CF diff : 15.059073559574559
ES Troyes AC and Arsenal diff : 10.61256949086824
Real Madrid CF and Arsenal diff : 20.453749900236172
```

```
In [15]: from matplotlib.colors import ListedColormap, BoundaryNorm
         import numpy
         import matplotlib.pyplot as plt
         import matplotlib.patches as mpatches
         from mpl_toolkits.mplot3d import Axes3D

         plt.rcParams['figure.figsize'] = (30,30)

         def plot_labelled_scatter_3D(X, y, class_labels, team_maps):
             num_labels = len(class_labels)

             x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
             y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
             z_min, z_max = X[:, 2].min() - 1, X[:, 2].max() + 1
             marker_array = ['o', '^', '*']
             color_array = ['#FFFF00', '#00AAFF', '#000000', '#FF00AA']
             cmap_bold = ListedColormap(color_array)
             bnorm = BoundaryNorm(numpy.arange(0, num_labels + 1, 1), ncolors=num_labels)
             fig = plt.figure()
             ax = fig.add_subplot(111, projection='3d')

             ax.scatter(X[:, 0], X[:, 1], X[:,2], s=65, c=y, cmap=cmap_bold, norm = bnorm, alpha
             for i, x, y, z in zip(range(len(X[:,0])), X[:, 0], X[:, 1], X[:, 2]):
                 if i in team_maps.values():
                     for team in team_maps.items():
                         if team[1] == i:
                             team_name = team[0]

                     ax.text(x, y, z, team_name)

             plt.xlim(x_min, x_max)
             plt.ylim(y_min, y_max)
             ax.set_zlim(z_min, z_max)
```

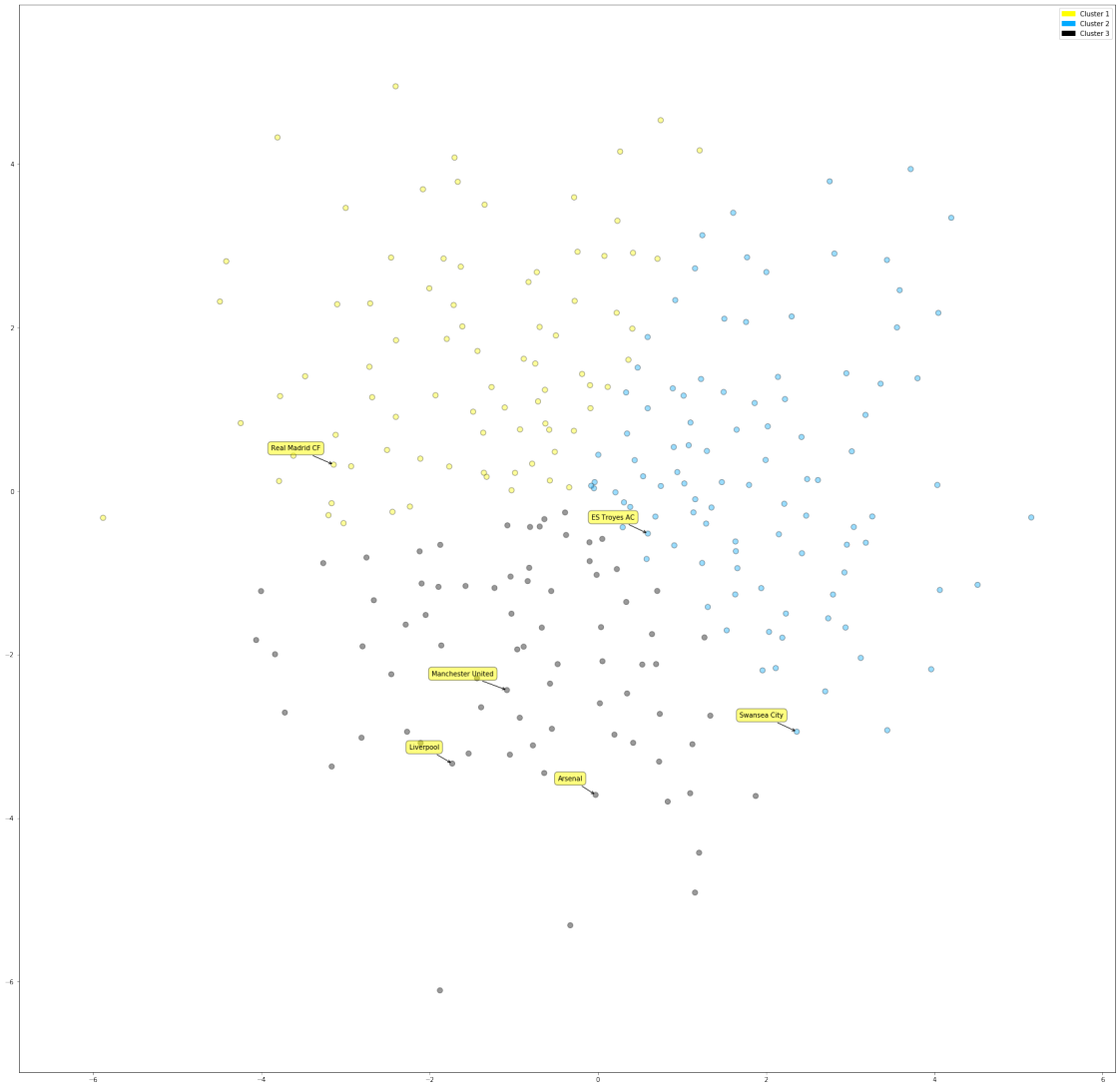
```
    h = []
    for c in range(0, num_labels):
        h.append(mpatches.Patch(color=color_array[c], label=class_labels[c]))
    plt.legend(handles=h)

    plt.show()
```

In [16]:
```
from matplotlib.colors import ListedColormap, BoundaryNorm
import numpy
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
from mpl_toolkits.mplot3d import Axes3D

plt.rcParams['figure.figsize'] = (30,30)

def plot_labelled_scatter_3D(X, y, class_labels, team_maps):
    num_labels = len(class_labels)

    x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
    y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
    z_min, z_max = X[:, 2].min() - 1, X[:, 2].max() + 1
    marker_array = ['o', '^', '*']
    color_array = ['#FFFF00', '#00AAFF', '#000000', '#FF00AA']
    cmap_bold = ListedColormap(color_array)
    bnorm = BoundaryNorm(numpy.arange(0, num_labels + 1, 1), ncolors=num_labels)
    fig = plt.figure()
    ax = fig.add_subplot(111, projection='3d')

    ax.scatter(X[:, 0], X[:, 1], X[:,2], s=65, c=y, cmap=cmap_bold, norm = bnorm, alpha
    for i, x, y, z in zip(range(len(X[:,0])), X[:, 0], X[:, 1], X[:, 2]):
        if i in team_maps.values():
            for team in team_maps.items():
                if team[1] == i:
                    team_name = team[0]

            ax.text(x, y, z, team_name)

    plt.xlim(x_min, x_max)
    plt.ylim(y_min, y_max)
    ax.set_zlim(z_min, z_max)

    h = []
    for c in range(0, num_labels):
        h.append(mpatches.Patch(color=color_array[c], label=class_labels[c]))
    plt.legend(handles=h)

    plt.show()
```

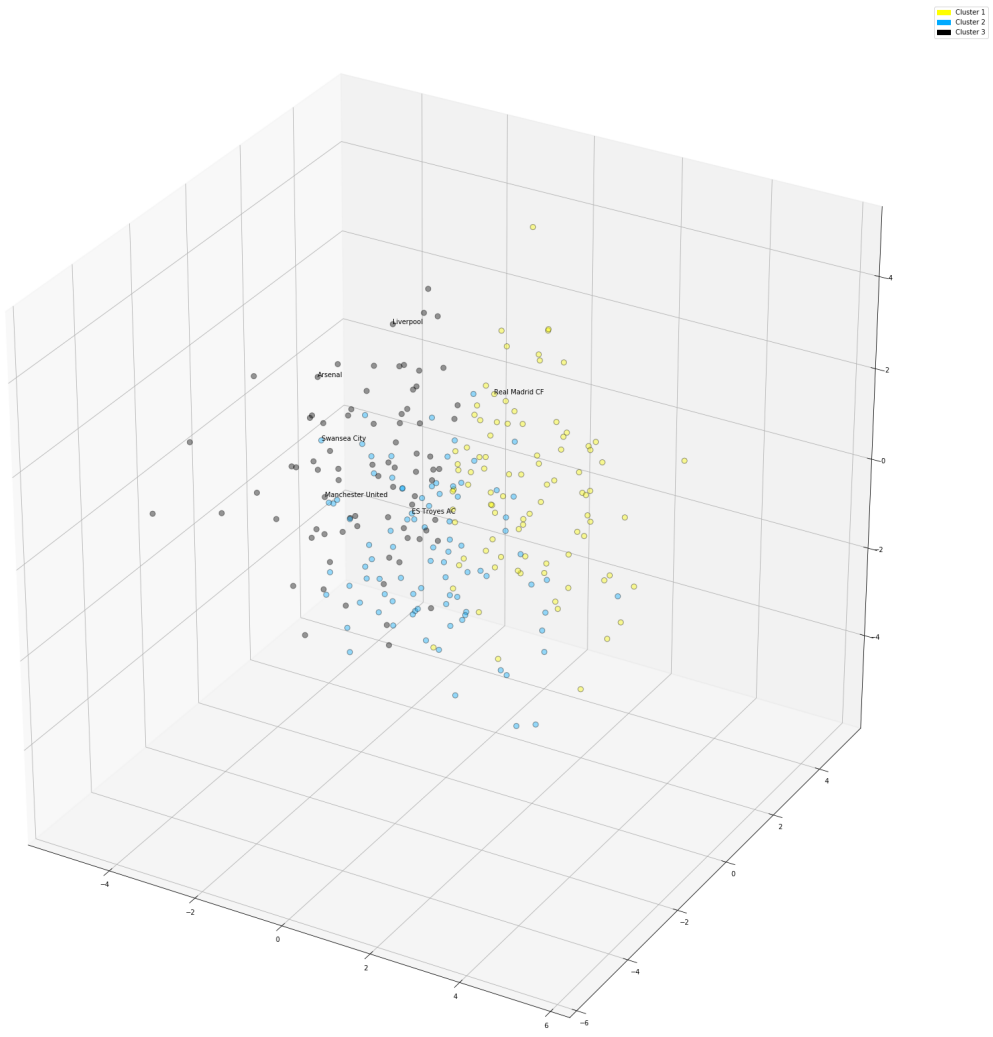
```
def plot_labelled_scatter(X, y, class_labels, team_maps):
    num_labels = len(class_labels)

    x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
    y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1

    marker_array = ['o', '^', '*']
    color_array = ['#FFFF00', '#00AAFF', '#000000', '#FF00AA']
    cmap_bold = ListedColormap(color_array)
    bnorm = BoundaryNorm(numpy.arange(0, num_labels + 1, 1), ncolors=num_labels)
    plt.figure()

    plt.scatter(X[:, 0], X[:, 1], s=65, c=y, cmap=cmap_bold, norm = bnorm, alpha = 0.40
    for i, x, y in zip(range(len(X[:,0])), X[:, 0], X[:, 1]):
        if i in team_maps.values():
            for team in team_maps.items():
                if team[1] == i:
                    team_name = team[0]

            plt.annotate(
                team_name,
                xy=(x, y), xytext=(-20, 20),
                textcoords='offset points', ha='right', va='bottom',
                bbox=dict(boxstyle='round,pad=0.5', fc='yellow', alpha=0.5),
                arrowprops=dict(arrowstyle = '->', connectionstyle='arc3,rad=0'))

    plt.xlim(x_min, x_max)
    plt.ylim(y_min, y_max)

    h = []
    for c in range(0, num_labels):
        h.append(mpatches.Patch(color=color_array[c], label=class_labels[c]))
    plt.legend(handles=h)

    plt.show()
```

1 Kmeans

In [17]:
```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler

kmeans = KMeans(n_clusters = 3, random_state=10)
kmeans.fit(x_mds)
plot_labelled_scatter(x_mds, kmeans.labels_, ['Cluster 1', 'Cluster 2', 'Cluster 3'], t
```




```
In [24]: kmeans = KMeans(n_clusters = 3, random_state=10)
         kmeans.fit(x_mds_2)
         plot_labelled_scatter_3D(x_mds_2, kmeans.labels_, ['Cluster 1', 'Cluster 2', 'Cluster 3
```




```
In [18]: old_df.iloc[[246, 114, 1244, 132, 394]]
```

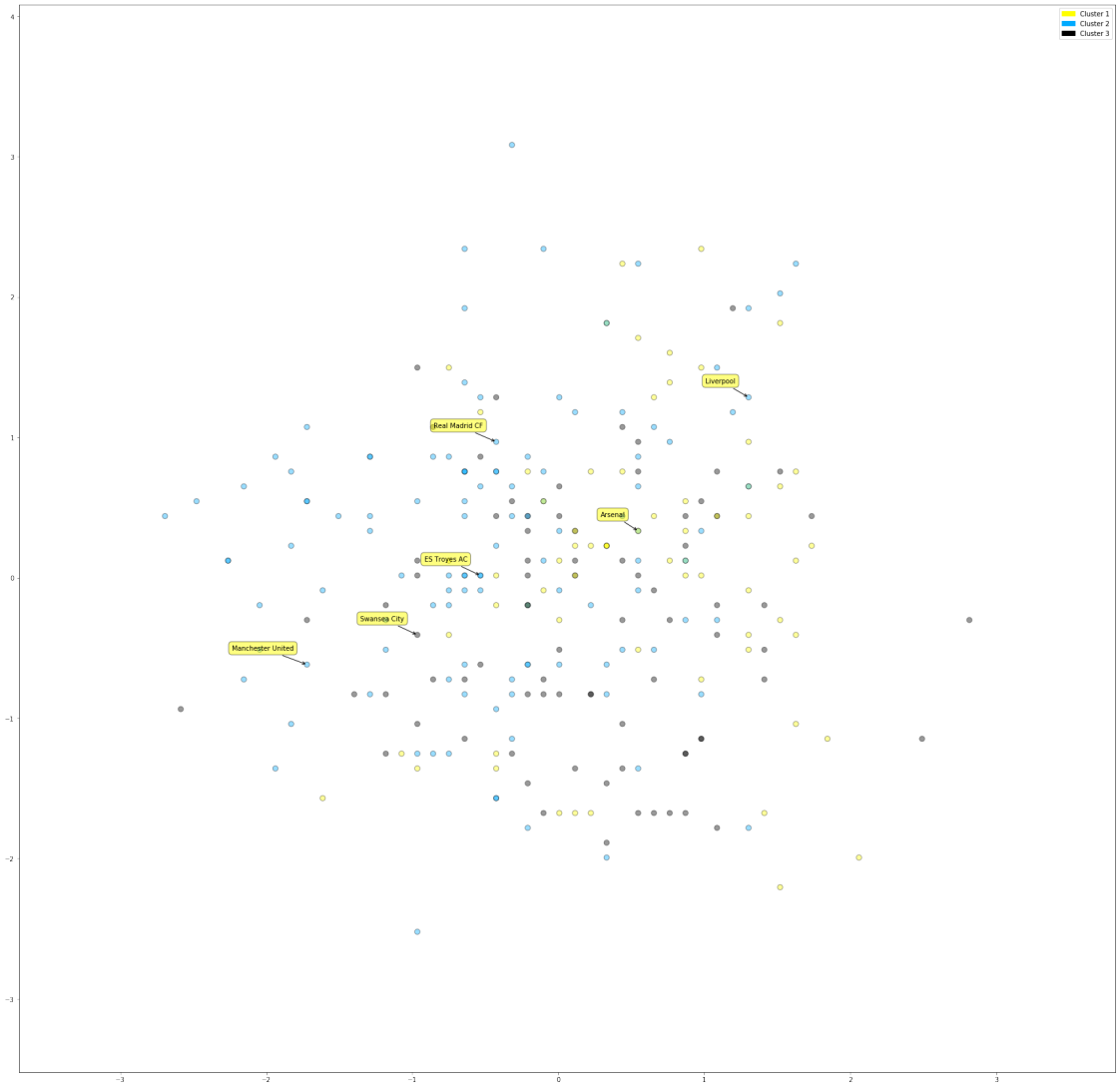
Out[18]:

	date	team_long_name	buildUpPlaySpeed	buildUpPlayDribbling	buildUpPlayPassing	chanceCreationPassing
246	2015-09-10 00:00:00	Swansea City	45	44.0	42	34
114	2015-09-10 00:00:00	Arsenal	59	51.0	30	28
1244	2015-09-10 00:00:00	Real Madrid CF	50	57.0	46	61
132	2015-09-10 00:00:00	Liverpool	66	60.0	45	34
394	2015-09-10 00:00:00	Olympique de Marseille	51			

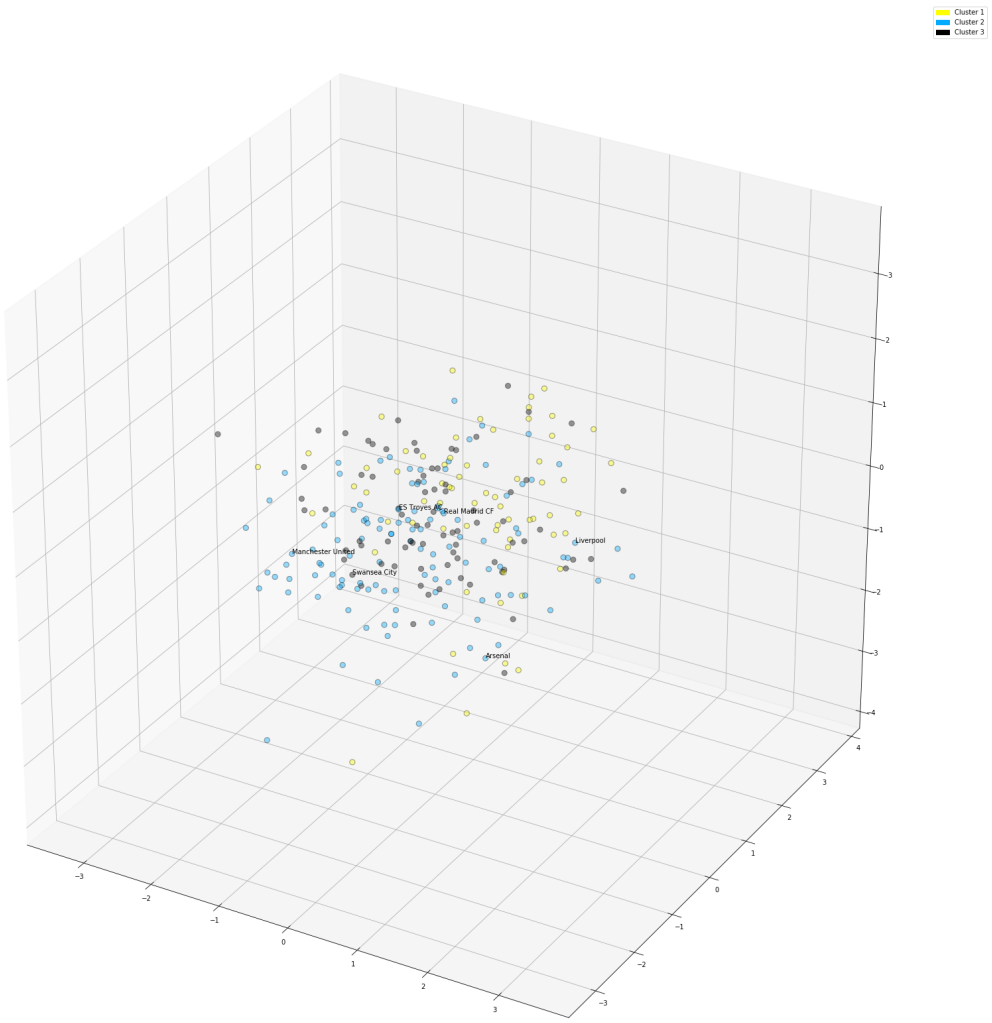
```
394                    77.0                  41                       49

      chanceCreationCrossing  chanceCreationShooting  defencePressure  \
246                       36                      55               31
114                       44                      46               51
1244                      41                      63               52
132                       34                      46               51
394                       69                      31               48

      defenceAggression  defenceTeamWidth
246                  47                42
114                  44                52
1244                 60                63
132                  52                61
394                  48                60
```

2 Spectral Clustering

```
In [51]: from sklearn import cluster

         spectral = cluster.SpectralClustering(n_clusters=3, affinity='nearest_neighbors')
         spectral.fit(x_normalized)
         plot_labelled_scatter(x_normalized, spectral.labels_, ['Cluster 1', 'Cluster 2', 'Clust
```




```
In [50]: plot_labelled_scatter_3D(x_normalized, spectral.labels_, ['Cluster 1', 'Cluster 2', 'Cl
```


Cluster 1
Cluster 2
Cluster 3
ES Troyes AC
Real Madrid CF
Liverpool
Manchester United
Swansea City
Arsenal