Abdul Basit_DSC540_Project_Milestone_5

File: Abdul Basit_DSC540_Project_Milestone_5.ipynb Name: Abdul Basit Date: 05/30/2020 Course: DSC 540 Data Preparation Title: Club Soccer Prediction Analysis

Club Soccer Prediction Analysis

Merging the Data and Storing in a Database/Visualizing Data

Create 5 visualizations that demonstrate the data you have cleansed. You should have at least 2 visualizations that have data from more than one source.

In [1]:
```
from __future__ import print_function, division
from bs4 import BeautifulSoup
from matplotlib import pyplot as plt
from sklearn import preprocessing

import seaborn as sns
import statsmodels.formula.api as smf
import pandas as pd
import plotly.offline as py
import plotly.graph_objs as go
import string
import os
import numpy as np
import sys
import math
import scipy.stats
import density
import random
import hypothesis
import scatter
import requests
import re, string
import random
import sklearn
import twitter
import urllib.request, urllib.parse, urllib.error
import json
import copy
import io
import datetime

from pandas import DataFrame
from urllib.error import HTTPError,URLError
from pandas import DataFrame
from urllib.request import urlopen as uReq
from requests import get
```

```
from prettytable import PrettyTable

import sqlite3

!apt-get update
!apt-get install -y default-jdk
!pip install tabula-py xlrd lxml
!pip install Textblob
!pip install -r requirements.txt
!pip install virtualenv
!apt-get update
!apt-get install -y default-jdk
!pip install tabula-py xlrd lxml
!pip install plotly
!pip install cufflinks
!pip install PTable

py.init_notebook_mode(connected=True)
```

```
C:\Users\basiab1\AppData\Local\Continuum\anaconda3\lib\site-packages\stats
models\tools\_testing.py:19: FutureWarning: pandas.util.testing is depreca
ted. Use the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
'apt-get' is not recognized as an internal or external command,
operable program or batch file.
'apt-get' is not recognized as an internal or external command,
operable program or batch file.
```

```
Requirement already satisfied: tabula-py in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (2.1.0)
Requirement already satisfied: xlrd in c:\users\basiab1\appdata\local\cont
inuum\anaconda3\lib\site-packages (1.1.0)
Requirement already satisfied: lxml in c:\users\basiab1\appdata\local\cont
inuum\anaconda3\lib\site-packages (4.2.5)
Requirement already satisfied: distro in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (from tabula-py) (1.5.0)
Requirement already satisfied: pandas>=0.25.3 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from tabula-py) (1.0.3)
Requirement already satisfied: numpy in c:\users\basiab1\appdata\local\con
tinuum\anaconda3\lib\site-packages (from tabula-py) (1.15.4)
Requirement already satisfied: python-dateutil>=2.6.1 in c:\users\basiab1\
appdata\local\continuum\anaconda3\lib\site-packages (from pandas>=0.25.3->
tabula-py) (2.7.5)
Requirement already satisfied: pytz>=2017.2 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from pandas>=0.25.3->tabula-py)
 (2018.7)
Requirement already satisfied: six>=1.5 in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (from python-dateutil>=2.6.1->pandas
>=0.25.3->tabula-py) (1.11.0)
Requirement already satisfied: Textblob in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (0.15.3)
Requirement already satisfied: nltk>=3.1 in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (from Textblob) (3.3)
Requirement already satisfied: six in c:\users\basiab1\appdata\local\conti
nuum\anaconda3\lib\site-packages (from nltk>=3.1->Textblob) (1.11.0)
Collecting matplotlib==1.4.3 (from -r requirements.txt (line 1))
  Using cached https://files.pythonhosted.org/packages/bb/ac/485df0ecb15aa
6fec1991945dc0cabfeb724a64f6729e34bab3c6a766813/matplotlib-1.4.3.tar.gz
```

```
    Complete output from command python setup.py egg_info:
    ========================================================================
======
    Edit setup.cfg to change the build options

    BUILDING MATPLOTLIB
                matplotlib: yes [1.4.3]
                    python: yes [3.7.1 (default, Oct 28 2018, 08:39:03) [M
SC
                            v.1912 64 bit (AMD64)]]
                  platform: yes [win32]

    REQUIRED DEPENDENCIES AND EXTENSIONS
                     numpy: yes [version 1.15.4]
                       six: yes [using six version 1.11.0]
                  dateutil: yes [using dateutil version 2.7.5]
                      pytz: yes [using pytz version 2018.7]
                   tornado: yes [using tornado version 5.1.1]
                 pyparsing: yes [using pyparsing version 2.3.0]
                     pycxx: yes [Official versions of PyCXX are not compat
ible
                            with matplotlib on Python 3.x, since they lack
                            support for the buffer object.  Using local co
py]
                    libagg: yes [pkg-config information for 'libagg' could
 not
                            be found. Using local copy.]
                  freetype: no  [The C/C++ header for freetype (ft2build.h
)
                            could not be found.  You may need to install t
he
                            development package.]
                       png: no  [The C/C++ header for png (png.h) could no
t be
                            found.  You may need to install the developmen
t
                            package.]
                     qhull: yes [pkg-config information for 'qhull' could
not be
                            found. Using local copy.]

    OPTIONAL SUBPACKAGES
                sample_data: yes [installing]
                   toolkits: yes [installing]
                      tests: yes [using nose version 1.3.7 / using unittest
.mock]
             toolkits_tests: yes [using nose version 1.3.7 / using unittest
.mock]

    OPTIONAL BACKEND EXTENSIONS
                     macosx: no  [Mac OS-X only]
                     qt5agg: yes [installing, Qt: 5.9.6, PyQt: 5.9.6]
                     qt4agg: no  [PyQt4 not found]
                     pyside: no  [PySide not found]
                    gtk3agg: no  [Requires pygobject to be installed.]
                  gtk3cairo: no  [Requires cairocffi or pycairo to be insta
lled.]
```

```
                    gtkagg: no  [Requires pygtk]
                     tkagg: no  [The C/C++ header for Tk (tk.h) could not
be
                            found.  You may need to install the developmen
t
                            package.]
                     wxagg: no  [requires wxPython]
                       gtk: no  [Requires pygtk]
                       agg: yes [installing]
                     cairo: no  [cairocffi or pycairo not found]
                 windowing: yes [installing, installing]

    OPTIONAL LATEX DEPENDENCIES
                     dvipng: no
                ghostscript: no
                      latex: no
                    pdftops: no

    ============================================================================
======
                            * The following required packages can not be b
uilt:
                            * freetype, png

    ----------------------------------------
```

```
Command "python setup.py egg_info" failed with error code 1 in C:\Users\ba
siab1\AppData\Local\Temp\pip-install-0gmoxr7u\matplotlib\
```

```
Requirement already satisfied: virtualenv in c:\users\basiab1\appdata\loca
l\continuum\anaconda3\lib\site-packages (20.0.18)
Requirement already satisfied: importlib-metadata<2,>=0.12; python_version
 < "3.8" in c:\users\basiab1\appdata\local\continuum\anaconda3\lib\site-pa
ckages (from virtualenv) (1.6.0)
Requirement already satisfied: distlib<1,>=0.3.0 in c:\users\basiab1\appda
ta\local\continuum\anaconda3\lib\site-packages (from virtualenv) (0.3.0)
Requirement already satisfied: filelock<4,>=3.0.0 in c:\users\basiab1\appd
ata\local\continuum\anaconda3\lib\site-packages (from virtualenv) (3.0.10)
Requirement already satisfied: appdirs<2,>=1.4.3 in c:\users\basiab1\appda
ta\local\continuum\anaconda3\lib\site-packages (from virtualenv) (1.4.3)
Requirement already satisfied: six<2,>=1.9.0 in c:\users\basiab1\appdata\l
ocal\continuum\anaconda3\lib\site-packages (from virtualenv) (1.11.0)
Requirement already satisfied: zipp>=0.5 in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (from importlib-metadata<2,>=0.12;
python_version < "3.8"->virtualenv) (0.5.2)
```

```
'apt-get' is not recognized as an internal or external command,
operable program or batch file.
'apt-get' is not recognized as an internal or external command,
operable program or batch file.
```

```
Requirement already satisfied: tabula-py in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (2.1.0)
Requirement already satisfied: xlrd in c:\users\basiab1\appdata\local\cont
inuum\anaconda3\lib\site-packages (1.1.0)
Requirement already satisfied: lxml in c:\users\basiab1\appdata\local\cont
inuum\anaconda3\lib\site-packages (4.2.5)
Requirement already satisfied: distro in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (from tabula-py) (1.5.0)
```

```
Requirement already satisfied: numpy in c:\users\basiab1\appdata\local\con
tinuum\anaconda3\lib\site-packages (from tabula-py) (1.15.4)
Requirement already satisfied: pandas>=0.25.3 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from tabula-py) (1.0.3)
Requirement already satisfied: python-dateutil>=2.6.1 in c:\users\basiab1\
appdata\local\continuum\anaconda3\lib\site-packages (from pandas>=0.25.3->
tabula-py) (2.7.5)
Requirement already satisfied: pytz>=2017.2 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from pandas>=0.25.3->tabula-py)
 (2018.7)
Requirement already satisfied: six>=1.5 in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (from python-dateutil>=2.6.1->pandas
>=0.25.3->tabula-py) (1.11.0)
Requirement already satisfied: plotly in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (4.6.0)
Requirement already satisfied: six in c:\users\basiab1\appdata\local\conti
nuum\anaconda3\lib\site-packages (from plotly) (1.11.0)
Requirement already satisfied: retrying>=1.3.3 in c:\users\basiab1\appdata
\local\continuum\anaconda3\lib\site-packages (from plotly) (1.3.3)
Requirement already satisfied: cufflinks in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (0.17.3)
Requirement already satisfied: colorlover>=0.2.1 in c:\users\basiab1\appda
ta\local\continuum\anaconda3\lib\site-packages (from cufflinks) (0.3.0)
Requirement already satisfied: ipython>=5.3.0 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from cufflinks) (7.1.1)
Requirement already satisfied: setuptools>=34.4.1 in c:\users\basiab1\appd
ata\local\continuum\anaconda3\lib\site-packages (from cufflinks) (40.5.0)
Requirement already satisfied: ipywidgets>=7.0.0 in c:\users\basiab1\appda
ta\local\continuum\anaconda3\lib\site-packages (from cufflinks) (7.4.2)
Requirement already satisfied: six>=1.9.0 in c:\users\basiab1\appdata\loca
l\continuum\anaconda3\lib\site-packages (from cufflinks) (1.11.0)
Requirement already satisfied: pandas>=0.19.2 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from cufflinks) (1.0.3)
Requirement already satisfied: numpy>=1.9.2 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from cufflinks) (1.15.4)
Requirement already satisfied: plotly>=4.1.1 in c:\users\basiab1\appdata\l
ocal\continuum\anaconda3\lib\site-packages (from cufflinks) (4.6.0)
Requirement already satisfied: jedi>=0.10 in c:\users\basiab1\appdata\loca
l\continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflinks) (
0.13.1)
Requirement already satisfied: traitlets>=4.2 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflink
s) (4.3.2)
Requirement already satisfied: decorator in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflinks) (4
.3.0)
Requirement already satisfied: colorama; sys_platform == "win32" in c:\use
rs\basiab1\appdata\local\continuum\anaconda3\lib\site-packages (from ipyth
on>=5.3.0->cufflinks) (0.4.0)
Requirement already satisfied: pickleshare in c:\users\basiab1\appdata\loc
al\continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflinks)
(0.7.5)
Requirement already satisfied: prompt-toolkit<2.1.0,>=2.0.0 in c:\users\ba
siab1\appdata\local\continuum\anaconda3\lib\site-packages (from ipython>=5
.3.0->cufflinks) (2.0.7)
Requirement already satisfied: backcall in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflinks) (0.
```

```
1.0)
Requirement already satisfied: pygments in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (from ipython>=5.3.0->cufflinks) (2.
2.0)
Requirement already satisfied: ipykernel>=4.5.1 in c:\users\basiab1\appdat
a\local\continuum\anaconda3\lib\site-packages (from ipywidgets>=7.0.0->cuf
flinks) (5.1.0)
Requirement already satisfied: widgetsnbextension~=3.4.0 in c:\users\basia
b1\appdata\local\continuum\anaconda3\lib\site-packages (from ipywidgets>=7
.0.0->cufflinks) (3.4.2)
Requirement already satisfied: nbformat>=4.2.0 in c:\users\basiab1\appdata
\local\continuum\anaconda3\lib\site-packages (from ipywidgets>=7.0.0->cuff
links) (4.4.0)
Requirement already satisfied: pytz>=2017.2 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from pandas>=0.19.2->cufflinks)
 (2018.7)
Requirement already satisfied: python-dateutil>=2.6.1 in c:\users\basiab1\
appdata\local\continuum\anaconda3\lib\site-packages (from pandas>=0.19.2->
cufflinks) (2.7.5)
Requirement already satisfied: retrying>=1.3.3 in c:\users\basiab1\appdata
\local\continuum\anaconda3\lib\site-packages (from plotly>=4.1.1->cufflink
s) (1.3.3)
Requirement already satisfied: parso>=0.3.0 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from jedi>=0.10->ipython>=5.3.0
->cufflinks) (0.3.1)
Requirement already satisfied: ipython-genutils in c:\users\basiab1\appdat
a\local\continuum\anaconda3\lib\site-packages (from traitlets>=4.2->ipytho
n>=5.3.0->cufflinks) (0.2.0)
Requirement already satisfied: wcwidth in c:\users\basiab1\appdata\local\c
ontinuum\anaconda3\lib\site-packages (from prompt-toolkit<2.1.0,>=2.0.0->i
python>=5.3.0->cufflinks) (0.1.7)
Requirement already satisfied: jupyter-client in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from ipykernel>=4.5.1->ipywid
gets>=7.0.0->cufflinks) (5.2.3)
Requirement already satisfied: tornado>=4.2 in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from ipykernel>=4.5.1->ipywidge
ts>=7.0.0->cufflinks) (5.1.1)
Requirement already satisfied: notebook>=4.4.1 in c:\users\basiab1\appdata
\local\continuum\anaconda3\lib\site-packages (from widgetsnbextension~=3.4
.0->ipywidgets>=7.0.0->cufflinks) (5.7.0)
Requirement already satisfied: jupyter-core in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from nbformat>=4.2.0->ipywidget
s>=7.0.0->cufflinks) (4.4.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in c:\users\basiab1
\appdata\local\continuum\anaconda3\lib\site-packages (from nbformat>=4.2.0
->ipywidgets>=7.0.0->cufflinks) (2.6.0)
Requirement already satisfied: pyzmq>=13 in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (from jupyter-client->ipykernel>=4.
5.1->ipywidgets>=7.0.0->cufflinks) (17.1.2)
Requirement already satisfied: jinja2 in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (from notebook>=4.4.1->widgetsnbextens
ion~=3.4.0->ipywidgets>=7.0.0->cufflinks) (2.10)
Requirement already satisfied: nbconvert in c:\users\basiab1\appdata\local
\continuum\anaconda3\lib\site-packages (from notebook>=4.4.1->widgetsnbext
ension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (5.3.1)
Requirement already satisfied: terminado>=0.8.1 in c:\users\basiab1\appdat
a\local\continuum\anaconda3\lib\site-packages (from notebook>=4.4.1->widge
```

```
tsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.8.1)
Requirement already satisfied: prometheus-client in c:\users\basiab1\appda
ta\local\continuum\anaconda3\lib\site-packages (from notebook>=4.4.1->widg
etsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.4.2)
Requirement already satisfied: Send2Trash in c:\users\basiab1\appdata\loca
l\continuum\anaconda3\lib\site-packages (from notebook>=4.4.1->widgetsnbex
tension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (1.5.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\basiab1\appdat
a\local\continuum\anaconda3\lib\site-packages (from jinja2->notebook>=4.4.
1->widgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (1.1.0)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\basiab1\ap
pdata\local\continuum\anaconda3\lib\site-packages (from nbconvert->noteboo
k>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (1.4.2)
Requirement already satisfied: mistune>=0.7.4 in c:\users\basiab1\appdata\
local\continuum\anaconda3\lib\site-packages (from nbconvert->notebook>=4.4
.1->widgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.8.4)
Requirement already satisfied: entrypoints>=0.2.2 in c:\users\basiab1\appd
ata\local\continuum\anaconda3\lib\site-packages (from nbconvert->notebook>
=4.4.1->widgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.2.3)
Requirement already satisfied: bleach in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (from nbconvert->notebook>=4.4.1->widg
etsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (3.0.2)
Requirement already satisfied: testpath in c:\users\basiab1\appdata\local\
continuum\anaconda3\lib\site-packages (from nbconvert->notebook>=4.4.1->wi
dgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.4.2)
Requirement already satisfied: webencodings in c:\users\basiab1\appdata\lo
cal\continuum\anaconda3\lib\site-packages (from bleach->nbconvert->noteboo
k>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets>=7.0.0->cufflinks) (0.5.1)
Requirement already satisfied: PTable in c:\users\basiab1\appdata\local\co
ntinuum\anaconda3\lib\site-packages (0.9.2)
```

Create a database in Python using sqlite3

In [2]:
```
conn = sqlite3.connect('Soccer_Prediction_Milestone5.db')
```

In [3]:
```
c = conn.cursor()
```

In [4]:
```
c.execute('''
    CREATE TABLE SPIMATCHES (
        "date"  TEXT,
        "league_id"     INTEGER,
        "team1" TEXT,
        "team2" TEXT,
        "spi1"  REAL,
        "spi2"  REAL,
        "prob1" REAL,
        "prob2" REAL,
        "probtie"       REAL,
        "proj_score1"   REAL,
        "proj_score2"   REAL,
        "importance1"   REAL,
        "importance2"   REAL,
        "score1"        INTEGER,
        "score2"        INTEGER,
```

```
        "xg1"   REAL,
        "xg2"   REAL,
        "nsxg1" REAL,
        "nsxg2" REAL,
        "adj_score1"    REAL,
        "adj_score2"    REAL
)''')
```

In [5]:
```
conn.commit()
```

In [6]:
```
df = pd.read_csv(r'spi_matches.csv')
```

In [7]:
```
spi_read.to_sql('SPIMATCHES', conn, if_exists='append', index = False)
```

In [8]:
```
df
```

Out[8]:

	date	league_id	league	team1	team2	spi1	spi2	prob1	prob2	probti
0	2016-08-12	1843	French Ligue 1	Bastia	Paris Saint-Germain	51.16	85.68	0.0463	0.8380	0.1157
1	2016-08-12	1843	French Ligue 1	AS Monaco	Guingamp	68.85	56.48	0.5714	0.1669	0.2617
2	2016-08-13	2411	Barclays Premier League	Hull City	Leicester City	53.57	66.81	0.3459	0.3621	0.2921
3	2016-08-13	2411	Barclays Premier League	Crystal Palace	West Bromwich Albion	55.19	58.66	0.4214	0.2939	0.2847
4	2016-08-13	2411	Barclays Premier League	Everton	Tottenham Hotspur	68.02	73.25	0.3910	0.3401	0.2689
...	...	...	...	...	...	...	...	...	...	...
34575	2020-12-06	2105	Brasileiro Série A	São Paulo	Flamengo	58.05	73.73	0.3274	0.4304	0.2423
34576	2020-12-06	2105	Brasileiro Série A	Bahía	Santos	48.12	61.02	0.3586	0.3800	0.2614
34577	2020-12-06	2105	Brasileiro Série A	Fluminense	Fortaleza	50.90	45.02	0.5550	0.2097	0.2353
34578	2020-12-06	2105	Brasileiro Série A	Atletico Mineiro	Palmeiras	51.72	67.88	0.3232	0.4297	0.2472
34579	2020-12-06	2105	Brasileiro Série A	Atlético Paranaense	Sport Recife	56.11	40.01	0.6595	0.1314	0.2091

34580 rows × 22 columns

In [9]:
```
df.head()
```

Out[9]:

	date	league_id	league	team1	team2	spi1	spi2	prob1	prob2	probtie	...	imp
0	2016-08-12	1843	French Ligue 1	Bastia	Paris Saint-Germain	51.16	85.68	0.0463	0.8380	0.1157	...	32.4
1	2016-08-12	1843	French Ligue 1	AS Monaco	Guingamp	68.85	56.48	0.5714	0.1669	0.2617	...	53.7
2	2016-08-13	2411	Barclays Premier League	Hull City	Leicester City	53.57	66.81	0.3459	0.3621	0.2921	...	38.1
3	2016-08-13	2411	Barclays Premier League	Crystal Palace	West Bromwich Albion	55.19	58.66	0.4214	0.2939	0.2847	...	43.6
4	2016-08-13	2411	Barclays Premier League	Everton	Tottenham Hotspur	68.02	73.25	0.3910	0.3401	0.2689	...	31.9

5 rows × 22 columns

The data saved the score from each game (team1 vs team2), and some historical data of each team. SPI stands for Soccer Power Index, score1 and score2 are the scores of the game.

In [10]:
```
# How many rows and how many columns are in my dataframe?

df.shape
```

Out[10]: (34580, 22)

In [11]:
```
# Since the data has 34580 records (it's still updating so the data size m
ay change), so I would start from one team,
# for example Juventus (an Italian league team).

J_df = df[df['team1']=='Juventus']
```

In [12]:
```
J_df.shape
```

Out[12]: (97, 22)

In [14]:
```
J_df.head()
```

Out[14]:

	date	league_id	league	team1	team2	spi1	spi2	prob1	prob2	probtie	..
40	2016-	1854	Italy Serie	Juventus	Fiorentina	82.79	69.19	0.6808	0.1010	0.2182	..

	08-20		A								
127	2016-09-10	1854	Italy Serie A	Juventus	Sassuolo	83.18	59.69	0.7988	0.0400	0.1612	..
167	2016-09-14	1818	UEFA Champions League	Juventus	Sevilla FC	83.47	78.92	0.6447	0.1335	0.2217	..
248	2016-09-21	1854	Italy Serie A	Juventus	Cagliari	82.43	58.24	0.8095	0.0391	0.1514	..
403	2016-10-15	1854	Italy Serie A	Juventus	Udinese	86.66	51.86	0.8759	0.0220	0.1021	..

5 rows × 22 columns

In [15]:
```
# Apply filter for the fields that I am interested in, which is team SPI d
uring that specific date and the score of that specific game.

df1 = J_df[['date','league','team1','team2','spi1','spi2','score1','score2
']]
df1.shape
```

Out[15]: (97, 8)

Identify bad data

In [16]:
```
# Check last few rows of the data

df1.tail()
```

Out[16]:

	date	league	team1	team2	spi1	spi2	score1	score2
31403	2020-04-04	Italy Serie A	Juventus	Torino	84.28	54.27	NaN	NaN
31923	2020-04-19	Italy Serie A	Juventus	Atalanta	84.28	82.78	NaN	NaN
32211	2020-04-26	Italy Serie A	Juventus	Lazio	84.28	76.44	NaN	NaN
32658	2020-05-10	Italy Serie A	Juventus	Sampdoria	84.28	62.05	NaN	NaN
33035	2020-05-24	Italy Serie A	Juventus	AS Roma	84.28	75.15	NaN	NaN

It shows some missing values. These are match fixtures happening in future dates so there are no scores yet. Those data points need to be removed.

In [17]:
```
# Drop N/As

df1 = df1.dropna()
df1.shape
```

Out[17]: (90, 8)

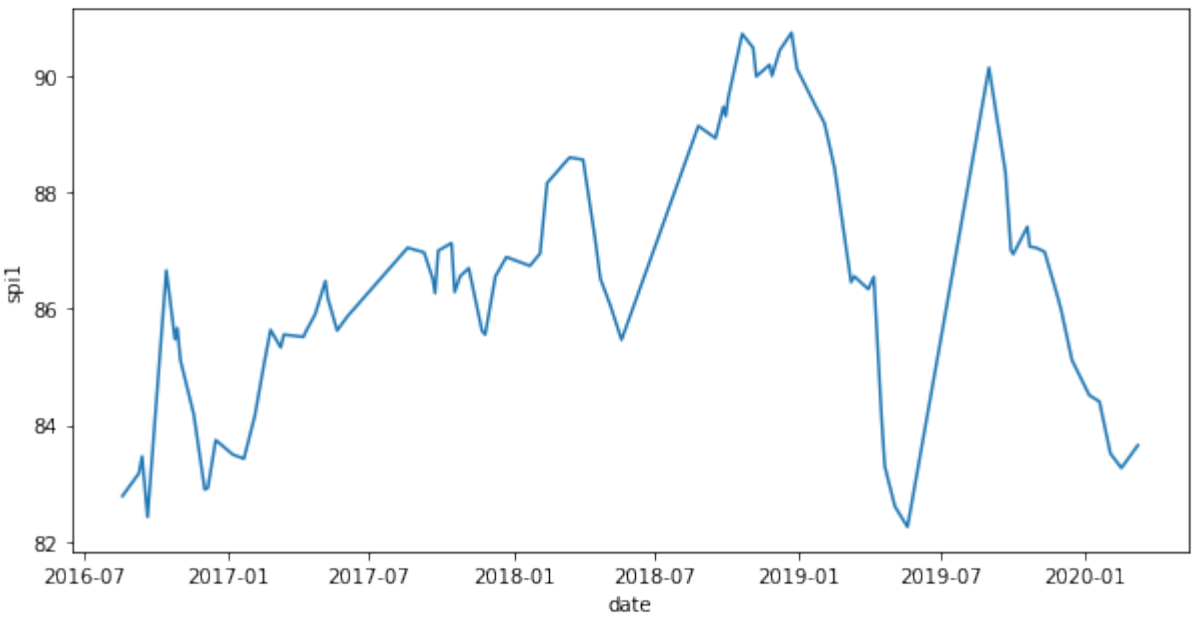
Format data into a more readable form

In [18]:
```
# Check the format of data, sometimes the date is not in the format we wan
t. It needs to be transformed.

df1.dtypes
```

Out[18]:
```
date       object
league     object
team1      object
team2      object
spi1      float64
spi2      float64
score1    float64
score2    float64
dtype: object
```

In [19]:
```
df1['date'] = pd.to_datetime(df1['date'])
```

In [20]:
```
# Plot the Juventus Soccor Power Index over time

plt.figure(figsize=(10,5))
sns.lineplot(data=df1, x="date", y="spi1")
plt.show()
```



The drop on May 2019 is becasue Juventus lost the game with Ajax in UEFA Champion League, and another two draws with Torino and Atalanta. And there's no updates between June and August since there's no game.

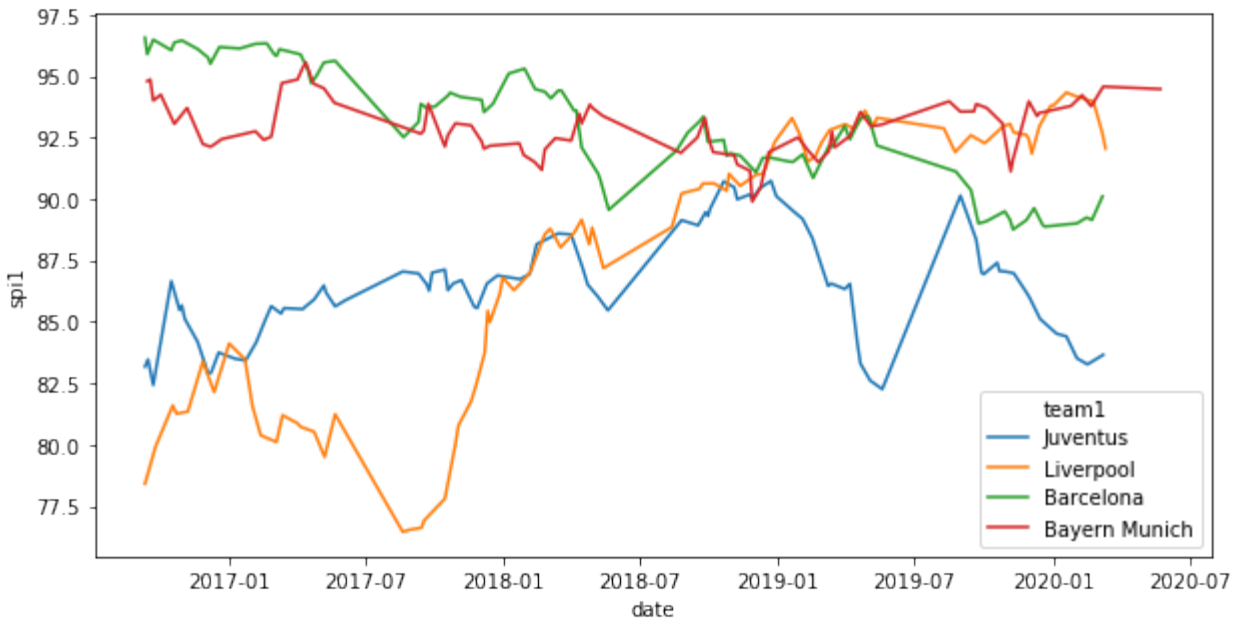
Conduct Fuzzy Matching

In [21]:
```
# Plot the spi score of couple top teams. Here I only selected Juventus, L
iverpool, Barcelona, and Bayern Munich. There are
# couple other top clubs with great performance, since it's just for a sim
ple illustration so I won't cover all of them.

df2 = df[df['team1'].isin(['Juventus','Liverpool','Barcelona','Bayern Muni
ch'])]
```

In [22]:
```
# Clean the data

df2 = df2[['date','league','team1','team2','spi1','spi2','score1','score2'
]]
df2 = df2.dropna()
df2['date'] = pd.to_datetime(df2['date'])
df2 = df2[df2['date']>'2016-09-01']
df2.shape
```

Out[22]: (348, 8)

In [23]:
```
# Plot the performance of the top teams

plt.figure(figsize=(10,5))
sns.lineplot(data=df2, x="date", y="spi1", hue='team1')
plt.show()
```



Some Inferences so far: 1) Huge improvement observed for Liverpool during 2018 and consistency through 2020 2) Bayern Munich showing consistent performance 3) Barcelona performance has declined since 2017 4) Juventus has too much variations in their performance, specially since 2019

Next let's take a look at Italy Serie A, which is the 4th soccer in the Europe.

In [24]:
```
df_I = df[df['league']=='Italy Serie A']
df_I = df_I[df_I['date']>'2017-07-01']
df_I = df_I[['date','team1','team2','spi1','spi2','score1','score2']]
```

In [25]:
```
df_I.head()
```

Out[25]:

	date	team1	team2	spi1	spi2	score1	score2
3315	2017-08-19	Juventus	Cagliari	87.05	57.81	3.0	0.0
3343	2017-08-19	Verona	Napoli	46.39	81.53	1.0	3.0
3396	2017-08-20	Atalanta	AS Roma	70.13	81.61	0.0	1.0
3409	2017-08-20	Internazionale	Fiorentina	77.03	65.92	3.0	0.0
3410	2017-08-20	Bologna	Torino	55.51	63.50	1.0	1.0

In [26]:
```
df_I.shape
```

Out[26]:
```
(1140, 7)
```

In [27]:
```
# Select the goal and goal lost for each team, rename the columns and stac
k them together.
df_I1 = df_I[['date','team1','score1','score2']]
df_I2 = df_I[['date','team2','score2','score1']]
df_I1.columns=['date','team','offense','defense']
df_I2.columns=['date','team','offense','defense']
```

In [28]:
```
frames = [df_I1,df_I2]
df_I3=pd.concat(frames)
df_I3.shape
```

Out[28]:
```
(2280, 4)
```

In [29]:
```
# Calculate a simple offense and defense index for each team.
df_score = df_I3.groupby(['team'],as_index=False).mean()
```

In [30]:
```
df_score
```

Out[30]:

	team	offense	defense
0	AC Milan	1.362745	1.098039
1	AS Roma	1.745098	1.088235
2	Atalanta	2.019802	1.178218
3	Benevento	0.868421	2.210526
4	Bologna	1.235294	1.470588
5	Brescia	0.846154	1.884615
6	Cagliari	1.089109	1.534653
7	Chievo Verona	0.802632	1.763158
8	Crotone	1.052632	1.736842
9	Empoli	1.342105	1.842105

10	Fiorentina	1.303922	1.245098
11	Frosinone	0.763158	1.815789
12	Genoa	1.009804	1.441176
13	Internazionale	1.702970	0.861386
14	Juventus	2.019608	0.764706
15	Lazio	2.009804	1.156863
16	Lecce	1.307692	2.153846
17	Napoli	1.882353	0.990196
18	Parma	1.158730	1.460317
19	Sampdoria	1.425743	1.534653
20	Sassuolo	1.217822	1.564356
21	Spal	1.009804	1.558824
22	Torino	1.326733	1.267327
23	Udinese	1.058824	1.500000
24	Verona	0.936508	1.650794

I want to put them into a scatterplot, so I want to fit them into a reasonable range. I will start with standarizing. I also multiplied defense index with -1, I want to makse sure the higher score means better performance.

In [31]:
```
df_score[['offense', 'defense']] = preprocessing.StandardScaler().fit_tran
sform(df_score[['offense', 'defense']])
df_score['defense'] = df_score['defense']*(-1)
```

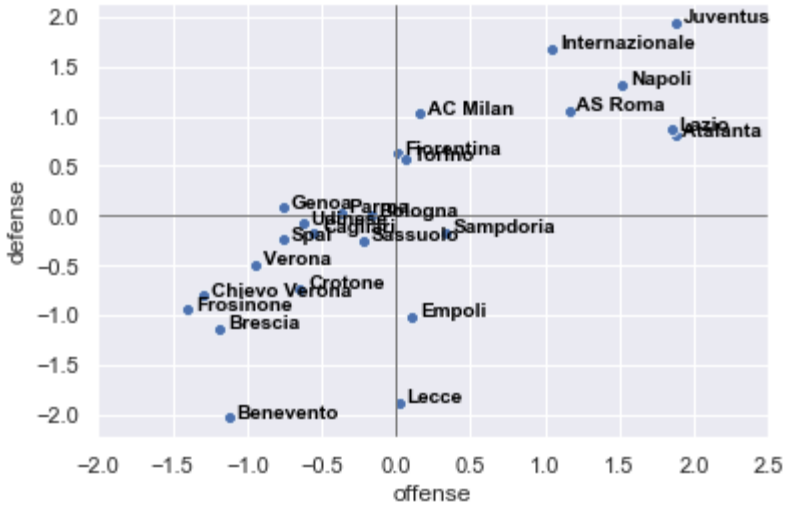
In [32]:
```
df_score
```

Out[32]:

	team	offense	defense
0	AC Milan	0.163698	1.023185
1	AS Roma	1.159624	1.050089
2	Atalanta	1.875153	0.803159
3	Benevento	-1.123882	-2.029706
4	Bologna	-0.168277	0.000835
5	Brescia	-1.181882	-1.135340
6	Cagliari	-0.549050	-0.174973
7	Chievo Verona	-1.295246	-0.802036
8	Crotone	-0.644063	-0.729820
9	Empoli	0.109937	-1.018683

10	Fiorentina	0.010479	0.619626
11	Frosinone	-1.398064	-0.946468
12	Genoa	-0.755618	0.081547
13	Internazionale	1.049892	1.672610
14	Juventus	1.874647	1.937920
15	Lazio	1.849111	0.861762
16	Lecce	0.020301	-1.874164
17	Napoli	1.517136	1.319129
18	Parma	-0.367705	0.029020
19	Sampdoria	0.327789	-0.174973
20	Sassuolo	-0.213788	-0.256484
21	Spal	-0.755618	-0.241301
22	Torino	0.069896	0.558626
23	Udinese	-0.627935	-0.079877
24	Verona	-0.946534	-0.493685

In [33]:
```
sns.set()
ax = sns.scatterplot(x="offense", y="defense", data=df_score)
ax.axhline(y=0, color='grey', linewidth=1)
ax.axvline(x=0, color='grey', linewidth=1)
ax.set_xlim(-2,2.5)
for line in range(0,df_score.shape[0]):
     ax.text(df_score.offense[line]+0.05, df_score.defense[line],
             df_score.team[line], horizontalalignment='left', size='small'
, color='black', weight='semibold')
```



Teams like Juventus, Internazionale (Inter Milan), and Napoli has the best performance on both offense and defense. On the other hand Benevento, Lecce, Frosinone have a worse performance on

both. Team like Sampdoria and Empoli are better on offense than defense, while Bologna and Genoa has better defense than offense.