Assignment4 (Score: 80.0 / 100.0)

1. Test cell (Score: 20.0 / 20.0)
2. Test cell (Score: 20.0 / 20.0)
3. Test cell (Score: 20.0 / 20.0)
4. Test cell (Score: 20.0 / 20.0)
5. Comment
6. Test cell (Score: 0.0 / 20.0)

Assignment 4¶

Description¶

In this assignment you must read in a file of metropolitan regions and associated sports teams from assets/wikipedia_data.html (assets/wikipedia_data.html) and answer some questions about each metropolitan region. I of these regions may have one or more teams from the "Big 4": NFL (football, in assets/nfl.csv (assets/nfl.csv)), MLB (baseball, in assets/mlb.csv (assets/mlb.csv)), NBA (basketball, in assets/nba.csv (assets/nba.csv) or NHL (hockey, in assets/nhl.csv (assets/nhl.csv)). Please keep in mind that all questions are from the perspective of the metropolitan reg and that this file is the "source of authority" for the location of a given sports team. Thus teams which are commonly kr by a different area (e.g. "Oakland Raiders") need to be mapped into the metropolitan region given (e.g. San Francisco E Area). This will require some human data understanding outside of the data you've been given (e.g. you will have to har code some names, and might need to google to find out where teams are)!

For each sport I would like you to answer the question: **what is the win/loss ratio's correlation with the population o the city it is in?** Win/Loss ratio refers to the number of wins over the number of wins plus the number of losses. Remer that to calculate the correlation with `pearsonr` (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html), so you are going to send in two order lists of values, the populations from the wikipedia_data.html file and the win/loss ratio for a given sport in the same ord Average the win/loss ratios for those cities which have multiple teams of a single sport. Each sport is worth an equal amount in this assignment (20%*4=80%) of the grade for this assignment. You should only use data **from year 2018** fo your analysis -- this is important!

Notes¶

1. Do not include data about the MLS or CFL in any of the work you are doing, we're only interested in the Big 4 in th assignment.
2. I highly suggest that you first tackle the four correlation questions in order, as they are all similar and worth the ma of grades for this assignment. This is by design!
3. It's fair game to talk with peers about high level strategy as well as the relationship between metropolitan areas an sports teams. However, do not post code solving aspects of the assignment (including such as dictionaries mappi areas to teams, or regexes which will clean up names).
4. There may be more teams than the assert statements test, remember to collapse multiple teams in one city into a single value!

As this assignment utilizes global variables in the skeleton code, to avoid having errors in your code you can either:

1. You can place all of your code within the function definitions for all of the questions (other than import statements)
2. You can create copies of all the global variables with the copy() method and proceed as usual.

Question 1¶

For this question, calculate the win/loss ratio's correlation with the population of the city it is in for the **NHL** using **2018** data.

In [1]:

Student's answer

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

def get_area(team):
    #print(team)
    #print(nhl_cities.index.values)
    for each in list(nhl_cities.index.values):
        #print(each)
        if team in each:
            #print(team)
           # print(nhl_cities.at[each, 'Metropolitan area'])
            #print(nhl_cities.at[each, 'Metropolitan area'])
            return nhl_cities.at[each, 'Metropolitan area']


nhl_df=pd.read_csv("assets/nhl.csv")
cities=pd.read_html("assets/wikipedia_data.html")[1]
#print(nhl_df['team'])
#print(cities[['Metropolitan area','NHL']])
#print(cities[cities['NHL'].str.contains('Devils')])
#print(cities.shape)
cities=cities.iloc[:-1,[0,3,5,6,7,8]]

population =cities[['Metropolitan area', 'Population (2016 est.)[8]']]
population['Metropolitan area']=population['Metropolitan area'].str.strip()
population.rename(columns={'Population (2016 est.)[8]':'Population'},inplace=True)
population = population.set_index('Metropolitan area')
cities['NHL'].replace(r'(.*)\[.*\].*|[——]',r'\1',regex=True, inplace=True)
cities['NHL'].replace('', np.nan, inplace=True)
cities.dropna(inplace=True)

nhl_cities = cities[['Metropolitan area', 'NHL']].set_index('NHL')

#print(nhl_cities)
nhl_df= nhl_df[nhl_df['year'] == 2018].drop([0, 9, 18, 26], axis=0).drop(['League','year','
                                                                          'GP','OL','GF','G
nhl_df['team'].replace(r'(.*)[\*].*',r"\1".strip(),regex=True, inplace=True)
nhl_df['Area']=nhl_df['team'].str.split(" ").str[-1:]
print(nhl_df['Area'].str[0])
#nhl_df["Area"]= nhl_df["Area"].str[0].str.cat(nhl_df['Area'].str[1], sep =" ",na_rep ="")
nhl_df["Area"]= nhl_df["Area"].str[0]
nhl_df['Area'] = nhl_df['Area'].apply(lambda x: get_area(x))

#print(nhl_df)
#print(nhl_df)
#nhl_df['W']=pd.to_numeric(nhl_df['W'])
#nhl_df['L']=pd.to_numeric(nhl_df['L'])
```

```
nhl_df[['W','L']] = nhl_df[['W','L']].apply(pd.to_numeric, axis=1)

nhl_df=nhl_df.groupby('Area').sum()
#print(nhl_df)

#nhl_df['Ratio']=nhl_df['W'] / (nhl_df['W']+ nhl_df['L'])
#nhl_df=nhl_df.assign(Ratio=lambda x: x['W'] / (x['W'] + x['L']))
nhl_df.eval("Ratio =W / (W + L)", inplace=True)

#nhl_df.set_index('Area',inplace=True)
nhl_df.drop(['W','L'], axis=1,inplace=True)
#print(nhl_df)
#print(population['Population'])
#print(nhl_df['Ratio'])
out_df = pd.merge(nhl_df, population, how="inner", left_index=True, right_index=True)
out_df['Population']=pd.to_numeric(out_df['Population'])
#print(out_df)
#print(out_df['Ratio'])

def nhl_correlation():
    # YOUR CODE HERE

    #raise NotImplementedError()

    population_by_region = out_df['Population'] # pass in metropolitan area population from
    win_loss_by_region =out_df['Ratio']  # pass in win/loss ratio from nhl_df in the same o

    assert len(population_by_region) == len(win_loss_by_region), "Q1: Your lists must be th
    assert len(population_by_region) == 28, "Q1: There should be 28 teams being analysed fo
    #Oakland Raiders
    return stats.pearsonr(population_by_region, win_loss_by_region)[0]

nhl_correlation()
```

```
1      Lightning
2         Bruins
3          Leafs
4       Panthers
5          Wings
6      Canadiens
7       Senators
8         Sabres
10      Capitals
11      Penguins
12        Flyers
13       Jackets
14        Devils
15    Hurricanes
16     Islanders
17       Rangers
19     Predators
20          Jets
21          Wild
22     Avalanche
23         Blues
24         Stars
25    Blackhawks
27       Knights
28         Ducks
29        Sharks
30         Kings
31        Flames
32        Oilers
33       Canucks
34       Coyotes
Name: Area, dtype: object
```

Out[1]:

```
0.01230899645574425
```

In [2]:

Grade cell: `cell-ebe0b2dfe1067e63` Score: 20.0 / 20.0

Question 2¶

For this question, calculate the win/loss ratio's correlation with the population of the city it is in for the **NBA** using **2018** data.

In [3]:

Student's answer

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

def gets_area(team):
    for each in list(nba_cities.index.values):
        if  team in each: return nba_cities.at[each, 'Metropolitan area']


nba_df=pd.read_csv("assets/nba.csv")
cities=pd.read_html("assets/wikipedia_data.html")[1]
cities=cities.iloc[:-1,[0,3,5,6,7,8]]

population =cities[['Metropolitan area', 'Population (2016 est.)[8]']]
population['Metropolitan area']=population['Metropolitan area'].str.strip()
population.rename(columns={'Population (2016 est.)[8]':'Population'},inplace=True)
population = population.set_index('Metropolitan area')

cities['NBA'].replace(r'(.*)\[.*\].*|[——]',r'\1',regex=True, inplace=True)
cities['NBA'].replace('—', np.nan, inplace=True)
cities['NBA']=cities['NBA'].str.strip('—').str.strip(' ')
cities['NBA'].replace('', np.nan, inplace=True)
cities.dropna(inplace=True)


nba_cities = cities[['Metropolitan area', 'NBA']].set_index('NBA')

nba_df=nba_df[nba_df['year']==2018].drop(['W/L%','GB','PS/G','PA/G','SRS', 'League','year']
nba_df['team'].replace(r'(.*)[\*]|[\(].*',r"\1".strip(),regex=True, inplace=True)
nba_df[['team','W','L']]=nba_df[['team','W','L']].apply(lambda x: x.str.strip())

nba_df['Area']=nba_df['team'].str.split(" ").str[-1:]
nba_df["Area"]= nba_df["Area"].str[0]
nba_df['Area'] = nba_df['Area'].apply(lambda x: gets_area(x))
#print(nba_df)
nba_df[['W','L']] = nba_df[['W','L']].apply(pd.to_numeric, axis=1)
nba_df=nba_df.groupby('Area').sum()
nba_df.eval("Ratio =W / (W + L)", inplace=True)
nba_df.drop(['W','L'], axis=1,inplace=True)

out_df = pd.merge(nba_df, population, how="inner", left_index=True, right_index=True)
out_df['Population']=pd.to_numeric(out_df['Population'])
#print(out_df)

def nba_correlation():
    # YOUR CODE HERE
    #raise NotImplementedError()

    population_by_region = out_df['Population'] # pass in metropolitan area population from
    win_loss_by_region = out_df['Ratio'] # pass in win/loss ratio from nba_df in the same o


    assert len(population_by_region) == len(win_loss_by_region), "Q2: Your lists must be th
    assert len(population_by_region) == 28, "Q2: There should be 28 teams being analysed fo

    return stats.pearsonr(population_by_region, win_loss_by_region)[0]
nba_correlation()
```

Out[3]:

```
-0.17657160252844614
```

In [4]:

Grade cell: `cell-e573b2b4a282b470` Score: 20.0 / 20.0

Question 3¶

For this question, calculate the win/loss ratio's correlation with the population of the city it is in for the **MLB** using **2018** data.

In [5]:

Student's answer

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

def gets_area(team):
    for each in list(mlb_cities.index.values):
        if  team in each: return mlb_cities.at[each, 'Metropolitan area']

mlb_df=pd.read_csv("assets/mlb.csv")
cities=pd.read_html("assets/wikipedia_data.html")[1]
cities=cities.iloc[:-1,[0,3,5,6,7,8]]
#cities.to_excel("cities_before.xlsx")
#print(cities)

population =cities[['Metropolitan area', 'Population (2016 est.)[8]']]
population['Metropolitan area']=population['Metropolitan area'].str.strip()
population.rename(columns={'Population (2016 est.)[8]':'Population'},inplace=True)
population = population.set_index('Metropolitan area')

cities['MLB'].replace(r'(.*)\[.*\].*|[——]',r'\1',regex=True, inplace=True)
cities['MLB'].replace('—', np.nan, inplace=True)
cities['MLB']=cities['MLB'].str.strip('—').str.strip(' ')
cities['MLB'].replace('', np.nan, inplace=True)
#cities.to_excel('asdsad.xlsx')
cities.dropna(inplace=True)

mlb_cities = cities[['Metropolitan area', 'MLB']].set_index('MLB')
#print(mlb_cities)

mlb_df=mlb_df[mlb_df['year']==2018].drop(['GB','W-L%','League','year'],axis=1)
#mlb_df.to_excel("mlb_df.xlsx")

mlb_df['team'].replace(r'(.*)[\*]|[\(].*',r"\1".strip(),regex=True, inplace=True)
mlb_df[['team']]=mlb_df[['team']].apply(lambda x: x.str.strip())
mlb_df['Area']=mlb_df['team'].str.split(" ").str[-1:]
mlb_df["Area"]= mlb_df["Area"].str[0]
mlb_df['Area'] = mlb_df['Area'].apply(lambda x: gets_area(x))
#print(mlb_df)
```

```
mlb_df.loc[0,'Area']='Boston'
#if mlb_df['team']== 'Boston Red Sox':
#    print(mlb_df)
#    mlb_df['Area']="Boston"

#mlb_df.to_excel("mlb_dfs.xlsx")
#print(mlb_df)
#mlb_df.rename(columns=lambda x: x.strip())
mlb_df[['W','L']] = mlb_df[['W','L']].apply(pd.to_numeric, axis=1)
mlb_df=mlb_df.groupby('Area').sum()
#print(len(mlb_df))
mlb_df.eval("Ratio =W / (W + L)", inplace=True)
mlb_df.drop(['W','L'], axis=1,inplace=True)


#mlb_df.to_excel("mlb-modified.xlsx")
#print(len(mlb_df))

#print(mlb_df)
out_df = pd.merge(mlb_df, population, how="inner", left_index=True, right_index=True)
out_df['Population']=pd.to_numeric(out_df['Population'])

#print(out_df)
def mlb_correlation():
    # YOUR CODE HERE
    #raise NotImplementedError()

    population_by_region = out_df['Population'] # pass in metropolitan area population from
    win_loss_by_region =out_df['Ratio'] # pass in win/loss ratio from mlb_df in the same or
    assert len(population_by_region) == len(win_loss_by_region), "Q3: Your lists must be th
    assert len(population_by_region) == 26, "Q3: There should be 26 teams being analysed fo
    return stats.pearsonr(population_by_region, win_loss_by_region)[0]
mlb_correlation()
```

Out[5]:

```
0.1505230448710485
```

In [6]:

Grade cell: `cell-764d4476f425c5a2` Score: 20.0 / 20.0

Question 4¶

For this question, calculate the win/loss ratio's correlation with the population of the city it is in for the **NFL** using **2018** data.

In [7]:

Student's answer

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re
```

```
def gets_area(team):
    for each in list(nfl_cities.index.values):
        if  team in each: return nfl_cities.at[each, 'Metropolitan area']

nfl_df=pd.read_csv("assets/nfl.csv")
cities=pd.read_html("assets/wikipedia_data.html")[1]
cities=cities.iloc[:-1,[0,3,5,6,7,8]]

population =cities[['Metropolitan area', 'Population (2016 est.)[8]']]
population['Metropolitan area']=population['Metropolitan area'].str.strip()
population.rename(columns={'Population (2016 est.)[8]':'Population'},inplace=True)
population = population.set_index('Metropolitan area')


cities['NFL'].replace(r'(.*)\[.*\].*|[——]',r'\1',regex=True, inplace=True)
cities['NFL'].replace('—', np.nan, inplace=True)
cities['NFL']=cities['NFL'].str.strip('—').str.strip(' ')
cities['NFL'].replace('', np.nan, inplace=True)
#cities.to_excel('c1.xlsx')
cities.dropna(inplace=True)

nfl_cities = cities[['Metropolitan area', 'NFL']].set_index('NFL')
nfl_df=nfl_df[nfl_df['year']==2018].iloc[:,[1,11,13,14]]
nfl_df.drop([0, 5, 10,15,20, 25,30,35],axis=0,inplace=True)
nfl_df['team'].replace(r'(.*)[\*+]|[\(].*',r"\1".strip(),regex=True, inplace=True)
#nfl_df.to_excel("n1.xlsx")
nfl_df[['team']]=nfl_df[['team']].apply(lambda x: x.str.strip())
nfl_df['Area']=nfl_df['team'].str.split(" ").str[-1:]
nfl_df["Area"]= nfl_df["Area"].str[0]
nfl_df['Area'] = nfl_df['Area'].apply(lambda x: gets_area(x))
#nfl_df.to_excel("n2.xlsx")

nfl_df[['W','L']] = nfl_df[['W','L']].apply(pd.to_numeric, axis=1)
print(nfl_df)
nfl_df=nfl_df.groupby('Area').sum()
print(nfl_df)
nfl_df.eval("Ratio =W / (W + L)", inplace=True)
nfl_df.drop(['W','L','year'], axis=1,inplace=True)
#print(len(nfl_df))
out_df = pd.merge(nfl_df, population, how="inner", left_index=True, right_index=True)
out_df['Population']=pd.to_numeric(out_df['Population'])
#out_df.to_excel('Q4.xlsx')
print(out_df)
def nfl_correlation():
    # YOUR CODE HERE
    #raise NotImplementedError()

    population_by_region =out_df['Population'] # pass in metropolitan area population from
    win_loss_by_region =out_df['Ratio'] # pass in win/loss ratio from nfl_df in the same or

    assert len(population_by_region) == len(win_loss_by_region), "Q4: Your lists must be th
    assert len(population_by_region) == 29, "Q4: There should be 29 teams being analysed fo

    return stats.pearsonr(population_by_region, win_loss_by_region)[0]

#print(nfl_correlation())
```

```
     L   W                  team  year                    Area
1    5  11  New England Patriots  2018                  Boston
2    9   7        Miami Dolphins  2018   Miami–Fort Lauderdale
3   10   6         Buffalo Bills  2018                 Buffalo
4   12   4         New York Jets  2018           New York City
6    6  10      Baltimore Ravens  2018               Baltimore
7    6   9   Pittsburgh Steelers  2018              Pittsburgh
8    8   7      Cleveland Browns  2018               Cleveland
9   10   6    Cincinnati Bengals  2018              Cincinnati
11   5  11        Houston Texans  2018                 Houston
12   6  10    Indianapolis Colts  2018             Indianapolis
13   7   9      Tennessee Titans  2018                Nashville
14  11   5  Jacksonville Jaguars  2018             Jacksonville
16   4  12    Kansas City Chiefs  2018              Kansas City
17   4  12  Los Angeles Chargers  2018              Los Angeles
18  10   6        Denver Broncos  2018                   Denver
19  12   4       Oakland Raiders  2018  San Francisco Bay Area
21   6  10        Dallas Cowboys  2018       Dallas–Fort Worth
22   7   9   Philadelphia Eagles  2018             Philadelphia
23   9   7   Washington Redskins  2018        Washington, D.C.
24  11   5       New York Giants  2018           New York City
26   4  12         Chicago Bears  2018                  Chicago
27   7   8     Minnesota Vikings  2018  Minneapolis–Saint Paul
28   9   6     Green Bay Packers  2018                Green Bay
29  10   6         Detroit Lions  2018                  Detroit
31   3  13    New Orleans Saints  2018              New Orleans
32   9   7     Carolina Panthers  2018                Charlotte
33   9   7       Atlanta Falcons  2018                  Atlanta
34  11   5  Tampa Bay Buccaneers  2018          Tampa Bay Area
36   3  13      Los Angeles Rams  2018             Los Angeles
37   6  10      Seattle Seahawks  2018                 Seattle
38  12   4   San Francisco 49ers  2018  San Francisco Bay Area
39  13   3     Arizona Cardinals  2018                 Phoenix
                       L   W  year
Area
Atlanta                9   7  2018
Baltimore              6  10  2018
Boston                 5  11  2018
Buffalo               10   6  2018
Charlotte              9   7  2018
Chicago                4  12  2018
Cincinnati            10   6  2018
Cleveland              8   7  2018
Dallas–Fort Worth      6  10  2018
Denver                10   6  2018
Detroit               10   6  2018
Green Bay              9   6  2018
Houston                5  11  2018
Indianapolis           6  10  2018
Jacksonville          11   5  2018
Kansas City            4  12  2018
Los Angeles            7  25  4036
Miami–Fort Lauderdale  9   7  2018
Minneapolis–Saint Paul 7   8  2018
Nashville              7   9  2018
New Orleans            3  13  2018
New York City         23   9  4036
Philadelphia           7   9  2018
Phoenix               13   3  2018
```

```
Pittsburgh               6   9  2018
San Francisco Bay Area  24   8  4036
Seattle                  6  10  2018
Tampa Bay Area          11   5  2018
Washington, D.C.         9   7  2018
                          Ratio  Population
Atlanta                0.437500     5789700
Baltimore              0.625000     2798886
Boston                 0.687500     4794447
Buffalo                0.375000     1132804
Charlotte              0.437500     2474314
Chicago                0.750000     9512999
Cincinnati             0.375000     2165139
Cleveland              0.466667     2055612
Dallas–Fort Worth      0.625000     7233323
Denver                 0.375000     2853077
Detroit                0.375000     4297617
Green Bay              0.400000      318236
Houston                0.687500     6772470
Indianapolis           0.625000     2004230
Jacksonville           0.312500     1478212
Kansas City            0.750000     2104509
Los Angeles            0.781250    13310447
Miami–Fort Lauderdale  0.437500     6066387
Minneapolis–Saint Paul 0.533333     3551036
Nashville              0.562500     1865298
New Orleans            0.812500     1268883
New York City          0.281250    20153634
Philadelphia           0.562500     6070500
Phoenix                0.187500     4661537
Pittsburgh             0.600000     2342299
San Francisco Bay Area 0.250000     6657982
Seattle                0.625000     3798902
Tampa Bay Area         0.312500     3032171
Washington, D.C.       0.437500     6131977
```

In [8]:

Grade cell: `cell-de7b148b9554dbda` Score: 20.0 / 20.0

Question 5¶

In this question I would like you to explore the hypothesis that **given that an area has two sports teams in different sports, those teams will perform the same within their respective sports**. How I would like to see this explored is v series of paired t-tests (so use `ttest_rel` (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.h between all pairs of sports. Are there any sports where we can reject the null hypothesis? Again, average values where sport has multiple teams in one region. Remember, you will only be including, for each sport, cities which have teams engaged in that sport, drop others as appropriate. This question is worth 20% of the grade for this assignment.

In [9]:

Student's answer

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

mlb_df=pd.read_csv("assets/mlb.csv")
nhl_df=pd.read_csv("assets/nhl.csv")
nba_df=pd.read_csv("assets/nba.csv")
nfl_df=pd.read_csv("assets/nfl.csv")
cities=pd.read_html("assets/wikipedia_data.html")[1]
cities=cities.iloc[:-1,[0,3,5,6,7,8]]

def sports_team_performance():
    # YOUR CODE HERE
    raise NotImplementedError()

    # Note: p_values is a full dataframe, so df.loc["NFL","NBA"] should be the same as df.l
    # df.loc["NFL","NFL"] should return np.nan
    sports = ['NFL', 'NBA', 'NHL', 'MLB']
    p_values = pd.DataFrame({k:np.nan for k in sports}, index=sports)

    assert abs(p_values.loc["NBA", "NHL"] - 0.02) <= 1e-2, "The NBA-NHL p-value should be a
    assert abs(p_values.loc["MLB", "NFL"] - 0.80) <= 1e-2, "The MLB-NFL p-value should be a
    return p_values
```

Comments:
No response.

In [10]:

Grade cell: `cell-fb4b9cb5ff4570a6` Score: 0.0 / 20.0

```
You have failed this test due to an error. The traceback has been removed because it may cor

NotImplementedError:
```

This assignment was graded by mooc_adswpy:e5e20d3b91dd, v1.46.070623