Assignment 4

Description

In this assignment you must read in a file of metropolitan regions and associated sports teams from assets/wikipedia_data.html and answer some questions about each metropolitan region. Each of these regions may have one or more teams from the "Big 4": NFL (football, in assets/nfl.csv), MLB (baseball, in assets/mlb.csv), NBA (basketball, in assets/nba.csv or NHL (hockey, in assets/nhl.csv). Please keep in mind that all questions are from the perspective of the metropolitan region, and that this file is the "source of authority" for the location of a given sports team. Thus teams which are commonly known by a different area (e.g. "Oakland Raiders") need to be mapped into the metropolitan region given (e.g. San Francisco Bay Area). This will require some human data understanding outside of the data you've been given (e.g. you will have to hand-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 of the city it is in?** Win/Loss ratio refers to the number of wins over the number of wins plus the number of losses. Remember that to calculate the correlation with `pearsonr`, so you are going to send in two ordered lists of values, the populations from the wikipedia_data.html file and the win/loss ratio for a given sport in the same order. 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** for 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 this assignment.
2. I highly suggest that you first tackle the four correlation questions in order, as they are all similar and worth the majority 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 and sports teams. However, do not post code solving aspects of the assignment (including such as dictionaries mapping 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!

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 [9]:

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

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

def nhl_correlation():
    # opening & cleaning the cities dataframe
    cities=pd.read_html("assets/wikipedia_data.html")[1]
    cities=cities.iloc[:-1,[0,3,5,6,7,8]]
    cities.drop([14,15,18,19,20,21,23,24,25,27,28,32,33,38,40,41,42,44,45,46,48,49,50],0,inplace=True)
    cities['NHL'] = cities['NHL'].str.replace('\[\w*\s*\d*\]','')
    cities = cities.set_index('Metropolitan area')

    # opening & cleaning the nhl_ft dataframe
    nhl_df=pd.read_csv("assets/nhl.csv")
    nhl_df.drop([0,9,18,26],0,inplace=True)
    nhl_df['team'] = nhl_df['team'].str.replace('\*','')
    nhl_df = nhl_df[nhl_df['year'] == 2018]
    nhl_df = nhl_df.set_index('team')


    team_dict = {'Tampa Bay Lightning':'Tampa Bay Area',
                 'Boston Bruins':'Boston',
                 'Toronto Maple Leafs':'Toronto',
                 'Florida Panthers':'Miami-Fort Lauderdale',
                 'Detroit Red Wings':'Detroit',
                 'Montreal Canadiens':'Montreal',
                 'Ottawa Senators':'Ottawa',
                 'Buffalo Sabres':'Buffalo',
                 'Washington Capitals':'Washington, D.C.',
                 'Pittsburgh Penguins':'Pittsburgh',
                 'Philadelphia Flyers':'Philadelphia',
                 'Columbus Blue Jackets':'Columbus',
                 'New Jersey Devils':'New York City',
                 'Carolina Hurricanes':'Raleigh',
                 'New York Islanders':'New York City',
                 'New York Rangers':'New York City',
                 'Nashville Predators':'Nashville',
                 'Winnipeg Jets':'Winnipeg',
                 'Minnesota Wild':'Minneapolis-Saint Paul',
                 'Colorado Avalanche':'Denver',
                 'St. Louis Blues':'St. Louis',
                 'Dallas Stars':'Dallas-Fort Worth',
                 'Chicago Blackhawks':'Chicago',
                 'Vegas Golden Knights':'Las Vegas',
                 'Anaheim Ducks':'Los Angeles',
                 'San Jose Sharks':'San Francisco Bay Area',
                 'Los Angeles Kings':'Los Angeles',
                 'Calgary Flames':'Calgary',
                 'Edmonton Oilers':'Edmonton',
                 'Vancouver Canucks':'Vancouver',
                 'Arizona Coyotes':'Phoenix'}

    nhl_df['Metropolitan area'] = pd.Series(team_dict)
    nhl_df = nhl_df.set_index('Metropolitan area')

    # combining the two dataframes
    joined_df = pd.merge(left = nhl_df, right = cities, how = 'left', on = 'Metropolitan area')
    joined_df= joined_df.iloc[:, [1,2,14]]
    joined_df.rename(columns = {'Population (2016 est.)[8]':'Population'}, inplace = True)
    joined_df['W'] = joined_df['W'].astype(float)
    joined_df['L'] = joined_df['L'].astype(float)
    joined_df['Population'] = joined_df['Population'].astype(float)

    # adding a column with the win to loss ratio and determining the mean values based on metropolitan area
    joined_df['W/L'] = joined_df['W']/(joined_df['W'] + joined_df['L'])
    joined_df = joined_df.groupby('Metropolitan area').mean()
```

```
    # determining the correlation between the win to loss ratio and population
    population_by_region = joined_df['Population']
    win_loss_by_region = joined_df['W/L']
    return stats.pearsonr(population_by_region, win_loss_by_region)[0]

nhl_correlation()

    #population_by_region = [] # pass in metropolitan area population from cities
    #win_loss_by_region = [] # pass in win/loss ratio from nhl_df in the same order as cities["Metropolitan area"]

    #assert len(population_by_region) == len(win_loss_by_region), "Q1: Your lists must be the same length"
    #assert len(population_by_region) == 28, "Q1: There should be 28 teams being analysed for NHL"

    #return stats.pearsonr(population_by_region, win_loss_by_region)
```

Out[9]: 0.012486162921209907

In [ ]:

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 [10]:
```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

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]]

cities.drop([16,17,19,20,21,22,23,26,29,30,31,34,35,36,37,39,40,43,44,47,48,49,50],0,inplace=True)

def nba_correlation():
    # opening & cleaning the cities dataframe
    cities=pd.read_html("assets/wikipedia_data.html")[1]
    cities=cities.iloc[:-1,[0,3,5,6,7,8]]
    cities.drop([16,17,19,20,21,22,23,26,29,30,31,34,35,36,37,39,40,43,44,47,48,49,50],0,inplace=True)
    cities['NBA'] = cities['NBA'].str.replace('\[\w*\s*\d*\]','')
    cities = cities.set_index('Metropolitan area')

    # opening & cleaning the nba_ft dataframe
    nba_df=pd.read_csv("assets/nba.csv")
    nba_df['team'] = nba_df['team'].str.replace('\*','')
    nba_df['team'] = nba_df['team'].str.replace('\(\d*\)','')
    nba_df['team'] = nba_df['team'].str.strip()
    nba_df = nba_df[nba_df['year'] == 2018]
    nba_df = nba_df.set_index('team')


    team_dict = {'Toronto Raptors':'Toronto',
                 'Boston Celtics':'Boston',
                 'Philadelphia 76ers':'Philadelphia',
                 'Cleveland Cavaliers':'Cleveland',
                 'Indiana Pacers':'Indianapolis',
                 'Miami Heat':'Miami-Fort Lauderdale',
                 'Milwaukee Bucks':'Milwaukee',
                 'Washington Wizards':'Washington, D.C.',
                 'Detroit Pistons':'Detroit',
                 'Charlotte Hornets':'Charlotte',
                 'New York Knicks':'New York City',
                 'Brooklyn Nets':'New York City',
                 'Chicago Bulls':'Chicago',
                 'Orlando Magic':'Orlando',
                 'Atlanta Hawks':'Atlanta',
                 'Houston Rockets':'Houston',
                 'Golden State Warriors':'San Francisco Bay Area',
                 'Portland Trail Blazers':'Portland',
                 'Oklahoma City Thunder':'Oklahoma City',
                 'Utah Jazz':'Salt Lake City',
                 'New Orleans Pelicans':'New Orleans',
                 'San Antonio Spurs':'San Antonio',
                 'Minnesota Timberwolves':'Minneapolis-Saint Paul',
                 'Denver Nuggets':'Denver',
                 'Los Angeles Clippers':'Los Angeles',
                 'Los Angeles Lakers':'Los Angeles',
                 'Sacramento Kings':'Sacramento',
                 'Dallas Mavericks':'Dallas-Fort Worth',
                 'Memphis Grizzlies':'Memphis',
                 'Phoenix Suns':'Phoenix'}

    nba_df['Metropolitan area'] = pd.Series(team_dict)
    nba_df = nba_df.set_index('Metropolitan area')

    # combining the two dataframes
    joined_df = pd.merge(left = nba_df, right = cities, how = 'left', on = 'Metropolitan area')
    joined_df= joined_df.iloc[:, [0,1,9]]
    joined_df.rename(columns = {'Population (2016 est.)[8]':'Population'}, inplace = True)
    joined_df['W'] = joined_df['W'].astype(float)
    joined_df['L'] = joined_df['L'].astype(float)
    joined_df['Population'] = joined_df['Population'].astype(float)

    # adding a column with the win to loss ratio and determining the mean values based on metropolitan area
    joined_df['W/L'] = joined_df['W']/(joined_df['W'] + joined_df['L'])
    joined_df = joined_df.groupby('Metropolitan area').mean()

    # determining the correlation between the win to loss ratio and population
    population_by_region = joined_df['Population']
    win_loss_by_region = joined_df['W/L']

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

nba_correlation()

    #raise NotImplementedError()
    #population_by_region = [] # pass in metropolitan area population from cities
    #win_loss_by_region = [] # pass in win/loss ratio from nba_df in the same order as cities["Metropolitan area"]

    #assert len(population_by_region) == len(win_loss_by_region), "Q2: Your lists must be the same length"
    #assert len(population_by_region) == 28, "Q2: There should be 28 teams being analysed for NBA"

    #return stats.pearsonr(population_by_region, win_loss_by_region)
```

Out[10]: -0.17657160252844617

In [ ]:

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 [11]:
```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

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]]

def mlb_correlation():
    # opening & cleaning the cities dataframe
    cities=pd.read_html("assets/wikipedia_data.html")[1]
    cities=cities.iloc[:-1,[0,3,5,6,7,8]]
    cities.drop([24,25,26,28,29,30,31,32,33,34,35,36,37,38,39,41,42,43,44,45,46,47,48,49,50],0,inplace=True)
    cities['MLB'] = cities['MLB'].str.replace('\[\w*\s*\d*\]','')
    cities = cities.set_index('Metropolitan area')

    # opening & cleaning the mbl_ft dataframe
    mlb_df=pd.read_csv("assets/mlb.csv")
    mlb_df = mlb_df[mlb_df['year'] == 2018]
    mlb_df = mlb_df.set_index('team')


    team_dict = {'Boston Red Sox':'Boston',
                 'New York Yankees':'New York City',
                 'Tampa Bay Rays':'Tampa Bay Area',
                 'Toronto Blue Jays':'Toronto',
                 'Baltimore Orioles':'Baltimore',
                 'Cleveland Indians':'Cleveland',
                 'Minnesota Twins':'Minneapolis-Saint Paul',
                 'Detroit Tigers':'Detroit',
                 'Chicago White Sox':'Chicago',
                 'Kansas City Royals':'Kansas City',
                 'Houston Astros':'Houston',
                 'Oakland Athletics':'San Francisco Bay Area',
                 'Seattle Mariners':'Seattle',
                 'Los Angeles Angels':'Los Angeles',
                 'Texas Rangers':'Dallas-Fort Worth',
                 'Atlanta Braves':'Atlanta',
                 'Washington Nationals':'Washington, D.C.',
                 'Philadelphia Phillies':'Philadelphia',
                 'New York Mets':'New York City',
                 'Miami Marlins':'Miami-Fort Lauderdale',
                 'Milwaukee Brewers':'Milwaukee',
                 'Chicago Cubs':'Chicago',
                 'St. Louis Cardinals':'St. Louis',
                 'Pittsburgh Pirates':'Pittsburgh',
                 'Cincinnati Reds':'Cincinnati',
                 'Los Angeles Dodgers':'Los Angeles',
                 'Colorado Rockies':'Denver',
                 'Arizona Diamondbacks':'Phoenix',
                 'San Francisco Giants':'San Francisco Bay Area',
                 'San Diego Padres':'San Diego'}

    mlb_df['Metropolitan area'] = pd.Series(team_dict)
    mlb_df = mlb_df.set_index('Metropolitan area')

    # combining the two dataframes
    joined_df = pd.merge(left = mlb_df, right = cities, how = 'left', on = 'Metropolitan area')
    joined_df= joined_df.iloc[:, [0,1,6]]
    joined_df.rename(columns = {'Population (2016 est.)[8]':'Population'}, inplace = True)
    joined_df['W'] = joined_df['W'].astype(float)
    joined_df['L'] = joined_df['L'].astype(float)
    joined_df['Population'] = joined_df['Population'].astype(float)

    # adding a column with the win to loss ratio and determining the mean values based on metropolitan area
    joined_df['W/L'] = joined_df['W']/(joined_df['W'] + joined_df['L'])
    joined_df = joined_df.groupby('Metropolitan area').mean()

    # determining the correlation between the win to loss ratio and population
    population_by_region = joined_df['Population']
    win_loss_by_region = joined_df['W/L']

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

mlb_correlation()

    #raise NotImplementedError()



    #population_by_region = [] # pass in metropolitan area population from cities
    #win_loss_by_region = [] # pass in win/loss ratio from mlb_df in the same order as cities["Metropolitan area"]

    #assert len(population_by_region) == len(win_loss_by_region), "Q3: Your lists must be the same length"
    #assert len(population_by_region) == 26, "Q3: There should be 26 teams being analysed for MLB"

    #return stats.pearsonr(population_by_region, win_loss_by_region)
```

Out[11]: 0.15027698302669307

In [ ]:
```

```

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 [12]:
```
import pandas as pd
import numpy as np
import scipy.stats as stats
import re

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 nfl_correlation():
    # opening & cleaning the cities dataframe
    cities=pd.read_html("assets/wikipedia_data.html")[1]
    cities=cities.iloc[:-1,[0,3,5,6,7,8]]
    cities.drop([13,22,27,30,31,32,33,34,35,36,37,38,39,40,41,42,43,45,46,47,49,50],0,inplace=True)
    cities['NFL'] = cities['NFL'].str.replace('\[\w*\s*\d*\]','')
    cities = cities.set_index('Metropolitan area')

    # opening & cleaning the nbl_ft dataframe
    nfl_df=pd.read_csv("assets/nfl.csv")
    nfl_df.drop([0,5,10,15,20,25,30,35],0,inplace=True)
```

```
    nfl_df['team'] = nfl_df['team'].str.replace('\*','')
    nfl_df['team'] = nfl_df['team'].str.replace('\+','')
    nfl_df = nfl_df[nfl_df['year'] == 2018]
    nfl_df = nfl_df.set_index('team')

    team_dict = {'New England Patriots':'Boston',
                 'Miami Dolphins':'Miami-Fort Lauderdale',
                 'Buffalo Bills':'Buffalo',
                 'New York Jets':'New York City',
                 'Baltimore Ravens':'Baltimore',
                 'Pittsburgh Steelers':'Pittsburgh',
                 'Cleveland Browns':'Cleveland',
                 'Cincinnati Bengals':'Cincinnati',
                 'Houston Texans':'Houston',
                 'Indianapolis Colts':'Indianapolis',
                 'Tennessee Titans':'Nashville',
                 'Jacksonville Jaguars':'Jacksonville',
                 'Kansas City Chiefs':'Kansas City',
                 'Los Angeles Chargers':'Los Angeles',
                 'Denver Broncos':'Denver',
                 'Oakland Raiders':'San Francisco Bay Area',
                 'Dallas Cowboys':'Dallas-Fort Worth',
                 'Philadelphia Eagles':'Philadelphia',
                 'Washington Redskins':'Washington, D.C.',
                 'New York Giants':'New York City',
                 'Chicago Bears':'Chicago',
                 'Minnesota Vikings':'Minneapolis-Saint Paul',
                 'Green Bay Packers':'Green Bay',
                 'Detroit Lions':'Detroit',
                 'New Orleans Saints':'New Orleans',
                 'Carolina Panthers':'Charlotte',
                 'Atlanta Falcons':'Atlanta',
                 'Tampa Bay Buccaneers':'Tampa Bay Area',
                 'Los Angeles Rams':'Los Angeles',
                 'Seattle Seahawks':'Seattle',
                 'San Francisco 49ers':'San Francisco Bay Area',
                 'Arizona Cardinals':'Phoenix'}

    nfl_df['Metropolitan area'] = pd.Series(team_dict)
    nfl_df = nfl_df.set_index('Metropolitan area')

    # combining the two dataframes
    joined_df = pd.merge(left = nfl_df, right = cities, how = 'left', on = 'Metropolitan area')
    joined_df= joined_df.iloc[:, [1,11,14]]
    joined_df.rename(columns = {'Population (2016 est.)[8]':'Population'}, inplace = True)
    joined_df['W'] = joined_df['W'].astype(float)
    joined_df['L'] = joined_df['L'].astype(float)
    joined_df['Population'] = joined_df['Population'].astype(float)

    # adding a column with the win to loss ratio and determining the mean values based on metropolitan area
    joined_df['W/L'] = joined_df['W']/(joined_df['W'] + joined_df['L'])
    joined_df = joined_df.groupby('Metropolitan area').mean()

    # determining the correlation between the win to loss ratio and population
    population_by_region = joined_df['Population']
    win_loss_by_region = joined_df['W/L']

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

nfl_correlation()

    #raise NotImplementedError()



    #population_by_region = [] # pass in metropolitan area population from cities
    #win_loss_by_region = [] # pass in win/loss ratio from nfl_df in the same order as cities["Metropolitan area"]

    #assert len(population_by_region) == len(win_loss_by_region), "Q4: Your lists must be the same length"
    #assert len(population_by_region) == 29, "Q4: There should be 29 teams being analysed for NFL"

    #return stats.pearsonr(population_by_region, win_loss_by_region)
```

Out[12]: 0.004922112149349393

In [ ]:

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 with a series of paired t-tests (so use `ttest_rel`) between all pairs of sports. Are there any sports where we can reject the null hypothesis? Again, average values where a 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 [13]:

```
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():
    def nhl_correlation():
        cities=pd.read_html("assets/wikipedia_data.html")[1]
        cities=cities.iloc[:-1,[0,3,5,6,7,8]]
        cities.drop([14,15,18,19,20,21,23,24,25,27,28,32,33,38,40,41,42,44,45,46,48,49,50],0,inplace=True)
        cities['NHL'] = cities['NHL'].str.replace('\[\w*\s*\d*\]','')
        cities = cities.set_index('Metropolitan area')

        nhl_df=pd.read_csv("assets/nhl.csv")
        nhl_df.drop([0,9,18,26],0,inplace=True)
        nhl_df['team'] = nhl_df['team'].str.replace('\*','')
        nhl_df = nhl_df[nhl_df['year'] == 2018]
        nhl_df = nhl_df.set_index('team')

        team_dict = {'Tampa Bay Lightning':'Tampa Bay Area',
                     'Boston Bruins':'Boston',
                     'Toronto Maple Leafs':'Toronto',
                     'Florida Panthers':'Miami-Fort Lauderdale',
                     'Detroit Red Wings':'Detroit',
                     'Montreal Canadiens':'Montreal',
                     'Ottawa Senators':'Ottawa',
                     'Buffalo Sabres':'Buffalo',
```

```
                    'Washington Capitals':'Washington, D.C.',
                    'Pittsburgh Penguins':'Pittsburgh',
                    'Philadelphia Flyers':'Philadelphia',
                    'Columbus Blue Jackets':'Columbus',
                    'New Jersey Devils':'New York City',
                    'Carolina Hurricanes':'Raleigh',
                    'New York Islanders':'New York City',
                    'New York Rangers':'New York City',
                    'Nashville Predators':'Nashville',
                    'Winnipeg Jets':'Winnipeg',
                    'Minnesota Wild':'Minneapolis-Saint Paul',
                    'Colorado Avalanche':'Denver',
                    'St. Louis Blues':'St. Louis',
                    'Dallas Stars':'Dallas-Fort Worth',
                    'Chicago Blackhawks':'Chicago',
                    'Vegas Golden Knights':'Las Vegas',
                    'Anaheim Ducks':'Los Angeles',
                    'San Jose Sharks':'San Francisco Bay Area',
                    'Los Angeles Kings':'Los Angeles',
                    'Calgary Flames':'Calgary',
                    'Edmonton Oilers':'Edmonton',
                    'Vancouver Canucks':'Vancouver',
                    'Arizona Coyotes':'Phoenix'}

    nhl_df['Metropolitan area'] = pd.Series(team_dict)
    nhl_df = nhl_df.set_index('Metropolitan area')
    nhl_df = nhl_df.iloc[:, [1,2]]
    nhl_df['W'] = nhl_df['W'].astype(float)
    nhl_df['L'] = nhl_df['L'].astype(float)
    nhl_df['W/L'] = nhl_df['W']/(nhl_df['W'] + nhl_df['L'])
    nhl_df = nhl_df.groupby('Metropolitan area').mean()
    return nhl_df

def nba_correlation():

        cities=pd.read_html("assets/wikipedia_data.html")[1]
        cities=cities.iloc[:-1,[0,3,5,6,7,8]]
        cities.drop([16,17,19,20,21,22,23,26,29,30,31,34,35,36,37,39,40,43,44,47,48,49,50],0,inplace=True)
        cities['NBA'] = cities['NBA'].str.replace('\[\w*\s*\d*\]','')
        cities = cities.set_index('Metropolitan area')

        nba_df = pd.read_csv("assets/nba.csv")
        nba_df['team'] = nba_df['team'].str.replace('\*','')
        nba_df['team'] = nba_df['team'].str.replace('\(\d*\)','')
        nba_df['team'] = nba_df['team'].str.strip()
        nba_df = nba_df[nba_df['year'] == 2018]
        nba_df = nba_df.set_index('team')

        team_dict = {'Toronto Raptors':'Toronto',
                     'Boston Celtics':'Boston',
                     'Philadelphia 76ers':'Philadelphia',
                     'Cleveland Cavaliers':'Cleveland',
                     'Indiana Pacers':'Indianapolis',
                     'Miami Heat':'Miami-Fort Lauderdale',
                     'Milwaukee Bucks':'Milwaukee',
                     'Washington Wizards':'Washington, D.C.',
                     'Detroit Pistons':'Detroit',
                     'Charlotte Hornets':'Charlotte',
                     'New York Knicks':'New York City',
                     'Brooklyn Nets':'New York City',
                     'Chicago Bulls':'Chicago',
                     'Orlando Magic':'Orlando',
                     'Atlanta Hawks':'Atlanta',
                     'Houston Rockets':'Houston',
                     'Golden State Warriors':'San Francisco Bay Area',
                     'Portland Trail Blazers':'Portland',
                     'Oklahoma City Thunder':'Oklahoma City',
                     'Utah Jazz':'Salt Lake City',
                     'New Orleans Pelicans':'New Orleans',
                     'San Antonio Spurs':'San Antonio',
                     'Minnesota Timberwolves':'Minneapolis-Saint Paul',
                     'Denver Nuggets':'Denver',
                     'Los Angeles Clippers':'Los Angeles',
                     'Los Angeles Lakers':'Los Angeles',
                     'Sacramento Kings':'Sacramento',
                     'Dallas Mavericks':'Dallas-Fort Worth',
                     'Memphis Grizzlies':'Memphis',
                     'Phoenix Suns':'Phoenix'}

        nba_df['Metropolitan area'] = pd.Series(team_dict)
        nba_df = nba_df.set_index('Metropolitan area')
        nba_df = nba_df.iloc[:, [0,1]]
        nba_df['W'] = nba_df['W'].astype(float)
        nba_df['L'] = nba_df['L'].astype(float)
        nba_df['W/L'] = nba_df['W']/(nba_df['W'] + nba_df['L'])
        nba_df = nba_df.groupby('Metropolitan area').mean()
        return nba_df


def mlb_correlation():

    cities=pd.read_html("assets/wikipedia_data.html")[1]
    cities=cities.iloc[:-1,[0,3,5,6,7,8]]
    cities.drop([24,25,26,28,29,30,31,32,33,34,35,36,37,38,39,41,42,43,44,45,46,47,48,49,50],0,inplace=True)
    cities['MLB'] = cities['MLB'].str.replace('\[\w*\s*\d*\]','')
    cities = cities.set_index('Metropolitan area')

    mlb_df=pd.read_csv("assets/mlb.csv")
    mlb_df = mlb_df[mlb_df['year'] == 2018]
    mlb_df = mlb_df.set_index('team')

    team_dict = {'Boston Red Sox':'Boston',
                 'New York Yankees':'New York City',
                 'Tampa Bay Rays':'Tampa Bay Area',
                 'Toronto Blue Jays':'Toronto',
                 'Baltimore Orioles':'Baltimore',
                 'Cleveland Indians':'Cleveland',
                 'Minnesota Twins':'Minneapolis-Saint Paul',
                 'Detroit Tigers':'Detroit',
                 'Chicago White Sox':'Chicago',
                 'Kansas City Royals':'Kansas City',
                 'Houston Astros':'Houston',
                 'Oakland Athletics':'San Francisco Bay Area',
                 'Seattle Mariners':'Seattle',
                 'Los Angeles Angels':'Los Angeles',
                 'Texas Rangers':'Dallas-Fort Worth',
                 'Atlanta Braves':'Atlanta',
                 'Washington Nationals':'Washington, D.C.',
                 'Philadelphia Phillies':'Philadelphia',
                 'New York Mets':'New York City',
                 'Miami Marlins':'Miami-Fort Lauderdale',
                 'Milwaukee Brewers':'Milwaukee',
                 'Chicago Cubs':'Chicago',
                 'St. Louis Cardinals':'St. Louis',
```

```
                    'Pittsburgh Pirates':'Pittsburgh',
                    'Cincinnati Reds':'Cincinnati',
                    'Los Angeles Dodgers':'Los Angeles',
                    'Colorado Rockies':'Denver',
                    'Arizona Diamondbacks':'Phoenix',
                    'San Francisco Giants':'San Francisco Bay Area',
                    'San Diego Padres':'San Diego'}

        mlb_df['Metropolitan area'] = pd.Series(team_dict)
        mlb_df = mlb_df.set_index('Metropolitan area')
        mlb_df= mlb_df.iloc[:, [0,1]]
        mlb_df['W'] = mlb_df['W'].astype(float)
        mlb_df['L'] = mlb_df['L'].astype(float)
        mlb_df['W/L'] = mlb_df['W']/(mlb_df['W'] + mlb_df['L'])
        mlb_df = mlb_df.groupby('Metropolitan area').mean()
        return mlb_df

    def nfl_correlation():

        cities = pd.read_html("assets/wikipedia_data.html")[1]
        cities = cities.iloc[:-1,[0,3,5,6,7,8]]
        cities.drop([13,22,27,30,31,32,33,34,35,36,37,38,39,40,41,42,43,45,46,47,49,50],0,inplace=True)
        cities['NFL'] = cities['NFL'].str.replace('\[\w*\s*\d*\]','')
        cities = cities.set_index('Metropolitan area')

        nfl_df = pd.read_csv("assets/nfl.csv")
        nfl_df.drop([0,5,10,15,20,25,30,35],0,inplace=True)
        nfl_df['team'] = nfl_df['team'].str.replace('\*','')
        nfl_df['team'] = nfl_df['team'].str.replace('\+','')
        nfl_df = nfl_df[nfl_df['year'] == 2018]
        nfl_df = nfl_df.set_index('team')

        team_dict = {'New England Patriots':'Boston',
                     'Miami Dolphins':'Miami-Fort Lauderdale',
                     'Buffalo Bills':'Buffalo',
                     'New York Jets':'New York City',
                     'Baltimore Ravens':'Baltimore',
                     'Pittsburgh Steelers':'Pittsburgh',
                     'Cleveland Browns':'Cleveland',
                     'Cincinnati Bengals':'Cincinnati',
                     'Houston Texans':'Houston',
                     'Indianapolis Colts':'Indianapolis',
                     'Tennessee Titans':'Nashville',
                     'Jacksonville Jaguars':'Jacksonville',
                     'Kansas City Chiefs':'Kansas City',
                     'Los Angeles Chargers':'Los Angeles',
                     'Denver Broncos':'Denver',
                     'Oakland Raiders':'San Francisco Bay Area',
                     'Dallas Cowboys':'Dallas-Fort Worth',
                     'Philadelphia Eagles':'Philadelphia',
                     'Washington Redskins':'Washington, D.C.',
                     'New York Giants':'New York City',
                     'Chicago Bears':'Chicago',
                     'Minnesota Vikings':'Minneapolis-Saint Paul',
                     'Green Bay Packers':'Green Bay',
                     'Detroit Lions':'Detroit',
                     'New Orleans Saints':'New Orleans',
                     'Carolina Panthers':'Charlotte',
                     'Atlanta Falcons':'Atlanta',
                     'Tampa Bay Buccaneers':'Tampa Bay Area',
                     'Los Angeles Rams':'Los Angeles',
                     'Seattle Seahawks':'Seattle',
                     'San Francisco 49ers':'San Francisco Bay Area',
                     'Arizona Cardinals':'Phoenix'}

        nfl_df['Metropolitan area'] = pd.Series(team_dict)
        nfl_df = nfl_df.set_index('Metropolitan area')
        nfl_df = pd.merge(left = nfl_df, right = cities, how = 'left', on = 'Metropolitan area')
        nfl_df = nfl_df.iloc[:, [1,11]]
        nfl_df['W'] = nfl_df['W'].astype(float)
        nfl_df['L'] = nfl_df['L'].astype(float)
        nfl_df['W/L'] = nfl_df['W']/(nfl_df['W'] + nfl_df['L'])
        nfl_df = nfl_df.groupby('Metropolitan area').mean()
        return nfl_df

    nhl_nba = pd.merge(left = nhl_correlation(), right = nba_correlation(), how = 'inner', on = 'Metropolitan area')
    nhl_mlb = pd.merge(left = nhl_correlation(), right = mlb_correlation(), how = 'inner', on = 'Metropolitan area')
    nhl_nfl = pd.merge(left = nhl_correlation(), right = nfl_correlation(), how = 'inner', on = 'Metropolitan area')
    nba_mlb = pd.merge(left = nba_correlation(), right = mlb_correlation(), how = 'inner', on = 'Metropolitan area')
    nba_nfl = pd.merge(left = nba_correlation(), right = nfl_correlation(), how = 'inner', on = 'Metropolitan area')
    mlb_nfl = pd.merge(left = mlb_correlation(), right = nfl_correlation(), how = 'inner', on = 'Metropolitan area')

    p_nhl_nba = stats.ttest_rel(nhl_nba['W/L_x'], nhl_nba['W/L_y'])[1]
    p_nhl_mlb = stats.ttest_rel(nhl_mlb['W/L_x'], nhl_mlb['W/L_y'])[1]
    p_nhl_nfl = stats.ttest_rel(nhl_nfl['W/L_x'], nhl_nfl['W/L_y'])[1]
    p_nba_mlb = stats.ttest_rel(nba_mlb['W/L_x'], nba_mlb['W/L_y'])[1]
    p_nba_nfl = stats.ttest_rel(nba_nfl['W/L_x'], nba_nfl['W/L_y'])[1]
    p_mlb_nfl = stats.ttest_rel(mlb_nfl['W/L_x'], mlb_nfl['W/L_y'])[1]

    sports = ['NFL', 'NBA', 'NHL', 'MLB']

    p_values = pd.DataFrame([[np.nan, p_nba_nfl, p_nhl_nfl, p_mlb_nfl],
                             [p_nba_nfl, np.nan, p_nhl_nba, p_nba_mlb],
                             [p_nhl_nfl, p_nhl_nba, np.nan, p_nhl_mlb],
                             [p_mlb_nfl, p_nba_mlb, p_nhl_mlb, np.nan]],
                            index = sports, columns = sports)
    return p_values

sports_team_performance()

    #raise NotImplementedError()

   # Note: p_values is a full dataframe, so df.loc["NFL","NBA"] should be the same as df.loc["NBA","NFL"] and
    # 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 around 0.02"
    #assert abs(p_values.loc["MLB", "NFL"] - 0.80) <= 1e-2, "The MLB-NFL p-value should be around 0.80"
    #return p_values
```

Out[13]:

	NFL	NBA	NHL	MLB
NFL	NaN	0.941792	0.030883	0.802069
NBA	0.941792	NaN	0.022297	0.950540
NHL	0.030883	0.022297	NaN	0.000708
MLB	0.802069	0.950540	0.000708	NaN

In [ ]: