Chongqing University of Technology
Md Anower Hossain(an hao ming)
Student ID: 62017010084
Assignment_5_Python
7th Semester

Sub: Big data technology and practice

In [3]:

```
import csv
import numpy as np
```

World Cup Data Analysis

In [43]:

```
import pandas as pd
df = pd.read_csv("players.csv")
df.head(2)
```

Out[43]:

	surname	team	position	minutes	shots	passes	tackles	saves
0	Abdoun	Algeria	midfielder	16	0	6	0	0
1	Belhadj	Algeria	defender	270	1	146	8	0

In [44]:

```
import pandas as pd
df = pd.read_csv("Teams.csv")
df.head(2)
```

Out[44]:

	team	ranking	games	wins	draws	losses	goalsFor	goalsAgainst	yellowCards	redCards
0	Brazil	1	5	3	1	1	9	4	7	2
1	Spain	2	6	5	0	1	7	2	3	0

In [45]:

```
import pandas as pd
df = pd.read_csv("playersExt.csv")
df.head(2)
```

Out[45]:

	surname	team	ranking	games	wins	draws	losses	goalsFor	goalsAgainst	yellowCards	redCards	position	minutes
0	Abdoun	Algeria	30	3	0	1	2	0	2	4	2	midfielder	16
1	Belhadj	Algeria	30	3	0	1	2	0	2	4	2	defender	270

Problem 1

What player on a team with "ia" in the team name played less than 200 minutes and made more than 100 passes? Print the player surname.

In [4]:

```
with open('players.csv','r') as f:
    rows = csv.DictReader(f)
    for data in rows:
        if 'ia' in data['team'] and (int(data['minutes']) < 200 and int(data['passes'])>10
0):
            print(data['surname'])
```

```
Kuzmanovic
```

Problem 2

Which team has the highest ratio of goalsFor to goalsAgainst? Print the team only.

In [5]:

```
with open('Teams.csv','r') as f:
    rows = csv.DictReader(f)
    for data in rows:
        ratio=int(data['goalsFor'])/int(data['goalsAgainst'])
        if ratio==7:
            print(data['team'])
```

```
Portugal
```

Problem 3

How many players on a team with ranking <10 played more than 350 minutes?

In [6]:

```
###### From PlayersExt File

with open('playersExt.csv','r') as f:
    rows = csv.DictReader(f)
    count = 0
    for data in rows:
        if int(data['ranking'])<10 and int(data['minutes'])>350:
            count = count+1
    print(count,'from one file named PlayersExt')
```

```
54 from one file named PlayersExt
```

another solution of problem 3

In [7]:

```
####### From Team and Players files

with open('Teams.csv','r') as t:
    rows = csv.DictReader(t)
    countRank = 0
    for data_t in rows:
        if int(data_t['ranking'])<10 :
            countRank+=1


with open('players.csv','r') as f:
    rows = csv.DictReader(f)
    countMinutes = 0
    for data in rows:
        if int(data['minutes'])>350:
            countMinutes = countMinutes+1
```

```
print(countRank+countMinutes,'Players ')
```

```
129 Players
```

Titanic Data Analysis

Read data from file Titanic.csv

Problem 1

Write a loop that asks the user to enter an age, then returns the number of married women over that age who embarked in Cherbourg. Terminate the loop when the user enters a number that is less than 0.

In [ ]:

```
import pandas as pd
df = pd.read_csv("Titanic.csv")
#df.head(5)
```

In [100]:

```
with open('Titanic.csv','r') as f:

    married_women = 0
    while True:
        df = pd.read_csv("C:\Titanic.csv")

        with open('Titanic.csv','r') as f:
            rows = csv.DictReader(f)

            input_age = input('Enter age: ')

            if input_age < "0":
                break
            married_women = 0
            for data in rows:
                if 'Mrs.' in data['first'] and 'Cherbourg' in data['embarked'] and data['a
ge']>input_age:
                    married_women = married_women+1
            print(married_women,'married women whose age more than',input_age,)
```

```
Enter age: 55
2 married women whose age more than 55
Enter age: 33
12 married women whose age more than 33
Enter age: 36
12 married women whose age more than 36
Enter age: 80
0 married women whose age more than 80
Enter age: 54
2 married women whose age more than 54
Enter age: -7
```

World Cup Data Visualization

Problem 1

Create a scatterplot of players showing passes made (y-axis) versus minutes played (x-axis). Color each player based on their position (goalkeeper, defender, midfielder, forward).

In [2]:

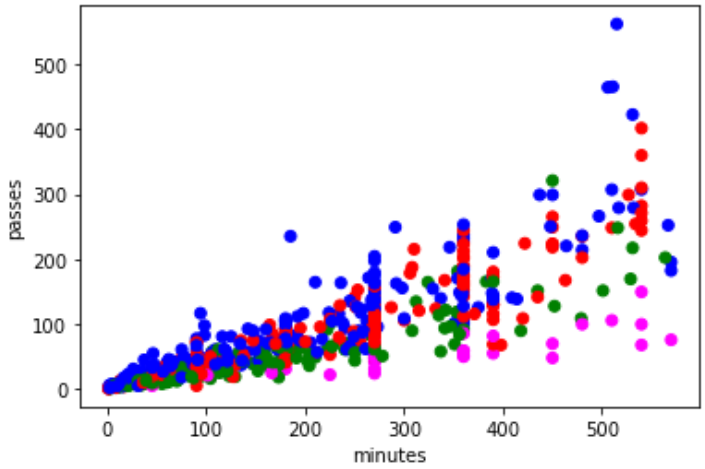
```
import pandas as pd
df = pd.read_csv("players.csv")
df.head(2)
```

Out[2]:

	surname	team	position	minutes	shots	passes	tackles	saves
0	Abdoun	Algeria	midfielder	16	0	6	0	0
1	Belhadj	Algeria	defender	270	1	146	8	0

In [4]:

```
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
df = pd.read_csv("players.csv")

with open('Players.csv','r') as p:
    rows = csv.DictReader(p)

    passes = []
    minutes = []
    colors = []
    for data_t in rows:
        passes.append(float(data_t['passes']))
        minutes.append(float(data_t['minutes']))

        if 'midfielder' in data_t['position'] :
            colors.append('blue')
        elif 'defender' in data_t['position'] :
            colors.append('red')
        elif 'goalkeeper' in data_t['position'] :
            colors.append('magenta')
        elif 'forward' in data_t['position'] :
            colors.append('green')
        else: colors.append('yellow')

    plt.xlabel('minutes')
    plt.ylabel('passes')
    plt.scatter(minutes,passes, c=colors)
    plt.show()
```



another solution of World Cup Data Visualization problem 1

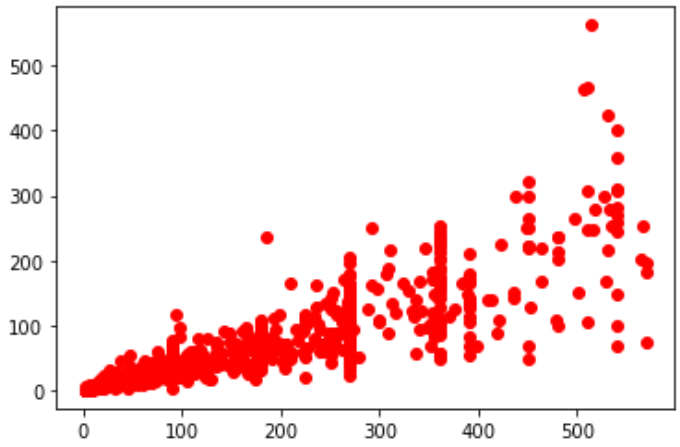
In [51]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
df_players = pd.read_csv('Players.csv')

newp = df_players
passes = newp.passes
minutes = newp.minutes
```

```
import matplotlib.cm as cm
plt.scatter(minutes,passes, color= 'red')
```

Out[51]:

```
<matplotlib.collections.PathCollection at 0x18053d8ffa0>
```



Problem 2

Create a pie chart showing the relative percentage of teams with 0, 1, and 2 red cards.

In [48]:

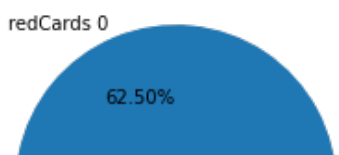
```
import pandas as pd
df = pd.read_csv("Teams.csv")
df.head(2)
```

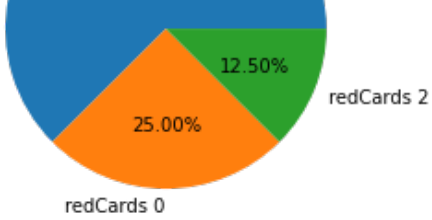
Out[48]:

	team	ranking	games	wins	draws	losses	goalsFor	goalsAgainst	yellowCards	redCards
0	Brazil	1	5	3	1	1	9	4	7	2
1	Spain	2	6	5	0	1	7	2	3	0

In [5]:

```
# Read Players.csv and Teams.csv into lists of dictionarie
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
df = pd.read_csv("Teams.csv")

with open('Teams.csv','r') as p:
    rows = csv.DictReader(p)

    redCards_0 = 0
    redCards_1 = 0
    redCards_2 = 0
    for data_readCards in rows:
        if int(data_readCards['redCards']) == 0 :
            redCards_0+=1
        elif int(data_readCards['redCards']) == 1 :
            redCards_1+=1
        elif int(data_readCards['redCards']) == 2 :
            redCards_2+=1

    plt.pie([redCards_0, redCards_1,redCards_2], labels=['redCards 0','redCards 0','redCards 2'], autopct='%1.2f%%')
    plt.show()
```





Titanic Data Visualization

Problem 1

Create a bar chart showing the average fare paid by passengers in each class. The three bars should be labeled 'first', 'second', 'third'.

In [36]:

```
import pandas as pd
df = pd.read_csv("Titanic.csv")
df.head(2)
```

Out[36]:

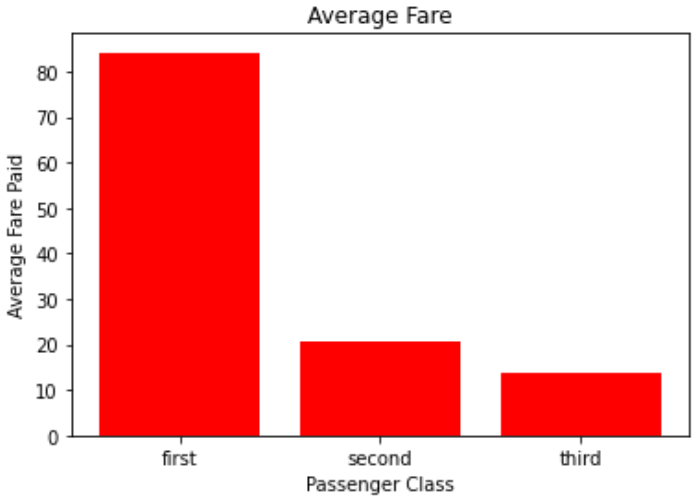
	last	first	gender	age	class	fare	embarked	survived
0	Braund	Mr. Owen Harris	M	22.0	3	7.2500	Southampton	no
1	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	F	38.0	1	71.2833	Cherbourg	yes

In [35]:

```
# Read Players.csv and Teams.csv into lists of dictionaries
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
df = pd.read_csv("Titanic.csv")

with open('Titanic.csv','r') as f:
    rows = csv.DictReader(f)

    bars = []
    heights = []
    for r in rows:
        class_sum = df.groupby('class')['fare'].apply(lambda x: x.sum())
        class_count = df.groupby('class')['fare'].apply(lambda x: x.count())

    print('sum of fare of seperate classes\n',class_sum)
    print('sum of classes attribute\n',class_count)

    heights = class_sum/class_count
    print('averages of fare by each classes\n',heights)

    bars = ['first', 'second', 'third']
    plt.xlabel("Passenger Class")
    plt.ylabel ("Average Fare Paid")
    plt.title("Average Fare")
    plt.bar(bars, heights, label = "Avg Fair Paid", color='red')
    plt.show()
```

```
sum of fare of seperate classes
 class
1    18177.4125
2     3801.8417
3     6714.6951
Name: fare, dtype: float64
sum of classes attribute
 class
1    216
2    184
```

```
3    491
Name: fare, dtype: int64
averages of fare by each classes
 class
1    84.154687
2    20.662183
3    13.675550
Name: fare, dtype: float64
```



World Cup with Pandas

In [1]:

```
import pandas as pd
f = open('Players.csv','r')
players = pd.read_csv(f)
f = open('Teams.csv','r')
teams = pd.read_csv(f)
```

Problem 1

What player on a team with "ia" in the team name played less than 200 minutes and made more than 100 passes? Print the player surname.

In [130]:

```
# if 'ia' in i and data_in_minutes<200:
import pandas as pd
df = pd.read_csv("players.csv")

data_in_minutes = df.minutes
data_in_passes = df.passes
data_in_team = df.team
data_in_surname = df.surname

ans = df[(data_in_minutes < 200) & (data_in_passes >100) & data_in_team.str.contains('ia')
& data_in_surname]
final_ans = ans.surname
print(final_ans)
```

```
431    Kuzmanovic
Name: surname, dtype: object
```

problem 1 another solution

In [131]:

```
with open('players.csv','r') as f:
    rows = csv.DictReader(f)
    for data in rows:
        if 'ia' in data['team'] and (int(data['minutes']) < 200 and int(data['passes'])>10
```

```
0):
            print(data['surname'])
```

```
Kuzmanovic
```

Titanic with Pandas

Problem 1

List the average fare paid by passengers in each of the embarkation cities.

In [132]:

```
import pandas as pd
Titanic_df = pd.read_csv("Titanic.csv")
ans = Titanic_df.groupby('embarked')['fare'].apply(lambda x: x.sum())

ans2 = Titanic_df.groupby('embarked')['fare'].apply(lambda x: x.count())

average_fare = ans/ans2
print('Listed the average fare paid by passengers in each of the embarkation cities',averag
e_fare)
```

```
Listed the average fare paid by passengers in each of the embarkation cities embarked
Cherbourg      59.954144
Queenstown     13.276030
Southampton    27.243651
Name: fare, dtype: float64
```

problem 1 another solution

In [83]:

```
# select embarked, avg(fare) from Titanic group by embarked order by avg(fare) desc
# avg = int(data['goalsFor'])/int(data['goalsAgainst'])

with open('Titanic.csv','r') as f:
    rows = csv.DictReader(f)
    df = pd.read_csv("Titanic.csv")

    Cherbourg_list = []
    count_Cherbourg =0

    Southampton_list = []
    count_Southampton = 0

    Queenstown_list = []
    count_Queenstown = 0


    #uniq = df.groupby('embarked')
    uniq = df.groupby('embarked')['fare'].apply(lambda x: x.sum())
    #uniq = df.groupby('embarked').apply(sum)

    print(uniq)

    print(' ')

    for data in rows:
        #fare = count(fare)

        if 'Cherbourg' in data['embarked']:
            count_Cherbourg = count_Cherbourg+1
            Cherbourg_list.append(float(data['fare']))

        elif 'Southampton' in data['embarked']:
            count_Southampton = count_Southampton+1
            Southampton_list.append(float(data['fare']))
```

```
        elif 'Queenstown' in data['embarked']:
            count_Queenstown = count_Queenstown+1
            Queenstown_list.append(float(data['fare']))


    Cherbourg_sum = 0
    Southampton_sum = 0
    Queenstown_sum =0


    for num in Cherbourg_list:
        Cherbourg_sum += float(num)
    Cherbourg_result = Cherbourg_sum/count_Cherbourg

    for num1 in Southampton_list:
        Southampton_sum += float(num1)
    Southampton_result = Southampton_sum/count_Southampton

    for num2 in Queenstown_list:
        Queenstown_sum += float(num2)
    Queenstown_result = Queenstown_sum/count_Queenstown

    print('Cherbourg avg fare  = ',Cherbourg_result)
    print('Southampton avg fare = ',Southampton_result)
    print('Queenstown avg fare = ',Queenstown_result)
```

```
embarked
Cherbourg      10072.2962
Queenstown      1022.2543
Southampton    17599.3988
Name: fare, dtype: float64

Cherbourg avg fare  =  59.95414404761905
Southampton avg fare =  27.243651393188795
Queenstown avg fare =  13.276029870129872
```