Timur Akdag

Sleep Efficiency

Q1:

```
import os

sleep_sheet = 'Sleep_Efficiency.csv'
input_dir = r'C:\Users\tiakd\.spyder-py3\Homework 4'
sleep_file = os.path.join(input_dir, sleep_sheet)

try:
    with open(sleep_file) as f:
        print('Q1')
        slp_dataframe = pd.read_csv(f)
```

I read the csv file into a Pandas dataframe using the pd.read_csv function.

Q2 and Q3:

In the code I combined answers for this question.

```
ID has 452 unique entries total.
ID has 0 entries missing

Age has 61 unique entries total.
Age has 0 entries missing

Gender has 2 unique entries total.
Gender has 0 entries missing

Bedtime has 424 unique entries total.
Bedtime has 0 entries missing

Wakeup time has 434 unique entries total.
Wakeup time has 0 entries missing

Sleep duration has 9 unique entries total.
Sleep duration has 0 entries missing

Sleep efficiency has 50 unique entries total.
Sleep efficiency has 0 entries missing

REM sleep percentage has 11 unique entries total.
REM sleep percentage has 0 entries missing

Deep sleep percentage has 18 unique entries total.
Deep sleep percentage has 0 entries missing

Light sleep percentage has 21 unique entries total.
Light sleep percentage has 0 entries missing
```

For most of the categories, no data is missing.

```
Awakenings has 6 unique entries total.
Awakenings has 20 entries missing

Caffeine consumption has 7 unique entries total.
Caffeine consumption has 25 entries missing

Alcohol consumption has 7 unique entries total.
Alcohol consumption has 16 entries missing

Smoking status has 2 unique entries total.
Smoking status has 0 entries missing

Exercise frequency has 7 unique entries total.
Exercise frequency has 6 entries missing
```

The last few categories of data are missing values.

Q4:

```
same_gen = slp_dataframe.groupby(['Gender','Age'])
for x in list_miss_col:
    series_null = np.where(slp_dataframe[x].isnull()==True)[0]

    for y in series_null:
        #Get the row where the data is null
        row = slp_dataframe.iloc[y]
        #Get the Age of Gender of that person
        age = row['Age']
        gen = row['Gender']

        vals_to_avg = []
        count = 0
        minus_checked = False
        while True:

            if (gen,age+count) in same_gen.groups.keys():
                temp = same_gen.get_group((gen,age+count))
            elif (gen,age-count) in same_gen.groups.keys():
                temp = same_gen.get_group((gen,age-count))
                minus_checked = True
            else:
                #The queried age does not exist for the person's gender.
                count+= 1
                continue

            non_nul = np.where(temp[x].isnull()==False)[0]

            if len(vals_to_avg) == 5:
                break

            for z in non_nul:
                #Z is the location of the non null value in temp
                vals_to_avg.append(temp.iloc[z][x])

                if len(vals_to_avg) == 5:
                    break
            if minus_checked and count > 0:
                count+= 1
                minus_checked = False
            else:
                count+=1
        #Broken out of while loop.  Assign the missing values
        slp_dataframe.at[y,x] = np.average(vals_to_avg)
```

This is the logic I have for question four. I attempted to get five people in age, and gender. The only flaw is which people get selected. The way I wrote this, the nearest neighbor by Age and Sex will always be the same people. For example, say the missing value is for someone that is female and age 40. I grab 5 people that are closest in age to this person. I average the values, and this is the value they receive. The problem would be if another female who is age 40 is missing the same data. The same 5 people I used the first time will be used again.

Q5:

Combined

Metrics	Group 1	Group 2	Group 3	Group 4	Group 5
Age	(10.5, 1.29)	(15.5, 1.29)	(25.8, 2.95)	(45.5, 7.86)	(64.9, 1.7)
Sleep duration	(8.38, 0.95)	(7.62, 0.75)	(7.5, 0.78)	(7.41, 0.89)	(7.52, 0.89)
Sleep efficiency	(0.54, 0.02)	(0.64, 0.04)	(0.77, 0.14)	(0.81, 0.13)	(0.76, 0.13)
REM sleep percentage	(18.0, 0.0)	(19.0, 2.0)	(23.75, 4.19)	(22.82, 3.87)	(22.67, 3.43)
Deep sleep percentage	(35.0, 0.0)	(32.0, 6.0)	(52.81, 16.59)	(54.18, 14.56)	(48.71, 17.47)
Light sleep percentage	(45.0, 0.0)	(47.25, 4.5)	(25.57, 15.53)	(23.27, 14.23)	(29.05, 17.7)
Awakenings	(2.5, 1.29)	(2.25, 0.96)	(1.66, 1.39)	(1.58, 1.31)	(2.14, 1.53)
Caffeine consumption	(0.0, 0.0)	(12.5, 25.0)	(31.86, 34.78)	(21.06, 27.46)	(21.43, 24.09)
Alcohol consumption	(2.75, 1.71)	(2.85, 2.18)	(1.42, 1.82)	(1.09, 1.49)	(1.05, 1.4)
Smoking status	(1.0, 0.0)	(0.0, 0.0)	(0.39, 0.49)	(0.34, 0.47)	(0.38, 0.5)
Exercise frequency	(0.0, 0.0)	(0.0, 0.0)	(1.82, 1.24)	(1.93, 1.47)	(1.48, 1.4)

Women

Metrics	Group 1	Group 2	Group 3	Group 4	Group 5
Age	(10.5, 1.29)	(15.5, 1.29)	(26.64, 2.95)	(41.6, 7.24)	(64.0, 1.26)
Sleep duration	(8.38, 0.95)	(7.62, 0.75)	(7.55, 0.87)	(7.44, 0.88)	(7.59, 0.83)
Sleep efficiency	(0.54, 0.02)	(0.64, 0.04)	(0.79, 0.14)	(0.81, 0.14)	(0.76, 0.15)
REM sleep percentage	(18.0, 0.0)	(19.0, 2.0)	(25.11, 3.25)	(23.37, 3.93)	(21.45, 2.84)
Deep sleep percentage	(35.0, 0.0)	(32.0, 6.0)	(53.98, 14.47)	(53.06, 15.25)	(46.27, 19.29)
Light sleep percentage	(45.0, 0.0)	(47.25, 4.5)	(23.83, 14.22)	(24.08, 15.3)	(31.55, 19.02)
Awakenings	(2.5, 1.29)	(2.25, 0.96)	(1.46, 1.35)	(1.52, 1.31)	(1.82, 1.6)
Caffeine consumption	(0.0, 0.0)	(12.5, 25.0)	(37.5, 24.18)	(29.54, 20.11)	(34.09, 23.11)
Alcohol consumption	(2.75, 1.71)	(2.85, 2.18)	(0.93, 1.59)	(1.17, 1.48)	(0.91, 1.58)
Smoking status	(1.0, 0.0)	(0.0, 0.0)	(0.39, 0.49)	(0.24, 0.43)	(0.27, 0.47)
Exercise frequency	(0.0, 0.0)	(0.0, 0.0)	(1.52, 1.33)	(1.7, 1.66)	(0.45, 0.93)

Men

Metrics	Group 1	Group 2	Group 3	Group 4	Group 5
Age	(0, 0)	(0, 0)	(24.82, 2.65)	(48.76, 6.82)	(65.9, 1.6)
Sleep duration	(0, 0)	(0, 0)	(7.44, 0.67)	(7.39, 0.91)	(7.45, 0.98)
Sleep efficiency	(0, 0)	(0, 0)	(0.76, 0.14)	(0.81, 0.12)	(0.75, 0.12)
REM sleep percentage	(0, 0)	(0, 0)	(22.16, 4.62)	(22.36, 3.77)	(24.0, 3.65)
Deep sleep percentage	(0, 0)	(0, 0)	(51.43, 18.84)	(55.12, 13.94)	(51.4, 15.78)
Light sleep percentage	(0, 0)	(0, 0)	(27.62, 16.85)	(22.59, 13.27)	(26.3, 16.67)
Awakenings	(0, 0)	(0, 0)	(1.89, 1.42)	(1.63, 1.32)	(2.5, 1.43)
Caffeine consumption	(0, 0)	(0, 0)	(25.21, 43.43)	(13.97, 30.65)	(7.5, 16.87)
Alcohol consumption	(0, 0)	(0, 0)	(2.01, 1.93)	(1.02, 1.51)	(1.2, 1.23)
Smoking status	(0, 0)	(0, 0)	(0.38, 0.49)	(0.43, 0.5)	(0.5, 0.53)
Exercise frequency	(0, 0)	(0, 0)	(2.17, 1.03)	(2.12, 1.27)	(2.6, 0.84)

Q6:

Female children ages 1-12 sleep the most on average. Men ages 31-60 sleep the least on average. The most awakenings are tied between female children 1-12 and male older adults ages 61 and higher. The least awakenings are women ages 18-30. The data set is lacking data for men groups 1-2, therefore they are not eligible.

Q7:

The maximum sleep efficiency on average is tied. Interestingly the tie is between men and women ages 31-60. The minimum sleep efficiency on average is female children ages 1-12. The maximum deep sleep percentage on average was men ages 31-60. The minimum deep sleep percentage on average was women ages 13-17.

Q8:

Exercise does not seem to have a large effect on sleep. Groups 1 and 2 do not have any exercise data, so we do not analyze those groups. Group 3 exercises for 1.82 hours on average, with a sleep duration of 7.5 Group 4 exercises for 1.93 hours, with a sleep duration of 7.41. Group 5 exercises for 1.48 hours, with a sleep duration of 7.52. The differences between these values is small. I do not think exercise has a large impact on sleep duration.

Q9:

This answer is fairly surprising. I do not see a major difference on average between smoking status and sleep duration. In a similar way to question 8 the differences between sleep duration and smoking status is small. I converted smoking status where Yes is equal to 1.

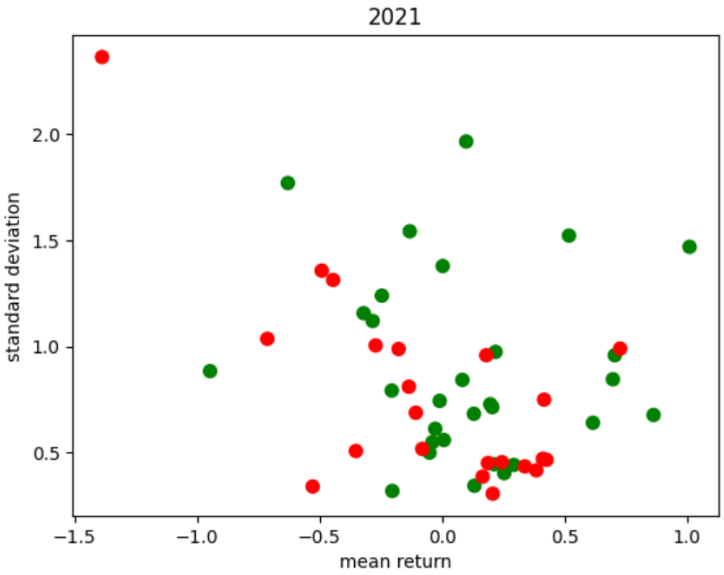
Q10:

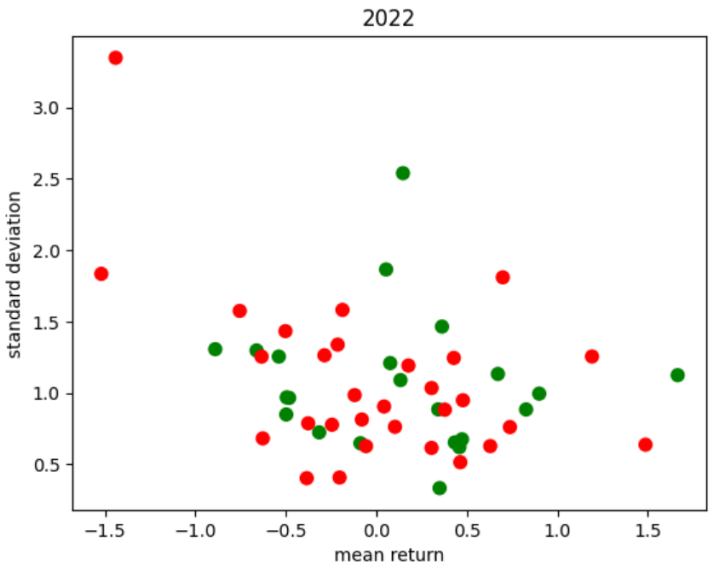
I will start by looking at children ages 1-12. This group on average sleeps longer than all other groups. It is very interesting comparing the sleep data between children and teenagers. The average REM, deep and light sleep percentage is very close between groups. I am surprised to see that children have any value in alcohol and smoking status. Looking at the table from young adults to older adults the sleep patterns are fairly similar in sleep percentage. The largest percentage is deep sleep for all of these groups at around 52%.

There are a couple of surprising takeaways from this table. Caffeine consumption does not seem to have a significant impact on sleep numbers. Consumption of caffeine spikes drastically from teenagers to young adults. From young adults to older adults caffeine consumption is over 20. If you look at sleep duration between young adults and groups 3-5, the difference is small. The only major difference is sleep efficiency. Young adults have lower sleep efficiency than the older groups. This is despite the older groups having more caffeine.

Labels

Examine_labels





Q1: One pattern I see is in general a green dot will appear when the mean return is between -0.5 and higher. There are few instances where a green point appears when the mean return is less than -0.5. There is not a clear pattern otherwise, the green and red points are very entwined.

Q2: Points of the same color are lumped fairly close together. The green has more instances where the points are scattered. Unlike red which is a bit more condensed.

Q3: The pattern that repeats from year 1 to year 2 is when the mean return is -0.5 and higher. Similar to year 1 the points in year 2 are intermingled with each other.

Q4: I do not believe the nearest-neighbor classifier will perform well. The cluster of points is condensed around each other. There are also a lot more green points in year 1 compared to year 2. Since there is not a clear distinction between year 1 and year 2 I think nearest-neighbor will not perform well.

Weekly Labeling

Q1:

```
In [16]: hold_shares = 100 / all_grp.get_group((2022,1)).iloc[0]['Open']
         lst_week =   all_grp.get_group((2022,52))
         hold_profit = hold_shares * lst_week.iloc[len(lst_week)-1]['Adj Close']
```

all_grp is a group by year and week. I got the opening price for the first day of the week in 2022. I then got the adjusted closing price of the last day of the week in 2022. The final profit is the number of shares bought originally times the closing price.

Q2:

```
In [19]: print('Q2')
         if yearly_profit[2022] > hold_profit:
             print('Label strategy better')
         elif yearly_profit[2022] == hold_profit:
             print('Buy and hold is as good as label strategy')
         else:
             print('Buy and hold results in a larger amount')
```

```
Q2
Buy and hold results in a larger amount
```

In the algorithm for calculating the profit, I created a dictionary. The dictionary is based on year and the value is profit made based on the strategy. Buy and hold performed better than using the labels.