ISM 4403 Homework Week 14

### Tasks:

Create a new Excel spreadsheet from the following table and export it as a CSV file without quotes.

**Chart 14.1**

**Data Set 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | Height (inches) | gender | Hair color | Eye Color | Age |
| 1 | 67 | male | brown | brown | 25 |
| 2 | 64 | female | brown | green | 23 |
| 3 | 74 | male | blond | blue | 27 |
| 4 | 73 | Male | brown | brown | 35 |
| 5 | 60 | female | red | green | 40 |
| 6 | 61 | female | brown | green | 45 |
| 7 | 73 | female | blond | blue |  |
| 8 | 70 | female | brown | blue | 50 |
| 9 | 56 | female | blond | brown | 60 |
| 10 | 57 | male | blond | brown | 18 |
| 11 | 64 |  | brown | brown | 25 |
| 12 | 69 | male | brown | green | 23 |
| 13 | 69 | female | blond |  | 27 |
| 14 | 70 | female | brown | brown | 35 |
| 15 | 71 | female | red | green | 40 |
| 16 | 60 | female |  | green | 45 |
| 17 | 80 | male | blond | blue | 41 |
| 18 | 75 | male | brown | blue | 50 |
| 19 | 78 | male | blond | brown | 60 |
| 20 | 69 | male | blond | brown | 18 |
| 21 |  | female | brown | brown | 25 |
| 22 | 66 | male | brown | green | 23 |
| 23 | 74 | female | blond | blue | 27 |
| 24 | 72 | male | brown | brown | 35 |
| 25 | 68 | female | red | green | 40 |
| 26 | 64 | female | brown | green | 45 |
| 27 | 63 | female | blond | blue | 41 |
| 28 |  | male | brown | blue | 50 |
| 29 | 70 | male | blond | brown | 60 |
| 30 | 62 | male | blond | brown | 18 |

**Data Set 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | Height (inches) | Hair color | Eye Color | Age |
| 1 | 73 | blond | brown | 25 |
| 2 | 70 | brown | green | 23 |
| 3 | 56 | blond | blue | 27 |
| 4 | 57 | brown | brown |  |
| 5 | 64 | red | green | 35 |
| 6 | 69 | brown | green | 40 |
| 7 | 69 | blond | blue | 45 |
| 8 | 67 | brown | blue | 50 |
| 9 | 64 | blond | brown | 60 |
| 10 | 74 | blond | brown | 18 |
| 11 | 73 | brown | brown | 25 |
| 12 | 60 | brown | green | 23 |
| 13 | 61 | blond |  | 27 |
| 14 | 70 | brown | brown | 35 |
| 15 | 71 |  | green | 40 |
| 16 | 60 | red | green | 45 |
| 17 | 80 | blond | blue | 41 |
| 18 | 75 | brown | blue | 50 |
| 19 | 78 | brown | brown | 60 |
| 20 | 69 | brown | brown | 18 |
| 21 | 66 | brown | brown | 25 |
| 22 | 74 | blond | green | 23 |
| 23 |  | blond | blue | 27 |

Using Jupyter Notebooks

Import the the sheet into a data structure in Python.

1. Import the two data sheets into Python merging them into a single sheet.

Please note you must handle the fact that one data sheet is missing a variable. Please generate data to replace it as gender is an import variable. This data should be generated using your best judgement as a data scientist.

1. Convert the structure to a DataFrame. Use this data structure for the remainder of this assignment.
2. Using pandas functions resolve any issues with missing data as done in previous homeworks (4).
3. Using pandas functions calculate the following.
   1. The mean height.
   2. The sum of all heights.
   3. The mode for gender, hair color, and eye color.

**Paste your code here**

# lab 13

import pandas as pd

import numpy as np

df1 = pd.read\_csv('flat\_file\_df1.csv')

df2 = pd.read\_csv('flat\_file\_df2.csv')

# had to drop gender because df2 is missing the whole column

df1.drop( columns='gender', inplace = True)

# now we must join them

df = pd.merge(df1, df2, on = 'age', how = 'left')

# hair\_color

df['hair\_color\_x'].value\_counts()

df['hair\_color\_x'].fillna(value = 'brown', inplace = True)

# eye\_color

df['eye\_color\_x'].value\_counts()

df['eye\_color\_x'].fillna(value = 'brown', inplace = True)

# height

print(df.height\_x)

df['height\_x'].fillna(df.height\_x.mean, inplace = True)

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height\_x'].sum()

df.median()

df.mode()

df.describe()

df['hair\_color\_x'].mode()

df['eye\_color\_x'].mode()

**End of Paste**

# eye\_color

df['eye\_color\_x'].value\_counts()

df['eye\_color\_x'].fillna(value = 'brown', inplace = True)

# height

print(df.height\_x)

df['height\_x'].fillna(df.height.mean, inplace = True)

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height\_x'].sum()

df.median()

df.mode()

df.describe()

df['hair\_color\_x'].mode()

df['eye\_color\_x'].mode()

0 67.0

1 67.0

2 67.0

3 64.0

4 64.0

5 64.0

6 74.0

7 74.0

8 74.0

9 73.0

10 73.0

11 60.0

12 60.0

13 61.0

14 61.0

15 73.0

16 70.0

17 70.0

18 56.0

19 56.0

20 57.0

21 57.0

22 64.0

23 64.0

24 64.0

25 69.0

26 69.0

27 69.0

28 69.0

29 69.0

36 60.0

37 80.0

38 75.0

39 75.0

40 78.0

41 78.0

42 69.0

43 69.0

44 NaN

45 NaN

46 NaN

47 66.0

48 66.0

49 66.0

50 74.0

51 74.0

52 74.0

53 72.0

54 72.0

55 68.0

56 68.0

57 64.0

58 64.0

59 63.0

60 NaN

61 NaN

62 70.0

63 70.0

64 62.0

65 62.0

Name: height\_x, Length: 66, dtype: float64

Traceback (most recent call last):

File "<ipython-input-19-8d94d85b9bd8>", line 12, in <module>

df['height\_x'].fillna(df.height.mean, inplace = True)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\generic.py", line 5067, in \_\_getattr\_\_

return object.\_\_getattribute\_\_(self, name)

AttributeError: 'DataFrame' object has no attribute 'height'

df['hair\_color\_x'].value\_counts()

df['hair\_color\_x'].fillna(value = 'brown', inplace = True)

# eye\_color

df['eye\_color\_x'].value\_counts()

df['eye\_color\_x'].fillna(value = 'brown', inplace = True)

# height

print(df.height\_x)

df['height\_x'].fillna(df.height\_x.mean, inplace = True)

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height\_x'].sum()

df.median()

df.mode()

df.describe()

df['hair\_color\_x'].mode()

df['eye\_color\_x'].mode()

0 67.0

1 67.0

2 67.0

3 64.0

4 64.0

5 64.0

6 74.0

7 74.0

8 74.0

9 73.0

10 73.0

11 60.0

12 60.0

13 61.0

14 61.0

15 73.0

16 70.0

17 70.0

18 56.0

19 56.0

20 57.0

21 57.0

22 64.0

23 64.0

24 64.0

25 69.0

26 69.0

27 69.0

28 69.0

29 69.0

36 60.0

37 80.0

38 75.0

39 75.0

40 78.0

41 78.0

42 69.0

43 69.0

44 NaN

45 NaN

46 NaN

47 66.0

48 66.0

49 66.0

50 74.0

51 74.0

52 74.0

53 72.0

54 72.0

55 68.0

56 68.0

57 64.0

58 64.0

59 63.0

60 NaN

61 NaN

62 70.0

63 70.0

64 62.0

65 62.0

Name: height\_x, Length: 66, dtype: float64

Traceback (most recent call last):

File "<ipython-input-20-36bc77d62ad6>", line 20, in <module>

df['height\_x'].sum()

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\generic.py", line 10931, in stat\_func

numeric\_only=numeric\_only, min\_count=min\_count)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\series.py", line 3630, in \_reduce

return op(delegate, skipna=skipna, \*\*kwds)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 76, in \_f

return f(\*args, \*\*kwargs)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 435, in nansum

the\_sum = values.sum(axis, dtype=dtype\_sum)

File "C:\Users\Owner\Anaconda3\lib\site-packages\numpy\core\\_methods.py", line 38, in \_sum

return umr\_sum(a, axis, dtype, out, keepdims, initial, where)

TypeError: unsupported operand type(s) for +: 'float' and 'method'

df['hair\_color\_x'].mode()

Out[21]:

0 brown

dtype: object

df['eye\_color\_x'].mode()

Out[22]:

0 brown

dtype: object

df.describe()

Out[23]:

id\_x age id\_y height\_y

count 66.000000 65.000000 66.000000 63.000000

mean 14.954545 34.538462 12.166667 68.349206

std 8.757295 12.756220 6.657750 6.154355

min 1.000000 18.000000 1.000000 56.000000

25% 8.000000 25.000000 7.000000 64.000000

50% 14.500000 27.000000 12.000000 69.000000

75% 22.000000 45.000000 18.000000 73.000000

max 30.000000 60.000000 23.000000 80.000000

df.isnull().sum()

Out[24]:

id\_x 0

height\_x 0

hair\_color\_x 0

eye\_color\_x 0

age 1

id\_y 0

height\_y 3

hair\_color\_y 3

eye\_color\_y 3

dtype: int64

import pandas as pd

import numpy as np

df1 = pd.read\_csv('flat\_file\_df1.csv')

df2 = pd.read\_csv('flat\_file\_df2.csv')

# had to drop gender because df2 is missing the whole column

df1.drop( columns='gender', inplace = True)

# now we must join them

df = pd.merge(df1, df2, on = 'age', how = 'left')

# hair\_color

df['hair\_color\_x'].value\_counts()

df['hair\_color\_x'].fillna(value = 'brown', inplace = True)

# eye\_color

df['eye\_color\_x'].value\_counts()

df['eye\_color\_x'].fillna(value = 'brown', inplace = True)

# height

print(df.height\_x)

df['height\_x'].fillna(df.height\_x.mean, inplace = True)

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height\_x'].sum()

df.median()

df.mode()

df.describe()

df['hair\_color\_x'].mode()

df['eye\_color\_x'].mode()

0 67.0

1 67.0

2 67.0

3 64.0

4 64.0

5 64.0

6 74.0

7 74.0

8 74.0

9 73.0

10 73.0

11 60.0

12 60.0

13 61.0

14 61.0

15 73.0

16 70.0

17 70.0

18 56.0

19 56.0

20 57.0

21 57.0

22 64.0

23 64.0

24 64.0

25 69.0

26 69.0

27 69.0

28 69.0

29 69.0

36 60.0

37 80.0

38 75.0

39 75.0

40 78.0

41 78.0

42 69.0

43 69.0

44 NaN

45 NaN

46 NaN

47 66.0

48 66.0

49 66.0

50 74.0

51 74.0

52 74.0

53 72.0

54 72.0

55 68.0

56 68.0

57 64.0

58 64.0

59 63.0

60 NaN

61 NaN

62 70.0

63 70.0

64 62.0

65 62.0

Name: height\_x, Length: 66, dtype: float64

Traceback (most recent call last):

File "<ipython-input-25-9be96943187e>", line 33, in <module>

df['height\_x'].sum()

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\generic.py", line 10931, in stat\_func

numeric\_only=numeric\_only, min\_count=min\_count)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\series.py", line 3630, in \_reduce

return op(delegate, skipna=skipna, \*\*kwds)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 76, in \_f

return f(\*args, \*\*kwargs)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 435, in nansum

the\_sum = values.sum(axis, dtype=dtype\_sum)

File "C:\Users\Owner\Anaconda3\lib\site-packages\numpy\core\\_methods.py", line 38, in \_sum

return umr\_sum(a, axis, dtype, out, keepdims, initial, where)

TypeError: unsupported operand type(s) for +: 'float' and 'method'

Removing all variables...

import pandas as pd

import numpy as np

df1 = pd.read\_csv('flat\_file\_df1.csv')

df2 = pd.read\_csv('flat\_file\_df2.csv')

# had to drop gender because df2 is missing the whole column

df1.drop( columns='gender', inplace = True)

# now we must join them

df = pd.merge(df1, df2, on = 'age', how = 'left')

# hair\_color

df['hair\_color\_x'].value\_counts()

df['hair\_color\_x'].fillna(value = 'brown', inplace = True)

# eye\_color

df['eye\_color\_x'].value\_counts()

df['eye\_color\_x'].fillna(value = 'brown', inplace = True)

# height

print(df.height\_x)

df['height\_x'].fillna(df.height\_x.mean, inplace = True)

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height\_x'].sum()

df.median()

df.mode()

df.describe()

df['hair\_color\_x'].mode()

df['eye\_color\_x'].mode()

0 67.0

1 67.0

2 67.0

3 64.0

4 64.0

5 64.0

6 74.0

7 74.0

8 74.0

9 73.0

10 73.0

11 60.0

12 60.0

13 61.0

14 61.0

15 73.0

16 70.0

17 70.0

18 56.0

19 56.0

20 57.0

21 57.0

22 64.0

23 64.0

24 64.0

25 69.0

26 69.0

27 69.0

28 69.0

29 69.0

36 60.0

37 80.0

38 75.0

39 75.0

40 78.0

41 78.0

42 69.0

43 69.0

44 NaN

45 NaN

46 NaN

47 66.0

48 66.0

49 66.0

50 74.0

51 74.0

52 74.0

53 72.0

54 72.0

55 68.0

56 68.0

57 64.0

58 64.0

59 63.0

60 NaN

61 NaN

62 70.0

63 70.0

64 62.0

65 62.0

Name: height\_x, Length: 66, dtype: float64

Traceback (most recent call last):

File "<ipython-input-26-9be96943187e>", line 33, in <module>

df['height\_x'].sum()

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\generic.py", line 10931, in stat\_func

numeric\_only=numeric\_only, min\_count=min\_count)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\series.py", line 3630, in \_reduce

return op(delegate, skipna=skipna, \*\*kwds)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 76, in \_f

return f(\*args, \*\*kwargs)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\core\nanops.py", line 435, in nansum

the\_sum = values.sum(axis, dtype=dtype\_sum)

File "C:\Users\Owner\Anaconda3\lib\site-packages\numpy\core\\_methods.py", line 38, in \_sum

return umr\_sum(a, axis, dtype, out, keepdims, initial, where)

TypeError: unsupported operand type(s) for +: 'float' and 'method'

**Paste your results here**

**End of Paste**