ISM 4403 Homework Week 13

### Tasks:

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

**Chart 13.1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

Using Jupyter Notebooks

Import the the sheet into a data structure in Python.

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**

import pandas as pd

import numpy as np

df = pd.read\_csv('lab12\_flat\_file.csv')

# gender

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# height

print(df.height)

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

**End of Paste**

**Paste your results here**

# hair\_color

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

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

# eye\_color

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

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

<bound method DataFrame.to\_numpy of id height gender hair\_color eye\_color age

0 1 67.0 male brown brown 25.0

1 2 64.0 female brown green 23.0

2 3 74.0 male blond blue 27.0

3 4 73.0 Male brown brown 35.0

4 5 60.0 female red green 40.0

5 6 61.0 female brown green 45.0

6 7 73.0 female blond blue NaN

7 8 70.0 female brown blue 50.0

8 9 56.0 female blond brown 60.0

9 10 57.0 male blond brown 18.0

10 11 64.0 NaN brown brown 25.0

11 12 69.0 male brown green 23.0

12 13 69.0 female blond NaN 27.0

13 14 70.0 female brown brown 35.0

14 15 71.0 female red green 40.0

15 16 60.0 female NaN green 45.0

16 17 80.0 male blond blue 41.0

17 18 75.0 male brown blue 50.0

18 19 78.0 male blond brown 60.0

19 20 69.0 male blond brown 18.0

20 21 NaN female brown brown 25.0

21 22 66.0 male brown green 23.0

22 23 74.0 female blond blue 27.0

23 24 72.0 male brown brown 35.0

24 25 68.0 female red green 40.0

25 26 64.0 female brown green 45.0

26 27 63.0 female blond blue 41.0

27 28 NaN male brown blue 50.0

28 29 70.0 male blond brown 60.0

29 30 62.0 male blond brown 18.0>

Traceback (most recent call last):

File "<ipython-input-2-6000892294d5>", line 8, in <module>

np.NAN(dfa)

TypeError: 'float' object is not callable

df['gender'].mode()

Out[3]:

0 female

dtype: object

import pandas as pd

import numpy as np

df = pd.read\_csv('lab11\_flat\_file.csv')

# gender

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

Out[4]:

0 brown

dtype: object

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

Out[6]:

id height age

count 30.000000 28.000000 29.000000

mean 15.500000 67.821429 36.241379

std 8.803408 6.116484 13.015527

min 1.000000 56.000000 18.000000

25% 8.250000 63.750000 25.000000

50% 15.500000 69.000000 35.000000

75% 22.750000 72.250000 45.000000

max 30.000000 80.000000 60.000000

df['gender'].mode()

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

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

Out[7]:

0 brown

dtype: object

df['height'].sum()

Out[8]: 1899.0

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

File "<ipython-input-9-857b029b6ebc>", line 1

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

^

SyntaxError: invalid syntax

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

Traceback (most recent call last):

File "<ipython-input-10-3aef3a6d4f73>", line 1, in <module>

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

NameError: name 'height' is not defined

print(df.height)

0 67.0

1 64.0

2 74.0

3 73.0

4 60.0

5 61.0

6 73.0

7 70.0

8 56.0

9 57.0

10 64.0

11 69.0

12 69.0

13 70.0

14 71.0

15 60.0

16 80.0

17 75.0

18 78.0

19 69.0

20 NaN

21 66.0

22 74.0

23 72.0

24 68.0

25 64.0

26 63.0

27 NaN

28 70.0

29 62.0

Name: height, dtype: float64

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

File "<ipython-input-12-af911022e85f>", line 1

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

^

SyntaxError: EOL while scanning string literal

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

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

Removing all variables...

import pandas as pd

import numpy as np

df = pd.read\_csv('lab11\_flat\_file.csv')

# gender

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# height

print(df.height)

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

Traceback (most recent call last):

File "<ipython-input-15-d6bcec6c640d>", line 4, in <module>

df = pd.read\_csv('lab11\_flat\_file.csv')

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\io\parsers.py", line 702, in parser\_f

return \_read(filepath\_or\_buffer, kwds)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\io\parsers.py", line 429, in \_read

parser = TextFileReader(filepath\_or\_buffer, \*\*kwds)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\io\parsers.py", line 895, in \_\_init\_\_

self.\_make\_engine(self.engine)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\io\parsers.py", line 1122, in \_make\_engine

self.\_engine = CParserWrapper(self.f, \*\*self.options)

File "C:\Users\Owner\Anaconda3\lib\site-packages\pandas\io\parsers.py", line 1853, in \_\_init\_\_

self.\_reader = parsers.TextReader(src, \*\*kwds)

File "pandas/\_libs/parsers.pyx", line 387, in pandas.\_libs.parsers.TextReader.\_\_cinit\_\_

File "pandas/\_libs/parsers.pyx", line 705, in pandas.\_libs.parsers.TextReader.\_setup\_parser\_source

FileNotFoundError: [Errno 2] File b'lab11\_flat\_file.csv' does not exist: b'lab11\_flat\_file.csv'

import pandas as pd

import numpy as np

df = pd.read\_csv('lab12\_flat\_file.csv')

# gender

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# height

print(df.height)

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

0 67.0

1 64.0

2 74.0

3 73.0

4 60.0

5 61.0

6 73.0

7 70.0

8 56.0

9 57.0

10 64.0

11 69.0

12 69.0

13 70.0

14 71.0

15 60.0

16 80.0

17 75.0

18 78.0

19 69.0

20 NaN

21 66.0

22 74.0

23 72.0

24 68.0

25 64.0

26 63.0

27 NaN

28 70.0

29 62.0

Name: height, dtype: float64

Traceback (most recent call last):

File "<ipython-input-16-e2b4eb619f6d>", line 33, in <module>

df['height'].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...

Removing all variables...

Removing all variables...

import pandas as pd

import numpy as np

df = pd.read\_csv('lab12\_flat\_file.csv')

# gender

df['gender'].value\_counts()

df['gender'].replace('Male', 'male', inplace = True)

df['gender'].fillna(value = 'female', inplace = True)

# hair\_color

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

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

# eye\_color

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

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

# height

print(df.height)

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

# is the missing data gone?

df.isnull().sum()

# yes

# part two

df.mean()

df['height'].sum()

df.median()

df.mode()

df.describe()

df['gender'].mode()

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

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

0 67.0

1 64.0

2 74.0

3 73.0

4 60.0

5 61.0

6 73.0

7 70.0

8 56.0

9 57.0

10 64.0

11 69.0

12 69.0

13 70.0

14 71.0

15 60.0

16 80.0

17 75.0

18 78.0

19 69.0

20 NaN

21 66.0

22 74.0

23 72.0

24 68.0

25 64.0

26 63.0

27 NaN

28 70.0

29 62.0

Name: height, dtype: float64

Traceback (most recent call last):

File "<ipython-input-17-e2b4eb619f6d>", line 33, in <module>

df['height'].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'

**End of Paste**

**Rubric**

20% for completing step 1

20% for completing step 2

20% for completing step a

20% for completing step b

20% for completing step c