PANDAS DATA FRAME TASKS

Mauzum shamil am

20bsc132

**SAMPLE DATA FRAME:**

exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],

'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],

'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],

'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}

labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']

**QUESTIONS:**

1) Write a Pandas program to create and display a Data Frame from a specified dictionary data which has the index labels.

ans)

Python code:

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j' ]**

**df = pd.DataFrame(exam\_data , index=labels)**

**print(df)**

**Sample Output:**

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

*d 3 James no NaN*

*e 2 Emily no 9.0*

*f 3 Michael yes 20.0*

*g 1 Matthew yes 14.5*

*h 1 Laura no NaN*

*i 2 Kevin no 8.0*

*j 1 Jonas yes 19.0*

2) Write a Pandas program to get the first 3 rows of a given DataFrame

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("First three rows of the data frame:")**

**print(df.iloc[:3])**

*Sample Output:*

*First three rows of the data frame:*

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

3) Write a Pandas program to select the 'name' and 'score' columns from the following DataFrame

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("Select specific columns:")**

**print(df[['name', 'score']])**

*Sample Output:*

*Select specific columns:*

*name score*

*a Anastasia 12.5*

*b Dima 9.0*

*c Katherine 16.5*

*d James NaN*

*e Emily 9.0*

*f Michael 20.0*

*g Matthew 14.5*

*h Laura NaN*

*i Kevin 8.0*

*j Jonas 19.0*

4) Write a Pandas program to select the rows where the number of attempts in the examination is greater than 2.

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts' : [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("Number of attempts in the examination is greater than 2:")**

**print(df[df['attempts'] > 2])**

Sample Output:

Number of attempts in the examination is greater than 2:

name score attempts qualify

b Dima 9.0 3 no

d James NaN 3 no

f Michael 20.0 3 yes

5) Write a Pandas program to select the rows where the score is missing,

i.e. is NaN

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("Rows where score is missing:")**

**print(df[df['score'].isnull()])**

Sample Output:

*Rows where score is missing:*

*attempts name qualify score*

*d 3 James no NaN*

*h 1 Laura no NaN*

6) Write a Pandas program to select the rows the score is between 15 and 20 (inclusive).

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("Rows where score between 15 and 20 (inclusive):")**

**print(df[df['score'].between(15, 20)])**

Sample Output:

*Rows where score between 15 and 20 (inclusive):*

*attempts name qualify score*

*c 2 Katherine yes 16.5*

*f 3 Michael yes 20.0*

*j 1 Jonas yes 19.0*

7) Write a Pandas program to change the score in row’d’ to 11.5.

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("\nOriginal data frame:")**

**print(df)**

**print("\nChange the score in row 'd' to 11.5:")**

**df.loc['d', 'score'] = 11.5**

**print(df)**

Sample Output:

*Original data frame:*

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

*d 3 James no NaN*

*e 2 Emily no 9.0*

*f 3 Michael yes 20.0*

*g 1 Matthew yes 14.5*

*h 1 Laura no NaN*

*i 2 Kevin no 8.0*

*j 1 Jonas yes 19.0*

*Change the score in row 'd' to 11.5:*

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

*d 3 James no 11.5*

*e 2 Emily no 9.0*

*f 3 Michael yes 20.0*

*g 1 Matthew yes 14.5*

*h 1 Laura no NaN*

*i 2 Kevin no 8.0*

*j 1 Jonas yes 19.0*

8) Write a Pandas program to calculate the sum of the examination attempts by the students.

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("\nSum of the examination attempts by the students:")**

**print(df['attempts'].sum())**

Sample Output:

*Sum of the examination attempts by the students:*

*19*

9) Write a Pandas program to change the name 'James' to 'Suresh' in name column of the data frame.

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("Original rows:")**

**print(df)**

**print("\nChange the name 'James' to ‘Suresh’:")**

**df['name'] = df['name'].replace('James', 'Suresh')**

**print(df)**

Sample Output:

*Original rows:*

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

*d 3 James no NaN*

*e 2 Emily no 9.0*

*f 3 Michael yes 20.0*

*g 1 Matthew yes 14.5*

*h 1 Laura no NaN*

*i 2 Kevin no 8.0*

*j 1 Jonas yes 19.0*

*Change the name 'James' to \‘Suresh\’:*

*attempts name qualify score*

*a 1 Anastasia yes 12.5*

*b 3 Dima no 9.0*

*c 2 Katherine yes 16.5*

*d 3 Suresh no NaN*

*e 2 Emily no 9.0*

*f 3 Michael yes 20.0*

*g 1 Matthew yes 14.5*

*h 1 Laura no NaN*

*i 2 Kevin no 8.0*

*j 1 Jonas yes 19.0*

10) Write a Pandas program to calculate the mean score for each different student in data frame.

ans)

**import pandas as pd**

**import numpy as np**

**exam\_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],**

**'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],**

**'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],**

**'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}**

**labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']**

**df = pd.DataFrame(exam\_data , index=labels)**

**print("\nMean score for each different student in data frame:")**

**print(df['score'].mean())**

Sample Output:

*Mean score for each different student in data frame:*

*13.5625*