04-Dataframe

September 16, 2020

1 Trabajo con Dataframe

```
[13]: import pandas as pd
      import numpy as np

      # Load Data
      studentsHeader = ['nota', 'genero', 'asistencia']
      df_students = pd.read_csv('notas.txt', sep=',', header=None,␣
       ↪names=studentsHeader)
      df_students
```

```
[13]:      nota  genero  asistencia
      0     4.7     0.0         0.1
      1     3.9     1.0         0.3
      2     1.5     0.0         0.0
      3     5.0     0.0         0.1
      4     3.9     0.0         0.9
      ..    …       …           …
      995   6.2     0.0         0.1
      996   6.3     0.0         0.1
      997   6.7     0.0         0.3
      998   7.0     0.0         0.8
      999   4.2     0.0         0.3

      [1000 rows x 3 columns]
```

```
[2]: df_students.groupby("genero").size()
```

```
[2]: genero
     0.0    505
     1.0    495
     dtype: int64
```

```
[3]: df_students.groupby('nota').size()
```

```
[3]: nota
     1.0    10
```

```
1.1    18
1.2    19
1.3    18
1.4    14
       ..
6.6    16
6.7    12
6.8    13
6.9    11
7.0     7
Length: 61, dtype: int64
```

[ ]:

```
df_students.describe()
```

[14]:

```
d_genero = { 0: 'Femenino', 1: 'Masculino'}
df_students['desc_genero'] = students['genero'].map(d_genero)
df_students
```

[14]:

```
     nota  genero  asistencia desc_genero
0     4.7     0.0         0.1    Femenino
1     3.9     1.0         0.3   Masculino
2     1.5     0.0         0.0    Femenino
3     5.0     0.0         0.1    Femenino
4     3.9     0.0         0.9    Femenino
..    …      …          …          …
995   6.2     0.0         0.1    Femenino
996   6.3     0.0         0.1    Femenino
997   6.7     0.0         0.3    Femenino
998   7.0     0.0         0.8    Femenino
999   4.2     0.0         0.3    Femenino

[1000 rows x 4 columns]
```

[15]:

```
df_students.groupby('desc_genero').size()
```

[15]:

```
desc_genero
Femenino     505
Masculino    495
dtype: int64
```

[17]:

```
df_students.describe()
```

[17]:

```
              nota       genero   asistencia
count  1000.000000  1000.000000  1000.000000
mean      3.978800     0.495000     0.505800
std       1.748407     0.500225     0.297144
min       1.000000     0.000000     0.000000
```

```
25%        2.400000     0.000000     0.200000
50%        3.900000     0.000000     0.500000
75%        5.600000     1.000000     0.800000
max        7.000000     1.000000     1.000000
```

```
[18]: df_students.head(10)
```

```
[18]:    nota  genero  asistencia desc_genero
      0   4.7     0.0         0.1    Femenino
      1   3.9     1.0         0.3   Masculino
      2   1.5     0.0         0.0    Femenino
      3   5.0     0.0         0.1    Femenino
      4   3.9     0.0         0.9    Femenino
      5   4.0     0.0         0.9    Femenino
      6   2.9     0.0         0.4    Femenino
      7   1.8     0.0         0.1    Femenino
      8   1.8     0.0         0.6    Femenino
      9   1.2     0.0         0.4    Femenino
```

```
[19]: df_students.columns
```

```
[19]: Index(['nota', 'genero', 'asistencia', 'desc_genero'], dtype='object')
```

```
[20]: pd.unique(df_students['nota'])
```

```
[20]: array([4.7, 3.9, 1.5, 5. , 4. , 2.9, 1.8, 1.2, 3. , 2.4, 1.3, 6.5, 3.7,
             5.4, 4.4, 5.2, 2.2, 3.3, 5.1, 4.1, 4.8, 4.3, 2.7, 6.2, 4.5, 3.1,
             6. , 5.9, 2.8, 6.3, 1.1, 5.6, 3.2, 6.9, 1.6, 3.6, 2. , 2.3, 6.8,
             5.3, 5.8, 5.5, 1.9, 6.6, 6.1, 6.7, 7. , 3.4, 3.8, 6.4, 4.6, 1.7,
             2.6, 3.5, 2.5, 1.4, 2.1, 1. , 5.7, 4.2, 4.9])
```

1.1 Diferencias entre len() y nunique()

```
[21]: lstNotas = df_students['nota']
      lstNotas
```

```
[21]: 0      4.7
      1      3.9
      2      1.5
      3      5.0
      4      3.9
            …
      995    6.2
      996    6.3
      997    6.7
      998    7.0
```

```
999    4.2
Name: nota, Length: 1000, dtype: float64
```

[23]:

```python
print(len(lstNotas))
print(df_students['nota'].nunique())
```

```
1000
61
```

[26]:

```python
print("Max :", df_students['nota'].max())
print("Min :", df_students['nota'].min())
print("Promedio :", df_students['nota'].mean())
print("Desviación estándar :", df_students['nota'].std())
print("Count :", df_students['nota'].count())
```

```
Max : 7.0
Min : 1.0
Promedio : 3.9788
Desviación estándar : 1.7484071286423786
Count : 1000
```

[32]:

```python
grupo_genero = df_students.groupby('desc_genero')
grupo_genero.describe()
```

[32]:

```
              nota                                                 genero      \
             count      mean       std  min  25%  50%  75%  max   count mean
desc_genero
Femenino     505.0  3.967129  1.777513  1.0  2.3  4.0  5.5  7.0   505.0  0.0
Masculino    495.0  3.990707  1.719922  1.0  2.5  3.9  5.6  7.0   495.0  1.0

             …         asistencia                                          \
             …  75%  max      count      mean       std  min  25%  50%  75%
desc_genero  …
Femenino     …  0.0  0.0      505.0  0.498416  0.291747  0.0  0.3  0.5  0.7
Masculino    …  1.0  1.0      495.0  0.513333  0.302660  0.0  0.2  0.5  0.8


             max
desc_genero
Femenino     1.0
Masculino    1.0

[2 rows x 24 columns]
```

[33]:

```python
# Regresa la media de cada columna numérica por genero
grupo_genero.mean()
```

```
[33]:                  nota  genero  asistencia
      desc_genero
      Femenino     3.967129     0.0    0.498416
      Masculino    3.990707     1.0    0.513333
```

```
[36]: df_students.dtypes
```

```
[36]: nota            float64
      genero          float64
      asistencia      float64
      desc_genero      object
      dtype: object
```

```
[37]: df_students.shape
```

```
[37]: (1000, 4)
```

```
[38]: df_students.tail()
```

```
[38]:      nota  genero  asistencia desc_genero
      995   6.2     0.0         0.1    Femenino
      996   6.3     0.0         0.1    Femenino
      997   6.7     0.0         0.3    Femenino
      998   7.0     0.0         0.8    Femenino
      999   4.2     0.0         0.3    Femenino
```

```
[40]: avg_by_gender = df_students.groupby('desc_genero')['nota'].mean()
      avg_by_gender
```

```
[40]: desc_genero
      Femenino     3.967129
      Masculino    3.990707
      Name: nota, dtype: float64
```

```
[41]: result = df_students.groupby('desc_genero')['nota'].max()
      result
```

```
[41]: desc_genero
      Femenino     7.0
      Masculino    7.0
      Name: nota, dtype: float64
```

```
[ ]:
```