40. Pandas III & Matplotlib – Επισκόπηση και Πρακτική Εξάσκηση

[ΕΠΑΝΑΛΗΨΗ ΜΑΘΗΜΑΤΩΝ 24-26]

40.0.1 Λύσεις των προηγούμενων ασκήσεων

Άσκηση 1

Δημιουργήστε ένα DataFrame που περιέχει τυχαία δεδομένα για τις στήλες "Όνομα", "Ηλικία", "Βαθμολογία" για 10 φοιτητές. Χρησιμοποιήστε ένα for loop για να το κάνετε αυτό. Η στήλη "Βαθμολογία" θα πρέπει να περιέχει τυχαίους αριθμούς μεταξύ 0 και 100. Θα εισάγετε επίσης τη βιβλιοθήκη numpy, και σαν πρώτη γραμμή του προγράμματός σας, εισάγετε «np.random.seed(42)» (ο αριθμός 42 είναι τυχαίος και χρησιμοποιείται ώστε αν και λαμβάνουμε τυχαία δεδομένα, να είναι πάντα τα ίδια).

Λύση 1

```
import pandas as pd
import numpy as np

np.random.seed(42)

data = {'Όνομα': [f"Φοιτητής {i+1}" for i in range(10)],
        'Ηλικία': np.random.randint(18, 25, 10),
        'Βαθμολογία': np.random.randint(0, 101, 10)}

df = pd.DataFrame(data)
print(df)
```

	Όνομα	Ηλικία	Βαθμολογία
0	Φοιτητής 1	24	86
1	Φοιτητής 2	21	74
2	Φοιτητής 3	22	74
3	Φοιτητής 4	24	87
4	Φοιτητής 5	20	99
5	Φοιτητής 6	22	23
6	Φοιτητής 7	22	2
7	Φοιτητής 8	24	21
8	Φοιτητής 9	19	52
9	Φοιτητής 10	20	1

Άσκηση 2

Χρησιμοποιήστε το παραπάνω dataframe, για να υπολογίστε τον μέσο όρο της "Βαθμολογίας" για τους φοιτητές που είναι 20 χρόνων και έχουν βαθμολογία πάνω από 50.

Λύση 2

```
import pandas as pd
import numpy as np

np.random.seed(42)
```

```
data = {'Όνομα': [f"Φοιτητής {i+1}" for i in range(10)],
        'Ηλικία': np.random.randint(18, 25, 10),
        'Βαθμολογία': np.random.randint(0, 101, 10)}
df = pd.DataFrame(data)

average_grade = df[(df['Ηλικία'] == 20) & (df['Βαθμολογία'] >
50)]['Βαθμολογία'].mean()
print(f"Ο μέσος όρος της Βαθμολογίας για τους 20χρονους με
βαθμολογία πάνω από 50 είναι: {average_grade}")
```

Άσκηση 3

Χρησιμοποιήστε το παραπάνω dataframe, για να σχεδιάστε ένα ιστόγραμμα για τις ηλικίες των φοιτητών, με διαίρεση σε 3 διαστήματα ηλικίας: 18-20, 21-23, 24-26 (θα χρησιμοποιήσετε την παράμετρο “bins” για τη διαίρεση των διαστημάτων). Για διευκόλυνσή σας, η παράμετρος θα είναι bins = [18, 20, 23, 26]. Προσθέστε τον κώδικα στο αρχείο .py που δημιουργήσατε προηγουμένως.

Λύση 3

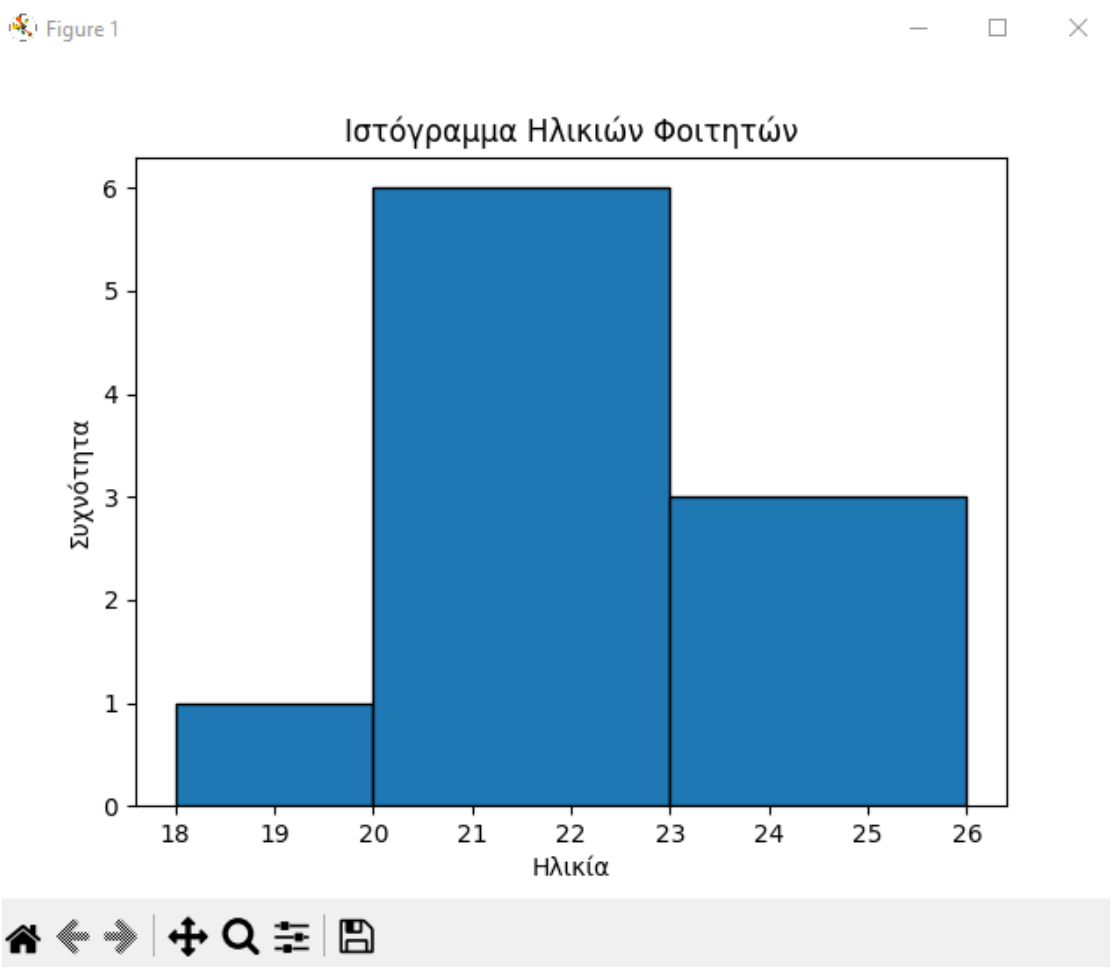
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(42)

data = {'Όνομα': [f"Φοιτητής {i+1}" for i in range(10)],
        'Ηλικία': np.random.randint(18, 25, 10),
        'Βαθμολογία': np.random.randint(0, 101, 10)}
df = pd.DataFrame(data)

average_grade = df[(df['Ηλικία'] == 20) & (df['Βαθμολογία'] >
50)]['Βαθμολογία'].mean()
print(f"Ο μέσος όρος της Βαθμολογίας για τους 20χρονους με
βαθμολογία πάνω από 50 είναι: {average_grade}")
```

```
bins = [18, 20, 23, 26]
plt.hist(df['Ηλικία'], bins=bins, edgecolor='black')
plt.xlabel('Ηλικία')
plt.ylabel('Συχνότητα')
plt.title('Ιστόγραμμα Ηλικιών Φοιτητών')
plt.show()
```



40.1.0 Επιστήμη των δεδομένων – Επισκόπηση

Η Επιστήμη των Δεδομένων αναφέρεται στην εφαρμογή τεχνικών, μεθόδων και αλγορίθμων για την ανάλυση και ερμηνεία δεδομένων, με σκοπό την εξαγωγή σημαντικών πληροφοριών και τη λήψη αποφάσεων. Η Python είναι ιδιαίτερα δημοφιλής στην Επιστήμη των Δεδομένων, και διαθέτει δύο βασικές βιβλιοθήκες που επιτρέπουν την αποτελεσματική εργασία με δεδομένα: την Pandas και την Matplotlib.

40.1.1 Επισκόπηση της Pandas

Η Pandas είναι μια ισχυρή βιβλιοθήκη της Python που προσφέρει δομές δεδομένων υψηλού επιπέδου και εργαλεία για ανάλυση δεδομένων. Η κύρια δομή δεδομένων της Pandas είναι το DataFrame, το οποίο επιτρέπει την οργάνωση και την ανάλυση δεδομένων σε μορφή πίνακα. Η Pandas είναι εξαιρετικά χρήσιμη για τον χειρισμό δεδομένων που προέρχονται από διάφορες πηγές όπως αρχεία CSV, Excel, βάσεις δεδομένων, κ.ά. Οι λειτουργίες της Pandas περιλαμβάνουν φιλτράρισμα, ομαδοποίηση, συγχώνευση και πολλές άλλες προχωρημένες επεξεργασίες δεδομένων.

Για να συνεχίσουμε με την πρακτική μας εξάσκηση για την ανακάλυψη των δυνατοτήτων της βιβλιοθήκης, θα χρειαστούμε ένα αρχείο σε μορφή csv, το οποίο μπορούμε να κατεβάσουμε από δω.

40.1.2 Πρακτική εξάσκηση 1

1. Γράψτε ένα πρόγραμμα για να πάρουμε τις στήλες του DataFrame (αρχείο movies.csv).

```
import pandas as pd
import numpy as np
df = pd.read_csv('movies.csv')
# Για να μην λάβουμε σφάλμα, μπορούμε σαν δεύτερο όρισμα
παραπάνω
# να θέσουμε: dtype={'column_name': 'desired_dtype'})
print("Στήλες του DataFrame:")
print(df.columns)
```

```
>>>
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex1.py ====

Warning (from warnings module):
  File "C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex1.py", line 7
    df = pd.read_csv('movies.csv')
DtypeWarning: Columns (10) have mixed types. Specify dtype option on import or s
et low_memory=False.
Στήλες του DataFrame:
Index(['adult', 'belongs_to_collection', 'budget', 'genres', 'homepage', 'id',
       'imdb_id', 'original_language', 'original_title', 'overview',
       'popularity', 'poster_path', 'production_companies',
       'production_countries', 'release_date', 'revenue', 'runtime',
       'spoken_languages', 'status', 'tagline', 'title', 'video',
       'vote_average', 'vote_count'],
      dtype='object')
>>>
```

1.1 Πώς μπορώ να πάρω τα data types κάθε στήλης;

```
import pandas as pd

# Υποθέτουμε ότι ήδη έχουμε διαβάσει το DataFrame και το
# έχουμε ονομάσει (df)
df = pd.read_csv('movies.csv')

column_data_types = df.dtypes

# Τα data types κάθε στήλης είναι:
print("Data Types κάθε στήλης:")
print(column_data_types)
```

```
Warning (from warnings module):
  File "C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex1.1.py", line 1
0
    df = pd.read_csv('movies.csv')
DtypeWarning: Columns (10) have mixed types. Specify dtype option on import or s
et low_memory=False.
Data Types κάθε στήλης:
adult                     object
belongs_to_collection     object
budget                    object
genres                    object
homepage                  object
id                        object
imdb_id                   object
original_language         object
original_title            object
overview                  object
popularity                object
poster_path               object
production_companies      object
production_countries      object
release_date              object
revenue                  float64
runtime                  float64
spoken_languages          object
status                    object
tagline                   object
title                     object
video                     object
vote_average             float64
vote_count               float64
dtype: object
>>>
```

40.1.3 Πρακτική εξάσκηση 2

2. Γράψτε ένα πρόγραμμα με Pandas για να πάρουμε τις πληροφορίες του DataFrame (αρχείο movies.csv), συμπεριλαμβανομένων των τύπων δεδομένων και της χρήσης μνήμης.

```
import pandas as pd
df = pd.read_csv('movies.csv', dtype = {'popularity':
'object'})
df.info()
```

```
Warning (from warnings module):
  File "C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex2.py", line 8
    df = pd.read_csv('movies.csv')
DtypeWarning: Columns (10) have mixed types. Specify dtype option on import or s
et low_memory=False.
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45466 entries, 0 to 45465
Data columns (total 24 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   adult                  45466 non-null  object
 1   belongs_to_collection  4494 non-null   object
 2   budget                 45466 non-null  object
 3   genres                 45466 non-null  object
 4   homepage               7782 non-null   object
 5   id                     45466 non-null  object
 6   imdb_id                45449 non-null  object
 7   original_language      45455 non-null  object
 8   original_title         45466 non-null  object
 9   overview               44512 non-null  object
 10  popularity             45461 non-null  object
 11  poster_path            45080 non-null  object
 12  production_companies   45463 non-null  object
 13  production_countries   45463 non-null  object
 14  release_date           45379 non-null  object
 15  revenue                45460 non-null  float64
 16  runtime                45203 non-null  float64
 17  spoken_languages       45460 non-null  object
 18  status                 45379 non-null  object
 19  tagline                20412 non-null  object
 20  title                  45460 non-null  object
 21  video                  45460 non-null  object
 22  vote_average           45460 non-null  float64
 23  vote_count             45460 non-null  float64
dtypes: float64(4), object(20)
memory usage: 8.3+ MB
Πληροφορίες DataFrame:
None
```

40.1.4 Πρακτική εξάσκηση 3

3. Γράψτε ένα προγραμματάκι για να πάρουμε τις λεπτομέρειες της τρίτης ταινίας του DataFrame.

```python
import pandas as pd
df = pd.read_csv('movies.csv', dtype={'popularity':'object'})
third_movie = df.iloc[2]
print("Πληροφορίες για την τρίτη ταινία:")
print(third_movie)
```

```
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex3.py ====
Πληροφορίες για την τρίτη ταινία:
adult                                                          False
belongs_to_collection    {'id': 119050, 'name': 'Grumpy Old Men Collect...
budget                                                             0
genres                   [{'id': 10749, 'name': 'Romance'}, {'id': 35, ...
homepage                                                         NaN
id                                                             15602
imdb_id                                                    tt0113228
original_language                                                 en
original_title                                        Grumpier Old Men
overview                 A family wedding reignites the ancient feud be...
popularity                                                   11.7129
poster_path                              /6ksm1sjKMFLbO7UY2i6G1ju9SML.jpg
production_companies     [{'name': 'Warner Bros.', 'id': 6194}, {'name'...
production_countries     [{'iso_3166_1': 'US', 'name': 'United States o...
release_date                                               1995-12-22
revenue                                                           0.0
runtime                                                         101.0
spoken_languages                 [{'iso_639_1': 'en', 'name': 'English'}]
status                                                       Released
tagline                  Still Yelling. Still Fighting. Still Ready for...
title                                                 Grumpier Old Men
video                                                          False
vote_average                                                     6.5
vote_count                                                      92.0
Name: 2, dtype: object
```

40.1.5 Πρακτική εξάσκηση 4

4. Γράψτε ένα πρόγραμμα Pandas για να μετρήσουμε τον αριθμό των γραμμών και των στηλών του DataFrame.

```
import pandas as pd
df = pd.read_csv('movies.csv', dtype={'popularity':'object'})
result = df.shape
print("Αριθμός γραμμών και στηλών του DataFrame:")
print(result)
```

```
>>>
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex4.py ====
Αριθμός γραμμών και στηλών του DataFrame:
(45466, 24)
>>>
```

40.1.6 Πρακτική εξάσκηση 5

5. Γράψτε ένα πρόγραμμα Pandas για να δείτε τις λεπτομέρειες της ταινίας με τίτλο «Grumpier Old Men».

```
import pandas as pd
df = pd.read_csv('movies.csv', low_memory=False)
# Θέτουμε το index στον τίτλο
df = df.set_index('title')
#Λεπτομέρειες της ταινίας 'Grumpier Old Men'
result = df.loc['Grumpier Old Men']
print("Πληροφορίες για την ταινία 'Grumpier Old Men:")
print(result)
```

```
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex5.py ====
Πληροφορίες για την ταινία 'Grumpier Old Men:
adult                                                          False
belongs_to_collection    {'id': 119050, 'name': 'Grumpy Old Men Collect...
budget                                                             0
genres                   [{'id': 10749, 'name': 'Romance'}, {'id': 35, ...
homepage                                                         NaN
id                                                             15602
imdb_id                                                    tt0113228
original_language                                                 en
original_title                                       Grumpier Old Men
overview                 A family wedding reignites the ancient feud be...
popularity                                                   11.7129
poster_path                                  /6ksmlsjKMFLbO7UY2i6G1ju9SML.jpg
production_companies     [{'name': 'Warner Bros.', 'id': 6194}, {'name'...
production_countries     [{'iso_3166_1': 'US', 'name': 'United States o...
release_date                                               1995-12-22
revenue                                                           0.0
runtime                                                         101.0
spoken_languages                 [{'iso_639_1': 'en', 'name': 'English'}]
status                                                       Released
tagline                  Still Yelling. Still Fighting. Still Ready for...
video                                                          False
vote_average                                                     6.5
vote_count                                                      92.0
Name: Grumpier Old Men, dtype: object
```

40.1.7 Πρακτική εξάσκηση 6

6. Γράψτε ένα πρόγραμμα Pandas για να δημιουργήσετε ένα μικρότερο DataFrame με ένα υποσύνολο όλων των χαρακτηριστικών.

```
import pandas as pd
df = pd.read_csv('movies.csv')
# Δημιουργία μικρότερου dataframe
small_df = df[['title', 'release_date', 'budget', 'revenue',
'runtime']]
print("Συνοπτικότερο DataFrame:")
print(small_df.head())
```

```
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex6.py ====

Warning (from warnings module):
  File "C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex6.py", line 7
    df = pd.read_csv('movies.csv')
DtypeWarning: Columns (10) have mixed types. Specify dtype option on import or s
et low_memory=False.
Συνοπτικότερο DataFrame:
                         title release_date    budget      revenue  runtime
0                    Toy Story   1995-10-30  30000000  373554033.0     81.0
1                      Jumanji   1995-12-15  65000000  262797249.0    104.0
2             Grumpier Old Men   1995-12-22         0          0.0    101.0
3            Waiting to Exhale   1995-12-22  16000000   81452156.0    127.0
4  Father of the Bride Part II   1995-02-10         0   76578911.0    106.0
```

40.1.8 Πρακτική εξάσκηση 7

7. Γράψτε ένα πρόγραμμα Pandas για να ταξινομήσετε το DataFrame με βάση την ημερομηνία έκδοσης.

```
import pandas as pd
df = pd.read_csv('movies.csv', dtype={'popularity':'object'})
# Δημιουργία μικρότερου dataframe
small_df = df[['title', 'release_date', 'budget', 'revenue',
'runtime']]
result = small_df.sort_values('release_date')
print("Δεδομένα βασισμένα στην ημ/νία έκδοσης της ταινίας.")
```

```
print(result)
```

```
==== RESTART: C:/Users/NK/AppData/Local/Programs/Python/Python312/40.ex7.py ====
Δεδομένα βασισμένα στην ημ/νία έκδοσης της ταινίας.
                                                title  ... runtime
19730                                             NaN  ...     NaN
29503                                             NaN  ...     NaN
34940                                Passage of Venus  ...     1.0
34937                        Sallie Gardner at a Gallop  ...     1.0
41602                                  Buffalo Running  ...     1.0
...                                                ...  ...     ...
45148                                  Engineering Red  ...    76.0
45203  All Superheroes Must Die 2: The Last Superhero  ...    74.0
45338                  The Land Where the Blues Began  ...     0.0
45410                                           Aprel  ...     NaN
45461                                          Subdue  ...    90.0

[45466 rows x 5 columns]
```

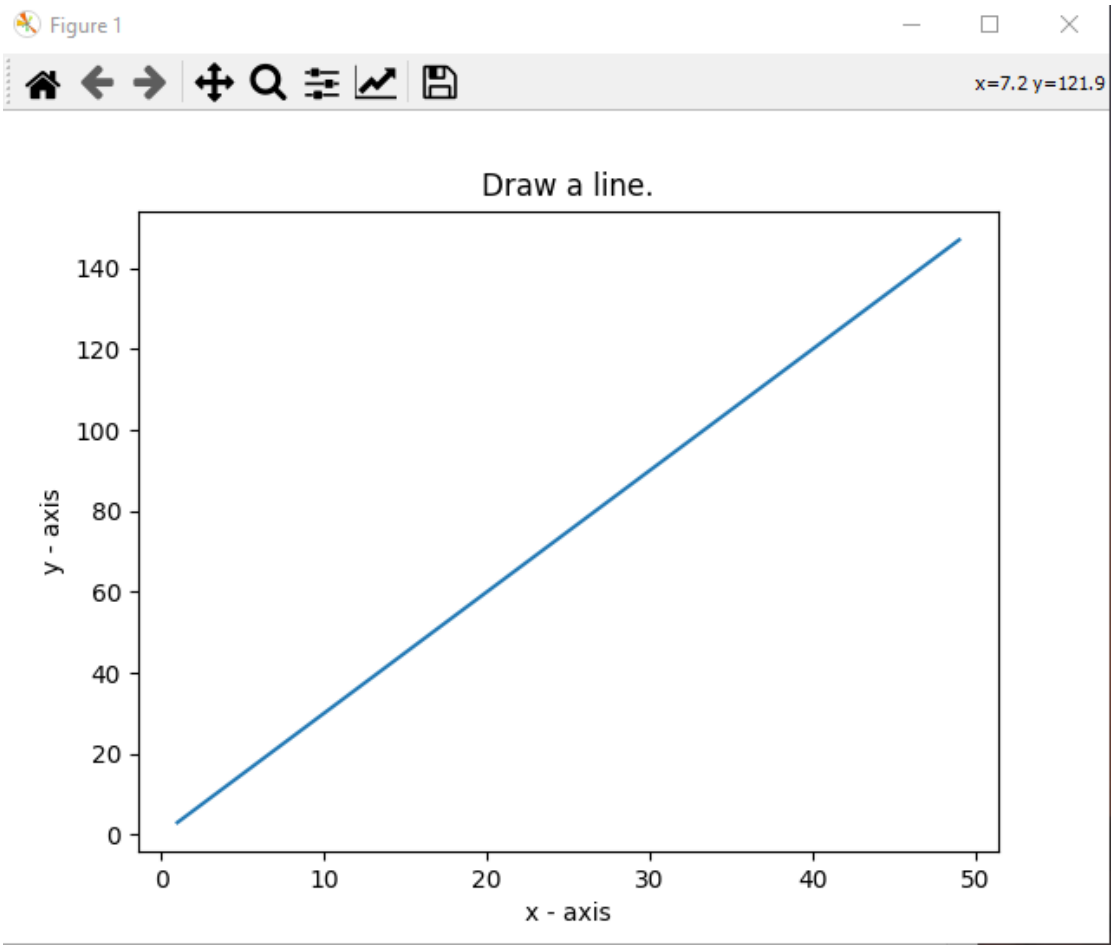
40.1.9 Επισκόπηση του Matplotlib

Το Matplotlib είναι μια βιβλιοθήκη οπτικοποίησης δεδομένων και μας παρέχει εργαλεία για τη δημιουργία γραφικών παραστάσεων. Με το Matplotlib, μπορούμε να δημιουργήσουμε γραφικές αναπαραστάσεις διάφορων τύπων. Είναι εξαιρετικά χρήσιμο για την επικοινωνία και την αντιληπτική ανάλυση των δεδομένων. Το Matplotlib είναι ευέλικτο και προσφέρει πλούσιες επιλογές παραμετροποίησης για την προσαρμογή των γραφικών μας.

40.1.10 Πρακτική εξάσκηση 8

8. Γράψτε ένα πρόγραμμα, χρησιμοποιώντας τη βιβλιοθήκη matplotlib, για να σχεδιάσετε μια γραμμή με κατάλληλη ετικέτα στον άξονα x, τον άξονα y και έναν τίτλο.

```
import matplotlib.pyplot as plt
X = range(1, 50)
Y = [value * 3 for value in X]
print("Τιμές του X:")
print(*range(1,50))
print("Τιμές του Y (Τριπλάσιο του X):")
print(Y)
# Σχεδιασμός γραμμών στους άξονες.
plt.plot(X, Y)
# Ετικέτα στον άξονα X.
plt.xlabel('x - axis')
# Ετικέτα στον άξονα Y.
plt.ylabel('y - axis')
# Τίτλος
plt.title('Draw a line.')
# Εμφάνιση του γραφήματος.
plt.show()
```



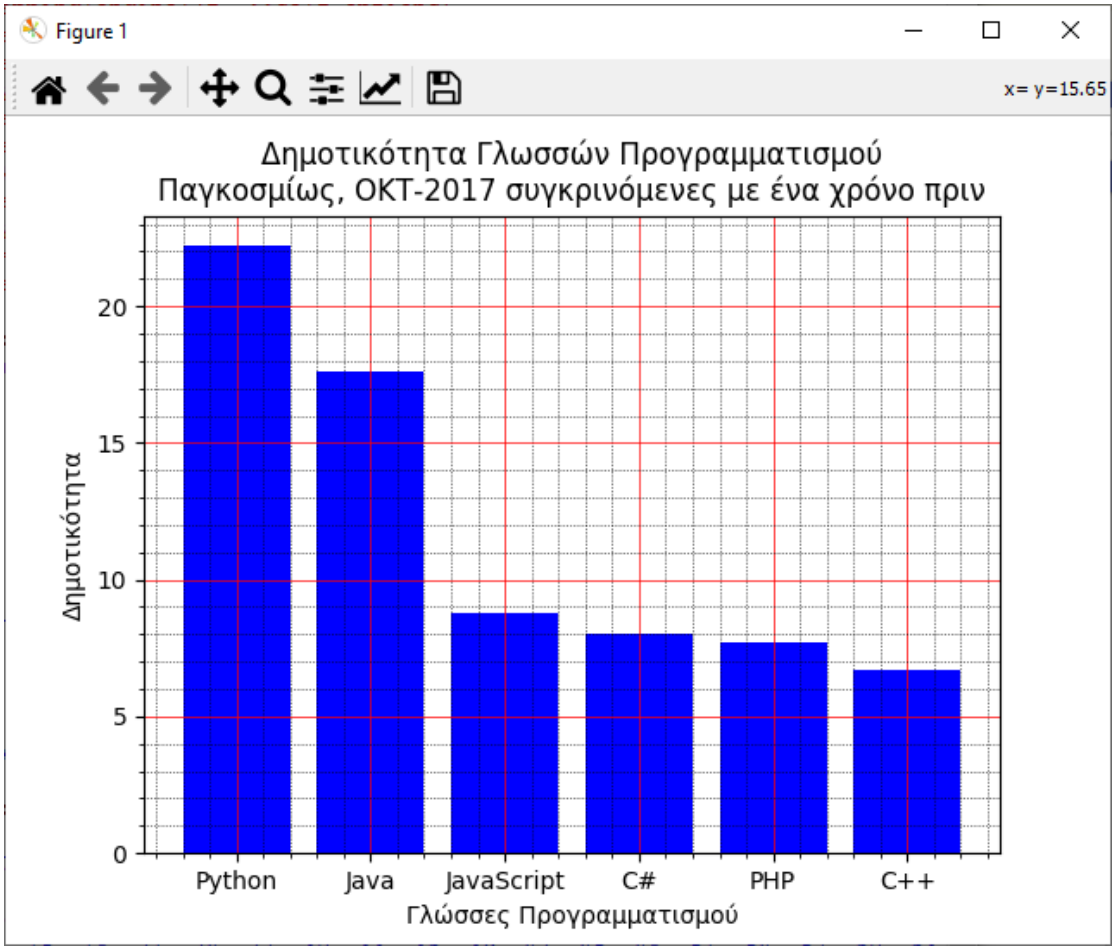
40.1.11 Πρακτική εξάσκηση 9

9. Γράψτε ένα πρόγραμμα για να εμφανίσετε ένα γράφημα ράβδων της δημοτικότητας των Γλωσσών προγραμματισμού. Χρησιμοποιήστε τα παρακάτω δεδομένα:

Γλώσσες προγραμματισμού: Python, Java, JavaScript, C#, PHP, C++

Δημοτικότητα: 25.95, 21.42, 8.26, 7.62, 7.37, 6.31.

```
import matplotlib.pyplot as plt
x = ['Python', 'Java', 'JavaScript', 'C#', 'PHP', 'C++']
popularity = [22.2, 17.6, 8.8, 8, 7.7, 6.7]
x_pos = [i for i, _ in enumerate(x)]
plt.bar(x_pos, popularity, color='blue')
plt.xlabel("Γλώσσες Προγραμματισμού")
plt.ylabel("Δημοτικότητα")
plt.title("Δημοτικότητα Γλωσσών Προγραμματισμού\n" +
"Παγκοσμίως, ΟΚΤ-2017 συγκρινόμενες με ένα χρόνο πριν")
plt.xticks(x_pos, x)
# Άνοιγμα του πλέγματος
plt.minorticks_on()
plt.grid(which='major', linestyle='-', linewidth='0.5',
color='red')
# Παραμετροποίηση του μικρού πλέγματος
plt.grid(which='minor', linestyle=':', linewidth='0.5',
color='black')
plt.show()
```



40.1.12 Πρακτική εξάσκηση 10

10. Γράψτε ένα πρόγραμμα για να εμφανίσετε ένα γράφημα πίτας της δημοτικότητας των Γλωσσών προγραμματισμού. Χρησιμοποιήστε τα παρακάτω δεδομένα:

Γλώσσες προγραμματισμού: Python, Java, JavaScript, C#, PHP, C++

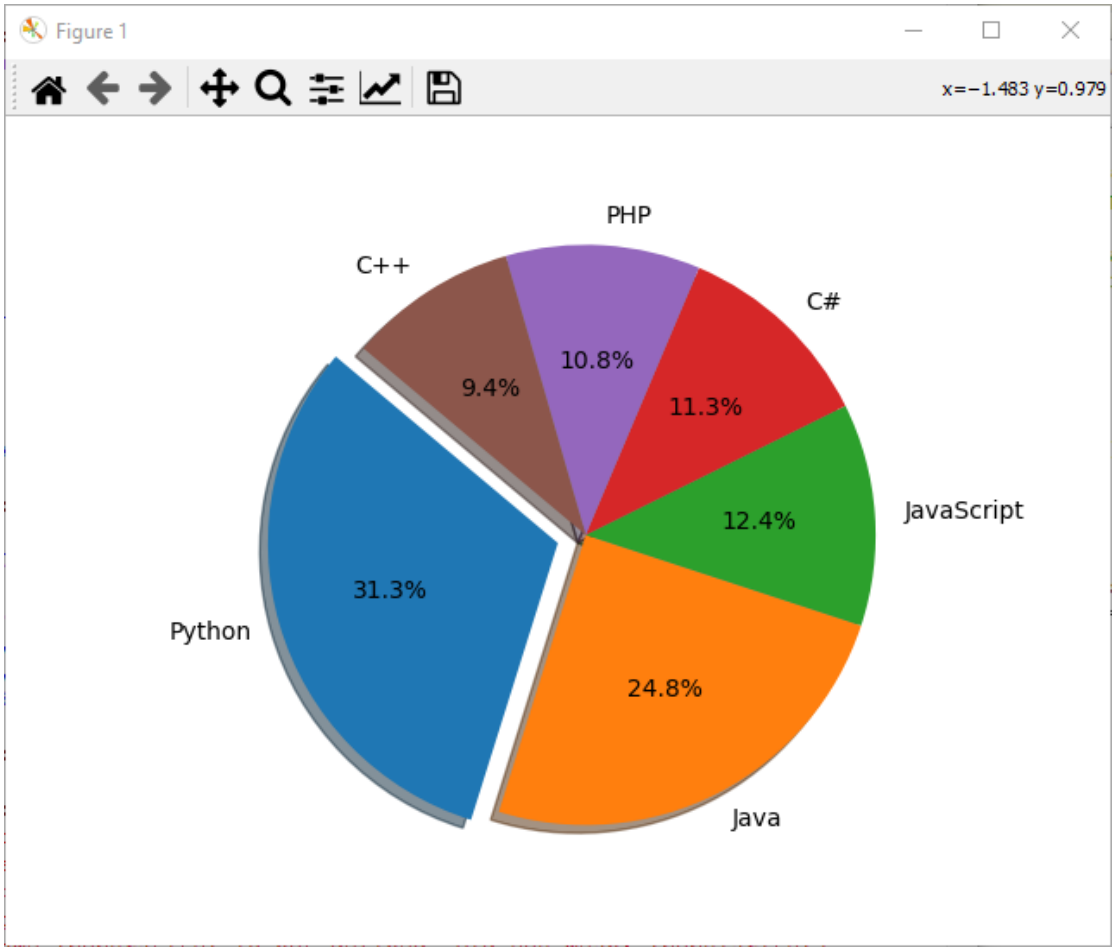
Δημοτικότητα: 25.95, 21.42, 8.26, 7.62, 7.37, 6.31.

```
import matplotlib.pyplot as plt

languages = ['Python', 'Java', 'JavaScript', 'C#', 'PHP',
'C++']
popularity = [22.2, 17.6, 8.8, 8, 7.7, 6.7]
colors = ["#1f77b4", "#ff7f0e", "#2ca02c", "#d62728",
"#9467bd", "#8c564b"]
# Απομάκρυνση του πρώτου πιτακίου
```

```
explode = (0.1, 0, 0, 0,0,0)
# Σχεδίαση
plt.pie(popularity, explode=explode, labels=languages,
colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)

plt.axis('equal')
plt.show()
```



Συνεχίζουμε την επανάληψή μας με τις παρακάτω ασκήσεις.

40.2.0 Ασκήσεις

Χρησιμοποιήστε το αρχείο movies.csv για να λύσετε τις παρακάτω δύο ασκήσεις.

1. Γράψτε ένα πρόγραμμα με τη βιβλιοθήκη Pandas για ανεύρεση των ταινιών, που κυκλοφόρησαν μετά την 1/1/1995.

2. Γράψτε ένα πρόγραμμα για να αποκτήσετε εκείνες τις ταινίες των οποίων τα έσοδα ξεπερνούν τα 2 εκατομμύρια και τα ξοδεύουν λιγότερο από 1 εκατομμύριο.

3. Γράψτε ένα πρόγραμμα με το Pyplotlib για να σχεδιάσετε ένα γράφημα σαν το παρακάτω. Δημιουργήστε μια τυχαία διανομή (X = randn(200) και σαν τίτλους των αξόνων μπορείτε να έχετε X και Y. Το γράφημα να δείχνει σαν το παρακάτω: