Pandas Python Library

Επιμέλεια: Α.Δρακόπουλος

Πίνακας περιεχομένων

Τι είναι το Pandas; 3
Βασικά Χαρακτηριστικά του Pandas 3
Συναρτήσεις του Pandas και Παραδείγματα Χρήσης 3
1. Δημιουργία και Ανάγνωση Δεδομένων 3
Δημιουργία DataFrame 3
Ανάγνωση από CSV 3
2. Πρόσβαση και Επιλογή Δεδομένων 3
Πρόσβαση σε Στήλες 3
Πρόσβαση σε Γραμμές 4
Πρόσβαση σε Τμήματα 4
3. Επεξεργασία Δεδομένων 4
Προσθήκη Νέας Στήλης 4
Τροποποίηση Τιμών 4
Διαγραφή Στήλης ή Γραμμής 4
4. Χειρισμός Δεδομένων που Λείπουν 4
Εντοπισμός Τιμών που Λείπουν 4
Αντικατάσταση Τιμών που Λείπουν 4
Διαγραφή Τιμών που Λείπουν 4
5. Ομαδοποίηση Δεδομένων 5
Ομαδοποίηση με GroupBy 5
6. Συγχώνευση και Ένωση Δεδομένων 5
Συγχώνευση 5
Ένωση (Concatenation) 5
7. Στατιστική Ανάλυση 5
Υπολογισμοί 5
8. Αποθήκευση Δεδομένων 5
Αποθήκευση σε CSV 5
Αποθήκευση σε Excel 6
Σύνοψη 6

Τι είναι το Pandas;

Το **Pandas** είναι μια βιβλιοθήκη της Python που παρέχει εργαλεία και δομές δεδομένων για αποδοτική ανάλυση και διαχείριση δεδομένων. Είναι ιδανική για την εργασία με δεδομένα σε μορφή **πίνακα** (dataframe), όπως π.χ. πίνακες Excel, αρχεία CSV, βάσεις δεδομένων ή δεδομένα από APIs.

Η βιβλιοθήκη Pandas παρέχει δύο βασικές δομές δεδομένων:

1. Series: Ένας μονοδιάστατος πίνακας δεδομένων (παρόμοιος με τις λίστες ή arrays).
2. DataFrame: Ένας δισδιάστατος πίνακας δεδομένων (παρόμοιος με πίνακες Excel ή SQL πίνακες).

Βασικά Χαρακτηριστικά του Pandas

1. **Ανάγνωση και Εγγραφή Δεδομένων**: Υποστήριξη διαφόρων μορφών δεδομένων, όπως CSV, Excel, SQL, JSON κ.ά.
2. **Επεξεργασία Δεδομένων**: Φιλτράρισμα, επιλογή, αλλαγή, και διαγραφή δεδομένων.
3. **Ανάλυση Δεδομένων**: Ομαδοποίηση, συνδυασμός, και υπολογισμοί σε πίνακες.
4. **Χειρισμός Δεδομένων που Λείπουν**: Εύκολος εντοπισμός και αντικατάσταση τιμών που λείπουν.
5. **Ενοποίηση Δεδομένων**: Συγχώνευση και ένωση πολλών DataFrames.

Συναρτήσεις του Pandas και Παραδείγματα Χρήσης

Παρακάτω παρουσιάζονται οι βασικές λειτουργίες του Pandas με αναλυτικά παραδείγματα και σχόλια.

1. Δημιουργία και Ανάγνωση Δεδομένων

Δημιουργία DataFrame

```
import pandas as pd

# Δημιουργία DataFrame από λεξικό
data = {
    "Όνομα": ["Γιάννης", "Μαρία", "Κώστας"],
    "Ηλικία": [25, 30, 35],
    "Πόλη": ["Αθήνα", "Θεσσαλονίκη", "Πάτρα"]
}
df = pd.DataFrame(data)
print(df)
```

Ανάγνωση από CSV

```
# Ανάγνωση δεδομένων από αρχείο CSV
df = pd.read_csv("data.csv")
print(df.head())  # Εμφάνιση των πρώτων 5 γραμμών
```

2. Πρόσβαση και Επιλογή Δεδομένων

Πρόσβαση σε Στήλες

```
# Επιλογή στήλης
print(df["Όνομα"])
```

```
# Επιλογή πολλαπλών στηλών
print(df[["Όνομα", "Ηλικία"]])
```

Πρόσβαση σε Γραμμές

```
# Επιλογή γραμμής με βάση το index
print(df.iloc[0])  # Πρώτη γραμμή

# Επιλογή γραμμής με βάση συνθήκη
print(df[df["Ηλικία"] > 30])  # Φιλτράρισμα με συνθήκη
```

Πρόσβαση σε Τμήματα

```
# Επιλογή τμήματος του DataFrame
print(df.loc[0:1, ["Όνομα", "Πόλη"]])  # Γραμμές 0-1, στήλες "Όνομα" και "Πόλη"
```

3. Επεξεργασία Δεδομένων

Προσθήκη Νέας Στήλης

```
# Δημιουργία νέας στήλης
df["Μισθός"] = [2000, 2500, 3000]
print(df)
```

Τροποποίηση Τιμών

```
# Αύξηση ηλικίας κατά 1
df["Ηλικία"] = df["Ηλικία"] + 1
print(df)
```

Διαγραφή Στήλης ή Γραμμής

```
# Διαγραφή στήλης
df = df.drop(columns=["Μισθός"])

# Διαγραφή γραμμής
df = df.drop(index=1)  # Διαγραφή της 2ης γραμμής
print(df)
```

4. Χειρισμός Δεδομένων που Λείπουν

Εντοπισμός Τιμών που Λείπουν

```
# Επιστροφή λογικού πίνακα με τις θέσεις των NaN
print(df.isnull())
```

Αντικατάσταση Τιμών που Λείπουν

```
# Αντικατάσταση NaN με 0
df["Ηλικία"].fillna(0, inplace=True)
```

Διαγραφή Τιμών που Λείπουν

```
# Διαγραφή γραμμών με NaN
df = df.dropna()
```

5. Ομαδοποίηση Δεδομένων

Ομαδοποίηση με GroupBy

```
# Δημιουργία DataFrame
data = {
    "Τμήμα": ["A", "A", "B", "B", "A"],
    "Μισθός": [1000, 1200, 1500, 1700, 1100]
}
df = pd.DataFrame(data)

# Ομαδοποίηση και υπολογισμός μέσου μισθού ανά τμήμα
grouped = df.groupby("Τμήμα")["Μισθός"].mean()
print(grouped)
```

6. Συγχώνευση και Ένωση Δεδομένων

Συγχώνευση

```
# Δημιουργία δύο DataFrames
df1 = pd.DataFrame({"ID": [1, 2], "Όνομα": ["Γιάννης", "Μαρία"]})
df2 = pd.DataFrame({"ID": [1, 2], "Μισθός": [2000, 2500]})

# Συγχώνευση με βάση το ID
merged = pd.merge(df1, df2, on="ID")
print(merged)
```

Ένωση (Concatenation)

```
# Δημιουργία δύο DataFrames
df1 = pd.DataFrame({"Όνομα": ["Γιάννης", "Μαρία"], "Ηλικία": [25, 30]})
df2 = pd.DataFrame({"Όνομα": ["Κώστας", "Ελένη"], "Ηλικία": [35, 40]})

# Κατακόρυφη ένωση
combined = pd.concat([df1, df2])
print(combined)
```

7. Στατιστική Ανάλυση

Υπολογισμοί

```
# Υπολογισμός μέσου όρου
mean_age = df["Ηλικία"].mean()
print("Μέσος όρος ηλικίας:", mean_age)

# Υπολογισμός μέγιστης τιμής
max_salary = df["Μισθός"].max()
print("Μέγιστος μισθός:", max_salary)
```

8. Αποθήκευση Δεδομένων

Αποθήκευση σε CSV

```
# Αποθήκευση σε αρχείο CSV
df.to_csv("output.csv", index=False)
```

Αποθήκευση σε Excel

```
# Αποθήκευση σε αρχείο Excel
df.to_excel("output.xlsx", index=False)
```

Σύνοψη

Το **Pandas** είναι μια εξαιρετικά ισχυρή και ευέλικτη βιβλιοθήκη για την επεξεργασία και ανάλυση δεδομένων. Είναι απαραίτητο εργαλείο για κάθε επιστήμονα δεδομένων ή αναλυτή, παρέχοντας εύκολες και γρήγορες λύσεις για εργασία με πίνακες και δεδομένα σε πραγματικές εφαρμογές.