Quanskill

Python Cơ Bản Cho Data Science

Kiểu Dữ Liệu Cơ Bản: String, Int và Float

Quanskill

Mục Lục

Giới Thiệu

Tại Sao Kiểu Dữ Liệu Quan Trọng?

Quanskill

- Trong Data Science, chúng ta xử lý nhiều loại dữ liệu khác nhau
- Hiểu kiểu dữ liệu giúp chọn đúng phương pháp phân tích
- Python có các kiểu dữ liệu cơ bản phù hợp cho từng mục đích

Ba Kiểu Dữ Liệu Cơ Bản

- **String (str)**: Văn bản, nhãn, danh mục
- **Integer (int)**: Số nguyên, đếm, chỉ số
- **Float**: Số thập phân, đo lường, tính toán

Kiểm Tra Kiểu Dữ Liệu

Quanskill

Python cung cấp hàm `type()` để kiểm tra kiểu dữ liệu:

```
# Check data types in Python
name = "Alice"
age = 25
height = 1.75

print("Type of name:", type(name))
print("Type of age:", type(age))
print("Type of height:", type(height))
```

Output:

```
Type of name: <class 'str'>
Type of age: <class 'int'>
Type of height: <class 'float'>
```

String (Chuỗi)

String Là Gì?

Quanskill

- String (chuỗi) là kiểu dữ liệu văn bản
- Được bao quanh bởi dấu ngoặc kép ""hoặc đơn ' '
- Rất quan trọng trong phân tích dữ liệu văn bản (text analytics)

Ví Dụ Trong Data Science

- Tên khách hàng: "Nguyễn Văn A"
- Danh mục sản phẩm: "Electronics", "Clothing"
- Bình luận khách hàng: "Sản phẩm rất tốt!"
- Địa chỉ email: "user@example.com"

Tạo String

Quanskill

```
# Creating strings - different ways
customer_name = "John Doe"
product_category = 'Electronics'
review = """This product is amazing!
I would definitely recommend it."""

print("Customer:", customer_name)
print("Category:", product_category)
print("Review:", review)
```

Output:

```
Customer: John Doe
Category: Electronics
Review: This product is amazing!
I would definitely recommend it.
```

String Operations - Nối Chuỗi

Quanskill

```python
# String concatenation for data processing
first_name = "Maria"
last_name = "Garcia"

# Method 1: Using + operator
full_name = first_name + " " + last_name
print("Full name:", full_name)

# Method 2: Using f-strings (recommended)
customer_id = 12345
message = f"Welcome {full_name}! Your ID is {customer_id}"
print("Message:", message)
```

Output:

```
Full name: Maria Garcia
Message: Welcome Maria Garcia! Your ID is 12345
```

String Methods - Xử Lý Dữ Liệu

Quanskill

```python
# Useful string methods for data cleaning
email = "  JOHN.DOE@GMAIL.COM  "

# Clean and standardize the email
clean_email = email.strip().lower()
print("Original:", f"'{email}'")
print("Cleaned:", clean_email)

# Extract information
domain = clean_email.split("@")[1]
print("Domain:", domain)

# Check content
print("Contains gmail:", "gmail" in clean_email)
```

Output:

```
Original: '  JOHN.DOE@GMAIL.COM  '
Cleaned: john.doe@gmail.com
Domain: gmail.com
Contains gmail: True
```

Integer (Số Nguyên)

Integer Là Gì?

Quanskill

- Integer (int) là số nguyên: ..., -2, -1, 0, 1, 2, ...
- Không có phần thập phân
- Dùng để đếm, chỉ số, ID, ranking

Ví Dụ Trong Data Science

- Số lượng sản phẩm: 150
- ID khách hàng: 12345
- Tuổi: 25
- Số lần click: 1250
- Ranking: 1, 2, 3, ...

Tạo và Sử Dụng Integer

Quanskill

```python
# Creating integers for data analysis
total_customers = 1000
new_signups = 45
churned_customers = 12

print("Total customers:", total_customers)
print("Data type:", type(total_customers))

# Calculate retention
active_customers = total_customers + new_signups - churned_customers
print("Active customers:", active_customers)

# Calculate percentages (will be float)
churn_rate = churned_customers / total_customers * 100
print("Churn rate:", churn_rate, "%")
```

Output:

```
Total customers: 1000
Data type: <class 'int'>
Active customers: 1033
Churn rate: 1.2 %
```

Integer Operations - Phép Toán

Quanskill

```python
# Integer operations for data analysis
sales_q1 = 250
sales_q2 = 300
sales_q3 = 275
sales_q4 = 425

# Basic arithmetic
total_sales = sales_q1 + sales_q2 + sales_q3 + sales_q4
average_sales = total_sales // 4  # Integer division
remainder = total_sales % 4       # Remainder

print("Total sales:", total_sales)
print("Average (integer division):", average_sales)
print("Remainder:", remainder)

# Comparison
best_quarter = max(sales_q1, sales_q2, sales_q3, sales_q4)
print("Best quarter sales:", best_quarter)
```

Output:

```
Total sales: 1250
Average (integer division): 312
Remainder: 2
Best quarter sales: 425
```

Float (Số Thập Phân)

Float Là Gì?

Quanskill

- Float là số thập phân (floating point number)
- Có phần nguyên và phần thập phân
- Dùng cho các phép đo lường chính xác, tỷ lệ phần trăm

Ví Dụ Trong Data Science

- Giá sản phẩm: 29.99
- Tỷ lệ conversion: 0.035 (3.5%)
- Chiều cao: 1.75 (mét)
- Điểm rating: 4.5/5.0
- Temperature: 25.6°C

Tạo và Sử Dụng Float

Quanskill

```python
# Creating floats for precise calculations
product_price = 29.99
discount_rate = 0.15  # 15%
tax_rate = 0.08       # 8%

print("Original price: $", product_price)
print("Data type:", type(product_price))

# Calculate final price
discount_amount = product_price * discount_rate
discounted_price = product_price - discount_amount
tax_amount = discounted_price * tax_rate
final_price = discounted_price + tax_amount

print("Discount amount: $", round(discount_amount, 2))
print("Final price: $", round(final_price, 2))
```

Output:

```
Original price: $ 29.99
Data type: <class 'float'>
Discount amount: $ 4.5
Final price: $ 27.59
```

Float Operations - Tính Toán Chính Xác

Quanskill

```python
# Float operations for statistical analysis
temperatures = [23.5, 25.1, 22.8, 26.3, 24.7]

# Calculate statistics
total_temp = sum(temperatures)
count = len(temperatures)
average_temp = total_temp / count

print("Temperature readings:", temperatures)
print("Total:", total_temp)
print("Count:", count)
print("Average:", round(average_temp, 2), " C ")

# Find min and max
min_temp = min(temperatures)
max_temp = max(temperatures)
range_temp = max_temp - min_temp

print("Min temperature:", min_temp, " C ")
print("Max temperature:", max_temp, " C ")
print("Temperature range:", range_temp, " C ")
```

Output:

```
Temperature readings: [23.5, 25.1, 22.8, 26.3, 24.7]
Total: 122.4
Count: 5
Average: 24.48 °C
Min temperature: 22.8 °C
Max temperature: 26.3 °C
```

Chuyển Đổi Kiểu Dữ Liệu

Type Conversion - Chuyển Đổi Kiểu

Quanskill

```python
# Type conversion in data processing
user_input = "123"       # String from user input
price_str = "45.99"      # String from CSV file
count_float = 10.0       # Float from calculation

print("Original types:")
print("user_input:", type(user_input), "value:", user_input)
print("price_str:", type(price_str), "value:", price_str)
print("count_float:", type(count_float), "value:", count_float)

# Convert types
user_id = int(user_input)            # String to int
product_price = float(price_str)     # String to float
item_count = int(count_float)        # Float to int

print("\nAfter conversion:")
print("user_id:", type(user_id), "value:", user_id)
print("product_price:", type(product_price), "value:", product_price)
print("item_count:", type(item_count), "value:", item_count)
```

Xử Lý Lỗi Khi Chuyển Đổi

Quanskill

```python
# Handling conversion errors safely
data_inputs = ["123", "45.99", "invalid", "67.5", "abc"]

print("Processing data inputs:")
for i, input_value in enumerate(data_inputs):
    print(f"\nInput {i+1}: '{input_value}'")

    try:
        # Try to convert to float first
        as_float = float(input_value)
        print("  As float:", as_float)

        # Check if it's actually an integer
        if as_float.is_integer():
            as_int = int(as_float)
            print("  As integer:", as_int)

    except ValueError:
        print("  Error: Cannot convert to number")
        print("  Keeping as string:", input_value)
```

Ứng Dụng Thực Tế

Ví Dụ: Phân Tích Dữ Liệu Khách Hàng

Quanskill

```python
# Customer data analysis example
customers = [
    {"name": "Alice Johnson", "age": 28, "spent": 1250.50},
    {"name": "Bob Smith", "age": 35, "spent": 890.25},
    {"name": "Carol Davis", "age": 42, "spent": 2100.75}
]

print("Customer Analysis Report")
print("=" * 30)

total_spent = 0.0
total_customers = len(customers)

for customer in customers:
    name = customer["name"]          # String
    age = customer["age"]            # Integer
    spent = customer["spent"]        # Float

    print(f"\nCustomer: {name}")
    print(f"Age: {age} years old")
    print(f"Total spent: ${spent}")

    total_spent += spent

average_spent = total_spent / total_customers
print(f"\nSummary:")
print(f"Total customers: {total_customers}")
print(f"Total revenue: ${round(total_spent, 2)}")
print(f"Average per customer: ${round(average_spent, 2)}")
```

Output: Báo Cáo Phân Tích

Quanskill

Output:

```
Customer Analysis Report
========================

Customer: Alice Johnson
Age: 28 years old
Total spent: $1250.5

Customer: Bob Smith
Age: 35 years old
Total spent: $890.25

Customer: Carol Davis
Age: 42 years old
Total spent: $2100.75

Summary:
Total customers: 3
Total revenue: $4241.5
Average per customer: $1413
```

Tóm Tắt và Luyện Tập

Tóm Tắt Kiểu Dữ Liệu

Quanskill

Bảng 1: So sánh các kiểu dữ liệu cơ bản

Kiểu	Mô tả	Ví dụ
String (str)	Văn bản, ký tự	"Alice", 'Product A'
Integer (int)	Số nguyên	25, 100, -5
Float	Số thập phân	29.99, 3.14, -0.5

Điểm Quan Trọng

- Sử dụng `type()` để kiểm tra kiểu dữ liệu
- Chuyển đổi giữa các kiểu: `int()`, `float()`, `str()`
- Xử lý lỗi khi chuyển đổi với `try-except`
- Chọn đúng kiểu dữ liệu cho từng mục đích

Bài Tập 1: Xử Lý Thông Tin Sản Phẩm

Tạo một chương trình xử lý thông tin sản phẩm:

- Tên sản phẩm: "Laptop Dell XPS"
- Giá gốc: 1299.99
- Số lượng: 50
- Tính giá sau giảm giá 10%

```
# Your code here:
product_name = "Laptop Dell XPS"
original_price = 1299.99
quantity = 50
discount_rate = 0.10

# Calculate discounted price
# Print product information
# Show data types
```

Câu Hỏi Thảo Luận

Quanskill

- Khi nào nên sử dụng Integer thay vì Float?
- Làm thế nào để xử lý dữ liệu đầu vào không hợp lệ?
- Tại sao việc hiểu kiểu dữ liệu quan trọng trong Data Science?

Questions?