BÀI THỰC HÀNH SỐ 2_3

Hiểu dữ liệu dữ liệu

Một số hàm hay sử dụng

1. **Hàm thống kê**

- Đếm số giá trị trong một trường
  df['Tentruong'].count()
- Tính trung bình
  df['Tentruong']. mean()
- Tính độ lệch của dữ liệu so với giá trị trung bình của dữ liệu
  df['Tentruong']. std()
- Tìm giá trị nhỏ nhất
  df['Tentruong']. min()
- Tìm gí trị lớn nhất
  df['Tentruong']. max()

Tìm giá trị tứ vị:Tứ phân vị là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu..

- Giá trị tứ phân vị thứ nhất Q1 bằng trung vị phần dưới
  df['Tentruong'].quantile(.25)
- Giá trị tứ phân vị thứ hai Q2 chính bằng giá trị trung vị
  df['Tentruong'].quantile(.5)
- Giá trị tứ phân vị thứ ba Q3 bằng trung vị phần trên
  df['Tentruong'].quantile(.75)
- Chúng ta có thể sử dụng hàm sau để thống kê dữ liệu thay cho các hàm trên.
  df.describe()
  Gộp nhóm

```
#thống kê trường Age và Salary theo gộp nhóm trường Skill
df.groupby('Skill')[['Age','Salary']].describe()
```

2. **Các hàm trực quan cơ bản**
   Sử dụng 1 trong 2 thư viện sau:

import seaborn as sns

import matplotlib.pyplot as plt

Vẽ đường thẳng

```
import seaborn as sns


x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]
sns.lineplot(x, y)
```

vẽ biểu đồ bar

```
import matplotlib.pyplot as plt

# Dữ liệu cho biểu đồ
x = ['A','B','C','D']
y = [1, 2, 3, 4]

# Tạo biểu đồ
plt.bar(x, y)
plt.xlabel('Nhãn trục x')
plt.ylabel('Nhãn trục y')
plt.title('Tiêu đề biểu đồ')

# Hiển thị biểu đồ
plt.show()
```

Các loại biểu đố thể hiện sự phân bố dữ liệu:

```python
import pandas as pd
import seaborn as sns

# đọc dữ liệu từ tệp dữ liệu EmployeeSalary.csv vào dataframe
df = pd.read_csv('EmployeeSalary.csv')

# vẽ biểu đồ boxplot cho cột 'salary'
sns.boxplot(x=df['salary'])

# vẽ histogram cho cột 'salary'
sns.histplot(data=df, x='salary')

# vẽ scatter plot giữa cột 'age' và 'salary'
sns.scatterplot(data=df, x='age', y='salary')
```

3. **Phân tích tương quan**

   Sử dụng hàm corr()

   Ví dụ phân tích sự tương quan giữa Salary và TrainedYear

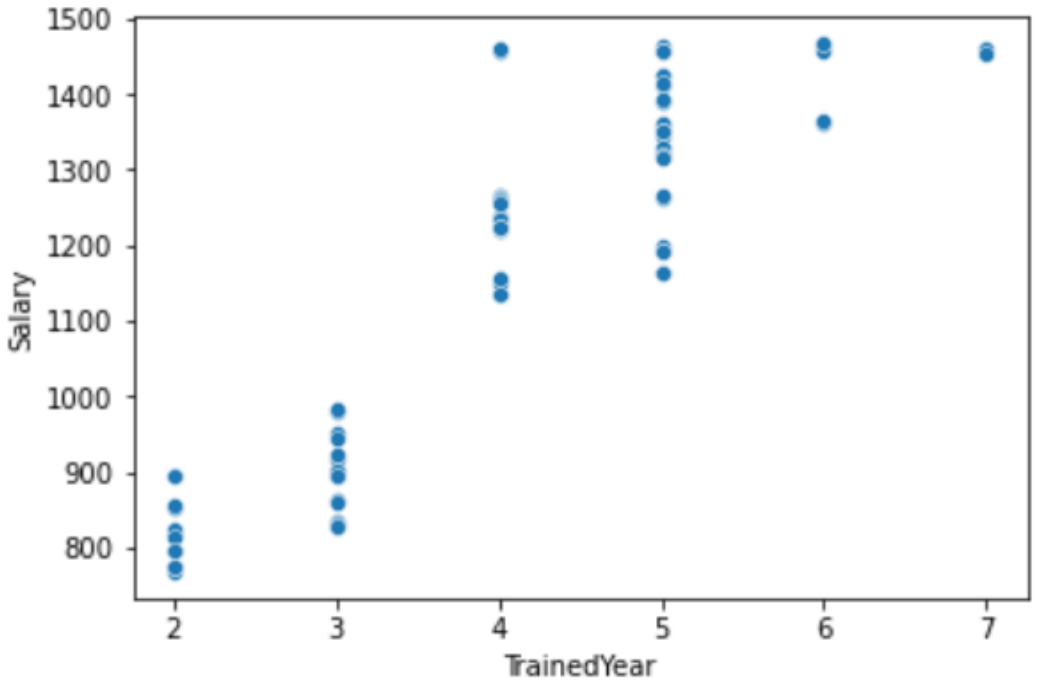
```python
df[['Salary','TrainedYear']].corr()
```

	Salary	TrainedYear
Salary	1.000000	0.908832
TrainedYear	0.908832	1.000000

Hoặc dùng biểu đồ scatter

```
sns.scatterplot(x='TrainedYear', y='Salary', data=df)
```

```
<AxesSubplot:xlabel='TrainedYear', ylabel='Salary'>
```



Bài 1: Sử dụng file dữ liệu JOB_INFO.xlsx thực hiện:

- Đọc file dữ liệu
- Hiển thị dữ liệu
- Thống kê dữ liệu
- Sử dụng các hàm và biểu đồ để hiểu dữ liệu: đo độ tập trung, phân tán và tóm tắt dữ liệu. (Tất cả các loại biểu đồ cho các cột trong bảng dữ liệu)
- Đo sự tương quan của trường Salarry và WorkingYear, Salarry và Age
- Đo sự tương quan của trường Salarry, TrainedYear, WorkingYear và Age

Bài 2: Sử dụng tệp dữ liệu ThiTHPT2018.csv thực hiện các yêu cầu sau:

1. Tính trung bình của các môn Toán, Văn. Anh
2. Tính độ lệch chuẩn, trung vị của các môn
3. Giá trị thường xuyên của các môn thi

4. Tính tứ vị phân của môn toán, Văn, Anh
5. Thống kê trường Toán,Văn, Anh theo gộp nhóm trường Ten Tinh
6. Vẽ các loại biểu đồ cho các môn học. (Vẽ các môn trên 1 biểu đồ và biểu đồ từng môn)

Bài 3 T**ự tìm hiểu tệp dữ liệu** CollegeRecruitingData