pandas

August 31, 2024

0.1 1. Getting Familiar with Pandas

```python
import pandas as pd
import numpy as np

# 1. Getting Familiar with Pandas
# Creating DataFrames and Series
data_dict = {'Name': ['Alice', 'Bob', 'Charlie'], 'Age': [25, 30, 35]}
df = pd.DataFrame(data_dict)
series = pd.Series([1, 2, 3, 4])
```

```python
# Selecting data and modifying DataFrames
selected_data = df[df['Age'] > 28]
df['Age'] = df['Age'] + 1  # Incrementing age by 1
```

```python
# 2. Data Handling with Pandas
# Reading data from a CSV file
df.to_csv('sample_data.csv', index=False)
df_loaded = pd.read_csv('sample_data.csv')
```

```python
# Handling missing data
df_loaded.loc[1, 'Age'] = np.nan  # Introduce a missing value
df_filled = df_loaded.fillna(df_loaded['Age'].mean())  # Fill missing values
 ↪with the mean
```

```python
# Removing duplicates
df_filled = df_filled.drop_duplicates()
```

```python
# Data type conversion
df_filled['Age'] = df_filled['Age'].astype(int)
```

```python

```

0.2 3. Data Analysis with Pandas

```python
import pandas as pd
import numpy as np
df=pd.DataFrame(np.arange(15))
# Assuming df is your DataFrame
summary_stats = df.describe()
df1 = pd.DataFrame({'key': ['A', 'B', 'C'], 'value1': [1, 2, 3]})
df2 = pd.DataFrame({'key': ['A', 'B', 'D'], 'value2': [4, 5, 6]})
print(summary_stats)
print(df)
```

```
                0
count   15.000000
mean     7.000000
std      4.472136
min      0.000000
25%      3.500000
50%      7.000000
75%     10.500000
max     14.000000
         0
0        0
1        1
2        2
3        3
4        4
5        5
6        6
7        7
8        8
9        9
10      10
11      11
12      12
13      13
14      14
```

```python
mean_value = df[0].mean()
median_value = df[0].median()
std_dev = df[0].std()
print(mean_value)
print(median_value)
print(std_dev)
```

```
7.0
7.0
4.47213595499958
```

```
[20]: aggregated_data = grouped_df[0].sum()
      print(aggregated_data)
```

```
0
0      0
1      1
2      2
3      3
4      4
5      5
6      6
7      7
8      8
9      9
10    10
11    11
12    12
13    13
14    14
Name: 0, dtype: int32
```

```
[21]: merged_df = pd.merge(df1, df2, on='key', how='inner')
      print(merged_df)
```

```
  key  value1  value2
0   A       1       4
1   B       2       5
```

```
[29]: df1 = pd.DataFrame({'value1': [1, 2, 3]}, index=['A', 'B', 'C'])
      df2 = pd.DataFrame({'value2': [4, 5, 6]}, index=['A', 'B', 'D'])
      joined_df = df1.join(df2, how='inner')
      print(joined_df)
```

```
   value1  value2
A       1       4
B       2       5
```

```
[25]: concatenated_df = pd.concat([df1, df2], axis=0)  # For row-wise concatenation
      print(concatenated_df)
```

```
  key  value1  value2
0   A     1.0     NaN
1   B     2.0     NaN
2   C     3.0     NaN
0   A     NaN     4.0
1   B     NaN     5.0
2   D     NaN     6.0
```

[ ]: