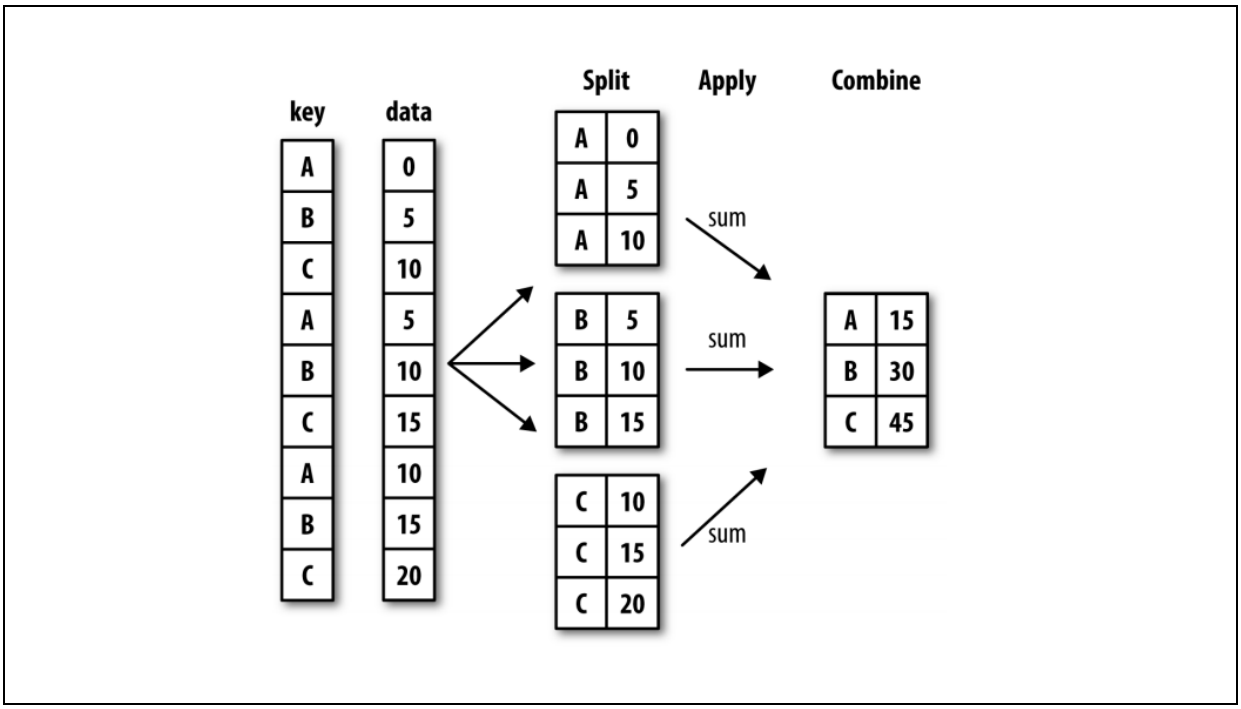
16. Pandas – DataFrame – Groupby

Contents

16. Pandas – DataFrame - Groupby

1. Groupby Introduction

- ✓ Groupby is very common and important operations in Data analysis,
- ✓ There are mainly 3 steps in Groupby operation,
  - Splitting the data into groups based on some criteria.
  - Applying operations to each group independently.
  - Combining the results.



- ✓ For example,
  - If we apply groupby on Product_Name then it groups the data into based on name of the product.
- ✓ The groupby(p) method returns a GroupBy object.

2. groupby(p)

- ✓ groupby(p) is predefined method in DataFrame.
- ✓ We should access this method by using DataFrame object.
- ✓ This method returns GroupBy object.

Program Name: Creating products DataFrame
demo1.py

```
import pandas as pd

d = {
        "Product": ["Samsung", "Nokia", "Samsung", "Motorola",
        "Nokia", "Samsung", "Samsung"],

        "Orders": [2, 4, 3, 4, 6, 7, 3]
}

df1 = pd.DataFrame(d)

print(df1)
```

Output

```
    Product  Orders
0   Samsung       2
1     Nokia       4
2   Samsung       3
3  Motorola       4
4     Nokia       6
5   Samsung       7
6   Samsung       3
```

Program Name: Creating products DataFrame, applying groupby
demo2.py

```
import pandas as pd

d = {
        "Product": ["Samsung", "Nokia", "Samsung", "Motorola",
        "Nokia", "Samsung", "Samsung"],

        "Orders": [2, 4, 3, 4, 6, 7, 3]
}

df1 = pd.DataFrame(d)

grouped = df1.groupby(["Product"])
result = grouped.sum()

print(df1)
print()
print(result)
```

Output

```
    Product  Orders
0   Samsung       2
1     Nokia       4
2   Samsung       3
3  Motorola       4
4     Nokia       6
5   Samsung       7
6   Samsung       3

           Orders
Product
Motorola        4
Nokia          10
Samsung        15
```

Program Name: Get the number of product on dates
Input file: demo3.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
grouped = df1.groupby(["Mail_Id"])
result = grouped.size()

print(result)
```

Output

```
Mail_Id
daniel@gmail.com       3
kedar@gmail.com        3
nirekshan@gmail.com    4
partha@gmail.com       1
prasad@gmail.com       3
shahid@gmail.com       1
dtype: int64
```

Program Name: Get the number of product on count
Input file: demo4.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
grouped = df1.groupby(["Product_Name"])
result = grouped.size()

print(result)
```

Output

```
Product_Name
Kindle Paper White    3
LG Washing Machine    1
Samsung               3
Sofa set              1
hTC mobile            2
iPad                  1
iPhone 8              3
iPhone 9              1
dtype: int64
```

Program Name: How many products sold out on each month?

Input file: demo5.py

sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

cols = ['Date', 'Product_Name']
grouped = df1.groupby(cols)['Date']
result = grouped.count()

print(result)
```

Output

```
Date        Product_Name
09/01/2019  iPad                1
09/02/2019  LG Washing Machine  1
09/06/2019  Kindle Paper White  2
            Sofa set            1
            iPhone 8            1
10/31/2019  Kindle Paper White  1
11/02/2019  Samsung             2
            hTC mobile          1
11/06/2019  Samsung             1
            iPhone 8            2
            iPhone 9            1
11/10/2019  hTC mobile          1
Name: Date, dtype: int64
```

Program Name: Product wise each month sales
Input file: demo6.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

cols = ['Product_Name', 'Date']
grouped = df1.groupby(cols)['Date']
result = grouped.count()

print(result)
```

Output

```
Product_Name         Date
Kindle Paper White   09/06/2019    2
                     10/31/2019    1
LG Washing Machine   09/02/2019    1
Samsung              11/02/2019    2
                     11/06/2019    1
Sofa set             09/06/2019    1
hTC mobile           11/02/2019    1
                     11/10/2019    1
iPad                 09/01/2019    1
iPhone 8             09/06/2019    1
                     11/06/2019    2
iPhone 9             11/06/2019    1
Name: Date, dtype: int64
```

Program Name: Customer wise each month sales
Input file: demo7.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

cols = ['Mail_Id', 'Date']
grouped = df1.groupby(cols)['Mail_Id']
result = grouped.count()

print(result)
```

Output

```
Mail_Id              Date
daniel@gmail.com     09/06/2019    2
                     10/31/2019    1
kedar@gmail.com      09/06/2019    1
                     11/06/2019    2
nirekshan@gmail.com  09/06/2019    1
                     11/02/2019    2
                     11/06/2019    1
partha@gmail.com     11/06/2019    1
prasad@gmail.com     09/02/2019    1
                     11/02/2019    1
                     11/10/2019    1
shahid@gmail.com     09/01/2019    1
Name: Mail_Id, dtype: int64
```

Program Name: Customer wise product details
Input file: demo8.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

col = ['Mail_Id', 'Product_Name']
grouped = df1.groupby(col)['Product_Cost']
result = grouped.sum()

print(result)
```

Output

```
Mail_Id               Product_Name
Daniel@gmail.com      Kindle Paper White    20000
                      Sofa set              50000
kedar@gmail.com       iPhone 8              60000
nirekshan@gmail.com   Kindle Paper White    20000
                      Samsung               30000
partha@gmail.com      iPhone 9              30000
prasad@gmail.com      LG Washing Machine    25000
                      hTC mobile            30000
shahid@gmail.com      iPad                  70000
Name: Product_Cost, dtype: int64
```

Program Name: Customer wise product details
Input file: demo9.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

col = ['Mail_Id', 'Product_Name']
grouped = df1.groupby(col, as_index = False)['Product_Cost']
result = grouped.sum()

print(result)
```

Output

	Mail_Id	Product_Name	Product_Cost
0	Daniel@gmail.com	Kindle Paper White	20000
1	Daniel@gmail.com	Sofa set	50000
2	kedar@gmail.com	iPhone 8	60000
3	nirekshan@gmail.com	Kindle Paper White	20000
4	nirekshan@gmail.com	Samsung	30000
5	partha@gmail.com	iPhone 9	30000
6	prasad@gmail.com	LG Washing Machine	25000
7	prasad@gmail.com	hTC mobile	30000
8	shahid@gmail.com	iPad	70000

Program Name: Customer wise product sales
Input file: demo10.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
col = ['Mail_Id']
grouped = df1.groupby(col)['Product_Cost']
result = grouped.sum()

print(result)
```

Output

```
Mail_Id
Daniel@gmail.com        70000
kedar@gmail.com         60000
nirekshan@gmail.com     50000
partha@gmail.com        30000
prasad@gmail.com        55000
shahid@gmail.com        70000
Name: Product_Cost, dtype: int64
```

Program Name: Product wise whole sales
Input file: demo11.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
col = ['Product_Name']
grouped = df1.groupby(col)['Product_Cost']
result = grouped.sum()

print(result)
```

Output

```
Product_Name
Kindle Paper White    40000
LG Washing Machine    25000
Samsung               30000
Sofa set              50000
hTC mobile            30000
iPad                  70000
iPhone 8              60000
iPhone 9              30000
Name: Product_Cost, dtype: int64
```

Program Name: Day wise product sales
Input file: demo12.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
col = ['Date']
grouped = df1.groupby(col)['Product_Cost']
result = grouped.sum()

print(result)
```

Output

```
Date
09/01/2019     70000
09/02/2019     25000
09/06/2019    100000
10/31/2019     10000
11/02/2019     35000
11/06/2019     80000
11/10/2019     15000
Name: Product_Cost, dtype: int64
```

Program Name: Describe method
Input file: demo13.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")
df2 = df1['Product_Cost'].describe()

print(df2)
```

Output

```
count       15.000000
mean     22333.333333
std      16888.993316
min      10000.000000
25%      10000.000000
50%      20000.000000
75%      22500.000000
max      70000.000000
Name: Product_Cost, dtype: float64
```

Program Name: Applying a single function to columns in groups
Input file: demo14.py
sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

d = {
     'Product_Cost' : sum
}

cols = ['Date', 'Product_Name']
grouped = df1.groupby(cols)
result = grouped.agg(d)

print(result)
```

Output

```
                                Product_Cost
Date       Product_Name
09/01/2019 iPad                        70000
09/02/2019 LG Washing Machine          25000
09/06/2019 Kindle Paper White          30000
           Sofa set                    50000
           iPhone 8                    20000
10/31/2019 Kindle Paper White          10000
11/02/2019 Samsung                     20000
           hTC mobile                  15000
11/06/2019 Samsung                     10000
           iPhone 8                    40000
           iPhone 9                    30000
11/10/2019 hTC mobile                  15000
```

Program Name: Applying a single function to columns in groups

Name: demo15.py

Input file: sales5.py

```
import pandas as pd

df1 = pd.read_csv("sales5.csv")

d = {
    'Product_Cost': sum,
    'Product_Name': "count",
}

cols = ['Date', 'Product_Name']

grouped = df1.groupby(cols)
result = grouped.agg(d)

print(result)
```

Output

```
                              Product_Cost  Product_Name
Date       Product_Name
09/01/2019 iPad                     70000             1
09/02/2019 LG Washing Machine       25000             1
09/06/2019 Kindle Paper White       30000             2
           Sofa set                 50000             1
           iPhone 8                 20000             1
10/31/2019 Kindle Paper White       10000             1
11/02/2019 Samsung                  20000             2
           hTC mobile               15000             1
11/06/2019 Samsung                  10000             1
           iPhone 8                 40000             2
           iPhone 9                 30000             1
11/10/2019 hTC mobile               15000             1
```

Program Name: Applying a single function to columns in groups

Input file: demo16.py

Input file: sales5.py

```
import pandas as pd

df1 = pd.read_csv('sales5.csv')

d = {
     'Product_Cost': [min, max, sum]
}

col = ['Product_Cost']
grouped = df1.groupby(col)
result = grouped.agg(d)

print(result)
```

Output

```
             Product_Cost
                      min    max    sum
Product_Cost
10000                10000  10000  50000
15000                15000  15000  30000
20000                20000  20000  80000
25000                25000  25000  25000
30000                30000  30000  30000
50000                50000  50000  50000
70000                70000  70000  70000
```