EX N0:10

**Data Wrangling**

Question

Do the following wrangling operations, 1) Group by 2) Remove duplicates 3) Merge

Aim

The aim of this exercise is to demonstrate data wrangling operations including grouping, removing duplicates, and merging of DataFrames using pandas

Program:

import pandas as pd

# Creating data

car\_selling\_data = {

'Brand': ['Maruti', 'Maruti', 'Maruti', 'Maruti', 'Hyundai', 'Hyundai', 'Toyota', 'Mahindra', 'Mahindra', 'Ford', 'Toyota', 'Ford'],

'Year': [2010, 2011, 2009, 2013, 2010, 2011, 2011, 2010, 2013, 2010, 2010, 2011],

'Sold': [6, 7, 9, 8, 3, 5, 2, 8, 7, 2, 4, 2]

}

# Creating DataFrame

df = pd.DataFrame(car\_selling\_data)

print("\nOriginal DataFrame:\n", df)

# Grouping the data by year

grouped = df.groupby("Year")

print("\nGroup by year 2010:\n", grouped.get\_group(2010))

import pandas as pd

# Student data

student\_data = {

'Name': ['Amit', 'Praveen', 'Jagadesh', 'Rahul', 'Vishal', 'Suraj', 'Rishab', 'Sathish', 'Amit', 'Rahul', 'Praveen', 'Amit'],

'Roll no': [123, 54, 29, 36, 59, 38, 12, 45, 34, 36, 54, 23],

'Email': ['xxxx@gmail.com', 'xxxxxx@gmail.com', 'xxxxxx@gmail.com', 'xx@gmail.com', 'xxxxx@gmail.com', 'xxxxx@gmail.com', 'xxxx@gmail.com', 'xxxxx@gmail.com', 'xxxxxxx@gmail.com', 'xxxxxxxxxx@gmail.com', 'xxxxxxxxxx@gmail.com', 'xxxxxxxxxx@gmail.com']

}

# Creating DataFrame

df = pd.DataFrame(student\_data)

print("\nOriginal DataFrame:\n", df)

# Removing duplicate rows based on 'Roll no'

non\_duplicate = df[~df.duplicated('Roll no')]

# Printing non-duplicate values

print("\nRemoved duplicated rows:\n", non\_duplicate)

# Creating DataFrame

details = pd.DataFrame({

'ID': [101, 102, 103, 104, 105, 106, 107, 108, 109, 110],

'NAME': ['Arun', 'Praveen', 'Harish', 'Pooja', 'Rahul', 'Naresh', 'Saurabh', 'Anush', 'Dinesh', 'Mohit'],

'BRANCH': ['Mech', 'Mech', 'CSE', 'CSE', 'CSE', 'EEE', 'EEE', 'ECE', 'ECE', 'IT']

})

print("\nOriginal DataFrame 1:\n", details)

# Creating DataFrame

fees\_status = pd.DataFrame({

'ID': [101, 102, 103, 104, 105, 106, 107, 108, 109, 110],

'PENDING': ['5000', '250', 'NIL', '9000', '15000', 'NIL', '4500', '1800', '250', 'NIL']

})

print("\nOriginal DataFrame:\n", fees\_status)

# Merging DataFrames

print("\nMerged DataFrame:\n", pd.merge(details, fees\_status, on='ID'))

Output:

Original DataFrame:

Brand Year Sold

0 Maruti 2010 6

1 Maruti 2011 7

2 Maruti 2009 9

3 Maruti 2013 8

4 Hyundai 2010 3

5 Hyundai 2011 5

6 Toyota 2011 2

7 Mahindra 2010 8

8 Mahindra 2013 7

9 Ford 2010 2

10 Toyota 2010 4

11 Ford 2011 2

Group by year 2010:

Brand Year Sold

0 Maruti 2010 6

4 Hyundai 2010 3

7 Mahindra 2010 8

9 Ford 2010 2

10 Toyota 2010 4

Original DataFrame:

Name Roll no Email

0 Amit 123 xxxx@gmail.com

1 Praveen 54 xxxxxx@gmail.com

2 Jagadesh 29 xxxxxx@gmail.com

3 Rahul 36 xx@gmail.com

4 Vishal 59 xxxxx@gmail.com

5 Suraj 38 xxxxx@gmail.com

6 Rishab 12 xxxx@gmail.com

7 Sathish 45 xxxxx@gmail.com

9 Rahul 36 xxxxxxxxxxx@gmail.com

10 Praveen 54 xxxxxxxxxxx@gmail.com

Removed duplicated rows:

Name Roll no Email

0 Amit 123 xxxx@gmail.com

1 Praveen 54 xxxxxx@gmail.com

2 Jagadesh 29 xxxxxx@gmail.com

3 Rahul 36 xx@gmail.com

4 Vishal 59 xxxxx@gmail.com

5 Suraj 38 xxxxx@gmail.com

6 Rishab 12 xxxx@gmail.com

7 Sathish 45 xxxxx@gmail.com

Original DataFrame 1:

ID NAME BRANCH

0 101 Arun Mech

1 102 Praveen Mech

2 103 Harish CSE

3 104 Pooja CSE

4 105 Rahul CSE

5 106 Naresh EEE

6 107 Saurabh EEE

7 108 Anush ECE

8 109 Dinesh ECE

9 110 Mohit IT

Original DataFrame:

ID PENDING

0 101 5000

1 102 250

2 103 NIL

3 104 9000

4 105 15000

5 106 NIL

6 107 4500

7 108 1800

8 109 250

9 110 NIL

Merged DataFrame:

ID NAME BRANCH PENDING

0 101 Arun Mech 5000

1 102 Praveen Mech 250

2 103 Harish CSE NIL

3 104 Pooja CSE 9000

4 105 Rahul CSE 15000

5 106 Naresh EEE NIL

6 107 Saurabh EEE 4500

7 108 Anush ECE 1800

8 109 Dinesh ECE 250

9 110 Mohit IT NIL