Exp1 A_42

February 7, 2026

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
print(pd.__version__)
print(np.__version__)
```

```
2.3.3
2.3.5
```

[3]:
```
data={
    "Age": [33,35,np.nan,37,38],
    "Salary": [70000, np.nan, 90000, 50000, np.nan],
    "Experience": [3,5,2,np.nan,1],
    "Rating": [4.2,4.3,4.7,4.1,np.nan]
}
df=pd.DataFrame(data)
print(df)
```

```
    Age   Salary  Experience  Rating
0  33.0  70000.0         3.0     4.2
1  35.0      NaN         5.0     4.3
2   NaN  90000.0         2.0     4.7
3  37.0  50000.0         NaN     4.1
4  38.0      NaN         1.0     NaN
```

[4]:
```
print(df.head())
```

```
    Age   Salary  Experience  Rating
0  33.0  70000.0         3.0     4.2
1  35.0      NaN         5.0     4.3
2   NaN  90000.0         2.0     4.7
3  37.0  50000.0         NaN     4.1
4  38.0      NaN         1.0     NaN
```

[5]:
```
print(df.tail())
```

```
    Age   Salary  Experience  Rating
0  33.0  70000.0         3.0     4.2
1  35.0      NaN         5.0     4.3
2   NaN  90000.0         2.0     4.7
```

```
3  37.0  50000.0         NaN     4.1
4  38.0      NaN         1.0     NaN
```

[6]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Age         4 non-null      float64
 1   Salary      3 non-null      float64
 2   Experience  4 non-null      float64
 3   Rating      4 non-null      float64
dtypes: float64(4)
memory usage: 292.0 bytes
None
```

[7]:
```
print(df.describe())
```

```
             Age   Salary  Experience    Rating
count   4.000000      3.0    4.000000  4.000000
mean   35.750000  70000.0    2.750000  4.325000
std     2.217356  20000.0    1.707825  0.262996
min    33.000000  50000.0    1.000000  4.100000
25%    34.500000  60000.0    1.750000  4.175000
50%    36.000000  70000.0    2.500000  4.250000
75%    37.250000  80000.0    3.500000  4.400000
max    38.000000  90000.0    5.000000  4.700000
```

[8]:
```
#check null values column wise
print(df.isnull())
```

```
     Age  Salary  Experience  Rating
0  False   False       False   False
1  False    True       False   False
2   True   False       False   False
3  False   False        True   False
4  False    True       False    True
```

[9]:
```
#count missing values
print(df.isnull().sum())
```

```
Age           1
Salary        2
Experience    1
Rating        1
dtype: int64
```

[10]:
```
#total missing values in dataset
print(df.isnull().sum().sum())
```

```
5
```

[11]:
```
#fill with mean
df_filled = df.fillna(df.mean())
print(df_filled)
```

```
     Age   Salary  Experience  Rating
0  33.00  70000.0        3.00   4.200
1  35.00  70000.0        5.00   4.300
2  35.75  90000.0        2.00   4.700
3  37.00  50000.0        2.75   4.100
4  38.00  70000.0        1.00   4.325
```

[12]:
```
#fill with median
df_median = df.fillna(df.median())
print(df_median)
```

```
    Age   Salary  Experience  Rating
0  33.0  70000.0         3.0    4.20
1  35.0  70000.0         5.0    4.30
2  36.0  90000.0         2.0    4.70
3  37.0  50000.0         2.5    4.10
4  38.0  70000.0         1.0    4.25
```

[13]:
```
print(df_filled.isnull().sum())
```

```
Age           0
Salary        0
Experience    0
Rating        0
dtype: int64
```

[14]:
```
corr_matrix = df_filled.corr()
print(corr_matrix)
```

```
                 Age    Salary  Experience    Rating
Age         1.000000 -0.230144   -0.555650  0.046443
Salary     -0.230144  1.000000   -0.179284  0.931381
Experience -0.555650 -0.179284    1.000000 -0.273665
Rating      0.046443  0.931381   -0.273665  1.000000
```

[15]:
```
#second dataset
data ={
    "ID" : [1,2,3,4,5],
    "Age": [17, 21, np.nan, 20, 19],
    "Gender": ["Male", "Female", "Female", np.nan, "Male"],
    "Study_hours": [3,4,np.nan,5,2],
```

```
    "Attendance_percent": [75, 80, 68, np.nan, 80],
    "Marks": [ 85,82, np.nan, 90, 70]
}

df = pd.DataFrame(data)

#save as csv
df.to_csv("student_performance.csv", index = False)

print(df)
```

```
   ID   Age  Gender  Study_hours  Attendance_percent  Marks
0   1  17.0    Male          3.0                75.0   85.0
1   2  21.0  Female          4.0                80.0   82.0
2   3   NaN  Female          NaN                68.0    NaN
3   4  20.0     NaN          5.0                 NaN   90.0
4   5  19.0    Male          2.0                80.0   70.0
```

[16]:
```
df = pd.read_csv("student_performance.csv")
print(df)
```

```
   ID   Age  Gender  Study_hours  Attendance_percent  Marks
0   1  17.0    Male          3.0                75.0   85.0
1   2  21.0  Female          4.0                80.0   82.0
2   3   NaN  Female          NaN                68.0    NaN
3   4  20.0     NaN          5.0                 NaN   90.0
4   5  19.0    Male          2.0                80.0   70.0
```

[17]:
```
print(df.head())
```

```
   ID   Age  Gender  Study_hours  Attendance_percent  Marks
0   1  17.0    Male          3.0                75.0   85.0
1   2  21.0  Female          4.0                80.0   82.0
2   3   NaN  Female          NaN                68.0    NaN
3   4  20.0     NaN          5.0                 NaN   90.0
4   5  19.0    Male          2.0                80.0   70.0
```

[18]:
```
print(df.describe())
```

```
             ID        Age  Study_hours  Attendance_percent  Marks
count  5.000000   4.000000     4.000000            4.000000   4.00
mean   3.000000  19.250000     3.500000           75.750000  81.75
std    1.581139   1.707825     1.290994            5.678908   8.50
min    1.000000  17.000000     2.000000           68.000000  70.00
25%    2.000000  18.500000     2.750000           73.250000  79.00
50%    3.000000  19.500000     3.500000           77.500000  83.50
75%    4.000000  20.250000     4.250000           80.000000  86.25
max    5.000000  21.000000     5.000000           80.000000  90.00
```

[19]:
```
print(df.isnull())
```

```
      ID    Age  Gender  Study_hours  Attendance_percent  Marks
0  False  False   False        False               False  False
1  False  False   False        False               False  False
2  False   True   False         True               False   True
3  False  False    True        False                True  False
4  False  False   False        False               False  False
```

[20]:
```
# Numerical columns
num_cols = ["Age", "Study_hours", "Attendance_percent", "Marks"]
df[num_cols] = df[num_cols].fillna(df[num_cols].mean())

# Categorical column
df["Gender"] = df["Gender"].fillna(df["Gender"].mode()[0])

print(df)
```

```
   ID    Age  Gender  Study_hours  Attendance_percent  Marks
0   1  17.00    Male          3.0               75.00  85.00
1   2  21.00  Female          4.0               80.00  82.00
2   3  19.25  Female          3.5               68.00  81.75
3   4  20.00  Female          5.0               75.75  90.00
4   5  19.00    Male          2.0               80.00  70.00
```

[21]:
```
correlation_matrix = df[num_cols].corr()
print(correlation_matrix)
```

```
                        Age  Study_hours  Attendance_percent     Marks
Age                1.000000     0.529150            0.277102  0.051665
Study_hours        0.529150     1.000000           -0.176182  0.865726
Attendance_percent 0.277102    -0.176182            1.000000 -0.354338
Marks              0.051665     0.865726           -0.354338  1.000000
```

[22]:
```
#third dataset
df = pd.read_csv(r"C:\Users\wania\OneDrive\Desktop\ML_LAB\foodpanda_dataset.
↪csv")
```

[23]:
```
import os
os.getcwd()
```

[23]:
```
'C:\\Users\\wania\\Downloads'
```

[24]:
```
df.head()
```

[24]:
```
  customer_id  gender     age      city signup_date order_id order_date  \
0       C5663    Male   Adult  Peshawar   1/14/2024    O9663  8/23/2023
1       C2831    Male   Adult    Multan    7/7/2024    O6831  8/23/2023
2       C2851   Other  Senior    Multan   6/20/2025    O6851  8/23/2023
```

```
3       C1694  Female  Senior  Peshawar    9/5/2023    O5694  8/23/2023
4       C4339   Other  Senior    Lahore  12/29/2023    O8339  8/24/2023

  restaurant_name dish_name category  quantity    price payment_method  \
0      McDonald's    Burger  Italian         5  1478.27           Cash
1             KFC    Burger  Italian         3   956.04         Wallet
2       Pizza Hut     Fries  Italian         2   882.51           Cash
3          Subway     Pizza  Dessert         4   231.30           Card
4             KFC  Sandwich  Dessert         1  1156.69           Cash

   order_frequency last_order_date  loyalty_points   churned  rating  \
0               38       7/19/2025             238    Active       3
1               24      11/25/2024              81    Active       2
2               42       5/10/2025              82  Inactive       3
3               27       7/24/2025              45  Inactive       2
4               35      12/21/2024             418  Inactive       3

  rating_date delivery_status
0  10/14/2024       Cancelled
1   8/21/2025         Delayed
2   9/19/2024         Delayed
3   6/29/2025         Delayed
4    3/6/2025       Cancelled
```

[25]:
```
df.tail()
```

[25]:
```
     customer_id  gender       age       city signup_date order_id order_date  \
5995       C6849    Male     Adult     Multan  11/25/2024   O10849  8/22/2025
5996       C3787  Female     Adult  Islamabad   1/28/2025    O7787  8/22/2025
5997       C2841   Other  Teenager  Islamabad  10/19/2023    O6841  8/22/2025
5998       C1624    Male     Adult  Islamabad   6/17/2024    O5624  8/22/2025
5999       C2068  Female     Adult     Multan   3/15/2025    O6068  8/22/2025

     restaurant_name dish_name   category  quantity    price payment_method  \
5995       Pizza Hut    Burger    Italian         4   875.71           Cash
5996             KFC     Pizza    Italian         5  1118.26           Cash
5997             KFC  Sandwich    Italian         4  1005.83           Card
5998             KFC     Fries  Fast Food         4  1226.10           Card
5999     Burger King     Fries  Fast Food         3  1131.27           Card

      order_frequency last_order_date  loyalty_points   churned  rating  \
5995               28      11/29/2024             166    Active       5
5996               12       6/8/2025              193  Inactive       3
5997               31      12/30/2024             278    Active       4
5998               37      12/27/2024              55  Inactive       2
5999                2       6/13/2025              41  Inactive       1
```

```
     rating_date delivery_status
5995  12/30/2024       Cancelled
5996    2/9/2025         Delayed
5997   3/23/2025       Cancelled
5998   3/15/2025       Delivered
5999   7/15/2025         Delayed
```

[26]:
```
print(df.columns)
```

```
Index(['customer_id', 'gender', 'age', 'city', 'signup_date', 'order_id',
       'order_date', 'restaurant_name', 'dish_name', 'category', 'quantity',
       'price', 'payment_method', 'order_frequency', 'last_order_date',
       'loyalty_points', 'churned', 'rating', 'rating_date',
       'delivery_status'],
      dtype='object')
```

[27]:
```
df.isnull().sum()
```

[27]:
```
customer_id        0
gender             0
age                0
city               0
signup_date        0
order_id           0
order_date         0
restaurant_name    0
dish_name          0
category           0
quantity           0
price              0
payment_method     0
order_frequency    0
last_order_date    0
loyalty_points     0
churned            0
rating             0
rating_date        0
delivery_status    0
dtype: int64
```

[28]:
```
num_cols = df.select_dtypes(include=np.number).columns
cat_cols = df.select_dtypes(exclude=np.number).columns
```

[29]:
```
df[num_cols].corr()
```

[29]:
```
                quantity     price  order_frequency  loyalty_points    rating
quantity        1.000000 -0.007424         0.011473        0.006136 -0.017778
price          -0.007424  1.000000         0.011608       -0.013781 -0.003585
```

```
order_frequency  0.011473  0.011608         1.000000        0.006744 -0.015788
loyalty_points   0.006136 -0.013781         0.006744        1.000000  0.008666
rating          -0.017778 -0.003585        -0.015788        0.008666  1.000000
```

[ ]:

[ ]: