import pandas as pd

# location of data 'order\_data.xlsx'

data = pd.read\_excel(r"C:\Users\tina\OneDrive\Desktop\tina-project\order\_data.xlsx")

# some missing 'Purchase\_Date' information.

def impute\_purchase\_date(row):

user\_id = row['User\_ID']

product\_id = row['Product\_ID']

# I guess the purchase date based on other purchases

# by the same user. For example, what was the average purchase date

# for this user?

# (Replace this actual logic to find a suitable substitute)

average\_purchase\_date = data[data['User\_ID'] == user\_id]['Purchase\_Date'].mean()

# Let's use that average date as a guess for the missing one

return average\_purchase\_date

# Now I can use this function to fill in the missing 'Purchase\_Date' values

data['Purchase\_Date'] = data.apply(lambda row: impute\_purchase\_date(row)

if pd.isna(row['Purchase\_Date']) else row['Purchase\_Date'],

axis=1)

# Finally, I can save this cleaned data to a new Excel file

data.to\_excel('cleaned\_data.xlsx', index=False) # Adjust filename as needed

print("Cleaned data exported to cleaned\_data.xlsx")

User\_ID Product\_ID Purchase\_Date Price Category

0 1001 P1001 2021-01-01 00:00:00 49.99 Electronics

1 1001 P1002 2021-01-03 00:00:00 19.99 Home and Kitchen

2 1001 P1003 2021-01-10 00:00:00 NaN Electronics

3 1001 P1004 2021-01-15 00:00:00 9.99 NaN

4 1002 P1001 2021-02-02 00:00:00 NaN Electronics

5 1002 P1002 2021-02-03 00:00:00 19.99 NaN

6 1002 P1003 2021-02-08 00:00:00 149.99 Electronics

7 1003 P1002 2021-03-05 00:00:00 19.99 Home and Kitchen

8 1003 P1004 2021-03-06 00:00:00 NaN Beauty

9 1004 P1001 2021-04-02 00:00:00 49.99 Electronics

10 1004 P1002 2021-04-03 00:00:00 NaN Home and Kitchen

11 1004 P1004 2021-04-08 00:00:00 9.99 Beauty

12 1005 P1001 2021-05-01 00:00:00 49.99 Electronics

13 1005 P1003 2021-05-03 00:00:00 NaN Electronics

14 1005 P1004 2021-05-08 00:00:00 9.99 NaN

15 1006 P1002 2021-06-03 00:00:00 19.99 Home and Kitchen

16 1006 P1003 2021-06-03 00:00:00 149.99 Electronics

17 1007 P1001 2021-07-05 00:00:00 49.99 NaN

18 1007 P1004 2021-07-06 00:00:00 9.99 Beauty

19 1008 P1001 2021-08-02 00:00:00 49.99 Electronics

20 1008 P1002 2021-08-03 00:00:00 19.99 Home and Kitchen

21 1008 P1003 2021-08-08 00:00:00 NaN Electronics

22 1008 P1004 2021-08-15 00:00:00 9.99 Beauty

23 1009 P1001 2021-09-01 00:00:00 49.99 Electronics

24 1009 P1002 2021-09-08 16:00:00 19.99 Home and Kitchen

25 1009 P1003 2021-09-10 00:00:00 149.99 NaN

26 1009 P1004 2021-09-15 00:00:00 NaN Beauty

27 1010 P1001 2021-10-01 00:00:00 49.99 Electronics

28 1010 P1002 2021-10-03 00:00:00 19.99 Home and Kitchen

29 1010 P1003 2021-10-10 00:00:00 149.99 NaN

30 1010 P1004 2021-10-15 00:00:00 9.99 NaN

Cleaned data exported to cleaned\_data.xlsx