jupyter Untitled1 Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help

Trusted

Code JupyterLab Python 3 (ipykernel)

[1]:
```
#import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
#read files by pandas

df1 = pd.read_csv(r"G:\AI\project2\ecommerce_customer_data_custom_ratios.csv")

df2 = pd.read_csv(r"G:\AI\project2\ecommerce_customer_data_large.csv")
```

[3]:
```
#display table 1
df1
```

[3]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
0	46251	2020-09-08 09:38:32	Electronics	12	3	740	Credit Card	37	0.0	Christine Hernandez	37	Male	0
1	46251	2022-03-05 12:56:35	Home	468	4	2739	PayPal	37	0.0	Christine Hernandez	37	Male	0
2	46251	2022-05-23 18:18:01	Home	288	2	3196	PayPal	37	0.0	Christine Hernandez	37	Male	0
3	46251	2020-11-12 13:13:29	Clothing	196	1	3509	PayPal	37	0.0	Christine Hernandez	37	Male	0
4	13593	2020-11-27 17:55:11	Home	449	1	3452	Credit Card	49	0.0	James Grant	49	Female	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
249995	33308	2023-08-10 13:39:06	Clothing	279	2	2187	PayPal	55	1.0	Michelle Flores	55	Male	1
249996	48835	2021-11-23 01:30:42	Home	27	1	3615	Credit Card	42	1.0	Jeremy Rush	42	Female	1
249997	21019	2020-07-02 14:04:48	Home	17	5	2466	Cash	41	0.0	Tina Craig	41	Male	0
249998	49234	2020-12-30 02:02:40	Books	398	2	3668	Crypto	34	0.0	Jennifer Cooper	34	Female	1
249999	16971	2021-03-13 16:28:35	Electronics	425	4	2370	Cash	36	1.0	Justin Lawson	36	Female	1

250000 rows × 13 columns

[4]:
```
#display table 2
df2
```

[4]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
0	44605	2023-05-03 21:30:02	Home	177	1	2427	PayPal	31	1.0	John Rivera	31	Female	0
1	44605	2021-05-16 13:57:44	Electronics	174	3	2448	PayPal	31	1.0	John Rivera	31	Female	0
2	44605	2020-07-13 06:16:57	Books	413	1	2345	Credit Card	31	1.0	John Rivera	31	Female	0
3	44605	2023-01-17 13:14:36	Electronics	396	3	937	Cash	31	0.0	John Rivera	31	Female	0
4	44605	2021-05-01 11:29:27	Books	259	4	2598	PayPal	31	1.0	John Rivera	31	Female	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
249995	33807	2023-01-24 12:32:18	Home	436	1	3664	Cash	63	0.0	Gabriel Williams	63	Male	0
249996	20455	2021-06-04 05:45:25	Electronics	233	1	4374	Credit Card	66	1.0	Barry Foster	66	Female	0
249997	28055	2022-11-10 17:11:57	Electronics	441	5	5296	Cash	63	NaN	Lisa Johnson	63	Female	0
249998	15023	2021-06-27 14:42:12	Electronics	44	2	2517	Cash	64	1.0	Melissa Fernandez	64	Male	0

249999	4148	2020-09-07 05:12:19	Home	307	5	3634	Cash	32	0.0	Angela Norton	32	Male	0

250000 rows × 13 columns

[5]:
```
#take information from tables
df1.info()
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Customer ID             250000 non-null  int64
 1   Purchase Date           250000 non-null  object
 2   Product Category        250000 non-null  object
 3   Product Price           250000 non-null  int64
 4   Quantity                250000 non-null  int64
 5   Total Purchase Amount   250000 non-null  int64
 6   Payment Method          250000 non-null  object
 7   Customer Age            250000 non-null  int64
 8   Returns                 202404 non-null  float64
 9   Customer Name           250000 non-null  object
 10  Age                     250000 non-null  int64
 11  Gender                  250000 non-null  object
 12  Churn                   250000 non-null  int64
dtypes: float64(1), int64(7), object(5)
memory usage: 24.8+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Customer ID             250000 non-null  int64
 1   Purchase Date           250000 non-null  object
 2   Product Category        250000 non-null  object
 3   Product Price           250000 non-null  int64
 4   Quantity                250000 non-null  int64
 5   Total Purchase Amount   250000 non-null  int64
 6   Payment Method          250000 non-null  object
 7   Customer Age            250000 non-null  int64
 8   Returns                 202618 non-null  float64
 9   Customer Name           250000 non-null  object
 10  Age                     250000 non-null  int64
 11  Gender                  250000 non-null  object
 12  Churn                   250000 non-null  int64
dtypes: float64(1), int64(7), object(5)
memory usage: 24.8+ MB
```

[6]:
```
# make describe to table 1
df1.describe()
```

[6]:

	Customer ID	Product Price	Quantity	Total Purchase Amount	Customer Age	Returns	Age	Churn
count	250000.00000	250000.000000	250000.000000	250000.000000	250000.000000	202404.000000	250000.000000	250000.000000
mean	25004.03624	254.659512	2.998896	2725.370732	43.940528	0.497861	43.940528	0.199496
std	14428.27959	141.568577	1.414694	1442.933565	15.350246	0.499997	15.350246	0.399622
min	1.00000	10.000000	1.000000	100.000000	18.000000	0.000000	18.000000	0.000000
25%	12497.75000	132.000000	2.000000	1477.000000	31.000000	0.000000	31.000000	0.000000
50%	25018.00000	255.000000	3.000000	2724.000000	44.000000	0.000000	44.000000	0.000000
75%	37506.00000	377.000000	4.000000	3974.000000	57.000000	1.000000	57.000000	0.000000
max	50000.00000	500.000000	5.000000	5350.000000	70.000000	1.000000	70.000000	1.000000

[7]:
```
# make describe to table 2

df2.describe()
```

[7]:

	Customer ID	Product Price	Quantity	Total Purchase Amount	Customer Age	Returns	Age	Churn
count	250000.000000	250000.000000	250000.000000	250000.000000	250000.000000	202618.000000	250000.000000	250000.00000
mean	25017.632092	254.742724	3.004936	2725.385196	43.798276	0.500824	43.798276	0.20052
std	14412.515718	141.738104	1.414737	1442.576095	15.364915	0.500001	15.364915	0.40039
min	1.000000	10.000000	1.000000	100.000000	18.000000	0.000000	18.000000	0.00000
25%	12590.000000	132.000000	2.000000	1476.000000	30.000000	0.000000	30.000000	0.00000
50%	25011.000000	255.000000	3.000000	2725.000000	44.000000	1.000000	44.000000	0.00000
75%	37441.250000	377.000000	4.000000	3975.000000	57.000000	1.000000	57.000000	0.00000
max	50000.000000	500.000000	5.000000	5350.000000	70.000000	1.000000	70.000000	1.00000

[8]:
```
#display first 10 rows
df1.head(10)
```

[8]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn

0	46251	2020-09-08 09:38:32	Electronics	12	3	740	Credit Card	37	0.0	Christine Hernandez	37	Male	0
1	46251	2022-03-05 12:56:35	Home	468	4	2739	PayPal	37	0.0	Christine Hernandez	37	Male	0
2	46251	2022-05-23 18:18:01	Home	288	2	3196	PayPal	37	0.0	Christine Hernandez	37	Male	0
3	46251	2020-11-12 13:13:29	Clothing	196	1	3509	PayPal	37	0.0	Christine Hernandez	37	Male	0
4	13593	2020-11-27 17:55:11	Home	449	1	3452	Credit Card	49	0.0	James Grant	49	Female	1
5	13593	2023-03-07 14:17:42	Home	250	4	575	PayPal	49	1.0	James Grant	49	Female	1
6	13593	2023-04-15 03:02:33	Electronics	73	1	1896	Credit Card	49	0.0	James Grant	49	Female	1
7	13593	2021-03-27 21:23:28	Books	337	2	2937	Cash	49	0.0	James Grant	49	Female	1
8	13593	2020-05-05 20:14:00	Clothing	182	2	3363	PayPal	49	1.0	James Grant	49	Female	1
9	28805	2023-09-13 04:24:00	Electronics	394	2	1993	Credit Card	19	0.0	Jose Collier	19	Male	0

[9]:
```
#display random 10 rows
df1.sample(10)
```

[9]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
19428	45706	2023-04-13 12:44:47	Books	271	1	666	Credit Card	65	0.0	Deanna Page	65	Female	0
23091	15924	2022-11-29 17:16:54	Clothing	258	1	1280	Credit Card	21	0.0	Adrian Anderson	21	Female	0
143176	27611	2022-05-10 02:59:23	Books	410	5	4486	Crypto	41	0.0	Thomas Dixon	41	Male	1
247524	21495	2020-06-14 08:19:02	Clothing	138	3	1200	Credit Card	58	1.0	Clayton Hull	58	Female	0
83155	2610	2021-09-16 15:52:00	Electronics	249	5	2971	Credit Card	62	1.0	Colleen Ramirez	62	Female	0
43098	45063	2022-04-14 19:22:30	Clothing	243	5	754	Credit Card	52	0.0	Michael Jacobs	52	Female	0
209326	26597	2022-07-09 13:44:39	Clothing	332	4	3000	Credit Card	35	0.0	Deborah Long	35	Male	0
212139	9384	2021-02-14 15:38:38	Clothing	62	5	2575	Credit Card	29	1.0	Stacy Greene	29	Female	1
134359	37062	2021-03-01 16:40:48	Clothing	447	1	3726	Crypto	27	0.0	Julie Barrett	27	Female	1
179006	29074	2023-04-19 09:45:04	Books	356	2	1126	PayPal	49	1.0	Bonnie Jones	49	Female	0

[10]:
```
#display last 10 rows

df2.tail(10)
```

[10]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
249990	150	2023-08-23 01:47:29	Home	196	3	1743	PayPal	68	0.0	Stephanie Morris	68	Male	0
249991	16247	2020-06-24 17:29:27	Electronics	368	1	1720	Credit Card	22	0.0	Mike Thompson	22	Male	0
249992	48538	2020-12-13 02:09:15	Electronics	239	1	2356	Cash	61	1.0	Alexis Nelson	61	Male	1
249993	39806	2021-12-07 05:42:38	Electronics	215	3	2349	Cash	60	1.0	Dana Brown	60	Female	0
249994	39806	2021-08-01 04:43:12	Electronics	225	5	5293	Credit Card	60	0.0	Dana Brown	60	Female	0
249995	33807	2023-01-24 12:32:18	Home	436	1	3664	Cash	63	0.0	Gabriel Williams	63	Male	0
249996	20455	2021-06-04 05:45:25	Electronics	233	1	4374	Credit Card	66	1.0	Barry Foster	66	Female	0
249997	28055	2022-11-10 17:11:57	Electronics	441	5	5296	Cash	63	NaN	Lisa Johnson	63	Female	0
249998	15023	2021-06-27 14:42:12	Electronics	44	2	2517	Cash	64	1.0	Melissa Fernandez	64	Male	0
249999	4148	2020-09-07	Home	307	5	3634	Cash	32	0.0	Angela Norton	32	Male	0

05:12:19

[11]:
```
#dispaly sum of nan is table1
df1.isna().sum()
```

[11]:
```
Customer ID                0
Purchase Date              0
Product Category           0
Product Price              0
Quantity                   0
Total Purchase Amount      0
Payment Method             0
Customer Age               0
Returns                47596
Customer Name              0
Age                        0
Gender                     0
Churn                      0
dtype: int64
```

[12]:
```
#dispaly sum of nan is table2
df2.isna().sum()
```

[12]:
```
Customer ID                0
Purchase Date              0
Product Category           0
Product Price              0
Quantity                   0
Total Purchase Amount      0
Payment Method             0
Customer Age               0
Returns                47382
Customer Name              0
Age                        0
Gender                     0
Churn                      0
dtype: int64
```

[52]:
```
#reblace nan with 0
df1['Returns'] = df1['Returns'].fillna(0)
df1.sample(5)
```

[52]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
246095	35346	2020-01-04 19:22:03	Books	265	5	1980	Crypto	24	1.0	April Dorsey	24	Male	1
94661	35037	2020-04-24 23:01:32	Electronics	89	1	4218	Credit Card	57	1.0	Mrs. Jennifer Lucas	57	Female	1
3126	5980	2021-03-07 17:23:25	Clothing	138	1	5045	Cash	61	0.0	Jamie Rodriguez	61	Female	0
240205	41727	2023-02-09 07:21:19	Clothing	17	3	2441	Credit Card	25	0.0	Mary Mccarthy	25	Male	0
173569	20193	2021-12-25 03:56:02	Electronics	137	1	2548	Cash	68	0.0	Ashley Webb	68	Female	0

[50]:
```
#reblace nan with 0
df2['Returns'] = df2['Returns'].fillna(0)
df2.sample(5)
```

[50]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
174154	33619	2023-05-21 21:08:10	Clothing	18	3	4478	Cash	47	1.0	Brian Martin	47	Female	0
199608	2641	2021-01-11 09:48:10	Electronics	286	5	5058	Credit Card	43	0.0	Michael Williams	43	Male	1
136483	5978	2021-12-12 19:27:59	Books	236	3	346	PayPal	46	0.0	Henry Day	46	Female	1
18513	7241	2020-06-16 17:08:38	Electronics	163	3	2747	Cash	41	1.0	Sara Briggs	41	Male	1
53361	6189	2021-04-16 20:34:11	Books	321	1	2865	Credit Card	47	1.0	Phillip Hebert	47	Female	1

[15]:
```
df1['Purchase Date'] = pd.to_datetime(df1['Purchase Date'])

df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Customer ID           250000 non-null  int64
 1   Purchase Date         250000 non-null  datetime64[ns]
 2   Product Category      250000 non-null  object
 3   Product Price         250000 non-null  int64
 4   Quantity              250000 non-null  int64
 5   Total Purchase Amount 250000 non-null  int64
```

```
 6   Payment Method         250000 non-null  object
 7   Customer Age           250000 non-null  int64
 8   Returns                250000 non-null  float64
 9   Customer Name          250000 non-null  object
 10  Age                    250000 non-null  int64
 11  Gender                 250000 non-null  object
 12  Churn                  250000 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(7), object(4)
memory usage: 24.8+ MB
```

[16]:
```
df2['Purchase Date'] = pd.to_datetime(df2['Purchase Date'])

df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Customer ID            250000 non-null  int64
 1   Purchase Date          250000 non-null  datetime64[ns]
 2   Product Category       250000 non-null  object
 3   Product Price          250000 non-null  int64
 4   Quantity               250000 non-null  int64
 5   Total Purchase Amount  250000 non-null  int64
 6   Payment Method         250000 non-null  object
 7   Customer Age           250000 non-null  int64
 8   Returns                250000 non-null  float64
 9   Customer Name          250000 non-null  object
 10  Age                    250000 non-null  int64
 11  Gender                 250000 non-null  object
 12  Churn                  250000 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(7), object(4)
memory usage: 24.8+ MB
```

[17]:
```
#dim of table 1
df1.shape
```

[17]:
```
(250000, 13)
```

[18]:
```
#dim of table 2

df2.shape
```

[18]:
```
(250000, 13)
```

[19]:
```
#duplicated in table 1
df1.duplicated().sum()
```

[19]:
```
0
```

[20]:
```
#duplicated in table 1

df2.duplicated().sum()
```

[20]:
```
0
```

[21]:
```
#number of values in column of Customer Name
df1['Customer Name'].value_counts()
```

[21]:
```
Customer Name
Michael Smith       107
John Smith          103
Jennifer Smith      102
Michael Johnson      98
Lisa Smith           97
                   ...
Haley Harris          1
Joseph Joyce          1
Jonathan Price MD     1
Sarah Melton          1
Justin Lawson         1
Name: count, Length: 39920, dtype: int64
```

[62]:
```
#download table1
df1.to_csv(r'G:\AI\project2\ecommerce_customer_data_custom_ratios-NEW.csv')
```

[58]:
```
#download table2

df2.to_csv(r"G:\AI\project2\ecommerce_customer_data_large-NEW.csv")
```

[ ]:
```

```