notebooks-02-data-cleaning

January 6, 2026

0.1 DATA CLEANING & KPI BASE TABLE

[10]:

```
import pandas as pd
import numpy as np
from faker import Faker
from datetime import timedelta
```

[13]:

```
merchants = pd.read_csv(
    r"D:\Shopify-Revenue-Growth-Analytics\Data\merchants.csv",
    parse_dates=["signup_date", "churn_date"],
    dayfirst=True
)

orders = pd.read_csv(
    r"D:\Shopify-Revenue-Growth-Analytics\Data\orders.csv",
    parse_dates=["order_date"],
    dayfirst=True
)
```

[27]:

```
print("Merchants")
merchants.head()
```

```
Merchants
```

[27]:

```
  merchant_id signup_date    country plan_type industry churned churn_date
0      M00001  2023-04-13  Australia     Basic     Food     Yes 2023-06-14
1      M00002  2024-03-11      India   Shopify     Food     Yes 2024-06-11
2      M00003  2023-09-28         US     Basic   Beauty      No        NaT
3      M00004  2023-04-17      India   Shopify     Food      No        NaT
4      M00005  2023-03-13     Canada     Basic   Beauty      No        NaT
```

[28]:

```
print("Orders")
orders.head()
```

```
Orders
```

[28]:

```
                               order_id merchant_id  order_date  order_value  \
0  2889ce8a-0fc7-4825-9fbe-dc82836e870e      M00001  2023-06-07        71.70
```

```
1  a76930d9-0876-4ec2-a851-03c709fad1fe      M00001  2023-06-11         33.30
2  6f5d7660-15a0-4904-93e5-82c5cc3eb6ab      M00001  2023-04-21         31.42
3  cf562f8e-fac2-4209-84f4-d33ff71a84d3      M00001  2023-04-30         64.98
4  79fbad9d-95a5-4231-a657-6edddcf13b31      M00001  2023-05-14         44.17

  channel payment_method
0     Web         Wallet
1     Web           Card
2  Mobile         Wallet
3     Web         Wallet
4  Mobile           Card
```

[14]:

```
gmv = orders.groupby("merchant_id")["order_value"].sum().reset_index()
gmv.columns = ["merchant_id", "total_gmv"]
```

[15]:

```
merchant_kpi = merchants.merge(gmv, on="merchant_id", how="left")
merchant_kpi["total_gmv"] = merchant_kpi["total_gmv"].fillna(0)
```

[18]:

```
merchant_kpi.head()
```

[18]:

```
  merchant_id signup_date    country plan_type industry churned churn_date  \
0      M00001  2023-04-13  Australia     Basic     Food     Yes 2023-06-14
1      M00002  2024-03-11      India   Shopify     Food     Yes 2024-06-11
2      M00003  2023-09-28         US     Basic   Beauty      No        NaT
3      M00004  2023-04-17      India   Shopify     Food      No        NaT
4      M00005  2023-03-13     Canada     Basic   Beauty      No        NaT

   total_gmv
0    3104.35
1    5336.55
2    2656.85
3   17345.86
4    4364.08
```

[19]:

```
merchant_kpi.info()
merchant_kpi.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   merchant_id  1200 non-null   object
 1   signup_date  1200 non-null   datetime64[ns]
 2   country      1200 non-null   object
 3   plan_type    1200 non-null   object
 4   industry     1200 non-null   object
 5   churned      1200 non-null   object
 6   churn_date   357 non-null    datetime64[ns]
 7   total_gmv    1200 non-null   float64
dtypes: datetime64[ns](2), float64(1), object(5)
memory usage: 75.1+ KB
```

```
[19]:               signup_date                     churn_date     total_gmv
       count                 1200                            357   1200.000000
       mean   2023-12-26 06:49:12  2024-07-04 11:33:46.890756352  10097.258158
       min    2023-01-01 00:00:00            2023-01-29 00:00:00      0.000000
       25%    2023-06-17 00:00:00            2024-01-09 00:00:00   4056.442500
       50%    2023-12-26 12:00:00            2024-07-13 00:00:00   7266.680000
       75%    2024-07-03 00:00:00            2024-12-19 00:00:00  12712.050000
       max    2024-12-30 00:00:00            2025-11-30 00:00:00  52861.490000
       std                    NaN                            NaN   9678.223845
```

```
[20]: merchant_kpi.to_csv("D:/Shopify-Revenue-Growth-Analytics/Data/merchant_kpi.
        ↪csv", index=False)
```