

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from lifetimes import BetaGeoFitter, GammaGammaFitter
from lifetimes.utils import (
    summary_data_from_transaction_data,
    calibration_and_holdout_data,
)
from lifetimes.plotting import plot_cumulative_transactions

plt.style.use("ggplot")
sns.set_palette("husl")

# =====
# 0:
# =====
file_path = "/Users/sousekilyu/Downloads/online_retail_II.xlsx"

print(">> 0:    ...")
#   Sheet
df_0910 = pd.read_excel(file_path, sheet_name="Year 2009-2010")
df_1011 = pd.read_excel(file_path, sheet_name="Year 2010-2011")
df_raw = pd.concat([df_0910, df_1011])
df_raw.head
```

```
>> 0: ...
<bound method NDFrame.head of
   Invoice StockCode ... Customer ID
0      489434    85048 ... 13085.0 United Kingdom
1      489434    79323P ... 13085.0 United Kingdom
2      489434    79323W ... 13085.0 United Kingdom
3      489434    22041 ... 13085.0 United Kingdom
4      489434    21232 ... 13085.0 United Kingdom
...
541905  581587    22899 ... 12680.0 France
541906  581587    23254 ... 12680.0 France
541907  581587    23255 ... 12680.0 France
541908  581587    22138 ... 12680.0 France
541909  581587      POST ... 12680.0 France
```

```
[1067371 rows x 8 columns]>
```

```
# =====
# 1: (Data Cleaning)
# =====
print("=> 1: ...")
df_cleaned = df_raw.copy()

# ID (C)
df_cleaned = df_cleaned[df_cleaned["Customer ID"].notna()]
df_cleaned = df_cleaned[~df_cleaned["Invoice"].astype(str).str.startswith("C")]
df_cleaned = df_cleaned[(df_cleaned["Quantity"] > 0) & (df_cleaned["Price"] > 0)]

#
df_cleaned["TotalRevenue"] = df_cleaned["Quantity"] * df_cleaned["Price"]
df_cleaned["InvoiceDate"] = pd.to_datetime(df_cleaned["InvoiceDate"])

print(f" : {len(df_cleaned)}")

>> 1: ...
: 805549
```