online retail analysis

dataset

https://archive.ics.uci.edu/dataset/352/online+retail

This **is a transnational dataset** that contains all the transactions **occurring between 01/12/2010 and 09/12/2011 for** a **UK-based and registered non-store online retail**. The company mainly **sells unique all-occasion gifts**. Many customers of the company are **wholesalers***.

* a wholesaler is a person or business that sells items to retail stores that will then sell them to individual customers for a higher price.

let's get started >>

exploratory data analysis

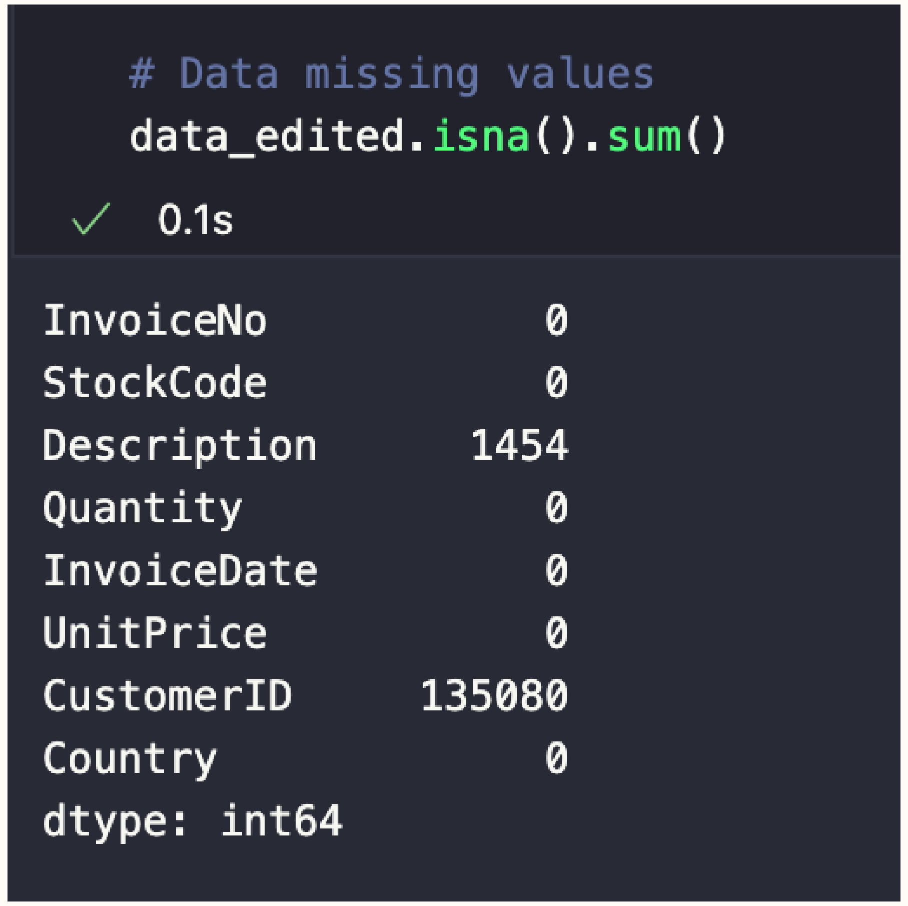
the dataset has

8 Columns including :

- InvoiceNo
- StockCode
- Description
- Quantity
- InvoiceDate
- UnitPrice
- CustomerID
- Country

and **541,909 rows**

but wait...
it has **missing values**



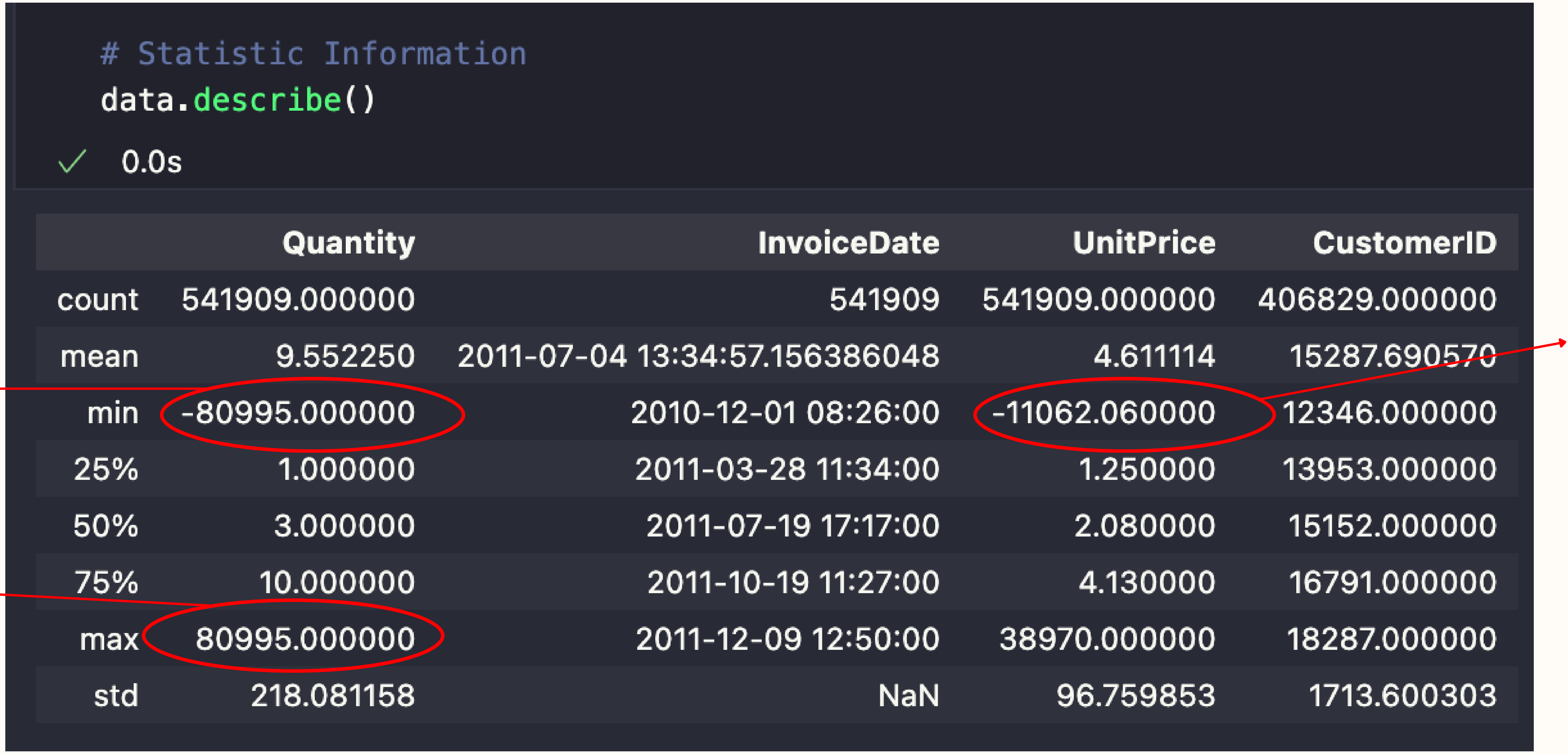
```
# Data missing values
data_edited.isna().sum()
```

✓ 0.1s

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

and it is not the only problem

illogical values within



```
# Statistic Information
data.describe()
```

✓ 0.0s

```
          Quantity                    InvoiceDate      UnitPrice     CustomerID
count  541909.000000                       541909  541909.000000  406829.000000
mean        9.552250  2011-07-04 13:34:57.156386048       4.611114   15287.690570
min    -80995.000000          2010-12-01 08:26:00  -11062.060000   12346.000000
25%         1.000000          2011-03-28 11:34:00       1.250000   13953.000000
50%         3.000000          2011-07-19 17:17:00       2.080000   15152.000000
75%        10.000000          2011-10-19 11:27:00       4.130000   16791.000000
max     80995.000000          2011-12-09 12:50:00   38970.000000   18287.000000
std       218.081158                          NaN      96.759853    1713.600303
```

minus items?

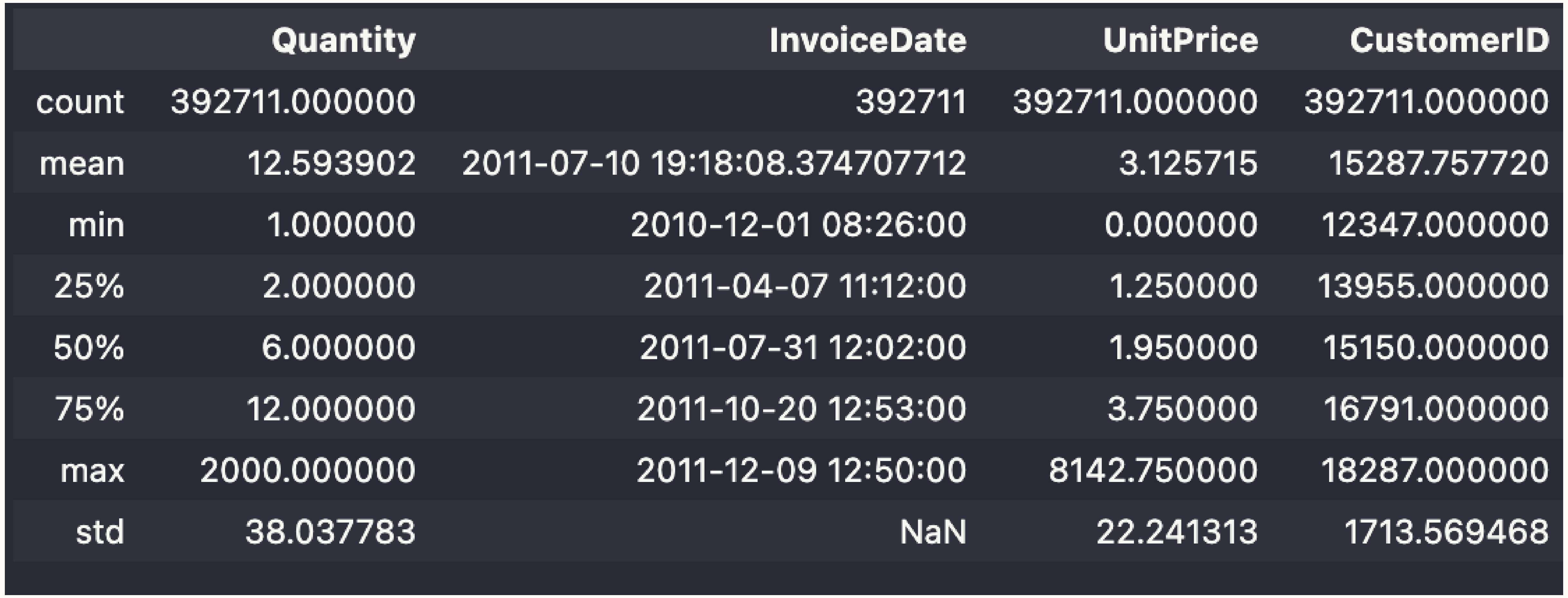
deficit price???

80k items in one buy??

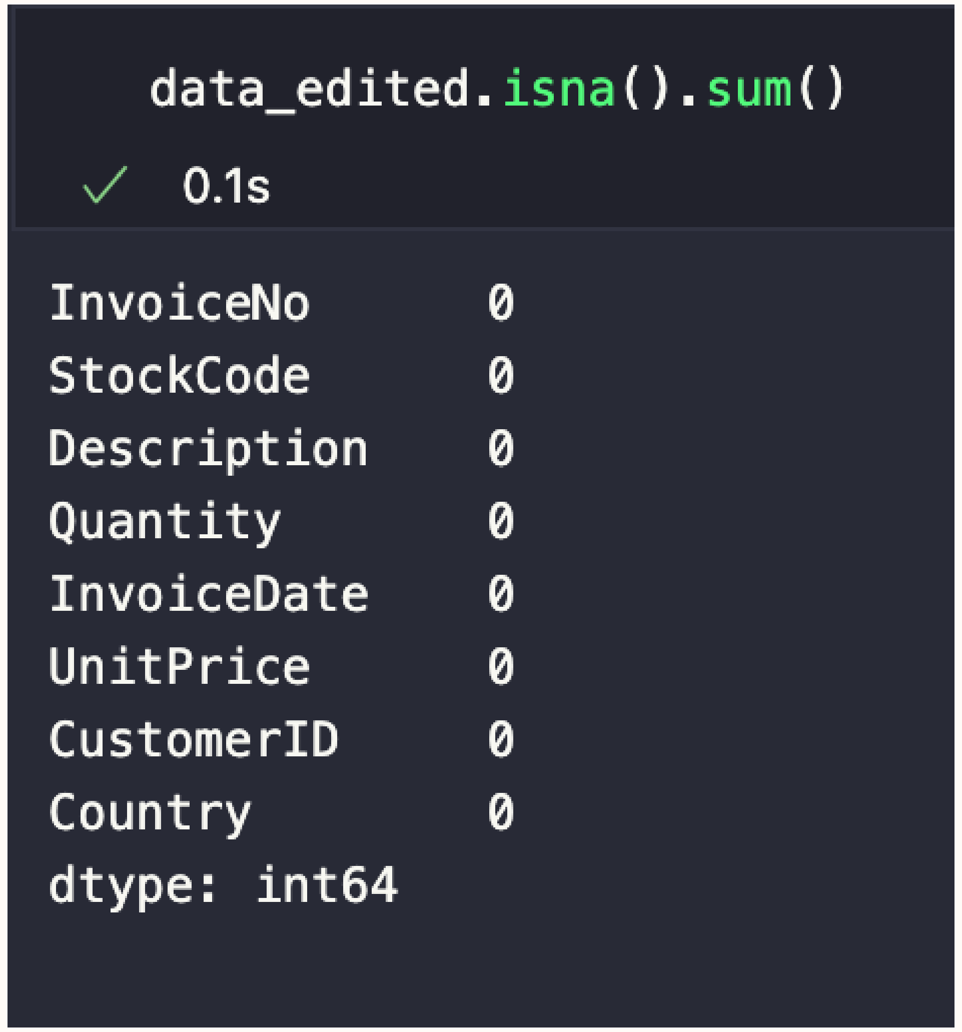
solution?

- illogical values **(remove)**
  - quantity < 0
  - unitprice < 0
  - excessive quantity **(remove if greater than a certain number)**
- missing values **(remove)**

after cleaning



	Quantity	InvoiceDate	UnitPrice	CustomerID
count	392711.000000	392711	392711.000000	392711.000000
mean	12.593902	2011-07-10 19:18:08.374707712	3.125715	15287.757720
min	1.000000	2010-12-01 08:26:00	0.000000	12347.000000
25%	2.000000	2011-04-07 11:12:00	1.250000	13955.000000
50%	6.000000	2011-07-31 12:02:00	1.950000	15150.000000
75%	12.000000	2011-10-20 12:53:00	3.750000	16791.000000
max	2000.000000	2011-12-09 12:50:00	8142.750000	18287.000000
std	38.037783	NaN	22.241313	1713.569468



```
data_edited.isna().sum()
```

✓ 0.1s

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

8 columns and **392,711 rows** (**14,9198 rows** difference)

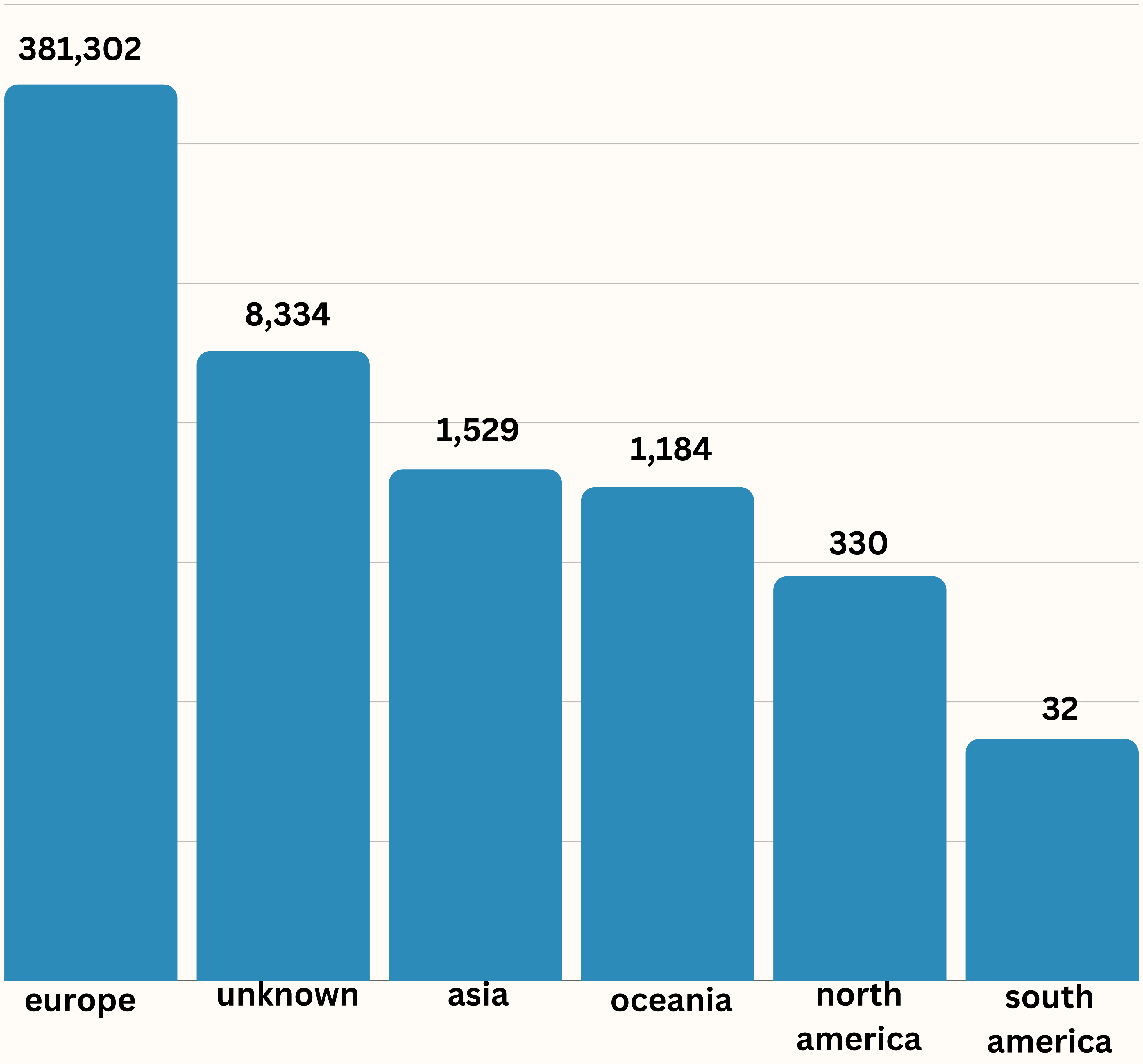
data insight

retailer based on continent

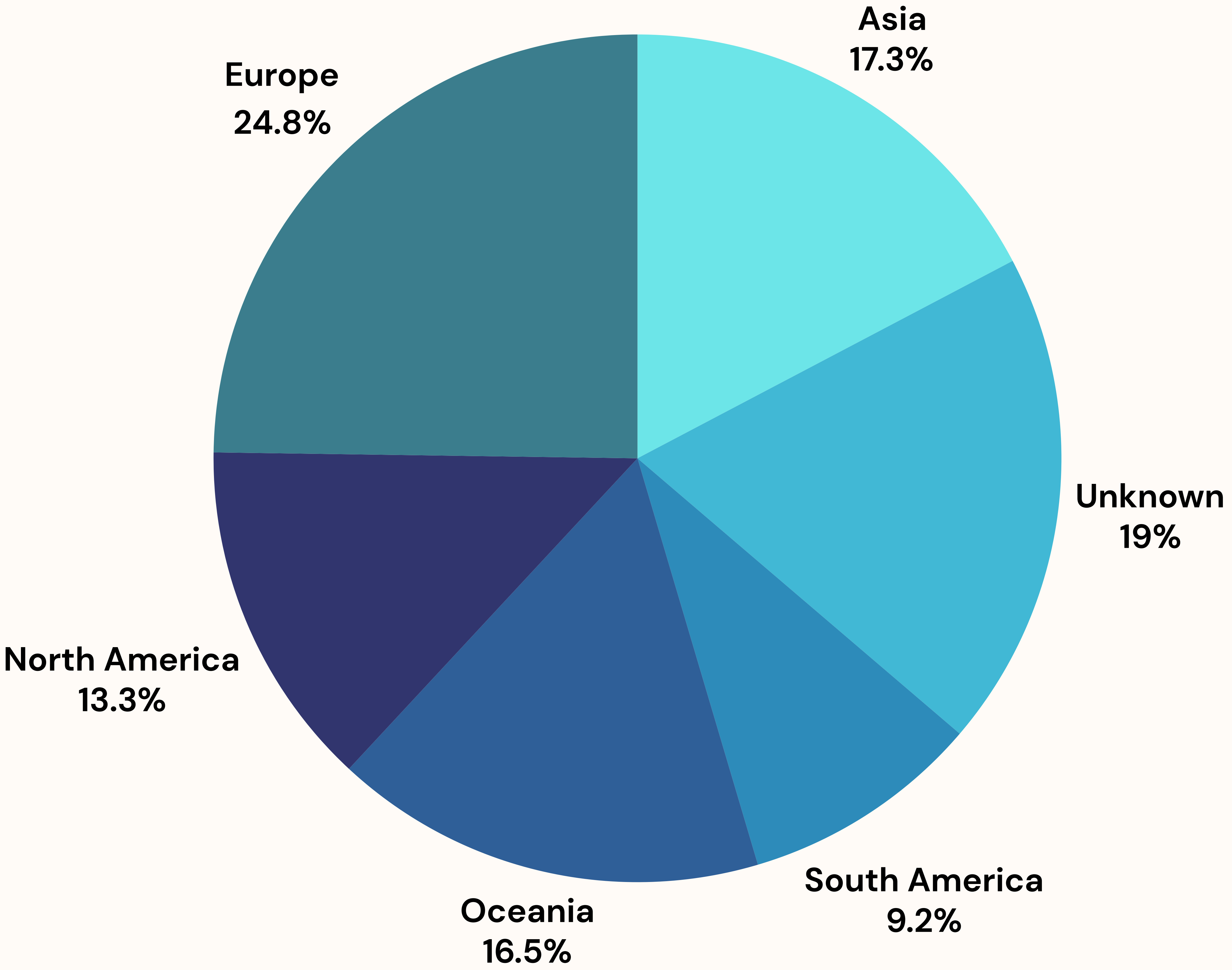
because it is UK-based, which within europe, it makes **europe** the highest retailer based on the continent with **381,302**.

followed by :
asia with **1,529**
oceania with **1,184**
north america with **330**
south america with **32**
and
8,334 unknown continent

Online Retail Analysis

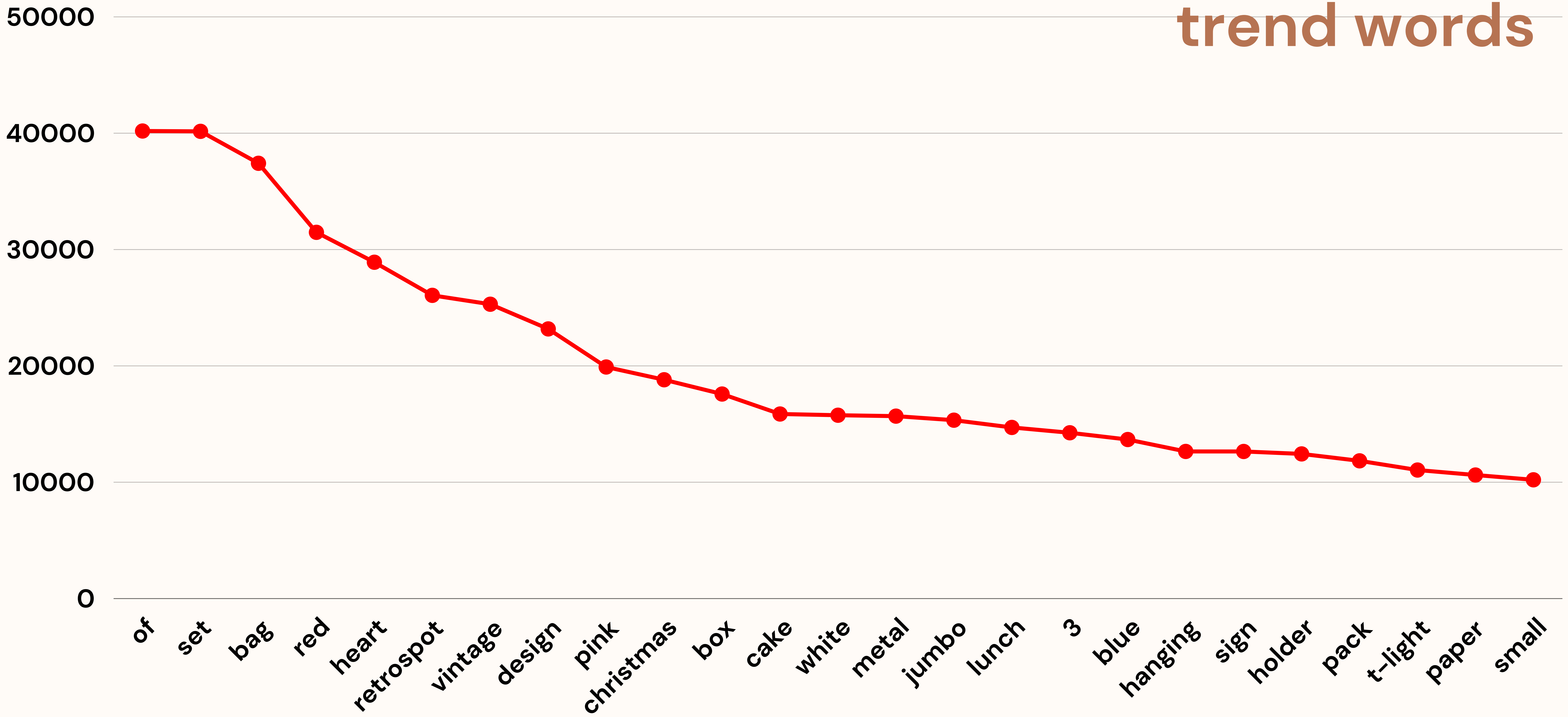


expense for each continent



europe with the highest expense, **24.8%**

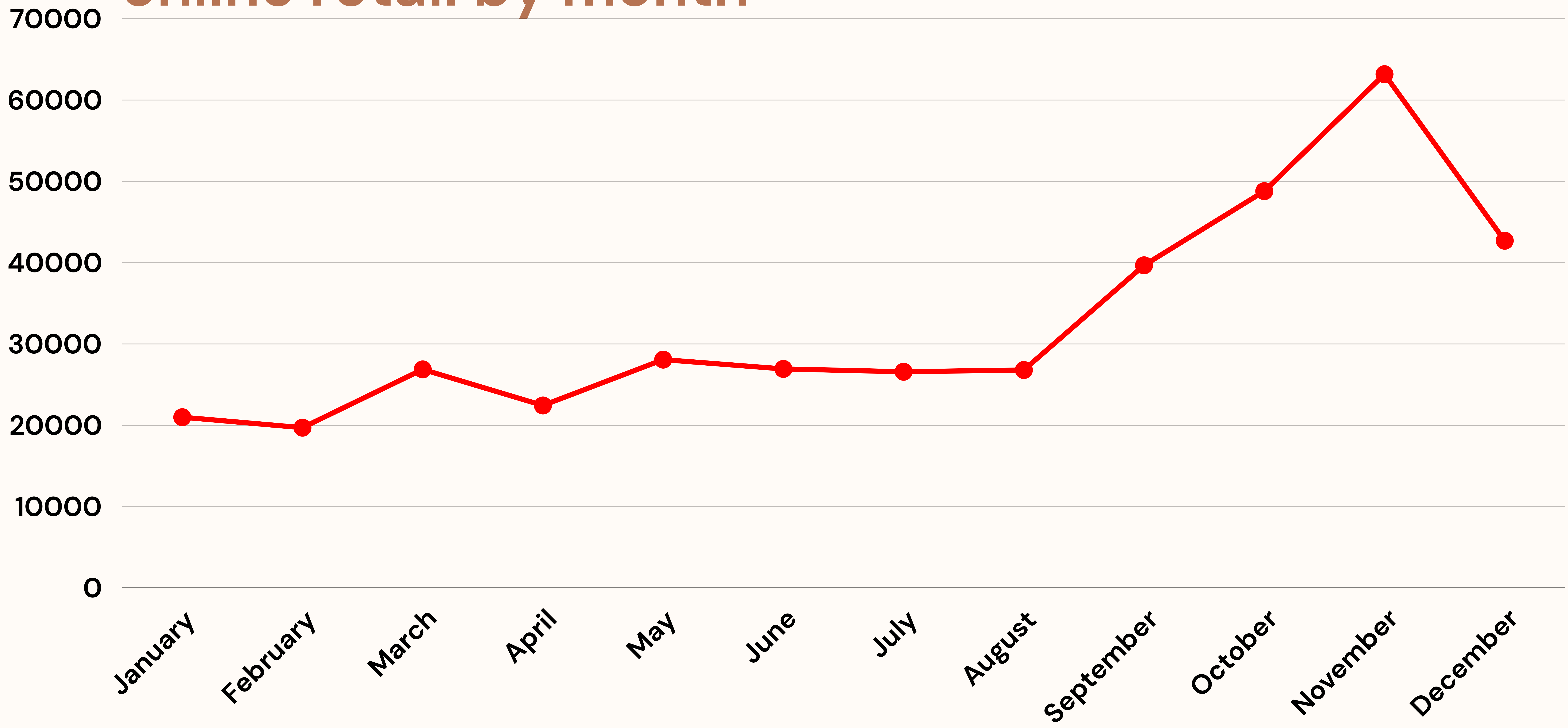
trend words



25 of the **frequently repeated words** in retail

Online Retail Analysis

online retail by month



Online Retail Analysis

4 months at the end of the year with **peak retail**