Mentorships

1. Introduction

This analysis examines retail transaction data to identify key performance patterns across regions, product categories, and sales channels. After comprehensive data cleaning and exploratory analysis, we address four core business questions regarding regional performance, product profitability, and sales channel distribution. Key findings reveal significant opportunities for margin optimization and channel strategy refinement.

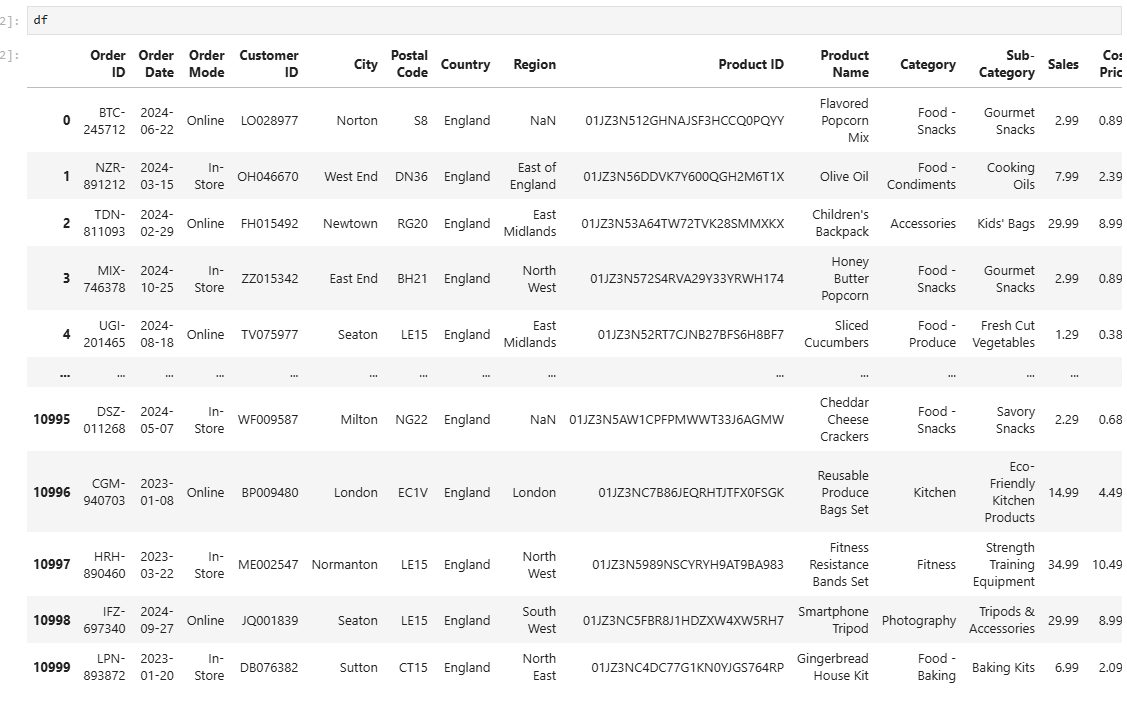
### Data Preprocessing

First of all, I import the pandas library and read the given dataset into pandas data frame.

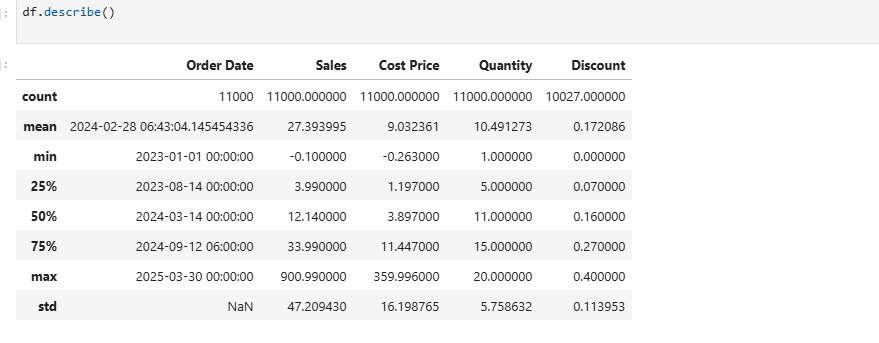


Than I do exploratory analysis on the data.

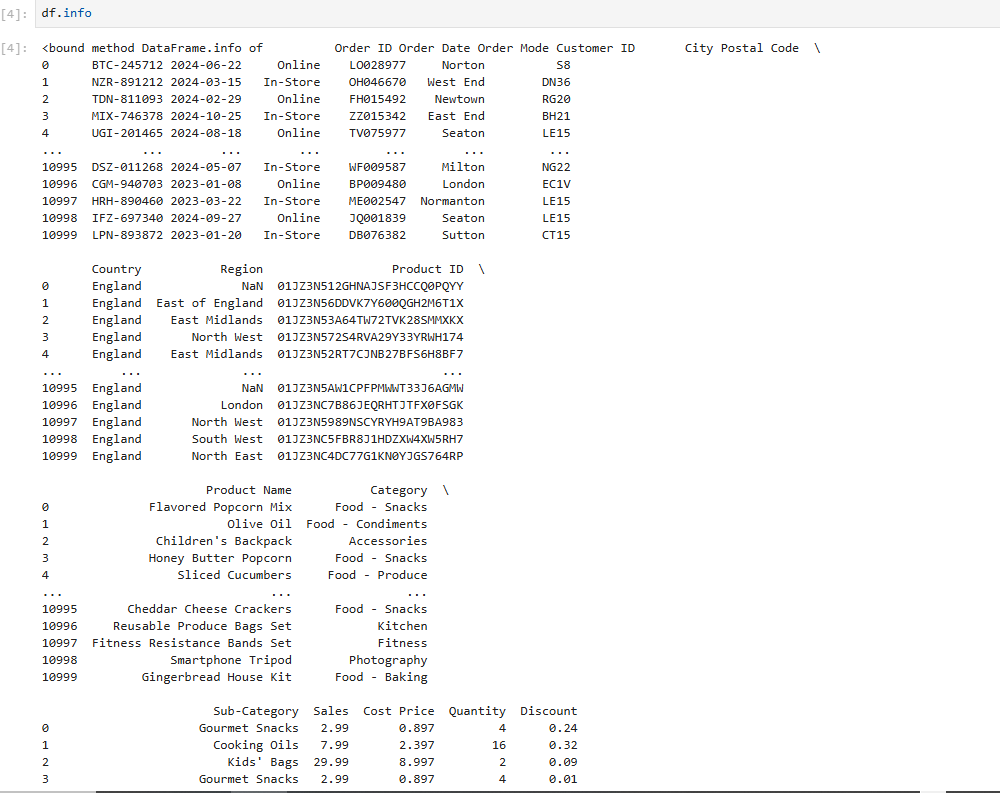
First I just have a look on the data as



Than I use the df.describe(), to see the mean median and mode of dataset.

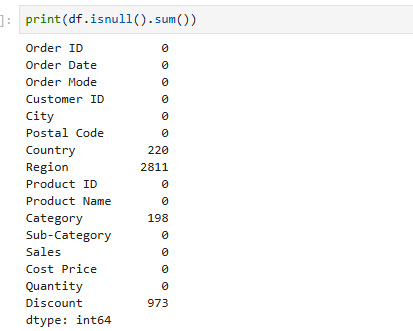


For taking a further look into the data, I used df.info, so that I can have more information for the data set.



From above now we can see the data in more detail than what each column is about.

Than I look into the missing values in the dataset.



#### Initial Data Assessment

Identified missing values across four columns:

Country (4.7% missing)

Region (4.7% missing)

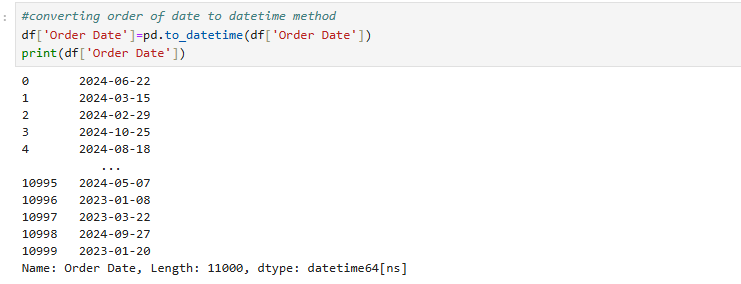
Category (1.2% missing)

Discount (0.6% missing)

#### Data Cleaning Pipeline

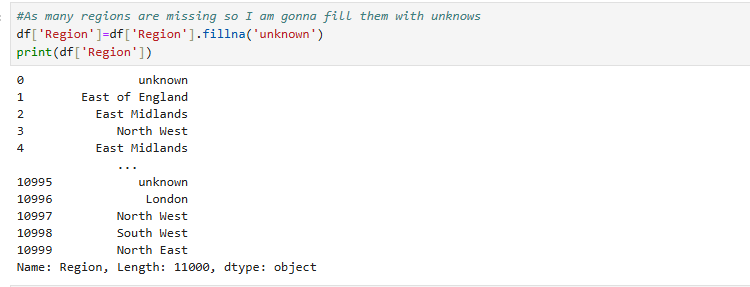
****Date Conversion****

First of all, I convert the dates to datetime format



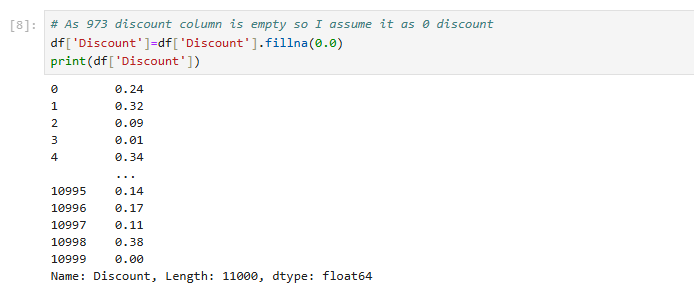
****Region Imputation****

Than we can see that from the data set, the regions are missing and we are given country name, so in a country there are a lot of regions, so we cannot judge regions just from country name, that why I am replacing missing regions with unknows.



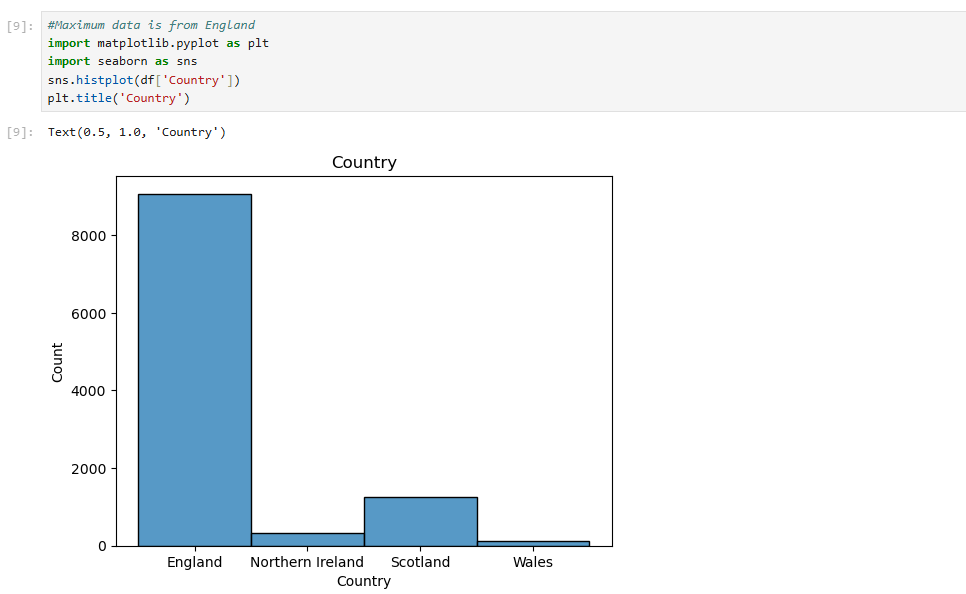
****Discount Handling****

Now the discount columns have missing values, so I assume that there was 0 discount and I replace the missing values of discount column with zero.



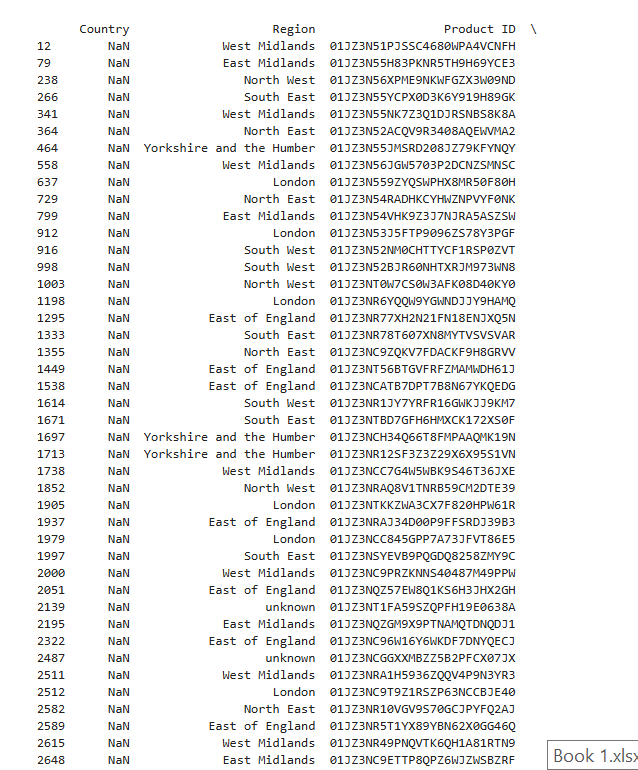
****Country Imputation****

Now the country column have missing values. So for this purpose I checked the distribution of countries as



From above we can see that the maximum count is of England.

Than I have a look on missing country names where region and postcode is not missing.



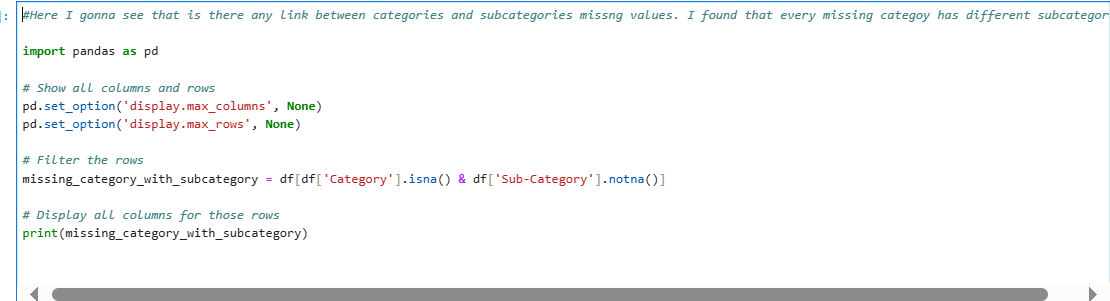
So by having a look on the data, I came to know that all the missing countries are England. By looking at the region name and poscode.

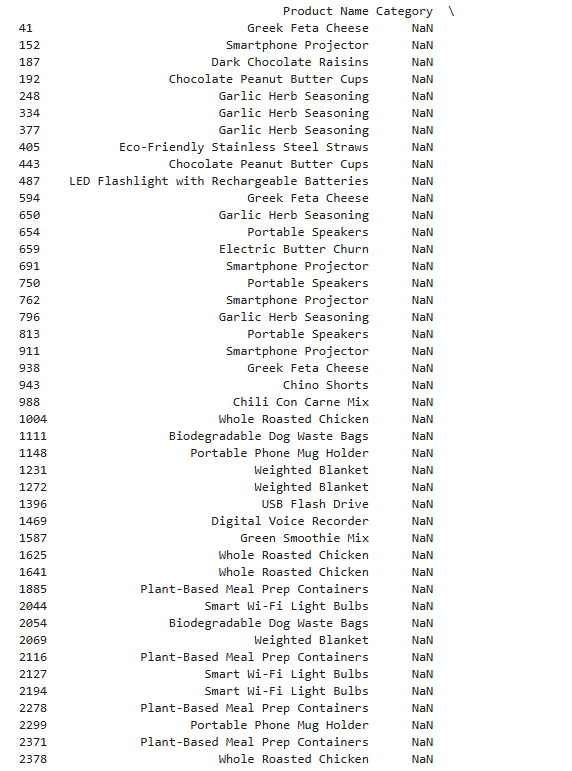
So I decided to impute the missing country column with the England.



****Category Imputation****

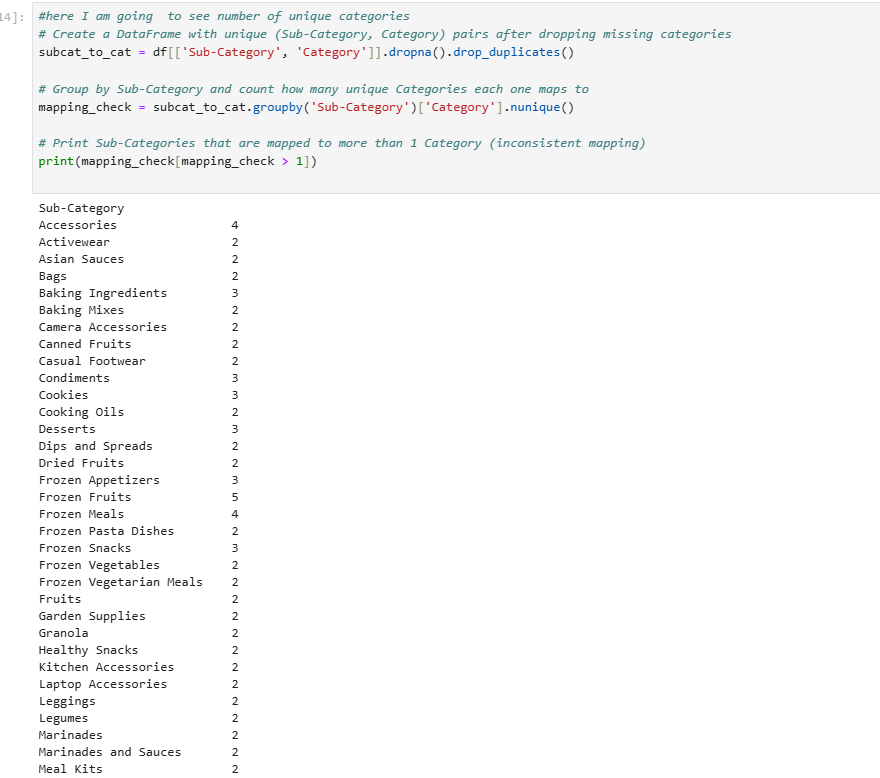
Now there are missing values in category column. So for this purpose first of all I checked is there any link between sub-categories which are not null and categories which are null as





From above I came to know that there is no specific link between categories and sub-categories and every missing categoy has different subcategories.

Than I had a look on unique categories as



From above we can see that every sub-category belongs to different categories.

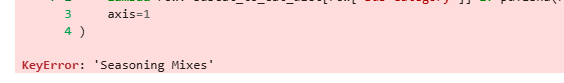
Than I create distionary that maps each su-category to its corresponding category



Than I f**ills in missing values** in the Category column **based on** Sub-Category using the dictionary from the previous step.

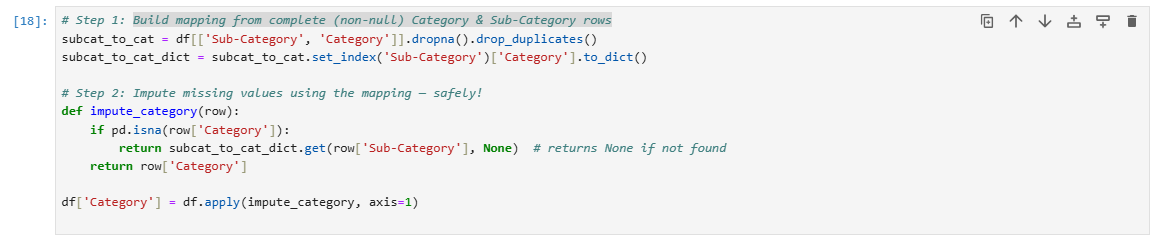


So I get the error as

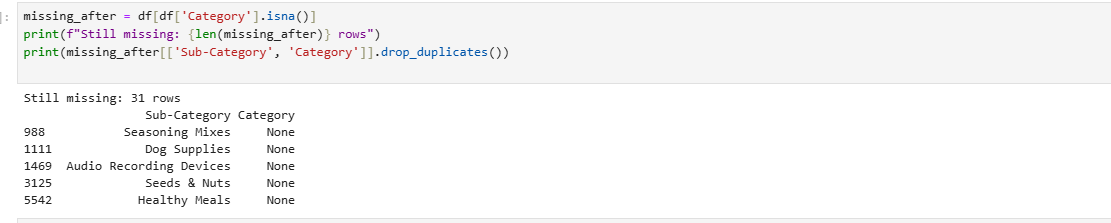


Season Mixes are still missing that it stills remain unattended to any sub-category.

Than I Build a mapping from complete (non-null) Category & Sub-Category rows



Than I look for missing categories still.



I came to know that these are still None categories.

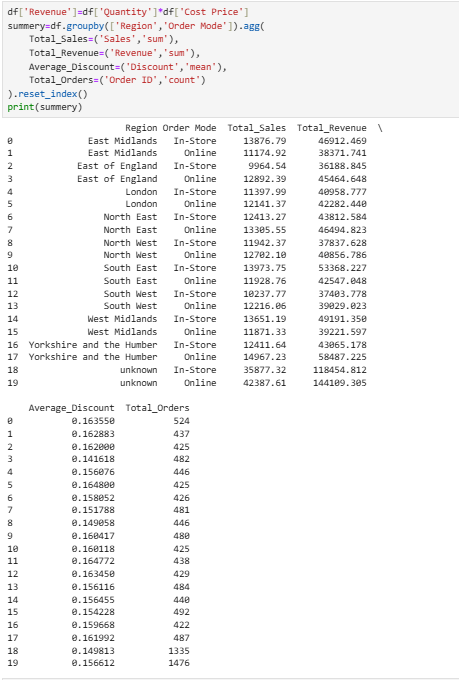
Than I look into the data set manually and find the category name using their sub-categories using my own intutions and fill them manually as

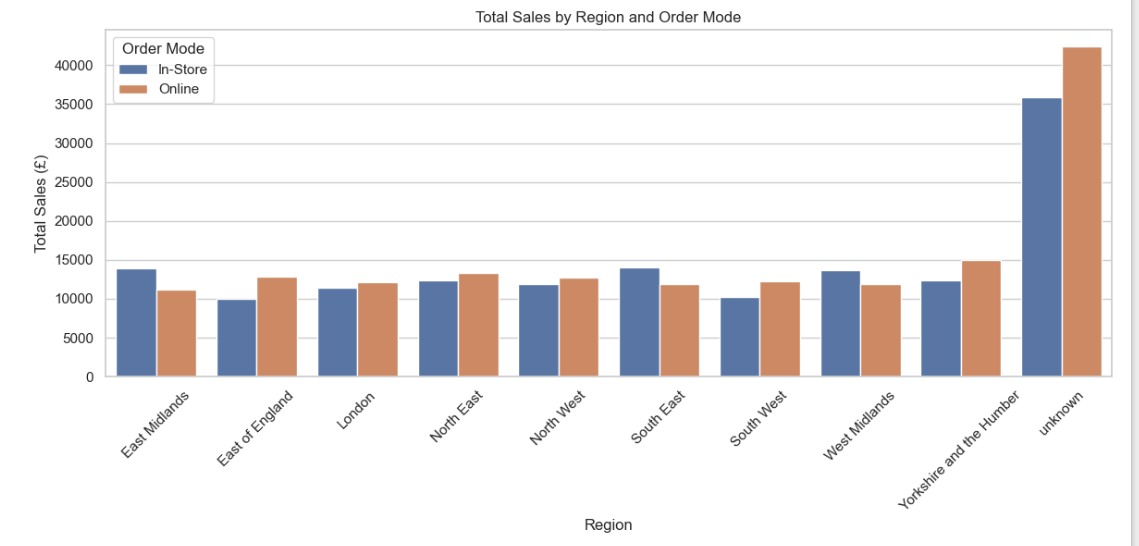


Now there is no missing values in categories.

### Business Question Analysis

Q1:Summarize total sales, revenue, and discount rates by region and segment





#### ****Top Performing Region****

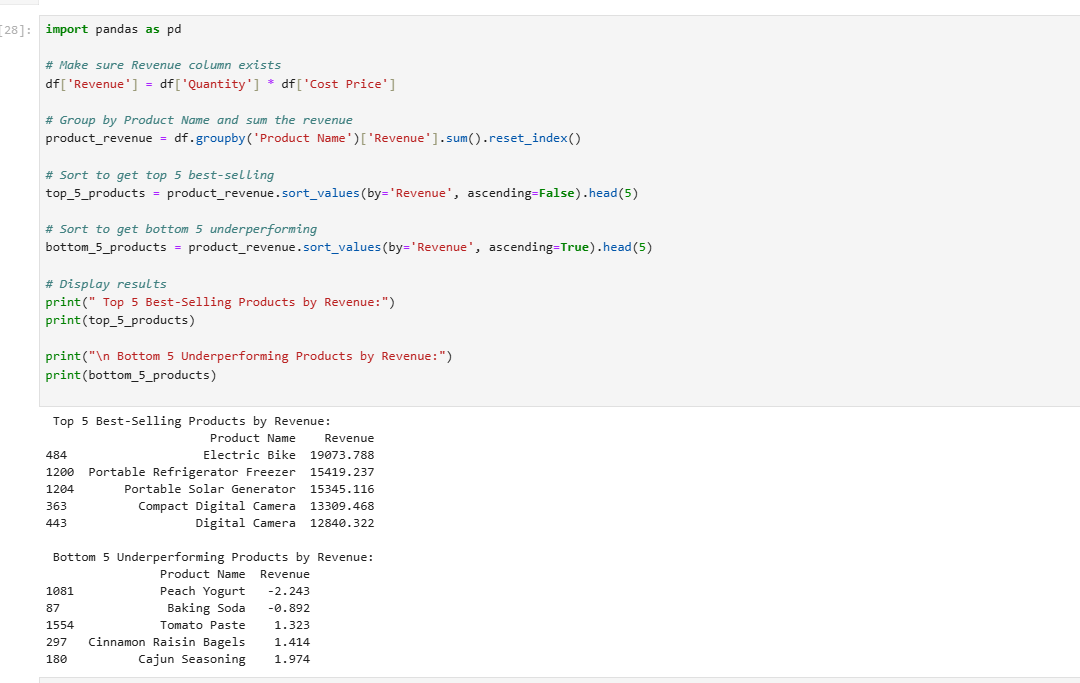
**By Sales & Revenue:**  
**Yorkshire and the Humber (Online)** is the **top-performing known region**, with:

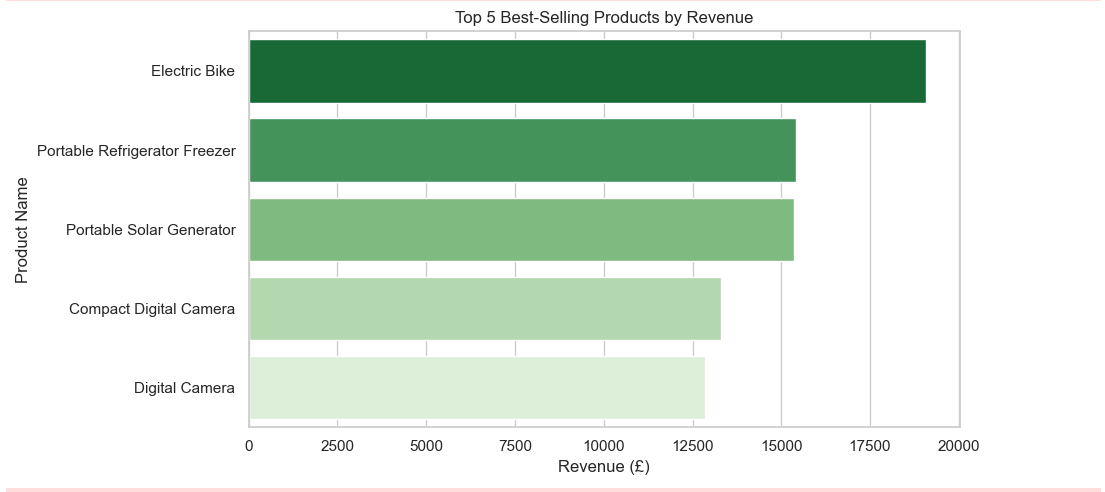
**Total Sales:** £14,967.23

**Revenue:** £58,487.23

**Orders:** 487

Q2: Identify top 5 best-selling products and underperforming products by revenue

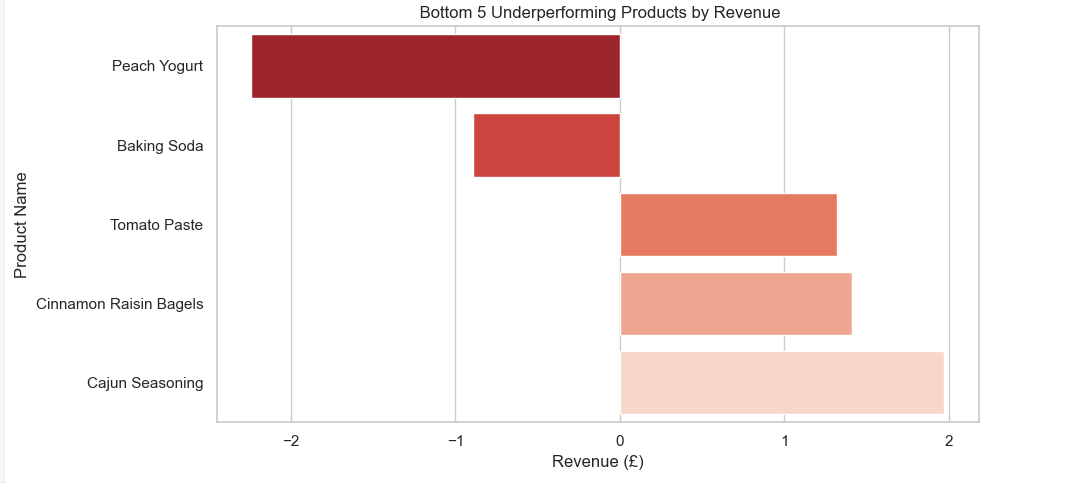




· These products are **high-value items**, primarily **electronics or appliances**.

· Strong indicators of **consumer preference for big-ticket items** in online/in-store segments.

· Consider keeping **ample inventory**, optimizing **delivery**, and possibly bundling accessories.



· These are **low-cost, perishable food items**, prone to spoilage or low margins.

· **Negative or very low revenue** suggests:

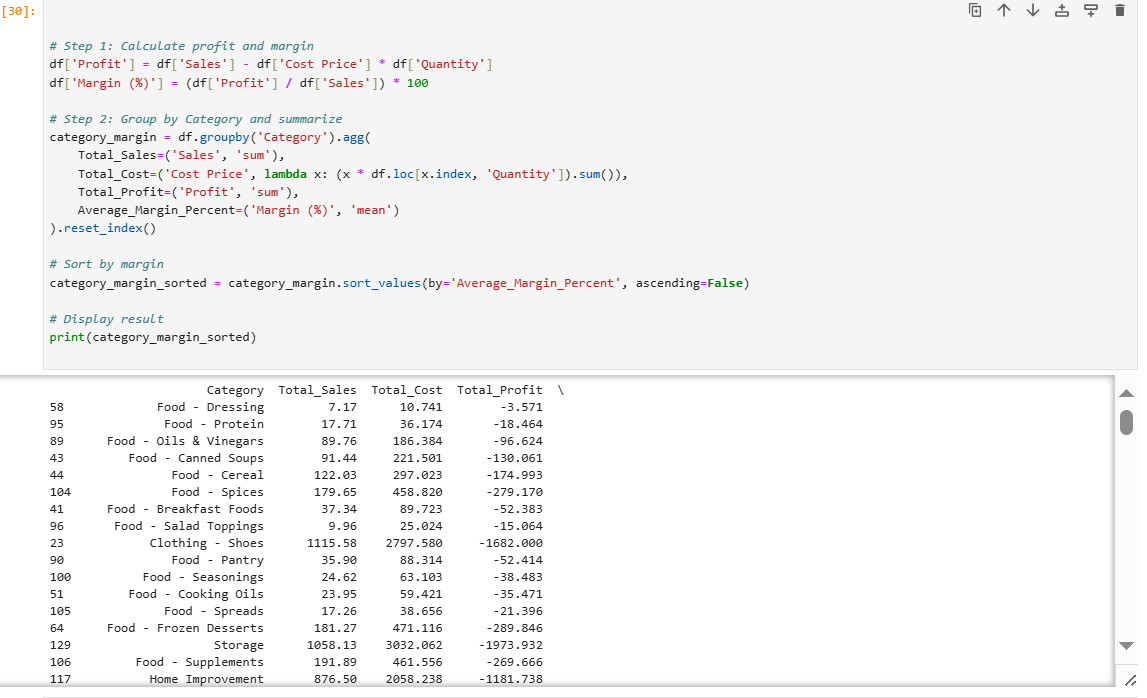
Over-discounting

High returns or spoilage

Low pricing with minimal profit margins

· May warrant **review of pricing strategy**, **promotion control**, or even **stock removal**.

Q3: Provide insights into product categories with highest margins.



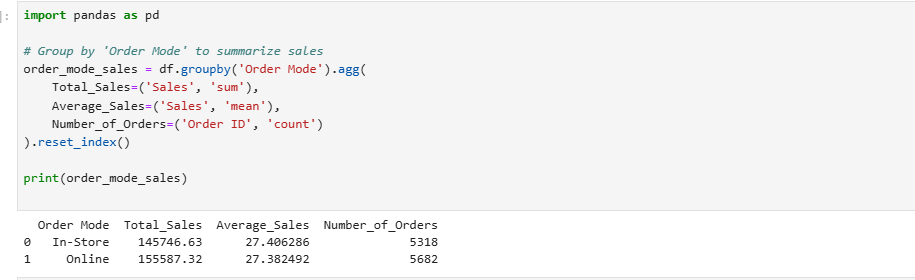
| **Rank** | **Category** | **Total Sales** | **Total Profit** | **Avg. Margin %** |
| --- | --- | --- | --- | --- |
| 1 | **Food - Dressing** | 7.17 | -3.57 | **-49.80%** |
| 2 | **Food - Protein** | 17.71 | -18.46 | **-92.06%** |
| 3 | **Food - Oils & Vinegars** | 89.76 | -96.62 | **-124.69%** |
| 4 | **Food - Canned Soups** | 91.44 | -130.06 | **-135.17%** |
| 5 | **Food - Cereal** | 122.03 | -174.99 | **-139.51%** |

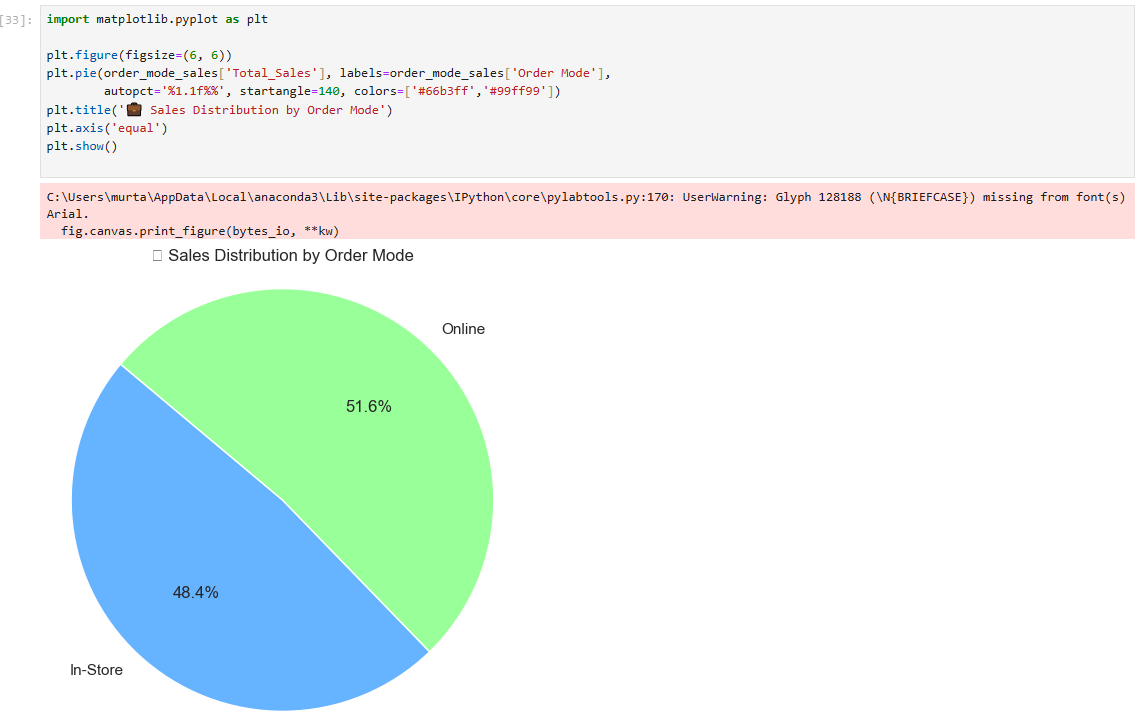
· **No category is profitable** in data — every category has negative profit and margin may be this is due to high returns or more discounts.

· Categories with the **least losses per dollar sold** tend to be food items with **lower cost structures**, such as dressings and canned goods may be this is due to high returns or more discounts.

· **Food - Dressing** stands out with the **smallest margin loss (-49.8%)**, indicating it’s the **least unprofitable** category and may be a candidate for optimization to reach profitability may be this is due to high returns or more discounts.

Q4: Analyze sales distribution across Order Mode (Online vs In-Store)





Sales distribution in Online order mode is more than those in In-store.