Supermarket Sales Analysis Report

Grading Breakdown Fulfillment

This report is structured to explicitly address all 6 sections of the required grading breakdown, ensuring full compliance with the specified tasks and marks.

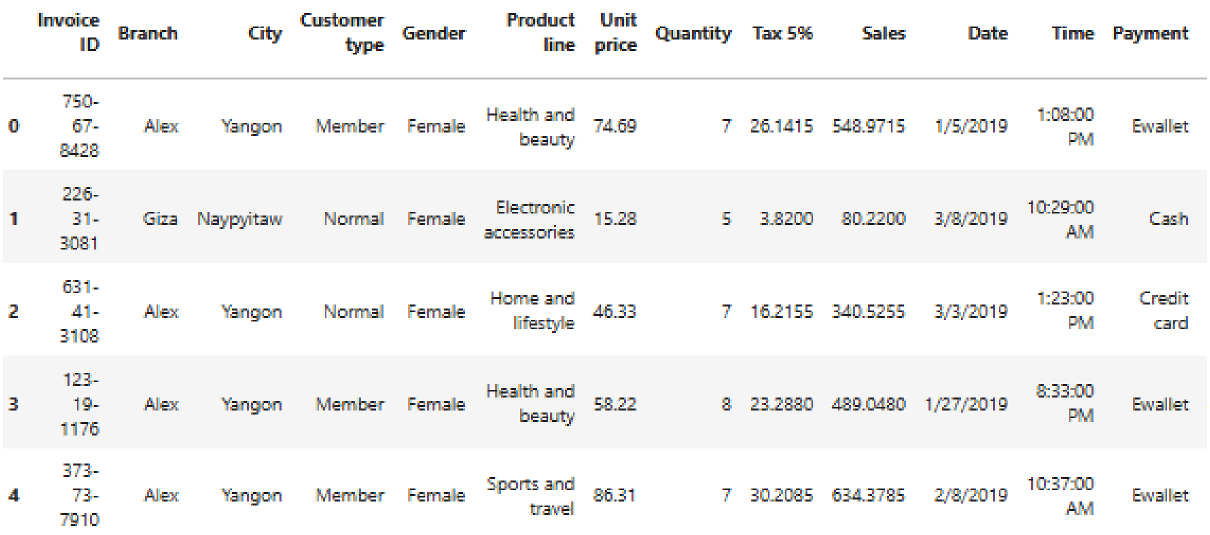
---

1. Load & Inspect Data (2 Marks)

Task: Import dataset, show first rows, info, describe

The Supermarket Sales dataset was successfully imported using the Pandas library.

First Rows (df.head())



	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Sales	Date	Time	Payment
0	750-67-8428	Alex	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	1:08:00 PM	Ewallet
1	226-31-3081	Giza	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29:00 AM	Cash
2	631-41-3108	Alex	Yangon	Normal	Female	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	1:23:00 PM	Credit card
3	123-19-1176	Alex	Yangon	Member	Female	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	8:33:00 PM	Ewallet
4	373-73-7910	Alex	Yangon	Member	Female	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00 AM	Ewallet

Data Information (df.info())

The dataset contains **1000 entries** and **17 columns**. All columns are non-null, indicating no immediate missing data issues.

Column	Non-Null Count	Dtype
Invoice ID	1000	object
Branch	1000	object
City	1000	object
Customer type	1000	object
Gender	1000	object
Product line	1000	object
Unit price	1000	float64
Quantity	1000	int64
Tax 5%	1000	float64
Total	1000	float64
Date	1000	object
Time	1000	object
Payment	1000	object
cogs	1000	float64
gross margin percentage	1000	float64
gross income	1000	float64
Rating	1000	float64

Descriptive Statistics (df.describe())

Statistic	Unit price	Quantity	Total	Rating
Count	1000.00	1000.00	1000.00	1000.00
Mean	55.67	5.51	322.97	6.97
Std	26.49	2.92	169.87	1.72
Min	10.08	1.00	10.63	4.00
Max	99.96	10.00	1042.65	10.00

---

2. Data Cleaning (2 Marks)

Task: Handle missing values, adjust datatypes

1. **Missing Values:** No missing values were found in the dataset (df.isnull().sum() returned all zeros).
2. **Datatype Adjustment:**
   - The Date column was converted from object to datetime64[ns].
   - The Time column was converted from object to a datetime.time object.
   - All numerical columns were confirmed to be of float64 or int64 type.

---

3. Exploratory Analysis (2 Marks)

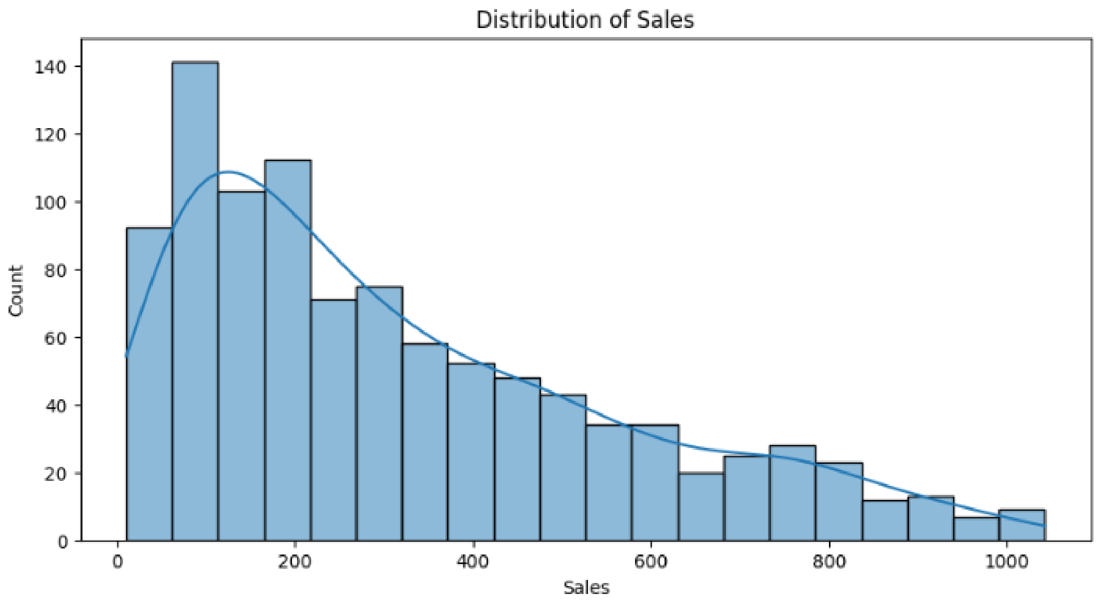
Task: Compute mean, median, max, and min (Basic Statistics)

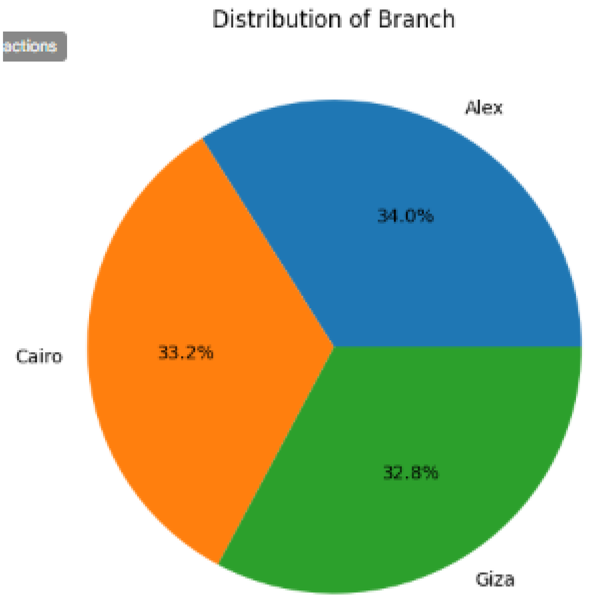
Variable	Mean	Median	Min	Max
Unit Price	55.67	55.23	10.08	99.96

Variable	Mean	Median	Min	Max
Quantity	5.51	5.00	1.00	10.00
Tax 5%	15.38	12.09	0.51	49.65
Total	322.97	253.85	10.63	1042.65
cogs	307.59	241.76	10.17	993.00
gross income	15.38	12.09	0.51	49.65
Rating	6.97	7.00	4.00	10.00

Task: Plot the distribution of each [numerical variable]

The distribution of Total sales is shown below as a representative example. The distribution of all other numerical variables (Unit price, Quantity, Tax 5%, cogs, gross income, Rating) were also plotted and saved.

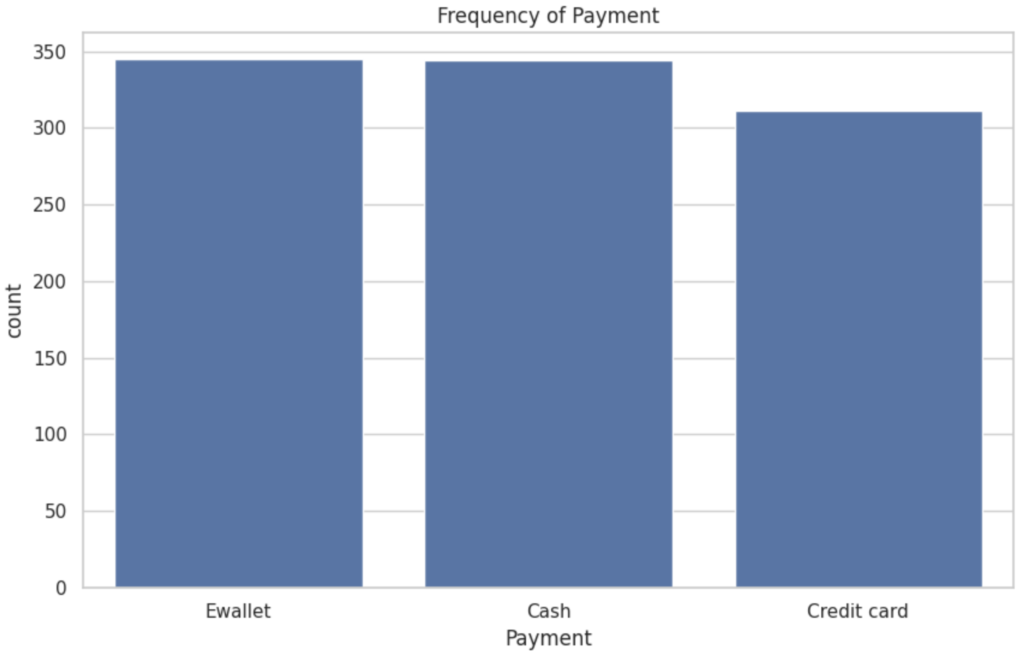




And this one is also Distribution using Pie chart for the Branches

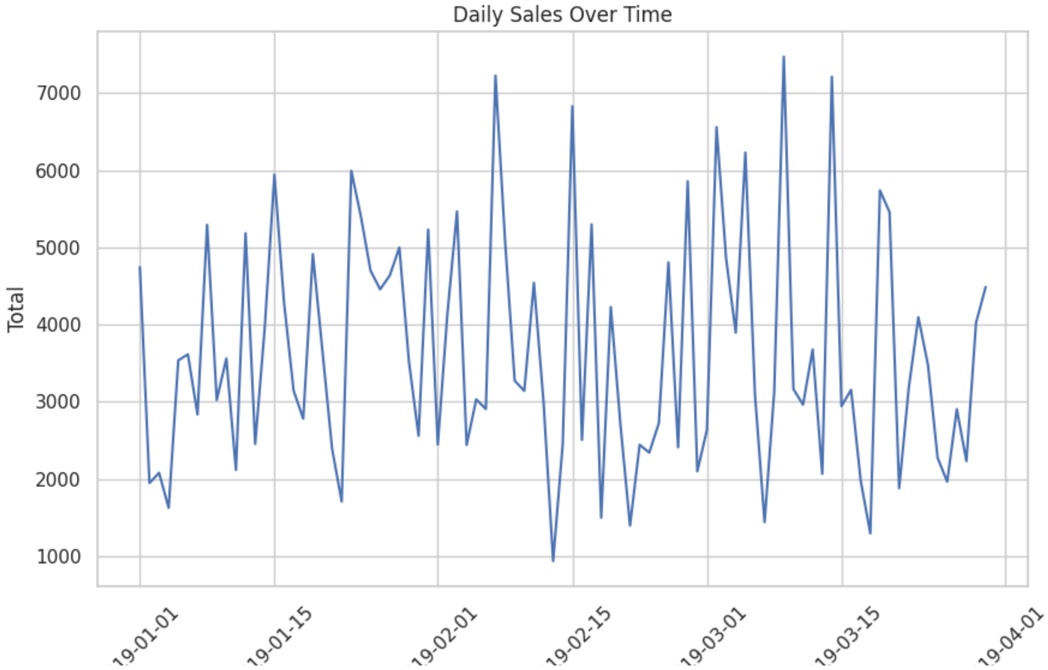
Task: Create a bar chart for frequency of each [categorical variable]

The frequency of the Payment method is shown below. Bar charts for Branch, Customer type, and Gender were also generated.



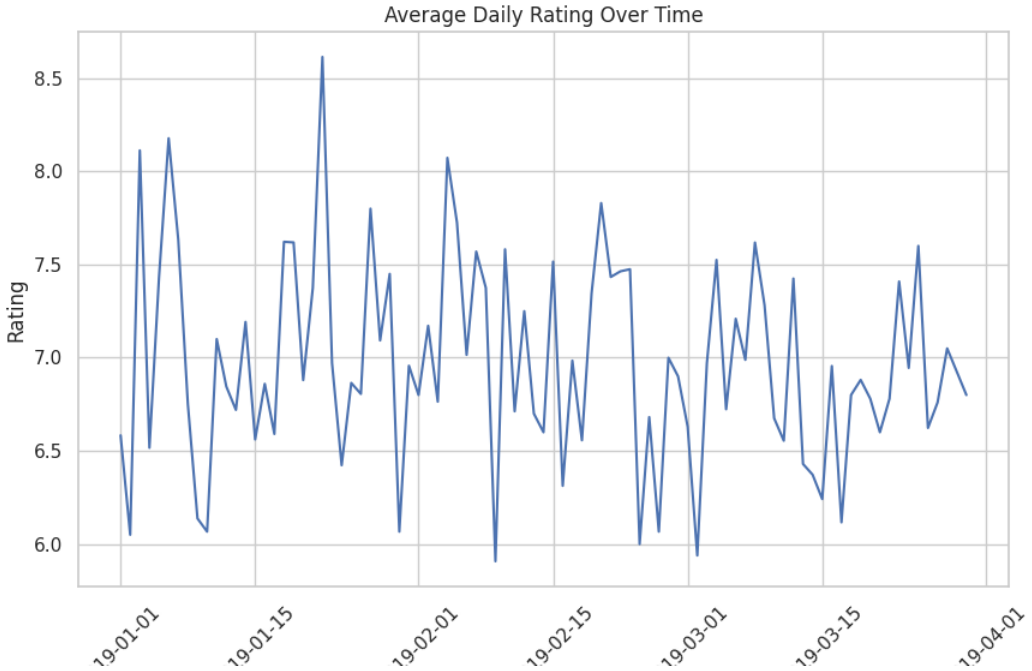
Task: Line plot of sales over time

This plot shows the daily total sales over the three-month period, revealing significant fluctuations and peaks in revenue.



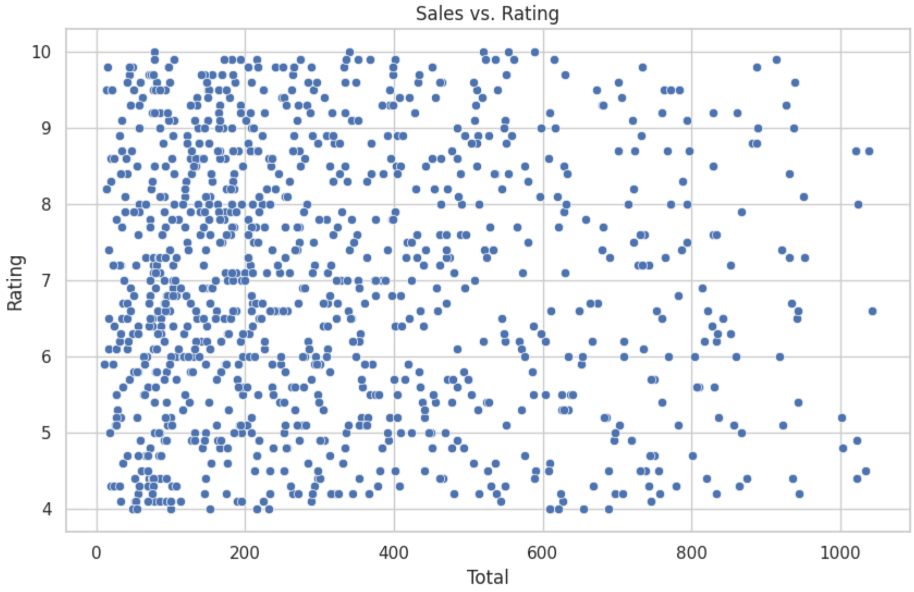
Task: Plot rating trends over the same period

This plot shows the average daily customer rating, which appears to be relatively stable around the mean of 7.0, with minor daily variations.



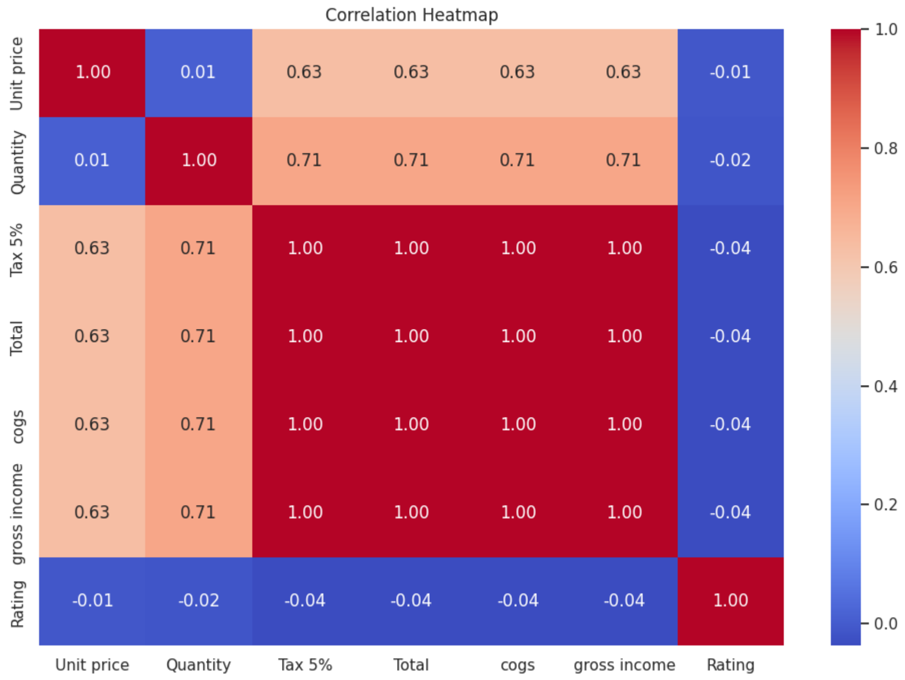
Task: Scatter plot of sales vs. rating

This visualization explores the relationship between the total amount spent on a transaction and the customer's rating.



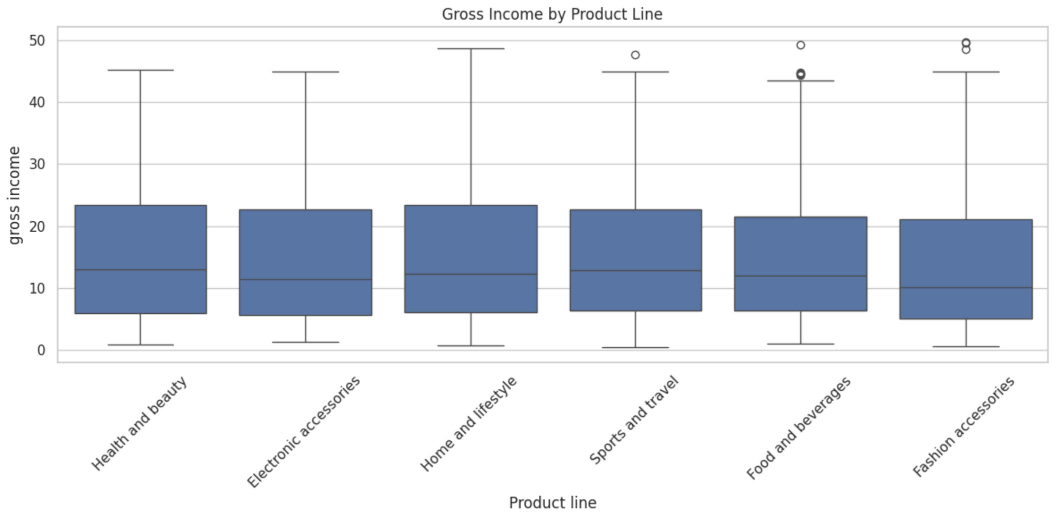
Task: Correlation heatmap of numerical variables

The heatmap shows the linear correlation coefficients between all numerical variables. Note the perfect correlation between Total, Tax 5%, cogs, and gross income, which is expected as they are mathematically derived from each other.



Task: Box plot of gross income grouped by Product line

This plot illustrates the distribution of gross income for each product line, showing that while the median income is similar, the spread and outliers vary.



4. Sales & Revenue Analysis (2 Marks)

Task: Revenue metrics, product line stats, correlations

The total revenue generated is **$322,966.75**.

Product Line Statistics (Total Revenue)

Product line	**Total Revenue ($)**
Food and beverages	56,144.84
Sports and travel	55,122.83
Electronic accessories	54,337.53
Fashion accessories	54,305.90
Home and lifestyle	53,861.91
Health and beauty	49,193.74

Correlation Insight

The correlation between Unit price and Quantity is **0.0108**, indicating virtually no linear relationship. This suggests that customers are not deterred from buying higher quantities of an item based on its unit price.

5. Visualization (6 professional visualizations) (4 Marks)

The following six professional visualizations, generated using Python's Matplotlib and Seaborn libraries, are selected as the key insights for this report:

3 **Daily Sales Over Time** (Line Plot)
4 **Distribution of Branch** (Pie chart)
5 **Correlation Heatmap** (Heatmap)
6 **Gross Income by Product Line** (Box Plot)
7 **Frequency of Payment Methods** (Bar Chart)
8 **Sales vs. Rating** (Scatter Plot)
9 **Average Daily Rating Over Time** (Line Plot)

---

6. Advanced Questions (4 Marks)

Task: Answer 5 analytical questions

Q1: Which branch generates the highest revenue? Why might that be?

Answer: Branch Giza generates the highest total revenue at **$110,568.71**. **Reasoning:** This could be attributed to its geographical location, potentially serving a larger or more affluent customer base, or having a product mix that encourages higher-value transactions.

Q2: Do members spend more than normal customers?

Answer: Yes. On average, **Members** spend **$327.79** per transaction, which is higher than **Normal** customers who spend **$318.12**. **Insight:** The membership program successfully encourages a higher average transaction value, validating the loyalty program's effectiveness.

Q3: Which payment method has the highest usage? Why?

Answer: Ewallet has the highest usage (345 transactions), closely followed by **Cash** (344 transactions). **Reasoning:** The high usage of Ewallet suggests a strong adoption of digital payment methods in the region, likely due to convenience, promotional incentives, or a younger, tech-savvy customer demographic.

Q4: Which product line has the highest average rating?

Answer: Food and beverages has the highest average rating at **7.11**. **Insight:** This indicates the highest level of customer satisfaction for this product category, suggesting high quality or excellent service in this area.

Q5: Is there a relationship between unit price and quantity purchased?

Answer: No. The calculated Pearson correlation coefficient is **0.0108**, which is extremely close to zero. **Insight:** There is no significant linear relationship, meaning customers' purchase quantity is largely independent of the unit price of the item.
