

# Sales Analysis Report - Online Store Data



Author: Ana Bojescu

This notebook presents a comprehensive analysis of the dataset `online_store_data.csv`. My main objective is to transform raw data into actionable insights that help identify:



Best-selling products,



The most and least successful product categories,



Key factors that drive product performance (such as price, rating, stock availability).

To perform this analysis, I will use:

- pandas for data cleaning and processing,

- matplotlib and seaborn for building insightful visualizations,

- plotly.express for creating interactive and visually appealing plots with minimal code,

- and well-structured Markdown commentary to guide interpretation and support business decision-making.

## Step 1: Importing libraries

```
In [169... import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import plotly.express as px
```

## Load dataset

```
In [170... df = pd.read_csv("PDV5_03-Task_3-online_store_data.csv")
```

## Preview of the Dataset

To get a quick sense of what the data looks like, I display the first few rows using

```
df.head()
```

```
In [171... df.head()
```

Out[171...

	product_name	category	price	quantity_in_stock	quantity_sold	brand	rating
0	iPad mini (2021) (64GB, Blue)	Tablets	499.00	266.0	807.0	Apple	8.75
1	K55 RGB (Red)	Keyboards	49.99	0.0	1117.0	Corsair	8.50
2	Fenix 7 (Purple)	Watches	699.99	0.0	530.0	Garmin	9.20
3	Mate 40 (256GB, Orange)	Smartphones	899.00	182.0	1030.0	Huawei	8.50
4	Rival 600 (Red)	Mice	79.99	35.0	1439.0	SteelSeries	9.25

## Step 2: Exploratory Data Analysis 🔍

I start by inspecting the overall dataset structure and familiarizing myself with the types of data available.

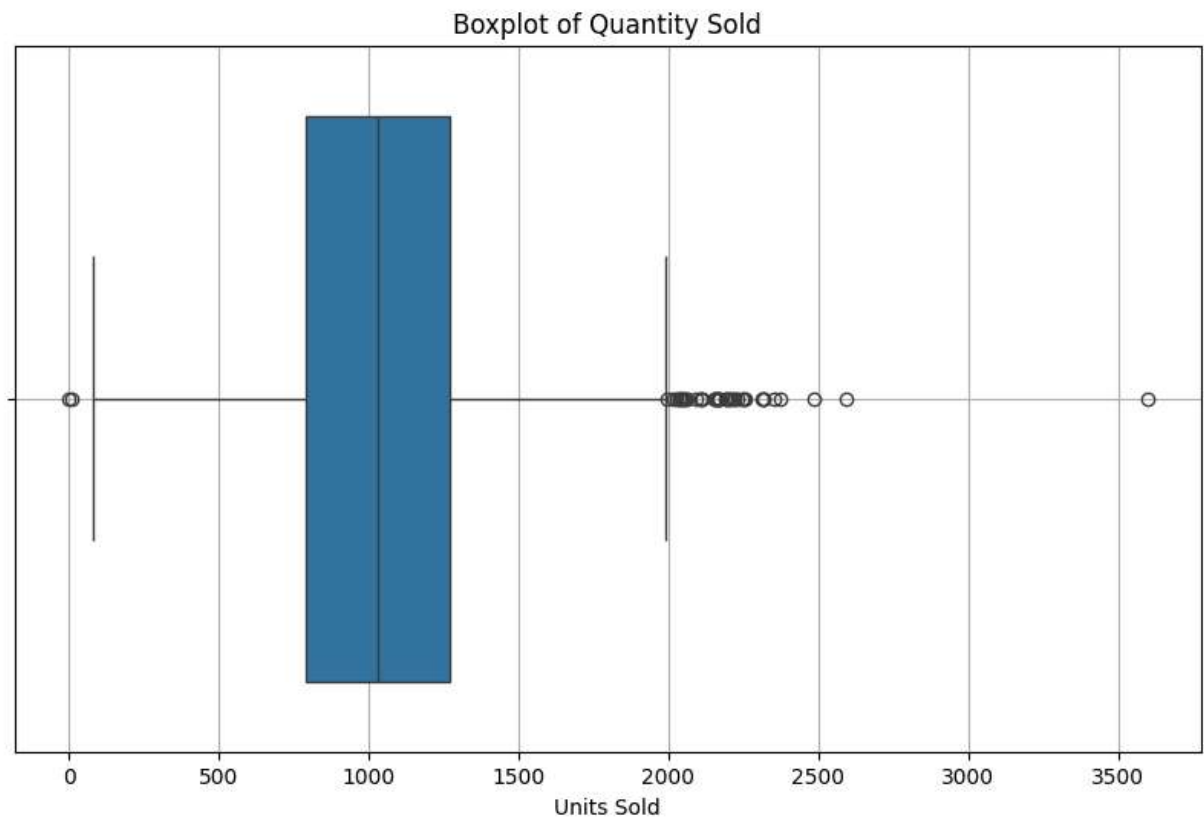
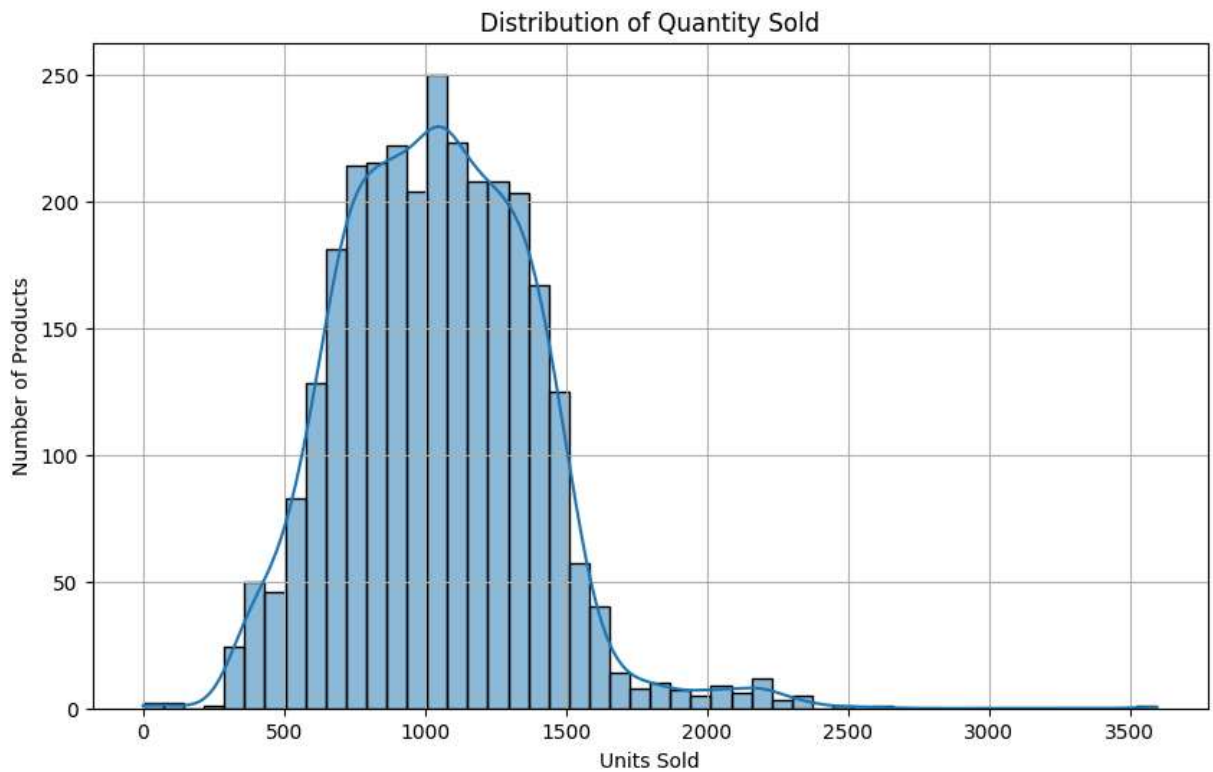
## Step 3: Distribution of Quantity Sold 📦

I examine how units sold are distributed across all products. This helps me identify outliers and top-performing items

In [172...

```
plt.figure(figsize=(10, 6))
sns.histplot(df['quantity_sold'], bins=50, kde=True)
plt.title('Distribution of Quantity Sold')
plt.xlabel('Units Sold')
plt.ylabel('Number of Products')
plt.grid(True)
plt.show()

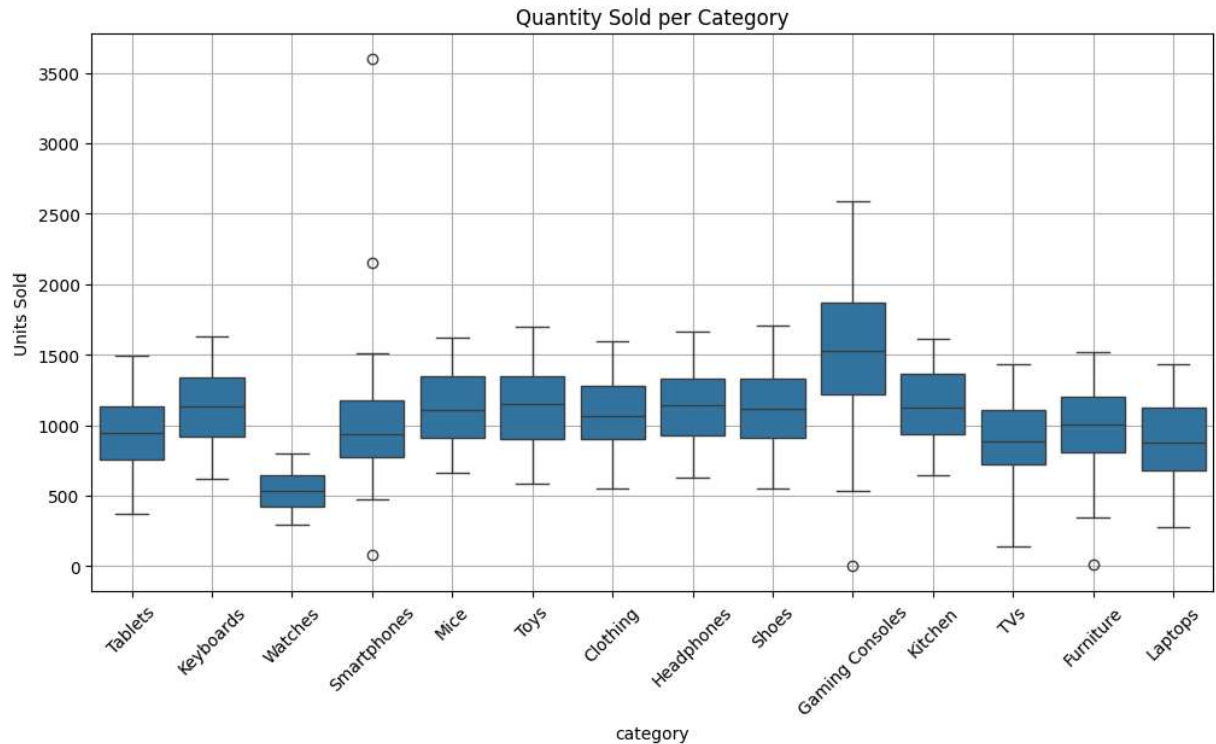
plt.figure(figsize=(10, 6))
sns.boxplot(x=df['quantity_sold'])
plt.title('Boxplot of Quantity Sold')
plt.xlabel('Units Sold')
plt.grid(True)
plt.show()
```



## Step 4: Sales Performance by Category 📁

I compare units sold across categories using boxplots to reveal which categories are performing the best.

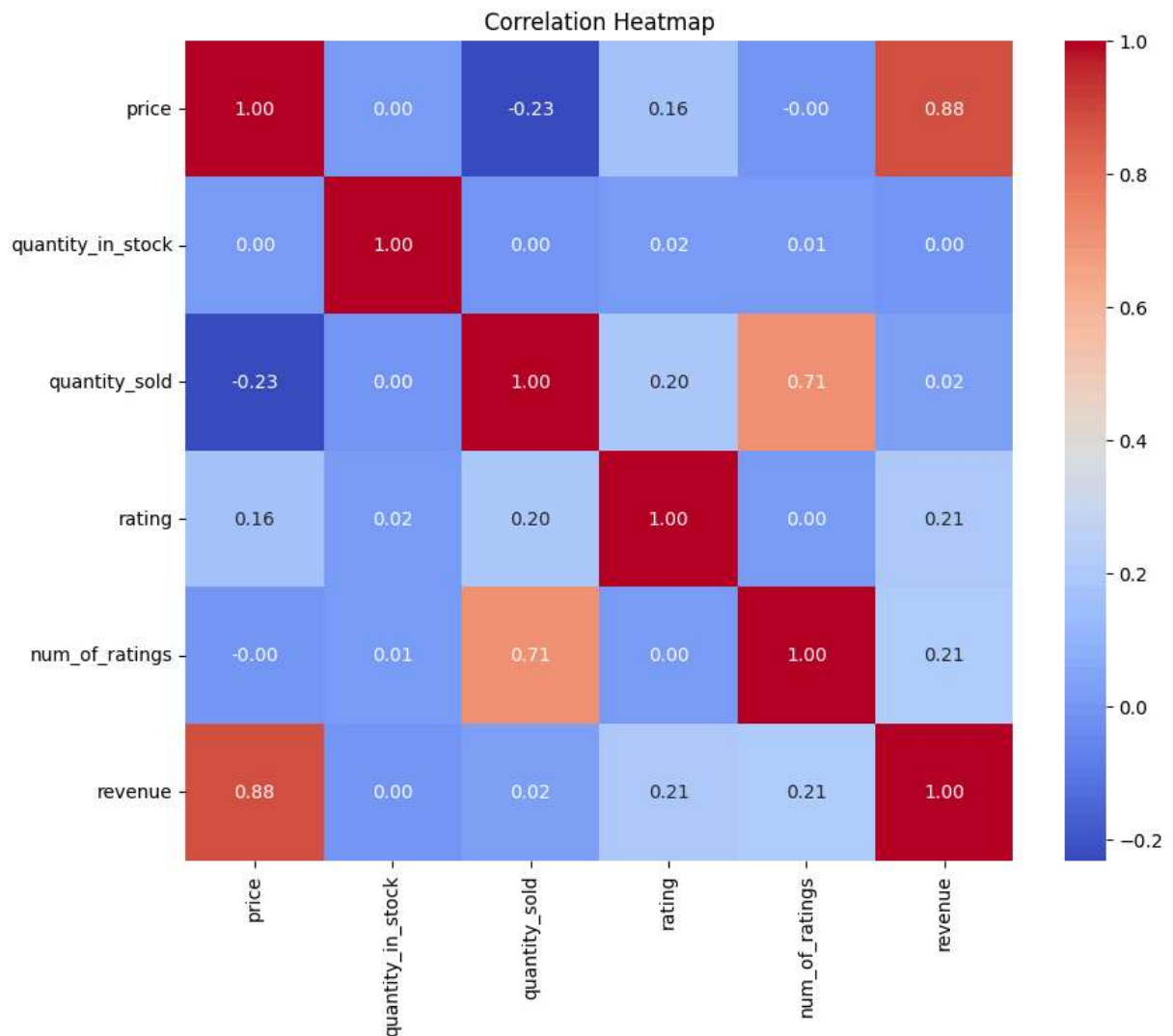
```
In [173... plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='category', y='quantity_sold')
plt.title('Quantity Sold per Category')
plt.xticks(rotation=45)
plt.ylabel('Units Sold')
plt.grid(True)
plt.show()
```



## Step 5: Correlation Analysis

I calculate and visualize the correlations between numeric variables like price, rating, stock, and revenue to better understand which factors are driving sales.

```
In [174... correlation_matrix = df.corr(numeric_only=True)
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



## Step 6: Key Success Factors - Scatter Analysis

I'll visualize how specific features (such as price and rating) relate to quantity sold, and I'll optionally enhance the plots by adding color or size mappings to enrich the visual insight.

```
In [175... plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='price', y='quantity_sold', size='revenue', hue='rating')
plt.title('Price vs Quantity Sold (Size = Revenue, Color = Rating)')
plt.xlabel('Price')
plt.ylabel('Quantity Sold')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.show()
```




## Interactive: Price vs Quantity Sold


In [176...


```
fig = px.scatter(
    df,
    x='price',
    y='quantity_sold',
    size='revenue',
    color='category',
    hover_data=['product_name', 'rating'],
    title='Interactive: Price vs Quantity Sold'
)
fig.show()
```


## Final Conclusions

After conducting this analysis, here's what I've discovered:

 A few products clearly stand out as top sellers — this became evident through both histograms and boxplots.

 Categories like Smartphones and Tablets appear to dominate in terms of sales volume, based on category-level comparisons.

 There is a strong correlation between revenue, units sold, user ratings, and price — suggesting these are major drivers of success.

 In general, high-rated products with moderate pricing tend to perform significantly better — a trend confirmed by the scatter plots.



## Business Implications

-I recommend focusing marketing efforts on categories that consistently perform well and have high user satisfaction.

-Introducing dynamic pricing for high-rated but lower-volume products could help increase their sales potential.

-It's important to actively monitor stock levels for popular items — running out of stock could mean missed revenue opportunities.

\*By going beyond the basic analysis — through annotations, interactive visuals, and extra insights — I can make the report not just informative, but also engaging and actionable. These enhancements help the team spot key patterns faster and make better decisions with confidence.