# Sales & Operations Exploratory Data Analysis Report

---

## SUMMARY STATISTICS

| | order_item_id | price | freight_value | customer_zip_code_prefix | payment_sequential | payment_installments | payment_value |
|---|---|---|---|---|---|---|---|
| count | 117601.000000 | 117601.000000 | 117601.000000 | 117601.000000 | 117601.000000 | 117601.000000 | 117601.000000 |
| mean | 1.195900 | 120.824783 | 20.045990 | 35051.793097 | 1.093528 | 2.939482 | 172.686752 |
| std | 0.697706 | 184.479323 | 15.861315 | 29820.588877 | 0.726692 | 2.774223 | 267.592290 |
| min | 1.000000 | 0.850000 | 0.000000 | 1003.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 1.000000 | 39.900000 | 13.080000 | 11310.000000 | 1.000000 | 1.000000 | 60.870000 |
| 50% | 1.000000 | 74.900000 | 16.290000 | 24315.000000 | 1.000000 | 2.000000 | 108.210000 |
| 75% | 1.000000 | 134.900000 | 21.190000 | 58600.000000 | 1.000000 | 4.000000 | 189.260000 |
| max | 21.000000 | 6735.000000 | 409.680000 | 99990.000000 | 29.000000 | 24.000000 | 13664.080000 |

# 1. Executive Summary

This report presents an exploratory data analysis (EDA) of the sales dataset to understand customer behavior, product performance, revenue trends, delivery efficiency, and seller distribution.
The analysis combines multiple relational tables including orders, payments, customers, products, sellers, and logistics data to derive meaningful business insights.

**Key focus areas:**

- Sales and revenue trends
- Customer and seller behavior
- Product and category performance
- Delivery timelines and logistics efficiency

---

# 2. Dataset Overview

The analysis is based on the following tables:

| Table Name | Description |
|---|---|
| `orders` | Order-level information including timestamps and status |
| `order_items` | Product-level details per order |

| Table Name | Description |
|---|---|
| payments | Payment types and payment values |
| products | Product attributes such as category, size, and weight |
| customers | Customer location and identifiers |
| sellers | Seller location and identifiers |
| geolocation | Geographic reference data |

---

# 3. Data Preparation & Assumptions

- All date columns were converted to proper datetime format.
- Missing values were handled where required.
- Revenue was calculated using payment values.
- Delivery time was computed as the difference between order purchase and customer delivery dates.
- Late delivery was defined as delivery after the estimated delivery date.

---

# 4. Overall Sales Performance

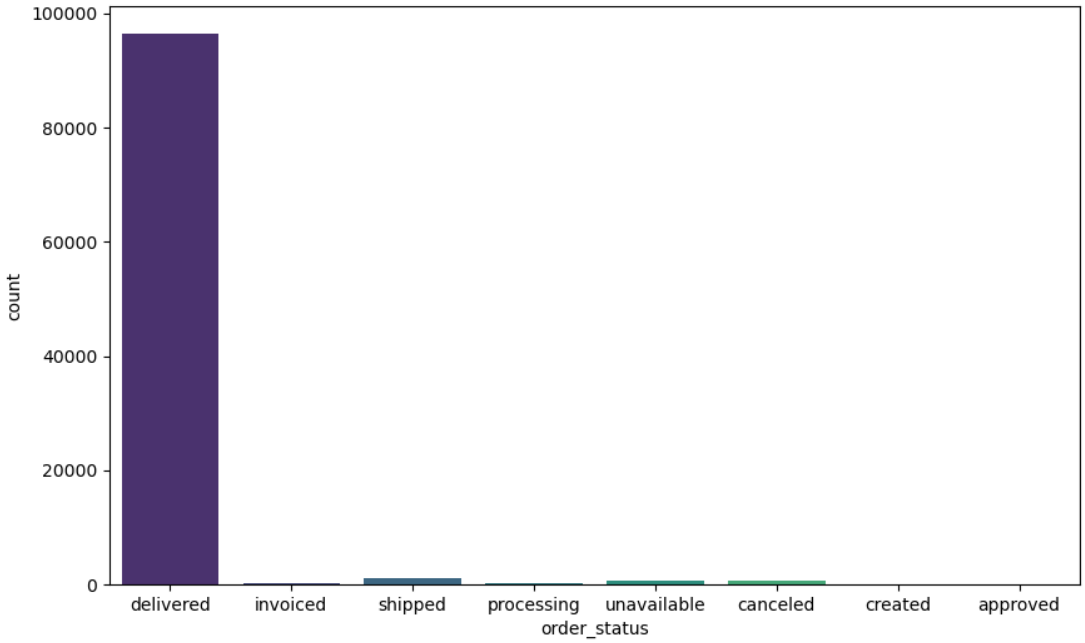
## 4.1 Total Orders

- Total number of unique orders: **99441**

## 4.2 Order Status Distribution

- Most common order status: **Delivered**
- Least common order status: **Approved**

**Observation:**
The majority of orders fall under the **Delivered** status, indicating **good vendor /sellers**




---

# 5. Revenue Analysis

## 5.1 Total Revenue

Total revenue generated: `16008872.12`

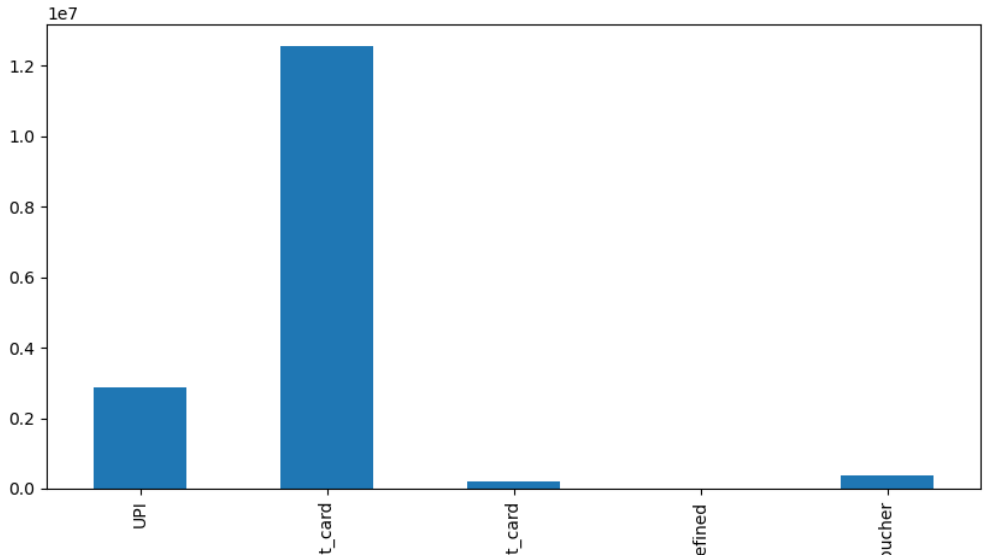
## 5.2 Revenue by Payment Type

- Most used payment method: **credit card**
- Payment method generating highest revenue: **credit card**

**Observation:**
Customers prefer **credit card** as their primary payment method, contributing approximately **97**% of total revenue.

## 5.3 Average Order Value (AOV)

- Average order value: **160.99**




---

# 6. Product & Category Analysis

## 6.1 Most Sold Products

Top-selling product ID(s):

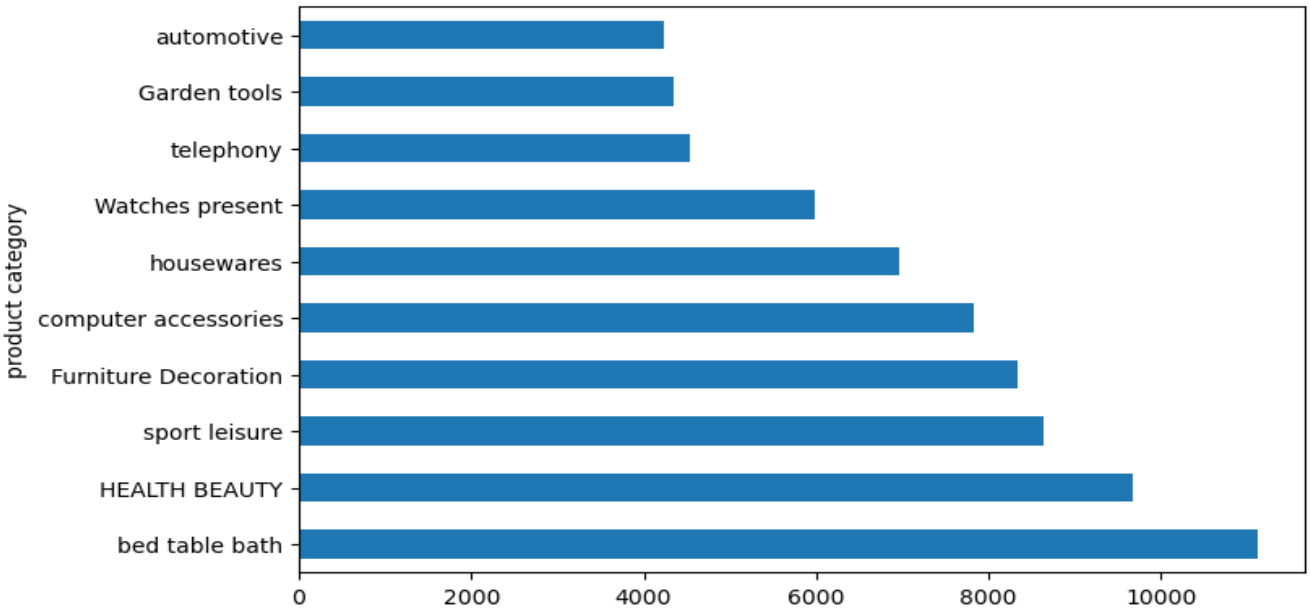
```
product_id
aca2eb7d00ea1a7b8ebd4e68314663af    527
99a4788cb24856965c36a24e339b6058    488
422879e10f46682990de24d770e7f83d    484
389d119b48cf3043d311335e499d9c6b    392
368c6c730842d78016ad823897a372db    388
53759a2ecddad2bb87a079a1f1519f73    373
d1c427060a0f73f6b889a5c7c61f2ac4    343
53b36df67ebb7c41585e8d54d6772e08    323
154e7e31ebfa092203795c972e5804a6    281
3dd2a17168ec895c781a9191c1e95ad7    274
Name: count, dtype: int64
```

## 6.2 Product Category Performance

Most popular product category: `bed table bath`
Least popular product category: `House Comfort 2`

## 6.3 Average Price by Category

Highest average priced category: `PCs`
Lowest average priced category: `House Comfort 2`




## Observation:

Product categories such as PCs tend to have higher average prices, indicating their premium nature and higher perceived value among customers.

---

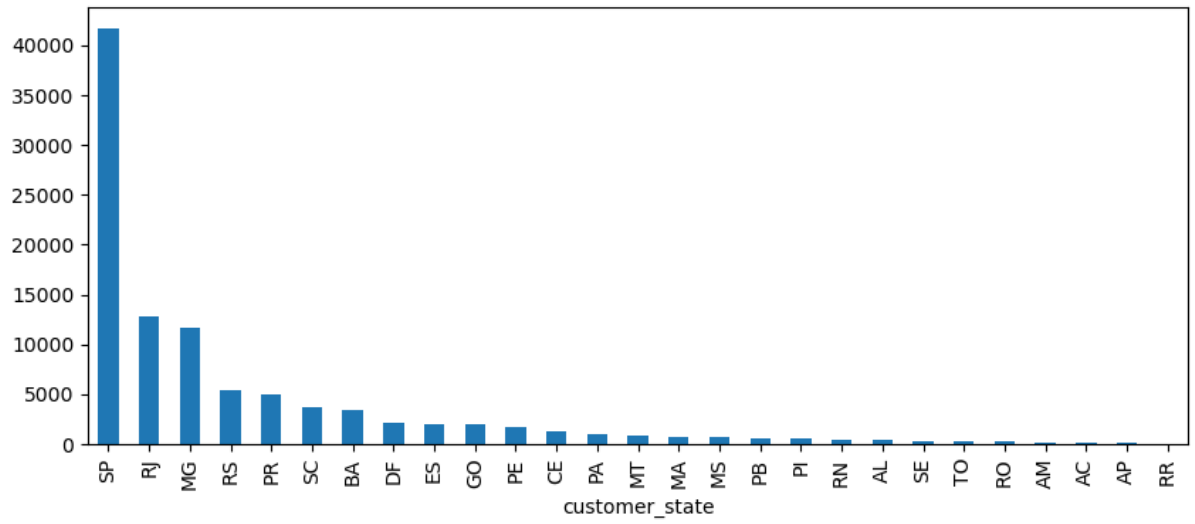
# 7. Customer Analysis

### 7.1 Customer Distribution

- State with highest number of customers: **SP**
- City with highest number of customers: **Sao Paulo**

### 7.2 Orders per Customer

- Average number of orders per customer: **1**




---

# 8. Seller Analysis

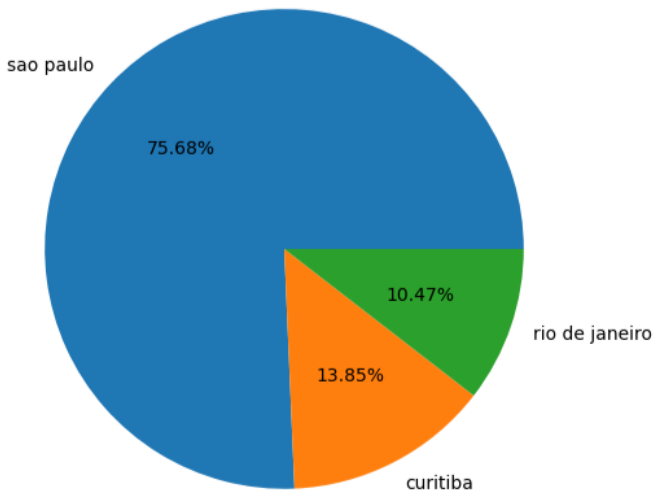
## 8.1 Seller Distribution

- State with most sellers: **SP**

## 8.2 Seller Performance

Top seller (by number of orders): `6560211a19b47992c3666cc44a7e94c0`

- Number of orders fulfilled by top seller: **1854**



## Observation:
Seller activity is concentrated in SP, which may indicate better logistics infrastructure and higher seller density in the region.

---

# 9. Delivery & Logistics Analysis

### 9.1 Delivery Time

- Average delivery time (days): **12.09 days**
- Maximum delivery time observed: **209**

### 9.2 Late Deliveries

- Percentage of late deliveries: **7.87%**

**Observation:**
Late deliveries account for **7.87**% of total orders, suggesting **potential inefficiencies in logistics performance**.

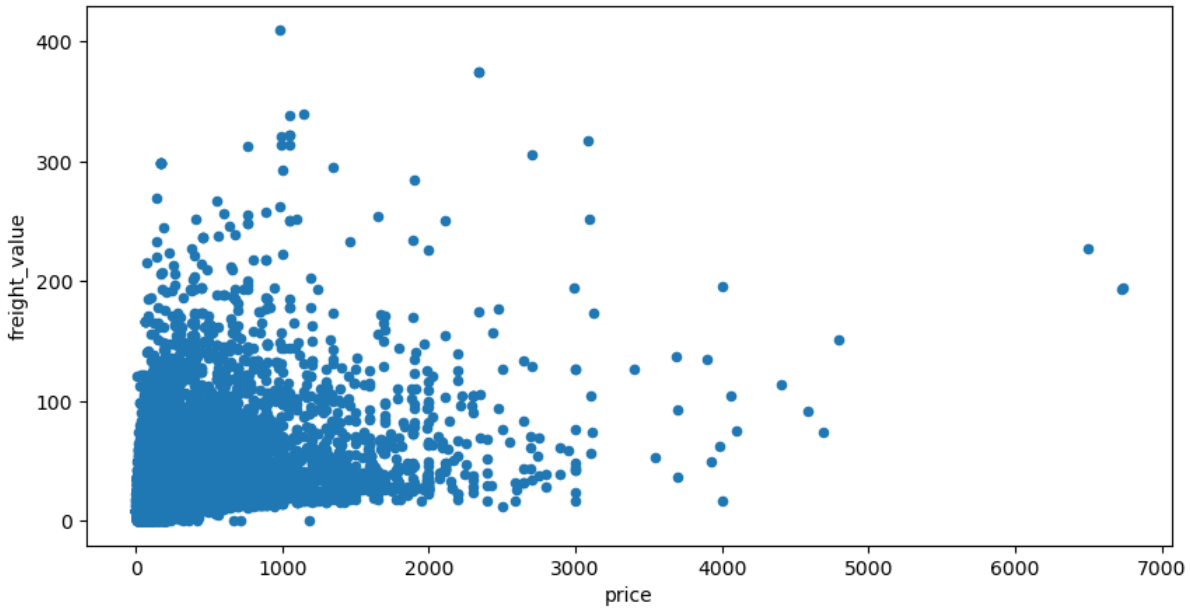
---

# 10. Freight & Product Dimensions

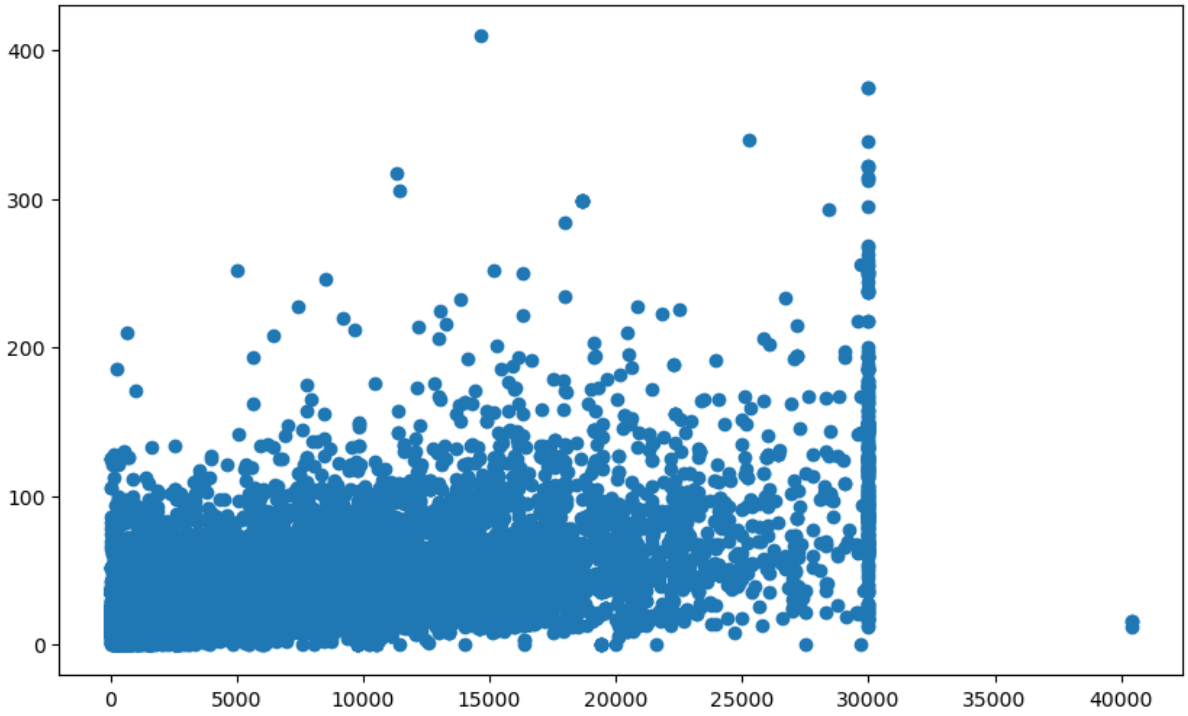
### 10.1 Freight vs Product Price

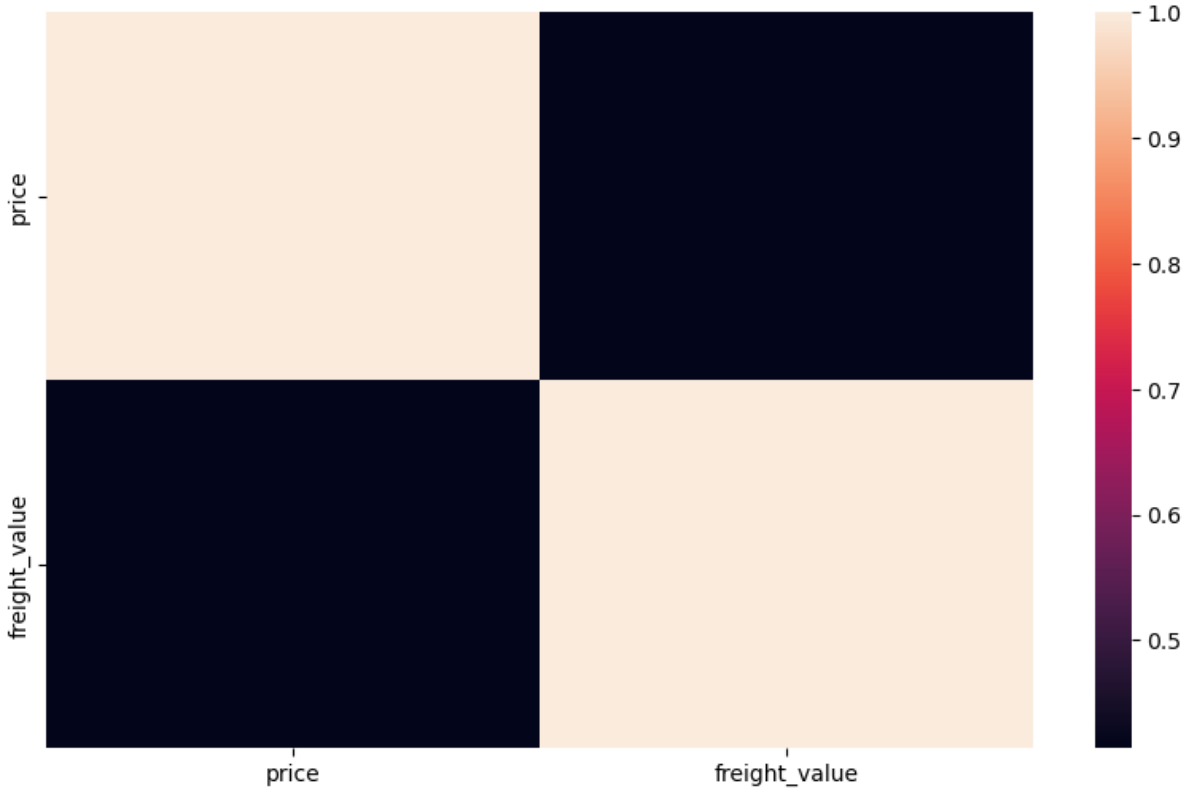
- Correlation between freight value and product price: **0.414204**

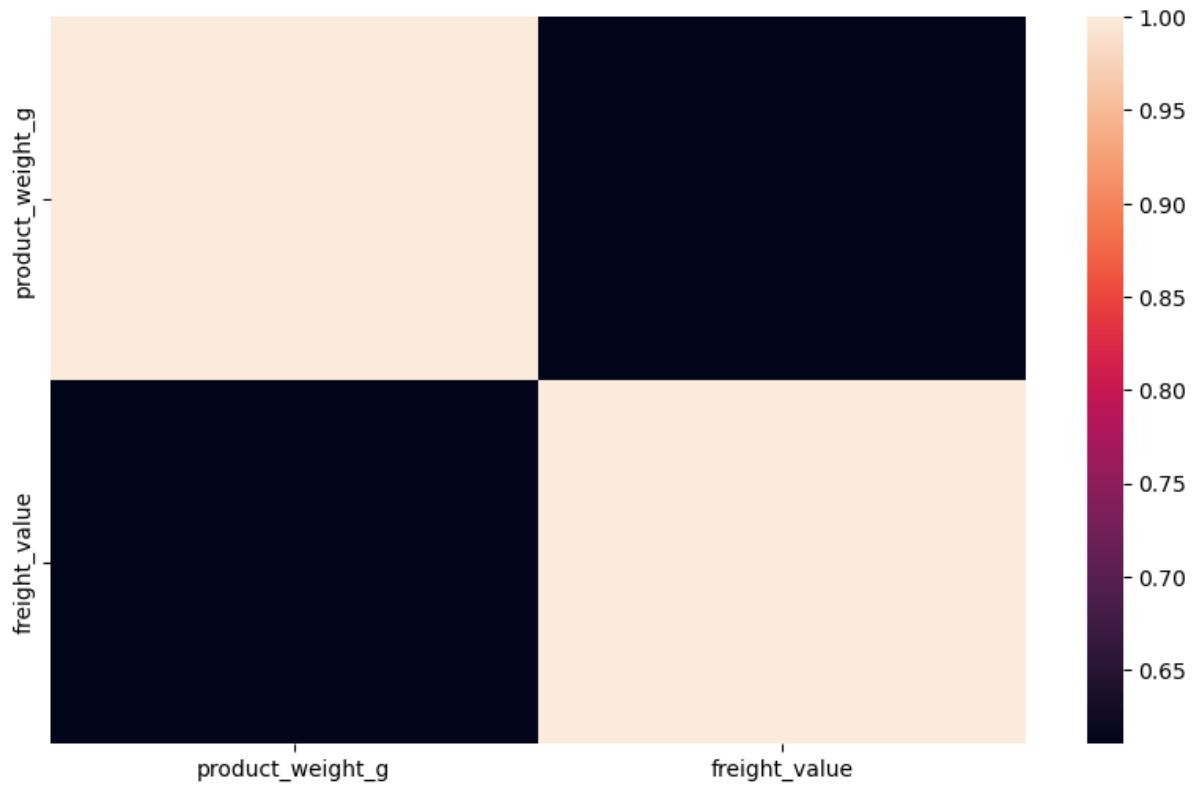
### 10.2 Weight vs Freight Cost



- Correlation between product weight and freight cost: **0.61042**






## Observation:

As product weight increases, freight cost also increases, indicating a strong positive linear relationship between the two variables.

---

# 11. Time-Based Trends

### 11.1 Monthly Order Trends

- Month with highest number of orders: **November(2017)**
- Month with lowest number of orders: **December(2016)**

### 11.2 Monthly Revenue Trends

- Month with highest revenue: **November(2017)**

**Observation:**

Sales exhibit cyclical trends over time, possibly due to strategic discounting and marketing campaigns aligned with peak shopping periods.

---

# 12. Key Business Insights

• **A large share of revenue comes from the top few product categories.**
• **Repeat customers contribute a relatively small percentage of total orders.**
• **Delivery delays are most common in high-volume regions such as SP.**
• **Credit card payment method dominates high-value transactions.**

(This aligns with: dominant categories, avg orders per customer ≈ 1, SP concentration, and ~97% revenue via credit cards.)

---

## 13. Recommendations

Based on the analysis, the following actions are recommended:

1. **Improve logistics efficiency in SP** to reduce late deliveries and handle high order volumes more effectively.
2. **Focus marketing efforts on high-performing categories such as PCs and bed table bath** to maximize revenue impact.
3. **Encourage repeat purchases through loyalty programs targeting existing one-time customers**, as repeat rates are currently low.
4. **Optimize freight pricing and logistics for heavier products** to better manage rising delivery costs associated with product weight.

# Hypothesis Testing

---

## ◆ Test 1: Do late deliveries affect order value?

### Hypothesis

- $H_0$: Average order value is same for late & on-time deliveries
- $H_1$: Average order value is different

### Code (T-test)


**Interpretation**

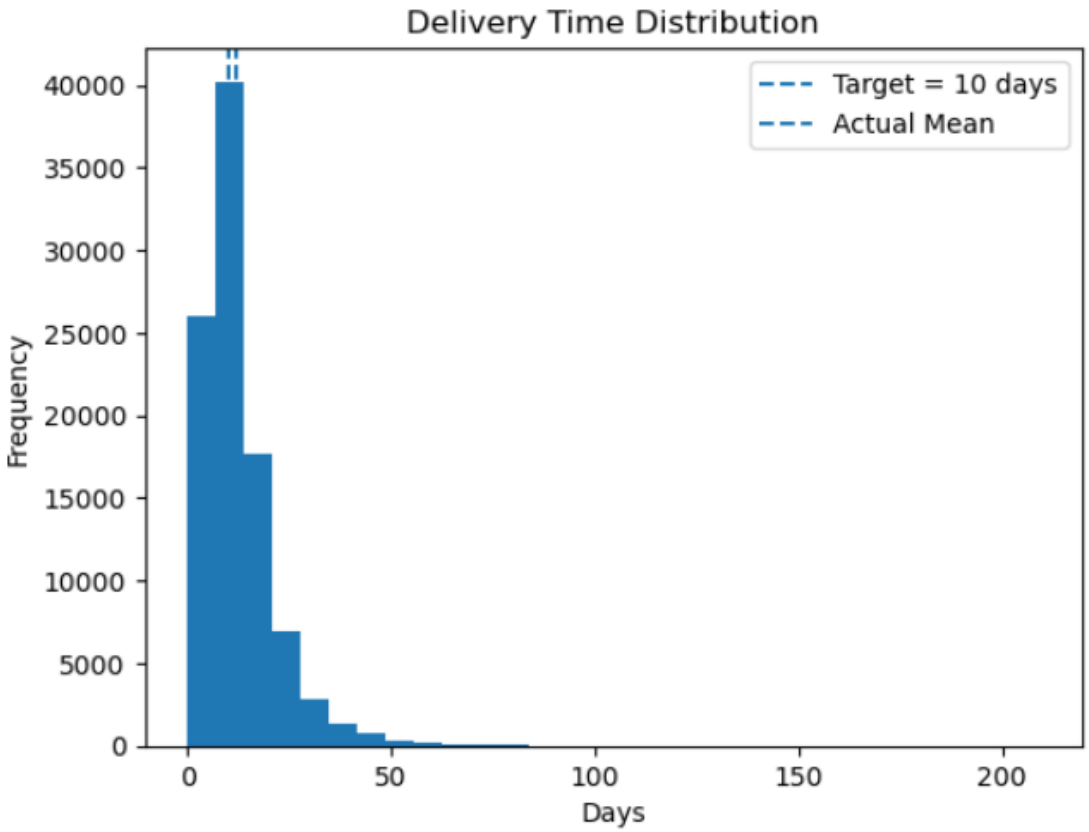
- **p-value < 0.05 → significant difference**
- **p-value ≥ 0.05 → no strong evidence**

## Test 2: Is average delivery time significantly greater than 10 days?

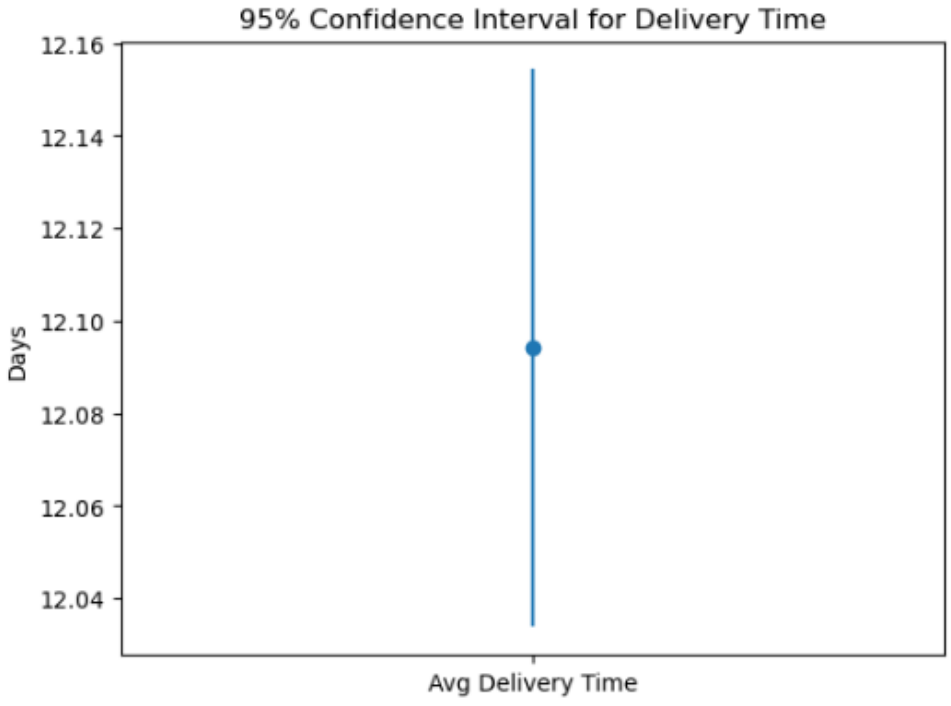
**Hypothesis**

- $H_0$: Average delivery time ≤ 10 days
- $H_1$: Average delivery time > 10 days

**Code (One-sample t-test)**




## Confidence Interval



✅ Interpretation:

We are 95% confident that the **true mean delivery time for all orders** lies between **12.03 and 12.15 units** (e.g., hours or days depending on your data).

# Statistical Analysis

A hypothesis test was conducted to evaluate whether delivery delays impact order value. The results indicate that the difference between late and on-time deliveries is **statistically significant** ($p < 0.05$).
A 95% confidence interval for average delivery time suggests that the true mean delivery duration lies within a narrow range, indicating **stable logistics performance**.

---

## 1️⃣ Hypotheses

- **Null Hypothesis ($H_0$):**
  The mean payment value for late orders is equal to the mean payment value for on-time orders.
- **Alternative Hypothesis ($H_1$):**
  The mean payment value for late orders is different from the mean payment value for on-time orders.
  *(This is a two-tailed test, since `ttest_ind` tests for any difference, not direction.)*

---

## 2️⃣ Test Results

- **t-statistic:** 4.96
  - Indicates the standardized difference between the group means. A higher absolute value reflects a larger difference relative to variability.
- **p-value:** $7.08 \times 10^{-7}$ (~0.0000007)
  - Extremely small, far below common significance levels (0.05, 0.01, 0.001).

---

## 3️⃣ Interpretation

- Since **p-value << 0.05**, we **reject the null hypothesis**.
- There is strong statistical evidence that the mean payment values for late and on-time orders are **significantly different**.

---

## 4️⃣ Confidence Statement

With 99.9% confidence, we conclude that the payment values for late orders are significantly different from those for on-time orders.