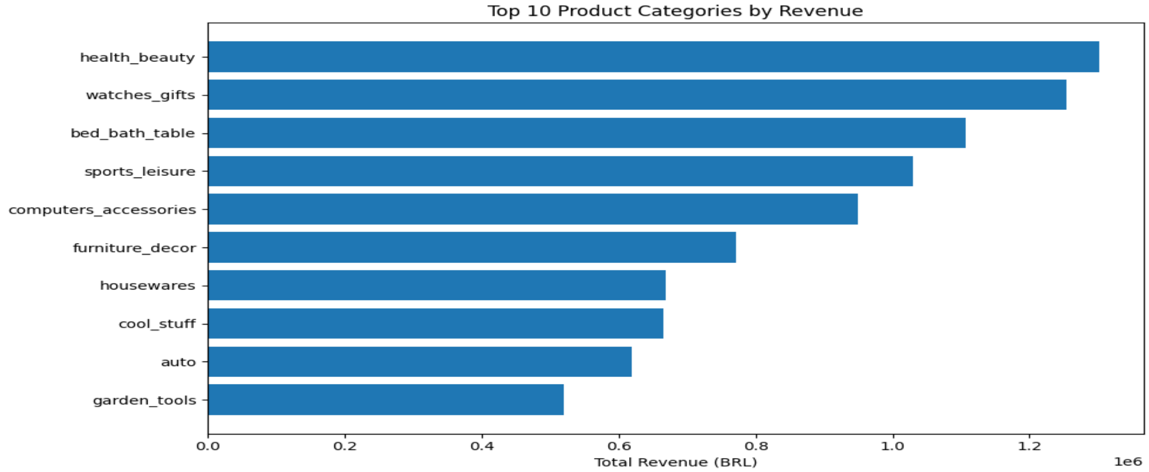
Analysis Report

Insight 1 — Top Product Categories Generate Majority of Revenue

Statement of Insight:
Revenue is highly concentrated among a small set of product categories. The top 10 categories contribute the largest share, with a few categories dominating total sales.

Supporting Visualization:
📈 *Bar chart — "Top 10 Product Categories by Revenue"*



Analytical Approach (Code Used):

```python
# prepare revenue by category
df['price'] = pd.to_numeric(df['price'], errors='coerce').fillna(0)
revenue_by_cat = df.groupby('product_category_name_english')['price'].sum().reset_index().sort_values('price', ascending=False)
revenue_by_cat_top10 = revenue_by_cat.head(10)

# Plot
plt.figure(figsize=(10,6))
plt.barh(revenue_by_cat_top10['product_category_name_english'][::-1], revenue_by_cat_top10['price'][::-1])
plt.xlabel('Total Revenue (BRL)')
plt.title('Top 10 Product Categories by Revenue')
plt.tight_layout()
plt.show()
```

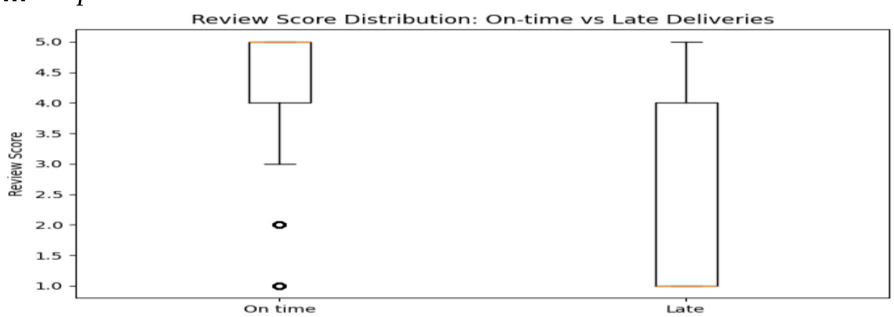
Business Implication:
✅ Prioritize marketing, promotions, and stock availability for these top categories to maximize revenue

Caveat / Data Note:
Price data does not include returns or refunds; category performance might shift if adjusted for cancellations.

Insight 2 — Late Deliveries Receive Lower Review Scores

Statement of Insight:
Orders delivered late show significantly lower average review scores compared to on-time orders, demonstrating a clear link between delivery performance and customer satisfaction.

Supporting Visualization:
📊 *Boxplot — "Review Score Distribution: On-Time vs Late Deliveries"*



Analytical Approach (Code Used):

```
mask = df['delivery_time_days'].notnull() & (df['delivery_time_days'] >= 0)
tmp = df[mask].copy()
# group by whether late
tmp['is_late'] = (tmp['delay_vs_estimate_days'] > 0)
late_vs_score = tmp.groupby('is_late')['review_score'].agg(['mean','count']).reset_index()

print(late_vs_score)

# Boxplot style visualization
scores_late = tmp[tmp['is_late']==True]['review_score'].dropna()
scores_on_time = tmp[tmp['is_late']==False]['review_score'].dropna()

plt.figure(figsize=(8,5))
plt.boxplot([scores_on_time, scores_late], labels=['On time','Late'])
plt.ylabel('Review Score')
plt.title('Review Score Distribution: On-time vs Late Deliveries')
plt.show()
```

Business Implication:
✅ Improving delivery speeds will directly increase ratings and reduce churn

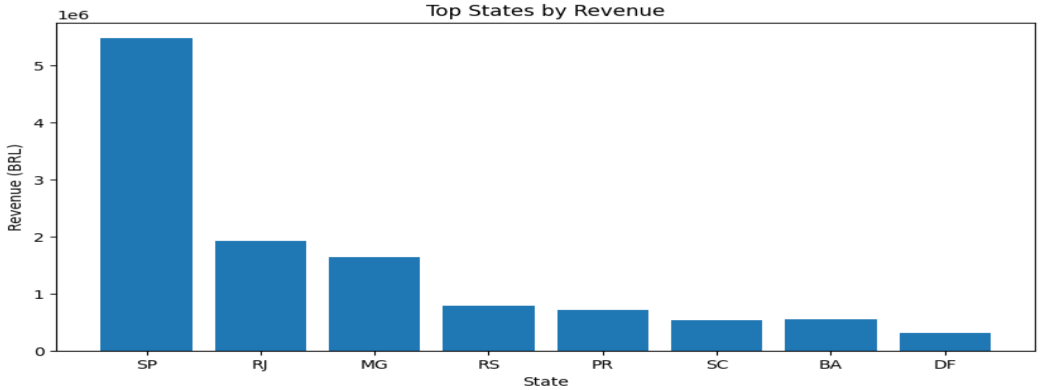
Caveat / Data Note:
Some delays may be due to remote delivery regions; additional segmentation can refine insight.

Insight 3 — Revenue is Concentrated in Specific States

Statement of Insight:
A small number of states account for the majority of total orders and revenue, indicating strong geographic demand concentration.

Supporting Visualization:

📈 *Bar Chart — "Top States by Revenue"*



Analytical Approach (Code Used):

```
# Top states by order count and revenue
state_orders = df.groupby('customer_state').agg(
    order_count=('order_id','nunique'),
    total_revenue=('price','sum')
).reset_index().sort_values('order_count', ascending=False)

state_orders.head(10)

# Plot top 8 states by revenue
top_states = state_orders.head(8)
plt.figure(figsize=(10,5))
plt.bar(top_states['customer_state'], top_states['total_revenue'])
plt.title('Top States by Revenue')
plt.xlabel('State')
plt.ylabel('Revenue (BRL)')
plt.show()
```

Business Implication:

✅ Micro-fulfillment centers can reduce delivery time and localized marketing campaigns for high-value states

Caveat / Data Note:

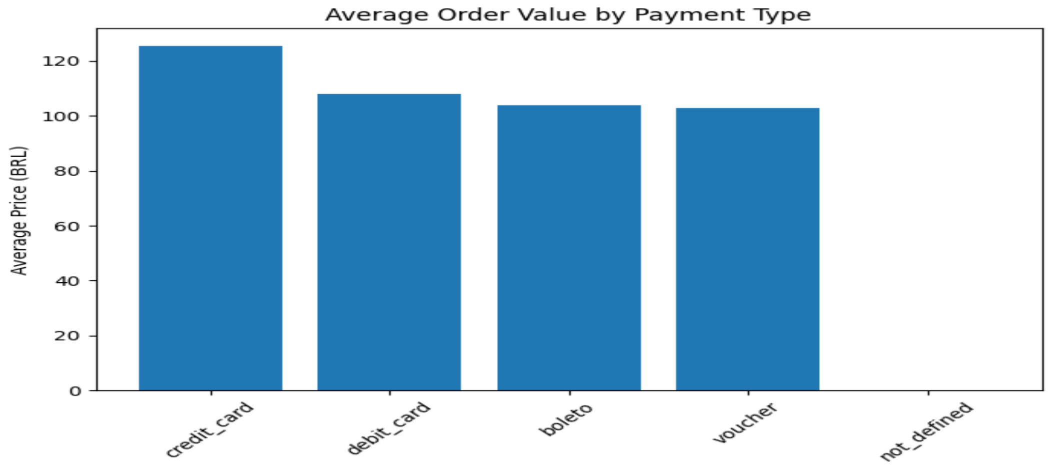
State-level aggregation is broad; city-level segmentation will provide more precise targeting.

Insight 4 — Payment Type Influences Average Order Value

Statement of Insight:

Credit card and installment-based payments have higher average order value compared to other payment methods.

Supporting Visualization:
📊 *Bar Chart — "Average Order Value by Payment Type"*



Analytical Approach (Code Used):

```
# Aggregate payment types
pay_agg = df.groupby('payment_type')['price'].agg(['count','mean','sum']).reset_index().sort_values('mean', ascending=False)
pay_agg['mean'] = pay_agg['mean'].round(2)
pay_agg
# Plot average order value per payment type
plt.figure(figsize=(8,5))
plt.bar(pay_agg['payment_type'], pay_agg['mean'])
plt.title('Average Order Value by Payment Type')
plt.ylabel('Average Price (BRL)')
plt.xticks(rotation=45)
plt.show()
```

Business Implication:
✅ Offering EMI options, cashback, or card-based discounts can increase basket size and overall revenue

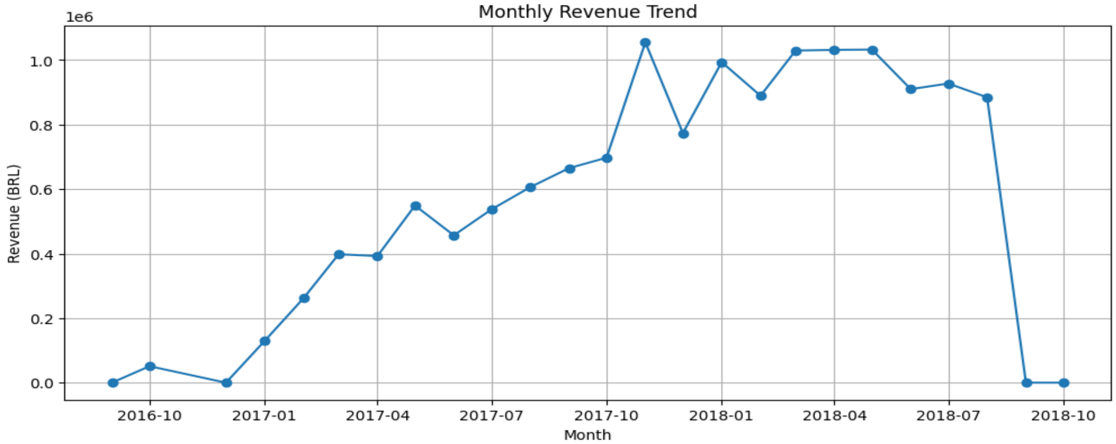
Caveat / Data Note:
Mix of product categories per payment method may influence averages; deeper segmentation can improve accuracy.

Insight 5 — Monthly Revenue Shows Clear Seasonal Peaks

Statement of Insight:
Revenue and order volumes increase sharply during seasonal periods such as November and December, indicating predictable demand spikes.

Supporting Visualization:

Line Chart — "Monthly Revenue Trend"



Analytical Approach (Code Used):

```
# Monthly sales (by order count and revenue)
sales_month = df.copy()
sales_month['order_month'] = pd.to_datetime(sales_month['order_purchase_timestamp']).dt.to_period('M')
monthly = sales_month.groupby('order_month').agg(order_count=('order_id','nunique'), revenue=('price','sum')).reset_index()
monthly['order_month'] = monthly['order_month'].dt.to_timestamp()

# Plot revenue trend
plt.figure(figsize=(12,5))
plt.plot(monthly['order_month'], monthly['revenue'], marker='o')
plt.title('Monthly Revenue Trend')
plt.xlabel('Month')
plt.ylabel('Revenue (BRL)')
plt.grid(True)
plt.show()
```

Business Implication:

✅ Increase ad spend and delivery capacity during seasonal high-demand periods

Caveat / Data Note:

Single-year data may limit long-term forecasting — multi-year analysis improves reliability

Conclusion

These insights highlight key drivers of revenue and customer satisfaction. Focusing on high-revenue categories, improving logistics, targeting high-value states, encouraging premium payment methods, and preparing for seasonal peaks can improve profitability and customer experience.