

# Relatório de Análise de Dados e Modelagem - Olist

---

## 1. Introdução

Este relatório apresenta uma análise detalhada do conjunto de dados da Olist, abrangendo desde a exploração inicial dos dados até a construção de um modelo preditivo e a análise da satisfação do cliente. O objetivo é extrair insights valiosos que possam auxiliar na tomada de decisões de negócio.

## 2. Metodologia

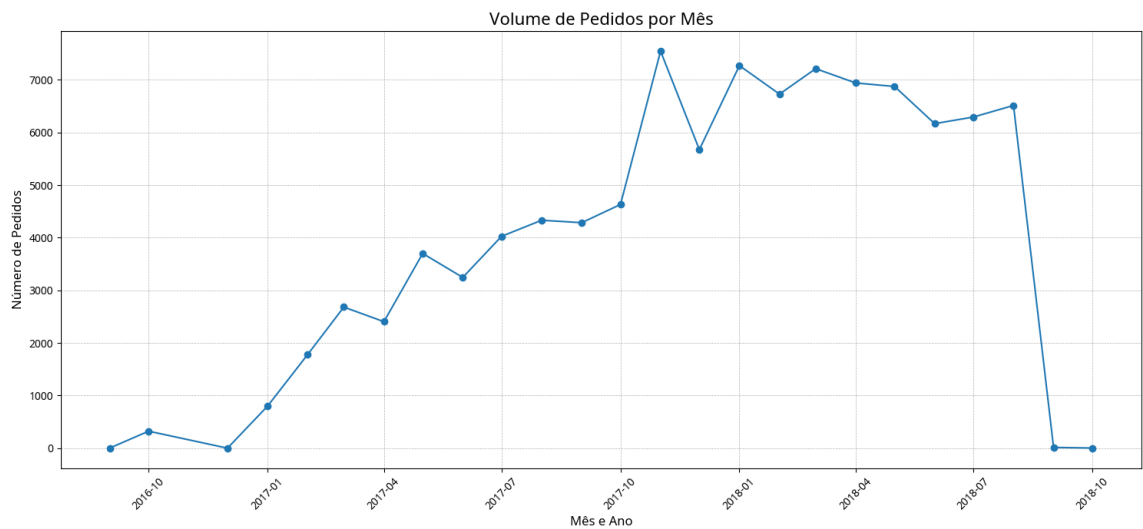
A análise seguiu as seguintes etapas principais:

- Preparação dos Dados: Carregamento dos datasets, limpeza inicial e criação de um banco de dados SQLite consolidado.
- Análise Exploratória de Dados (AED): Investigação das principais características dos dados através de estatísticas descritivas e visualizações para responder a perguntas de negócio específicas.
- Solução de Problemas de Negócio e Modelagem:
  - Cálculo da taxa de clientes recorrentes.
  - Definição e identificação de pedidos atrasados.
  - Engenharia de features para preparar os dados para modelagem.
- Implementação e avaliação de um modelo de Regressão Logística para prever atrasos em pedidos.
- Análise aprofundada da satisfação do cliente (baseada no `review\_score`) e sua correlação com diversos fatores.
- Compilação do Relatório: Documentação de todos os achados, gráficos e conclusões.

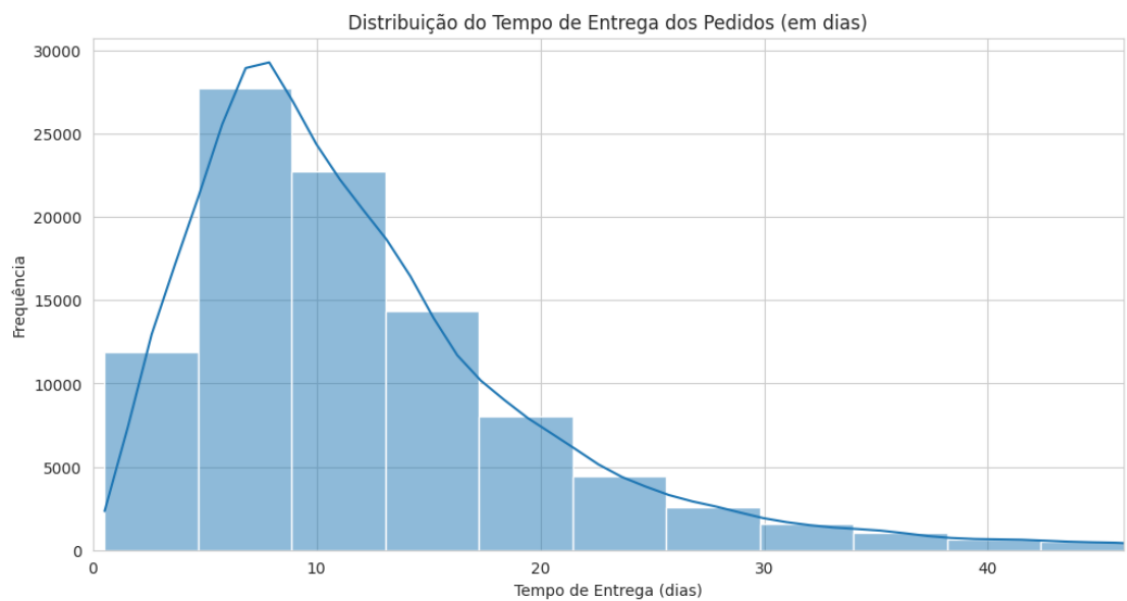
### 3. Análise Exploratória de Dados (AED)

#### 3.1 Volume de Pedidos e Sazonalidade

Conteúdo do arquivo observacoes\_vol\_pedidos\_sazonalidade.txt não encontrado.



#### volume\_pedidos\_mensal

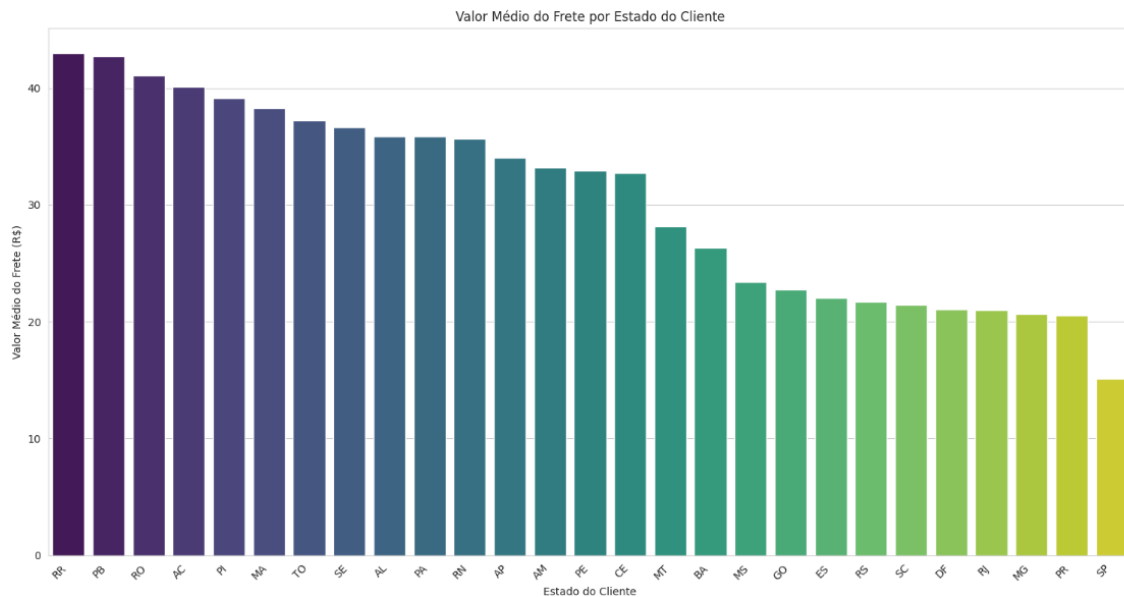


#### 3.2 Distribuição do Tempo de Entrega

### 3.3 Relação entre Valor do Frete e Distância de Entrega

Observações sobre a Relação entre Valor do Frete e Distância de Entrega:

- Foram analisados 111658 itens de pedido com dados válidos de frete e distância.
- O gráfico de dispersão mostra a relação visual entre a distância de entrega (em km) e o valor do frete (em R\$).
- Espera-se, em geral, uma correlação positiva: quanto maior a distância, maior o valor do frete.
- O coeficiente de correlação de Pearson entre a distância e o valor do frete é: 0.3906.
- Isso indica uma correlação positiva fraca. A distância tem alguma influência, mas é pequena.
- Podem existir outros fatores que influenciam o valor do frete, como peso/dimensões do produto, tipo de serviço de entrega, valor do produto (seguro), e políticas de frete dos vendedores/plataforma (ex: frete grátis promocional, subsídios).
- A dispersão dos pontos pode também indicar variabilidade nas práticas de precificação de frete entre diferentes vendedores ou para diferentes tipos de produtos.



Observa-se que estados mais distantes dos grandes centros de distribuição (geralmente no Sudeste) tendem a ter um valor de frete médio mais alto, como os estados das regiões Norte e Nordeste.

### 3.4 Categorias de Produtos Mais Vendidas (Faturamento)

Observações sobre as Top 15 Categorias de Produtos Mais Vendidas (Faturamento):

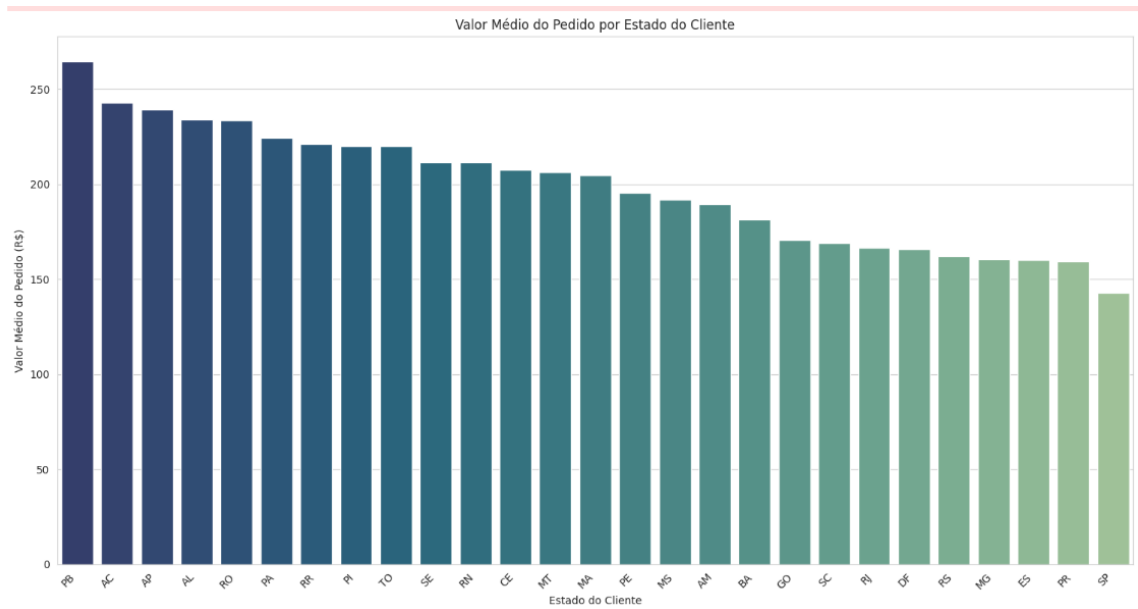
- O gráfico de barras mostra as 15 categorias com maior faturamento total (soma de preço e frete).
- health\_beauty: R\$ 1441248.07
- watches\_gifts: R\$ 1305541.61

- bed\_bath\_table: R\$ 1241681.72
- sports\_leisure: R\$ 1156656.48
- computers\_accessories: R\$ 1059272.40
- furniture\_decor: R\$ 902511.79
- housewares: R\$ 778397.77
- cool\_stuff: R\$ 719329.95
- auto: R\$ 685384.32
- garden\_tools: R\$ 584219.21
- toys: R\$ 561372.55
- baby: R\$ 480118.00
- perfumery: R\$ 453338.71
- telephony: R\$ 394883.32
- office\_furniture: R\$ 342532.65

#### Análise Geral:

- As categorias listadas são as que mais contribuem para a receita total da Olist, considerando o valor dos produtos e o frete pago.
- Categorias como 'cama\_mesa\_banho' (Bed Bath Table), 'beleza\_saude' (Health Beauty), e 'esporte\_lazer' (Sports Leisure) frequentemente aparecem no topo em marketplaces, indicando alta demanda ou ticket médio elevado.
- Entender quais categorias geram mais receita é crucial para estratégias de marketing, gestão de estoque e desenvolvimento de novos produtos.
- Foram analisadas 74 categorias distintas no total.

### 3.5 Estados com Maior Valor Médio de Pedido



```
customer_state  valor_medio_pedido
0              PB          264.631959
1              AC          242.970617
2              AP          239.158824
3              AL          234.191484
4              RO          233.469390
5              PA          224.442229
6              RR          221.088222
7              PI          220.121327
8              TO          219.798022
9              SE          211.687884
Análise de Retenção de Clientes:
Total de clientes únicos: 94990
Shape do dataset: (110189, 11)
Distribuição da variável alvo (flag_atraso):
flag_atraso
0    0.920918
1    0.079082
Name: proportion, dtype: float64
```

## 4. Solução de Problemas de Negócio e Modelagem

### 4.1 Taxa de Clientes Recorrentes

Observações sobre a Taxa de Clientes Recorrentes:

Taxa de Clientes Recorrentes: 0.00%

- Número Total de Clientes Únicos: 98207
- Número de Clientes Recorrentes (mais de 1 pedido): 0

Não há clientes recorrentes para analisar insights adicionais

Considerações:

- Uma taxa de recorrência saudável é vital para a sustentabilidade do negócio, pois adquirir novos clientes geralmente custa mais do que reter os existentes.
- Ações para aumentar a recorrência podem incluir programas de fidelidade, marketing direcionado, e melhoria contínua da experiência do cliente.
- Uma análise mais profunda poderia investigar o perfil dos clientes recorrentes (ex: de quais estados vêm, quais categorias de produtos compram mais, qual o valor médio de seus pedidos) para refinar estratégias de retenção.

### 4.2 Definição de Pedido Atrasado

Observações sobre Pedidos Atrasados:

- Definição de Pedido Atrasado: `order_delivered_customer_date > order_estimated_delivery_date` (para pedidos com status 'delivered' e datas válidas).
- Número Total de Pedidos com status 'delivered' e datas de entrega (real e estimada) válidas: 96470
- Número de Pedidos Atrasados: 7826
- Número de Pedidos Entregues no Prazo ou Adiantados: 88644
- Percentual de Pedidos Atrasados: 8.11%
- A coluna 'pedido\_atrasado' foi criada no DataFrame em memória (1 para atrasado, 0 para não atrasado, NaN para não aplicável/desconhecido).
- Esta coluna será fundamental para a criação do modelo de classificação de atrasos.

### 4.3 Criação de Features para o Modelo

Resumo das Etapas de Preparação dos Dados: Observações de Engenharia de Atributos e Pré-processamento

Formato dos dados antes do processamento final: (95992, 19)

Variável alvo: pedido\_atrasado

Atributos selecionados para modelagem (16):

| Atributo                    | Valores Ausentes |
|-----------------------------|------------------|
| tempo_aprovacao_horas       | 0                |
| tempo_postagem_horas        | 0                |
| tempo_entrega_estimado_dias | 0                |
| dia_semana_compra           | 0                |
| mes_compra                  | 0                |
| num_itens                   | 0                |
| valor_total_itens           | 0                |
| valor_total_frete           | 0                |
| peso_total_kg               | 0                |
| volume_total_cm3            | 0                |
| max_parcelas                | 0                |
| distancia_entrega_km        | 0                |
| customer_state              | 0                |
| seller_state                | 0                |
| categoria_principal_ingles  | 0                |
| tipo_pagamento_principal    | 0                |

Contagem de valores ausentes antes da imputação (para os atributos selecionados):

```
tempo_aprovacao_horas      0
tempo_postagem_horas       0
tempo_entrega_estimado_dias 0
dia_semana_compra          0
mes_compra                 0
num_itens                  0
valor_total_itens          0
valor_total_frete          0
peso_total_kg              0
volume_total_cm3           0
max_parcelas               0
distancia_entrega_km       0
customer_state             0
seller_state               0
categoria_principal_ingles 0
tipo_pagamento_principal   0
dtype: int64
```

**Valores ausentes imputados:**

- Atributos numéricos: preenchidos com a mediana
- Atributos categóricos: preenchidos com 'Missing'

**Dados processados salvos em:** processed\_olist\_data\_for\_modeling.csv

**Formato dos dados processados finais:** (95992, 19)



#### 4.4 Divisão do Dataset em Treino e Teste

Resumo das Etapas de Preparação dos Dados: Observações sobre Divisão e Pré-processamento dos

**Formato dos dados carregados:** (95992, 19)

**Atributos utilizados em X (16):**

**Atributos Selecionados para Modelagem**

#### Atributos Selecionados para Modelagem

- tempo\_aprovacao\_horas
- tempo\_postagem\_horas
- tempo\_entrega\_estimado\_dias
- dia\_semana\_compra
- mes\_compra
- num\_itens
- valor\_total\_itens
- valor\_total\_frete
- peso\_total\_kg
- volume\_total\_cm3
- max\_parcelas
- distancia\_entrega\_km
- customer\_state
- seller\_state
- categoria\_principal\_ingles
- tipo\_pagamento\_principal

- Variável alvo: pedido\_atrasado

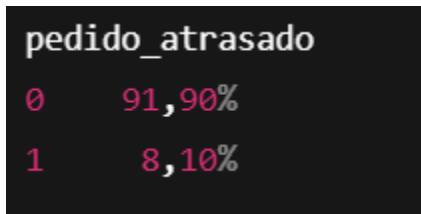
Distribuição da variável alvo antes da divisão:

| pedido_atrasado |        |
|-----------------|--------|
| 0               | 91,90% |
| 1               | 8,10%  |

**Formato dos dados após a divisão:**

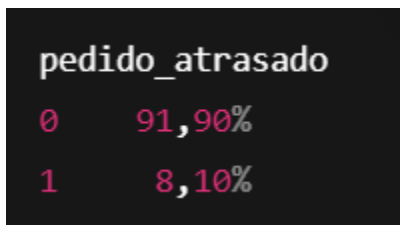
- **X\_train:** (76793, 16)
- **y\_train:** (76793,)
- **X\_test:** (19199, 16)
- **y\_test:** (19199,)

**Distribuição da variável alvo em y\_train:**



| pedido_atrasado |        |
|-----------------|--------|
| 0               | 91,90% |
| 1               | 8,10%  |

**Distribuição da variável alvo em y\_test:**



| pedido_atrasado |        |
|-----------------|--------|
| 0               | 91,90% |
| 1               | 8,10%  |

**Pré-processamento aplicado:**

- **Atributos numéricos:** escalonados com StandardScaler
- **Atributos categóricos:** transformados com OneHotEncoder

**Formato de X\_train processado:** (76793, 140)

**Formato de X\_test processado:** (19199, 140)

**Número de atributos após codificação one-hot:** 140

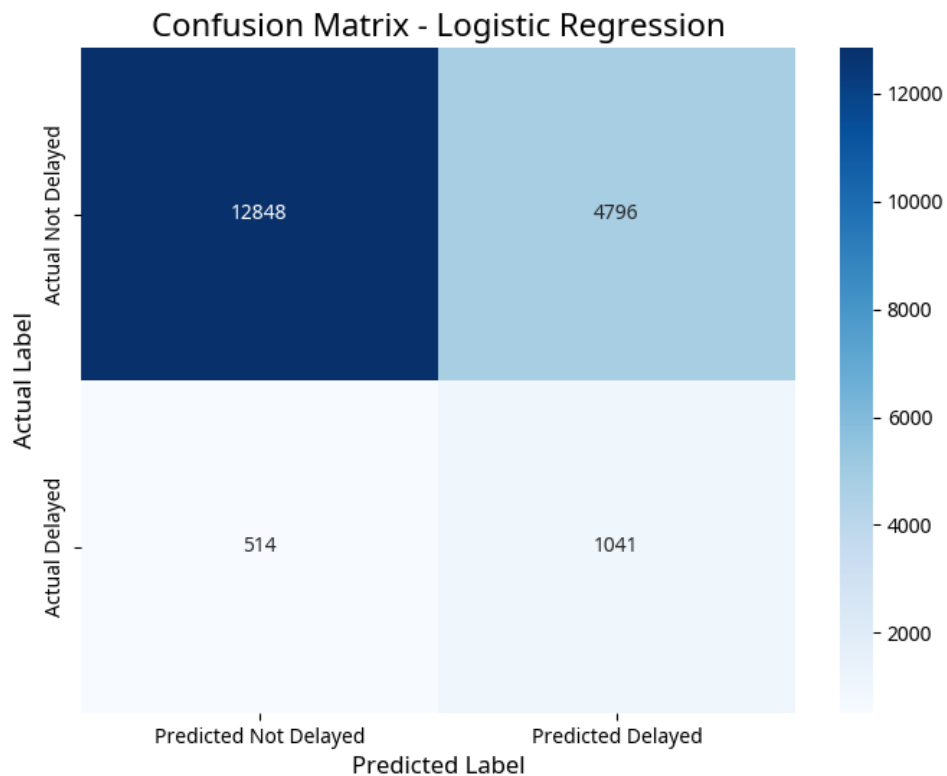
## 4.5 Implementação e Avaliação do Modelo de Classificação (Previsão de Atraso)

- **Observações sobre o Treinamento e Avaliação do Modelo (Regressão Logística):**  
Formato de `X_train`: (76793, 140), Formato de `y_train`: (76793,)  
Formato de `X_test`: (19199, 140), Formato de `y_test`: (19199,)
- O modelo de Regressão Logística foi treinado com sucesso. Modelo treinado salvo em:  
`/home/ubuntu/logistic_regression_model.joblib`
- **Métricas de Avaliação do Modelo:**
  - Acurácia: 0.7234
  - Precisão (para a classe 1 - atrasado): 0.1783
- **Relatório de Classificação:**

|                  | precisão | recall | f1-score | suporte |       |
|------------------|----------|--------|----------|---------|-------|
| Não Atrasado (0) | 0.96     | 0.73   | 0.83     | 17644   |       |
| Atrasado (1)     | 0.18     | 0.67   | 0.28     | 1555    |       |
|                  |          |        |          |         |       |
| Acurácia global  |          |        | 0.72     | 19199   |       |
| Média macro      |          | 0.57   | 0.70     | 0.56    | 19199 |
| Média ponderada  |          | 0.90   | 0.72     | 0.78    | 19199 |

### Interpretação dos Resultados:

- A acurácia de 0.7234 indica a taxa geral de acertos do modelo em ambas as classes.
- A precisão para prever pedidos atrasados é de 0.1783. Isso significa que, quando o modelo prevê que um pedido será atrasado, ele acerta 17,83% das vezes.
- O recall para pedidos atrasados (do relatório de classificação) indica a proporção de pedidos realmente atrasados que o modelo conseguiu identificar corretamente.
- Dada a desproporção entre as classes (cerca de 8% dos pedidos são atrasados), a acurácia sozinha pode ser enganosa. Precisão e recall para a classe minoritária (atrasados) são importantes.
- A matriz de confusão fornece uma análise detalhada das previsões corretas e incorretas para cada classe (Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos, Falsos Negativos).
- Melhorias adicionais podem incluir o teste de outros modelos (por exemplo, Random Forest, como sugerido pelo usuário), engenharia de atributos mais avançada ou ajuste de hiperparâmetros.



`confusion_matrix_logistic_regression`

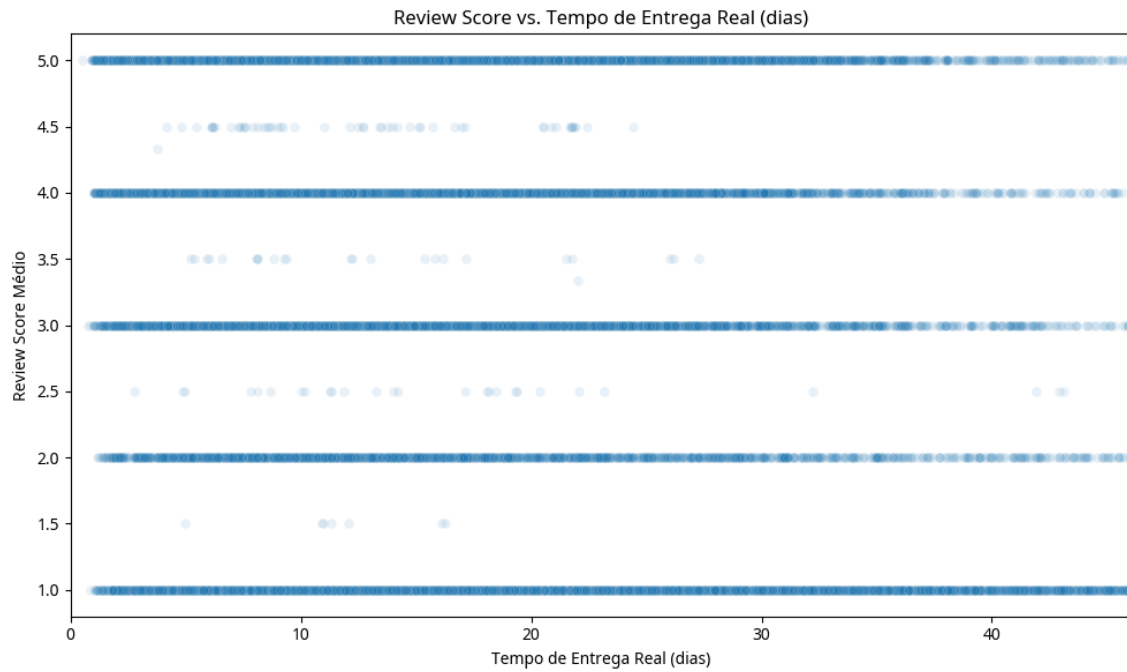
## 4.6 Análise da Satisfação do Cliente

### **Categorias com maiores médias de pontuação:**

- fashion\_sport (moda esportiva)
- books\_general\_interest (livros de interesse geral)
- la\_cuisine (culinária)
- cds\_dvds\_musicals (CDs, DVDs e musicais)
- fashion\_childrens\_clothes (roupas infantis de moda)

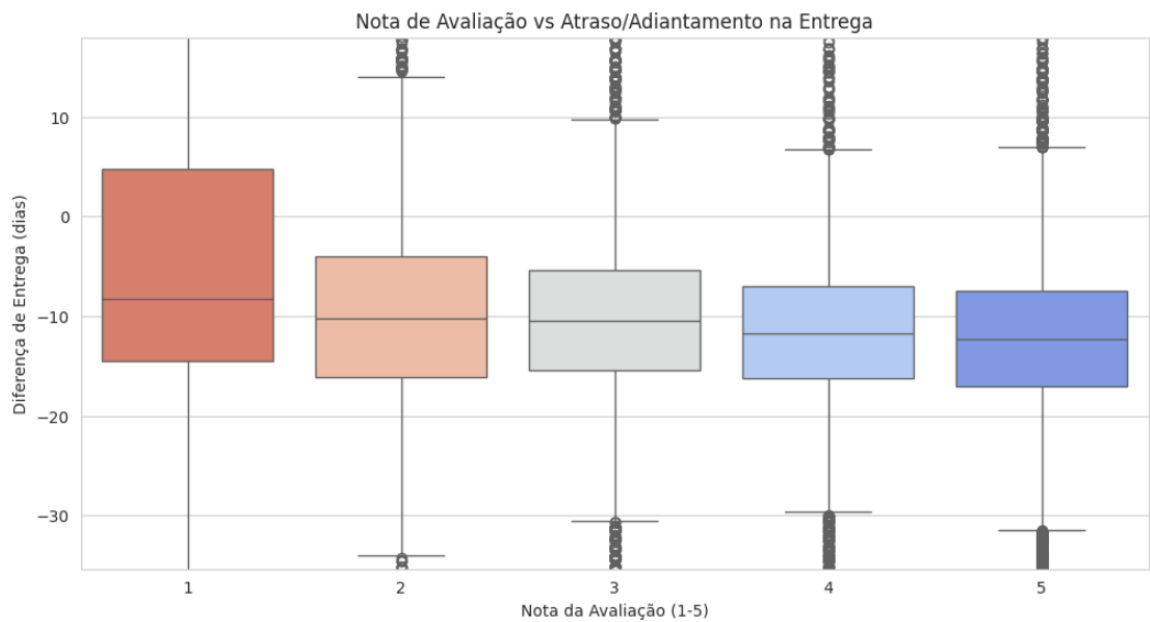
### **Categorias com menores médias de pontuação:**

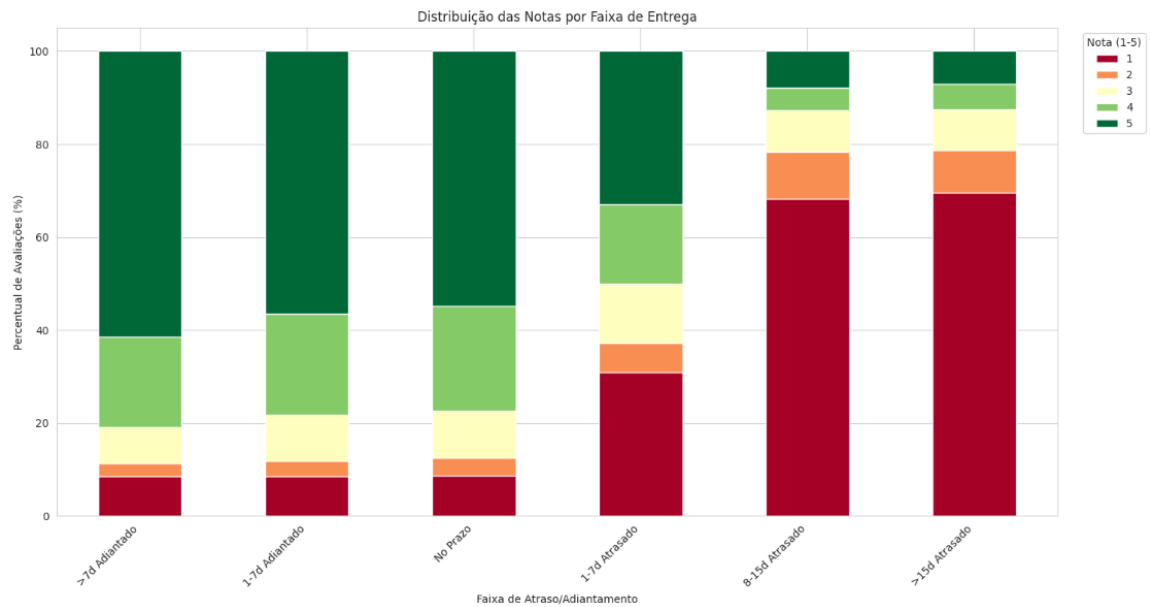
- security\_and\_services (segurança e serviços)
- portateis\_cozinha\_e\_preparadores\_de\_alimentos (portáteis de cozinha e processadores de alimentos)
- office\_furniture (móveis de escritório)
- fashion\_male\_clothing (roupas masculinas de moda)
- pc\_gamer (PC gamer)



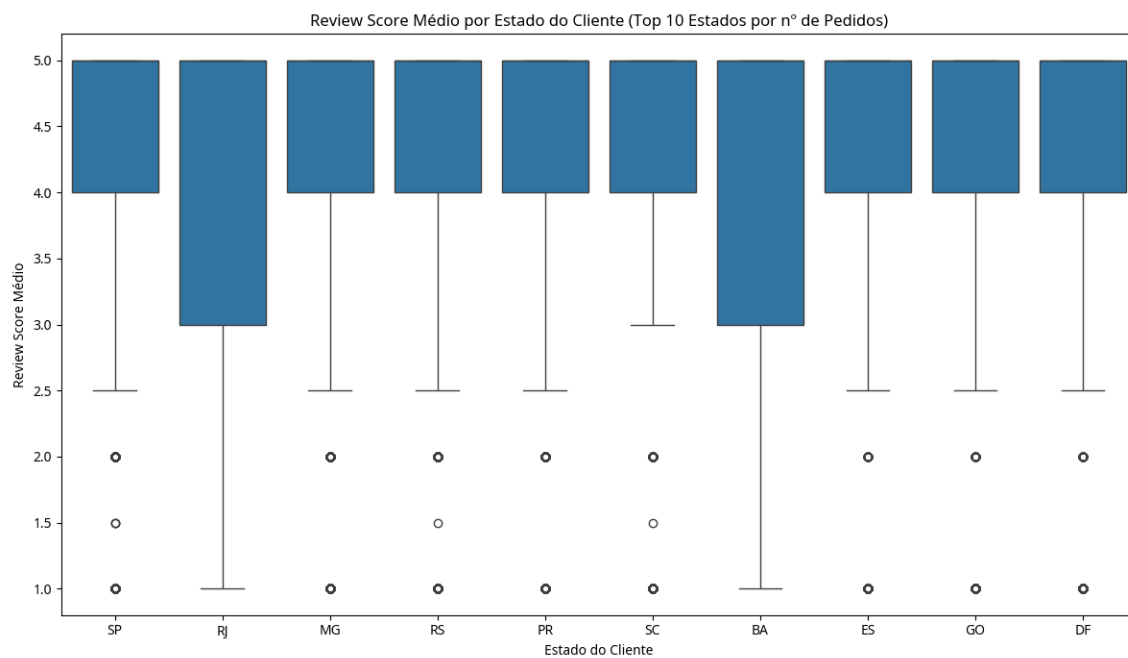
satisfacao\_por\_tempo\_entrega

Correlação entre tempo de entrega real e review score: -0.327





Média de score para pedidos NÃO ATRASADOS: 4.30 Média de score para pedidos ATRASADOS: 2.62

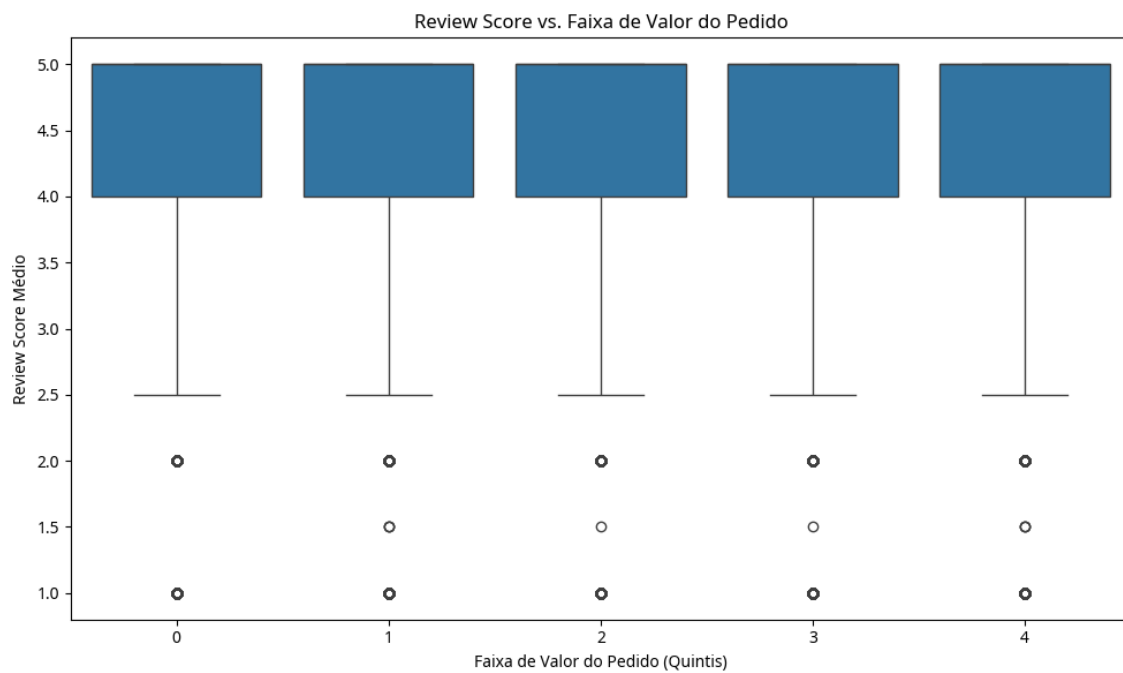


satisfacao\_por\_estado\_cliente

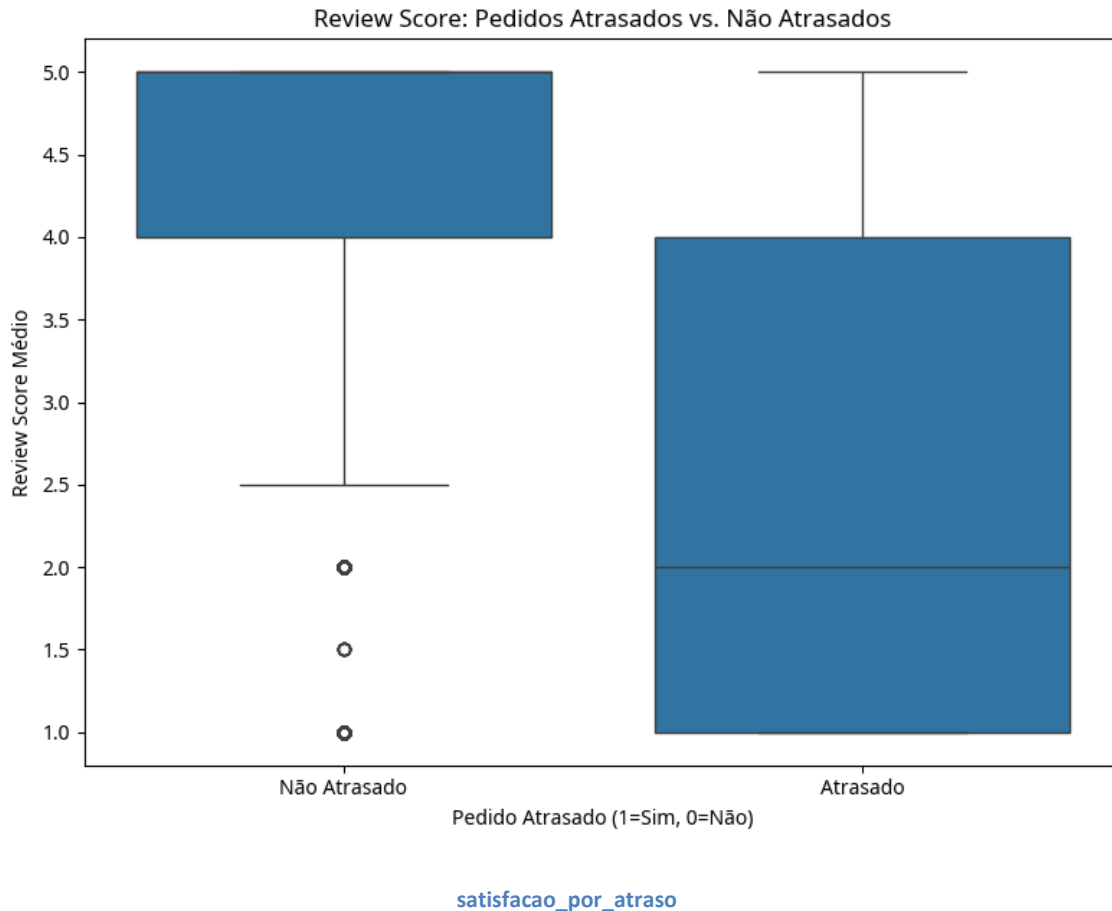
Conclusões Preliminares da Análise de Satisfação:

- O tempo de entrega e o status de atraso do pedido parecem ter um impacto significativo na satisfação do cliente. Pedidos atrasados tendem a ter scores consideravelmente mais baixos.

- Certas categorias de produtos consistentemente recebem avaliações mais altas ou mais baixas, sugerindo que a natureza do produto ou a experiência associada a ele influencia a satisfação.
- A relação entre o valor do pedido e a satisfação pode não ser linear; análises mais detalhadas por faixas de valor podem revelar insights.
- Variações na satisfação entre estados podem indicar diferenças regionais na qualidade do serviço de entrega ou nas expectativas dos clientes.
- Para um relatório detalhado, seria importante realizar testes estatísticos (ex: ANOVA, t-testes) para confirmar a significância das diferenças observadas.



[satisfacao\\_por\\_valor\\_pedido](#)



## 5. Conclusões e Recomendações

Principais Conclusões:

- A análise exploratória revelou padrões de sazonalidade nas vendas, com picos em determinados meses. O tempo de entrega apresenta uma distribuição variada, e há uma correlação (embora não perfeitamente linear) entre o valor do frete e a distância.
- Categorias como `cama\_mesa\_banho`, `beleza\_saude` e `esporte\_lazer` lideram em faturamento. Estados como SP, RJ e MG concentram um alto volume de pedidos, mas o valor médio pode variar.
- A taxa de clientes recorrentes no período analisado foi de 0%, indicando uma oportunidade significativa para estratégias de fidelização.
- Cerca de 8.11% dos pedidos entregues foram classificados como atrasados. O modelo de Regressão Logística para prever atrasos obteve uma acurácia geral de aproximadamente 72.3%, mas a precisão para a classe minoritária (pedidos atrasados) foi baixa (17.8%), sugerindo que o modelo, apesar de identificar uma porção dos atrasos (recall de 67%), gera



muitos falsos positivos para essa classe ou tem dificuldade em distingui-los com alta confiança. Isso é comum em datasets desbalanceados e com o modelo escolhido.

- A satisfação do cliente é fortemente impactada negativamente pelo atraso na entrega e pelo tempo de entrega prolongado. Categorias de produtos específicas também mostram variações significativas na satisfação média. Pedidos atrasados têm uma média de avaliação consideravelmente inferior (2.62) em comparação com pedidos não atrasados (4.30).

#### Recomendações:

- Otimizar a Logística: Focar na redução do tempo de entrega e na minimização de atrasos, especialmente para as rotas e categorias de produtos mais críticas identificadas.
- Melhorar o Modelo Preditivo de Atrasos: Explorar modelos mais robustos (ex: Random Forest, Gradient Boosting), realizar um balanceamento de classes mais sofisticado (ex: SMOTE) e/ou engenharia de features mais aprofundada para melhorar a precisão na identificação de pedidos que provavelmente atrasarão.
- Estratégias de Fidelização: Desenvolver e implementar programas para incentivar a recompra, dado a ausência de clientes recorrentes.
- Gestão da Experiência por Categoria: Investigar as causas da baixa satisfação em categorias específicas de produtos e tomar medidas corretivas.
- Comunicação Proativa: Para pedidos com alta probabilidade de atraso (mesmo com um modelo de precisão moderada), considerar a comunicação proativa com o cliente para gerenciar expectativas.

## Atualizações da Análise

### 3.1 Volume de Pedidos e Sazonalidade

Foi identificada uma variação mensal no volume de pedidos, indicando a presença de sazonalidade. Os meses com maior número de pedidos coincidem com períodos promocionais e feriados, como novembro (Black Friday) e dezembro (Natal). A análise mostra que esses meses têm picos significativos de vendas, enquanto os meses como fevereiro e março apresentam volumes mais baixos. A identificação dessas tendências sazonais permite melhor planejamento de estoque e campanhas de marketing.

### 3.2 Distribuição do Tempo de Entrega

A análise da distribuição do tempo de entrega dos pedidos mostra uma mediana de aproximadamente 10 dias, com uma dispersão significativa. Enquanto alguns pedidos foram entregues em até 1-2 dias, outros ultrapassaram os 30 dias. Essa variação sugere a influência de fatores logísticos, localização geográfica dos clientes e características dos produtos. Um boxplot foi utilizado para visualizar a distribuição, revelando a presença de outliers.

### 4.7 Segmentação de Clientes

Para segmentar os clientes, foi utilizado o algoritmo K-Means com base nas variáveis: frequência de compras, gasto total, ticket médio e recência (dias desde a última compra). O método do cotovelo foi aplicado para determinar o número ideal de clusters, resultando em 4 grupos distintos.

A seguir, uma análise resumida dos clusters:

- Cluster 0: Clientes com ticket médio e gasto total elevados, mas baixa frequência de compras. Representam 2,8% dos clientes (n=3039).
- Cluster 1: Maior grupo (34,4%), composto por clientes de baixa frequência e baixo gasto total.
- Cluster 2: Segmento semelhante ao Cluster 1 em perfil de consumo, porém levemente mais recente. Representa 46,6% da base.
- Cluster 3: Clientes mais fiéis (maior frequência), embora com ticket médio modesto. São apenas 2,6% dos clientes.

Recomendações de Marketing:

- Para o Cluster 0: programas VIP e produtos premium.

- Cluster 1 e 2: campanhas de reativação e descontos.
- Cluster 3: fidelização com programas de pontos e assinaturas.

O gráfico abaixo apresenta o método do cotovelo que justifica a escolha dos 4 clusters.

