

Prática em Ciência de Dados 3 - Projeto 1

Maria Victória Brandão Barros - 12608692

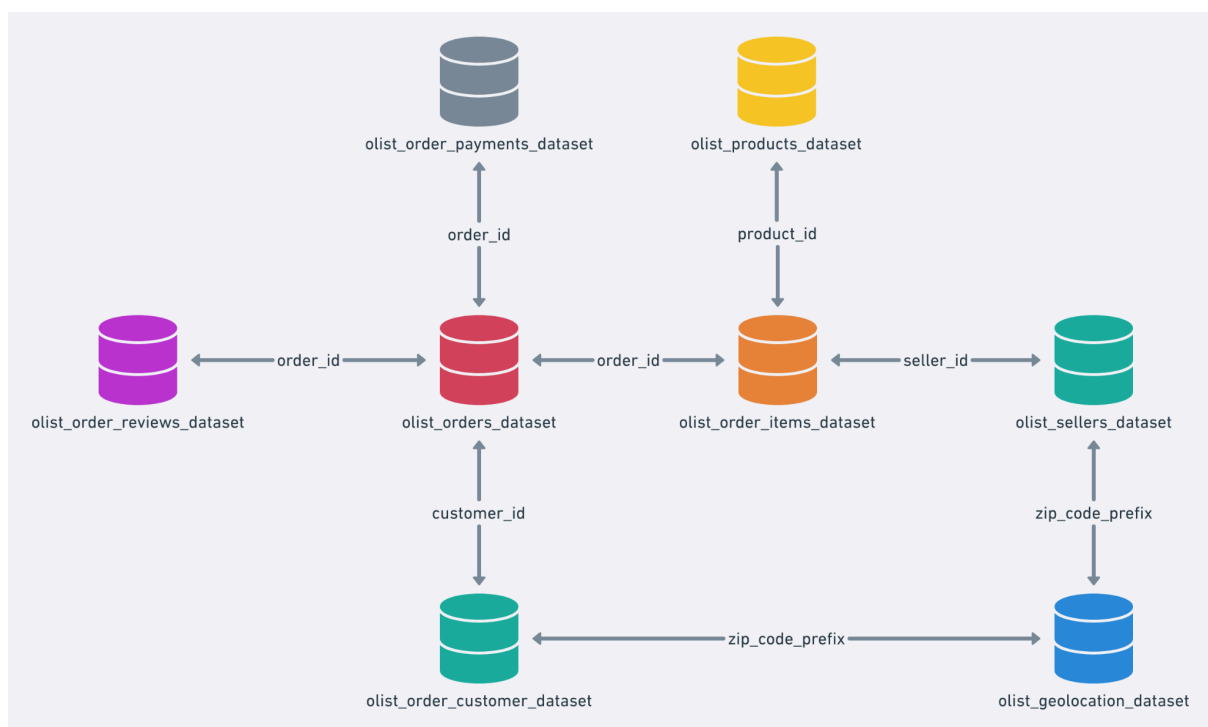
Lucas Szavara Nusp 12690087

1. Introdução

Neste projeto foram aplicadas técnicas de ciência de dados para analisar um conjunto de dados pré estabelecido, com a finalidade de responder questões apresentadas pela professora e pelos demais alunos, além de explorar os principais fatores que contribuem para uma avaliação positiva.

2. Os dados

O conjunto de dados trabalhado apresenta informações sobre 100 mil pedidos realizados na loja on-line Olist entre os anos de 2016 e 2018. Existem ao todo 9 tabelas com dados variando de tipo de item do pedido, satisfação do cliente, localização do destinatário, forma de pagamento, entre outros. As tabelas se relacionam da seguinte forma:



A tabela `product_category_name_translation` não se relaciona com nenhuma outra tabela.

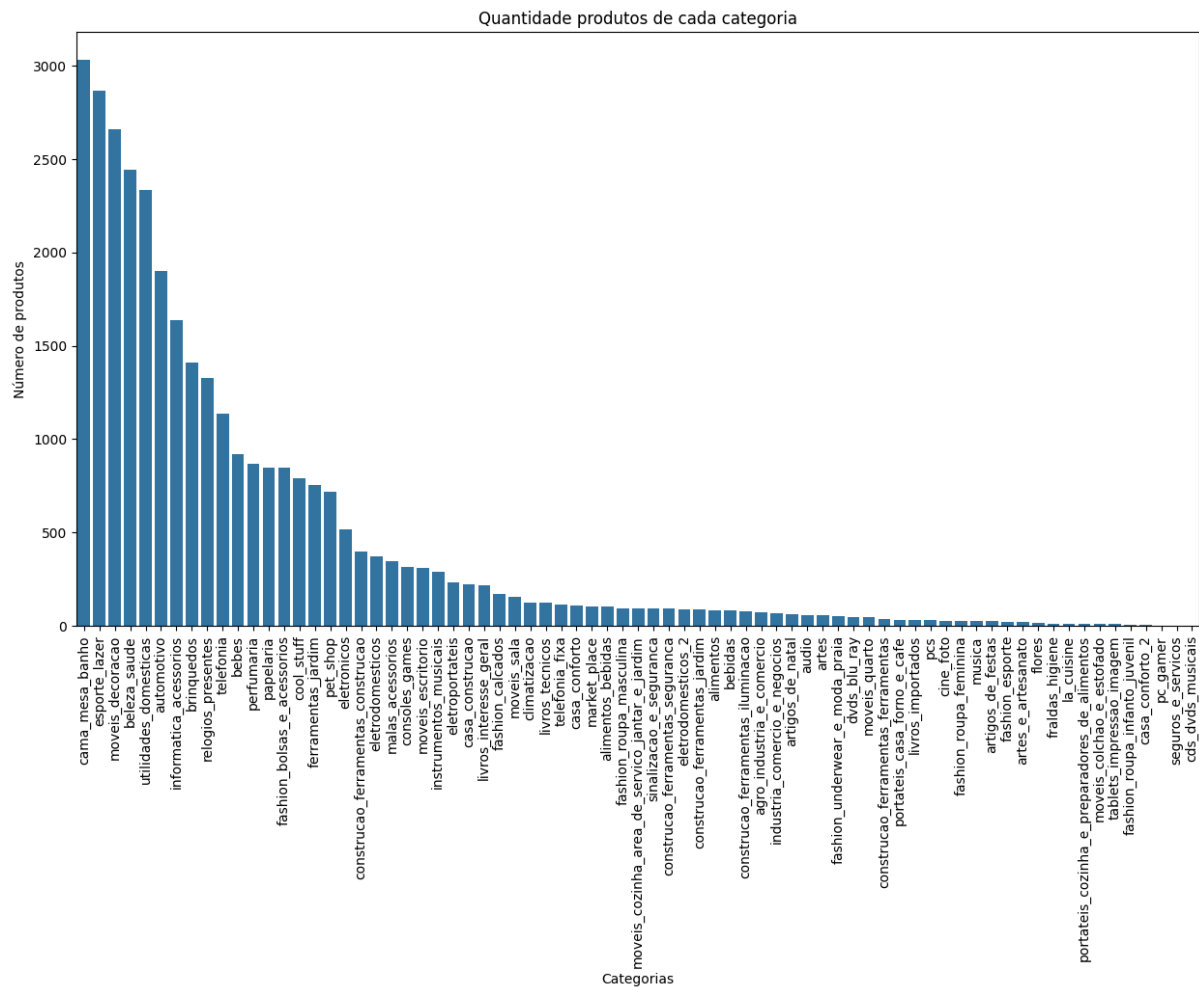
3. Questões propostas

3.1. Quantas categorias de produtos foram comercializadas e quais os atributos mais relevantes?

Ao todo a base de dados conta com 73 diferentes categorias de produtos. As categorias de produtos mais relevantes serão discutidas na seção de técnicas de machine learning.

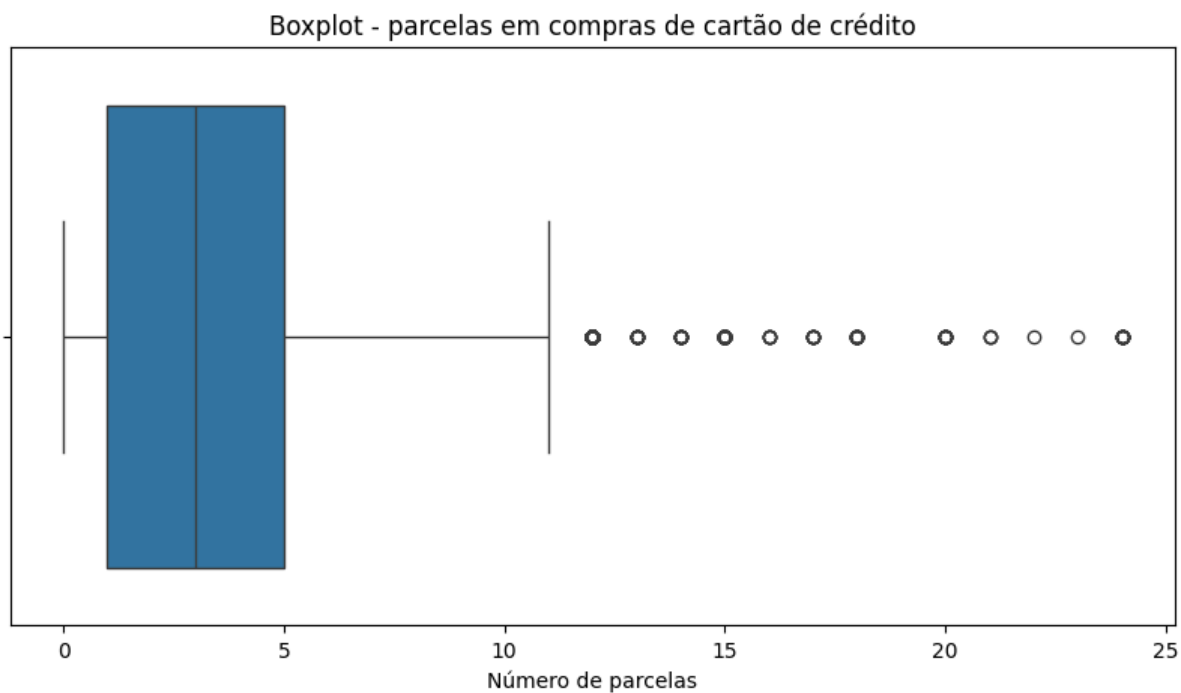
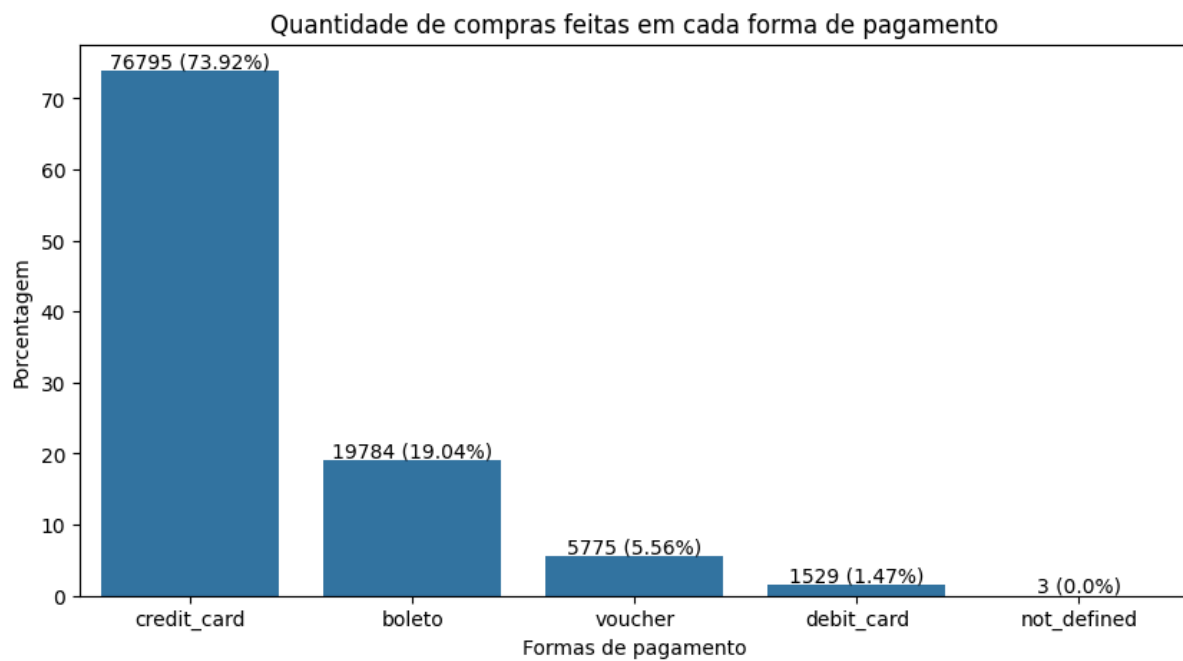
3.2. Qual a categoria de produtos mais vendida?

A categoria mais vendida é a “cama, mesa e banho”, com 11.115 produtos listados na base de dados.

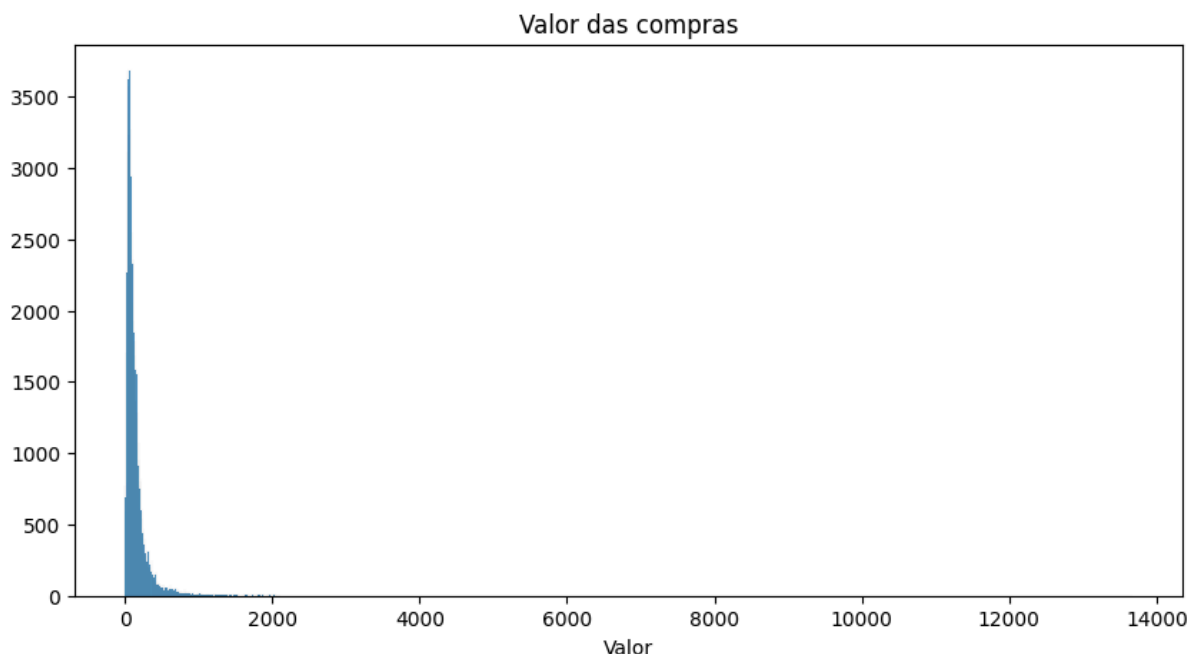


3.3. Qual a porcentagem de compras feitas com cartão de crédito? Neste caso, qual é o número médio de parcelas?

73,2% das compras foram feitas com cartão de crédito, com um número médio de 3,5 parcelas.

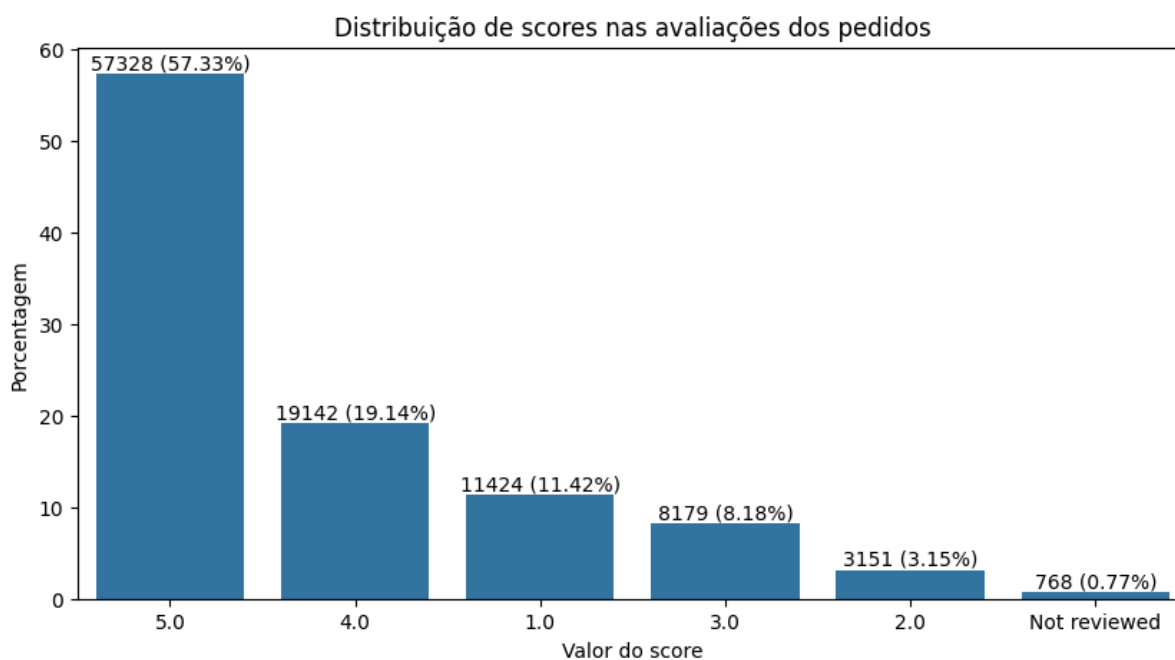


- 3.4. O valor das compras fica em média em torno de qual valor em reais?**
O valor médio das compras é de 154,10 reais.



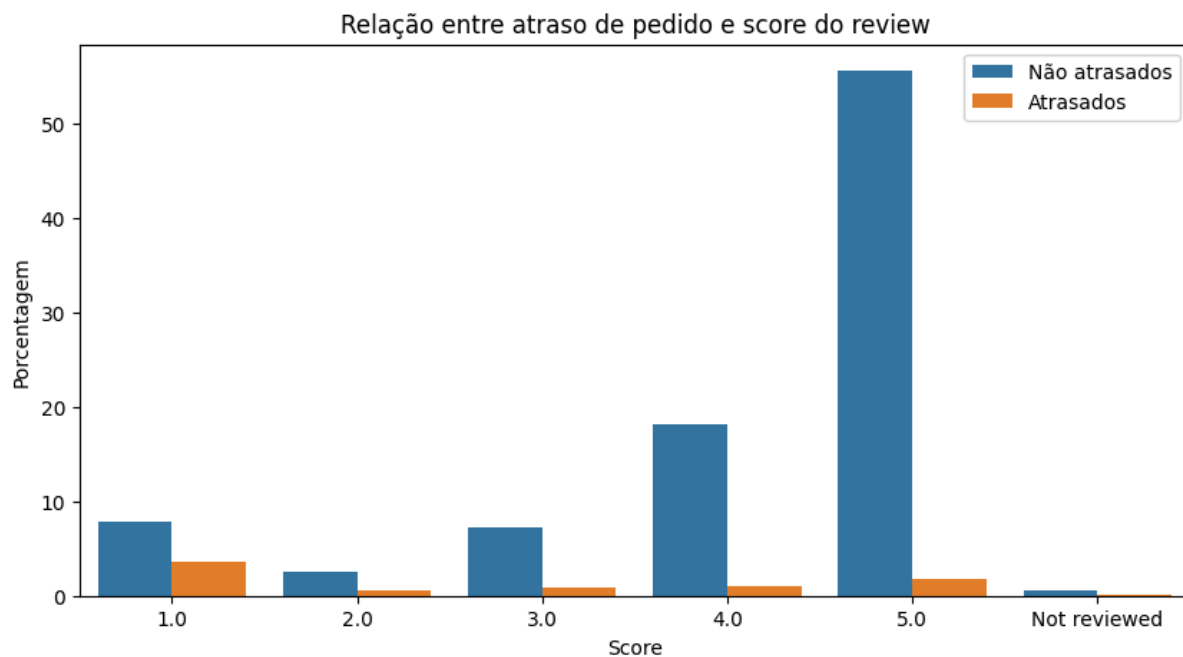
3.5. Qual a porcentagem de clientes satisfeitos com a realização da compra? (review score = 4 ou 5)

76,47% dos clientes classificaram seus pedidos como 4 ou 5 estrelas.



3.6. Existe relação entre o tempo de entrega e avaliação feita pelo cliente?

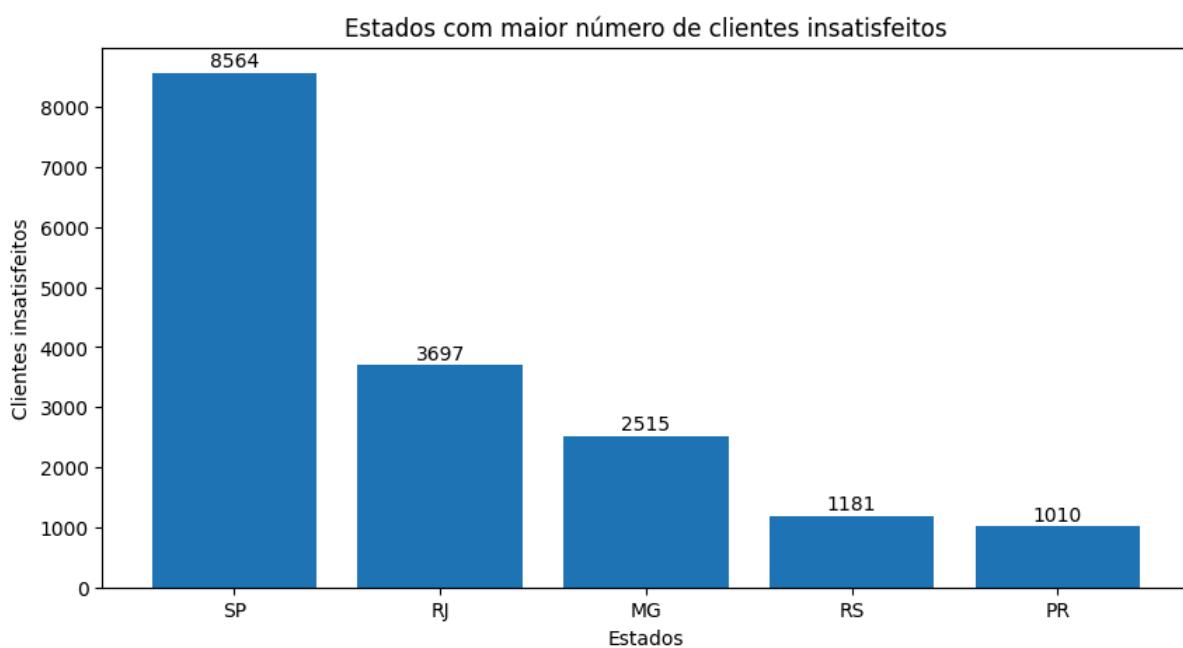
É possível perceber uma relação entre o pedido chegar ou não atrasado e a avaliação do cliente. Das avaliações de nota 1, uma porcentagem muito grande são pedidos que chegaram atrasados, enquanto avaliações de notas maiores possuem uma proporção menor de pedidos com atraso.



3.7. Quais estados do Brasil possuem mais clientes insatisfeitos?

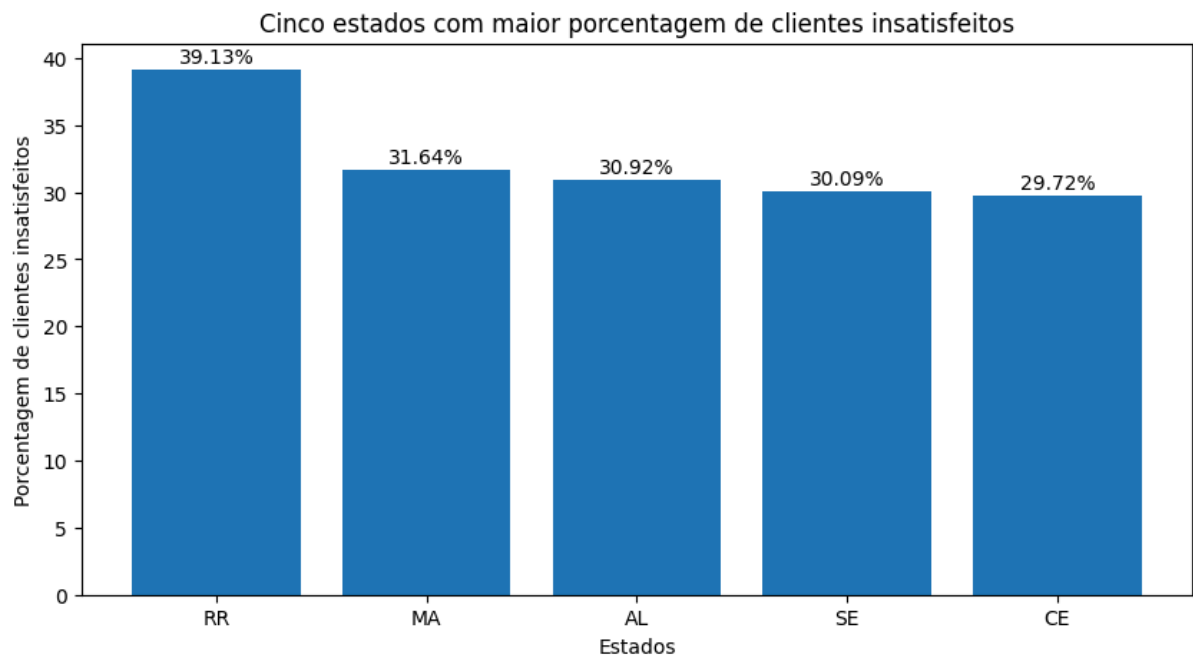
Em número absoluto, os estados com maior número de clientes insatisfeitos (review score menor que 4) são:

- São Paulo - 8.564
- Rio de Janeiro - 3.697
- Minas Gerais - 2.515
- Rio Grande do Sul - 1.181
- Paraná - 1.010



Quando levamos em consideração a quantidade de pedidos de cada estado, os estados com maior proporção de clientes insatisfeitos são:

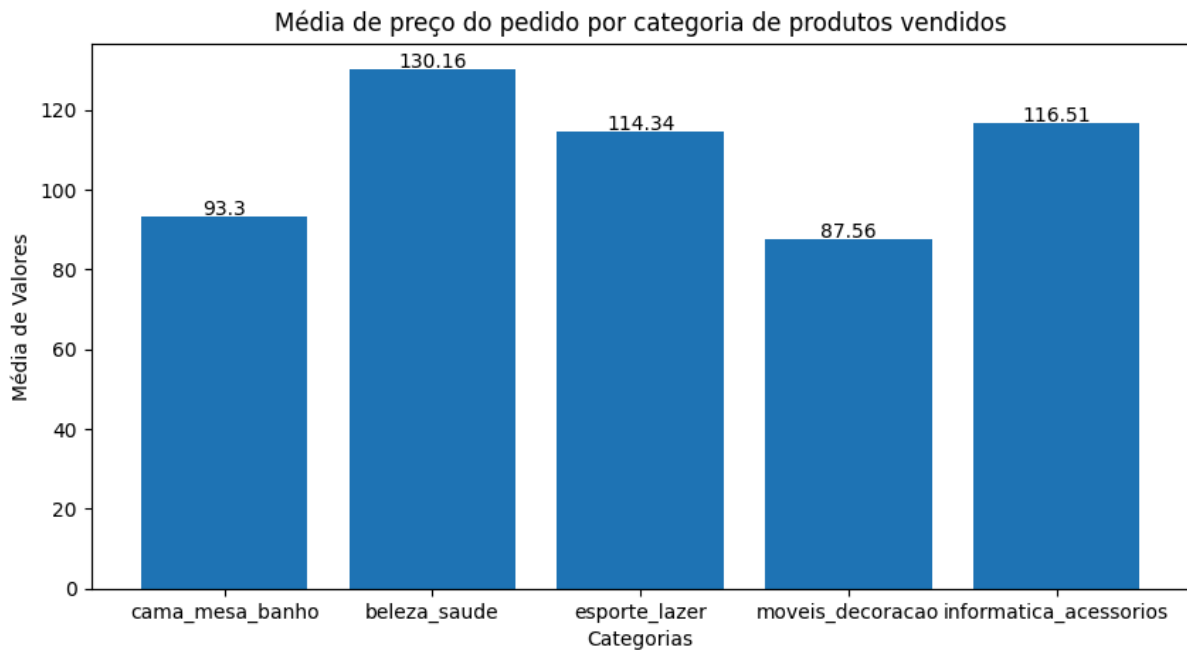
- Roraima - 39,13%
- Maranhão - 31,63%
- Alagoas - 30,91%
- Sergipe - 30,08%
- Ceará - 29,72%



3.8. Média do preço do pedido por categoria de produtos vendidos

A média de valor de cada pedido por categoria de produto, considerando as cinco categorias mais vendidas, é:

- Cama, mesa e banho - 93,3 reais
- Beleza e saúde - 130,16 reais
- Esporte e lazer - 114,34 reais
- Móveis e decoração - 87,56 reais
- Informática e acessórios - 116,51 reais



4. Técnicas de Machine Learning

Essa seção do relatório é dedicada às técnicas de machine learning desenvolvidas durante o projeto. As técnicas têm o objetivo de prever a nota de um dado pedido de acordo com as covariáveis estabelecidas, sendo elas:

- Número de pagamentos utilizados
- Número médio das parcelas do pedido
- Total do pagamento
- Indicador de atraso no pedido
- Tempo de entrega
- Interação entre tempo de entrega e Indicador de atraso
- Dias de atraso (0 caso o pedido tenha sido entregue na data correta)
- Valor mensal médio a ser pago
- Indicador de pedido cancelado
- Total de produtos distintos no pedido
- Total de vendedores distintos no pedido
- Porcentagem de produtos vendidos por vendedores do mesmo estado que o cliente
- Comprimento médio do nome dos produtos
- Comprimento médio da descrição dos produtos
- Número médio de fotos dos produtos
- Peso médio dos produtos
- Comprimento médio dos produtos
- Altura médio dos produtos
- Largura médio dos produtos
- Preço total dos produtos
- Frete total dos produtos
- Porcentagem dos pagamentos feitos com cada tipo de pagamento
- Porcentagem das categorias dos produtos
- Estado do cliente

Os dados foram separados em 90% para treino e 10% para teste. Foram utilizadas quatro diferentes técnicas para predição, e os resultados obtidos estão listados a seguir.

4.1. Naive Bayes

O classificador de Bayes ingênuo, por não ser muito robusto para problemas complexos, atingiu uma acurácia de apenas 57%.

4.2. KNN

Utilizando o Grid Search CV, o melhor valor para K numa faixa de 3 a 10 é 9. Com esse valor, a acurácia do modelo é 55%

4.3. Random Forest

Os melhores valores dos parâmetros encontrados pelo Grid Search CV foram:

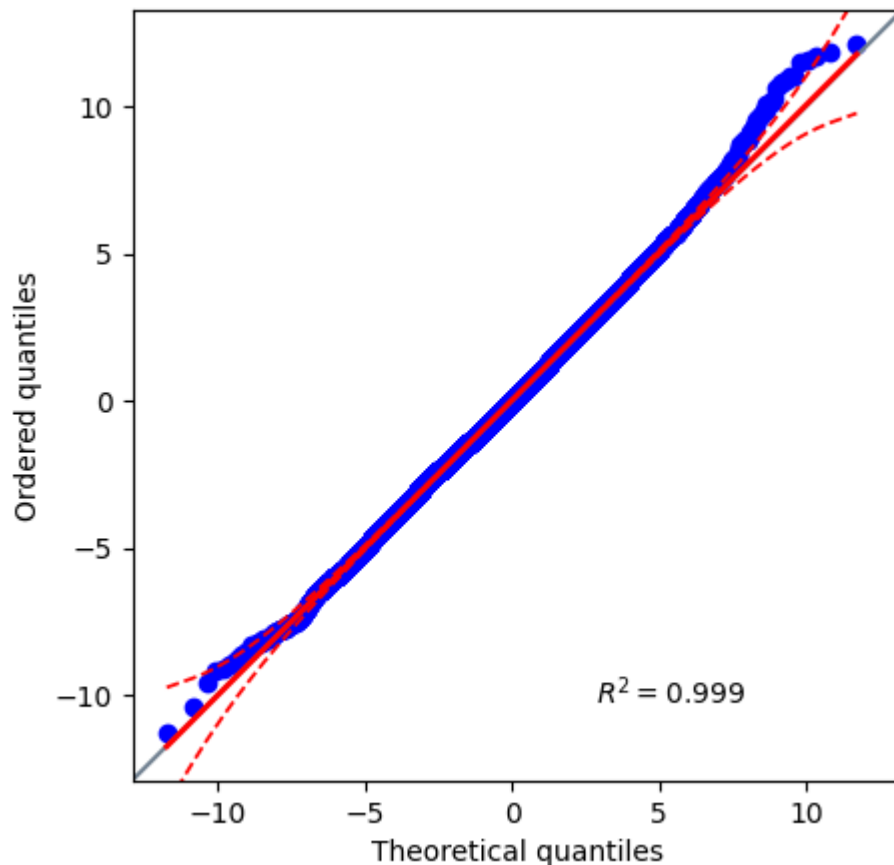
- Profundidade máxima: 10 (de 1 a 11)
- Mínimo de amostras no nó-folha: 1 (de 1 a 5)
- Mínimo de amostras para dividir um nó: 2 (de 2 a 5)
- Número de árvores: 150 (de 150 a 200)

O valor da acurácia para esses parâmetros é de 62%.

4.4. Regressão Ordinal por variável latente

Foi testado usando variáveis das distribuições logística, normal e gumbel (gumbel_r no scipy). Todas tiveram resultados muito similares, pela facilidade de interpretação, foi preferível a logística.

A verificação desses modelos foi feita em duas etapas. Primeiramente analisamos os resíduos substitutos (Liu e Zhang (2018)), usando o gráfico quantil-quantil e um gráfico de pontos para algumas covariáveis:



Apesar de apresentar algumas variações do esperado, essas divergências são pequenas ou explicadas pela quantidade de observações com determinados valores nas covariáveis, isto é, é esperado uma maior presença de resíduos outliers nos valores mais comuns de uma covariável. Após essa etapa, calculamos as métricas de predição do modelo, obtendo:

- Acurácia de 61,09%
- MSE de 1.35

Dessa forma temos um modelo com resultados muito próximos do melhor modelo preditivo observado, mas com uma verificação de ajuste mais completa e com maior interpretabilidade dos parâmetros, permitindo identificar as covariáveis com maior peso no modelo, desconsiderando as covariáveis com valor P abaixo de 0,05:

Covariável	Coeficiente	RC	Valor P
Porcentagem de produtos da categoria PC Gamer comparado com Utilidades domésticas	-1,4421	0,2364	0,043
Porcentagem de produtos da categoria Portáteis cozinha e	-1,4034	0,2458	0,005

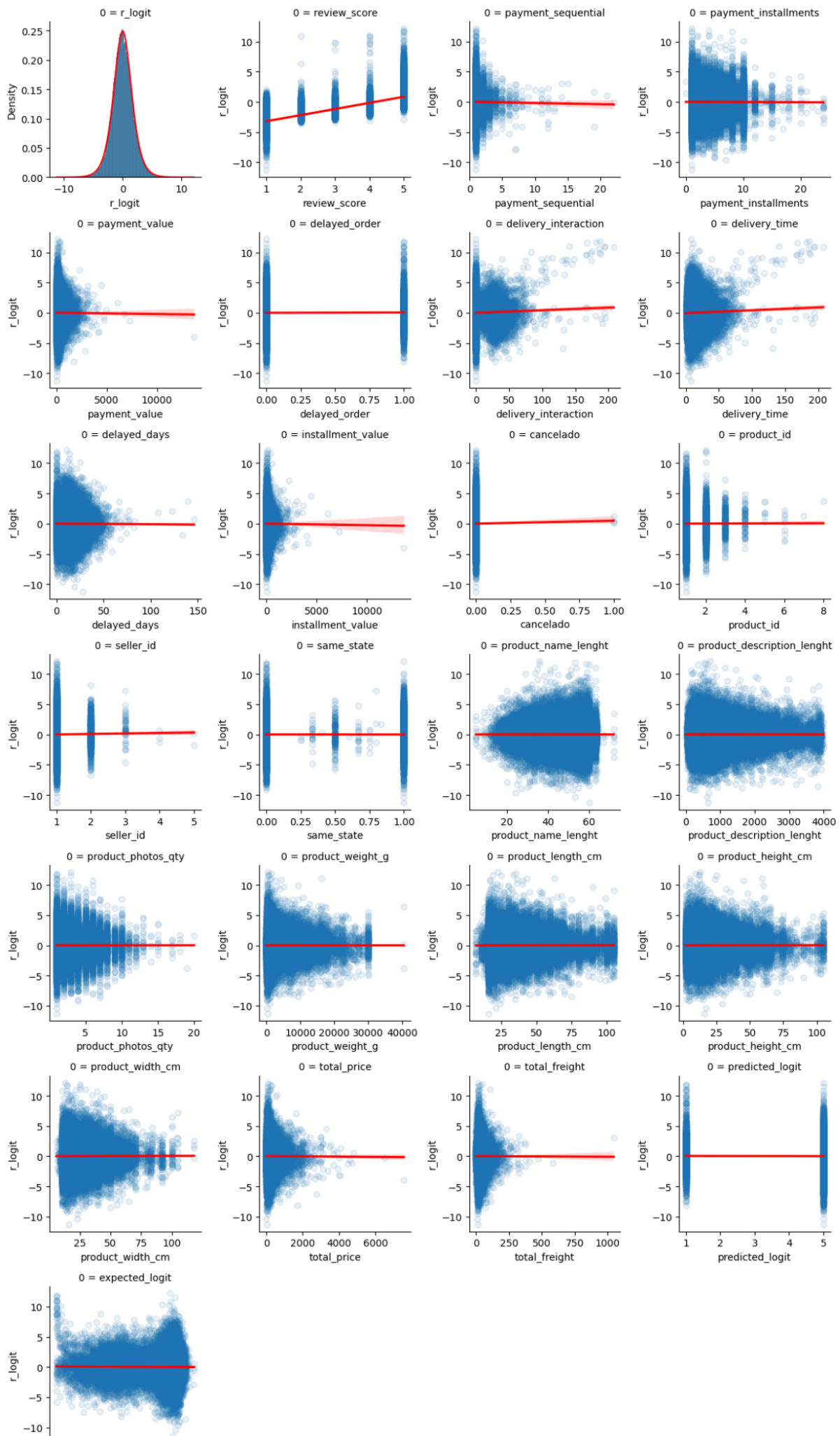
preparadores de alimento comparado com Utilidades domésticas			
Número de vendedores	-1,3501	0,2592	<0,001
Pedido atrasado	-1,049	0,3502	<0,001
Cliente mora no Amazonas, comparado com clientes do Tocantins	0,7593	2,1368	<0,001

Em que RC representa a razão de chances da avaliação do cliente ser maior que determinado valor, por exemplo, dado um pedido i , de um cliente no Tocantins, com chance igual a c de ser avaliado com pelo menos a nota j . O mesmo pedido, para o Amazonas, considerando todas as outras covariáveis constantes, terá chance 2,1368 maior de ser avaliado com pelo menos a nota j . Matematicamente, seja X_i as covariáveis de um pedido para um cliente no Tocantins e X_m um pedido similar, mas para ser entregue no Amazonas. Então:

$$\frac{P(Y \geq j | X_m)}{P(Y < j | X_m)} = 2,1368 * \frac{P(Y \geq j | X_i)}{P(Y < j | X_i)}$$

Em outras palavras, se o pedido no Tocantins tem 2 avaliações pelo menos iguais a 3 para cada 1 avaliação abaixo de 3, o mesmo pedido no Amazonas terá 2,1368 * 2 avaliações iguais ou melhores que 3 para cada avaliação pior do que 3.

Considerando todos os fatores dentro do controle do vendedor, uma estratégia que pode aumentar suas avaliações seria realizar entregas mais rapidamente e com previsões mais assertivas, de produtos com menor valor. Além disso, pedidos pagos com maior porcentagem de voucher costumam ter avaliações piores que outros métodos de pagamento, enquanto pedidos pagos com cartão de débito e boleto tem avaliações melhores, é difícil entretanto afirmar que essa relação é causal e pode ter relação com os motivos pelos quais os clientes escolhem cada método, por exemplo, clientes com voucher já estão frustrados com a loja, enquanto clientes que pagam com boleto ou cartão de débito podem estar comprando produtos que impactam menos na economia doméstica, logo, se sentindo menos frustrados com defeitos ou outros problemas.



5. Referências

Liu, D., & Zhang, H. (2018). Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. *Journal of the American Statistical Association*, 113(522), 845–854. <https://doi.org/10.1080/01621459.2017.1292915>