eEDB-002/2024-3

Análise Preditiva e Clusterização de Avaliações de Clientes para Otimização do Atendimento

Uma Abordagem Baseada análise de sentimento

Grupo 02:
Aline Bini
Ana Lívia Franco
Ana Pris
João Squinelato
Marcelo Pena
Thais Carvalho

Índice:

Introdução

Em um mercado cada vez mais competitivo, o atendimento ao cliente deixou de ser apenas um diferencial para se tornar uma verdadeira exigência. Técnicas avançadas, ferramentas de suporte e novas abordagens foram desenvolvidas para reduzir o tempo de resposta e aumentar a satisfação dos consumidores.
Entretanto, mesmo com todas as inovações que permitem a automatização de processos, muitos desafios no atendimento ao cliente ainda precisam ser enfrentados pelas empresas. Neste trabalho, exploraremos os principais desafios no atendimento ao cliente, com foco na redução do tempo de espera para resolução de tickets.
O **objetivo** é diminuir o tempo em fila de tickets para o atendimento, uma vez que dado o grande volume de atendimentos diários, seria humanamente impossível filtrá-los em tempo hábil. A importância está justamente na redução de custos, por meio da otimização do atendimento, como também na redução da taxa de churn em virtude de um atendimento rápido e eficiente.

Fonte de Dados

Neste trabalho, utilizamos o conjunto de dados de revisões de pedidos da empresa Olist, disponível publicamente no kaggle (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce). A Olist é uma plataforma brasileira de marketplace que conecta vendedores a diversos canais de vendas. O dataset foi escolhido por fornecer uma grande quantidade de dados de avaliações de clientes, contendo informações detalhadas sobre a experiência de compra, como a pontuação de review, comentários e datas de criação e resposta das revisões.

A escolha deste conjunto de dados se justifica pelo seu volume e diversidade, o que permite a criação de modelos robustos para análise de sentimentos e clusterização de tickets. Além disso, a natureza dos dados é ideal para modelar desafios do atendimento ao cliente, como otimização de tempo de resposta e identificação de padrões de satisfação e insatisfação. Essas características tornam o dataset adequado para testar e validar técnicas de processamento de linguagem natural (NLP), análise de sentimento e aprendizado não supervisionado, que são fundamentais para a proposta de otimização do atendimento ao cliente.

Camadas de Processamento

Camada raw

rwzd_olist_order_reviews

Armazenamento dos dados brutos do arquivo olist_order_reviews_dataset.csv.

- Colunas:
  - 'review_id'
  - 'order_id'
  - 'review_score'
  - 'review_comment_title'
  - 'review_comment_message'
  - 'review_creation_date'
  - 'review_answer_timestamp'
- Quantidade de registros: 100000

Camada trusted

trzd_olist_order_reviews

- Transformações baseadas na camada *Raw* para:
- Filtrar avaliações não nulas
- Removendo avaliações que citam palavras referentes a entregas, uma vez que o foco é avaliação dos produtos
- Quantidade de registros 20641

Camada delivery

dlzd_olist_order_reviews

Transformações baseadas na camada *Trusted* para:

- Obter tokens a partir dos textos de review_comment_title e *review_comment_message*
- Gerar nova coluna *review_comment_title_and_message* com os tokens do título e comentário de uma mesma avaliação
- Criação da coluna *review_sentiment*, onde:

-1 é atribuído para avaliações com *review_score* menores que 3

0 para avaliações com *review_score* igual a 3

+1 para avaliações com *review_score* maior que 3

Padronizar tokens:

- Lower case
- Remover palavras comuns e conectores
- Desconsiderar de palavras com 1 caractere e mais de 15 caracteres

- Desconsiderar de espaços em branco, números e caracteres especiais
- Quantidade de registros: 20641

Utilização do pacote *simple_preprocess* da lib gensim:
https://tedboy.github.io/nlps/generated/generated/gensim.utils.simple_preprocess.html

dlzd_olist_order_reviews_clean
Transformações baseadas na tabela *dlzd_olist_order_reviews* para:
Remoção de stopwords presentes nos tokens das colunas:

- review_comment_title
- review_comment_message
- review_comment_title_and_message
- Quantidade de registros: 20641

dlzd_olist_order_reviews_training
Randomicamente e estratificado pra manter a proporção 80(treino):20
Amostra da tabela *dlzd_olist_order_reviews* para uso no treinamento dos modelo:
Colunas filtradas: review_sentiment e review_comment_title_and_message
Quantidade de registros: 16743

dlzd_olist_order_reviews_test
Amostra da tabela *dlzd_olist_order_reviews* para uso no treinamento dos modelo:
Colunas filtradas: *review_sentiment* e *review_comment_title_and_message*
Quantidade de registros: 4186

Análise Exploratória sobre os Tokens

Análises sobre os tokens - notebook token_analysis
Análises realizadas sobre os dados da camada Delivery, desconsiderando as stopwords para:

- Total de palavras (incluindo suas repetições): 100550
- Total de palavras únicas: 9584
- Diversidade lexical: 0.0953 (pouca variação lexical e muitas repetições)

Notamos que até 3 repetições ocorrem erros de digitação e palavras de contexto muito específico, nesse sentido:
Total de palavras que aparecem até 3 vezes: 7107
Total de palavras que ocorrem mais de 3 vezes: 2477

25 palavras mais frequentes x repetições (desconsiderando stop words):

```
[('produto', 891),
 ('qualidade', 151),
 ('recomendo', 145),
 ('comprei', 128),
 ('gostei', 117),
 ('achei', 100),
 ('compra', 98),
 ('loja', 80),
 ('frete', 76),
 ('problema', 69),
 ('foto', 66),
 ('correios', 65),
 ('pedido', 57),
 ('material', 57),
 ('produtos', 57),
 ('caixa', 56),
 ('relogio', 55),
 ('embalagem', 54),
 ('melhor', 53),
 ('sao', 52),
 ('esperava', 52),
 ('site', 51),
 ('cor', 50),
 ('tamanho', 48),
 ('preco', 46)]
```

25 palavras mais frequentes e avaliações negativas:

```
[('produto', 891),
 ('qualidade', 151),
 ('recomendo', 145),
 ('comprei', 128),
 ('gostei', 117),
 ('achei', 100),
 ('compra', 98),
 ('loja', 80),
 ('frete', 76),
 ('problema', 69),
 ('foto', 66),
 ('correios', 65),
 ('pedido', 57),
 ('material', 57),
 ('produtos', 57),
 ('caixa', 56),
 ('relogio', 55),
 ('embalagem', 54),
 ('melhor', 53),
 ('sao', 52),
 ('esperava', 52),
 ('site', 51),
 ('cor', 50),
 ('tamanho', 48),
 ('preco', 46)]
```

25 palavras mais frequentes em avaliações neutras:

```
[('produto', 722),
 ('nao', 705),
 ('qualidade', 141),
 ('recomendo', 131),
 ('porem', 128),
 ('gostei', 104),
 ('comprei', 96),
 ('achei', 84),
 ('compra', 81),
 ('foto', 62),
 ('loja', 60),
 ('material', 55),
 ('ja', 54),
 ('melhor', 52),
 ('esperava', 52),
 ('problema', 50),
 ('relogio', 49),
 ('cor', 47),
 ('caixa', 46),
 ('embalagem', 45),
 ('sao', 45),
 ('tamanho', 45),
 ('site', 44),
 ('otimo', 44),
 ('defeito', 41)]
```

25 palavras mais frequentes para avaliações positivas:

```
[('produto', 3708),
 ('recomendo', 2805),
 ('otimo', 2136),
 ('excelente', 1309),
 ('nao', 993),
 ('qualidade', 965),
 ('gostei', 927),
 ('super', 913),
 ('otima', 747),
 ('loja', 734),
 ('compra', 533),
 ('adorei', 436),
 ('perfeito', 397),
 ('comprar', 396),
 ('parabens', 385),
 ('amei', 374),
 ('lindo', 367),
 ('atendimento', 353),
 ('rapido', 343),
 ('certo', 333),
 ('satisfeito', 330),
 ('lannister', 311),
 ('satisfeita', 276),
 ('bonito', 265),
 ('conforme', 262)]
```

25 palavras mais frequentes aos pares, independente da ordem (dataset completo - sem stop words):

```
[(('produto', 'qualidade'), 1179),
 (('produto', 'recomendo'), 1117),
 (('otimo', 'produto'), 865),
 (('recomendo', 'super'), 692),
 (('excelente', 'produto'), 586),
 (('loja', 'produto'), 579),
 (('comprei', 'produto'), 560),
 (('gostei', 'produto'), 526),
 (('compra', 'produto'), 440),
 (('otimo', 'recomendo'), 357),
 (('loja', 'recomendo'), 320),
 (('produto', 'site'), 305),
 (('otima', 'qualidade'), 295),
 (('defeito', 'produto'), 294),
 (('qualidade', 'recomendo'), 288),
 (('produto', 'super'), 278),
 (('diferente', 'produto'), 271),
 (('errado', 'produto'), 263),
 (('foto', 'produto'), 256),
 (('otima', 'produto'), 251)]
```

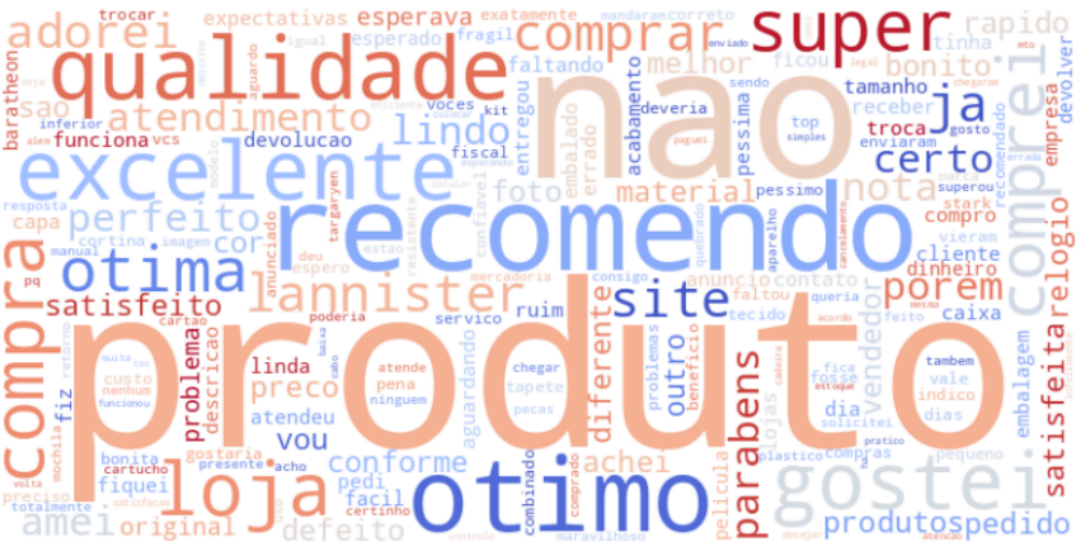
25 bigramas mais frequentes (duas palavras consecutivas - sem stop words - analise dos adjetivos):

```
('otimo', 'produto'): 639 vezes
('super', 'recomendo'): 562 vezes
('produto', 'nao'): 525 vezes
('recomendo', 'produto'): 463 vezes
('gostei', 'produto'): 289 vezes
('excelente', 'produto'): 279 vezes
('produto', 'qualidade'): 275 vezes
('otima', 'qualidade'): 251 vezes
('produto', 'excelente'): 246 vezes
('nao', 'recomendo'): 242 vezes
('recomendo', 'otimo'): 242 vezes
('produto', 'recomendo'): 235 vezes
('nao', 'gostei'): 230 vezes
('produto', 'produto'): 217 vezes
('produto', 'otimo'): 208 vezes
('recomendo', 'recomendo'): 162 vezes
('produto', 'defeito'): 155 vezes
('qualidade', 'produto'): 155 vezes
('recomendo', 'loja'): 153 vezes
('otimo', 'otimo'): 145 vezes
('produto', 'otima'): 141 vezes
('lojas', 'lannister'): 137 vezes
('nao', 'funciona'): 128 vezes
('pessima', 'qualidade'): 118 vezes
('nota', 'fiscal'): 114 vezes
```

25 trigramas mais frequentes:

```
('produto', 'otima', 'qualidade'): 110 vezes
('super', 'recomendo', 'produto'): 86 vezes
('recomendo', 'otimo', 'produto'): 85 vezes
('otimo', 'produto', 'recomendo'): 66 vezes
('produto', 'excelente', 'qualidade'): 58 vezes
('produto', 'super', 'recomendo'): 53 vezes
('nao', 'gostei', 'produto'): 51 vezes
('produto', 'pessima', 'qualidade'): 50 vezes
('super', 'recomendo', 'otimo'): 46 vezes
('recomendo', 'super', 'recomendo'): 44 vezes
('otimo', 'otimo', 'produto'): 44 vezes
('otimo', 'produto', 'excelente'): 40 vezes
('nao', 'recomendo', 'produto'): 39 vezes
('otimo', 'custo', 'beneficio'): 37 vezes
('produto', 'otimo', 'produto'): 37 vezes
('produto', 'atendeu', 'expectativas'): 34 vezes
('produto', 'nao', 'funciona'): 34 vezes
('produto', 'nao', 'original'): 33 vezes
('produto', 'recomendo', 'produto'): 33 vezes
('otima', 'qualidade', 'recomendo'): 32 vezes
('otimo', 'produto', 'produto'): 32 vezes
('excelente', 'produto', 'recomendo'): 32 vezes
('recomendo', 'produto', 'nao'): 31 vezes
('recomendo', 'loja', 'produto'): 30 vezes
('recomendo', 'produto', 'qualidade'): 29 vezes
```

Nuvem de palavras:

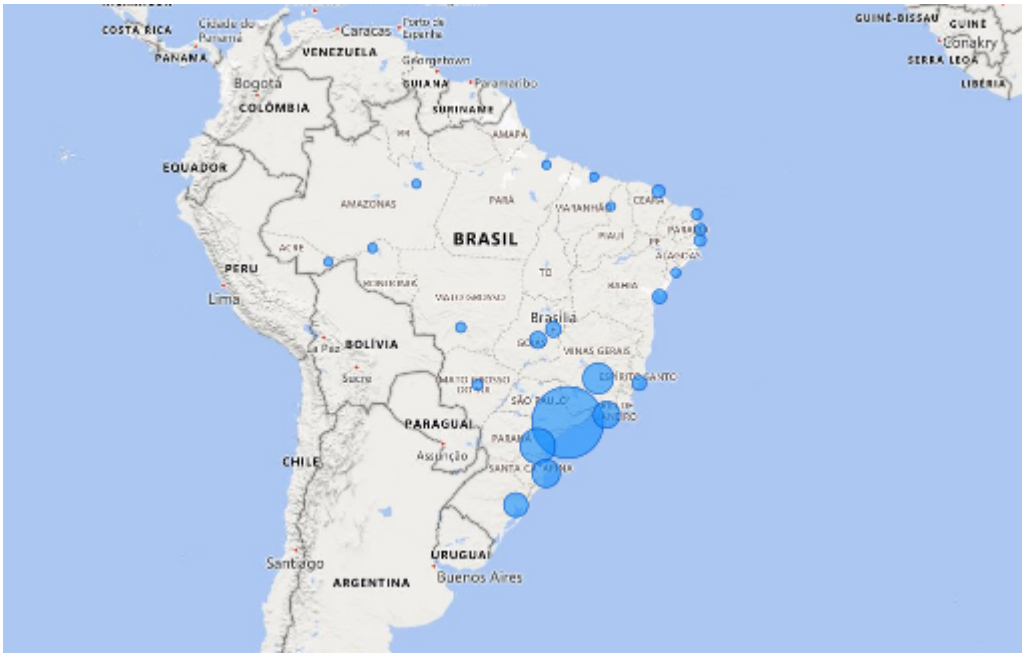


Nuvem de palavras avaliações negativas:

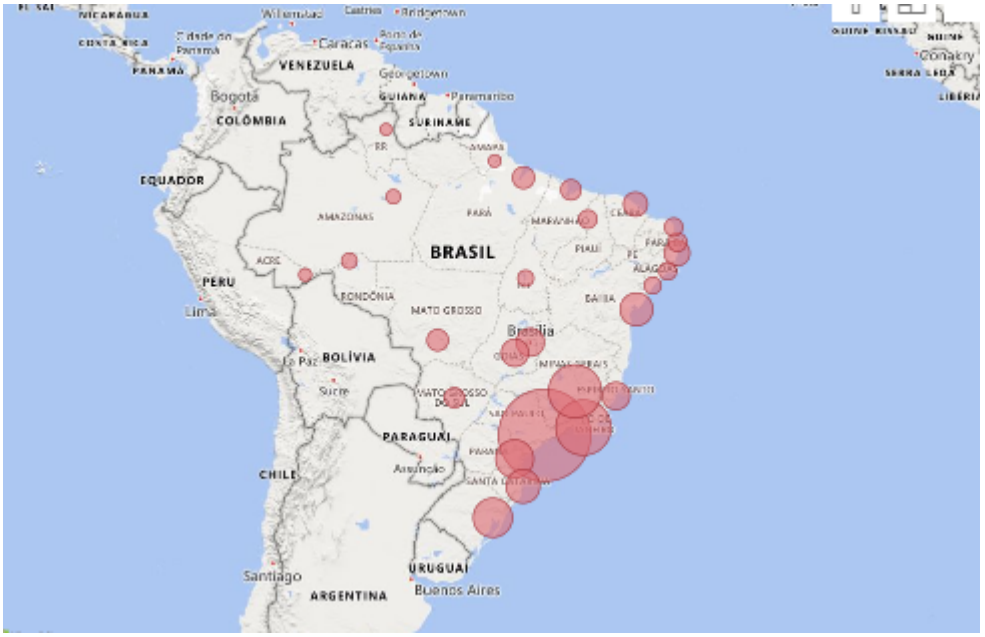
Nuvem de palavras para avaliações neutras:

Nuvem de palavras para avaliações positivas:

Mapa de vendedores:

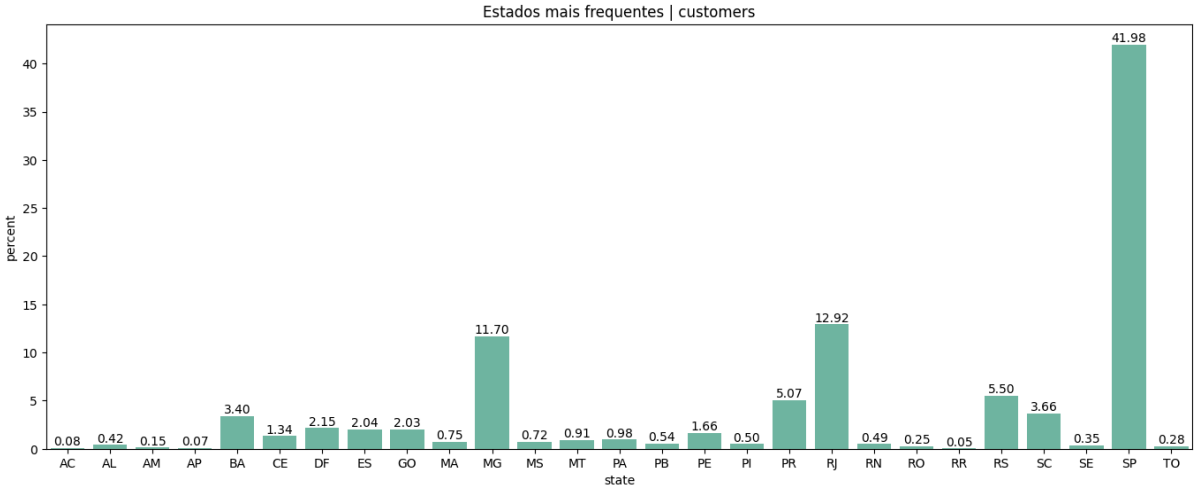


Mapa de parceiros:

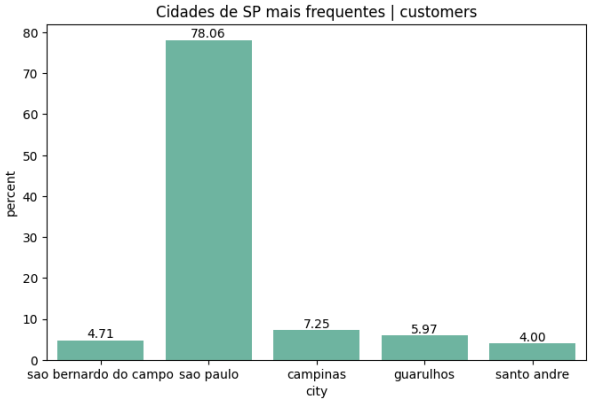


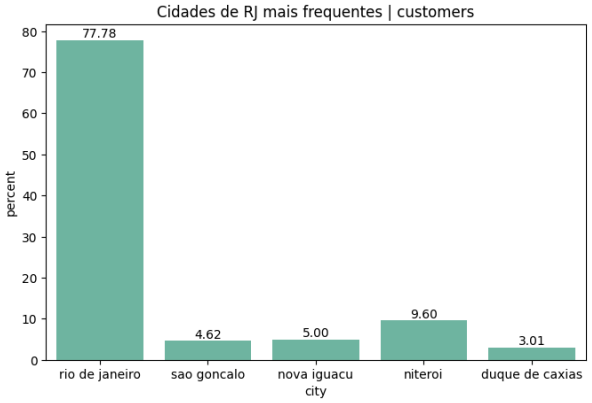
Análise Exploratória Univariada

Customers: estados de origem

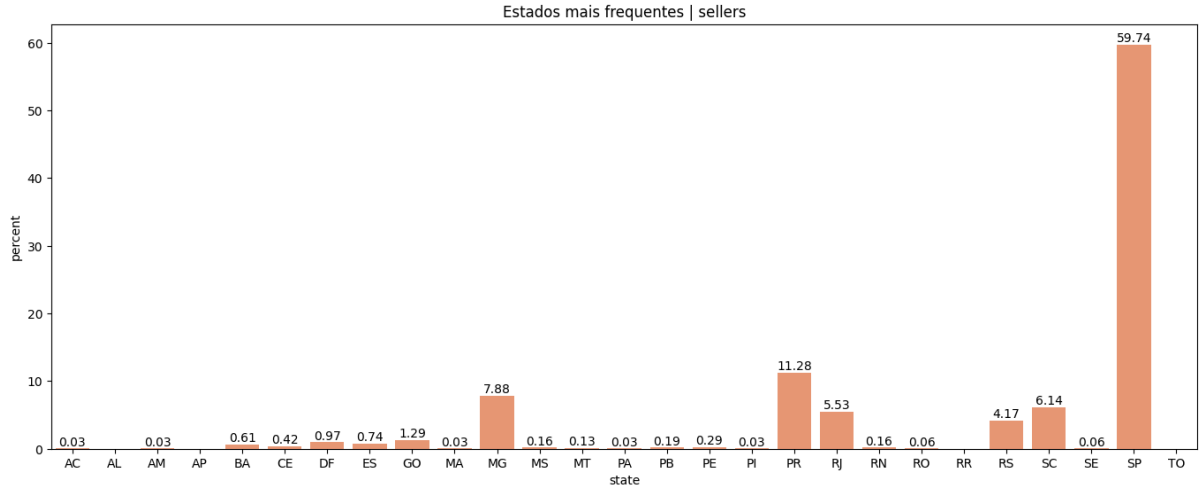


Customers: dentre os estados de origem mais frequentes, as cidades de origem

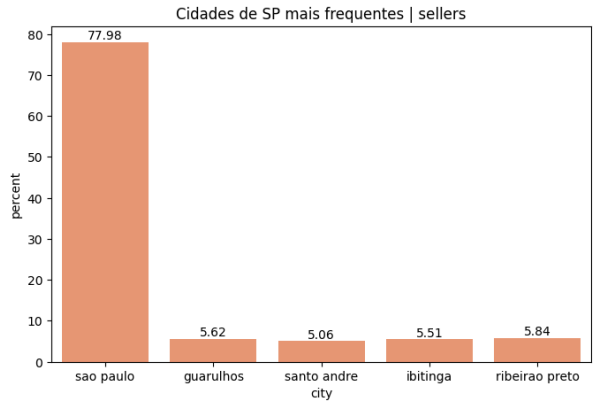


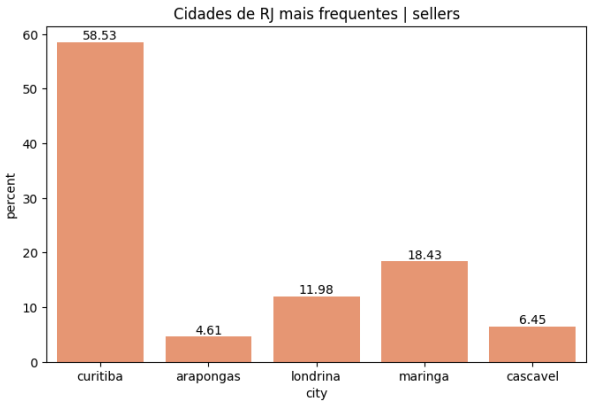


Sellers: estados de origem

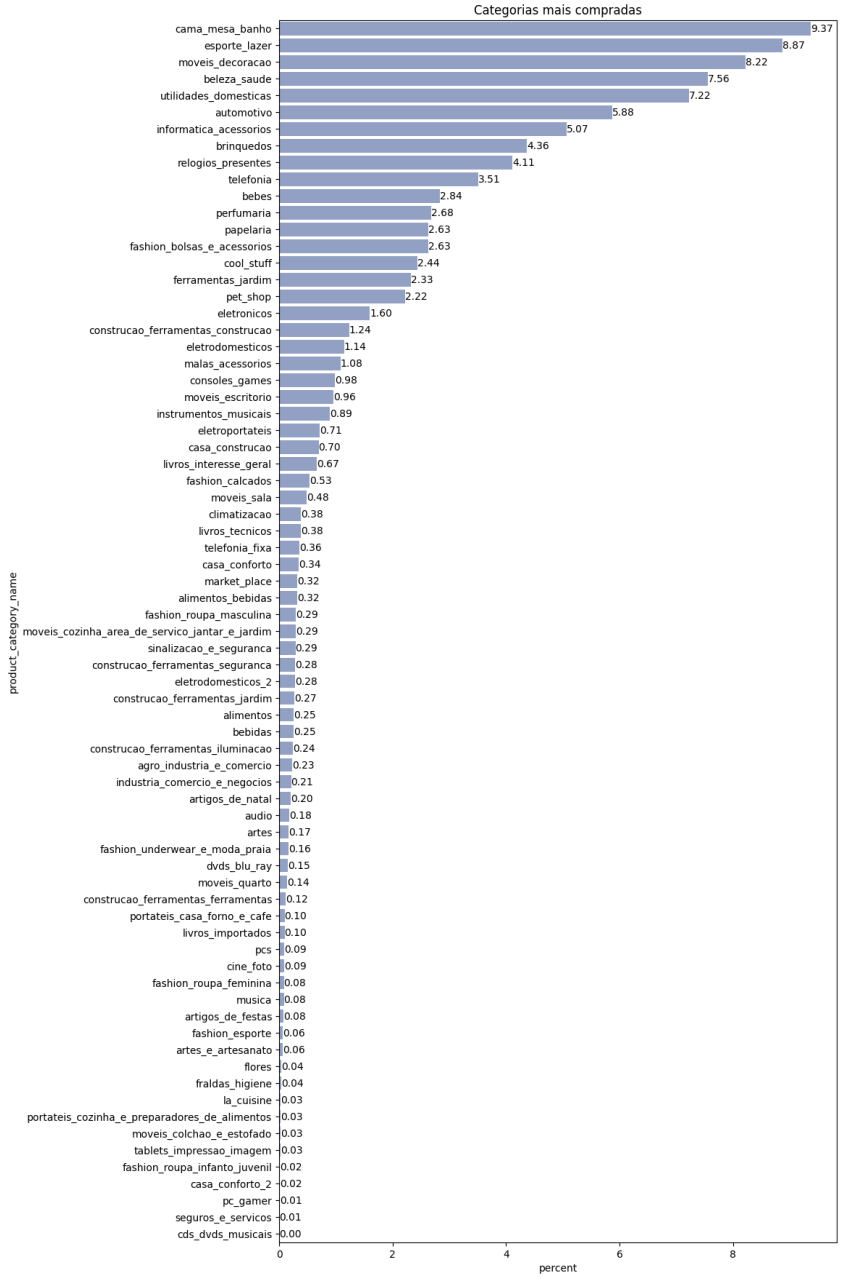


Sellers: dentre os estados de origem mais frequentes, as cidades de origem

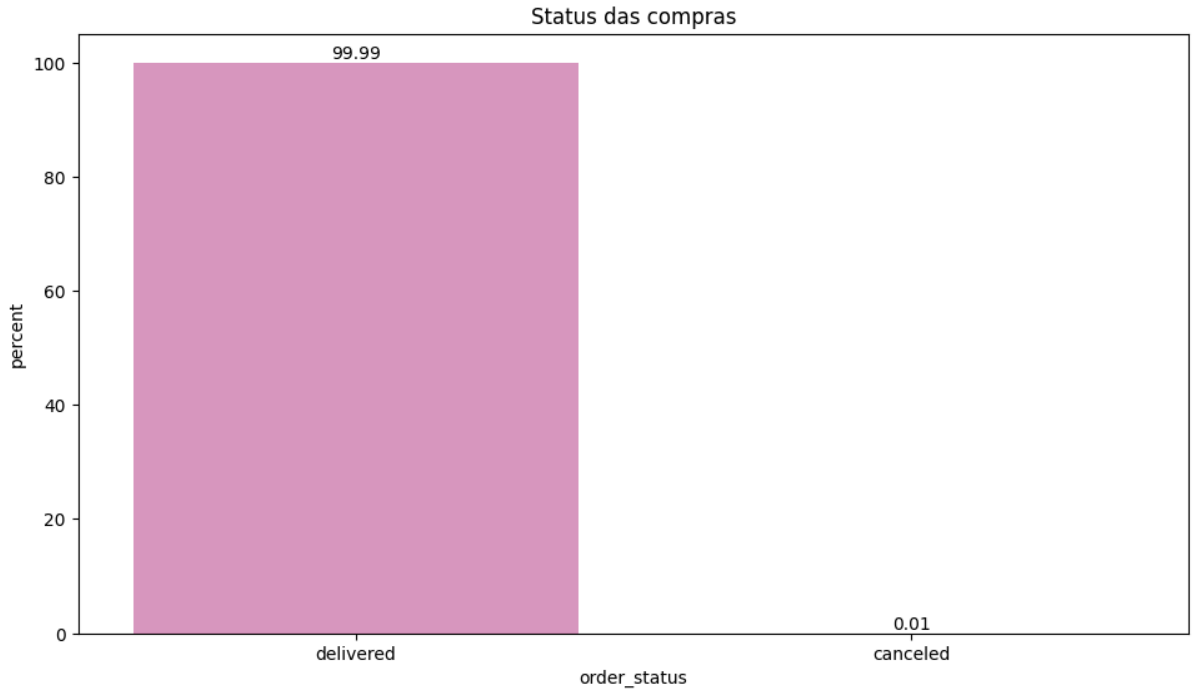




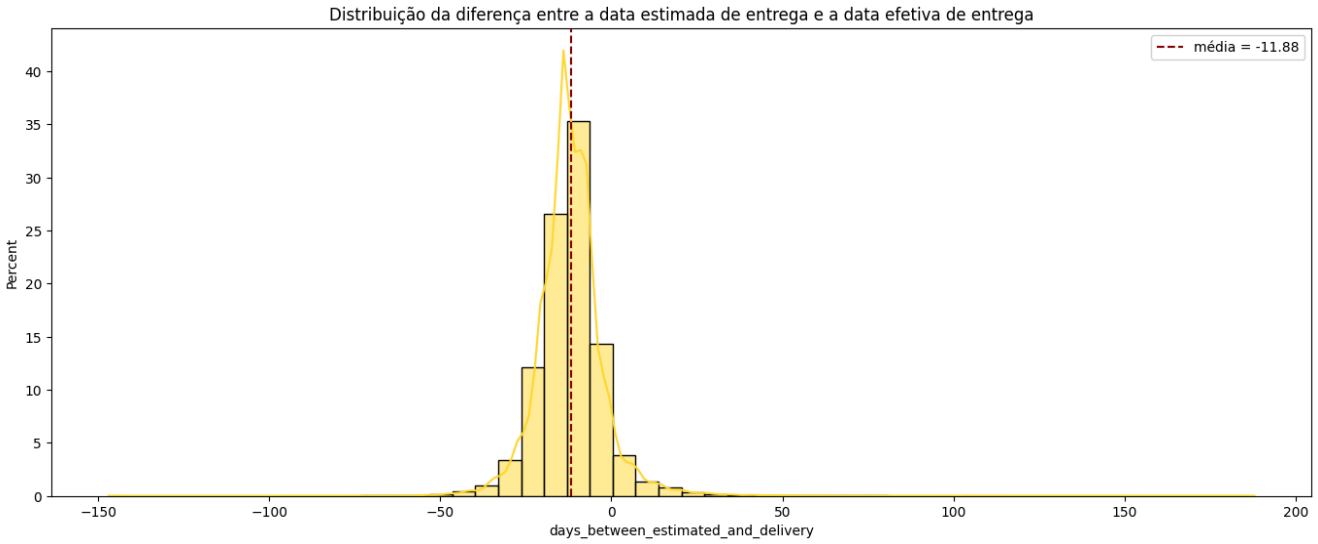
Products: share de cada categoria



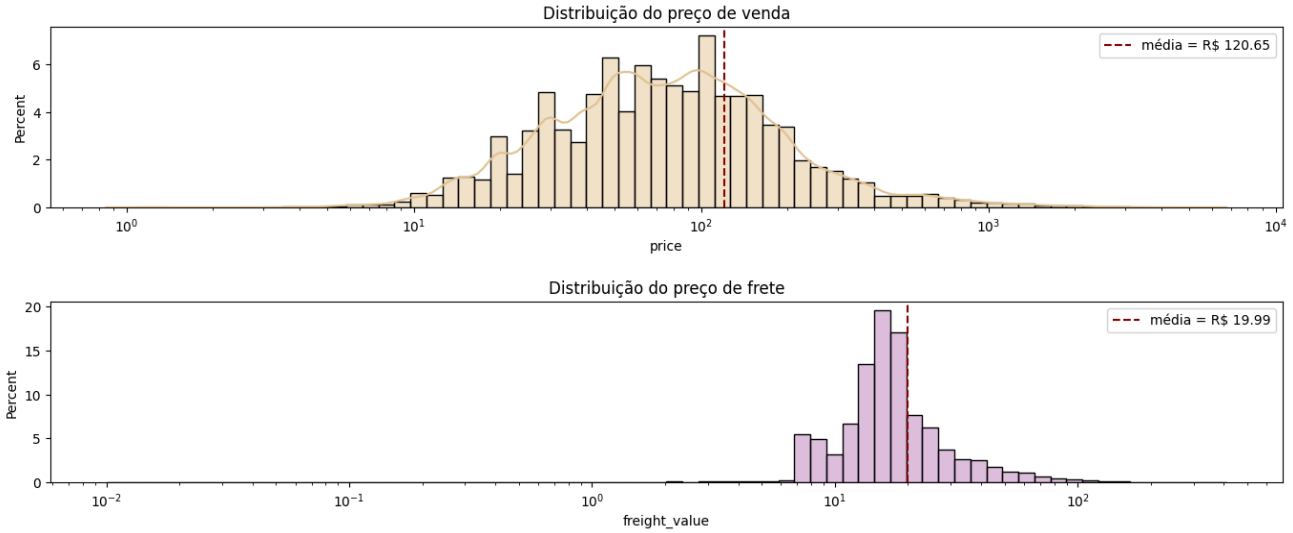
Orders: status das compras



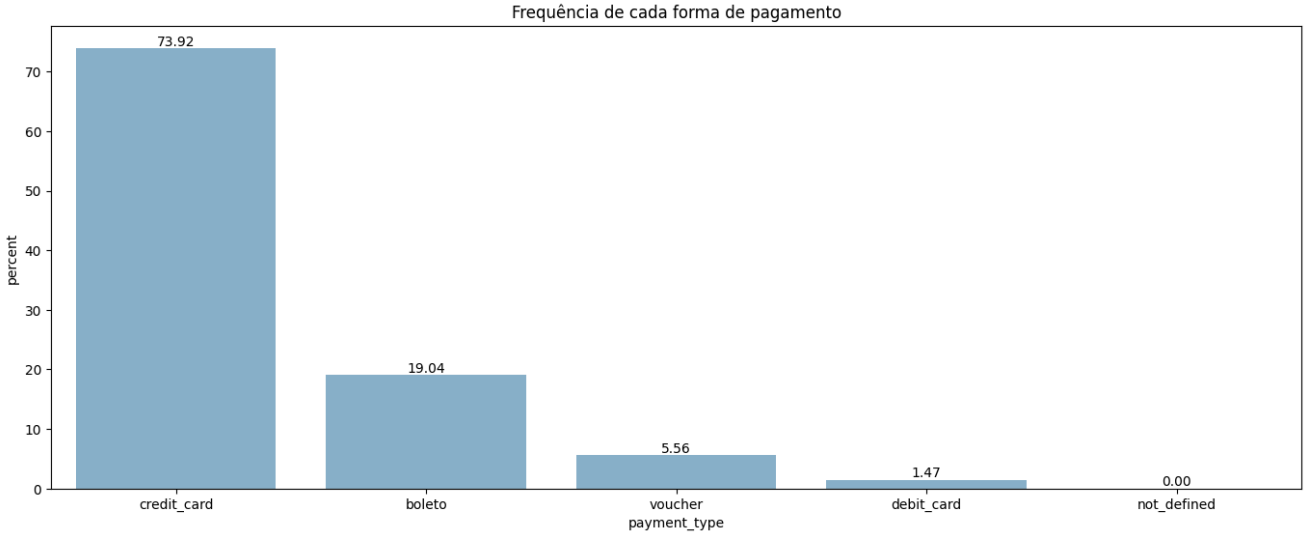
Orders: data estimada de entrega e data efetiva da entrega



Order items: preço de venda e preço de frete

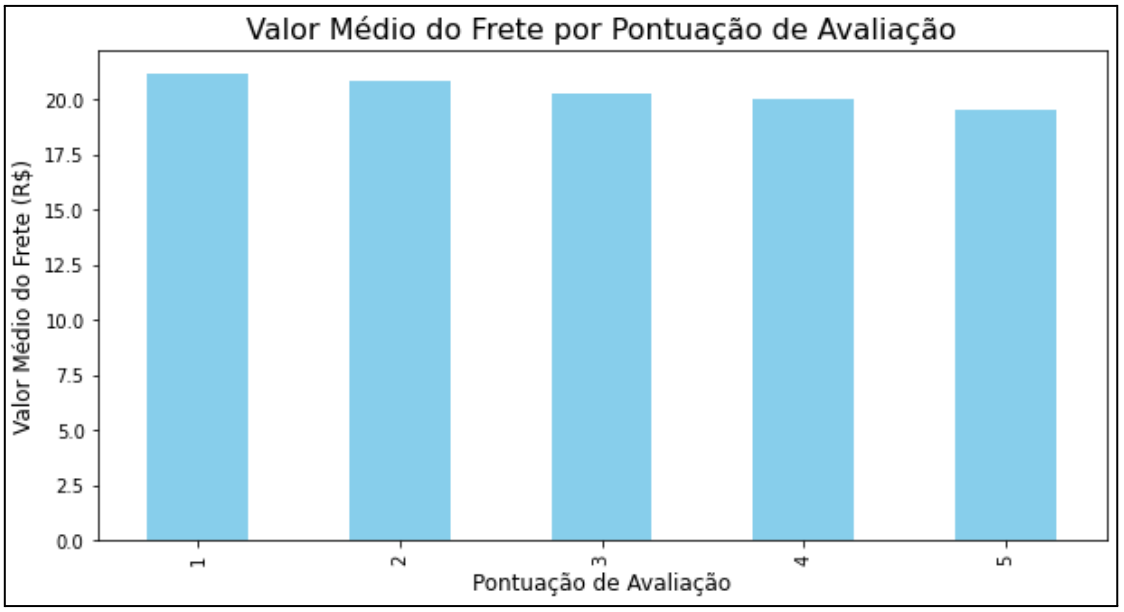


Order payments: share das formas de pagamentos

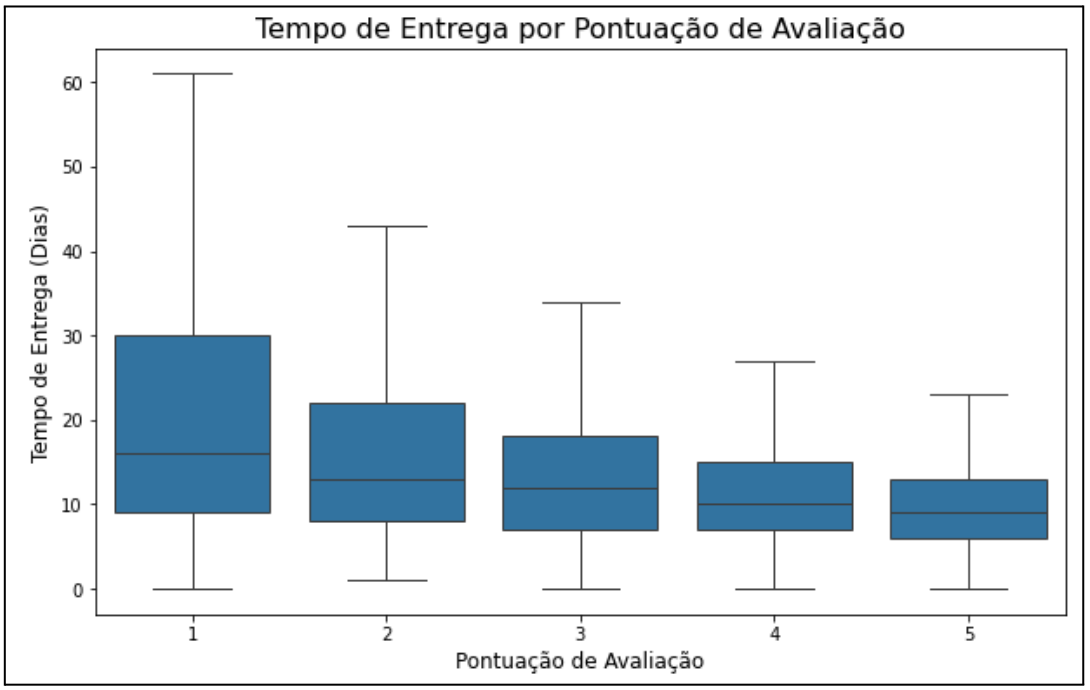


Análise Exploratória Bivariada

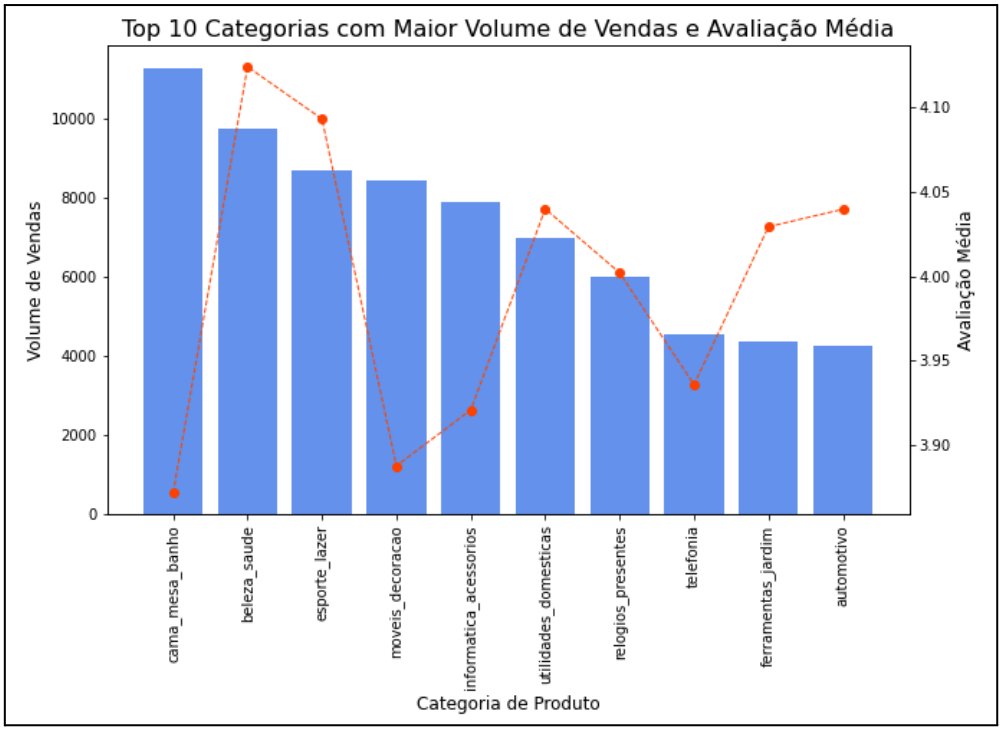
Análise da relação entre valor do frete e a pontuação de avaliação:



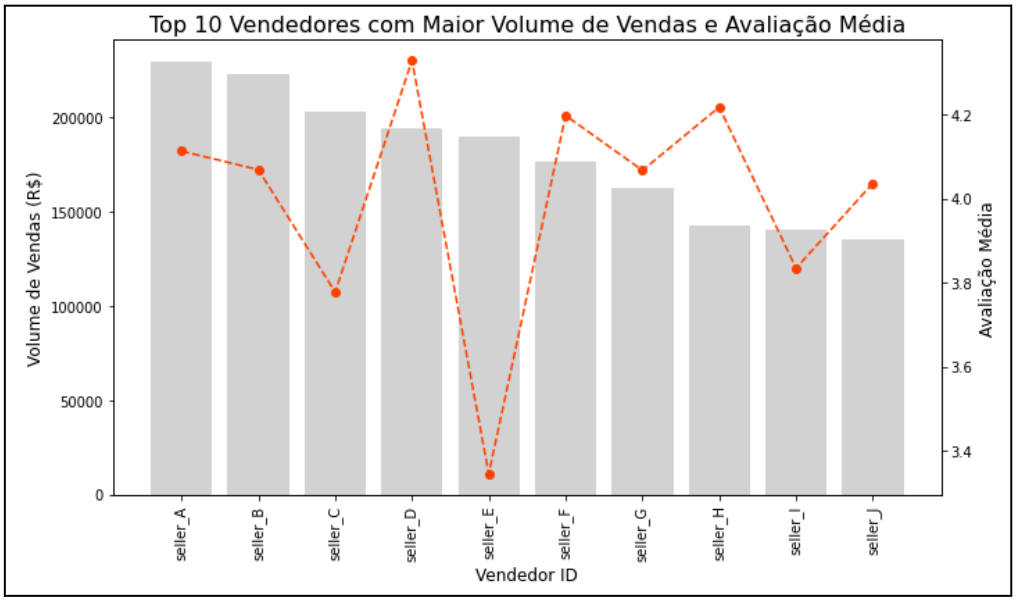
Impacto do tempo de entrega na avaliação do cliente:



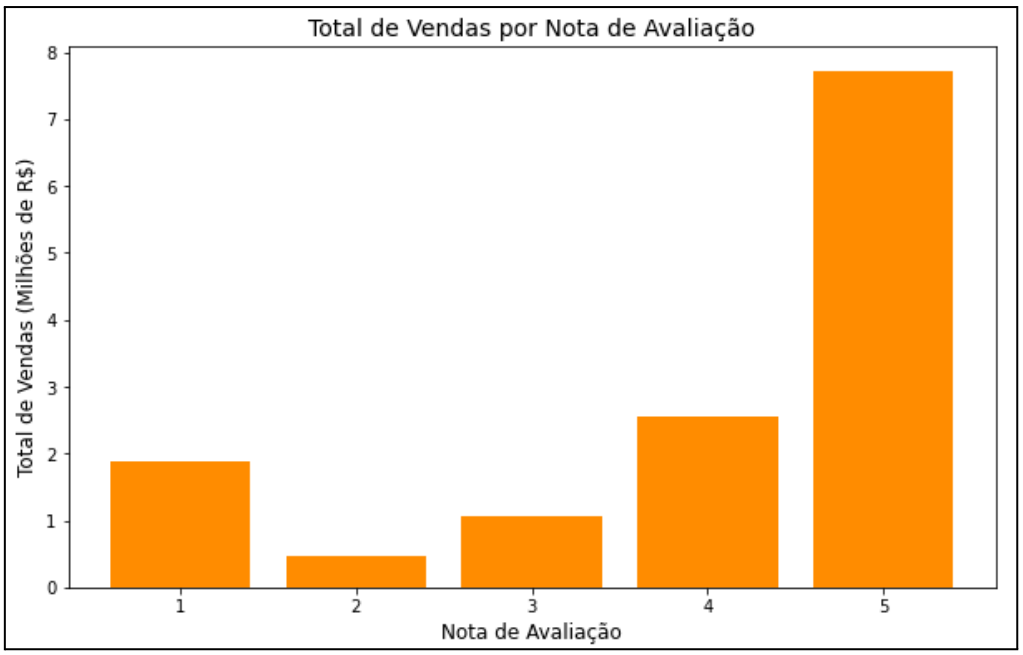
Análise por categoria de produto e sua relação com volume de vendas e avaliações:



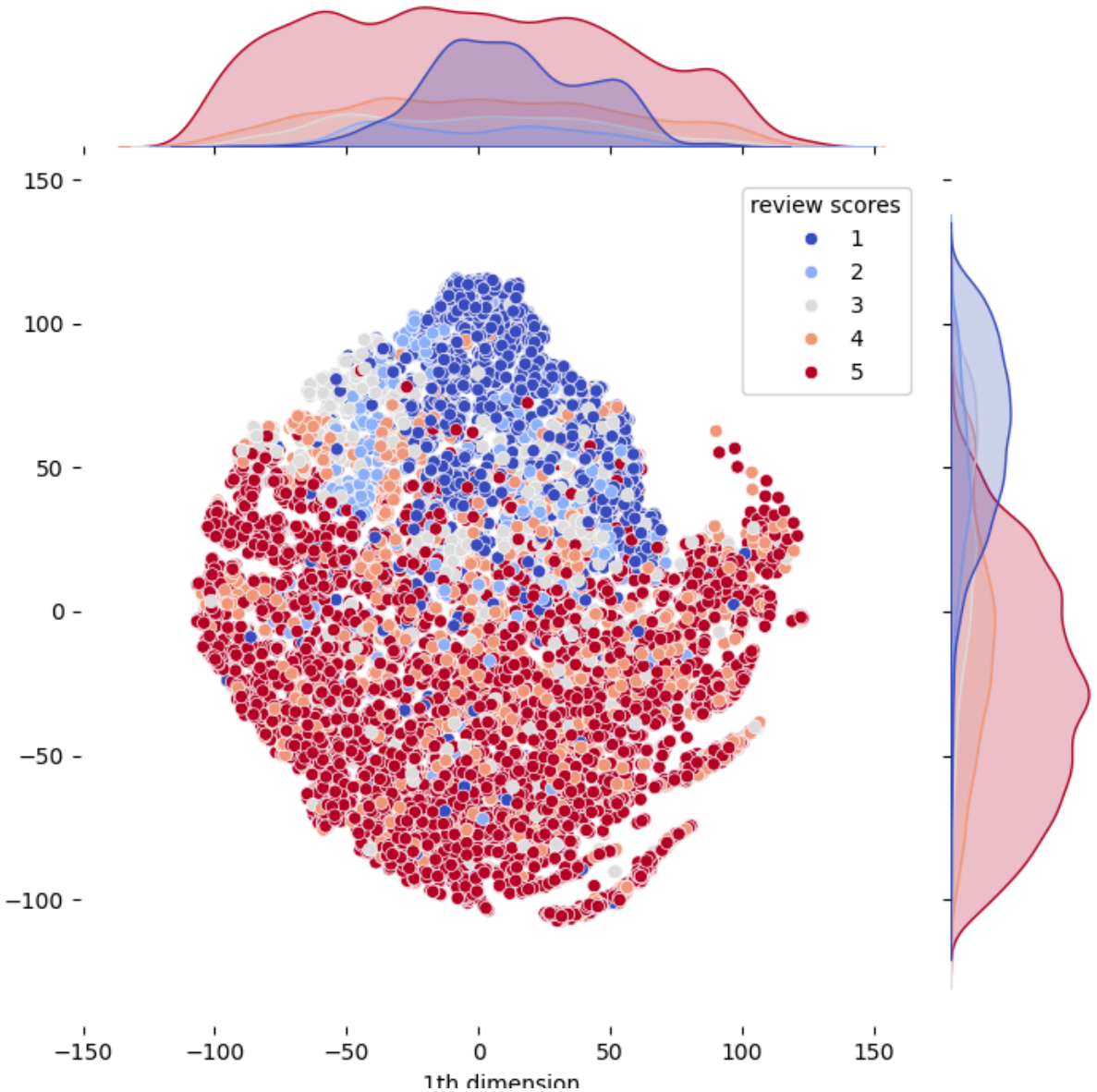
Análise por vendedores e sua relação com volume de vendas e avaliações:



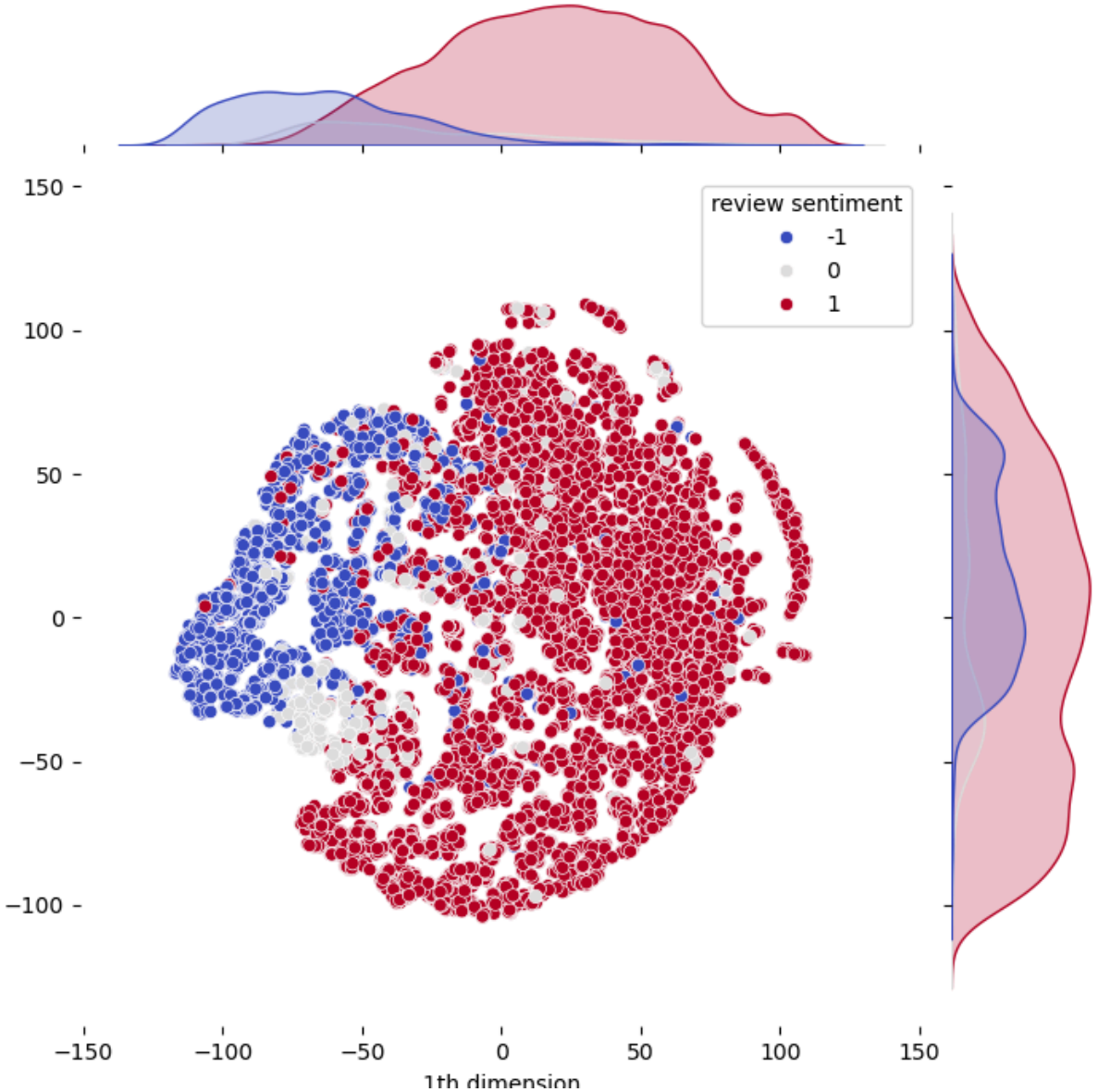
Análise por notas das avaliações e faturamento:


Total de Vendas por Nota de Avaliação
Total de Vendas (Milhões de R$)
0
1
2
3
4
5
6
7
8
1
2
3
4
5
Nota de Avaliação

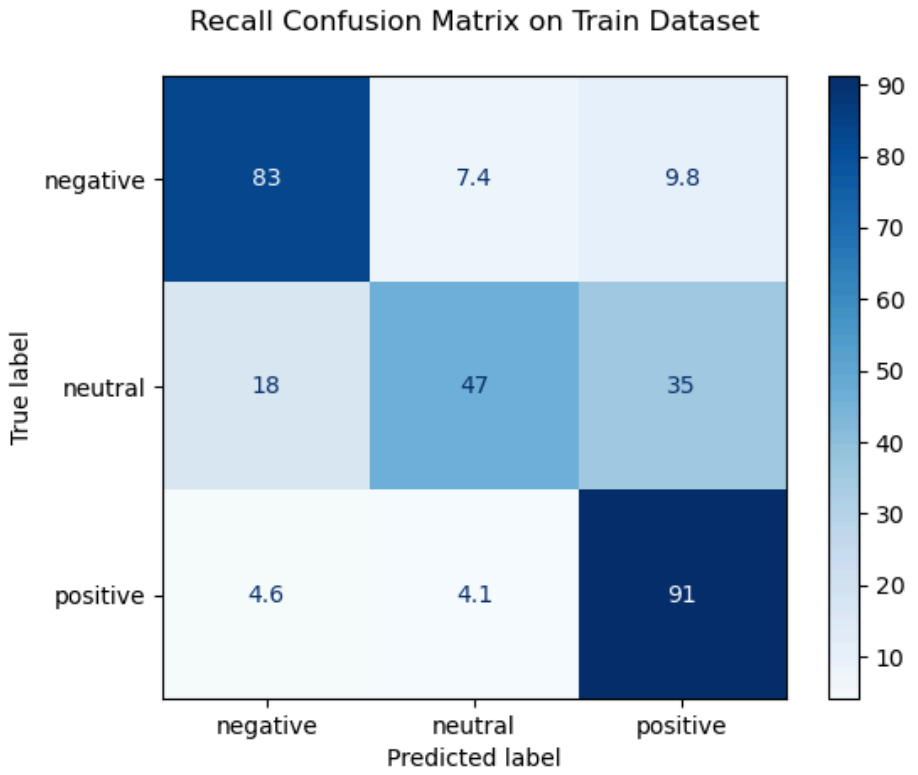
t-SNEs com review scores



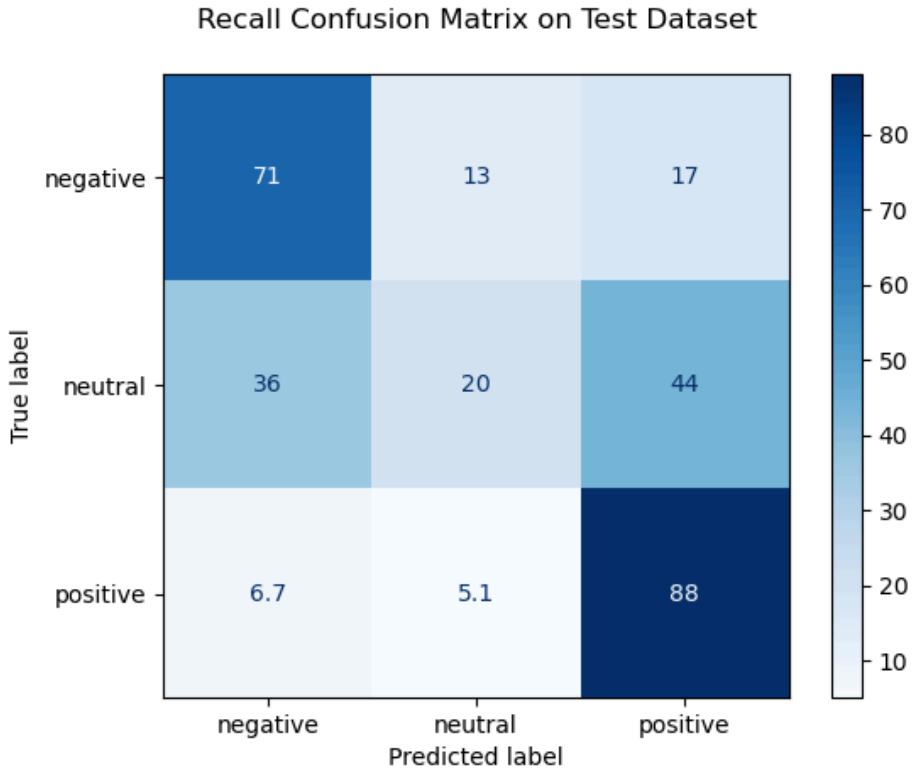
t-SNEs com review sentiment



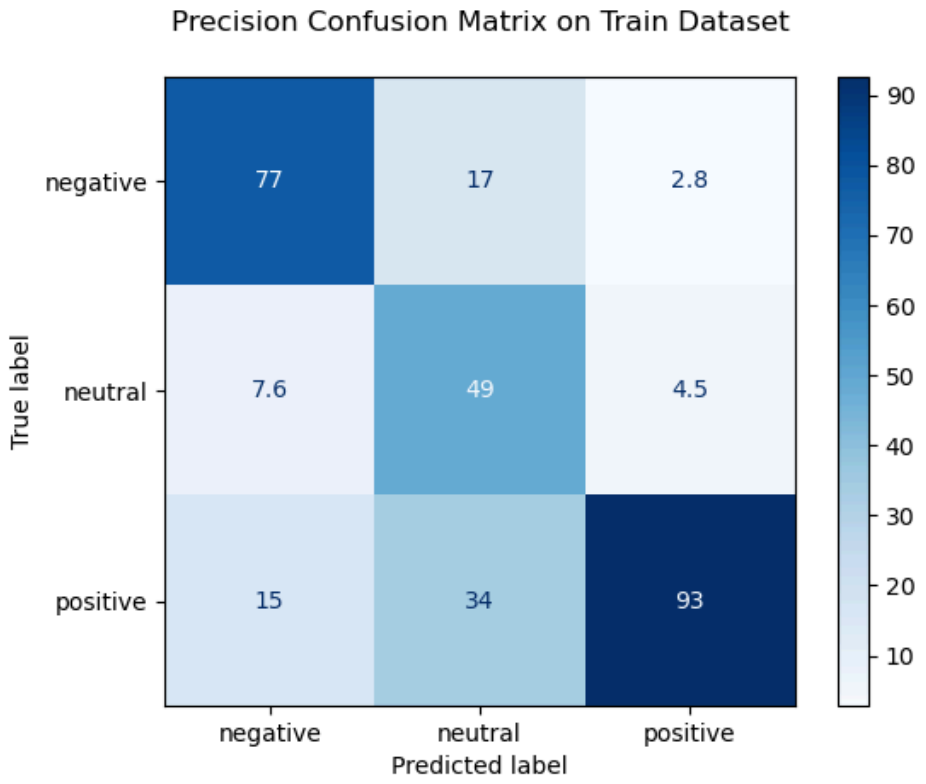
Métricas de recall para o conjunto de treinamento



Métricas de recall para o conjunto de teste



Métricas de precisão para o conjunto de treinamento



Métricas de precisão para o conjunto de teste