Ficha Técnica do Projeto
Projeto: Projeto 4 - DataLab
Analista de Dados: Bruna Derner
Ferramentas utilizadas: SQL, Google Colab, VSCode, Google Apresentações e Loom.

1. Objetivo do Projeto

Investigar os determinantes da nota de avaliação (rating) atribuída pelos clientes a produtos comercializados em um e-commerce. Especificamente, analisa‑se como variáveis como preço (cheio e com desconto), desconto percentual, popularidade (número de avaliações), categorias de produto e sentimentos extraídos das resenhas influenciam o rating final concedido pelos consumidores. O intuito é compreender quais fatores exercem maior impacto na percepção de qualidade e satisfação do usuário, auxiliando gestores a promover estratégias de precificação, marketing e atendimento mais efetivas.

2. BigQuery

2.1. Importação de dados

Os dados foram carregados no ambiente do BigQuery, garantindo que todas as tabelas relevantes (como amz_prod e Amazon Review) estivessem disponíveis para as etapas seguintes de tratamento. Nessa fase inicial, confirmou-se a integridade dos formatos de arquivo e a compatibilidade entre os esquemas de cada tabela, assegurando que nenhuma informação essencial fosse perdida durante o processo de upload.

2.2. Identificar e Tratar Nulos

Na tabela amz_prod, foram detectados quatro registros sem informação na coluna about_product, que foram removidos ou preenchidos conforme a estratégia de limpeza estabelecida. Na tabela de avaliações (Amazon Review), identificaram-se 466 valores ausentes distribuídos entre as colunas img_link e product_link, além de duas ocorrências sem valor em rating_count. Os valores da amz_review foram mantidos. Quando utilizados em análise em python retirados utilizando código, para não perder as demais informações.

2.3. Tratar Duplicados

Na tabela amz_prod, havia um total de 106 linhas totalmente duplicadas, isto é, em que todas as colunas apresentavam exatamente os mesmos valores. Essas duplicatas foram consolidadas, mantendo apenas uma instância de cada produto e descartando as entradas redundantes. Havia também 12 registros de product_id duplicados que não foram identificados no tratamento das linhas duplicadas porque algum campo da tabela podia apresentar valor nulo em img_link, product_link ou about_product. Na tabela amz_review, identificaram-se registros de product_id diferentes, mas com o mesmo review_id, essa variação aconteceu porque eram o mesmo produto, com cores diferentes. Então as avaliações reais eram: 1189. Nesses casos, avaliou-se se a distinção era relevante para a análise: quando

a cor representava uma característica substancial do produto, manteve-se o registro.

Linha	product_id	qtde_reviews
1	B01486F4G6	2
2	B096MSW6CT	3
3	B083342NKJ	2
4	B0B5B6PQCT	2
5	B0B5LVS732	2
6	B01M6453MB	2
7	B09MT84WV5	2
8	B08CF3D7QR	2
9	B00J5DYCCA	2
10	B009P2LIL4	2
11	B07DJLFMPS	2

2.4. Dados fora do escopo

Algumas colunas não serão utilizadas na análise, como user_name, img_link e product_link, mas essas informações foram mantidas na tabela.

2.5. Dados discrepantes

2.5.1. Variáveis Categóricas

Nas variáveis categóricas, identificaram-se inconsistências de nomenclatura (variações de escrita ou categorias duplicadas). Para uniformizar, definiu-se um conjunto padronizado de categorias e reclassificou-se cada entrada.

2.5.2. Variáveis Numéricas

Não foram encontrados dados discrepantes numéricos. As imagens abaixo mostram a distribuição dos dados, na primeira avalia-se rating, a segunda rating count e a última porcentagem de desconto.

Linha	minimo	maximo	media	mediana
1	2.0	5.0	4.0918518518518514	4.1

Linha	minimo	maximo	media	mediana
1	2	426973	17656.862759643882	4703

Linha	minimo	maximo	media	mediana
1	0.0	0.94	0.46700740740740715	0.49

2.6. Criar novas variáveis

No modelo, foi incluída a variável de categoria, a partir da qual extraímos tanto a primeira categoria quanto a última categoria de cada produto. Também calculamos os quartis para as

variáveis de desconto, de número de avaliações (rating_count) e de avaliação em si (rating). As demais variáveis foram criadas diretamente no Ecolab, onde realizamos a análise de sentimento para classificar o teor de cada avaliação: identificamos se o comentário apontava um problema no produto, se tratava de um feedback geral sobre o produto, se mencionava o serviço, se se referia à qualidade do produto ou à entrega. Todos esses aspectos de sentimento foram, então, registrados como variáveis classificatórias.

2.7. Unir tabelas

Após a preparação individual de cada tabela, realizou-se a integração entre amz_prod e Amazon Review por meio da coluna product_id, consolidando informações de produto e avaliação em um único conjunto de dados.

2.8. Quartil

Foram calculados quartis de rating, rating_count e discount_percentage.

3. Jupyter Notebook

3.1. Análise de Sentimento e NLP segmentando tema da avaliação

Antes de formularmos as hipóteses relativas aos determinantes do rating, conduzimos uma etapa dedicada à análise de sentimentos nas avaliações textuais dos usuários, utilizando técnicas de Processamento de Linguagem Natural (NLP). O principal objetivo dessa fase foi transformar as opiniões expressas em texto — que, por si só, são informações qualitativas — em uma métrica quantitativa capaz de ser inserida nos modelos de regressão. Em outras palavras, pretendíamos extrair, de cada review, uma medida numérica representativa do teor positivo ou negativo do comentário, de modo a investigar, posteriormente, se esse componente emocional influenciaria a nota atribuída ao produto.

O procedimento começou pelo pré-processamento das resenhas. Inicialmente, todas as avaliações foram convertidas para letras minúsculas para garantir consistência na comparação de termos. Em seguida, removemos pontuação, símbolos especiais e expressões muito curtas ou ambíguas, de modo a eliminar ruídos como emojis, tags HTML ou URLs que pudessem interferir na contagem de palavras relevantes. As stopwords – isto é, palavras comuns da língua que não carregam carga semântica significativa para o sentimento, como “o”, “a”, “de”, “para” – foram filtradas para que não inflassem indevidamente as contagens de termos. Por último, aplicamos técnicas de lematização (em alguns casos, optamos pela normalização via root de palavras quando o uso de lematização mostrou-se mais adequado), garantindo que variações de um mesmo radical fossem reconhecidas como uma única unidade semântica. Por exemplo, “funciona”, “funcionando” e “funcionou” passaram a ser tratados uniformemente como “funcionar”.

Com o texto devidamente limpo, empregamos um léxico pré-definido de palavras classificadas como positivas ou negativas. Esse dicionário continha termos que, em geral, expressam satisfação (como “excelente”, “perfeito”, “recomendo”) e termos que denotam insatisfação ou problemas (por exemplo, “defeituoso”, “ruim”, “não funciona”). Para cada resenha, calculamos o número total de ocorrências de palavras positivas e, de forma análoga, o número de palavras negativas. A partir desses valores, geramos uma variável chamada sent_positivo, que corresponde à razão entre o número de termos positivos e a quantidade total de palavras examinadas na avaliação. Assim, uma review em que 20% das palavras fossem positivas teria sent_positivo igual a 0,20, indicando que uma parte significativa do texto expressava bom desempenho do produto ou satisfação do usuário. Dessa forma, a variável sent_positivo — construída a partir do léxico de palavras positivas — tornou-se a

principal métrica de sentimento utilizada na análise subsequente. Ela capturou, em forma numérica, a intensidade de elogios ou impressões favoráveis que cada usuário expressou ao avaliar o produto.

3.2. Hipóteses e Resultados

3.2.1. Hipótese 1: Produtos com mais avaliações tendem a ter ratings mais altos

O modelo de regressão linear simples indica que existe uma relação estatisticamente significativa entre o número de avaliações (rating_count) e a nota média (rating), mas essa relação é extremamente fraca. O modelo explica apenas 0,9% da variância do rating (R^2 = 0,009), o que revela que a popularidade, medida pelo número de avaliações, tem um efeito prático muito pequeno sobre a nota média.

Embora o coeficiente estimado seja positivo, indicando que produtos com maior contagem de avaliações tendem a exibir ratings ligeiramente superiores, a magnitude desse efeito é mínima. Este padrão pode refletir dois fenômenos simultâneos: um efeito de sinalização de popularidade, em que consumidores confiam em produtos amplamente avaliados, e um efeito reputacional acumulado, em que bons ratings iniciais atraem mais compradores e reforçam a média.

Contudo, a análise dos resíduos do modelo revela limitações importantes. Há presença de autocorrelação e desvio em relação à normalidade, o que compromete a validade estatística da regressão simples. Em suma, embora o número de avaliações contribua positivamente para o rating, essa variável, isoladamente, não é um bom preditor da nota média.

3.2.2. Hipótese 2: A categoria do produto afeta o rating médio

O modelo de regressão linear avaliou o impacto das categorias de produto sobre a nota média (rating). Embora o modelo seja estatisticamente significativo no conjunto ($p < 0,001$), seu poder explicativo é baixo: as categorias de produto explicam apenas 3,8% da variação nos ratings ($R^2 = 0,038$).

Além disso, nenhuma categoria apresenta um efeito individual estatisticamente significativo sobre a nota média, o que sugere que, isoladamente, as diferenças entre categorias não produzem um impacto consistente no rating.

Observa-se também que o modelo apresenta limitações importantes: os resíduos indicam autocorrelação (Durbin-Watson = 0,246) e violação da normalidade, o que compromete a robustez estatística das conclusões.

Em resumo, embora a hipótese de que a categoria do produto afete o rating médio seja confirmada em termos globais, esse efeito é fraco e não se traduz em diferenças significativas entre categorias específicas.

3.2.3. Hipótese 3: Existe uma associação entre a porcentagem de desconto e o rating.

O modelo de regressão linear avaliou a relação entre o percentual de desconto (discount_percentage) e a nota média (rating). O resultado foi estatisticamente significativo ($p < 0,001$), porém com um poder explicativo modesto: o percentual de desconto explica apenas 2,8% da variação no rating ($R^2 = 0,028$). O coeficiente estimado é negativo e significativo, indicando que produtos com maiores descontos tendem a receber avaliações médias ligeiramente mais baixas. Este padrão sugere que ofertas mais agressivas podem estar associadas a percepções de menor qualidade ou expectativas não atendidas. No entanto, o modelo apresenta limitações importantes: há presença de autocorrelação nos resíduos (Durbin-Watson = 0,277) e violação da normalidade, o que reduz a robustez das inferências.

Em resumo, confirma-se a existência de uma associação negativa entre a porcentagem de desconto e o rating médio, ainda que o efeito prático seja pequeno e o modelo tenha limitações metodológicas.

3.2.4. Hipótese 5: Produtos com valores mais altos tendem a ter avaliações maiores que os mais baratos.

A análise comparou as avaliações médias (rating) entre produtos de preço alto e preço baixo. A diferença observada foi mínima: 4,09 para preços altos e 4,08 para preços baixos. Tanto o teste t ($t = 0,38$; $p = 0,703$) quanto a regressão linear (coef. = -0,0068; $p = 0,703$) indicaram que não há diferença estatisticamente significativa entre os grupos. O modelo apresentou um R^2 praticamente nulo (0,0%), reforçando que o preço, por si só, não explica a variação na nota média dos produtos. Embora o modelo apresente limitações, como autocorrelação e não normalidade dos resíduos, o resultado geral sugere que a variação no preço não impacta de forma significativa o rating.

3.2.5. Hipótese 7: Q1 tem mais sentimentos negativos

A análise revela que 40,6% das avaliações no primeiro quartil (Q1) apresentam sentimento negativo, em comparação com 20,8% nas demais avaliações. Essa diferença é estatisticamente significativa ($Z = 6,90$; $p < 0,0001$), indicando uma presença consideravelmente maior de avaliações negativas entre os produtos de pior desempenho.

A regressão logística confirma esse padrão: avaliações com sentimento negativo estão significativamente associadas a uma maior probabilidade de pertencerem ao Q1 (coef. = 0,96; $p < 0,001$). Embora o Pseudo R^2 seja modesto (3,25%), o modelo é estatisticamente robusto e indica que o sentimento negativo contribui de forma relevante para a probabilidade de um produto figurar entre os piores avaliados.

3.2.6. Hipótese 8: Determinadas categorias dominam o Q1

A análise indica que determinadas categorias de produto estão associadas a uma maior probabilidade de avaliações figurarem no primeiro quartil (Q1), que corresponde às avaliações mais negativas. Especificamente, as categorias Electronics (coef. = 0,56; $p = 0,003$) e Home&Kitchen (coef. = 0,79; $p < 0,001$) apresentam maior propensão a avaliações negativas. Já a categoria Outros apresentou um coeficiente negativo, mas não significativo ($p = 0,096$). Além disso, o sentimento positivo está fortemente associado a uma menor chance de uma avaliação estar no Q1 (coef. = -0,87; $p < 0,001$), reforçando que aspectos emocionais influenciam a avaliação dos produtos. O modelo é estatisticamente significativo ($p < 0,0001$) e apresenta um Pseudo R^2 de 5,3%, indicando uma melhoria modesta em relação ao modelo nulo, com relevância prática.

3.2.7. Hipótese 10: Proporção de sentimentos positivos varia por categoria

A análise indica que existe uma associação estatisticamente significativa entre a categoria do produto e a presença de sentimento positivo (elogio) nas avaliações (Qui-quadrado = 469,41; $p = 0,0480$). A regressão logística, embora marginalmente significativa no conjunto ($p = 0,065$), revela que algumas categorias específicas apresentam menor probabilidade de conter elogios nas avaliações. Notadamente, a categoria Smartphones apresenta uma associação negativa e estatisticamente significativa com a presença de elogios (coef. = -0,91; $p = 0,032$), sugerindo que avaliações para esses produtos têm menor chance de conter sentimentos

positivos explícitos. O modelo apresenta um Pseudo R^2 de 2,4%, indicando uma explicação modesta da variância, mas com relevância prática.

3.2.8. Regressão Linear Múltipla

A análise mostra que diversas variáveis estão associadas de maneira relevante à nota média dos produtos (rating). A regressão linear confirma esse padrão: fatores como preço com desconto (log), quantidade de avaliações (log), percentual de desconto, sentimento positivo e categoria do produto influenciam significativamente o rating. O sentimento positivo exerce o maior impacto individual positivo (coef. = 0,16; $p < 0,001$), indicando que avaliações com tom favorável contribuem para notas mais altas. Por outro lado, um maior percentual de desconto está associado a uma redução no rating (coef. = -0,15; $p = 0,001$), sugerindo que descontos agressivos podem gerar percepções mais críticas por parte dos consumidores. Além disso, produtos das categorias Electronics (coef. = -0,10; $p < 0,001$) e Home&Kitchen (coef. = -0,14; $p < 0,001$) tendem a receber avaliações ligeiramente inferiores, em comparação com a categoria de referência. O modelo é estatisticamente significativo ($p < 0,0001$) e apresenta um R^2 de 17,5%, explicando uma parcela relevante da variação nas avaliações. Esses resultados reforçam que tanto características objetivas do produto quanto o conteúdo emocional das avaliações contribuem de forma consistente para a nota atribuída pelos consumidores.

Anexo 1 - Links de Interesse

bruna-derner.proj04

https://lookerstudio.google.com/reporting/8c59f8a9-be3a-4aff-95fe-07d49d74c5b5

https://drive.google.com/file/d/1mScWPZe9yXcXWu2murBxx9gKmwPqrn1s/view?usp=sharing