PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2024/2
PROJETO	Segmentação de Clientes - Renner
COMPONENTES DO GRUPO	Lucas Gomes Lorenzo Corrêa Lazzarotto Luiz Eduardo de Souza Pedro Devincenzi Ferreira

Breve descrição do problema

Com o crescente aumento do uso de e-commerce, diversidade de produtos e serviços disponíveis, e características distintas de clientes que utilizam este tipo de plataforma, estratégias que funcionavam de maneira eficiente tempos atrás, podem não responder tão bem às demandas atuais.
Conseguir identificar de maneira correta as diferenças fundamentais entre grupos de clientes e a partir disso, ser capaz de elaborar estratégias ideais para cada grupo é um desafio contemporâneo e dinâmico, mas que tem se tornado um diferencial importante na retenção e incremento de receitas.

Breve descrição da solução proposta

Como proposta de trabalho, o grupo planeja desenvolver uma ferramenta chamada “Renner Rethink”, que visa analisar grupos de clientes sob diversos prismas e agrupá-los através de técnicas de Clusterização, para identificar características que diferenciem os grupos.
Os entregáveis serão feitos através de páginas web com as seguintes funções:
- Detalhamentos das análises e tratamento dos dados
- Modelagem interativa, onde a Renner pode executar testes e avaliar resultados sobre os modelos
- Relatório em PDF com toda a explicação do projeto baseado no CRISP-DM

Fases da Metodologia CRISP-DM

Fase	Atividades	Conclusão
Compreensão do Negócio	Entendimento do problema proposto e em qual contexto está inserido. Principais características e particularidades da empresa parceira e do mercado trabalhado.	100%
Compreensão dos Dados	Avaliar os dados disponibilizados para verificar erros e desvios, além de iniciar o entendimento se estes dados são capazes de gerar insumos para a resolução do problema	80%
Preparação dos Dados	Com os dados higienizados, fazer agregações e tratamentos para utilização na etapa de modelagem.	80%
Modelagem	Efetuar testes com diversas ferramentas que, em tese, podem responder ao problema proposto. Selecionar os mais promissores e para estes, escolher um ou mais modelos considerados ideais.	0%
Avaliação	Definir as métricas que serão utilizadas para comparar modelos e avaliar os resultados. Deve-se ser capaz de identificar se os resultados encontrados estão coerentes com o negócio.	0%

Resumo do que foi concluído até o momento

• Análise dos Dados	❶❷❸❹❺❻❼❽❾❿
• Desenvolvimento do template do Front-end	❶❷❸④⑤⑥⑦⑧⑨⑩
• Tratamento de Dados	❶❷❸❹❺❻❼⑧⑨⑩
• Definição dos prismas de análise	❶❷❸❹❺⑥⑦⑧⑨⑩
• Escolha das técnicas de clusterização	①②③④⑤⑥⑦⑧⑨⑩
• Avaliação dos clusters e sua explicabilidade	①②③④⑤⑥⑦⑧⑨⑩
• Relatório	❶❷❸❹❺❻⑦⑧⑨⑩

Autocrítica

Nas primeiras semanas o ritmo do grupo foi prejudicado pela demora no envio dos dados por parte da empresa parceira.
Vencida essa fase, o grupo está alinhado e conseguindo avançar nas atividades propostas durante a semana. Como ponto positivo destacamos a participação do grupo no período de aula, onde todos os integrantes estão presentes para participar das discussões e alinhamentos para a próxima semana de trabalho.
Há dificuldade em antecipar atividades, impedindo que haja algum tipo de margem nas datas de entrega, beirando algum tipo de atras, mas que não tem se refletido nas entregas.

Dado o cenário, comprometimento do grupo e proposta alinhada com a Renner, entendemos que o grupo conseguirá entregar 100% do que foi planejado.

Nota até o momento: 8,0

-x-

RELATÓRIO

1. Compreensão do Negócio

Background

O mercado de e-commerce está cada vez mais concorrido, com maior oferta e diversidade de opções para clientes, que podem rapidamente encontrar informações relevantes sobre suas demandas. Entender o comportamento do público, definir estratégias competitivas e ser mais acurado na identificação de quais campanhas podem ser atrativas para cada grupo pode ser uma peça-chave na prospecção e fidelização de clientes.

Como um dos players mais importantes no país, o grupo Renner busca ser referência no uso de dados para inteligência de mercado, alinhando tecnologia aos produtos de moda, decoração, financeiro e serviços.

Objetivos de negócio e critérios de sucesso

Objetivo: Encontrar agrupamentos de clientes com base no seu estágio de vida e padrão de consumo, buscando identificar quais produtos e canais são mais aderentes ao seu gosto. A hipótese do grupo é que estes segmentos refletem necessidades e preferências distintas, permitindo campanhas de marketing mais direcionadas e personalizadas, além de estratégias de vendas otimizadas.

Critérios de Sucesso:

Definimos como critérios de sucesso, dois prismas a serem analisados:

- Técnico:
  - Utilização de métricas consolidadas para avaliação de distância ou densidade de clusters.
  - Devemos ser capazes de representar segmentos distintos de clientes, baseado em características comuns entre indivíduos de um mesmo cluster.
- Negócio:
  - Os clusters devem ser explicáveis e coerentes, com caracterização clara de similaridade e diferenciação.
  - Os resultados devem ser alinhados com as expectativas da Renner.

Inventário de recursos

- Pessoas:
  - Luiz Eduardo de Souza – Projeto III
  - Lucas Gomes – Projeto II
  - Lorenzo Lazzarotto – Projeto II
  - Pedro Ferreira – Projeto I
- Fontes de Dados
  - Arquivos em formato CSV disponibilizados pela Renner
    - Tabela de Clientes, contendo registros pessoais de cada cliente.
    - Tabela de Navegação no site e app, contendo informações sobre produtos visitados, colocados no carrinho e lista de desejos, e itens comprados.

        - Tabela de Transação, contendo datas de venda, valor, categoria do item vendido e plataforma.
- Equipamentos
    - Computadores pessoais de cada um dos integrantes do grupo.
    - Repositório no GitHub

Requisitos, suposições e restrições

Requisitos:

- Entrega será composta por todos os passos necessários até a modelagem e avaliação.
- Entrega de um relatório final contendo todas as etapas de desenvolvimento do projeto.
- Serão compartilhados com a Renner documentos e códigos produzidos pelo grupo e que sejam suficientes para o entendimento da solução proposta.
- Reuniões para apresentação de resultados parciais.
- Eventuais esclarecimentos de dúvidas pertinentes ao desenvolvimento do projeto, devendo essas serem sanadas em conjunto com o orientador da disciplina e a equipe da Renner.
- Cumprimento dos prazos e datas acordadas com o objetivo de não atrasar as entregas finais.

Suposições:

- Os dados recebidos serão considerados curados e corretos pela equipe da Renner.
- Os dados disponibilizados foram anonimizados até o nível onde não se pode identificar dados pessoais de um cliente, bem como personificar qualquer navegação ou transação.

Restrições:

- Não será considerada parte da entrega a etapa de deploy dos modelos obtidos, podendo esta ser feita caso seja de interesse única e exclusivamente do grupo.
- Não serão consideradas soluções que necessitem investimento financeiro para serem implementadas, salvo se em comum acordo com todos os integrantes do grupo e com o orientador da disciplina.
- Todo o trabalho estará em conformidade com a Lei Geral de Proteção de Dados (LGPD).
- Os dados disponibilizados pelo grupo Renner serão utilizados somente para fins deste projeto e não serão compartilhados fora do contexto da disciplina.

Terminologia

E-commerce: Comércio eletrônico, refere-se às vendas pela internet de produtos e serviços.

GitHub: Plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o Git.

Clusterização: Técnica de agrupamento de dados para criar grupos de clientes com comportamentos e características semelhantes.

Segmentação de Mercado: Divisão do mercado em grupos distintos com base em características comuns.

PCA: Técnica de redução de dimensionalidade chamada Análise de Componentes Principais.

RFM: O RFM, ou RFV, é uma metodologia de segmentação utilizada para agrupar consumidores de acordo com o seu comportamento de compra com base nos dados de Recência, Frequência e Valor Gasto (Monetário).

Streamlit: Biblioteca open-source em Python que permite a criação de aplicativos web para análise de dados de forma extremamente rápida.

Objetivos de mineração e critérios de sucesso

Clusterização de clientes com base nos dados disponíveis, de forma que possam ser segmentados de maneira consistente em número de grupos capaz de identificar diferenças significativas entre as características específicas de cada grupo.

Critérios de Sucesso:

- Os clusters devem ter quantidades de indivíduos suficientes para que campanhas de vendas ou marketing façam sentido, sem perder a especificidade e características do cluster.
- Serão aplicadas métricas reconhecidas e comumente utilizadas para avaliar a qualidade da separação dos clusters.
- As características de cada cluster devem ser coerentes e alinhadas com o que a Renner busca.

Plano de Projeto

1. Análise exploratória de dados
   1.1. Verificar a qualidade dos dados recebidos
2. Definição da proposta de solução
   2.1. Avaliar a disponibilidade dos dados necessários para desenvolver a solução
   2.2. Verificar a necessidade de buscar complementação de dados
3. Tratamento dos dados
   3.1. Criar dataset “gold” para desenvolvimento do projeto
4. Análise exploratória sobre os dataset “gold”
   4.1. Avaliar se o dataset tratado possui os dados necessários para atendimento do projeto.

Avaliação inicial de técnicas e ferramentas

Para clusterização, serão considerados algoritmos clássicos para essa tarefa, como:

- K-means: Algoritomo baseado em distâncias
- DB-scan: Algortimo baseado em densidade
- Clusterização Hierarquica: Baseado em dados divisivos e/ou aglomerativos
- Deep learning: Algoritmos baseados em aprendizado profundo
- Análise de RFM: Índice de Recência, Frequencia e valor monetário.

Para a seleção e/ou redução de features, serão considerados:

- PCA: Análise de componentes principais

Outras técnicas podem ser utilizadas conforme a evolução do projeto.

2. Compreensão dos Dados

Coleta dos dados

Os dados foram disponibilizados pela própria Renner, sendo compostos por 3 arquivos em formato .CSV, representando registros de transação, clientes e navegação no app ou site da empresa.

Caso necessário, o grupo buscará dados relevantes para solução do problema em outras fontes confiáveis.

Descrição dos dados

- Clientes.csv
  - 34707 registros
  - Colunas:
    - Id_cliente: Chave primária de cliente
    - Gênero: Indica o gênero indicado pelo cliente
      - F: Feminino
      - M: Masculino
      - I: Indefinido
    - Idade: Número que representa a idade no momento da extração da base de dados
    - Bairro: Bairro cadastrado pelo cliente
    - Cidade: Cidade cadastrada pelo cliente
    - Data última compra Renner: Registro da última compra feita pelo cliente
    - Data primeira compra Renner: Registro da primeira compra feita pelo cliente
- Navegação.csv
  - 171834 registros
  - Colunas:
    - Id_cliente: Chave primária de cliente
    - Nome_evendo: Ações que o cliente fez sobre um item
    - Data_evento: Data em que o evento ocorreu
    - Codigo_item: Chave primária do item em questão
- Transação.csv
  - 763347 Registros
  - Colunas:
    - Data_venda:
    - Id_cliente: Chave primária de cliente
    - Codigo_item: Chave primária do item
    - Valor: Valor total da transação
    - Nome_divisão: Divisão que o item pertence (Categoria do item)
    - Tipo_venda: Se venda online ou offline

Análise exploratória dos dados

A análise exploratória sobre os dados teve como objetivo principal entender a distribuição, características e particularidades dos dados nos diferentes conjuntos de dados apresentados pela organização parceira, assim garantindo um melhor entendimento dos mesmos e possibilitando o grupo a escolher as melhores abordagens para serem aplicadas tanto ne preparação, quanto na modelagem do problema.

Abaixo, algumas observações feitas pelo grupo:

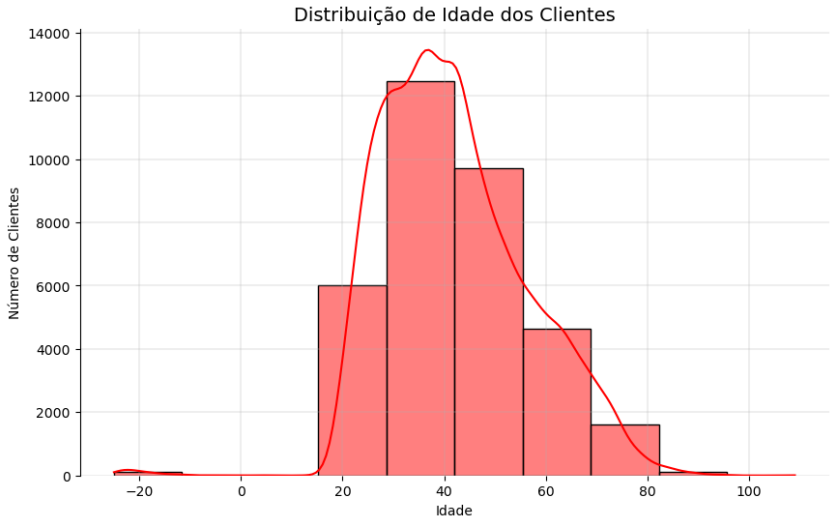

Figura 1 - Distribuição das idades dos clientes

Existem casos em que a idade está negativa. Uma análise sobre estes valores indicou que podem ser facilmente tratados tomando-se o valor absoluto das idades.

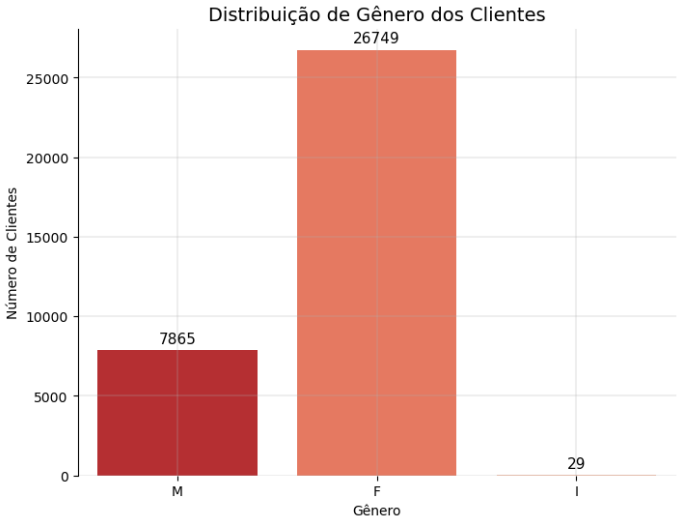

Figura 2 - Distribuição do Gênero dos clientes

Há grande predominância de clientes do sexo feminino, o que era esperado considerando as tendências de clientes do negócio da organização parceira.

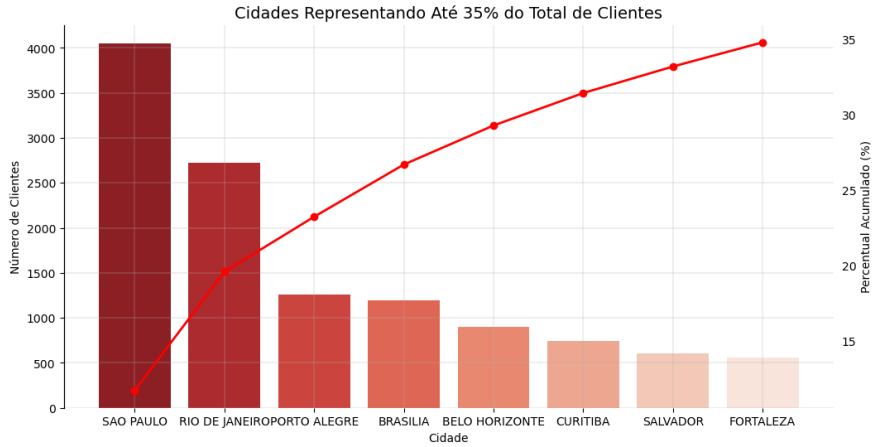

Figura 3 - Número de Clientes por cidade

De todo o conjunto de dados, São Paulo e Rio de Janeiro representam 20% do número total de clientes. Adicionando-se Porto Alegre e Brasília obtém-se pouco mais de 25%, o que demonstra como as duas primeiras cidades possuem grande representatividade no dataset.

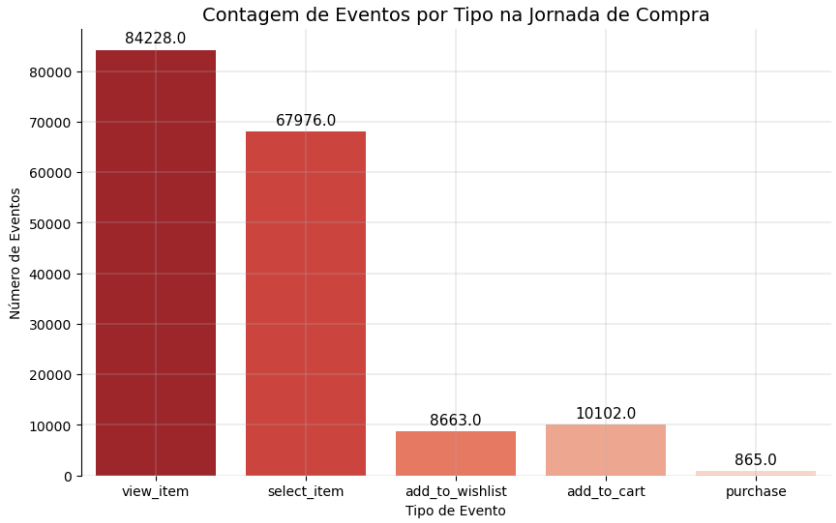

Figura 4 - Contagem de Eventos Online

No conjunto de dados de navegação, é predominante a presença dos eventos view_item e select_item, tendo os demais eventos poucos registros.

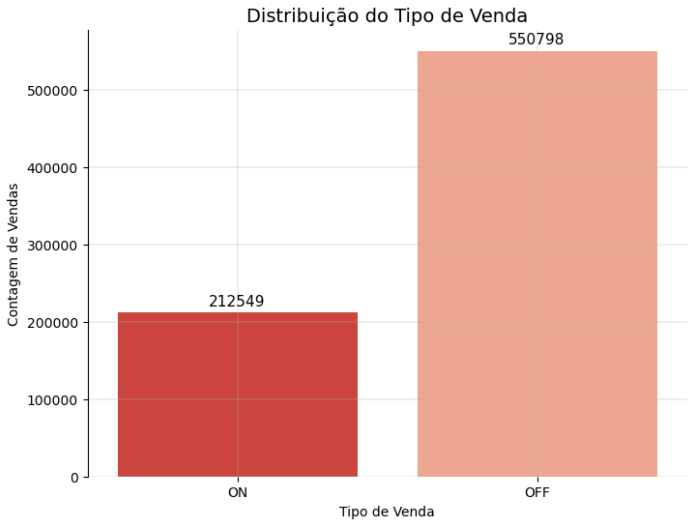

Figura 5 - Contagem da Quantidade de Vendas por Tipo (Online ou Offline)

Há grande diferença de registros quando comparadas compras Online e Offline. As vendas online representam em torno de um terço do total de vendas.

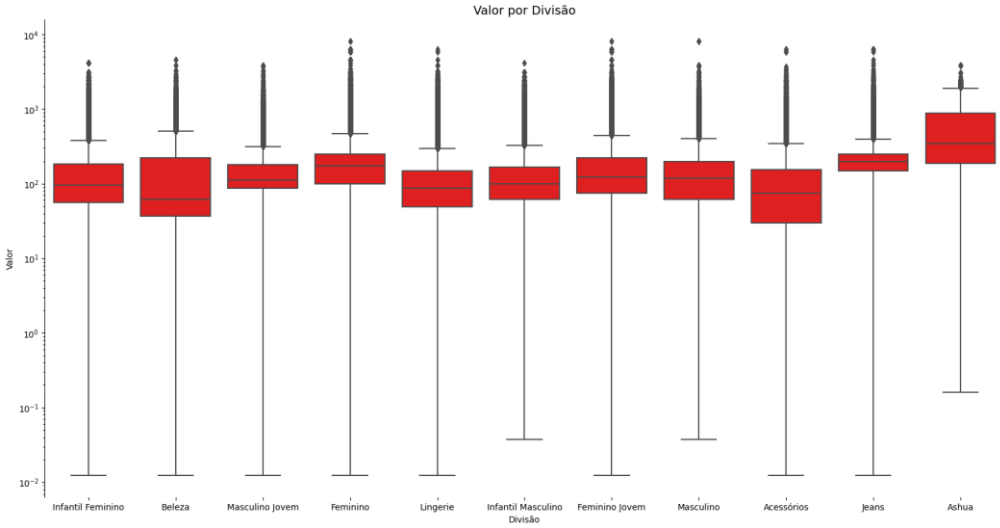

Figura 6 - Análise do valor vendido por categoria de produto

Quando avaliado os valores das vendas por categoria, percebemos que há certo equilíbrio nos dados, ainda que existam muitos outliers que deverão ser tratados.

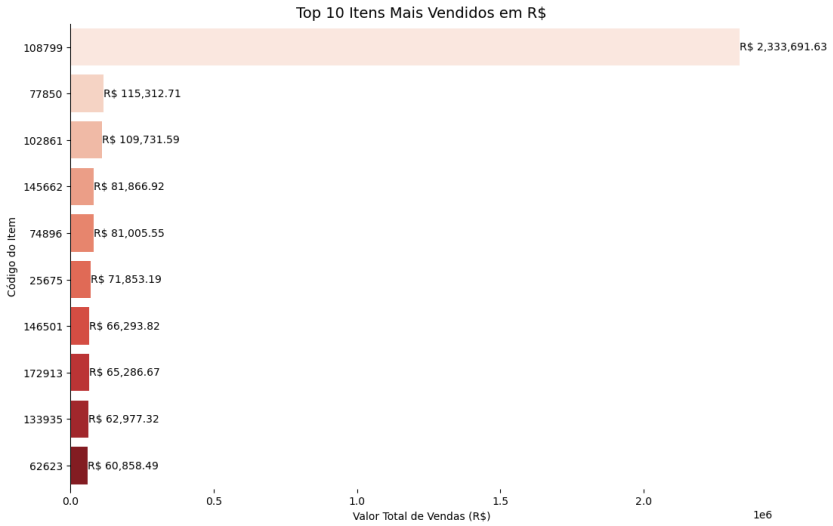

Figura 7 - 10 itens com maior total de vendas em reais

Avaliando os 10 itens com maior valor total em vendas, podemos perceber que há um item com código 108799 que possui valor total acima de 2 milhões de reais. O segundo item que mais vende (Código 77850) possui um total de pouco mais de 115 mil reais. Supõe-se que possa haver algum tipo de erro para o primeiro caso.

Verificação de qualidade dos dados

Critérios utilizados para avaliar a qualidade dos dados:

- Dados faltantes: Na tabela de clientes, encontramos linhas com a coluna idade sem valor. Para a tabela de navegação, encontramos linhas com a coluna id_cliente sem valor, o que impossibilita a ligação destas linhas com as demais tabelas - não identifica o cliente. Já para a tabela de transação, linhas com a coluna nome_divisao sem valor foram encontradas.
- Coerência dos dados: Na tabela de clientes, encontramos linhas com a coluna idade negativa ou muito próxima de zero, e idades muito altas (> 100 anos), na coluna gênero o valor ‘i’ foi encontrado em algumas linhas.
- Completude dos dados: A grande maioria dos dados pode ser considerada completa e útil para as abordagens apresentada pelo grupo.
- Diversidade: Analisando as três tabelas de modo geral, a quantidade de features que nos ajudam a entender o problema e encontrar uma solução foi considerada baixa pela equipe. Acreditamos que uma análise usando dados externos será necessária.

3. Preparação dos Dados

Limpeza dos dados

Para os dados em que a idade estava negativa, tomamos o valor absoluto destas idades e a distribuição dos valores ficou conforme abaixo:

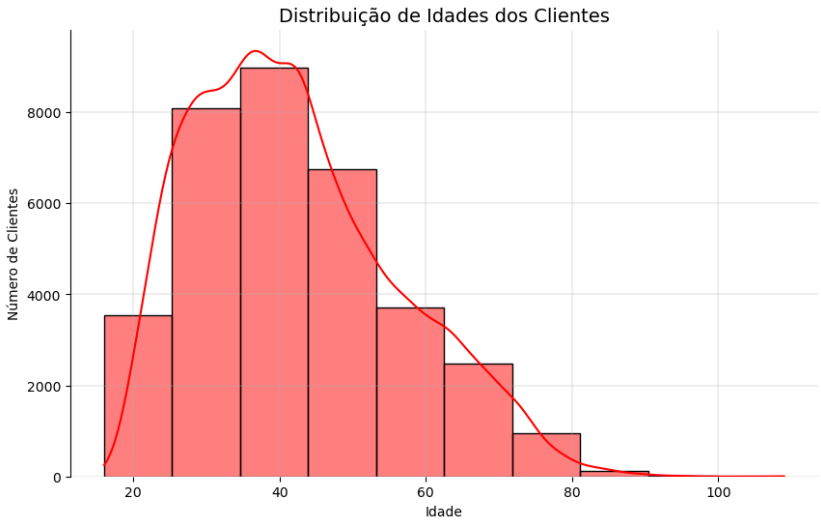

Figura 8 - Distribuição das idades após tratamento

Verificamos que existiam 23 clientes cuja data da primeira compra na Renner estava registrada como posterior a data da última compra. Fizemos a inversão dos registros para corrigir o problema.

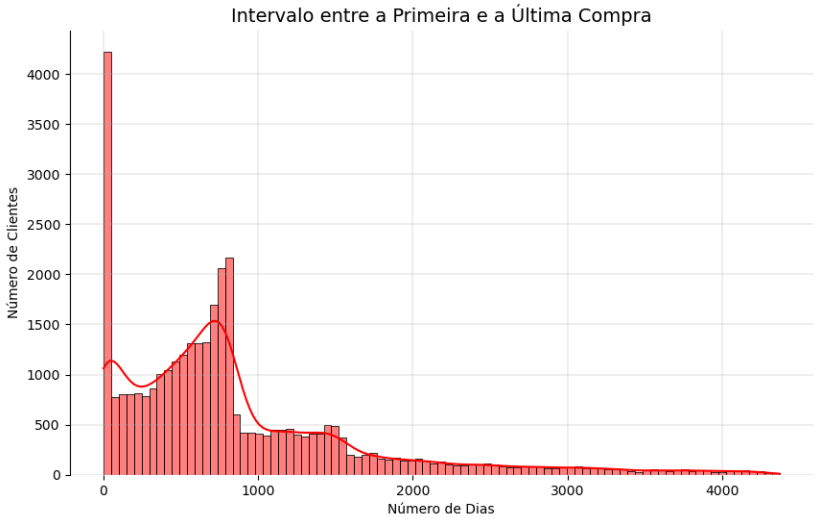

Figura 9 - Distribuição do número de dias entre a primeira e a última compra registrada

Criação de atributos e registros

Foi verificado que um mesmo item possuía diversos valores de venda e para tentar entender melhor esse comportamento, criamos um dataset auxiliar em que capturamos o preço mínimo, médio, máximo e a moda do preço para cada item. Com essas informações, pudemos calcular o desvio padrão, amplitude e coeficiente de variação de cada um dos itens, totalizando 186.127 itens.

Com objetivo de reduzir outliers, aplicamos as seguintes regras em sequência:

- Descartar Itens com moda do preço menor do que R$1,00.
- Descartar itens com desvio padrão calculado a partir de 1,5.

Na base de clientes criamos métricas para cada indivíduo, com dados de contagem de comprar por modalidade e por tipo de produto. Calculamos a diferença entre a primeira e a última compra, total de compras, tempo médio entre compras e ticket médio. Além disso, fizemos uma classificação se este cliente é de uma capital ou não.

Para os registros de navegação, aplicamos somente dados de contagem de cada evento por cliente.

Integração de dados

Como recebemos bases com 3 perspectivas diferentes (Cliente, Navegação, Transação), utilizamos como chave primária em todas elas o campo id_cliente, dessa forma, foi possível unir todas as tabelas e criar um dataset completo com foco nos clientes, seus dados e características agregadas.

Após as junções, totalizamos 23.664 clientes diferentes.

Descrição do dataset final

O conjunto de dados final ficou com 23.664 linhas (clientes diferentes) e 31 colunas. São elas:

- id_cliente
- gênero
- idade
- bairro
- cidade
- data_ultima_compra_renner
- data_primeira_compra_renner
- capital
- intervalo_pri_ult_compra
- total_compras
- compras_ON
- compras_OFF
- ticket_medio
- produtos_diferentes
- intervalo_medio
- Acessórios
- Ashua
- Beleza
- Feminino
- Feminino Jovem
- Infantil Feminino
- Infantil Masculino
- Jeans
- Lingerie
- Masculino
- Masculino Jovem
- add_to_cart
- add_to_wishlist
- purchase
- select_item
- view_item

4. Modelagem

Ao longo desta seção, devem ser descritas técnicas de modelagem escolhidas e utilizadas para atingir os objetivos de negócios e de mineração, porém com uma perspectiva de dados. Além de ser um processo iterativo, também é importante salientar que é importante descrever mesmo os modelos que ao final do projeto não serão aproveitados, pois o aprendizado pode ser relevante para escolhas em projetos futuros.

Técnicas e suposições de modelagem

Descrição das suposições sobre os dados e suposições sobre as técnicas de modelagem utilizadas.

Projeto de testes e experimentos

Projeto de como os modelos serão **avaliados** em uma perspectiva de **Ciência de Dados**. Descrição da configuração dos experimentos: quais dados serão utilizados em qual/quais modelo(s), como o modelo será testado ou avaliado, plano para criação de dados de teste (se houver), descrição do planejamento de como especialistas do domínio avaliarão os resultados etc.

Descrição dos modelos

Visão geral dos modelos produzidos e dos processos que levaram à sua produção. Descrição dos tipos dos modelos e a relação deles com os objetivos de mineração. Configuração de parâmetros utilizada. Descrição **detalhada** dos modelos. Descrição do comportamento e da interpretação dos modelos.

Avaliação dos modelos

Resultados dos experimentos realizados de acordo com o que foi projetado na etapa de **projeto de testes e experimentos**.

5. Avaliação

Nesta seção, os resultados são avaliados sob a **perspectiva de negócio**. Para isto, os objetivos de negócio e os critérios de sucesso são revisitados e analisados em face os resultados encontrados.

Avaliação dos resultados do projeto

Comparação detalhada entre cada critério de sucesso de objetivo de negócio e resultados encontrados na modelagem. Conclusão geral sobre o

Revisão do processo e conclusões gerais

Reflexões gerais sobre o processo realizado ao longo do projeto. Os objetivos de negócio foram atingidos? Existem novos objetivos de negócio ao final do projeto? Conclusões sobre projetos futuros.

6. Autocrítica

Nas primeiras semanas o ritmo do grupo foi prejudicado pela demora no envio dos dados por parte da empresa parceira.

Vencida essa fase, o grupo está alinhado e conseguindo avançar nas atividades propostas durante a semana. Como ponto positivo destacamos a participação do grupo no período de aula, onde todos os integrantes estão presentes para participar das discussões e alinhamentos para a próxima semana de trabalho.

Há dificuldade em antecipar atividades, impedindo que haja algum tipo de margem nas datas de entrega, beirando algum tipo de atras, mas que não tem se refletido nas entregas.

Dado o cenário, comprometimento do grupo e proposta alinhada com a Renner, entendemos que o grupo conseguirá entregar 100% do que foi planejado.

Nota até o momento: **8,0**