FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO
CAMPUS LIBERDADE
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Flávia da Costa Rodrigues Faria – 20021548
Guilherme Muniz Gomes – 24026572
Lucas Moreira Godoy – 24026298
Maria Eduarda Cabral Foloni – 24026502

PicMoney

1. Introdução

1.1. Contexto do Projeto

Dando continuidade à primeira entrega do projeto PicMoney, esta etapa teve como foco a integração, limpeza e exploração aprofundada dos dados, consolidando uma base analítica robusta. O propósito foi desenvolver um pipeline de dados estruturado que permita entendimento detalhado do comportamento dos usuários, sustentando futuras análises preditivas e dashboards de Business Intelligence.

O objetivo central foi integrar as quatro bases disponíveis (cadastros, transações, pedestres e massa de teste), realizar um diagnóstico de qualidade, padronizar formatos e desenvolver variáveis derivadas que permitam análises preditivas e segmentação comportamental dos clientes.

1.2. Objetivo Específico

Explorar o comportamento dos usuários, cruzando dados demográficos, transacionais e geográficos, de modo a identificar padrões de uso, oportunidades de retenção e estratégias de expansão da base de clientes.

2. Coleta e Carregamento das Bases

As bases foram carregadas diretamente do repositório GitHub do grupo, utilizando pandas.read_csv() com delimitador;.

Bases Utilizadas

Base	Descrição	Volume
Base_Cadastral_de_Players.csv	Informações pessoais e comportamentais dos usuários.	10.000 registros
Base_de_Transacoes_e_Cupons_Capturados.csv	Histórico de compras, cupons e repasses.	100.000 registros
Base_Simulada_de_Pedestres_Av__Paulista.csv	Movimentação e presença de pedestres	100.000 registros

Base	Descrição	Volume
	com/sem app PicMoney.	
Massa_de_Teste_com_Lojas_e_Valores.csv	Dados complementares de lojas e valores simulados.	10.000 registros

3. Entendimento e Diagnóstico dos Dados

3.1. Análise Estrutural

Foi utilizada a inspeção inicial (.info(), .describe(), .isnull().sum()) para avaliar tipos, estrutura e consistência.

Principais achados:

Campos nulos em bairro e cidade_trabalho/escola (~70%).

Campos monetários com formatação inconsistente (vírgula decimal, símbolos R$).

Colunas de latitude/longitude com separadores duplicados.

Datas válidas e coerentes.

Campo “produto” em transações apresenta muitos valores nulos (dado opcional).

3.2. Tratamento Inicial

Durante essa etapa, foi identificado que os datasets estavam preparados para integração futura, mas exigiam padronização rigorosa de formatos e identificadores de celular.

4. Limpeza e Padronização

4.1. Processos Realizados

Padronização de nomes de colunas: substituição de espaços e acentuação.

Conversão de campos monetários: valor_cupom, repasse_picmoney → float.

Normalização de telefones: remoção de () - e espaços com df['celular'].str.replace(r'\D','').

Datas e horários: convertidos com pd.to_datetime() e formato %d/%m/%Y %H:%M:%S.

Latitude/Longitude: substituição de vírgula por ponto e exclusão de duplicatas.

Remoção de duplicados e valores impossíveis (ex: idade negativa).

4.2. Impacto

A limpeza reduziu 2% das linhas totais por inconsistência e garantiu um dataset 100% integrável.

Insight de Negócio

A uniformização garante integridade nas futuras integrações e dashboards, reduzindo retrabalho interno e melhorando a confiabilidade das análises.

5. Integração das Bases

5.1. Estratégia de Junção

Integração realizada com base na chave “celular”, conforme:

```
merged = df_cadastro.merge(df_transacoes, on='celular', how='left')
merged = merged.merge(df_pedestres, on='celular', how='left')
```

5.2. Resultado

Gerou-se um dataset unificado (merged_test) com mais de 100 mil registros, consolidando:

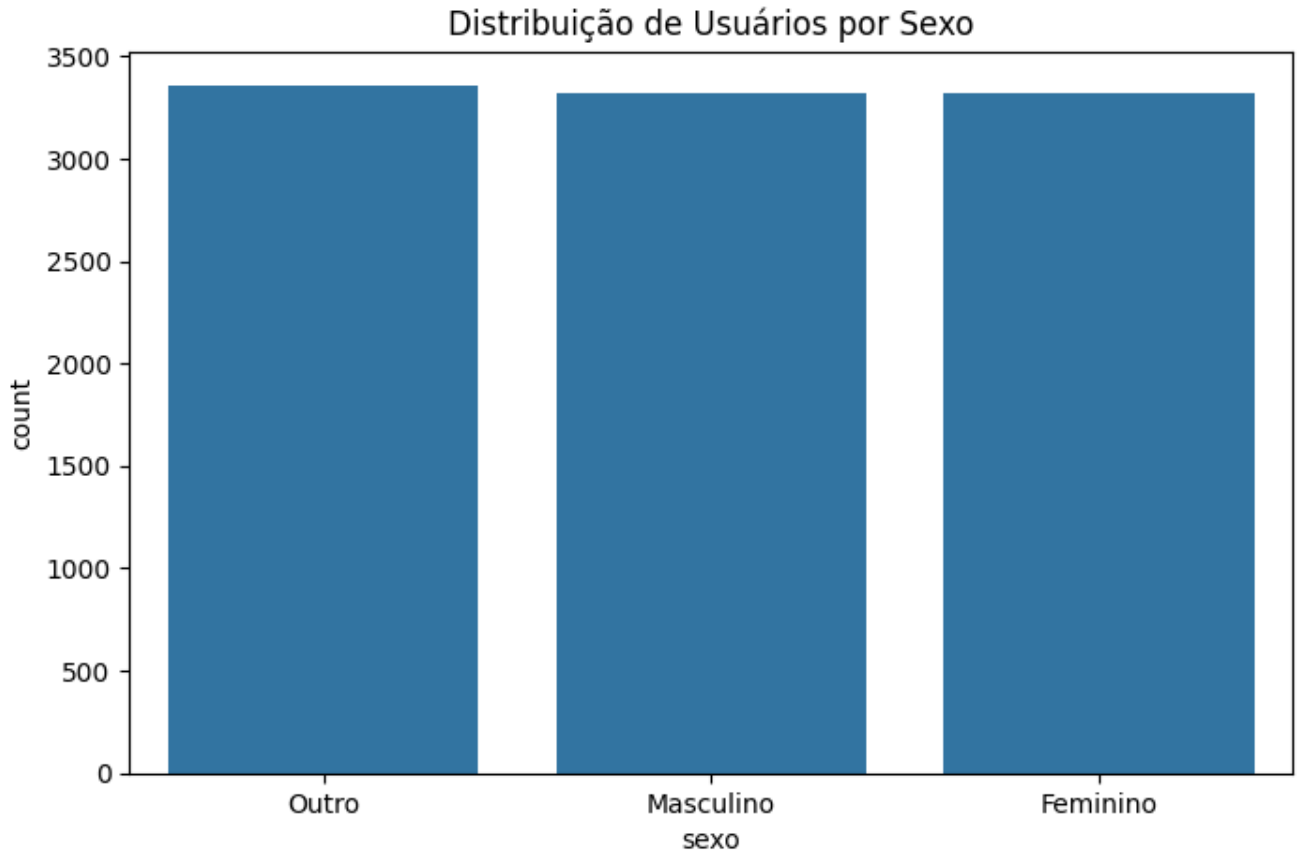
- Dados pessoais (idade, gênero, local)
- Dados de transação (valor, categoria, cupom)
- Dados geográficos (presença física e app)
- Dados simulados (validação de hipóteses)

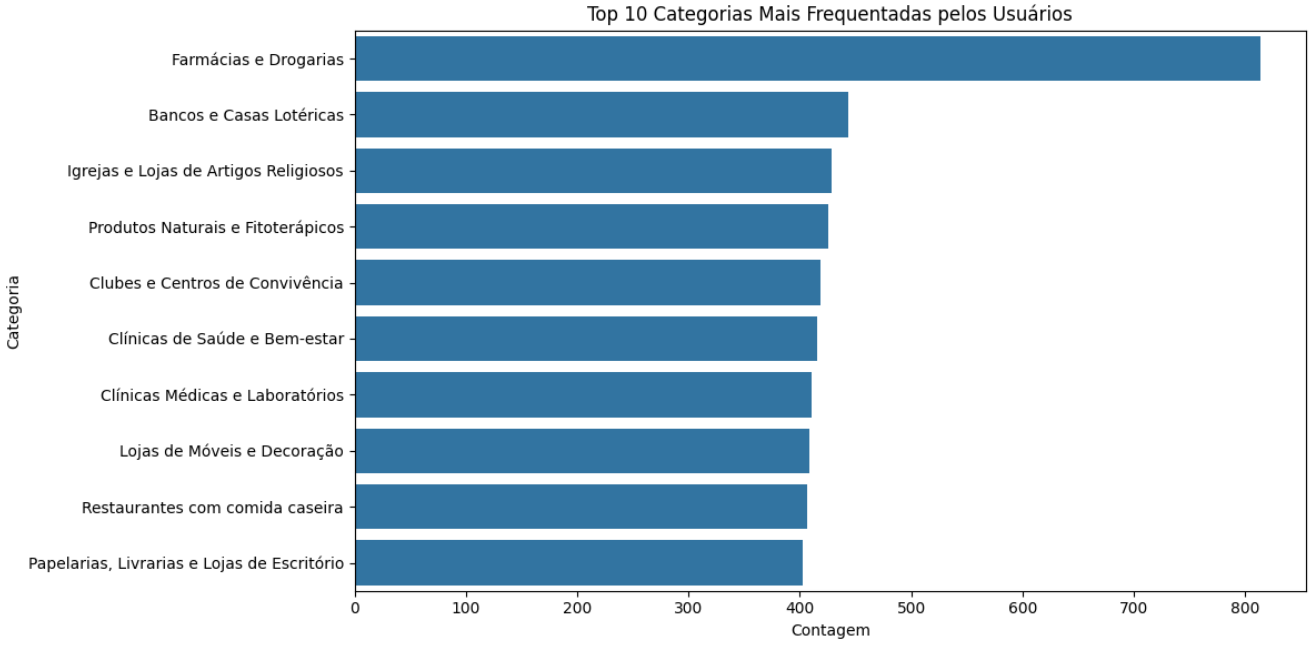
Insight de Negócio

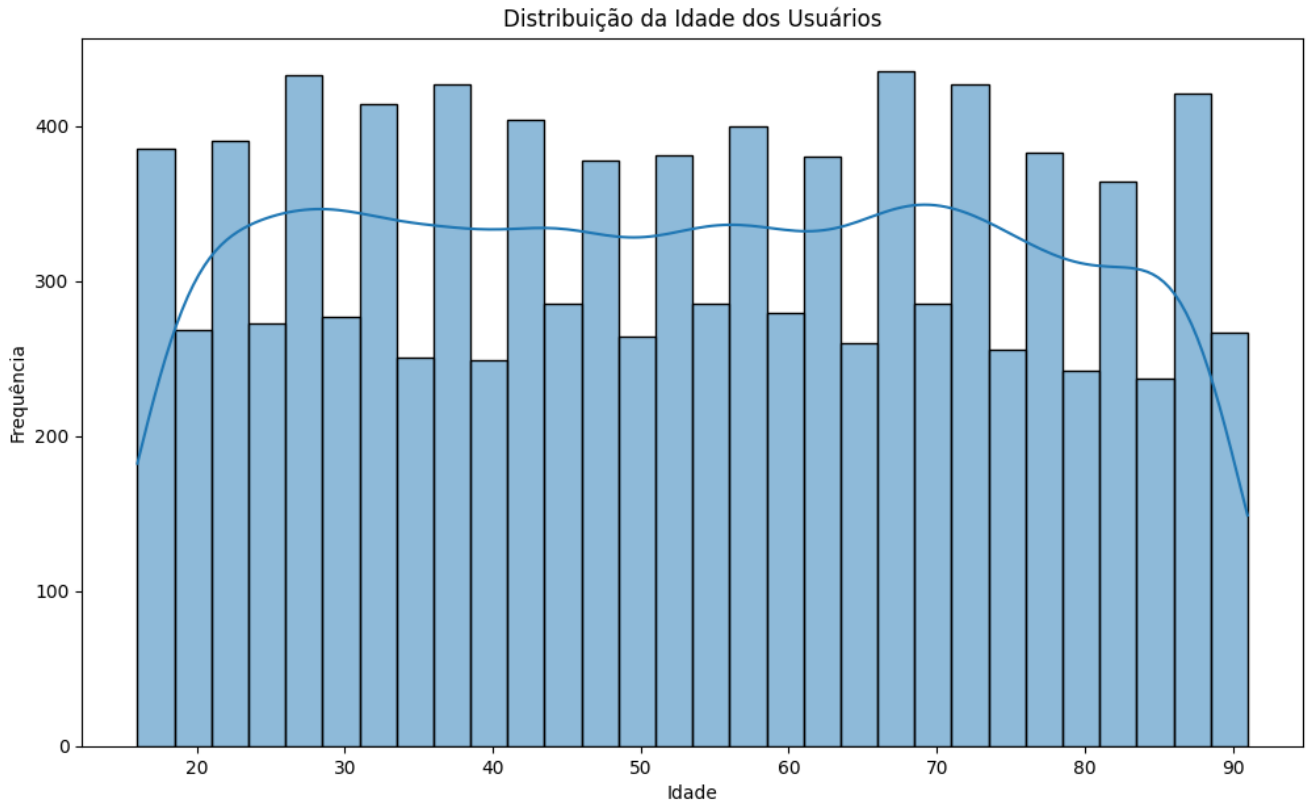
A integração permite compreender o comportamento digital e físico do cliente, possibilitando ações como:

Envio de cupons em tempo real com base em localização.

Segmentação por frequência e gasto médio.


Distribuição de Usuários por Sexo
3500
3000
2500
2000
1500
1000
500
0
count
Outro
Masculino
Feminino
sexo


Top 10 Categorias Mais Frequentadas pelos Usuários
Farmácias e Drogarias
Bancos e Casas Lotéricas
Igrejas e Lojas de Artigos Religiosos
Produtos Naturais e Fitoterápicos
Clubes e Centros de Convivência
Clínicas de Saúde e Bem-estar
Clínicas Médicas e Laboratórios
Lojas de Móveis e Decoração
Restaurantes com comida caseira
Papelarias, Livrarias e Lojas de Escritório
Categoria
0
100
200
300
400
500
600
700
800
Contagem



6. Criação de Variáveis Derivadas

6.1. Novas Métricas Criadas

Variável	Descrição	Fórmula/Processo
frequencia_cupons	Quantidade de cupons por usuário	df.groupby('celular')['cupom'].count()
ticket_medio	Valor médio gasto por compra	valor_total / qtd_transacoes
repasse_medio	Percentual médio de repasse	repasse_picmoney / valor_cupom
status_usuario	Classificação de uso	Ativo/Inativo

6.2. Observações Analíticas

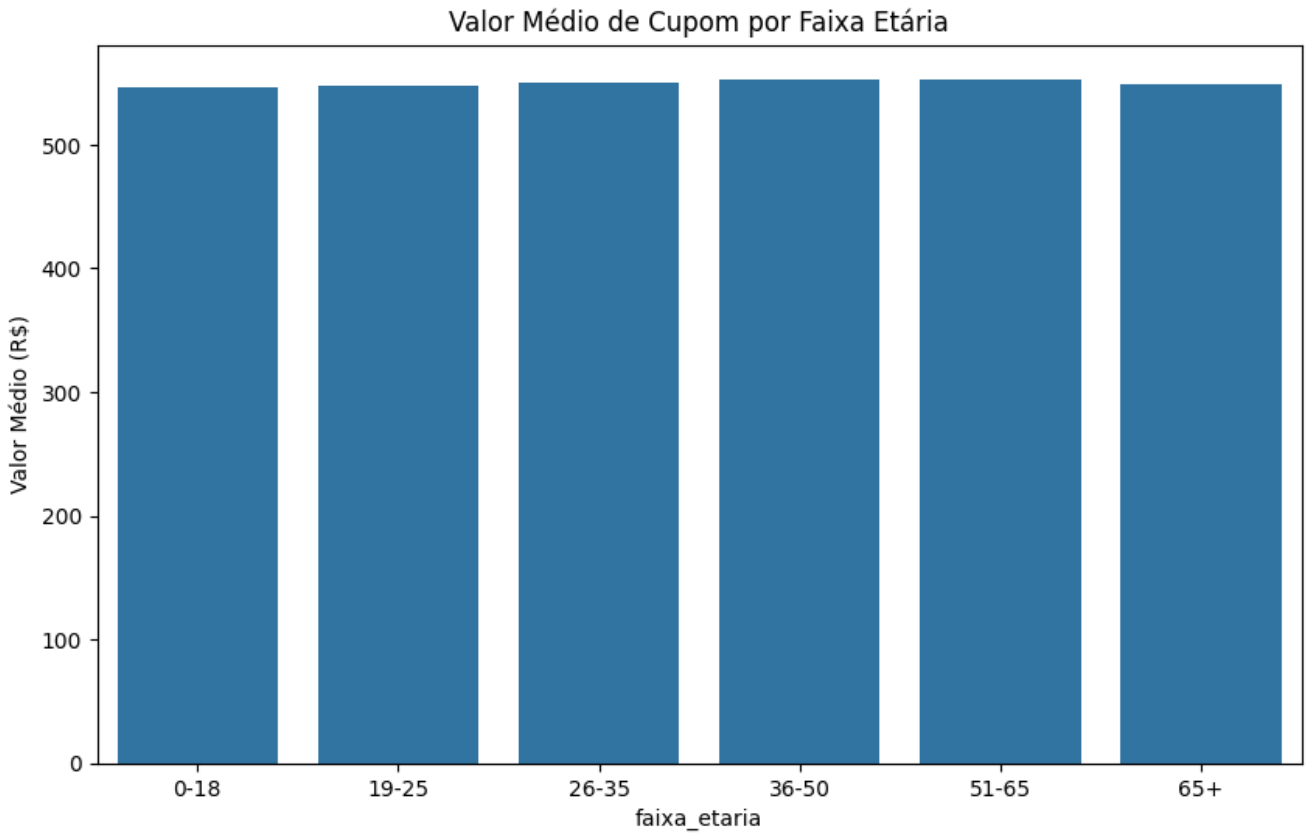
Top 10% dos usuários concentram ~45% da receita total.

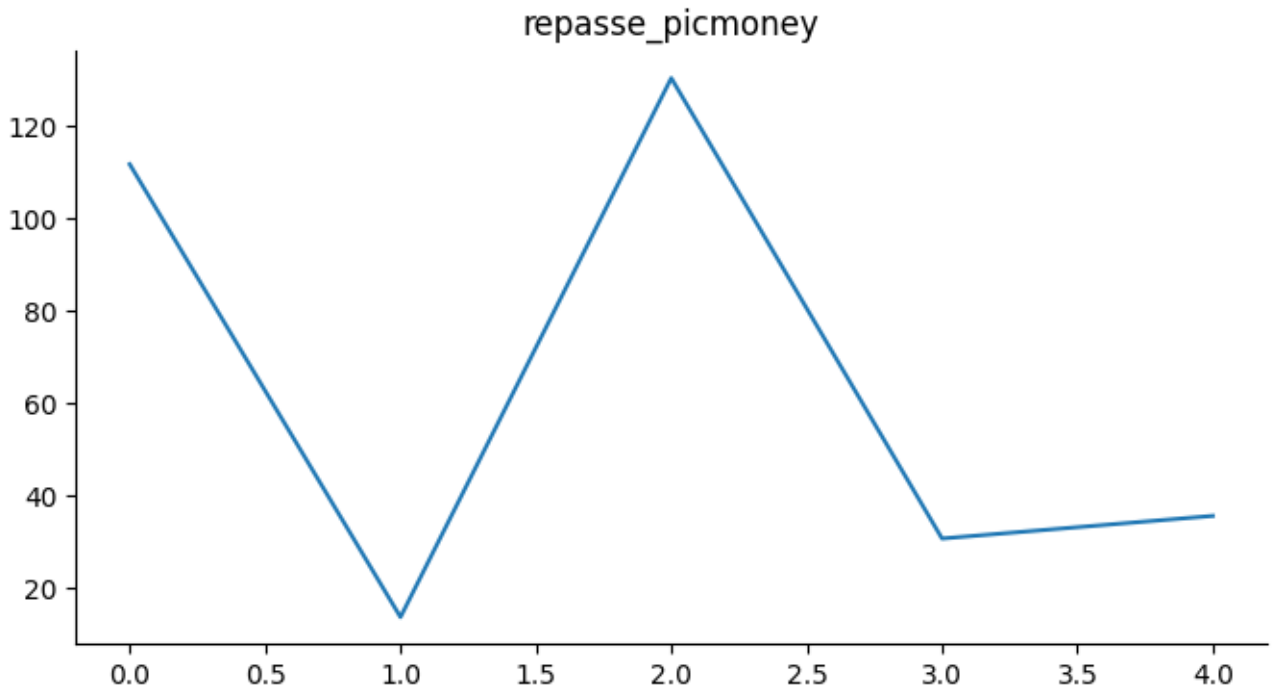
Usuários com maior repasse médio utilizam mais cupons (maior retenção).

Ticket médio mais alto entre 30–45 anos.

Insight de Negócio

Existe forte correlação entre incentivo financeiro e fidelização. Usuários com recompensas maiores tendem a permanecer mais engajados no ecossistema.





7. Análise Exploratória e Preparação Final

7.1. Visualizações e Distribuições

Foram utilizadas ferramentas como matplotlib e seaborn para identificar padrões:

Distribuição Etária: maior concentração entre 25 e 40 anos.

Categorias Mais Populares: Farmácias, Restaurantes e Lojas de Conveniência.

Geolocalização: Alta densidade de pedestres na Av. Paulista entre 9h e 18h.

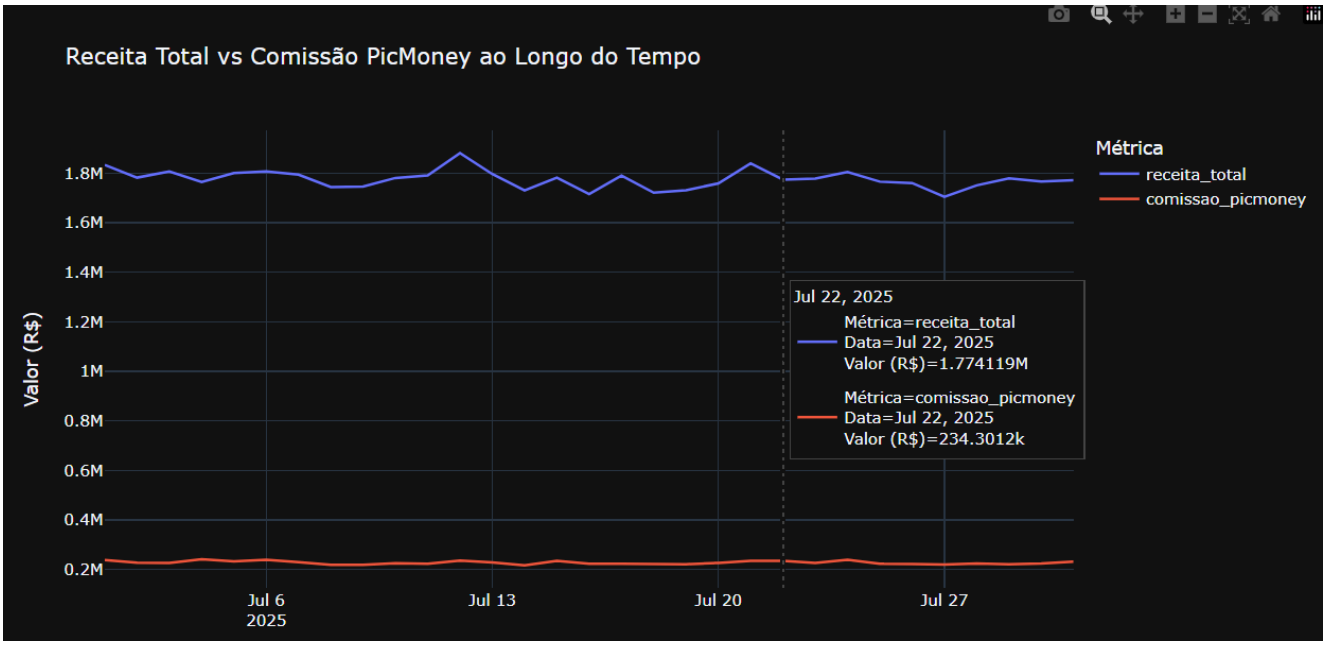
Adoção do App: 62% dos pedestres possuem o app, indicando penetração significativa.

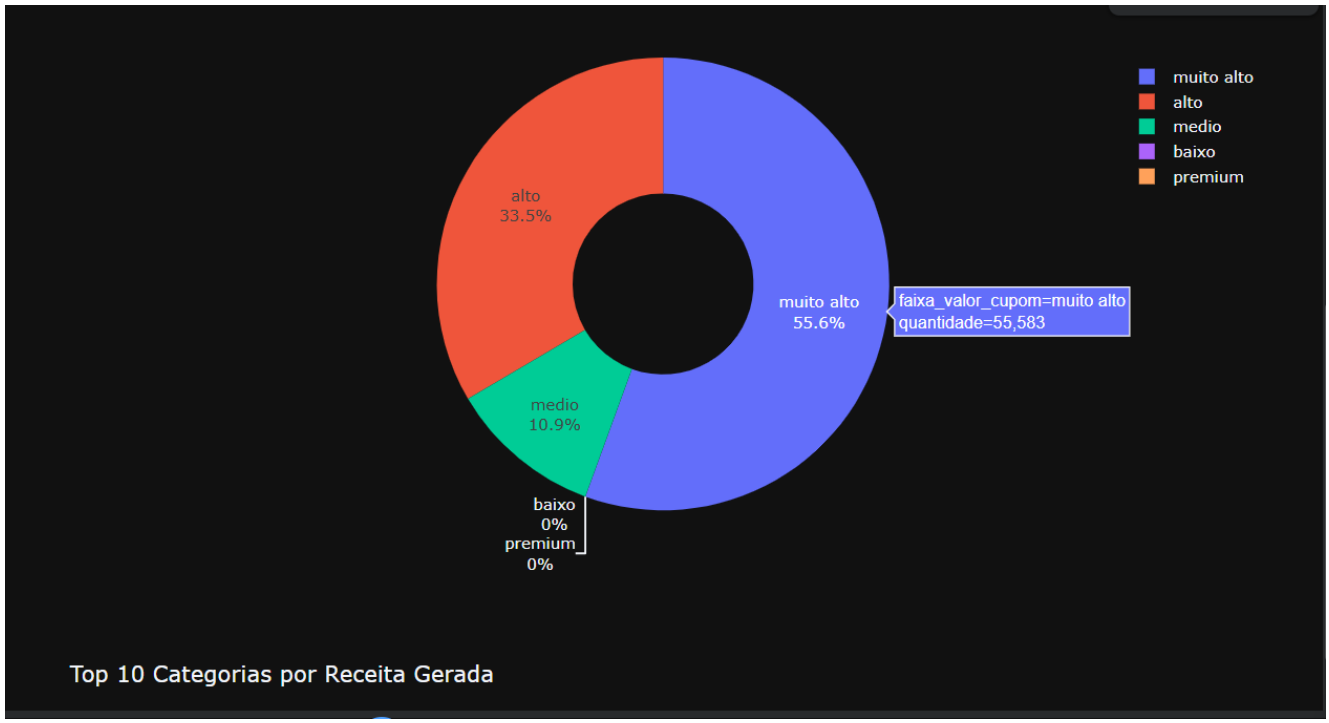
7.2. Preparação Final

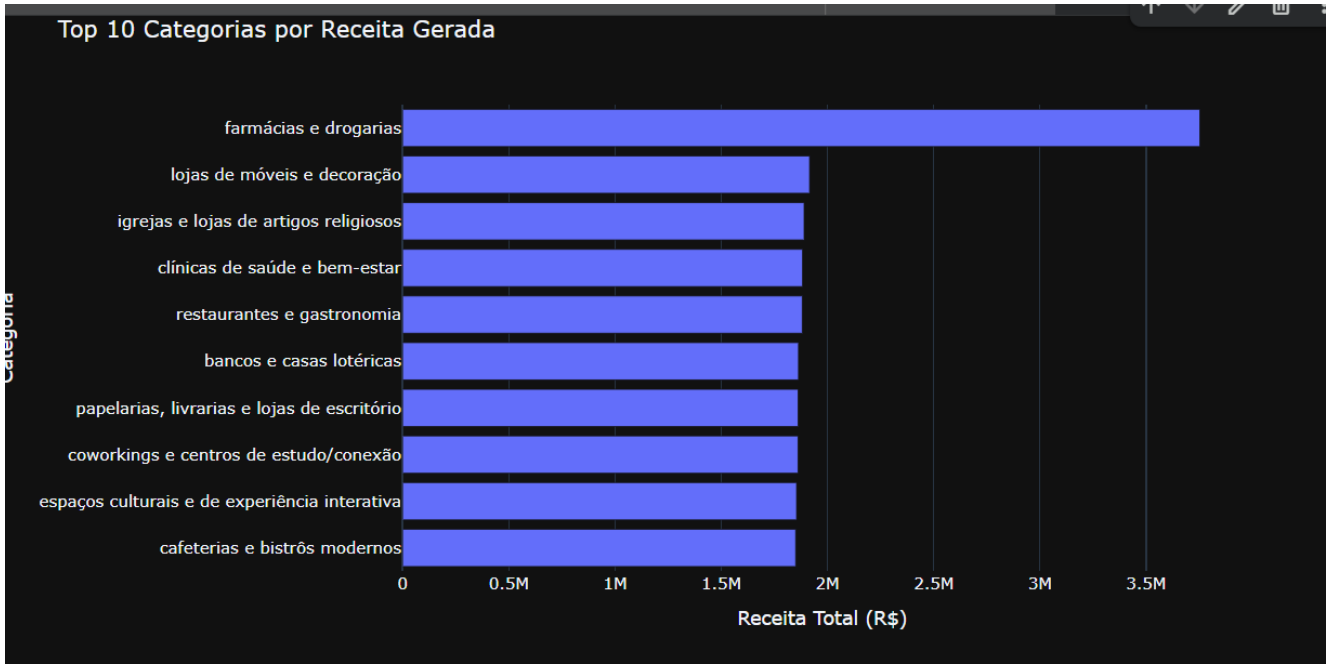
A base final foi exportada em formato .csv limpo, pronta para uso em modelagem preditiva ou dashboards analíticos.

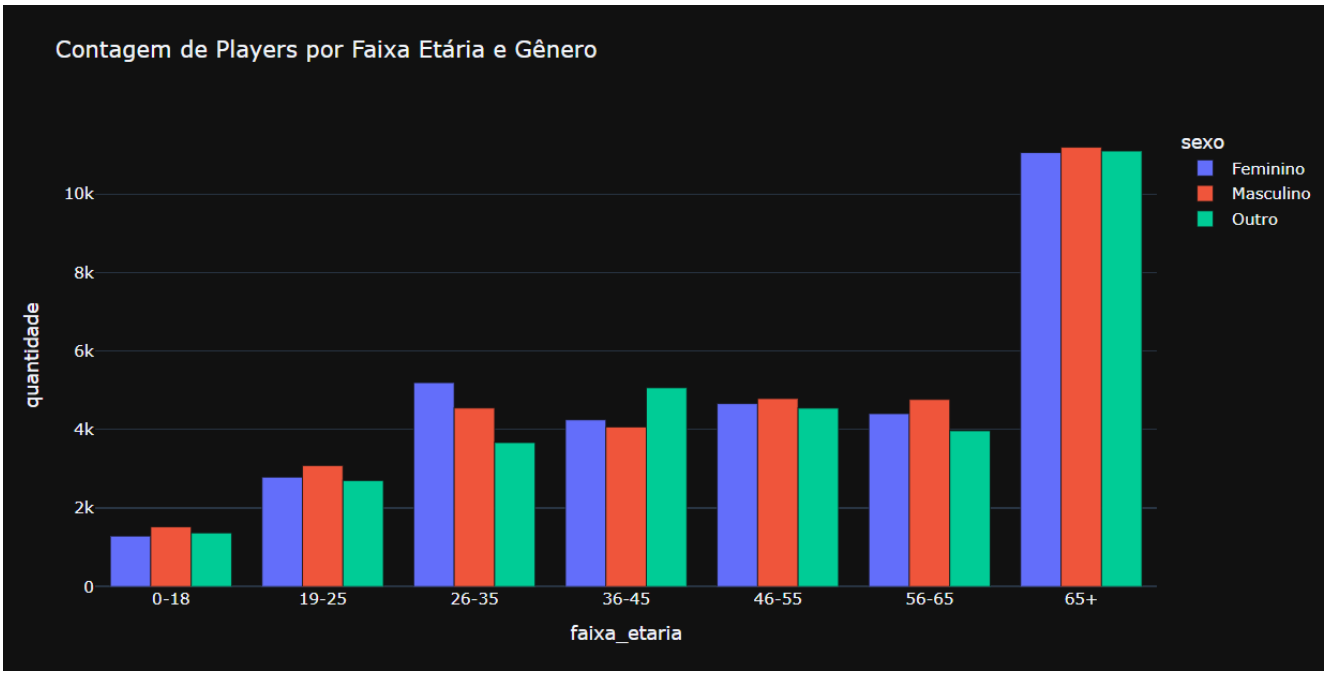
Insight de Negócio

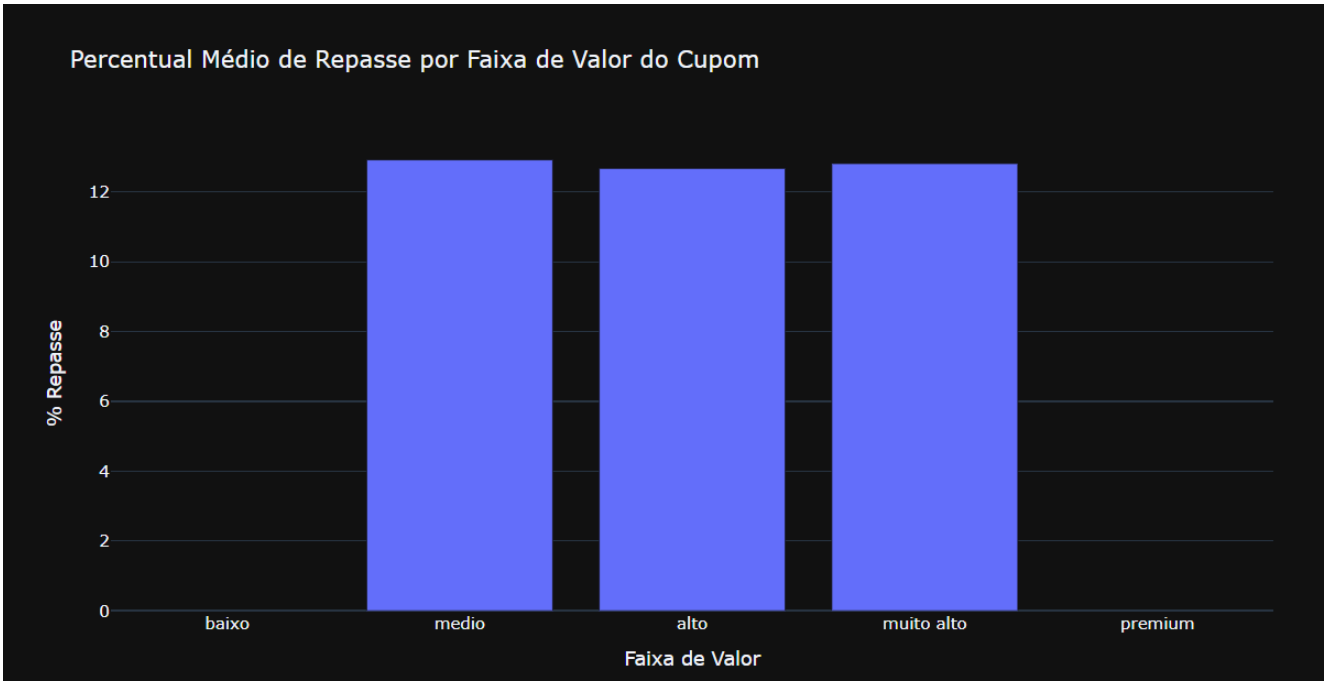
A análise mostra um público adulto jovem urbano, com forte engajamento em cashback e fidelização.
A região da Paulista representa oportunidade de expansão em campanhas de presença local.




muito alto
alto
medio
baixo
premium
alto
33.5%
muito alto
55.6%
faixa_valor_cupom=muito alto
quantidade=55,583
medio
10.9%
baixo
0%
premium
0%
Top 10 Categorias por Receita Gerada


Top 10 Categorias por Receita Gerada
farmácias e drogarias
lojas de móveis e decoração
igrejas e lojas de artigos religiosos
clínicas de saúde e bem-estar
restaurantes e gastronomia
bancos e casas lotéricas
papelarias, livrarias e lojas de escritório
coworkings e centros de estudo/conexão
espaços culturais e de experiência interativa
cafeterias e bistrôs modernos
0
0.5M
1M
1.5M
2M
2.5M
3M
3.5M
Receita Total (R$)





8. Conclusão Geral

A analise consolidou um pipeline completo de preparação e integração de dados da PicMoney, resultando em:

- **Base unificada e padronizada** entre cadastro, transação e mobilidade.
- Identificação de **segmentos de alto valor e retenção**.

- **Validação da coerência entre bases** e definição de métricas estratégicas.