Projeto Integrador - Análise de Dados PicMoney

Seguindo o modelo CRISP-DM

Integrantes do Grupo:

Nicolas Morales / 24025897,
Eduardo Chen Zou / 24025817,

Fabiano Henrique Chou / 24025991,

Bernardo Seijas Cavalcante 24026290.

Data: [22/09/2025]

Introdução

Este relatório apresenta a Análise de dados da PicMoney utilizando o modelo CRISP-DM. O objetivo é coletar, descrever, explorar e avaliar a qualidade das bases fornecidas, preparando terreno para futuras análises e modelos preditivos.

De acordo com a metodologia CRISP-DM, o ciclo de vida de mineração de dados é dividido em seis fases: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Neste trabalho, o foco concentrou-se até a etapa de Verificação da Qualidade dos Dados, preparando a base para fases futuras.

Etapa 1 – Coleta de Dados

Foram utilizadas quatro bases de dados distintas:

- **Base de Transações – Cupons Capturados** (100.000 linhas, 12 colunas).
- **Base Simulada – Pedestres Av. Paulista** (100.000 linhas, 15 colunas).
- **Massa de Teste – Lojas e Valores** (10.000 linhas, 11 colunas).
- ***Base Cadastral de Players*** *(10.000 linhas, 11 colunas)*

Durante a coleta, foi necessário realizar um processo de **preparação e tratamento inicial dos arquivos Excel**, pois alguns campos vinham desconfigurados. Identificamos problemas como:

- **Formato de datas e horários**: apareciam em padrões diferentes (DD/MM/AAAA e AAAA-MM-DD, além de horários em texto), exigindo conversão para um formato unificado.
- **Idioma/região do Excel**: os arquivos estavam com regionalização incorreta, fazendo com que vírgulas e pontos fossem interpretados de forma errada nos valores numéricos. Atualizamos as configurações de idioma e realizamos a padronização.
- **Caracteres especiais e acentuação**: alguns registros vinham com símbolos trocados devido à codificação. Foi feito o ajuste para UTF-8 e a revisão da acentuação.

Após esses ajustes, os dados foram considerados prontos para as próximas etapas da metodologia CRISP-DM, garantindo consistência mínima e qualidade suficiente para as análises descritivas e exploratórias.

Coleta de Dados no Colab

```python
import pandas as pd

base_cupons = pd.read_excel("PicMoney-Base_de_Transa__es_-_Cupons_Capturados-100000 linhas (1).xlsx")
base_pedestres = pd.read_excel("PicMoney-Base_Simulada_-_Pedestres_Av__Paulista-100000 linhas (1).xlsx")
base_lojas = pd.read_excel("PicMoney-Massa_de_Teste_com_Lojas_e_Valores-10000 linhas (1).xlsx")
base_players = pd.read_excel("PicMoney-Base_Cadastral_de_Players-10_000 linhas (1).xlsx")

print("Cupons:", base_cupons.shape)
print("Pedestres:", base_pedestres.shape)
print("Lojas:", base_lojas.shape)
print("Players:", base_players.shape)

base_cupons.head()
```

```
Cupons: (100000, 12)
Pedestres: (100000, 15)
Lojas: (10000, 11)
Players: (10000, 11)
```

	celular	data	hora	nome_estabelecimento	bairro_estabelecimento	categoria_estabelecimento	id_campanha	id_cupom	tipo_cupom	produto	valor_cupom	repasse_picmoney
0	(61) 96497-8673	2025-07-10	16:15:00	Habib's	República	Lojas de Eletrônicos e Games	CAM2768	CUP542835	Cashback	Não tem	229.64	11.48
1	(11) 94231-6424	2025-07-15	08:15:00	Smart Fit	Vila Prudente	Lojas de Eletrônicos e Games	CAM6679	CUP291620	Cashback	Não tem	356.33	17.82
2	(11) 97965-2178	2025-07-20	16:45:00	Outback	Tucuruvi	Igrejas e Lojas de Artigos Religiosos	CAM6473	CUP670811	Produto	Tempora	719.06	27.61
3	(11) 93418-4646	2025-07-20	15:45:00	Subway	Penha	Fisioterapia e Terapias Complementares	CAM8293	CUP590364	Produto	Magnam	798.34	25.85
4	(11) 97973-1725	2025-07-07	11:00:00	Octavio Café	Santo Amaro	Clínicas Médicas e Laboratórios	CAM5588	CUP528033	Produto	Quam	718.45	28.85

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=F4IKK8M0-Hsc&line=8&uniqifier=1

O código acima foi utilizado para realizar a coleta das quatro bases de dados no ambiente Colab. Cada arquivo Excel foi lido utilizando a biblioteca Pandas, e a função shape retornou o número de registros e colunas de cada conjunto. Em seguida, o comando head() exibiu as cinco primeiras linhas da base de cupons, confirmando que os dados foram carregados corretamente e possuíam colunas como celular, data, hora, nome do estabelecimento, categoria e valor do cupom.

Etapa 2 – Descrição dos Dados

Nesta etapa, foram analisadas estatísticas descritivas como média, mediana, Desvio padrão, valores mínimos e máximos.

Também foram identificadas variáveis categóricas (ex.: sexo, tipo de cupom) e numéricas (ex.: idade, valores de compra)

```
print("----> MÉDIA <----")
print("Idade média:", base_players["idade"].mean())
print("Valor médio de compra:", base_lojas["valor_compra"].mean())
print("Valor médio de cupom:", base_cupons["valor_cupom"].mean())
print("Repasse médio PicMoney:", base_cupons["repasse_picmoney"].mean())
```

```
----> MÉDIA <----
Idade média: 52.7935
Valor médio de compra: 549.6840749999999
Valor médio de cupom: 550.4895853
Repasse médio PicMoney: 70.4747391
```

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=M86KdFZ94MID&line=7&uniqifier=1

```
print("----> MEDIANA <----")
print("Idade mediana:", base_players["idade"].median())
print("Valor mediano de compra:", base_lojas["valor_compra"].median())
print("Valor mediano de cupom:", base_cupons["valor_cupom"].median())
print("Repasse mediano PicMoney:", base_cupons["repasse_picmoney"].median())
```

```
=== MEDIANA ===
Idade mediana: 53.0
Valor mediano de compra: 548.885
Valor mediano de cupom: 550.265
Repasse mediano PicMoney: 32.815
```

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=Q6cma5UL4QIJ&line=3&uniqifier=1

As medidas de tendência central — média e mediana — permitem compreender o comportamento típico dos dados.

Na base Players, a idade média foi de aproximadamente 53 anos, enquanto a mediana também foi de 53 anos, indicando uma distribuição equilibrada, sem grandes distorções causadas por valores extremos.

Na base Lojas, o valor médio de compra foi de cerca de R$ 550, com mediana muito próxima (R$ 549), o que mostra consistência nos gastos mais comuns dos clientes.

Já nos Cupons, tanto a média (R$ 550) quanto a mediana (R$ 550) apresentaram valores semelhantes, reforçando a regularidade das campanhas nesse aspecto.

Por outro lado, no repasse à PicMoney, observou-se uma discrepância: a média foi de R$ 70, enquanto a mediana ficou em torno de R$ 33. Isso revela que a maioria das transações gera repasses mais baixos, mas alguns casos específicos, de valor elevado, acabam puxando a média para cima.

```
print("----> DESVIO PADRÃO <---")
print("Desvio padrão da idade:", base_players["idade"].std())
print("Desvio padrão do valor de compra:", base_lojas["valor_compra"].std())
print("Desvio padrão do valor do cupom:", base_cupons["valor_cupom"].std())
print("Desvio padrão do repasse:", base_cupons["repasse_picmoney"].std())
```

```
=== DESVIO PADRÃO ===
Desvio padrão da idade: 21.568963344768523
Desvio padrão do valor de compra: 260.76200310787374
Desvio padrão do valor do cupom: 259.4101427179635
Desvio padrão do repasse: 90.8235194557047
```

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=gN2AYVkR4WSv&line=2&uniqifier=1

O desvio padrão mede a dispersão dos dados em relação à média.

Na base Players, o valor de 21,6 anos mostra que, apesar da média ser de 53, há uma variação ampla de idades, confirmando a presença de diferentes faixas etárias.

Nos valores de compra, o desvio padrão foi de aproximadamente R$ 261, revelando que alguns clientes gastam muito mais ou muito menos que a média de R$ 550.

Nos cupons, o desvio padrão de R$ 259 reforça a existência de grande variação entre os valores das campanhas.

Já no repasse, a dispersão foi de cerca de R$ 91, indicando que há estabelecimentos ou campanhas que contribuem de forma desproporcional para a receita da PicMoney

```
print("----> MODA <----")

print("Sexo (Players):", base_players["sexo"].mode().iloc[0])
print("Sexo (Pedestres):", base_pedestres["sexo"].mode().iloc[0])
print("Idade (Players):", base_players["idade"].mode().iloc[0])
print("Valor de compra (Lojas):", base_lojas["valor_compra"].mode().iloc[0])
print("Valor de cupom (Cupons):", base_cupons["valor_cupom"].mode().iloc[0])
print("Repasse PicMoney (Cupons):", base_cupons["repasse_picmoney"].mode().iloc[0])
print("Idade (Pedestres):", base_pedestres["idade"].mode().iloc[0])
```

```
----> MODA <----
Sexo (Players): Outro
Sexo (Pedestres): Outro
Idade (Players): 67
Valor de compra (Lojas): 186.36
Valor de cupom (Cupons): 371.39
Repasse PicMoney (Cupons): 7.76
Idade (Pedestres): 19
```

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=3B2CPBVU4f91&line=9&uniqifier=1

A moda representa o valor mais frequente em um conjunto de dados.

Na base Players, o sexo mais comum foi classificado como “Outro”, enquanto a idade modal foi de 67 anos, sugerindo que há uma concentração de usuários nessa faixa etária, mesmo que a média seja mais baixa (≈53 anos).

Na base Pedestres, observou-se o mesmo padrão no sexo (“Outro”), mas a idade modal foi de apenas 19 anos, indicando uma forte presença de jovens no conjunto analisado.

Na base Lojas, o valor de compra mais frequente foi de aproximadamente R$ 186,36, possivelmente refletindo um preço recorrente em determinados tipos de estabelecimentos.

Por fim, nos Cupons, os valores mais comuns foram R$ 371,39 para o cupom e R$ 7,76 para o repasse à PicMoney, mostrando que a maioria das transações ocorre em valores menores, apesar de existirem campanhas maiores que elevam a média.

```python
print("---> FREQUÊNCIAS <----")

print('\n[  frequencia do sexo ]')
print(base_players["sexo"].value_counts())

print('\n[ frequencia do tipo de cupom: ]')
print(base_cupons["tipo_cupom"].value_counts())

print("\n[ Top 10 Categorias de Estabelecimentos (Cupons): ]")
print(base_cupons["categoria_estabelecimento"].value_counts().head(10))
```

https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=wOW17TJ74kn9&line=9&uniqifie=1

```
---> FREQUÊNCIAS <----

[  frequencia do sexo ]
sexo
Outro        3356
Masculino    3322
Feminino     3322
Name: count, dtype: int64

[ frequencia do tipo de cupom: ]
tipo_cupom
Cashback    33556
Produto     33328
Desconto    33116
Name: count, dtype: int64

[ Top 10 Categorias de Estabelecimentos (Cupons): ]
categoria_estabelecimento
Farmácias e Drogarias                          6816
Lojas de Móveis e Decoração                    3479
Igrejas e Lojas de Artigos Religiosos          3441
Clínicas de Saúde e Bem-estar                  3415
Restaurantes e Gastronomia                     3410
Papelarias, Livrarias e Lojas de Escritório    3399
Supermercados e Mercados Express               3378
Coworkings e Centros de Estudo/Conexão         3375
Bancos e Casas Lotéricas                       3354
Espaços Culturais e de Experiência Interativa  3353
Name: count, dtype: int64
```

- O código foi usado para contar quantas vezes cada categoria aparece nas bases.
- Sexo (Players): os resultados mostraram um equilíbrio, com praticamente o mesmo número de usuários classificados como Outro (3.356), Masculino (3.322) e Feminino (3.322). Isso indica diversidade entre os cadastrados, sem predominância clara de um grupo.
- Tipo de Cupom (Cupons): as frequências também ficaram muito próximas: Cashback (33.556), Produto (33.328) e Desconto (33.116). Esse equilíbrio sugere que os clientes utilizam igualmente os diferentes tipos de benefícios oferecidos pela plataforma.
-
- Categorias de Estabelecimento (Cupons): as dez categorias mais recorrentes foram lideradas por Farmácias e Drogarias (6.816), seguidas por setores como Lojas de Móveis e Decoração (3.479), Clínicas de Saúde e Bem-estar (3.415) e Restaurantes e Gastronomia (3.410). Isso mostra que os cupons PicMoney têm maior adesão em áreas ligadas à saúde, bem-estar e consumo cotidiano, mas também alcançam cultura, serviços e até espaços religiosos.

Etapa 3 – Exploração dos Dados

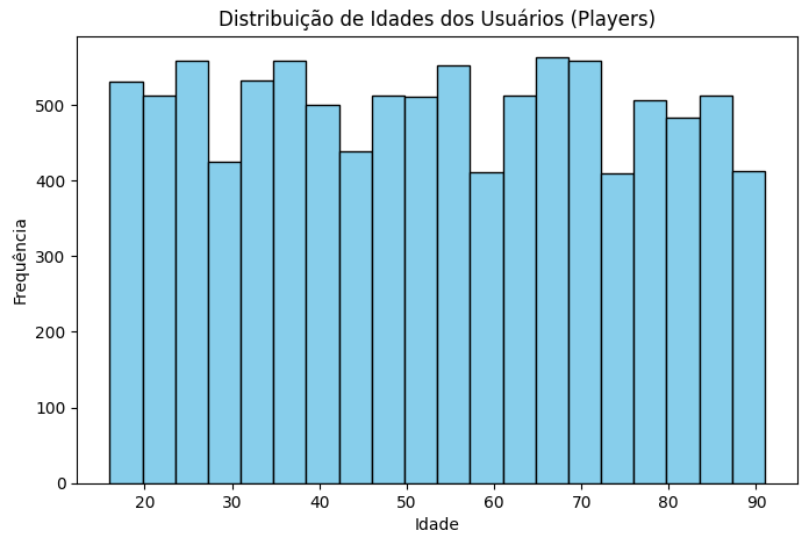
Foram geradas visualizações e análises para identificar padrões e relações entre as variáveis. Exemplos de gráficos incluem:

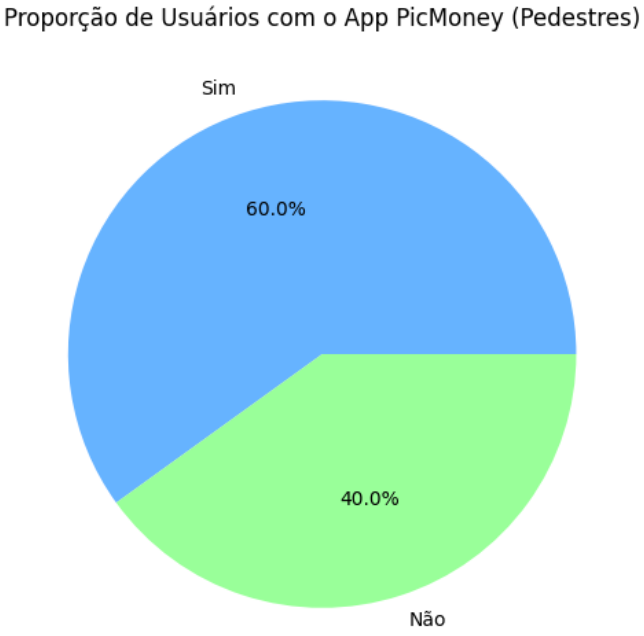
- Distribuição de idades dos usuários (base Players).
- Categorias de estabelecimentos mais frequentes (base Cupons).
- Ticket médio das compras (base Lojas).
- Proporção de usuários com e sem o app instalado (base Pedestres)

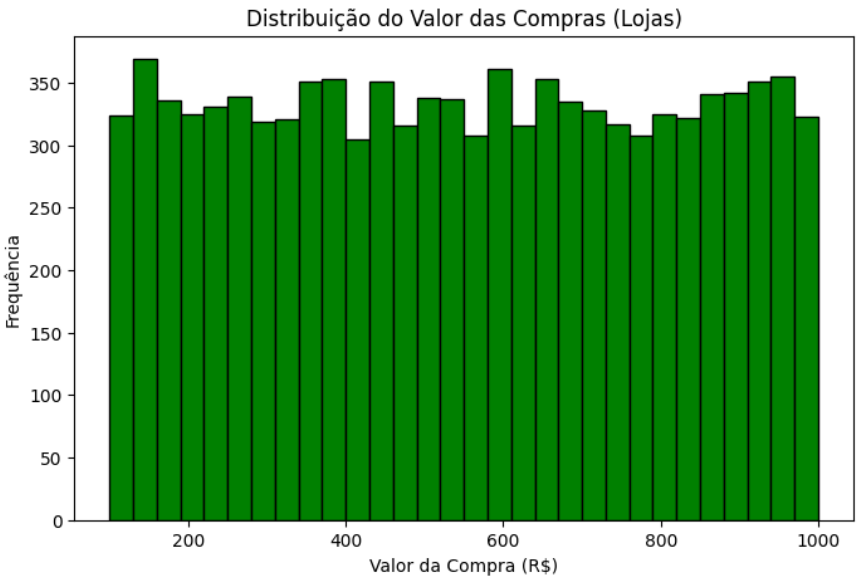
https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=tChLOa_2RAJP&line=20&uniqifier=1

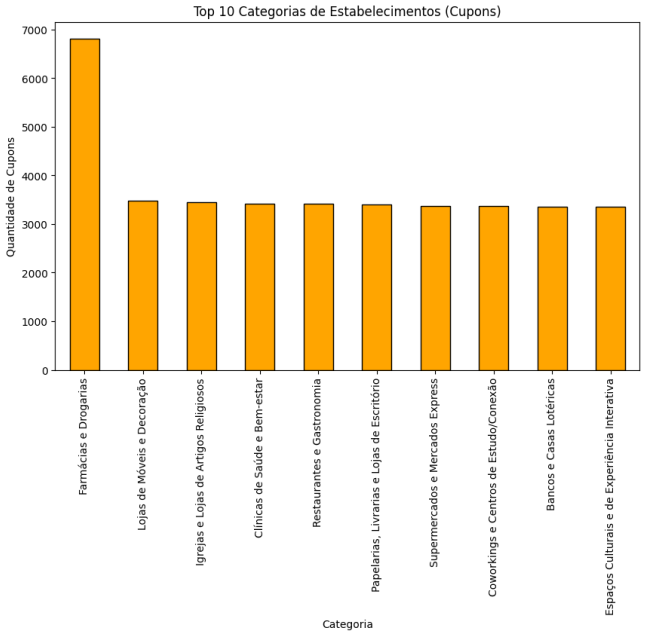
```python
import matplotlib.pyplot as plt

# 1. Distribuição de idades (Players)
plt.figure(figsize=(8,5))
base_players["idade"].plot(kind="hist", bins=20, color="skyblue", edgecolor="black")
plt.title("Distribuição de Idades dos Usuários (Players)")
plt.xlabel("Idade")
plt.ylabel("Frequência")
plt.show()
```

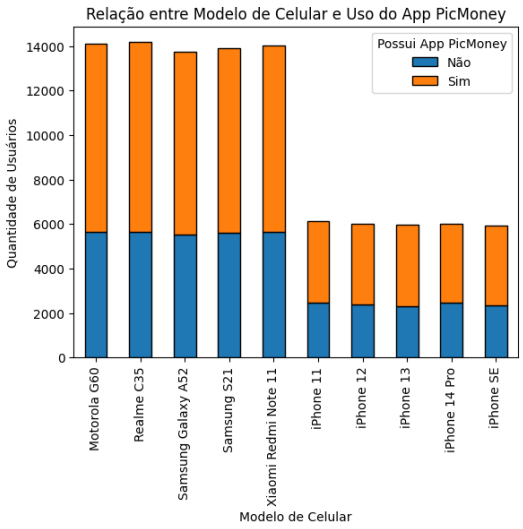




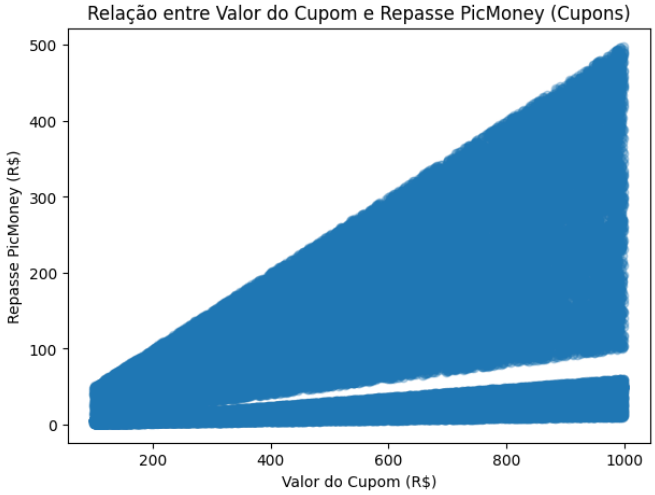




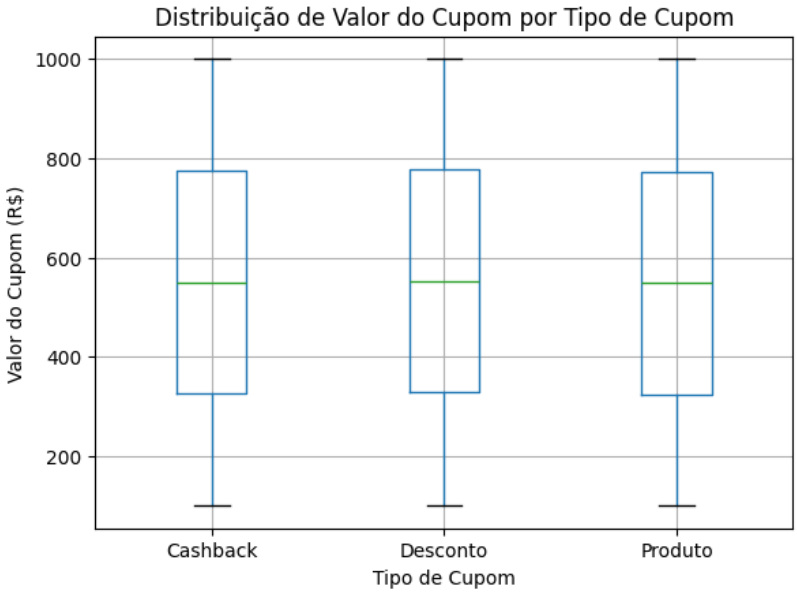
- O histograma de idades (Players) mostra que há uma distribuição ampla de usuários, variando dos 18 até os 90 anos. A frequência se mantém relativamente equilibrada entre as faixas etárias, o que indica que a plataforma é utilizada por públicos de diferentes idades, sem concentração muito forte em apenas uma faixa.
- O gráfico de barras das categorias de estabelecimentos (Cupons) evidencia que Farmácias e Drogarias são o setor com maior número de cupons emitidos, seguido por Lojas de Móveis e Decoração e Igrejas/Artigos Religiosos. Isso demonstra que os cupons da PicMoney têm forte aplicação em saúde, consumo doméstico e em setores específicos como o religioso.
- A distribuição do valor das compras (Lojas) mostra uma frequência relativamente constante de transações em diferentes faixas de valor, variando de aproximadamente R$ 100 até R$ 1.000. Isso indica que não há uma concentração clara em um único ticket médio, mas sim uma dispersão uniforme de valores de compra.
- O gráfico de pizza (Pedestres) revela que 60% dos usuários já possuem o app PicMoney instalado, enquanto 40% ainda não utilizam o aplicativo. Esse resultado demonstra uma taxa de adesão positiva, mas também mostra que existe um espaço significativo para expandir a base de usuários5.



2https://colab.research.google.com/drive/1v8hzLDCuuQ7QPXgJNOkBeEt3gaOEJg2f#scrollTo=NzYu_H9-rIbm&line=1&uniqifier=1



3https://colab.research.google.com/drive/1v8hzLDCuuQ7QPXgJNOkBeEt3gaOEJg2f#scrollTo=-VwX5BdyuV8S&line=1&uniqifier=1



1https://colab.research.google.com/drive/1v8hzLDCuuQ7QPXgJNOkBeEt3gaOEJg2f#scrollTo=UuOT2Dxnuuly&line=5&uniqifier=1

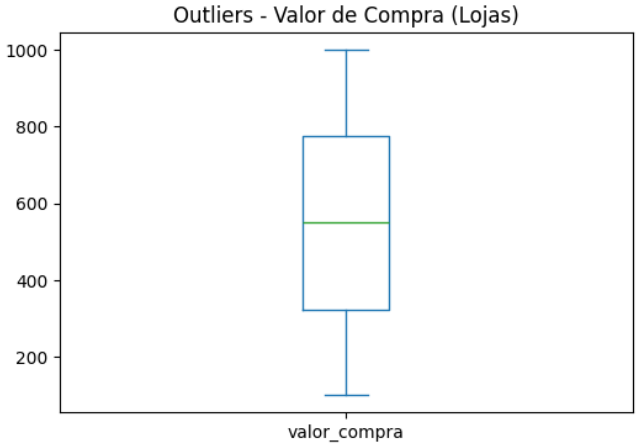
- O gráfico mostra a distribuição de usuários por modelo de celular e a proporção daqueles que possuem ou não o aplicativo instalado. Observa-se que modelos como Motorola G60, Realme C35 e Samsung Galaxy A52 concentram maior número de usuários, sendo que em todos os modelos há significativa presença de pessoas que ainda não utilizam o app. Essa análise é útil para identificar perfis tecnológicos e direcionar campanhas específicas de incentivo à instalação do aplicativo.

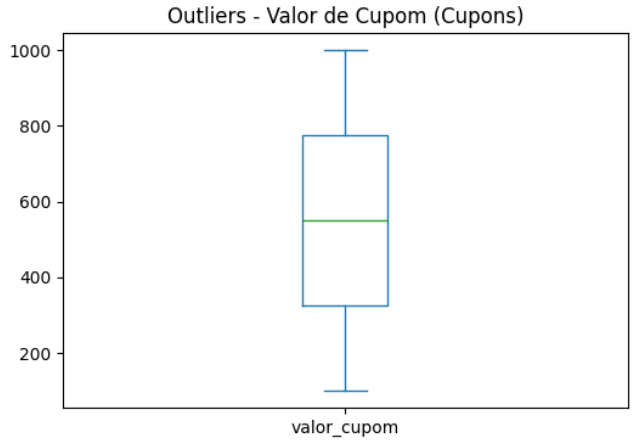
- O gráfico de dispersão revela uma tendência positiva: à medida que o valor do cupom aumenta, também cresce o repasse realizado para a PicMoney. A relação, no entanto, não é perfeitamente linear, apresentando diferentes faixas de comportamento, possivelmente ligadas a regras distintas de campanhas promocionais. Essa evidência mostra que políticas de cupom têm impacto direto na receita gerada pela plataforma.

- O boxplot evidencia que os três tipos de cupom (Cashback, Desconto e Produto) apresentam distribuições semelhantes em termos de valores, com medianas próximas de R$ 550. Ainda assim, há variabilidade em cada categoria, com cupons que vão de valores baixos até o limite máximo próximo de R$ 1.000. Esse resultado indica que a PicMoney distribui benefícios de forma relativamente equilibrada entre os diferentes tipos de cupom, permitindo atender a perfis diversos de consumidores.

Etapa 4 – Verificação da Qualidade

Nesta etapa foram avaliados valores ausentes, registros duplicados, outliers e consistência dos dados. As principais estratégias sugeridas incluem remoção de duplicados, substituição de valores faltantes e análise de outliers.





https://colab.research.google.com/drive/1iDxQRzovRu5TIN5X08XVF--Eo4Tixvhl#scrollTo=HhIzn_yuL8iK&line=22&uniqifier=1

```python
# 1. Valores nulos
print("----> VALORES NULOS <----")
print("Cupons:\n", base_cupons.isna().sum())
print("\nPedestres:\n", base_pedestres.isna().sum())
print("\nLojas:\n", base_lojas.isna().sum())
print("\nPlayers:\n", base_players.isna().sum())

# 2. Registros duplicados
print("\n----> DUPLICADOS <----")
print("Cupons:", base_cupons.duplicated().sum())
print("Pedestres:", base_pedestres.duplicated().sum())
print("Lojas:", base_lojas.duplicated().sum())
print("Players:", base_players.duplicated().sum())

# 3. Outliers (exemplo em valores de compra e cupom)
import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
base_lojas["valor_compra"].plot(kind="box", title="Outliers - Valor de Compra (Lojas)")
plt.show()

plt.figure(figsize=(6,4))
base_cupons["valor_cupom"].plot(kind="box", title="Outliers - Valor de Cupom (Cupons)")
plt.show()
```

```
----> DUPLICADOS <----
Cupons: 0
Pedestres: 0
Lojas: 0
Players: 0
```

Na verificação de qualidade, constatamos 0 registros duplicados quando consideradas todas as colunas, o que indica bases consistentes em nível de linha. Em contrapartida, a base Pedestres apresentou 40.043 valores ausentes em três campos relacionados ao histórico de compras (“data_ultima_compra”, “ultimo_tipo_cupom” e “ultimo_valor_capturado”), o que é esperado, pois nem todos os pedestres possuem evento de compra anterior registrado. As demais bases (Cupons, Lojas e Players) não apresentaram valores ausentes.

A análise gráfica por boxplots evidenciou que tanto os valores de compra quanto os de cupons não apresentam outliers graves, mas exibem alta variação natural entre clientes. Isso sugere que os dados são consistentes, porém refletem diferentes perfis de consumo (clientes de ticket mais baixo e clientes de ticket mais alto).

Conclusão e Próximos Passos

A análise inicial das bases da PicMoney trouxe uma visão geral sobre a estrutura dos dados e suas principais características. As estatísticas descritivas ajudaram a identificar valores médios em torno de R$ 550 tanto para compras quanto para cupons, com variação considerável nos dados, como mostrado pelos desvios-padrão. Os gráficos confirmaram essa dispersão, evidenciando que não há um único perfil predominante, mas sim diferentes faixas de valores e comportamentos de consumo.

A exploração dos dados também destacou a diversidade nas variáveis categóricas: os tipos de cupons (Cashback, Produto e Desconto) estão distribuídos de maneira equilibrada, e categorias como Farmácias e Drogarias, Lojas de Móveis e Decoração e Restaurantes apareceram entre as mais relevantes. Além disso, a proporção de usuários que já possuem o aplicativo instalado (cerca de 60%) indica uma boa taxa de adesão, mas ainda deixa espaço para crescimento.

Na verificação da qualidade, observamos que não houve registros duplicados e que a maior parte das bases não apresenta valores nulos. No entanto, a base de Pedestres concentrou ausências significativas em variáveis ligadas ao histórico de compras, o que reforça a necessidade de tratamento. Também foram identificados valores extremos nos campos monetários, ainda que dentro dos limites esperados, sugerindo a importância de revisar esses pontos antes de análises mais avançadas.

Como próximos passos, será fundamental aplicar técnicas de preparação e limpeza dos dados, tratando os valores ausentes e padronizando informações. A partir disso, será possível avançar para etapas mais sofisticadas do CRISP-DM, como modelagem, avaliação e implantação, que permitirão gerar segmentações de usuários, identificar padrões de comportamento e desenvolver estratégias de marketing mais direcionadas para a PicMoney.