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: 10/11/2025

2ª Entrega – Preparação dos Dados

Nesta segunda entrega do Projeto Interdisciplinar com parceira da PicMoney, avançamos na metodologia CRISP-DM até a etapa de Preparação dos Dados. Após compreender o negócio e analisar as bases brutas na primeira entrega, o foco agora foi organizar, limpar, integrar e formatar os conjuntos de dados, garantindo consistência, completude e padronização, facilitando as análises. As ações foram desenvolvidas no ambiente Google Colab, utilizando a linguagem Python e a biblioteca pandas.

O processo iniciou-se com a seleção das quatro bases de dados fornecidas: Cupons Capturados, Pedestres da Av. Paulista, Lojas e Valores, e Base Cadastral de Players. Cada uma delas foi importada por meio da função read_excel(), permitindo a leitura direta dos arquivos em formato XLSX. Essa etapa foi essencial para confirmar a estrutura e o volume de informações de cada base, além de possibilitar a verificação inicial de colunas, tipos de dados e possíveis inconsistências.

```
import pandas as pd

base_transacoes = 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")

base_transacoes.head()
```

	celular	data	hora	nome_estabelecimento	bairro_estabelecimento	categoria_estabelecimento	id_ca
0	(61) 96497-8673	2025-07-10	16:15:00	Habib's	República	Lojas de Eletrônicos e Games	CA
	(11)	2025-					

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=FEzixtDxMZT5&line=1&uniqifier=1

Na sequência, foi realizada a limpeza e uniformização dos dados. Essa fase envolveu a padronização dos nomes das colunas (conversão para letras minúsculas e substituição de espaços por sublinhados), remoção de duplicatas e ajuste de campos de texto e data. Também foi aplicada uma função para normalizar valores e corrigir divergências causadas por acentuação e espaços extras. O objetivo foi criar um conjunto de dados homogêneo, eliminando ruídos e facilitando as manipulações.

```python
def limpar_base(df):
    df = df.copy()
    df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_")
    df = df.drop_duplicates()

    for col in df.select_dtypes(include="object").columns:
        df[col] = df[col].astype(str).str.strip()

    for col in df.columns:
        if "data" in col:
            df[col] = pd.to_datetime(df[col], errors="coerce")

    return df
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=pmQ0i7JaEvzR&line=3&uniqifier=1

```python
transacoes.head()
#teste de limpeza
```

	celular	data	hora	nome_estabelecimento	bairro_estabelecimento	categoria_estabelecimento	id_car
0	(61) 96497-8673	2025-07-10	16:15:00	Habib's	República	Lojas de Eletrônicos e Games	CA
	(11)						

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=P85SCzykFE3I&line=1&uniqifier=1

Com as bases já limpas, passamos à derivação de novas variáveis. Foram criadas colunas adicionais que enriquecem as análises, como o valor total das transações (quantidade x valor_unitário) e a faixa etária dos usuários, calculada a partir da coluna de idade. Essas variáveis derivadas permitem observar padrões de comportamento e consumo entre diferentes grupos de clientes.

ETAPA 3 — Derivar Dados

```python
if "quantidade" in transacoes.columns and "valor_unitario" in transacoes.columns:
    transacoes["valor_total"] = transacoes["quantidade"] * transacoes["valor_unitario"]
```

```python
if "idade" in players.columns:
    players["faixa_etaria"] = pd.cut(
        players["idade"],
        bins=[0, 18, 30, 45, 60, 120],
        labels=["0-18", "19-30", "31-45", "46-60", "60+"]
    )
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=D7x5uv4WHv85&line=3&uniqifier=1

A etapa seguinte foi a integração dos dados. Essa fase consolidou as informações de diferentes fontes em uma única base unificada, garantindo que todas as relações entre players, transações, lojas e pedestres fossem mantidas. O processo de integração foi realizado através de operações de merge, com base nas chaves principais como 'celular' e 'nome_estabelecimento'. Após cada junção, foram feitas verificações de cobertura e checagens de consistência para garantir que os relacionamentos estavam corretos e sem perda de dados relevantes.

```
print("Colunas Transações:\n", transacoes.columns.tolist(), "\n")
print("Colunas Pedestres:\n", pedestres.columns.tolist(), "\n")
print("Colunas Lojas:\n", lojas.columns.tolist(), "\n")
print("Colunas Players:\n", players.columns.tolist(), "\n")
```

```
Colunas Transações:
 ['celular', 'data', 'hora', 'nome_estabelecimento', 'bairro_estabelecimento', 'categoria_estabelecimento'

Colunas Pedestres:
 ['celular', 'data', 'horario', 'local', 'latitude', 'longitude', 'tipo_celular', 'modelo_celular', 'possu

Colunas Lojas:
 ['numero_celular', 'data_captura', 'tipo_cupom', 'tipo_loja', 'local_captura', 'latitude', 'longitude', '

Colunas Players:
 ['celular', 'data_nascimento', 'idade', 'sexo', 'cidade_residencial', 'bairro_residencial', 'cidade_traba
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=jnb_GIUWH29f&line=2&uniqifier=1

```
print("Exemplo de celulares e nomes normalizados:")
print(t[["celular","nome_estabelecimento","nome_estab_norm"]].head(3))
print(lj[["nome_loja","nome_loja_norm"]].head(3))
```

```
Exemplo de celulares e nomes normalizados:
       celular nome_estabelecimento nome_estab_norm
0  61964978673              Habib's         habib's
1  11942316424            Smart Fit       smart fit
2  11979652178              Outback         outback
       nome_loja nome_loja_norm
0  Pão de Açúcar  pao de acucar
1  Pão de Açúcar  pao de acucar
2        Kalunga        kalunga
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=spEZ5fjGPUoL&line=1&uniqifier=1

```
Checagem de nulos em campos-chave:
celular: 0 nulos
nome_estab_norm: 0 nulos
nome_loja_norm: 77684 nulos

Cobertura por tabela (amostras):
Players (idade): 100.0%
Lojas (tipo_loja): 22.3%
Pedestres (dt_pedestre): 0.1%

Top 20 estabelecimentos sem correspondência em Lojas:
nome_estab_norm
drogasil                  4351
droga raia                4284
sabin                     3221
lavoisier                 3186
octavio cafe              3165
forever 21                3149
sesc carmo                3122
cafe cultura              3118
sesc paulista             3110
selfit                    3102
extra                     3094
fleury                    3089
starbucks                 3088
clube pinheiros           3084
carrefour express         3077
just run                  3070
rascal                    2920
churrascaria boi preto    2919
madero                    2888
acai no ponto             2780
Name: count, dtype: int64
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=xxCCU07R48rs&line=6&uniqifier=1

Por fim, a etapa de formatação consolidou todo o processo. Foram selecionadas as colunas mais importantes, convertidos os tipos de dados numéricos e temporais, e os registros foram organizados em ordem cronológica. Essa fase também incluiu a exibição de uma amostra da base final formatada, com seus tipos de dados, para comprovar a coerência e a preparação adequada para análises exploratórias e modelagens preditivas.

ETAPA 5 Formatar os Dados

```
# Seleção de colunas finais
colunas_principais = [
    # Transações
    "celular", "data", "hora", "dt_transacao", "produto",
    "nome_estabelecimento", "nome_estab_norm", "categoria_estabelecimento",
    "id_campanha", "id_cupom", "tipo_cupom", "valor_cupom", "repasse_picmoney",
    # Players
    "idade", "sexo", "cidade_residencial", "bairro_residencial",
    "cidade_trabalho", "bairro_trabalho", "cidade_escola", "bairro_escola",
    "categoria_frequentada", "faixa_etaria",
    # Lojas
    "nome_loja_norm", "tipo_loja", "local_captura", "endereco_loja",
    "latitude", "longitude", "valor_compra", "data_captura",
    # Pedestres (último evento)
    "dt_pedestre", "local", "tipo_celular", "modelo_celular",
    "possui_app_picmoney", "data_ultima_compra", "ultimo_tipo_cupom",
    "ultimo_valor_capturado", "ultimo_tipo_loja"
]

colunas_existentes = [c for c in colunas_principais if c in base_final.columns]
base_pronta = base_final[colunas_existentes].copy()
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=ogNcpacX5JuQ&line=3&uniqifier=1

```
# Ordenar por tempo da transação, se existir
if "dt_transacao" in base_pronta.columns:
    base_pronta = base_pronta.sort_values("dt_transacao")

print("Base formatada: linhas =", len(base_pronta), " | colunas =", len(base_pronta.columns))
print("Colunas finais:")
print(base_pronta.columns.tolist() )

# Exportações
base_pronta.to_excel("Base_PicMoney_Preparada.xlsx", index=False)
base_pronta.to_csv("Base_PicMoney_Preparada.csv", index=False, sep=";")
print("Arquevios salvos:Base_PicMoney_Preparada.xlsx e Base_PicMoney_Preparada.csv")
```

```
Base formatada: linhas = 100011  | colunas = 38
Colunas finais:
['celular', 'data', 'hora', 'dt_transacao', 'produto', 'nome_estabelecimento', 'nome_estab_norm', 'categor
Arquivos salvos: Base_PicMoney_Preparada.xlsx e Base_PicMoney_Preparada.csv
```

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=ogNcpacX5JuQ&line=3&uniqifier=1

```python
colunas_exemplo = [
    "celular", "dt_transacao", "nome_estabelecimento",
    "tipo_loja", "valor_cupom", "idade", "faixa_etaria"
]
colunas_existentes = [c for c in colunas_exemplo if c in base_final.columns]

base_formatada = base_final[colunas_existentes].copy()

if "valor_cupom" in base_formatada.columns:
    base_formatada["valor_cupom"] = pd.to_numeric(base_formatada["valor_cupom"], errors="coerce")
if "dt_transacao" in base_formatada.columns:
    base_formatada["dt_transacao"] = pd.to_datetime(base_formatada["dt_transacao"], errors="coerce")
if "dt_transacao" in base_formatada.columns:
    base_formatada = base_formatada.sort_values("dt_transacao", ascending=False)

print("Amostra de dados formatados e prontos para análise:")
display(base_formatada.head(10))

print("\nTipos de dados após formatação:")
print(base_formatada.dtypes)
```

```
Amostra de dados formatados e prontos para análise:
```

1 to 10 of 10 entries Filter

index	celular	dt_transacao	nome_estabelecimento	tipo_loja	idade	faixa_etaria
47290	11958258969	2025-07-31 22:00:00	Casas Bahia	NaN	18	0-18
65987	11977053050	2025-07-31 22:00:00	Magazine Luiza	NaN	17	0-18
48906	11959637151	2025-07-31 22:00:00	Pão de Açúcar	mercado express	25	19-30
81879	11991754451	2025-07-31 22:00:00	Octavio Café	NaN	65	60+
15248	11926139264	2025-07-31 22:00:00	McDonald's	NaN	65	60+
28574	11939858171	2025-07-31 22:00:00	Just Run	NaN	53	46-60

https://colab.research.google.com/drive/1UjAo4lqS92Va_FG_EUfHeVkcaTm2WnMf?authuser=1#scrollTo=M-2SyI176nks&line=1&uniqifier=1

Conclusão

A segunda entrega do projeto consolidou de forma definitiva o processo de preparação dos dados dentro do ciclo CRISP-DM. Todas as etapas desenvolvidas — seleção, limpeza, derivação, integração e formatação — garantiram que as quatro bases originais da PicMoney fossem tratadas e unificadas de maneira limpa, estruturada e padronizada, eliminando duplicidades, inconsistências e ruídos que poderiam comprometer análises futuras.

O uso do Google Colab e da biblioteca pandas foi essencial para assegurar automação, reprodutibilidade e rastreabilidade de cada transformação aplicada, tornando o processo transparente e confiável. Além disso, o trabalho sistemático de preparação criou uma base sólida para todas as próximas etapas do projeto, especialmente as fases de análise exploratória e modelagem.

Com as etapas concluídas, foi possível gerar a base final integrada, que servirá como ponto de partida para a criação de dashboards gerenciais e analíticos voltados para a equipe de decisão para a PicMoney. Essa base permitirá visualizar indicadores-chave como volume de transações por estabelecimento, perfil demográfico dos usuários, tipos de cupons mais utilizados e repasses médios à empresa.

A partir dessas informações, torna-se possível construir painéis interativos e estratégicos, que fornecem insights em tempo real, apoiam a tomada de decisão baseada em dados e ampliam a inteligência analítica da plataforma PicMoney.

Em síntese, a segunda entrega representa o marco técnico mais importante do projeto até o momento — o momento em que as informações dispersas foram transformadas em conhecimento estruturado. Com os dados preparados e validados, o projeto segue fortalecido para a próxima fase: a construção dos dashboards e a análise exploratória dos resultados, que demonstrarão o verdadeiro potencial das informações organizadas durante este ciclo.