Projeto Interdisciplinar: Ciência de dados
Entrega 2

Guilherme Barioni RA:24026140
Iury Xavier da Silva Mangueira RA:24026311
Lilian Mercedes Paye Conde RA:24026462
Marcus Miranda Duque RA:24026080
Murilo de Souza Vieira RA:24025726

Preparação dos dados

Ao carregar as planilhas na Dashboard, os dados precisam ser devidamente tratados antes da visualização. Com base na metodologia CRISP-DM, desenvolvemos e aplicamos scripts responsáveis por cada etapa de preparação, garantindo consistência, padronização e qualidade das informações utilizadas nas análises.

Selecionar os dados

Foram selecionadas as quatro planilhas fornecidas pela PicMoney, previamente analisadas na entrega anterior, sendo elas a base de transação de cupons capturados, a base cadastral, a base simulada de pedestres e a massa de teste com lojas parceiras. Essas fontes constituem o núcleo de dados do projeto, permitindo a integração e o cruzamento de informações entre usuários, locais e transações.

Limpeza/Uniformização dos dados

Base de transações – Cupons capturados

Reaplicamos os scripts da entrega anterior, redistribuindo as lojas parceiras da PicMoney em 12 novas categorias que representam corretamente suas funções comerciais.

Massa de teste com lojas e valores

Aplicamos novamente os scripts de categorização, assegurando que todas as lojas estivessem alocadas em suas respectivas categorias.
Além disso, criamos uma função para corrigir coordenadas geográficas (latitude e longitude), ajustando os valores para o intervalo válido entre -90 e 90. Os dados tratados foram armazenados nas novas colunas “latitude_limpa” e “longitude_limpa”.

Base Simulada – Pedestres Av. Paulista

As coordenadas foram convertidas utilizando a mesma função criada anteriormente, entre o intervalo de -90 e 90. Na coluna “ultimo_tipo_loja”, as categorias foram renomeadas de acordo com o padrão adotado na Base Cadastral, garantindo uniformidade entre as planilhas.

Base cadastral de Players

Removemos usuários duplicados utilizando o número de celular como chave primária, assegurando a integridade do banco.

Formatar os dados

Base de transações – Cupons capturados

As colunas referentes à data e hora de captura foram convertidas de string para datetime (função `to_datetime` do pandas), permitindo análises temporais precisas.

Massa de teste com lojas e valores

A coluna *“data”* também foi convertida para o formato datetime, facilitando operações de filtragem e agrupamento.

Base Simulada – Pedestres Av. Paulista

As colunas *“data”* e *“horario”* foram transformadas em datetime, viabilizando cruzamentos com outras bases temporais.

Base Cadastral de Players

A coluna *“data_nascimento”* foi convertida para datetime, e a coluna *“idade”* para int, padronizando os tipos de dados para análises estatísticas.

Derivar os dados

Base de transações – Cupons capturados

Para facilitar a análise dos cupons capturados, foi criada duas novas colunas que contém o dia da semana correspondente a data de captura e em qual período do dia, a captura foi realizada com o dia sendo dividido em três partes manhã, tarde e noite. A função .day_name() registra os nomes em inglês, para melhor compreensão e análise, os dias da semana foram traduzidos.

Também foi criada uma nova coluna que representa somente a hora em que o cupom foi resgatado, ela acaba facilitando comparações entre varias entradas na tabela.

Massa de teste com lojas e valores

Adicionamos uma coluna com o dia da semana da captura dos cupons.
Além disso, constatamos que os endereços originais não correspondiam às coordenadas fornecidas. Para corrigir isso, utilizamos a API do CEP Aberto, que retorna o endereço real com base na latitude e longitude.

Base Simulada – Pedestres Av. Paulista

Foram criadas as colunas “Dia da Semana” e “Período do Dia”, extraídas das colunas “data” e “horario”, mantendo o mesmo padrão de derivação das demais bases.

Integrar os dados

Para consolidar a análise, integramos a Base Cadastral e a Base de Pedestres, utilizando o número de celular como chave de junção (*merge*).
Com isso, garantimos que os usuários presentes em ambas as bases fossem corretamente identificados.
Também adicionamos a coluna *“possui_app_picmoney”*, marcando com “Sim” os usuários que possuem o aplicativo da PicMoney instalado.

Conclusão

A segunda entrega consolidou a etapa de Preparação dos Dados do projeto, garantindo que todas as bases estivessem limpas, padronizadas, integradas e prontas para análise.

A aplicação da metodologia CRISP-DM orientou o processo de forma sistemática, desde a seleção até a integração das fontes, assegurando qualidade, consistência e confiabilidade dos dados.

Com as tabelas tratadas e enriquecidas, o conjunto está agora preparado para a fase de análise exploratória e modelagem preditiva, que permitirá extrair insights relevantes sobre o comportamento dos usuários, a performance das lojas parceiras e as dinâmicas de transações na plataforma PicMoney.