FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO

CAMPUS LIBERDADE

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL CARVALHO MOTA - 24026334

GUILHERME DE LIMA SIQUEIRA - 21010649

RODRIGO LUIZ MENEZES DOS REIS - 24025708

VITORIA LETICIA MACIEL DA SILVA - 24026593

PICMONEY

VERIFICAÇÃO DA QUALIDADE DOS DADOS

São Paulo, 2025

Sumário

1 INTRODUÇÃO

Nesta etapa, foi realizada uma verificação da qualidade dos dados nas bases selecionadas para o projeto. O objetivo foi identificar e, se necessário, corrigir problemas como valores ausentes, dados duplicados e inconsistências para garantir a fidedignidade das análises subsequentes.

2 Verificação de Consistência Semântica (Categoria vs. Estabelecimento)

Método: realizamos uma verificação por amostragem para validar a coerência lógica entre o nome do estabelecimento e a sua respectiva categoria. Foram selecionados exemplos de estabelecimentos conhecidos para verificar se a sua classificação estava correta.

Resultado: A verificação revelou **inconsistências semânticas graves** na base de dados. Foi constatado que diversos estabelecimentos estão alocados em categorias que não correspondem à sua área de atuação. Por exemplo, a rede de restaurantes **"Outback Steakhouse" foi encontrada classificada como "Igrejas e Lojas de Artigos Religiosos"**. No COLAB foram listados outros casos que ocorrem as mesmas inconsistências.

Impacto e Ação Tomada: Esta inconsistência tem um **impacto crítico** em todas as análises baseadas em categorias, como "Top 5 Categorias Mais Consumidas" e "Receita por Categoria", que foram realizadas na fase de exploração. Os resultados dessas análises ficam **comprometidos e não refletem a realidade**, pois as transações estão sendo agrupadas em categorias erradas.

Ação Tomada: por serem dados extremamente necessários para diversas das nossas análises, conseguimos corrigir esse erro e estamos utilizando a base alterada com essa correção.

3 Verificação de Coerência Geográfica (Latitude e Longitude)

Método: Foi realizada uma verificação por amostragem das coordenadas (latitude e longitude) presentes nas bases de dados, como a PicMoney-BasePedestresAvPaulista. As coordenadas dos primeiros registros de cada base foram extraídas e consultadas em um serviço de mapas (Google Maps) para validar sua correspondência com as localizações esperadas em São Paulo.

Resultado: A verificação confirmou que grande parte das coordenadas presentes na base de dados são **inválidas e inconsistentes**. Os pontos geográficos resultantes da consulta não correspondem a nenhuma localização coerente em São Paulo, gerando erro sempre que procuradas. Por exemplo, as coordenadas [Latitude=-23.567.430.342.750.400, Longitude=-4.664.844.333.528.140] que deveriam estar na Pamplona, na verdade gera erro de “não encontrado”.

Impacto e Ação Tomada: A invalidez dos dados de latitude e longitude **impede a realização de qualquer análise geoespacial de precisão**, como a criação de mapas de calor (heatmaps), a clusterização de

transações por localização exata ou o cruzamento com dados de pedestres. O uso desses dados levaria a conclusões e visualizações completamente equivocadas.

Ação Tomada: Diante desta limitação crítica, foi decidido **não utilizar as colunas de latitude e longitude** para as análises. Como alternativa, as análises com viés geográfico foram baseadas em informações de localização textuais e mais confiáveis, como a coluna bairro_estabelecimento. Embora menos granular, esta abordagem garante a integridade e a validade das conclusões.

4 Verificação de Duplicatas (Telefone na Base de Pedestres)

Método: Foi realizada uma verificação de duplicatas na primeira coluna do arquivo PicMoney-BasePedestresAvPaulista.csv. Conforme identificado, esta coluna contém o número de telefone de players, que deveria funcionar como um identificador único para cada indivíduo. O processo consistiu em contar a frequência de cada número de telefone presente na base de dados para identificar valores repetidos.

Resultado: A análise confirmou a existência de **valores duplicados** na coluna de telefone. Foram identificados múltiplos registros que, apesar de representarem diferentes eventos de captura de dados, estão associados ao mesmo número de telefone. No COLAB são mostrados exemplos das duplicatas.

Impacto e Ação Tomada: A presença de duplicatas em uma coluna de identificação é uma **falha de integridade de dados significativa**. Isso indica que o mesmo indivíduo foi registrado múltiplas vezes ou que existem erros no processo de coleta de dados. O impacto direto é a impossibilidade de usar o telefone como uma chave primária confiável para cruzar esta base de dados com outras. Além disso, qualquer contagem de "pedestres únicos" a partir deste arquivo seria imprecisa.

Ação Tomada: A base de pedestres será tratada como um registro de "encontros" ou "avistamentos", e não como um cadastro de indivíduos únicos. Além disso, não será possível cruzar nenhuma base de dados, uma vez que o dado que seria tratado como chave primária, tem duplicatas. Também pretendemos fazer uma limpeza na base para fins de utilização dos dados, como a contagem de players.

5 Conclusão

Em conjunto, essas inconsistências levam à conclusão de que a qualidade dos dados é **baixa** e **inadequada** para gerar insights de negócio confiáveis sem um prévio e extensivo trabalho de limpeza. As análises exploratórias realizadas, embora metodologicamente corretas, devem ser interpretadas como uma **demonstração de potencial analítico**, e não como um retrato fiel da realidade operacional da empresa PicMoney.