Unidade Curricular

Processamento para Big Data

Ano Letivo 2024/2025 | 4° Trimestre

2° Ano da Licenciatura em Ciência de Dados

Grupo 6

Bárbara Andreia Castelão Alexandre, n° 112408

Carlos Filipe Romano Vidigal, n° 106585

José Ricardo de Almeida Valério, n° 112255

Leonor Oliveira Caratão, n° 123462

Índice

1. Introdução e Contextualização

Este trabalho teve como objetivo a aplicação prática de técnicas de processamento de dados em larga escala utilizando a plataforma Apache Spark, com foco na linguagem Python (PySpark). O projeto seguiu uma abordagem modular com três notebooks distintos, abordando desde a ingestão e tratamento dos dados até a aplicação de algoritmos de análise não supervisionada.

1.1 Dataset e Contexto de Negócio

O repositório utilizado foi o "*Big Sales Data*" do portal Kaggle. Nele foi escolhido o dataset "*Liquor_Sales.csv*", com registos de vendas de bebidas alcoólicas no estado de Iowa, EUA (2012-2020), disponibilizado pelo *Iowa Department of Revenue, Alcoholic Beverages Division*[1]. Este órgão regula a comercialização, fiscalização e distribuição de bebidas alcoólicas no estado[2].
O dataset possui 24 variáveis que abrangem informações sobre vendas (valores em dólares e volumes), características dos produtos (categorias e embalagens), informações geográficas (endereços, cidades, condados), dados temporais (datas das transações) e informações sobre fornecedores e lojas. Esta diversidade de variáveis permitiu uma análise abrangente do comportamento comercial no setor de bebidas alcoólicas.

1.2 Objetivos do Projeto

O principal objetivo foi segmentar as vendas com base nas características dos produtos e volumes, identificando grupos com padrões comerciais semelhantes. Esta segmentação pode ser útil para apoiar decisões de marketing, gestão de stock e análise de mercado.

2. Metodologia e Desenvolvimento

2.1 Notebook 1: Importação e Tratamento de Dados

Devido ao volume considerável dos dados e impossibilidade de execução computacional, foi implementada uma estratégia de amostragem aleatória para criar um subset representativo de aproximadamente 589 mil registos (3% do dataset original). Esta abordagem permitiu manter a representatividade estatística enquanto otimizava o desempenho computacional para as análises subsequentes por parte dos elementos do grupo.

Correção de Inconsistências

Um aspeto crítico do tratamento de dados foi a identificação e correção de inconsistências. Descobrimos que colunas como `*County*`, `*City*`, `*Vendor Name*`, `*Category Name*` e `*Store Name*` apresentavam valores semanticamente iguais mas escritos de formas diferentes (ex: "ADAMS" vs "Adams", "DES MOINES" vs "Des Moines"). Este tipo de problema é muito comum em bases de dados reais e, se não for tratado, pode comprometer análises de agregação, segmentação e machine learning. Para resolver, convertemos todos os valores para *uppercase* e uniformizando as categorias garantimos que cada entidade fosse representada de forma única.

[1] Fonte oficial do dataset: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy/about_data
[2] Vídeo fonte de entendimento de negócio: https://www.youtube.com/watch?v=LinQMfvX-aY

Esta transformação foi aplicada de forma sistemática a todas as colunas categóricas relevantes. Além das diferenças de capitalização, encontrámos também variantes ortográficas e erros de codificação, como a presença do caracter especial "" em nomes de produtos e fornecedores. Corrigimos os valores afetados, substituindo-os pelas versões corretas. Também foram detetados valores incoerentes (registos com `*Sales (Dollars)*` e `*Bottle Volume*` iguais a zero por exemplo) que foram removidos posteriormente. Este trabalho de "limpeza" foi moroso, mas essencial para garantir a qualidade dos dados.

Engenharia de Features Temporais

A partir da coluna `*Date*`, foram extraídas variáveis temporais como `*Year*`, `*Month*`, `*Day*`, `*WeekOfYear*`, e `*DayOfWeek*`, permitindo análises sazonais e comportamentais.

Qualidade dos Dados

Por fim, exportámos a amostra limpa para *Parquet*. A escolha deste formato foi sugerida pois este formato é muito mais eficiente para leitura e processamento em Spark, além de ocupar menos espaço em disco. Também guardámos um ficheiro *csv* único a fim de gerarmos um relatório de perfil com a biblioteca `*ydata_profiling*`, seguindo recomendações dos docentes. Este relatório permitiu identificar mais rapidamente padrões, outliers e possíveis problemas.

2.2 Notebook 2: Análise Exploratória

O segundo notebook centrou-se na compreensão da estrutura e distribuição dos dados através de análises estatísticas e visualizações.

Variáveis Derivadas

Foram criadas as seguintes variáveis que facilitaram o processo de exploração:

- **Unit Margin (Dollars)**: Diferença entre preço de venda e custo, identificando produtos com maior rentabilidade;
- **Price per Liter (Dollars)**: Normalização do preço por volume para comparação justa entre produtos;
- **Distance to Gov Division (Km)**: Calculada usando a fórmula de Haversine [3]e implementada através de uma User Defined Function (UDF), representando a distância de cada loja à sede da *Iowa Department of Revenue, Alcoholic Beverages Division*.

Estatísticas Descritivas

Foram calculadas estatísticas descritivas para as variáveis numéricas, incluindo a moda (não disponível nativamente no método `*describe()*` do Spark), implementada via UDF. Esta análise revelou:

- Valores médios, desvios-padrão, mínimos e máximos;
- Volumes de vendas atípicos;
- Identificação de outliers significativos (1 com Pack de 312 unidades quando o padrão é 12; 1 registo com margem-unitária negativa);
- Distribuições assimétricas em várias variáveis;

[3] https://pt.wikipedia.org/wiki/Fórmula_de_haversine

Nesta fase, o grupo tomou a decisão crucial de não remover mais valores extremos neste projeto. Essa decisão resulta de uma reavaliação metodológica após uma primeira abordagem com o método de Tukey (baseado no intervalo interquartil). Embora essa técnica seja útil para identificar valores extremos, a sua aplicação resultou na exclusão de uma parte substancial do conjunto de dados (mais de metade), o que implicaria a perda significativa de informações potencialmente relevantes, nomeadamente associadas a produtos de luxo ou compras de elevado valor - características comuns no setor de bebidas alcoólicas. Verificámos se existiam erros de inserção nesses casos mas eram transações legítimas.
Optou-se, portanto, por preservá-los, reconhecendo o seu valor informativo para uma análise mais completa e representativa. Estes valores extremos podem refletir padrões de consumo distintos e contribuir para a identificação de segmentos específicos na etapa de clustering, como pretendido neste projeto. A sua remoção poderia comprometer a detecção de grupos de elevado interesse comercial.
Em substituição, foi adotada, no final do notebook, uma transformação logarítmica nas variáveis quantitativas e decidimos realizar as próximas análises exploratórias sobre os nossos dados amostrais intactos.

Correlações e Visualizações

Com scatterplots, boxplots, heatmaps e matrizes de correlação, identificaram-se relações entre variáveis, redundâncias e padrões. Destaque para fortes correlações entre custo, preço, volume e vendas.

Análises Temporais e Geográficas

Estudou-se a evolução de vendas por ano, mês e dia, revelando sazonalidades e preferências temporais. Análises espaciais mostraram que a localização não tem relação direta com o volume de vendas.

Outras Visualizações

- **Sunburst**: Estrutura hierárquica "categoria-fornecedor" por volume vendido, permitindo identificar concentrações de mercado e dependências na cadeia de fornecimento
- **Heatmap temporal**: Vendas de garrafas por semana ao longo dos anos, revelando padrões sazonais consistentes com picos no final do ano e impactos atípicos em 2020 (muito provavelmente relacionados à pandemia)

Exportou-se novamente o dataset com as transformações em formato Parquet.

2.3 Notebook 3: Análise Não Supervisionada - K-means

Features Selecionadas

Selecionaram-se variáveis log-transformadas e que foram normalizadas usando *StandardScaler*: `*log_pack*`; `*log_bottle_volume_ml*`; `*log_unit_margin_dollars*`; `*log_price_per_liter_dollars*` e `*log_bottles_sold*`.

Determinação do K ótimo e avaliação

Aplicado o método do cotovelo, k=3 mostrou ser o ponto de inflexão ideal, com Silhouette Score de 0.396. Embora este valor indique uma separação moderada (não próxima ao ideal de 1), é aceitável

considerando a complexidade e natureza dos dados comerciais, onde fronteiras claras entre segmentos podem não existir naturalmente.

Redução da Dimensionalidade (PCA) e Visualização dos Clusters

Após treinar o modelo, analisámos a distribuição dos registos por cluster, os centros de cada grupo e as médias das principais variáveis de negócio. Para compreender a estrutura dos clusters, foram implementadas duas técnicas de redução de dimensionalidade:

1. **PCA (Principal Component Analysis)**: Redução para 2 componentes principais para visualização 2D

2. **t-SNE (t-Distributed Stochastic Neighbor Embedding)**: Aplicação em 2D e 3D para visualização mais sofisticada das relações locais entre pontos

Estas visualizações confirmaram a existência de três grupos distintos com algumas sobreposições, validando a escolha de k=3.

3. Resultados e Perfil dos Clusters

- **Cluster 0 - "Vendas de valor médio e volume típico"**:
  Representa as vendas mais típicas, com produtos comuns, margens baixas e volumes médios. Pode representar bebidas populares com grande rotação, ou seja, bebidas que são compradas muitas vezes por diferentes clientes, em diferentes momentos.

- **Cluster 1 - "Vendas de volume baixo e Produtos Premium"**:
  Agrupa produtos premium, com margens altas, poucas unidades por venda e preços elevados. Pode representar licores caros importados, edições especiais ou compras de nicho.

- **Cluster 2 - "Vendas de grandes volumes e transações com foco em quantidade"**:
  Corresponde a vendas em massa, com grandes volumes, preços acessíveis e valor total elevado. Pode representar promoções, eventos ou compras para grandes grupos.

Os resultados obtidos têm possível aplicabilidade direta em contextos empresariais:

- **Gestão de Inventário**: Otimização de stock baseada nos padrões de cada segmento
- **Marketing Direcionado**: Estratégias personalizadas para cada cluster de produtos/clientes
- **Estratégia de Preços**: Ajuste de preços baseado na sensibilidade de cada segmento
- **Expansão Geográfica**: Insights para abertura de novas lojas baseados na distribuição dos clusters

4. Conclusão

Este projeto consolidou práticas de engenharia de dados e análise de dados em Big Data com Spark. Desde a ingestão, tratamento e análise exploratória até à modelação, o trabalho revelou-se um desafio realista e educativo. As técnicas aplicadas permitiram extrair informação de valor do dataset, identificar padrões e propor segmentações para decisões de negócio.

O uso do PySpark foi essencial para lidar com o volume de dados, e as visualizações trouxeram clareza à interpretação dos clusters. Apesar de limitações computacionais o projeto cumpriu os seus objetivos.