Universidade Federal de Minas Gerais

Mineração de Dados - Trabalho Prático 1 - Padrões Frequentes

Análise da Cesta de Produtos Instacart (Kaggle)

Professor: Wagner Meira

Aluno: Pedro Lóes

Data: 18-11-2021

Conteúdo

1 - Entendimento do Negócio

O problema foi retirado da competição Análise da Cesta de compras Instacar na plataforma Kaggle. A competição foi iniciada a 4 anos e já foi encerrada mas o repositório continua ativo para que os usuários da plataforma possam baixar os dados e desenvolver projetos.

O Instacar é um aplicativo de compras e entregas dos produtos de mercearia e supermercado. Depois de selecionados os produtos e executada a ordem de compra no aplicativo Instacart, agentes de compras revisam o pedido, fazem as compras e entregam para os clientes.

1.1 Objetivos do Negócio

No aspecto serviço prestado ao cliente o principal objetivo é otimizar o processo de compras de produtos em mercearias e supermercados oferecendo serviços de shopping virtual e entrega. Desta forma o cliente ganha tempo, comodidade, segurança e qualidade em seu processo de compras podendo ainda automatizá-lo.

Os principais objetivos, tendo em vista os interesses da empresa, são maximizar vendas, bem como fidelizar e captar novos clientes. Desta forma a empresa aumentará o seu retorno, bem como ampliará e solidificará suas áreas de atuação.

1.2 Avaliação da Situação

A competição disponibilizou um banco de dados relacional no formato plano com separador de vírgulas contendo as informações sobre as ordens dos clientes ao longo do tempo. O banco de dados foi anonimizado e a única informação sobre os clientes é referente a ordem dos produtos. Os dados de supermercados e mercearias também foram anonimizadas para que suas informações pessoais identificáveis não pudessem ser recuperadas.

Banco de Dados

O banco de dados relacional disponibilizado pela empresa possui 5 tabelas no formato de arquivos planos com separador de vírgulas.

- Corredores
  - Coluna com os identificadores dos corredores onde os produtos podem ser encontrados no supermercado.
  - Coluna com os nomes dos corredores.
- Departamentos
  - Coluna com os identificadores dos departamentos em que os produtos foram classificados.
  - Coluna com os nomes dos corredores.
- Ordem de Produtos
  - Coluna com os identificadores das ordens dos clientes.
  - Coluna com os identificadores dos produtos das ordens.
  - Coluna com a ordem em que os produtos foram adicionados ao carrinho.
  - Coluna com indicadores sobre a recorrência de produtos em ordens anteriores.
- Ordens
  - Coluna com os identificadores das ordens dos clientes.
  - Coluna com os identificadores dos clientes.
  - Coluna com identificadores sobre a amostra ser do tipo treino ou teste.

  - Coluna com os identificadores dos números das ordens.
  - Coluna com os identificadores do dia da semana.
  - Coluna com os identificadores da hora em que os produtos foram comprados.
  - Coluna com o número de dias após a última ordem.
  - Coluna com indicadores sobre a recorrência das compras.

- Produtos
  - Coluna com os identificadores dos produtos.
  - Coluna com os nomes dos produtos.
  - Coluna com os identificadores dos corredores.
  - Coluna com os identificadores de departamentos.

A equipe de ciência de dados utiliza as informações de transações para desenvolver modelos que prevêem quais produtos os clientes comprarão novamente, tentarão comprar pela primeira vez ou quais anúncios devem ser mostrados para os clientes para que outros produtos sejam adicionados ao carrinho durante uma sessão.

As técnicas de mineração de dados já utilizadas pela empresa são XGBoost, word2vec e Annoy para predizer se o usuário comprará de novo produtos já comprados e para recomendar produtos semelhantes ao longo do processo de compras.

Para a execução do projeto foi utilizada uma máquina com processador Intel I7-9700 com 3GhZ e 16 GB de memória do tipo RAM no sistema Windows 10. A plataforma de desenvolvimento integrado RStudio, bem como a linguagem R foram utilizadas para carregar, preparar, explorar, desenhar, modelar os dados e produzir o relatório.

1.3 Objetivos da Mineração de Dados

A competição consistiu na predição de quais produtos comprados anteriormente estarão na próxima ordem de compra de um dado cliente. Porém o problema foi adaptado neste projeto para determinar quais conjuntos de padrões frequentes de produtos podem ser observados nesta base de dados e quais regras e associações podem ser derivadas desta técnica de mineração de dados.

Para tanto, serão utilizadas técnicas de análise descritiva dos dados e o algoritmo Apriori para identificar padrões frequentes. Finalmente serão geradas regras de associação com suas respectivas confianças com intuito de gerar entendimento sobre o comportamento dos clientes no que diz respeito aos seus padrões de compras e gerar inteligência para o negócio do ponto de vista de marketing do tipo cross selling, bem como sistemas de recomendação para a plataforma de comércio eletrônico da empresa.

1.4 Descrição do Projeto

Este projeto foi elaborado para aplicar a técnicas de mineração de padrões frequentes abordadas na disciplina Mineração de Dados do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais. Técnicas de análise descritiva e exploratória também serão utilizadas para explorar os dados. O objetivo final do projeto foi produzir um relatório segundo as especificações **CRISP** que foi submetido para análise como o primeiro trabalho prático da disciplina.

2 - Entendimento dos Dados

A etapa de entendimento dos dados compreende a extração dos dados, a descrição das meta informações e estruturas dos dados, a verificação da qualidade dos dados para identificar problemas como dados faltantes ou ausência de integridade nas bases e a análise exploratória para investigar tendências e ou padrões aparentes.

2.1 Coleta dos Dados

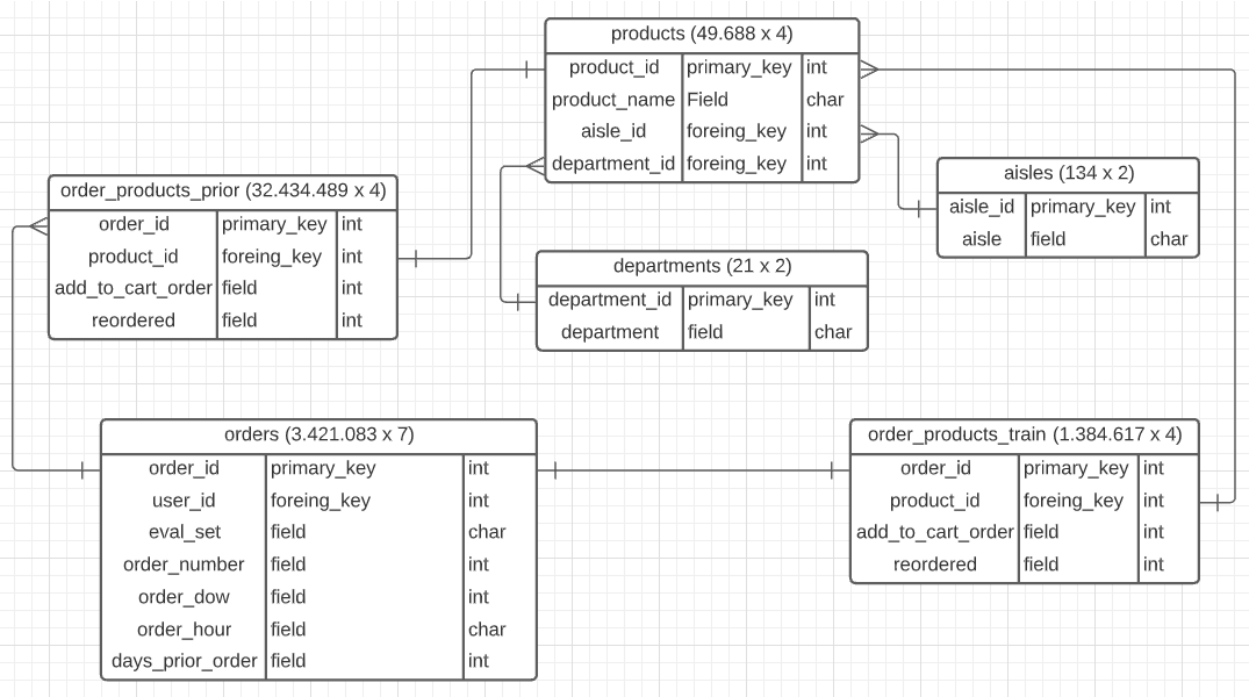
O Banco de dados serão coletados na plataforma Kaggle por meio do download de um aqruivo .zip contendo as 5 tabelas utilizando o comando.

```
>_ kaggle competitions download -c instacart-market-basket-analysis
```

2.2 Descrição dos Dados

A descrição dos dados compreendeu levantamento de meta informações sobre o banco de dados tais como, formato dos arquivos de cada tabela, número de observações e número de atributos de cada tabela, tipo de dado dos atributos de cada tabela e identificação de chaves primárias e estrangeiras para junção das tabelas.

Esquema Relacional das Tabelas



- O arquivo plano **order_products_prior.csv** com separador de vírgulas contém informações sobre 32.434.489 ordens. Os atributos permitem a recuperação de uma chave primária identificadora da ordem, uma chave estrangeira identificadora do produto, a ordem de adição do produto ao carrinho de uma determinada compra e quantas vezes esta ordem já foi reordenada.
- O arquivo plano **orders.csv** com separador de vírgulas contém informações sobre 3.421.083 ordens. Os atributos permitem a recuperação de uma chave primária identificadora da ordem, uma chave estrangeira identificadora do cliente, a categoria da observação no que diz respeito ao particionamento, o número da ordem, o dia da semana, a hora e o número de dias após a última ordem.
- O arquivo plano **products.csv** com separador de vírgulas contém informações sobre 49.688 produtos. Os atributos permitem a recuperação de uma chave primária identificadora do produto, duas chaves estrangeiras para identificar corredores e departamentos e um atributo para recuperar o nome dos produtos.

- O arquivo plano **departamentos.csv** com separador de vírgulas contém informações sobre 21 departamentos e seus respectivos nomes.
- O arquivo plano **corredores.csv** com separador de vírgulas contém informações de 134 corredores da loja onde os produtos podem ser encontrados e seus respectivos nomes.
- O arquivo plano **order_products_train.csv** com separador de vírgulas contém informações sobre 1.384.617 ordens. Os atributos permitem a recuperação de uma chave primária identificadora da ordem, uma chave estrangeira identificadora do produto, a ordem de adição do produto ao carrinho de uma determinada compra e quantas vezes esta ordem já foi reordenada. Devido aos seu tamanho e as informações que disponibiliza, essa foi a principal tabela utilizada no projeto.

Valores Únicos Banco de Dados de Treino

Atributo	Número de Observações Únicas
order_id	131209
product_id	39123

- O banco de dados de treino contém informações sobre 131.209 ordens distintas com 39.123 produtos distintos.

2.3 Verificação da Qualidade dos Dados

A verificação da qualidade dos dados consistiu na identificação de dados faltantes e na verificação da integridade das estruturas das tabelas e das informações nelas contidas.

Para avaliar a integridade dos dados foram construídos gráficos e tabelas com intuito de identificar se o tipo de dado de cada observação é compatível com o tipo de dado do respectivo atributo a que pertence. Também foi verificado se os valores de cada atributo possuíam consistência. Finalmente foi verificado se as tabelas possuíam algum tipo de incoerência no formato dos arquivos através da inspeção dos resultados das importações comparados com os esquemas esperados.

Para identificar dados faltantes foram produzidos gráficos e tabelas dos atributos de cada arquivo plano da da base de dados com intuito de identificar a quantidade de dados faltantes em cada campo de cada tabela, identificar como esse valores faltantes foram representados, avaliar em que categoria de dados faltantes a que esses valores pertenciam e assim determinar se tais valores deveriam ser removidos ou sugerir técnicas de imputação apropriadas.

Dados Faltantes

Atributo	Número de Observações Faltantes
oder_id	0
product_id	0
add_to_cart_order	0
reordered	0

- Nenhum atributo apresentou dados faltantes.

Primeiras 10 Observações

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1
1	13176	6	0
1	47209	7	0
1	22035	8	1
36	39612	1	0
36	19660	2	1

- Nenhum atributo apresentou observações com tipo de dado inconsistente com o esquema relacional.

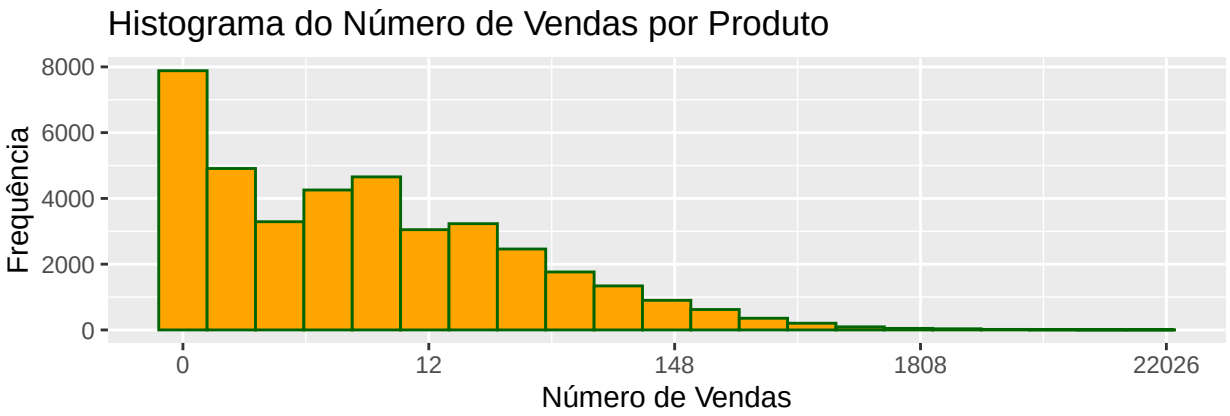
2.4 Exploração dos Dados

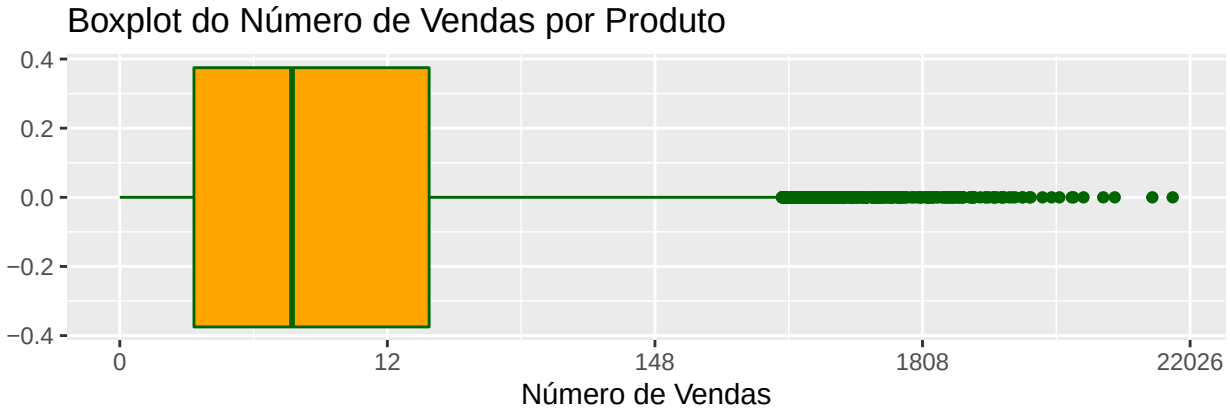
A exploração dos dados coinsistiu na produção e análise de gráficos, tabelas e testes para identificar padrões aparentes que colaborassem na etapa de modelagem, determinar os dados relevantes para atender aos objetivos do negócio, identificar valores atípicos e sugerir possíveis técnicas de engenharia de variáveis para melhorar a capacidade preditiva dos modelos.

Número de Vendas por Produto

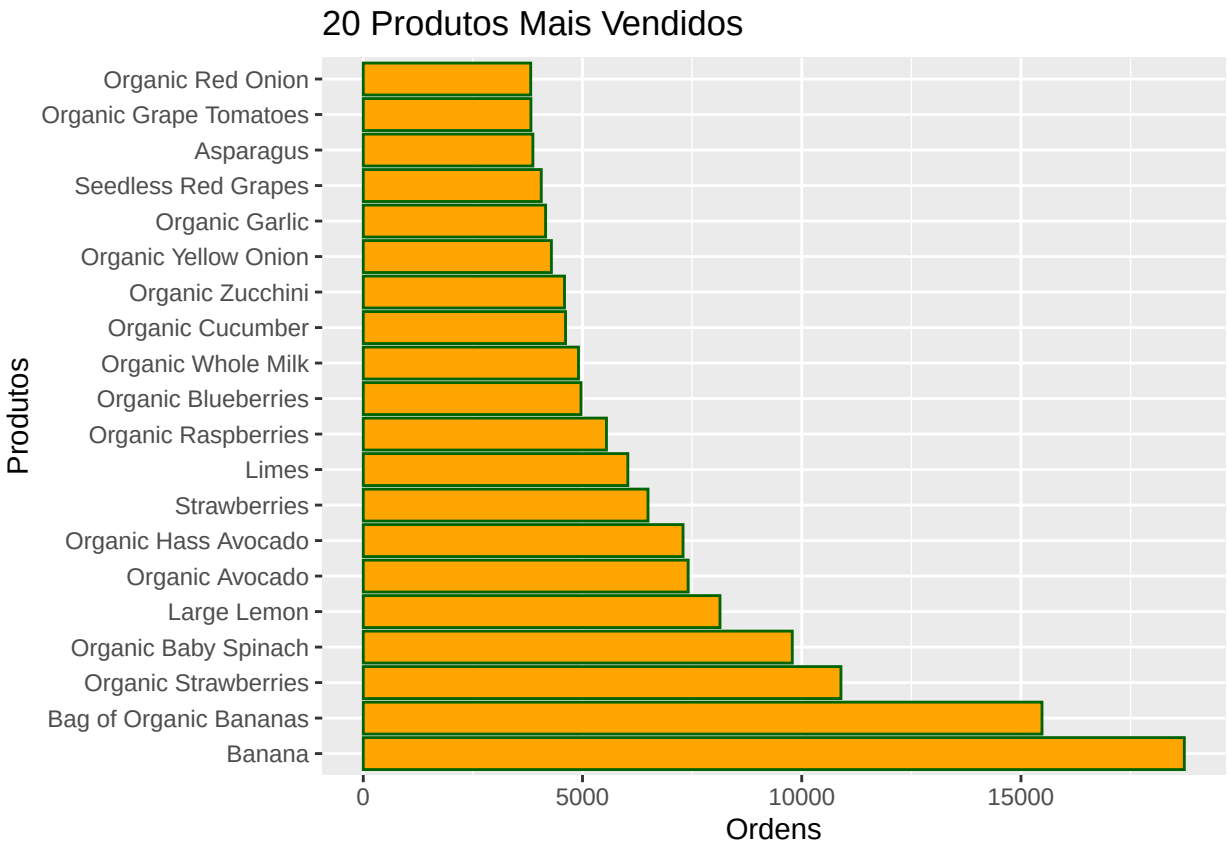
Estatística	Valor
Mínimo	1.00000
1º Quartil	2.00000
Mediana	5.00000
Média	35.39138
3º Quartil	18.00000
Máximo	18726.00000
Desvio Padrão	222.53341

- 7.884 produtos só foram comprados 1 vez e apenas o produto Banana atingiu o número máximo de vendas com um total de 18.726 ordens.
- A mediana de 5 vendas por produto destoa bastante da média de 35 vendas por produto. Tal fato indicou assimetria negativa com cauda à direita. Ou seja, a distribuição do número de venda por produto se concentrou em até 20 produtos, porém diversos produtos apresentaram valores atípicos com magnitude maior que 2 desvio padrão em relação à média.





- O histograma do número de vendas por produto ilustrou as estatísticas de quartis, medidas de centralidade e dispersão do número de vendas dos 39.113 produtos. O histograma acrescenta a informação sobre o peso da cauda, sendo possível observar que apesar da maioria dos produtos apresentar entre 1 e 20 vendas, existem milhares de produtos que apresentaram mais que 20 vendas e centenas de produtos que apresentaram mais que 200 vendas.
- O boxplot acrescenta a informação de que apesar da mediana de 5 vendas por produto compensar os valores atípicos e estimar melhor a centralidade dos dados, foi possível observar que a caixa não parece apresentar estado de inércia ou equilibrado no valor 5. Também foi possível constatar que dezenas de produtos apresentaram mais do que 500 vendas.

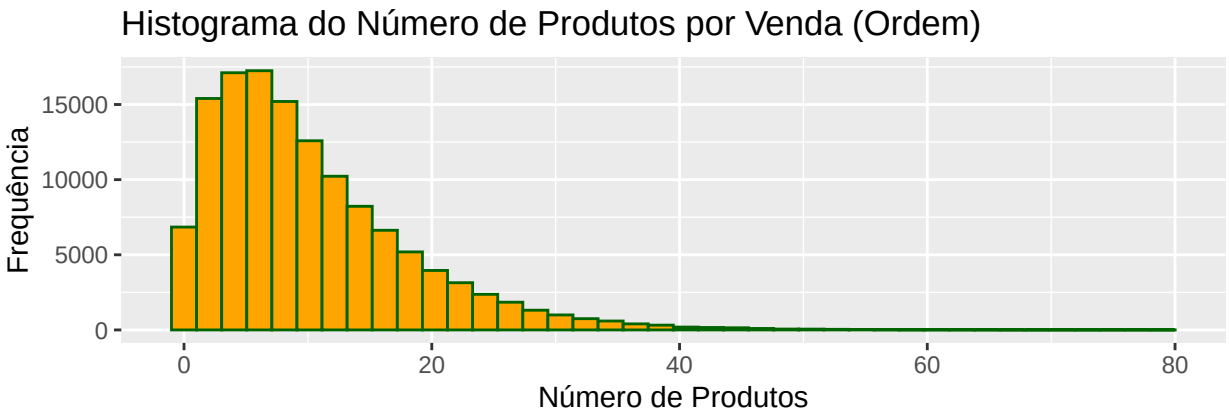


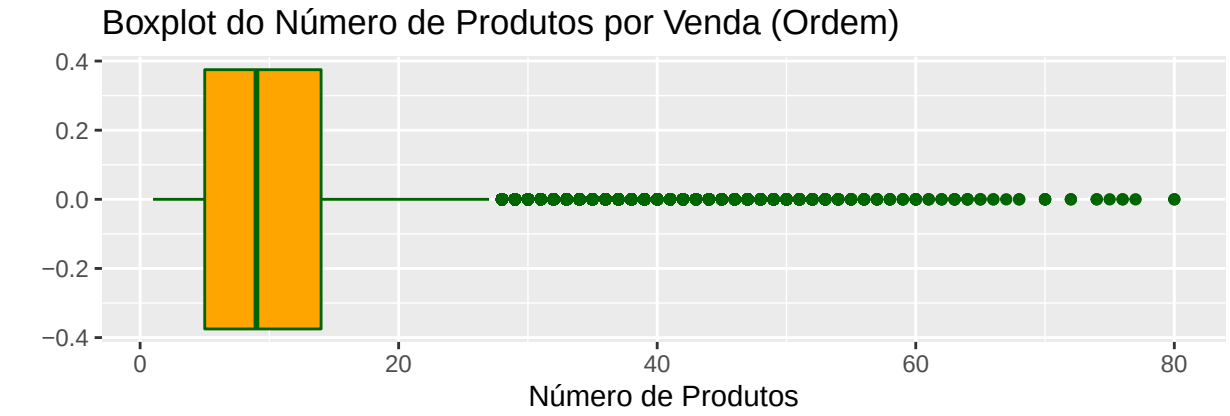
- Os cinco produtos mais vendidos foram bananas, bananas orgânicas, morangos orgânicos, espinafre orgânico e limão.
- Frutas e legumes orgânicos compõem a maioria dos 20 produtos mais vendidos.

Número de Produtos por Venda (Ordem)

Estatística	Valor
Mínimo	1.000000
1º Quartil	5.000000
Mediana	9.000000
Média	10.552759
3º Quartil	14.000000
Máximo	80.000000
Desvio Padrão	7.932847

- O número de produtos por venda(ordem) varia entre 1 e 80 produtos com maioria entre 1 e 15 produtos por venda nas 131.199 vendas (ordens).
- A maioria das ordens apresentou vendas com 5 a 15 produtos.
- A mediana de 9 produtos apresentou valores semelhantes indicando que o espalhamento dos dados ao longo da amplitude é regular. Tal fato pode ser comprovado pela magnitude do desvio em relação à amplitude dos dados. Ou seja, apesar de existirem vendas com até 80 produtos, a maioria dos dados está concentrada em torno da média com poucos valores atípicos.





- O histograma do número de produtos por ordem indicou distribuição assimétrica à direita. Apesar de apresentar valores atípicos, a cauda à direita é leve com poucas vendas que contêm mais de 40 produtos. A diminuição do número de produtos por venda foi gradual, atingindo um número de vendas baixo depois de 40 produtos por venda.

O boxplot indicou que a mediana estava centralizada entre o 1º e 3º quartis. Alguns valores atípicos podem ser observados acima de 27 produtos por venda (ordem). Apenas 6 vendas contiveram mais do que 70 produtos.

3 Preparação dos Dados

A etapa de preparação dos dados consistiu na seleção dos conjuntos de dados e atributos relevantes e apropriados para modelagem de padrões frequentes bem como a engenharia de atributos, formatação, limpeza e integração dos dados.

3.1 Seleção de Dados

A análise exploratória mostrou que os arquivos planos **order_products_train.csv** e **products.csv** separados por vírgula possuem informações relevantes e apropriadas para a implementação de técnicas de padrões frequentes.

Os atributos **order_id** e **product_id** do arquivo **order_products_train.csv** que contém as informações de identificadores das transações e produtos foram selecionados para serem utilizados pelo algoritmo apriori para mineração de padrões frequentes.

O atributo **product_name** do arquivo **products** contem as informações dos nomes dos produtos e foi selecionado para recuperar os nomes dos produtos que estão identificados pelo atributo **product_id** do arquivo **order_products_train.csv**.

Amostra do Arquivo order_products_train.csv

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1
1	13176	6	0

- Apenas os 2 primeiros atributos desta tabela foram utilizados.

Amostra do Arquivo products.csv

product_id	product_name	aisle_id	department_id
11	Peach Mango Juice	31	7
12	Chocolate Fudge Layer Cake	119	1
13	Saline Nasal Mist	11	11
14	Fresh Scent Dishwasher Cleaner	74	17
15	Overnight Diapers Size 6	56	18
16	Mint Chocolate Flavored Syrup	103	19

- Apenas os dois primeiros atributos desta tabela foram utilizados.

3.2 Limpeza dos Dados

Todos os arquivos contidos na fonte de dados da competição já foram limpos e tradados e portanto esta etapa não precisou ser realizada neste projeto. Uma possível limpeza seria redução do nome dos produtos que em alguns casos apresenta descrição com frases extensas. Porém as principais palavras que descrevem os produtos nas frases não apresentam padrão regular de posicionamento e portanto a redução envolveria técnicas de processamento natural da linguagem para identificação de nomes que foge do escopo deste trabalho.

3.4 Junção dos Dados

O primeiro atributo, que representa a chave primária da tabela **products** será utilizado para junção com a tabela **order_products_train** que contem a chave estrangeira identificador do produto e o segundo atributo será utilizado para recuperar os nomes dos produtos.

Amostra da Tabela com Junção

order_id	product_name
36	Organic Half & Half
36	Super Greens Salad
36	Cage Free Extra Large Grade AA Eggs
36	Prosciutto, Americano
36	Organic Garnet Sweet Potato (Yam)
36	Asparagus

A etapa de modelagem consistiu no uso da técnica Apriori de mineração de padrões frequentes para identificar conjuntos de produtos comprados frequentemente na empresa Instacart. Desta forma a empresa poderá sugerir produtos para um cliente baseado em seus padrões de compras comparados aos padrões de compras de outros clientes em ordens anteriores.

Algoritmo Apriori

O algoritmo apriori foi utilizado na versão do pacote **arules** da linguagem R implementado em na linguagem C++ para garantir velocidade na execução em bancos de dados com milhões de transações. O algoritmo basicamente calcula:

- O suporte de um grupo de produtos representa o percentual de vezes em que esse conjunto de produtos pode ser observado considerando todas as compras.
- A confiança de um grupo de produtos que representa o percentual de vezes em que duas coleções de itens são compradas juntas dado que a primeira coleção tenha sido comprada.
- A elevação de uma regra de associação que representa a correlação de um conjunto de produtos com outro conjunto de produtos em um regra. Se o lift é 1 os conjuntos não possuem relação, se é maior que 1 os conjuntos são dependentes e se é menor do que 1, um conjunto terá efeito negativo sobre a presença do outro conjunto.

Importação dos Dados Classe Transaction

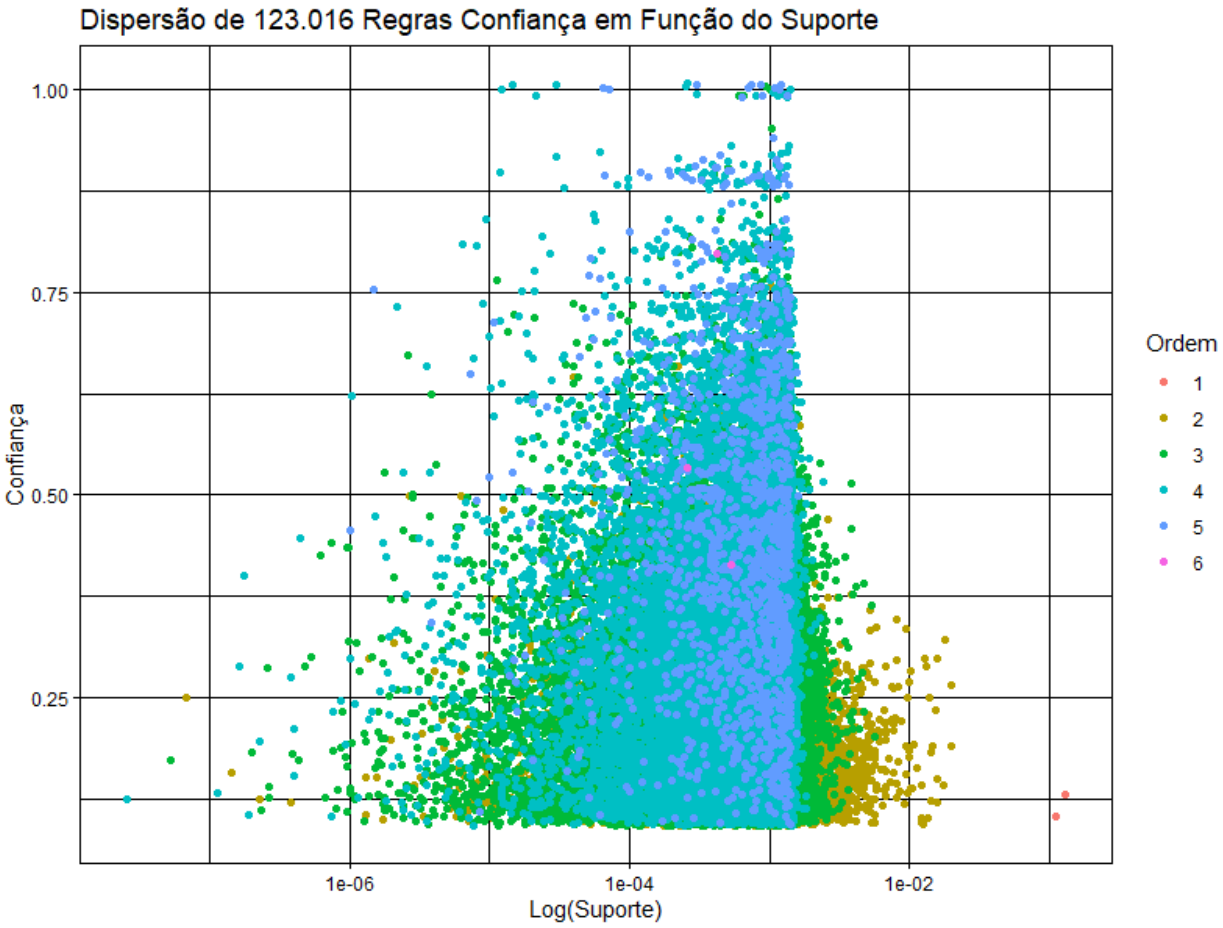
Para que o algoritmo funcione adequadamente os dados devem estar no formato do tipo de dado classe especificamente construído para representar as transações do mundo real. Este objeto da classe transaction possui atributos e métodos necessários para implementação do algoritmo.

Esta classe basicamente disponibiliza uma matriz binária esparsa com compras nas linhas e produtos nas colunas indicando a presença ou ausência de determinados produtos em cada compra. Também podem ser encontradas metas sobre as compras tais como, o número de transações e produtos, os nomes das transações e produtos e uma tabela da distribuição do número de produtos por compras.

Parametrização

Suportes maiores como 0.1 resultaram em nenhuma regra de associação. Para determinar a extensão dos parâmetros de suporte e confiança foi elaborado um modelo base. Outros 3 modelos foram desenvolvidos com diferentes especificações dos parâmetros.

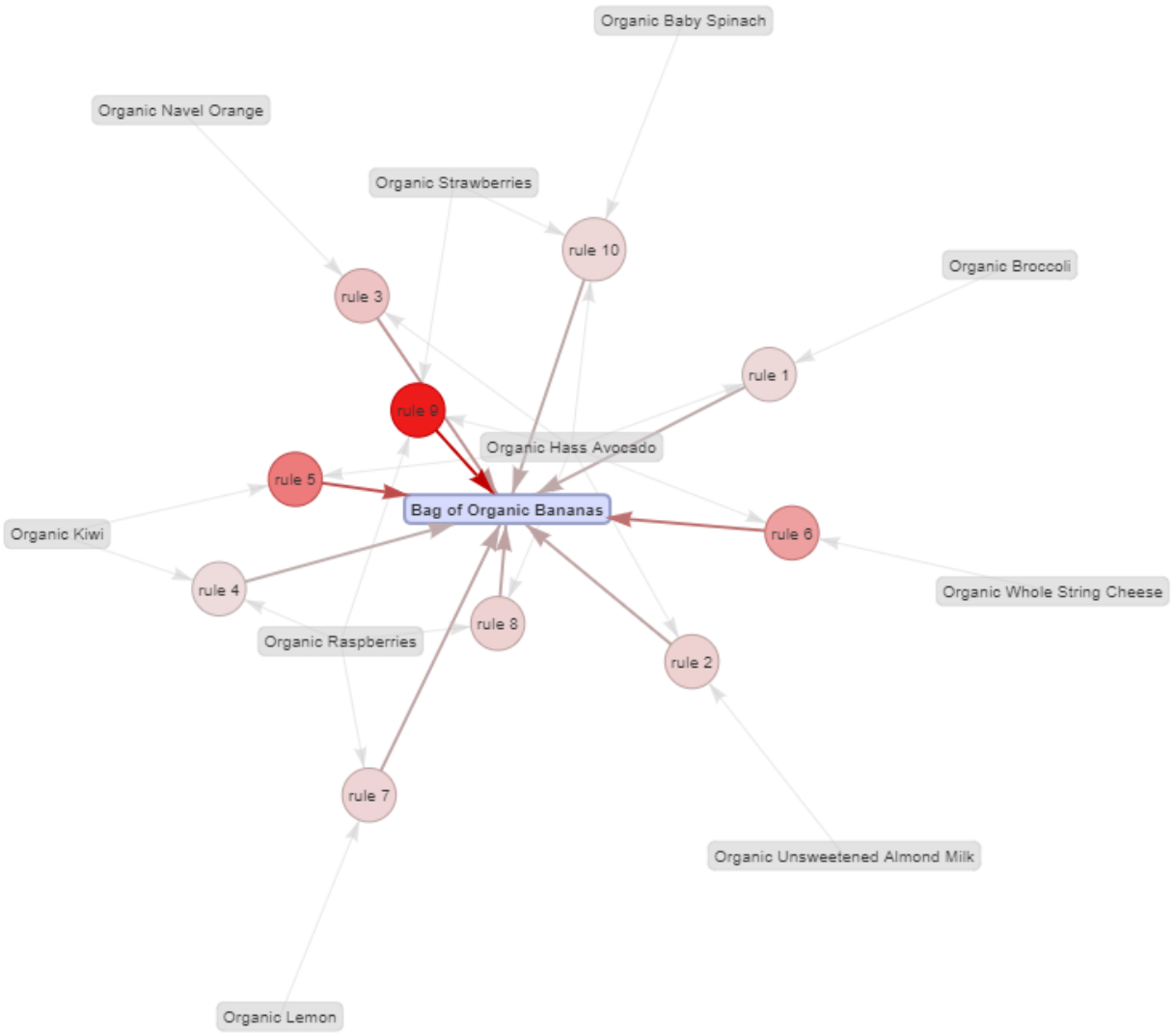
O modelo base considerou o suporte de 1e-8 e confiança mínima de 0.1 utilizados como linha de base para explorar as regiões com maior confiança entre os suportes definidos.



- Os pontos foram agitados para evitar sobreposição. A ordem de uma regra, ou seja, no caso de ordem 3, 2 produtos estão associados a um produto, foram coloridas para separar os tamanhos de conjuntos das de cada regra.
- A confiança exibe uma relação de correlação positiva com o suporte de 1e-8 até 1e-3. A partir deste ponto é possível observar a queda da confiança à medida que o suporte aumenta, ou seja, uma relação de correlação negativa.
- As regras de ordem 3, 4, 5 apresentaram as maiores confianças no intervalo de suporte entre 1e-5 e 1e-3.

4.1 Modelo 1

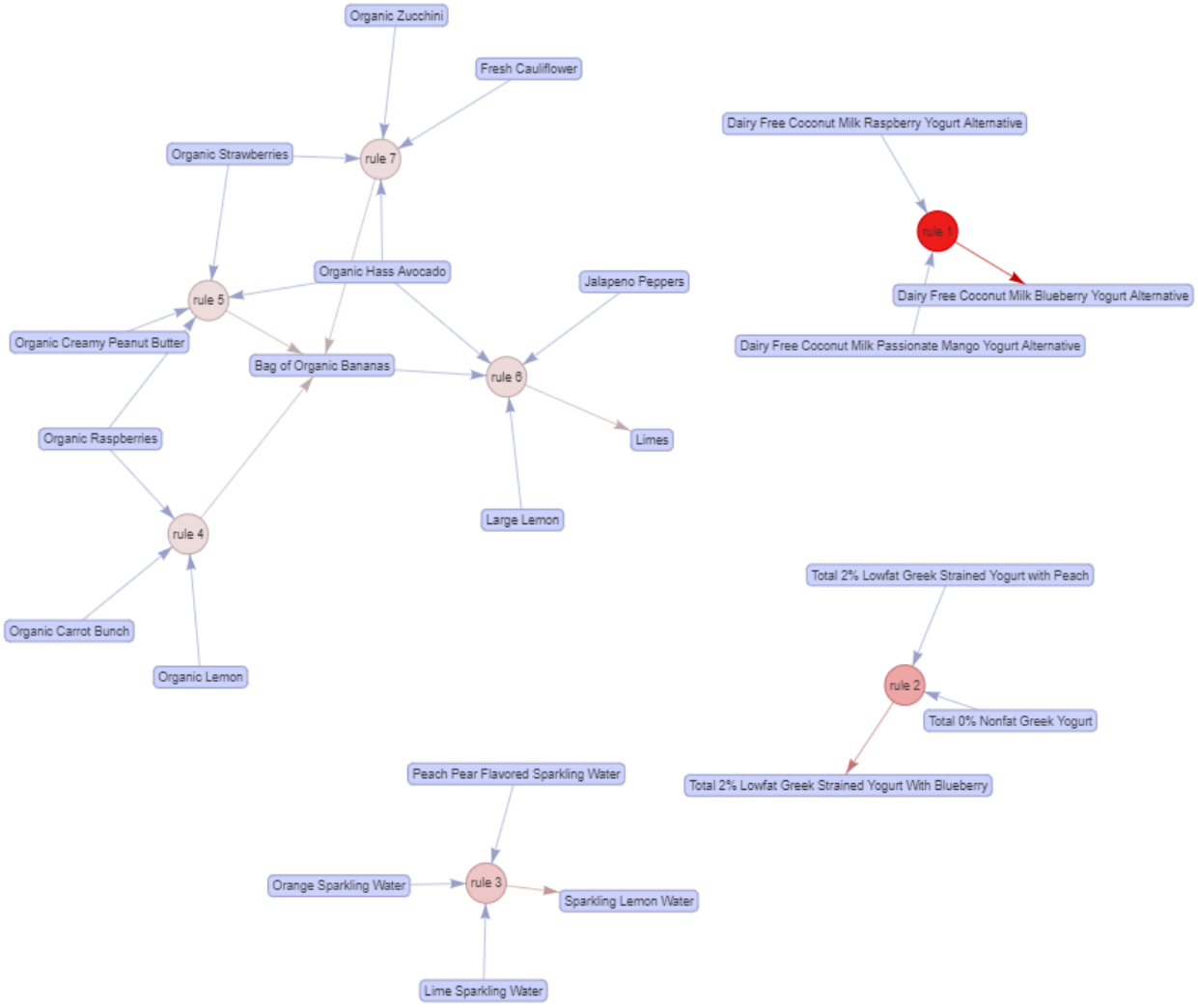
O modelo 1 considerou o suporte mínimo de 0.001 e confiança mínima de 0.51 para filtrar as regras com maior confiança e maior suporte.



- Com confiança mínima de 0.51 e suporte mínimo de 0.001 o algoritmo encontrou 10 regras com conjuntos de 2 produtos associados a um saco de Bananas Orgânicas.

4.2 Modelo 2

O modelo 2 considerou o suporte mínimo de 0.000125 e confiança mínima de 0.95 para filtrar as regras com maior confiança e com suportes maiores.



- Com confiança mínima de 1 e suporte mínimo de 0.000125 o algoritmo encontrou 7 regras com 2 conjuntos de 2 produtos associados a um outro produto, 2 conjuntos de 3 produtos associados a um outro produto e 3 conjuntos de 4 produtos associados a um outro produto.

4.3 Modelo 3

O modelo 3 considerou o suporte mínimo de 0.0001 e confiança mínima de 0.8 para filtrar as regras com confianças e suportes aceitáveis, mas ao mesmo tempo gerar um número maior de regras para sistemas de recomendação na plataforma de comércio eletrônico da empresa.

- O modelo gerou 1117 regras distribuídas da seguinte forma:
  - 5 produtos estão associados a um outro produto.
  - 195 conjuntos de 2 produtos estão associados a um outro produto.
  - 660 conjuntos de 3 produtos estão associados a um outro produto.
  - 254 conjuntos de 4 produtos estão associados a um outro produto.
  - 3 conjuntos de 6 produtos estão associados a um outro produto.

5 - Avaliação

A avaliação dos modelos considerou a análise dos resultados de cada um dos 3 modelos bem como suas a coerência com os objetivos do negócio.

5.1 Modelo 1

Todas as regras que indicam um saco de bananas orgânicas são compostas de conjuntos de produtos também orgânicos. Tal fato revela um padrão de compras de clientes que preferem comprar todos os produtos orgânicos, o que faze sentido já que comprar apenas um determinado produto orgânico e os outros não orgânicos não removeria mas apenas diminuiria a presença de agrotóxicos na dieta.

O modelo 1 pode gerar entendimento sobre um grupo de clientes que preferem comprar todos os produtos o mais natural possível, mesmo que esses produtos possuam preços superiores aos seus similares não orgânicos. Para o negócio isso poderia implicar em anúncios de outros produtos orgânicos para os clientes com este tipo de dieta.

5.2 Modelo 2

Itens diferentes de produtos orgânicos foram encontrados associados a produtos orgânicos como na regra 6 que associa banana orgânica, abacate orgânico, pimenta jalapeno e limão grande com limas. Tal fato pode estar relacionado a um tipo de receita específica como guacamole em que itens que não são encontrados na categoria de orgânicos são comprados para produzir uma receita corretamente.

A regra 3 indica que produtos com clientes que compram água com gás nos sabores pêssego, laranja e lima também tendem a comprar água com gás no sabor limão. Tal fato pode indicar que alguns produtos do mesmo tipo são comprados conjuntamente em múltiplos sabores.

O modelo 2 pode gerar entendimento sobre as compras de produtos para a realização de uma receita ou sobre a diversidade de sabores para produtos de um mesmo tipo. Para o negócio tal fato pode indicar a identificação das receitas que são compostas por estes conjuntos de produtos encontrados no padrão de compras dos clientes e sugerir opções de ingredientes que compõe a receita para clientes que compram apenas alguns dos ingredientes e ainda não produzem as receitas.

5.2 Modelo 3

O modelo 3 com mais regras e com bom índice de confiança poderia ser utilizado para recomendar produtos de forma geral na plataforma de comércio eletrônico da Instacart. Desta forma clientes com itens já adicionados ao carrinho de compras poderiam receber sugestões de produtos para serem adicionados ao carrinho no processo de comprar.

6 - Implantação

6.1 Modelo 1

O modelo 1 poderia ser utilizado pela equipe de marketing direcionado para sugerir novos produtos orgânicos para consumidores que apresentam padrões de carrinho de compras de somente produtos orgânicos, ou mesmo produtos orgânicos antigos para clientes do grupo de orgânicos que compram produtos não orgânicos em função de desconhecerem as opções orgânicas.

6.2 Modelo 2

O modelo 2 poderia ser utilizado pela equipe de marketing direcionado para sugerir todos os produtos de uma receita para clientes que não conhecem as receitas mas compram alguns dos produtos contidos nas mesmas. Juntamente com a recomendação a receita poderia ser incluída no anúncio como fator motivador para a realização da compra.

6.3 Modelo 3

O modelo 3 com mais de 100 regras poderia ser utilizado pela equipe de desenvolvimento do comércio eletrônico com objetivo de ordenar a sugestão de itens para serem acrescentados no processo de compras baseado nas regras de associação. Desta forma um número maior de produtos do portfólio da empresa seria recomendado com base na inteligência artificial gerada pelas regras de associação na mineração de padrões de compras frequentes.

7 - Fontes e Referências

- Instacart
- kaggle
- Zaki, Mohammed J., and Wagner Meira. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press.
- Lucidcharts
- Tidyverse
- Naniar
- Arules
- ArulesViz
- DataCamp
- Learn by Merketing