UNIVERSIDADE DE SÃO PAULO

Instacart: agrupamento de clientes e análise de cesta de compras

Autores:
Cristiano di Maio Chiaramelli
Douglas Seiti Kodama
Filipe Mariano Freire da Silva
Marcelo Tabacnik
Raul Wagner Martins Costa
Thauan Leandro Gonçalves
Vitor Giovani Dellinocente

Professora:
Solange Oliveira Rezende

28 de Novembro de 2017

1 Introdução

Este projeto foi realizado com o intuito de obter conhecimento útil a partir da base de dados da plataforma Instacart, que fornece a seus usuários a possibilidade de fazer pedidos de mantimentos e recebê-los no mesmo dia a partir de estabelecimentos próximos ao usuário. Os detalhes da base de dados e sua organização estão descritos na Seção 2.

Uma das análises feitas sobre a base de dados está relacionada ao agrupamento de clientes, cujo objetivo é associar cada cliente a um grupo, de forma a obter a maior semelhança entre os clientes de um mesmo grupo. Tal estudo está detalhado na Seção 3.

Outra investigação realizada refere-se às regras de associação oriundas dos pedidos feitos por clientes. Tais regras são importantes para detectar padrões de compra e promover o marketing direcionado na plataforma. Tal investigação está especificada na Seção 4.

Por fim, na Seção 5, são elaboradas algumas considerações finais sobre como o conhecimento obtido poderia ser usado para prover melhorias à plataforma Instacart.

2 Detalhes da base de dados

A base de dados utilizada está organizada em 6 relações contendo informações sobre clientes e seus pedidos, produtos, corredores e departamentos. A Tabela 1 mostra a quantidade de itens para cada um desses grupos.

Tabela 1: Grupos e suas quantidades.

Clientes	206 208
Pedidos	3 421 083
Produtos	49 688
Corredores	134
Departamentos	21

Na Figura 1, é possível observar que a probabilidade do número de compras decresce exponencialmente, isto é, há uma probabilidade maior de selecionar um cliente com poucos pedidos do que um com muitos.

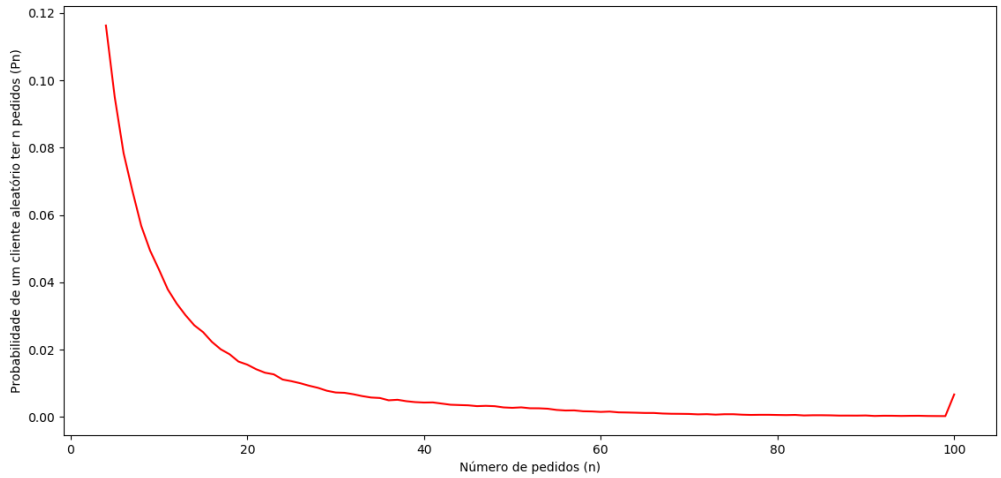


Figura 1: Distribuição de probabilidade do número de pedidos por cliente, isto é, a probabilidade de selecionar um cliente aleatoriamente na base de dados e ele ter um certo número de pedidos.

Um comportamento similar está presente na distribuição de produtos por pedido, ilustrada na Figura 2. Há um pico global na faixa 10 ± 10 produtos (na qual grande parte dos pedidos se situa), após o qual a probabilidade decresce exponencialmente. Isso implica que há uma probabilidade muito alta de um pedido ter poucos produtos e uma probabilidade muito baixa de ter muitos.

Por fim, a Figura 3 revela um comportamento senoidal na distribuição de probabilidade de produtos em um certo dia e horário da semana. É possível perceber que a grande maioria dos pedidos é feita entre o dia e a tarde, isto é, em torno das 10 as 15 horas.

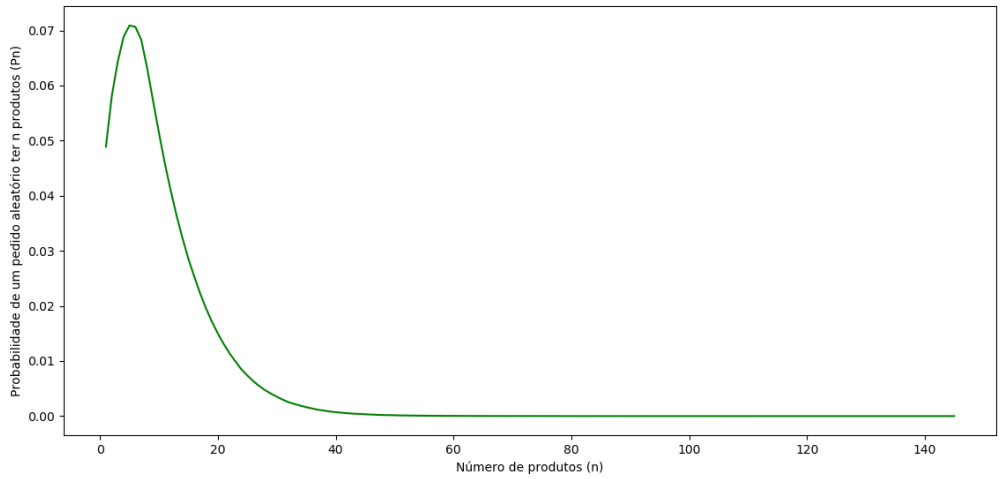


Figura 2: Distribuição de probabilidade do número de produtos por pedido, isto é, a probabilidade de selecionar um pedido aleatoriamente na base de dados e ele ter um certo número de produtos.

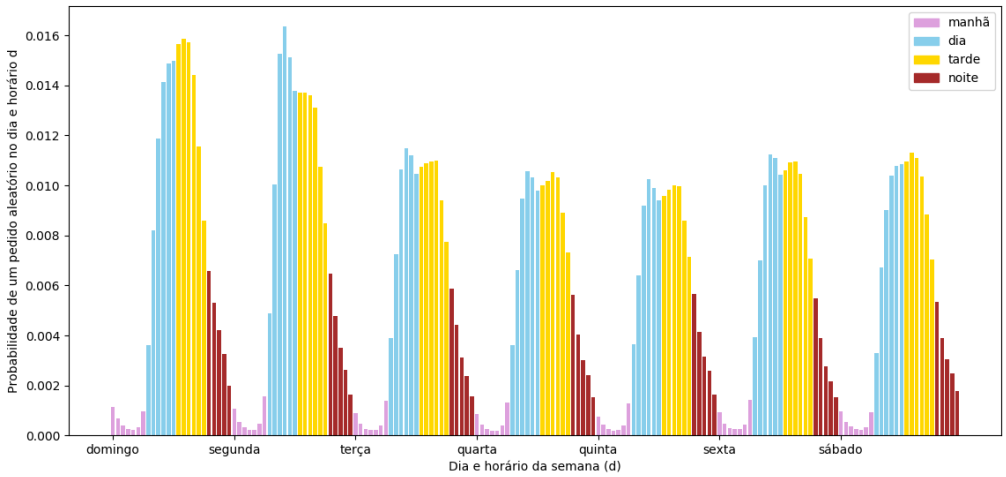


Figura 3: Distribuição de probabilidade do número de pedidos realizados por dia e horário da semana.

3 Agrupamento de clientes

O agrupamento tem como objetivo associar cada cliente a um grupo, de forma que cada cliente em seu grupo compartilhe características semelhantes em relação aos outros clientes desse grupo. Tais características (ou atributos) são escolhidas de modo a melhorar a separação dos grupos (ou comunidades) e obter um agrupamento compreensível.

Outro detalhe importante no agrupamento é definir uma função que determine a similaridade entre dois clientes. Tal função é altamente dependente dos atributos escolhidos e pode influenciar bastante na formação das comunidades.

Por fim, é preciso definir o tipo de agrupamento que será utilizado, se divisivo ou hierárquico. Os algoritmos divisivos, como o K-means, buscam separar os dados em partições disjuntas, enquanto que os algoritmos hierárquicos, como o de Ward, procuram construir árvores nas quais cada nó representa um agrupamento possível dos dados.

3.1 Pré-processamento

Na etapa de pré-processamento, procura-se organizar, selecionar e modificar todos dados disponíveis de modo que apenas as informações necessárias sejam passadas ao algoritmo de agrupamento. Dessa forma, 3 atividades são feitas durante o pré-processamento, são elas:

- Todos arquivos em disco da base de dados são trazidos à memória principal. O conteúdo desses arquivos é então organizado em classes representativas de: clientes, pedidos, produtos, corredores e departamentos; cada uma dessas classes contendo diversos atributos que a caracterizam.

- Escolhe-se alguns atributos da classe cliente para serem usados no agrupamento. Um exemplo de conjunto de atributos utilizado neste projeto é: número de pedidos do cliente, recorrência média (em dias) do cliente para fazer um pedido e horário de compra da semana mais comum do cliente.

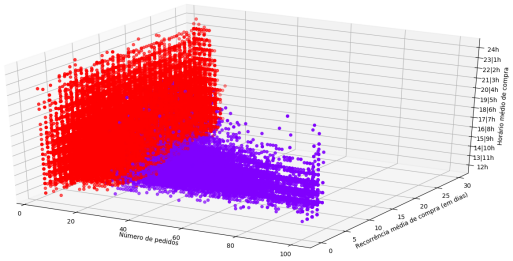
- Baseado nos atributos escolhidos, cria-se uma matriz de dimensão m por n, sendo m o número de clientes a serem agrupados e n o número de atributos escolhidos. Este é o conjunto final pré-processado que será passado para o algoritmo de agrupamento escolhido.

Após o pré-processamento, pode-se realizar o agrupamento através de algum algoritmo. Neste projeto, foi escolhido o K-means pela sua simplicidade de implementação e compreensão, bem como aplicabilidade para grandes bases de dados. Foram utilizados três atributos dos clientes, aqueles já descritos anteriormente em 3.1, bem como um número de grupos (ou *clusters*) variável.

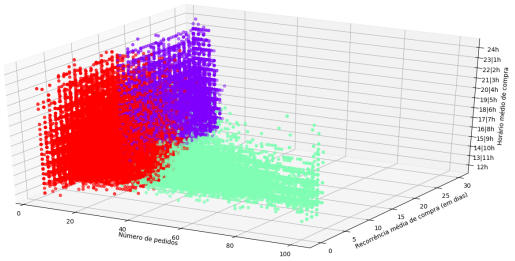
A Figura 4 ilustra alguns exemplos de agrupamento feito pelo K-means com diferentes número de grupos.

3.2 Pós-processamento

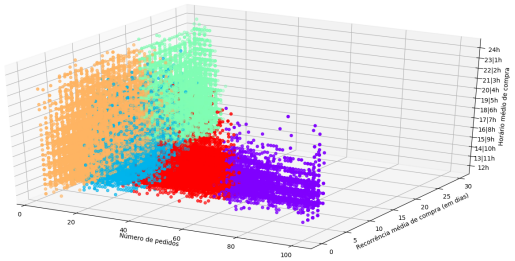
A fase de pós-processamento é responsável por realizar uma avaliação do conhecimento obtido, sendo comparado ao conhecimento do especialista. Responsabiliza-se também por identificar



(a) K = 2



(b) K = 3



(c) K = 5

Figura 4: K-means com diferentes números de *clusters*.

erros nas fases de pré-processamento e de reconhecimento de padrões, assim como validar e dar confiabilidade aos dados obtidos.

O conhecimento deve sair desta fase com garantia de sua validade e utilidade no processo de tomada de decisão, responsável pela fase de utilização do conhecimento. Assim sendo, o principal objetivo a que propomos utilizando o agrupamento foi a identificação de padrões de compras entre os clientes com base nos atributos citados na seção 3.1. Assim, vamos descrever os resultados obtidos levando em consideração inicialmente a influência que os atributos exercem sobre os resultados, e possíveis relações entre si e, por fim, descreveremos os agrupamentos obtidos e as descrições dos elementos de cada grupo.

3.2.1 Relações entre atributos

1. *Horário médio das compras*: percebemos que o horário predominante de compras está compreendido entre as 9 horas da manhã e as 15 horas da tarde, principalmente em padrões de compras com recorrência média baixa de compras;

2. *Recorrência média de compra e números de pedidos*: percebemos claramente a relação entre o número de compras e a recorrência média dos clientes. Os clientes que realizam mais compras são os clientes que possuem um tempo de recorrência menor, o que segue a intuição e o conhecimento do especialista (padrões comuns de compras);

3. *Número de pedidos, recorrência média de compra e horário médio de compra*: a principal relação que identificamos foi que clientes com baixo número de compras (com recorrência média baixa) se distribuem mais em relação ao horário que realizam as compras. Ao contrário, clientes com alto número de compras (com recorrência média alta) realizam compras predominantemente no período entre 9 horas da manhã e 15 horas da tarde, como citado no item 1.

3.2.2 Padrões de compras

Sabemos a influência que o número K (número de *clusters*, grupos) possui no agrupamento e identificação dos padrões. Iremos nos basear principalmente na Figura 4(b), a qual acreditamos possuir padrões de compras mais claros de serem descritos. Para tal, vamos utilizar duas categorias de classificação: Tipo de compras e Tempo de sistema.

No que diz respeito ao Tipo de compras, os clientes podem realizar compras maiores (com muitos itens) e menores (com poucos itens). Isso é responsável por influenciar o número de compras dos clientes e a recorrência média de compra.

No que diz respeito ao Tempo de sistema, os clientes podem ser classificados como recentes (que começaram a usar o sistema há pouco tempo) e antigos (que já usam há mais tempo). Isso permite diferenciar os clientes que possuem poucas compras por serem recentes ou por padrão intrínseco de compra.

- *Padrão de compra 1 (verde)*: este padrão representa os clientes que possuem o padrões de compras pequenas, que se caracteriza por comprar com grande frequência mas levar poucos produtos por vez. Isto justifica o número altíssimo de compras que tais clientes realizaram;

- *Padrão de compra 2 (roxo)*: este padrão representa os cliente que possuem padrões de compras grandes, que se caracteriza pela baixa frequência de compras mas levando vários itens por vez. Devido ao padrão de compra (grande e esparso), concluímos que tais clientes provavelmente já possuem uma tendência a emprego fixo (e até famílias), o que justifica a variação de compras em relação ao horário do dia em que são realizadas;

- *Padrão de compra 3 (vermelho)*: este padrão representa clientes que possuem padrões de compras semelhantes aos padrões de compras 1, mas que ainda são clientes recentes (ainda não realizaram muitas compras no sistema). Percebe-se a tendência de transição entre os dois padrões quando se analisa a figura 4(b) onde o número de pedidos está por volta dos 30, ficando mais claro ainda analisando a Figura 4(c) (transição azul-vermelho).

3.3 Utilização do conhecimento

Após as validações adequadas realizadas pela fase de pós-processamento, pode-se agora integrar o conhecimento obtido em uma plataforma que possibilite e facilite a tomada de decisão, satisfazendo assim as necessidades iniciais do projeto.

Dentre as principais atividades que podem ser realizadas com o conhecimento obtido, estão a promoção do marketing direcionado, levando em consideração às características dos clientes de cada grupo, o uso de promoções específicas, a identificação de prováveis hábitos, tomando iniciativa para explorá-los, entre outros.

4 Regras de associação

A associação tem por objetivo descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados levando em conta suas características. Um ponto importante nas regras de associação é definir os valores de suporte e confiança. O suporte é o número de vezes que um elemento X aparece em um certo conjunto de dados D sobre o número total de elementos desse conjunto, enquanto que a confiança é o suporte de uma regra de associação $(X \implies Y)$ sobre o suporte de X. Nesse projeto, encontramos as associações existentes entre produtos comprados, e portanto, utilizamos apenas os dados presentes na tabela de pedidos e de produtos.

4.1 Pré-processamento

Nessa etapa organizou-se os dados de entrada, removendo atributos que não são relevantes para gerar regras de associação. Para isso removeu-se os atributos "add_to_cart_order" e "reordered" da tabela de pedidos. A tabela de pedidos tem portanto apenas dois atributos, order_id e product_id, relacionando cada produto comprado à uma compra.
Além disso, foi necessário reduzir o tamanho do conjunto de entrada, devido à limitação da memória utilizada pelos algoritmos escolhidos. Assim, podou-se de maneira aleatória a base de dados, com o intuito de reduzir a influencia disso nos resultados. Assim sendo, o conjunto de entrada possui dados referentes à 10.000 compras e 105.343 produtos comprados, o que representa uma média de 10,5 itens por compra. A tabela de Produtos apenas mapeia os $product_id$ para $product_name$, e portanto, não foi alterada durante a fase de pré-processamento.

4.2 Extração de padrões

Para a encontrar as associações usamos dois algoritmos distintos, a fim de poder comparar os seus resultados.

- O primeiro algoritmo é o FP-Growth, que consiste em montar uma árvore chamada de FP-Tree, cada nó da árvore contém um produto com o número de vezes que esse produto apareceu no *dataset*, e cada aresta representa uma associação com um outro produto. O algoritmo consiste em ir expandindo essa árvore e obtendo o suporte e a confiança das regras de associação. É uma maneira muito eficiente e rápida de encontrar uma solução, pois não temos a necessidade de recalcular o suporte e a confiança pra todas as associação a cada iteração, essa tarefa é realizada em O(n) utilizando a árvore, além disso o consumo de memória é bem menor pois o algoritmo mantém em memória uma versão compacta do *dataset*.Além disso o algoritmo é pouco paralelizável pois seus dados são muito dependentes.

- Já o outro algoritmo utilizado é o Apriori, que utiliza uma estratégia de busca *BFS*. Durante o primeiro passo, o algoritmo calcula o valor do suporte de cada item na base. Os itens cujo suporte é menor que o suporte mínimo são podados. No segundo passo, o Apriori calculará os suporte das possíveis duplas, combinando os itens ainda não podados. Seguindo o mesmo raciocínio, as duplas cujo suporte calculado não satisfaz o suporte mínimo serão podadas. O algoritmo continuará a repetir esses passos, até que não existam mais relações a calcular.

A fim de filtrar os resultados, precisamos tomar muito cuidado com as características da base. Como dito anteriormente, a mesma possui cerca de 17.000 produtos diferentes, e uma média de 10,5 produtos por compra, o que resulta em uma probabilidade média de apenas 0,06% de um produto qualquer estar presente em uma compra. Assim sendo, precisamos utilizar valores pequenos, de 0,3% para o suporte mínimo, e 25% para a confiança mínima de cada relação, podando as que não satisfazem tais valores. Vale ressaltar que essa escolha foi feita tanto para o algoritmo Apriori quanto para o FP-Growth.

A fim de filtrar os resultados, precisou-se tomar muito cuidado com as características da base. A mesma possui cerca de 17.000 produtos diferentes, e uma média de 10,5 produtos por compra, o que resulta em uma probabilidade média de apenas 0,06% de um produto qualquer estar presente em uma compra. Assim sendo, é necessário utilizar valores pequenos, de 0,3% para o suporte mínimo, e 25% para a confiança mínima de cada relação, podando as que não satisfazem tais valores.

4.3 FP-Growth X Apriori

Para a devida comparação, leva-se em conta que as bases usadas para ambos os algoritmos foi a mesma, suporte mínimo igual e confiança mínima iguais. A maioria das associações encontradas pelo Apriori foram encontradas pelo FP-Growth, porém esse segundo algoritmo encontrou muito mais regras. Em quesito de resultados, foram bem satisfatórios pois quando era encontrada uma regra usando ambos os algoritmos, sua confiança e suportes eram iguais. Quesitos de vantagens e desvantagens, o FP-Growth é melhor que o Apriori pois possui pouco consumo de memória e o tempo de execução é bem menor. O problema é que graças a construção da árvore os dados se tornam muito dependentes, assim esse algoritmo é de difícil paralelização, principalmente quando feito em uma organização com memória compartilhada. Além disso a cada iteração o Apriori recalcula todos os suportes novamente ($O(n^2)$), já o FP-Growth não, realizando essa tarefa em O(n).

4.4 Pós-processamento

A etapa de pós processamento é necessária para facilitar a interpretação dos dados. Do resultado encontrado pelo algoritmo Apriori, removeu-se o atributo $lift$ que não nos interessa. Assim, o arquivo de saída possui apenas os atributos $antecedants$, $consequents$, $support$, e $confidence$, representando o conjunto de produtos que implicam na compra de outro conjunto de produtos, com um valor de suporte e confiança. É nessa fase final que fez-se o mapeamento entre $product_id$ e $product_name$, utilizando a tabela de Produtos. Além disso, o resultado dos algoritmos retorna os IDs dos produtos. Assim, foi necessário fazer uma ”tradução”, substituindo os IDs pelos nomes.

4.5 Resultados

Foram selecionados apenas alguns dos resultados mais significativos dos nossos testes, onde a compra de alguns produtos leva à compra de outro.

Vale ressaltar que os valores de suporte e confiança são relativamente pequenos graças à grande diversidade de produtos presentes na base, fato que já foi elucidado anteriormente. Outro fato interessante que pode ser destacado dos resultados, é a presença majoritária dos

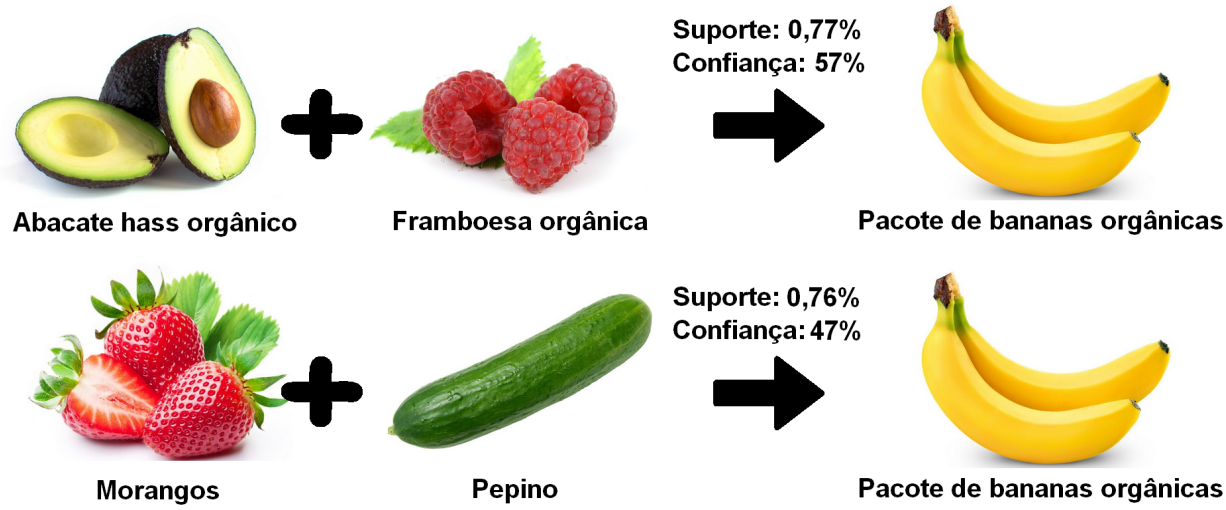


Figura 5: Algumas associações encontradas pelos nossos testes, onde a compra de dois produtos estimula a compra de outro

produtos *Pacote de bananas orgânicas* e *Banana* dentre as associações mais significativas. Esses produtos são, de fato, os mais comprados da base. Abaixo na Tabela 7 podemos ver todas as associações do tipo $(X \implies Y)$ geradas. Além dessas associações obteve-se resultados com associações do tipo $(X, Y \implies Z)$ que estão ilustradas na Tabela 8.

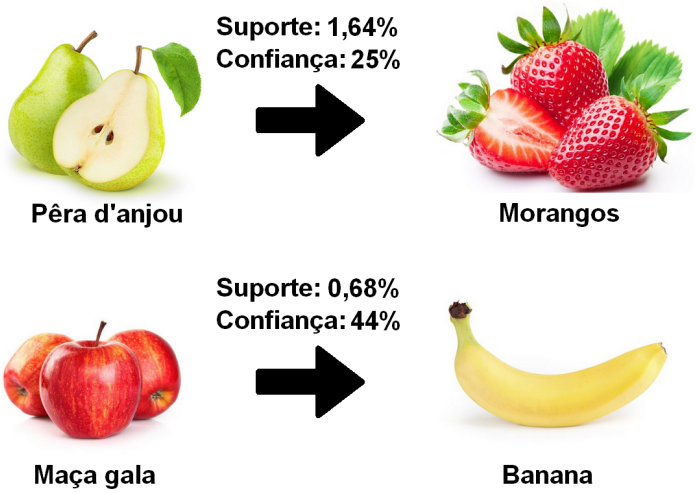


Figura 6: Algumas associações encontradas pelos nossos testes, onde a compra de um produto estimula a compra de outro

Antecedants	Consequents	Support	Confidence
['Michigan Organic Kale']	['Banana']	0.0209	0.258373205742
['Bunched Cilantro']	['Limes']	0.0145	0.310344827586
['Blueberries']	['Banana']	0.0168	0.27380952381
['Fresh Cauliflower']	['Bag of Organic Bananas']	0.0196	0.255102040816
['Organic Lemon']	['Bag of Organic Bananas']	0.0285	0.315789473684
['Organic Avocado']	['Banana']	0.0559	0.320214669052
['Honeycrisp Apple']	['Banana']	0.0288	0.361111111111
['Organic Cilantro']	['Limes']	0.0262	0.270992366412
['Organic Granny Smith Apple']	['Bag of Organic Bananas']	0.0161	0.291925465839
['Honey Nut Cheerios']	['Banana']	0.0097	0.309278350515
['Unsweetened Original Almond Breeze Almond Milk']	['Banana']	0.0113	0.29203539823
['Jalapeno Peppers']	['Limes']	0.0137	0.343065693431
['Organic Tomato Cluster']	['Bag of Organic Bananas']	0.0196	0.25
['Organic Large Extra Fancy Fuji Apple']	['Bag of Organic Bananas']	0.0229	0.349344978166
['Organic Whole Milk']	['Banana']	0.0372	0.25
["Organic D'Anjou Pears"]	['Bag of Organic Bananas']	0.0164	0.359756097561
['100% Whole Wheat Bread']	['Banana']	0.0158	0.303797468354
['2% Reduced Fat Milk']	['Banana']	0.0121	0.297520661157
['Granny Smith Apples']	['Banana']	0.0113	0.300884955752
['Organic Blueberries']	['Bag of Organic Bananas']	0.036	0.258333333333
['Organic Navel Orange']	['Bag of Organic Bananas']	0.015	0.42
['Bunched Cilantro']	['Banana']	0.0145	0.296551724138
['Limes', 'Large Lemon']	['Banana']	0.0133	0.255639097744
['Limes', 'Banana']	['Large Lemon']	0.0096	0.354166666667
['Sparkling Lemon Water']	['Sparkling Water Grapefruit']	0.0106	0.292452830189
['Organic Kiwi']	['Organic Strawberries']	0.0139	0.280575539568
['Red Vine Tomato']	['Banana']	0.0156	0.333333333333
['Organic Small Bunch Celery']	['Bag of Organic Bananas']	0.0215	0.251162790698
['Yellow Onions']	['Banana']	0.0282	0.27304964539
['Green Bell Pepper']	['Banana']	0.0198	0.262626262626
['Broccoli Crown']	['Banana']	0.0228	0.254385964912
['Organic Unsweetened Almond Milk']	['Bag of Organic Bananas']	0.0164	0.335365853659
['Boneless Skinless Chicken Breasts']	['Banana']	0.0155	0.316129032258

Figura 7: Tabela com regras de associação um pra um, com suporte mínimo de 0.003 (0.3%) e confiança mínima de 0.25 (25%). Resultados obtidos a partir da execução do algoritmo Apriori e FP-Growth.

Antecedants	Consequents	Support	Confidence
['Bag of Organic Bananas', 'Organic Cucumber']	['Organic Strawberries']	0.01	0.37
['Organic Strawberries', 'Organic Cucumber']	['Bag of Organic Bananas']	0.0078	0.474358974359
['Limes', 'Large Lemon']	['Organic Avocado']	0.0133	0.315789473684
['Large Lemon', 'Organic Avocado']	['Limes']	0.0111	0.378378378378
['Limes', 'Organic Avocado']	['Large Lemon']	0.0096	0.4375
['Bag of Organic Bananas', 'Organic Lemon']	['Organic Hass Avocado']	0.009	0.333333333333
['Organic Hass Avocado', 'Organic Lemon']	['Bag of Organic Bananas']	0.0064	0.46875
['Organic Baby Spinach', 'Banana']	['Organic Avocado']	0.0158	0.253164556962
['Organic Baby Spinach', 'Organic Avocado']	['Banana']	0.0107	0.373831775701
['Limes', 'Organic Baby Spinach']	['Organic Avocado']	0.0088	0.340909090909
['Limes', 'Organic Avocado']	['Organic Baby Spinach']	0.0096	0.3125
['Organic Baby Spinach', 'Organic Avocado']	['Limes']	0.0107	0.280373831776
['Bag of Organic Bananas', 'Organic Blueberries']	['Organic Strawberries']	0.0093	0.365591397849
['Organic Strawberries', 'Organic Blueberries']	['Bag of Organic Bananas']	0.01	0.34
['Organic Strawberries', 'Organic Baby Spinach']	['Bag of Organic Bananas']	0.0115	0.339130434783
['Bag of Organic Bananas', 'Organic Large Extra Fancy Fuji Apple']	['Organic Strawberries']	0.008	0.375
['Organic Strawberries', 'Organic Large Extra Fancy Fuji Apple']	['Bag of Organic Bananas']	0.0056	0.535714285714
['Bag of Organic Bananas', 'Organic Whole Milk']	['Organic Strawberries']	0.0086	0.383720930233
['Organic Whole Milk', 'Organic Strawberries']	['Bag of Organic Bananas']	0.0089	0.370786516854
['Bag of Organic Bananas', 'Organic Raspberries']	['Organic Hass Avocado']	0.0156	0.282051282051
['Organic Hass Avocado', 'Organic Raspberries']	['Bag of Organic Bananas']	0.0077	0.571428571429
['Limes', 'Banana']	['Organic Avocado']	0.0096	0.333333333333
['Limes', 'Organic Avocado']	['Banana']	0.0096	0.333333333333
['Bag of Organic Bananas', 'Organic Zucchini']	['Organic Strawberries']	0.0082	0.390243902439
['Organic Strawberries', 'Organic Zucchini']	['Bag of Organic Bananas']	0.0071	0.450704225352
['Bag of Organic Bananas', 'Organic Cucumber']	['Organic Hass Avocado']	0.01	0.3
['Organic Hass Avocado', 'Organic Cucumber']	['Bag of Organic Bananas']	0.0059	0.508474576271
['Organic Hass Avocado', 'Organic Baby Spinach']	['Bag of Organic Bananas']	0.0089	0.38202247191
['Large Lemon', 'Organic Avocado']	['Banana']	0.0111	0.306306306306
['Bag of Organic Bananas', 'Organic Raspberries']	['Organic Strawberries']	0.0156	0.391025641026
['Organic Strawberries', 'Organic Raspberries']	['Bag of Organic Bananas']	0.0121	0.504132231405
['Bag of Organic Bananas', 'Organic Hass Avocado']	['Organic Strawberries']	0.0192	0.317708333333
['Organic Hass Avocado', 'Organic Strawberries']	['Bag of Organic Bananas']	0.0105	0.580952380952
['Bag of Organic Bananas', 'Large Lemon']	['Organic Strawberries']	0.0095	0.315789473684
['Large Lemon', 'Organic Strawberries']	['Bag of Organic Bananas']	0.0079	0.379746835443

Figura 8: Tabela com regras de associação dois pra um, com suporte mínimo de 0.003 (0.3%) e confiança mínima de 0.25 (25%). Resultados obtidos a partir da execução do algoritmo Apriori e FP-Growth.

5 Considerações finais

Conforme o objetivo inicial deste projeto, o estudo da base de dados do Instacart gerou resultados que podem ser de interesse para, por exemplo, uma futura estratégia de marketing da empresa.

O agrupamento dos clientes baseado nas três características já citadas: horário médio das compras feitas, recorrência média de compra e número de pedidos mostrou que clientes que possuem mais compras tendem a fazer compras menores e mais frequentes que clientes com menos compras, independente do tempo de cadastro dele.

Também foi notado que clientes que possuem menor número de pedidos tendem a possuir um horário de compras mais disperso em comparação com clientes que possuem um número maior de pedidos.

A respeito da associação de produtos, foi possível concluir que muitas das compras foram associadas com bananas. A instacart poderia se aproveitar e criar uma estratégia de marketing voltada ao dito produto.

Outro ponto que poderia gerar resultados interessantes na fase de associação seria ter aglomerado os produtos por departamento (Frutas, Bebidas, Laticínios, etc). Assim, encontraríamos as possíveis relações entre as classes de produtos, e não entre os produtos em si.