**PRÉ-PROCESSAMENTO**

É importante ressaltar que tenho mais experiência em python, logo fiz o trabalho em python.

Primeiramente no pré-processamento, ao abrir o arquivo notei 3 problemas principais. O visual studio me comentou do problema de sintaxe da falta de uma vírgula, eu vi que existiam caracteres especiais ao invés de letras padrão, além de o valor de compra sempre ser “999”.

Então pessoalmente adicionei uma vírgula no lugar faltante no arquivo JSON.

Depois criei um script “arruma_valor_compra” que ordena a compra de 1 até o fim do arquivo JSON.

E daí criei um script “arruma_caracteres_especiais” que removia os caracteres especiais e deixava consistente a formatação das letras por todo o arquivo.

Com esse pré-processamento, acredito então que havia tornado os dados consistentes o suficientes para a mineração de dados em si.

**RESPOSTAS**

**(utilizando apriori)**

Temos uma relação bem forte com queijo mussarela e os produtos da padaria de 1 a 1, tanto panificados ou não (considerando um threshold baixo de 0.1 e confidence de 0.3):

| | antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|---|
| 21 | (Queijo Mussarela) | (Pastel Frango) | 0.03252 | 0.102564 | 0.742081 |
| 55 | (Queijo Mussarela) | (Pão Gajeta) | 0.03252 | 0.102564 | 0.901099 |
| 33 | (Queijo Mussarela) | (Pastel Queijo) | 0.03252 | 0.102564 | 0.573427 |
| 8 | (Queijo Mussarela) | (Café Melita) | 0.03252 | 0.102564 | 0.742081 |
| 15 | (Queijo Mussarela) | (Doce Goiabada) | 0.04065 | 0.128205 | 0.876068 |

Indicando que quem compra queijo mussarela, quase sempre estará comprando alguma outra coisa para acompanhar o queijo mussarela, como “Doce Goiabada” com um lift de 0.876068 e “Pão Gajeta” com lift de 0.901099, valores de lift bem altos, por mais que a confidence seja baixa. Aumentando o threshold de 0.1 para 0.3 podemos ver uma mudança:

| | antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|---|
| 0 | (Café Iguaçú) | (Refri - Coca Cola) | 0.032520 | 0.307692 | 2.226244 |
| 1 | (Doce Leite) | (Café Melita) | 0.032520 | 0.307692 | 2.226244 |
| 4 | (Refri - Pepsi) | (Pastel Presunto e Queijo) | 0.040650 | 0.312500 | 7.687500 |
| 8 | (Presunto Sadia) | (Pão Francês) | 0.056911 | 0.350000 | 1.537500 |
| 2 | (Café Melita) | (Pastel Queijo) | 0.048780 | 0.352941 | 1.973262 |

Isso indica por mais que o queijo tenha um lift alto e suporte ok, a sua confidence é muito baixa em vários produtos, demonstrando que muitas pessoas que compram queijo mussarela compram algo a mais, mas raramente compram a mesma coisa.

Além disso, com essa confidence de 0.3 e threshold de 0.3, podemos ver uma relação mais forte nos produtos, um exemplo de alto suporte sendo o presunto sadia com pão francês, que indica que quem compra presunto, quer um pão para colocar o presunto dentro, de fato.

O produto mais forte porém considerando esse script:

```
regras_1a1_fortes = regras_1a1[regras_1a1['confidence'] > 0.5]
produtos_mais_influentes = regras_1a1_fortes['antecedents'].apply(lambda x: list(x)[0]).value_counts()
produto_mais_influente = produtos_mais_influentes.idxmax()
```

Que verifica os produtos de maior confidence, conta quantos antecedentes eles têm e depois pega o produto de maior confidence e quantidade de antecedentes, acaba por ser o “Pastel Presunto e Queijo”, como em sua aparição:

| antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|
| (Pastel Presunto e Queijo) | (Refri - Pepsi) | 0.040650 | 1.000000 | 7.687500 |

Que mostra uma relação extremamente forte com pastel de presunto e queijo com um refri pepsi, onde 100% das vezes que o pastel é comprado, o refri pepsi também é comprado com confidence de 1. É interessante se ter em mente também o suporte, que por mais que não muito alto, ainda indica um produto bem influente em poucos casos. Porém, se modificar a confidence para 0.3, vemos que um produto mais universalmente importante é “Presunto Perdigão”, como no caso de:

| antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|
| (Presunto Perdigão) | (Pão Francês) | 0.065041 | 0.500000 | 2.196429 |
| (Presunto Perdigão) | (Queijo Mussarela) | 0.056911 | 0.437500 | 1.379808 |

Isso mostra também uma relação de confidence altos, mas com suporte maior e mais abrangente. Além disso, sendo uma associação de regra bem intuitiva, visto que quem compra presunto não só costuma querer o pão, mas algo para acompanhar o presunto neste pão.

No que diz respeito às regras dos doces, usando uma confidence de 0.2 e uma threshold de 0.3, vemos o seguinte:

| antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|
| (Doce Leite) | (Café Melita) | 0.03252 | 0.307692 | 2.226244 |
| (Doce Goiabada, Café Melita) | (Queijo Mussarela) | 0.02439 | 0.750000 | 2.365385 |
| (Doce Goiabada, Queijo Mussarela) | (Café Melita) | 0.02439 | 0.600000 | 4.341176 |
| (Café Melita, Queijo Mussarela) | (Doce Goiabada) | 0.02439 | 0.750000 | 5.125000 |

Que mostra que “Café Melita” e “Doce Goiabada” parecem estar largamente associados à compra de doces, visto que todas as regras de maior confidence estão associadas à compra de algum doce mais o café. Outro ponto interessante é que se diminuirmos o threshold para 0.25, vemos que no grande esquema dos doces, o “Café

Melita” só é importante por causa do “Doce Goiabada”:

| antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|
| (Doce Leite) | (Café Melita) | 0.03252 | 0.307692 | 2.226244 |
| (Doce Amendoim) | (Presunto Sadia) | 0.03252 | 0.266667 | 1.640000 |
| (Doce Amendoim) | (Pão Francês) | 0.03252 | 0.266667 | 1.171429 |
| (Doce Goiabada) | (Queijo Mussarela) | 0.04065 | 0.277778 | 0.876068 |
| (Doce Goiabada) | (Refri - Fanta) | 0.04065 | 0.277778 | 2.009804 |
| (Refri - Fanta) | (Doce Goiabada) | 0.04065 | 0.294118 | 2.009804 |
| (Café Melita, Queijo Mussarela) | (Doce Goiabada) | 0.02439 | 0.750000 | 5.125000 |
| (Café Melita, Doce Goiabada) | (Queijo Mussarela) | 0.02439 | 0.750000 | 2.365385 |
| (Queijo Mussarela, Doce Goiabada) | (Café Melita) | 0.02439 | 0.600000 | 4.341176 |

Vemos que os doces tem mais regras de associação como o esperado, contudo, um ponto a salientar, é que as regras dos doces todas tem um suporte baixo, indicando que não fazem grande parte de todas as compras dessa padaria, porém contém algumas das regras de confidence mais altas da padaria, como nas 3 últimas linhas do dataframe.

Isso pode mostrar de maneira errônea que essas três regras são muito importantes, porém, são 3 regras de associação praticamente repetidas entre elas mesmas, com suporte baixo em comparação com as outras regras.

Para acrescentar também, vemos que ”Doce Goiabada” aparenta ser um dos doces mais fortes, visto seu suporte alto em comparação aos outros doces e sua aparição em várias regras de associação. Principalmente na compra consequente de Fanta. Esse dataframe demonstra que o ”Doce Goiabada” está por trás das regras de maior confidence e nas associações de maior suporte.

Podemos ver também que a associação de “Queijo Mussarela” também se dá por conta do “Doce Goiabada”, que sempre aparece em suas associações.

**CONSIDERAÇÕES PESSOAIS**

Mineração de dados me parece ser algo de tentativa e erro (como no caso de achar support e confidence adequados), ao ponto que toda vez que eu visitava meu código para verificar novas regras com valores diferentes, eu sempre percebia que cada informação dependia de um contexto específico. Tudo parece ter um “sweet spot”, além de necessitar de grande força interpretativa, por que nem toda regra com confidence alta é relevante, como nem toda regra com support alto é relevante, tudo tem sua característica de análise única. Sinto que não explorei tudo que poderia nesse trabalho, mas acho também que isso virá com a prática.