Análise de comentários em E-commerces

Jailson Dias
Ramom Pereira

Base de Dados

- Kaggle
- Review em Inglês
- Base mais de 23K comentários
- Watson Natural Language Understanding

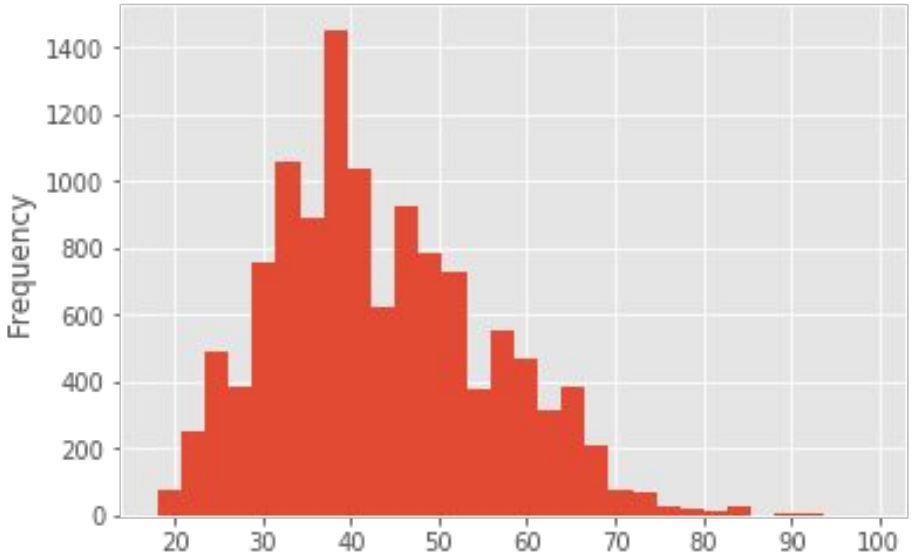
Especificação da Base

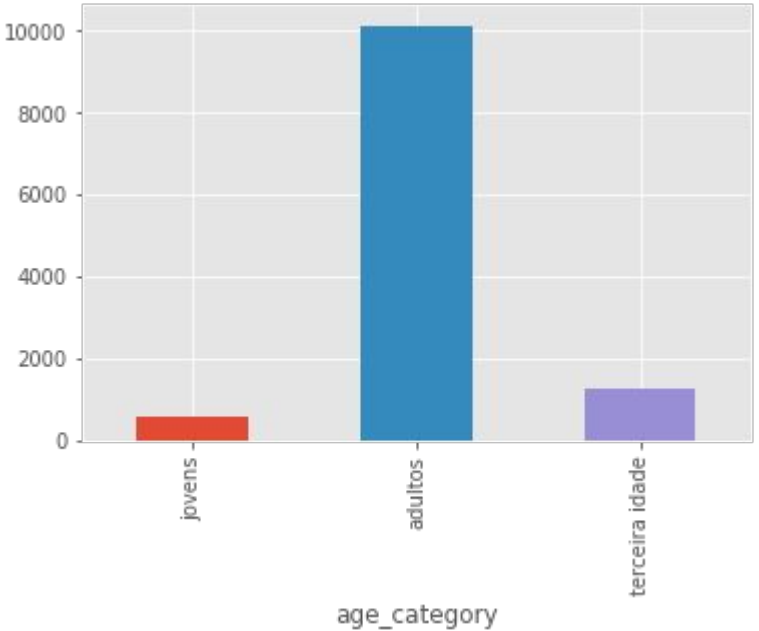
1. Clothing ID -> ID da roupa que foi avaliada pelo usuário
2. Age -> Idade do usuário que avaliou a roupa
3. Title -> Titulo do comentário do usuário
4. Review Text -> Comentário do usuário sobre o produto
5. Rating -> Nota que o usuário deu entre 1 e 5
6. Recommended IND -> Boleano que representa se o usuário recomenda ou não o produto
7. Positive Feedback Count -> Quantidade de usuários que acharam este comentário útil
8. Division Name -> Nome da divisão que o produto pertence
9. Department Name -> Nome do departamento que o produto pertence
10. Class Name -> Nome da classe que o produto pertence

Pré processamento dos dados

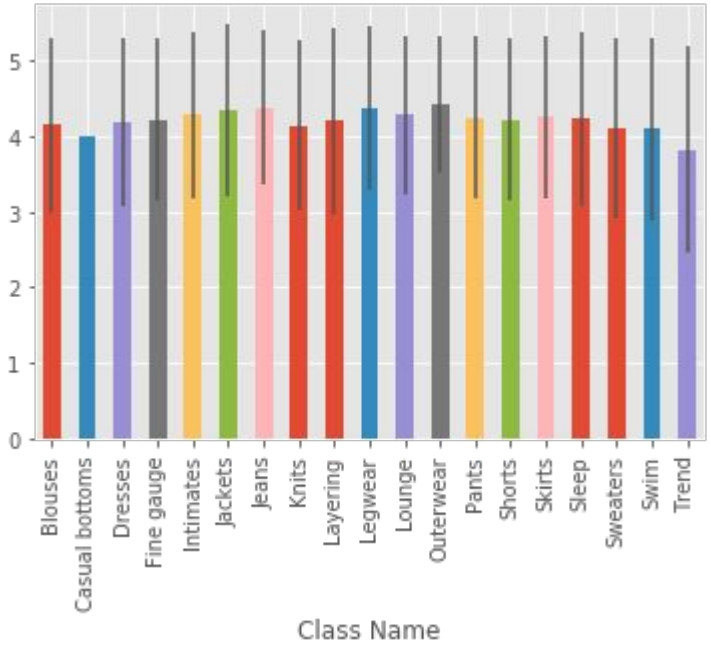
- Remover "Review Text" nulos → 845 comentários
- Treinar o Watson para a base de dados → 100 comentários
- Criar a feature de comentário falando de Loja ou produto
- Criar a feature de comentário falando positivamente, negativamente ou neutro
- Separar idade em categorias de jovem, adultos e terceira idade
- Separar Rating em muito baixo, razoável e bom
- Ajustando o tipo das features de string para categórico e booleano

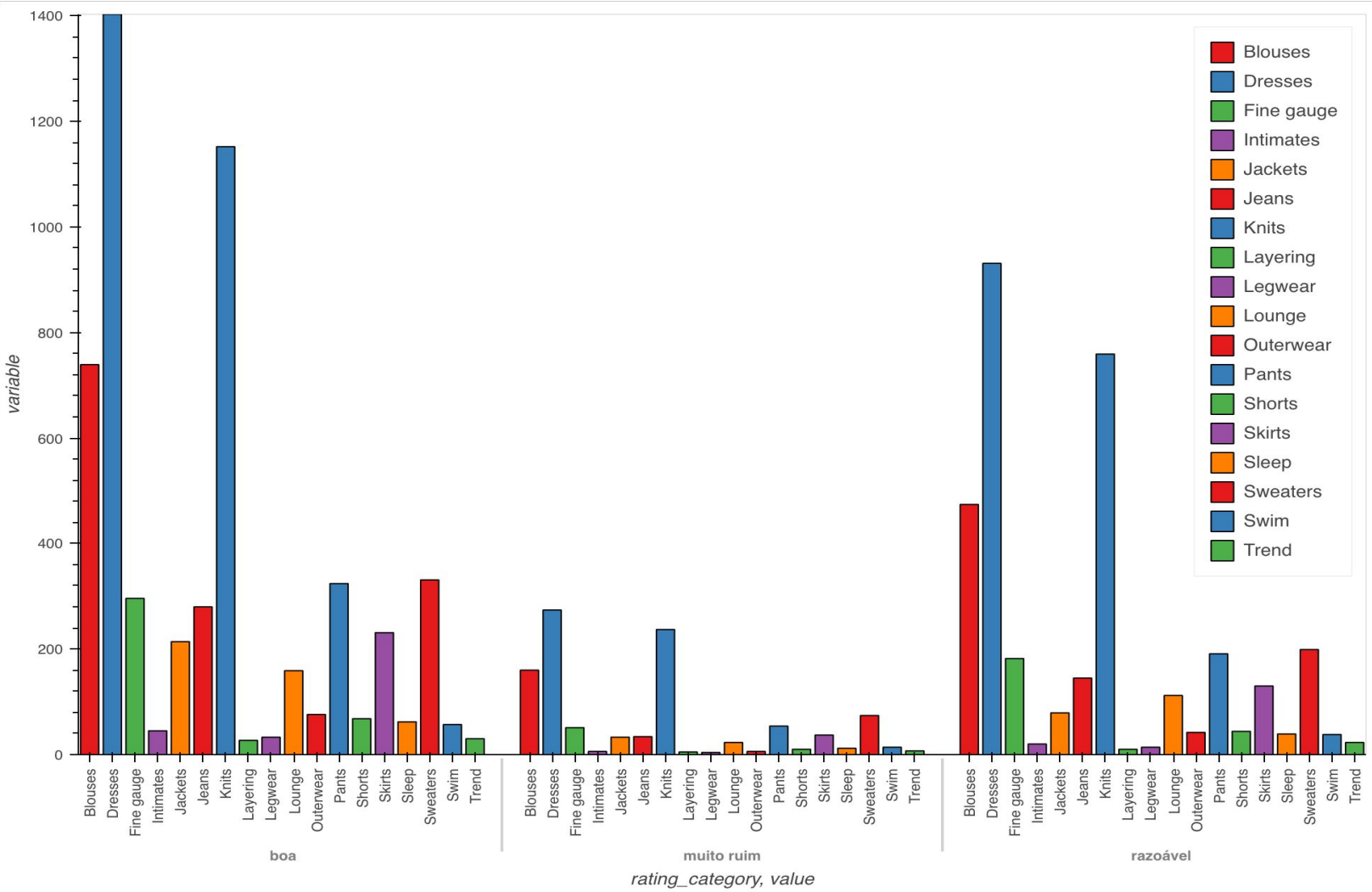
Visualização - Idade



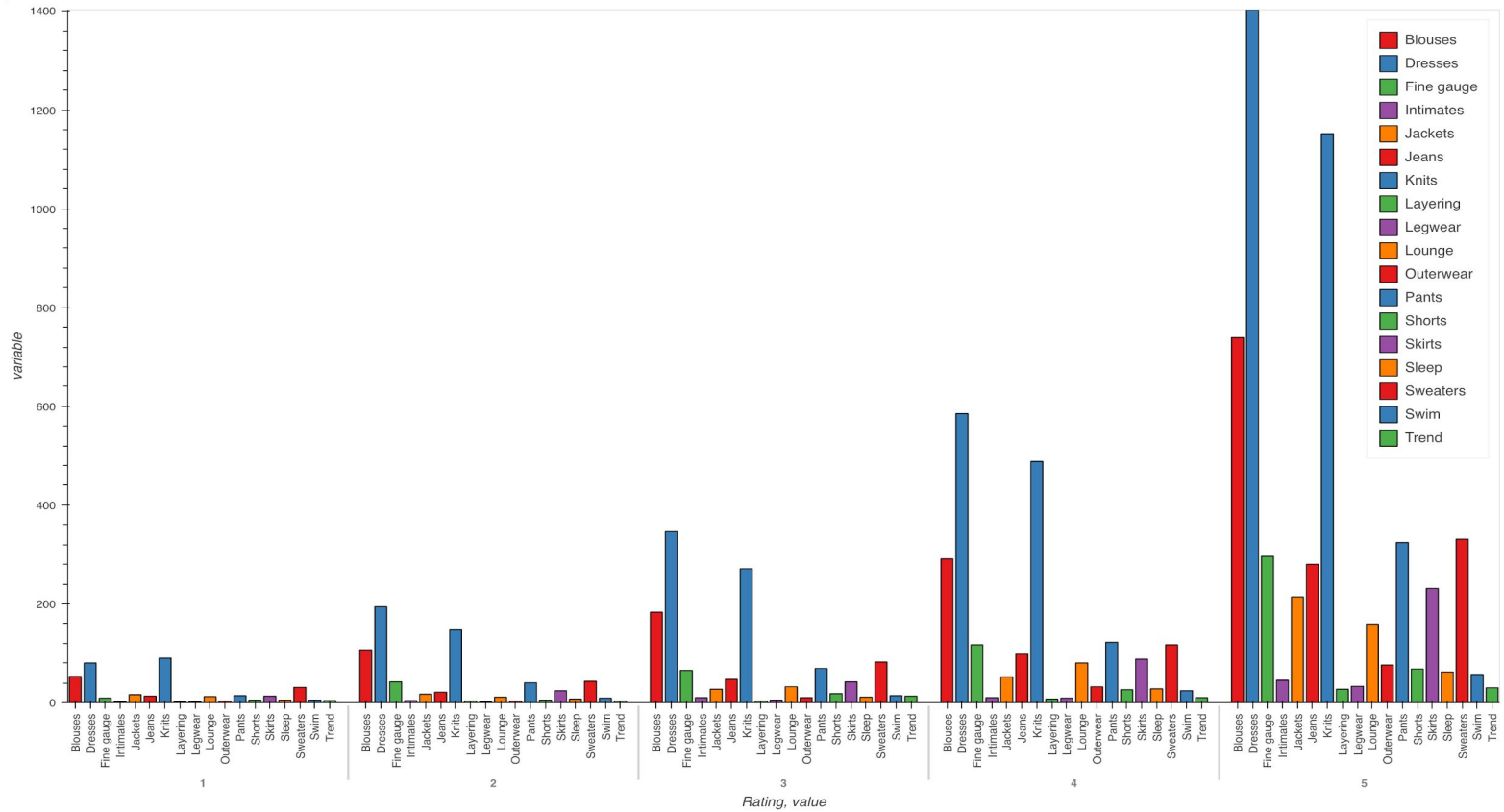


Visualização - Rating médio por tipo de roupa

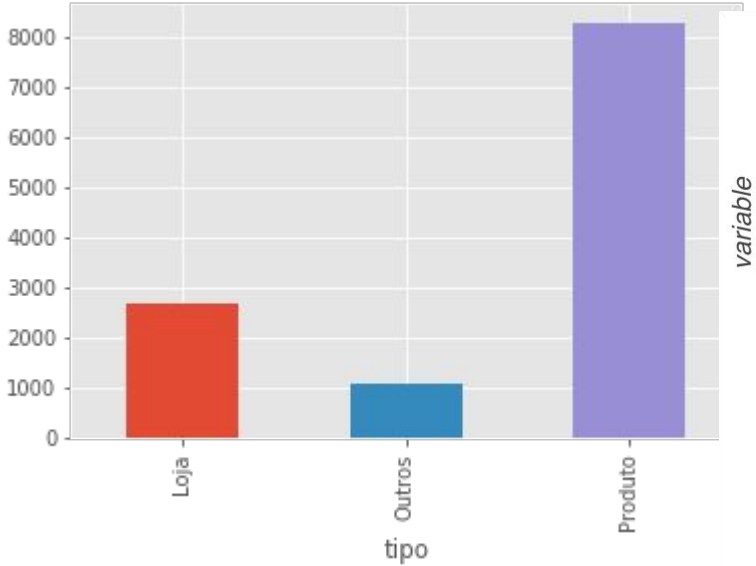


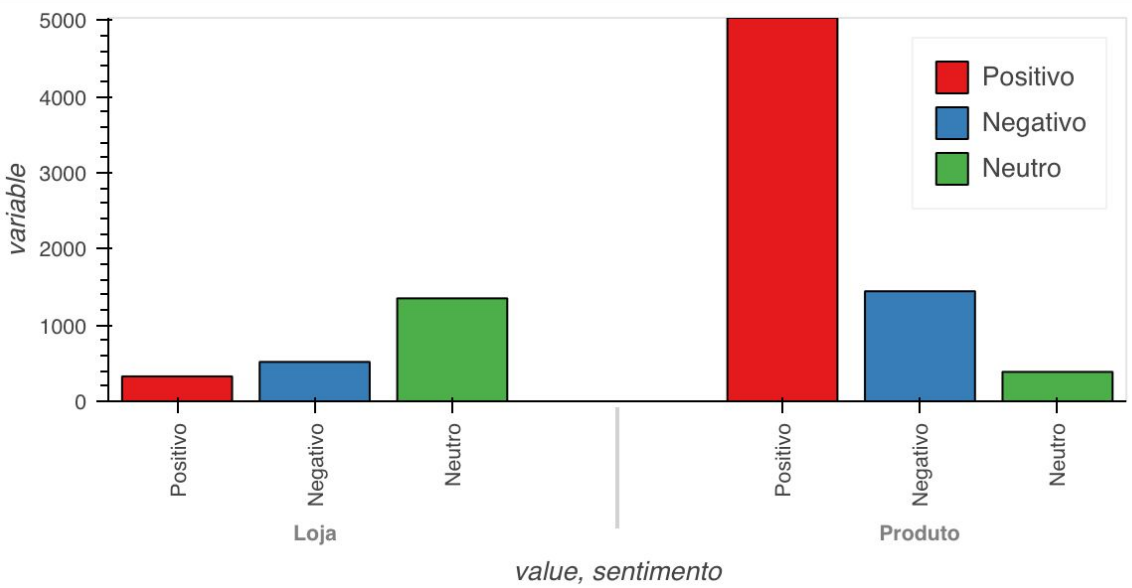

1400
1200
1000
800
600
400
200
0
variable
Blouses
Dresses
Fine gauge
Intimates
Jackets
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
Pants
Shorts
Skirts
Sleep
Sweaters
Swim
Trend
boa
muito ruim
razoável
rating_category, value

Visualização - Rating por tipo de roupa

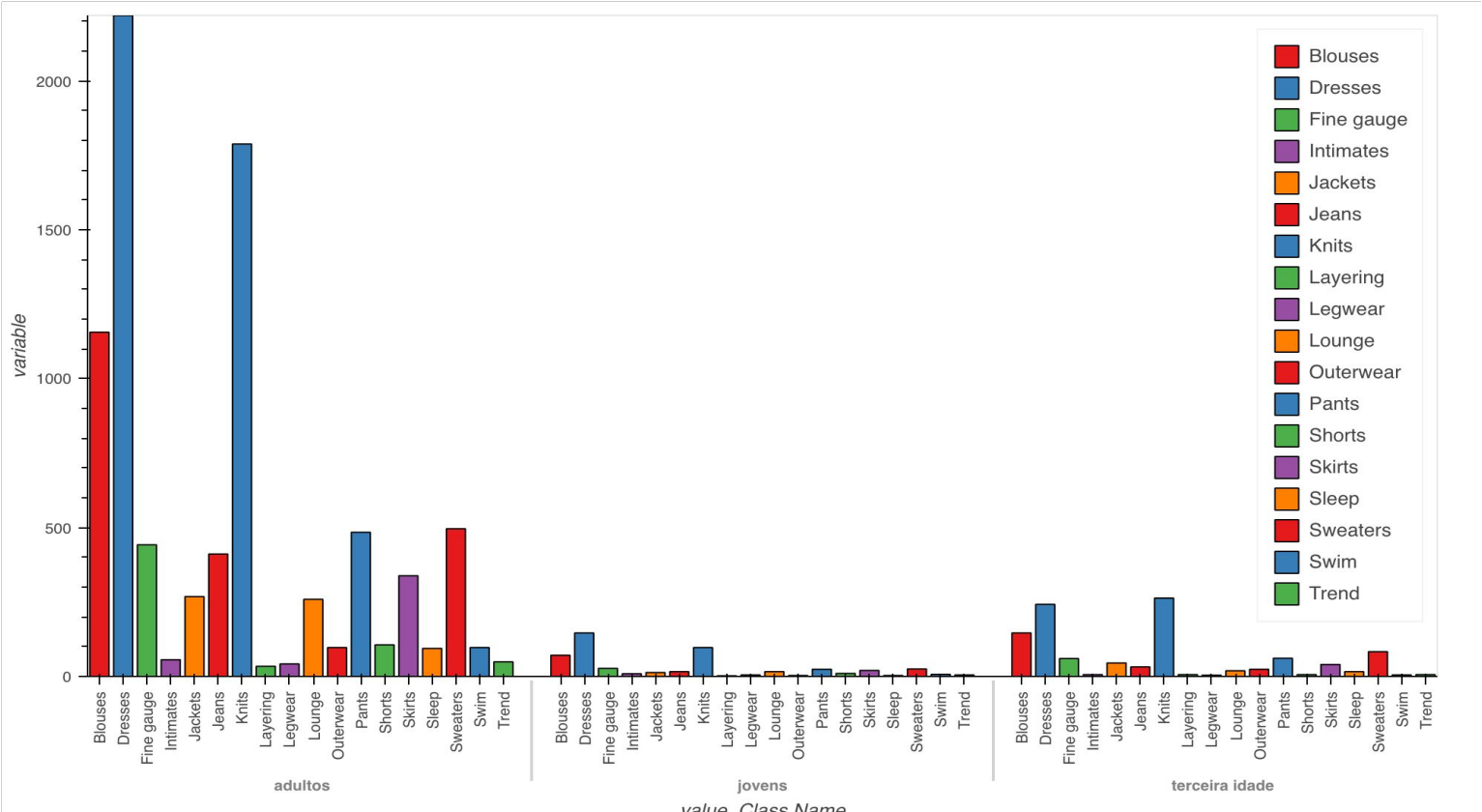


Visualização - Comentários

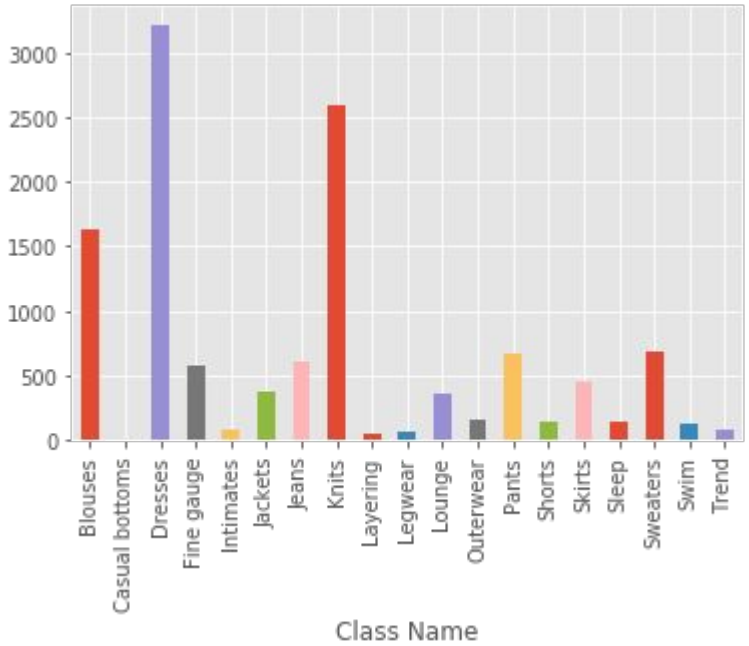




Visualização - Compras por faixa etária



Visualização - Compras por tipo de roupa

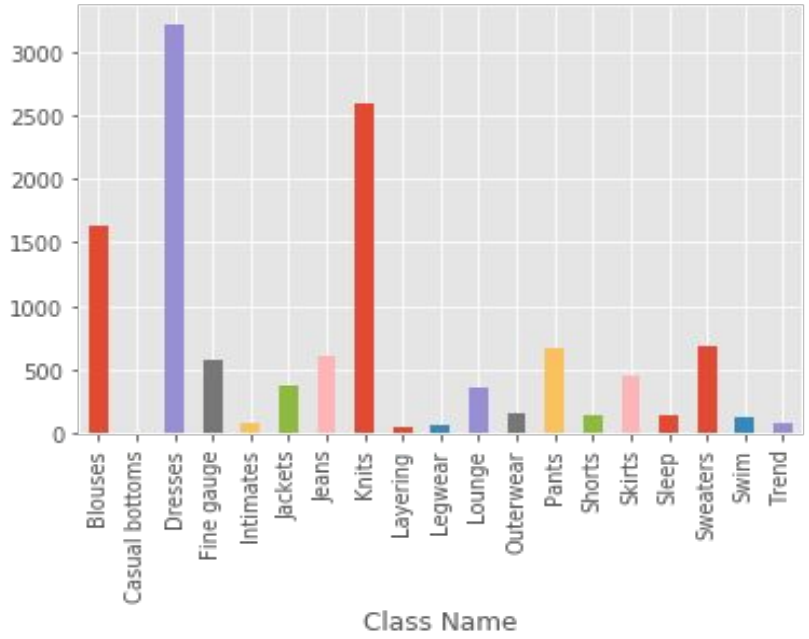


Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;
2. Pessoas jovens entre 18 e 25 anos preferem vestidos;
3. Produtos bem avaliados têm maiores quantidades de comentários

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;



Hipóteses

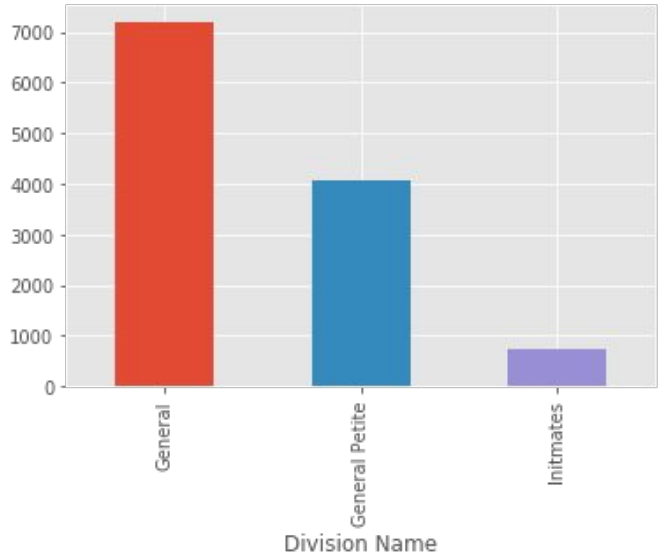
1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

- Naive Bayes: 12,39%
- KNN: 19,97%
- Árvore de decisão: 26,64%

Problema: muitas classes, 20 no total

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;



Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

   Considerando o Division Name que tem apenas 3 classes

- Naive Bayes: 60,81%
- KNN: 47,61%
- Árvore de decisão: 60,11%

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

```
Arvore de decisão
acc 0.6011111111111112
MSE 0.4030555555555556
matrix de confusão
 [[ 0.   98.1   1.9 ]
 [ 0.05 97.44  2.51]
 [ 0.08 97.34  2.58]]
```

```
Bayes
acc 0.6080555555555556
MSE 0.3919444444444443
matrix de confusão
 [[  0. 100.   0.]
 [  0. 100.   0.]
 [  0. 100.   0.]]
```

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

```
KNN
acc 0.4761111111111111
MSE 0.6655555555555556
matrix de confusão
 [[ 8.57 63.81 27.62]
 [ 9.5  58.89 31.61]
 [ 9.33 56.79 33.89]]
```

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

   Considerando o Division Name que tem apenas 3 classes
   Oversampling

- Naive Bayes: 32,94%
- KNN: 31,20%
- Árvore de decisão: 33,36%

Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

```
Arvore de decisão
acc 0.3336109954185756
MSE 1.046230737192836 3
matrix de confusão
 [[32.65 31.97 35.37]
 [30.69 32.92 36.4 ]
 [30.84 34.88 34.27]]
```

```
Bayes
acc 0.3294460641399417
MSE 1.0616409829237818
matrix de confusão
 [[30.61 15.65 53.74]
 [27.07 23.66 49.27]
 [28.64 21.66 49.69]]
```

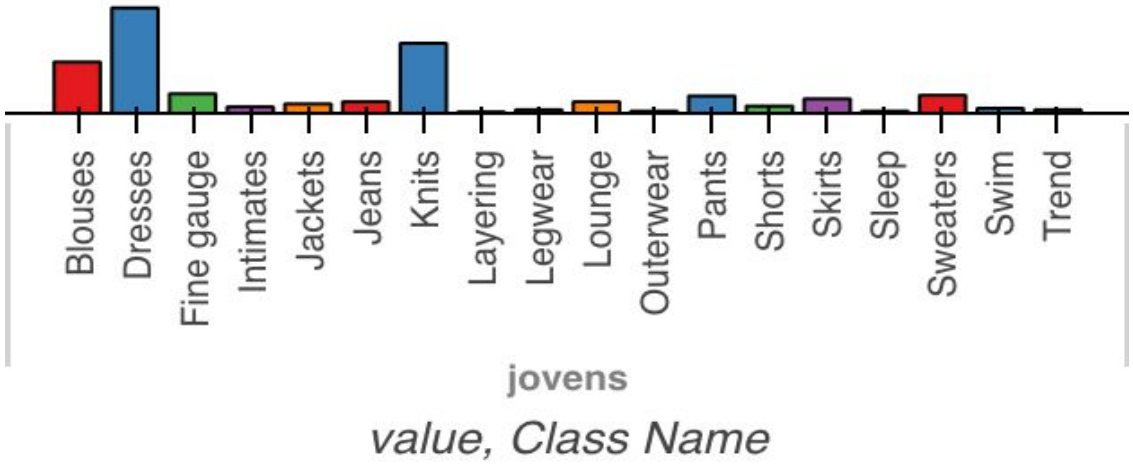
Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;

```
KNN
acc 0.3119533527696793
MSE 1.1778425655976676
matrix de confusão
 [[55.1  36.05  8.84]
 [46.76 37.72 15.52]
 [46.39 38.19 15.42]]
```

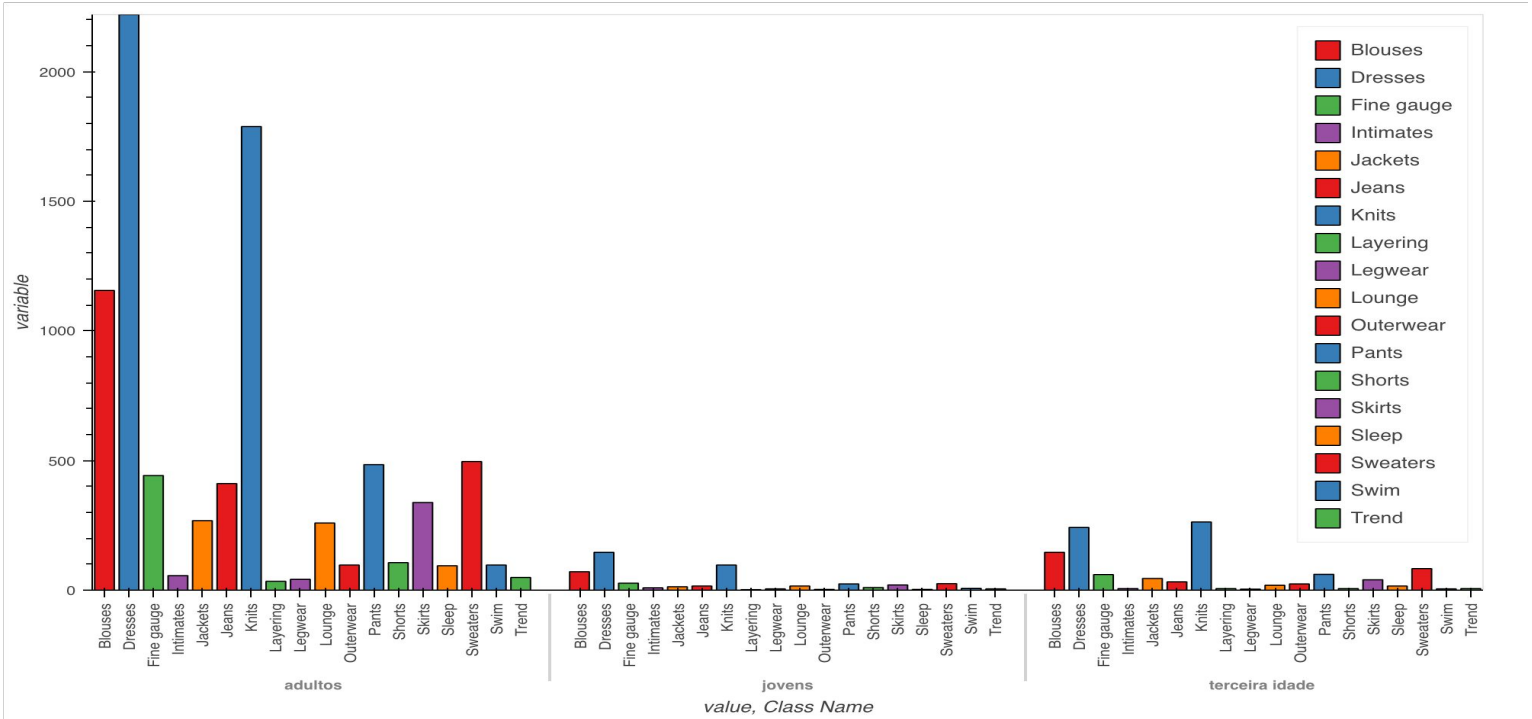
Hipóteses

2. Pessoas jovens entre 18 e 25 anos preferem vestidos;



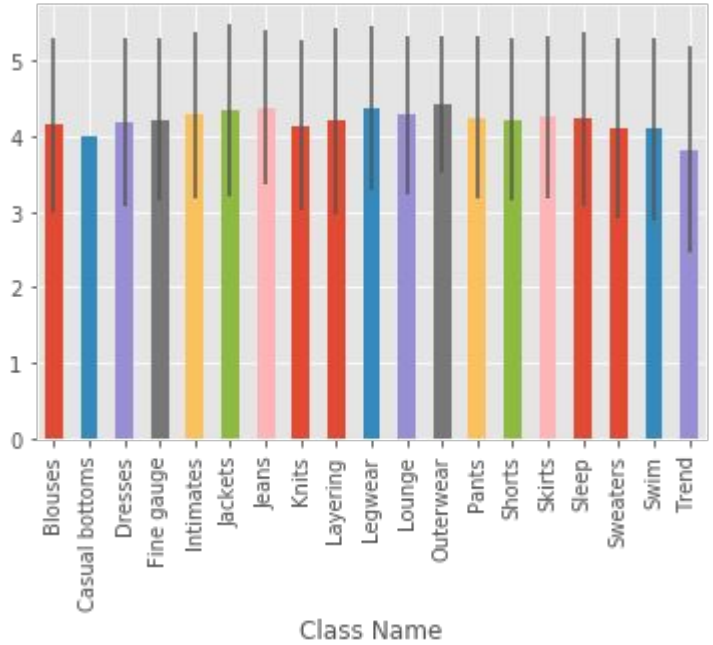
Hipóteses

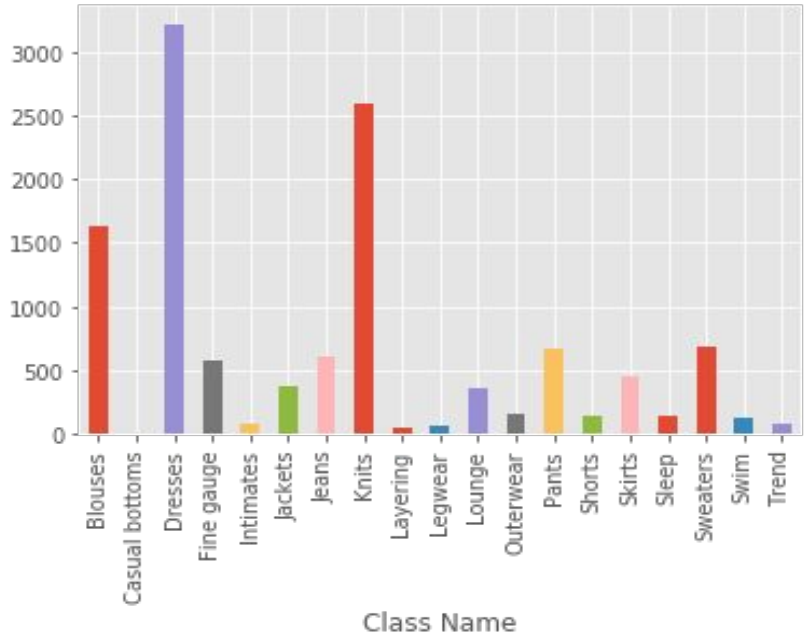
2. Pessoas jovens entre 18 e 25 anos preferem vestidos;



Hipóteses

3. Produtos bem avaliados têm maiores quantidades de comentários





Conclusão

- Comentários mais sobre os produtos
- As pessoas tendem a avaliar mais positivamente os produtos que as lojas
- As pessoas entre 30 e 45 anos compram mais roupas online que os mais jovens
- Pessoas da terceira idade também estão muito presentes nas compras online