Análise da relação entre preços e volume de avaliação em sites de comparação de preços

Carolina Bromfman, Livia Reis e Michel Wachslicht
Orientador: Admar Neto

1. Objetivo e Motivação

Tendo como ponto de partida o escopo de Marketing, esse projeto foi desenvolvido de forma a auxiliar empresas que possuam um ambiente de avaliação na web de seus produtos/serviço, a compreender melhor os consumidores. Assumindo como premissa que avaliações são um referencial da percepção do consumidor sobre um certo produto e que tal percepção esteja atrelada ao cumprimento de expectativas, tem-se que um preço elevado pode gerar uma maior percepção de valor que, caso não seja suprida, pode gerar insatisfação e mais avaliações negativas. Tal reflexão inicial é apenas uma hipótese do impacto do preço, desse modo, o objetivo do projeto é auxiliar empresas a entenderem de que forma a precificação pode influenciar as avaliações de seus produtos. Por outro lado, no que tange os consumidores, investiga-se possíveis vieses causados pelos preços dos produtos presentes nas avaliações em sites de review.

2. Revisão da Literatura

Segundo a análise econométrica de dados textuais apresentada no artigo *"Deriving the Pricing Power of Product Features by Mining Consumer Reviews"* (ARCHAK N.; GHOSE A.; IPEIROTIS P.G.; 2011), conclui-se que o efeito exercido pelo preço e pelo volume de avaliações sobre as vendas é contrário. Isto é, mais especificamente, através do uso de *Text Mining*, por meio de NLP, somada a metodologia Crowdsourcing, uma maneira semi-automatizada de usar a inteligência humana, ao invés de totalmente automatizada; os autores concluem, no artigo, que enquanto o preço tem relação negativa sobre as vendas, o volume tem relação positiva sobre elas.

No que diz respeito à influência social que permeia o ato da compra, conforme "*Arousal, valence, and volume: how the influence of online review characteristics differs with respect to utilitarian and hedonic products*" (REN, J.; NICKERSON, J.V.; 2019), há evidências de que o alto volume de avaliações aumenta a visibilidade do produto e pode o tornar mais socialmente desejável. Indo além, pode-se utilizar a teoria de Bens de Veblen para entender o fenômeno de enfoque aqui, a qual diz respeito aos produtos que possuem a procura proporcional ao seu preço elevado, uma vez que o valor percebido nesses produtos estão relacionados ao status que os mesmos proporcionam.

Tendo em vista que o conceito de Bens de Veblen se aplica para bens de luxo, optou-se por usar hotéis como objeto para as análises. Isso porque assumiu-se o pressuposto de que bens de luxo estão relacionados com características como requinte e qualidade, as quais estão presentes na hotelaria (BELLAICHE; MEI-POCHTLER; HANISCH, 2010).

3. Construção da Base de Dados

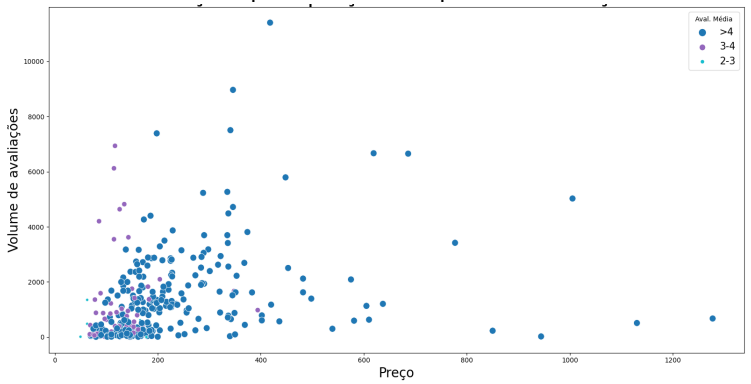
Através da biblioteca Selenium, presente na linguagem de programação Python, realizou-se um *Web Scraping* do site Google Hotels, do qual coletou-se os dados

de 379 hotéis do estado do Rio de Janeiro. Tal site tem como vantagem, além da ampla gama de hotéis e do fato de possuir uma página para cada hotel, o que auxilia na coleta das informações. Em contrapartida, o site de *reviews* contempla imóveis do Airbnb, os quais foram excluídos da coleta, por não possuírem avaliações escritas.
Os dados estáticos como nome dos hotéis, número de avaliações, quantidade de estrelas e preço foram recolhidos com base nas classes de cada item, presentes na linguagem de marcação utilizada na construção de sites (HTML). Para elementos não estáticos, como botões de troca de página e mudança de aba, utilizou-se o XPath. Além disso, também fez-se o uso da biblioteca unidecode para remover acentuação e outros símbolos utilizados pela língua brasileira, e presentes nas avaliações coletadas, os quais não são compreendidos pelo Python. Em suma, foram coletados os seguintes dados: nome do hotel, preço, quantidade de avaliações, avaliação média, avaliações escritas por consumidores e as respectivas notas dadas por estes.

4. Análises Descritivas

Enxergar os dados que foram coletados é de extrema importância para o desenvolvimento do projeto, tendo-se obtido os dados de interesse via *Web Scraping*, como descrito previamente. Isso porque uma análise descritiva ajuda a se ter uma visão geral dos dados quantitativos coletados, sendo esses o preço, volume de avaliações e avaliação média do hotel (1-5 estrelas).
Primeiramente, é necessário ressaltar que foram retirados dois *outliers* dessa análise, o Copacabana Palace e o Fasano. Os seus preços divergiam muito dos demais, custando mais de R$1700 reais, fato que impacta desde a dispersão dos dados até a linearidade dos mesmo. Ainda, ressalta-se que nenhuma conclusão é formulada a partir das análises descritivas a seguir.
Conforme mostra a Imagem 1 dos anexos, tanto o preço, quanto o volume possuem uma certa dispersão dos dados, justificado por alguns hotéis em que o preço é mais elevado que o restante. Tal fato indica baixa homogeneidade da amostra e, possivelmente, uma gama de características e padrões de comportamento diferentes para/com certos hotéis. A Imagem 2, histograma do preço, mostra essa grande dispersão dos preços.
Outra variável de extrema importância para a pesquisa é o volume de avaliações. Como a tabela (Imagem 1) demonstra, a variação é muito grande, o que pode dificultar a análise, uma vez que quando hotéis recebem poucas avaliações, a sua nota pode ficar viesada. Contudo, mesmo com a amostra pequena, apenas 10% dos hotéis possuem menos de 100 avaliações.
Para tentar, descritivamente, enxergar uma relação entre o preço e a quantidade de avaliações de cada hotel, foi feito o gráfico a seguir. A partir dele não se pode notar uma tendência, porém, percebe-se que o aumento da dispersão de pontos nos preços mais elevados, sendo isso um indício da hipótese de heterogeneidade de comportamento citada anteriormente.

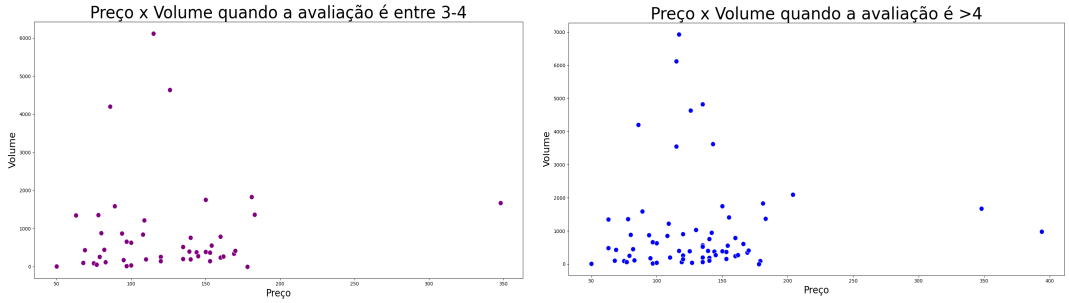
Imagem 4: Gráfico de dispersão entre preço e volume de avaliações



Fonte: Autoria Própria

Além do mais, o gráfico está colorido para a avaliação média, pois essa pesquisa também tenta enxergar se outros fatores como a avaliação do hotel podem impactar na relação preço e volume. Os dois gráficos a seguir mostram a relação do preço com o volume, porém agora separado por avaliação média. O primeiro quando a avaliação está entre 3 e 4, e o segundo, quando esta é maior do que 4. Há somente um hotel que possui avaliação menor do que 3. É possível, então, perceber que a avaliação média que o hotel recebeu não tem impacto aparente na relação do preço e volume de avaliações.

Imagem 5: Gráfico de dispersão



Fonte: Autoria Própria

Ademais, a fim de entender qual seria o modelo mais adequado para a regressão linear que seria feita em sequência, testou-se as quatro formas de modelagem - *lin-lin*, *lin-log*, *log-lin* e *log-log* - e dentre elas a que mais se aproximou de um modelo linear foi a *log-log* (Imagem 6). Vale ressaltar que uma hipótese para que as extremidades tenham assumido um formato mais exponencial é a concentração de hotéis com valores medianos ou *outliers* de preços elevados.
Para investigar mais a fundo a hipótese, limitou-se o logaritmo dos preços (eixo x do gráfico) para valores até 10^2,6, desse modo, foi possível se aproximar ainda mais de um modelo linear, o que pode indicar que a hipótese anteposta é razoável (Imagem 7). Vale ressaltar que o modelo log-log captura a elasticidade da variável resposta (volume), com a variável independente (preço) e, dado a horizontalidade da linha, pode indicar uma relação elástica, ou seja, o número de avaliações varia em resposta às mudanças de preço.

5. Metodologia

5.1. Análise das Avaliações

A fim de construir um modelo que abrangesse a percepção do consumidor, optou-se por realizar uma análise de sentimentos das avaliações escritas coletadas, por meio de técnicas de NLP, as quais permitem que uma máquina entenda a linguagem humana. Primeiramente, de modo a entender quais eram as principais palavras citadas nas avaliações escritas, utilizou-se o método *Wordcloud*.

Imagem 7: Nuvem de *Tags*

Fonte: Autoria Própria

Baseado nas palavras mais recorrentes nas avaliações, foram determinados os temas para classificação de tais. Por meio do método de classificação de texto *Zero Shot*, baseado em NLI, linguagem natural que determina a relação lógica entre dois textos, definiu-se a probabilidade de cada rótulo - comida, localização, preço, limpeza, infraestrutura, cômodo e atendimento - ser o assunto central tratado em cada avaliação. Nesse sentido, foi possível compreender qual é o principal fator presente na observação do consumidor sobre determinado hotel.

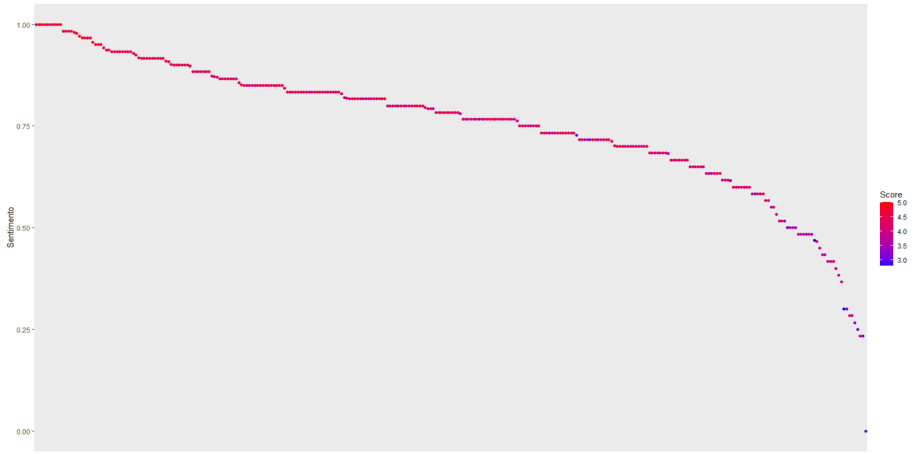
Apoiando-se na mesma lógica, através dos rótulos "positivo" e "negativo", também foi possível compreender se a percepção do consumidor, em sua avaliação, foi favorável ao hotel. Vale ressaltar que o método *Zero Shot* implica no uso de modelos pré prontos utilizando a biblioteca *Transformers*.

Depois de rodar o modelo, foi possível listar a importância das classes em cada avaliação. Assim, pode-se identificar quais classes apareceram mais vezes como mais importantes, como foi feito na Imagem 8 dos anexos. A partir dele, foi identificado que o atendimento é a classe com maior importância nas avaliações; enquanto a comida é a de menor importância. Vale ressaltar que tal resultado não era o esperado pelo grupo, o qual imaginava que preço seria a variável mais relevante, porém uma hipótese para isso não ter ocorrido é que o preço funciona como uma régua para a expectativa do cliente sobre o atendimento do hotel. Desse modo, quando o preço é alto mas o serviço não é de qualidade, o consumidor irá pontuar esse segundo aspecto em sua avaliação, ao invés do preço elevado.

Analogamente, pode-se utilizar o mesmo raciocínio para entender se as avaliações são majoritariamente positivas ou negativas. O resultado está

presente na Imagem 9, em que se identificou que a grande maioria das avaliações são entendidas como positivas.
Buscando entender se as avaliações escritas representam a nota com a qual o hotel é atribuído, o seguinte gráfico foi criado. Nele é visto que, apesar de tender à realidade, alguns hotéis possuem alta taxa de comentários positivos, mesmo que possua nota baixa.

Imagem 10: Porcentagem de avaliações positivas por hotel pelo nota do hotel



Fonte: Autoria Própria

Além disso, foi treinado um modelo de Random Forest para para que fosse estudado a importância das variáveis no impacto do volume de avaliações. Apesar de criar um modelo de previsão não ser um foco, é um bom método para se entender quais variáveis são mais relevantes. Segundo a Imagem 11, é interessante notar que a variável Nome é a mais importante, e isso se dá, possivelmente, pelo fato de cada hotel ter muitas características específicas que não estão sendo abrangidas pelas outras variáveis.

5.2. Regressão

Enfim, com o propósito de responder nossa questão inicial, optou-se pela regressão linear múltipla, uma vez que a variável resposta, volume de avaliações, é métrica. Dessa forma, a fim de cumprir com a análise inicial proposta, a principal variável independente é o preço em reais, sendo assim, busca-se interpretar o coeficiente da mesma (β_1).

Além das variáveis já citadas, de modo tanto a contemplar a percepção do consumidor, quanto a evitar a exclusão de variável relevante, adicionou-se as seguintes variáveis e interações:

- Número de estrelas;
- Avaliação majoritariamente positiva ou negativa? (*dummy*: assume 1 quando for, majoritariamente, positiva e 0 negativa);
- Comida é fator mais relevante nas avaliações? (*dummy*);
- Localização é fator mais relevante nas avaliações? (*dummy*);
- Preço é fator mais relevante nas avaliações? (*dummy*);
- Atendimento é fator mais relevante nas avaliações? (*dummy*);

- Limpeza é fator mais relevante nas avaliações? (*dummy*);
- Interação entre Preço e Relevância do preço;
- Interação entre Número de estrelas e Avaliação é positiva ou negativa.

Segue em sequência a construção da equação:

$$\ln(Volume) = \beta_0 + \beta_1 \ln(Preço) + \beta_2 Estrela + \beta_3 Percep + \beta_4 R_{Comida} + \beta_5 R_{Loc} + \beta_6 R_{Preço} + \beta_7 R_{Atend} + \beta_8 R_{Limpeza} + \beta_9 R_{Comodo} + \beta_{10} Preço * R_{Preço} + \beta_{11} Estrela * Percep$$

Vale ressaltar que as *dummys* foram criadas a partir da premissa que, para ser o fator mais relevante, ou seja, o que assumirá o número 1, ele precisa assumir a maior probabilidade na avaliação, dado que a somatória das probabilidades de todos os fatores para cada avaliação deve dar 100%. Os demais fatores que não assumirem a posição de maior relevância recebem o valor 0. Nessa lógica, a variável "Perceb" recebe o valor 1 quando a probabilidade da avaliação ser positiva é maior que 50%.

6. Limitações

O estudo possui limitações tanto no momento da coleta de dados, quanto na modelagem. Durante esse primeiro momento, a principal limitação a ser constatada tem relação com a data em que os preços foram coletados e as avaliações foram feitas. Foi necessário assumir como premissa preços estáticos, o que não ocorre na prática, porém tentou-se compensar tal incoerência com a variável de importância do preço na avaliação. Ainda no período de coleta das avaliações, enfrentou-se um problema relacionado à opção de "ler mais", desse modo, não foi possível coletar todas as avaliações por completo. Porém, a parte coletada foi suficiente para a execução das análises utilizando NLP. Também é possível citar o fato dos hotéis coletados serem apenas da região do Rio de Janeiro.
Em termos da modelagem, é relevante citar o fato do *Zero Shot* não ser completamente otimizado para a língua portuguesa, também havendo restrição de informações características de cada hotel.

7. Resultados e Conclusão

Após a realização da regressão, a qual pode ser observada na Imagem 12 dos anexos, constatou-se que todas as variáveis independentes, com exceção das interações, têm efeito positivo sobre o volume de avaliações. É relevante citar dois principais principais coeficientes, β_1 *e* β_3, ou seja, os coeficientes do preço e da percepção do consumidor sobre o hotel com base em sua avaliação.
O primeiro indica que uma variação de 1% no preço resultará em aumento de 1% no volume de avaliações, fato que é condizente com o pressuposto de elasticidade do modelo de regressão log-log. Já o segundo coeficiente indica que, quando a percepção passa de negativa (valor 0 da variável) para positiva (valor 1 da variável), a quantidade de avaliações aumenta em 200%. Tal efeito pode indicar que consumidores sentem-se mais motivados a avaliar quando sua experiência em certo hotel foi positiva.
Sendo assim, devido aos resultados encontrados, é recomendável que empresas do ramo de hotelaria não deixem de lado a variação no volume de avaliações dada

uma mudança de preço. Isso porque, como citado na revisão da literatura, quanto maior o número de avaliações, maior será a visibilidade do seu serviço e, assim, mais desejável ele será. Ademais, é interessante que essas empresas também incentivem a avaliação do estabelecimento, visto que a motivação para escrever um comentário é duplicada quando a experiência é positiva.
Em termos da teoria, uma constatação interessante é que mesmo que o preço tenha, diretamente, uma influência negativa sobre as vendas, uma vez que ele aumenta o volume de avaliação, ele tem um efeito indiretamente positivo sobre as vendas.

8. Referência Bibliográfica

ARCHAK N.; GHOSE A.; IPEIROTIS P.G. (2011) **Deriving The Pricing Power Of Product Features By Mining Consumer Reviews.** Management Sci.

REN, J.; NICKERSON, J.V. (2019) **Arousal, Valence, And Volume: How The Influence Of Online Review Characteristics Differs With Respect To Utilitarian And Hedonic Products.** European Journal of Information.

BELLAICHE, J. M.; MEI-POCHTLER, A; HANISCH, D. (2010). **The New World Of Luxury: Cough Between Growing Momentum And Lasting Change.** The Boston Consulting Group.

The Goods That Become More Desirable The More Expensive They Get. Jake Courage. Disponível em: https://blog.42courses.com/home/2018/7/10/the-goods-that-become-more-desirable-the-more-expensive-they-get

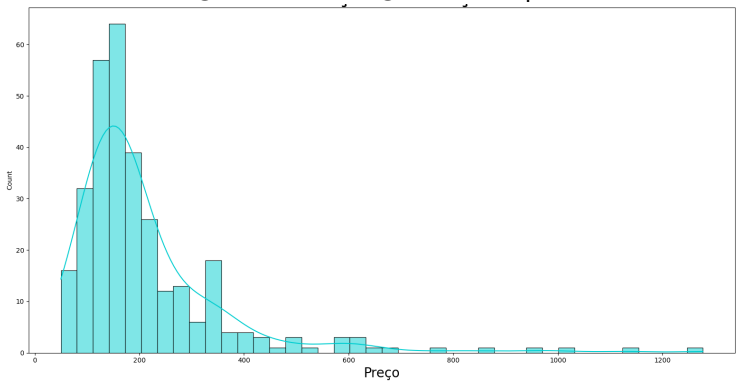
9. Anexos

Imagem 1: Tabela sumário

	Preço	Volume	Aval. média
Média	214	1335	4,2
Mediana	168	868	2,3
Max	1277	11406	5
Min	50	2	2,8
Q1 (25%)	130	243	4,1
Q3 (75%)	236	1828	4,5
Desvio Padrão	157,6	1587,8	0,39

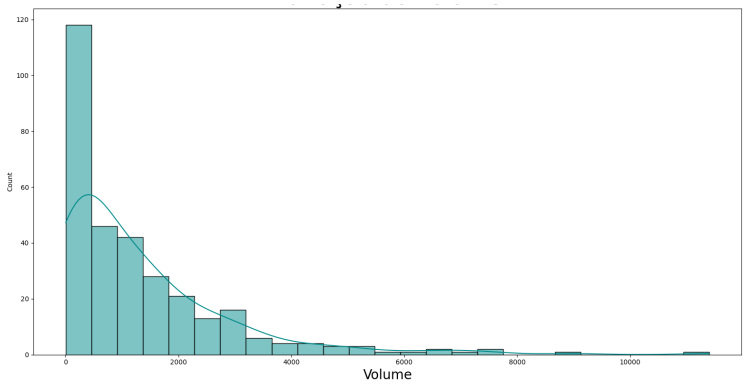
Fonte: Autoria Própria

Imagem 2: Histograma de preço



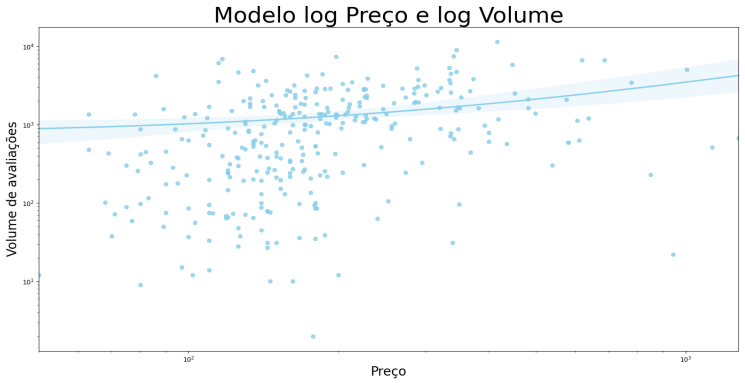
Fonte: Autoria Própria

Imagem 3: Histograma do volume



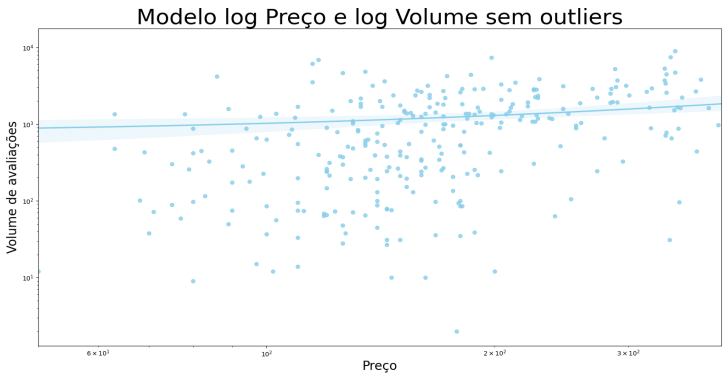
Fonte: Autoria Própria

Imagem 6: Modelo *log-log*



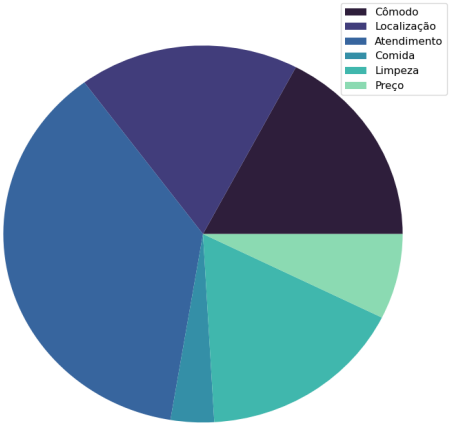
Fonte: Autoria Própria

Imagem 7: Modelo *log-log* com um corte dos valores maiores



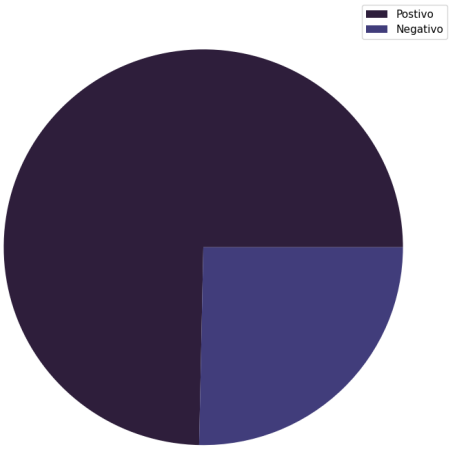
Fonte: Autoria Própria

Imagem 8: Frequência em que as classes apareceram como mais importantes



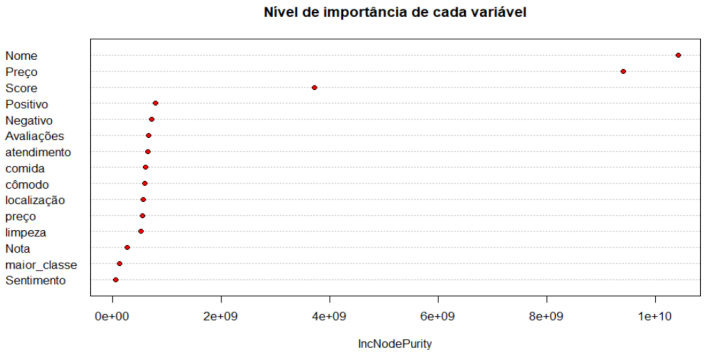
Fonte: Autoria Própria

Imagem 9: Frequência da polaridade das avaliações



Fonte: Autoria Própria

Imagem 11: Nível de importância de cada variável



Fonte: Autoria Própria

Imagem 12: Saída da regressão em R

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.0125 -0.8032  0.1956  0.8264  2.8548

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.8489269  0.1743874   4.868 1.14e-06 ***
Preço_log      1.0118025  0.0189605  53.364  < 2e-16 ***
Score          0.1570384  0.0428490   3.665 0.000248 ***
Percep         2.0390421  0.2106131   9.681  < 2e-16 ***
R_Comida       0.3156335  0.0548566   5.754 8.87e-09 ***
R_Loc          0.1755152  0.0279597   6.277 3.52e-10 ***
R_Preço        0.2285336  0.0578705   3.949 7.88e-05 ***
R_Atend        0.2755093  0.0277585   9.925  < 2e-16 ***
R_Limpeza      0.1426529  0.0317767   4.489 7.19e-06 ***
R_Comodo              NA         NA      NA       NA
R_Preço:Preço -0.0006304  0.0002034  -3.099 0.001946 **
Score:Percep  -0.5861745  0.0503267 -11.647  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.154 on 17477 degrees of freedom
  (60 observations deleted due to missingness)
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1923
F-statistic: 417.2 on 10 and 17477 DF,  p-value: < 2.2e-16
```

Fonte: Autoria Própria