オンラインネットショッププラットフォーム「Olist」の公開データセット分析

濱岡豊研究会18期　割谷　菜那子　栗林　瞭

目次

- 本分析の目的
- 二次データ
- 先行研究
- 分析の方針
- 単純集計、仮説設定
- 分析過程、結果
- 重回帰分析　過程、結果
- 考察
- 参考文献

本分析の目的

本分析の目的

- 「Olist」と呼ばれる、主に小売業者をターゲットとしたインターネット上におけるオンラインショップを開設することができるプラットフォームサービスが主にブラジルにて展開されている。

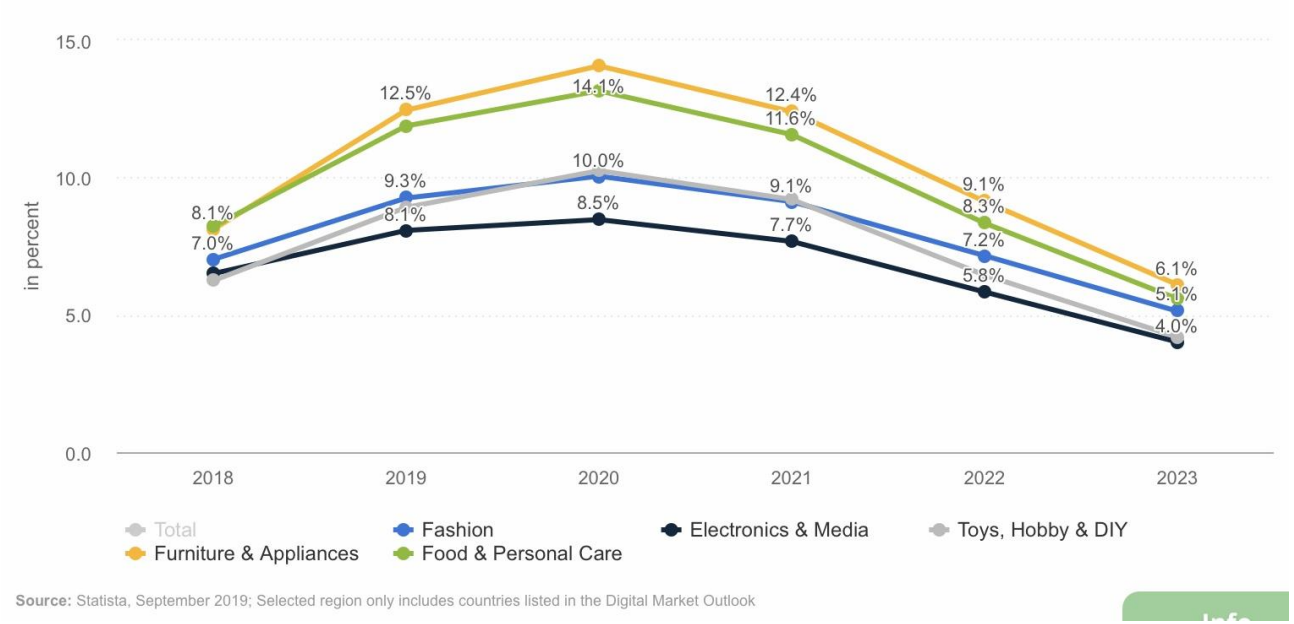
- そのOlistが自らのサービスを利用して商品を購買している顧客データを、「Kaggle」と呼ばれる企業や研究者がデータを投稿し各自で分析を行うことができるサービスに投稿していた。

本分析の目的

- 本分析はKaggleに公開されているOlistの顧客データにおける「olist_products_dataset.csv」および「olist_order_reviews_dataset.csv」というOlistで購買した製品の詳細な情報が表となっているデータに着目し、消費者が求める最も理想的な商品の理論値を分析、提言を行う。
- 具体的には、製品の評価、製品説明の長さ（あるいは製品名の長さ）がどれだけ売上に影響するかについて分析を行う。
- 分析する製品のジャンルは家具に限定する。

2次データ

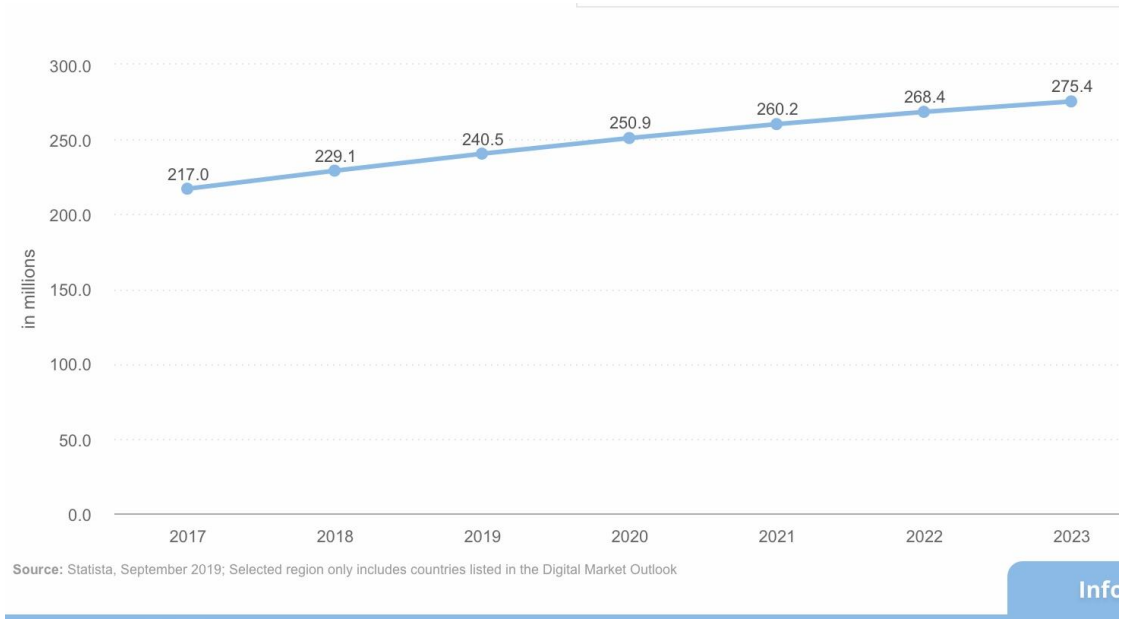
南アメリカにおけるeCommerceの収入成長率(2016)



- 予想も含め、山なりな形
- 現時点では「家具、電化製品」が一番成長率が高い

https://www.statista.com/outlook/243/103/ecommerce/south-america#market-revenue (2019年11月29日アクセス)

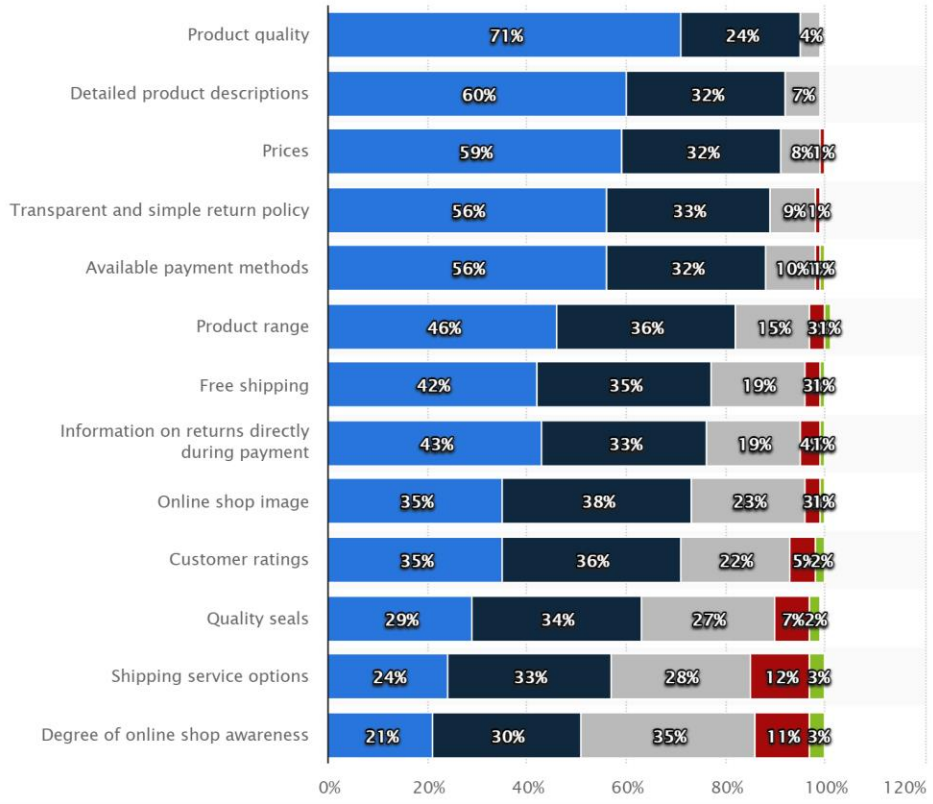
南アメリカにおけるeCommerceの利用者数(2019)



- 利用者は**増加傾向**にある（予想も含め）

https://www.statista.com/outlook/243/103/ecommerce/south-america#market-revenue (2019年11月29日アクセス)

ドイツの消費者が考えるオンラインサイトを選ぶ要因(2016)



- 製品の品質、詳しい製品情報、値段が消費者が求めている要因となった。
  - 青:とても重要
  - 紺:どちらかというと重要
  - 灰:どちらでもない
  - 赤:どちらかというと重要ではない
  - 緑:あまり重要ではない

→本分析では詳しい製品情報に着目して分析する。

https://www.statista.com/statistics/451736/criteria-for-the-selection-of-online-shops-in-germany/(2019年11月27日アクセス)

先行研究

他の消費者の製品評価、口コミについて

- 青木(2005)

eコマースにおけるインターネット上での通信販売は、商品が破損してしまう物理的リスクと、商品選択のミスによる金銭支出の無駄による経済的リスクをはらんでいる。

上記のリスクを消費者は緩和させるために消費者は他の消費者の評価を口コミという形で獲得する可能性がある。

→消費者が製品情報を獲得しようとする時、他の消費者の製品評価や口コミを参照すると考えられる。

製品評価、口コミを重要視する事例

- ヤマダ電機の家電口コミ＆比較レビューサイト「ピーチクパーク」

分析の方針

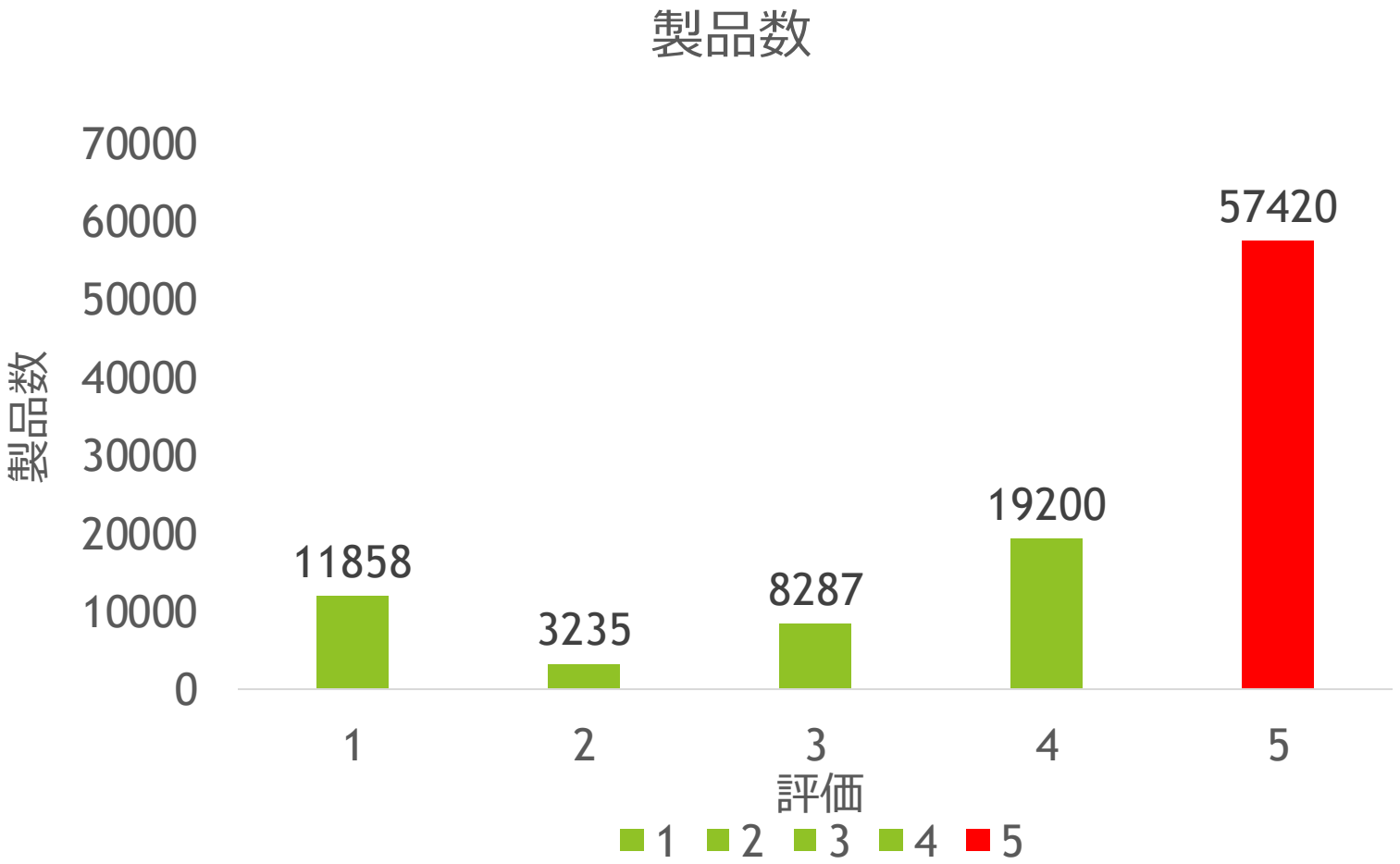
分析の方針

- 本分析で用いるデータセット「olist_products_dataset.csv」と「olist_order_reviews_dataset.csv」は購買された製品の様々な要素ごとのデータが記載されており、その中でも特に製品情報に関するデータに着目し、重回帰分析を行い、理論値を算出する。

- 算出されたデータを元に考察、提言を行う。

単純集計、仮説設定

製品の評価



Min. :1
1sr Qu. :2
Median :3
Mean :3
3rd Qu. :4
Max. :5

- 製品の評価として、5が最も多く、次点で4

→製品の評価が高ければ高いほど、商品の購買につながるのではないか？

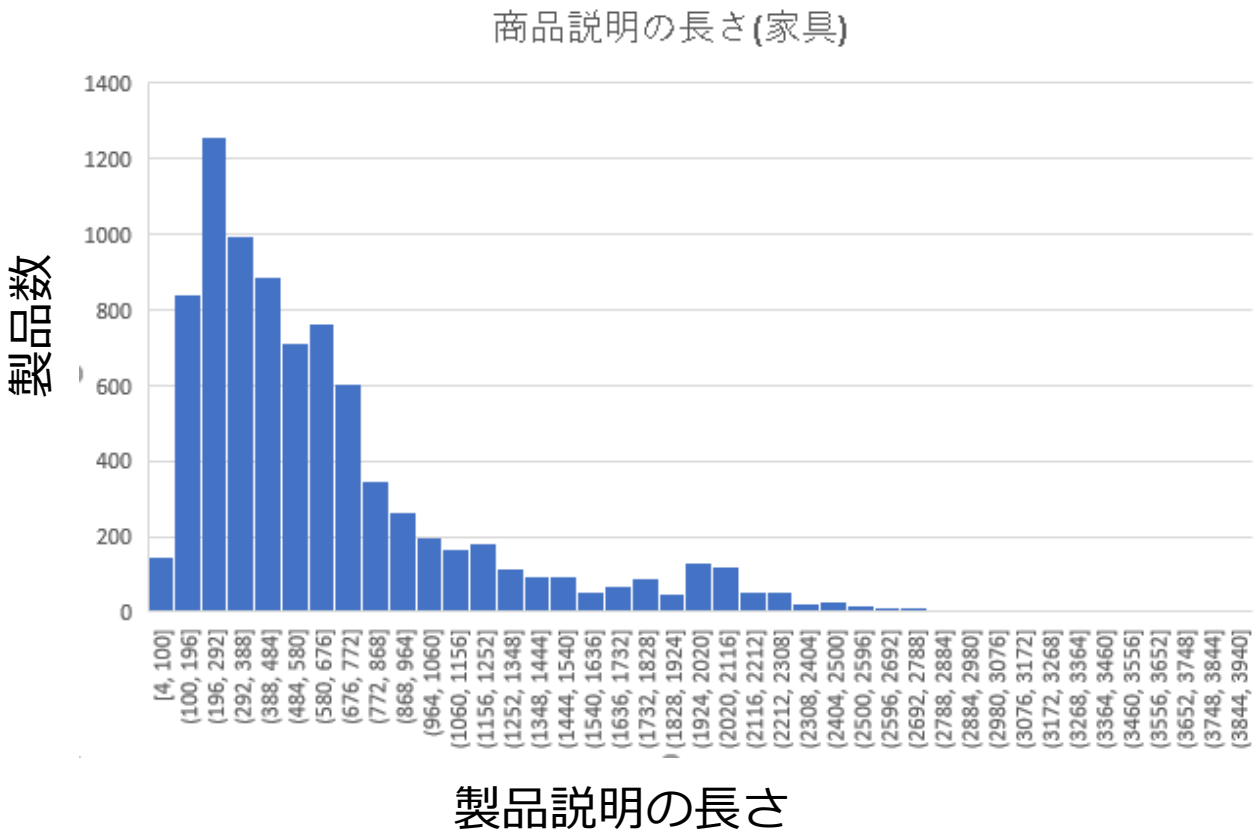
- 評価1の方が評価3よりも高い理由は何か？

製品カテゴリー

カテゴリー（英語）	製品数
Bed Table Bath	1411
Furniture décor	94
Housewares	116
Kitchen Dining Laundry Garden furniture	10
Office furniture	2657
Home comfort	3029
Furniture Mattress and Upholstery	104
Furniture living room	45
Furniture bedroom	309
Home comfort 2	111

- 製品カテゴリー数が多いので、分析対象にするカテゴリーを絞った
- 家具は製品情報量に左右されると考え、家具という分類に置かれそうな製品カテゴリーを抽出
- 家具であればサイズ、色、機能など商品そのものの属性が多いため、分析がしやすいと考えた。
- 製品の多様性が広すぎるとの指摘があったため、Office furnitureを分析対象とする。

商品説明の長さ



Min. :4.0
1st Qu. :726.8
Median :1503.5
Mean :1617.7
3rd Qu. :2322.2
Max. :3992.0
NA's :1

- 先行研究において消費者が製品情報を獲得する時、製品説明を読む行動から製品説明の長さは購買意欲につながるのではないか？

分析過程、結果

分析過程①

- ジャンルを「Office Furniture」に絞って分析を行うにあたって、本分析で用いるデータセット「olist_products_dataset.csv」と「olist_order_reviews_dataset.csv」には「products」のデータには製品をカテゴライズできるが、「order_reviews」のほうはできない問題が発生。

- 2つのデータセットの共通点はファイル内の「products_id」のみ。「product_category_name」は「products」にのみ記載されているため。

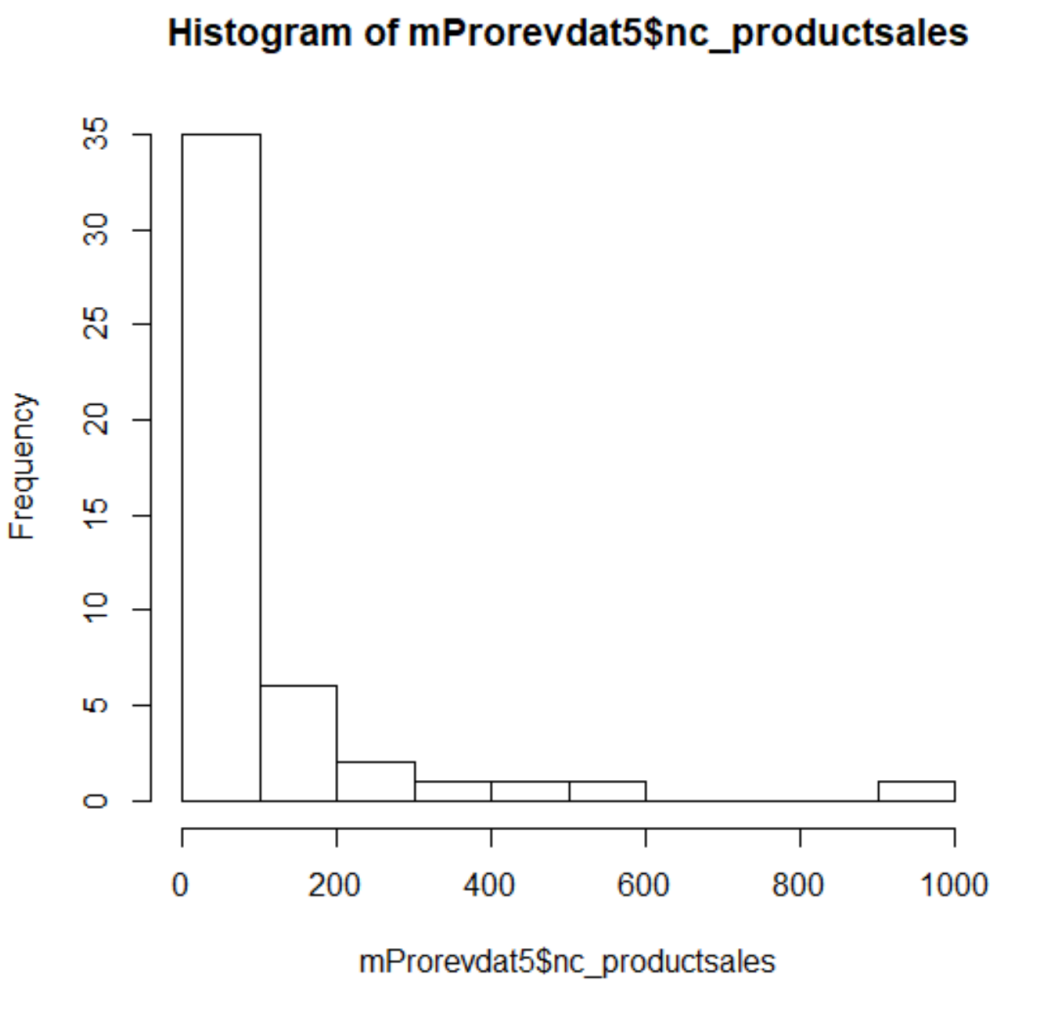
→分析を行う前に2つのデータセット内の「Office Furniture」のみのデータを出力しないと分析ができないことが判明。

分析過程②

- 解決策として、ExcelのVLOOKUP関数を用いてOffice Furnitureのproducts_idのみを検索、新しくデータセットを出力した。

- 出力後のデータセットの個数は4939個。被説明変数を累計販売個数、説明変数はproduct_description_length(商品説明の長さ)、review_score(製品レビューの評価)に設定し、重回帰分析を行った。

重回帰分析結果①



累計販売個数に偏りがあるため、log(累計販売個数)で分析を行う

重回帰分析結果②

- Prorevdat<-fread(file="bunsekidata.csv",sep=",",encoding="UTF-8",stringsAsFactors =F,data.table = FALSE)
- result2 <-lm(log(累計販売個数) ~ product_description_lenght+review_score, data = Prorevdat, model=TRUE, x=TRUE, y=TRUE, qr=TRUE, singular.ok=TRUE, contrasts=NULL, offset=NULL)
- summary(result2)

重回帰分析結果③

係数：	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.030978	0.082964	24.480	< 2e-16***
product_description_lenght	-0.00013	4.54E-05	-2.771	0.006***
review_score	0.027055	0.017931	1.509	0.131

有意水準：　0%=*** 0.1%=** 1%=* 5%=.

累計販売個数＝10^(2.030978-0.00013(product_description_length)+0.027055(review_score))

Residual standard error: 1.763 on 4935 degrees of freedom
Multiple R-squared: 0.002
Adjusted R-squared: 0.002
P-value: 0.007
AIC:19620.630
N:4939

重回帰分析　過程、再分析

分析過程①(データセットの作成)

- 前回のご指摘により、製品に対して評価をつける、という消費者行動と製品を購買する、という消費者行動は因果関係が逆であることが判明したため、別の説明変数を設定。

- 再分析に用いる説明変数は商品説明の長さ、価格、商品名の長さに設定。被説明変数は「Office Furniture」のみの累計販売個数とした。

- そして仮説を3点設定した。1点目は「H1:商品説明の長さが長いほど、累計販売個数は増加する」、2点目は「H2:製品の販売価格が安いほど累計販売個数は増加する」、3点目は「H3:製品名の長さが長いほど、累計販売個数は増加する」とした。

- 前回のデータセットでは製品レビューの有無に関わらず評価がつけられたもののみだったため、データセットについても変更を行った。

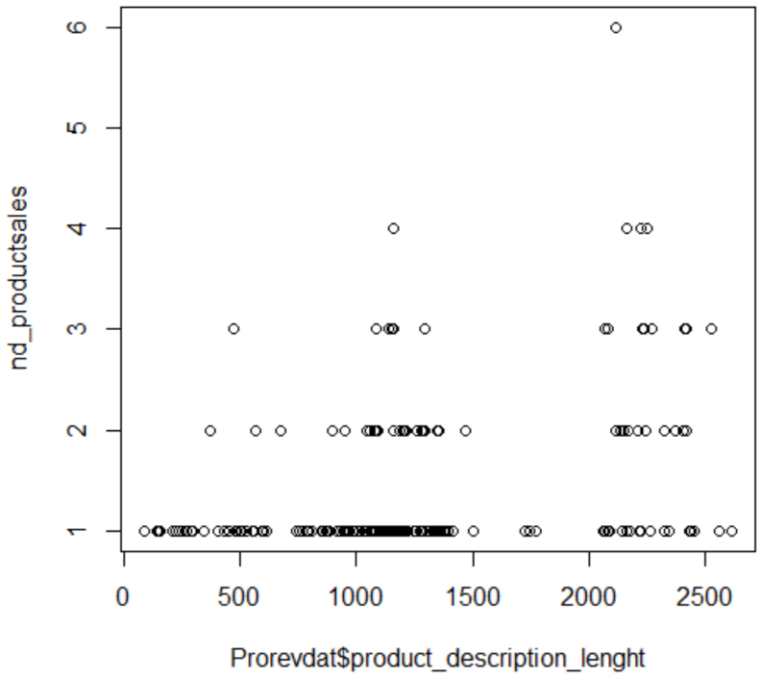
分析過程②(データセットの作成)

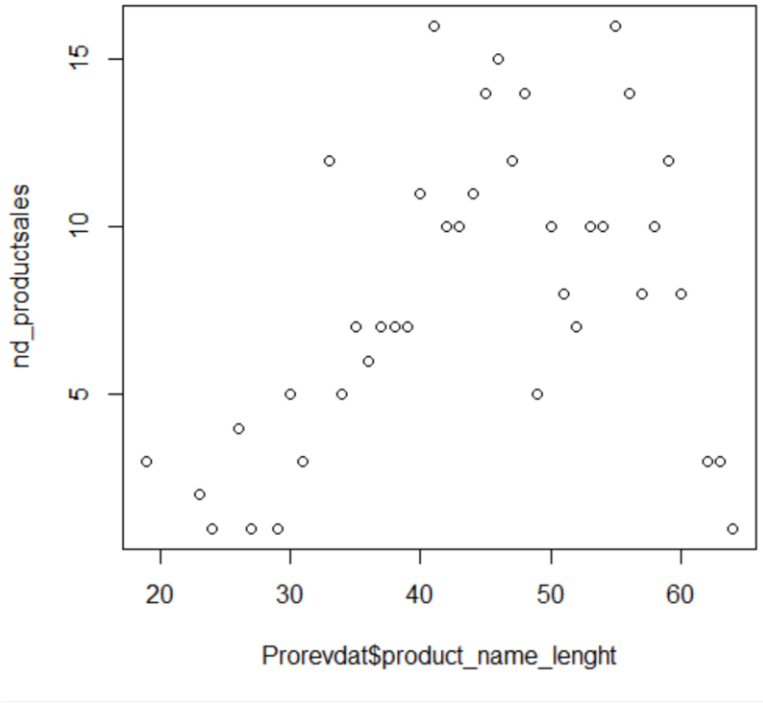
- 用いたデータセットは「olist_order_items_dataset.csv」と「olist_products_dataset.csv」を使用した。

- また「olist_order_items_dataset.csv」のデータから価格の要素を紐づける際に、一製品の中で複数の値段が設定されていた(セール、まとめ買いなどの割引？)ため、分析を行いやすくするために価格は対象製品ごとの最大値に固定した。

- 再分析においては価格を1つに統一しているため、前回の分析からの母数は大幅に減り出力後のデータセットの個数は309個となった。

変数分布①



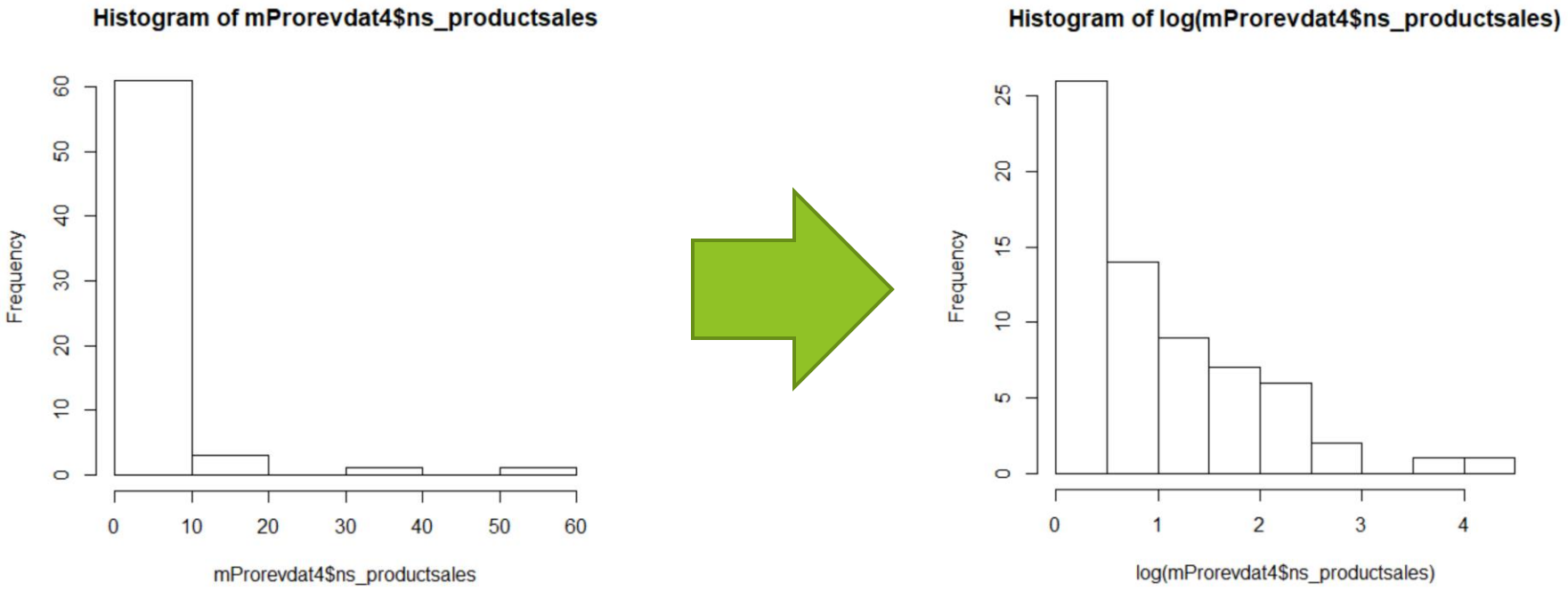


Price変数は、何故かRで「figure margins too large」と出てしまい、Rにてグラフを描くことができなかった。

変数分布②


price
製品数
製品価格(レアル)
0
100
200
300
400
500
600
700
800
900
1000
0
50
100
150
200
250
300
350

再分析結果①



被説明変数である累計販売個数に偏りがあるので、log(累計販売個数)を使用し、改善

再分析結果②

- 被説明変数を累計販売個数として、説明変数を以下の通りにした：
  - 製品説明の長さ
  - 価格
- なお、「製品名の長さ」も加え、さらなる追加分析も行った
- 製品説明の長さの影響は、逆U字型になっている可能性があるので、二乗した結果も載せる

再分析結果③ 製品説明の長さ、価格

- Prorevdat<-fread(file="Productsdata.csv",sep=",",encoding="UTF-8",stringsAsFactors =F,data.table = FALSE)
- Result2 <-lm(log(累計販売個数) ~ (product_description_lenght+price, data = Prorevdat, model=TRUE, x=TRUE, y=TRUE, qr=TRUE, singular.ok=TRUE, contrasts=NULL, offset=NULL)
- summary(result2)

再分析結果④　製品説明の長さ、価格

累計販売個数=10^((1.91E+00)+(8.76E-05)(製品説明の長さ)+(6.70E-05)(価格))

係数:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.91E+00	2.36E-01	8.081	1.5e-14 ***
product_description_lenght	8.76E-05	1.48E-04	0.594	0.553
price	6.70E-05	6.81E-04	0.098	0.922

有意水準：　0%=*** 0.1%=** 1%=* 5%=.

Residual standard error: 1.597 on 306 degrees of freedom
Multiple R-squared: 0.001
Adjusted R-squared: -0.005
p-value: 0.836
AIC:1171.142
N:309

再分析結果⑥ 製品説明の長さ、価格(二乗)

- Prorevdat<-fread(file="Productsdata.csv",sep=",",encoding="UTF-8",stringsAsFactors =F,data.table = FALSE)
- x <-((Prorevdat$product_description_lenght)^2)
- result2 <-lm(log(累計販売個数) ~ x+price, data = Prorevdat, model=TRUE, x=TRUE, y=TRUE, qr=TRUE, singular.ok=TRUE, contrasts=NULL, offset=NULL)
- summary(result2)

再分析結果⑦　製品説明の長さ、価格(二乗)

累計販売個数=10^((1.97E+00)+((2.60E-08)(製品説明の長さ))^2+((6.63E-05)(価格))^2)

係数:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.97E+00	1.65E-01	11.925	<2e-16 ***
price_description_lenght	2.60E-08	4.96E-08	0.526	0.600
price	6.63E-05	6.81E-04	0.097	0.922

有意水準：　0%=*** 0.1%=** 1%=* 5%=.

Residual standard error: 1.597 on 306 degrees of freedom
Multiple R-squared: 0.001
Adjusted R-squared: -0.006
p-value: 0.869
AIC:1171.219
N:309

再分析結果⑨　製品説明の長さ、価格、製品名の長さ

- result4 <-lm(log(累計販売個数) ~ (product_description_lenght)^2+price+(product_name_lenght)^2, data = Prorevdat, model=TRUE, x=TRUE, y=TRUE, qr=TRUE, singular.ok=TRUE, contrasts=NULL, offset=NULL)
- summary(result4)

再分析結果⑩　製品説明の長さ、価格、製品名の長さ

累計販売個数=10^((2.39E+00)+(6.91E-05)(製品説明の長さ)+(8.59E-05)(価格)-(1.00E-2)(製品名の長さ))

係数:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.39E+00	5.29E-01	4.523	8.73e-06 ***
product_description_lenght	6.91E-05	1.49E-04	0.465	0.642
price	8.59E-05	6.81E-04	0.126	0.900
product_name_lenght	-1.00E-02	9.78E-03	-1.027	0.305

有意水準：　0%=*** 0.1%=** 1%=* 5%=.

Residual standard error: 1.597 on 305 degrees of freedom
Multiple R-squared: 0.005
Adjusted R-squared: -0.005
p-value: 0.703
AIC:1172.076
N:309

再分析結果⑫　製品説明の長さ、価格、製品名の長さ(二乗)

- Prorevdat<-fread(file="Productsdata.csv",sep=",",encoding="UTF-8",stringsAsFactors =F,data.table = FALSE)
- x <-((Prorevdat$product_description_lenght)^2)
- y <-((Prorevdat$product_name_lenght)^2)
- result4 <-lm(log(累計販売個数) ~ x+y+price, data=Prorevdat, model=TRUE, x=TRUE, y=TRUE, qr=TRUE, singular.ok=TRUE, contrasts=NULL, offset=NULL)
- summary(result4)

再分析結果⑬　製品説明の長さ、価格、製品名の長さ(二乗)

累計販売個数=10^((2.14E+00)+((2.01E-08)(製品説明の長さ))^2-(7.41E-05)(価格)+((7.37E-05)(製品名の長さ))^2)

係数:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.14E+00	3.11E-01	6.898	3.05e-11 ***
product_description_lenght	2.01E-08	5.04E-08	0.398	0.691
product_name_lenght	-7.41E-05	1.11E-04	-0.671	0.503
price	7.37E-05	6.82E-04	0.108	0.914

有意水準：　0%=*** 0.1%=** 1%=* 5%=.
Residual standard error: 1.599 on 305 degrees of freedom
Multiple R-squared: 0.002
Adjusted R-squared: -0.007
p-value: 0.866
AIC:1172.764
N:309

モデルの適合度結果まとめ

分析	Adjusted R-squared	Multiple R-squared	AIC
製品説明の長さ、価格	-0.005	0.001	1171.142
製品説明の長さ、価格(二乗)	-0.006	0.001	1171.219
製品説明の長さ、価格、製品名の長さ	-0.005	0.005	1172.076
製品説明の長さ、価格、製品名の長さ(二乗)	-0.007	0.002	1172.764

再分析結果まとめ

- 二乗しない方が全体的にP値が低かったので、二乗しない以下のモデルを採択とする（有意でないものが多いが）：
  - 累計販売個数=10^((1.91E+00)+(8.76E-05)(製品説明の長さ)+(6.70E-05)(価格))
  - 累計販売個数=10^((2.39E+00)+(6.91E-05)(製品説明の長さ)+(8.59E-05)(価格)-(1.00E-2)(製品名の長さ))
- 回帰係数が低い理由としては、相関が低い、サンプルサイズが小さいなどが考えられる。

分析結果の考察、提言

考察① 商品説明の長さ

- H1:商品説明の長さが長いほど、累計販売個数は増加する→棄却

- 消費者がオンラインにおけるeコマースを用いた取引において、物理的リスクと経済的リスクを回避するために消費者は購買しようとする商品の情報を取得する行動をおこすと考え本仮説を設定したが、棄却となった。

- 商品説明がただ長いだけでは消費者の購買意欲にはつながらず、逆に多すぎると消費者の興味が薄れてしまうのではないかと考える。またデータセット内にもある通り、商品説明だけでなく画像枚数や別のデータセットにある販売者評価など他の影響によるものが大きいのではないかと考えられる。

考察② 価格

- H2:製品の販売価格が安いほど累計販売個数は増加する→棄却

- 本分析では購買している製品のジャンルを「Office Furniture」に絞ったため、いわゆるオフィスで用いる用具において価格が安価に設定されている文房具類などが累計販売個数を考慮した際に増加すると考えたが、棄却となった。

- 本分析に用いたデータセットでは、製品のジャンルは判別できるものの詳細な中身については情報が記載されていなかった。実際のデータセット内のデータにおいても一製品の価格が1レアルから900レアル以上まで幅があり、分析には適していなかったのではないかと考えられる(1レアルは日本円で約27円)。

考察③ 商品名の長さ

- H3:製品名の長さが長いほど、累計販売個数は増加する→棄却

- 製品購買後の評価に関する被説明変数を削除したため、先行研究より新たに消費者が情報を取得する手段として商品名の長さが長いほど、情報の取得料が大きくなると考えたため設定したが、棄却となった。

- 本仮説も製品名の長さはデータセット上にて数値化されたいたものの、具体的にどのような製品名になっていたかどうかまでは判別ができなかったため、有意に働かない結果になったと考えられる。

最後に

- 本分析では「Olist」と呼ばれるブラジルのプラットフォームサービスが公開しているデータセットから消費者視点から考える購買意欲につながる要素を分析した。分析における仮説はすべて棄却という形に終わってしまったが、今後卒業論文に着手するにあたって本分析で学んだ分析方法、複数のデータセットの紐づけ方などを活かしていきたい。

- Acknowledgements
データを公開したOlist社、およびデータの分析を可能にしてくれたkaggle.com に感謝いたします。
We thanks to Olist for releasing this dataset. We also appreciate Kaggle.com for enabling analyze this data.

参考文献

・Amos Tversky,Itamar Simonson(1993) Context-dependent Preference

https://www.jstor.org/stable/2632953?seq=1#metadata_info_tab_contents

(2019年12月3日アクセス)

・青木均(2005)「インターネット通販と消費者の知覚リスク」pp.69-82

・Customer criteria for selecting online shops in Germany 2016, Statista

https://www.statista.com/statistics/451736/criteria-for-the-selection-of-online-shops-in-germany/(2019年11月27日アクセス)

・eCommerce South America, Statista

https://www.statista.com/outlook/243/103/ecommerce/south-america#market-revenue (2019年11月29日アクセス)

・Kaggle.com

https://www.kaggle.com/olistbr/brazilian-ecommerce#olist_products_dataset.csv (2019年11月26日アクセス)