Análise de Regressão Linear
Grupo 4

Heitor Augustaitis de Oliveira
Lucas Moura de Carvalho
Bruno Sergio Procopio Junior
Thales Simão do Amaral Camargo
Matheus Taipina Benini
Bruno de Oliveira Feitosa

28/09/2020

São Paulo–SP
2020

Sumário

1 Introdução

1.1 Motivação

Este relatório tem como objetivo mostrar as análises feitas pelo grupo, utilizando a linguagem R e a técnica de regressão linear, sobre a base de dados *summer-products-with-rating-and-performance_2020-08*. A preparação do ambiente de trabalho e as ferramentas empregadas serão apresentadas, assim como os comandos utilizados para processar os dados e suas respectivas saídas.

Uma questão natural é se os produtos seguem a economia de escala, onde quanto mais se vende mais se diminui o preço dele. A relação entre as unidades vendidas e a sua nota geral pode ser usada como teste de sanidade da suposição de que quanto mais se vende mais avaliado é um item. Outra questão interessante é o comportamento do preço do varejo (do consumidor final) em relação ao preço. Por fim as colunas em relação ao vendedor foram selecionadas para adicionar variedade a análise, vendedores com notas maiores tem maior nota de produtos, menores preços?

1.2 Preparação do ambiente

Para as nossas análises, utilizamos a linguagem R e as ferramentas R Studio e Jupyter Notebook. Tanto o script em R (.r) quanto o notebook (.ipynb) utilizado estão disponíveis aqui. Duas bibliotecas além do "base" foram utilizadas, a **dplyr** que é uma coleção de ferramentas para facilitar o manuseio de objetos como *data frames* e o **ggplot2** que é excelente para a criação de gráficos.

Execução dos comandos de carregamento da base dados e seleção dos campos a serem trabalhados:

```
# Le o dataset com as colunas desejadas
dataset <- read.csv(file="./summer-products-with-rating-and-performance_2020-08.csv", header=TRUE, sep=",")
dataset <- as.data.frame(dataset)
base    <- select(dataset, price, units_sold, rating, rating_count, retail_price,  merchant_rating, merchant_rating_count)

# Nomes alternativos
colnames(base) <- c("Price", "Units Sold", "Rating", "Rating C", "Retail Price", "Merch R", "Merch C")
```

A base de dados veio sobre formato *.csv* e sobre ela escolhemos sete colunas que condizem com as motivações da escolha da base, eles foram renomeadas de acordo com a tabela abaixo.

Coluna original	Coluna Nova	Significado
price	Price	preço do item
units_sold	Units Sold	número de unidades vendidas
rating	Rating	Sumário das notas dadas ao produto
rating_count	Rating C	quantidade de notas dadas ao produto
retail_price	Retail Price	preço de varejo
merchant_rating	Merch R	sumário das notas dadas ao vendedor
merchant_rating_count	March C	quantidade das notas dadas ao vendedor

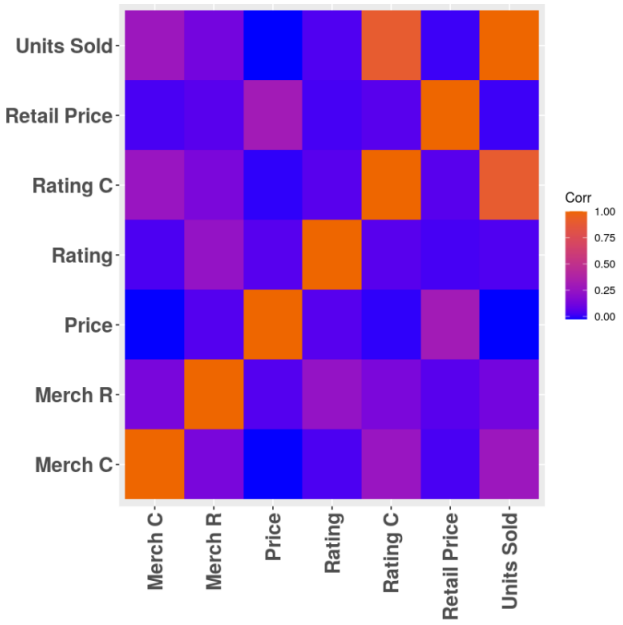
Tabela de colunas

2 Análise Inicial

2.1 Correlações

A primeira análise foi feita com uma matriz de correlação das colunas, a baixo está o seu mapa de calor.

```
# Mapa de calor de correlacao
cor_map = data.frame(rows = rep(colnames(base), each = ncol(base)), cols = rep(colnames(base), each = 1, times=ncol(bas
e)), Corr = c(cor(base)), stringsAsFactors=FALSE)

cor_heat = ggplot(cor_map, aes(rows, cols)) + geom_tile(aes(fill = Corr))
cor_heat = cor_heat + scale_fill_gradient(low = "#0000FF", high = "#EE6600") + theme(axis.title.y=element_blank(), axis.ti
tle.x=element_blank(), axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size=16, face="bold"), axis.text.y = e
lement_text(vjust = 0.5, hjust=1, size=16, face="bold"))
cor_heat

# Eliminares os valores com correlacao quase perfeita e muita baixa
cor_map = subset(cor_map, (abs(Corr) < 0.99 & abs(Corr) > 0.2))
```

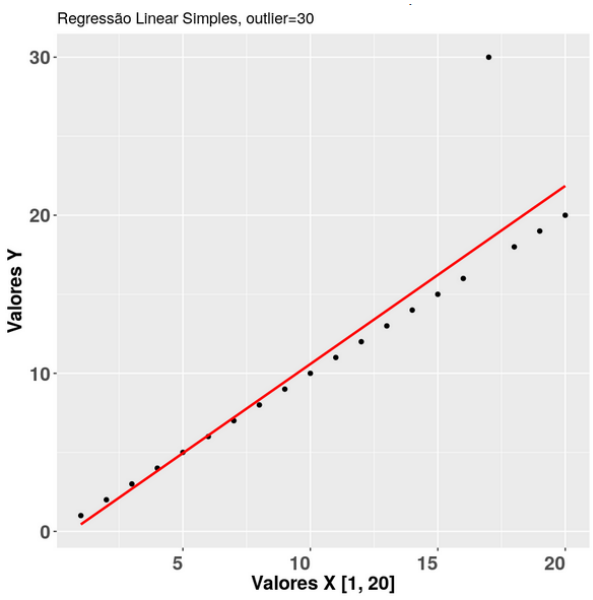


Matriz de Correlação

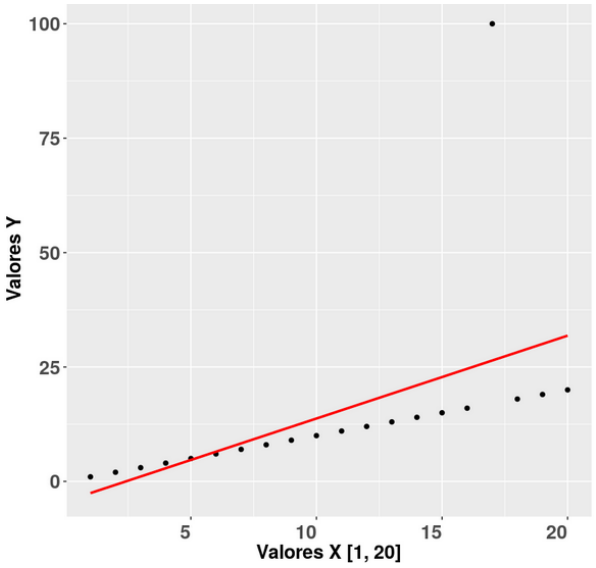
Observando a matriz, nota-se uma correlação com valor alto além das identidades, que é o de Rating C com Units Sold, o que era de se esperar. Dessa matriz foram eliminados os valores com correlação menor que 0.2 já que não há linearidade aparente significante.

2.2 Tratamento de outliers

Para uma regressão linear, a existência de outliers pode ser especialmente nociva uma vez que as posições dos valores é levadae em consideração de maneira linear e ascendente. Nas imagens abaixo fica nítido esse problema, onde dos pontos da curva $f(x) = x$ há apenas um outlier $(18,\ 30)$ na priemira (1) e $(18,\ 100)$ na segunda (2), que jogam a reta do modelo para longe dos outros pontos.



1. Regressão com outlier = 30



2. Regressão com outlier = 100

Foi criada uma função que remove linhas do dataset, se baseando em uma das colunas. É calculado o intervalo interquartil e são eleminados os valores além dos limites superior e inferior.

```r
## Seçao de outliers
# Quartis
Q1 = 0.25
Q2 = 0.5
Q3 = 0.75
Q4 = 1.0

# Calcula Inter Quartile Range (IQR) da lista
iqrange <- function(arr)
{
    quantile(arr, Q3) - quantile(arr, Q1)
}

# Limpa outliers se baseando em uma coluna
limpaOutlier <- function(df, col)
{
    IQR = iqrange(unlist(df[[col]]))
    lower = Q1 - 1.5*IQR
    upper = Q3 + 1.5*IQR
    subset(df, df[[col]] > lower & df[[col]] < upper)
}
```

3. Código para outliers

$$IQQ = Q1 - Q2$$

$$LS = Q3 + 1.5IQQ$$

$$LI = Q1 - 1.5IQQ$$

4. Definições de IQQ, LS, LI

Sobre a matriz de correlação foram aplicados os tratamentos de outliers e foram selecionados apenas os pares que sofreram um aumento na correlação.

```r
# Calcula novo cor_map aplicando remocao de outliers 2 a 2

cor_map$Corr_New <- NA
cor_rows = cor_map$"rows"
cor_cols = cor_map$"cols"

for(i in 1:(nrow(cor_map)))
{
    sub_base = limpaOutlier(base, cor_rows[[i]])
    sub_base = limpaOutlier(sub_base, cor_cols[[i]])
    sub_base = select(sub_base, cor_rows[i], cor_cols[[i]])
    cor_map[i, "Corr_New"] = cor(sub_base)[2]
}
cor_map_new <- subset(cor_map, abs(cor_map$Corr_New) > abs(cor_map$Corr))
cor_map
cor_map_new
```

5. Tratamento de outliers 2 a 2

	rows	cols	Corr	Corr_New
	<chr>	<chr>	<dbl>	<dbl>
5	Price	Retail Price	0.3047476	0.3992272
14	Units Sold	Merch C	0.2728973	0.2758841
28	Rating C	Merch C	0.2581676	0.2826965

6. Resultados

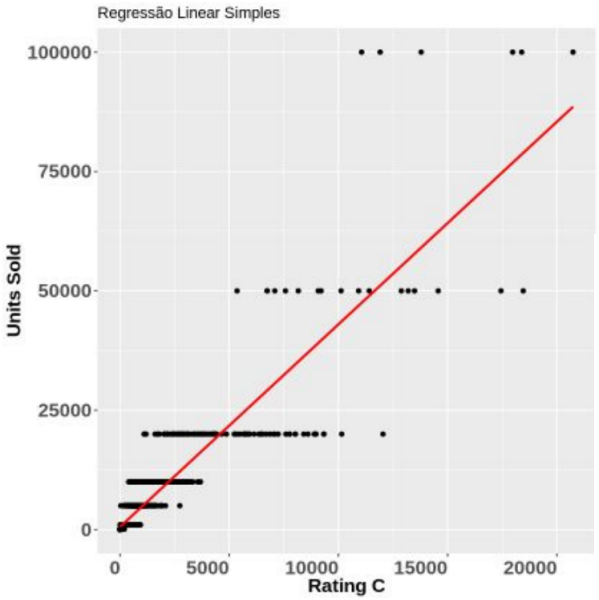
3 Análise de Regressão

Baseando-se na alta correlação encontrada entre Rating Count e Units Sold, foi feita a análise de regressão linear para essas duas variáveis, sendo Units Sold a variável dependente e Rating Count a variável independente.

É possível observar alguns resultados positivos da análise, como o valor 0.809 para o R^2 e o baixo valor para p-value, mostrando que a variável possui significância estatística para o modelo.

```r
xc = "Rating C"
yc = "Units Sold"

# Normal
lmGraph(base, NULL, base[[xc]], base[[yc]], xcol_name=xc, ycol_name=yc)

# Sem outlier
sub_base = limpaOutlier(base, xc)
sub_base = limpaOutlier(sub_base, yc)
lmGraph(sub_base, NULL, sub_base[[xc]], sub_base[[yc]], xcol_name=xc, ycol_name=yc)
```

A figura a seguir (7) mostra o gráfico de dispersão entre as variáveis Rating Count e Units Sold. Observa-se uma linearidade entre as variáveis, apesar do comportamento estranho da variável Units Sold, que possui os dados agrupados em valores específicos. Isso se deve ao fato de que, provavelmente, os valores dessa coluna foram arrendondados para faixas específicas de valores, o que causou a aparência atípica dos dados.

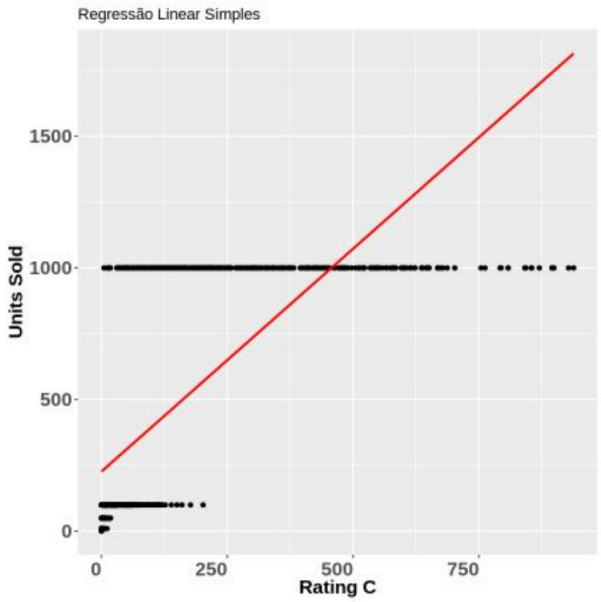


7. Regressão Rating C e Units Sold

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
   Min     1Q Median     3Q    Max
-31643   -694   -524    146  52510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  565.052    113.026   4.999  6.4e-07 ***
xcol           4.242      0.052  81.582  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4090 on 1571 degrees of freedom
Multiple R-squared:  0.809,     Adjusted R-squared:  0.8089
F-statistic:  6656 on 1 and 1571 DF,  p-value: < 2.2e-16
```

8. Sumário

O mesmo procedimento de regressão linear foi executado , porém retirando os outliers das duas variáveis utilizadas. O resultado foi um R^2 muito menor, e um p-value muito alto, o que retira a significância estatística da variável. Isso foi feito pois os valores da parte superior da ordenada são pequenos em número porém algumas vezes maiores em magnetude que os valores na parte inferior.



9. Sem outliers

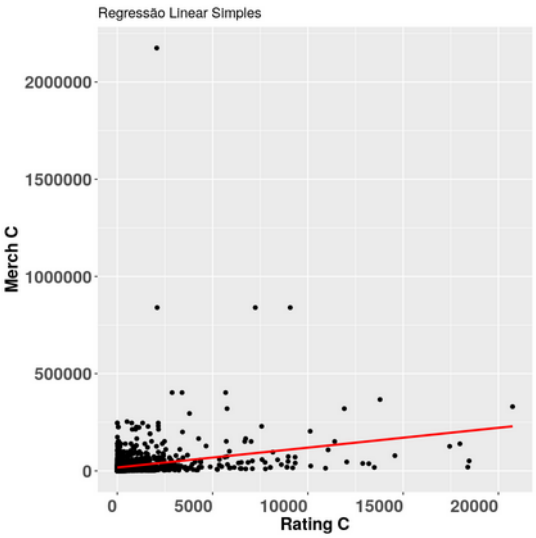
```
Call:
lm(formula = ycol ~ xcol)

Residuals:
   Min     1Q Median     3Q    Max
-814.1 -201.1 -152.0  253.8  764.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 226.62234   12.16664   18.63   <2e-16 ***
xcol          1.69240    0.05554   30.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325 on 1051 degrees of freedom
Multiple R-squared:  0.4691,    Adjusted R-squared:  0.4686
F-statistic: 928.5 on 1 and 1051 DF,  p-value: < 2.2e-16
```

10. Sumário

Veja que na segunda imagem (9) que sem aqueles valores o modelo centra-se no nas vendas feitas na linha de 1000 Units Sold que é bem larga e é desbalanceada de uma reta totalmente horizontal pelo grosso de valores na parte inferior. O fato da reta não

passar mais pelo centro dos valores inferiores indica um desbalanceamente por falta dos valores superiores que ficavam mais a frente no eixo da abscissa.

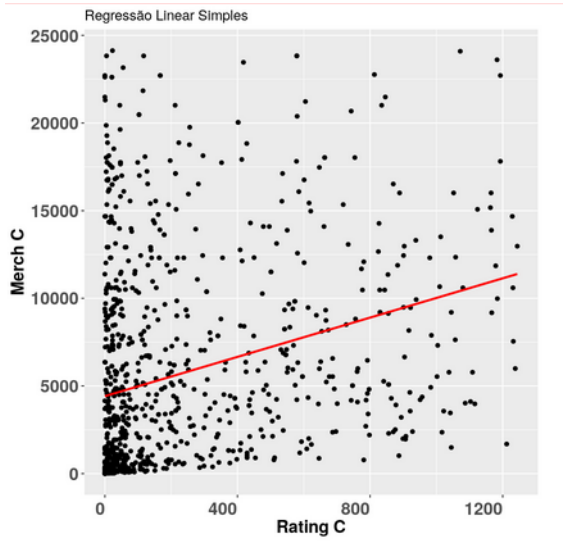
Sobre os três pares do mapa de correlação foram feitas regressões, tanto antes quanto depois da remoção dos outliers, resultando em valores não satisfatórios. Houve aumento do valores de R^2, mas mesmo assim permaneceram abaixo de 0.2, porem o valor dos coeficientes lineares encotrados nos revelam algumas informações interessantes.



11. Merch C e Rating C

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
    Min      1Q  Median      3Q     Max
-185989  -17547  -14710   -1753 2136236

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.741e+04  2.096e+03   8.308   <2e-16 ***
xcol        1.021e+01  9.641e-01  10.592   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75840 on 1571 degrees of freedom
Multiple R-squared:  0.06665,   Adjusted R-squared:  0.06606
F-statistic: 112.2 on 1 and 1571 DF,  p-value: < 2.2e-16
```

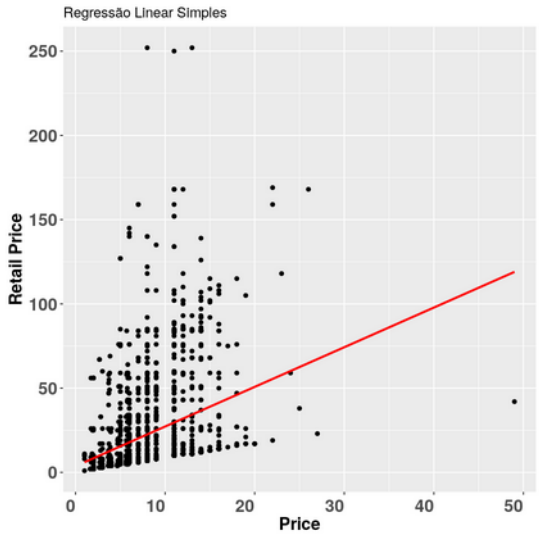
12. Sumário



13. Sem outliers

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
   Min     1Q Median     3Q    Max
 -9521  -4308  -2169   2755  19587

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4416.3955   218.1329  20.246   <2e-16 ***
xcol           5.6030     0.5953   9.413   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5672 on 1020 degrees of freedom
Multiple R-squared:  0.07992,   Adjusted R-squared:  0.07902
F-statistic:  88.6 on 1 and 1020 DF,  p-value: < 2.2e-16
```

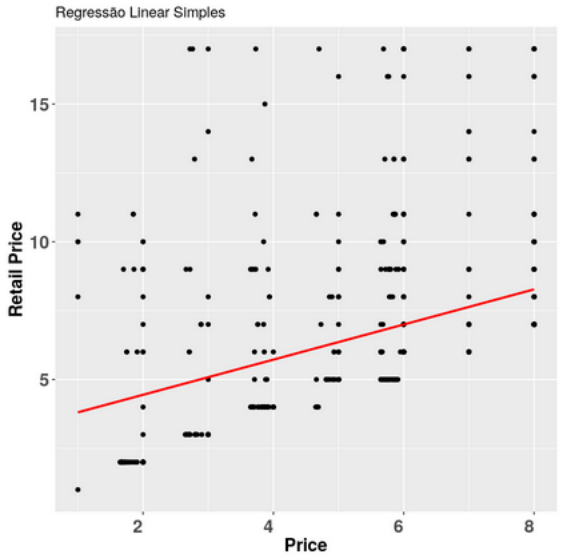
14. Sumário



15. Retail Price e Price

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
    Min      1Q  Median      3Q     Max
-76.990 -15.523 -10.465   3.947 229.477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7002     1.7081   2.166   0.0304 *
xcol          2.3529     0.1855  12.682   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.92 on 1571 degrees of freedom
Multiple R-squared:  0.09287,   Adjusted R-squared:  0.09229
F-statistic: 160.8 on 1 and 1571 DF,  p-value: < 2.2e-16
```

16. Sumário

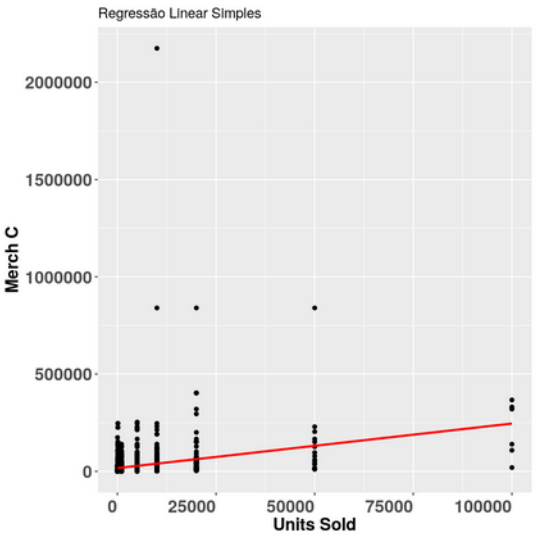


17. Sem outliers

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8062 -1.6339 -1.2718  0.5535 12.0965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.16827    0.33109   9.569   <2e-16 ***
xcol         0.63794    0.05456  11.692   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.962 on 721 degrees of freedom
Multiple R-squared:  0.1594,    Adjusted R-squared:  0.1582
F-statistic: 136.7 on 1 and 721 DF,  p-value: < 2.2e-16
```

18. Sumário

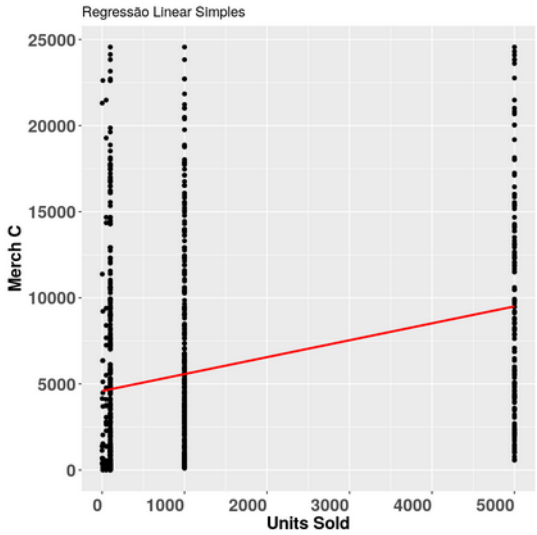
Para essa regressão o valor original da tangente era de 2.3 que é maior do que 1, sendo assim os preços de varejo, ou seja os preços do consumidor final, tendem a subir quanto maiores forem os preços do item. Agora, na segunda análise esse valor cai ao se retirarem os preços altos de vende e revenda, com uma tangente de 0.63, quanto mais aumenta o preço do item, menos o consumdir final tende a pagar a mais por ele.



19. Merch C e Units Sold

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
    Min      1Q  Median      3Q     Max
-226199  -16739  -14032   -1445 2135312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.656e+04  2.099e+03   7.892 5.55e-15 ***
xcol        2.289e+00  2.036e-01  11.243  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75520 on 1571 degrees of freedom
Multiple R-squared:  0.07447,   Adjusted R-squared:  0.07388
F-statistic: 126.4 on 1 and 1571 DF,  p-value: < 2.2e-16
```

20. Sumário



21. Sem outliers

```
Call:
lm(formula = ycol ~ xcol)

Residuals:
   Min     1Q Median     3Q    Max
 -8930  -4441  -2039   2927  19876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4589.4025   217.1400  21.136   <2e-16 ***
xcol           0.9824     0.1075   9.135   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5791 on 1013 degrees of freedom
Multiple R-squared:  0.07611,   Adjusted R-squared:  0.0752
F-statistic: 83.45 on 1 and 1013 DF,  p-value: < 2.2e-16
```

22. Sumário

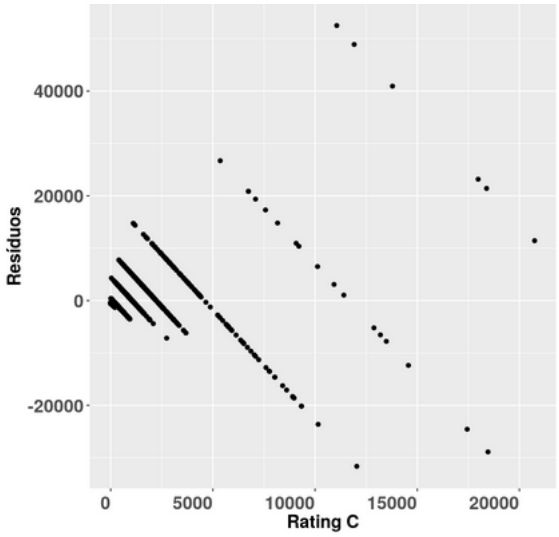
Foi feita uma regressão multipla com para Units Sold, que é uma variável de muito interesse, com todas as outras colunas exceto a Rating C que tem uma correlação muito alta. Os resultados também foram insatisfatórios, sem poder explicativo algum.

```
fit<-lm(base$"Units Sold" ~ base$"Price" + base$"Rating" + base$"Retail Price" + base$"Merch C" + base$"Merch R")
summary(fit)
```

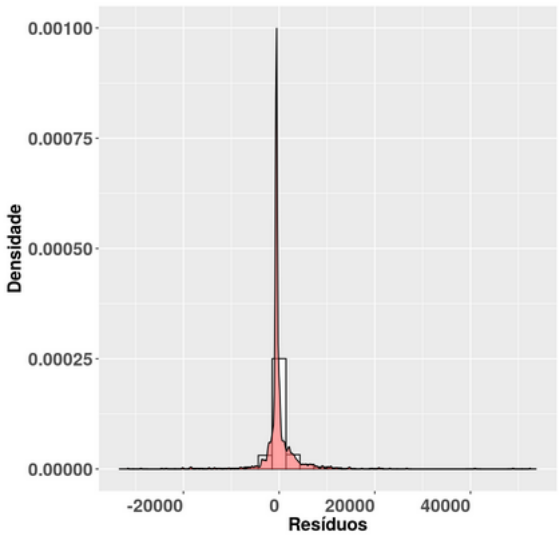
```
Call:
lm(formula = base$"Units Sold" ~ base$Price + base$Rating + base$"Retail Price" +
    base$"Merch C" + base$"Merch R")

Residuals:
   Min     1Q Median     3Q    Max
-62426  -3747  -2711    836  96352

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.251e+04  4.549e+03  -2.750 0.006031 **
base$Price          -6.113e+01  6.058e+01  -1.009 0.313103
base$Rating          2.121e+02  4.526e+02   0.468 0.639494
base$"Retail Price"  2.459e+00  7.841e+00   0.314 0.753909
base$"Merch C"       3.101e-02  2.915e-03  10.637  < 2e-16 ***
base$"Merch R"       3.886e+03  1.150e+03   3.379 0.000745 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8976 on 1567 degrees of freedom
Multiple R-squared:  0.08259,   Adjusted R-squared:  0.07966
F-statistic: 28.21 on 5 and 1567 DF,  p-value: < 2.2e-16
```

Figura 23: Regressão Multipla Units Sold

Por fim foram feitos o plot dos resíduos da regressão com outliers de Units Sold e Rating C, pelo Rating C e um histograma. A variância dos resíduos não é constante e aumenta tanto quanto aumenta Rating C, o que indica um modelo inadequado. O histograma aparente ter uma distribuição normal dos resíduos. Um teste de Shapiro-Wilk revela os resultados de 0.55379 e o $p - value < 2.2e - 16$, revelando que o modelo não é muito normal, assim rejeitamos a hipótse de normalidade.



24. Resíduos e Rating C



25. Histograma dos resíduos

4 Referências

Sales of summer clothes in E-commerce Wish. **Kaggle**, 2020. Disponível em: <https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish>. Acesso em: 3 de set. de 2020.

HAIR, Joseph F. et al. Análise multivariada de dados. Bookman editora, 2009.

BUSSAB, Wilton O.; MORETTIN, Pedro A. Estatística Básica, 5ª Edição, São Paulo. Editora Saraiva, 2006.

Sales of summer clothes in E-commerce Wish. **Kaggle**, 2020. Disponível em: <https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish>. Acesso em: 3 de set. de 2020.

Teste de Shapiro-Wilk. **Portal Action**. Disponível em: < http://www.portalaction.com.br/inferencia/64-teste-de-shapiro-wilk/ >. Acesso em: 3 de set. de 2020.