# BLACK FRIDAY

## Introdução a Modelos Dinâmicos

Relatório de análise do dataset

Docentes
Professora Diana Mendes (Coordenadora)
Professora Conceição Figueiredo

Elaborado por:

André Simões Novo nº 93343
Luís Miguel dos Santos Pereira nº 98398
Sebastião Manuel Inácio Rosalino nº 98437

Licenciatura em Ciência de Dados 2º Ano - Turma CDB1
Ano Letivo 2021/2022 - 1º Semestre

iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

Índice

# 1. Enquadramento do caso em estudo

## 1.1. Conceito de Black Friday

O presente trabalho tem por objeto a avaliação do comportamento dos consumidores no Black Friday, através da análise de um *dataset* de informação relativa a compras efetuadas durante esse período especial de descontos no comércio a retalho.

Importa começar por explicar, sucintamente, em que consiste a Black Friday. Trata-se de um conceito de vendas que nasceu nos Estados Unidos da América criado para incentivar as compras de Natal. Acontece sempre na última sexta-feira de novembro, depois do Dia de Acão de Graças (*Thanksgiving*, no original). O Thanksgiving é o feriado nacional mais celebrado nos EUA e é um dia tradicionalmente passado com a família. O Dia de Ação de Graças é sempre na última quinta-feira de novembro e as deslocações internas dos americanos para visitar a família afastavam-nos das lojas e, por essa razão, era um período de vendas muito fraco no comércio a retalho.

Graças ao Black Friday, um fim de semana que era tradicionalmente muito fraco para os retalhistas transformou-se num período de compras muito intenso, por força dos descontos muito generosos oferecidos para atrair os clientes. Este conceito resultou de tal forma que a ideia se espalhou um pouco por todo o mundo. Se antes da sua criação o dia era considerado "negro" pela fraca faturação, agora pode continuar a ser considerado "sombrio" mas para os clientes. Todos querem aproveitar as promoções e descontos o que leva a muita euforia que culmina na formação de longas filas, atropelamentos, e até comportamentos pouco cívicos dos consumidores, para além das fraudes associadas aos descontos oferecidos.

Perante esta realidade, o Black Friday passou a assumir um papel relevante no planeamento de vendas das empresas de grande consumo, as quais passaram a investir fortemente em campanhas de marketing neste período e a usar cada vez mais informação analítica para compreender os comportamentos dos clientes nos processos de compras, face aos estímulos criados. Neste contexto, a exploração de informação em grandes volumes sobre as vendas e sobre os clientes poderá dar importantes vantagens competitivas para as empresas que atuam nestes setores, criando, assim, um campo para a utilização da ciência de dados de extrema relevância.

## 1.2. Dataset utilizado e descrição das variáveis

O presente trabalho foi realizado por recurso a um dataset disponibilizado pelos docentes, em formato csv, com a seguinte designação: **BlackFriday.csv**.

Este dataset contém informação anonimizada sobre compras realizadas em 3 cidades, por consumidores categorizados pelo género, idade, profissão, estado civil e antiguidade de residência da cidade onde realizaram as compras. As compras, por sua vez, estão segmentadas por 3 categorias de produtos.

O dataset apresenta a seguinte estrutura de variáveis:

| Variável | Tipo de variável | Breve descrição |
|---|---|---|
| User_ID | Qualitativa Nominal | O id do comprador |
| Product_ID | Qualitativa Nominal | O id do produto comprado |
| Gender | Qualitativa Nominal | O sexo do comprador |
| Age | Qualitativa Ordinal | A faixa etária do comprador |
| Occupation | Qualitativa Nominal | A profissão do comprador (por código) |
| City_Category | Qualitativa Nominal | A categoria de cidade do comprador |
| Stay_In_Current_City_Years | Qualitativa Ordinal | Tempo em anos de residência |
| Marital_Status | Qualitativa Nominal | Estado civil |
| Product_Category_1 | Qualitativa Nominal | Categoria principal do produto |
| Product_Category_2 | Qualitativa Nominal | Subcategoria do produto |
| Product_Category_3 | Qualititativa Nominal | Subcategoria do produto |
| Purchase | Quantitativa Discreta | Dinheiro gasto |

## 1.3. Problema/questões a resolver e responder

Considerando a importância da Black Friday para as vendas das empresas no período pré-Natal, o objetivo do presente trabalho passa por poder analisar o conjunto de informação disponível no *dataset* para compreender o comportamento e as preferências dos consumidores neste período especial de compras, com estruturação da informação por alguns atributos relevantes desses consumidores e com segmentação pelos produtos vendidos.

Este estudo poderá ser importante para definir os níveis de stocks das empresas, a formação dos preços, o planeamento da força de vendas, entre outras dimensões

relevantes dos processos de venda, para além de permitir ajustar a oferta de produtos às reais preferências dos clientes.

Nesse sentido, com este trabalho procura-se compreender/responder a algumas das seguintes questões:

*Qual a relação entre o género e as compras?*

*Será que existe associação entre a faixa etária e as compras?*

*Será que a ocupação influência a preferência pelo Black Friday?*

*Será que a categoria de cidade influência as vendas?*

*Será que o estado civil influência o consumo?*

*Será que a permanência prolongada numa cidade influência o consumo?*

Neste relatório, foram investigadas diferentes preferências dos clientes em diferentes cidades, com diferentes profissões, casados e solteiros, mulher ou homem e tendo em conta o tempo de residência na cidade. Foram, ainda, usados modelos de regressão linear para testar a significância estatística dos resultados obtidos.

Foi concretizada a segmentação do *dataset* numa subamostra e utilizados modelos para explorar a relação entre o valor total da compra por pessoa e as suas informações sociodemográficas (foi usado o critério de filtragem de clientes residentes na sua cidade há 4 ou mais anos.). Por fim, foi efetuada uma análise preditiva sobre um *dataset* de treino e adotado um modelo para prever o valor da compra (variável Purchase) nos dados pertencentes ao *dataset* de teste.

No presente relatório apresenta-se alguns dos resultados obtidos ao logo da realização do trabalho, sendo que o código utilizado consta do ficheiro de R remetido em conjunto com este documento. Por economia de processo, optou-se, na maioria das situações, por não transcrever para este relatório o código criado em R.

## 2. Importação e limpeza de dados

A informação do *dataset* foi inicialmente analisada e trabalhada em Excel, com o propósito de estudar e compreender a estrutura dos dados.

Exemplo de um processo de análise realizado em Excel:

| | Género | |
|---|---|---|
| | M | F |
| n.º | 4225 | 1666 |
| Gastos totais | 3853044357 | 1164624021 |
| % gastos | 76,78953782 | 23,21046218 |
| % gastos por individuo | 911963,1614 | 699054,0342 |

| Género | Numero | Gastos totais | % Gastos | % Gastos por inividuo |
|---|---|---|---|---|
| M | 4225 | 3853044357 | 76,78953782 | 911963,1614 |
| F | 1666 | 1164624021 | 23,21046218 | 699054,0342 |

Após a análise dos dados em Excel, procedeu-se à importação da base de dados para R.

Foram também criadas duas tabelas bddistinctuser e bddistinctprod, contendo as informações referentes a clientes e produtos unícos.

Primeiras linhas da tabela bddistinctuser, sendo o Purchase o total de dinheiro gasto por um cliente em todas as suas compras

| | User_ID | Gender | Age | Occupation | City_Category | Stay_In_Current_City_Years | Marital_Status | Product_Category_1 | Product_Category_2 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000001 | F | 0-17 | 10 | A | 2 | 0 | 1 | 6 | 228003 |
| 2 | 1000002 | M | 55+ | 16 | C | 4+ | 0 | 8 | 17 | 628598 |
| 3 | 1000003 | M | 26-35 | 15 | A | 3 | 0 | 1 | 2 | 300754 |
| 4 | 1000004 | M | 46-50 | 7 | B | 2 | 1 | 1 | 8 | 186999 |

Primeiras linhas da tabela bddistinctprod, sendo o Purchase o total de dinheiro gerado por um produto em todas as suas vendas

| | Product_ID | Product_Category_1 | Product_Category_2 | Purchase |
|---|---|---|---|---|
| 1 | P00248942 | 1 | 6 | 9287185 |
| 2 | P00085442 | 12 | 14 | 485991 |
| 3 | P00193542 | 1 | 2 | 8666527 |
| 4 | P00184942 | 1 | 8 | 24060871 |

Após importação do ficheiro e das tabelas acima referidas procedeu-se à limpeza dos dados. Começou-se por eliminar toda a coluna Product_Category_3 (visto que continha 69% de NAs: Not Available) e posteriormente todas os registos da coluna Product_Category_2 que possuíam NAs (31% dos registos desta coluna foram eliminados). Todas estas limpezas permitiram reduzir os registos, passando de 537.577 para 370.591 linhas.

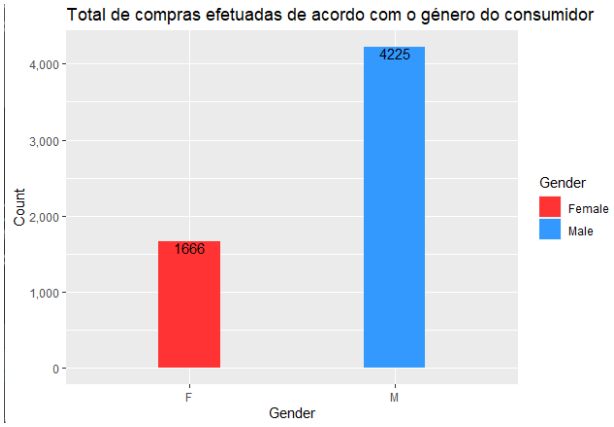
## 3. Estatísticas Descritivas e *Data Visualization* das variáveis

Analisando os dados, considerou-se útil obter conhecimento sobre duas perspetivas. Por um lado, sobre o perfil dos consumidores em análise e, por outro, sobre as categorias de produtos objeto de compra.

### 3.1. Análise dos consumidores

#### a) Por género

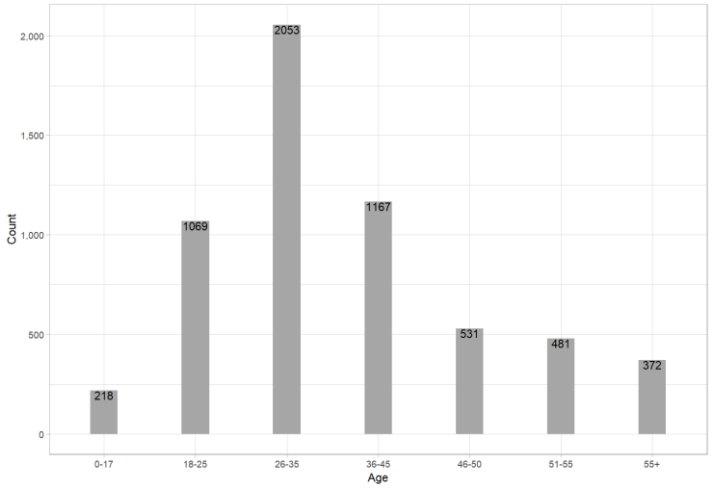
```
Clientes_final$Gender :
        Frequency Percent Cum. percent
F            1666    28.3         28.3
M            4225    71.7        100.0
  Total      5891   100.0        100.0
```




Pode verificar-se que dos 5.891 clientes analisados, cerca de 72% são do sexo masculino e apenas 28% do feminino.

### b) Por idade

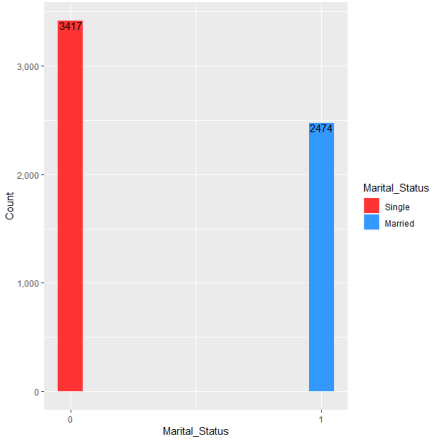
```
Clientes_final$Age :
        Frequency Percent Cum. percent
0-17          218     3.7         3.7
18-25        1069    18.1        21.8
26-35        2053    34.8        56.7
36-45        1167    19.8        76.5
46-50         531     9.0        85.5
51-55         481     8.2        93.7
55+           372     6.3       100.0
  Total      5891   100.0       100.0
```




Em termos de faixas etárias, verifica-se que a maior concentração de consumidores a utilizar as promoções do Black Friday se dá entre os 26 e os 35 anos (com cerca de 35%), seguido de forma equivalente pelas faixas entre os 36-45 anos e os 18-25 anos.

### c) Por estado civil

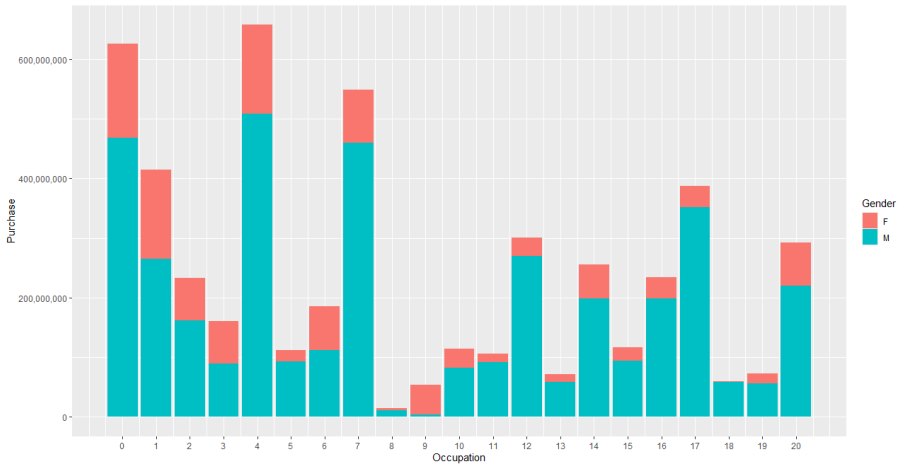
```
Clientes_final$Marital_Status :
        Frequency Percent Cum. percent
0            3417      58          58
1            2474      42         100
  Total      5891     100         100
```



Retira-se a informação de que os não casados (58%) poderão ser mais propensos a comprar no Black Friday do que os casados (42%).

### d) Por ocupação profissional

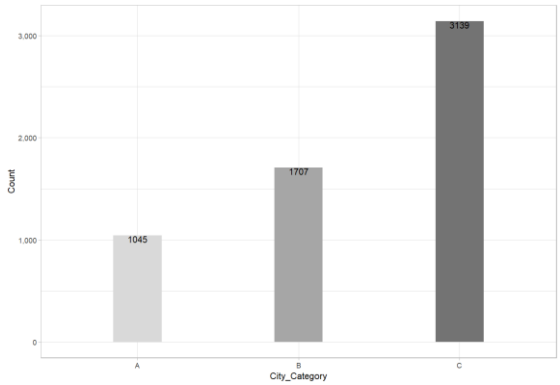
```
Clientes_final$Occupation :
        Frequency Percent Cum. percent
4             740    12.6         12.6
0             688    11.7         24.2
7             669    11.4         35.6
1             517     8.8         44.4
17            491     8.3         52.7
12            376     6.4         59.1
14            294     5.0         64.1
20            273     4.6         68.7
2             256     4.3         73.1
16            235     4.0         77.0
6             228     3.9         80.9
10            192     3.3         84.2
3             170     2.9         87.1
15            140     2.4         89.4
13            140     2.4         91.8
11            128     2.2         94.0
5             111     1.9         95.9
9              88     1.5         97.4
19             71     1.2         98.6
18             67     1.1         99.7
8              17     0.3        100.0
  Total      5891   100.0        100.0
```




A análise desta informação não permite retirar informação relevante, na medida em que não se conhecem as profissões em causa (estão anonimizadas).

### e) Por categoria de cidade

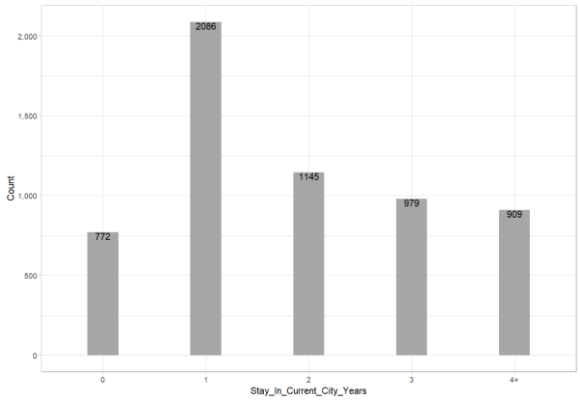
```
Clientes_final$City_Category :
        Frequency Percent Cum. percent
A            1045    17.7         17.7
B            1707    29.0         46.7
C            3139    53.3        100.0
  Total      5891   100.0        100.0
```




A base de dados em análise integra consumidores de 3 categorias de cidades, não especificamente identificadas, mas ordenadas por dimensão sendo a cidade A a mais pequena, a B a média e a C a maior. Verificou-se que essa ordem se mantém nos registos das compras.

### f) Por antiguidade na cidade

```
Clientes_final$Stay_In_Current_City_Years_Clean :
        Frequency Percent Cum. percent
0             772    13.1         13.1
1            2086    35.4         48.5
2            1145    19.4         68.0
3             979    16.6         84.6
4+            909    15.4        100.0
  Total      5891   100.0        100.0
```




Verificou-se que os clientes que mais efetuam compras habituam na sua cidade de residência há 1 ano. Por outro lado, os clientes que registam um menor número de

compras efetuadas habitam na sua cidade de residência há menos de 1 ano. Os clientes com residência há mais anos (2 ou mais anos) recorrem menos ao Black Friday.

### g) Estatísticas de compras por clientes

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 5891 | 851751.6 | 932997.8 | 512612 | 665292.5 | 490810.3 | 44108 | 10536783 | 10492675 | 2.46 | 8.82 | 12155.87 |

*(Variável Purchase da tabela Clientes_final)*

Estatísticas relevantes sobre os clientes:

Número de clientes únicos: 5891

Média de compras por cliente (em unidades monetárias: um): 851.751,6

Desvio padrão de compras por cliente (um): 932.997,8

Mediana de compras por cliente (um): 512.612

Mínimo de compras por cliente (um): 44.108

Máximo de compras por cliente (um): 10.536.783

Diferença entre o máximo e o mínimo de compras por cliente (um): 10.492.675

Skewness da variável: 2,46

Kurtosis da variável: 8,82

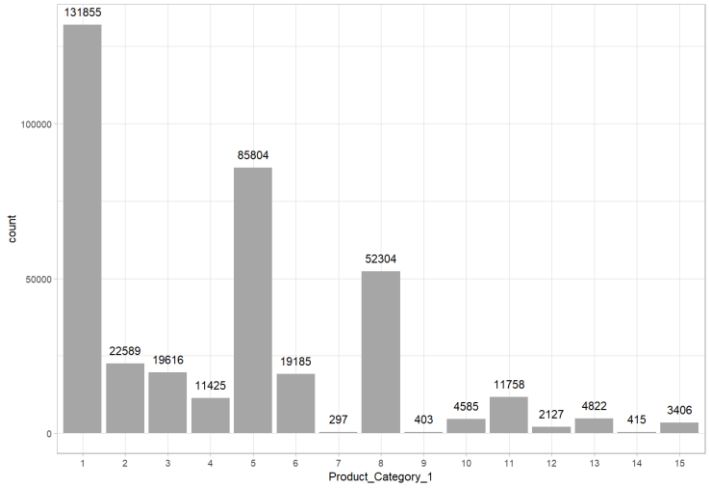
Erro padrão da variável: 12.155,87

## 3.2. Análise por categorias de produtos

### a) Categoria 1

bd$Product_Category_1 :

| | Frequency | Percent | Cum. percent |
|---|---|---|---|
| 1 | 131855 | 35.6 | 35.6 |
| 2 | 22589 | 6.1 | 41.7 |
| 3 | 19616 | 5.3 | 47.0 |
| 4 | 11425 | 3.1 | 50.1 |
| 5 | 85804 | 23.2 | 73.2 |
| 6 | 19185 | 5.2 | 78.4 |
| 7 | 297 | 0.1 | 78.5 |
| 8 | 52304 | 14.1 | 92.6 |
| 9 | 403 | 0.1 | 92.7 |
| 10 | 4585 | 1.2 | 93.9 |
| 11 | 11758 | 3.2 | 97.1 |
| 12 | 2127 | 0.6 | 97.7 |
| 13 | 4822 | 1.3 | 99.0 |
| 14 | 415 | 0.1 | 99.1 |
| 15 | 3406 | 0.9 | 100.0 |
| Total | 370591 | 100.0 | 100.0 |

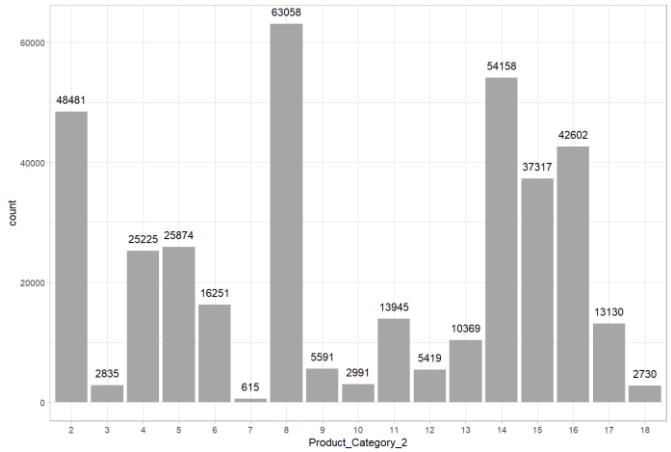



A categoria principal de produtos mais vendida é, com larga margem face às demais, a "1" com 121.855 unidades vendidos. Por outro lado, a categoria de produtos menos vendida é a categoria 7 com 297 unidades.

## b) Categoria 2

bd$Product_Category_2 :

| | Frequency | Percent | Cum. percent |
|---|---|---|---|
| 2 | 48481 | 13.1 | 13.1 |
| 3 | 2835 | 0.8 | 13.8 |
| 4 | 25225 | 6.8 | 20.7 |
| 5 | 25874 | 7.0 | 27.6 |
| 6 | 16251 | 4.4 | 32.0 |
| 7 | 615 | 0.2 | 32.2 |
| 8 | 63058 | 17.0 | 49.2 |
| 9 | 5591 | 1.5 | 50.7 |
| 10 | 2991 | 0.8 | 51.5 |
| 11 | 13945 | 3.8 | 55.3 |
| 12 | 5419 | 1.5 | 56.7 |
| 13 | 10369 | 2.8 | 59.5 |
| 14 | 54158 | 14.6 | 74.2 |
| 15 | 37317 | 10.1 | 84.2 |
| 16 | 42602 | 11.5 | 95.7 |
| 17 | 13130 | 3.5 | 99.3 |
| 18 | 2730 | 0.7 | 100.0 |
| Total | 370591 | 100.0 | 100.0 |




A subcategoria de produtos mais vendida é a categoria 8 (63058 unidades). É de salientar também a relevância, no que toca às vendas, das subcategorias 14 (54158 un), 2 (48481 un) e 16 (42602 un). A subcategoria de produtos menos vendida é a categoria 7, contando apenas com 615 unidades vendidas.

### c) Estatísticas de compras por produtos

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 3623 | 1384948 | 2608237 | 435264 | 795741.4 | 577717.3 | 405 | 27532426 | 27532021 | 4.43 | 26.89 | 43332.41 |

*(Variável Purchase da tabela Produtos_final)*

Estatísticas relevantes sobre compras por produto:

Número de produtos únicos: 3623

Média do valor de compra por produto (em unidades monetárias: um): 1.384.948

Desvio padrão do valor de compra (um): 2.608.237

Mediana do valor de compra (um): 435.264

Valor mínimo de compra (um): 405

Valor máximo de compra: 27.532.426

Diferença entre o valor máximo e o mínimo: 27.532.021

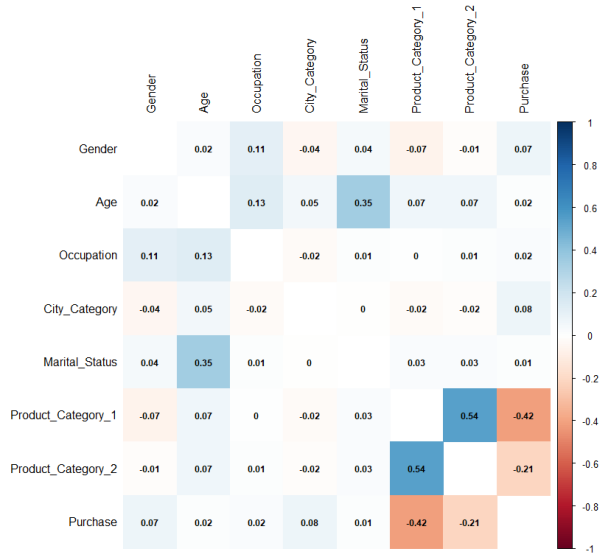
Skewness da variável: 4,43

Kurtosis da variável: 26,89

Erro padrão da variável: 43.332,41

## 4. Analise de correlação e causalidade entre as variáveis

Apresenta-se, de seguida, um gráfico de correlação entre todas as variáveis:




Correlações positivas mais fortes:

1º - Parece existir uma correlação positiva e moderadamente forte (0,54) entre as categorias de produto 1 e produto 2. No entanto, esta correlação poderá ser falsa, porque não são variáveis quantitativas, são variáveis nominais (não faz sentido concluir que à medida que avançamos na categoria de produto 1 a categoria de produto 2 aumenta, já que estes números são códigos e representam nomes).

2º - Existe uma correlação positiva moderada entre o estado civil e a idade (0.35), o que, sendo natural, não releva para o presente estudo.

3º - Parece existir uma correlação negativa e moderadamente forte entre as categorias purchase e as categorias de produto 1 (-0.42) e também de produto 2 (-0,21). No entanto, estas correlações poderão não ser significantes, porque não são variáveis quantitativas, são variáveis nominais (e não faz sentido concluir que à medida que avançamos na categoria de produto 1 o purchase aumenta, já que a categoria de produto 1 é um código representativo de um nome).

## 5. Pré-processamento dos dados e manipulação de variáveis

A variável dependente é o Purchase sendo todas as restantes (Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2) variáveis independentes.

Foi então criada uma subamostra contendo apenas os clientes que estão há 4 anos ou mais na sua cidade de residência (56.865 registos), cujo código é de seguida apresentado:

```
samplecity = bd %>% dplyr::filter(bd$Stay_In_Current_City_Years == "4+") %>%
dplyr::select(!c(Stay_In_Current_City_Years))
```

Primeiras 6 linhas da subamostra

Filter

| | User_ID | Product_ID | Gender | Age | Occupation | City_Category | Marital_Status | Product_Category_1 | Product_Category_2 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000008 | P00249542 | M | 26-35 | 12 | C | 1 | 1 | 5 | 19614 |
| 2 | 1000008 | P00220442 | M | 26-35 | 12 | C | 1 | 5 | 14 | 8584 |
| 3 | 1000008 | P00303442 | M | 26-35 | 12 | C | 1 | 1 | 8 | 11927 |
| 4 | 1000010 | P00085942 | F | 36-45 | 1 | B | 1 | 2 | 4 | 16352 |
| 5 | 1000010 | P00118742 | F | 36-45 | 1 | B | 1 | 5 | 11 | 8886 |
| 6 | 1000010 | P00058342 | F | 36-45 | 1 | B | 1 | 3 | 4 | 10946 |

Showing 1 to 6 of 56,865 entries, 10 total columns

## 6. Aplicação de algoritmos de aprendizagem supervisionada sobre o conjunto de dados

Neste processo são utilizados dois indicadores, a saber:

- **AIC** (Critério de Informação de Akaike), que avalia a qualidade de um modelo estatístico. Fornece, portanto, uma métrica para comparação e seleção de modelos, em que menores valores de AIC representam uma maior qualidade e simplicidade.
- **MAPE** (Erro Absoluto Médio em Percentagem), que se consubstancia numa métrica que avalia a precisão da previsão de um modelo estatístico.

Para efeitos de aprendizagem supervisionada (regressão linear) foram realizados 15 modelos diferentes sob o conjunto de dados (subamostra contem todos os clientes residentes na sua cidade à 4+ anos mais), cujos AICs são apresentados de seguida:

| | df | AIC |
|---|---|---|
| model1 | 62 | 1079172.26 |
| model2 | 62 | 48295.49 |
| model3 | 62 | 58685.03 |
| model4 | 62 | 48377.24 |
| model5 | 62 | 49852.66 |
| model6 | 62 | -8156.18 |
| model7 | 62 | -30699.69 |
| model8 | 62 | 29054.55 |
| model9A | 81 | 48025.20 |
| model9B | 81 | 28929.18 |
| model10A | 63 | 48267.24 |
| model10B | 63 | 29036.85 |
| model11A | 82 | 47931.66 |
| model11B | 82 | 29119.05 |
| model12A | 93 | 47790.34 |
| model12B | 93 | 28754.25 |
| model13A | 92 | 47874.16 |
| model13B | 92 | 28821.14 |
| model14A | 84 | 47927.06 |
| model14B | 84 | 29060.75 |
| model15A | 112 | 47647.57 |
| model15B | 112 | 28788.23 |

Apresenta-se de seguida os MAPEs de todos os modelos estudados:

| | V1 |
|---|---|
| MAPE1 | Inf |
| MAPE2 | 0.3234072 |
| MAPE3 | 0.3245964 |
| MAPE4 | 0.3233955 |
| MAPE5 | 0.3358864 |
| MAPE6 | 0.3247967 |
| MAPE7 | 0.3233919 |
| MAPE8 | 0.9128409 |
| MAPE9A | 0.3219366 |
| MAPE9B | 0.9175145 |
| MAPE10A | 0.3232441 |
| MAPE10B | 0.9174769 |
| MAPE11A | 0.3215184 |
| MAPE11B | 0.9184293 |
| MAPE12A | 0.3204490 |
| MAPE12B | 0.9175017 |
| MAPE13A | 0.3208223 |
| MAPE13B | 0.9138388 |
| MAPE14A | 0.3214786 |
| MAPE14B | 0.9196270 |
| MAPE15A | 0.3196102 |
| MAPE15B | 0.9105370 |

Dos 15 modelos utilizamos foram selecionados para futura comparação os 3 melhores modelos.

São estes os modelos 12A, 13A e 15A.

### a) O modelo 12A é construído da seguinte maneira:

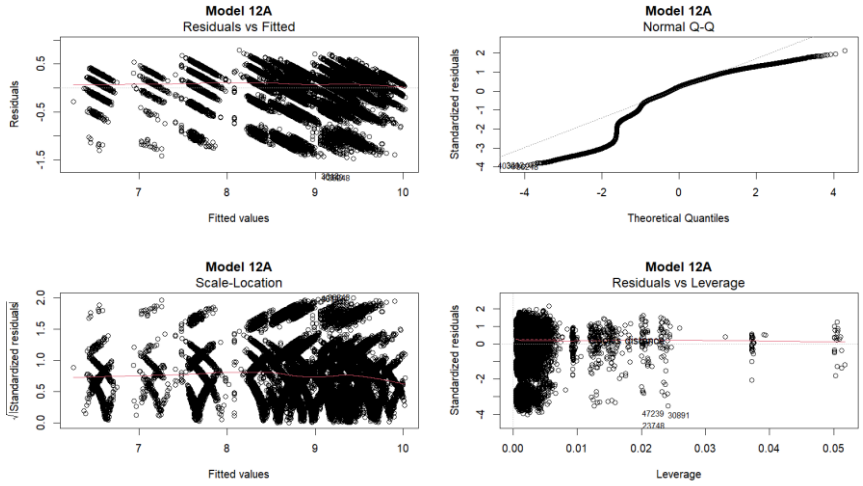
```
model12A = lm(formula = log(Purchase) ~ Age + Occupation + Gender +
        City_Category + Marital_Status + Product_Category_1 +
        Product_Category_2 + Occupation*Marital_Status + Age*City_Category ,
data = samplecity)
```

Sendo este o seu desempenho:

```
Residual standard error: 0.36804293284908007 on 56773 degrees of freedom
Multiple R-squared:  0.69314343584481675,       Adjusted R-squared:  0.69265158325048271
F-statistic: 1409.2503401012646 on 91 and 56773 DF,  p-value: < 2.22044604925031e-16
```

Este modelo logaritmiza a variável Purchase e interseta as variáveis Occupation com Marital_Status e Age com City_Category.

Para o modelo 12A os gráficos dos resíduos são os seguintes:




### b) O modelo 13A é construído da seguinte maneira:

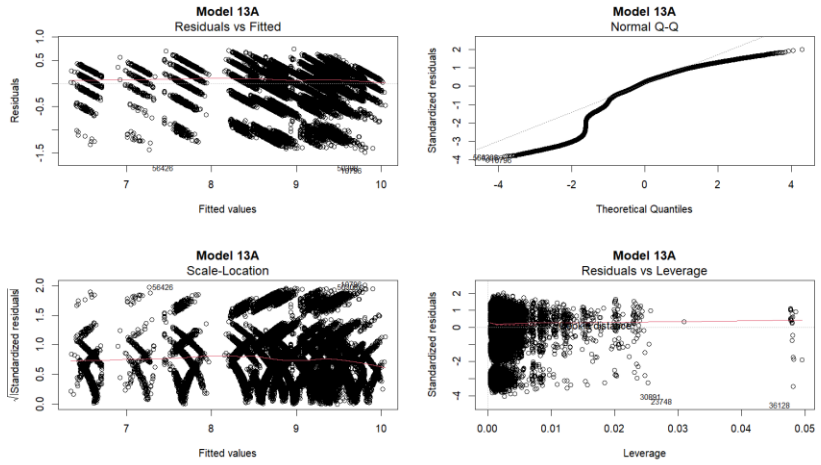
```
model13A = lm(formula = log(Purchase) ~ Age + Occupation + Gender +
        City_Category + Marital_Status + Product_Category_1 +
        Product_Category_2 + Occupation*Gender + Age*City_Category , data =
samplecity)
```

Este modelo logaritmiza a variável Purchase e interseta as variáveis Occupation com Gender e Age com City_Category.

Sendo este o seu desempenho:

```
Residual standard error: 0.36831753064473349 on 56774 degrees of freedom
Multiple R-squared:  0.69267995901374446,       Adjusted R-squared:  0.69219278524249761
F-statistic: 1421.8334399265245 on 90 and 56774 DF,  p-value: < 2.22044604925031e-16
```

Para o modelo 13A os gráficos dos resíduos são os seguintes:




### c) <u>O modelo 15A é construído da seguinte maneira:</u>

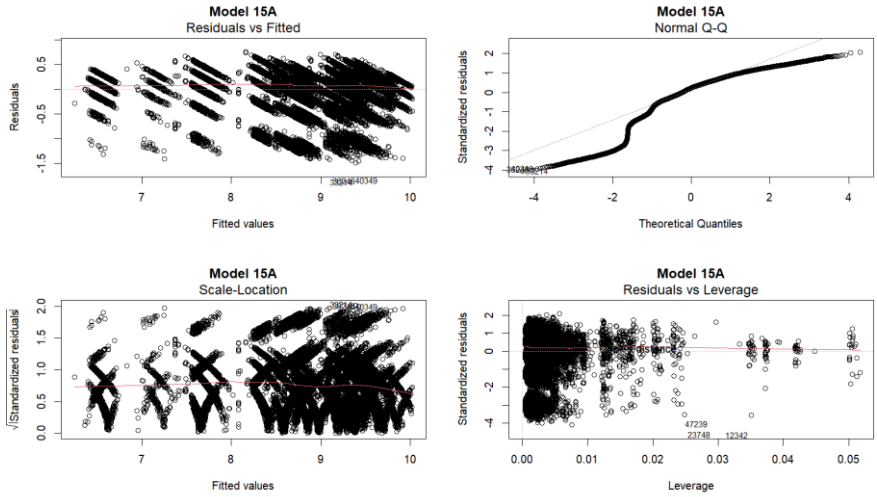
```
model15A = lm(formula = log(Purchase) ~ Age + Occupation + Gender +
        City_Category + Marital_Status + Product_Category_1 +
        Product_Category_2 + Occupation*Marital_Status +
Gender*City_Category*Age , data = samplecity)
```

Este modelo logaritmiza a variável Purchase e interseta as variáveis Occupation com Marital_Status e Gender com City_Category.

Sendo este o seu desempenho:

```
Residual standard error: 0.36751991103154769 on 56754 degrees of freedom
Multiple R-squared:  0.69411736004481583,       Adjusted R-squared:  0.69352450156091916
F-statistic: 1170.7977180028379 on 110 and 56754 DF,  p-value: < 2.22044604925031e-16
```

Para o modelo 15A os gráficos dos resíduos são os seguintes:




## 7. Divisão do *dataset* em conjuntos de treino e de teste.

Para posterior análise da capacidade preditiva dos modelos em uso, procedeu-se à divisão do *dataset* definido anteriormente em *dataset* de treino e teste.

Foi definido que o *dataset* de treino conteria 90% dos dados e que o *dataset* de teste os restantes 10%.

O *dataset* de treino passou a conter 51178 registos e o *dataset* de teste 5687.

Primeiras 6 linhas do *dataset* de treino:

| | User_ID | Product_ID | Gender | Age | Occupation | City_Category | Marital_Status | Product_Category_1 | Product_Category_2 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1001755 | P00210042 | F | 18-25 | 4 | B | 0 | 8 | 14 | 8032 |
| 2 | 1000710 | P00155442 | M | 26-35 | 20 | A | 0 | 1 | 11 | 15568 |
| 3 | 1003464 | P00130642 | F | 26-35 | 14 | B | 1 | 11 | 16 | 4529 |
| 4 | 1001667 | P00256042 | M | 51-55 | 16 | B | 0 | 6 | 8 | 16703 |
| 5 | 1003885 | P00115742 | M | 26-35 | 12 | B | 1 | 6 | 10 | 20460 |
| 6 | 1003111 | P00053942 | M | 18-25 | 4 | C | 0 | 1 | 2 | 19136 |

Showing 1 to 6 of 51,178 entries, 10 total columns

Primeiras 6 linhas do *dataset* de teste:

| | User_ID | Product_ID | Gender | Age | Occupation | City_Category | Marital_Status | Product_Category_1 | Product_Category_2 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|
| 51179 | 1002181 | P00113242 | M | 26-35 | 0 | B | 1 | 1 | 6 | 19610 |
| 51180 | 1002181 | P00344242 | M | 26-35 | 0 | B | 1 | 8 | 10 | 7891 |
| 51181 | 1002181 | P00215242 | M | 26-35 | 0 | B | 1 | 1 | 2 | 11470 |
| 51182 | 1002181 | P00057742 | M | 26-35 | 0 | B | 1 | 2 | 8 | 12914 |
| 51183 | 1002181 | P00084242 | M | 26-35 | 0 | B | 1 | 8 | 14 | 2140 |
| 51184 | 1002181 | P00000742 | M | 26-35 | 0 | B | 1 | 5 | 14 | 6873 |

Showing 1 to 6 of 5,687 entries, 10 total columns

Código utilizado para proceder às divisões do *dataset* em treino (*trainA) e teste (testeA).*

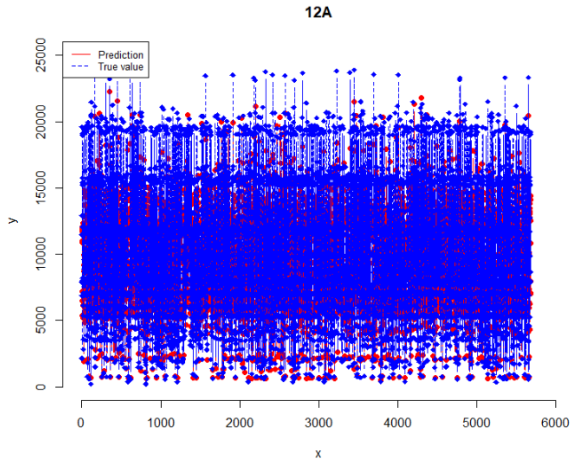
```
set.seed(7)
trainA = sample_frac(samplecity, 0.9)
sample_id = as.numeric(rownames(trainA))
testA = samplecity[-sample_id,]
```

## 8. Validação do modelo escolhido e previsão da variável target (sobre o dataset de teste)

Já com todos os treinos efetuados usando os três modelos selecionados sobre o *dataset* de treino, procedeu-se à previsão da variável *target* (Purchase) do *dataset* de teste, sendo estes os gráficos representativos dos resultados:

**Modelo 12A:**

**Predição vs Realidade:**

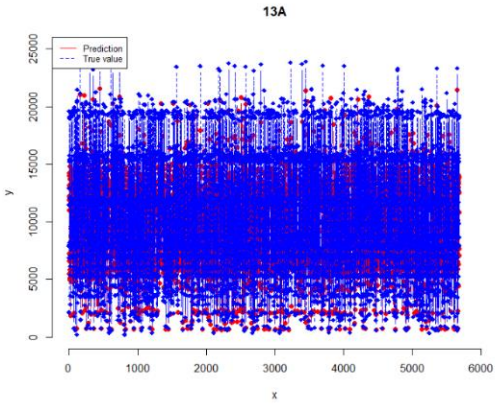



Desempenho do modelo de treino 12A:

```
Residual standard error: 0.36841372585573479 on 51086 degrees of freedom
Multiple R-squared:  0.69275055107611638,       Adjusted R-squared:  0.69220324457625204
F-statistic: 1265.7451560464465 on 91 and 51086 DF,  p-value: < 2.22044604925031e-16
```

**Modelo 13A:**

**Predição vs Realidade**:

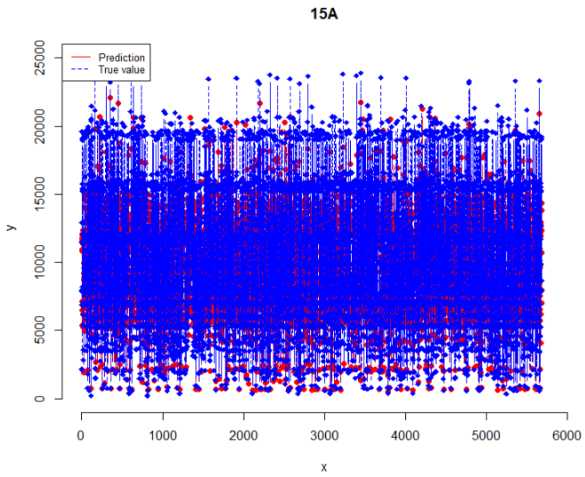



Desempenho do modelo de treino 13A:

```
Residual standard error: 0.36865460511877007 on 51087 degrees of freedom
Multiple R-squared:  0.69234262083064779,       Adjusted R-squared:  0.69180062063245173
F-statistic:  1277.384442173606 on 90 and 51087 DF,  p-value: < 2.22044604925031e-16
```

**Modelo 15A:**

**Predição vs Realidade:**




Desempenho do modelo de treino 15A:

```
Residual standard error: 0.36789800919406879 on 51067 degrees of freedom
Multiple R-squared:  0.69372409621648123,       Adjusted R-squared:  0.69306436783188474
F-statistic: 1051.5298604907273 on 110 and 51067 DF,  p-value: < 2.22044604925031e-16
```

Conclui-se que devido à grande dimensão da subamostra escolhida, os gráficos obtidos são de difícil compreensão/interpretação dada a concentração de pontos.

No entanto, é também possível verificar que na maioria dos gráficos os pontos dos valores reais e os pontos das previsões encontram-se sobrepostos o que confirma a qualidade dos modelos. É de realçar também a verificação do pressuposto relativo ao teste Breusch-Godfrey para todos os modelos de treino.

## 9. Avaliação da performance da previsão feita (sobre o conjunto de teste)

Importa, primeiramente, introduzir um novo conceito utilizado para a avaliação da performance da tarefa preditiva. O **RMSE** (Erro Médio Quadrático) é uma métrica utilizada para medir as diferenças entre os valores predizídos por um modelo e as observações reais.

A performance dos modelos sobre o conjunto de teste é essencialmente avaliada com dois indicadores, o **MAPE** e o **RMSE**. De seguida apresentamos as duas medidas para os 3 modelos estudados.

- Para o indicador MAPE:

| | Valores MAPE |
|---|---|
| Modelo 12A | 0.3171495 |
| Modelo 13A | 0.3177265 |
| Modelo 15A | 0.3159385 |

- Para o indicador RMSE:

| | Valores RMSE |
|---|---|
| Modelo 12A | 11381.692763202815 |
| Modelo 13A | 11381.692687262641 |
| Modelo 15A | 11381.693419633677 |

Conclui-se que, no que se refere ao indicador MAPE o modelo que apresenta uma melhor qualidade é o modelo 15A e, por outro lado, quanto ao indicador RMSE o modelo que apresenta menor erro nas predições efetuadas para o conjunto de teste é o modelo 13A.

## 10. Conclusões e análise de resultados

Para a retirada de conclusões foi selecionado o modelo 15A, na medida em que se revelou ser o que melhor qualidade de previsão apresenta.

Os modelos apresentam performances boas no que que toca ao R quadrado ajustado, pelo que podemos considerar que as análises produzidas são preditoras de comportamentos futuros.

Segue de seguida este indicador para os 15 modelos utilizados:

| | R-quadrado ajustado |
|---|---|
| model1 | 0.6206118 |
| model2 | 0.6897403 |
| model3 | 0.6873781 |
| model4 | 0.7027596 |
| model6 | 0.9832561 |
| model7 | 0.9999994 |
| model8 | 0.7494068 |
| model9A | 0.6913145 |
| model9B | 0.7510752 |
| model10A | 0.6898998 |
| model10B | 0.7503145 |
| model11A | 0.6918273 |
| model11B | 0.7505310 |
| model12A | 0.6926516 |
| model12B | 0.7522059 |
| model13A | 0.6921928 |
| model13B | 0.7527946 |
| model14A | 0.6918630 |
| model14B | 0.7503419 |
| model15A | 0.6935245 |
| model15B | 0.7530442 |

Passando à previsão out-sample (testes sobre o modelo treino) os modelos apresentam também bons desempenhos com indicadores de R quadrados médios razoáveis, verificando todos o pressuposto Breusch-Godfrey.

Ao nível da previsão in-sample (testes sobre a subamostra na integra), como já referido na questão 6, apresentam também bons resultados com indicadores significativos (em média), dos quais se destacam os seguintes (utilizando o modelo 15A):

a. No que toca à faixa etária, a que mais gasta é a dos 0 aos 17 anos. A faixa etária que menos gasta é a dos 18 aos 25 anos.
b. Em termos de ocupação a que mais gasta é a ocupação 2. A ocupação que menos gasta é a ocupação 5 (gastando em média menos 10%).

c. No que toca ao género, o que mais gasta é o feminino. O género masculino gasta em média menos 49% que o género feminino.
d. A categoria de cidade que mais gasta é a A. A categoria de cidade que menos gasta é a cidade B (gastando em média 51% que a cidade A).
e. No que toca à categoria ao estado civil, os solteiros registam um maior valor de compras. Os casados gastam em média 4.6% menos que os solteiros.
f. No que toca às categorias dos produtos, 23 das 30 combinações presentes apresentam significância no modelo estudado.
g. No que toca às interseções do modelo, quase todas mostram significância para o modelo.

Em síntese, pode afirmar-se que os modelos apresentam boa capacidade preditiva em relação consumo futuro no período do Black Friday, tendo por base os dados sociodemográficos constantes do *dataset* estudado.