Mineração de Dados em Larga Escala

Aula prática 4

1) Comentário dos resultados obtidos de *resumetable(df):*

```
> resumetable(df)
Dataset Shape:  129971x13
                     Name    dtypes Missing Uniques
1                 country character      63      44
2             description character       0  119955
3             designation character   37465   37980
4                  points   integer       0      21
5                   price   numeric    8996     391
6                province character      63     426
7                region_1 character   21247    1230
8                region_2 character   79460      18
9             taster_name character   26244      20
10 taster_twitter_handle character   31213      16
11                  title character       0  118840
12                variety character       1     708
13                 winery character       0   16757
```

Figura 1 - Resultado da função resumetable

A função *resumetable* é uma função descritiva de uma dataframe R onde são apresentadas algumas estatísticas das features, nomeadamente o Nome, tipo, valores em falta e valores únicos de cada feature.

Após observação da dataframe construída, conclui-se que se trata de um conjunto de dados relacionado com provas de vários tipos de vinhos. Este conjunto de dados inclui informação acerca do país (country) onde foi feita a prova, a descrição (description) e designação (designation) do vinho, bem como a respetiva pontuação (points) atribuída pelo provador (taster_name) identificado pelo nome e conta de twitter (taster_twitter_handle).

Quanto ao vinho, este é caracterizado pela província (province) e região (region_1 e region_2), o seu nome (title), tipo (variety), e a adega (winery).

Da observação dos valores em falta, podemos aferir que em certas features, este conjunto de dados é esparso nomeadamente nas regiões, preço, designação e identificação dos provadores, tanto o seu nome como a handle do twitter.

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

2) Descrição estatística dos dados numéricos

```
#Statistics of numerical data:
#              points          price
#count  129971.000000  120975.000000
#mean       88.447138      35.363389
#std         3.039730      41.022218
#min        80.000000       4.000000
#25%        86.000000      17.000000
#50%        88.000000      25.000000
#75%        91.000000      42.000000
#max       100.000000    3300.000000
```

Figura 2 - Resumo estatístico dos dados numéricos

- Points: Esta variável compreende-se entre 80 e 100, com um desvio padrão de 3.03, que não é muito elevado. Em simultâneo, nota-se que a média e mediana apresentam valores muito próximos, 88.45 e 88, respetivamente, o que mostra que os pontos atribuídos têm aproximadamente uma distribuição simétrica. As pontuações entre 80 e 91 estendem-se a três quartis (75%), sendo que o último corresponde a pontuações entre 92 e 100.

- Price: Ao observar a contagem de valores não nulos do preço, comprova-se que nem todos os dados relativos a provas de vinhos não apresentam valor uma vez que o conjunto de dados tem 129971 observações. Sendo que o mínimo e o máximo são respetivamente 4 e 3300 dólares, serve de melhor análise os preços por quartil. Até aos 75% dos valores de preço, estes são imperativamente inferiores a 42 dólares. Uma vez que os restantes 25% compreendem-se entre 43 e 3300 podemos ditar que existe um grande desvio padrão de valores, como registado pela função estatística, 41.02. Este valor mostra que a dispersão de valores é muito elevada e consequentemente os dados estão mais afastados da média de valores, que é de 35.36.

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

3) Interpretação da distribuição de *Points*.

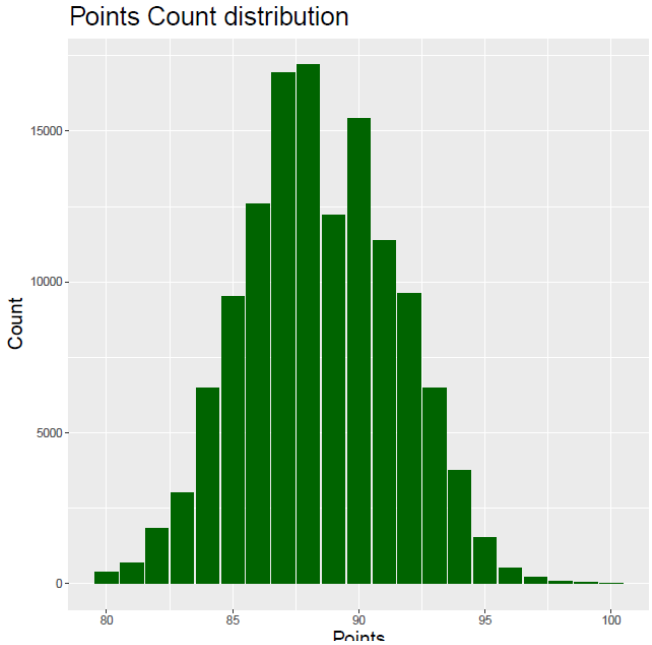

Figura 3- Histograma de pontuação

Do gráfico apresentado acima, podemos afirmar que a pontuação com maior representação neste conjunto de dados acontece aproximadamente no intervalo de [87, 93]. Conclui-se também que a distribuição da pontuação atribuída a cada um dos vinhos segue aproximadamente uma distribuição normal, onde os valores máximos e mínimos são pouco frequentes, e que, tal como comprovado anteriormente, o desvio padrão desta distribuição não é elevado, uma vez que os valores da média e da moda são próximos.

Aplicando um agrupamento por categorias dada a pontuação dos vinhos e gerando o gráfico dai resultante, concluímos que as pontuações entre 80 e 82 constituem apenas 2.3% das avaliações dadas pelos provadores, e pontuações acima dos 97 apenas 0.1%. Tal como mencionado anteriormente, 93% da distribuição das pontuações acontece entre os 83 e 93.

$$pontuação = \begin{cases} 0, & se\ [80,82] \\ 1, & se\ [83,86] \\ 2, & se\ [87,89] \\ 3, & se\ [90,93] \\ 4, & se\ [94,97] \\ 5, & se\ [98,100] \end{cases}$$

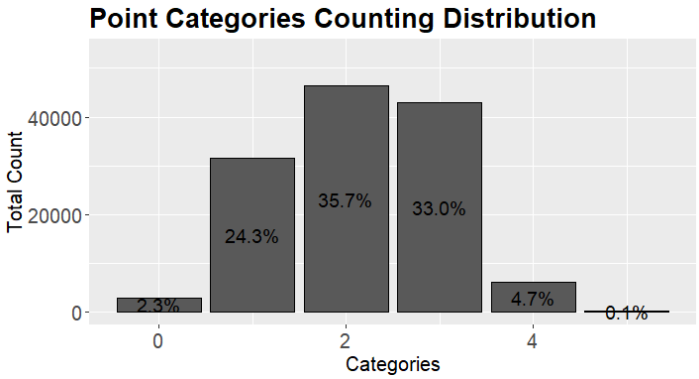

Figura 4 - Distribuição categorias

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

4) Cálculo dos *outliers*.

```
> CalcOutliers(df$points)
Lower cut values:  79
Upper cut values:  98
Identified lowest outliers:  0
Identified upper outliers:  129
Identified outliers:  129
Non-outlier observations:  129842
Total percentual of outliers:  0.0994 %
```

Figura 5 - Resultado de CalcOutliers

O cálculo dos *outliers,* utilizando a função *CalcOutliers*, identifica os valores de corte com base em três vezes o valor do desvio padrão do atributo numérico passado como argumento. A margem de corte inferior e superior é calculada com a média dos dados, subtraída ou somada do valor de corte, respetivamente. A contagem de *outliers,* bem como a sua distinção como inferiores e superiores, são guardadas e mostradas no *output* da função. Por fim, é mostrada a percentagem de *outliers* existentes no conjunto de dados original.

Desta interpretação, conclui-se que o conjunto de dados tem apenas *outliers* superiores, que corresponde a valores que superam 98, que compreende os 0.1% das pontuações atribuídas, observadas no gráfico da figura 3.

5) Distribuição de preços

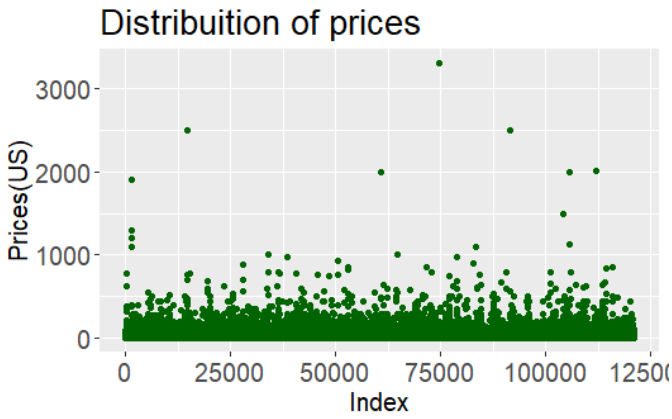

Figura 7 - Distribuição preços

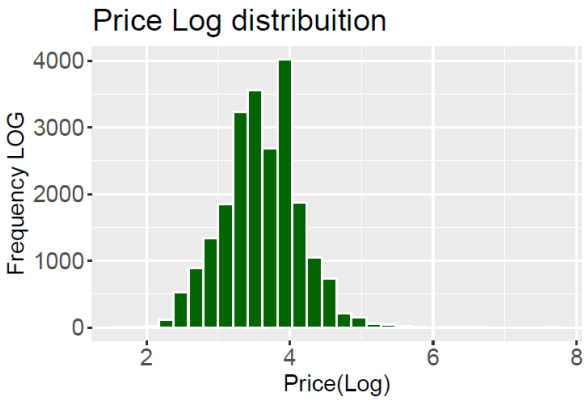

Figura 6 - Distribuição logarítmica do preço

Observamos da figura 4, distribuição de preços, uma grande concentração de valores abaixo dos 500 dólares, com um número muito menor a superar as 1000. Atendendo ao grande domínio de valores de preços para os vinhos, observamos da figura 5, um gráfico logarítmico com os valores discretizados em 30 *bins,* uma melhor compreensão da distribuição dos valores, verificando que a distribuição do histograma dos *bins* criados segue uma distribuição muito próxima da normal.

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

6) *Outliers* preços

```
> CalcOutliers(df$price %>% na.omit())
Lower cut values:  -88
Upper cut values:  158
Identified lowest outliers:  0
Identified upper outliers:  1177
Identified outliers:  1177
Non-outlier observations:  119798
Total percentual of Outliers:  0.9825 %
```

Referente aos *outliers* dos preços, para os seus cálculos é necessário descartar primeiros os valores em falta. Os resultados da função *CalcOutliers* mostram que existem apenas *outliers* superiores, que correspondem a 0.9825% dos dados existentes para esta característica no conjunto de dados. Este resultado é justificado pelo desvio padrão desta variável ser elevado (42.02), sendo a média dos valores muito afastada do valor máximo (35.36 a 3000). Relembrando que o valor mínimo dos preços é 4, a existência de *outliers* inferior seria muito pouco provável, o que foi comprovada pela função.

7) Gama de valores de vinhos abaixo de 300 dólares

Sugeriu-se a validação da distribuição de preços abaixo dos 300 dólares:

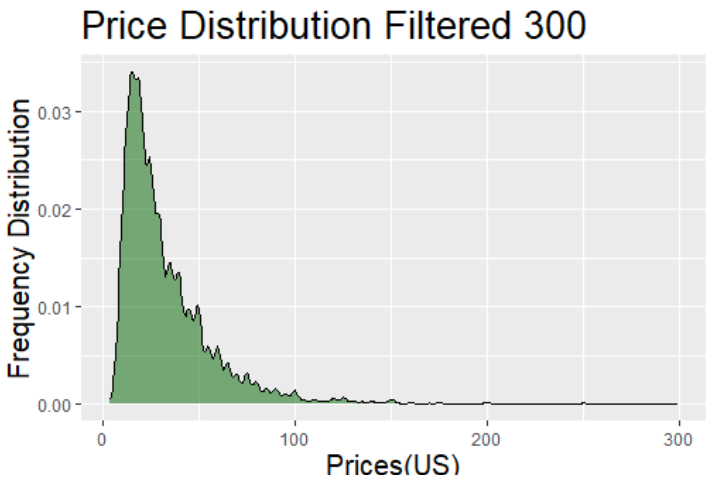

Figura 8 - Preço abaixo 300 Dollars

Da observação do gráfico na figura 6, comprovamos que a grande maioria dos vinhos se situa entre os 0 e 50 dólares, contribuindo entre 1 a 3% para a distribuição de preços na gama dos 4 aos 300 dólares. À medida que o preço evolui, a distribuição decresce cada vez mais, sendo quase mínima para os valores no extremo superior.

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

8) Características que distinguem o vinho mais caro do vinho com maior pontuação

O gráfico apresentado pela figura 7 representa a distribuição das pontuações em função do preço.

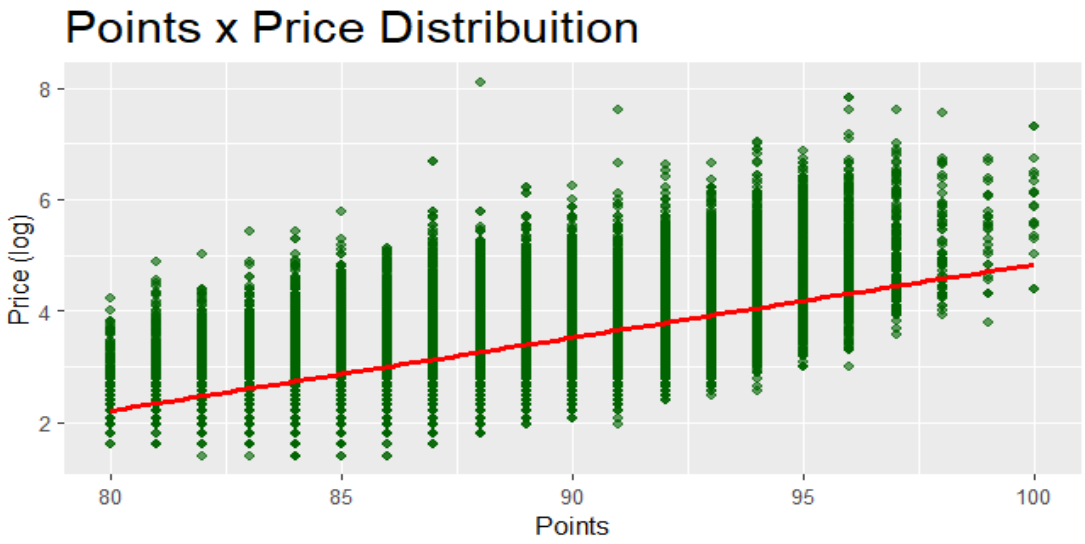

Figura 9- Pontuação vs preço

Deste gráfico, observamos que a evolução de pontuação acompanha o crescimento do preço. Ao observar em detalhe o *dataset* podemos verificar que os vinhos com pontuação mais elevada pertencem a uma gama muito restrita de províncias, enfatizando a região de Bordeaux. No entanto, observando os vinhos mais caros e os vinhos com maior pontuação, a província, a região e a *variety* do vinho também aparentam refletir diferenças na pontuação. Desta forma, o *dataset* do sistema de recomendação também podia ser construído a partir destas características: *variety, price*¸ *province*, *region*.

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

9) Comparação dos países de origem dos vinhos

Da observação dos seguintes gráficos concluimos:

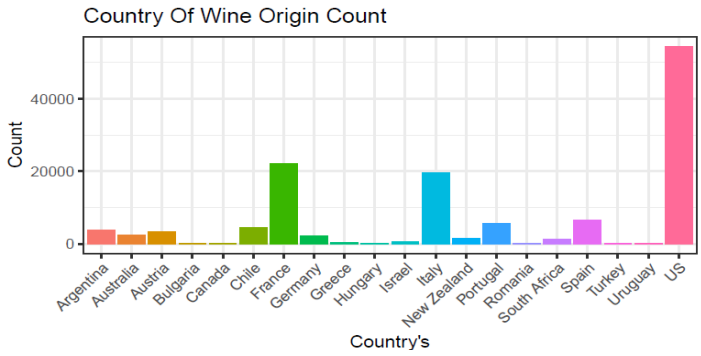

Figura 10 - Contagem do país de origem

Os vinhos mais representados no conjunto de dados pertencem aos Estados Unidos da América, França e Itália, respetivamente.

Analisando agora os pontos atribuidos por país de origem dos vinhos:

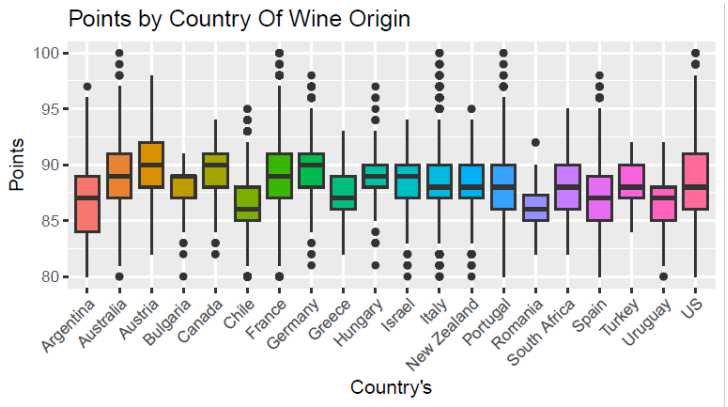

Figura 11 - Boxplot país de origem

Notamos que a Argentina, Roménia e Chile são as que têm um primeiro quartil menor, pelo que indicam que os primeiros 25% vinhos da amostra destes países obtiveram as menores pontuações. Os Estados Unidos tem vinhos com pontuações nos dois extremos, com o melhor e pior pontuação e primeiro e último quartis mais distantes, no entanto, são também o país mais representado pelo que a amostra é maior. Ao observar o grafico abaixo, *boxplot* dos preços, os mais baratos são produzidos na Roménia, com França a produzir os mais caros, que coincidentemente são também os mais bem pontuados.

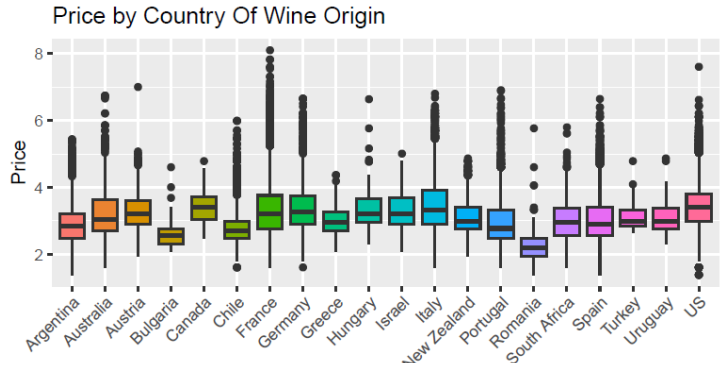

Figura 12 - Boxplot do preço por país de origem

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

10) Comparação dos provadores dos vinhos

A destacar algumas conclusões (no seguimento do raciocínio dos exercícios anteriores).

Dos gráficos seguintes, observa-se (Fig. 8) 19 provedores de vinho, sendo 'Roger Voss' quem mais provas fez, e aproximadamente 25000 amostras sem identificação do provedor; (Fig. 9) 'Anne' e 'Matt' com a mediana de provas mais elevada. Conclui-se também que o Roger Voss foi quem provou vinhos mais caros, no entanto, partilha a atribuição da pontuação mais elevada com Kerin O'Keele, quarto colocado quanto a número de provas de vinhos realizadas.

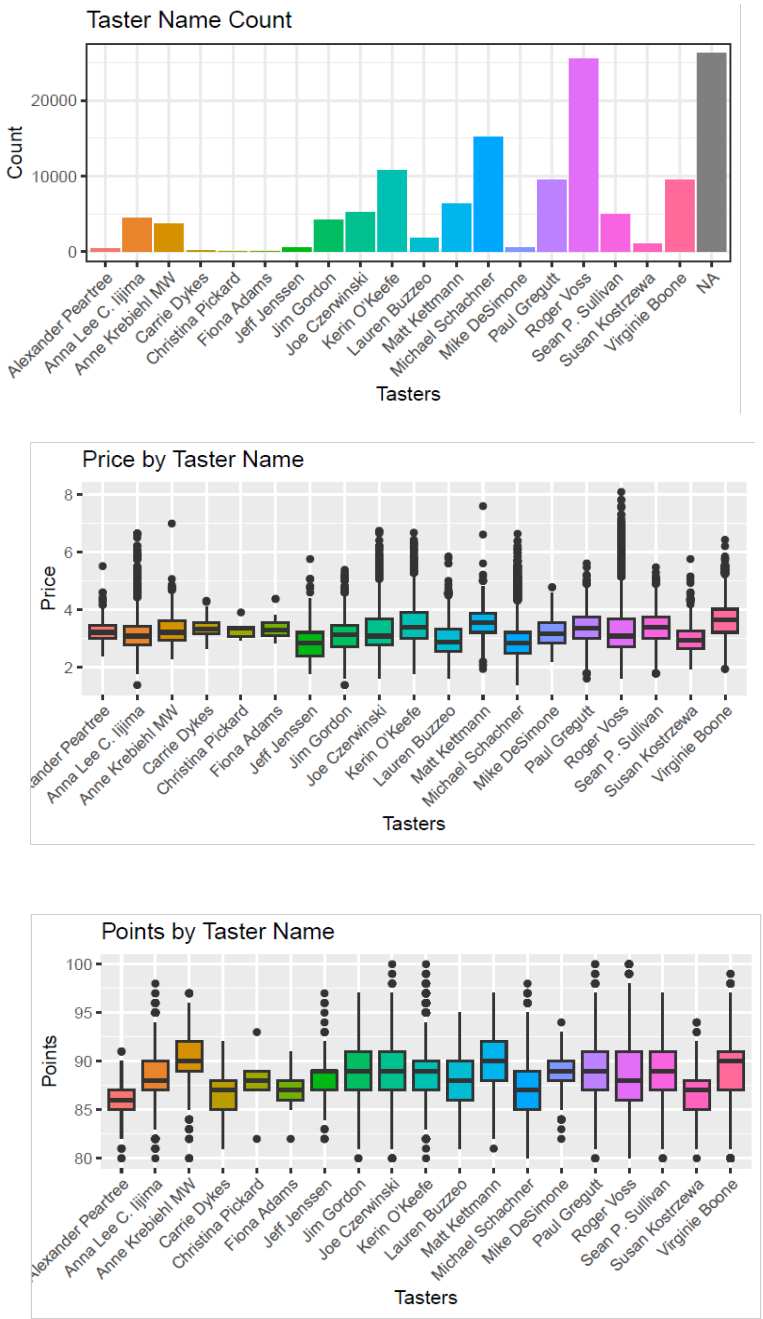

Figura 13 - Contagem e boxplots de preço e pontuação por taster

Grupo 4:

Nuno Gomes Nº 18364

Rafael Carvalho Nº 47663

Ricardo Ramos Nº 46638

11) Comparação das províncias de origem dos vinhos

Verifica-se que o número de vinhos varia de região para região, destacando-se maior quantidade na Califórnia (USA). No entanto, os vinhos mais caros são provenientes da região de Bordeaux e descartando os valores considerados discrepantes pelo *boxplot*, podemos concluir que Burguyndy e Champagne têm os vinhos tipicamente melhor pontuados dada a proximidade da mediana ao terceiro quartil.

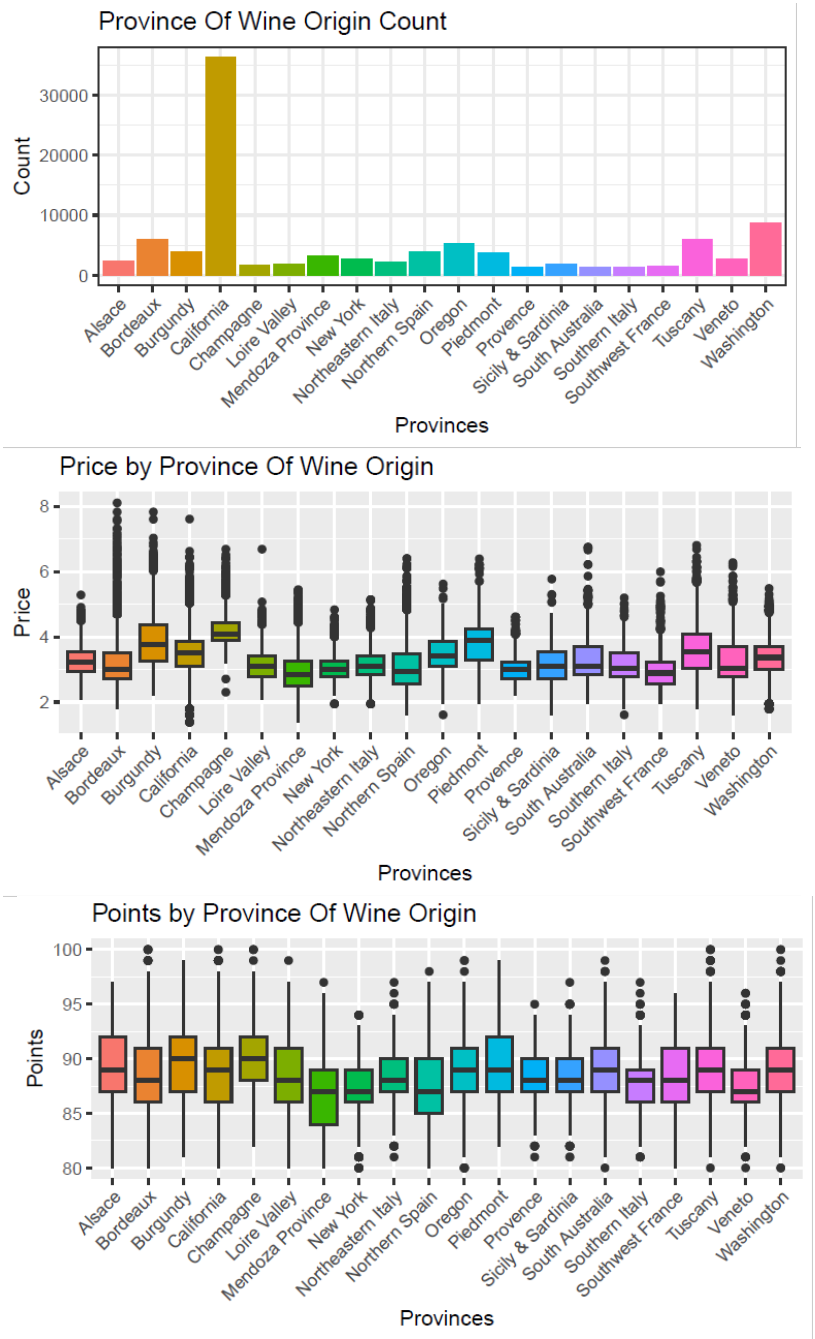

Figura 14 - Contagem e boxplots de preço e pontuação por província de origem

Grupo 4:

Nuno Gomes Nº 18364

Rafael Carvalho Nº 47663

Ricardo Ramos Nº 46638

12) Comparação das variedades dos vinhos

Pelos gráficos observamos que a variedade de vinhos mais presentes são os franceses, com o Pinot Noir em destaque, seguindo-se do Chardonnay, Cabernet Sauvignon e Bordeaux-style Red Blend. Olhando aos preços, considerando os vinhos mais caros são precisamente estes, com menor variação de preços, à exceção do Cabernet Sauvignon, onde o primeiro e terceiro quartil são mais distantes.

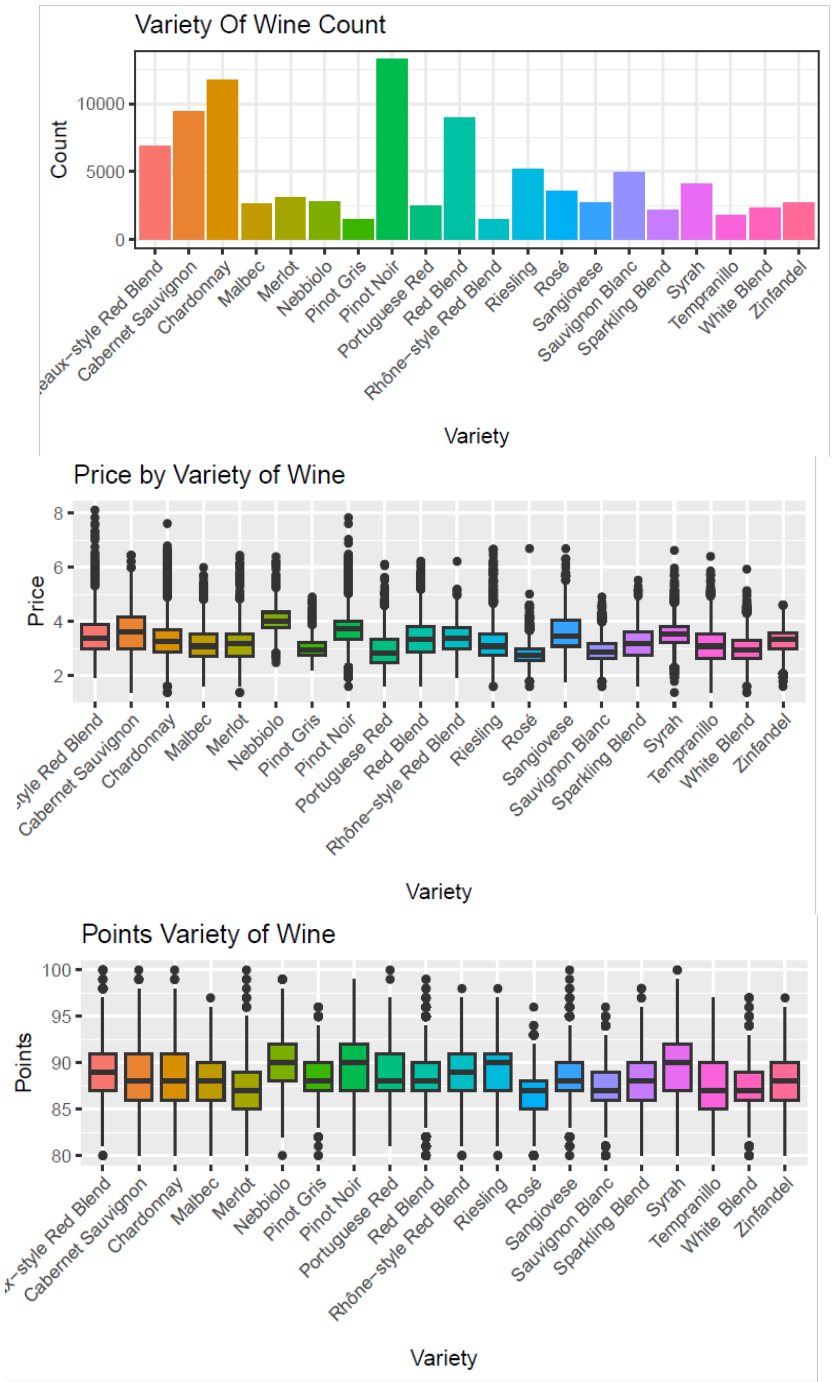

Figura 15 - Contagem e boxplots de preço e pontuação por variedade

Grupo 4:

Nuno Gomes Nº 18364

Rafael Carvalho Nº 47663

Ricardo Ramos Nº 46638

13) Comparação das adegas

As adegas mais representadas no conjunto de dados são a “Wines & Winemakers”, que é uma adega portuguesa, Siduri e Testarossa, localizadas na Califórnia, que coincide com a forte representação vista anteriormente. Ambas são conhecidas por produzir vinhos Pinot Noir e Chardonnay, também dos mais representados no conjunto de dados. Verifica-se então a forte influência da larga representação dos Estados Unidos da América, e do estado da Califórnia.

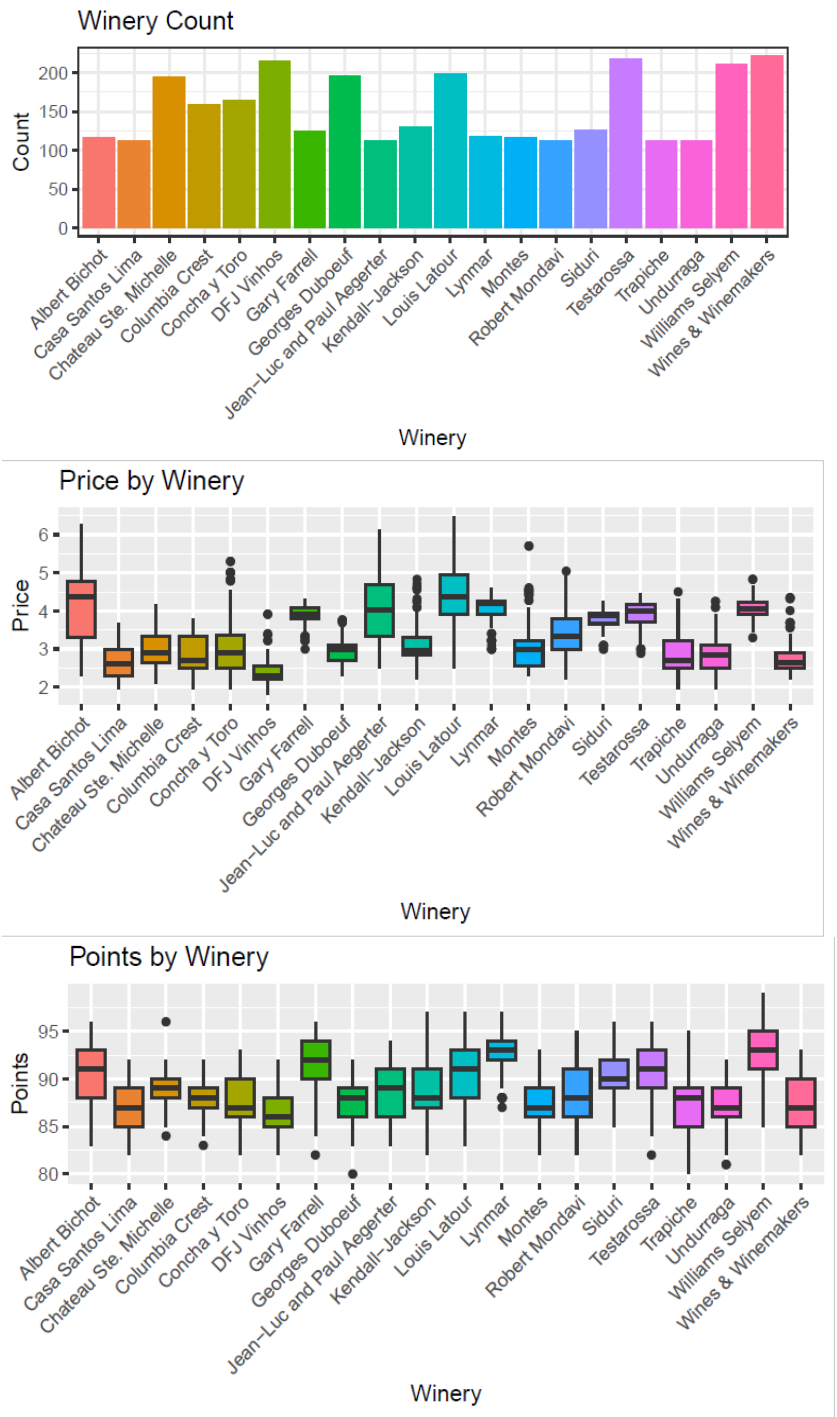

Figura 16 - Contagem e boxplots de preço e pontuação por adega

Grupo 4:

Nuno Gomes Nº 18364

Rafael Carvalho Nº 47663

Ricardo Ramos Nº 46638

14) Wordcloud

A *wordcloud* permite representar o conjunto de palavras presentes nos dados, neste caso, nas descrições feitas pelos provadores dos vinhos. Quanto maiores as palavras, maior a frequência absoluta no texto. Desta forma, podemos concluir, como esperado, dado o teor do conjunto de dados, que as palavras mais frequentes são essencialmente adjetivos descritivos do "*wine*", nomeadamente: *"palate", "flavors", "fruit", "acidity"*

Figura 17 - Wordcloud de descrições

Elaborou-se também a *wordcloud* dos títulos (titles) dos vinhos, onde notamos com maior destaque vinhos de diferentes anos, com destaque a "*valley", "pinot", "red",* e "*sauvignon"*.

Figura 18 - Wordcloud de titles

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

15) Correção dos códigos e explicação do Recommender System

Pretende-se implementar um algoritmo de recomendação com base em *Collaborative Filtering,* que se trata de um método de fazer previsões automáticas com base em interesses de um *user* com base em interesses e gostos de outros *users* semelhantes.
Este *Recommender System* é implementado com base na semelhança entre dois utilizadores segundo o algoritmo *k-nearest neighbours* (KNN), onde o *rating* r_{ui} é calculado com base nos ratings de outros utilizadores '*u*' a um item '*i*'.
Observando o código fornecido, notou-se inicialmente que no método *get_top_terms* não estavam a ser considerados os termos na cláusula do método *group_by,* como demonstrado:

Antes:

```
group_by(category, doc_id) %>%
summarize(total_count = sum(term_count)) %>%
group_by(category) %>%
```

Depois:

```
group_by(terms, category, doc_id) %>%
summarize(total_count = sum(term_count)) %>%
group_by(terms) %>%
```

Figura 19 - Alteração get_top_terms

Para guardar os histogramas de termos mais recorrentes nos diferentes países, ajustou-se o código R:

```
p_list <- lapply(unique(df_top_terms$category), function(cat) {
  df_plot <- df_top_terms %>%
    filter(category == cat) %>%
    mutate(terms = reorder(terms, rank))
  ggplot(df_plot, aes(x = terms, y = total_count, fill = category)) +
    geom_col() +
    ggtitle(paste0("Wines from ", cat, " N-grams")) +
    theme(plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
          axis.text.x = element_text(angle = 90, hjust = 1)) +
    labs(x = NULL, y = NULL, fill = NULL) +
    scale_fill_manual(values = c("#F8766D", "#00BFC4", "#E76BF3", "#7CAE00",
  ggsave(paste("../plots/",cat, ".pdf"))
})
```

Figura 20 - Alteração para guardar histogramas de cada category

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

Contudo, mostrando o exemplo dos termos para Portugal, é difícil compreender qual o termo mais frequente, devido aos vários termos que aparecem.

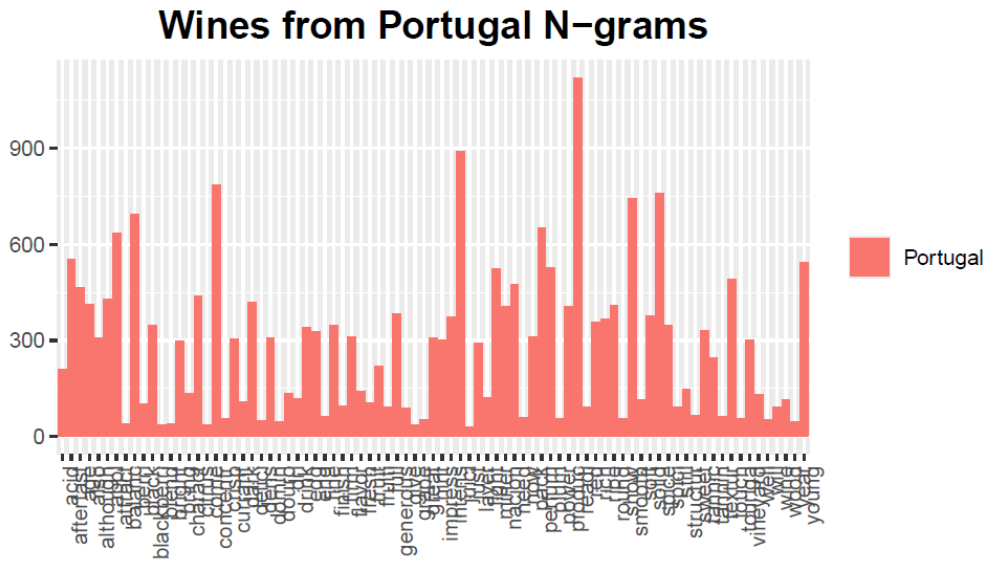

Figura 21 - Exemplo de histograma de termos para Portugal

Por fim, foi necessário alterar o código que trata do *collaborative filtering:*

```
## THIS CODE BELOW IS PURPOSELY WRONG. CORRECT IT - Corrected
example <- c(83, 269, 605, 103, 54)
for (n in 1:5) {
  #query_index <- sample(nrow(wine_pivot_matrix), 1)
  query_index <- example[n]
  #indices <- knn_model$nn.idx(wine_pivot[query_index,], k = 6)
  #distances <- knn_model$nn.dists(wine_pivot[query_index,], k = 6)
  indices <- knn_model$nn.idx[query_index, 1:6]
  distances <- knn_model$nn.dists[query_index, 1:6]
  print(paste0("Recommendation for ## ", rownames(wine_pivot)[query_index], " ##:"))
  for (i in 2:6) {
    print(paste0(i-1, ": ", rownames(wine_pivot)[indices[i]], " with distance: ", distances[i]))
  }
  cat("\n")
}
```

Figura 22 - Alteração knn para collaborative filtering

Grupo 4:
Nuno Gomes Nº 18364
Rafael Carvalho Nº 47663
Ricardo Ramos Nº 46638

Obtemos então o resultado:

```
[1] "Recommendation for ## Çalkarası ##:"
[1] "1: Malagouzia-Chardonnay with distance: 121.622366364086"
[1] "2: Merlot-Shiraz with distance: 121.622366364086"
[1] "3: Verdil with distance: 121.622366364086"
[1] "4: Loureiro-Arinto with distance: 121.622366364086"
[1] "5: Kuntra with distance: 121.622366364086"

[1] "Recommendation for ## Marsanne ##:"
[1] "1: Grenache-Mourvèdre with distance: 125.976188226188"
[1] "2: Cabernet Sauvignon-Sangiovese with distance: 127.526467840994"
[1] "3: Semillon-Chardonnay with distance: 129.417927660738"
[1] "4: Counoise with distance: 129.417927660738"
[1] "5: Souzao with distance: 129.464280788177"

[1] "Recommendation for ## Weissburgunder ##:"
[1] "1: Austrian Red Blend with distance: 309.867713710222"
[1] "2: Grauburgunder with distance: 310.436789056967"
[1] "3: Zweigelt with distance: 321.339384452015"
[1] "4: Scheurebe with distance: 322.436040169209"
[1] "5: Pinot Blanc with distance: 332.63944444398"

[1] "Recommendation for ## Cerceal ##:"
[1] "1: Baga-Touriga Nacional with distance: 91.0439454329611"
[1] "2: Maria Gomes with distance: 125.900754564856"
[1] "3: Baga with distance: 126.039676292825"
[1] "4: Cercial with distance: 127.291005181042"
[1] "5: Maria Gomes-Bical with distance: 127.322425361756"

[1] "Recommendation for ## Bombino Nero ##:"
[1] "1: Greco Bianco with distance: 0"
[1] "2: Bombino Nero with distance: 0"
[1] "3: Piedirosso with distance: 0"
[1] "4: Tintilia  with distance: 0"
[1] "5: Magliocco with distance: 1"
```

Figura 23 - Recomendações para cada user

Considerando apenas vinhos “baratos”, isto é, abaixo dos 100 dólares, obtemos:

```
[1] "Recommendation for ## Carcajolu ##:"
[1] "1: Carcajolu with distance: 0"
[1] "2: Jacquère with distance: 1"
[1] "3: Poulsard with distance: 1"
[1] "4: Merlot-Grenache with distance: 1"
[1] "5: Altesse with distance: 2"

[1] "Recommendation for ## Marzemino ##:"
[1] "1: Marzemino with distance: 0"
[1] "2: Nosiola with distance: 1"
[1] "3: Groppello with distance: 1"
[1] "4: Rebo with distance: 1"
[1] "5: Verduzzo Friulano  with distance: 2"

[1] "Recommendation for ## Zinfandel ##:"
[1] "1: Viognier-Marsanne with distance: 88.1192374002408"
[1] "2: Muscat Canelli with distance: 88.2156448709638"
[1] "3: Mourvèdre-Syrah with distance: 88.4590300647707"
[1] "4: Arneis with distance: 124.450793488832"
[1] "5: Sagrantino with distance: 124.458828533777"

[1] "Recommendation for ## Chambourcin ##:"
[1] "1: Norton with distance: 1"
[1] "2: Touriga with distance: 3"
[1] "3: Prieto Picudo with distance: 122.331516789419"
[1] "4: Gros Plant with distance: 122.331516789419"
[1] "5: Jacquez with distance: 122.331516789419"

[1] "Recommendation for ## Bovale ##:"
[1] "1: Insolia with distance: 0"
[1] "2: Nuragus with distance: 0"
[1] "3: Bovale with distance: 0"
[1] "4: Grecanico with distance: 1"
[1] "5: Cannonau with distance: 1"
```

Figura 24 - Recomendações para cada user, vinhos baratos