Mineração de Dados em Larga Escala

Aula prática 4

# Uma imagem com texto, captura de ecrã, Tipo de letra, número Descrição gerada automaticamenteComentário dos resultados obtidos de *resumetable(df):*

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.

# Descrição estatística dos dados numéricos

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

* 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.

# Uma imagem com diagrama, captura de ecrã, texto, Gráfico Descrição gerada automaticamenteInterpretação da distribuição de *Points*.

Figura 1- Histograma de pontuação

Do gráfico apresentado na figura 1, 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,

|  |  |
| --- | --- |
|  | Figura 2 - Distribuição categorias |

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.

# Cálculo dos *outliers*.



Figura 3 - 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.

# Distribuição de preços



Figura 4 - Distribuição logarítmica do preço

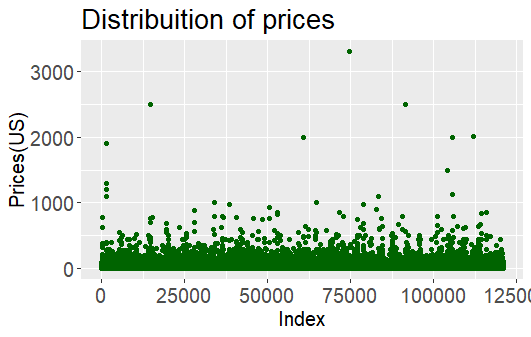


Figura 5 - Distribuição preços

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.

# *Uma imagem com texto, Tipo de letra, captura de ecrã Descrição gerada automaticamenteOutliers* preços

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.

# Gama de valores de vinhos abaixo de 300 dólares

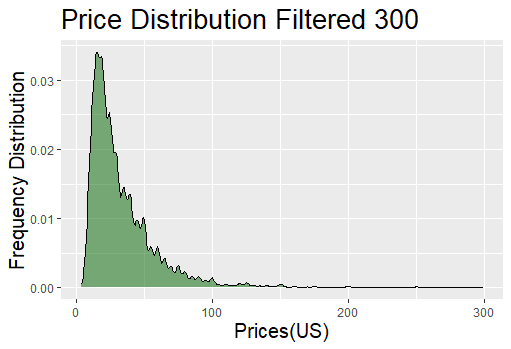
Sugeriu-se a validação da distribuição de preços abaixo dos 300 dólares:

Figura 6 - 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.

# 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.

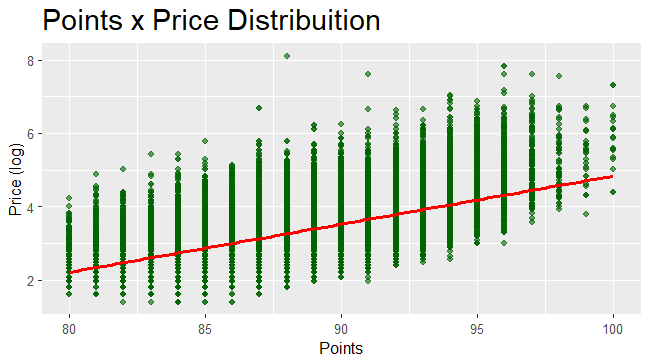


Figura 7- Pontuação vs preço

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*.

# Comparação dos países de origem dos vinhos

Da observação dos seguintes gráficos concluimos:

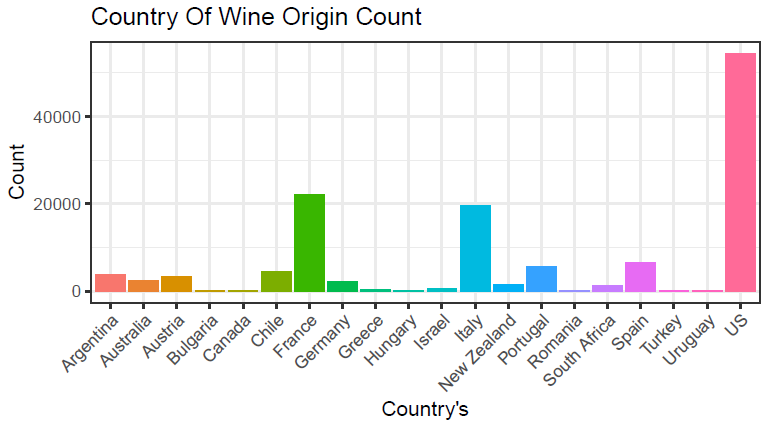


Figura 8 - Contagem do país de origem

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:

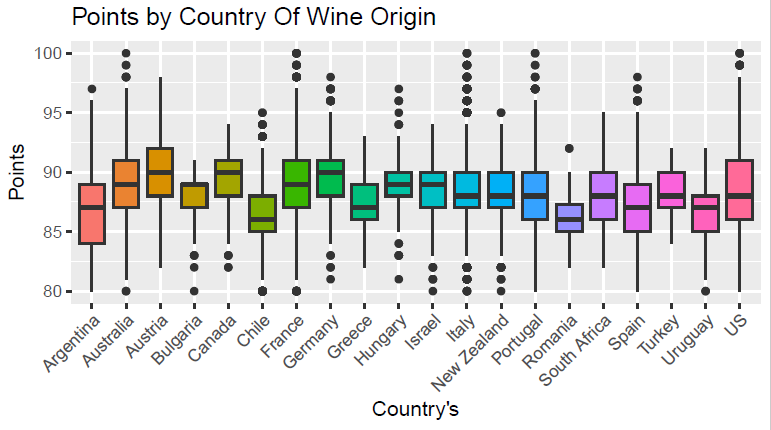


Figura 9 - Boxplot país de origem

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.

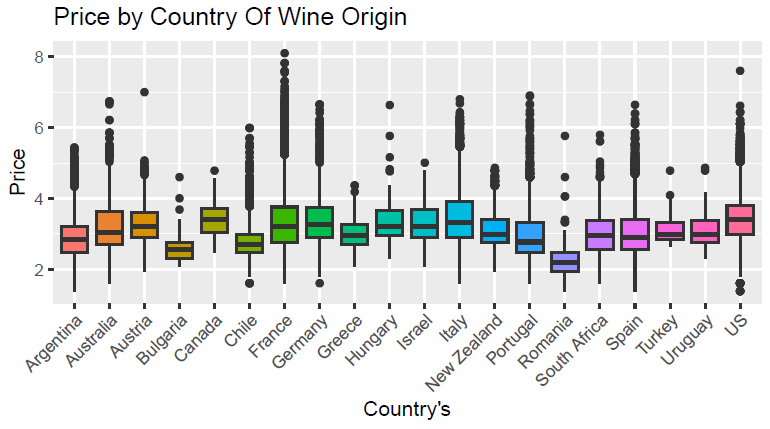


Figura 10 - Boxplot do preço por país de origem

# 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.

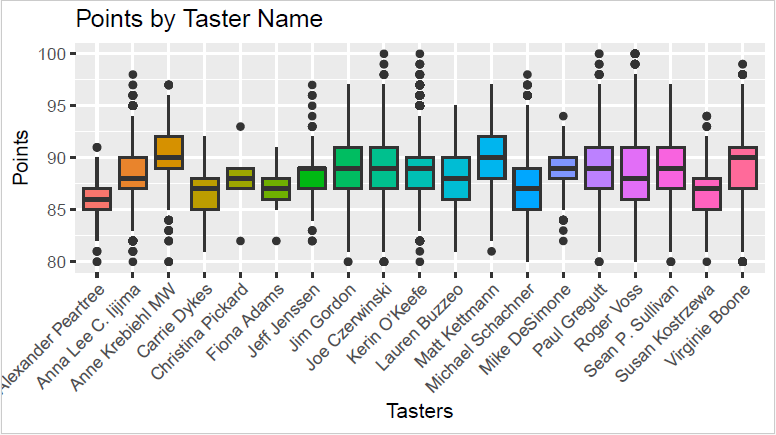
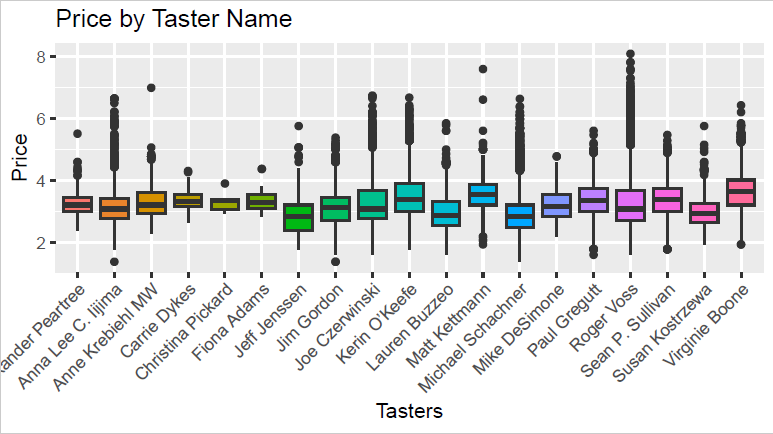
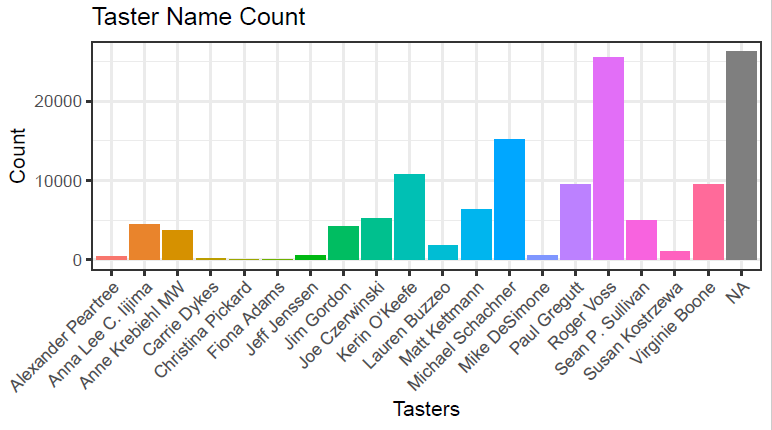


Figura 11 - Contagem e boxplots de preço e pontuação por taster name

# 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.

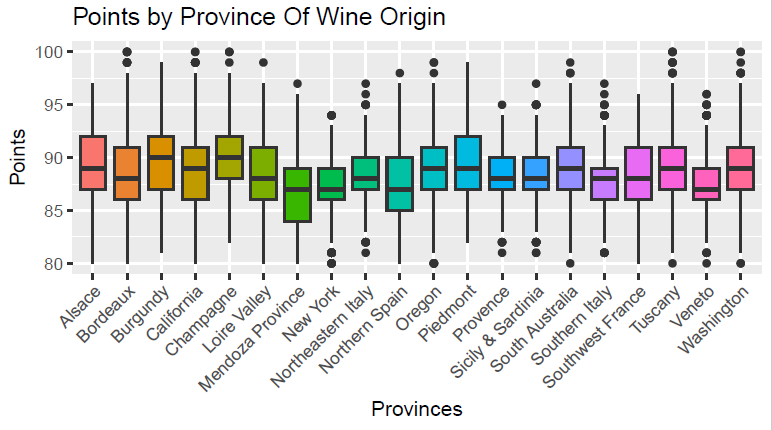
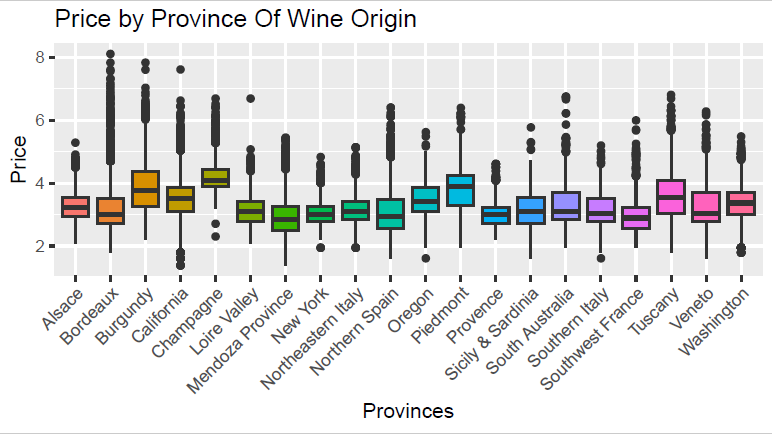
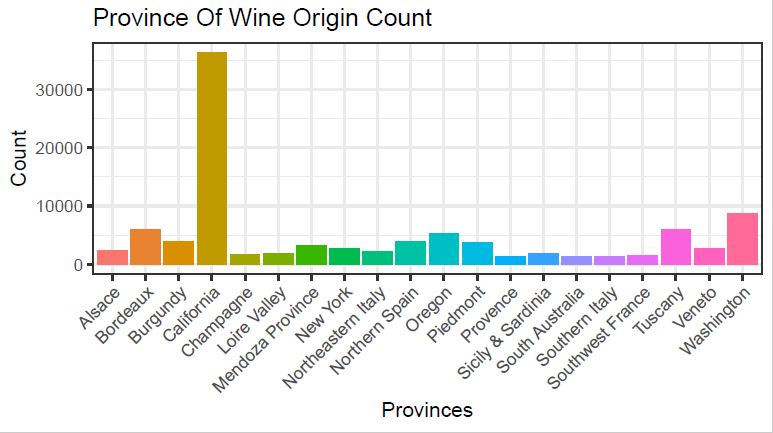
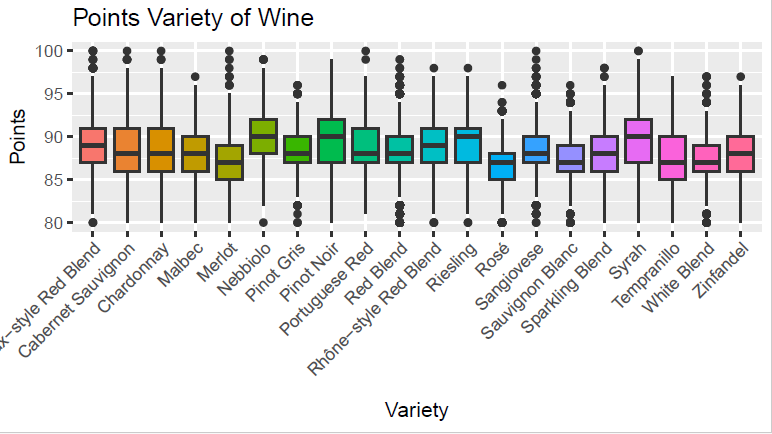
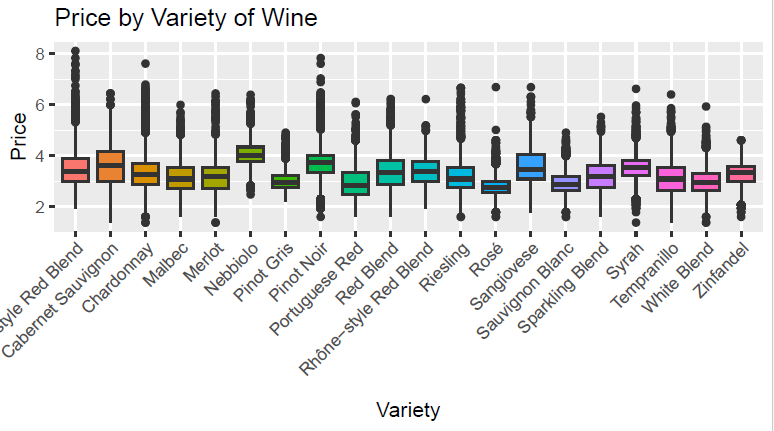
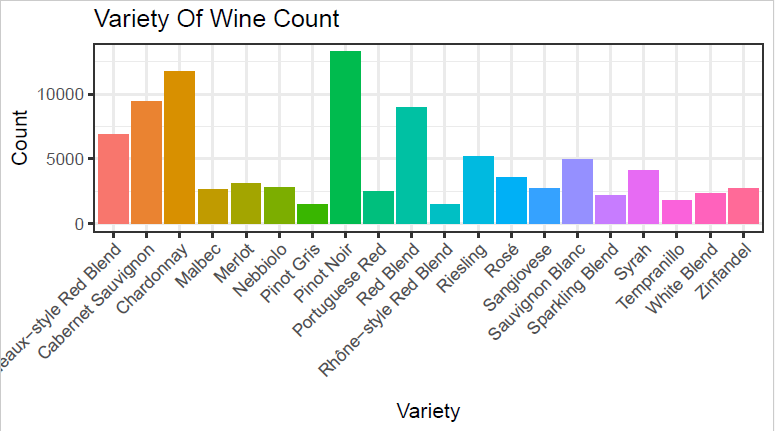


Figura 12 - Contagem e boxplots de preço e pontuação por província de origem

# 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 13 - Contagem e boxplots de preço e pontuação por variedade



# 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.

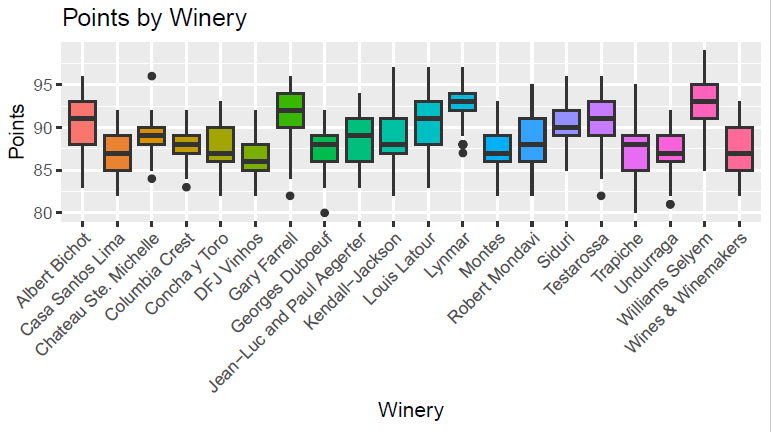
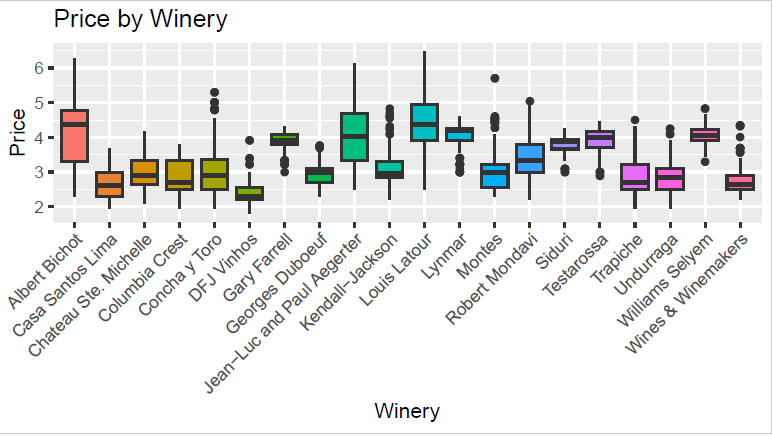
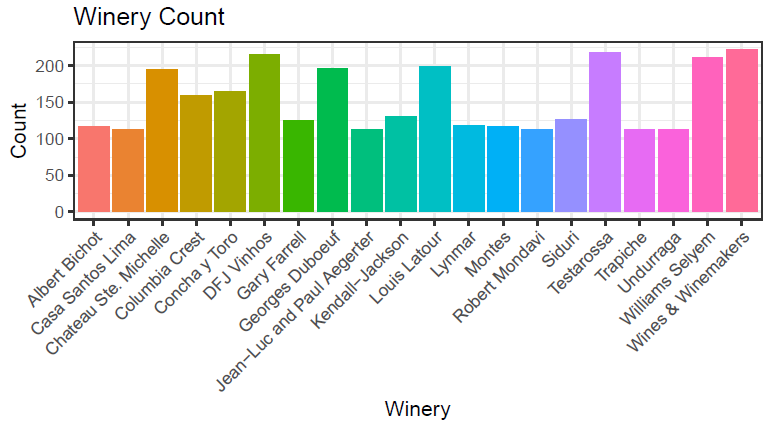


Figura 14 - Contagem e boxplots de preço e pontuação por adega

# 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 15 - 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 16 - Wordcloud de titles

# 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 rui*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:**

**Depois:**

Figura 17 - Alteração get\_top\_terms

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

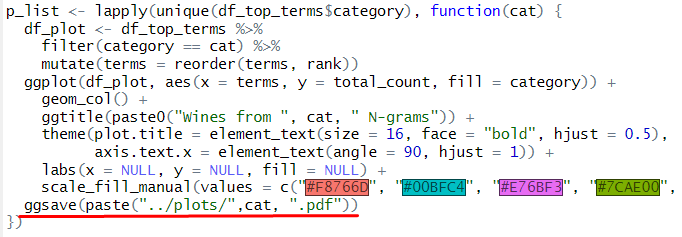


Figura 18 - Alteração para guardar histogramas de cada category

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

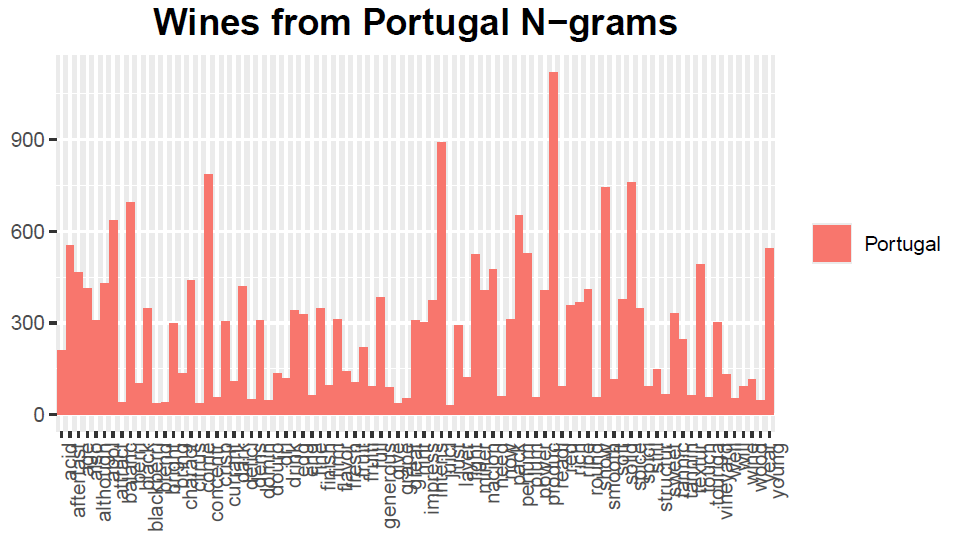


Figura 19 - Exemplo de histograma de termos para Portugal

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

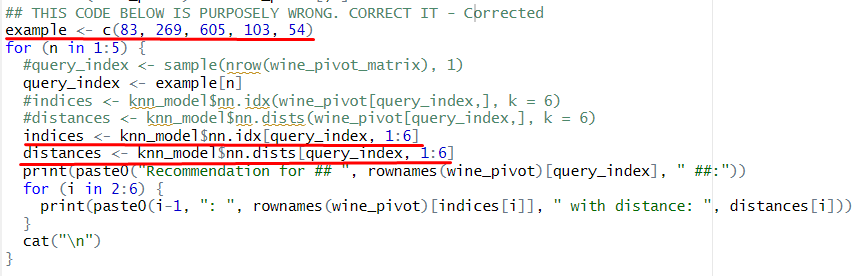


Figura 20 - Alteração knn para collaborative filtering

Obtemos então o resultado:

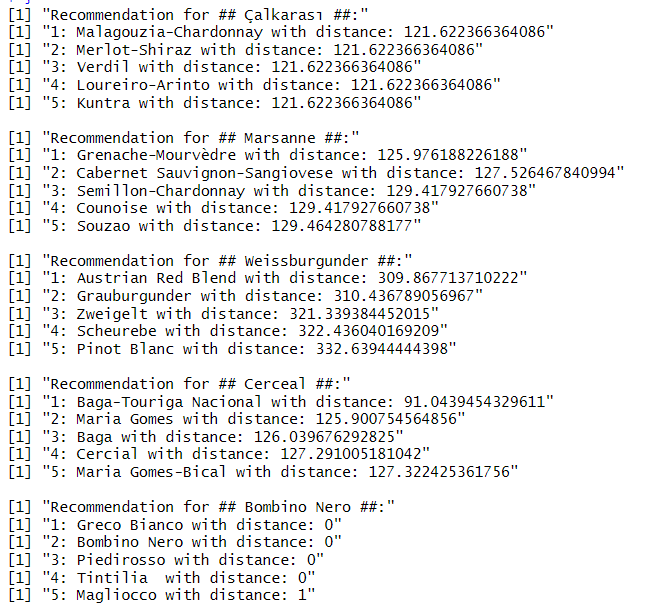


Figura 21 - Recomendações para cada user

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

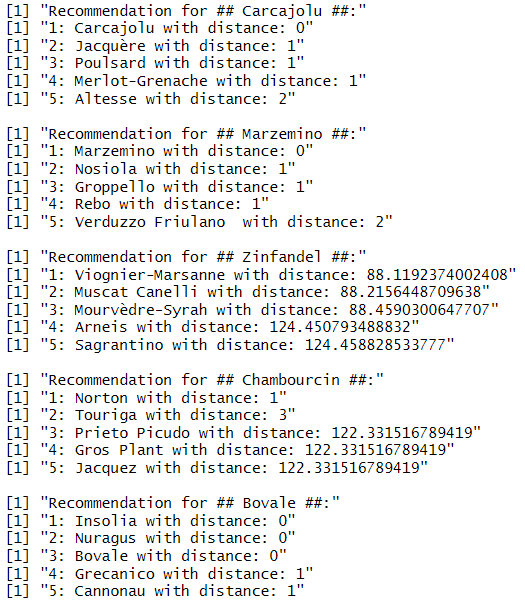


Figura 22 - Recomendações para cada user, vinhos baratos