Mineração de Dados em Larga Escala

Aula prática 4

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

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

Descrição gerada automaticamente

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 construida, 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) atribuida pelo provador (taster\_name) identificado pelo nome e conta de twitter (taster\_twitter\_handle).

Quanto ao vinho, este é caracterizado pela provincia (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, 42.02. 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.

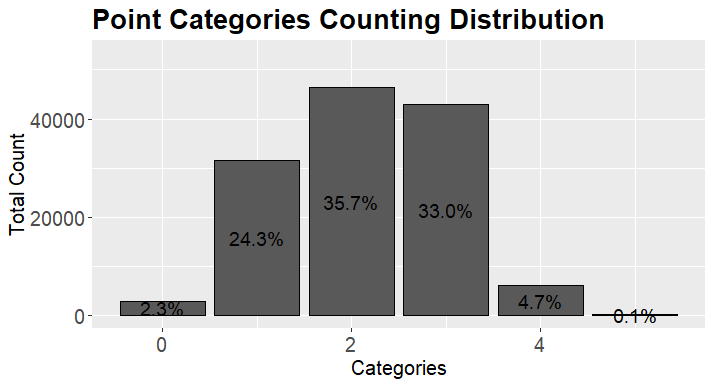
# Interpretação da distribuição de *Points*.

Uma imagem com diagrama, captura de ecrã, texto, Gráfico

Descrição gerada automaticamente

Da figura acima, podemos afirmar que a pontuação com maior representação neste conjunto de dados acontece aproximadamente no intervalo de [87, 93]. Do gráfico 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.

Ao aplicarmos um agrupamento por categorias dada a pontuação dos vinhos, seguindo:



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.

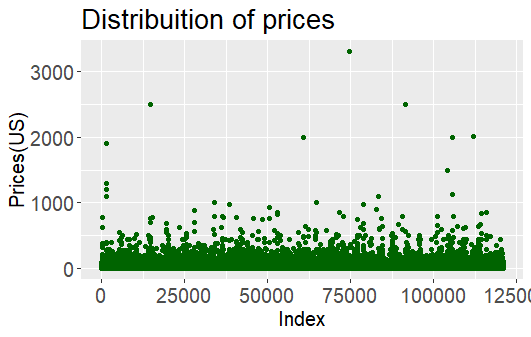
# Uma imagem com texto, Tipo de letra, captura de ecrã, número Descrição gerada automaticamenteCálculo dos *outliers*.

O cálculo dos *outliers* foi feito utilizando a função *CalcOutliers*, onde estes valores de corte são identificados 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 anterior.

# Distribuição de preços

Uma imagem com texto, captura de ecrã, diagrama, Gráfico

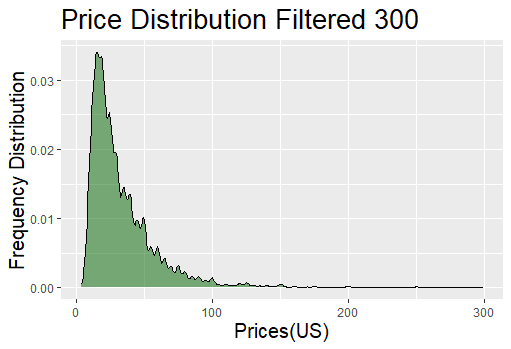
Descrição gerada automaticamente

A distribuição de preços mostra 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, um gráfico logarítmico, com os valores discretizados em 30 *bins* permite uma melhor compreensão da distribuição dos valores, onde verificamos 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

Para determinar os *outliers* dos preços 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 em todo o conjunto de dados. Compreende-se este resultado uma vez que o desvio padrão desta variável é elevado (42.02), e a média dos valores é muito afastada do valor máximo (35.36 até 3000). Relembrando que o valor mínimo dos preços é 4, a existência de *outliers* inferior seria muito pouco provável, 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:

Da observação deste gráfico, podemos comprovar que a grande maioria dos vinhos se situa entre os 0 e 50 dólares, contribuindo entre 1 até 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.