InfoVis final report 2018/2019

Inês Vilhena 84593
Henrique Almeida 84725
Tomás Oliveira 84773

Grupo 3, IST Taguspark

ABSTRACT
O nosso projeto consiste em mostrar interativamente os dados de um *dataset* sobre vinhos. O *dataset* tem aproximadamente 130000 observações e o objetivo é permitir ao utilizador explorar os dados à sua vontade, controlando os anos, o país, a adega, e vendo os padrões. Os dados foram retirados do site WineEnthusiast e o conjunto de dados está disponível no site Kaggle.com.

Author Keywords
Wine; dataset; big data; infovis; college; project; visualization; idioms; interactivity; information.

ACM Classification Keywords
Visualization: Visualization application domains: Information visualization.

INTRODUCTION
No âmbito da disciplina de Visualização de Informação do Instituto Superior Técnico o nosso grupo escolheu fazer um projeto para explorar informação sobre vinhos. Sendo uma área na qual poucas pessoas são conhecedoras, queremos que a nossa visualização permita a toda a gente escolher um bom vinho. Para além disso, tendo como base informação recolhida do website WineEnthusiasts, o nosso objetivo é descobrir padrões e relações nos dados que não seriam óbvios numa abordagem mais superficial, como por exemplo se um determinado especialista tem preferência por certos países, ou se uma adega específica produz vinhos demasiado caros para a qualidade dos mesmos. As nossas tarefas principais foram especificamente as seguintes:

- comparar os vinhos de acordo com o preço e pontuação;
- identificar os melhores vinhos;
- computar a pontuação média dos vinhos de cada adega e de cada país;
- analisar as pontuações dadas por cada especialista;
- comparar países de acordo com os seus melhores vinhos e pontuação média de vinhos por ano.

Podemos ver que o nosso projeto é mais direcionado para o utilizador que quer comparar e identificar vinhos, possivelmente para decidir qual o melhor a adquirir para uma determinada situação. Alguns exemplos de perguntas que poderá ver respondidas são:

- Que adega faz o vinho com o preço mais razoável para a qualidade?
- Há alguma relação entre o preço e a pontuação de um vinho?
- Como é que o preço e a pontuação de certos vinhos mudam ao longo dos anos?
- Em que ano(s) foram feitos os melhores vinhos?
- Qual é a adega com melhor pontuação?
- Os críticos têm alguma parcialidade em relação a certos tipos de vinho ou países?
- Que países fazem os melhores vinhos?
- Quais são os países com os melhores vinhos abaixo de um certo valor?

São estas e outras perguntas que procuramos responder com este projeto de uma forma interativa e agradável de utilizar.

RELATED WORK
Para nos inspirarmos para o nosso trabalho fomos consultar os projetos já feitos de anos anteriores e as aulas teóricas dadas da disciplina. Todas as visualizações consultadas são sobre assuntos muito distintos do nosso e das quais não pudemos retirar ideias diretamente. Contudo retirámos ideias de idiomas que poderíamos adaptar para o nosso próprio projeto e ideias de layout gerais, como por exemplo o layout do projeto de 2016/2017 “Global Peace Index”, onde vimos o *scatterplot, line chart* e mapa que posteriormente utilizámos.

***Websites* com objetivos semelhantes**
Existem *websites* cujo objetivo é similar ao nosso mas alcançado de forma mais complexa e detalhada. Três bons exemplos são

- http://www.snooth.com/
- https://www.wine-searcher.com/
- https://www.winesdirect.com/

O nosso projeto não utilizou um conjunto de dados tão extenso nem tem tantas páginas diferentes: o nosso objetivo era, em uma única visualização, o utilizador conseguir explorar diferentes dados com diferentes filtros de maneira fácil e intuitiva, e achámos que isso ainda não existia.

THE DATA
Os nossos dados foram encontrados no site Kaggle.com: https://www.kaggle.com/zynicide/wine-reviews/home, sendo este *dataset* uma compilação de dados retirados do site Wine Enthusiast (https://www.wineenthusiast.com). O conjunto de dados em estado bruto não estava completamente de acordo com o que procurávamos: havia

por exemplo informação que não foi utilizada e dados que tiveram de ser calculados através de outros. Assim, começámos por limpar e eliminar o desnecessário: eliminámos o atributo da conta do *Twitter* dos especialistas (*tasters*, críticos) e a region_1 e region_2 (regiões ainda mais específicas de onde o vinho é originário), pois no caso destes últimos atributos muito poucos vinhos apresentavam essa informação.
De seguida tratámos os espaços em branco substituindo-os por "null", "-1" ou "unknown" para poderem ser processados e não causarem erros na visualização.
Uma informação que achámos útil mas que não vinha originalmente foi o ano do vinho: como não existia o atributo foi necessário encontrá-lo no título da *review* de cada vinho para as 130000 *reviews* do dataset. Depois de isolarmos o ano descartámos o restante texto.

- para o gráfico horizontal com as diferentes formas criámos um ficheiro com o tipo de vinho, o ano, o preço e a pontuação
- para o choropleth foi necessário um *dataset* com o ID do país, o nome do país e a pontuação média dos seus vinhos.

entre outros.
Numa visão geral o conjunto de dados é muito completo e fornece muita informação, só faltou a informação do ano do vinho que teve de ser manualmente encontrada.

VISUALIZATION

Overall Description

A nossa solução é uma visualização com um total de 6 idiomas que têm algum tipo de interação entre eles.

Figure 1 - Layout geral com que a visualização começa.

Embora tenhamos dito no Checkpoint II que íamos agrupar as diferentes variedades de vinho, conseguimos que no idioma *Horizontal Shapes Chart* tivéssemos um mecanismo de filtração e seleção das mesmas, pelo que não foi preciso fazer essa agregação em todo o projeto (foi necessário no *Network Chart*, onde apenas pudemos representar as 30 variedades mais populares).
Tivemos contudo de criar algumas medidas derivadas: por exemplo, tendo em conta a adega, a pontuação média dos seus vinhos ou a pontuação máxima dos seus vinhos para cada ano. Todas estes valores médios, máximos e afins foram criados em diferentes ficheiros para serem utilizados por cada idioma:

- um conjunto de dados com os países, anos e pontuações máximas (para o line chart)
- um conjunto de dados para os vinhos abaixo de 20€ e a respetiva adega e pontuação (bubble chart)
- dataset idêntico mas para os vinhos abaixo de 15€ (bubble chart)

No canto superior direito temos um *line chart* com possibilidade de escolher dois países: clicando num dos botões que diz "Select Country" aparece esta página:

Figure 2 - Mapa choropleth com o rato por cima de um país.

Através deste Choropleth conseguimos desde logo ver quais os países com a pontuação média dos seus vinhos mais alta: quanto mais escura for a cor de um país, mais alta é a

pontuação. Por exemplo França tem uma pontuação média mais alta que Portugal e Espanha, no entanto mais baixa que Inglaterra. Os países que não têm cor não existem no nosso *dataset* e como tal não estão representados no idioma.

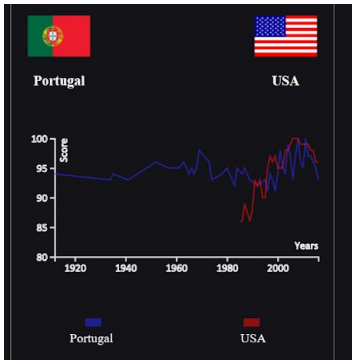


Figure 3 - Line chart de comparação de dois países.

O *line chart* compara ao longo dos anos a pontuação máxima dos dois países selecionados, como pode ser visto aqui. Não temos indicação dos anos nem opção para os escolher, é um indicador mais geral do como um país tem vindo a evoluir ao longo do tempo avaliando o seu melhor vinho em cada ano.

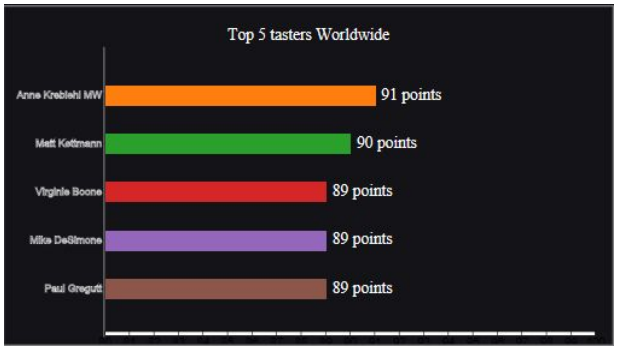


Figure 4 - Horizontal bar chart

O *horizontal bar chart* ilustra os 5 *tasters* que atribuíram a melhor pontuação tendo em conta o país de origem do vinho, ou então worldwide (média global das pontuações atribuídas a todos os vinhos). Por exemplo, no que diz respeito à pontuação global, podemos verificar que os 5 *tasters* que atribuíram a pontuação mais elevada têm uma classificação média semelhante.

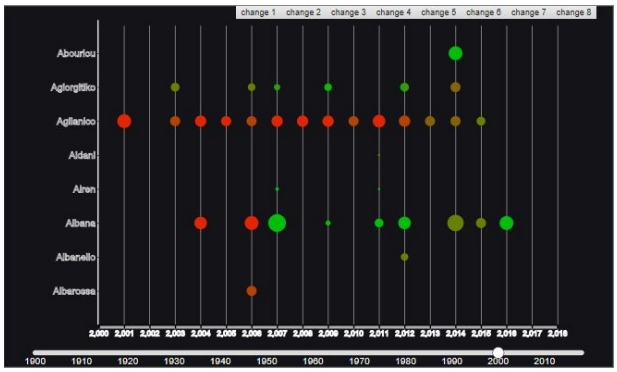


Figure 5 - Horizontal shape chart

Através do idioma acima representado comparamos as várias variedades de vinho existentes no *dataset*. Podemos constatar a evolução do preço e da pontuação de 8 variedades (selecionáveis) ao longo dos anos (ajustando o intervalo de tempo pretendido por meio de um *slider*). Uma variedade num certo ano é representada por uma *bubble* que possui uma cor e um tamanho. Quanto maior for a *bubble* mais elevada a pontuação atribuída à variedade num certo ano. O preço da variedade num dado ano é representado usando 6 cores: cores próximas do verde representam preços mais diminutos, enquanto que cores na vizinhança do vermelho refletem preços mais elevados. Para classificar em 6 cores discretizámos os valores com base nos vários preços obtendo o 6-quantil de todos os preços: [0, 16], [16, 19.62], [19.62, 23.9], [23.9, 30], [30, 40], [40, max]. Na figura acima podemos, por exemplo, verificar que a variedade *Aglianico* tem uma subida no preço ao longo dos anos e, ao mesmo tempo, vê a sua pontuação diminuída.

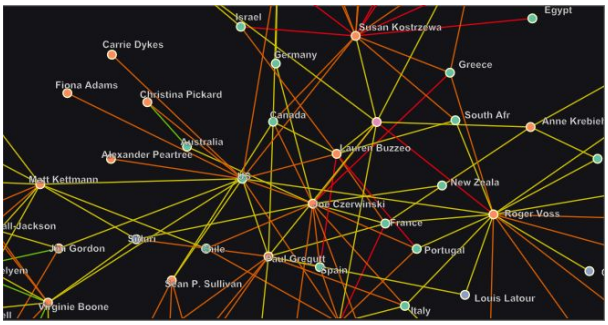


Figure 6 - Network chart

Com o intuito de representar possíveis parcialidades dos críticos utilizámos uma *Network chart*. Neste idioma existem 3 tipos de elementos: *tasters,* países e adegas. Cada um destes elementos está necessariamente ligado a pelo menos um outro. Esta ligação tem uma cor que representa a pontuação dada por um elemento *taster* a um país ou a uma certa adega. Quanto mais perto do verde estiver a cor da ligação entre dois elementos melhor a pontuação atribuída por parte de um nó *taster* a um nó país ou adega. Ao fazer *mouseover* sobre um certo nó todos os elementos, nós e

ligações em que este não está contido são escurecidos, salientando assim as ligações de que este nó faz parte. Por outro lado, se fizermos um *click* sobre um nó país alteramos os dados que estão a ser representados no idioma da figura 4, o *Horizontal bar chart*, mudando o país retratado no que ao *top 5* de avaliadores diz respeito. Aquando do *click* também é mudado o domínio do *Scatterplot* (fig. 11) passando este a representar a relação entre o preço e pontuação do país do nó que foi *clickado*.

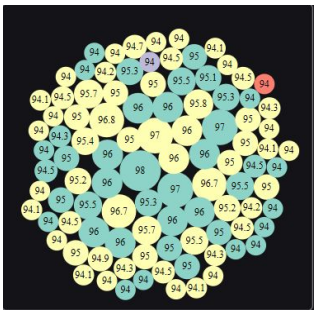


Figure 7

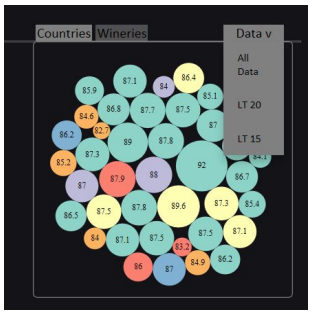


Figure 8

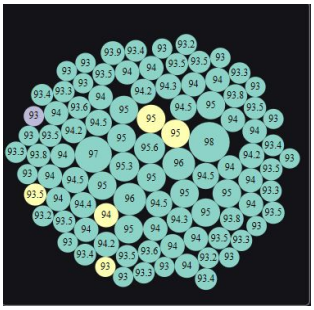


Figure 9

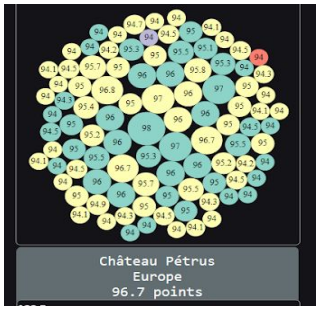


Figure 10

Usando os vários *bubble chart* acima representados conseguimos mostrar as pontuações atribuídas aos diferentes vinhos tendo em conta dois critérios: país de origem e adega. Quanto ao país de origem, patente na Figura 8, comparamos os vários países com base na classificação média (tamanho da *bubble*) e representamos também o continente do país em questão (cor da *bubble*). Quanto à adega, patente na Figura 7, comparamos as diferentes adegas com base na classificação média (tamanho da *bubble*) e representamos igualmente o continente através da cor da *bubble*. Ao fazer *mouseover* sobre uma *bubble*, neste caso um país, existe uma interação com o idioma *Horizontal Shapes*. Esta interação consiste no realce das *shapes* de variedades pertencentes ao país em que foi efetuado o *mouseover*. Ao fazer *mouseover* temos também acesso à informação escrita que aquela *bubble* transmite através de um *tooltip* que mostra, no caso da Figura 10, o nome da adega, do continente onde ela está situada e a pontuação média da mesma. Existem 3 visualizações que permitem observar diferentes subconjuntos dos dados das pontuações da adega: conjunto total, vinhos com preço inferior a 20 eur e 15 eur. A seleção da visualização pretendida é feita através de um menu, como se pode observar na Figura 8.

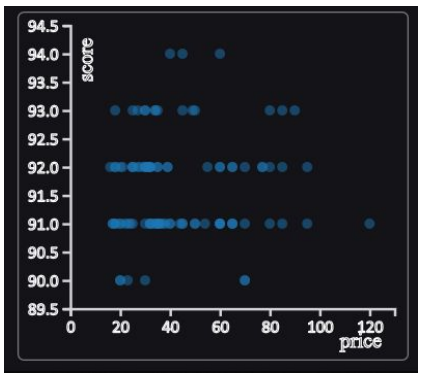


Figura 11: Scatterplot dos vinhos do Canadá.

O *Scatterplot* é o último dos nossos idiomas, localizado no canto inferior direito. É um gráfico que pretende mostrar a relação entre o preço e a pontuação dos vinhos de um determinado país, selecionado no *Bubble Chart*. Neste caso podemos observar que o Canadá tem alguns vinhos com um preço bastante elevado para a pontuação, mas também existem vinhos com uma pontuação excelente e mais em conta.

Rationale

Os idiomas utilizados foram os que considerámos ser os mais esteticamente apelativos de entre os mais úteis para o nosso projeto. Por exemplo o *Bubble Chart* apresenta de forma colorida quais os países com maior pontuação média e o *Network Chart* é uma maneira diferente e interativa de ligar os críticos aos países que eles avaliam. As várias ligações coloridas de um nó *taster* a vários países quando *highlighted* foi a melhor maneira encontrada de mostrar este possível *bias* pois mostram todas as avaliações do *taster* em simultâneo. Acreditamos que todas as técnicas utilizadas, mais ou menos interativas, cumprem o objetivo e transmitem a informação pretendida ao utilizador.

Utilizámos alguns visual encodings no projeto como um todo que permitem ver rapidamente algumas informações relevantes. Por exemplo as cores do *Horizontal Shapes Chart* estão numa escala de vermelho a verde (caro a barato) devido à percepção geral de que o vermelho representa perigo/valores indesejados e o verde tranquilidade/valores ótimos. A mesma escala é utilizada para o *Network Chart*, em que as linhas verdes indicam uma pontuação elevada e as linhas mais próximas do vermelho indicam pontuações menores.

Tal como as cores, também o tamanho das bolhas e das formas tem um significado: quanto maior o tamanho da forma maior a pontuação atribuída. Esta decisão acarreta problemas pois é difícil analisar com rigor a diferença de tamanho entre diferentes *bubbles*, por isso adicionámos

uma *tooltip* que indica a pontuação da associada à *bubble* em questão, isto quando é feito *mouseover* sobre esta.

Assim como as cores tentámos que esse significado fosse intuitivo e de fácil compreensão para o utilizador. O resto das cores utilizadas no projeto foram apenas escolhidas por questões estéticas.

O *Horizontal Shape Chart* foi sofrendo alterações à medida que o desenvolvimento da aplicação foi avançando. Numa primeira abordagem (*VIS Sketch*) escolhemos *visual encodings* ligeiramente diferentes dos que implementámos na versão final: cada uma das 8 variedades em comparação teria uma forma geométrica diferente e os preços das variedades num certo ano seriam representados pela altura relativa das formas dentro da linha correspondente à variedade em questão, sendo as cores usadas para representar o continente da variedade. Esta mudança deveu-se à complexidade acrescida em termos de programação.

Ao longo do tempo fomos também implementando algumas versões diferentes do *line chart* (como animações no *mouseover*, eixos com valores e possibilidade de escolha dos anos), e acabámos por decidir utilizar uma versão mais simples e menos confusa. No entanto sempre considerámos o *line chart* como um bom idioma para mostrar a progressão de um país ao longo dos anos, pois indica de forma clara as tendências do mesmo. Contudo por repetição de informação que seria apresentada não incluímos o segundo *line chart* que foi planeado no Checkpoint IV, que apresentava o vinho mais caro por cada país ao longo dos anos.

O *Horizontal Bar Chart,* o *Bubble Chart* e o *Scatterplot* não sofreram alterações ao longo do desenvolvimento do projeto: conseguimos implementar a ideia que tivemos desde o início. Um idioma que infelizmente não nos foi possível implementar por complexidade do mesmo foi um que se localizaria por baixo dos *line charts*, que comparava dois países, a mesma variedade de uvas em vários anos diferentes, e comparava através da pontuação média com uma escala de cores. Seria excessivamente complexo para apresentar algo que os outros idiomas existentes já fazem entre eles, por isso também não prosseguimos com essa ideia.

Demonstrate the potential

A nossa tarefa "Computar a pontuação média dos vinhos de cada adega e de cada país" é facilmente respondida com o *Bubble Chart*:

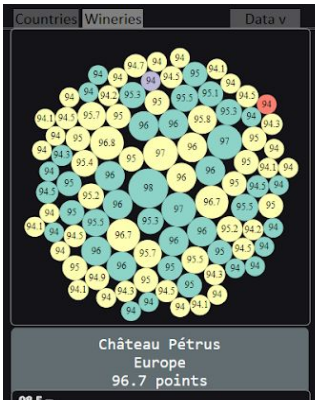


Figure 12 - Bubble Chart a mostrar a pontuação média por adega.

Temos apenas de selecionar "Wineries" e o utilizador consegue ver qual é a adega com melhor ou pior pontuação média (tamanho da bubble) e vê a informação da adega em baixo, num *tooltip*. Uma das conclusões que podemos retirar do *Bubble Chart* é a de que o melhor adega com vinho com preço médio abaixo dos 15$ está radicada nos EUA. Esta conclusão é de facto inesperada porque os EUA são talvez o país com mais poder de compra e em competição com, por exemplo, Portugal "ganham" no melhor vinho mais barato.

A questão "Comparar países de acordo com os seus melhores vinhos e pontuação média de vinhos por ano" é respondida num idioma, o *line chart*:

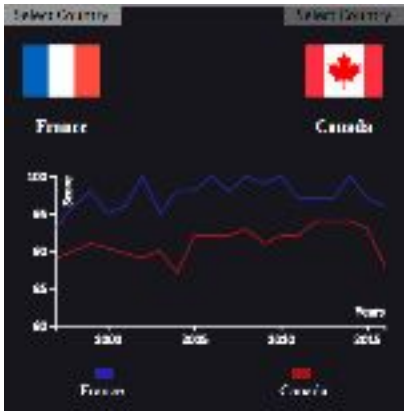


Figure 13 - Line Chart para a tarefa em questão.

Este permite-nos comparar os melhores vinhos de cada país ao longo dos anos. A comparação dos países por pontuação média por ano não é possível mostrar explicitamente, contudo acreditamos não ser a informação mais relevante para um utilizador que queira comprar um vinho. Para além

disso a comparação dos países e sua pontuação média está exibida no *Bubble Chart*.

A tarefa "Analisar as pontuações de cada *taster*" pode ser vista individualmente por país. Vindo ao *Network Chart*:

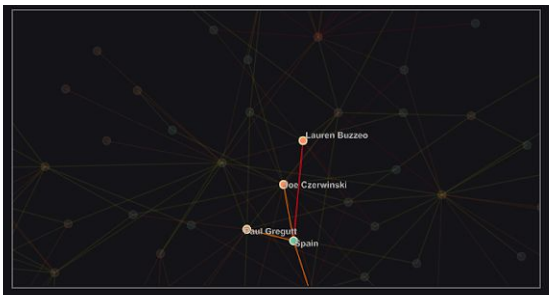
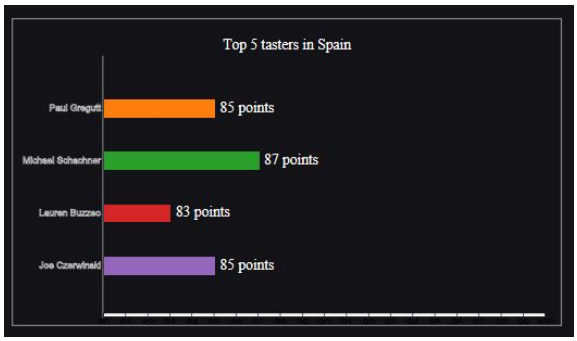
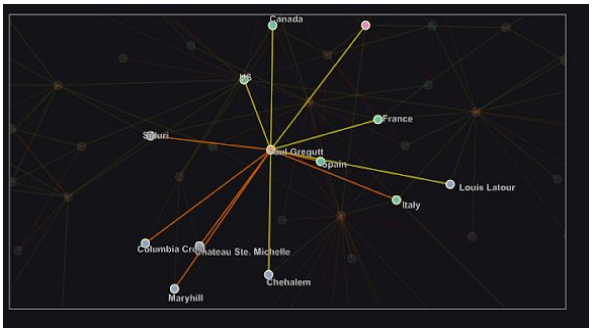


Figure 14 - Network Chart para a tarefa em questão.

Selecionamos por exemplo o país Espanha clicando nele (só o *mouseover* já permite ver os *tasters* que avaliam vinhos desse país), e depois de clicarmos no país vemos o *Horizontal Bar Chart*:



que mostra os mesmos especialistas que o *Network Chart*, mas com um valor numérico. Neste caso podemos ver que o *taster* Paul Gregutt atribuiu pontuações aos vinhos espanhóis que em média deu 85 pontos. Caso queiramos ver as pontuações de um só taster, passando o rato por cima de uma barra podemos ver esse mesmo taster highlighted:



IMPLEMENTATION DETAILS

Os maiores desafios que tivemos de ultrapassar foram a implementação do idioma *Horizontal shape chart* e a implementação de interação entre diferentes visualizações. As dificuldades da primeira tiveram que ver com o facto deste idioma ser implementado praticamente de raiz e de necessitar de mudanças de domínio quer no eixo das abcissas quer no eixo das ordenadas. No que às interações diz respeito encontramos particulares dificuldades na interação com a visualização *Bubble chart* isto porque este gráfico tem uma física associada e qualquer alteração no "ambiente" faz com que a visualização tome comportamentos arbitrários.

CONCLUSIONS AND FUTURE WORK

No desenvolvimento deste projeto aprendemos a importância do idioma certo para transmitir informação, pois a mesma informação em dois gráficos diferentes pode ser bem ou mal transmitida e recebida. Aprendemos também como a cor é relevante para passar eficazmente a ideia de uma tendência, como é o caso da cor das formas do *Horizontal Shape Chart* (como a escala transmite uma sensação de muito/pouco ou bom/mau).

Não respondemos a todas as questões a que nos propusemos, como é o caso da comparação dos países por pontuação média ao longo dos anos, mas acreditamos que as tarefas que foram respondidas são suficientes para se ter uma boa noção do conjunto de dados. Mais especificamente onde comprar os melhores vinhos, os sítios com os vinhos mais baratos, entre outras informações úteis e relevantes à possível compra de um vinho. µ

Caso tivéssemos mais 1 mês para concluir o trabalho e 3000$ para investir neste trabalho faríamos várias coisas:

- Fazer uma recolha da web de mais informações sobre cada variedade, vinho e adega que não constem do *dataset* inicial. Esta pesquisa consumiria consideráveis em termos de tempo sendo que para acelerar este processo podia ser utilizado parte dos fundos disponíveis para ter acesso a bases de dados que não estão disponíveis gratuitamente. Estes novos dados podiam permitir que novas associações fossem descobertas;
- Utilizar métodos de *supervised* e *unsupervised learning* para obter, por exemplo, regras de associação que, utilizando novos idiomas, pudessem ser visualizadas;
- Simplificar as visualizações que estão a ser utilizadas. Ao fazer uma reflexão em jeito de conclusão sobre o estado final do projeto, notámos uma redundância em certos idiomas, pelo que seria alocado mais tempo na simplificação da informação apresentada, agregando certos idiomas num só, de modo a tornar a informação mais explícita e de fácil leitura.