|  |  |
| --- | --- |
| Znalezione obrazy dla zapytania wine  World of winecraft  Differentiation of the wines across United States – final report of the project for Spatial Data Capture coursework | Joseph BARRAT  Yiting CHEN  Zijian DENG  Piotr KULAK |

University College London

London 2018

**1. Introduction**

**1.1 Statement of the problem and motivation for the project**

The purpose of the study was to offer information of the distribution of the high-quality wines produced in the United States and identify the value of those top wines like a master sommelier though blind tasting. The aim of this project is realised by an interactive map website developed by our group within Spatial Data Capture coursework. It can not only spatially visualise the top wines but also show the general information about the wines by state including their variety, winery, designation, region etc. The report also includes incisive graphics and information for wine related industry. Overall, this provides a snapshot of the U.S wine industry of both long and short-term history.

**1.2 Significance of the study**

The wine industry of the United States is “at the tail end of its largest growth stage in history” (McMillan et al, 2018). According to a series of statistics for the U.S wine market of recent years, the wine market is at a growth period as a result of the trends of demographic shift of the United States. Euromonitor International (2017) indicates that the ageing of the most loyal and target consumers of wine results in the differences for the wine industry and market. The significance of millennials and generation X is flourishing and the demand for high-quality and premium wines at an affordable price level is highlighted as their consumption concepts are different with older generation’s (Euromonitor International, 2017). Although the growth is decelerating in recent years, premiumization within wine segment, perceived as making not-so-common wines available in wider distribution and at more affordable prices, is still the leading trend across country. The study of better-quality wines is vital for the winery owners and related businesses to have an intuitive understanding of the top wines in the U.S in order to adapt to the changing consumer patterns. With the impact of imports on the American wine industry, it is helpful to provide overall information of local products for wine consumers through an interactive session.

In addition, the trends are that these younger age groups are proving to be quite enthusiastic consumers of wines without spending time trying to learn detailed traits of vintages and varieties which is too time-consuming and experience-based. This project offers adequate information for those enthusiastic consumers to access not only basic information of the U.S wines but also the description and traits collected from master qualified sommeliers and tasters.

**1.3 Scope of the Study**

This study was limited to the perceptions of a range of master sommeliers collected from WineEnthusiast (2018) website as to the reviews and descriptions of the high-quality wines of the U.S that are evaluated based on the price and points assigned to them. For the purpose of the study, the high-quality or top wines are defined as the wines that has points over 80 out of 100. This study was restricted to the all variety of wines in the U.S wineries including Red Blend, Montepulciano, Pinot Grigio, Riesling, Sangiovese etc. For the purpose of this project, what constitutes master sommeliers is defined by the WineEnthusiast (2018). Dataset for this study were scraped from Kaggle.com and it contains two datasets of wine reviews the WineEnthusiast during the November of 2017.

One of the challenges of this project is advanced interaction techniques as the website should allow users to access to tasting notes by selecting a state, a price range or value range. Secondly, it is necessary to visualise the interactive text of wine descriptions efficiently for users to getting detailed information of the wines. Finally, the website has to allow users to search wines using the key words in the reviews, which requires compatible techniques.

**1.4 Review of related projects**

Kerren, Kyusakova and Paradis (2012) mentioned a series of interactive visualisation techniques based on a number of wine reviews. They developed an interactive information visualisation tool to analyse the corpus of wine reviews or tasting notes from the Wine Advocate (Kerren et al, 2012). Users can compare many wines from their visual, olfactory, and gustatory features. This tool also allows users to find information of the varieties, vintages, and countries of the wines. Their research focused on text visualisation approaches and it can provide a good experience for those case studies of interactive visualisation of textual data. However, the tool their created paid more attention to the analysis of lexical, grammatical, and discursive patterns of the wine reviews. This tool focus on visualisation for linguistic exploration and is considered not suitable or intuitive for users who mainly aim to access information of wines.

**2. Aim of the project**

The aim of this project is to create a predictive model to identify wines like a master sommelier could through blind tasting. The website is designed as an interactive map to visualise and compare the type of wines produced within the United States. This interactive map will spatially visualise the top wineries of the United States by state. The first step was gathering data to train a model. Deep learning is utilised to predict the wine variety using key words in the description or review of wines. This model theoretically can provide information for users and identify the wine based on the description that a professional sommelier could give. It will also visualise the cluster of the high-quality wineries of the US which is likely to support the growth of wine sector. This will provide vital insight into a states’ capability to produce wine and quality at which it stands across the United States.

In general terms, the aim of the project is to make an interactive visualisation map of the U.S wines concise and easy to use. End-user can display wine reviews that pertain to state, can choose the price range and points range. The website also uses deep learning (sentiment analysis) to identify the wine variety using words in the wine reviews.

**3. Data collection, handling and management**

Data used in this project have been initially found on [www.kaggle.com](http://www.kaggle.com) which serves as platform for sharing datasets, educate about data science and is developed social community for scientific research base. The author of the dataset, scraped the wine reviews from WineEnthusiast portal which serves as a main portal for wine-lovers and is recognizable internationally. Datasets have been scraped from the website using Python script and the script is available at <https://github.com/zackthoutt/wine-deep-learning>.

Data management in our case meant first of all assessment of the data, its credibility, reliability and usability. Two separate datasets have been downloaded and merged through SQL, however, it turned out that after this operation there are many duplicate records. This led to next phase: data cleansing and enrichment. First task was to eliminate all the records that did not pertain to USA. From those, we deleted records where no state has been assigned or other crucial value was missing. We deleted also few records for wines that have been assigned as collaboration between two or more states. Newly created dataset needed to be cleaned from duplicates and other erroneous information.

There were simple duplicates where all the records have been the same.



Another type of duplicate was that one field was missing as a result of different structure of two merged datasets.



Another type of duplicate could be characterized as incomplete record and those had to be merged and corrected manually.   
  
  
  
There was also a problem with individual characters that do not normally occur in English and have its origin mainly in French, Spanish or German. They have occurred after merging the datasets and these were for instance: é, ù, **û,** ï, ü, ç etc. Those characters have been represented as odd letters like *A~* which had to be transferred into é or *£}* which had to be transferred into ï and so on. The biggest problem with this issue was that those misspelled letters had to be spotted manually and corrected individually. The great help here was just to paste a fragment of the review into internet browser to look for the actual wine and look for the faulty letter. However, it was long-time process which finally led to neat and tidy final dataset that could be processed further in order to produce some quality analysis. Overall, after cleansing the dataset of more than 120 thousand records, 69 thousand of valid records left.

Data contains following information: review of the wine given by professional taster, designation of the wine which is a name of the site where vineyard is located (geographic designation), points (in range between 80 and 100), price (given in American dollars), state where vineyard is located, more precise region within state, winery name and market wine name (not always given).

In general terms, data management for this dataset and project as a whole have been limited to operations within MySQL and Excel as data was not very complex and the data type did not require complicated operations in order to provide compatibility or stability.

Additional dataset that was used in the project was climatic data collected from <http://www.noaa.gov/>. The data was averaged on state level and applied to the construction of the correlation charts and data contained average humidity, average rainfall and average sunshine exposure.

**4. Data Analysis, methods, major findings**

The data involved in this project was identified as having four main variables with which three routes of analysis can be taken. The first variable, and key to this project, is the spatial element. The “province” variable allows for all other data analysis within this dataset to be allocated to a mapping framework for visualization purposes. The second and third variables are the reviewer’s score and the price of the wine. From these two variables it is possible to gain some insight into the quality of the wine, how it compares to others and how much it costs to purchase. The fourth and final variable from this dataset, is the review of the wine itself. Each review provides a description of the flavours perceived during the wine tasting. This data therefore allows for three spatial elements of analysis.

However, with the score and cost variables, there are a few things we can do to make the data more accessible, more readable and overall better for analysis and mapping. It is commonplace within the wine industry where more expensive bottles are associated with higher quality wines. However, this website will consider the general user. Thus, rather than having cost data mapped out, it seemed it would be more value to a wine enthusiast to have the value for money mapped out. This can be fairly easily achieved with indexing. The score for the wine is based on a scale from 1-100 (1 being the worst and 100 being the best). For the price to be on a comparable scale to this, it needs to be indexed. This can be achieved by indexing the cost of the wines. The cost of each bottle of wine will be divided by the highest price wine and multiplied by 100 as explained in the equation below.

Where:

Xn = Single value of a bottle of wine

Xm = Maximum value or most costly wine

I = Index Value

This isn’t perfect because the highest price wine will be at 100 and the desired trait is best value. Whilst some other factors may be considered in value for money, this report and website argues that a lower costing wine, with the same quality score as a higher costing wine, will have a greater value for money. Therefore, by inverting the index the lowest price wine will have the higher value. Using the same equation as above except subtracting the index value from 100.

The result is a higher value for the lower costing wines. This can then be averaged with the score to produce a value for money value. This value can then be averaged per state and used as a guide to what states offer the best value for money when it comes to high quality wine. Using SQL functions, key figures can be extrapolated from these variables, such as the average cost of wine per state, the average quality of wine per state as well as an overall value for money per state value, which can be achieved by averaging the value for money within a given state. These can then be spatially located on the website for wine enthusiasts to use

The remaining variable of this dataset will be used in a largely qualitative context rather than quantitative. To achieve this, a simple yet effective python script was used which would split the words in a review into individual items and then count the occurrence of each review. See the below example of script used for Iowa:

*from collections import Counter*

*review\_data* = " An inventive blend of Catawba and Marachel Foch a French-American hybrid between Goldriesling and an American vine called vitis riparia this rosé has a heavily perfumed nose and a slightly clumsy combination of sweetness and spicy red Still the wine has a fresh finish and some structure that recommends it to myriad food pairings " \

" Sweet butter and caramel notes linger amidst smoke and char on this unabashedly oaky full-bodied Iowa Seyval Blanc Savory pressed apple flavors are pleasant but subdued " \

" Anyone who doesn't like barnyard smells will want to steer clear of this funky example from Iowa The crisp palate shows some cherry notes and little tannin Finishes with some oily ashy notes " \

" This exotic white blend from Iowa offers aromas of orange and citrus with a minerally zip and on the palate is a bit on the sweet and curvy side Not terribly inspired but still pretty it will pair well with Asian cuisine " \

*split\_reviews = review\_data.split()*

*Counter = Counter(split\_reviews)*

*most\_occur = Counter.most\_common(50)*

*print(most\_occur)*

To achieve this, a separate dataset "ReviewData\_Subset.xls" had to be created with no one punctuation asides from " and " \ to show the start and end of a review. From here, it was possible to manually enter the review data into the python script and separate nouns and adverbs from the verbs pertaining to the flavour.

This study goes further, to extend our analysis further to predict whether favourable climatic conditions would produce better quality wine. According to Santos et. Al (2010) the productivity of yield is controlled by factors such as precipitation, humidity and solar radiation as well as a few others. In this vein of literature, it is reasonable to suggest that there may also be a link between these climatic factors and quality. For the purposes of this study, focus will solely be on the named factors. Using python, we produced a tool that would run simple linear regression against the climatic data and the wine quality data to produce an R Squared value that would be indicative of any relationship present between the datasets.

**4.1 The Results**

Figure 1 and table 1 shows the three key data variables used for this website and report. We can see that the states of Oregon, Rhode Island, and Washington break past the 92-point mark when it comes to value for money, placing in the top three for the average value for money. Rhode Island comes out on top with the highest average quality wines with low average costs. Washington and Oregon, place 2nd and 3rd, producing high quality wines, yet with higher costing wine. New York manages to also break the 92-point mark, but with lower quality wine and much cheaper prices. Whilst California is also producing good value for money wines, albeit slightly lower than the previously mentioned wines, the customer would be expected to pay the highest costs for a less superior product. Nevada and Kentucky are the two lowest ranking states in terms of value for money. Their wines are two of the costliest, falling below only California in terms of costs, and yet Nevada and Kentucky scored on average 84.25 and 83 on quality. Whilst these aren’t the lowest scores, States producing similar or worse quality wine, such as Iowa and Missouri, having significantly lower costs which in turn aid in bringing their value for money up in the rankings. In all the graph works well in telling the user where they might have to pay more for inferior products, or what they might have to compromise on to get a little more for their money. For instance, a user of this will now know that they could get a similar higher quality wine in Washington for a cheaper cost than in California. Alternatively, a user will know that in Iowa or Ohio, they won’t necessarily get great quality wine, but their money will go further.

|  |  |  |  |
| --- | --- | --- | --- |
| **State** | **Average Score** | **Average Value For Money Score** | **Avg Price ($)** |
| Arizona | 84.88 | 90.49 | 29.2 |
| California | 88.63 | 91.72 | 38.99 |
| Colorado | 86.12 | 90.86 | 32.99 |
| Connecticut | 84 | 90.98 | 15.33 |
| Hawaii | 87 | 91.63 | 28 |
| Idaho | 86.61 | 91.92 | 20.77 |
| Illinois | 84.5 | 90.09 | 32.33 |
| Iowa | 83 | 90.47 | 15.5 |
| Kentucky | 83 | 89.17 | 35 |
| Massachusetts | 84.86 | 91.01 | 21.29 |
| Michigan | 86.15 | 90.91 | 32.43 |
| Missouri | 83.33 | 90.51 | 17.33 |
| Nevada | 84.25 | 89.61 | 37.75 |
| New Jersey | 84 | 90.3 | 25.5 |
| New Mexico | 83.96 | 90.48 | 22.42 |
| New York | 87.18 | 92.07 | 22.83 |
| North Carolina | 84.57 | 90.88 | 21 |
| Oregon | 89.05 | 92.09 | 36.55 |
| Ohio | 84.08 | 91.13 | 13.75 |
| Pennsylvania | 84.56 | 90.72 | 23.33 |
| Rhode Island | 90 | 93.33 | 25 |
| Texas | 85.22 | 90.9 | 25.6 |
| Vermont | 84.33 | 90.06 | 31.67 |
| Virginia | 85.62 | 91.01 | 27.01 |
| Washington | 88.95 | 92.31 | 32.44 |

Table 1. Average value for money score for each state

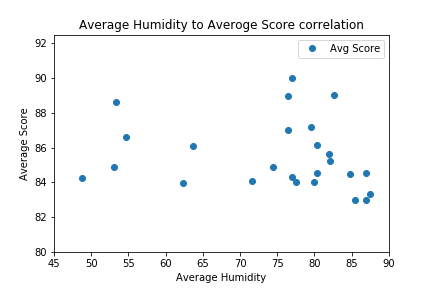
The results have also been broken down into top quality wine per state and best value for money by state. This can be seen in table 2, which show the user what winery is producing the best bottle of wine and which is producing the best value for money wine.

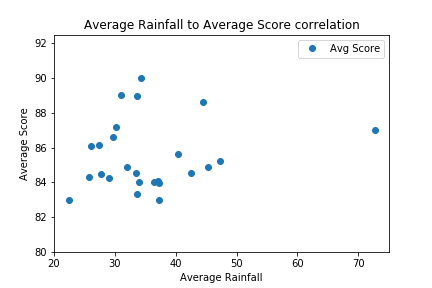
|  |  |  |
| --- | --- | --- |
| **State** | **Top Scoring Winery** | **Best Value Winery** |
| **Arizona** | Page Springs (88) | Page Springs (92.43) |
| **California** | Vérité/Cardinale (100) | Failla (96.96) |
| **Colorado** | Creekside Cellars (91) | The Infinite Monkey Theorem (93.33) |
| **Connecticut** | Hopkins Vineyard (84) | Hopkins Vineyard (91.26) |
| **Hawaii** | Ulupalakua (87) | Ulupalakua (91.63) |
| **Idaho** | Koenig Vineyards/Cinder (91) | Colter's Creek (94.33) |
| **Illinois** | Illinois Sparkling Co. (86) | Illinois Sparkling Co. (91.26) |
| **Iowa** | Jasper (84) | Jasper (91.13) |
| **Kentucky** | Elk Creek Vineyards (83) | Elk Creek Vineyards (89.16) |
| **Massachusetts** | Still River (89) | Still River (92.83) |
| **Michigan** | Bel Lago (91) | Chateau Grand Traverse (94.13) |
| **Missouri** | Mount Pleasant (86) | Les Bourgeois (91.96) |
| **Nevada** | Nevada Ridge (85) | Nevada Ridge (91.03) |
| **New Jersey** | Heritage/Unionville (87) | Heritage (91.96) |
| **New Mexico** | La Chiripada (88) | La Chiripada (92.66) |
| **New York** | Wölffer/Hermann J. Wiemer/ Fox Run (94) | Hermann J. Wiemer (95.23) |
| **North Carolina** | Raffaldini (90) | Raffaldini (93.8) |
| **Ohio** | Debonne (87) | Debonne (92.76) |
| **Oregon** | Cayuse (99) | The Eyrie Vineyards (96.2) |
| **Pennsylvania** | Chaddsford (91) | Chaddsford (93.3) |
| **Rhode Island** | Sakonnet (90) | Sakonnet (93.33) |
| **Texas** | Llano Estacado (89) | McPherson (93.13) |
| **Vermont** | Snow Farm (85) | Snow Farm (90.33) |
| **Virginia** | King Family/Linden (92) | Lovingston (94.16) |
| **Washington** | Charles Smith/Cayuse (100) | Rulo (96.66) |

Table 2. Top scoring winery and best value for money winery within states

|  |  |  |
| --- | --- | --- |
| Rainfall | Sunshine | Humidity |
| 0.024 | 0.073 | 0.030 |

Table 3. - R Square value correlation with Review Score





**4.2 Limitations of the data**

This data has a few serious limitations that should be taken into account when considering this analysis. Firstly, states such as California produce a far greater number of wines than other states on this list. There is no comparison to states such as Hawaii which only have one wine in the data set, and therefore high quality wines produced in California and even good value for money wines, may get diluted by the thousands of records in the dataset. In some cases, this may not be considered a good distribution of data for analysis, however it seems certain climates and state specific social culture provide a better propagation of wine development and hence can’t be accounted for. Dataset covers about half of the states so there is big field for improvement of the dataset. Overrepresentation of California implicates necessity of generalized analysis which is not relevant for all states.

The reviews for the wines are based on an individual’s palette and therefore, what one person may consider a quality wine may not be the definition of quality to another individual. Therefore, these reviews are largely down to the perspective of different reviewers and each one may not be judged equally or fairly and can even depend a person’s experience within wine tasting. This may also come into play with the quality correlations and climatic factors. As the reviews are subjective to the individual reviewer, quality of the wine may be reflected by individual preferences.

**5. Technical integration between elements**

In this project, there are many elements that will be technically integrated to come up with the main website. According to Fedele, Panteghini and Vergalli (2011), in order for an analytical element to be effective in conveying the required message, it is important to come up with technical tools and to integrate them properly to come up with a unitary whole. As much as technical tools like data analysis should be applied effectively in the website, it is important to consider the users and their different levels of understanding. Impact assessment is one of the major stages in website project planning. One of the steps is to look at the target audience and to design tools that they will be conversant with. Users of data are sometimes very many and it is only appropriate to use simple techniques that will be understandable and easier to conceptualize.

In our project, the main goal is to come up with a website that will contain different dataset relating to wine use and their appropriate analyses. The goal here is to bring information closer to the users. Usually, people find it hard to collect data and sometimes it would be of use to collect data that people would find useful and bring it close to them. In the event that this data is scattered or contained in big data systems, it becomes very complicated for some of the users to extract it. However, when it is presented in a simplified and understandable manner, users can read it at all times and use it in their day to day lives (Tee, 2016). Most of the commonly used datasets in the website are those related to population, consumption and other social phenomena.

**5.1 Integration of data analysis tools**

As stated by Idzorek and Kowara (2013), data analysis is the process of breaking the complexity of data in order to present something that is understandable to the users. Basically, in this project, the first step was to collect all the data relating to wines consumption and storing in in cloud hosted databases. These include system tools like OneDrive that are used to store data that can be used online for different analyses. In order to ensure that this data can be used by different populations, it will be sorted into different variables and presented independently. One of the technical aspects of this project is to ensure that from the website, different users of data are able to extract useful information based on their fields of interest.

It is important to remember that the goal of the project is to bring information close to the users. There are different sides of the information that could not all be required by the users. Apart from there being different datasets, there are different analysis tools that that can be understood by different users. General public and learners from secondary schools and colleges would only conceptualize analysis tools like means, ranges and other descriptive statistics. In order to ensure that all these groups remain sorted, it is good to ensure that all these tools are well blended. The utmost goal is to attract as many users as possible.

**5.2 Integration of data visualization**

Data visualization tools refer to such application as graphs and other visual elements. Just like it was the case for data analysis, there are different data visualization tools that can be integrated at different levels to ensure that the information contained in the data is represented in a usable manner. For general data users and students in secondary schools and colleges, simplified data visualizations like bar graphs, percentage graphs and pie charts are the most applicable. For advanced data users, more sophisticated representations like scatter plots, stem and leaf diagrams, box plots etc. These are used to provide advanced data analysis results in order to satisfy this class of users.

An interactive map was a key part of our website as we wanted to offer users the ability to interact with a map to find results regarding differing variables. To do this we used a website (<http://createaclickablemap.com/>) which provided the necessary application to create interactive maps for our project. We were able to use this to import our data and create our maps including Score Map, Price Map, Value for Money Map and Flavour Profile Map.

**5.3 Integration of HTML frameworks**

These are those frameworks that are integrated into the website to make it more real and to give it a friendly user interface. Interaction within a website can sometime be complicated when the users fail to understand some factors related to the data sets. Things like location and size of data are sometimes important to be understood by users. In order to ensure this feature is well integrated within the website, the project will make use of Github repository.

**5.4 Integration of exploratory analysis**

Apart from presenting different data analytics tools, it is important to combine with narrations so that users of data can understand the meanings of different statistics and representations. It has already been pointed out that the website will make integration between data analysis and data representation. Once these are perfectly done, it will be possible for users to get more insight on the data and be able to apply the findings of these analyses in their different areas of interest. In order to make this data and findings even more useful, the project will integrate it with narrations and analyses. These are explanations that are aimed at providing more information regarding the data analysis results.

**5.5 Integration of design and aesthetic features**

Finally, it is important to understand that without proper presentation, a website cannot attract the eye of the user. It is important to start by pointing out that in the current day environment, products, however useful, should be forced to the users in order to add value to them. In order to ensure that the data that is contained in the website is presented to users and that it is availed when it is needed, the project will have designed in a manner that users will find it and use it (Tee, 2016). The first design feature that will be adopted is simplicity. Simplicity in web design means that it is simple to find anything on the website and to use it. Rather than using a sophisticated user interface, the project will be based on a simple interface with clear fonts and diagrams that are easy to access.

**6. Summary**

Through profound analysis it was possible to show an interesting fragment of wine industry in USA. It was not meant to be economic description of the industry but description of the topic for people who are truly interested in wine and those who have not been into wine and would like to start their adventure in this huge subject.

By far the most developed states when it comes to wines are located on the west of USA. California, Washington and Oregon states produce the biggest amount of top-quality wines. Similarly, we can observe that top scoring wineries are also located within those states (plus Rhode Island). On the East Coast the biggest producer is New York (by number of produced wines). However, by looking at the list of the scores of the best value for money single wines it does not only come to those 3 states but is divided fairly equally across the United States.

We assume that development within alcoholic beverages industry, especially wineries and breweries across the USA will be growing steadily. Recent years have shown that popularity of wine has its concrete effects such as boost of wineries, developing dedicated university courses or television programs.

Although climate has big impact on raw material (yield and condition) for production but final score is rather influenced by particular tester’s likes and it is very subjective and vary from tester to tester.

Project itself was a great chance to make the topic of wineries wider known and positively acquired by people from outside the industry.

All the data and code can be found at <https://github.com/winecraftUSA/winecraftUSA.github.io>, website is stored at <https://winecraftusa.github.io/>.

**References**

Fedele, A., Panteghini, P. and Vergalli, S. (2011) Optimal application of web design tools to create user interfaced websites for social; *International journal of technolog*y, 12(4), pp.438-468

Idzorek, T. and Kowara, M. (2013) Factor-Based web design and its application in big data systems; Journal of advanced technology; 69(3), pp.19-29

McMillan, R., EVP and Founder Silicon Valley Bank Wine Division (2018) State of the Wine Industry 2018 [Online]. Available from:<https://www.svb.com/uploadedFiles/Content/Trends_and_Insights/Reports/Wine_Report/SVB-2018-wine-report.pdf> (Accessed: 16 May 2018)

Euromonitor International (2017) Wine in the US [Online]. Available from:<http://www.euromonitor.com/wine-in-the-us/report> (Accessed: 16 May 2018)

Thach, L. (2018) Overview of the US wine industry in 2018: stable growth forecasted - based on 2017 stats [Online]. Available from:<https://lizthachmw.com/2018/01/31/overview-of-the-us-wine-industry-in-2018-stable-growth-forecasted-based-on-2017-stats/> (Accessed: 16 May 2018)

Kerren, A., Kyusakova, M. and Paradis, C. (2012) From culture to text to interactive visualisation of wine reviews [Online]. Available from:<https://pdfs.semanticscholar.org/90c1/49a110108fa3201d0d1f5d5ab44e313eb50c.pdf> (Accessed: 16 May 2018)

Tee, K. (2016) Optimization and integration of different online tools for web design and use; *The Journal of web analytics*