Wine Quality

Paulo Jorge Nevado Pinto
103234
paulojnpinto02@ua.pt

Tiago Gomes Carvalho
104142
tiagogcarvalho@ua.pt

***Abstract*-In this report, it is explained how to implement several views to represent data related to the red and white Portuguese wine quality. Human-computer interactivity was an essential factor that was considered.**

***Keywords*-Human-computer, view, overplotting, trending line, persona, low quality prototype, functional prototype, d3.js**

I. Introduction

This project is related to the first assignment of the Information Visualization course, in which it was, initially, proposed to pick a theme and a dataset. Using the data from the datasets, the students had to apply diverse visualization techniques to represent the data with appropriate views.

The project was a team effort and all the code related to it is in our GitHub repository.

Figure 1 – "Wine Quality Analysis" logo

II. Motivation and Objectives

Wine is most consumed drinks all over the world, and Portugal is one of the countries that consumes most wine '*per capita*' in the world and that consumption has been increasing yearly [1]. Besides that, Portugal is a country with a high wine productivity over other European countries and is considered have a great average wine quality. However, people lack ways to deeply analyze the quality of that wine and the how it's influence by other factors. That means that there are no systems, currently available and at reach, that can satisfy these needs. This is where our platform comes to present one possible solution to solve this problem. Our solution, "Wine Quality Analysis", intents to grant the user data-rich information presented in easy-to-read views.

During the planification phase, we designed the system to try to give a solution to the problems above described, mainly to be able effectively present how the quality of the wine is influenced by other attributes and to be capable of comparing different view's results and take great conclusions about the comparisons.

III. Users and Personas

Since we are dealing with human-centered design and development, it is important to sketch some personas reflecting the type of users we are expecting to intensively use our product.

To achieve a user-friendly final product, we started to create two personas. Each persona describes a different use case and questions that should be answerable through our application.

A. First Persona

Her name is Sara Morais, 32yo. She is the owner of a small vineyard in Portugal. She is constantly seeking to improve the quality of her wines and she wants to understand the science behind wine quality to make informed decisions during the wine-making process.

This persona helps us better understand questions such as, for example, "w*hich physicochemical properties*

have the strongest correlation with wine quality". This type of user is also interested in comparing data so they can take notes and draw conclusions.

B. *Second Persona*

His name is Fernando Ferreira, 49yo. He is a financial analyst and an avid wine collector. As a wine enthusiast, he has been exploring wines from different regions for the past seven years. He enjoys understanding the intricacies of the wines he collects and tastes. Besides that, he wants to know how the different physicochemical properties influence the quality of the wine he tastes.

With this persona, we can better develop the system to answers such as, for example "*which type of wine generally has more residual sugar, red or white?*", since some users want to take a deep look at some specific data.

IV. Questions

Considering the personas described above, we wrote some questions that our product should be able to answer.

We started by defining some important and more generic questions such as the following:

1. How is a certain attribute distributed in each wine dataset?
2. Which are the attributes with most influence in wine quality? Do they differ from red wine to white wine?
3. Are there attributes with near to none influence on the wine quality?
4. How does a certain attribute influence the final wine quality?
5. For a certain wine quality, what are the typical attribute values for each wine type?

These questions will later be the pillar supporting the importance of each view that was created, because all of them will be able to be answered by analyzing the results presented in the views.

V. Dataset

The data in study was extracted from two datasets from Kaggle, a public platform containing multiple datasets, regarding most topics [2]. The datasets were then processed to produce two new data files, containing the correlation between all the features of the datasets, that will be better explained later in the reports.

The initial datasets are related the two variants of '*vinho verde'*, a unique product from the Minho (northwest) region of Portugal [3]. The data was collected from May 2004 to February 2007, containing only certified values by CVRVV. Both consist of a set of rows, all characterized by the following features:

- Fixed Acidity ($g_{(tartaric\ acid)}/dm^3$)
- Volatile Acidity ($g_{(acetic\ acid)}/dm^3$)
- Citric Acid (g/dm^3)
- Residual Sugar (g/dm^3)
- Chlorides ($g_{(sodium\ chloride)}/dm^3$)
- Free Sulfur Dioxide (mg/dm^3)
- Total Sulfur Dioxide (mg/dm^3)
- Density (g/cm^3)
- Potential of Hydrogen (pH)
- Sulphates ($g_{(potassium\ sulphate)}/dm^3$)
- Alcohol (vol.%)
- Quality

The datasets were reconsidered during the preprocessing stage, being reduced to only be contemplated by the most common physicochemical tests. Different amounts of data were then registered, consisting of 1599 red wine sets and 4898 white wine sets.

In order to compare the wine's attributes, the correlation index between features was required, so those values were previously calculated and saved on two new files, one for each wine type. This was done mostly to reduce the complexity in the calculations processed between views and to enhance the user experience.

VI. Low Fidelity Prototype

The low fidelity prototype was created to be a simple diagram of an early-stage design concept of the final product. In this step, a very basic usable model was done using Balsamiq, a web-based user interface design tool for generating digital sketches of ideas and concepts for an applications or websites. The process of creating this prototype was centered on trying to answer some of the questions above described using some visualization techniques.

The main mindset during this step was trying to design a web application capable of being used easily by anyone who wanted to find answers to the questions created and that had no previous great knowledge or experience using these kinds of systems.

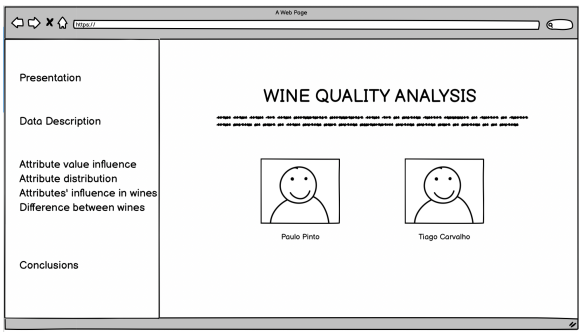

Figure 2 – Low fidelity prototype home view

The initial designed consisted of a single page application, leaving the most of interactions with the system to be located at the sidebar, the system's component responsible for guiding the user through the views. This provided to be a great decision, that would lately be implemented in the functional prototype. Without being too complicated for the user, this sidebar was planned to allow the user to navigate the application freely and without any specific order.

The first question to be solved was "*How is a certain attribute distributed in each wine dataset?"*. We decided to create a Bar Chart, with registered values of a certain wine's attribute sorted in an ascending order over the count of records for each value. The user would be able to select which attribute would be presented in the chart and which wine would to be considered for the data.

For the questions "*Which are the attributes with most influence in wine quality? Do they differ from red wine to white wine?*" and "*Are there attributes with near to none influence on the wine quality?*", we created a Correlogram, a chart of correlation matrix, since it is very useful to highlight the most correlated variables in a data table. In our plot we decided on using correlation coefficients that were colored according to their value. The user would be able to select which wine type to be presented in the view.

To answer the question "*How does a certain attribute influence the final wine quality?*", we decided to rely on Scatter Chart, since an easy-to-read chart, common chosen to graphically represent and explain the relationship between two continuous variables of the same dataset. Users familiar with the Cartesian coordinate system, normally find this chart to be really helpful when trying to compare two different variables that, at naked eyes, seem to have no correlation.

Last but not least, to provide an answer to the question "*For a certain wine quality, what are the typical attribute values for each wine type?*", we settled on creating a Radar Chart. This kind of charts displays multivariate data stacked at an axis with the same central point. With the use of a radar chart, it was possible to feature three or more attributes at once and finally directly compare both wines at the same view. The spider web designed is not always the easiest to analyze, but in this case where the comparison between the two datasets was the predominant factor on answering the question, we found this chart to be very helpful.

This prototype was lately used by same similar students and some teachers and proved to be mostly effective on answering all the question, but some problems were detected during those uses. Unfortunately, some elements would need to be reformed to better fit the user's needs and expected behaviors, but those were saved for the functional prototype.

VII. Visualization Solution

Upon studying the registered results from the usability tests of the low fidelity prototype, we took some notes and decided on how we would want the system to answer the questions above described and how to fix the problems that the users reported.

As said before, to avoid unnecessary complexity, a single page application with the alternating simple components, according to the desired view.

VIII. Technologies

To develop the application, it was used React Native with the support of some libraries such as "*react-d3-library*", the key imported library, responsible for providing the system with amazing methods to fully explore most visualization techniques.

IX. Functional Prototype

The solution consists of a total of five possible views, that can be accessed through the sidebar, pressing the easy-to-use buttons. Regardless of the view, the sidebar will always be locked on the left side of the screen, so that the user can rapidly change between views without much trouble.

A. Home view

This is the initial view, the one seen upon entering the platform.

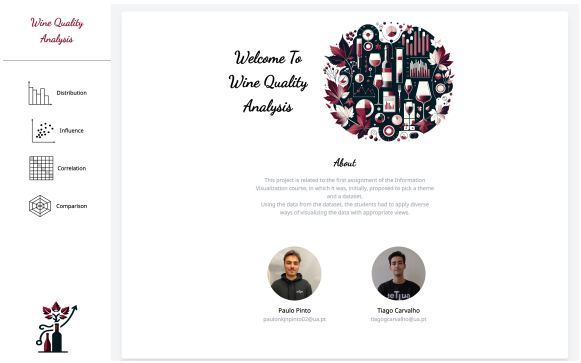

Figure 3 – Home view

It presents the project's name together with a thematic image related to the concept of this project. Those elements are followed by an "About" section, dedicated to explaining this project objective and the answers the system was designed to answer. At the bottom of this view, there's some information about the project authors, their photos, and socials.

B. Attribute Distribution View

This is the view that was sketched in the low fidelity prototype. Its functionality is answering the question "*How is a certain attribute distributed in each wine dataset?"*. It contemplates a Bar Chart developed using *React*'s *D3.js* library.

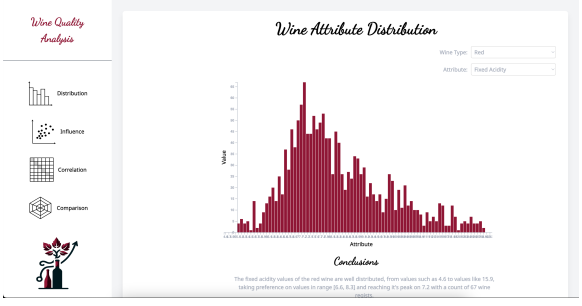

Figure 4 – Wine attribute distribution for red wine fixed acidity

This chart presents the existing values for a certain wine's attribute over the total number of times each value was registered. This way it's possible to examine that attribute's distribution in the dataset. As visible in the Figure 4, the view is initially created, presenting values, regarding the fixed acidity for the red wine variant.

Besides answering to the main questions this visualization, also allows the user to obtain information regarding other things, such as the range of values in study for that type of wine, the most common values or sets of values and, in some attributes, its visible to the naked eye which is the value with the most entries in the dataset. This view is also helpful to determine if the dataset offers a good set of a values for each attribute, helping the user determine if the dataset is reliable for study.

C. Attribute Influence on Wine Quality View

This was the second chart to be created and, this time a Scatter Chart. It was also developed using the same libraries.

Figure 5 –Red wine fixed acidity influence on wine quality

This view helps the user answer to the question "*How does a certain attribute influence the final wine quality?*", since it stabilizes the relation between a selected attribute and the quality related to each value in the records. With this chart the user can identify different pattern, associated with the most common values of the attribute in study to each wine quality. The majority of the information that can be analyzed on our chart is related to data plotting and trending lines. These are two concepts related to this type of charts and helps the user examine which is the range of values that normally is related to one quality.

Figure 6 –Red wine alcohol influence on wine quality

In the Figure 5, its visible that low values for the chlorides are favorable for the wine quality. In the Figure 6, related the alcohol values, its visible some sort of a trending line since low wine qualities are related to low values of alcohol and greater qualities are more associated with higher alcohol values [4].

However, this type of observation can sometimes lead to false affirmations and results because a relationship between two variables in a scatter plot, does not mean that changes in one variable are responsible for changes in the other. This gives rise to the common phrase in statistics that correlation does not imply causation.

D. Attributes Correlation

For next view, we have a Correlogram using heatmap color themed circles. This is plot uses the pre transformed data above described. In this plot, we aim to solve the questions "*Which are the attributes with most influence in wine quality? Do they differ from red wine to white wine?*" and "*Are there attributes with near to none influence on the wine quality?*".

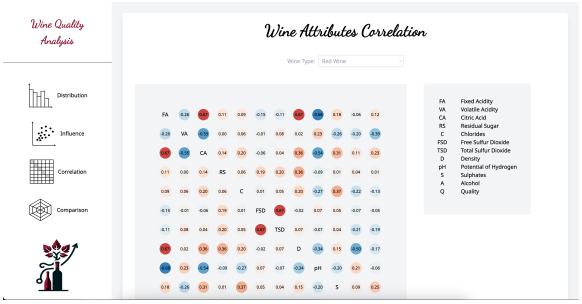

Figure 7 – Red wine attributes correlation

Since it depends directly on the correlation coefficients that can be calculated from the data rows of each dataset, we decided that it would be more user-friendly to pre-process those coefficients and providing them to the system. This strategy saves the user from waiting the long time it takes to fully read the datasets, calculate each coefficient, and produce the correlation matrix. It also avoids unnecessary and repeated operations and improves the overall user experience.

Now considering the plot, it consists of a correlogram that, for each cell, besides the value, presents a circle colored with one of the colors from the traditional heatmap pallet, dark blue to red. The cells valued with negative numbers will get blueish colors and the ones valued with positive values will get reddish circles. The more the value is closer to zero, the close the color will get to white.

This kind of plots is always a risky move when presenting that because some people might find it hard to read or to analyze. However, the plot presented proved to provide a valuable resource and an advantage to this project and to the data being presented. For those trained or used to working around this sort of plots, find this plot to be one of those you can more easily transmit the essential data and provide the user with the current answer in a quicker way.

E. Wine Attributes Comparison

The final chart we implemented in this project is, last but not least, the Radar Chart, which was chosen to answer the question "*For a certain wine quality, what are the typical attribute values for each wine type?*". This chart, named by some, as one of the best charts to use when the goal is comparing few similar objects with a limited set of attributes. In this particular case, we would be comparing the wines, more specifically, the mean values for each wine's attributes.

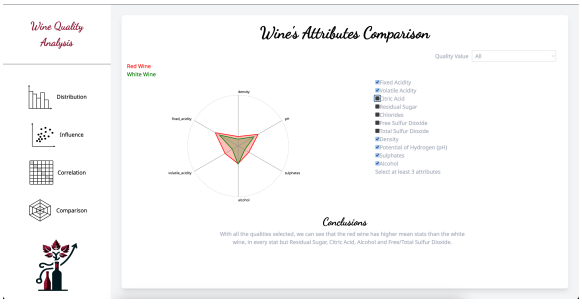

Figure 8 – Wines comparison on fixed acidity, volatile acidity, density, pH, sulphates, and alcohol

In addition, the user might want to only select to see the mean values of the wines related to a certain quality. To do so, there's a dropdown containing all

the registered wine qualities, that can be used to select to select the wine quality the user wants to study.

This chart showed some interesting results to extent that some of those results were very different to some wine qualities.

The user also might want to restrict the number of attributes in study, so we decided to implement multiple checkboxes, that can be used to add or remove attributes in the radar chart.

F. Conclusions Section

As clearly visible in every single component, chart, plot or page, there's a section reserved for conclusions in the bottom.

These conclusions take into consideration the visualization being presented to the user and explain some general aspect. Those conclusions might not be the ones that the user might be looking, but those aim to help new users understand the idea being the chart and show some values the user might also want to know.

X. Conclusion

We are very proud to have made this research and developed this product. It provided to be a great final result, having a functional prototype able to solve the questions created and satisfy the needs of a great number of possible interested user.

With this project we both revisited the importance of evaluating our solution from the first steps to the final implementations and we recognize that it was due to some great advises from our colleagues that we were able to produce the necessary adjustments to better fit everyone's interests.

Both students contributed equality to create the reported produce in every stage of the development, so we consider both contributions to be 50%.

XI. Future Work

In the future, we basically need to enhance our solution. For now, you don't feel the need to create new visualizations. However, some plots and chart were in our opinion lacking crucial elements.

For example, in the Bar Chart and Radar Chart, we feel they lack a label and a way to track which value the user wants to analyze at a time.

Some stylizations would also come in hard, since some elements of the application don't fit in the general theme and can leave the user pretty confuse.

XII. Appendix

The files and the code used to develop this application are all available in github.com/tiagosora/vi-wine-quality and the functional prototype is deployed in tiagosora.github.io/vi-wine-quality.

XIII. Acknowledgement

We would like to acknowledge the helpful availability of the teachers Beatriz Santos and Paulo Dias and for giving us great advises on out to implement our solution.

XIV. References

[1] Portugal drinks the most wine in the world, by TPN/Lusa, in News, Portugal, 12 May 2023, News Portugal

[2] Wine Quality Dataset, Wine Quality Prediction - Classification Prediction, yasserh, wine-quality-dataset

[3] Modeling wine preferences by data mining from physicochemical properties, Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, Department of Information Systems/R&D Centre Algoritmi, University of Minho, 9 June 2009

[4] A Complete Guide to Scatter Plots, Mike Yi, Chartio Charts, Data Tutorials, 2021