Exploratory Data Analysis: Insights into Spanish Wine

By Will Johnson

Index

1 | Introduction

The main objective of this project was to complete an exploratory data analysis of data collected from the online platform Vivino, regarding Spanish wine ratings and quality.

Vivino is the World's largest online marketplace and community platform for wine. Users can find information about wine they may wish to consume, rate the wine and also find links to local distributors.

The project is concerned with analysing data from Vivino on Spanish wines below 50€.

The 'big picture' idea would be to expand this exploration and analysis to a larger dataset collected over a longer period of time.

1.1 | Context

Data was collected using self defined APIs that collected two dataframes of 20,000 white wines and 20,000 red wines. Both data frames were exported to .csv files and are included in the 'data' folder. These are:

- **white_wine_raw.csv**
- **red_wine_raw.csv**

1.2 | Hypotheses

Before delving into the exploration and analysis of the wine data from Vivino, it is essential to clarify the key questions, hypotheses, and assumptions that guided the direction of this project.

For the purpose of this project, average user submitted evaluations have been used as a proxy for wine quality. For this project the term quality is better conceived as consumer perception.

The primary and overarching question the project aimed to address was:

1. **Is it possible to use this dataset to identify wines that offer a high quality-to-price ratio?**
    a. It is possible that certain price ranges consistently provide a better quality-to-price ratio. We will explore whether wines in these price brackets tend to offer better overall value compared to others.
    b. Can we identify wines that offer a high quality-to-price ratio?
2. **Are there specific price points that offer better value?**
    a. It is possible that certain price ranges consistently provide a better quality-to-price ratio. We explore whether wines in these price brackets tend to offer better value compared to others.
    b. Can we identify wines that offer a high quality-to-price ratio?
    c. We assume that some wines provide exceptional quality for their price and aim to identify these wines through our analysis.
    d. Are there notable trends in wine ratings based on price categories?
3. **What is the relationship between wine price and user ratings?**
    a. We hypothesise that there is a correlation between the price of a wine and its rating. Higher-priced wines may not always receive higher ratings, and conversely, well-rated wines might not always be the most expensive.

In addition to these primary hypotheses, we will also consider other factors that may influence the findings, such as regional differences in wine quality and pricing, as well as potential biases in user ratings. This broader analysis could reveal additional insights into the value-for-money proposition of wines.

2 | Data Conditioning

The data frames were processed before moving onto any exploration or analysis of the data and the relevant .csv file for the the project is:

- **wine_df_final.csv**

Initial view of data frames:

- white_wine_raw.csv (20000,16)
- red_wine_raw.csv (20000,16)

2.1 | Cleaning and exploration

1. Examined which variables were in the data frames
2. Removed duplicate entries.
3. Determined if there were any columns that were not needed for analysis and removed as required.
4. Was unnecessary to remove whole columns due to many NaN values with the exception of grape varieties column for white wines.
5. Added the grape variety names manually for wines that were missing this information.
6. Made sure formats were correct for columns
7. Checked outliers and decided to leave them in for the analyses.

2.2 | Grouping

1. Firstly, a column (‘red/white’) was added to each data frame to categorise the wines as either red or white.
2. The data frames were then merged into a single data frame.
3. Two new columns were created:
    a. ‘grape_count’: a count of the number of different grape varieties in each wine.
    b. ‘comunidad’: The community the wine was from, based on the information in ‘region’ column.
4. Data frame saved to .csv file as cleaned dataset ready to be analysed.

2.3 | Final Dataset

The final dataset was massively reduced in size due to an excessive amount of duplicate entries in the initial data frames. From the 40000 collected entries, little over 10% were left after cleaning.

File:
wine_df_final.csv

Shape:
4051 rows, 16 columns

Columns:
Vintage_name, wine_name, year, region, bodega, varietal_name, grape_varieties, price, acidity, intensity, sweetness, tannin, wine_ratings_count, wine_ratings_average, red/white, grape_count, communidad.

3 | Analysis

Analysis was carried out in two stages:

- Visual analysis
- Statistical analysis

The next section aims to answer the questions set out in section 1.2.

3.1 | Is it possible to use this dataset to identify wines that offer a high quality-to-price ratio?

Prior to starting the analysis to address this question, the distributions of wine_ratings_average and price were assessed for the entire dataset (A.1, A.2)

As previously stated, wine_ratings_average was used as a proxy for perceived quality of the wines. As such, a side-by-side boxplot was generated for the price of wines in each rating category (A.3) with the intention to:

- Highlight ratings with large variation in price
- Visually compare the median prices for each rating category
- Gain a clearer perspective of how price varied in relation to perceived quality
- Visualise outliers to find wines that were either low priced, or highly priced for their given quality rating compared to the majority of their rating group.
    - Outliers below deemed interesting
    - Outliers above deemed wines to potentially avoid

As observed in A.3, 83 outliers above were highlighted in the 3.8, 3.9 and 4.0, which indicated the possibility that wines of these qualities are often overpriced. Conversely, 2 wines were highlighted as outliers for being low priced for their quality and are therefore recommended for further investigation (R.1, R.2). These two were deemed likely to be of fantastic value for their quality.

Despite some overlap between the notches of the boxes at the higher end of the ratings categories, ANOVA testing deemed there to be statistically significant differences between the means of prices in each rating category. This begins to illustrate the idea that there is a link between quality and price and leads into the second key question.

3.2 | Are there specific price points that offer better value?

The price variable was categorised into 5 price-point sub categories:

- 0.00 - 9.99
- 10.00 - 19.00
- 20.00 - 29.99
- 30.00 - 39.99
- 40.00 - 49.99

Side-by-side boxplots were generated to visualise the average rating for each of these price-points, and assessed in a similar manner as before. Outliers above were considered wines with an uncharacteristically high average rating, and therefore likely high quality, for their given price-point. These wines were again filtered from the dataset and can be reviewed in the appendix (R3, R4, R5, R6, R7, R8)

As similar to the previous questions analysis, notches in the boxplots showed little overlapping between the lowest three price-points, perhaps suggesting there to be statistically significant differences between the medians of these price-points.

ANOVA test and Mann-Whitney U tests between each grouping deemed there to be statistically significant differences between the mean ratings of all groups. This suggests that there are significant differences in wine quality for all increases of price-point.

While each price-point increase appears to offer better quality, the 20.00 - 29.99 group showed two positive traits for being a candidate for the best group for quality/value:

- The smallest IQR, indicating more certainty about the quality of the wines
- Most outliers above, indicating higher possibility of finding a higher quality wine than the price-point would initially suggest.

3.3 | What is the relationship between wine price and user ratings?

To assess the relationship between price and perceived quality, the wine_ratings_averages were treated as a numeric variable. A dispersion diagram was generated (A.5) which showed indication of a moderate positive correlation, indicating that as price increases, the quality of a wine also increases. This was confirmed as a statistically significant relationship with a reported Pearson's correlation statistic of 0.65, which is classified as moderate to strong correlation.

To check this pattern across all communities in the dataset, correlation matrices were made. These also included the numeric variables that had data on the profile measures acidity, intensity, sweetness and tannin. Heatmaps were produced for the entire dataset (A6) and then for each community. This highlighted the variation in the strength of the price, quality relationship between communities.

The Islas Canarias showed the strongest correlation between price and rating (0.85), with Las Baleares in second (0.73). This makes these communities strong contenders for the most likely wines to have the most certain link between an increase in price and an increase in the quality of the wine. At the other end of the rankings was Madrid with a correlation stat of only 0.05 suggesting very little correlation between an increase in price and an increase in quality.

The profile measures data was also analysed in this section as an extension to investigate any correlation with price. As shown in A.6 the correlations with profile measures was quite low for both price and ratings, with only intensity showing some weak positive correlation with rating. This perhaps indicates how users perceive stronger intensity as being a characteristic of higher quality wines. However, the correlation is weak and therefore would require more investigation through a larger dataset.

4 | Conclusions

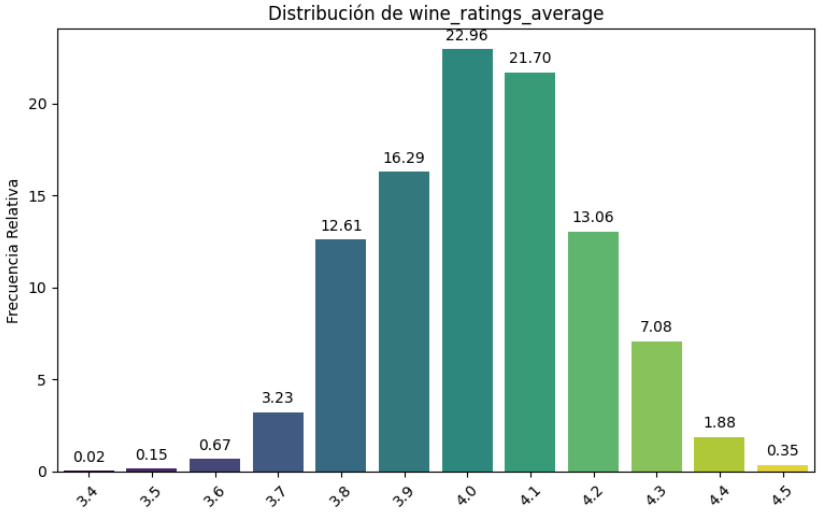
1. This dataset most definitely can be used to identify wines that are of high quality but low in price, offering excellent value to the consumer.
    a. This sort of identification analysis could be particularly useful to wine bars looking to find products that have a good price-point and are likely to sell well.
    b. Equally, this information would be very interesting to anyone interested in wine and is not a feature of Vivino that I have found, which could lend itself to a user experience update of the application/platform.

2. The dataset was particularly useful in considering whether certain price-points offer genuine differences in quality. This could allow consumers to make better reasoned decisions about how they choose to buy wine. The analysis highlighted three key ideas:
    a. Increments of 10€ showed significant increases in average wine quality. This can provide some confidence and reassurance that paying more will likely reward you with an increase in quality.
    b. The most reliable quality can be found in the 20.00 - 29.99€ price-point. This again can be used to provide extra reliability and certainty when looking to buy unknown wines.
    c. The same price-point had the most outliers for wines rated well above, which indicates the possibility of purchasing a wine of much higher quality than expected.

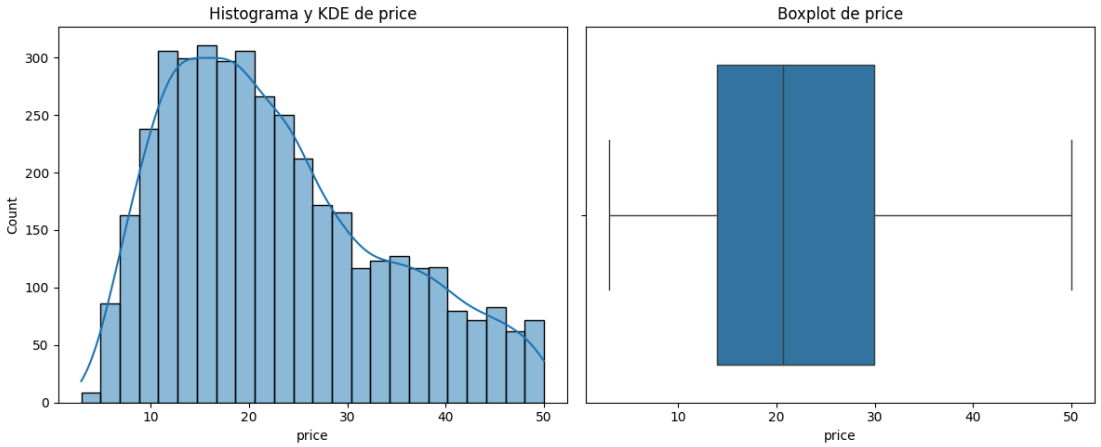
3. The relationship between price and quality is a highly debated topic in regards to wine. However, from correlation statistics of this dataset we can conclude that there is a positive relationship between the price of wine and its quality.

Appendix: Graphic Visualisations

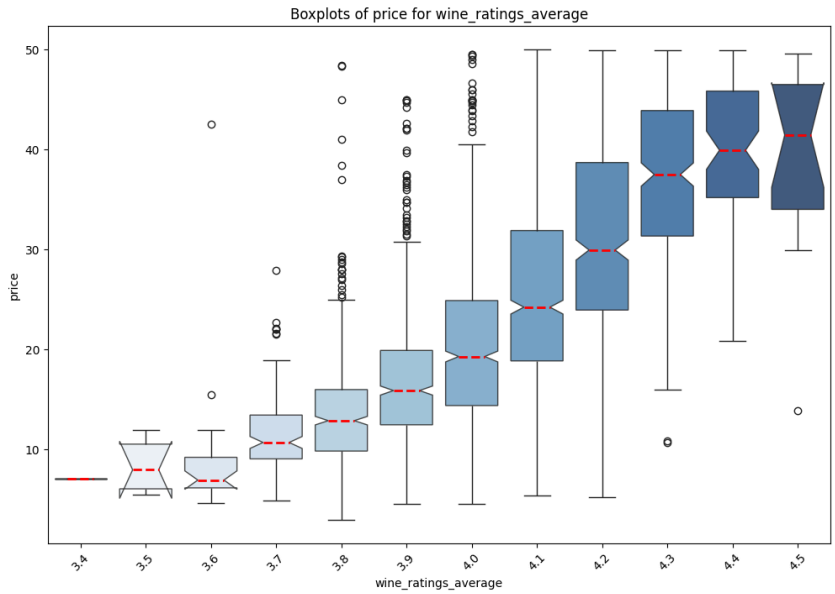
A.1 Distribution of ratings



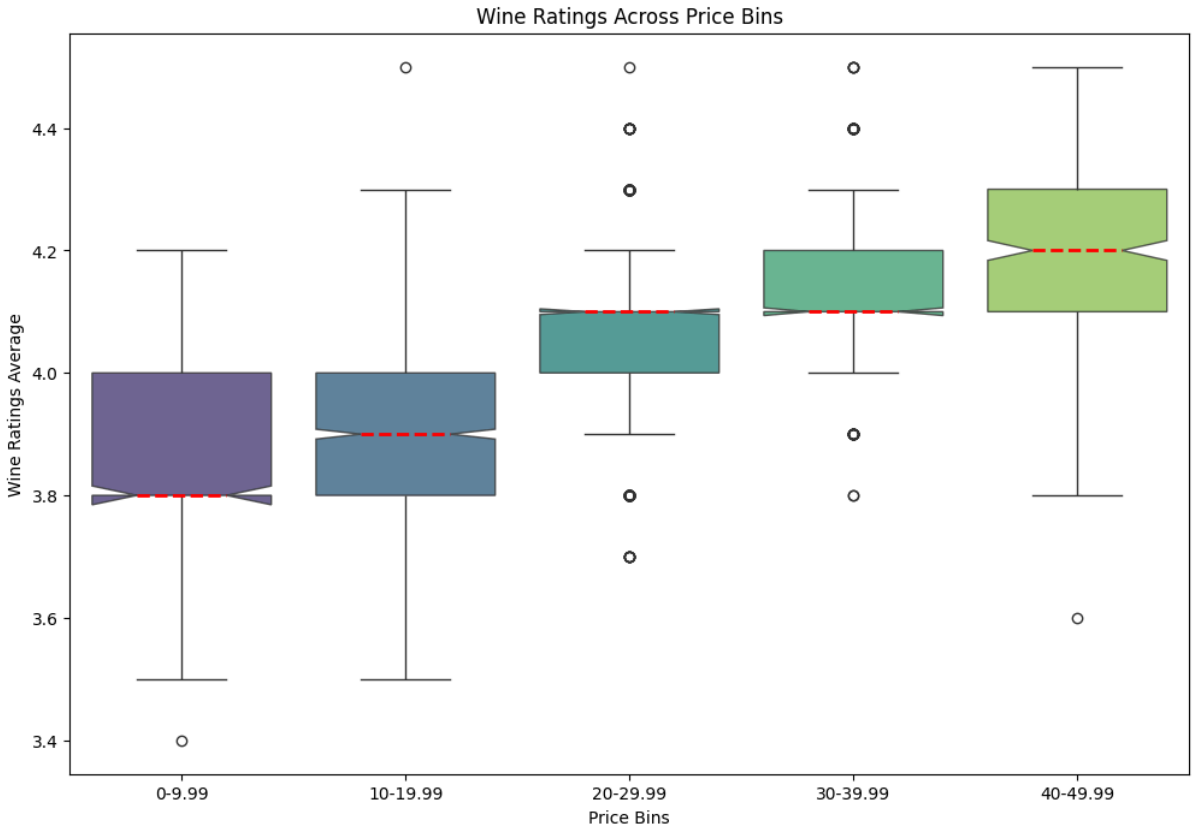
A.2 Distribution of price



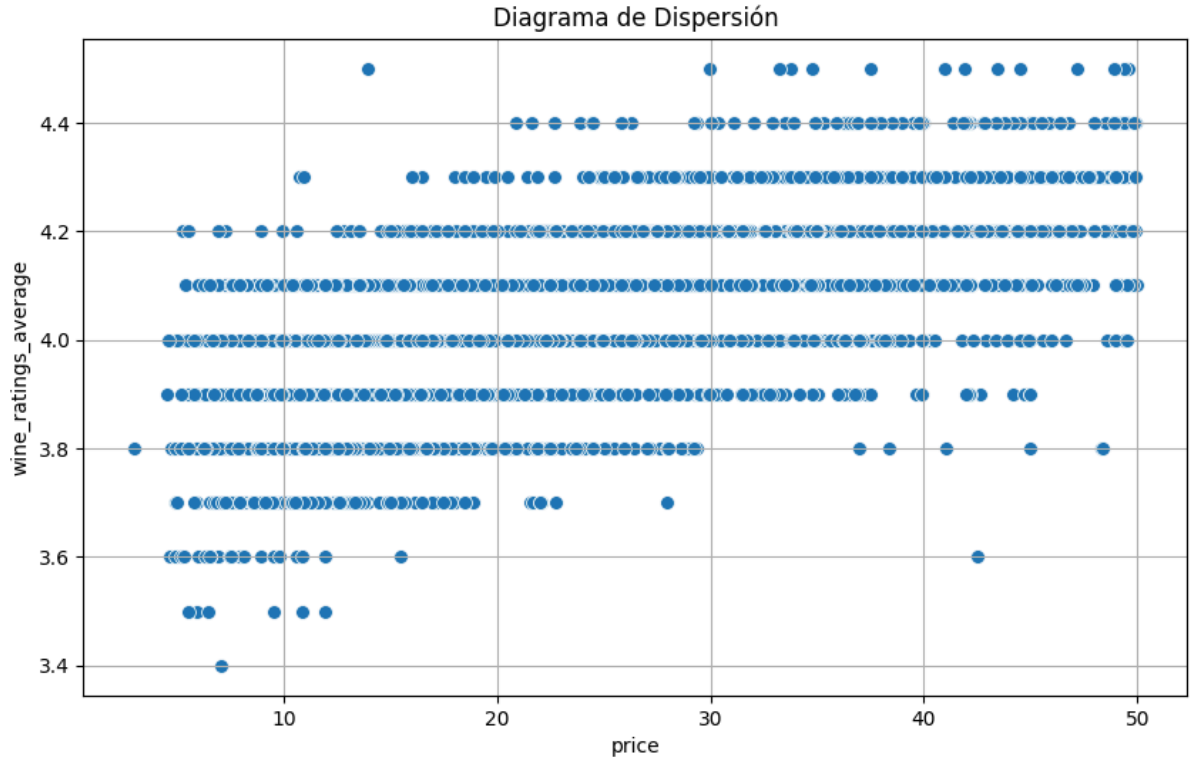
A.3 Side by side boxplot of prices for ratings



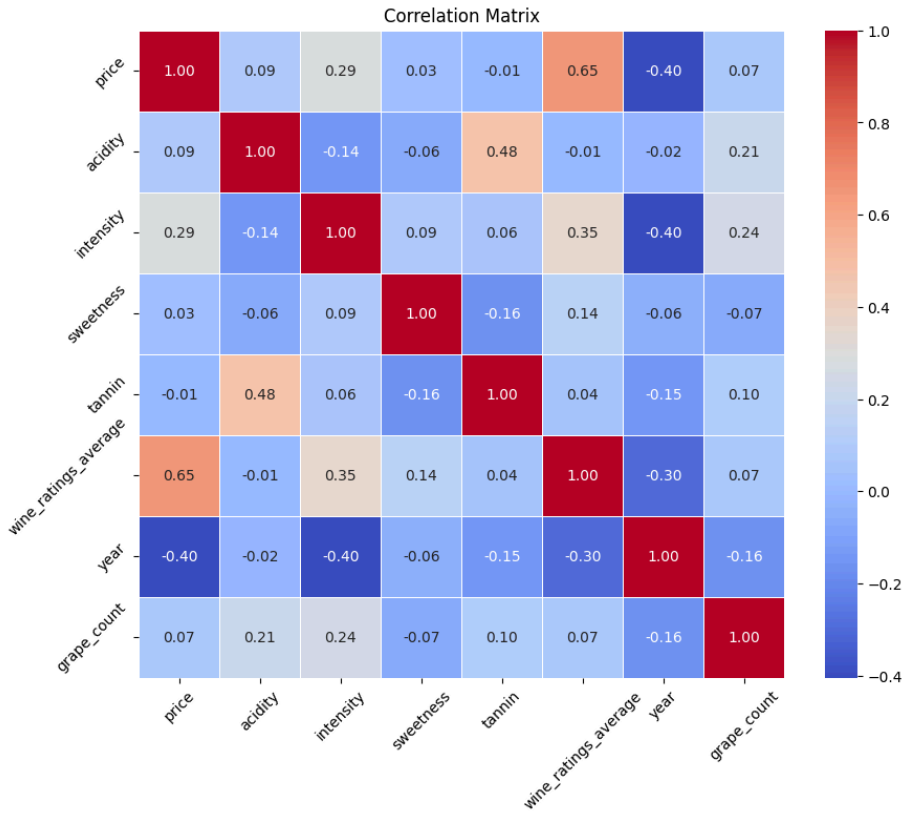
A.4 Side by side boxplot of ratings for price-points



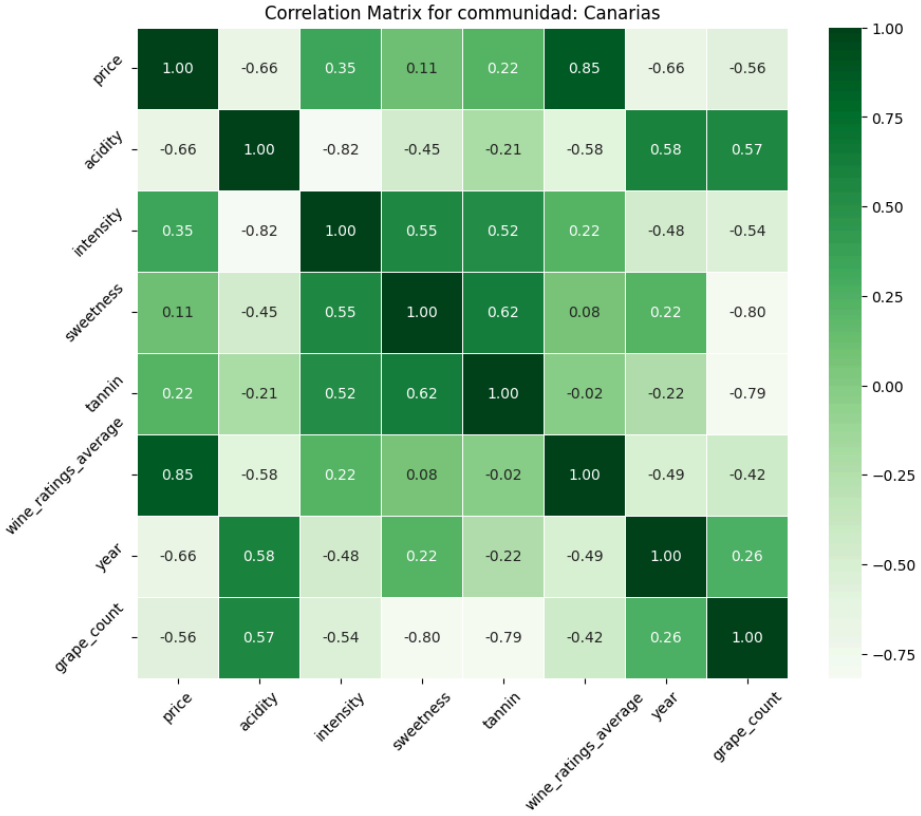
A.5 Dispersion diagram of price and ratings



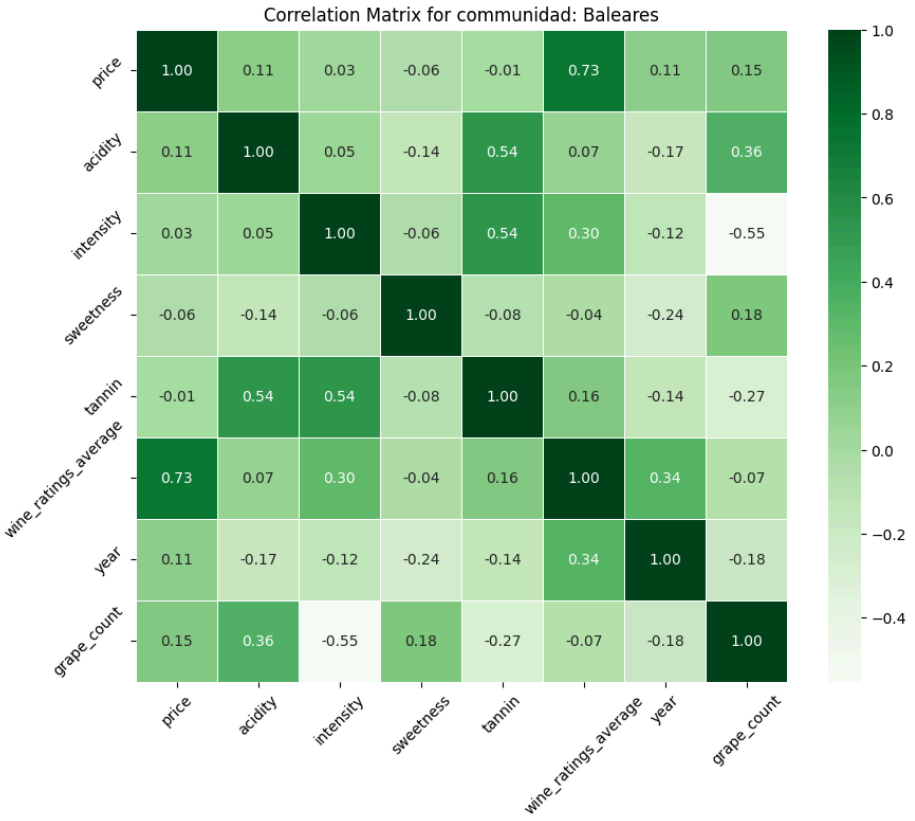
A.6 Heatmap for correlation in whole dataset



A.7 Heatmap for correlation in Canarias



A.8 Heatmap for correlation in Baleares



A.9 Heatmap for correlation in Madrid

Table of wine recommendations

	R1	R2	R3	R4	R5	R6	R7	R8
Name	**Terras Mancas Blanco**	**Airaz**	**Godina Garnacha**	**Tinto**	**Familia Chavarri Reserva**			
Year	**2022**	**2020**	**2021**	**2022**	**2016**			
Style	**White**	**Red**	**Red**	**Red**	**Red**			
Bodega	**Terras Mancas**	**Abrera**	**Morca**	**6º Elemento**	**Larchago**			
Community	**Galicia**	**Aragon**	**Aragon**	**Valencia**	**La Rioja**			
Grapes	**Albariño, Loureiro, Godello, Treixadura**	**Tempranillo, Garnacha**	**Garnacha**	**Bobal**	**Tempranillo, Graciano, Garnacha**			
Ratings count	**379**	**38**	**4412**	**2039**	**1214**			
Average rating	**4.3**	**4.5**	**4.3**	**4.3**	**4.3**			
Price	**10.82€**	**13.90**	**18.90**	**19.85**	**16.50**			