My Vivino Blog

Introduction

Revolutionizing the wine enthusiasts' journey into the captivating world of wines is Vivino, the renowned wine app and online marketplace. Boasting an impressive database housing over 15 million wines and a thriving community of 61 million users, Vivino has firmly established itself as the go-to platform for wine aficionados in search of instant access to wine information, ratings, and reviews. However, within this vast collection of wines lies a challenge – uncovering novel and intriguing bottles that align with each user's distinct tastes. Hence, our mission revolves around crafting an innovative recommendation system for Vivino. This system will empower users with personalized suggestions tailored precisely to their individual preferences. Through the application of cutting-edge algorithms and data analysis techniques, our goal is to elevate the wine exploration experience on Vivino. Together, we will enable users to effortlessly and confidently uncover extraordinary and memorable wine selections.

Data Collection

Gathering information from the Vivino website presented quite a challenge. The website's ever-changing content, known as its dynamic nature, posed a hurdle for traditional data-scraping tools like Beautiful Soup.
To resolve this, we turned to Selenium, an open-source web scraping tool. Selenium empowers us to automate interactions with web browsers, effectively navigating the dynamic website and gathering the necessary data. Our data scraping script was executed six times, each time targeting a different wine type. Subsequently, we merged these six tables into a unified dataset.
We successfully retrieved the following data from the Vivino website:

Column name	Description
Winery	The wine's brand name.
Name	The distinctive label or varietal of the wine.
Vintage	The year for vintage wines, or "NaN" for non-vintage (NV) wines.
Region	The geographical area in the country where the wine originated.
Country	The country of wine production.

Ratings	The average user rating on the Vivino website.
Number of Ratings	How frequently users reviewed a specific wine on Vivino.
Wine Type	Categorized as Red, Dessert, Sparkling, Fortified, or Rosé.
Price	The cost of each wine drink.

Data Cleaning & Validations

In our Vivino project, we noticed some missing values in the vintage feature when examining the dataset. These gaps weren't removed since they corresponded to non-vintage wines. However, we decided to exclude wines with fewer than 100 reviews as our primary aim was to build a recommendation system. This step helped us conserve memory that might have otherwise been consumed by a dense matrix.

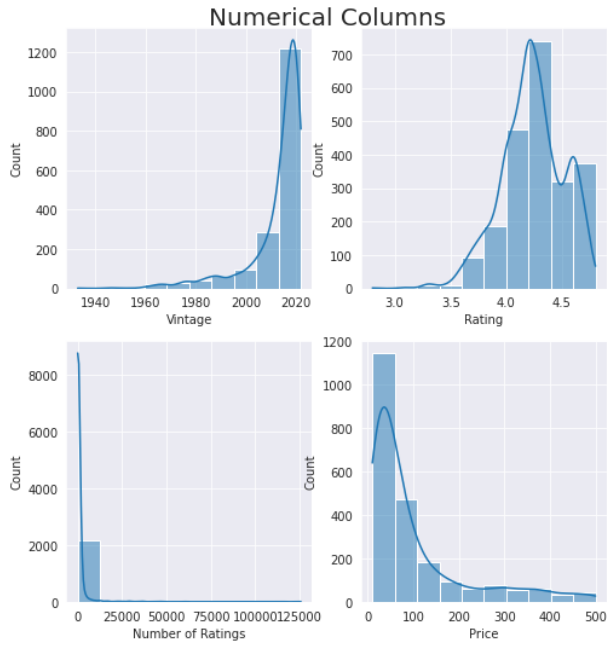
Following the data cleaning process, we were left with a total of 1260 rows, significantly fewer than the initial dataset comprising over 2219 wines. Nevertheless, this refined dataset remained substantial for training our recommendation system.

Data Exploration

- Exploring with Histograms

Histograms are like visual snapshots of data, giving us a glimpse into how our information is spread out. They're handy for spotting if data is balanced or leans in one direction. When data is evenly distributed, it forms a symmetrical shape, just like the mean, median, and mode all being the same. But when things aren't so balanced, the distribution takes on a lopsided look, either leaning left or right.
Now, let's connect this to our project. We used histograms to illustrate our numerical columns:

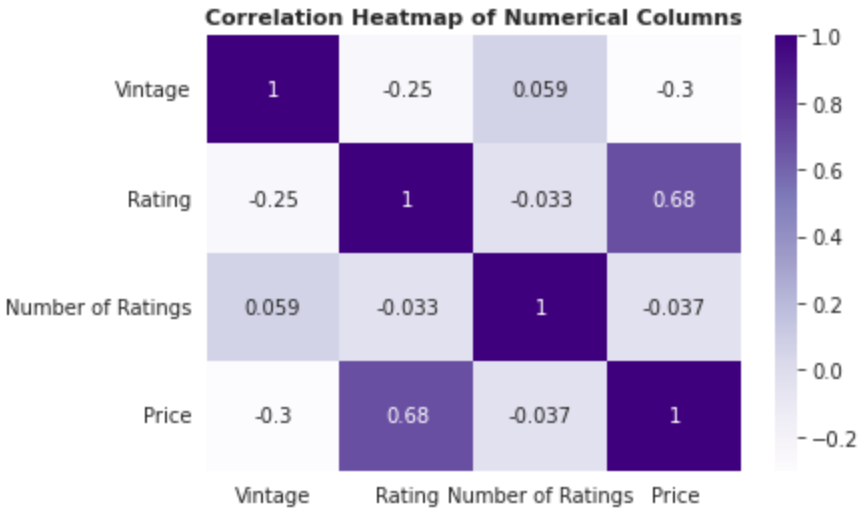
1. Vintage data skews to the right, indicating that most wines are fairly recent, predominantly from 2020, with a sprinkle of older ones.
2. Rating data appears close to a normal distribution. This suggests that a majority of wines hold around a 4.1 rating.
3. Price data skews left, hinting that most wines fall below the $100 mark. However, there are some outliers with significantly higher prices.
4. The Number of Ratings is one-sided, reaching a limit. It doesn't provide information about unrated wines.



- Visualizing Data Relationships (Scatter Matrix)

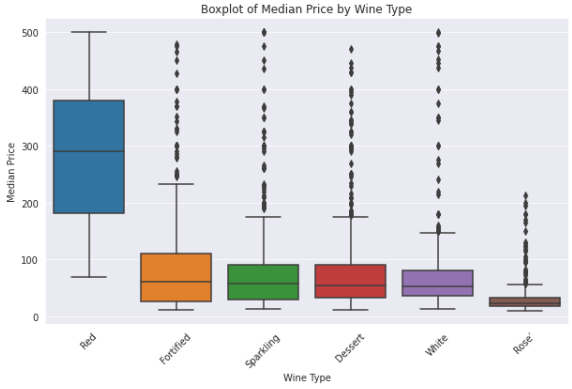
Let's explore how different attributes in our dataset relate to each other using a heatmap. Correlation measures the strength of the relationship between two attributes and ranges from -1 to 1. A correlation of 1 signifies a perfect positive relationship, meaning they move together. Conversely, a correlation of -1 indicates a perfect negative relationship, implying they move in opposite directions. A correlation of 0 suggests no relationship between the two attributes.

In our dataset, the most notable correlation exists between the Price and Rating columns in the Vivino dataset, indicating a robust positive connection. In simpler terms, wines with higher ratings tend to come with higher price tags. As for other attributes, we observe subtle negative correlations, implying weak connections. For instance, wines with more ratings tend to be priced lower. However, there's an interesting exception with the Number of Ratings and Vintage attributes, which exhibit a subtle positive correlation. This means that as a wine's vintage increases, its number of ratings is likely to rise slightly.



Data Visualization

- Analyzing Wine Prices with Boxplots



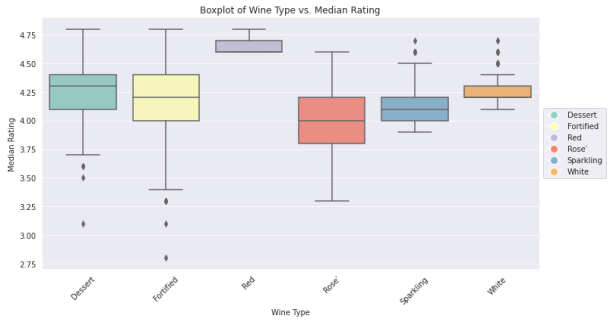
To understand how wine prices are distributed, we turn to boxplots. These graphical representations provide insights into the dataset's shape and key statistics.

Our boxplot analysis reveals that Red wine, on average, commands a price of nearly $300, which is at least double the cost of other wine varieties. This places Red wine in the higher price range. Furthermore, the boxplot illustrates that the pricing of Red wine closely follows a normal distribution, indicating a fairly even spread around the mean.

In contrast, the pricing of other wine types exhibits a mostly positive skew (right-skewed). This suggests that their prices are skewed to the right, with more values concentrated on the higher end of the price spectrum. Notably, Sparkling wines stand out as the exception, displaying a normal distribution. However, it's worth noting that Sparkling wines do have several extreme values (outliers) on the higher-priced side.

Among all the wine types, Rosé wine emerges as the most budget-friendly option, boasting the lowest median price.

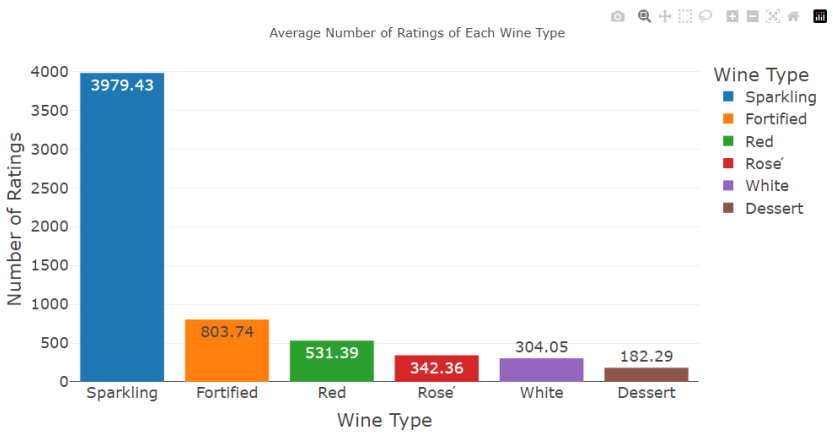
- Analyzing Wine Ratings with Boxplots



Analyzing the data through boxplots, we observe interesting insights regarding wine ratings:

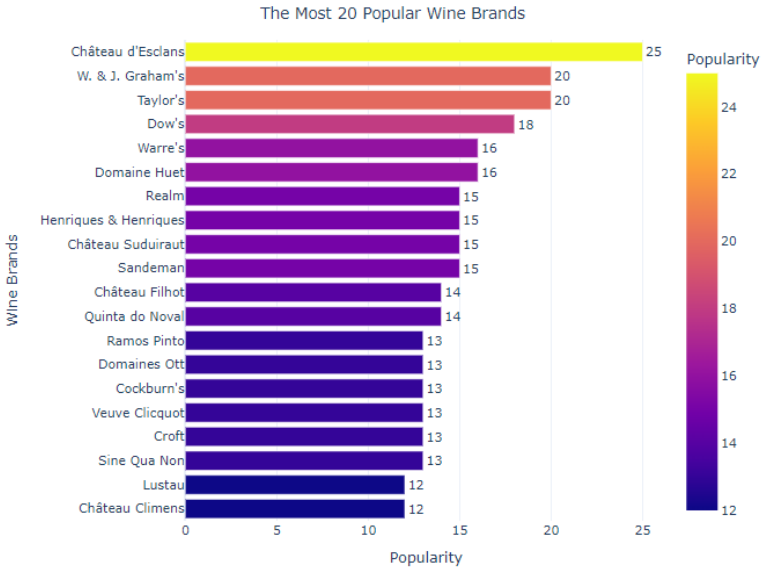
- Red wines lead the pack with a median rating of 4.6, signifying that at least half of them boast a rating of 4.6 or higher. Notably, the first quartile (25th percentile) aligns with the median, indicating a symmetrical distribution. In simple terms, there are an equal number of Red wines with ratings below and above 4.6.
- White wines secure the second spot, displaying a median rating of 4.0. This implies that half of the white wines achieve a rating of 4.0 or better. Similarly, the first quartile for white wines coincides with the median, suggesting a symmetrical distribution as well.
- In contrast, other wine categories exhibit a left-skewed pattern, implying a higher prevalence of lower ratings compared to higher ones. The sole exception is Sparkling wines, which exhibit a normal distribution with a minor presence of outliers. This peculiar trend could be attributed to the versatility of Red wines, making them a favored choice for both standalone enjoyment and food pairing.

- Analyzing Average Number of Ratings of Each Wine Type - Using Bar Plot



The popularity of each wine type can be gauged by the number of reviews it receives. In this regard, Sparkling wines take the lead, followed by Fortified wines and Red wines. This observation suggests that Sparkling wines hold the spot for the most popular wine type. It's worth noting, though, that the number of reviews doesn't always reflect a wine's quality. Instead, it often signifies its level of recognition. Sparkling wines tend to shine in this aspect as they frequently grace celebrations and special occasions, making them more widely known than other varieties. Additionally, it's important to remember that a wine's price doesn't necessarily correlate with its quality, even though pricier wines often offer superior taste due to their higher-grade grapes and lengthier aging process.

- Analyzing the most popular brands of wine



Wine producers often craft diverse wine selections, known as wine labels or wine varietals. A wine label represents a specific wine produced under a particular brand, distinguishable by its unique name, vintage year, and production region. Conversely, a wine varietal signifies a distinct wine crafted from a specific grape variety.

According to the data, the three most favored wine brands are:

1. Château d'Esclans (offering 25 labels/varietals)
2. W.&J. Graham's (providing 20 labels/varietals)
3. Taylor's (also presenting 20 labels/varietals)

Dow's and Warre's come next with 18 and 16 labels/varietals, respectively. Notably, these brands originate from different countries. Château d'Esclans hails from France, W.&J. Graham's and Taylor's represent Portugal, Dow's has its roots in England, and Warre's proudly originates from Portugal. This signifies the presence of a global wine market and showcases the international appeal of various wine types.

Additionally, it's intriguing to observe the diversity of wine types produced by these brands. Château d'Esclans offers both Red and White wines, whereas W.&J. Graham's and Taylor's exclusively produce Red wines. Dow's boasts a repertoire of Red and White wines, along with Sparkling wines. This reflects a dynamic demand for a wide array of wine varieties, demonstrating that people appreciate different wines for various culinary experiences and occasions.

- DIstribution of Countries wines production

Leading the global wine production are France, followed by the United States, Portugal, Italy, and Spain. In Africa, South Africa and morroco stands as the sole wine-producing nations.
Here's a closer look at these top five wine-producing countries:
France: Renowned as the world's top wine producer, France boasts an impressive repertoire of over 500 wine labels and grape varieties. The heart of its wine industry lies in the southern regions, benefiting from an ideal grape-growing climate.

United States: Securing the second position globally, the United States offers a diverse array of over 200 wine labels and grape varieties. Its vineyards span across the nation, with prominent wine regions in California, Oregon, Washington, and New York.

Portugal: Portugal claims the third spot globally, presenting a rich tapestry of more than 150 wine labels and grape varieties. The Douro Valley takes center stage in Portugal's wine industry, thriving under an optimal grape-friendly climate.

Italy: Italy ranks fourth in global wine production, showcasing a portfolio of over 100 wine labels and grape varieties. The Piedmont region serves as Italy's wine nucleus, flourishing in an environment conducive to grape cultivation.

Spain: Securing the fifth position worldwide, Spain offers an array of over 85 wine labels and grape varieties. The epicenter of Spain's wine industry resides in the Rioja region, where conditions are perfectly suited for grape cultivation.

The Machine Learning Section

Building the Recommendation System

Our objective is to build a wine suggestion system by harnessing data obtained from the Vivino website (the dataset we have collected and analized from the beginning of this reading).
To achieve this, we've developed a class known as WineRecommender. Within this class, we've implemented two distinct recommendation methods: KNeighbors and Similarity.
In the KNeighbors approach, we leverage the NearestNeighbors module from the scikit-learn library to identify wines that closely align with the user's selection.
Alternatively, the Similarity method gauges wine similarity using the cosine similarity metric, considering factors like ***Rating, Number of Ratings,*** and ***Price***. The results are then elegantly organized in a tabular format.

Model Testing

Initializing the Class

To initiate the class;

```
wr=WineRecommender(vivino_data)
```

Note: vivino_data is the DataFrame of wines from the Vivino website.

To use the `KNeighbors`, we use the code below:

```
KNeighbors=wr.recommend(pos=500,                    recommend_by="price",
method="KNeighbors")
KNeighbors
```

The code resets the index of the vivino_data DataFrame, checks if a specific index (500) is valid, and if so, uses a recommendation function (`KNeighbors)` to suggest a wine based on price, displaying the top recommendation. Where the pos parameter is the index of any wine. The values of the index start from 0 up to 1259. The recommend_by is the column you want to base your prediction on and the method is the type of method you want to use for the recommendation.
To use the `Similarity`, we use the code below:

```
similarity_algo = wr.recommend(winery = "La Ferme Rouge" ,name="Le Gris
2021", method="Similarity")
similarity_algo
```

Where the winery parameter is the brand name of the wine, name is the unique name of the wine, and method is the type of method

Our Results:

For KNeighbors Method

```
In [48]: KNeighbors=wr.recommend(pos=520, recommend_by="price", method="KNeighbors")
         KNeighbors

         Showing Recommendations for

         Winery                La Ferme Rouge
         Name                    Le Gris 2021
         Vintage                         2021
         Region                          Zaër
         Country                      Morocco
         Rating                           3.8
         Number of Ratings                 34
         Wine Type                       Rose´
         Price                          211.99
         Name: 1212, dtype: object

Out[48]: Winery                La Ferme Rouge
         Name                    Le Gris 2021
         Vintage                         2021
         Region                          Zaër
         Country                      Morocco
         Rating                           3.8
         Number of Ratings                 34
         Wine Type                       Rose´
         Price                          211.99
         Name: 1212, dtype: object
```

We picked out "La Ferme Rouge Le Gris 2021" at random using the `pos` parameter and inquired about suggestions focusing on the price. As this chosen wine falls under the Rosé category and price $211.99, it's natural that all the suggested wines would closely align with both the price point and the Rosé wine type. This indicates that the recommendation system is performing as intended, delivering pertinent suggestions in response to user input.

For Similarity Method:

```
Showing Recommendations for

Winery              La Ferme Rouge
Name                  Le Gris 2021
Vintage                       2021
Region                        Zaër
Country                    Morocco
Rating                         3.8
Number of Ratings               34
Wine Type                    Rosé
Price                      211.99
Name: 1212, dtype: object
```

Out[50]:

	Winery	Name	Vintage	Region	Country	Rating	Number of Ratings	Wine Type	Price
878	Roederer Estate	L'Ermitage Brut 2012	2012.0	Anderson Valley	United States	4.1	399	Sparkling	129.00
1003	Vincent Joudart	Special Club Brut Millésime Champagne 2011	2011.0	Champagne	France	4.0	103	Sparkling	139.95
2213	Château Filhot	Sauternes (Grand Cru Classé) 1976	1976.0	Sauternes	France	3.7	29	Dessert	179.32
1539	Croft	Vintage Port 2003	2003.0	Porto	Portugal	4.0	196	Fortified	199.99
910	J. de Telmont	Blanc de Blancs Brut Champagne 2012	2012.0	Champagne	France	4.1	41	Sparkling	149.97
1793	Trimbach	Gewurztraminer Alsace Vendanges Tardives 2011	2011.0	Alsace	France	4.1	267	Dessert	114.99
1810	Château Rieussec	Sauternes (Premier Grand Cru Classé) 2011	2011.0	Sauternes	France	4.0	1005	Dessert	119.99
2162	Dobogó	Mylitta Late Harvest 2007	2007.0	Tokaj	Hungary	4.0	37	Dessert	119.99
774	Louis Roederer	Blanc de Blancs Brut Champagne (Vintage) 2014	2014.0	Champagne	France	4.2	656	Sparkling	104.99
1095	Kenzo Estate	Yui Rosé 2021	2021.0	Napa Valley	United States	4.0	128	Rosé	104.99

In this metthod, we've employed the very wine we used earlier using the `KNeighbors` technique. However, this time around, we've specified the `winery` and `name` as parameters. What happens next is an automatic quest for wines that bear a similarity in terms of ***Rating***, ***Price***, and ***Number of Ratings***. It's evident that the suggested wines offer a wider array of options, particularly when it comes to Wine Type and Price.
Let's dive deeper into how these two methods differ:
`KNeighbors` method starts by picking a random wine and suggests others that resemble it, focusing on ***Price***, ***Rating***, or ***Number of Ratings***. When we specify `winery` and `name`, we pinpoint a particular wine and then find similar ones based on Rating, Price, and Number of Ratings.
For instance, take ***"La Ferme Rouge Le Gris 2021"*** With `KNeighbors`, it mainly suggests wines priced around $211.99. In contrast, using `winery` and `name` parameters, it offers a broader range of wine recommendations, providing more diversity.

In summary,

After analyzing the Vivino dataset, we've uncovered some interesting insights:

- The wine market is vast and enjoys widespread popularity. People indulge in wine for various reasons, whether it's the flavor, the occasion, or the unique regional touch of where the wine originates.
- Sparkling wines are more commonly preferred over Red wines. However, Red wines tend to boast higher quality and come with a heftier price tag. This might be because producing Red wines involves a lengthier production process compared to Sparkling wines.

- Our team has developed a personalized recommendation system. It takes into consideration individual preferences, including factors like ***Price, Rating, or the Number of Ratings.***

We firmly believe that our data analysis and tailored recommendation system will prove invaluable to wine enthusiasts and businesses operating within the wine industry.
Authored by Olajuwon Aina and Winifred Chinenye Fidelis.