**Wine Recommender System Data Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Shane Nelson  [sn888@drexel.edu](mailto:sn888@drexel.edu) | Allie Schneider  [as5664@drexel.edu](mailto:as5664@drexel.edu) | Francis Villamater  [fv48@drexel.edu](mailto:fv48@drexel.edu) | Will Wu  [ww437@drexel.edu](mailto:ww437@drexel.edu) |

College of Computing and Informatics, Drexel University, Philadelphia, PA 19104

***Abstract*** – **The goal of this project is to construct a recommender system which can be used by both novice consumers and experienced wine sommeliers alike to provide recommendations for new wines to try. The obtained dataset contains ratings for thousands of wines and the commonly used characteristics to categorize them including, but not limited to, variety, country/province/region, winery, price, designation, and even a user-inputted description. Although it is not immediately clear whether all features are readily usable or helpful as-is, this paper explores the data acquisition, analysis, and pre-processing steps taken to ensure the data is suitable for accomplishing project goals.**

***Keywords –* Wine, WineEnthusiast, Recommender System, Python, Machine Learning**

***GitHub Repository - https://github.com/snelson97/DSCI591\_G8SAFW***

1. Introduction

Wine is one of the most popularly consumed alcoholic beverages in the world, with multiple different varieties being produced by thousands of different wineries found in many countries around the globe. Although it may be easy for an experienced wine taster to identify and select a wine based on their individual knowledge, the average consumer would almost certainly not consider themself a wine connoisseur and may become overwhelmed with the myriad combinations of region, winery, variety, and description of a wine.

Of the different types of recommender systems which exist, a content-based system and a knowledge-based system both stand out as being types of systems which would be particularly relevant. The goal for a content-based system is to utilize features in the data to characterize the qualities of a wine (the “content”), then take user reviews/ratings and identify a single, “target” user for which the system could predict their rating of un-reviewed wines in order to provide recommendations for new wines to try based on the highest predicted ratings. The deliverable of the content-based system is to take wine review data for a target user as input and provide a list of recommended, previously un-reviewed, wines as output. Alternatively, one type of knowledge-based recommender, constraint-based, takes input search criteria from a user, (for example description keyword(s), country/region, winery, price, etc.) and provides an output of related wines based on the search criteria and ranked by other user ratings. Similarly, a case-based knowledge recommender takes an input wine name/ID and provides a list of similar wines to try based on similarity calculations. The deliverable of the knowledge-based system is to take input search criteria or a specific example of wine and provide relevant, related wines based on similar characteristics. Another potential option is to use clustering to group wines based on shared characteristic combinations and provide recommendations based on wines within the same cluster ranked by average user rating.

The system is applicable to a wide range of users from novice wine consumers to well-versed wine sommeliers. A novice user can use this system to help them to find new wines that they enjoy. The system can also aid novice wine consumers in cultivating a better understanding of their palate for wine. Additionally, the recommender system can provide guidance and education that will help the novice wine consumer understand the complexities of wine by offering insights into different grape varieties, regions, and styles. This knowledge empowers novice wine consumers to make more informed decisions and enhances their overall appreciation of wine. Finally, the system can save a novice wine consumer valuable time and energy in their wine selection process. Instead of feeling overwhelmed by an extensive wine list or relying on guesswork, they can rely on the system’s recommendations, which are tailored to their preferences and curated based on expert knowledge. A well-versed wine sommelier can use this system to expand the breadth of their experiences. The system can serve as a valuable tool for discovering new and obscure wines, expanding the wine sommelier’s knowledge beyond the existing repertoire. The system may also bring to light different perspectives and suggestions by helping the wine sommelier to explore different wine styles and regions. The system may empower wine sommeliers to elevate their wine selection skills, provide personalized recommendations, and deliver exceptional experiences to their clientele.

2. The Dataset

Our Wine Recommender System uses a dataset that was acquired from Kaggle.com[1]. The dataset was scraped from WineEnthusiast in 2017 by a Kaggle user and includes nearly 120,000 unique wine reviews provided by wine sommeliers. Each of the 120,000 reviews included in this dataset have been reviewed by a wine enthusiast who recorded information about their review. The information recorded from each wine review includes the country the wine was made, a description of the wine, a designation for each wine, a points or score of how much the wine was enjoyed, the price of each bottle of wine, the province and region the wine was made, the reviewer’s name and twitter handle, the title of the wine, the variety of the wine, and the winery that made the wine for a total of 13 attributes.

This dataset has been selected for the use of the Wine Recommender System because of the wide range of information that was recorded for each wine that was reviewed. The points or score of how much the wine was enjoyed by the wine enthusiast gives the system a numeric value for the quality and taste of the wine. The inclusion of a description of the wine straight from a wine enthusiast will help the system to understand how wines differ from each other. The variety of each wine is one of the foremost identifying features of a wine and is used to distinguish wines based on the types of grapes used in production. The country, province and region of each wine allows the system to compare wines that were made thousands and thousands of miles apart, as well as compare wines that were made within the same geographic area. Finally, the price of each bottle of wine will allow the system to determine how accessible each wine is to the average consumer.

WineEnthusiast (winemag.com) started as a monthly print magazine and has expanded into an acclaimed, multifaceted media brand offering of-the-moment content in the print and digital publishing space[2]. Wine Enthusiast has over 4 million readers and considers itself as the most influential voice in wine and drinks journalism today. WineEnthusiast offers perspectives, stories and insights on wine and drinks. WineEnthusiast has a global network of editors, writers, and tasters which allow for an accessible but expert view on the world of wine. WineEnthusiast offers 10 annual glossy magazine editions, a website (winemag.com), a biweekly podcast, a wine review buying guide, and virtual and in-person events.

3. Exploratory Data Analysis

|  |  |  |
| --- | --- | --- |
|  | **Points** | **Price ($)** |
| Count | 118,971 | 110,706 |
| Mean | 88.44 | 35.58 |
| SD | 3.09 | 41.88 |
| Minimum | 80 | 4 |
| 25th %ile | 86 | 17 |
| 50th %ile | 88 | 25 |
| 75th %ile | 91 | 42 |
| Maximum | 100 | 3,300 |

**Table 1: Descriptive Statistics for Numerical Attributes (SD: Standard Deviation)**

Starting with the target attribute, ‘points’ is a discrete variable with a range from 80-100 increasing by 1, for a total of 21 levels. The most common value in ‘points’ is ‘88’ at 15,141 (12.7%), followed by ‘87’ at 15,126 (12.7%), and ‘90’ in third at 13,787 (11.6%). When looking at the distribution on a histogram (Figure 1), within this range from 80-100, the distribution is fairly normal with, if anything, a very slight right skew since the median, 88, is marginally less than the mean, 88.44.

A picture containing plot, line, diagram, screenshot

Description automatically generated

**Figure 1: Distribution of Points**

Analysis of the ‘description’ attribute is limited considering it was used as a free-text field, but it contains 118,938 unique values, indicating that there is a different description for nearly every wine in the dataset. The ‘winery’ attribute, a categorical free-text field, contains 16,757 unique values and provides information about the specific winery/company which produced the wine. The attribute ‘designation’ is a subcategory of winery and is also a free-text field with 37,979 unique values and information about the specific vineyard within a winery from where the grapes were obtained. The value ‘Reserve’ is most common at 1,870 (2.2%). However, a quick scan of this attribute will reveal that not all values are written in English. The top five values are ‘Reserve’, ‘Estate’, ‘Reserva’, ‘Riserva’, ‘Estate Grown’, respectively; here we can see three variations of the word ‘Reserve’, making this a tricky attribute to use. The ‘variety’ attribute has 707 unique values and indicates the type of grapes used to produce the wine.

A picture containing text, line, diagram, plot

Description automatically generated

**Figure 2: Distribution of Price**

The ‘price’ attribute is visibly heavily skewed to the right as seen in the histogram (Figure 2), which is supported by the fact that the mean ($35.36) is more than 10 price points higher than the median ($25) and price ranging from $0-$3,300. Under our attribute ‘country’, data is heavily dominated by ‘US’ at 50,298, comprising just over 42% of the entire dataset. The second and third most common values are ‘France’ at 19,771 (17%) and ‘Italy’ at 17,830 (15%), respectively. There are only 42 unique values, but the top 3 values already make up almost 75% of the dataset, which indicates that a more specific location feature would be beneficial. The ‘province’ feature is much better, with 422 unique provinces providing a better level of detail than the country feature, however once again the data is still dominated by California wines (>30%). California is much larger geographically than any other province included in the dataset, and it would be unfair to consider Northern California and Southern California wines to be from the same location, therefore we must continue to explore the region features and their usability. ‘Region\_2 ‘is only populated for provinces of California, New York, Oregon, and Washington and provides a great level of detail compared to just using the province feature, however there is a significant number of null values for all countries outside of the US and also for some excluded US states. ‘Region\_1’ is an extremely specific feature with over 1,200 unique values, however there are still a large number of null values (>10%).

Finally, the ‘taster\_name’ feature has 19 unique values and is not necessarily something that we would use as input for any machine learning model, but it would definitely be required for a content-based recommender so that a target user can be identified and recommendations can be made based on wines which have not yet been reviewed by the user. The ‘title’ feature, similar to taster name, has 118,840 unique values and does not provide much usefulness for modeling purposes but would be valuable for the end-user when making recommendations. Without the wine title, it would not be possible to recommend specific wines and it would also not be possible to distinguish between two wines with the same features, so although this feature will be excluded from modeling, it must be kept in the final data frame for labeling purposes.

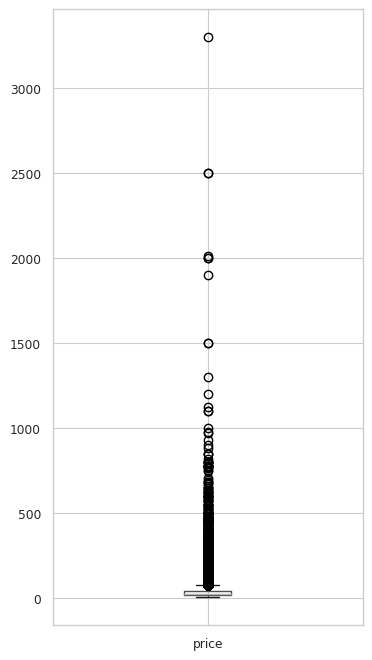
**Missing Values**

We encountered 8,265 missing values in the price column. Since this represents

approximately 7% of the total 118,971 rows, we have decided to remove these rows with missing price values. Even after dropping them, we will still have over 110,000 rows available for analysis. Similar to the approach taken with the price column, we have also chosen to drop the missing values in the country and variety columns. By removing the missing values in the country column, any associated missing values in the province column will be automatically eliminated. Furthermore, the variety feature only contains a single missing value (which could be a result from data entry error) and we will be dropping it as well.

**Outliers**

Regarding outliers, our primary concern lies within the price feature.



**Figure 3: Price Before Transformation**

Figure 3 highlights the presence of numerous extreme outliers in the price column, significantly impacting the statistical outcomes of the column. To address this potential issue, we have chosen to employ a logarithmic transformation on the column. This transformation proportionally reduces all values, eliminates outliers, adjusts skewness, equalizes variance, and lowers the standard deviation. Consequently, the emphasis shifts towards relative differences between values rather than absolute differences, which can prove beneficial in the development of our model. Figure 4 showcases a box plot of the price feature subsequent to the applied transformation. The column exhibits a more normal distribution, with the outliers successfully eliminated.

5. Results and Discussion

Now that the exploratory data analysis has been completed and each of the variables have been analyzed in their raw format, some pre-processing steps can be taken to ensure that the data is cleaned and transformed as needed and all additional relevant variables have been created. As previously mentioned, the price feature must be updated using a log transformation to reduce the right skew and create a more normal distribution than initially. We have also chosen to update the range of the ‘points’

A picture containing text, diagram, plot, line

Description automatically generated

**Figure 4: Price After Log Transformation**

feature for easier interpretability by end-users of our system. Currently, a user may consider a rating of 82 to be ‘high’ because of their limited knowledge of the range or distribution of other wine ratings, therefore we find it beneficial to re-scale this feature from 0-11, which is a more traditional range. This is a purely ‘cosmetic’ update and will not affect the modeling or feature correlation whatsoever. Figure 5 visualizes the correlation between the points and price features after transforming each, displaying a strong, positive relationship (Pearson Correlation = 0.62).

A picture containing text, screenshot, line, diagram

Description automatically generated

**Figure 5: Correlation Between Price and Points After Transformation**

Additionally, we have added another column ‘point\_range’, which groups wines into ‘high’ or ‘low’ ratings based on the median rating of 88. Any wines below 88 are considered ‘low’ while anything above is considered ‘high’, which would be required for the implementation of any classification algorithms.

A picture containing text, screenshot, display, rectangle

Description automatically generated

**Figure 6: Distribution of High/Low Point Range**

Figure 6 displays the distribution of the newly created ‘point\_range’ feature which is relatively balanced and would not require any over/under-sampling for classification modeling.

Due to the previously discussed difficulties with using just one of the country, province, or region features, we have also created a new ‘location’ feature which is a copy of province with region\_2 used in place of any of the 4 U.S. states for which this feature is populated. The resulting ‘location’ feature has 440 unique values and distinctly shows some regions which typically have higher or lower rated wines than others. Figure 7 shows the ‘Northern Spain’ region typically has lower rated wines, while other regions such as “Willamette Valley” or “Burgundy” generally have higher rated wines.

A picture containing plot, screenshot, line, diagram

Description automatically generated

**Figure 7: High/Low Distribution of 10 Most Common Locations**

Furthermore, although there was no update or transformation needed for the variety feature, we have included Figure 8 to show the distribution of high and low ratings for the top 10 most common wine varieties.

A picture containing plot, line, screenshot, font

Description automatically generated

**Figure 8: High/Low Distribution of 10 Most Common Varieties**

Wines such as Pinot Noir clearly receive a much higher rating on average, while wines such as Merlot and Rosé have notably more lower ratings than high.

Finally, the description feature is initially a string of text for each review, therefore we found it easiest to convert each description into a lowercase list of words with all punctuation removed. After removing stop words from each description, there are 46,638 unique keywords found in the dataset which could potentially be used for a knowledge-based recommender system.

6. Conclusion

**Summary**

In this project, we started with a dataset acquired from Kaggle, which consisted of nearly 120,000 unique wine reviews provided by wine sommeliers. The dataset encompassed 13 attributes, including information about the country, description, designation, points, price, province, region, reviewer details, title, variety, and winery of each wine. During the exploratory data analysis, we discovered multiple detailed and helpful columns, some of which could be readily used in modeling without much cleaning or transformation, while others required some pre-processing steps in order to be effectively deployed as features for future machine learning algorithms. To address missing values, we decided to remove rows with missing values in the price, country, province, and variety columns. This allowed us to retain over 110,000 rows for analysis, ensuring a substantial dataset for our Wine Recommender System. Regarding outliers, we focused primarily on the price feature and employed a logarithmic transformation on the price column, which reduced all values, eliminated outliers, adjusted skewness, equalized variance, and lowered the standard deviation. This transformation resulted in a more normalized distribution of prices and improved the overall statistical outcomes of the column. By the end of our EDA and pre-processing, we have created two entirely new categorical variables, ‘points\_range’ and ‘location’, while transforming two of our existing numeric variables as well. We ultimately retained 110,650 reviews after dropping null values and duplicate records, with 16 features prior to one-hot encoding for categorical features.

**Future Work**

Our Wine Recommender System, built upon the pre-processed dataset, has the potential to assist a wide range of users, from novice wine consumers to well-versed wine sommeliers. For novices, the system offers personalized recommendations based on their preferences, allowing them to discover new wines, cultivate their palate, and gain a better understanding of the complexities of wine. For sommeliers, the system serves as a valuable tool for expanding their wine repertoire, exploring new and obscure wines, and delivering exceptional experiences to their clientele.

Now that the data has been analyzed and pre-processed as needed, the next steps will be to implement a variety of different machine learning techniques and algorithms to build recommender systems. For a content-based recommender system, both regression and classification algorithms can be used since we have both a numeric ‘points’ attribute and our newly created ‘points\_range’ attribute which splits reviews into a High or Low category. To implement any of these models, we can use one-hot encoding for the categorical variables and traditional train-test splitting with hyperparameter tuning to evaluate our models and determine which would be the best to use in the “real word”. Alternatively, for a knowledge-based recommender system, we would need to create binary feature vectors and measure distance between two different wines to determine a similarity score and provide recommendations ranked based on similarity with average user rating as a tiebreaker. Finally, clustering is another option which could be used to provide recommendations based on other un-reviewed wines found in the same cluster.

References:

[1] Zackthoutt. (2017, November 27). Wine reviews. Kaggle. https://www.kaggle.com/datasets/zynicide/wine-reviews

[2] About Us. Wine Enthusiast. (2021, May 5). https://www.winemag.com/about-us/