**Wine Recommender System Analysis**

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

**Abstract** -

**Keywords** - Wine,

**GitHub Repository** -

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, type, 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 with the 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 (un-reviewed) wines as output. The knowledge based recommender takes input qualities 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. Alternatively, a case-based knowledge recommender takes an input wine name/ID and provides a list of similar wines to try. 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 create a collaborative filtering system, however that would not be concerned with qualities or types of wine at all, and would instead only return recommended wines based on those rated highly by similar users.

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.

1. **Dataset**

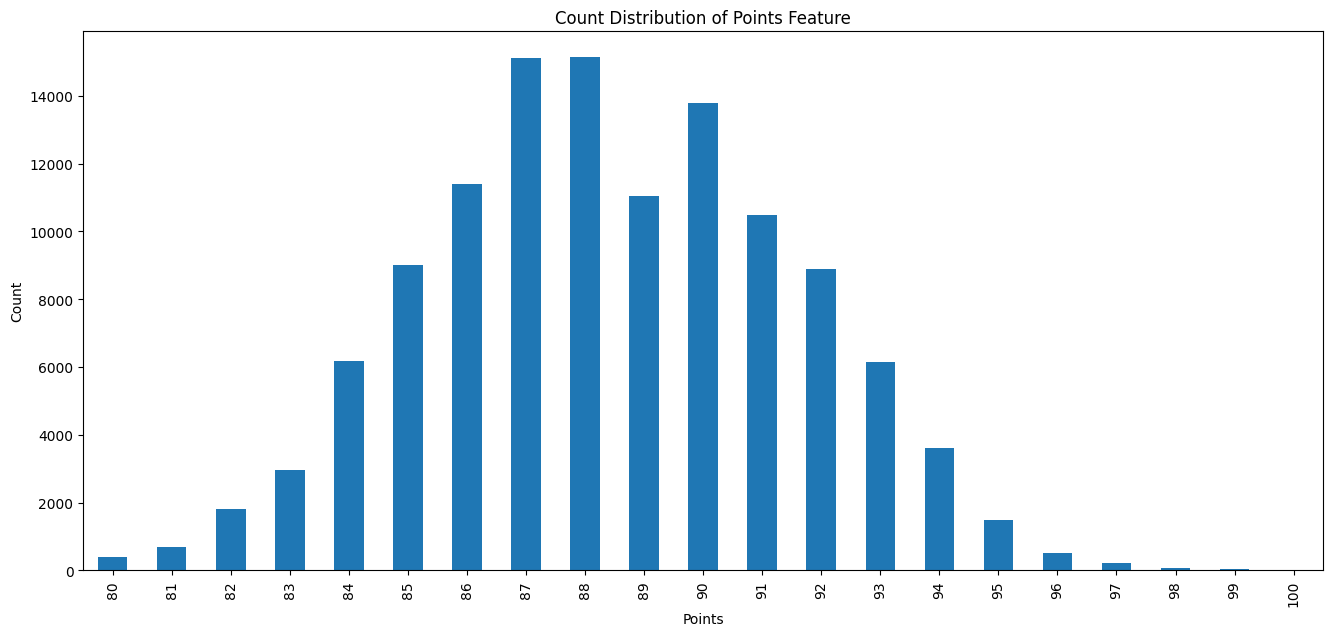
The Wine Recommender System uses a dataset that was acquired from Kaggle.com. The dataset was scraped from WineEnthusiast in 2017 by a Kaggle user, and includes nearly 120,000 unique wine reviews provided by wine sommeliers. Each of the 120,000 reviews included in this dataset have been reviewed by a 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 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 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 designation of wine variety gives the system the ability to categorize wine. 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 space1. 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. WineEnthusiasts 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.

1. **Exploratory Data Analysis**

Under our attribute “country”, data is heavily dominated by ‘US’ at 54,504, comprising just under 42% of the entire dataset. The second and third most common values are ‘France’ at 22,093 (17%) and ‘Italy’ at 19,540 (15%), respectively. As mentioned above, there are only 43 levels but the top 3 values already make up almost 74% of the dataset. Analysis of the ‘description’ attribute is limited considering it was used as a free-text field. However, as mentioned above, it contains 119,955 unique values. The attribute ‘designation’ is also a free-text field and categorical with 37,979 unique values. The value ‘Reserve’ is most common at 2,009 (1.5%). 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’.

Our ‘points’ attribute ranges from 80-100 increasing by 1, for a total of 21 levels. Most common value in ‘points’ was ‘88’ at ***Figure 1: Distribution of Points***

(13%), and ‘90’ in third at 15,410 (11.9%). When looking at the distribution on a histogram

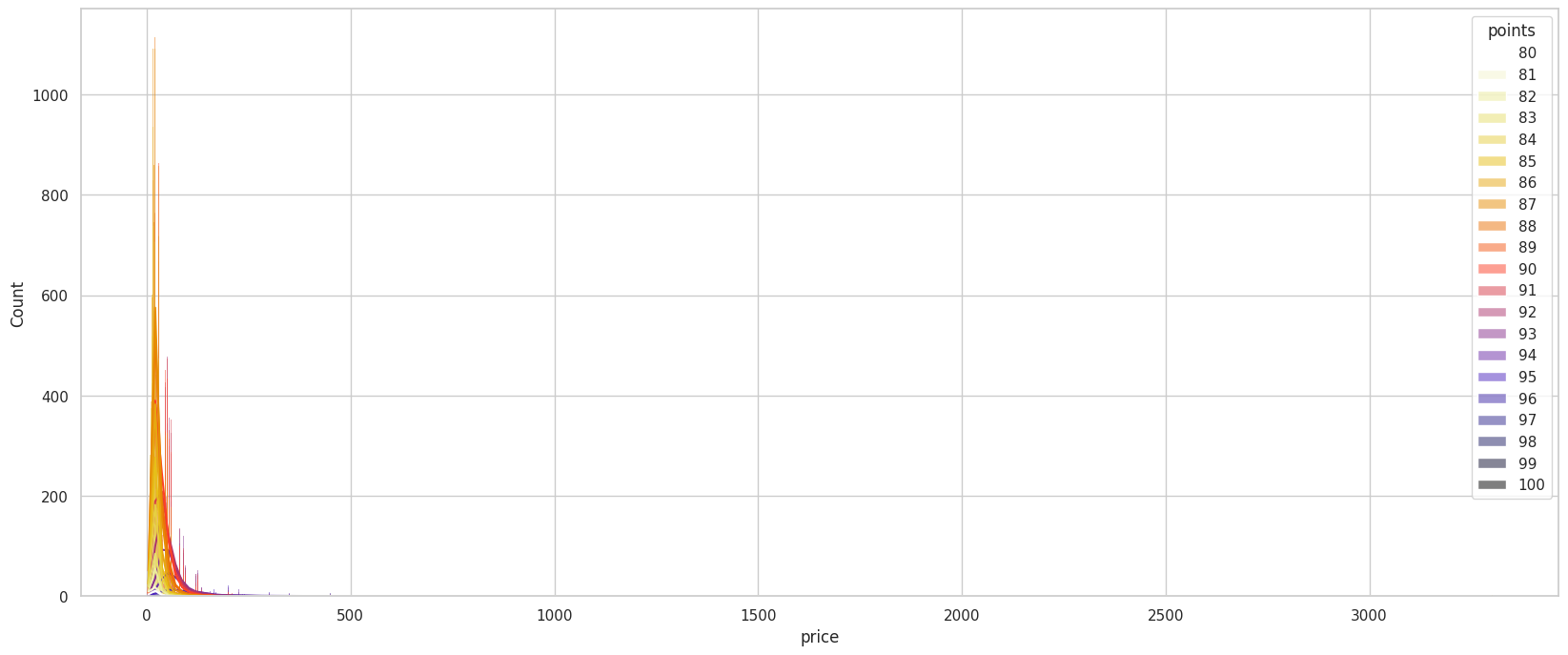
***Table 1: Descriptive Statistics for numerical attributes (SD: Standard Deviation)***

|  | Points | Price ($) |
| --- | --- | --- |
| Count | 129,971 | 120,975 |
| Mean | 88.44 | 35.36 |
| SD | 3.04 | 41.02 |
| Minimum | 80 | 4 |
| 25th %ile | 86 | 17 |
| 50th %ile | 88 | 25 |
| 75th %ile | 91 | 42 |
| Maximum | 100 | 3,300 |

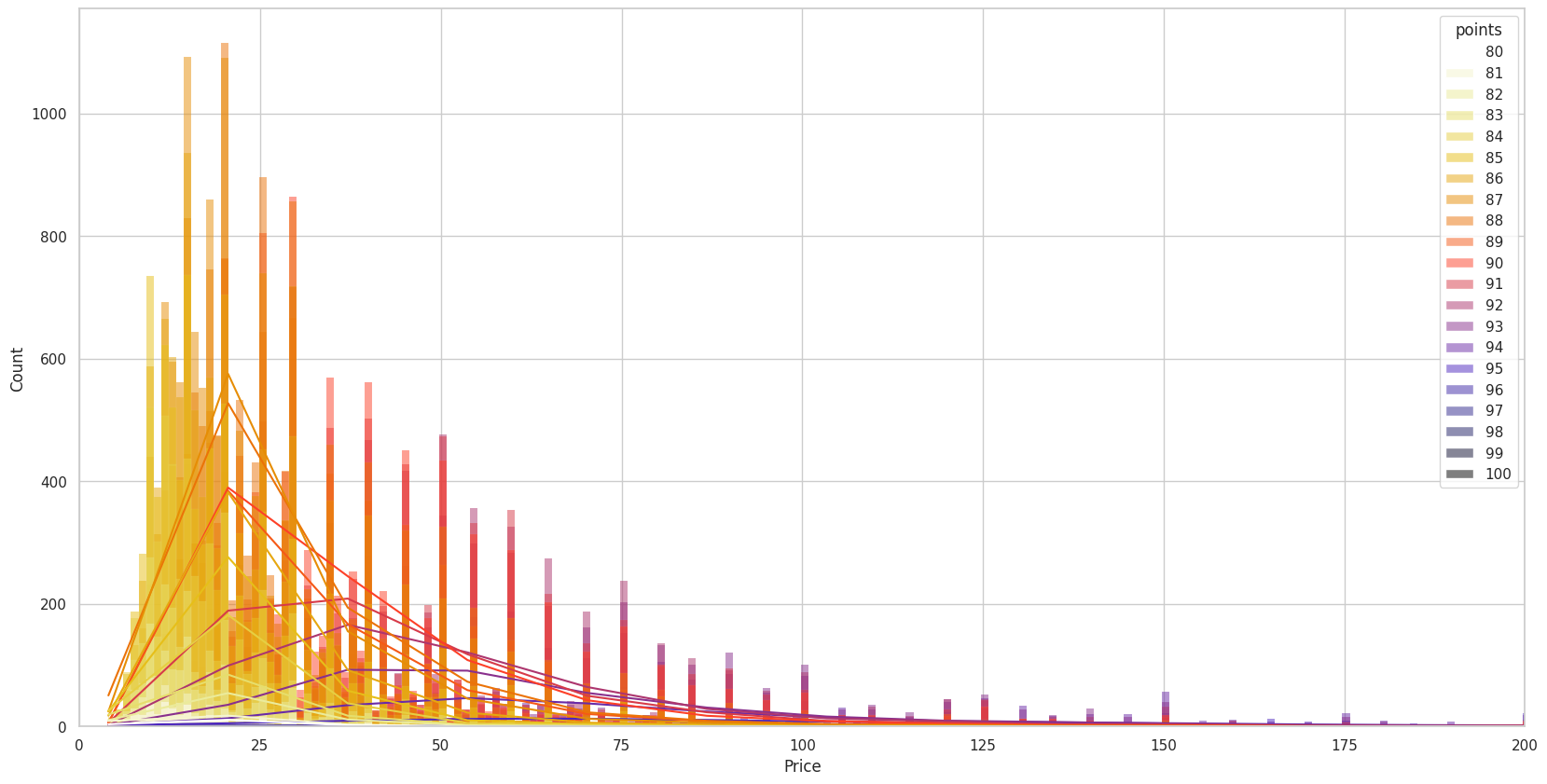
(Figure 1), within this range from 80-100, the distribution is fairly normal with, if anything, a very slight right skew since the median, 88, is marginally less than the mean, 88.48.

The ‘price’ attribute is heavily skewed to the right as seen in the histogram (Figure 2),

***Figure 2: Histogram of ‘price’***

and with greater detail from ranges 0-200 (Figure 3). This 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 54,504, comprising just under 42% of the entire dataset. The second and third most common values are ‘France’ at 22,093 (17%) and ‘Italy’ at 19,540 ***Figure 3: Histogram of ‘price’, range $0-$200***

(15%), respectively. As mentioned above, there are only 43 levels but the top 3 values already make up almost 74% of the dataset. Analysis of the ‘description’ attribute is limited considering it was used as a free-text field. However, as mentioned above, it contains 119,955 unique values. The attribute ‘designation’ is also a free-text field and categorical with 37,979 unique values. The value ‘Reserve’ is most

common at 2,009 (1.5%). 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’.

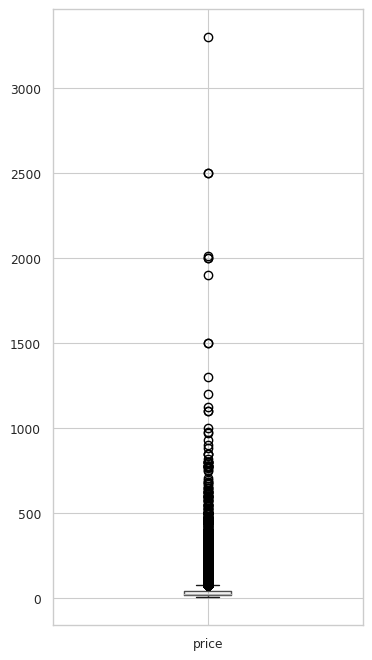
**Missing Values**

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

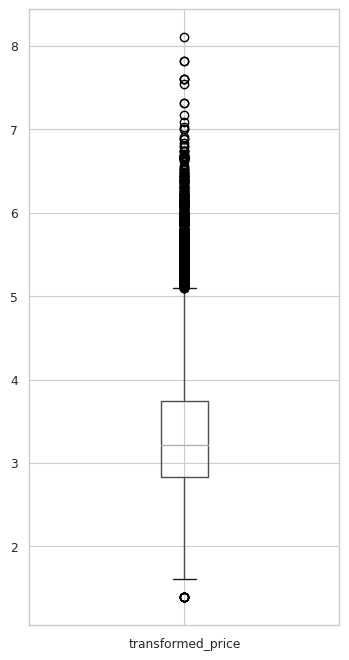
dropping it as well.

**Outliers**

Regarding outliers, our primary concern lies within the price feature. Figure 4 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 5 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.

****

***Figure 4: Price before transformation***



***Figure 5: Price after transformation***

1. **Results and Discussion**

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 observed that the dataset was dominated by wines from the United States, accounting for approximately 42% of the entire dataset, followed by France and Italy. The wine descriptions and designations were rich and varied, providing valuable insights into the qualities and characteristics of each wine. The points attribute, which represented the enjoyment score, showed a fairly normal distribution with a slight right skew. The price attribute, on the other hand, exhibited a significant right skew, indicating a wide range of prices for the wines in the dataset.

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. We 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.

1. **Conclusion**
   1. **Summary / Recommendations**

In conclusion, 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.

Overall, the Wine Recommender System that will be built in this project has the potential to revolutionize the way individuals discover and select wines, catering to both novices and experts alike, and fostering a deeper appreciation and enjoyment of this popular beverage.

* 1. **Future Work**

The insights gained from the dataset, coupled with the data preprocessing techniques applied, have laid the foundation for the development of an effective Wine Recommender System. Future work could involve implementing and evaluating different recommendation algorithms, such as content-based and knowledge-based approaches, to provide more accurate and tailored wine recommendations. Additionally, integrating collaborative filtering techniques could further enhance the system's ability to suggest wines based on the preferences of similar users.

1. **References**
2. Zackthoutt. (2017, November 27). Wine reviews. Kaggle. https://www.kaggle.com/datasets/zynicide/wine-reviews
3. About Us. Wine Enthusiast. (2021, May 5). https://www.winemag.com/about-us/