Runnan Zhang | rz2456

Iris Zhao | hz2580

Xinyi Qian | xq2190

Qingyuan Wang | qw2259

**Project Deliverable 2 - What Do You Want to Drink Today?**

This project is aiming at helping wine buyers search wine that best matches their needs and preferences. We start with our business goal and reverse engineer the process of designing such system to achieve our goal. The project includes 4 steps: 1) analyzing the data of existing wine buyers to design a questionnaire that can best reflect the needs and preferences of potential buyers. 2) Collecting the information from the questionnaire and query data from the database(wine dataset). 3) Returning a list of wine that best matches buyers’ interests. 4) Trying other techniques to find useful information to help wine buyers make decision.

Use of Analytical Techniques

1. Data cleaning: tidyr, mice package, cbind()
2. Data visualization: ggplot2
3. Text mining: freq\_terms() in the qdap package, wordcloud
4. Association rules: inspect() in the arules package and apriori()
5. R Dashboard: flexdashboard and shiny package.
6. Predictive models: decision tree and linear regression
7. User based recommender system: recommenderlab package

\*1) - 5) are the included in the final code, and 6) -7) can be found in working code.

Improvement in Data Cleaning

The raw data set consists of 129971 observations and 14 variables. First of all, we have converted the blank cells into NAs. We also eliminate irrelevant or duplicated variables such as ‘region\_2’ and ‘taster\_twitter\_handle’. We find the proportion of NA for all variables in the raw data set.

We remove observations with the proportion of NA less than 0.05 (i.e. country, variety). We also remove observations with NA in variable ‘designation’, ‘region\_1’, and ‘taster\_name’ because we cannot predict categorical variables to fill out the NAs. As a result, we end up dealing with missing values in variable ‘price’. Additionally, we extract the year from the wine’s title. During the extraction process, we build a function called ‘Numextract’, which extracts number between 1980 and 2020 from the wine’s title. We create a new variable ‘year’ and insert those numbers into the new variables. Afterward, we select variables ‘points’, ‘country’, ‘variety’, and ‘year’ as predictors to impute the missing values of price.

Eventually, we have our cleansed dataset that includes 54635 observations and 12 variables

Designing Questionnaire

The questions are designed based on the analysis of wine dataset. **Our goal is to find the wine that best satisfies buyers’ needs.** In order to achieve this goal, we need to design a **questionnaire** that most accurately conveys buyers’ needs. Words are carefully chosen based on the frequency of use and manner of speech. In other words, we choose words that are commonly used to describe wine.

1. What type of wine are you looking for?

❐ White ❐ Red

1. What flavor do you want?

❐ Fruit ❐ Rich ❐ Fresh ❐ Dry ❐ Sweet

1. What is your preferred wine age?

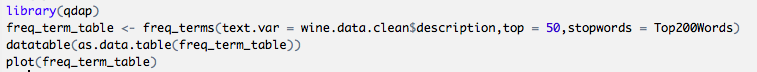
❐ 1995 and before ❐ 1995 - 2000 ❐ 2000 - 2005 ❐ 2005 - 2010 ❐ 2010 and after

1. What is the country of origin of your preferred wine?

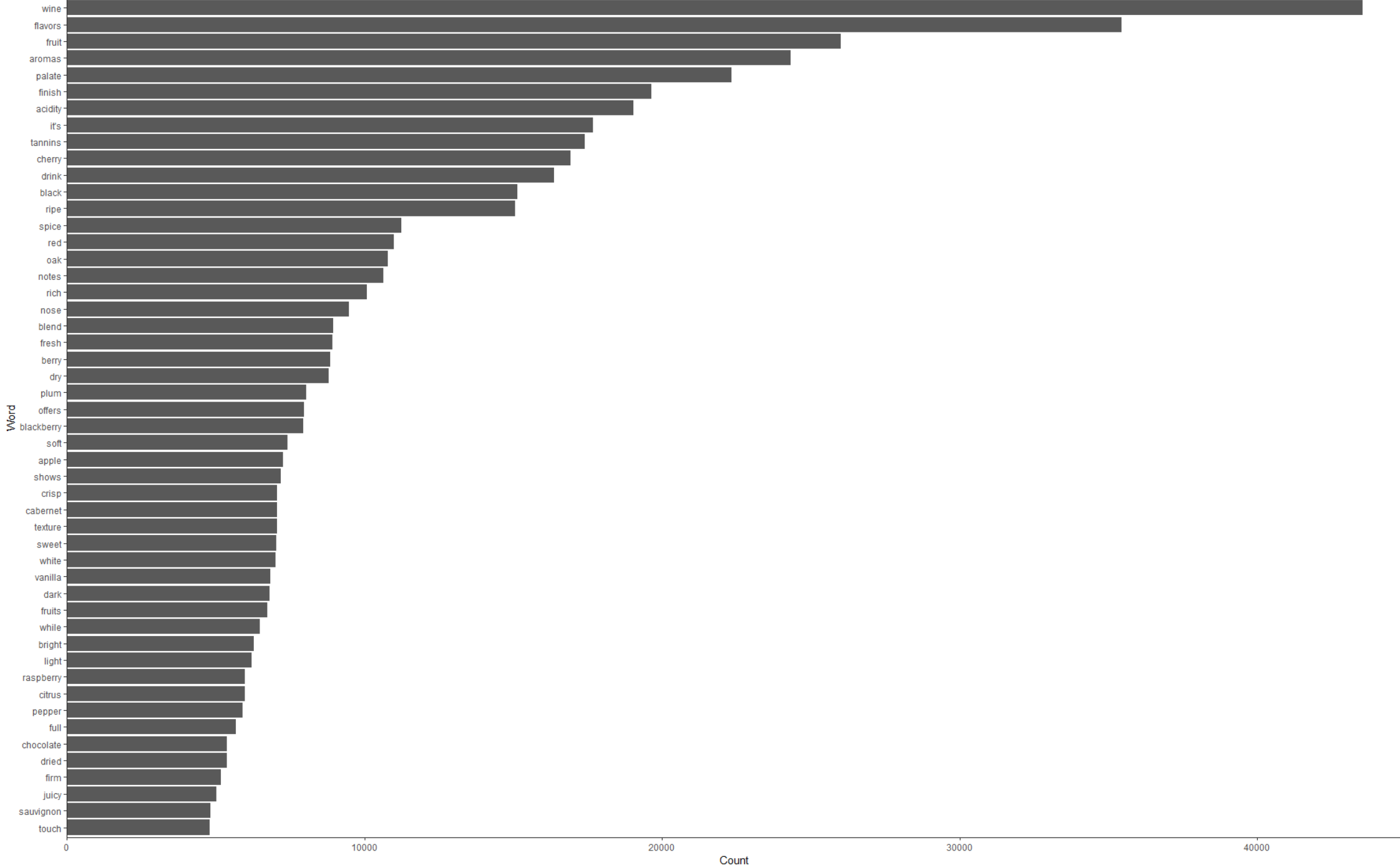
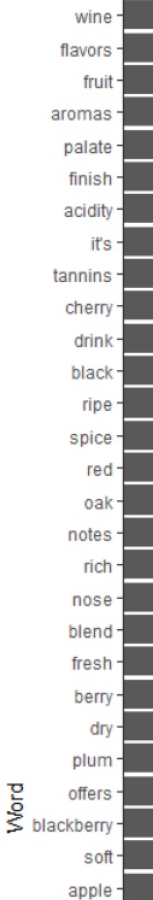
❐ U.S. ❐ France ❐ Italy ❐ Spain ❐ Other

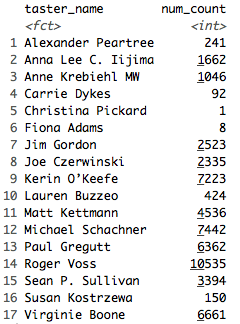
1. What is your budget? Please enter a number.

$ \_\_\_\_\_\_

Below is the detailed process of why and how we designed each option. Firstly, we make buyers to choose the wine type ( red wine or white wine) as a common practice. We use freq\_terms() to find the **top 50 words** that are used in the description and use business acumen to categorize the words into roughly four categories: ingredients(fruit, cherry, berry, etc.), aroma, acidity, taste/flavor(sweet, ripe, spice, etc.). We also eliminated the stop words and manually filtered words that are not that helpful in this case such as ‘wine’. 

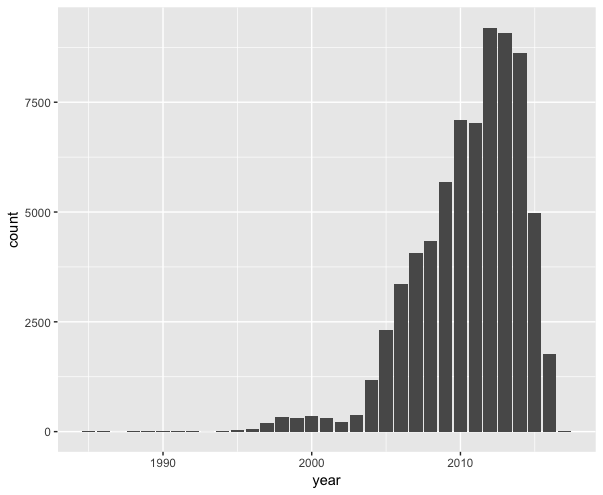
\*We also did word cloud as another visual format attached in the appendix.



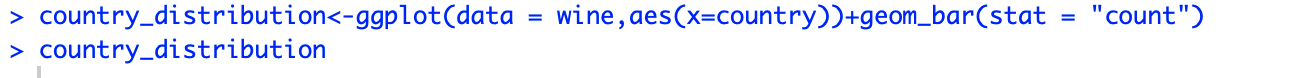
To be noted, there are only 17 wine tasters in this dataset, and some people tasted over 10,000 wine, and therefore it is reasonable assume that these 17 people are professional wine tasters. Presumably, descriptions of wine is accurate since they are given by professional wine tasters. 

For the second question, after doing some research on the choice of word in wine tasting and considering the ease of understanding, we decide to use fruit, rich, fresh, dry, sweet as choices to the question that asks about buyers’ preference of wine taste.

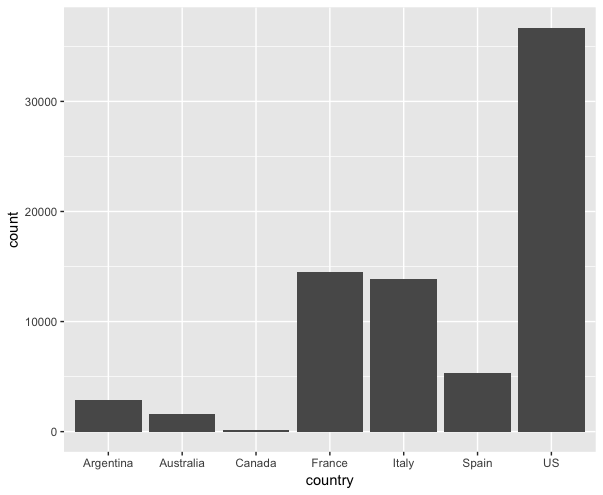
**Aging** is another factor offered to buyers to make selections. We first plot the distribution of the year and get the diagram:

We found that most of the wine is produced after 1995 and hence we use 1995, 2000, 2005, and 2010 to cut the timeframe. Buyers will be given five choices as listed in question four.



then, we apply the same steps to find where the wine is coming from, and get:



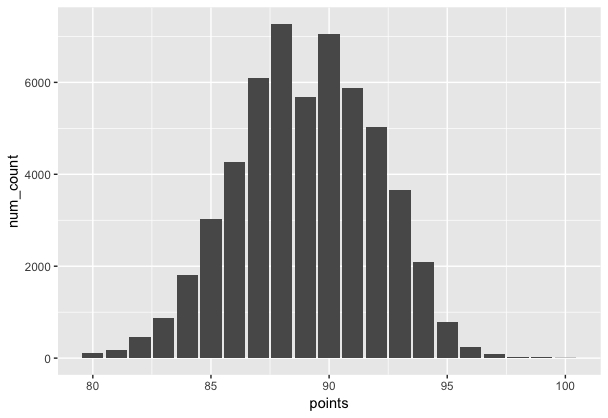
To be consistent with the number of choices given in the previous question, we pick the top four countries of origin as independent choices and combine the rest of the countries as choice “other”.

Lastly, we ask for the buyers’ budget. We will first find the best match from the database(wine dataset) according to the answers of the first four questions. The system will return a list of matching wine ranked from the most expensive to the cheapest. Now, we will start to take the budget into account. If the cheapest recommended wine is beyond budget, no result will be shown and buyers have to choose again.

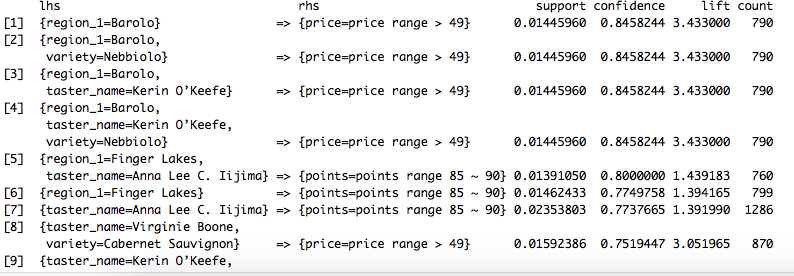
Recommender System and Association Rule

In addition to what we have done so far, we dive deeper into the dataset by implementing different techniques. We first consider recommending similar wine to the same user, however, there are 48863 unique wine out of 54635 observations. The number of repeated wine tasting that can be analyzed for such use is limited, so item-based collaborative filtering is not appropriate. Our dataset does not include information about wine tasters, hence user-based collaborative filtering is also not feasible. We have to move to other techniques.

We want to find associations among variables: points, price, region\_1, taster\_name, variety, year. The first step is to convert price and points to categorical variables, as all other variables are already categorical. We first take a look at distribution of price and found that most of the wine are below $100. Also, considering the interquartile range, we separate the price range into <20, 20 - 30, 30 - 49, >49. 

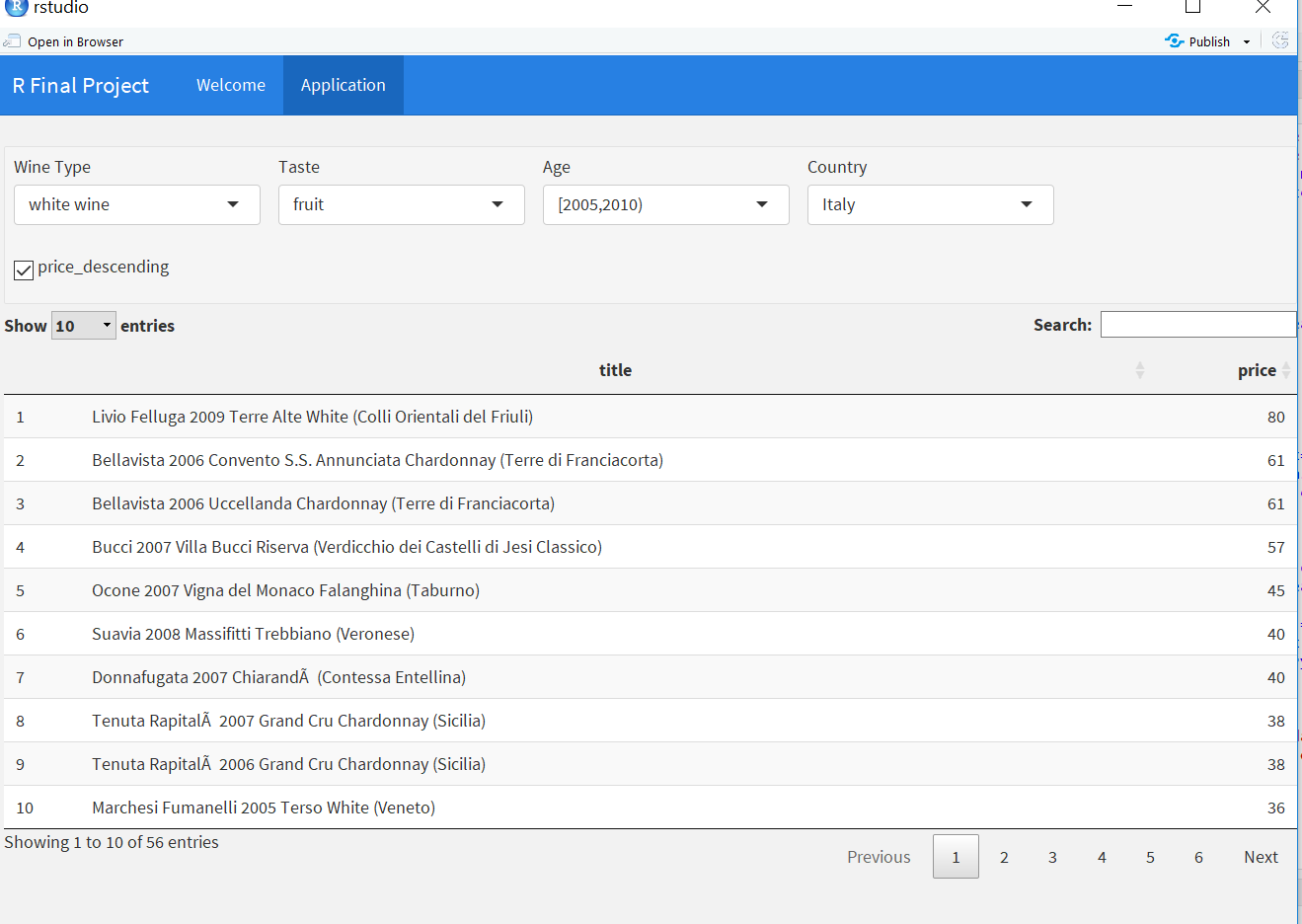


Same process can be applied to convert points into categorical data. We look at the distribution of the points and find:

We break the rating into four range: 80 - 85, 85 - 90, 90 - 95, and 95 - 100. After converting all the data into categorical, we run the apriori() function and find several associations. We look at lift as the key metric to determine if the association is strong enough. For example, output 5 indicates that wine produced in Barolo is very likely to receive points between 85-90 which is considered as the average/below average quality based on our rule of rating.

Results Display and Conclusion

Below is the example of our final application that meets our business objective that mentioned in the beginning, helping wine buyers search wine that best matches their needs and preferences. We built a dynamic dashboard in R to enable customized wine recommendation based on users selections (such as types, taste, age, and country of wine, etc.). The name and price of suggested wine will be recommended based on business users’ input and different filters. The price can be sorted from most expensive to the most affordable. We believe that our application will help business users to find the wines that they are looking for in the most efficient and user-friendly way.



In addition to applying the analytical techniques we learned in class, most importantly, we develop a consumer-centric mindset and design our functionality according to consumers’ needs. We gained first-hand experience and better understand the definition of data mining. There is a lot of information in the data and the use of data can vary depending on the business goals. For example, as we design the aging question in the questionnaire, we need the year of wine, and therefore we need to redo the data cleaning to extract the year of wine from the title variable. We get the information we need and gain insights from the information by using different processing approaches.

**Appendix**

Word cloud



Linear Regression

Decision Tree

