Feel Like Drinking?

What Do You Want to Drink Today

Xinyi Qian

Qingyuan Wang

Runnan Zhang

Iris zhao

Understanding of the Dataset

- Downloaded from kaggle
- The data was scraped from WINE ENTHUSIAST MAGAZINE during the week of June 15th, 2017.

129971
observations

14
variables

Analytical Procedures and Techniques

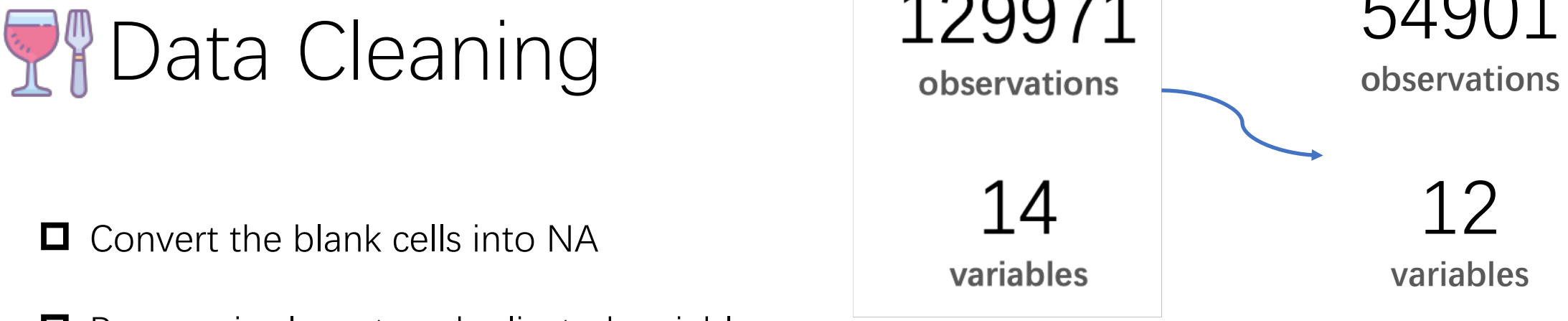
1) Analyzing the data of existing wine buyers to design a questionnaire that can best reflect the needs and preferences of potential buyers.

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.

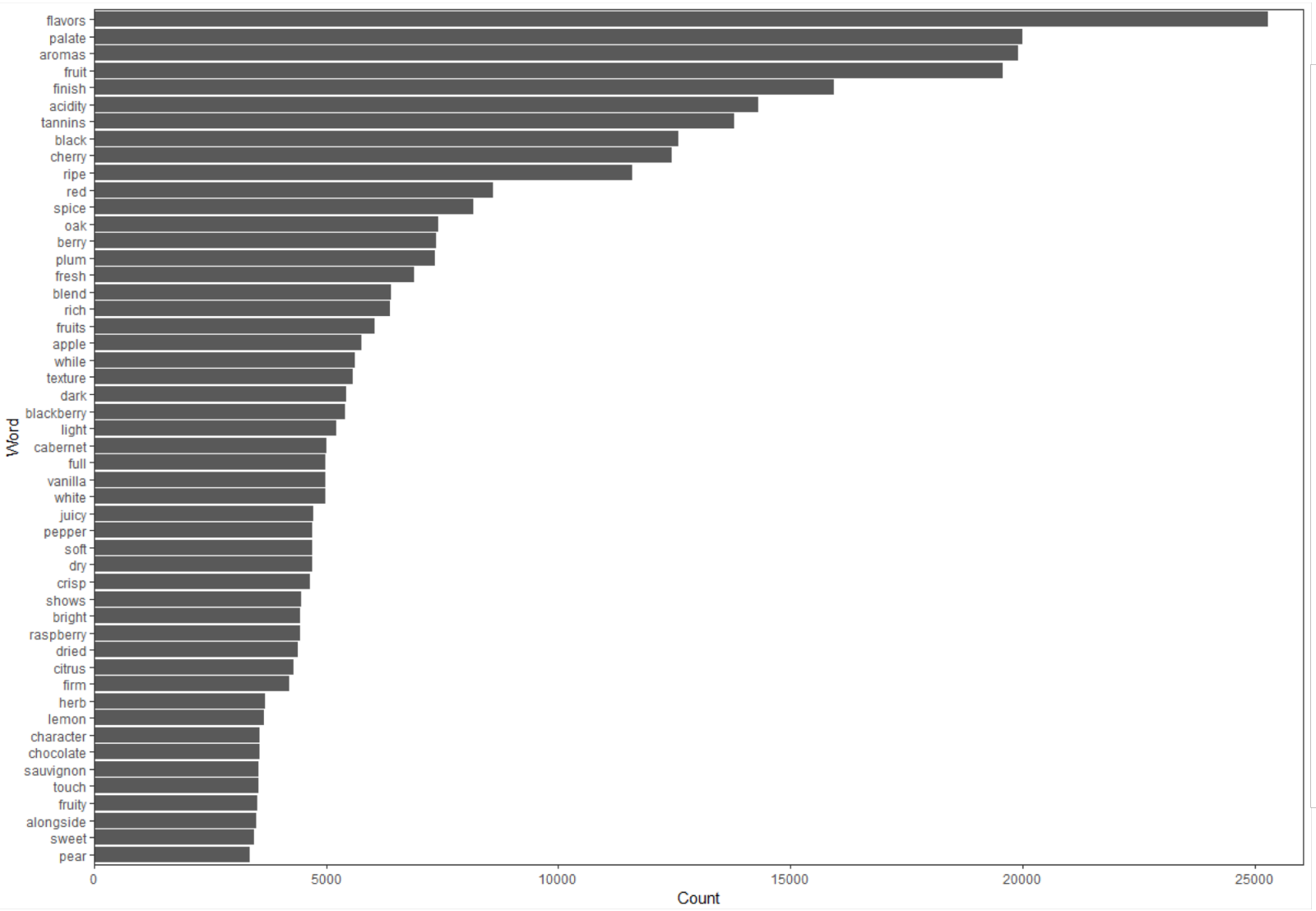
4) Trying other techniques to find useful information to help wine buyers make decision.

Data Cleaning



- Convert the blank cells into NA
- Remove irrelevant or duplicated variables
- Remove observations with NA in variable 'country', 'variety', 'designation', 'region_1', and 'taster_name'
- Extracting year from the wine's title and use it as a new variable added to the dataset
- Impute missing value
- Select variables as predictors for imputation

Most Common Words & Word Cloud



Questionnaire for Wine & Recommendation System Application

User-based Collaborative Filte

1. What type of wine are you looking for?

   ❒ White ❒ Red

2. What flavor do you want?

   ❒ Fruit ❒ Rich ❒ Fresh ❒ Dry ❒ Sweet

3. What is your preferred wine age?

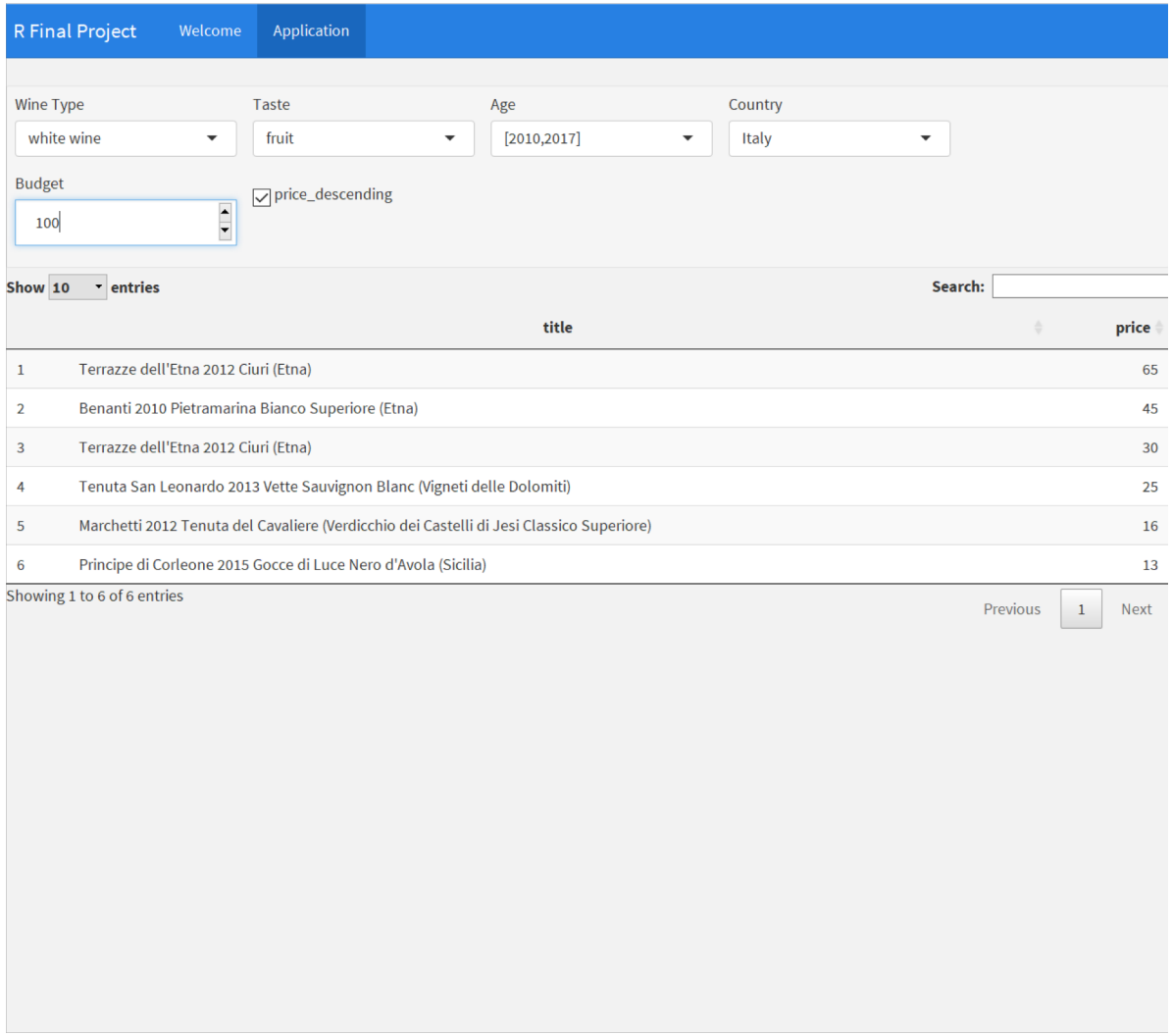
   ❒ 1995 and before ❒ 1995 - 2000 ❒ 2000 - 2005 ❒ 2005 - 2010 ❒ 2010 and after

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

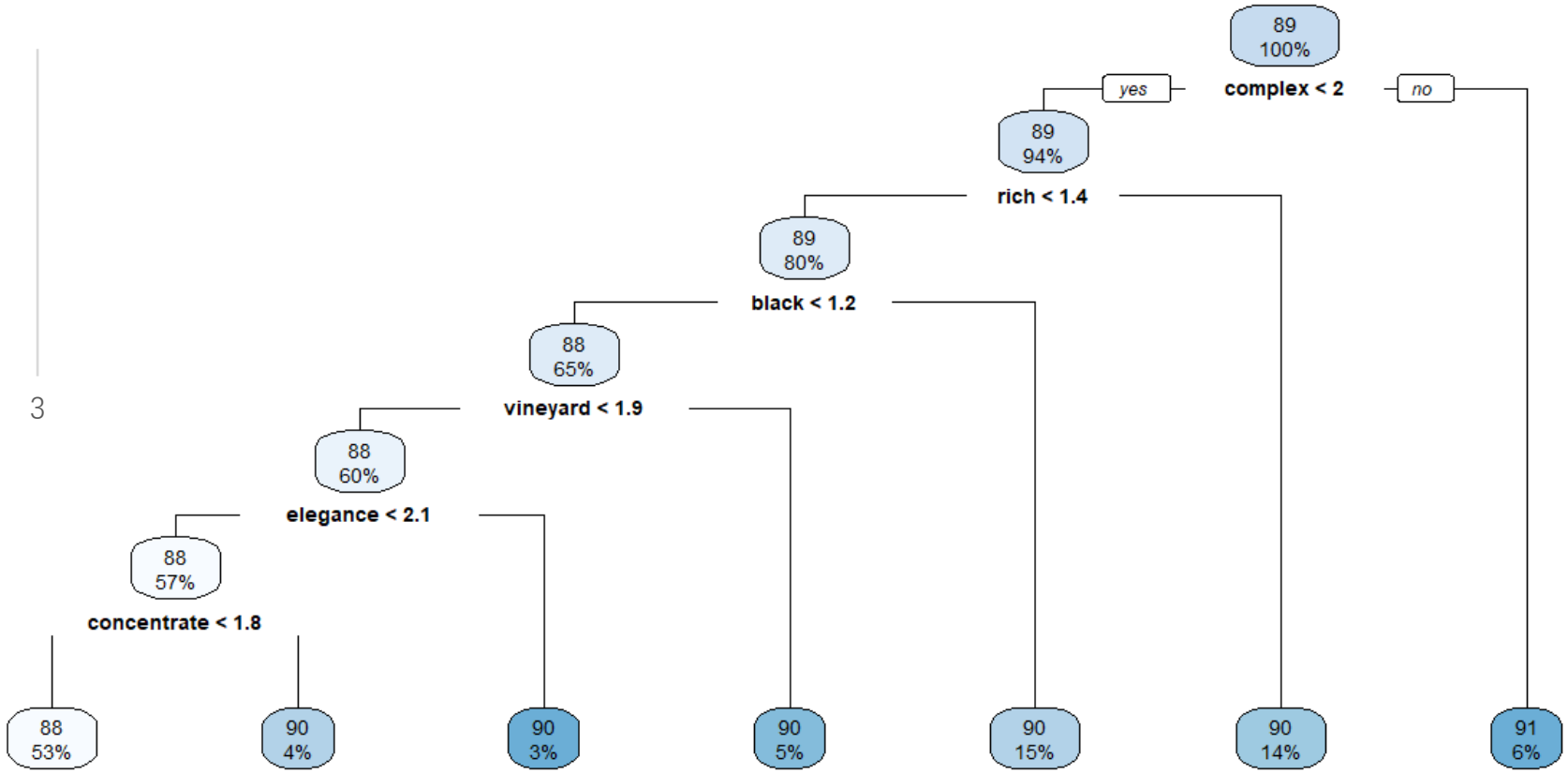
   ❒ U.S. ❒ France ❒ Italy ❒ Spain ❒ Other

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

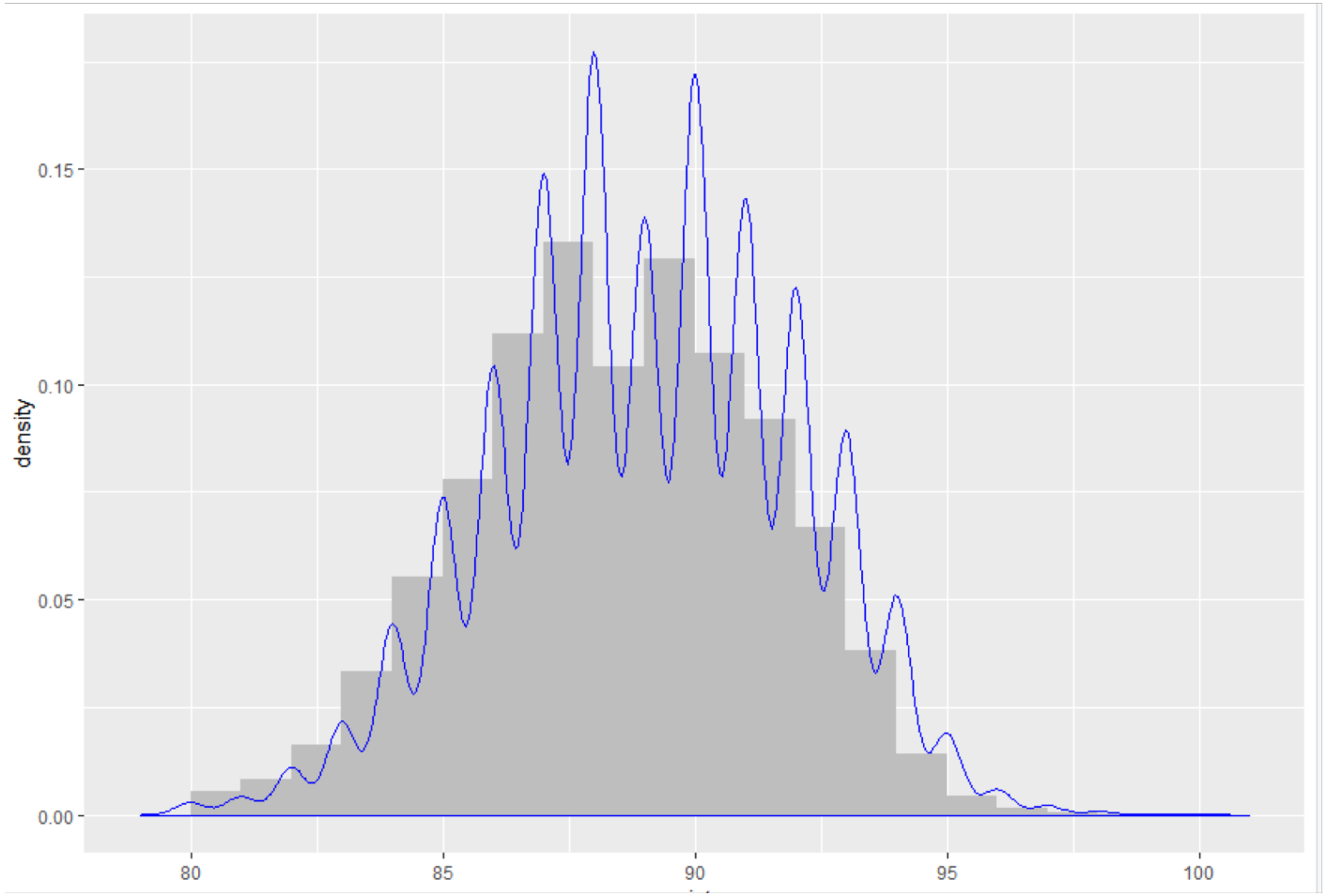
   $ _____



Most Influential Words



Length of Review & Points



cor = **0.4734295**

The longer, the better

Association rule

	lhs		rhs	support	confidence	lift	count
[1]	{region_1=Barolo}	=>	{price=price range > 49}	0.01445960	0.8458244	3.433000	790
[2]	{region_1=Barolo, variety=Nebbiolo}	=>	{price=price range > 49}	0.01445960	0.8458244	3.433000	790
[3]	{region_1=Barolo, taster_name=Kerin O'Keefe}	=>	{price=price range > 49}	0.01445960	0.8458244	3.433000	790
[4]	{region_1=Barolo, taster_name=Kerin O'Keefe, variety=Nebbiolo}	=>	{price=price range > 49}	0.01445960	0.8458244	3.433000	790
[5]	{region_1=Finger Lakes, taster_name=Anna Lee C. Iijima}	=>	{points=points range 85 ~ 90}	0.01391050	0.8000000	1.439183	760
[6]	{region_1=Finger Lakes}	=>	{points=points range 85 ~ 90}	0.01462433	0.7749758	1.394165	799
[7]	{taster_name=Anna Lee C. Iijima}	=>	{points=points range 85 ~ 90}	0.02353803	0.7737665	1.391990	1286
[8]	{taster_name=Virginie Boone, variety=Cabernet Sauvignon}	=>	{price=price range > 49}	0.01592386	0.7519447	3.051965	870
[9]	{taster_name=Kerin O'Keefe,						

Analytical Techniques

- Data cleaning:
  tidyr, mice package, cbind()
- Data visualization:
  ggplot2
- Text mining:
  freq_terms() in the qdap package, wordcloud
- Association rules:
  inspect() in the arules package and apriori()
- R Dashboard:
  flexdashboard and shiny package
- Predictive models:
  decision tree and linear regression
- User based recommender system:
  recommenderlab package