Data-driven Wine Selection

Machine Learning Summer 2019 Final Project

Introduction

- Do you feel overwhelmed when you order a bottle of wine at the restaurant or buy from supermarket or wine store?

- Wine selection is complicated and personal. The one you love doesn't necessarily depend on the rating or price. It depends on your personal taste.

- What if you can have a virtual personal sommelier who can recommend the right wine based on your taste or the wine you already tried?

- Just like how the sommelier would recommend wine for you. **This project aims to classify wine variety and identify similar wine based on taste by analyzing the descriptors written by sommeliers.**

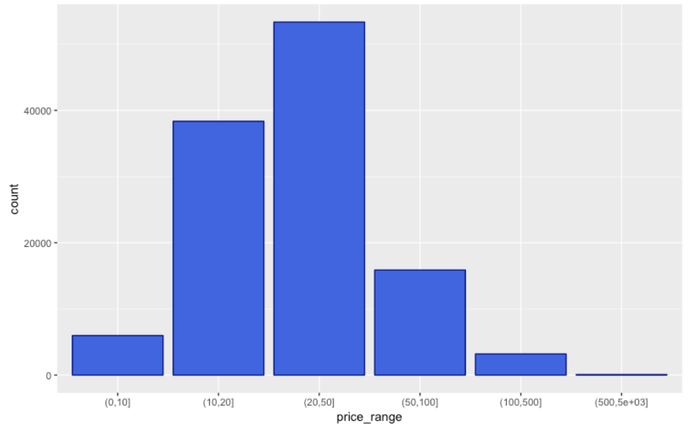
The Data

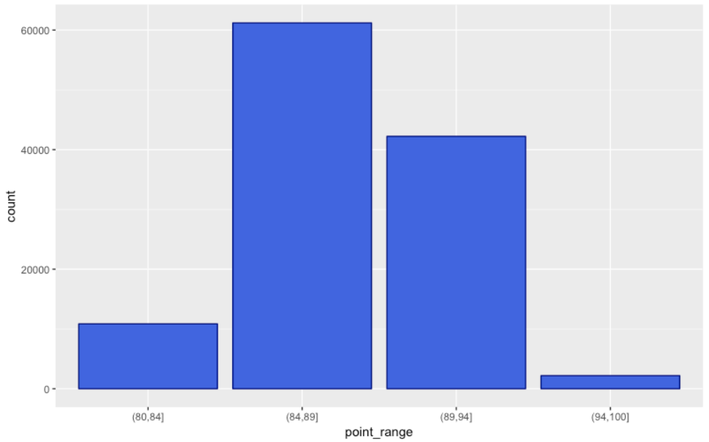
- The data source: it is a public dataset, scraped from WineEnthusiast by zackthoutt: https://www.kaggle.com/zynicide/wine-reviews.

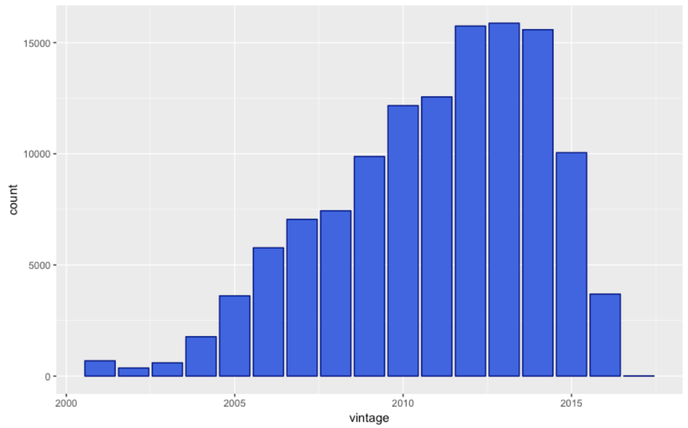
```
Observations: 129,971
Variables: 14
$ X                     <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,…
$ country               <fct> Italy, Portugal, US, US, US, Spain, Italy, France, Germany, France, US, France, U…
$ description           <fct> "Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't…
$ designation           <fct> Vulkà Bianco, Avidagos, , Reserve Late Harvest, Vintner's Reserve Wild Child Bloc…
$ points                <int> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 8…
$ price                 <dbl> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, 19, 30, 34, NA, 12, 24, 30, 13, 28, 32, 2…
$ province              <fct> Sicily & Sardinia, Douro, Oregon, Michigan, Oregon, Northern Spain, Sicily & Sard…
$ region_1              <fct> Etna, , Willamette Valley, Lake Michigan Shore, Willamette Valley, Navarra, Vitto…
$ region_2              <fct> , , Willamette Valley, , Willamette Valley, , , , , , Napa, , Sonoma, , Central C…
$ taster_name           <fct> Kerin O'Keefe, Roger Voss, Paul Gregutt, Alexander Peartree, Paul Gregutt, Michae…
$ taster_twitter_handle <fct> @kerinokeefe, @vossroger, @paulgwine , , @paulgwine , @wineschach, @kerinokeefe, …
$ title                 <fct> Nicosia 2013 Vulkà Bianco  (Etna), Quinta dos Avidagos 2011 Avidagos Red (Douro),…
$ variety               <fct> White Blend, Portuguese Red, Pinot Gris, Riesling, Pinot Noir, Tempranillo-Merlot…
$ winery                <fct> Nicosia, Quinta dos Avidagos, Rainstorm, St. Julian, Sweet Cheeks, Tandem, Terre …
```

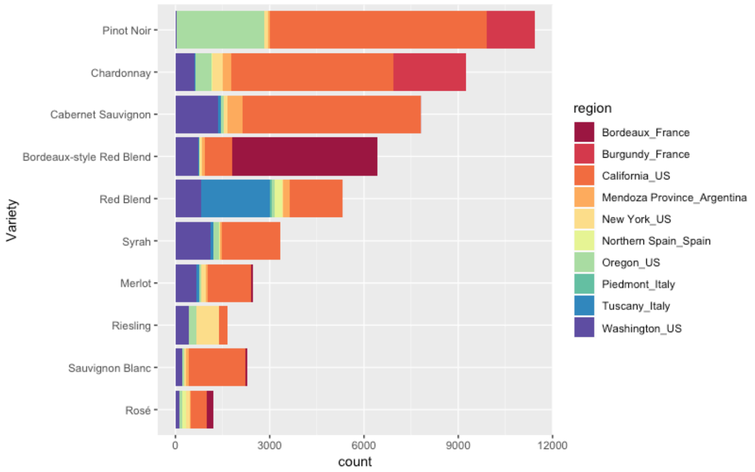
Exploratory Data Analysis - Overview

- Price range: most wine is between 10 to 100 USD
- Point range: most wine is between 84 to 94, which stands for “very good” and “outstanding
- Vintage distribution is left skewed. Vintage between 2010 to 2015 have the highest count
- Largest variety is pinot noir and California is the biggest region in many variety categories



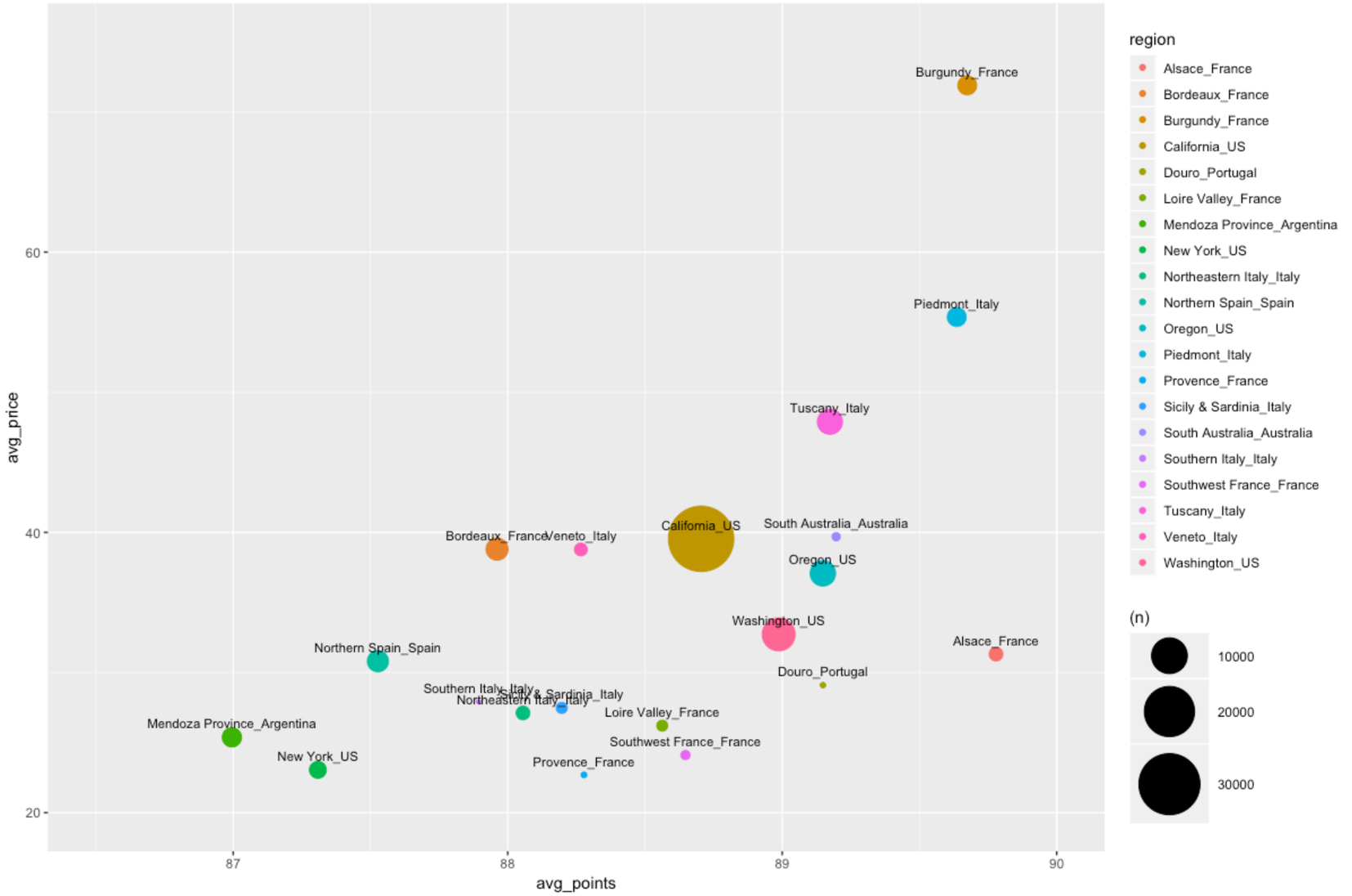






Exploratory Data Analysis - Overview_Top_Regions

- California, US has the largest count with average price & points
- Oregon and Washington, US have higher quality but lower price wine compared to California
- Burgundy, France has high end wine with highest price and points
- Alsace, France has good quality but affordable wine with highest points and lower price

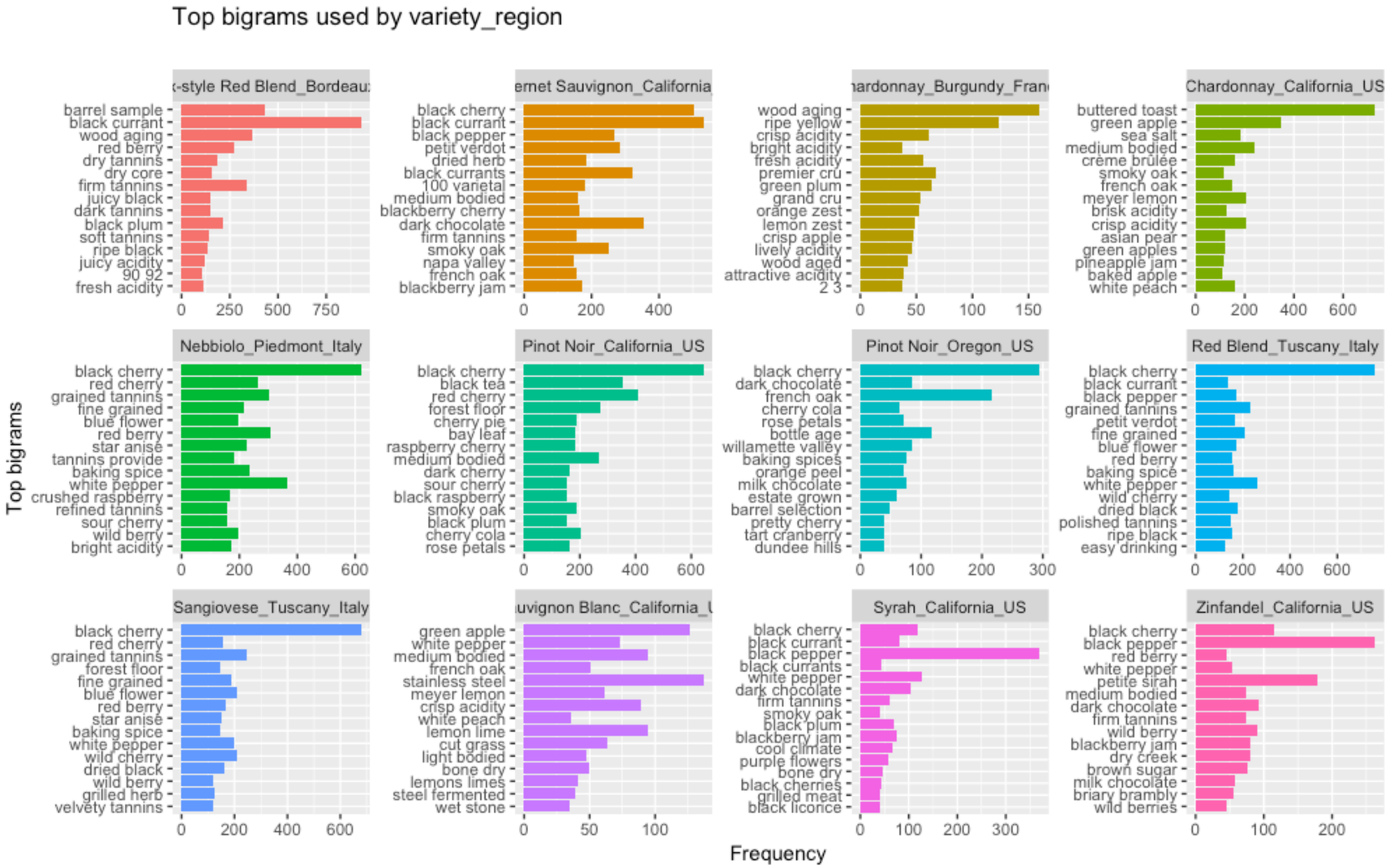


Word Cloud - Bag of Words vs N-Gram

- Bags-of-words analysis shows the top 50 single words after removing stop words
- N-Gram analysis includes sequences of adjacent words. Bigram includes 2 adjacent words
- It seems bigram words are more informative and descriptive compared to single words
- We can see the top terms that describe the distinct quality of wine such as black cherry, medium bodied, black current, firm tannins, bright acidity, baking spice, wood aging etc

Term_Frequency - N-Gram

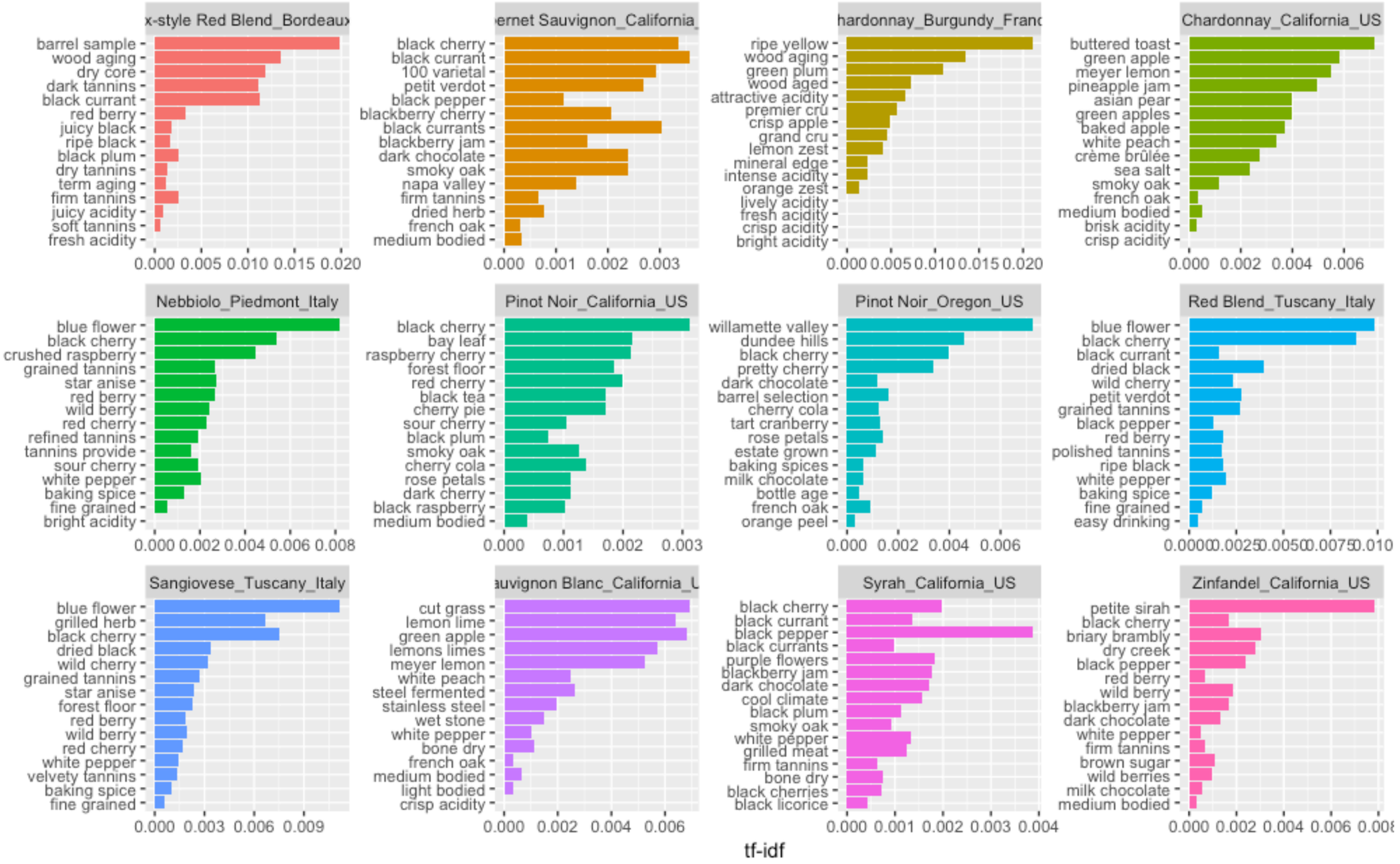
- Top 15 bigrams by variety region
- Each variety from different region has it unique descriptors.
- Chardonnay from Burgundy has terms like wood aging, ripe yellow, crisp acidity vs. from California, it is described as buttered toast, green apple, medium bodied
- Different type of wine have more different terms. Compared to Cabernet from California, Pinot Noir from Oregon has unique terms such as dark chocolate, rose petals, baking spices, orange peel etc



TF-IDF

- TFIDF, short for term frequency–inverse document frequency, is a numerical method that is intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value is proportional to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general

- After adjusting the term frequency, it shows term frequency based on the uniqueness of each variety region. For Pinot Noir from California, bay leaf and raspberry have higher rankings than n-gram frequency



Exploratory Data Analysis - Waterfall Table

Group	Observations	Variables
Original	129,971	14
Dropped	7104	0
Added	0	7
Total	122,867	21

Classification - Training/Test/Validation

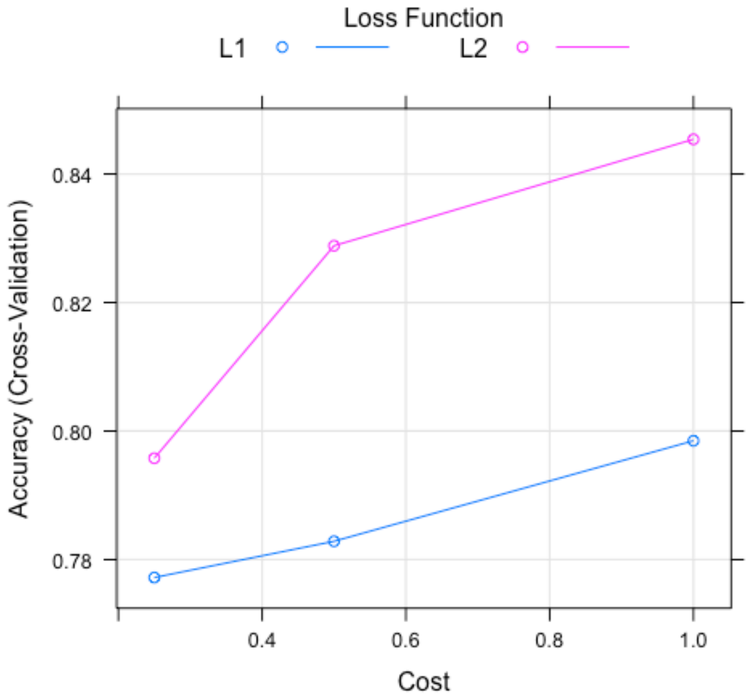
- The model is fitted to classify variety regions based on the text features per document
- To simplify the model, I selected top 10 variety regions to classify
- I used TFIDF for the modeling as it represents the uniqueness of each document better and will produce a more accurate prediction
- Due to the large amount of features and complicity of the model, the modeling dataset is a subset with 20000 rows
- Training and test dataset is split by 70/30
- Training and test are processed separately in exact same manner to ensure they have the same structure for modeling
- Training set has 14004 samples, 511 predictors and 10 classes; test set has 5996 samples 511 predictors
- Validation: 10-fold cross validation

The text processing for modeling is:

1. Tokenization
2. Lower casing
3. Stopword removal
4. Stemming
5. Transform to dfm
6. Remove sparse terms
7. Ensure test dfm has same features as train df

Modelling

- I read linear Support Vector Machine is recommended for text classification. See reference here.
- The final values used for the model is cost = 1 and loss = L2
- The final training model accuracy is 0.845




```
Overall Statistics

               Accuracy : 0.8412
                 95% CI : (0.8317, 0.8504)
    No Information Rate : 0.19
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8179

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bordeaux-style Red Blend_Bordeaux_France Class: Cabernet Sauvignon_California_US
Sensitivity                                               0.9894                                    0.8128
Specificity                                               0.9807                                    0.9543
Pos Pred Value                                            0.8809                                    0.7590
Neg Pred Value                                            0.9984                                    0.9664
Prevalence                                                0.1259                                    0.1506
Detection Rate                                            0.1246                                    0.1224
Detection Prevalence                                      0.1414                                    0.1613
Balanced Accuracy                                         0.9851                                    0.8835
```

- Compared to the test data, the overall accuracy rate is 0.841
- Bordeaux Red Blend has the highest accuracy rate 0.985
- Syrah California US has the lowest accuracy rate 0.775
- Some of the classes have quite low sensitivity

Wine Recommendation - Cosine Similarity

- To create a wine recommendation function, I use cosine similarity function to calculate the distance between document
- It is common to use cosine to evaluate the distance of text and document
- In the example, the reference wine is Cookies & Cream 2010 Cabernet Sauvignon and the most similar wine is Beaulieu Vineyeard 2010 Cabernet
- The accuracy of the result will need further investigation with tasting

```
> get_similar_wine(similarities, 10000)
                  Cookies & Cream 2010 Cabernet Sauvignon (California)
                                                                1.0000000
Beaulieu Vineyard 2010 Coastal Estates Cabernet Sauvignon (California)
                                                                0.3795163
                    Dante 2010 Reserve Cabernet Sauvignon (California)
                                                                0.3574675
                         Tom Gore 2013 Cabernet Sauvignon (California)
                                                                0.3356415
                         Bigvine 2006 Cabernet Sauvignon (Napa Valley)
                                                                0.3314776
```