

# 'Choose Your Wine' Application

## by Ekaterina Galin

### Part I. Introduction

With the wide choice of wine in the market, it is becoming more and more difficult to choose a bottle that delivers high value in terms of quality and price. This application helps wine lovers choose a bottle of good quality wine that better suits their requirements in terms of price, country of origin and affintage. With the help of this application, they can find a bottle of great value, as well as get information about the bottles presented in the dataset. Wine enthusiasts can see the countries that produce most wine, the most popular types of grape used in the wine making process, wine producers that got the highest score from the critics and so on.

### Part II. Application Overview

Let us see how the application works. To the left, there are 6 tabs, 3 sliders and 1 select box.

The screenshot shows the 'Select Wine Parameters' panel. It contains three sliders: 'Price' (range 10 to 50), 'Score' (range 80 to 100), and 'Year' (range 1,999 to 2,017). Below the sliders is a 'Select Country' box with a list of countries: US, Italy, France, Spain, Portugal, Chile, Austria, Australia, Germany, and Argentina.

The 3 sliders under 'Select Wine Parameters' help set the price range, score given by wine critics and the year of wine production. The Select Country Box gives a possibility to choose the country of origin. By default, all 10 biggest wine-producing countries are chosen when the application is launched. In order to delete a country, one needs to click on the country in the box and press 'Delete' button on the keyboard. In order to add the deleted country again, one needs to choose the country from the drop down menu and click on it. It is important to delete countries one by one with some time space, otherwise the application might crash.

The graphs and tables in the application are reactive, which means that respond to the changes in the sliders on the left-hand panel.

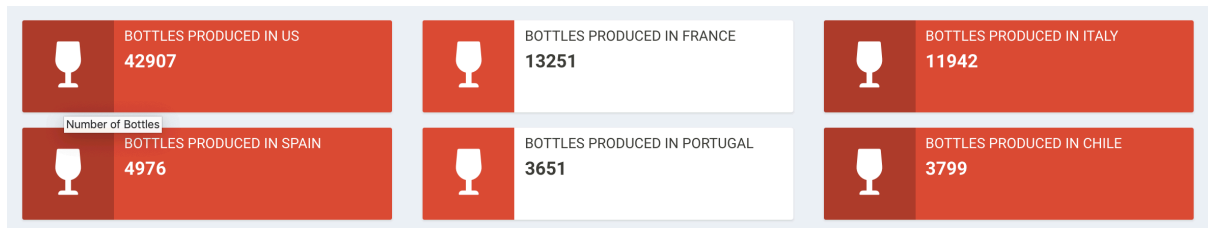
#### Tab 1. Introduction

This tab welcomes visitors, gives information about the purpose of the application, some guidelines on its usage and a link to the LinkedIn page of the author (which happens to be me). In this section, HTML tags are used to customize text, as well as add a picture and a hyperlink to the web page.

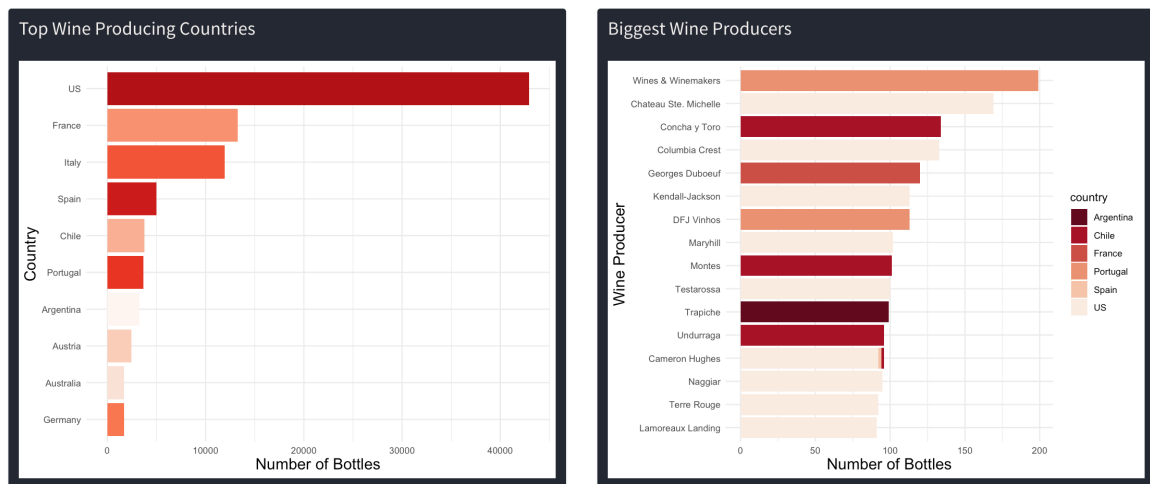
The screenshot shows the application's main menu with six tabs: Introduction, Fact File, Wine Comparison, Choose Your Wine, Wine Taste, and US States View.

#### Tab 2. Fact File

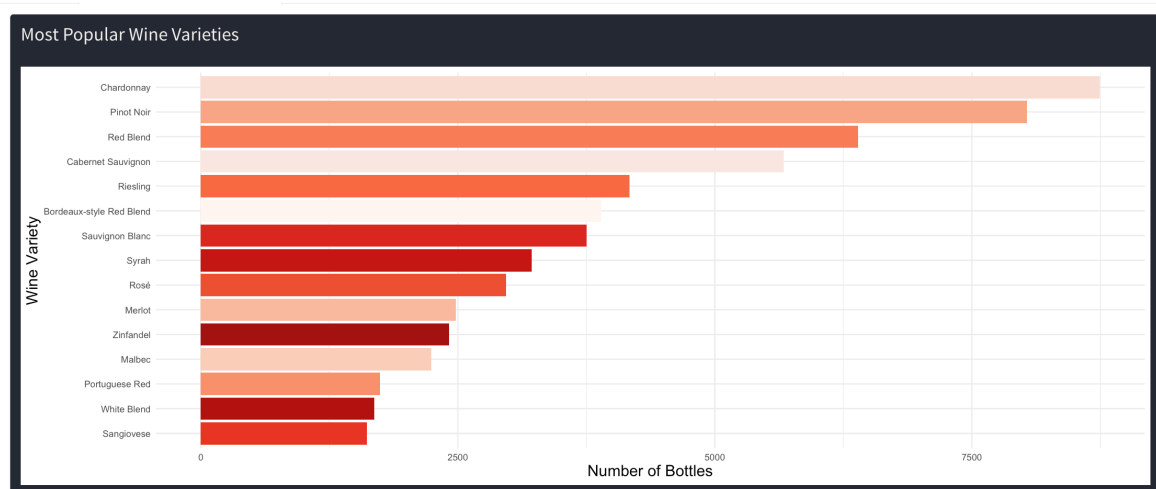
This tab gives general overview of the dataset and is further divided into 3 tabs. At the top of the page one can find information about the number of bottles produced in top 6 countries and represented in this dataset.



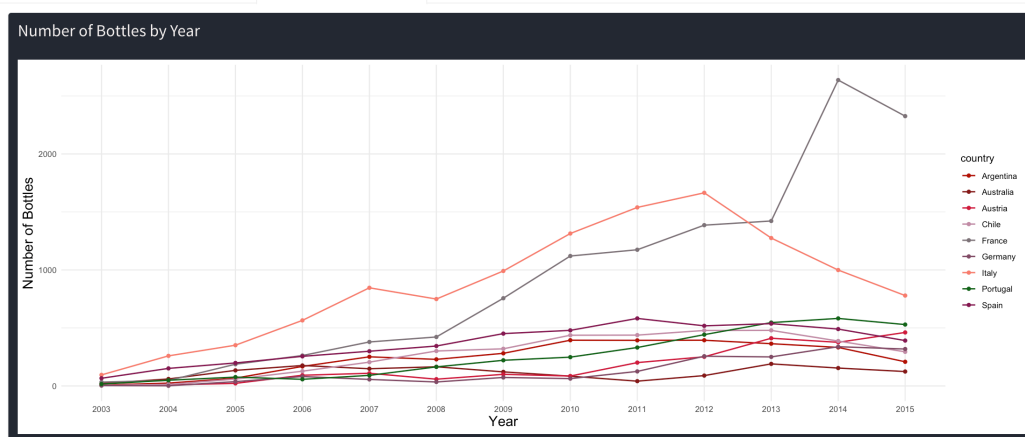
The **Overview Tab** gives information about the number of bottles produced in each of the 10 countries, as well as presents biggest wine producing companies which are colored according to the country of origin. Since the dataset is very big, the data on wine producers is filtered on the condition that there are more than 90 bottles produced by the company. The most difficult part in this code was to find the way to extend the color palette as in the original palette there were only 9 colors, while I required 12.



The **Most Popular Wine Varieties Tab** gives information about the number of bottles by each wine variety. Since the dataset is big, the data is filtered on the condition that there are more than 1500 bottles of wine of the same variety.



**Number of Bottles By Year Tab** shows the number of bottles produced each year in each country, starting from 2003. Information about bottles produced in the US is filtered out as it greatly exceeds the number of bottles produced in other countries and shifts the other lines.



As initially there was no information on the year of production, I had to extract this information from wine name ('title' column) and attach it as a separate column to the existing dataset. For example, I managed to extract the year 2013 from the following title: 'Nicosia 2013 Vulkà Bianco (Etna)'.

The code for this manipulation is the following:







```
yearExtract <- function(string) {
  t <- regmatches(string, regexec("[0-9]{4}", string))
  sapply(t, function(x) {
    if(length(x) > 0){
      return(as.numeric(x))
    } else {
      return(NA)
    }
  })
}

# Creating data frame with title and vintage year
title <- wine_data$title
distributor <- wine_data$distributor
wine_tidy2 <- data.frame(title, distributor)
wine_tidy2$vintage_year <- yearExtract(as.character(wine_tidy2$title))
str(wine_tidy2)
head(wine_tidy2)

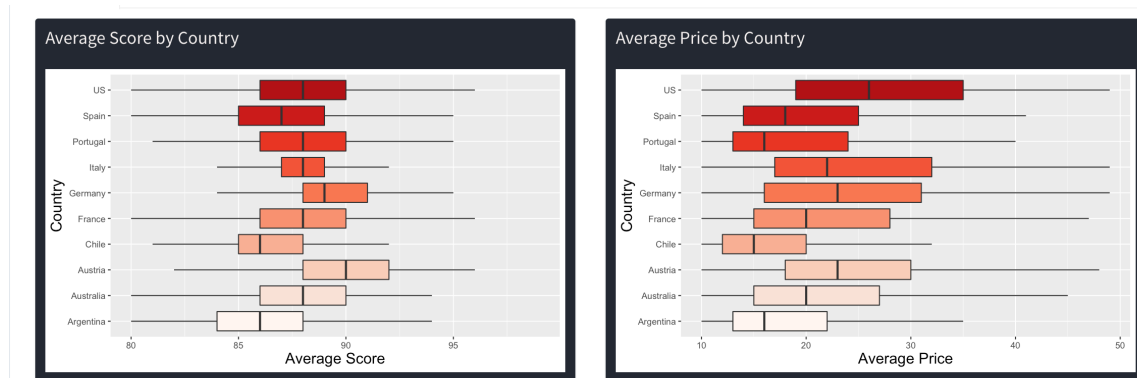
# Adding new year column to the existing wine dataset
wine_data_1 <- cbind(wine_data, year = wine_tidy2$vintage_year) |
` ``
```

### Tab 3. Wine Comparison

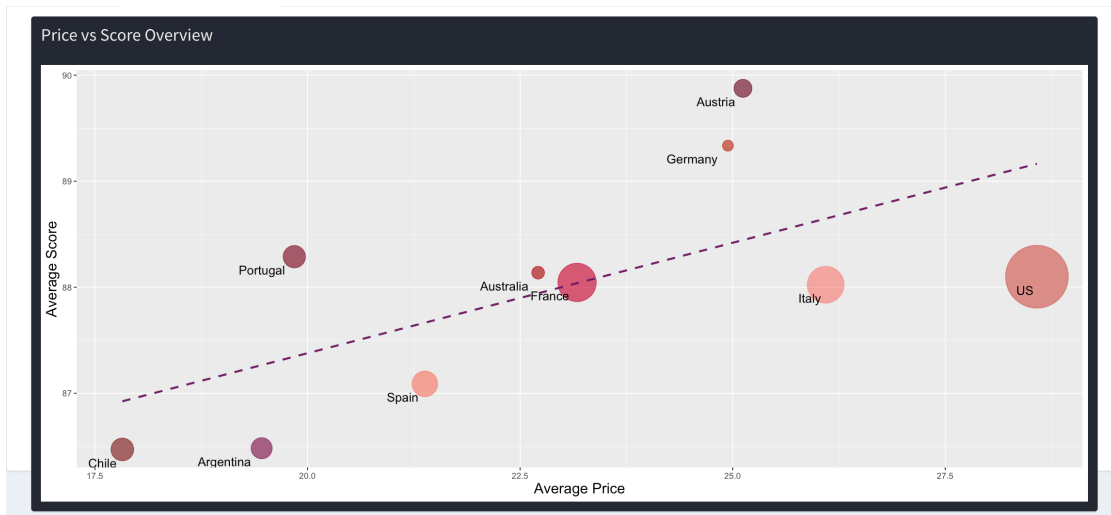
This tab gives information about the value wine delivers (score / price ratio). The value boxes on top serve as reference points as they give general information (average, minimum, maximum price & score) across all the 10 countries.

 AVERAGE PRICE (\$) <b>26</b>	 MINIMUM PRICE (\$) <b>10</b>	 MAXIMUM PRICE (\$) <b>50</b>
 AVERAGE SCORE <b>88</b>	 MINIMUM SCORE <b>80</b>	 MAXIMUM SCORE <b>99</b>

**Points & Price Tab** gives information about average score and price by country in the form of box plots. App users can compare this data with the average price and score of all the bottles represented in the dataset (value boxes on top of the page).



**Price vs Score Overview Tab** shows a scatter plot and represents correlation between mean of price and mean of score by country. It also shows a linear relationship between the two variables, fitting a linear model  $\text{lm}()$ . The plot clearly shows that Austrian & German wines have better value compared to wines produced by other countries (high average score and relatively low average price) as well as Portuguese wine that are cheap and highly scored. This was one of the most time-consuming graphs to build as I needed to scale the circles representing the number of bottles. However, the name of the country was also scaled, so it took me some time to figure out how to solve this problem.



**Top Rated Varieties & Wineries Tab** contains two tables and gives information about the average, minimum and maximum score of wine by winery and country as well as the number of bottles represented in the dataset.

Top Rated Wineries

Show  entries

Search:

	Winery	Average	Min	Max	N_bottles
1	Dr. Loosen	91	86	96	66
2	Foxen	91	85	96	62
3	Gary Farrell	91	82	95	73
4	Iron Horse	91	86	96	69
5	Marimar Estate	91	85	95	63

Showing 1 to 5 of 94 entries

Previous **1** 2 3 4 5 ... 19 Next

Top Rated Varieties

Show  entries

Search:

	Variety	Average	Min	Max	N_bottles
1	Blafränkisch	90	82	95	179
2	Grüner Veltliner	90	83	96	1039
3	Sangiovese Grosso	90	84	95	217
4	Aglianico	89	81	95	230
5	Alvarinho	89	83	93	111

Showing 1 to 5 of 79 entries

Previous **1** 2 3 4 5 ... 16 Next

#### Tab 4. Choose Your Wine

This tab helps choose wine by selecting different parameters (country of origin, year, price, score) with the help of sliders. It also gives information about the value the bottle of wine delivers. The value column is a mutated column and is calculated as score divided by price, with 9 being the highest score/ price ratio, and, thus, delivering best value. By default, the table is arranged by score descending. However, it could be arranged by price, year or value in a descending or ascending order.

#### Tab 5. Wine Title?

My initial idea was to apply a text mining technique to the 'Description' column which to visualize the adjectives that wine connoisseurs use to describe wine taste. However, the data set appeared to be so big, that the graph kept crashing the application. One by one I applied all the tips given by you (starting from filtering the data at the initial stage and finishing by changing parameters of the word cloud). However, the deployed application kept crashing.

Then I decided to apply the same technique to the 'Title' column to see which words are most frequently used in the titles of wines. Though, this graph does not make as much sense as the previous one, it works and shows how textual data from the 'Title' column was structured, tokenized and counted in order to create the word cloud.



### Tab 6. US States Overview

According to the information gathered from the previous manipulations, we found out that US is the biggest wine producer in the world. That is why this tab focuses on the wine produced in the US.

Application users can click on a US state and get information about the number of bottles produced in this state and represented in the dataset, number of unique wine varieties produces in the state, the average price and score by state. The information is easy to understand as it is visualised on a map and color code is used to present values.

This graph is the only one that is not reactive and does not depend on the changes of the sliders on the left-hand panel.

