Exploratory Data Analysis

Wine Reviews Dataset

Authors:

* David Jîtcă (407)
* Iordache Ștefan (411)

Contents

[1. Motivation 2](#_Toc30000723)

[2. Data Exploration – The Road to Wine Expert 3](#_Toc30000724)

[2.1. Dataset Description 3](#_Toc30000725)

[2.2. Duplicates 4](#_Toc30000726)

[2.3. Features 5](#_Toc30000727)

[2.3.1. Points 5](#_Toc30000728)

[2.3.2. Price 6](#_Toc30000729)

[2.3.3. Points – Prices Correlation 7](#_Toc30000730)

[2.3.4. Geo Analysis 8](#_Toc30000731)

[2.3.5. Title & Description 11](#_Toc30000732)

[3. Models 13](#_Toc30000733)

[3.1. Logistic Regression 13](#_Toc30000734)

[3.2. Support-Vector Machine (SVM) 14](#_Toc30000735)

[4. Conclusions & Ideas 15](#_Toc30000736)

# Motivation

We wondered how we could create a predictive model to identify which wines are we drinking without opening the bottles. Blind tasting is a discipline for master sommeliers but even normal people can do it with little to no help from our models.

In order to obtain such a model and some great results it was necessary to understand in depth the dataset that we are going to work with. Some exploratory data analysis was necessary and we tried our best to find which features are the most important and what kind of columns we can drop without noticing any other flavors in our wines from those expected.

Such a project might open a path to advanced systems. For example, using our models we can predict the right price for a bottle of wine before buying it or we can verify if a certain drink have the variety and the taste we would like based on what we are drinking. This model might be used in intelligent vending machines were we would like to buy the wine we desire, based on aromas, color or price.

Without any further introduction we are going to squeeze every bit of information out of the dataset we chose to work with, [Wine Reviews from Winemag](https://www.kaggle.com/zynicide/wine-reviews). It is a rich dataset and we are going to deeply understand it and extract complex features and conclusions.

In the next chapters we are going to search and crunch our computers through the data. Also, we are determined to test a large variety of learning algorithms based on different perspectives, such as Logistic Regression, Support Vector Machines (SVM), Regressors, Random Forests, etc.

# Data Exploration – The Road to Wine Expert

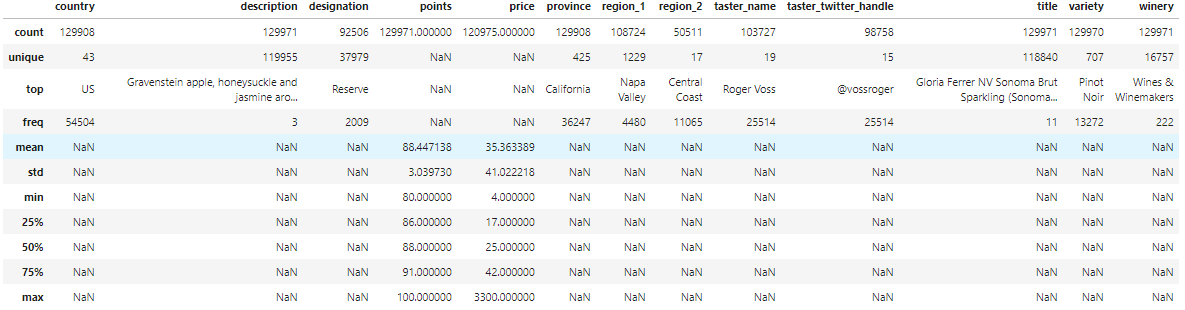
In order to begin our journey we need to slice our data and perform an in-depth autopsy on the dataset.

First of all, let’s see the shape of the data we are dealing with. The dataset contains 3 files by default (without further data extraction and scraping):

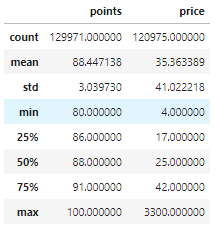
* *“winemag-data-first150k.csv”* (this is the first version)
* *“winemag-data-130k-v2.csv”* (this is the augmented version of the input, without duplicates and with better preprocessing on the scrape part)
* *“winemag-data-130k-v2.json”* (raw version of the second version, as a .json file)

We are going to use the second version of the dataset because it contains more data (one extra column) and it is better preprocessed, without containing any duplicates.

## Dataset Description



Dataset Description



Numeric values description

We can easily conclusion that the only continuous variable in our input is the price because of the diversity in the values. The maximum value for the price column is $3300, the minimum is $4, but we can also see that 75% of the wines are valued under $42.

The points are a good feature and those are integer values between 80 (for bad bottles) and 100 (for premium wines).

## Duplicates

The shape of the dataset is (129971, 13) which means that we are dealing with close to 130.000 records of data, with 13 extracted observations for each row.

|  |  |
| --- | --- |
| Column | Null Values |
| Country | 63 |
| Description | 0 |
| Designation | 37465 |
| Points | 0 |
| Price | 8996 |
| Province | 63 |
| Region\_1 | 21247 |
| Region\_2 | 79460 |
| Taster\_name | 26244 |
| Taster\_twitter\_handle | 31213 |
| Title | 0 |
| Variety | 1 |
| Winery | 0 |

The most missing values are in the following columns:

* Region\_1
* Region\_2
* Designation
* Taster\_twitter\_handle
* Taster\_name
* Price

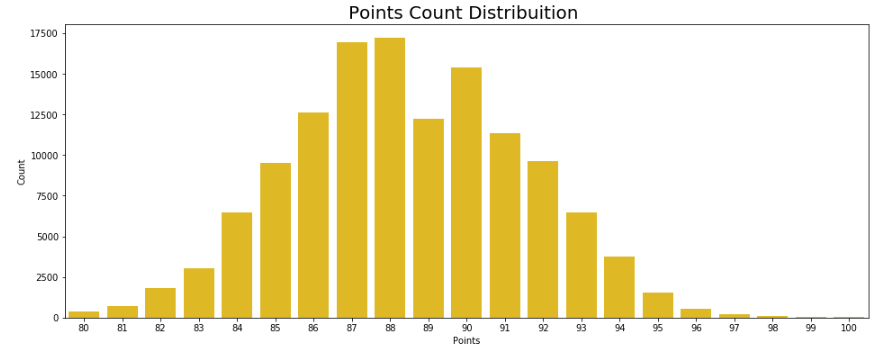
For now we are going to drop from our analysis the taster because we should build a model that does not fit on person names. Region\_1 & Region\_2 columns will be included later.

The most concerning part is that we deal with a large amount of wines without prices. We can’t build an accurate prediction model with undeclared prices. We are going to drop all of the rows without prices listed.

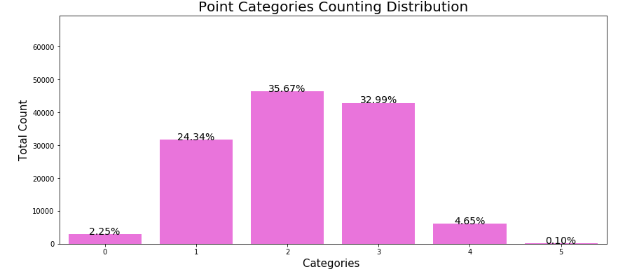
## Features

### Points

Let’s start with the distribution of the points column from our dataset. As stated before, all of our values are between 80 and 100. Just a few number of wines have less than 82 points and more than 95 points and this means that we deal with a natural behavior regarding the reviews.



Let’s group the points by ranges: 80-83, 83-87, 87-90, 90-94, 94-98, 98-100. Now we can clearly understand the data and confirm the facts exposed above.



We can also describe this column by using a formula for outliers. Using standard deviation we can define an interval for real data and outliers. We can consider the lower bound as mean – 3 \* standard deviation and the upper bound as mean + 3 \* standard deviation.

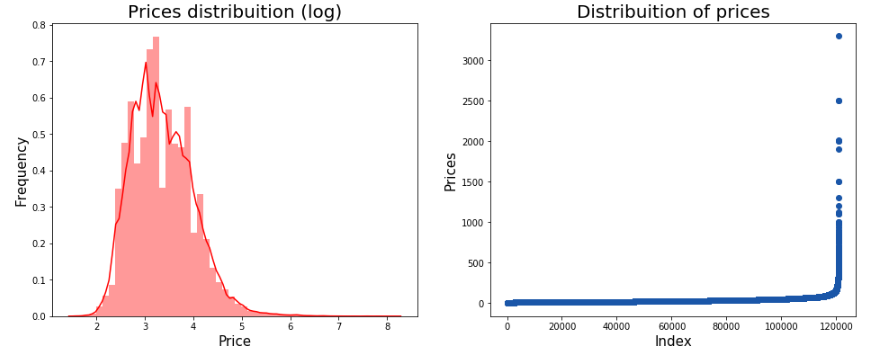
By applying the formula to the points column we managed to extract only 129 outliers from 130k entries. This means that only 0.0994% are outliers and only 129 wines have more than 98 points.

### Price

We are going to explore this column in a similar way as we did with points. First of all, the distribution might give us some insight.

The first figure is the log distribution of prices and it gives us an impression that the data is normally distributed.

Also, there are few high values, most wines being situated in a fair interval of prices.

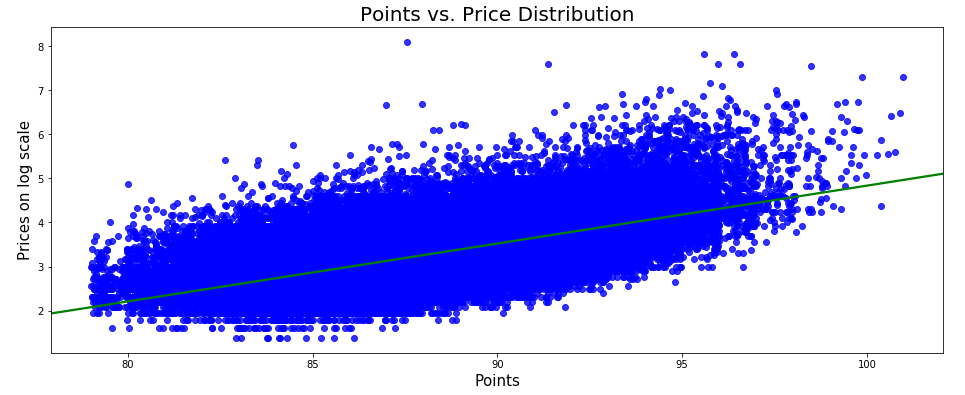


Now, let’s take a look at the outliers and their percentage. We obtained with the same formula defined 1177 outliers from a total of 130k reviews. This means that we have less than 1% of outliers and we can further use prices as a valuable feature.

### Points – Prices Correlation

Let’s dive in deep and visualize the correlation between points and price columns. This might give us a better understanding of the dataset and some ideas about what can we achieve with this dataset.

First of all, we are going to simply plot the distribution of points (X axis) and prices (Y axis, log scaled).



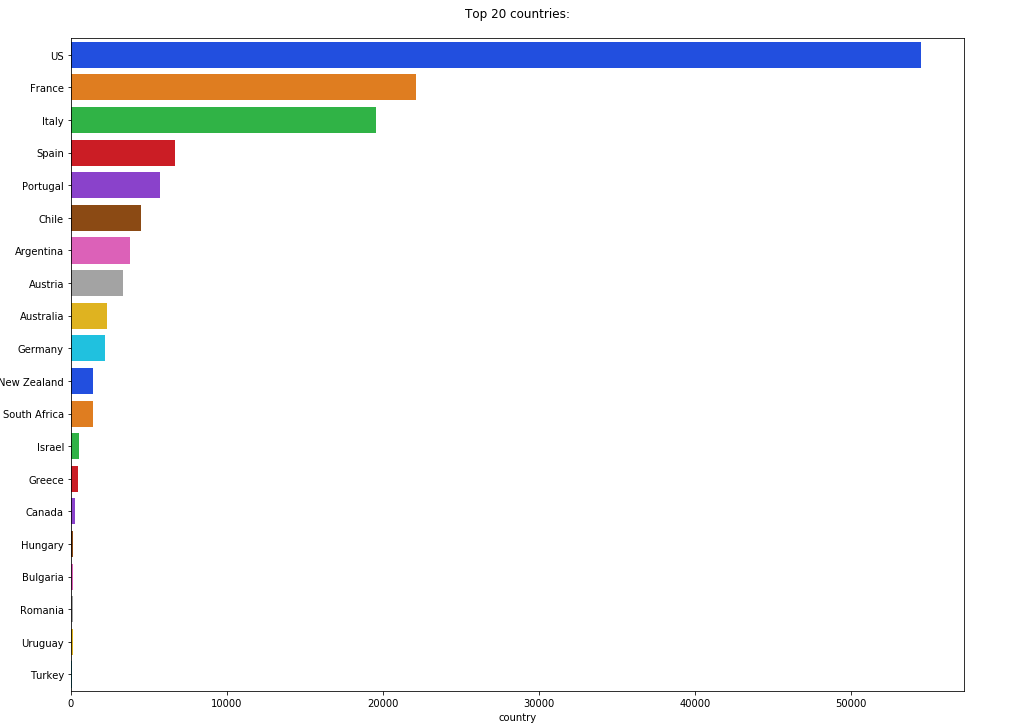
This scatter plot is very meaningful as it states some clear and interesting observations:

* The most expensive wines are rated between 87 and 90 points, so the quality doesn’t directly reflect in the price.
* The highest price isn’t of the wine with highest rating.
* We are looking at a possible regression line.

The results mentioned above clearly suggests the possibility of building a recommendation system to find the cheapest wines with similar quality and from the same variety.

### Geo Analysis

Let’s take a look at some geographical distributions. We are going to focus on the countries of provenience since we can’t deduce with high accuracy the regions for unlabeled data. We are going to find some interesting facts about wines in general and also about this particular dataset.

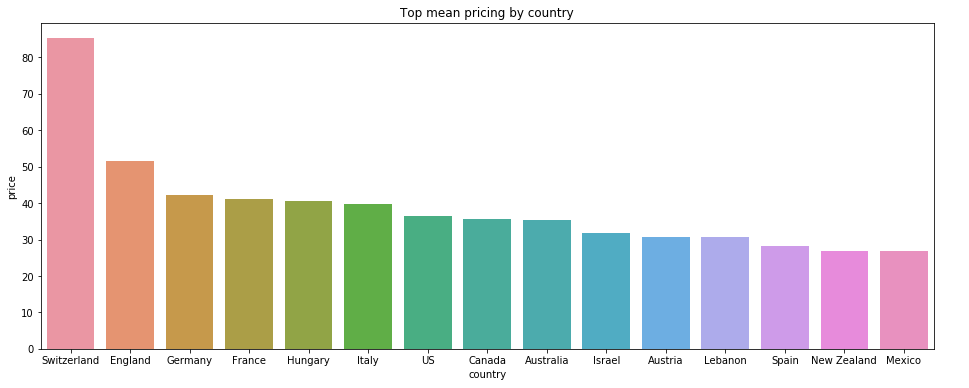


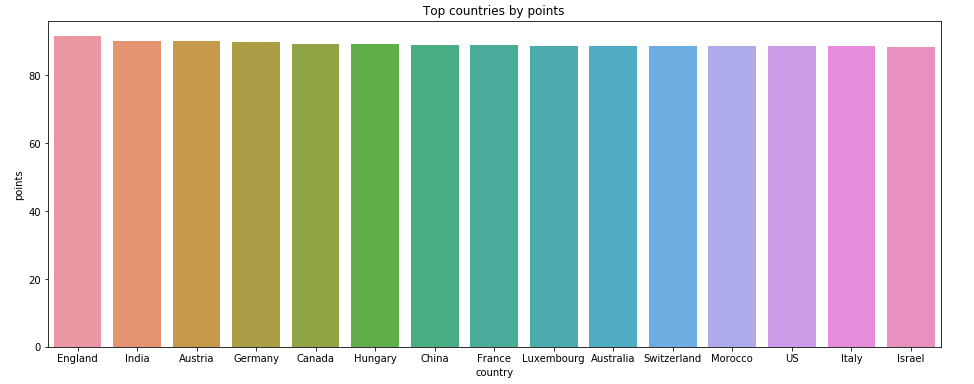
As we can see, there is a big surprise for us. We were expecting to discover France, Italy, Argentina and Chile as the wine manufacturers worldwide, but there is another clear winner for this set of data: United States (US).

After some investigations we found that the source of this dataset mostly consists of US based reviewers. Taking it further let’s try and inspect the relationship between countries, prices and points.

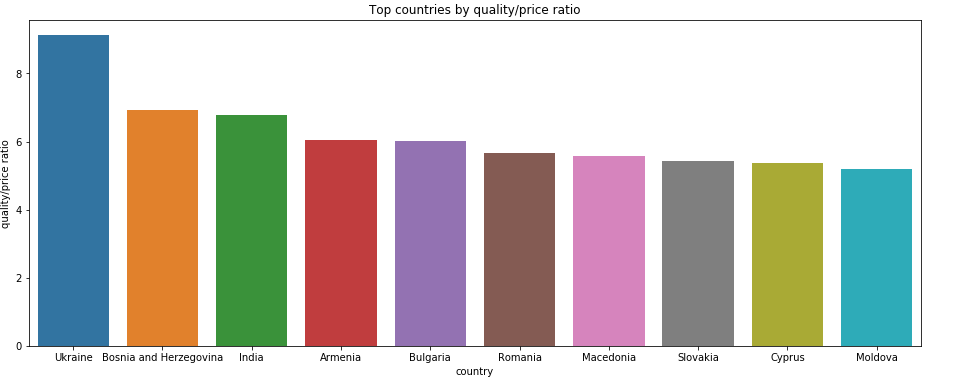
Here we have some big surprises. As we know Switzerland isn’t such a cheap country and wine prices does confirm the fact. With an average price on wine of over $80 it is, by far, the most expensive country when it comes to drinking alcohol. England is another country that offers a higher mean price than the international market.

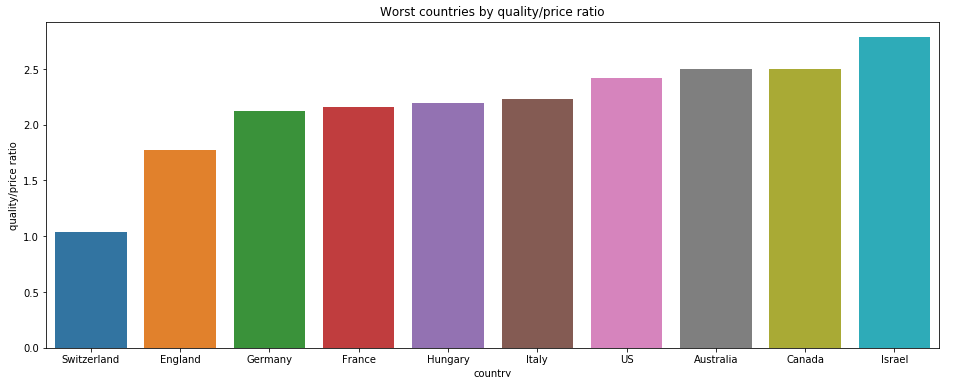
In contrast, we listed the countries by the mean number of points. England seems to be the winner but there are little to no differences when we compare it to other countries. Instead, India comes on the second place when we take into account only the quality. This is a big surprise because India does have medium prices.





Without further introduction we are going to prove if the deductions are right by plotting top countries by points/price ratio. A high score on this scale means that a country does produce good wines without robbing the visitors and clients.





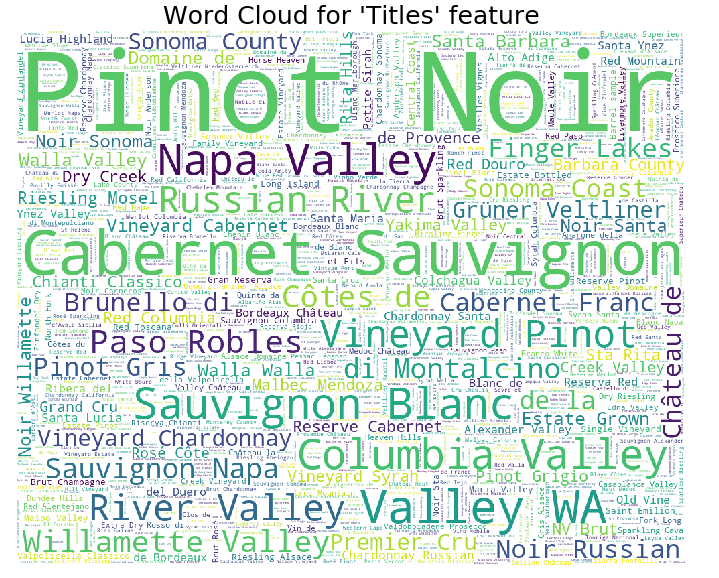
I guess that the graphics does speak by themselves. Ukraine, Bosnia & Herzegovina, India, Armenia, Bulgaria and Romania are the best countries if you want to taste the best wines of the world without paying extra cash.

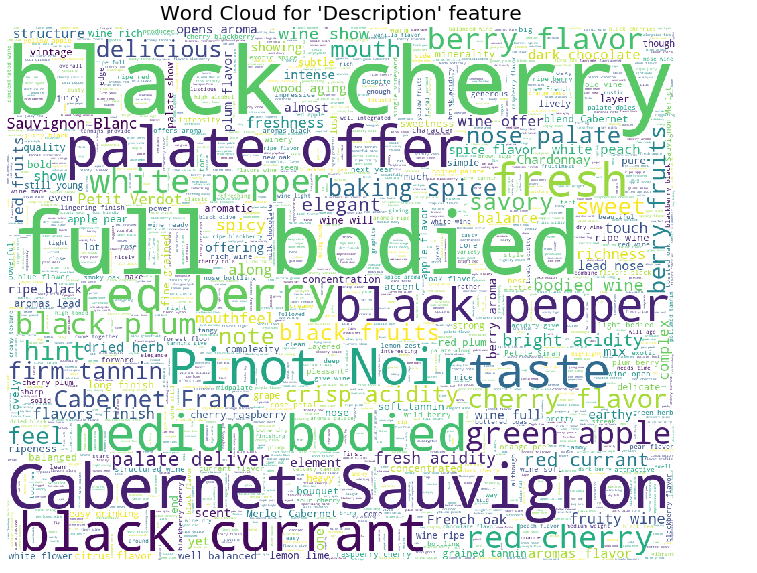
In contrast Switzerland and England are the worst choices from the selection. Well developed countries tend to raise the prices even if the quality remains the same. Those are the results we were looking for in order to train good networks and models.

### Title & Description

After we extracted some useful features and conclusions from numerical and geographical data we are ready to split the text parts: title & description, which contains far more data to be parsed and processed.

We are dealing with a large quantity of data and texts so we are going to visualize the most valuable words by using word clouds.

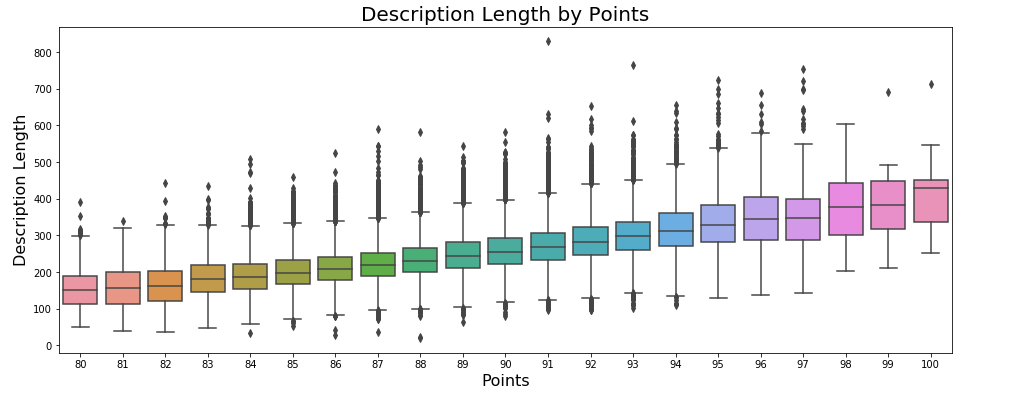




For title we got what we expected, wine variety: from Pinot Noir to Cabernet Sauvignon. For the description column we can see a little bit of a mix between varieties, fruits, aromas and specific keywords from this domain: “full bodied”, “palate”, “crisp acidity”.

Remember, those cloud words contain little to no processing. We deleted useless words called stopwords like certain verbs.

Another interesting fact is the relationship between the rating of a certain wine and the length of the review. It is a linear dependency between those two values and we can certainly use this indicator for a prediction system. Below we can visualize the box plot.



# Models

## Logistic Regression

We are going to predict the variety based on description using Logistic Regression. It is a statistical model that uses a logistic function to model class dependent variables, most cases being binary variables.

In our case we will try to predict the varieties for wines that have a presence of over 200 apparitions in our dataset. Due to computational limits we are going to use only 50k reviews but we can apply the same models very easily on the whole dataset later.

Our varieties are categorical so we can try create a model using Logistic Regression. The model is going to compute decision boundaries between classes and using only a fair amount of features (in our case the description) we managed to obtain about 58% accuracy.

As mentions, this method is sensitive to outliers and the description might be saturated with such data. Also, logistic loss doesn’t reach zero even if the point is classified with a high degree of confidence. This leads us to degradation in accuracy.

In order to obtain results we used ‘softmax’ as internal function, in extension to the classic binary classificatory. We are going to obtain probabilistic values, also leading to lower accuracy.

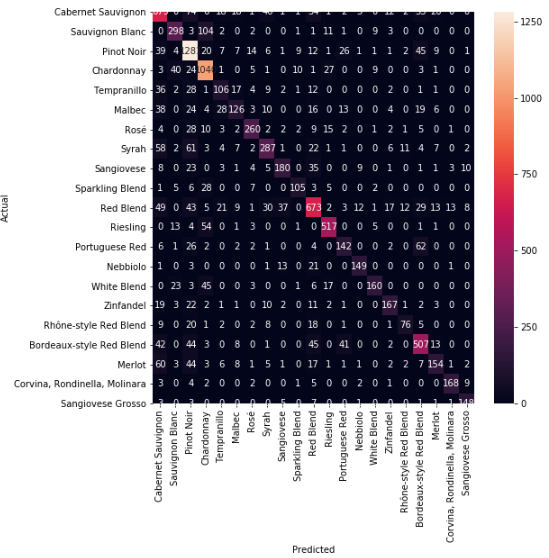
The model is pretty naive and it might be improved by using more data, encoded for easy usage.

## Support-Vector Machine (SVM)

In order to compare Logistic Regression to other techniques we are going to use a different approach: SVMs or Support-Vector Machines.

The objective of this algorithm is to find the hyperplane that has the maximum margin in an N-dimensional space, N being the number of features that distinctly classifies the data. It produces categorical values, not probabilities and by using the Radial Basis Function (RBF) we can accommodate higher dimensions, which is perfect for our dataset.

It is dedicated for classification tasks and it does give us better results. Those results were expressed as a confusion matrix.



# Conclusions & Ideas

We can finally conclude that we can build a system that predicts wines even from bare descriptions, made by specialists.

We can use simple to implement algorithms or we can dive in and use even neural networks for certain tasks. The dataset is rich in information and it can be combined with other datasets from Kaggle if we would like to create a more complex model.

By using Logistic Regression and SVMs we were able to obtain some good predictions with little to no preprocessing. Also, we would like to state again that those algorithms were fed with one column, the description part.

Also, it is important to mention that even if we did the training on one column we used the whole dataset to filter and analyze what we are dealing with.

As a general conclusion we would like to say:

Don’t drink and drive!

Cheers!