Exploratory Data Analysis

Wine Reviews Dataset

Authors:

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

Contents

[1. Motivation 2](#_Toc29971273)

[2. Data Exploration – The Road to Wine Expert 3](#_Toc29971274)

[2.2. Dataset Description 3](#_Toc29971275)

# 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

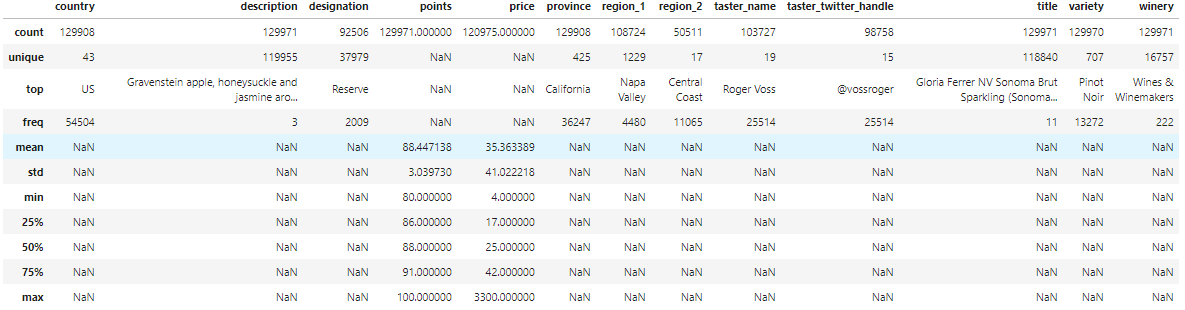
In order to begin our journey we need to slice our data and perform an in-depth autopsy on the dataset.

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):

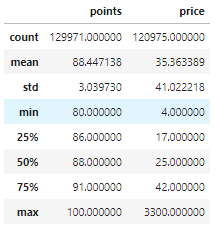
* *“winemag-data-first150k.csv”* (this is the first version)
* *“winemag-data-130k-v2.csv”* (this is the augmented version of the input, without duplicates and with better preprocessing on the scrape part)
* *“winemag-data-130k-v2.json”* (raw version of the second version, as a .json file)

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

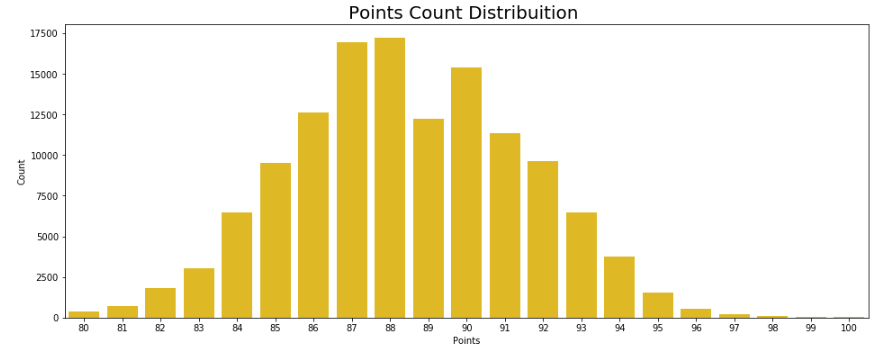
For now we are going to drop from our analysis the taster because we should build a model that does not fit on person names. Region\_1 & Region\_2 columns will be included later.

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.

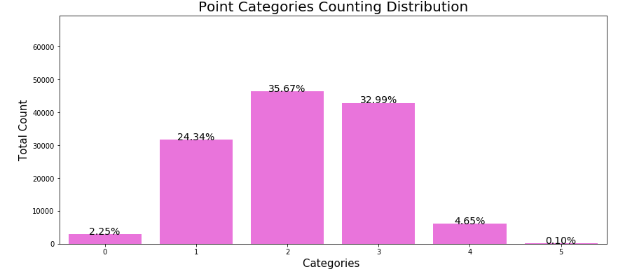
## Features

### 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.

