Wine rating and price: patterns, trends, and insights.

Sebastiano Denegri

December 12, 2020

Beer is made by men, wine by God.
— Martin Luther, circa 1500s

Exploratory Data Analysis for Machine Learning – IBM / Coursera.

Table of contents

1. Introduction

1.1. Scope of the project

In this project I've analyzed data about wine to better understand patterns and trends related to this product category; in particular, I want to find insights and patterns in relation to:

- Wine rating
- Wine price.

I've adopted a **descriptive analytic approach**, to unlock potential insights and hidden correlations between data attributes.

For the scope of this analysis, I used datasets from Kaggle.com, containing 4 files for each wine type: red, white, rose, and sparkling.

Source of data: Vivino.com.

1.2. Data Understanding – Dataset description & attributes' summary.

In order to build a useful dataset for our descriptive analysis, I concatenated the 4 sets of data (red, white, rose, and sparkling.) into one. Before doing this, I added one additional field ("Style") in each dataset.

First 5 observations:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style
0	Pomerol 2011	France	Pomerol	Château La Providence	4.2	100	95.00	2011	red
1	Lirac 2017	France	Lirac	Château Mont-Redon	4.3	100	15.50	2017	red
2	Erta e China Rosso di Toscana 2015	Italy	Toscana	Renzo Masi	3.9	100	7.45	2015	red
3	Bardolino 2019	Italy	Bardolino	Cavalchina	3.5	100	8.72	2019	red
4	Ried Scheibner Pinot Noir 2016	Austria	Carnuntum	Markowitsch	3.9	100	29.15	2016	red

Last 5 observations:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style
13829	Special Cuvée Brut Aÿ Champagne N.V.	France	Champagne	Bollinger	4.2	37765	46.00	N.V.	sparkling
13830	Brut Premier Champagne N.V.	France	Champagne Premier Cru	Louis Roederer	4.2	40004	36.48	N.V.	sparkling
13831	Impérial Brut Champagne N.V.	France	Champagne	Moët & Chandon	4.1	76037	40.61	N.V.	sparkling
13832	Brut (Carte Jaune) Champagne N.V.	France	Champagne	Veuve Clicquot	4.2	86839	43.60	N.V.	sparkling
13833	Brut Champagne N.V.	France	Champagne	Dom Pérignon	4.6	94287	170.00	N.V.	sparkling

The wine dataset contains **13,834 observations with 9 attributes: 6 object-types and 3 numeric-types, either floats or integers.**

Attributes and data types are as follows:

- Name: Name of wine. Data type: object (text). 10,934 categories (some observations share the same name).
- Country: Origin country of wine. Data type: object (text). 33 categories.
- Region: Origin region or province of wine. Data type: object (text). 861 categories.
- Winery: Origin winery. Data type: object (text). 3,505 categories.
- Rating: Average rating. Data type: float. Range: 2.2 - 4.9.
- "NumberOfRatings": Number of people who rated the wine. Data type: integer. Range: 25 - 94,287 reviews.
- Price: Price in EUR currency. Data type: float. Range: 3.15 - 3,410.79 euro.
- Year: Year of production. Data type: object (text). 34 categories.
- Style: Style of wine. Data type: object (text). 4 categories.

2. Methodology

2.1. Plan for Data Exploration

1. Data Cleaning.
   Purpose: preparing the data for the Descriptive Analysis.
   Steps:
   a. Missing data
   b. Duplicates
   c. Outlier Analysis
   d. Other inconsistent data.
2. Exploratory Data Analysis (EDA).
   Purpose: get an initial feeling of the data; determine if data needs further cleaning and transformations.
   Steps:
   a. Descriptive Statistics - Summarize main data characteristics.
   b. Data Mining - Uncover relationships between variables; extract/select important variables for potential modeling.
3. Data Preparation: preparing the data for potential modeling (feature engineering and variable transformation).

2.2. Data Cleaning

2.2.1 Missing data and duplicates

The dataset didn't contain neither missing values (in any of the attributes) nor duplicates, although we already noticed that some wines shared the same name.

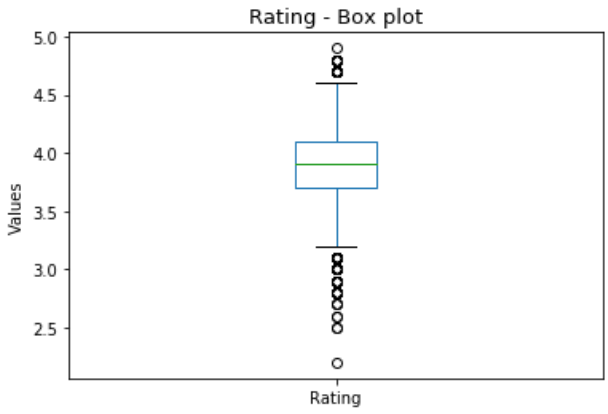
After performing a deeper analysis, the conclusion was that different observations with the same name can have same "Country", and "Year" values, but being produced by a different Winery (or being produced in a different Region in very few cases). I, then, created a new attribute, "Unique Name", merging Name, Winery, and Region values. Eventually, only 2 duplicates were found and discarded.

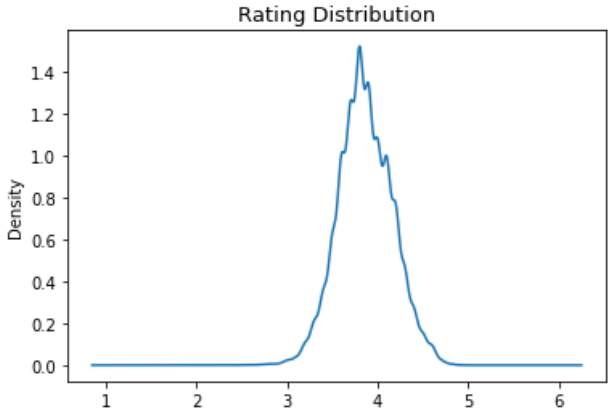
After dropping the duplicates, the dataset had **13,832 observations.**

2.2.2. Outlier Analysis and other inconsistent data

I performed an Outlier Analysis (using the IQR method) in the numeric data type columns: Rating, Number of Ratings, and Price.

- **Rating**





Data has (almost) a normal distribution, with some outliers:

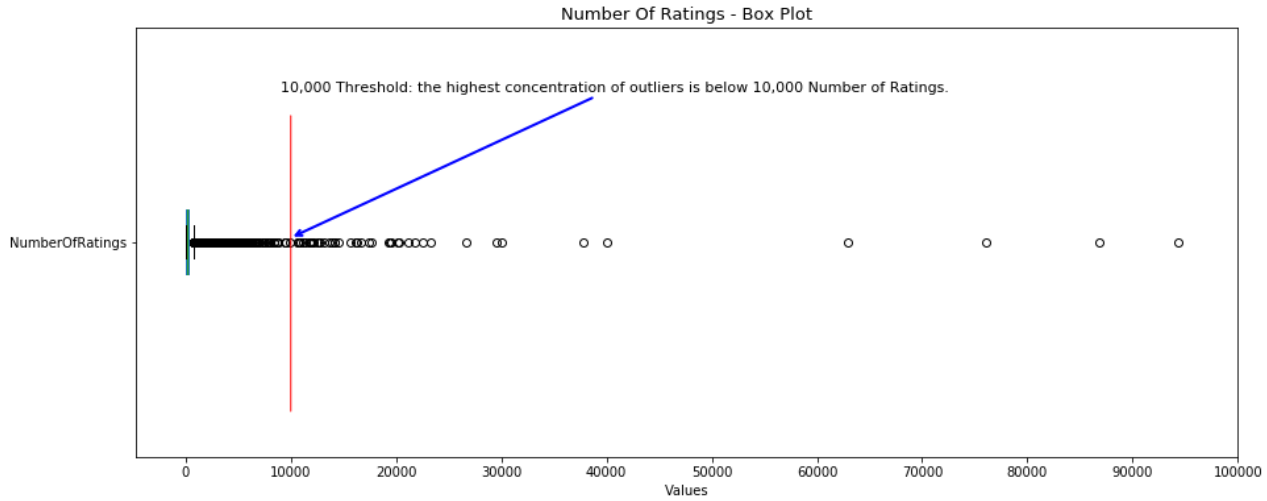
```
"Rating" attribute range: 2.2 - 4.9.
"Rating" non-outlier region: 3.100000000000001 - 4.699999999999999.

Numbers of outliers in "Rating" attribute: 144.
Percentage of "Rating" outliers: 1%.
```

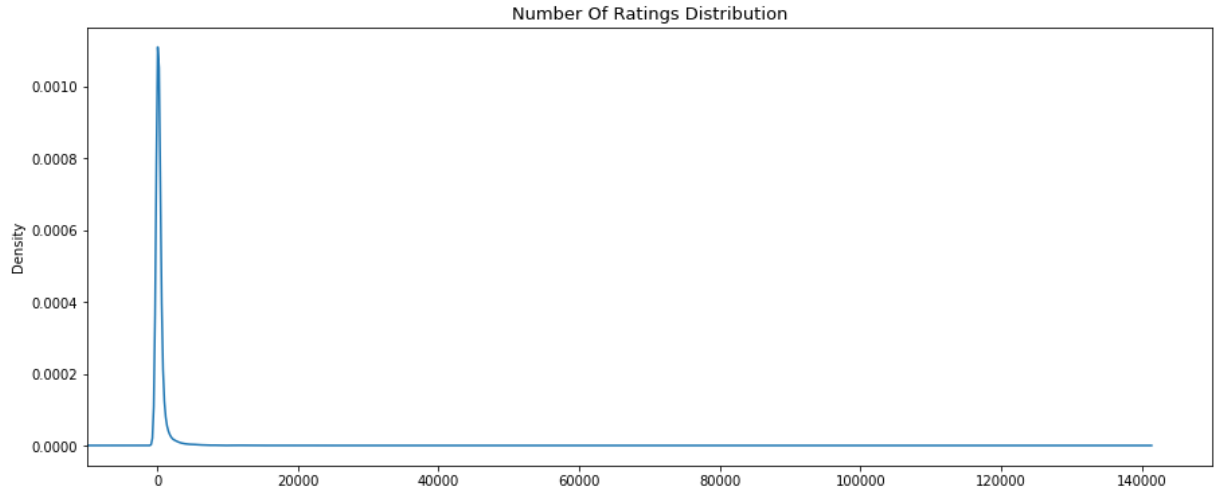
Rating outliers – sample:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style	Unique_Name
23	Virtus Tannat 2013	Brazil	Serra Gaúcha	Monte Paschoal	2.9	100	6.77	2013	red	Monte Paschoal - Virtus Tannat 2013 - Serra Ga...
253	Promontory 2013	United States	Napa Valley	Promontory	4.7	107	721.34	2013	red	Promontory - Promontory 2013 - Napa Valley
607	Tribu Merlot 2016	Argentina	Maipu	Trivento	3.1	119	8.67	2016	red	Trivento - Tribu Merlot 2016 - Maipu
632	Cabernet Sauvignon - Merlot 2015	Slovenia	Podravje	Puklavec Family Wines	3.0	120	7.65	2015	red	Puklavec Family Wines - Cabernet Sauvignon - M...
655	Sorì dei Mori Barbera d'Asti 2018	Italy	Barbera d'Asti	Vinchio - Vaglio Serra	3.1	121	8.90	2018	red	Vinchio - Vaglio Serra - Sorì dei Mori Barbera...

The "Rating" outliers seemed to be a normal part of the data distribution due to its natural variation.

- **Number of Ratings**





The data distribution appears to be very right-skewed, with several outliers beyond the max limit of the Interquartile range. The highest concentration of outliers is below the 10,000 number-of-reviews threshold.

```
"NumberOfRatings" attribute range: 25.0 – 94,287.
"NumberOfRatings" non-outlier region: -364 - 756.

Numbers of outliers in "NumberOfRatings" attribute: 1,544
Percentage of "NumberOfRatings" outliers: 11%.

Number of extreme outliers (> 10,000 Number of Ratings): 44
Percentage of extreme outliers (> 10,000 Number of Ratings): 0.32%
```

Number-of-Ratings outliers – sample:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style	Unique_Name
25	Badia a Passignano Gran Selezione Chianti Clas...	Italy	Chianti Classico	Antinori	4.2	1000	39.90	2016	red	Antinori - Badia a Passignano Gran Selezione C...
26	Saint-Émilion Grand Cru (Premier Grand Cru Cla...	France	Saint-Émilion Grand Cru	Château Figeac	4.4	1000	174.49	2012	red	Château Figeac - Saint-Émilion Grand Cru (Prem...
27	Carlo V Il Rosso dell'Imperatore 2013	Italy	Veneto	Colli Vicentini	3.7	1000	15.90	2013	red	Colli Vicentini - Carlo V Il Rosso dell'Impera...
28	Pinot Noir 2016	United States	Monterey	District 7	3.6	1001	15.98	2016	red	District 7 - Pinot Noir 2016 - Monterey
29	Marquis de Calon Saint-Estèphe 2015	France	Saint-Estèphe	Château Calon-Ségur	3.9	1001	43.28	2015	red	Château Calon-Ségur - Marquis de Calon Saint-E...

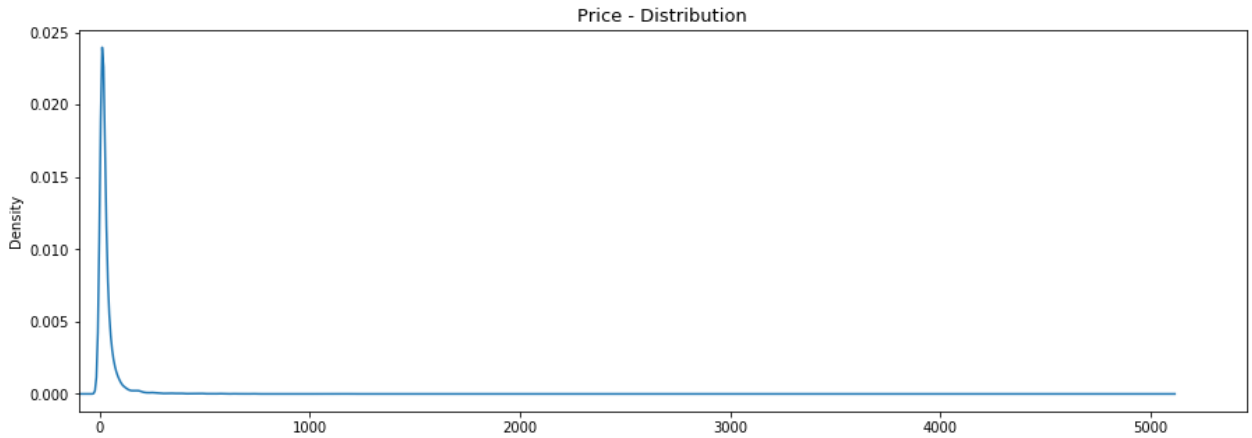
Number-of Ratings extreme outliers (number of ratings > 10,000) – sample:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style	Unique_Name
13831	Brut Champagne N.V.	France	Champagne	Dom Pérignon	4.6	94287	170.00	N.V.	sparkling	Dom Pérignon - Brut Champagne N.V. - Champagne
13830	Brut (Carte Jaune) Champagne N.V.	France	Champagne	Veuve Clicquot	4.2	86839	43.60	N.V.	sparkling	Veuve Clicquot - Brut (Carte Jaune) Champagne ...
13829	Impérial Brut Champagne N.V.	France	Champagne	Moët & Chandon	4.1	76037	40.61	N.V.	sparkling	Moët & Chandon - Impérial Brut Champagne N.V. ...
12427	Vinho Verde Branco N.V.	Portugal	Vinho Verde	Casal Garcia	3.7	62980	4.35	N.V.	white	Casal Garcia - Vinho Verde Branco N.V. - Vinho...
13828	Brut Premier Champagne N.V.	France	Champagne Premier Cru	Louis Roederer	4.2	40004	36.48	N.V.	sparkling	Louis Roederer - Brut Premier Champagne N.V. -...

The number of outliers is quite high: 11% of all observations. However, the number of extreme outliers (wines which have got more than 10,000 ratings) is negligible: 44 observations (0.32% of the whole dataset).

The data distribution is very right-skewed: all outliers are beyond the upper maximum limit. The wines with the highest number of ratings (the extreme outliers) are very popular wines produced by highly renowned wineries such as Dom Perignon, Veuve Clicquot, Moet & Chandon, Bollinger, Ferrari... it makes sense that very popular and renowned wines get a higher number of reviews compared to others. Therefore, these outliers don't seem to be mistakes or aberrations, but a normal part of the data distribution due to its natural variation.

- **Price**



Again, the attribute data distribution appears to be very right-skewed, with several outliers beyond the max limit of the Interquartile range. The most expensive bottle is almost 3,500 Euros: this is an extreme outlier far away from all other observations. The highest concentration of outliers is below the 500 euro price threshold.

```
"Price" attribute range: 3.15 - 3410.79.
"Price" non-outlier region: -24 - 66.4.

Numbers of outliers in "Price" attribute: 1,313
Percentage of "Price" outliers: 9%.

Number of extreme outliers (> 500 Euro price): 48
Percentage of extreme outliers (> 10,000 Number of Ratings): 0.35%
```

Price outliers – sample:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style	Unique_Name
0	Pomerol 2011	France	Pomerol	Château La Providence	4.2	100	95.00	2011	red	Château La Providence - Pomerol 2011 - Pomerol
11	Descendant 2016	Australia	Barossa	Torbreck	4.3	100	140.64	2016	red	Torbreck - Descendant 2016 - Barossa
26	Saint-Émilion Grand Cru (Premier Grand Cru Cla...	France	Saint-Émilion Grand Cru	Château Figeac	4.4	1000	174.49	2012	red	Château Figeac - Saint-Émilion Grand Cru (Prem...
32	Saint-Julien (Grand Cru Classé) 2010	France	Saint-Julien	Château Léoville Poyferré	4.4	1008	189.00	2010	red	Château Léoville Poyferré - Saint-Julien (Gran...
34	Vigna del Noce Barbera d'Asti 2007	Italy	Barbera d'Asti	Trinchero	4.1	101	76.90	2007	red	Trinchero - Vigna del Noce Barbera d'Asti 2007...

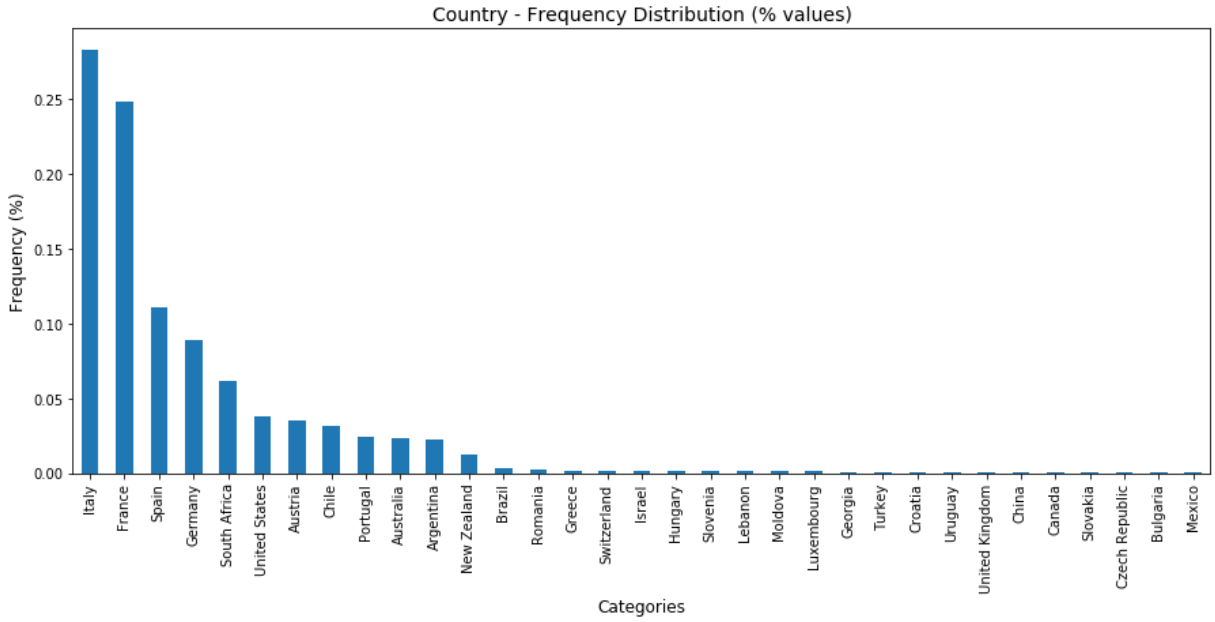
Price extreme outliers (price > 500) – sample:

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Style	Unique_Name
2344	Pomerol 2012	France	Pomerol	Pétrus	4.7	204	3410.79	2012	red	Pétrus - Pomerol 2012 - Pomerol
7466	Saint-Émilion Grand Cru (Premier Grand Cru Cla...	France	Saint-Émilion Grand Cru	Château Ausone	4.5	72	1599.95	2010	red	Château Ausone - Saint-Émilion Grand Cru (Prem...
5560	Pauillac (Premier Grand Cru Classé) 2010	France	Pauillac	Château Lafite Rothschild	4.4	457	1399.00	2010	red	Château Lafite Rothschild - Pauillac (Premier ...
4395	Pauillac (Premier Grand Cru Classé) 1992	France	Pauillac	Château Lafite Rothschild	4.5	346	1266.25	1992	red	Château Lafite Rothschild - Pauillac (Premier ...
4124	Pessac-Léognan (Premier Grand	France	Pessac-Léognan	Château Haut-Brion	4.6	327	1107.00	2010	red	Château Haut-Brion - Pessac-

Again, the number of "Price" outliers is quite high (9% of all observations), and the data distribution is right-skewed: all outliers are beyond the upper maximum limit. However, the number of extreme outliers (very expensive wines, above 500 Euros per bottle) is limited: 0.35% of all data set.

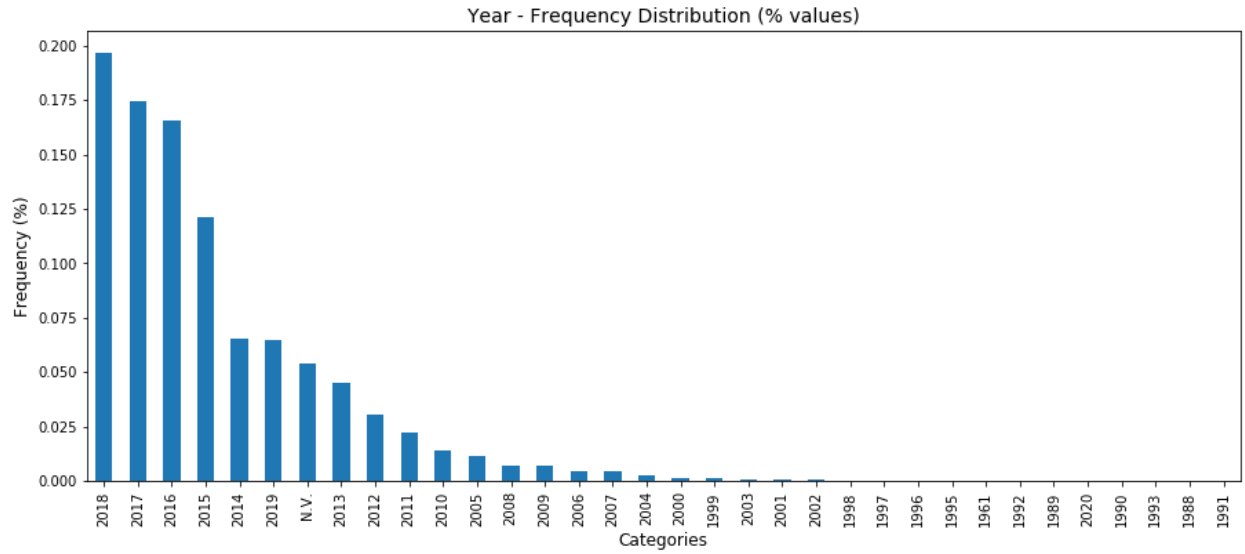
Again, the "Price" outliers seem to be a normal part of the data distribution: 9% expensive wines, and 0.35% extremely expensive wines is not an abnormal distribution for this product category.

- **Categorical-Attribute Analysis**




```
Number of categories: 33
```

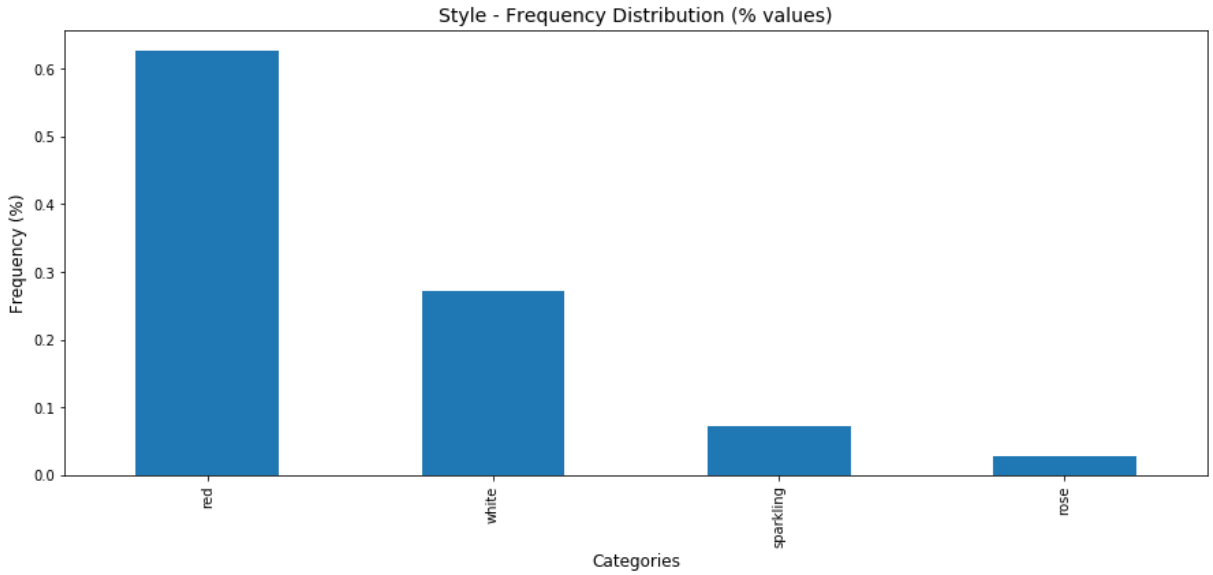
The attribute Country has 33 unique values, with a high concentration in few countries. Italy and France, only, account for more than 50% of the observations; Italy, France, and Spain account for more than 60% of total number of observations.




```
Number of categories: 34
```

The attribute Year ranges from 1988 until 2020, with 1 apparent anomaly: 1961, which has a very small number of observations. After further checking, I verified that 1961 wines are actually correctly recorded. The attribute includes N.V., non-vintage, that is wines produced by mixing harvest of two years or more. The N.V. category is the 7th category with the highest frequency, accounting for about 5% of total observations.

The frequency distribution is highly concentrated towards the last decade (starting from 2011), which accounts for 89% of observations.




```
Number of categories: 4
```

"Style" frequency distribution is, again, quite imbalanced: red wines account for more than 60% of observations, and Red + White wines account for about 90% of total dataset.

```
"Region" attribute - number of categories: 861

"Region" categories frequency (normalized values):
Rioja                                      0.027834
Stellenbosch                               0.024364
Pfalz                                      0.023930
Toscana                                    0.022195
Champagne                                  0.019014
...                                             ...
Würzburg                                   0.000072
Savigny-lès-Beaune 1er Cru 'Aux Clous'     0.000072
Bourgogne Epineuil                         0.000072
Parrina                                    0.000072
Terrazze Retiche di Sondrio                0.000072
Name: Region, Length: 861, dtype: float64
```

The attribute Region has 861 unique values, with 2.8% of entries belonging to the category with the highest frequency (Rioja). The attribute is interesting for a geographical-descriptive analysis, but, considering the huge number of categories and their frequency distribution, it doesn't seem to have any explanatory value towards the variables Rating and Price.

```
"Winery" attribute - number of categories: 3,505

"Winery" categories frequency (normalized values):

Markus Molitor        0.005278
Errazuriz             0.004193
Torres                0.003904
Joseph Drouhin        0.003615
Gaja                  0.003036
...                        ...
Edouard Leiber        0.000072
Callejo               0.000072
Burkheimer            0.000072
Tiberio               0.000072
Miles Mossop Wines    0.000072
Name: Winery, Length: 3505, dtype: float64
```

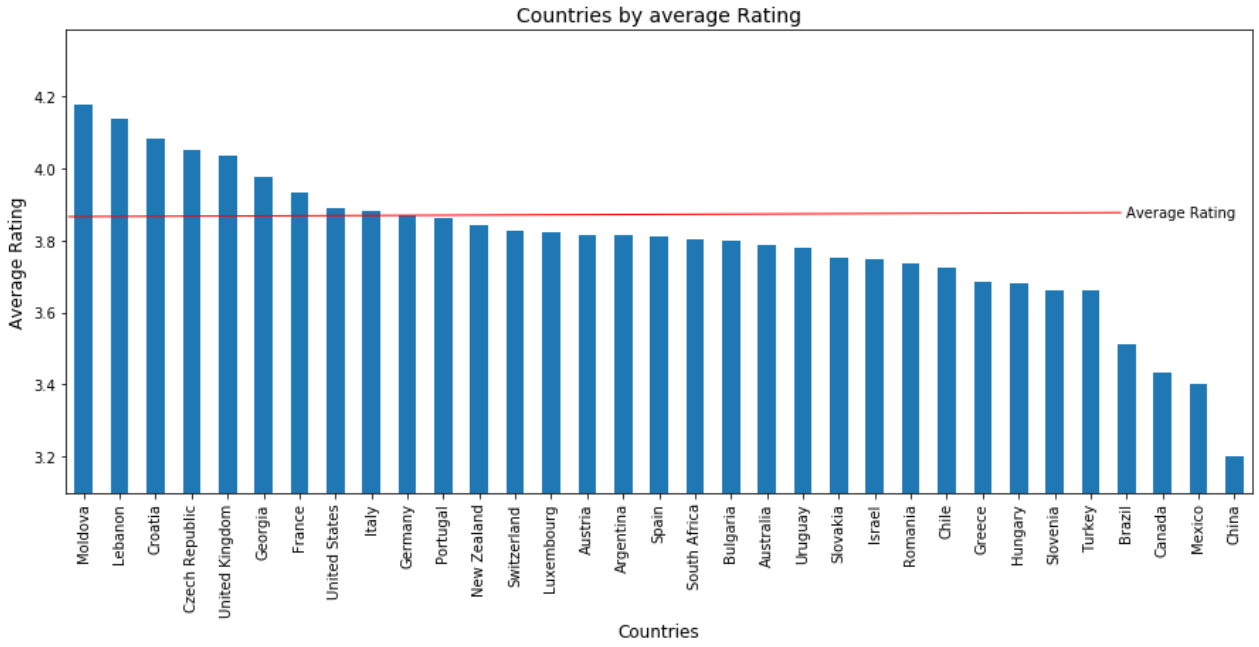
The "Winery attribute has 3,505 unique values, with 0.5% of entries belonging to the category with the highest frequency (Markus Molitor). The attribute might be interesting for a supplier-analysis, but, considering the huge number of categories and their frequency distribution, it doesn't seem to have any explanatory value towards the variables Rating and Price.

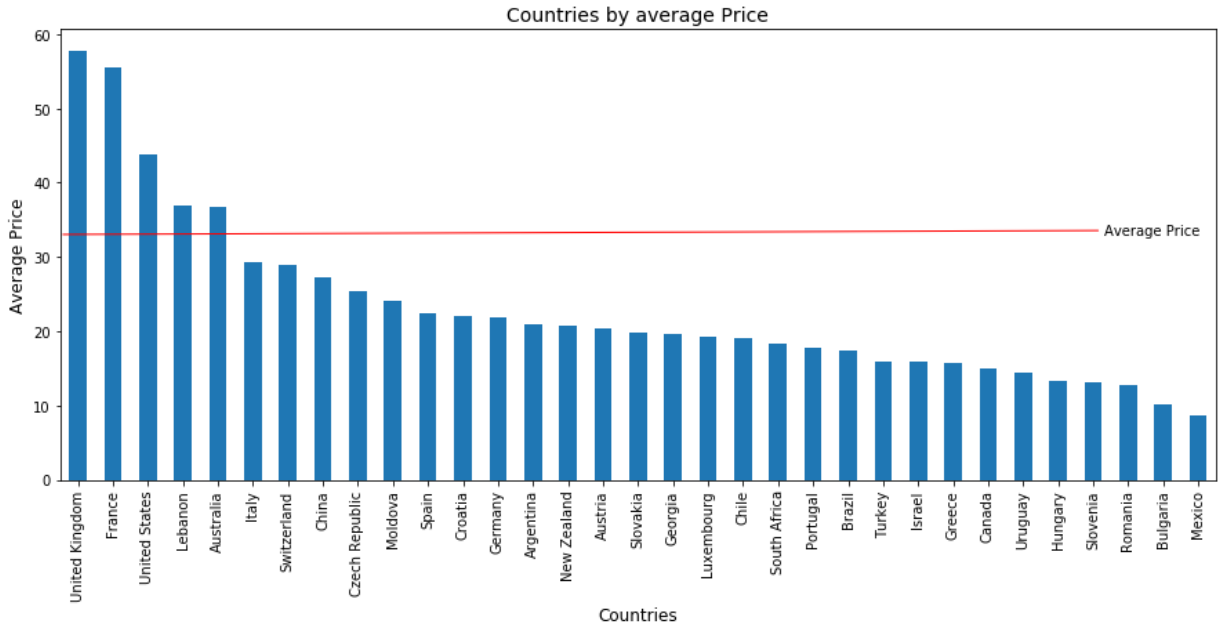
Besides an imbalanced distribution in some of the attributes, **no anomalies or mistakes are found.**

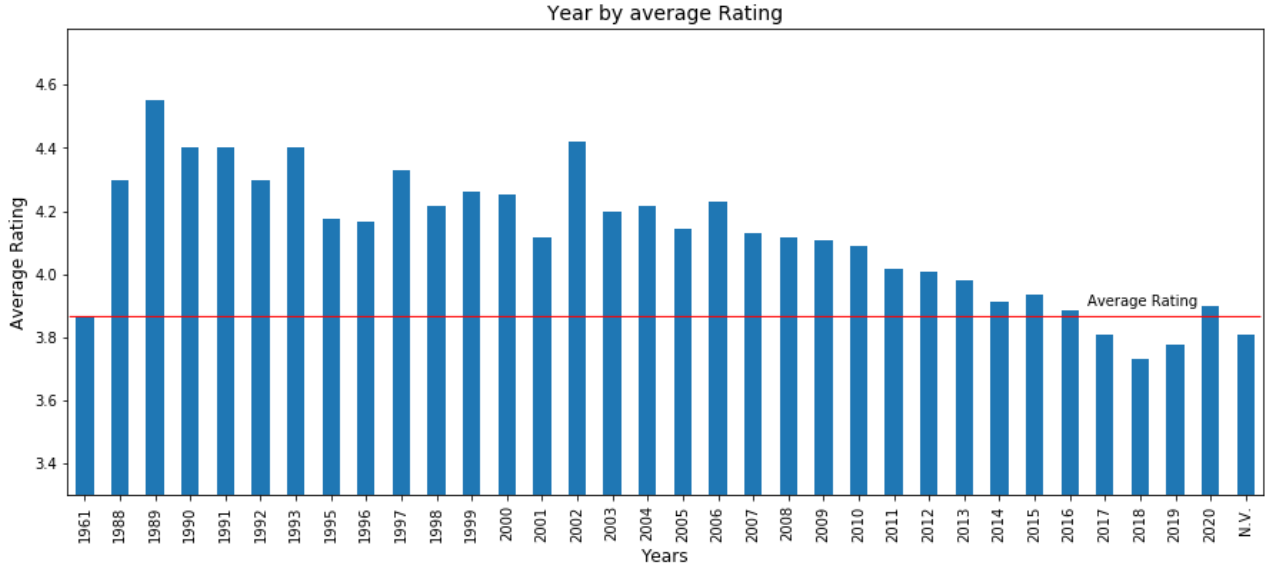
3. E.D.A. results: key findings and insights

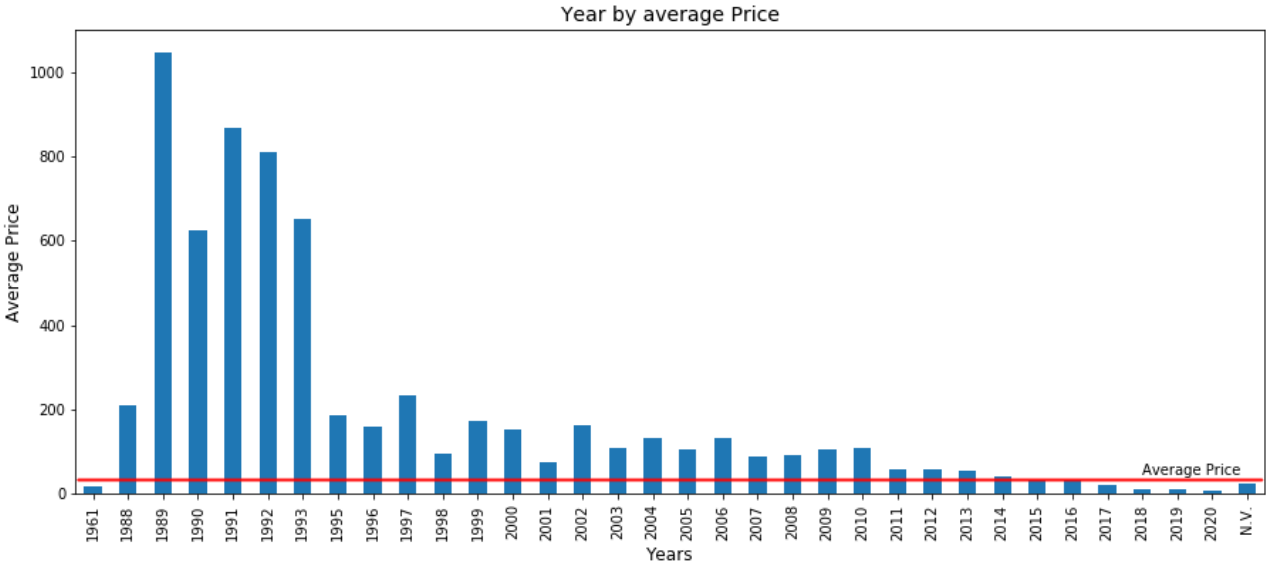
3.1. Main data characteristics

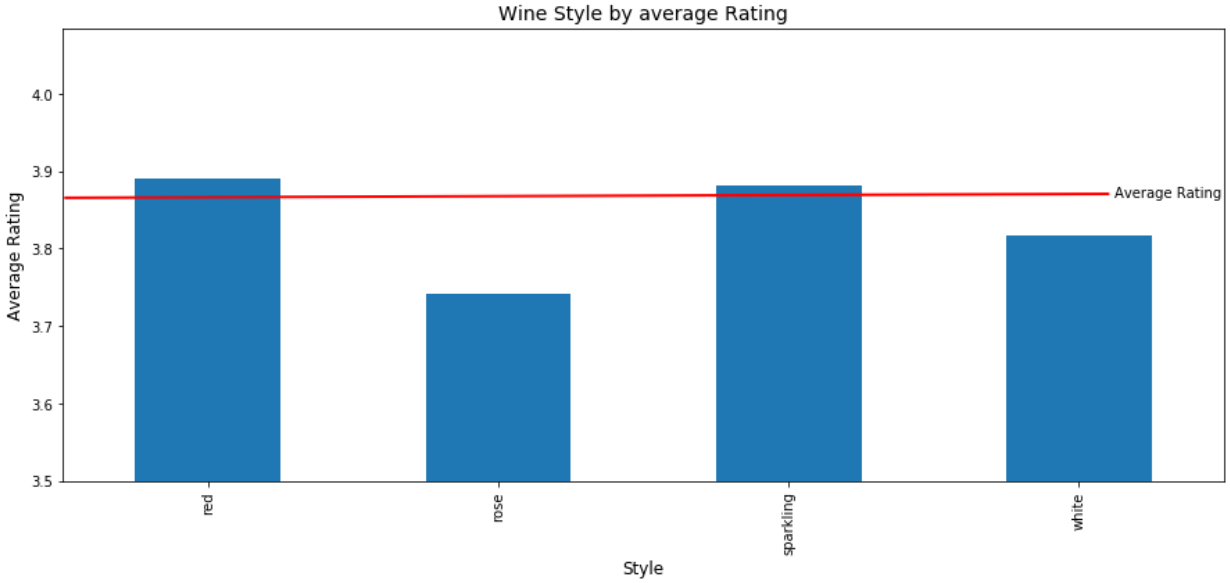
- Italy is the biggest wine producer: 28% of Vivino.com supply.
- As already observed during the outlier analysis, "Rating" data has approximately a normal distribution, whilst "NumberOfRatings" and "Price" data distributions are very right skewed.
- 20% of Vivino.com wines are produced in 2018 (the year with the highest frequency).
- Red wine is, by far, the most produced wine: 63% of the whole supply.

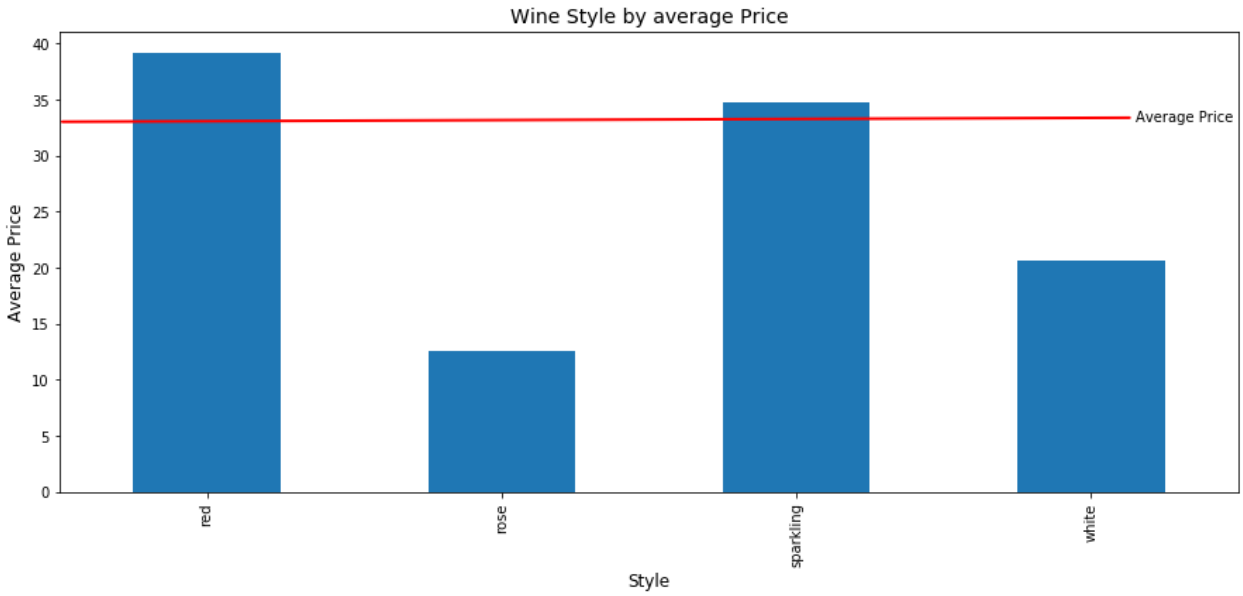



Countries by average Price
Average Price
Countries
Average Price
United Kingdom
France
United States
Lebanon
Australia
Italy
Switzerland
China
Czech Republic
Moldova
Spain
Croatia
Germany
Argentina
New Zealand
Austria
Slovakia
Georgia
Luxembourg
Chile
South Africa
Portugal
Brazil
Turkey
Israel
Greece
Canada
Uruguay
Hungary
Slovenia
Romania
Bulgaria
Mexico


Year by average Rating
Average Rating
Years
Average Rating
1961
1988
1989
1990
1991
1992
1993
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
N.V.


Year by average Price
Average Price
1000
800
600
400
200
0
Average Price
1961
1988
1989
1990
1991
1992
1993
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
N.V.
Years


Wine Style by average Rating
Average Rating
4.0
3.9
3.8
3.7
3.6
3.5
Average Rating
red
rose
sparkling
white
Style



- Moldova wines have the highest average rating score: 4.175. That's quite an unexpected result. It would be interesting to know if this difference between Moldova's wine average rating, and the population average rating, is statistically significant, or just due to chance, or to the natural variability of the data.

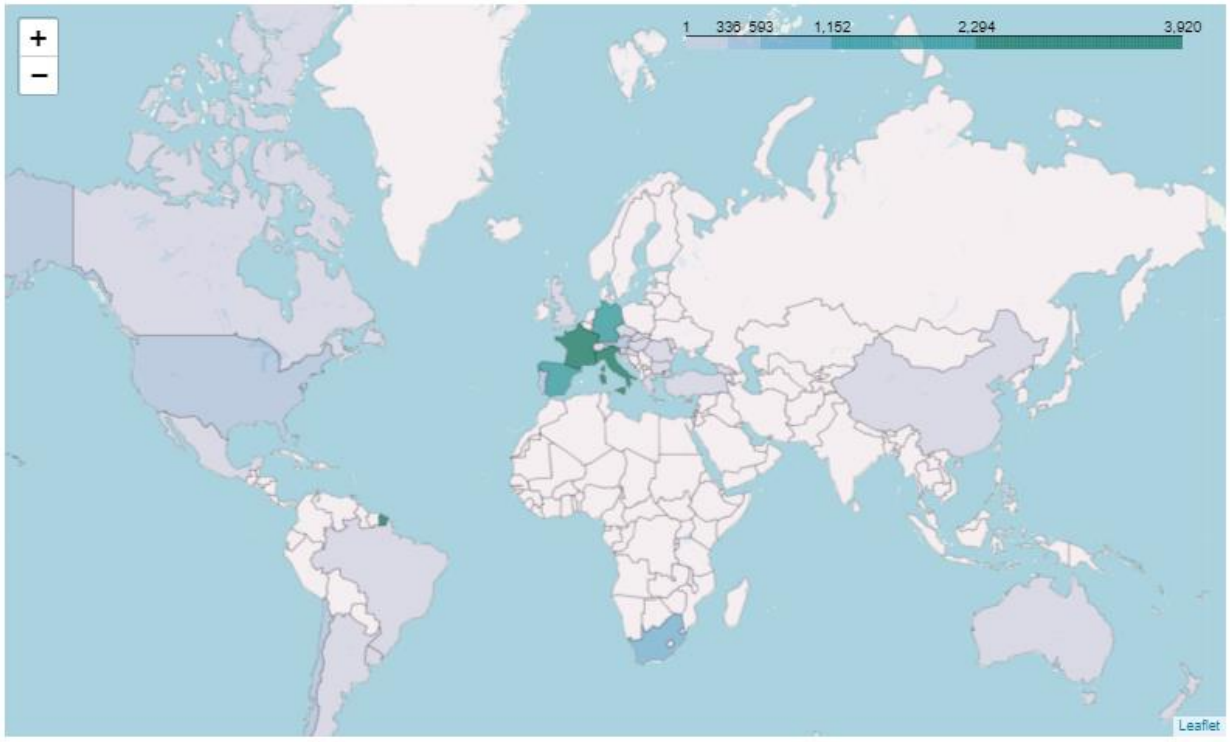
  Question 1: *Is the difference between Moldova's wine average rating and the population (whole dataset) average rating statistically significant?* (See Hypothesis-Testing section for the formal significance test).

- France is the only top producer whose mean price per bottle is above the average.
- The last 3 years (2017, 2018, 2019) are the only years whose mean rating score is below the average (alongside with the non-vintage wines).
- Wines produced from 1989 to 1993 have a way higher price than all other wines.

  Question 2: *Is the difference between 1989-1993 average wine price and population (whole dataset) average price statistically significant?*

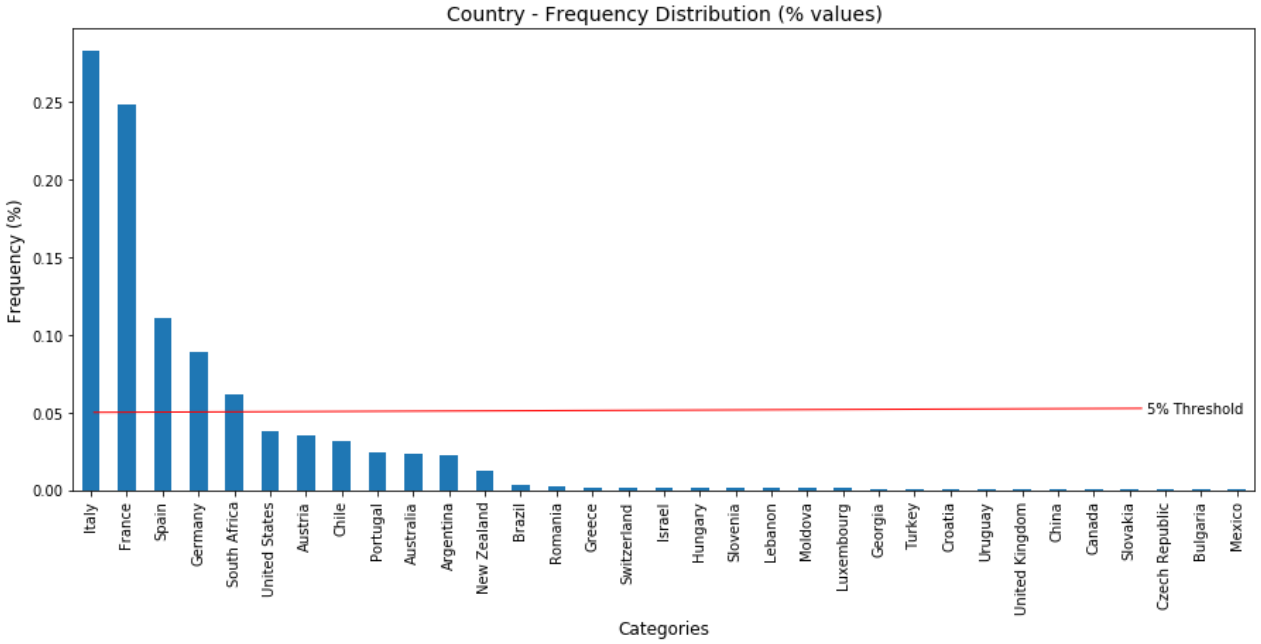
- Red wine is the top rated style, although there's not much difference in terms of average Rating, when it comes to wine styles.
- Red wines are the most expensive whilst Rose wines are the least pricey (average price).

World map with the wine producing countries.



Wine production is very concentrated in few countries: Italy, France, Spain, Germany, and South Africa.

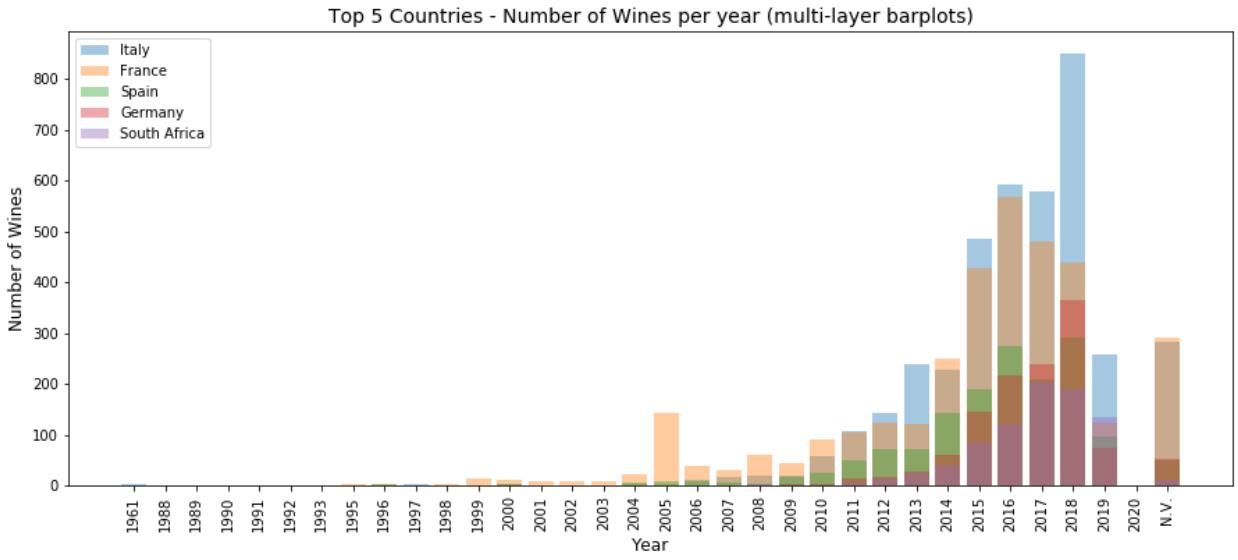
3.2. Top 5 countries analysis



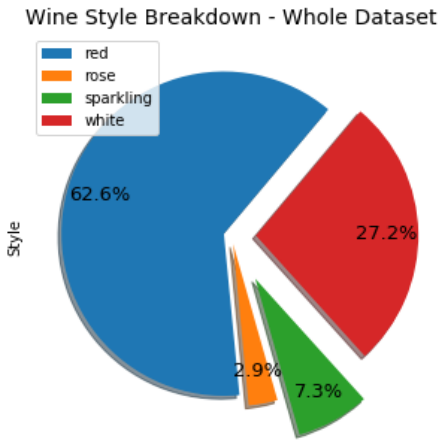
Top 5 countries in Vivino.com, with a share (in volume) greater than 5%:

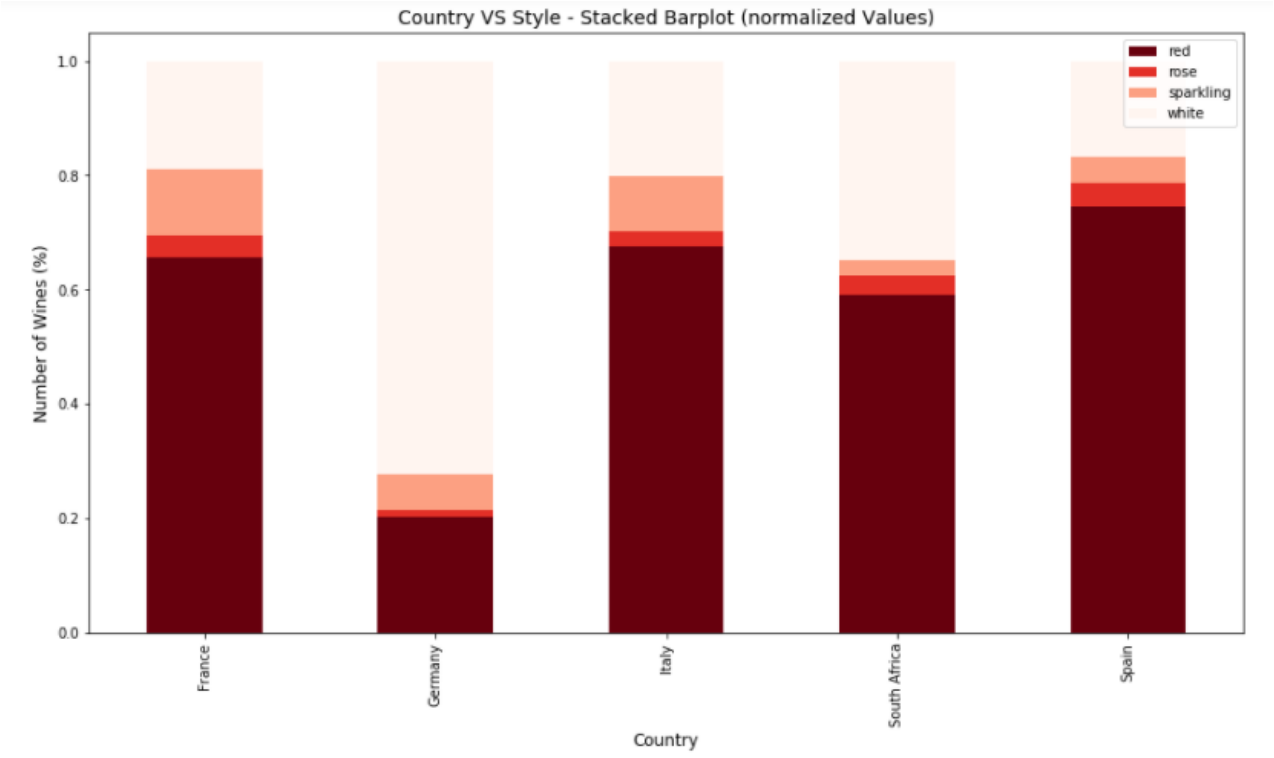
- Italy
- France
- Spain
- Germany
- South Africa

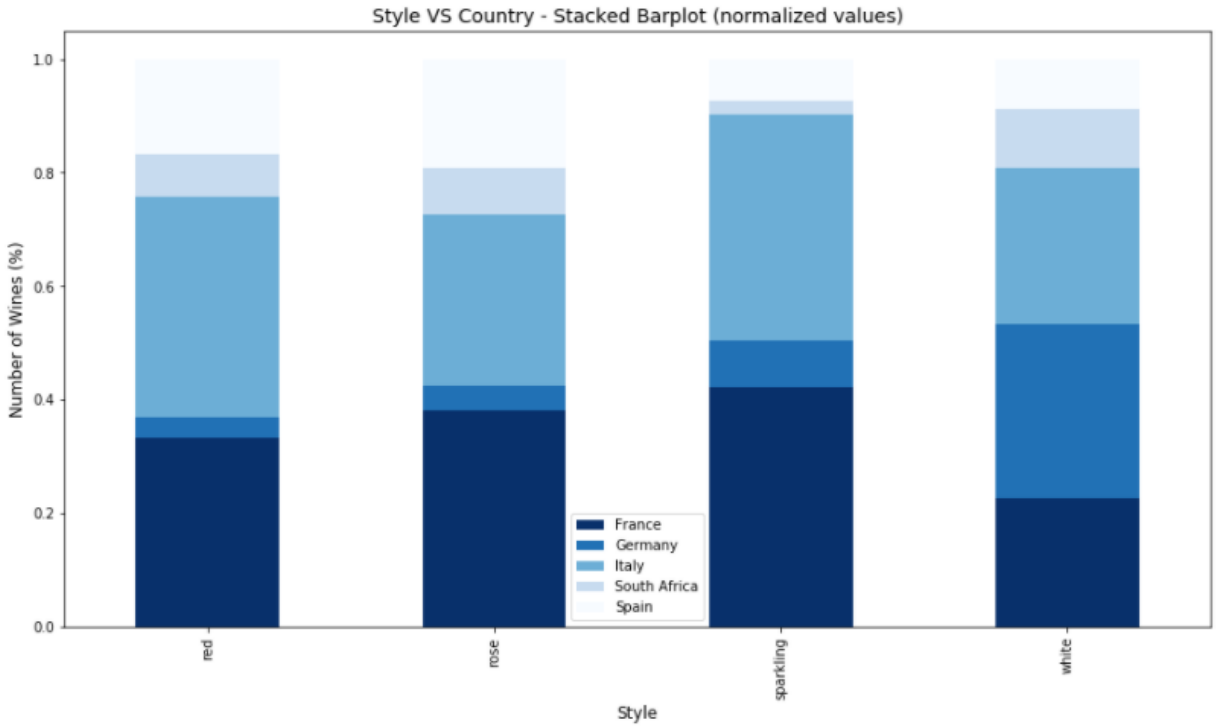
Share (volume) of top 5 countries: 79.25%.



- The number of wines in Vivino.com increases as the year increases as well, until 2018.
- Italy has been increasing its share since 2012, overtaking France as the biggest wine supplier in 2015.
- In 2018 (the year with the greatest production), Italy alone accounted for more than 30% of the wine supply.
- Germany has been steadily increasing its share as well since 2014.
- Since 2016, French wine share has been decreasing.







- Italy and France are the biggest producers for red, rose', and sparkling wines.
- Germany is the only country, amongst the top 5, which produces more white wines than red wines.
- German white wine share is greater than Italy and France's share.
- Italy and France have a very similar wine style breakdown.
- Spain has a considerable share on the red wines: 16.8%.

3.3. Supplier analysis

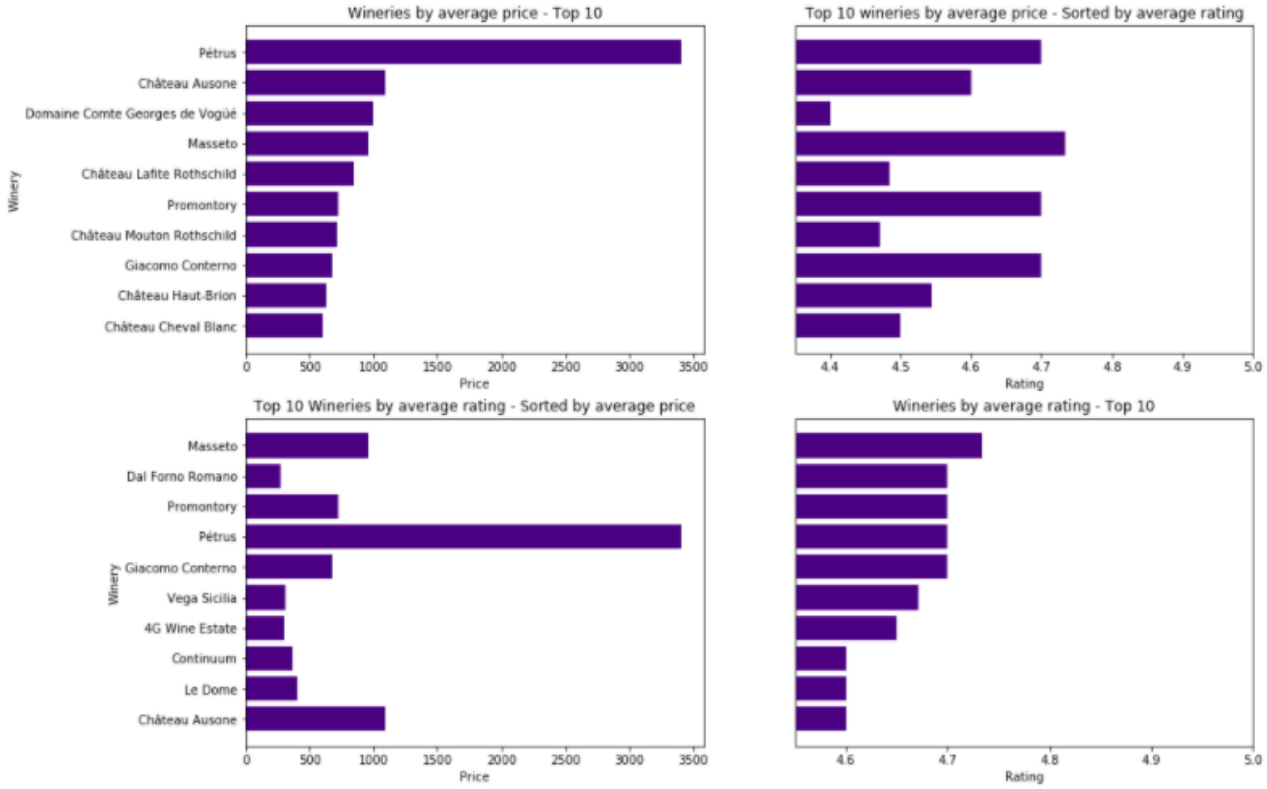
Top 10 Wineries (per number of wines):

	Winery	Country	number_of_wines	share (%)	avg_rating	avg_price	population_avgrating	population_avgprice	above_avgrating	above_avgprice
2293	Markus Molitor	Germany	73	0.53	4.009589	33.827123	3.865667	33.027764	True	True
1580	Errazuriz	Chile	58	0.42	3.884483	38.585517	3.865667	33.027764	True	True
3226	Torres	Spain	54	0.39	3.811111	24.800000	3.865667	33.027764	False	False
1979	Joseph Drouhin	France	50	0.36	3.980000	66.608200	3.865667	33.027764	True	True
1733	Gaja	Italy	42	0.30	4.366667	197.145952	3.865667	33.027764	True	True
2246	M. Chapoutier	France	42	0.30	4.011905	69.891905	3.865667	33.027764	True	True
1530	E. Guigal	France	36	0.26	3.994444	61.301667	3.865667	33.027764	True	True
2605	Paul Jaboulet Aîné	France	36	0.26	3.950000	66.965556	3.865667	33.027764	True	True
108	Antinori	Italy	36	0.26	4.163889	100.397778	3.865667	33.027764	True	True
3406	Von Winning	Germany	34	0.25	4.008824	24.943824	3.865667	33.027764	True	False

- Torres, Spain, is the only winery, out of the top 10 wineries, that has a rating score (average) lower than the overall rating average.
- Torres, Spain, and Von Winning, Germany, are the only 2 wineries, out of the top 10 wineries, which have an average wine price less than the overall price average.
- Gaja, Italy, (5th winery per number of wines) is the winery, amongst the top 10, with the highest average rating and price.

Question 3: *Is the difference between Gaja average rating/price and population (dataset) average rating/price statistically significant (or due to chance)?*

Let's visualize some bar plots about the 10 best rated and 10 most expensive wineries.



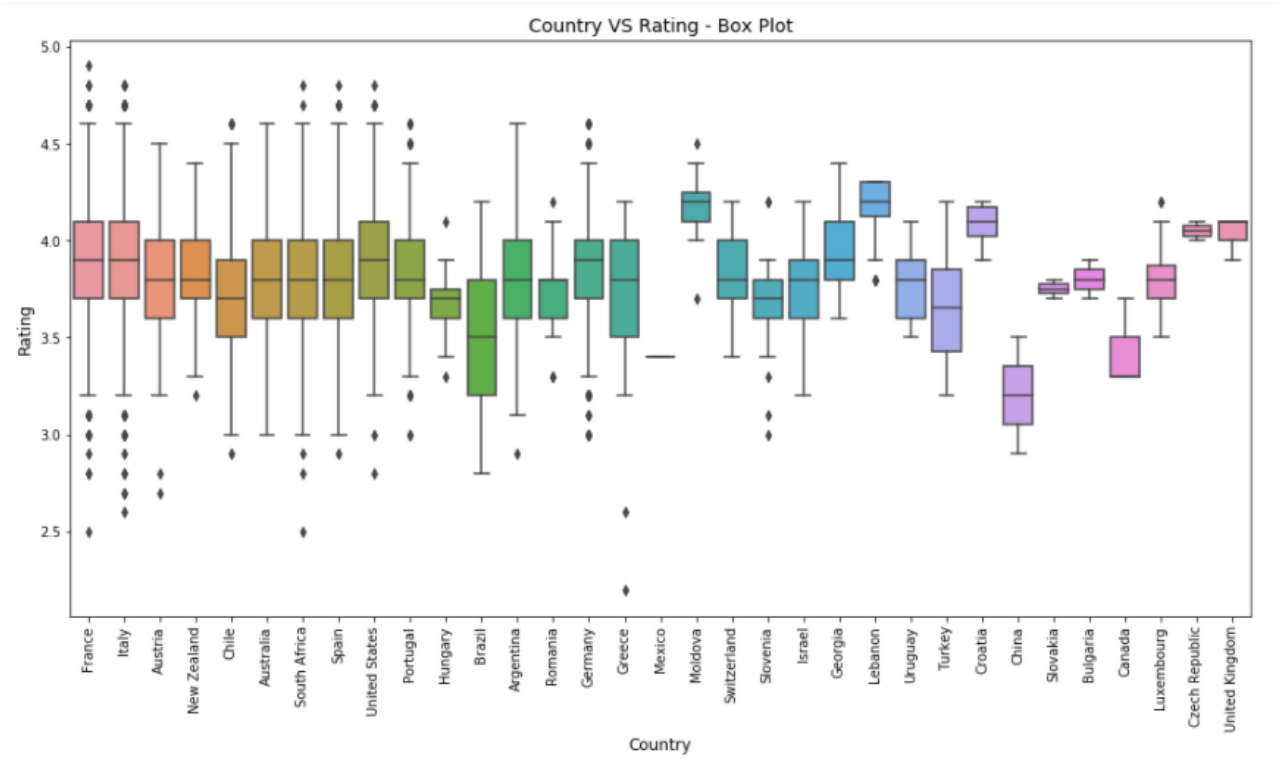
- The top 10 wineries, per number of wines, are neither amongst the best rated nor amongst the most expensive.
- There is a clear overlap amongst the top rated and the most expensive winery: 5 wineries (Giacomo Conterno, Promontory, Masseto, Château Ausone, and Pétrus) belong to both groups.
- Pétrus, France, is, by far, the most expensive winery (with only 1 observation); it is the most extreme Price outlier.

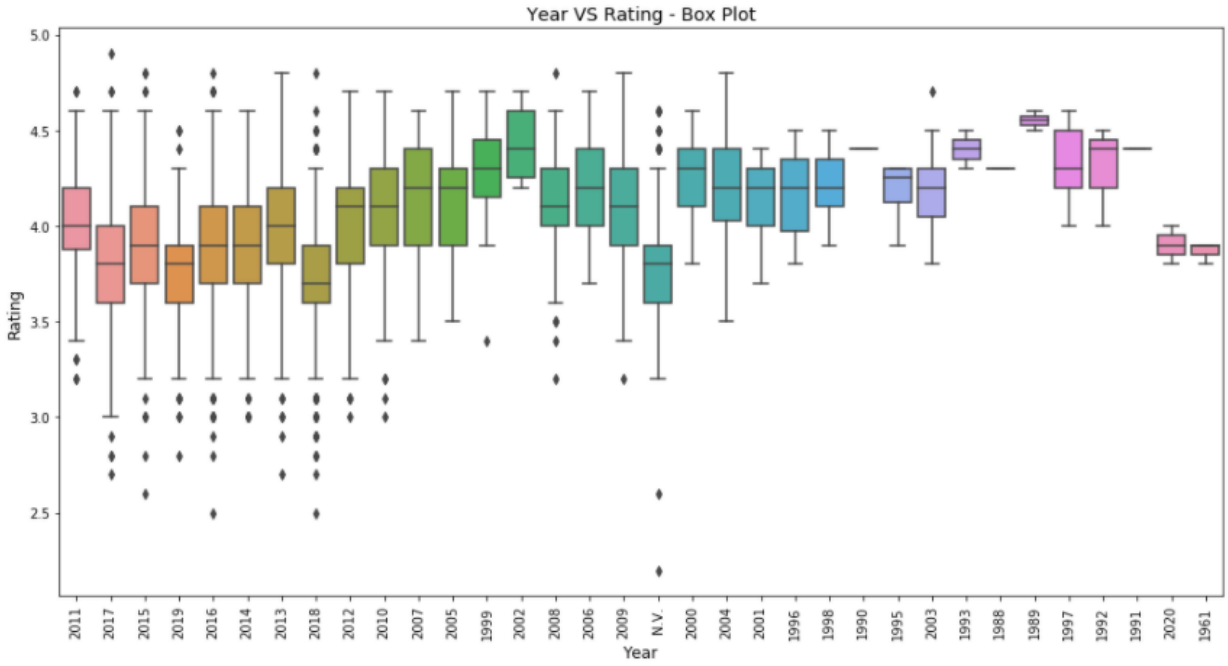
- Masseto, Italy, is the top rated winery, 4th by average wine price.
- 7 out of the 10 most expensive wineries are French.
- Amongst the top-10 rated wineries, both Italy and France have the highest number (3 each).

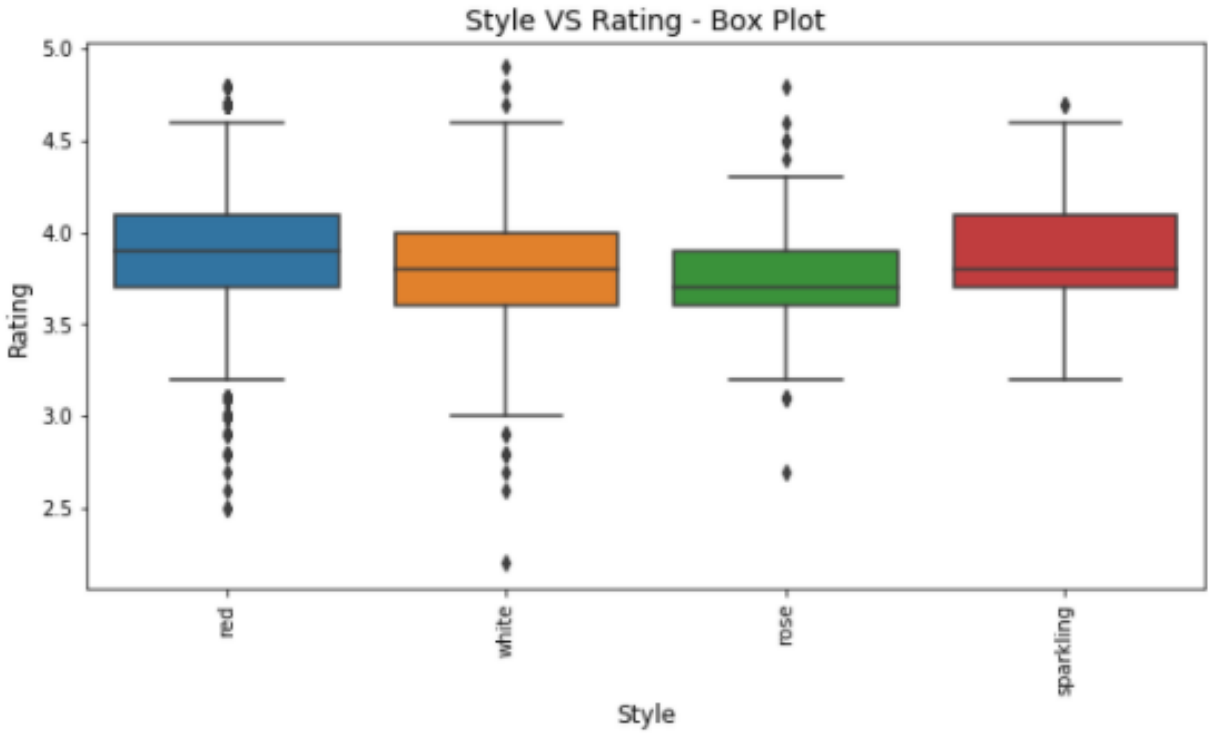
3.4. Data Mining

Uncovering correlations, patterns, and relationships between variables.

3.4.1 Target: rating.

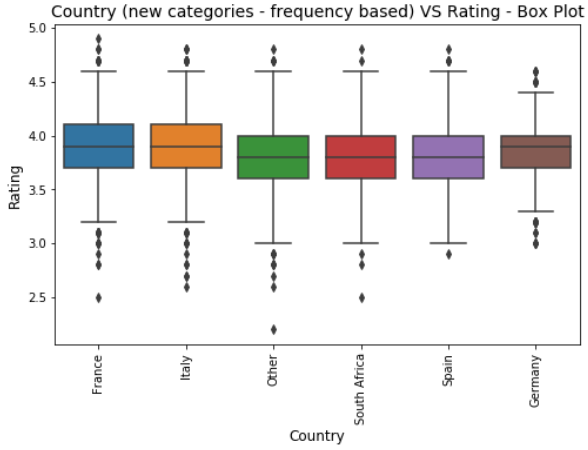


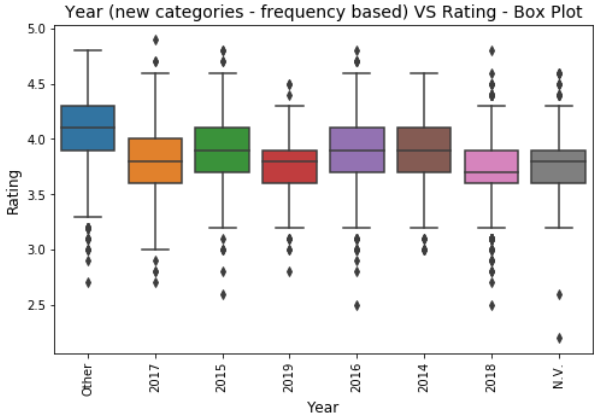




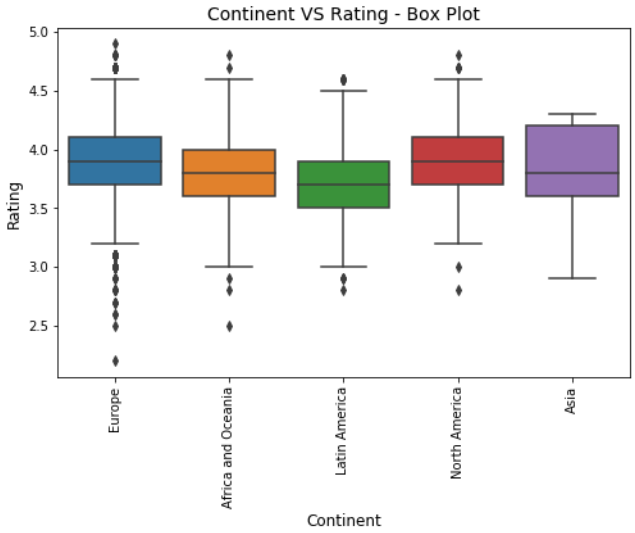
- It seems there's a significant overlap between the categories belonging to the Country, Year, and Style attributes, in relation to the Rating variable; that means that possible relationships between the features and the target are not very significant.

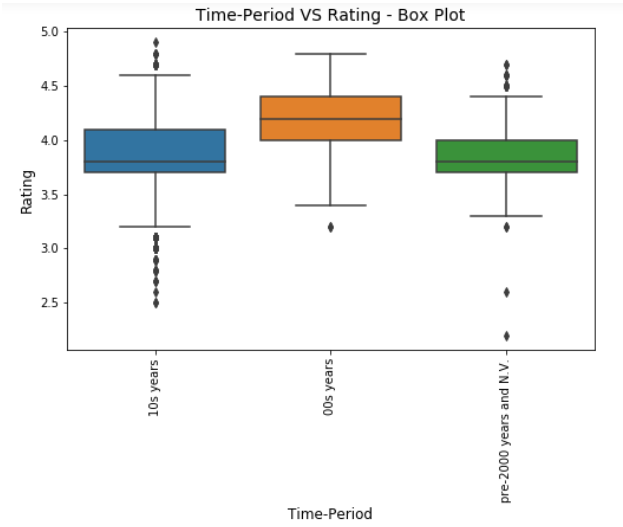
- **Country & Year new categories - frequency-based**





- **Country & Year New categories - Continents and Time Periods**



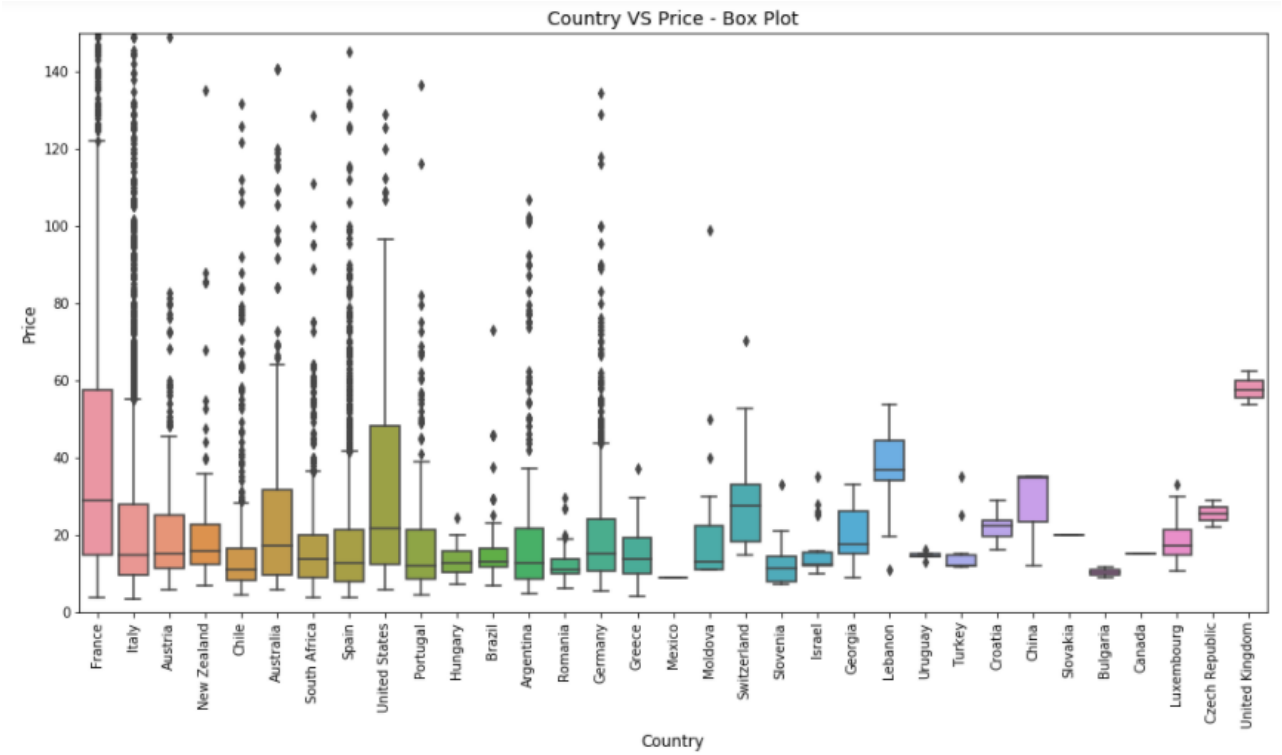


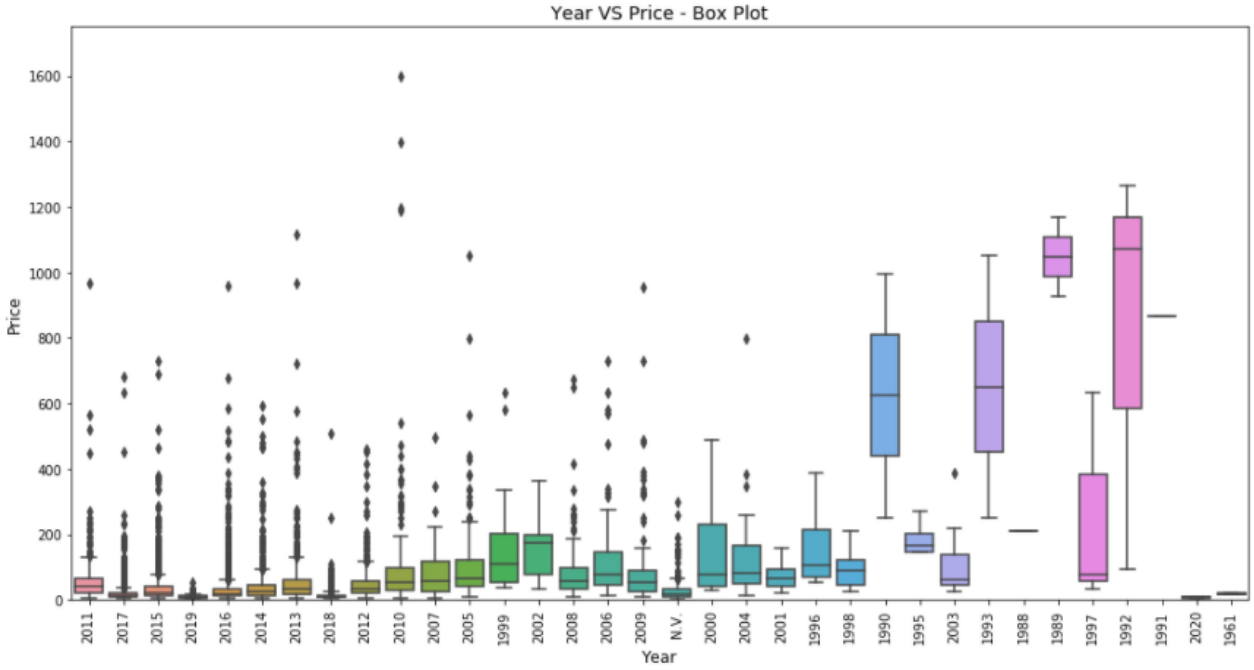
- Any relationships between "Country" and "Rating" still fail to appear, regardless of the value categorization.
- The 00s decade wines seem to have, overall, a higher rating than the other time-periods, although the category overlap is still significant.
- I am going to use the **Analysis of Variance - ANOVA** - method to compare the attribute's impact, and double-check whether the new attribute categorization, somehow, can increase the correlation between the variables.

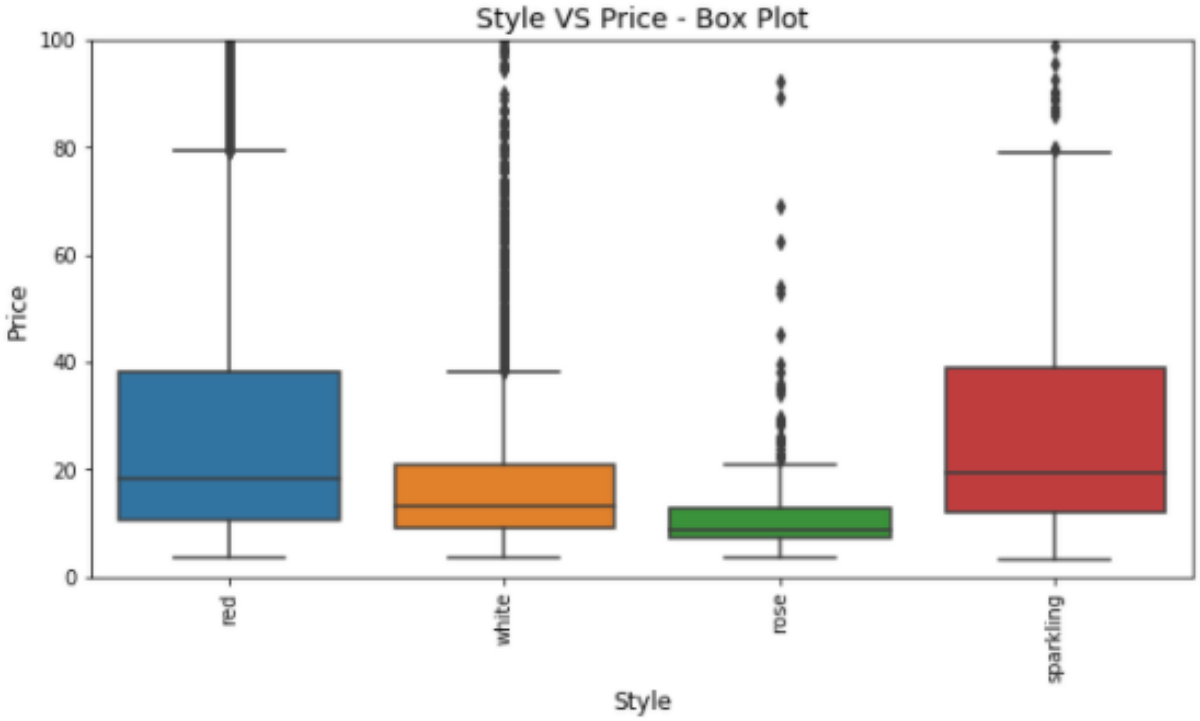
```
Dictionary with features and F-test scores - Target Variable: "Rating":
{'Country - Original Categories': 19.66,
 'Country - New Categories (freq-based)': 75.97,
 'Continent': 56.23,
 'Year - Original Categories': 67.42,
 'Year - New Categories (freq-based)': 284.81,
 'Time Period': 298.37,
 'Wine Style': 78.01}
```

- The F-test scores, by ANOVA, confirmed that Country and Style have the least impact on the wine Rating.
- Grouping the Years into time-periods proved to be the most impactful feature transformation on the Rating variable.

3.4.2. Target: price.

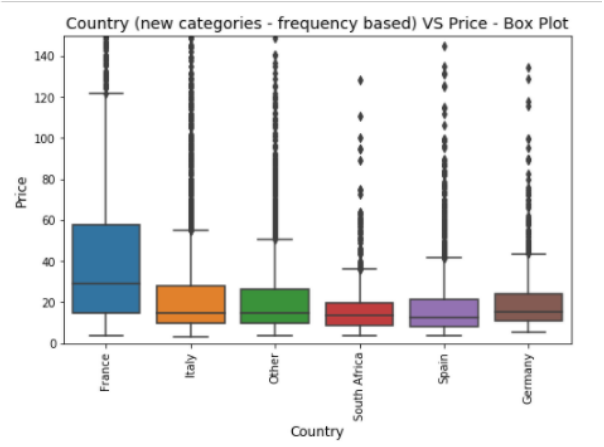


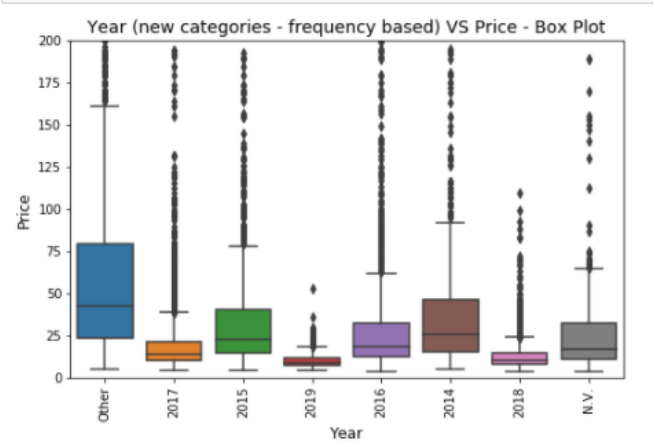




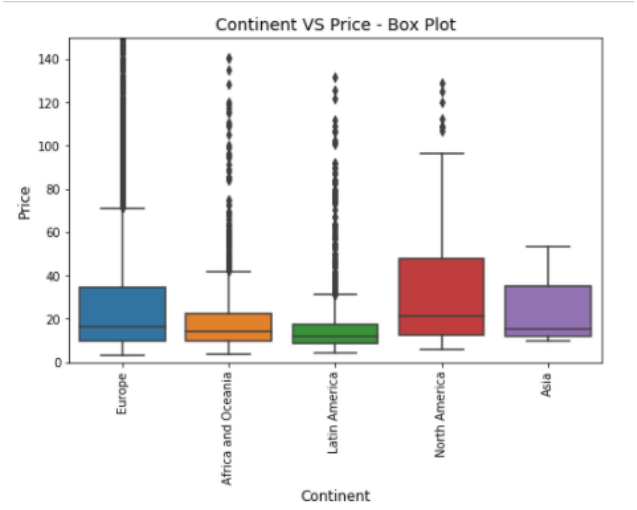
- It seems there's some overlapping between categories for both Country and Style attributes, in relation to the "Price" variable.
- The attribute Year might have a bigger impact in explaining the Price variable.

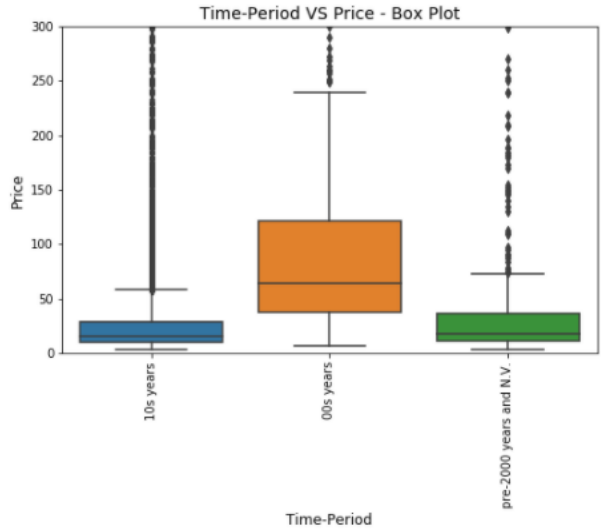
- **Country & Year new categories - frequency-based**





- **Country & Year New categories - Continents and Time Periods**





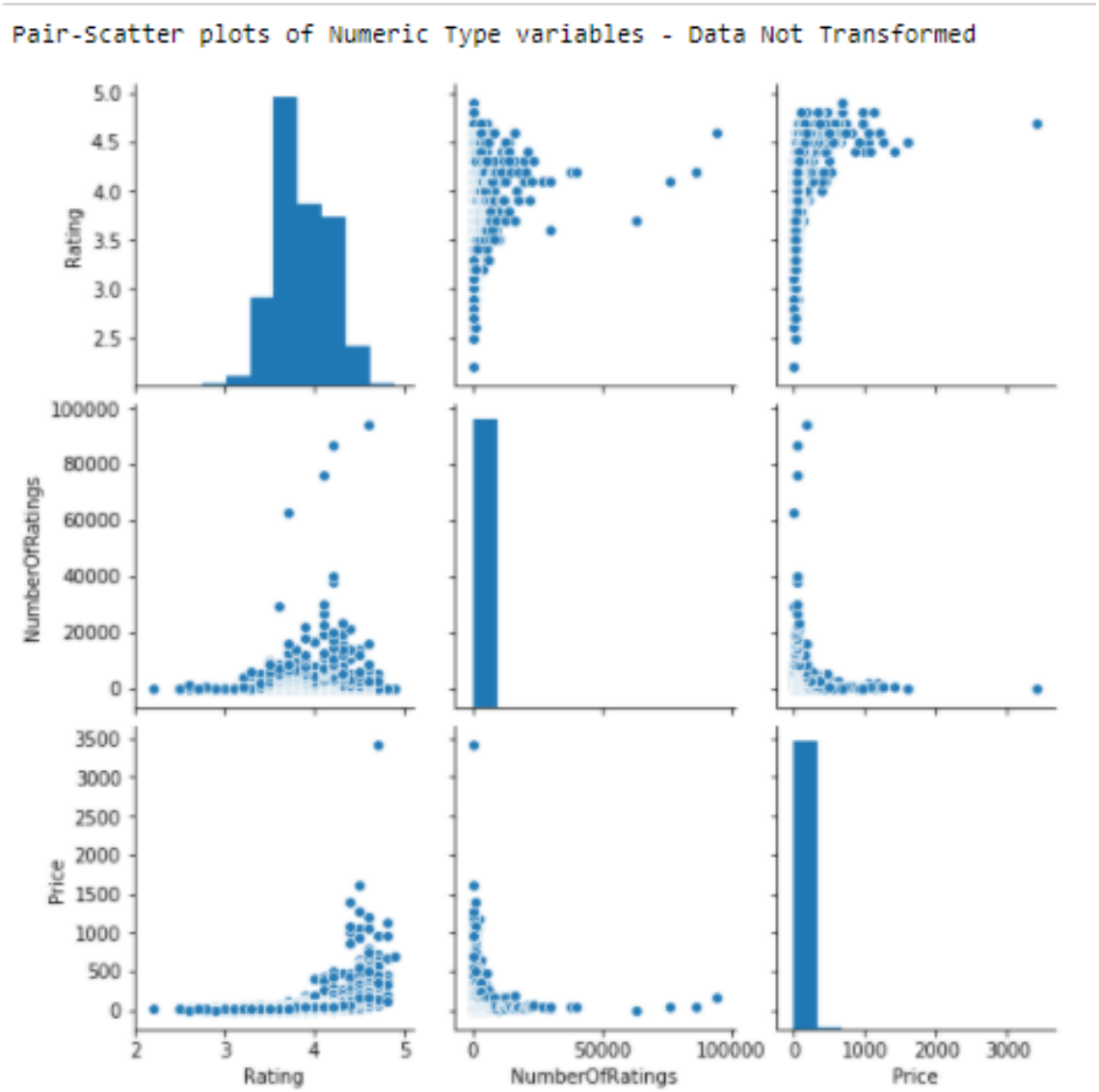
- As well as per the "Rating" variable, any relationships between "Country" and "Price" still fail to appear.
- The 00s decade wines seem to have, overall, a higher price than the other time-periods, although the overlapping of the other 2 categories is evident.

```
Dictionary with features and F-test scores - Target Variable: "Price":
{'Country - Original Categories': 17.78,
 'Country - New Categories (freq-based)': 100.93,
 'Continent': 18.59,
 'Year - Original Categories': 110.06,
```

```
'Year - New Categories (freq-based)': 228.04,
'Time Period': 370.46,
'Wine Style': 72.29}
```

- The F-test scores, by ANOVA, confirmed that Country and Style have the least impact on the wine Price.
- Again, grouping the Years into time-periods proved to be the most impactful feature transformation on the Price variable.

3.4.3. Numeric type attributes

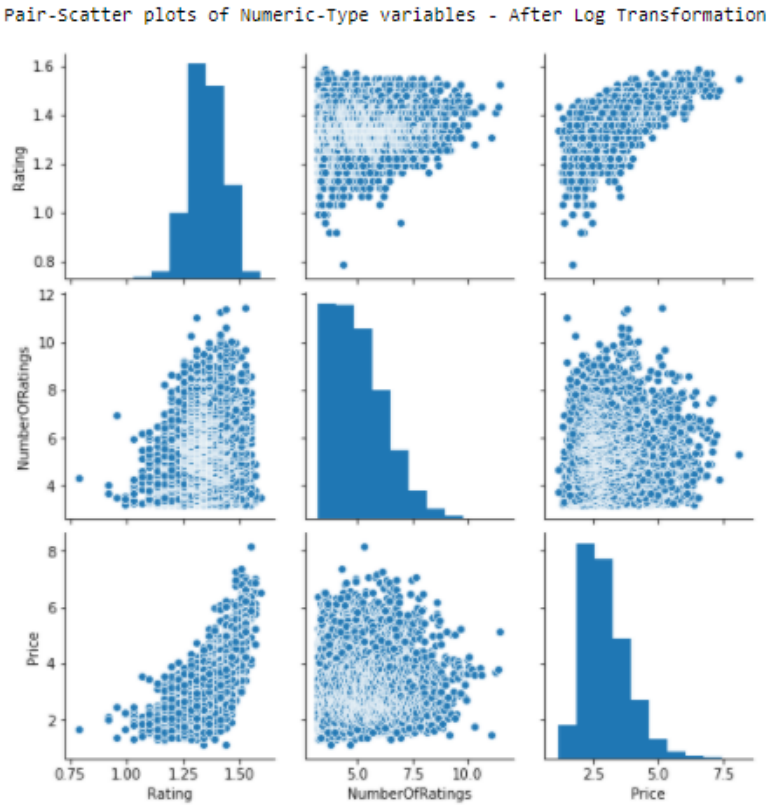


```
Data distribution skeweness score:
Rating               0.007989
Price               14.961501
NumberOfRatings     29.056535
```

- Price and Number-of-Ratings data distribution is very right-skewed (high positive skeweness score).
- Rating distribution is almost normal (very low skeweness score).
- No linear correlations between any of the features (see also dictionary with Pearson Correlation coefficients below).
- There might be potential curvi-linear relationships between some of the variables.

```
Dictionary with Pearson Correlation Coefficients:
{'Number of Ratings VS Rating': 0.07,
 'Price VS Rating': 0.45,
 'Number of Ratings VS Price': 0.02}
```

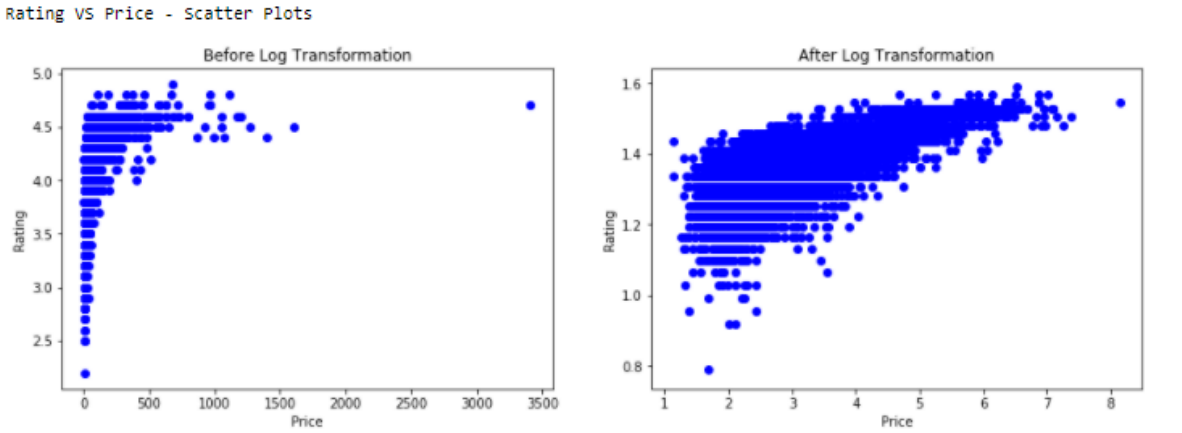
Logarithmic Transformation: data not transformed VS data log transformed

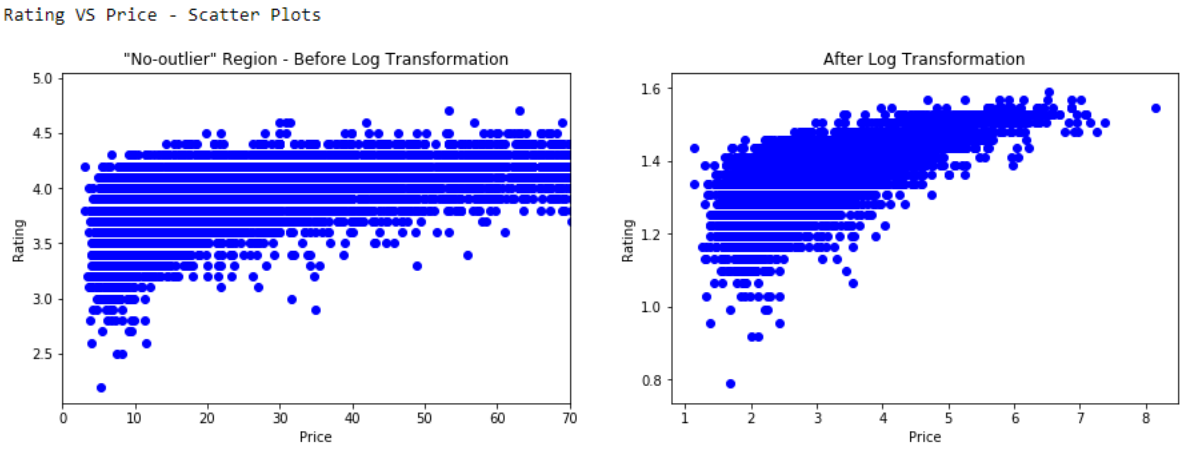


```
Data distribution skeweness score - After Log Transformation:
Rating            -0.265821
Price              1.039141
NumberOfRatings    0.748350

Dictionary with Pearson Correlation Coefficients - After Log Transformation:
{'Number of Ratings VS Rating': 0.12,
 'Price VS Rating': 0.72,
 'Number of Ratings VS Price': 0.07}
```

- After the Log Transformation, Price and NumberOfRatings distribution is "more normal", however Price distribution is still quite right-skewed.
- Rating distribution is more left-skewed than before the log transformation, but still "normal".
- "Rating" and "Price" variables are more linearly correlated, but I fail to see any other improvement in the linear correlation of the other variables.

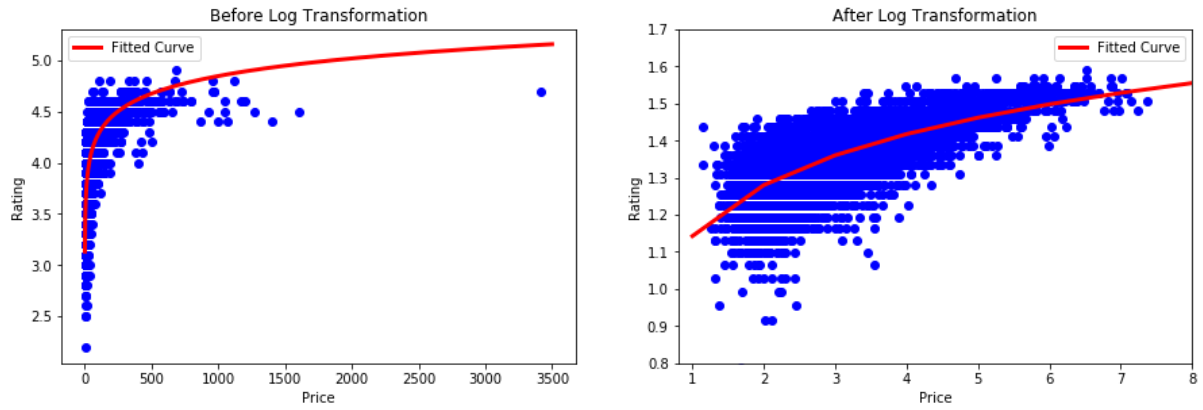
➢ Rating as a function of Price



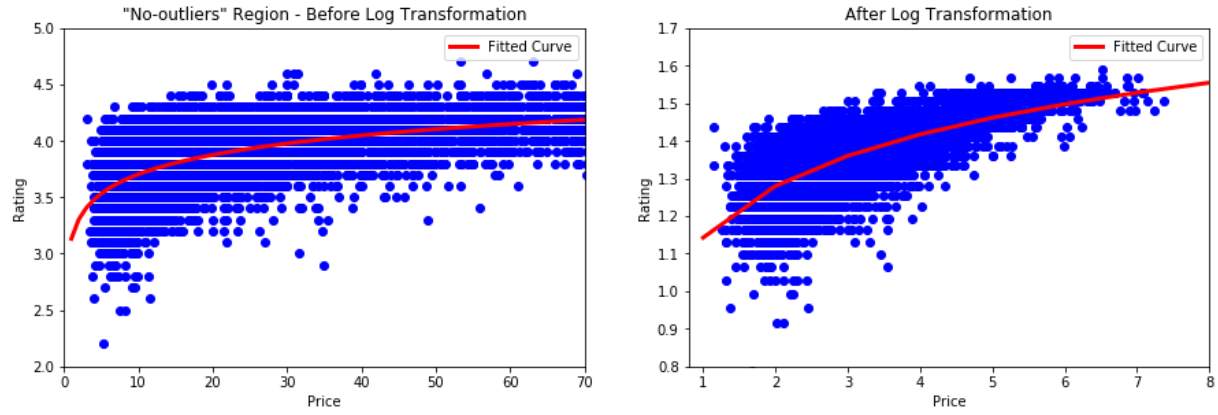


Fitted curve: logarithmic function.

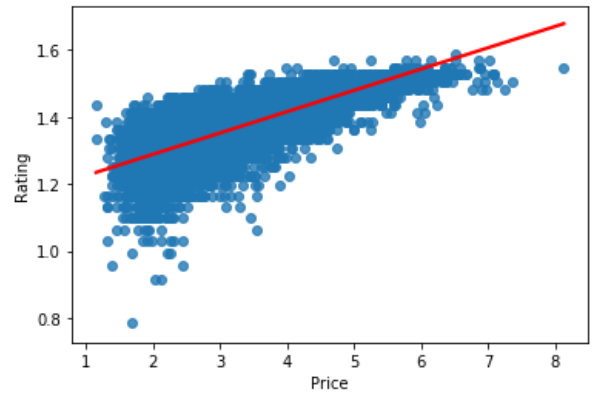
Rating VS Price - Scatter Plots with Fitted Curves



Rating VS Price - Scatter Plots with Fitted Curves



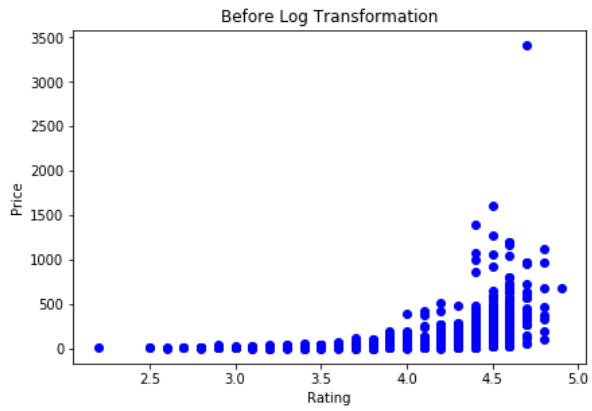
Rating VS Price - Scatter plot with Regression Line

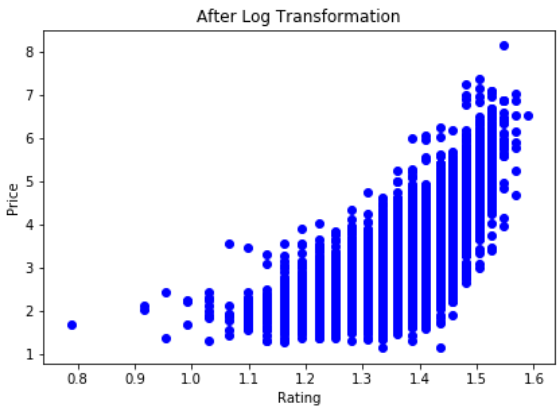


- A logarithmic regression seems to be a good fitted curve to explain the relationship of Rating as a function of Price.
- After log transformation, a linear correlation between the variables is also possible.

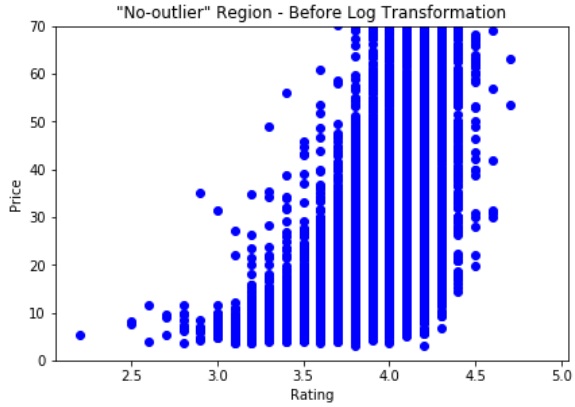
➢ Price as a function of Rating

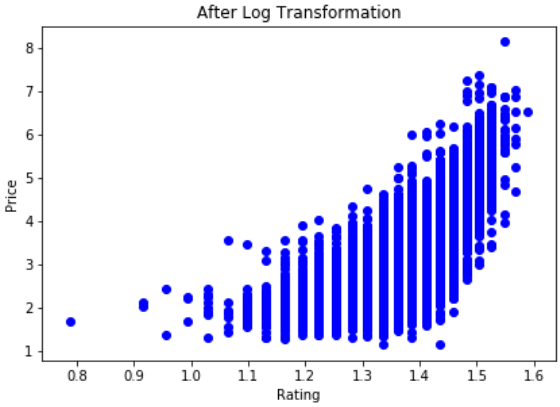
Price VS Rating - Scatter Plots



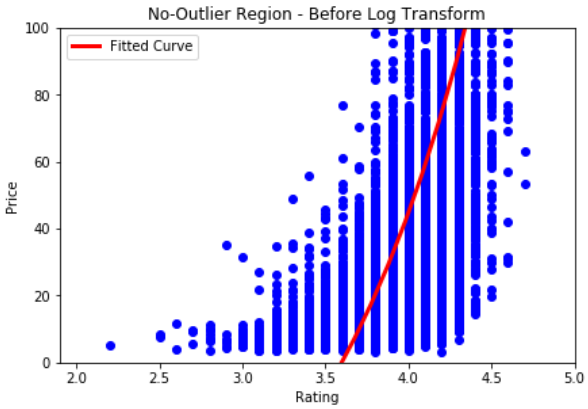


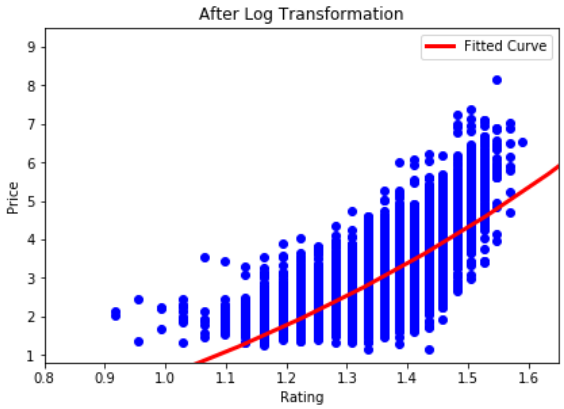
Price VS Rating - Scatter Plots



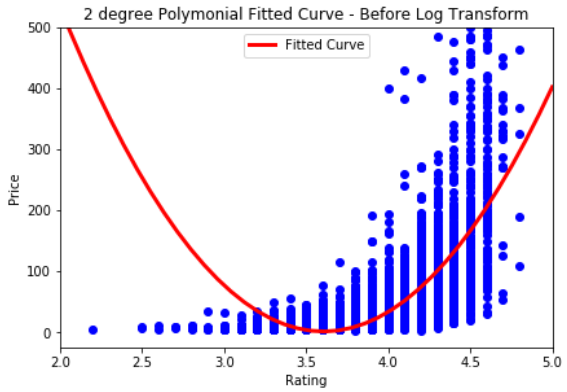


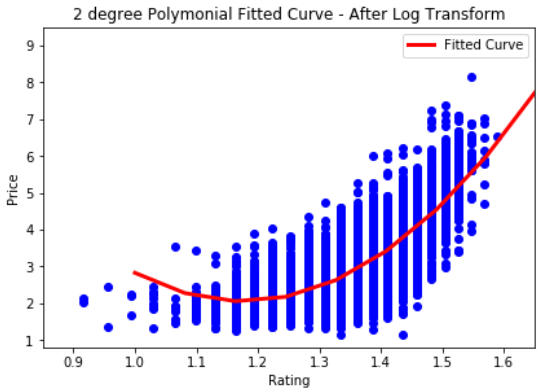
Fitted curve: exponential function
Price VS Rating - Scatter Plots with Fitted Curves



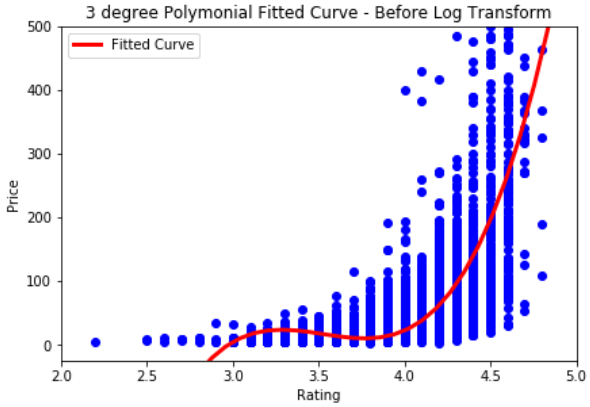


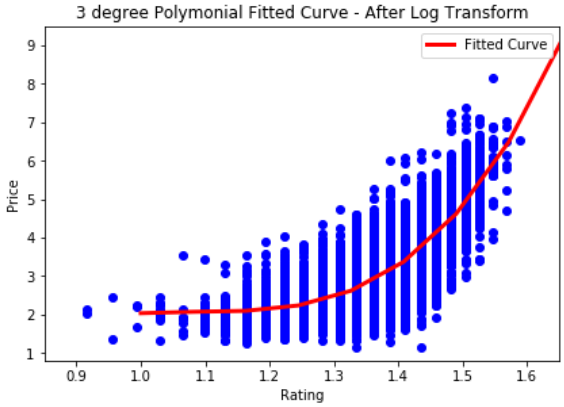
Price VS Rating - Scatter Plots with Fitted Curves



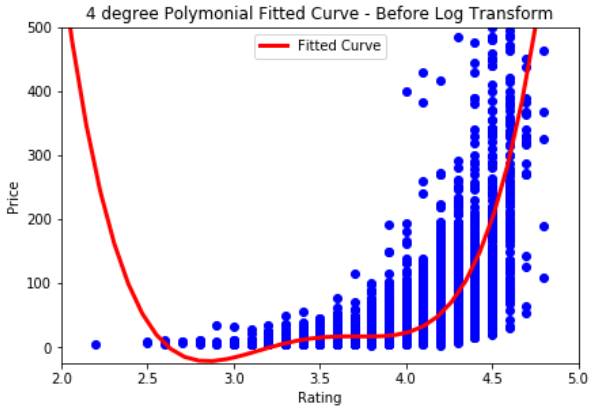


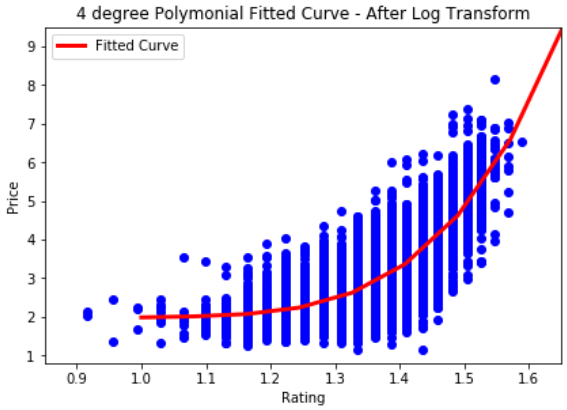
Price VS Rating - Scatter Plots with Fitted Curves



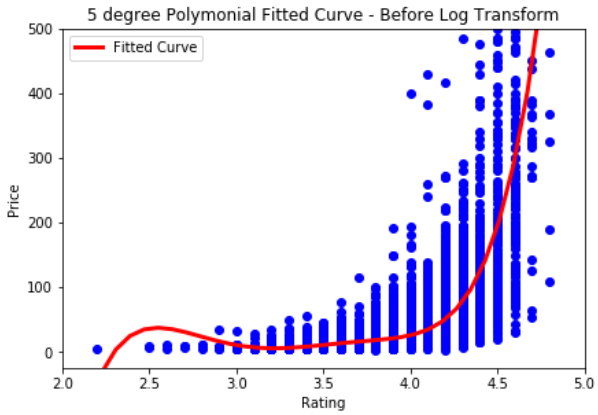


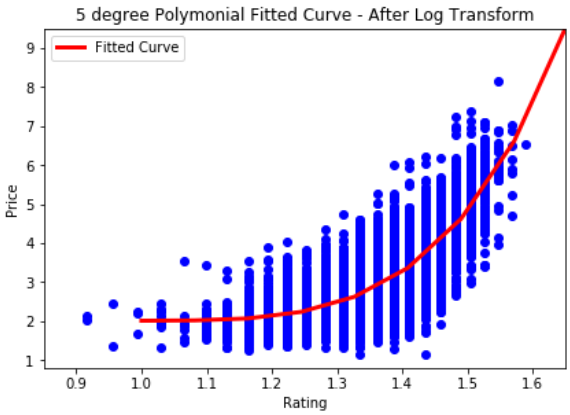
Price VS Rating - Scatter Plots with Fitted Curves



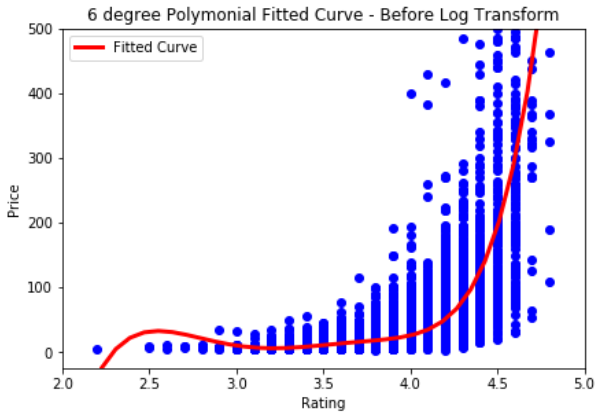


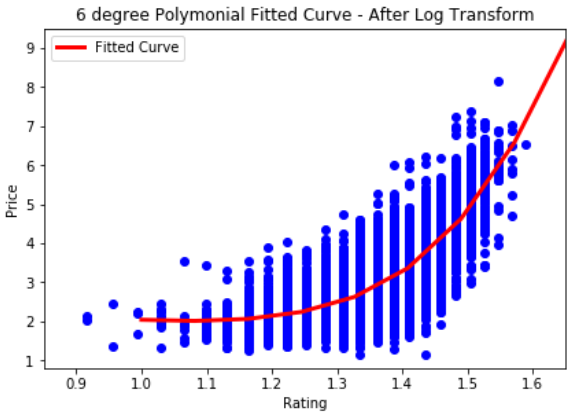
Price VS Rating - Scatter Plots with Fitted Curves



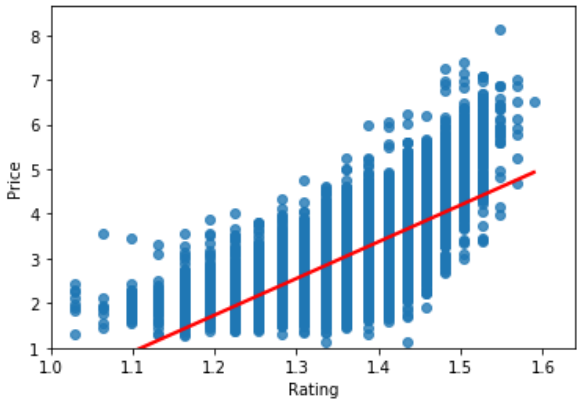


Price VS Rating - Scatter Plots with Fitted Curves





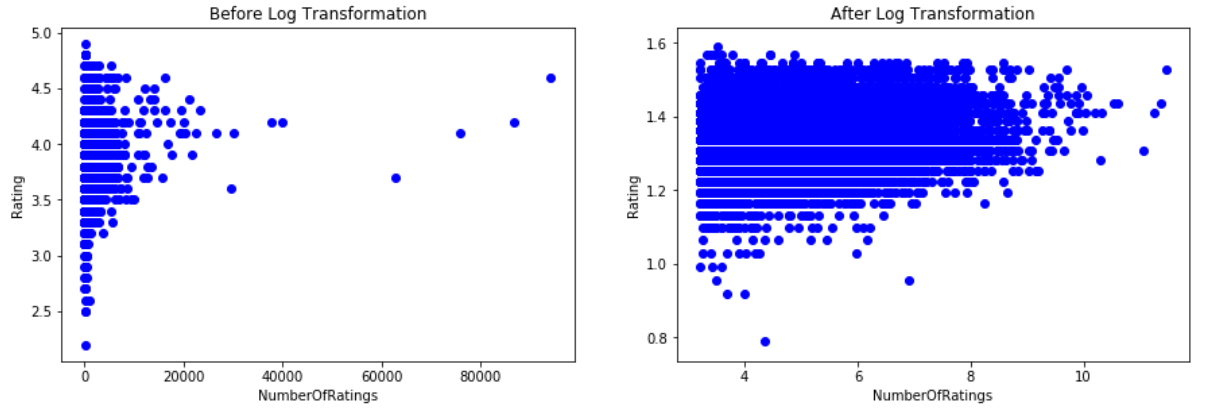
Price VS Rating - Scatter plot with Regression Line



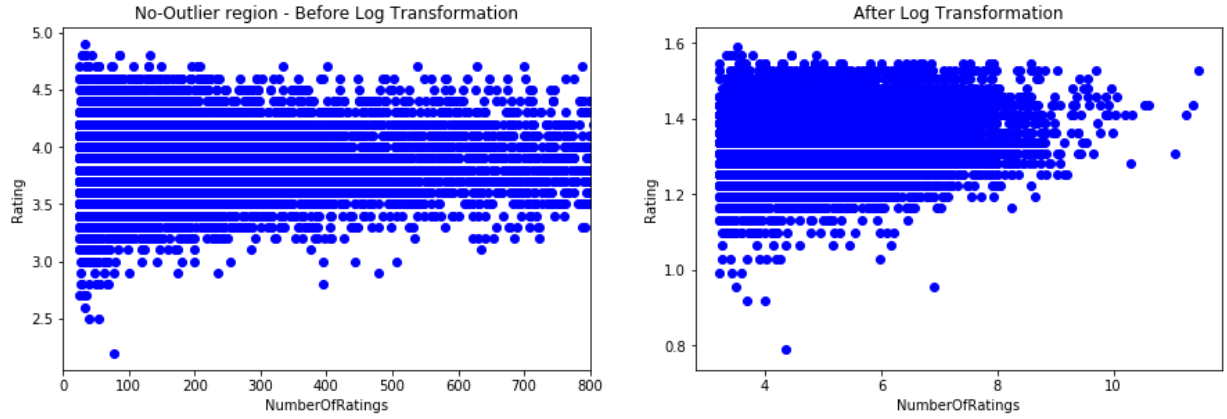
- Overall, data after the log transformation is better explained, either by an exponential or polynomial function, than data before the log transformation.
- A polynomial curve seem to be the best regression model to represent the relationship of Price as a function of Rating.
- After log transformation, a linear correlation between the variables is also possible.

➢ Rating as a function of Number of Ratings

Rating VS NumberOfRatings - Scatter Plots



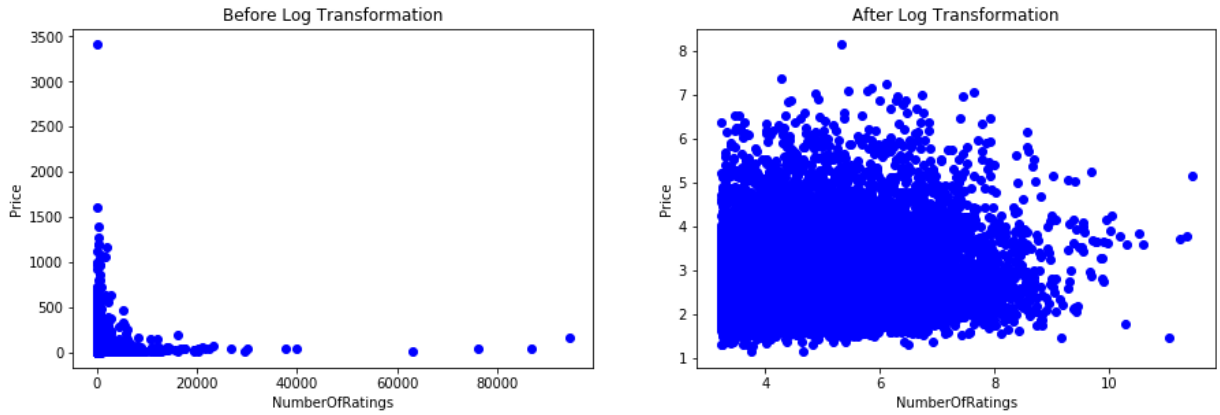
Rating VS NumberOfRatings - Scatter Plots



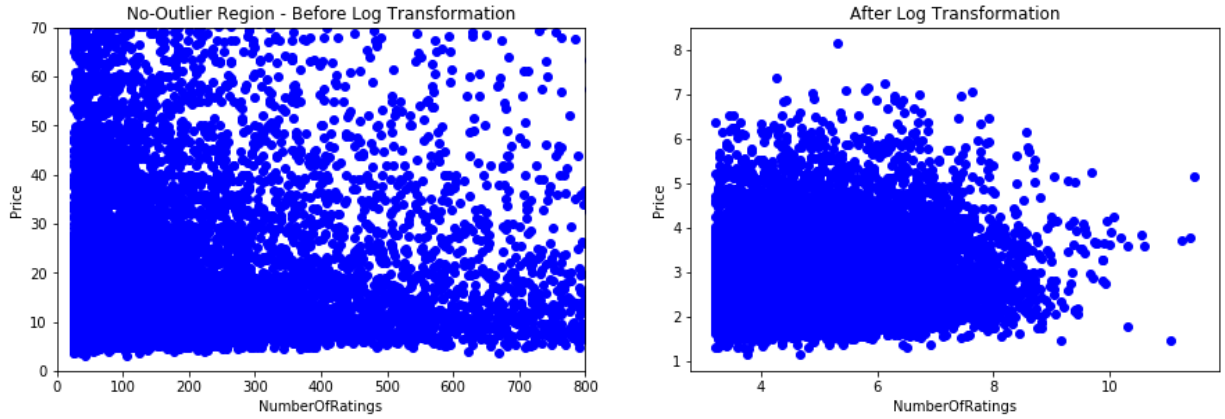
- It seems that the variable Number of Ratings cannot really explain the Rating target.

➢ Price as a function of Number of Ratings

Price VS NumberOfRatings - Scatter Plots



Price VS NumberOfRatings - Scatter Plots



- Again, it seems that the variable Number of Ratings cannot really explain the Price variable.

3.5. Feature Engineering & Variable Transformations

As already observed, Log-Transformation turned the data distribution into a more "normal" curve, and strengthened potential relationships, either linear or curvi-linear, between the variables Rating and Price.

Based on our analysis, the Time-periods variable was the most impactful attribute in relation to Rating and Price. Therefore, my feature selection will include:

- Rating (as explanatory variable for the Price target)
- Price (as explanatory variable for the Rating target)
- Time Periods

In order to build a dataset with features ready for potential modeling, I've applied the followings:

- **Logarithmic transformation**
- **Nominal encoding** (One-Hot encoding) for Time-Periods categorical variable.
- **Robust Scaler**, as scaling approach, in order to minimize the outliers' impact.

Features after the variable transformations:

```
array([[ 0.72191888,  1.50114416,  0.        ,  0.        ,  0.        ],
       [ 0.95113998, -0.02407561,  0.        ,  0.        ,  0.        ],
       [ 0.        , -0.64039824,  0.        ,  0.        ,  0.        ],
       ...,
       [ 0.48717386,  0.78619611,  0.        , -1.        ,  1.        ],
       [ 0.72191888,  0.84596097,  0.        , -1.        ,  1.        ],
       [ 1.60811564,  1.99068646,  0.        , -1.        ,  1.        ]])
```

4. Hypothesis Testing

1. **Question**: Moldova wines have the highest average rating score, that is 4.175 (VS 3.87 population average rating). Is the difference between Moldova's wine average rating and the population average rating statistically significant?

Null hypothesis (H0): The difference is not significant, but just due to random chance or natural variability of the data. Moldova wines' average rating score = 3.87.

Alternative hypothesis (H1): The difference is statistically significant. Moldova wines have actually a higher average rating score than the population average. Moldova wines' average rating score > 3.87.

2. **Question:** 1989-1993 wines have an average price of 794 euros, vs a population average price of 33 euros per wine. Is the difference between 1989-1993 average wine price and population average price statistically significant (or due to chance)?

Null hypothesis (H0): The difference is not significant, but just due to random chance or natural variability of the data. 1989-1993 wines' average price = 33.03 euros.

Alternative hypothesis (H1): The difference is statistically significant. 1989-1993 wines have actually a higher average price than the population average. 1989-1993 wines' average price > 33.03 euros.

3. **Question**: Gaja winery (the 5th winery per number of wines), is the winery, amongst the top 10, with the highest average rating (4.37 VS 3.87 (population average)) and highest average price (197.15 euro VS 33.03 euro (population average)). Is the difference between Gaja average rating/price and population average rating/price statistically significant?

Null hypothesis (H0): The difference is not significant, but just due to random chance or natural variability of the data. Gaja average rating = 3.87. Gaja average price = 33.03 euro.

Alternative hypothesis (H1): The difference is statistically significant. Gaja wines have actually a higher average rating and price than the population average. Gaja average rating > 3.87. Gaja average price > 33.03 euro.

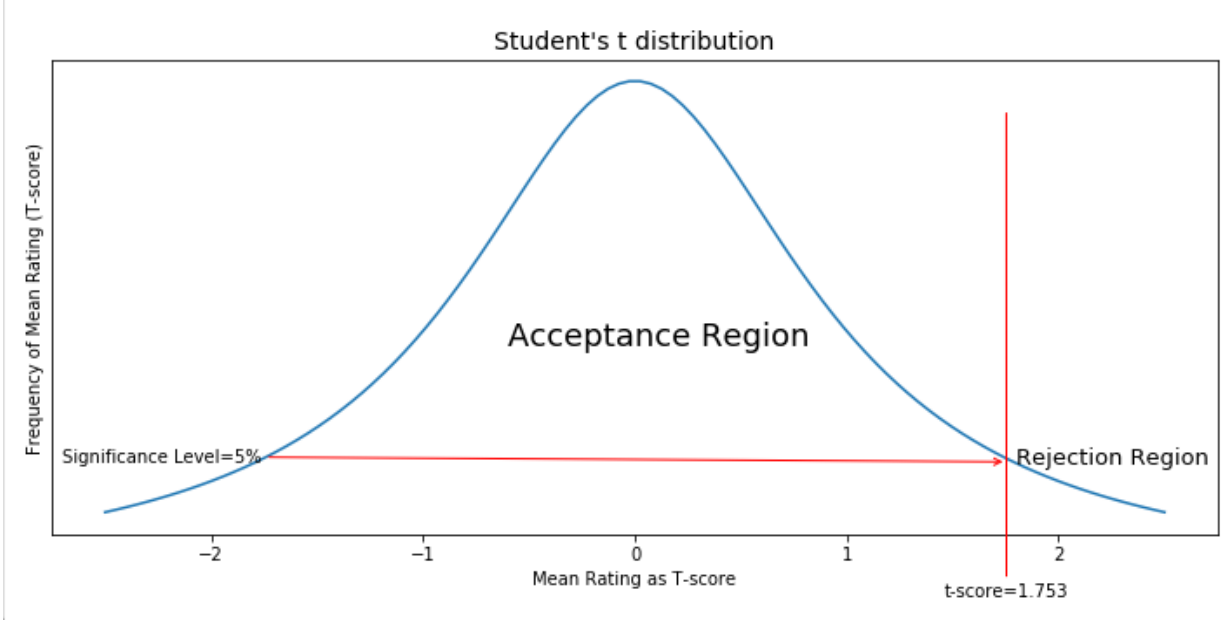
4.1 Significance Test

(For the scope of this project we conducted a significance test only for the first hypothesis.)

To conduct a Significance Test, I have followed a frequentist approach, that is I want to obtain a decision boundary to decide whether we accept or reject the Null Hypothesis (H0).

Since our sample size is less than 30 observations, and we know that the Rating variable is approximately Normally distributed, the appropriate test-statitsic is the **One Sample T-test** (Student's t Distribution).

- **Test Statistic:** 1-sample t-test (right tailed).
- **Significance Level**: 5% (P-Value = 0.05).
- **Confidence Interval**: 95%.
- **Rejection Region**: Since we are performing a 1-tailed test (right-tailed), the t-critical value is 1.753. T-scores higher than 1.753 will lead to the rejection of H0.
- **Acceptance Region:** T-scores lower than 1.753.




```
T-score: 6.35
Probability of sample's mean rating to happen: 0.0013%.
Null hypothesis (Ho) is rejected
```

The obtained t-score is greater than our t-critical value (6.35 > 1.753), that means we can consider the difference between the Moldova's wine mean rating and the population/dataset mean rating as statistically significant (or, in other words, Moldova's mean rating had less than 5% of probability to happen (actually only 0.0013%), if the Null Hypothesis was true, so we cannot consider it as just due to random chance). Therefore, **we can reject the null-hypothesis, with a confidence level of 95%,** and state that Moldova wines' mean rating is greater than 3.88.

5. Discussion

1. **Outlier policy.** In performing this analysis, I decided not to remove or transform the outliers. In variables like Price and Number-of-Ratings, the outliers are, approximately, 10% of the data. Since outliers are rare by definition, I didn't reckon the idea of removing the outliers as a legitimate choice, in this particular case. In the absence of a specific reason to delete the outliers (like if the outliers had been errors in the dataset), I considered the outliers as part of the natural distribution of data. This has, obviously, increased the data variability, although I managed to reach interesting insights despite of the presence of outliers. I just want to point out that, instead of removing or transforming the outliers, I dealt with them in 2 ways:
    a. Applying a logarithmic transformation on the data. The transformation made the data distribution more normal, and strengthened the correlation between the variables.
    b. While visually inspecting potential relationships between the variables, I focused on the "non-outlier" region, in order to not let the analysis be biased by extreme outliers.
2. **Variable Correlations.** Relationship between the variables Rating and Price can be either linear (after log transform), or curvilinear:
    a. Price could be used as explanatory variable for Rating, using a logarithmic regression.
    b. The curvilinear relationship between Price (as explanatory variable) and Rating (as target) could be either exponential or polynomial (degree: 2 - 6).
3. **Hypothesis Testing.** Regarding the significance test on the first hypothesis about wines from Moldova, the outcome is quite surprising: I, myself, a wine enthusiast and never heard about Moldova wines. The test's result is, anyway, unequivocal: my next wine purchase will definitely be a bottle from Moldova!
4. **Geospatial analysis.** It would be extremely interesting creating a Choropleth Map of the Wine Regions; for this purpose, a file (ideally either a geojson or topojson file) with regions' coordinates is needed.
5. **Further analyses.** Attributes such as Region and Winery are of no real value as explanatory variables for targets like Rating and Price. Although it might be interesting to analyze the attributes' relationships and impact in ad hoc single-country analyses.

6. Conclusion

In this project I've analyzed a dataset about wine, from Vivino.com. Quality of data was pretty good, with neither missing values nor, basically, duplicates. I've adopted a descriptive approach, to analyze data current status, find insights, and uncover relationships between variables.

During the Data Cleaning stage, I analyzed the data distribution, and performed outlier and frequency distribution (for categorical variables) analysis.

The Exploratory Data Analysis is made of 2 parts:

- Descriptive Statistics.
- Data Mining.

Descriptive Statistics. In this part I analyzed the data overall current status, discussing general data information, such as: top producing country stats, country ranking (rating and price), year distribution, product price, style breakdown, supplier analysis...

Data Mining. Uncovering correlations, patterns, and relationships between variables. Considering both Rating and Price as target variables, I analyzed the impact of the other attributes, categorical as well as numerical, using a variety of techniques: category boxplots, Analysis of Variance (ANOVA), scatter plots, logarithmic transformations, linear and curvi-linear regressions, Pearson Correlation Coefficients...

Results are as follows:

- The most impactful categorical variable over the targets (both Rating and Price) was the Year attribute; value categorization in time-periods ("pre-2000 years and N.V.", "00s years", "10s years") proved to be most significant variable transformation. Country and Style explanatory value, over the targets, is not significant.
- Rating and Price, after a logarithmic transformation, showed a linear tendency (Pearson Correlation > 0.7); that means that one variable can be used to explain/predict the other, using a linear regression model.
- A logarithmic regression model could be used to explain/predict Rating, using Price as the independent variable.
- An exponential regression model could be used to explain/predict Price, using Rating as the independent variable.
- Polynomial regressions could be used to explain/predict Price, using Rating as the independent variable (degrees from 3 to 6 seem to work better).
- The Number-of-Ratings variable doesn't seem to have any impact, or explanatory value, over the target variables.

Recommended next steps. Data is already transformed (using one-hot type of encoding and robust scaling approach) and ready for model development and testing. Following tests should be conducted to find the best regression model:

- Linear or curvilinear (logarithmic, exponential, polynomial) regressions
- After or before logarithmic transformation
- With or without Time-period variable.

The performed analysis proved to be very useful to understand the current status of data, and it's a robust initial exploration to further develop parametric machine learning models.

7. Appendix

Reference Notebook link:

https://github.com/SebastianoDenegri/eda_vivino_dataset/blob/main/EDA_wineds_vivino.ipynb

Linkedin Article Link:

https://www.linkedin.com/pulse/wine-rating-price-patterns-trends-insights-sebastiano-denegri