EDA of Wine Reviews

Vikas Ranjan

DSC-530 – Data Science, Bellevue University

Professor Shankar Parajulee

Aug. 08, 2020

Abstract

Wine, a much loved alcoholic drink has been produced and enjoyed since thousands of years. It is typically made from fermented grapes. Different varieties of grapes and strains of yeasts produce different styles of wine. This dataset consists of details of 129971 wines reviews produced across the globe by different wineries. We would be performing EDA on the wine reviews dataset by analyzing the variables in the dataset. We would be performing statistical functions to uncover some insights from the dataset.

Dataset Variables

1. Country
2. Description
3. Designation
4. Points
5. Price
6. Province
7. Region_1
8. Region_2
9. Taster_name
10. Taster_twitter_handle
11. Title
12. Variety
13. WineryWine

Data Cleanup

- This wine review dataset when loaded into a dataframe has 14 columns and 129971 Rows. Below is the snapshot of missing data in the dataset:

    - There are 9 columns that have missing values.
    - There is 1 column having greater than 50% missing value.
    - There is 1 column having greater than 40% missing value.
    - There is 1 column having greater than 30% missing value.
    - There are 4 columns having greater than 20% missing value.
    - There are 5 columns having greater than 10% missing value.

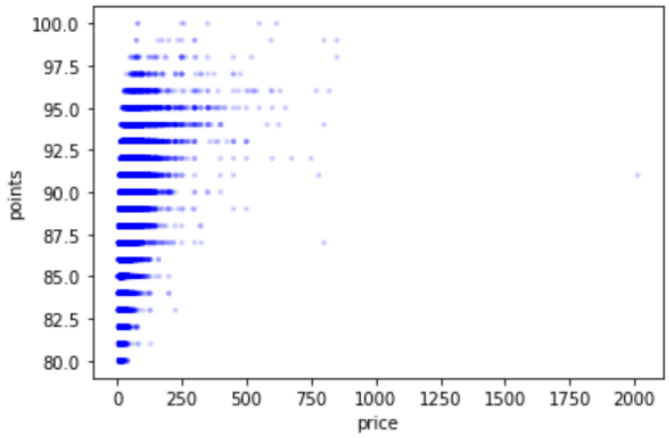
- Below is the snapshot for variable wise missing values.

	Zero Values	Missing Values	% of Total Values
region_2	0	79460	61.1
designation	0	37465	28.8
taster_twitter_handle	0	31213	24
taster_name	0	26244	20.2
region_1	0	21247	16.3
price	0	8996	6.9
country	0	63	0
province	0	63	0
variety	0	1	0

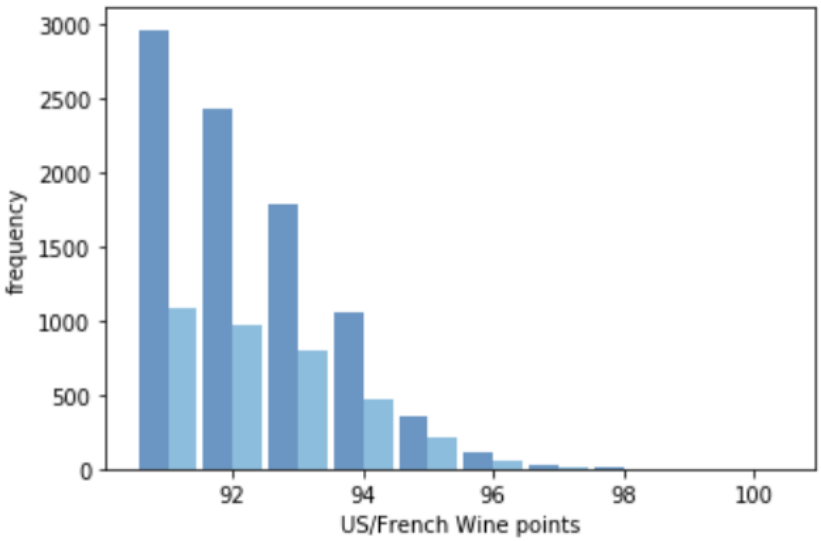
- There are duplicates in the description field, therefore we need to remove them.
- First variable “"Unnamed: 0” in the data frame is just the serial #, therefore it is safe to remove them as there is no benefit of it to provide insight about data.
- We also need to remove the region2 from data frame since majority of it is not having any value.
- We also need to null values from the dataframe.
- There are 47660 rows and 11 columns, after all the cleanup and removing all rows with null values.

Outcomes

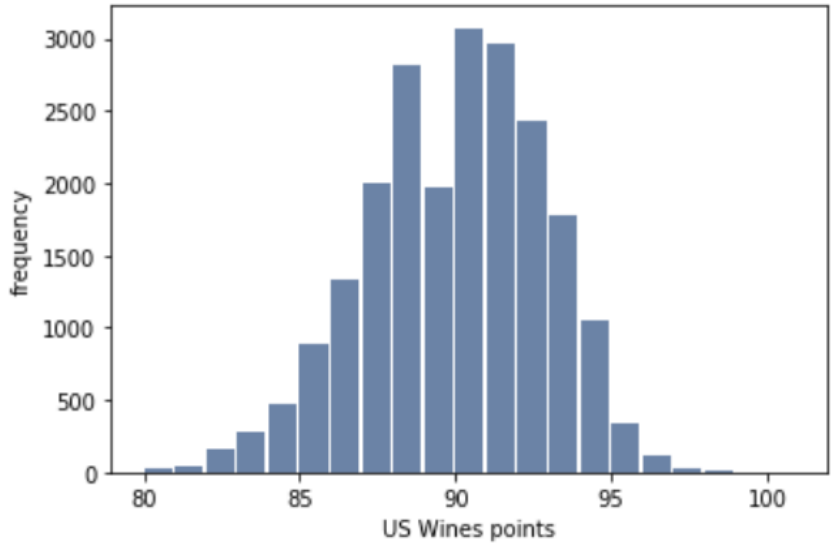
- Scatter Plot of price and points of the wines.



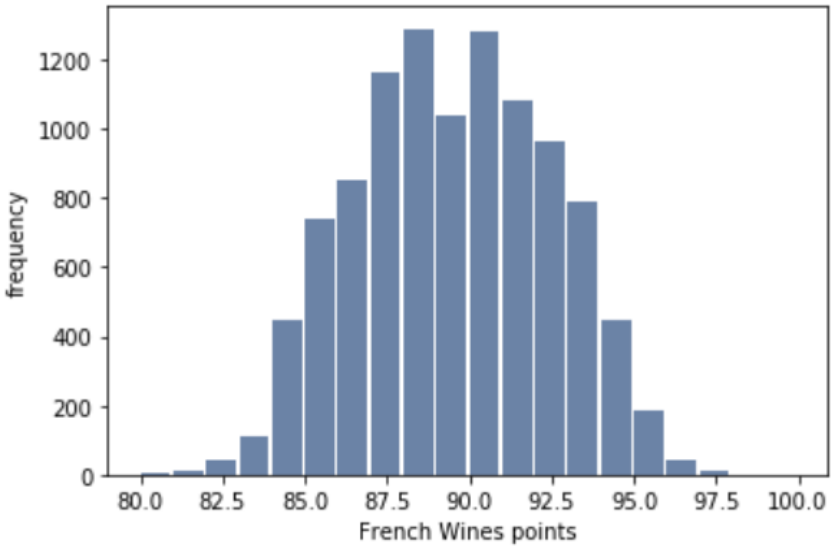
- Histogram of wine scores/price and comparison of US and French wines.



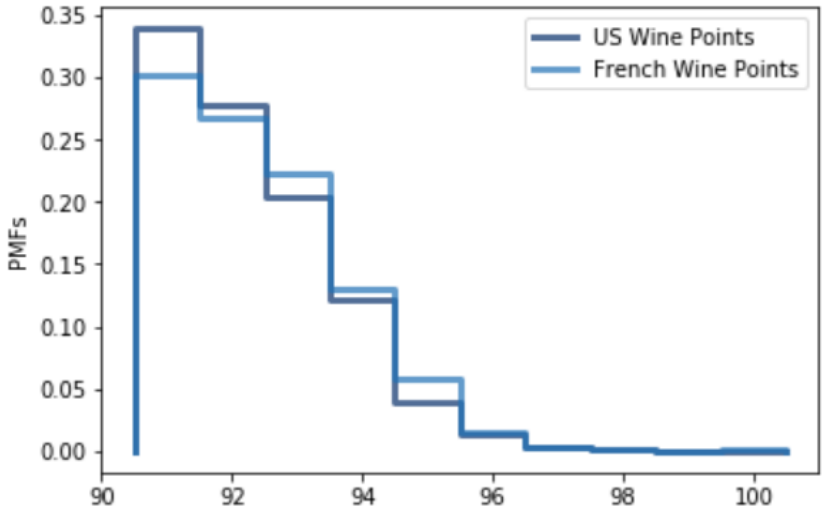
- Eliminating the outliers in the dataset, created dataframes for US Wines and French wines which are priced less than 250. Below histogram suggests that most of the US Wines are having a score of 91.



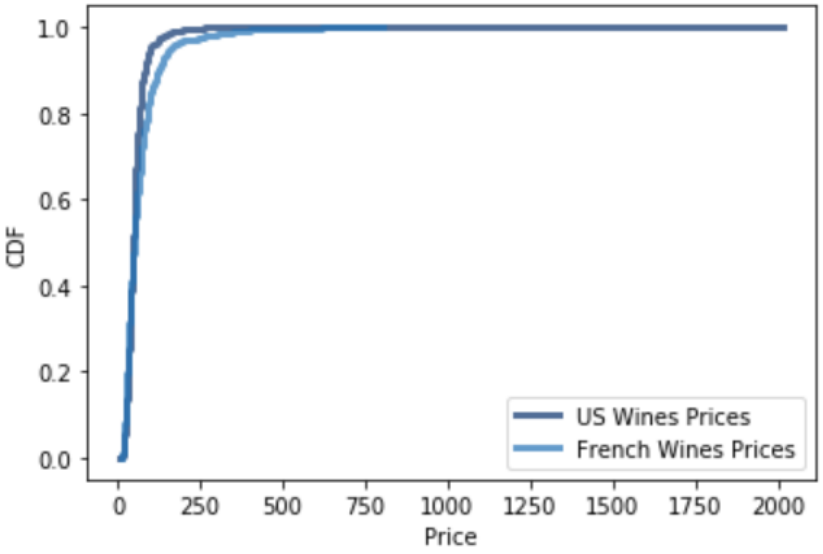
- Below histogram suggests that most of the French wines are having a score of 88.5.



- PMF of wine scores.



- CDF of price between US and French wines which have score greater than 90.



- Created a model to predict the price of the wine based on the score of the wine. And based on it, tried to predict what might be the price of a wine which has a score of 97.

Dep. Variable:	price	**R-squared:**	0.382
Model:	OLS	**Adj. R-squared:**	0.382
Method:	Least Squares	**F-statistic:**	2.941e+04
Date:	Sat, 08 Aug 2020	**Prob (F-statistic):**	0.00
Time:	08:47:19	**Log-Likelihood:**	-5.1033e+05
No. Observations:	47660	**AIC:**	1.021e+06
Df Residuals:	47658	**BIC:**	1.021e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9043.0587	99.442	90.938	0.000	8848.152	9237.966
points	0.6205	0.004	171.502	0.000	0.613	0.628

Omnibus:	801.058	**Durbin-Watson:**	1.851
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	682.472
Skew:	0.232	**Prob(JB):**	6.36e-149
Kurtosis:	2.641	**Cond. No.**	5.52e+04

-

After proceeding far along in this final assignment, I felt I should have chosen a dataset with more numeric variables. One of the biggest challenge I faced was that I only had 2 numeric variables (price and points) to play with. When I looked that this dataset, I made an assumption that French wines are expensive since they are better. However after performing the EDA, I noticed that US wines on an average are cheaper than French wines when compared with the ones with similar score.

Reference:

Downey, A. (2015). Think stats: Exploratory data analysis.