Codebook for Wine reviews Dataset

Our team

Team name : TeamWine

Team members:

a. Omer Sedakah - omer.sedakah@gmail.com. Air Force pilot, has background in programing (high school, BA) and using data.
b. Julia Korsunsky - juliakors16@gmail.com, Technological data analyst, GOI. has high school background in programming and knowledge in SQL. BA in Psychology and English Literature.
c. Limor Shilony - lshilony@gmail.com, Entrepreneur, co-founder of a digital health startup. Holds a Msc. in Management of Information Systems.

The Data

This data set can be downloaded from Kaggle data sets:

https://www.kaggle.com/zynicide/wine-reviews/downloads/winemag-data-130k-v2.csv/4

https://www.kaggle.com/zynicide/wine-reviews/downloads/winemag-data_first150k.csv/4

Indicate source credentials / data owner

The data was scraped from WineEnthusiast June 2017

Specify data copy rights (if any) and/or publication limitations

Please see link below

https://creativecommons.org/licenses/by-nc-sa/4.0/

Business questions

a. What is the highest rated wine by country / by region ?
b. What grape varieties are the highest rated wine made of ?
c. Are there common descriptions for the highest rated wines ?
d. Is there correlation between wine ratings and their price ?
e. Which winery produced the highest rated wine ?

8. Who (hypothetically) needs to review your business questions before you analyze?

Find an academic article(s) (patent or blog) that relates to data similar to yours.
https://medium.freecodecamp.org/using-data-science-to-understand-what-makes-wine-taste-good-669b496c67ee
The main conclusion on this article is that it was possible to train a machine to predict correctly (97% of the time) the quality of a wine based only on it's description. We believe that it is possible to apply the same module in our dataset.

Variables description

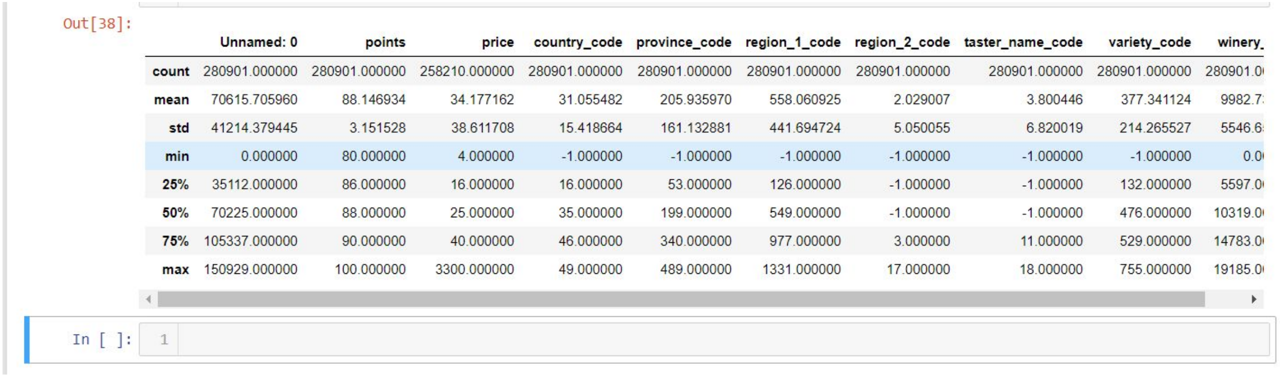
Variable Name	**Description**	**Type**	**Possible values**
country	Country of the wine	string	Country names
description	A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.	string	any

designation	The vineyard within the winery where the grapes that made the wine are from	string	any
points	The number of points WineEnthusiast rated the wine on a scale of 1-100	integer	1-100 whole numbers
price	The cost for a bottle of the wine	integer	whole numbers
province	The province or state that the wine is from	string	Province names
region_1	The wine growing area in a province or state	string	Region names
region_2	Sometimes there are more specific regions specified within a wine growing area	string	Sub region names
taster_name	Name of the person who tasted and reviewed the wine	string	any
taster_twitter_handle	Twitter handle for the person who tasted and reviewed the wine	string	any
title	The title of the wine review, which often contains the vintage	string	any

variety	The type of grapes used to make the wine	string	Grape type names
winery	The winery that made the wine	string	

Summary statistics

We converted all variables with finite number of values to categorical type, so we could run statistic descriptive functions on them.



Out[38]:

	Unnamed: 0	points	price	country_code	province_code	region_1_code	region_2_code	taster_name_code	variety_code	winery_
count	280901.000000	280901.000000	258210.000000	280901.000000	280901.000000	280901.000000	280901.000000	280901.000000	280901.000000	280901.0
mean	70615.705960	88.146934	34.177162	31.055482	205.935970	558.060925	2.029007	3.800446	377.341124	9982.7
std	41214.379445	3.151528	38.611708	15.418664	161.132881	441.694724	5.050055	6.820019	214.265527	5546.6
min	0.000000	80.000000	4.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.0
25%	35112.000000	86.000000	16.000000	16.000000	53.000000	126.000000	-1.000000	-1.000000	132.000000	5597.0
50%	70225.000000	88.000000	25.000000	35.000000	199.000000	549.000000	-1.000000	-1.000000	476.000000	10319.0
75%	105337.000000	90.000000	40.000000	46.000000	340.000000	977.000000	3.000000	11.000000	529.000000	14783.0
max	150929.000000	100.000000	3300.000000	49.000000	489.000000	1331.000000	17.000000	18.000000	755.000000	19185.0

In [ ]:

Github

user name for git account: TeamWine

Email: omer.sedakah@gmail.com

Repository link: https://github.com/TeamWine/IDC-BDA-Exercises