

Wine Tasting Reviews

This dataset is a compilation of different reviews for various wines from various wineries. The dataset was compiled by "zackthoutt" on Kaggle.com. We will be using his dataset to answer some questions during this exploratory analysis:

1. Which country produces the best wine? (Dictated by points)
2. Which tasters give higher scores? Or lower ones?
3. Do comments or descriptions affect the score of the wine? Does the length of a comment affect its score?
4. Which region of Sicily & Sardinia produce the best wine?
5. Which wine is the most expensive?

These questions will help us get a better understanding of the quality of wine in these countries, and can tell us which wines are best.

```
import pandas as pd
df = pd.read_csv('/content/wine.csv')
```

[+ Code](#)
[+ Text](#)


```
df.head()
```

	country	description	designation	points	price	province	taster_name	taster_twitter_handle	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime	NaN	87	14.0	Oregon	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette	Pinot Gris	Rainstorm

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129971 entries, 0 to 129970
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   country                129908 non-null object
1   description            129971 non-null object
2   designation            92506 non-null object
3   points                129971 non-null int64
4   price                 120975 non-null float64
5   province              129908 non-null object
6   taster_name           103727 non-null object
7   taster_twitter_handle  98758 non-null object
8   title                 129971 non-null object
9   variety               129970 non-null object
10  winery                 129971 non-null object
dtypes: float64(1), int64(1), object(9)
memory usage: 10.9+ MB
```

```
df.describe()
```




	points	price
count	129971.000000	120975.000000
mean	88.447138	35.363389
std	3.039730	41.022218
min	80.000000	4.000000
25%	86.000000	17.000000
50%	88.000000	25.000000
75%	91.000000	42.000000
max	100.000000	3300.000000

Here is some of our data. We can see there are several different types of wine. It's produced in many places, has various blends of ingredients, and has many reviews. Let's begin answering our first question: Which country produces the highest rated wine?

```
# Which country produces the highest rated wine?
```

```
df.groupby('country')['points'].mean().sort_values(ascending = False).head(5)
```




	points
country	
England	91.581081
India	90.222222
Austria	90.101345
Germany	89.851732
Canada	89.369650

dtype: float64

Using the code we wrote, we are able to narrow down and sort the various countries by both region and average point value. Based off the results of the code we ran, we can see that England has the highest rated wine, on average. But what about the people who rated these wines? What can we deduce from their ratings?

```
# Which taster gives the lowest scores (points), on average?
```

```
df.groupby('taster_name')['points'].mean().sort_values(ascending = True)
```



points	
taster_name	
Alexander Peartree	85.855422
Carrie Dykes	86.395683
Susan Kostrzewa	86.609217
Fiona Adams	86.888889
Michael Schachner	86.907493
Lauren Buzzeo	87.739510
Christina Pickard	87.833333
Jeff Jenssen	88.319756
Anna Lee C. Iijima	88.415629
Joe Czerwinski	88.536235
Jim Gordon	88.626287
Roger Voss	88.708003
Sean P. Sullivan	88.755739
Kerin O'Keefe	88.867947
Paul Gregutt	89.082564
Mike DeSimone	89.101167
Virginie Boone	89.213379
Matt Kettmann	90.008686
Anne Krebiehl MW	90.562551

dtype: float64

According to this bit of data, Alexander Peartree seems to be more critical of the wines than the other tasters. He rated lower scores on average than the others. Let's continue with analyzing the wine.

```
# Which variety of wine is the most expensive, on average?  
  
df.groupby('variety')['price'].mean().sort_values(ascending = False)
```



price	
variety	
Ramisco	495.000000
Terrantez	236.000000
Francisa	160.000000
Rosenmuskateller	150.000000
Malbec-Cabernet	113.333333
...	...
Roscetto	NaN
Sauvignon Blanc-Sauvignon Gris	NaN
Tempranillo-Malbec	NaN
Vital	NaN
Zelen	NaN

707 rows × 1 columns

dtype: float64

Regarding the most expensive wines, two in particular stand out. The Ramisco and Terrantez are significantly more expensive than all the other varieties. They seem to be a outliers among the more common varieties. Though as you can see from the line of code below, Ramisco is not the highest rated. Terrantez seems to have better value and a higher rating overall.

```
df.groupby('variety')['points'].mean().sort_values(ascending = False)
```



	points
variety	
Terrantez	95.000000
Tinta del Pais	95.000000
Gelber Traminer	95.000000
Bual	94.142857
Sercial	94.000000
...	...
Shiraz-Tempranillo	82.000000
Aidani	82.000000
Picapoll	82.000000
Airen	81.666667
Chancellor	80.500000

707 rows × 1 columns

dtype: float64

Which year of wines has the best score (points), on average?

```
df['year'] = df['title'].str.extract('(\d{4})')
```

```
df.groupby('year')['points'].mean().sort_values(ascending = False).head(5)
```



	points
year	
1969	98.0
1973	96.0
1952	95.5
1927	95.0
1945	95.0

dtype: float64

Here we see the years for the wine in the dataset. It seems wines from the year 1969 have the highest ratings overall. Not the oldest, nor the newest.

✓ Descriptions

Now we are reaching the descriptions from the wine tasters. Here we will see if the following things affect the rating:
the word "depth" being used, the word "fruity" being used, the word "herbal" being used, and the length of the description.

Do reviews with the word "depth" in them tend to get better than average or worse than average points?

```
df['depth'] = df['description'].str.contains('depth')
```

```
df.groupby('depth')['points'].mean().sort_values(ascending = False).head(10)
```

**points**

```
# Do reviews with the word "fruity" in them tend to get better than average or worse than average points?
```

```
df['fruity'] = df['description'].str.contains('fruity')
```

```
df.groupby('fruity')['points'].mean().sort_values(ascending = False).head(10)
```

**points****fruity**

fruity	points
False	88.509749
True	87.614521

dtype: float64

```
# Do reviews with the word "herbal" in them tend to get better than average or worse than average points?
```

```
df['herbal'] = df['description'].str.contains('herbal')
```

```
df.groupby('herbal')['points'].mean().sort_values(ascending = False).head(10)
```

**points****herbal**

herbal	points
False	88.489250
True	87.470019

dtype: float64

```
# What is the relationship between number of characters (description) and points?
```

```
df['description_length'] = df['description'].str.len()
```

```
correlation = df[['description_length', 'points']].corr()
```

```
print(correlation)
```



	description_length	points
description_length	1.00000	0.55776
points	0.55776	1.00000

After running those lines of code, we can now answer the previous questions:

1. Do reviews with the word "depth" in them tend to get better than average or worse than average points? (Yes)
2. Do reviews with the word "fruity" in them tend to get better than average or worse than average points? (No)
3. Do reviews with the word "herbal" in them tend to get better than average or worse than average points? (No)
4. What is the relationship between number of characters (description) and points? (There is a moderate correlation between length and points given)