Creating, Reading and Writing

In [12]:

```
import pandas as pd
```

In [13]:

```
pd.DataFrame({'Yes': [50, 21], 'No': [131, 2]})
```

Out[13]:

	Yes	No
0	50	131
1	21	2

In [14]:

```
pd.DataFrame({'Bob': ['I liked it.', 'It was awful.'], 'Sue': ['Pretty good.',
```

Out[14]:

	Bob	Sue
0	I liked it.	Pretty good.
1	It was awful.	Bland.

In [15]:

```
pd.DataFrame({'Bob': ['I liked it.', 'It was awful.'],
              'Sue': ['Pretty good.', 'Bland.']},
             index=['Product A', 'Product B'])
```

Out[15]:

	Bob	Sue
Product A	I liked it.	Pretty good.
Product B	It was awful.	Bland.

In [16]:

```
pd.Series([1, 2, 3, 4, 5])
```

Out[16]:

```
0    1
1    2
2    3
3    4
4    5
dtype: int64
```

In [17]:

```
pd.Series([30, 35, 40], index=['2015 Sales', '2016 Sales', '2017 Sales'], name
```

Out[17]:

```
2015 Sales    30
2016 Sales    35
2017 Sales    40
Name: Product A, dtype: int64
```

In [18]:
```
wine_reviews = pd.read_csv(r"D:\Data Analytics\Python\Kaggle Pandas\winemag-da
```

In [19]:
```
wine_reviews.shape
```

Out[19]:
```
(65499, 14)
```

In [20]:
```
wine_reviews.head()
```

Out[20]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	t
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	I
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	I

In [21]:
```
wine_reviews = pd.read_csv(r"D:\Data Analytics\Python\Kaggle Pandas\winemag-da
wine_reviews.head()
```

Out[21]:

	country	description	designation	points	price	province	region_1	region_2	taster_name
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O’Keefe
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt

Indexing, Selecting & Assigning

In [22]:
```
pd.set_option('display.max_rows', 5)
```

In [23]:
```
wine_reviews.country
```

Out[23]:
```
0           Italy
1        Portugal
           ...
65497          US
65498       Spain
Name: country, Length: 65499, dtype: object
```

In [24]:

```
wine_reviews['country']
```

Out[24]:

```
0           Italy
1        Portugal
           ...   
65497          US
65498       Spain
Name: country, Length: 65499, dtype: object
```

In [25]:

```
wine_reviews['country'][0]
```

Out[25]:

```
'Italy'
```

Indexing in pandas

In [26]:

```
wine_reviews.iloc[0]
```

Out[26]:

```
country                                                    Italy
description    Aromas include tropical fruit, broom, brimston...
                                     ...                        
variety                                              White Blend
winery                                                   Nicosia
Name: 0, Length: 13, dtype: object
```

In [27]:

```
wine_reviews.iloc[:, 0]
```

Out[27]:

```
0           Italy
1        Portugal
           ...   
65497          US
65498       Spain
Name: country, Length: 65499, dtype: object
```

In [28]:

```
wine_reviews.iloc[:3, 0]
```

Out[28]:

```
0       Italy
1    Portugal
2          US
Name: country, dtype: object
```

In [29]:

```
wine_reviews.iloc[1:3, 0]
```

Out[29]:

```
1    Portugal
2          US
Name: country, dtype: object
```

In [30]:
```
wine_reviews.iloc[[0, 1, 2], 0]
```

Out[30]:
```
0       Italy
1    Portugal
2          US
Name: country, dtype: object
```

In [31]:
```
wine_reviews.iloc[-5:]
```

Out[31]:

	country	description	designation	points	price	province	region_1	region_2	taster_nai
65494	France	Made from young vines from the Vaulorent porti...	Fourchaume Premier Cru	90	45.0	Burgundy	Chablis	NaN	Roger Vc
65495	Australia	This is a big, fat, almost sweet-tasting Caber...	NaN	90	22.0	South Australia	McLaren Vale	NaN	J Czerwin
65496	US	Much improved over the unripe 2005, Fritz's 20...	Estate	90	20.0	California	Dry Creek Valley	Sonoma	N
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	N
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Micha Schachr

Label-based selection

In [32]:
```
wine_reviews.loc[0, 'country']
```

Out[32]:
```
'Italy'
```

In [33]:
```
wine_reviews.loc[:, ['taster_name', 'taster_twitter_handle', 'points']]
```

Out[33]:

	taster_name	taster_twitter_handle	points
0	Kerin O’Keefe	@kerinokeefe	87
1	Roger Voss	@vossroger	87
...	...	...	...
65497	NaN	NaN	90
65498	Michael Schachner	@wineschach	90

65499 rows × 3 columns

Manipulating the index

In [34]:
```
wine_reviews.set_index("title")
```

Out[34]:

	country	description	designation	points	price	province	region_1	region_2	taster
title									
Nicosia 2013 Vulkà Bianco (Etna)	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	(
Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Rog
...	...	...	...	...	...	...	...	...	
Hendry 2004 Block 24 Primitivo (Napa Valley)	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	
Bodegas Dios Baco S.L. NV Manzanilla Sherry (Jerez)	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Sch

65499 rows × 12 columns

Conditional selection

In [35]:

```
wine_reviews.country == 'Italy'
```

Out[35]:

```
0         True
1        False
         ...
65497    False
65498    False
Name: country, Length: 65499, dtype: bool
```

In [36]:

```
wine_reviews.loc[wine_reviews.country == 'Italy']
```

Out[36]:

	country	description	designation	points	price	province	region_1	region_2	taster
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	C
6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	C
...	...	...	...	...	...	...	...	...	
65477	Italy	Made of 65% Merlot, 25% Cabernet Sauvignon, 5%...	Ruit Hora	88	30.0	Tuscany	Bolgheri	NaN	C
65478	Italy	Aromas suggesting French oak, coconut and spic...	NaN	88	36.0	Tuscany	Vino Nobile di Montepulciano	NaN	C

10005 rows × 13 columns

In [37]:
```
wine_reviews.loc[(wine_reviews.country == 'Italy') & (wine_reviews.points >= 9
```

Out[37]:

	country	description	designation	points	price	province	region_1	region_2	taste
120	Italy	Slightly backward, particularly given the vint...	Bricco Rocche Prapó	92	70.0	Piedmont	Barolo	NaN	
130	Italy	At the first it was quite muted and subdued, b...	Bricco Rocche Brunate	91	70.0	Piedmont	Barolo	NaN	
...	...	...	...	...	...	...	...	...	
65365	Italy	Stunning and sophisticated, it leads with inte...	Sanct Valentin	94	40.0	Northeastern Italy	Alto Adige	NaN	(
65399	Italy	Aesthetics and elegance are important values t...	Nectar Dei	91	65.0	Tuscany	Maremma	NaN	

3479 rows × 13 columns

In [38]:

```
wine_reviews.loc[(wine_reviews.country == 'Italy') | (wine_reviews.points >= 9
# Suppose we'll buy any wine that's made in Italy or which is rated above aver
```

Out[38]:

	country	description	designation	points	price	province	region_1	region_2	taster_nan
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Ker O'Kee
6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Ker O'Kee
...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	Na
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Micha Schachn

31430 rows × 13 columns

In [39]:
```
wine_reviews.loc[wine_reviews.country.isin(['Italy', 'France'])]
```

Out[39]:

	country	description	designation	points	price	province	region_1	region_2	taster_nam
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Ker O'Kee
6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Ker O'Kee
...	...	...	...	...	...	...	...	...	
65492	France	A rounded, fruity wine, packed with yellow pea...	Mont-de-Milieu Premier Cru	90	30.0	Burgundy	Chablis	NaN	Roger Vo
65494	France	Made from young vines from the Vaulorent porti...	Fourchaume Premier Cru	90	45.0	Burgundy	Chablis	NaN	Roger Vo

21179 rows × 13 columns

In [40]:

```
wine_reviews.loc[wine_reviews.price.notnull()]
```

Out[40]:

	country	description	designation	points	price	province	region_1	region_2	taster_n
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger '
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gre
...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Mic Schac

60829 rows × 13 columns

Assigning data

In [41]:

```
wine_reviews['critic'] = 'everyone'
wine_reviews['critic']
```

Out[41]:

```
0        everyone
1        everyone
           ...   
65497    everyone
65498    everyone
Name: critic, Length: 65499, dtype: object
```

In [42]:
```
# Or with an iterable of values:

wine_reviews['index_backwards'] = range(len(wine_reviews), 0, -1)
wine_reviews['index_backwards']
```

Out[42]:
```
0        65499
1        65498
         ...
65497        2
65498        1
Name: index_backwards, Length: 65499, dtype: int64
```

Summary Functions and Maps

Summary functions

In [43]:
```
wine_reviews.points.describe()
```

Out[43]:
```
count    65499.000000
mean        88.434037
               ...
75%         91.000000
max        100.000000
Name: points, Length: 8, dtype: float64
```

In [44]:
```
wine_reviews.taster_name.describe()
```

Out[44]:
```
count          51856
unique            19
top       Roger Voss
freq           13045
Name: taster_name, dtype: object
```

In [45]:
```
wine_reviews.points.mean()
```

Out[45]:
```
88.43403716087269
```

In [46]:
```
wine_reviews.taster_name.unique()
```

Out[46]:
```
array(['Kerin O’Keefe', 'Roger Voss', 'Paul Gregutt',
       'Alexander Peartree', 'Michael Schachner', 'Anna Lee C. Iijima',
       'Virginie Boone', 'Matt Kettmann', nan, 'Sean P. Sullivan',
       'Jim Gordon', 'Joe Czerwinski', 'Anne Krebiehl\xa0MW',
       'Lauren Buzzeo', 'Mike DeSimone', 'Jeff Jenssen',
       'Susan Kostrzewa', 'Carrie Dykes', 'Fiona Adams',
       'Christina Pickard'], dtype=object)
```

In [47]:

```
wine_reviews.taster_name.value_counts()
```

Out[47]:

```
taster_name
Roger Voss            13045
Michael Schachner      7752
                      ...
Fiona Adams              11
Christina Pickard         4
Name: count, Length: 19, dtype: int64
```

Maps

In [48]:

```
review_points_mean = wine_reviews.points.mean()
wine_reviews.points.map(lambda p: p - review_points_mean)
```

Out[48]:

```
0       -1.434037
1       -1.434037
           ...
65497    1.565963
65498    1.565963
Name: points, Length: 65499, dtype: float64
```

In [49]:
```
def remean_points(row):
    row.points = row.points - review_points_mean
    return row

wine_reviews.apply(remean_points, axis='columns')
```

Out[49]:

	country	description	designation	points	price	province	region_1	region_2	taster_
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	-1.434037	NaN	Sicily & Sardinia	Etna	NaN	O'
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	-1.434037	15.0	Douro	NaN	NaN	Roger
...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	1.565963	31.0	California	Napa Valley	Napa	
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	1.565963	10.0	Andalucia	Jerez	NaN	M Scha

65499 rows × 15 columns

In [50]:
```
wine_reviews.head(1)
```

Out[50]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	t
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	

In [51]:

```
review_points_mean = wine_reviews.points.mean()
wine_reviews.points - review_points_mean
```

Out[51]:

```
0       -1.434037
1       -1.434037
           ...
65497    1.565963
65498    1.565963
Name: points, Length: 65499, dtype: float64
```

In [52]:

```
wine_reviews.country + " - " + wine_reviews.region_1
```

Out[52]:

```
0            Italy - Etna
1                     NaN
                ...
65497    US - Napa Valley
65498       Spain - Jerez
Length: 65499, dtype: object
```

I'm an economical wine buyer. Which wine is the "best bargain"? Create a variable bargain_wine with the title of the wine with the highest points-to-price ratio in the dataset.

In [53]:

```
bargain_wine = wine_reviews.loc[(wine_reviews.points / wine_reviews.price).idx
bargain_wine
```

Out[53]:

```
'Bandit NV Merlot (California)'
```

There are only so many words you can use when describing a bottle of wine. Is a wine more likely to be "tropical" or "fruity"? Create a Series descriptor_counts counting how many times each of these two words appears in the description column in the dataset. (For simplicity, let's ignore the capitalized versions of these words.)

In [54]:

```
tropical = wine_reviews.description.map(lambda x: "tropical" in x).sum()
fruity = wine_reviews.description.map(lambda x: "fruity" in x).sum()
descriptor_counts = pd.Series([tropical, fruity], index=['tropical', 'fruity']
```

In [55]:

```
descriptor_counts
```

Out[55]:

```
tropical    1813
fruity      4632
dtype: int64
```

We'd like to host these wine reviews on our website, but a rating system ranging from 80 to 100 points is too hard to understand - we'd like to translate them into simple star ratings. A score of 95 or higher counts as 3 stars, a score of at least 85 but less than 95 is 2 stars. Any other score is 1 star.

Also, the Canadian Vintners Association bought a lot of ads on the site, so any wines from Canada should automatically get 3 stars, regardless of points.

Create a series star_ratings with the number of stars corresponding to each review in the

In [56]:
```
def rating(row):
    return '3 stars' if row.points >= 95 or row.country == 'Canada' else ('2 s

star_ratings = wine_reviews.apply(rating, axis='columns')
star_ratings
```

Out[56]:
```
0        2 stars
1        2 stars
          ...   
65497    2 stars
65498    2 stars
Length: 65499, dtype: object
```

Grouping and Sorting

Groupwise analysis

In [57]:
```
wine_reviews.groupby('points').points.count()
```

Out[57]:
```
points
80     155
81     305
      ... 
99      15
100      8
Name: points, Length: 21, dtype: int64
```

In [58]:
```
# to get the cheapest wine in each point value category
wine_reviews.groupby('points').price.min()
```

Out[58]:
```
points
80       5.0
81       5.0
       ...  
99      75.0
100    150.0
Name: price, Length: 21, dtype: float64
```

In [59]:
```
# here's one way of selecting the name of the first wine reviewed from each wi
wine_reviews.groupby('winery').apply(lambda df: df.title.iloc[0])
```

Out[59]:
```
winery
1+1=3                          1+1=3 NV Rosé Sparkling (Cava)
10 Knots                 10 Knots 2010 Viognier (Paso Robles)
                                     ...                    
àMaurice    àMaurice 2013 Fred Estate Syrah (Walla Walla V...
Štoka                         Štoka 2009 Izbrani Teran (Kras)
Length: 13549, dtype: object
```

In [60]:
```
# how we would pick out the best wine by country and province
wine_reviews.groupby(['country', 'province']).apply(lambda df: df.loc[df.point
```

Out[60]:

		country	description	designation	points	price	province	region_1	regic
country	**province**								
Argentina	**Mendoza Province**	Argentina	If you love massive Argentine reds with purity...	Finca Pedregal Single Vineyard Barrancas Maipú...	95	74.0	Mendoza Province	Mendoza	
	Other	Argentina	This single-vineyard Malbec blend from vineyar...	Chañar Punco	94	68.0	Other	Calchaquí Valley	
...	**...**	...	...	...	...	...	...	...	
Uruguay	**San Jose**	Uruguay	Baked, sweet, heavy aromas turn earthy with ti...	El Preciado Gran Reserva	87	50.0	San Jose	NaN	
	Uruguay	Uruguay	Cherry and berry aromas are ripe, healthy and ...	Blend 002 Limited Edition	91	22.0	Uruguay	NaN	

385 rows × 15 columns

In [61]:
```
# we can generate a simple statistical summary of the dataset
wine_reviews.groupby(['country']).price.agg([len, min, max])
```

Out[61]:

	len	min	max
country			
Argentina	1907	4.0	230.0
Armenia	1	14.0	14.0
...	...	...	...
Ukraine	5	6.0	10.0
Uruguay	61	10.0	120.0

41 rows × 3 columns

Multi-indexes

In [62]:

```
countries_reviewed = wine_reviews.groupby(['country', 'province']).description
countries_reviewed
```

Out[62]:

		len
country	**province**	
Argentina	**Mendoza Province**	1635
	Other	272
...	**...**	...
Uruguay	**San Jose**	3
	Uruguay	7

385 rows × 1 columns

In [63]:

```
mi = countries_reviewed.index
type(mi)
```

Out[63]:

```
pandas.core.indexes.multi.MultiIndex
```

In [64]:

```
countries_reviewed.reset_index()
```

Out[64]:

	country	province	len
0	Argentina	Mendoza Province	1635
1	Argentina	Other	272
...	...	...	...
383	Uruguay	San Jose	3
384	Uruguay	Uruguay	7

385 rows × 3 columns

Sorting

In [65]:

```
countries_reviewed = countries_reviewed.reset_index()
countries_reviewed.sort_values(by='len')
```

Out[65]:

	country	province	len
93	Croatia	Hrvatsko Primorje	1
291	Slovenia	Kras	1
...	...	...	...
375	US	Washington	4308
353	US	California	18122

385 rows × 3 columns

In [66]:

```
countries_reviewed.sort_values(by='len', ascending=False)
```

Out[66]:

	country	province	len
353	US	California	18122
375	US	Washington	4308
...	...	...	...
329	South Africa	Vlootenburg	1
139	Greece	Beotia	1

385 rows × 3 columns

In [67]:

```
countries_reviewed.sort_index()
```

Out[67]:

	country	province	len
0	Argentina	Mendoza Province	1635
1	Argentina	Other	272
...	...	...	...
383	Uruguay	San Jose	3
384	Uruguay	Uruguay	7

385 rows × 3 columns

In [68]:

```python
countries_reviewed.sort_values(by=['country', 'len'])
```

Out[68]:

	country	province	len
1	Argentina	Other	272
0	Argentina	Mendoza Province	1635
...	...	...	...
381	Uruguay	Montevideo	10
379	Uruguay	Canelones	24

385 rows × 3 columns

Create a Series whose index is the taster_twitter_handle category from the dataset, and whose values count how many reviews each person wrote.

In [69]:

```python
reviews_written = wine_reviews.groupby('taster_twitter_handle').size()

# or

# reviews_written = wine_reviews.groupby('taster_twitter_handle').taster_twitt
```

What are the minimum and maximum prices for each variety of wine? Create a DataFrame whose index is the variety category from the dataset and whose values are the min and max values thereof.

In [70]:

```python
price_extremes = wine_reviews.groupby(['variety']).price.agg([min, max])
price_extremes
```

Out[70]:

	min	max
variety		
Abouriou	75.0	75.0
Agiorgitiko	10.0	66.0
...	...	...
Çalkarası	19.0	19.0
Žilavka	15.0	15.0

590 rows × 2 columns

What are the most expensive wine varieties? Create a variable sorted_varieties containing a copy of the dataframe from the previous question where varieties are sorted in descending order based on minimum price, then on maximum price (to break ties).

In [71]:
```
sorted_varieties = price_extremes.sort_values(by=['min', 'max'], ascending = F
sorted_varieties
```

Out[71]:

variety	min	max
Terrantez	236.0	236.0
Bual	194.0	230.0
...	...	...
Tempranillo-Malbec	NaN	NaN
Zelen	NaN	NaN

590 rows × 2 columns

What combination of countries and varieties are most common? Create a Series whose index is a MultiIndexof {country, variety} pairs. For example, a pinot noir produced in the US should map to {"US", "Pinot Noir"}. Sort the values in the Series in descending order based on wine count.

In [72]:
```
country_variety_counts = wine_reviews.groupby(['country', 'variety']).variety.
```

Data Types and Missing Values

Dtypes

In [73]:
```
wine_reviews.price.dtype
```

Out[73]:
```
dtype('float64')
```

In [74]:
```
wine_reviews.dtypes
```

Out[74]:
```
country            object
description        object
                    ...
critic             object
index_backwards     int64
Length: 15, dtype: object
```

In [75]:
```
wine_reviews.points.astype('float64')
```

Out[75]:
```
0        87.0
1        87.0
         ...
65497    90.0
65498    90.0
Name: points, Length: 65499, dtype: float64
```

```
In [76]: wine_reviews.index.dtype
```

```
Out[76]: dtype('int64')
```

Missing data

```
In [77]: wine_reviews[pd.isnull(wine_reviews.country)]
```

Out[77]:

	country	description	designation	points	price	province	region_1	region_2	taster_nam
913	NaN	Amber in color, this wine has aromas of peach ...	Asureti Valley	87	30.0	NaN	NaN	NaN	Mik DeSimor
3131	NaN	Soft, fruity and juicy, this is a pleasant, si...	Partager	83	NaN	NaN	NaN	NaN	Roger Vos
...	...	...	...	...	...	...	...	...	
59670	NaN	The heady florality of damask rose is joined b...	Steintal	92	38.0	NaN	NaN	NaN	Anr Krebiehl MV
60678	NaN	This wine was made for grilled meats, with its...	Dry	86	17.0	NaN	NaN	NaN	Susa Kostrzew

32 rows × 15 columns

```
In [78]: wine_reviews.region_2.fillna("Unknown")
```

```
Out[78]: 0        Unknown
         1        Unknown
                    ...
         65497       Napa
         65498    Unknown
         Name: region_2, Length: 65499, dtype: object
```

In [79]:
```
wine_reviews.taster_twitter_handle.replace("@kerinokeefe", "@kerino")
```

Out[79]:
```
0             @kerino
1          @vossroger
                ...
65497             NaN
65498     @wineschach
Name: taster_twitter_handle, Length: 65499, dtype: object
```

Create a Series from entries in the points column, but convert the entries to strings. Hint: strings are str in native Python.

In [80]:
```
point_strings = wine_reviews.points.astype('str')
```

Sometimes the price column is null. How many reviews in the dataset are missing a price?

In [81]:
```
missing_price_reviews = wine_reviews[wine_reviews.price.isnull()]
n_missing_prices = len(missing_price_reviews)
n_missing_prices

# Cute alternative solution: if we sum a boolean series, True is treated as 1
#n_missing_prices = reviews.price.isnull().sum()
# or equivalently:
#n_missing_prices = pd.isnull(reviews.price).sum()
```

Out[81]:
```
4670
```

What are the most common wine-producing regions? Create a Series counting the number of times each value occurs in the region_1 field. This field is often missing data, so replace missing values with Unknown. Sort in descending order.

In [82]:
```
reviews_per_region = wine_reviews.region_1.fillna('Unknown').value_counts().so
reviews_per_region
```

Out[82]:
```
region_1
Unknown                    10755
Napa Valley                 2226
                           ...
Civitella d'Agliano            1
Vino da Mesa de Toledo         1
Name: count, Length: 1112, dtype: int64
```

Renaming and Combining

Renaming

In [83]:

```
wine_reviews.rename(columns={'points': 'score'})
```

Out[83]:

	country	description	designation	score	price	province	region_1	region_2	taster_nam
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Ker O'Kee
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Vos
...	...	...	...	...	...	...	...	...	.
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	Na
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Micha Schachn

65499 rows × 15 columns

In [84]:

```
wine_reviews.rename(index = {0: 'firstEntry', 1: 'secondEntry'})
```

Out[84]:

	country	description	designation	points	price	province	region_1	region_2	tast
firstEntry	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	
secondEntry	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	R
...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	S

65499 rows × 15 columns

In [85]:
```
wine_reviews.rename_axis("wines", axis='rows').rename_axis("fields", axis='col
```

Out[85]:

fields	country	description	designation	points	price	province	region_1	region_2	taster_nar
wines									
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Ke O'Kee
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Vo
...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	Na
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Micha Schachr

65499 rows × 15 columns

Combining

In [86]:
```
canadian_youtube = pd.read_csv(r"D:\Data Analytics\Python\Kaggle Pandas\CAvide
british_youtube = pd.read_csv(r"D:\Data Analytics\Python\Kaggle Pandas\GBvideo

pd.concat([canadian_youtube, british_youtube])
```

Out[86]:

	video_id	trending_date	title	channel_title	category_id	publish_time	
0	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	E
1	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	
...	...	...	...	...	...	...	
38914	-DRsfNObKIQ	18.14.06	Eleni Foureira - Fuego - Cyprus - LIVE - First...	Eurovision Song Contest	24	2018-05-08T20:32:32.000Z	E
38915	4YFo4bdMO8Q	18.14.06	KYLE - Ikuyo feat. 2 Chainz & Sophia Black [A...	SuperDuperKyle	10	2018-05-11T04:06:35.000Z	K

49508 rows × 16 columns

In [93]:
```
left = canadian_youtube.set_index(['title', 'trending_date'])
right = british_youtube.set_index(['title', 'trending_date'])
#left.join(right, lsuffix='_CAN', rsuffix='_UK')
#left.join(right)

# The lsuffix and rsuffix parameters are necessary here because the data has t
```

region_1 and region_2 are pretty uninformative names for locale columns in the dataset. Create a copy of reviews with these columns renamed to region and locale, respectively.

In [91]:
```
renamed = wine_reviews.rename(columns=dict(region_1='region', region_2='locale
renamed
```

Out[91]:

	country	description	designation	points	price	province	region	locale	taster_name	t
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O’Keefe	
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	
...	...	...	...	...	...	...	...	...	...	
65497	US	This wine wears its 15.8% alcohol better than ...	Block 24	90	31.0	California	Napa Valley	Napa	NaN	
65498	Spain	A unique take on Manzanilla Sherry, which is o...	Manzanilla	90	10.0	Andalucia	Jerez	NaN	Michael Schachner	

65499 rows × 15 columns

In [ ]: