COMP4220: Machine Learning, Spring 2021, Assignment 2

Due: Wednsday, Feb 24, 11pm

Please submit one pdf file for all questions.

You can type your answer for the first two questions in the below cell of each question using "Markdown" option!

**When turning in assignments after the due date, please clearly specify the number of late hours used.

Import libraries

In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

pandas programming assignment

P1. Visualization

p1.1 stack two series horizontally and show the result.

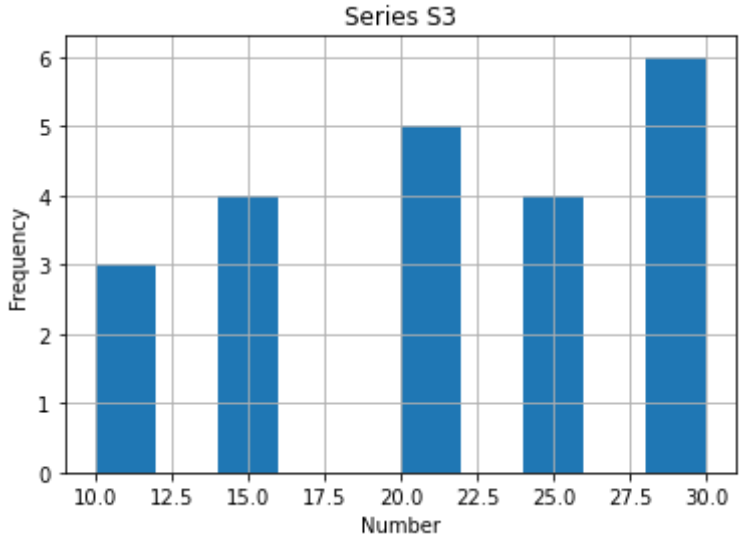
In [5]:
```
# Horizontal
s1 = pd.Series(range(0,30,2))
s2 = pd.Series(list('MACHINELEARNING'))
df1 = pd.concat([s1, s2], axis = 1)
print(df1)
```

```
     0  1
0    0  M
1    2  A
2    4  C
3    6  H
4    8  I
5   10  N
6   12  E
7   14  L
8   16  E
9   18  A
10  20  R
11  22  N
12  24  I
13  26  N
14  28  G
```

P1.2 show the histogram representing of following series.

In [6]:
```
s3 = pd.Series([10, 15, 20, 25, 30,10, 15, 20, 25,30,10, 15, 20, 25,30,
30,30,30,20,20,25,15])
s3.hist()
plt.ylabel('Frequency')
plt.xlabel('Number')
plt.title('Series S3')
```

Out[6]:
```
Text(0.5, 1.0, 'Series S3')
```



P1.3. show the boxplot representing of following dataframe.

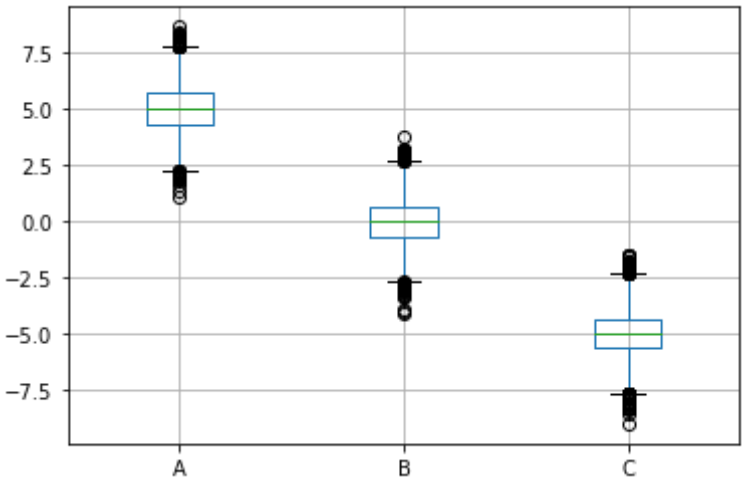
In [7]:
```
# Boxplot can be drawn by calling Series.plot.box() and DataFrame.plot.box(),
# or DataFrame.boxplot() to visualize the distribution of values within each column.



df2=pd.DataFrame(
    {"A": np.random.randn(10000)+5,"B": np.random.randn(10000),"C": np.random.randn(10000)-5},
    columns=["A", "B", "C"])

df2.boxplot()
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2672bf90550>
```



P2. Working with winedataset

In [9]:
```
# Load dataset
reviews = pd.read_csv("winemag-data-130k-v2.csv", index_col=0)
reviews.head()
```

Out[9]:

	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

	country	description	designation	points	price	province	region_1	region_2	taster_name
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

P2.1 Select the description column from reviews and assign the result to the variable desc.

In [10]:
```
desc = reviews['description']
desc
```

Out[10]:
```
0         Aromas include tropical fruit, broom, brimston...
1         This is ripe and fruity, a wine that is smooth...
2         Tart and snappy, the flavors of lime flesh and...
3         Pineapple rind, lemon pith and orange blossom ...
4         Much like the regular bottling from 2012, this...
                                ...
129966    Notes of honeysuckle and cantaloupe sweeten th...
129967    Citation is given as much as a decade of bottl...
129968    Well-drained gravel soil gives this wine its c...
129969    A dry style of Pinot Gris, this is crisp with ...
```

```
129970    Big, rich and off-dry, this is powered by inte...
Name: description, Length: 129971, dtype: object
```

P2.2 Select the first value from the description column of reviews, assigning it to variable first_description.

In [11]:
```
first_description = reviews['description'][0]
first_description
```

Out[11]:
```
"Aromas include tropical fruit, broom, brimstone and dried herb. The pa
late isn't overly expressive, offering unripened apple, citrus and drie
d sage alongside brisk acidity."
```

P2.3 Select the first row of data (the first record) from reviews, assigning it to the variable first_row.

In [13]:
```
first_row = reviews.iloc[0]
first_row
```

Out[13]:
```
country                                                        It
aly
description              Aromas include tropical fruit, broom, brimsto
n...
designation                                                Vulkà Bia
nco
points
87
price
NaN
province                                                Sicily & Sardi
nia
```

```
region_1                                                               E
tna
region_2
NaN
taster_name                                                    Kerin O’Ke
efe
taster_twitter_handle                                           @kerinoke
efe
title                                      Nicosia 2013 Vulkà Bianco  (Et
na)
variety                                                         White Bl
end
winery                                                             Nico
sia
Name: 0, dtype: object
```

P2.4 Select the first 10 values from the description column in reviews, assigning the result to variable first_descriptions.

In [14]:
```
first_descriptions = reviews['description'].head(10)
first_descriptions
```

Out[14]:
```
0    Aromas include tropical fruit, broom, brimston...
1    This is ripe and fruity, a wine that is smooth...
2    Tart and snappy, the flavors of lime flesh and...
3    Pineapple rind, lemon pith and orange blossom ...
4    Much like the regular bottling from 2012, this...
5    Blackberry and raspberry aromas show a typical...
6    Here's a bright, informal red that opens with ...
7    This dry and restrained wine offers spice in p...
8    Savory dried thyme notes accent sunnier flavor...
9    This has great depth of flavor with its fresh ...
Name: description, dtype: object
```

P2.5 Select the records with index labels 1, 2, 3, 5, and 8, assigning the result to the variable sample_reviews.

In [15]:
```
sample_reviews = reviews.iloc[[1,2,3,5,8]]
sample_reviews
```

Out[15]:

	country	description	designation	points	price	province	region_1	region_2	taster_nar
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Vo
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Greg
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexan Peart
5	Spain	Blackberry and raspberry aromas show a typical...	Ars In Vitro	87	15.0	Northern Spain	Navarra	NaN	Mich Schach
8	Germany	Savory dried thyme notes accent sunnier flavor...	Shine	87	12.0	Rheinhessen	NaN	NaN	Anna Lee Iiji

P2.6 Create a variable df containing the country, province, region_1, and region_2 columns of the records with the index labels 0, 1, 10, and 100. Show the result.

In [16]:
```
df = reviews.loc[[0,1,10,100],['country','province','region_1','region_2']]
df
```

Out[16]:

	country	province	region_1	region_2
0	Italy	Sicily & Sardinia	Etna	NaN
1	Portugal	Douro	NaN	NaN
10	US	California	Napa Valley	Napa
100	US	New York	Finger Lakes	Finger Lakes

P2.7 Create a variable df containing the country and variety columns of the first 100 records. Show the result.

In [17]:
```
df = reviews.iloc[0:100][['country','variety']]
df
```

Out[17]:

	country	variety
0	Italy	White Blend
1	Portugal	Portuguese Red
2	US	Pinot Gris
3	US	Riesling

	country	variety
4	US	Pinot Noir
...	...	...
95	France	Gamay
96	France	Gamay
97	US	Riesling
98	Italy	Sangiovese
99	US	Bordeaux-style Red Blend

100 rows × 2 columns

P2.8 Create a DataFrame italian_wines containing reviews of wines made in Italy.

In [18]:
```
italian_wines = reviews[reviews['country']=='Italy']
italian_wines
```

Out[18]:

	country	description	designation	points	price	province	region_1	region_2	taster_na
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	K O'Ke
6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	K O'Ke

	country	description	designation	points	price	province	region_1	region_2	taster_na
13	Italy	This is dominated by oak and oak-driven aromas...	Rosso	87	NaN	Sicily & Sardinia	Etna	NaN	K O'Ke
22	Italy	Delicate aromas recall white flower and citrus...	Ficiligno	87	19.0	Sicily & Sardinia	Sicilia	NaN	K O'Ke
24	Italy	Aromas of prune, blackcurrant, toast and oak c...	Aynat	87	35.0	Sicily & Sardinia	Sicilia	NaN	K O'Ke
...	...	...	...	...	...	...	...	...	
129929	Italy	This luminous sparkler has a sweet, fruit-forw...	NaN	91	38.0	Veneto	Prosecco Superiore di Cartizze	NaN	N
129943	Italy	A blend of Nero d'Avola and Syrah, this convey...	Adènzia	90	29.0	Sicily & Sardinia	Sicilia	NaN	K O'Ke
129947	Italy	A blend of 65% Cabernet Sauvignon, 30% Merlot ...	Symposio	90	20.0	Sicily & Sardinia	Terre Siciliane	NaN	K O'Ke

	country	description	designation	points	price	province	region_1	region_2	taster_na
129961	Italy	Intense aromas of wild cherry, baking spice, t...	NaN	90	30.0	Sicily & Sardinia	Sicilia	NaN	K O'Ke
129962	Italy	Blackberry, cassis, grilled herb and toasted a...	Sàgana Tenuta San Giacomo	90	40.0	Sicily & Sardinia	Sicilia	NaN	K O'Ke

19540 rows × 13 columns

P2.9 Create a DataFrame top_oceania_wines containing all reviews with at least 95 points (out of 100) for wines from Australia or New Zealand.

In [19]:
```
top_oceania_wines = reviews[(reviews['points']>=95)&((reviews['country']=='Australia')|(reviews['country']=='New Zealand'))]
top_oceania_wines
```

Out[19]:

	country	description	designation	points	price	province	region_1	region_2	t
345	Australia	This wine contains some material over 100 year...	Rare	100	350.0	Victoria	Rutherglen	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
346	Australia	This deep brown wine smells like a damp, mossy...	Rare	98	350.0	Victoria	Rutherglen	NaN	
348	Australia	Deep mahogany. Dried fig and black tea on the ...	Grand	97	100.0	Victoria	Rutherglen	NaN	
349	Australia	RunRig is always complex, and the 2012 doesn't...	RunRig	97	225.0	South Australia	Barossa	NaN	
356	Australia	Dusty, firm, powerful: just a few apt descript...	Georgia's Paddock	95	85.0	Victoria	Heathcote	NaN	
360	Australia	Bacon and tapenade elements merge easily on th...	Descendant	95	125.0	South Australia	Barossa Valley	NaN	
365	Australia	The Taylor family selected Clare Valley for it...	St. Andrews Single Vineyard Release	95	60.0	South Australia	Clare Valley	NaN	
14354	Australia	This wine's concentrated dark fruit shows in t...	Old Vine	95	60.0	South Australia	Barossa Valley	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
16538	Australia	Rich, dense and intense, this is a big, muscul...	The Family Tree	95	65.0	South Australia	Barossa Valley	NaN	
28573	Australia	Astralis has become one of Australia's top col...	Astralis	95	350.0	South Australia	Clarendon	NaN	
34502	Australia	This prodigious wine showcases Barossa's abili...	The Relic	98	135.0	South Australia	Barossa Valley	NaN	
34506	Australia	If Standish's Relic is the feminine side of Sh...	The Standish Single Vineyard	96	135.0	South Australia	Barossa Valley	NaN	
38988	Australia	Penfolds Bin 707 has leapt in quality over the...	Bin 707	95	200.0	South Australia	South Australia	NaN	
39059	Australia	The Taylor family selected Clare Valley for it...	St. Andrews Single Vineyard Release	95	60.0	South Australia	Clare Valley	NaN	
39961	Australia	As unevolved as they are, the dense and multil...	Grange	96	185.0	South Australia	South Australia	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
39962	Australia	Seamless luxury from stem to stern, this ‘baby...	RWT	95	70.0	South Australia	Barossa Valley	NaN	
45809	Australia	The 2007 Astralis impresses for its combinatio...	Astralis	95	225.0	South Australia	Clarendon	NaN	
56953	Australia	This inky, embryonic wine deserves to be cella...	Grange	99	850.0	South Australia	South Australia	NaN	
56956	Australia	You may have to scour the country to secure so...	Andelmonde	97	95.0	South Australia	Barossa Valley	NaN	
56957	Australia	Thorn Clarke has taken its Shiraz to a new lev...	Ron Thorn Single Vineyard	96	89.0	South Australia	Barossa	NaN	
56959	Australia	Is this the Yin to Grange's Yang? The wines ar...	Hill of Grace	96	820.0	South Australia	Eden Valley	NaN	
59977	Australia	This is a top example of the classic Australia...	The Peake	96	150.0	South Australia	McLaren Vale	NaN	
59984	Australia	This is a throwback to those brash, flavor-exu...	One	95	95.0	South Australia	Langhorne Creek	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
67096	Australia	Just a tiny serving of this dark nectar will l...	Calliope Rare	98	86.0	Victoria	Rutherglen	NaN	
67101	Australia	This Muscat is the color of dark coffee, with ...	Rare	95	300.0	Victoria	Rutherglen	NaN	
76392	Australia	When the alcohol levels are reined in to appro...	Georgia's Paddock	95	85.0	Victoria	Heathcote	NaN	
77028	Australia	This has all the size and weight you've come t...	Grange	98	850.0	South Australia	South Australia	NaN	
77036	Australia	RWT (unromantically derived from “Red Wine Tri...	RWT	96	150.0	South Australia	Barossa Valley	NaN	
77037	Australia	Winemaker Dave Powell is no longer with Torbre...	RunRig	96	225.0	South Australia	Barossa Valley	NaN	
77042	Australia	This is likely the most ageworthy Shiraz winem...	Eligo	95	100.0	South Australia	Barossa	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
77044	Australia	The fruit for this offering comes from the Gre...	R Reserve	95	105.0	South Australia	Barossa Valley	NaN	
77046	Australia	With aromas and flavors that range widely from...	The Factor	95	125.0	South Australia	Barossa Valley	NaN	
83357	Australia	A throwback to the monster Shiraz style of old...	Grange	96	500.0	South Australia	South Australia	NaN	
84815	Australia	The Factor is always one of Torbreck's biggest...	The Factor	95	125.0	South Australia	Barossa	NaN	
84816	Australia	Nashwauk is Kaesler's McLaren Vale project, fi...	Beacon	95	145.0	South Australia	McLaren Vale	NaN	
87128	Australia	This full-bodied, muscular Shiraz is built for...	Amery Vineyard Block 6	96	120.0	South Australia	McLaren Vale	NaN	
87137	Australia	Perhaps the best young wine I've tasted from M...	NaN	95	84.0	Western Australia	Margaret River	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
87143	Australia	This is wonderfully complex and aromatic, with...	St. Andrews Single Vineyard Release	95	60.0	South Australia	Clare Valley	NaN	
91851	New Zealand	This full-bodied, richly tannic wine delivers....	Homage	95	100.0	Hawke's Bay	NaN	NaN	
98386	Australia	One of the more approachable of the d'Arenberg...	Little Venice Single Vineyard	95	85.0	South Australia	McLaren Vale	NaN	
99318	Australia	From vines planted in 1912, this has been an i...	Mount Edelstone Vineyard	95	200.0	South Australia	Eden Valley	NaN	
99330	Australia	This Cabernet equivalent to Grange has explode...	Bin 707	95	500.0	South Australia	South Australia	NaN	
99340	Australia	This rich, opulent wine carries its massive oa...	Les Amis	95	185.0	South Australia	Barossa Valley	NaN	
109427	Australia	This wine is dark brown in hue with a greenish...	Rare	99	300.0	Victoria	Rutherglen	NaN	

	country	description	designation	points	price	province	region_1	region_2	t
109434	Australia	D'Arenberg's lineup of single-vineyard Shiraze...	The Swinging Malaysian Single Vineyard	96	85.0	South Australia	McLaren Vale	NaN	
122421	Australia	Despite this wine's weight and richness, it re...	Amon-Ra Unfiltered	96	110.0	South Australia	Barossa Valley	NaN	
122430	Australia	These blends are traditional in Australia—they...	Anaperenna	95	80.0	South Australia	Barossa Valley	NaN	
122507	New Zealand	This blend of Cabernet Sauvignon (62.5%), Merl...	SQM Gimblett Gravels Cabernets/Merlot	95	79.0	Hawke's Bay	NaN	NaN	
122939	Australia	Full-bodied and plush yet vibrant and imbued w...	The Factor	98	125.0	South Australia	Barossa Valley	NaN	

P2.10 What is the data type of the points column in the dataset?

In [20]:
```
pointtype = reviews.points.dtype
pointtype
```

Out[20]:
```
dtype('int64')
```

P2.11 Create a Series from entries in the points column, but convert the entries to strings.

In [21]:
```
# Hint: Convert a column of one type to another by using the astype fun
ction.
point_strings = reviews.points.astype(str)
```

P2.12 Sometimes the price column is null. How many reviews in the dataset are missing a price?

In [22]:
```
number_missing_prices = reviews.price.isnull().sum()
number_missing_prices
```

Out[22]: 8996

P2.13 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. Your output should look something like this:

In [23]:
```
reviews_per_region = reviews.region_1.fillna('Unknown').value_counts().
sort_values(ascending=False)
reviews_per_region
```

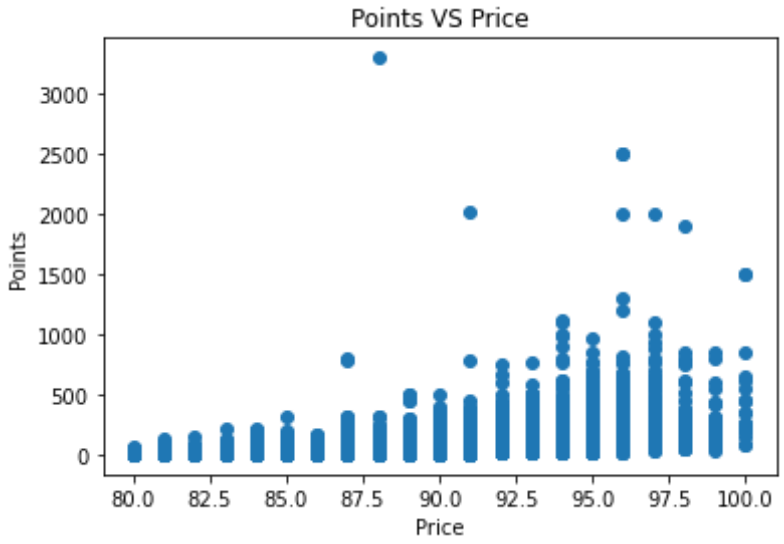
Out[23]:
```
Unknown                    21247
```

```
Napa Valley                          4480
Columbia Valley (WA)                 4124
Russian River Valley                 3091
California                           2629
                                     ...
Collioure                               1
Burgundy                                1
Paestum                                 1
Moscato di Pantelleria                  1
Sonoma-Santa Barbara-Mendocino          1
Name: region_1, Length: 1230, dtype: int64
```

P2.14 Scatter plots are commonly used to map the relationship between numerical variables. visualize the correlation between variables (points and price) using a scatter plot.

Add Title and Axis Labels to yout plot

In [70]:
```
plt.xlabel("Price")
plt.ylabel("Points")
plt.title('Points VS Price')
scatter_plot = plt.scatter(reviews.points, reviews.price)
```



In [ ]: