* Which of the following data would you use the analysis technique classification?
* **predicting the weather**
* predict the price of a stock
* simulating sales of a new product
* predicting the score on a test
* What are the ingredients to form a data science problem?
* **define what it is you're trying to tackle**
* **assess the situation with respect to the problem**
* **define your goals and objectives**
* What will be printed by the following code snippet?

x = [1, 2, 3]

y = x

x[1] = 42

print(y)

* **[1, 42, 3]**
* Given the following code:

arr = np.array([[1,2,3,4],[5,6,7,8],[9,10,11,12]])

Which of the 2 following commands produce the same result?

* **arr[0:1,1:3]**
* arr[2,1:3]
* arr[1:2,1:3]
* **arr[:1,1:3]**
* You are given the following lines of code:

arr = np.array([[1,2,3],[4,5,6],[7,8,9]])

slice = arr[:2,1:3]

slice[0,0]

What element in arr is equivalent to slice[0,0]?

* **arr[0,1]**
* What is the result of the following lines of code?

a=np.array(["cat","dog","fish"])

b=np.array(["dog","fish","rabbit"])

print(np.setdiff1d(b,a))

* ['cat']
* **['rabbit']**
* ['dog' 'fish']
* ['cat' 'dog' 'fish' 'rabbit']
* What is the output of the following broadcasting call?

A = np.array([1,2,3,4])

B = np.array([[1,2],[3,4]])

A + B

* array([[2, 4], [6, 8]])
* array([[2, 4, 3, 4], [4, 5, 3, 4]])
* array([[1, 2], [3, 4]])
* **Value Error**
* Which of the following statistics does the describe function show?
* **mean()**
* **min()**
* mode()
* **max()**
* **std()**
* What is the function call to find cells in a dataframe df with timestamp on 2007-02-04, given the dataframe has a parsed time column labelled ‘parsed\_time’?
* **df['parsed\_time'] == '2007-02-04'**
* df['parsed\_time'] = '2007-02-04'
* '2007-02-04' >= df['parsed\_time'] >= '2007-02-04'
* '2007-02-05' > df['parsed\_time'] > '2007-02-03'
* Suppose the pandas.core.series.Series ser has indices 'apple' and 'orange'. Which of the following is NOT a correct way to access data at a certain location in the ser Series?
* ser.loc['apple']
* **ser.loc['apple','orange']**
* ser.loc[['apple']]
* ser.loc[['apple','orange']]
* Suppose you have a pandas DataFrame called df which has a column called 'time'. Which function call will allow you to group the dataframe by 'time'?
* **df.groupby(['time'])**
* groupby(df['time'])
* df.aggregate(['time'])
* df['time'].group()
* Which is an example of conceptually driven data visualization?
* **Physicists use a visualization to teach students the well-known relationship between force and acceleration.**
* Doctors try to explore the relationship between a drug and the effect it has on their patients using data visualization.
* Realtors visualize a data set containing rental listings and the amount of interest they attract
* What are the qualities of good data visualization, according to Andy Kirk?
* **Trustworthy**
* **Accessible**
* **Elegant**
* Which graphing method should you use to visualize the correlation between two arrays?
* Histogram
* Barplot
* **Scatter plot**
* Line plot
* What is true between supervised and unsupervised approaches?
* In supervised, the target is unavailable. In unsupervised, the target is unavailable.
* In supervised, the target is provided. In unsupervised, the target is provided.
* In supervised, the target is unavailable. In unsupervised, the target is provided.
* **In supervised, the target is provided. In unsupervised, the target is unavailable.**
* In a decision tree, which nodes do NOT have test conditions?
* Root nodes
* Internal nodes
* **Leaf nodes**
* True or False?: Not specifying the parameter random\_state in the train\_test\_split function for every run will output the same result.
* True
* **False**
* When is a prediction task referred to as simple linear regression?
* **When there is only one input variable.**
* When there are two input variables.
* When there are more than two input variables.
* What are the 2 components of data retrieval mentioned in this class?
* The way you store specific data in a data management system.
* **The way you specify how to get the desired data out of the relational data store.**
* **The internal processing that occurs w/in a DBMS to compute/evaluate that a retrieval request**
* Assume the code: text = 'New York-based' What would be the output of text.split()?
* ['New', ' ', 'York-based']
* ['New', 'York', '-', 'based']
* ['New', 'York', 'based']
* **['New', 'York-based']**
* Given the following code

import json

status = ['like','excited','wow','dislike']

What is the correct way to use json to print the first three elements in the list status?

* print(json.dumps(status[4], indent = 1))
* print(json.dumps(status[3], indent = 1))
* print(json.dumps(status[1:4], indent = 1))
* **print(json.dumps(status[0:3], indent = 1))**
* If you find a peak distribution when plotting word frequency, what does this tell you about the vocabulary which produced that distribution?
* There are many unique words
* There is a large vocabulary
* **There is a focused topic**