Time left 0:59:46

What is the goal of exploratory data analysis?

Select one:

- a. Get a summary of the data, visualize and understand about the data
- b. Visualize and make the data clean
- c. Make the data clean, optimize a model, increase the predictiveness
- d. Understand about data and transform data into some forms

Time left 0:58:25

What conclusions can be drawn from a box plot in exploratory data analysis?

Select one or more:

- a. Is variability different between subgroups?
- b. Is the location concentration different between subgroups?
- c. Is there outliers?
- d. Is there any important feature (variable)?

Time left 0:58:15

Which method shows hierarchical data in a nested format?

Select one:

- a. Bar chart
- b. Treemap
- c. Population pyramid
- d. None of the other options

Time left 0:58:06

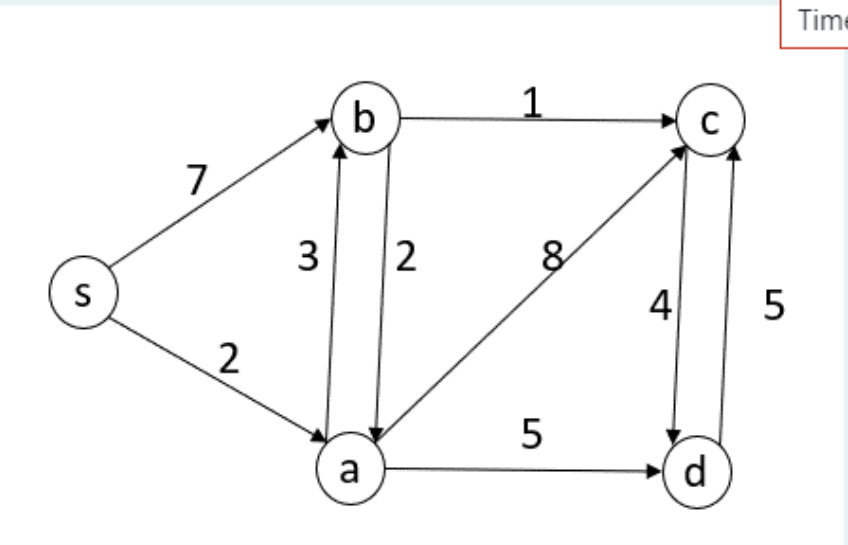
Is Hold-out a method for data preprocessing and understanding?

Select one:

- a. No, it is a method for training a model from a given dataset.
- b. No, it is a strategy for model assessment and selection.
- c. Yes, of course.

Using Dijkstra algorithm what is the length of shortest path from s to c?

Time left 0:57:48



Select one:

- a. 8
- b. 6
- c. There is no path from s to c
- d. 10

Time left 0:57:35

Does Scrapy natively support incremental crawling strategy?

Select one:

- a. Yes
- b. No

Time left 0:56:54

Can Google Openrefine import data on remote URL?

Select one:

- a. Yes
- b. No

Time left 0:56:31

Where is the difference between supervised learning and unsupervised learning?

Select one:

- a. From the type of the output which is often a real number in supervised learning
- b. From the aim of the algorithm, unsupervised learning often does not do prediction
- c. From the training data for which supervised learning often requires labels/responses for the training phase
- d. From the way we train a model, supervised learning means that we have to provide detailed steps for a machine to learn

Time left 0:56:12

What conclusions can be drawn from a histogram in exploratory data analysis?

Select one or more:

- [ ] a. Is the distribution of the data symmetric or skewed.
- [ ] b. The dispersion of the data.
- [ ] c. The distribution of the set of observations.
- [ ] d. The data centralization.
- [ ] e. Is there outliers in the data?

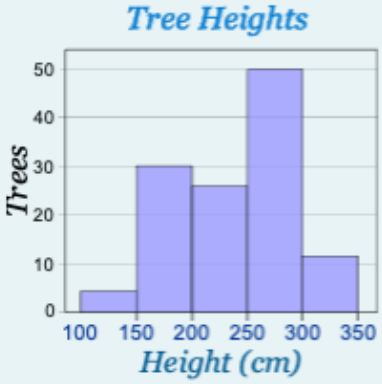
Time left 0:56:01

Which of the following accurately describes XPath?

Select one:

- a. XPath is the same as an XML file.
- b. XPath is a query language.
- c. XPath is a programming language.
- d. XPath can be read using a Word document.

Time left 0:55:47



Point out the correct statement.

Select one:

- a. Histogram of tree heights data
- b. A graph count tree of each height in data
- c. A column chart of tree heights data
- d. A graph count tree in data

Time left 0:55:38

Which of the following scenario may not be a good fit for HDFS?

Select one or more:

- [ ] a. Storing enormous small files.
- [ ] b. Storing data related to applications requiring low latency data access
- [ ] c. Scenarios requiring random writes to the same file
- [ ] d. None of the mentioned.

Time left 0:54:59

What information could you gain from a box-plot?

Select one or more:

- [ ] a. Skewness
- [ ] b. Probability distribution
- [ ] c. Lower/upper quartile
- [ ] d. Gap

Time left 0:54:45

What is not a problem of data quality at value level?

Select one:

- ○ a. Synonym
- ○ b. Missing value
- ○ c. Syntax violation

Time left 0:54:34

What is the main difference between Web-Scraper and Scrapy?

Select one:

- a. Scrapy is a library whereas Web-Scraper is stand-alone
- b. Scrapy relies on XPath, whereas Web-Scraper does not
- c. Scrapy is a library whereas Web-Scraper is a web-browser plugin
- d. Web-Scraper is more refined than Scrapy, because it relies on a selector hierarchy

Time left 0:54:17

You made a system to predict network attacks and you sure that it has a testing accuracy of 99%. However your boss says that your system is useless in practice. What may be the reasons?

Select one or more:

- [ ] a. Your boss does not have enough knowledge to understand your hard work and system.
- [ ] b. You are unlucky.
- [ ] c. The training set may be problematic.
- [ ] d. Accuracy may not reflect what your boss wants in this domain.
- [ ] e. Your evaluation of the system may be done incorrectly.

Time left 0:54:08

What is the role of a loss function?

Select one:

- a. To measure the loss/error when making future prediction
- b. To measure the error in some senses and to play as the objective function for learning from data
- c. No role in the data science process

Time left 0:53:58

Which is the most suitable statement about model selection?

Select one:

- a. The other statements are wrong.
- b. Model selection concerns on the best setting of the parameters for a model when learning from a training dataset. Sometimes it refers to selecting one from many models.
- c. Model selection only concerns on selection of the best one amongst different models when working with a given problem.

Time left 0:50:24

Overfitting may refer to the situation where

Select one:

- a. Too few training data for a machine to learn
- b. A method can predict inaccurately the behaviour of another method
- c. A method makes small error rate on the training data while having significantly larger error rate for future data
- d. Too many training data so that a machine can learn easily

Time left 0:49:21

Temperature is of which type?

Select one:

- a. Unordered continuous data
- b. Ordered discrete data
- c. Unordered discrete data
- d. Ordered continuous data

Time left 0:49:08

The three layers that make up the ..... architecture are the backend, the artist, and the scripting layers?

Select one:

- a. Matlab
- b. Matplotlib
- c. Pyplot
- d. Seaborn

Time left 0:48:59

Variety is a challenge related to big data, and refers to

Select one:

- a. The data that comes in continuously and fast
- b. The computation power that big data requires
- c. The data with high uncertainty due to the presence of fake/noisy information in some sources (particularly on the internet)
- d. The different kinds of data that must be handled: structured/unstructured data

Time left 0:48:51

What does Evaluation in the data science process include?

Select one:

- a. The evaluation of a system deployment in real life
- b. The analysis, assessment, comparison of the results from both offline and real-life scenarios if any

Time left 0:48:43

Given an uncompressed grayscale image of 256 levels, how many byte(s) per pixel does it need?

Select one:

- a. 1
- b. 3
- c. 24
- d. 8

Time left 0:47:46

What is the purpose of histogram equalization?

Select one:

- a. To reduce noise from images.
- b. To represent image content.
- c. To increase the brightness of an image.
- d. To enhance the contrast of an image.

Time left 0:47:40

Which Libs in Python should we use for exploratory data analysis?

Select one or more:

- a. SciPy and Numpy
- b. NLTK, Spacy
- c. Requests, Scrapy, BeautifulSoup
- d. Tensorflow, Keras, Scikit-learn
- e. Pandas
- f. Matplotlib

Time left 0:47:01

Is Business understanding a crucial step in the product-driven data science process?

Select one:

- a. No, it does not relate to Data Science
- b. No, we can ignore that step
- c. Yes, of course

Time left 0:46:52

Assume that you train a classifier on 10,000 training points and obtain a training accuracy of 99%. However, when you submit it to Kaggle, your accuracy is only 67%. Which of the following has a good chance of improving your performance on Kaggle?

Select one or more:

- a. Train on more data.
- b. Set your regularization coefficient (if any) to 0.
- c. Use a validation set to tune your hyperparameters.
- d. Remove randomly some parts of the training set when training your classifier.

Time left 0:46:42

Which kind of chart will be created with the following code?

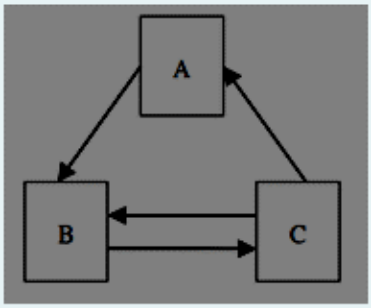
```
question.plot(kind='barh')
```

Select one:

- a. Bar Graph
- b. None of the other options
- c. Line graph
- d. Column Graph

Time left 0:46:29

Calculate Pagerank of A with damping factor d = 0.7



Select one:

- a. 0.3753
- b. 0.2245
- c. 0.3933
- d. 0.2314

Time left 0:42:45

Learning a decision tree by the ID3 algorithm will stop if

Select one:

- a. The tree is big enough
- b. The tree cannot classify correctly all the training data
- c. The tree classifies correctly all the training data
- d. The tree classifies correctly all the training data, or at any path all the attributes are used

Time left 0:42:36

In Scrapy way, how to store crawled data into databases?

Select one:

- a. Write a hook into item pipelines
- b. Write a hook into downloader
- c. Write a hook into spider middleware

Time left 0:42:28

..... function is responsible for consolidating the results produced by each of the Map() functions/tasks.

Select one:

- a. Map
- b. All of the mentioned
- c. Reducer
- d. Reduce

Time left 0:42:19

Point out the correct statement:

Select one:

- a. Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data.
- b. Hbase is a not relational database but it supports SQL.
- c. Pig is a relational database with SQL support.
- d. All of the mentioned.

Time left 0:42:11

Velocity is a challenge of the era of big data, and refers to

Select one:

- a. The speed of analysis
- b. The data that vary heavily
- c. The computation it requires massively
- d. The data that come continuously and fast

Time left 0:42:01

Can robots.txt practically stop unwanted web crawlers?

Select one:

- a. Yes
- b. No

Time left 0:40:35

What is not a cause of noises in data?

Select one:

- a. Different considerations between the time when the data was collected and when it is analyzed
- b. Faulty data collection instruments
- c. Human error at data entry

Time left 0:40:09

Why data in real world is dirty?

Select one or more:

- [ ] a. Incompete
- [ ] b. Integrated
- [ ] c. Noisy
- [ ] d. Inconsistent

Time left 0:39:57

Which statement is the most closely related to "The curse of dimensionality"?

Select one:

- a. The high dimensionality may pose difficulties for storage and computation
- b. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance.
- c. When the dimensionality increases, the difficulty of data analysis may not be affected significantly

Time left 0:39:47

What is the most famous algorithm to rank web pages in the search engine results?

Select one:

- a. Textrank
- b. Webrank
- c. Pagerank