**1. What does the term "in situ" mean in the context of big data?**  
A. Accelerometers.  
B. The sensors used in airplanes to measure altitude.  
C. Bringing the computation to the location of the data.  
D. In the situation.

**2. Which of the followings procedure best generalizes big data procedures such as (but not limited to) the map reduce process?**  
A. split->sort->merge  
B. split->do->merge  
C. split->map->shuffle and sort->reduce  
D. split->shuffle and sort->map->reduce

**3. What is the purpose of GROUP BY queries?**  
A. Enables calculations based on specific columns of the table.  
B. Enables queries within queries.  
C. Required before you can use functions like AVG, SUM, MIN, MAX, COUNT.

**6. Amazon has been collecting review data for a particular product. They have realized that almost 90% of the reviews were mostly a 5/5 rating. However, of the 90%, they realized that 50% of them were customers who did not have proof of purchase or customers who did not post serious reviews about the product. Of the following, which is true about the review data collected in this situation?**  
A. High Valence  
B. Low Valence  
C. Low Volume  
D. High Veracity  
E. High Volume  
F. Low Veracity

**7. What are the challenges of data with high variety?**  
A. Hard to perform emergent behavior analysis.  
B. The quality of data is low.  
C. Hard in utilizing group event detection.  
D. Hard to integrate.

**9. What does the 1 represent in the following line of code? ssc = StreamingContext(sc, 1)**  
A. To create only one partition to manage the stream.  
B. To specific debug output.  
C. To create one single context.  
D. A batch interval of 1 second.

**11. What kind of centrality identifies "hubness"?**  
A. Between-ness  
B. Closeness  
C. Degree  
D. Group

**12. What is the order of the three steps to Map Reduce?**  
A. Map -> Reduce -> Shuffle and Sort  
B. Shuffle and Sort -> Reduce -> Map  
C. Map -> Shuffle and Sort -> Reduce  
D. Shuffle and Sort -> Map -> Reduce

**13. What are the layers (parts) of Spark?**  
A. SparkSQL, Graphx, MLlib, Spark Core, Spark Streaming  
B. SparkSQL, Graphx, Spark Graph, Spark Core, Spark RDD  
C. SparkSQL, Spark Graph, Spark Core, Spark RDD, Spark Streaming  
D. SparkSQL, Graphx, MLlib, Spark Graph, Spark RDD, Spark Streaming

**14. What database would be best suited for the following scenario: An app development company is trying to implement a cloud based storage system for their new map-based app. The ud will manage the longitude and latitude of the data in order to track user location?**  
A. Solr  
B. Vertica  
C. Aerospike  
D. Redis

**15. The key reason mentioned in the video that biology applications need Big Data analytics is...**  
A. The integration of multiple data sources from different researchers and of different sources of information.  
B. The complexity of interactions that correlate to inform phenotypes.  
C. The new use of computational techniques to explore new areas of biology research more quickly than can be done with "live" or wetlab experiments.

**16. KNIME: In the confusion matrix, what is the difference between low\_humidity\_day and Prediction(low\_humidity\_day)?**  
A. low\_humidity\_day is the target class label, and Prediction(low\_humidity\_day) is the predicted class label.  
B. low\_humidity\_day is the predicted class label, and Prediction(low\_humidity\_day) is the target class label.  
C. There is no difference. The two are the same.

**19. Imputing missing data means to:**  
A. replace missing values with something reasonable.  
B. drop samples with missing values.  
C. replace missing values with outliers.  
D. merge samples with missing values.

**23. A team is working on a problem where the requirements are well known but they are short on time. They prefer to start the design phase when some of the requirements have been defined. Similarly, they want to start coding when the design for part of the system is ready. What model will best fit their needs?**  
A. Sashimi  
B. V-model  
C. Waterfall  
D. Spiral

**24. Which of the following are benefits for organization-generated data? (select all that apply):**  
A. Improved Safety  
B. Better Profit Margins  
C. Higher Sales  
D. High Velocity  
E. Customer Satisfaction

**25. Consider the following generic statement: db.<collection>.find(<query filter>, <projection>).<cursor modifier>. Which part of the statement would reflect that of the SELECT statement in SQL as illustrated in the lecture?**  
A. <query filter>  
B. <projection>  
C. <cursor modifier>  
D. <collection>

**26. Which of the following is NOT a category of machine learning?**  
A. Cluster Analysis  
B. Classification  
C. Regression  
D. Association Analysis  
E. Algorithm Prediction

**27. What is the purpose of the acronym BASE?**  
A. The same as ACID.  
B. To overcome CAP theorem.  
C. To impose properties on a BDMS in order to guarantee certain results.  
D. Enables stricter enforcement of ACID type design.

**28. How does lazy evaluation work in Spark?**  
A. Transformations are queued and executed at a certain threshold.  
B. Transformations are not executed until the action stage.  
C. Actions are queued and executed at a certain threshold.  
D. Actions are not executed until the transformation stage.

**29. What is K-means clustering?**  
A. Divide samples using k lines.  
B. Classify data by k decisions.  
C. Group samples into k clusters.  
D. Classify data by k actions.

**30. What is an example of open-source tools built for Hadoop and what does it do? Nếu có graph và scale thì quất luôn**   
A. Giraph, for SQL-like queries.  
B. Pig, for real-time and in-memory processing of big data.  
C. Zookeeper, analyze social graphs.  
D. Zookeeper, management system for animal-named related components.

**31. What does the following do? df.select("userid", "teamlevel").show(5)**  
A. Select the rows named "userid" and "teamlevel" and display first 5 rows.  
B. Display all rows except "userid" and "teamlevel".  
C. Select the columns named "userid" and "teamlevel" and display first 5 rows.  
D. Display all columns except "userid" and "teamlevel".

**32. The Cypher query language shares some commands in common with SQL. True or False?**  
A. True  
B. False

**33. This line of code creates a k-means model with 12 clusters: kmeans = KMeans(k=12, seed=1). What is the significance of "seed=1"?**  
A. This sets the seed to a specific value, which is necessary to reproduce the k-means results.  
B. This means that this is the first iteration of k-means. The seed value is incremented by 1 every time k-means is executed.  
C. This specifies that the first cluster centroid is set to sample #1.

**34. KNIME: In configuring the Numeric Binner node, what would happen if the definition for the humidity\_low bin is changed from ]-infinity... 25.0 [ to ]-infinity ... 25.0 ] (i.e., the last bracket is changed from [to]?**  
A. The definition for the humidity\_low bin would change from excluding 25.0 to including 25.0.  
B. The definition for the humidity\_low bin would change from having 25.0 as the endpoint to having 25.1 as the endpoint.  
C. Nothing would change.

**35. Of the options below, which query allows you to find the state with the most counties?**  
A. source="census.csv" | stats count by CENSUS2010POP | sort count  
B. stats count by STNAME | sort -count  
C. source="census.csv" | stats count by CTYNAME | sort num(count)  
D. source="census.csv" | stats count by STNAME | sort count desc

**36. Why is it necessary to scale the data (Step 4)?**  
A. Since the values of the features are on different scales, all features need to be scaled so that all values will be positive.  
B. Since the values of the features are on different scales, all features need to be scaled so that no one feature dominates the clustering results.  
C. Since the values of the features are on different scales, all features need to be scaled so that the cluster centers can be displayed on the same plot for easier analysis.

**37. KNIME: In configuring the Numeric Binner node, what would happen if the "Append new column" box is not checked?**  
A. The relative\_humidity\_3pm variable will become a categorical variable.  
B. The relative\_humidity\_3pm variable will remain unchanged, and a new unnamed categorical variable will be created.  
C. The relative\_humidity\_3pm variable will become undefined, and an error will occur.

**38. Where does the data for each worker node get sent to after a collect function is called?**  
A. Other Worker Nodes  
B. Spark Streaming  
C. Spark Context  
D. None; Stays in the Same Node  
E. Spark SQL

**39. What are the steps required for data analysis?**  
A. Investigate, Build Model, Evaluate  
B. Classification, Regression, Analysis  
C. Regression, Evaluate, Classification  
D. Select Technique, Build Model, Evaluate

**40. Which of these statements is true about samples and variables? Sample là instance ,have many var,var descry character**  
A. A sample is an instance or example of an entity in your data.  
B. A sample can have many variables to describe it.  
C. A variable describes a specific characteristic of an entity in your data.  
D. All of these statements are true.

**41. What are the two main categories of techniques for exploring data? Choose two.**  
A. Histogram  
B. Outliers  
C. Visualization  
D. Trends  
E. Correlations  
F. Summary statistics

**42. What does the following line of code do in postgres?**  
SELECT count(userid) FROM (SELECT buyclicks.userid, teamLevel, price FROM buyclicks JOIN gameclicks on buyclicks.userid = gameclicks.userid) temp WHERE price=3 and teamLevel=5;  
A. Displays the users who have bought items worth $3 and have had a team with level 5.  
B. This is an invalid line of code, the subquery is not formatted properly.  
C. Counts the users who exist between both gameclicks and buyclicks files.  
D. Finds the total number of user ids (repeats allowed) in buy-clicks that have bought items with prices worth $3 and was in a team with level 5 at some point in time.

**43. Suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number. Which constraint should we have placed on the Account Number field for the end of year collection?**  
A. Account should have at most n digits.  
B. If we had n duplicate Account Numbers then we will remove n-1 duplicate fields.  
C. There are no constraints.  
D. Account Number should be unique.

**44. Which Cypher command launches a Neo4j database search?**  
A. MATCH  
B. RETURN  
C. CREATE  
D. None of the others

**45. Which of the following is NOT an example of feature selection?**  
A. Adding an in-state feature based on an applicant's home state.  
B. Re-formatting an address field into separate street address, city, state, and zip code fields.  
C. Removing a feature with a lot of missing values.  
D. Replacing a missing value with the variable mean.

**46. This question is about "best paths". To find the most discussed email in an email network, would we be looking to minimize a function or maximize a function?**  
A. Maximize  
B. Minimize

**47. What are ziplists in Redis?**  
A. A special type of data type that can store up to 512 mb of image data.  
B. A look-up table that is stored as a value in the database. Look-up table points to actual values in memory.  
C. A compressed list that is stored within the value of the database.  
D. A special type of data type that can store hashes that point to multiple attributes.

**48. Which of the following are parts of the 5 P's of data science and what is the additional P introduced?**  
A. People, Purpose, Product, Process, Programmability, Platforms người ,mục tiêu,ra sản phẩm theo quá trình và thành program- plat  
B. People, Purpose, Product, Perception, Process, Platforms  
C. People, Purpose, Product, Perception, Process  
D. People, Purpose, Process, Programmability, Platforms

**49. Which of the following is a TRUE statement about classification?**  
A. Classification is a supervised task.  
B. Classification is an unsupervised task.  
C. In a classification problem, the target variable has only two possible outcomes.

**50. Neo4j is a...**  
A. Graph database  
B. Relational database  
C. None of the others

**51. What reasoning was given for the following: why is the "data storage to price ratio" relevant to big data?**  
A. Companies can't afford to own, maintain, and spend the energy to support large data storage unless the cost is sufficiently low.  
B. Larger storage means easier accessibility to big data for every user because it allows users to download in bulk.  
C. It isn't, it was just an arbitrary example on big data usage.  
D. Access of larger storage becomes easier for everyone, which means client-facing services require very large data storage.

**52. What is data fusion?**  
A. Extracting a global value from a data source.  
B. Extracting true sources from a data source.  
C. Extracting the true value of a data item.  
D. Another term for customer analytics.

**53. What kind of centrality would you want to analyze in a graph if you wanted to inject information that flows through the shortest path in a network and have it spread quickly?**  
A. Degree  
B. Group  
C. Closeness  
D. Between-ness

**54. According to Ilkay, why is exploring data crucial to better modeling? Data exploration...**  
A. leads to data understanding which allows an informed analysis of the data.  
B. enables a description of data which allows visualization.  
C. enables understanding of general trends, correlations, and outliers.  
D. enables histograms and others graphs as data visualization.

**55. Which of the following are part of the four main categories to acquire, access, and retrieve data? (select all that apply):**  
A. NoSQL Storage  
B. Remote Data  
C. Traditional Databases  
D. Web Services  
E. Text Files

**56. Rule confidence is used to:**  
A. Identify frequent item sets  
B. Determine the rule with the most items  
C. Measure the intuitiveness of a rule  
D. Prune rules by eliminating rules with low confidence

**57. Which of the following is an example of big data utilized in action today?**  
A. The Internet  
B. Wi-Fi Networks  
C. Social Media  
D. Individual, Unconnected Hospital Databases

**58. Cypher includes a 'diameter' command to find the longest path in a graph network. True or False?**  
A. False  
B. True

**59. Which of the following is a Cypher command used to combine two or more query results?**  
A. union  
B. combine  
C. merge  
D. return

**60. Spark: If we split the data using 70% for training data and 30% for test data, how many samples would the training set have (using seed 13234)?**  
A. 730  
B. 334  
C. 70

**61. In the video on "Inclusion and Exclusion Constraints" we learn that adding constraints can actually make our analysis job easier. For example, when we require that a given node be included on a path, which of the following impacts now make the analysis job easier? (Choose 2)**  
A. Changing the weights on the edges of the graph and/or subgraphs  
B. Splitting the task into 2 independent shortest path problems  
C. Reduction of the size of the graph

**62. Why is data science mainly about teamwork?**  
A. Analytic solutions are required.  
B. Engineering solutions are preferred.  
C. Data science requires a variety of expertise in different fields.  
D. Exhibition of curiosity is required.

**63. DataFrame in Apache Spark prevails over RDD and does not contain any feature of RDD. True or False?**  
A. True  
B. False

**64. What does it mean to have \_id:0 within our query statement?**  
A. Grab the first object in the results.  
B. Grab as many objects as possible.  
C. Does not have an effect, simple convention left for compatibility issues.  
D. Tell MongoDB not to return a document id.

**65. The following query will return a graph containing whatever loops might exist. match (n)-[r]-(n) return n, r. True or False?**  
A. True  
B. False

**66. In linear regression, the least squares method is used to:**  
A. Determine the distance between two pairs of samples.  
B. Determine whether the target is categorical or numerical.  
C. Determine the regression line that best fits the samples.  
D. Determine how to partition the data into training and test sets.

**67. What is a correlated subquery?**  
A. A type of query that contains a subquery that requires information from a query one level up.  
B. A type of query that contains a relationship between a variable attribute x and a variable attribute y. The two variables have a dependent relationship causing a correlation.  
C. A type of query that requires two tables in order to calculate values.

**68. What does the 'k' stand for in k-nearest-neighbors?**  
A. the number of samples in the dataset  
B. the number of nearest neighbors to consider in classifying a sample  
C. the distance between neighbors: All neighboring samples that are 'k' distance apart from the sample are considered in classifying that sample.  
D. the number of training datasets

**69. What is streaming?**  
A. Calculating results using real-time data otherwise known as streaming data.  
B. Using static data stored from a real-time source in order to process and guide the application.  
C. Utilizing real-time data to compute and change the state of an application continuously.  
D. Using sensors to manipulate the system, such as a smart car being able to drive by itself using sensors to detect road hazards.

**70. When the following command is executed, where is the file written and how can it be accessed? counts.coalesce(1).saveAsTextFile('hdfs:/user/udera/wordcount/outputDir')**  
A. HDFS and through the system directory with the "cd" terminal command.  
B. HDFS and through the "hadoop fs" command.  
C. The local file system and through the "hadoop fs" command.  
D. The local file system and through the directory with the "cd" terminal command.

**71. What does the following line of code do? words = lines.flatMap(lambda line: line.split(" "))**  
A. Each line in the document is split up into words.  
B. Each line in the document is split into various Spark partitions.  
C. Each word in each line is counted.  
D. Each word is merged into lines to be counted later.

**72. Suppose we have an email network where the edges of a graph represent the number of emails from one user to another. If I was going to ask if Maria had sent any emails that (either directly or through forwarding from others) reached Julio, I would ask if:**  
A. Julio's node was reachable from Maria's node.  
B. Maria's node was reachable from Julio's node.

**73. What is the main difference between KNIME and Spark MLlib?**  
A. KNIME requires programming, while Spark MLlib does not.  
B. KNIME requires programming in Java, while Spark MLlib requires programming in Python.  
C. KNIME is a graphical user interface-based machine learning tool, while Spark MLlib provides a programming-based distributed platform for scalable machine learning algorithms.  
D. KNIME originated in Germany, while Spark MLlib was created in California, USA.

**74. Which of the Vs results in challenges due to graphs created from varying kinds, formats, sources, and meanings of data?**  
A. Variety  
B. Valence  
C. Volume  
D. Velocity

**75. Which of the Vs BEST describes the result in constant increasing in the number of edges in a graph, sometimes causing challenges in knowing when one has found "an answer" to one's analysis question?**  
A. Variety  
B. Volume  
C. Velocity  
D. Valence

**76. An influencer in a network is defined as:**  
A. a node which can reach all other nodes quickly  
B. the biggest gossip in the network  
C. a node which has heavy weight edges to at least 1/2 of the nodes in the network

**77. The feature independence assumption in Naive Bayes simplifies the classification problem by:**  
A. assuming that the prior probabilities of all classes are independent of one another.  
B. assuming that classes are independent of the input features.  
C. ignoring the prior probabilities altogether.  
D. allowing the probability of each feature given the class to be estimated individually.

**78. What is the function of Spark SQL? (Choose 3)**  
A. Efficient data manipulation using SQL-like structure.  
B. Enables relational queries on Spark.  
C. Deploy business intelligence tools over Spark.  
D. Connect to a variety of databases.  
E. Better ability to manipulate big data.  
F. Better worker node interpolation.

**79. Fill in the blank with the best answer: CAP theorem states that \_\_\_\_\_\_\_ all at once within a distributed computer system?**  
A. it is impossible to have consistency, accuracy, and partial tolerance  
B. it is necessary to have consistency, accuracy, and partial tolerance  
C. it is necessary to have consistency, availability, and partition tolerance  
D. it is impossible to have consistency, availability, and partition tolerance

**80. Which of the following are the 2 core "key player" problems that centrality analytics can address? (choose 2)**  
A. A set of nodes which can reach (almost) all other nodes  
B. What is the shortest path through a network  
C. Which nodes' removal will maximally disrupt gián đoạn the network  
D. Which nodes have the highest ratio of out-degree nodes to in-degree nodes

**81. What would be the two possible solutions associated with "big data" information integration? (Choose 2)**  
A. Probabilistic Schema Mapping  
B. Customer Transactions  
C. Pay-as-you-go Model  
D. Mediated Schema  
E. Attribute Grouping

**82. Which of the following are the three types of analytics questions asked about communities? (more than one)**  
A. Static  
B. Evolution  
C. Prediction  
D. Connection

**83. Which one of the following statements/situations/conversations align with an agile mindset?**  
A. A manager providing a status update to a customer says: "We are done with requirements and design so we are 50% done."  
B. Manager: "The customer is suggesting another change in feature X which is complete as per specification. They agree that it was built as we agreed upon but it lacks some functionality and will result in user frustration. Should we implement the change or not?" Developer: "Yes, if it does not fulfill user needs, we should make the change."  
C. A software vendor says to a client: "We must define exactly what the user needs and put that in agreement."  
D. A manager e-mails her team: "Meeting face to face just for a 15-minute discussion is inefficient, let us all just dial-in from our desk."

**84. In this code snippet below from the Hands On exercise on importing data, '100L + row...' adds 100 to the value of every country ID. Which of the following statements are true regarding this decision? (Note: you may select more than one)** (Code snippet provided)  
A. Another option would have been to add 100 to the metropolis keys as they were imported, and leave the country keys as they were originally numbered.  
B. This step was needed to create unique keys between the country and the metropolis datasets.  
C. Another option would be to add 500 to the country keys.

**85. Which of the following are general requirements for a programming language in order to support big data models? Select all that apply.**  
A. Utilize Map Reduction Methods  
B. Support Big Data Operations  
C. Enable Adding of More Racks  
D. Optimization of Specific Data Types  
E. Handle Fault Tolerance

**86. What database would be best suited for the following scenario: A big wholesale company is trying to implement a search engine for their products?**  
A. Redis  
B. Aerospike  
C. Solr  
D. Vertica

**87. What is in-memory processing?**  
A. Having the pipeline completely in disk.  
B. Writing data to disk between pipeline steps.  
C. Writing data to memory between pipeline steps.  
D. Having the pipeline completely in memory.  
E. Having the input completely in disk.  
F. Having the input completely in memory.

**88. What is the lambda architecture?**  
A. A type of hybrid data processing architecture.  
B. A type of architecture that only contains part of the data processing method.  
C. A type of swappable data processing layer.  
D. An architecture that natively supports lambda calculus.

**89. Suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number. Suppose 100 people signup for our system and of the 100 people, 60 of them did not input an address. The system lists the values as NULL for these empty entries in the address field. Would this situation still have structure for our data?**

**Ý là người dùng là 100 người nhưng có 60 không nhập address nên là hỏi là nếu với các giá trị null thì có cấu truc kg**  
A. No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.  
B. Yes the data has structure because we have placed a structural constraint on the data, thus the data will always have the originally defined structure.

**90. What is the most populated county in the state of Texas?**  
A. Harris  
B. Dallas  
C. Travis  
D. Bexar

**91. What are the challenges to data with a high valence?**  
A. Complex Data Exploration Algorithms  
B. Difficult to Integrate  
C. Reliability of Data

**92. What are the three computing steps of a semi-join?**  
A. Project, Ship, Reduce  
B. Project, Decompose, Send  
C. Index, Join, Display  
D. Query, Join, Display

**93. Which classification algorithm uses a probabilistic approach?**  
A. naive bayes  
B. none of the others  
C. decision tree  
D. k-nearest-neighbors

**94. What is schema-on-read?**  
A. The process where formatted data is given structure when read.  
B. Another name for data lakes.  
C. Data is stored as raw data until it is read by an application where the application assigns structure.  
D. The process where data is pre-formatted prior to being read but the schema is loaded on read.

**95. If I want to find the diameter of a graph, I should start by finding the shortest path between each set of nodes. True or False?**  
A. True  
B. False

**96. Which method is used to avoid overfitting in decision trees?**  
A. Post-pruning  
B. None of these  
C. Pre-pruning  
D. Pre-pruning and post-pruning

**97. What is a characteristic of streaming data?**  
A. Data is unbounded in size but requires only finite time and space to process it.  
B. The data is unbounded in size and the size determines the time and space of processing the data.  
C. The data is finite and requires only finite time and space to process the data.  
D. Data is finite in size and size determines the time and space of processing the data.

**98. What is a subquery?**  
A. A query statement within another query.  
B. A short query than normal.  
C. An alternative query that acts as a substitute for another query.

**99. Consider the following generic statement: db.<collection>.find(<query filter>, <projection>).<cursor modifier>. Which part of the statement would reflect that of the WHERE statement in SQL as illustrated in the lecture?**  
A. <projection>  
B. <cursor modifier>  
C. <query filter>  
D. <collection>

**100. A model that generalizes well means that:**  
A. The model is overfitting.  
B. The model does a good job of fitting to the noise in the data.  
C. The model performs well on data not used in training.  
D. The model performs well on data used to adjust its parameters.

**101. What is the difference between low-level interfaces and high-level interfaces?**  
A. Low-level deals with storage and scheduling while high-level deals with interactivity.  
B. Low-level deals with interactivity while high-level deals with storage and scheduling.

**102. Which of the followings is NOT an example of regression?**  
A. Predicting the price of a stock  
B. Estimating the amount of rain  
C. Determining whether power usage will rise or fall  
D. Predicting the demand for a product

**103. Which of the following is the best description of why it is important to learn about the foundations for big data?**  
A. Foundations is all that is required to show a mastery of big data concepts.  
B. Foundations help you revisit calculus concepts required in the understanding of big data.  
C. Foundations allow for the understanding of practical concepts in Hadoop.  
D. Foundations stand the test of time.

**104. Which of the following are problems to look out for when integrating your project with Hadoop? (select all that apply):**  
A. Random Data Access  
B. Infrastructure Replacement  
C. Data Level Parallelism  
D. Task Level Parallelism  
E. Advanced Algorithms

**105. Use the following table named "user\_table" to answer. How would you go about querying the entire username column (however many)?** (Table provided)  
A. SELECT user\_table FROM username  
B. SELECT username FROM user\_table  
C. SELECT username FROM user\_table WHERE userId=1  
D. SELECT username FROM userld WHERE \*

**106. KNIME: Considering the Numeric Binner node again, what would happen if the "Append new column" box is not checked?**  
A. The relative\_humidity\_3pm variable will become a categorical variable.  
B. The relative\_humidity\_3pm variable will remain unchanged, and a new unnamed categorical variable will be created.  
C. The relative\_humidity\_3pm variable will become undefined, and an error will occur.

**107. What does the following filter line of code do? df.filter(df["teamlevel"] > 1)**  
A. Filter each row to show only team levels larger than 1.  
B. Filter each column to show only team levels larger than 1.  
C. Select the first two columns of the data and filter each column to show only team levels larger than 1.  
D. Select the first two columns of the data and displays only team levels greater than 1.