1. **Which of the following is an example of big data utilized in action today?** D

   A. The Internet
   B. Individual, Unconnected Hospital Databases
   C. Wi-Fi Networks
   D. Social Media

2. **What reasoning was given for the following: why is the "data storage to price ratio" relevant to big data?** D

   A. Larger storage means easier accessibility to big data for every user because it allows users to download in bulk.
   B. It isn't, it was just an arbitrary example of big data usage.
   C. Companies can't afford to own, maintain, and spend the energy to support large data storage unless the cost is sufficiently low.
   D. Lower prices mean larger storage becomes easier to access for everyone, creating bigger amounts of data for client-facing services to work with.

3. **What is the best description of personalized marketing enabled by big data?** B

   A. Being able to obtain and use customer information for groups of consumers and utilize them for marketing needs.
   B. Being able to use personalized data from every single customer for personalized marketing needs.
   C. Marketing to each customer on an individual level and suiting to their needs.

4. **Of the following, which are some examples of personalized marketing related to big data?** C

   A. News outlets gathering information from the internet in order to report them to the public.

B. A survey that asks your age and markets to you a specific brand.
C. Facebook revealing posts that cater towards similar interests.

5. **What is the workflow for working with big data?** C

A. Theory -> Models -> Precise Advice
B. Extrapolation -> Understanding -> Reproducing
C. Big Data -> Better Models -> Higher Precision

6. **Which is the most compelling reason why mobile advertising is related to big data?** C

A. Mobile advertising in and of itself is always associated with big data.
B. Since almost everyone owns a cell/mobile phone, the mobile advertising market is large and thus requires big data to contain all the information.
C. Mobile advertising benefits from data integration with location which requires big data.
D. Mobile advertising allows massive cellular/mobile texting to a wide audience, thus providing large amounts of data.

7. **What are the three types of diverse data sources?** A

A. Machine Data, Organizational Data, and People
B. Machine Data, Map Data, and Social Media
C. Information Networks, Map Data, and People
D. Sensor Data, Organizational Data, and Social Media

8. **What is an example of machine data?** A

A. Weather station sensor output.
B. Social Media
C. Sorted data from Amazon regarding customer info.

9. **What is an example of organizational data?** A

A. Disease data from Center for Disease Control.
B. Satellite Data
C. Social Media

10. **Of the three data sources, which is the hardest to implement and streamline into a model?** B

A. Machine Data
B. People
C. Organizational Data

11. **Which of the following summarizes the process of using data streams?** A

A. Integration -> Personalization -> Precision
B. Big Data -> Better Models -> Higher Precision
C. Theory -> Models -> Precise Advice
D. Extrapolation -> Understanding -> Reproducing

12. **Where does the real value of big data often come from?** C

A. Size of the data.
B. Having data-enabled decisions and actions from the insights of new data.
C. Combining streams of data and analyzing them for new insights.
D. Using the three major data sources: Machines, People, and Organizations.

13. **What does it mean for a device to be "smart"?** B

A. Having a specific processing speed in order to keep up with the demands of data processing.
B. Connect with other devices and have knowledge of the environment.
C. Must have a way to interact with the user.

14. **What does the term "in situ" mean in the context of big data?** A

A. Bringing the computation to the location of the data.
B. In the situation
C. The sensors used in airplanes to measure altitude.
D. Accelerometers.

15. **Which of the following are reasons mentioned for why data generated by people are hard to process? Choose all that apply.** ABC

A. Very unstructured data.
B. Skilled people to analyze the data are hard to come by.
C. The velocity of the data is very high.
D. They cannot be modeled and stored.

16. **What is the purpose of retrieval and storage; pre-processing; and analysis in order to convert multiple data sources into valuable data?** D

A. Since the multi-layered process is built into the Neo4j database connection.
B. Designed to work like the ETL process.
C. To enable ETL methods.
D. To allow scalable analytical solutions to big data.

17. **Which of the following are benefits of organization-generated data? Choose all that apply.** ABDE

A. Better Profit Margins
B. Customer Satisfaction
C. High Velocity
D. Improved Safety
E. Higher Sales

18. **What are data silos and why are they bad?** A

A. Data produced from an organization that is spread out. Bad because it creates unsynchronized and in-

visible data.
B. A giant centralized database to house all the data production within an organization. Bad because it hinders opportunity for data generation.
C. Highly unstructured data. Bad because it does not provide meaningful results for organizations.
D. A giant centralized database to house all the data produces within an organization. Bad because it is hard to maintain as highly structured data.

19. **Which of the following are benefits of data integration? Choose all that apply.** ABCDF

A. Unify your data system.
B. Adds value to big data.
C. Reduce data complexity.
D. Increase data collaboration.
E. Monitoring of data.
F. Increase data availability.

20. **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?** B

A. Low Volume
B. Low Veracity
C. High Veracity
D. High Volume
E. Low Valence
F. High Valence

21. **As mentioned in the slides, what are the challenges to data with a high valence?** A

A. Complex Data Exploration Algorithms
B. Difficult to Integrate
C. Reliability of Data

22. **Which of the following are the 6 V's in big data?** BCDEFG

A. Vision
B. Valence
C. Variety
D. Veracity
E. Value
F. Velocity
G. Volume

23. **What is the veracity of big data?** A

A. The abnormality or uncertainties of data.
B. The connectedness of data.
C. The size of the data.
D. The speed at which data is produced.

24. **What are the challenges of data with high variety?** C

A. Hard in utilizing group event detection.
B. Hard to perform emergent behavior analysis.
C. Hard to integrate.
D. The quality of data is low.

25. **Which of the following is the best way to describe why it is crucial to process data in real-time?** B

A. More accurate.
B. Prevents missed opportunities.
C. More expensive to batch process.
D. Batch processing is an older method that is not as accurate as real-time processing.

26. **What are the challenges with big data that has high volume?** C

A. Storage and Accessibility
B. Effectiveness and Cost
C. Cost, Scalability, and Performance
D. Speed Increase in Processing

27. **Which of the following are parts of the 5 P's of data science and what is the additional P introduced in the slides?** BCDEFG

A. Perception
B. Programmability
C. Purpose
D. Product
E. Platforms
F. Process
G. People

28. **Which of the following are part of the four main categories to acquire, access, and retrieve data?** ACDE

A. Remote Data
B. Web Services
C. Traditional Databases
D. Text Files
E. NoSQL Storage

29. **What are the steps required for data analysis?** A

A. Select Technique, Build Model, Evaluate
B. Investigate, Build Model, Evaluate
C. Classification, Regression, Analysis
D. Regression, Evaluate, Classification

30. **Of the following, which is a technique mentioned in the videos for building a model?** A

A. Analysis
B. Investigation
C. Validation
D. Evaluation

31.	**What is the first step in finding a right problem to tackle in data science?** **A. Ask the Right Questions** **B. Define the Problem** **C. Define Goals** **D. Assess the Situation**	B
32.	**What is the first step in determining a big data strategy?** **A. Organizational Buy-In** **B. Build In-House Expertise** **C. Business Objectives** **D. Collect Data**	C
33.	**According to Ilkay, why is exploring data crucial to better modeling?** **Data exploration... <complete the sentence>** **A. leads to data understanding which allows an informed analysis of the data.** **B. enables histograms and others graphs as data visualization.** **C. enables understanding of general trends, correlations, and outliers.** **D. enables a description of data which allows visualization.**	A
34.	**Why is data science mainly about teamwork?** **A. Data science requires a variety of expertise in different fields.** **B. Exhibition of curiosity is required.** **C. Analytic solutions are required.** **D. Engineering solutions are preferred.**	A
35.	**What are the ways to address data quality issues?**	BCDE

A. Data Wrangling
B. Generate best estimates for invalid values.
C. Remove data with missing values.
D. Merge duplicate records.
E. Remove outliers.

36. **What is done to the data in the preparation stage?** B

A. Identify Data Sets and Query Data
B. Understand Nature of Data and Preliminary Analysis.
C. Retrieve Data
D. Build Models
E. Select Analytical Techniques

37. **Which of the following is the best description of why it is important to learn about the foundations for big data?** D

A. Foundations help you revisit calculus concepts required in the understanding of big data.
B. Foundations stand the test of time.
C. Foundations is all that is required to show a mastery of big data concepts.
D. Foundations allow for the understanding of practical concepts in Hadoop.

38. **What is the benefit of a commodity cluster?** BC

A. Much faster than a traditional super computer.
B. Cost Effective
C. Enables fault tolerance
D. Prevents network connection failure.
E. Prevents individual component failures.

39. **What is a way to enable fault tolerance?** BC

A. System Wide Restart
B. Redundant Data Storage
C. Data-Parallel Job Restart

D. Distributed Computing
E. Better LAN Connection

40. **What are the specific benefit(s) to a distributed file system?** ABD

A. High Fault Tolerance
B. High Concurrency
C. Large Storage
D. Data Scalability

41. **Which of the following are general requirements for a programming language in order to support big data models?** ACDE

A. Optimization of Specific Data Types
B. Utilize Map Reduction Methods
C. Enable Adding of More Racks
D. Handle Fault Tolerance
E. Support Big Data Operations

42. **What does IaaS provide?** A

A. Hardware Only
B. Computing Environment
C. Software On-Demand

43. **What does PaaS provide?** C

A. Hardware Only
B. Software On-Demand
C. Computing Environment

44. **What does SaaS provide?** B

A. Computing Environment
B. Software On-Demand
C. Hardware Only

45. B

What are the two key components of HDFS and what are they used for?

A. FASTA for genome sequence and Rasters for geospatial data.
B. NameNode for metadata and DataNode for block storage.
C. NameNode for block storage and Data Node for metadata.

46. **What is the job of the NameNode?** A

A. Coordinate operations and assigns tasks to Data Nodes
B. Listens from DataNode for block creation, deletion, and replication.
C. For gene sequencing calculations.

47. **What is the order of the three steps to Map Reduce?** D

A. Shuffle and Sort -> Map -> Reduce
B. Map -> Reduce -> Shuffle and Sort
C. Shuffle and Sort -> Reduce -> Map
D. Map -> Shuffle and Sort -> Reduce

48. **What is a benefit of using pre-built Hadoop images?** C

A. Guaranteed hardware support.
B. Less software choices to choose from.
C. Quick prototyping, deploying, and validating of projects.
D. Quick prototyping, deploying, and guaranteed bug free.

49. **What is an example of open-source tools built for Hadoop and what does it do?** BC

A. Zookeeper, analyze social graphs.
B. Zookeeper, management system for animal named related components.

C. Giraph, for processing large-scale graphs.
D. Giraph, for SQL-like queries.
E. Pig, for real-time and in-memory processing of big data.

50. **What is the difference between low level interfaces and high level interfaces?** B

A. Low level deals with interactivity while high level deals with storage and scheduling.
B. Low level deals with storage and scheduling while high level deals with interactivity.

51. **Which of the following are problems to look out for when integrating your project with Hadoop?** ABCE

A. Advanced Alogrithms
B. Random Data Access
C. Task Level Parallelism
D. Data Level Parallelism
E. Infrastructure Replacement

52. **As covered in the slides, which of the following are the major goals of Hadoop?** ABCEF

A. Facilitate a Shared Environment
B. Enable Scalability
C. Provide Value for Data
D. Latency Sensitive Tasks
E. Optimized for a Variety of Data Types
F. Handle Fault Tolerance

53. **What is the purpose of YARN?** A

A. Allows various applications to run on the same Hadoop cluster.
B. Enables large scale data across clusters.
C. Implementation of Map Reduce.

54. C

What are the two main components for a data computation framework that were described in the slides?

A. Node Manager and Applications Master
B. Applications Master and Container
C. Resource Manager and Node Manager
D. Resource Manager and Container
E. Node Manager and Container

55. **Download the text to Alice's Adventures in Wonderland from http://www.gutenberg.org/files/11/11-0.txt (If it redirects you to a page with a welcome popup, click on the "Plain Text UTF-8" option on that page or just download the attachment below) and run wordcount on it. This can be done by using hadoop commands. How many times does the word Cheshire occur? (Do not include the word 'Cheshire with an apostrophe. The string -->'Cheshire<-- does not count)**

Enter a number: 6

56. **The set of example MapReduce applications includes wordmedian, which computes the median length of words in a text file. If you run wordmedian using words.txt (the Shakespeare text) as input, what is the median word length?**

Note that wordmedian prints the median length to the terminal at the end of the MapReduce job; the output file does not contain the median length.

Enter a number: 4

57. **(Questions 1-3 pertain to the video lecture "Exploring the Relational Data Model of CSV")**
What is the approximate population of La Paz county in the state of Arizona for the CENSUS2010POP (column H)? (Choose the best answer.) A

A. 20000
B. 10000
C. 25000
D. 15000

58. **What county in the state of Wyoming has the smallest estimated population?** B

A. Uinta
B. Niobrara
C. Platte
D. Sweetwater

59. **At 2:45 of the video, the Instructor creates a filter for all of the counties in California with a population greater than 1,000,000. However, included in the results is the entire state of California. This anomalous value might skew our analysis if, for example, we wanted to compute the average population of these results. What additional filter might work to resolve this problem?** A

A. Add a filter to detect and remove results which do not include the word "County" in column G.
B. Add a filter which finds all counties with population greater than 100,000 AND less than 10,000,000 for column H (CENSUS2010POP).
C. Add a filter where the value in column E is greater than 1,000,000.
D. None of the above

60. **(Questions 4 and 5 pertain to the video "Exploring Sensor Data")**
How often (in seconds) do the R5 measurements occur? A

A. 60
B. 50

C. 40
D. 30

61. **What is the field for rain accumulation?** D

A. Sm
B. Dn
C. Dx
D. Rc

62. **(Questions 6 and 7 pertain to the video lecture "Exploring the Array Data Model of an Image") What is the (Red, Green, Blue) pixel value for location 500, 2000?** D

A. (134, 145, 46)
B. (50, 156, 182)
C. (100, 123, 149)
D. (163, 118, 79)

63. **Is this value likely to be land or ocean?** B

A. Ocean
B. Land

64. **(Questions 8 and 9 pertain to the video lecture "Exploring the Semistructured Data Model of JSON") Given a tweet, what path would you most likely enter to obtain a count of the number of followers for a user?** A

A. user/followers_count
B. user/statuses_count
C. user/listed_count
D. None of the above

65. **Which of the following fields are nested within the 'entities' field (select all that apply)?** ABC

A. user_mentions

B. urls
C. symbols
D. views
E. events
F. tweets

66. **What is a possible pitfall of utilizing Excel as a way to manipulate small databases?** A

A. Excel does not enforce many principles of relational data models.
B. Excel does not allow algorithms for data manipulation.
C. Excel is a user program and thus cannot run on a server.

67. **What does the term "atomic" mean in the context of relational databases?** B

A. A tuple that cannot be reduced.
B. One unit of information that cannot be decomposed.
C. A column or row of data. Depends on the context.
D. Fixed schema of a particular database.

68. **What is the Pareto-Optimality problem?** B

A. Find the shortest path from source node to target node.
B. Find the best possible path given two or more optimization criteria where neither constraint can be fully optimized simultaneously.
C. Find the optimal path that requires going through specific nodes given by the user.

69. **What constitutes a community within a graph?** B

A. Many anomalous neighborhoods within the same vicinity.
B. A dense amount of edge connections between

nodes in a community and a few connections across communities.
C. A neighborhood defined by an integer constant K around a specific node. All K+1 nodes belong in another community.
D. High density of nodes at a certain location.

70. **Why are trees useful for semi-structured data such as XML and JSON?** A

A. Trees take advantage of the parent-child relationship of the data for easy navigation.
B. Computers can easily visualize the data with a tree structure.
C. It is not always the case that XML and JSON can be represented as trees.
D. They are only useful for XML data as tree-like structure is apparent with tags. While JSON does not contain a tree-like structure as it contains arrays.

71. **What is the general purpose of modeling data as vectors?** A

A. Results can be ordered by similarity using vector projection.
B. Enables image searching.
C. Enables weighting of the query.
D. The ability to normalize vectors allowing probability distributions.

72. **For the following questions 7, 8, and 9, 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 we collect data month by month. Each month, we would have a batch of data containing the fields listed above. At the end of the year, we want to summarize our registrant activities for the entire year, A

so we would remove redundancies in our data by removing any records with duplicate account numbers from month to month. What type of operation do we use in this scenario?

A. Union
B. Join
C. Subsetting
D. Not an Operation

73. **From the information given in question 7, what are the constraints, if any, which we have placed on the Account Number field for the end of year collection?** D

A. Account should have at most n digits.
B. There are no constraints.
C. If we had n duplicate Account Numbers then we will remove n-1 duplicate fields.
D. Account Number should be unique.

74. **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?** A

A. 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.
B. No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.

75. **What is true between data modeling and the formatting of the data?** A

A. The data does not necessarily need to be formatted in a way that represents the data model. Just so long as it can be extrapolated.

B. There is always one specific schema for storing model data that is the best and preferred method for the specific data representation.
C. There is a one to one correspondence between formatting data and data modeling. For every model of data, there is only one way to store the data.

76. **What is streaming?** C

A. Using static data stored from a real time source in order to process and guide the application.
B. Calculating results using real time data otherwise known as streaming data.
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.

77. **Of the following, what best describes the properties of working with streaming data?** ABEF

A. Independent computations that do not rely on previous or future data.
B. Does not ping the source interactively for a response upon receiving the data.
C. Always unbounded in sequence, in other words, data is not guaranteed to be in order.
D. Data is always utilized for streaming the application.
E. Small time windows for working with data.
F. Data manipulation is near real time.

78. **What is a characteristic of streaming data?** D

A. The data is finite and requires only finite time and space to process the data.
B. The data is unbounded in size and the size determines the time and space of processing the data.

C. Data is finite in size and size determines the time and space of processing the data.
D. Data is unbounded in size but requires only finite time and space to process it.

79. **What type of algorithm is required for analyzing streaming data?** B

A. Accurate and Consistent
B. Fast and Simple
C. Fast and Complex
D. Accurate and Memory Efficient

80. **What is lambda architecture?** B

A. A specific method for processing streaming data using special real time processes.
B. A method to process streaming data by utilizing batch processing and real time processing.
C. A specific hardware architecture for a server made specifically for processing real time data.

81. **Of the following, which best represents the challenge regarding the size and frequency of data?** C

A. There may not be data to produce the notion of size and frequency.
B. The size and frequency of the streaming data may be too small.
C. The size and frequency of the streaming data may be sporadic.

82. **What is the difference between data lakes and data warehouses?** A

A. Data lakes house raw data while data warehouses contain pre-formatted data.
B. Data lakes utilize hierarchical systems while data warehouses use object storage.

C. Data lakes contain only files while data warehouses contain only databases.

83. **What is schema-on-read?** C

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.

84. **The desired characteristics of a BDMS include (select all that apply):** ACDE

A. A flexible semi-structured data model
B. Support for ACID
C. A full query language
D. Continuous data ingestion
E. Support for common "Big Data" data types
F. Narrow range of query sizes

85. **Fill in the blank with the best answer: CAP theorem states that _____ all at once within a distributed computer system?** D

A. it is necessary to have consistency, availability, and partition tolerance
B. it is necessary to have consistency, accuracy, and partial tolerance
C. it is impossible to have consistency, accuracy, and partial tolerance
D. it is impossible to have consistency, availability, and partition tolerance

86. **What is the purpose of the acronym BASE?** A

A. To impose properties on a BDMS in order to guarantee certain results.

B. Enables stricter enforcement of ACID type design.
C. The same as ACID.
D. To overcome CAP theorem.

87. **What are ziplists in Redis?** C

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

88. **What is one of the main features of Aerospike?** B

A. Enables real time data streaming from external sources.
B. Support for geospatial data storage and geospatial queries.
C. Better equipped for string based search applications.
D. Images as values within the database.

89. **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 cloud will manage the longitude and latitude of the data in order to track user location.** C

A. Solr
B. Redis
C. Aerospike
D. Vertica

90. **What database would be best suited for the following scenario: A big wholesale company is trying to imple-** C

ment a search engine for their products.

A. Redis
B. Aerospike
C. Solr
D. Vertica

91. **Which of the following data types are supported by Redis? (select all that apply)** ABDE

A. Strings
B. Lists
C. Images
D. Hashes
E. Sorted Sets
F. Streaming Video

92. **What does it mean for a query language to be declarative?** B

A. The language specifies both the process of how to obtain the data and specifies what data to obtain.
B. The language specifies what data to obtain.
C. The language specifies the process of how to obtain the data.
D. A language specific declaration of data types in order to define the method of data retrieval.

93. **Use the following table named "user_table" to answer the next 2 problems.** A

userId | username | email
1 | admin | admin@corporate.moe
2 | h4xor | 1337@rawr.cte

How would you go about querying the entire username column (however many)?

A. SELECT username FROM user_table

B. SELECT user_table FROM username
C. SELECT username FROM user_table WHERE userId=1
D. SELECT username FROM userId WHERE *

94. **How would you go about querying the entire database table (please refer to question 2's table)?** B

A. SELECT username, email FROM userId
B. SELECT * FROM user_table
C. SELECT user_table FROM *
D. SELECT * FROM * WHERE user_table

95. **What is the global indexing table?** C

A. A global table that uses a specific technique called indexing and the table uses an index as the primary key.
B. An index table in order to keep track of data records within one machine.
C. An index table in order to keep track of a given data type that might exist within multiple machines.
D. An index table in order to keep track of a given data type that might exist within one machine.

96. **What are the three computing steps of a semi-join?** A

A. Project, Ship, Reduce
B. Project, Decompose, Send
C. Index, Join, Display
D. Query, Join, Display
E. None Applicable

97. **What is the purpose of a semi-join?** A

A. Increase the efficiency of sending data across multiple machines.
B. Another name for join: an operation to combine two tables by column.

C. Increase the speed of the join for trade-off of increased data transmission cost.

98. **What is a subquery?** A

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. **What is a correlated subquery?** B

A. 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.
B. A type of query that contains a subquery that requires information from a query one level up.
C. A type of query that requires two tables in order to calculate values.

100. **What is the purpose of GROUP BY queries?** A

A. Enables calculations based on specific columns of the table.
B. Required before you can use functions like AVG, SUM, MIN, MAX, COUNT.
C. Enables queries within queries.

101. **Consider the following generic statement for questions 10-12:**
db.<collection>.find(<query filter>, <projection>).<cursor modifier>
Which part of the statement would reflect that of the FROM statement in SQL as illustrated in the lecture? D

A. <query filter>
B. <projection>
C. <cursor modifier>
D. <collection>

102. **Which part of the statement would reflect that of the SELECT statement in SQL as illustrated in the lecture?** D

A. <collection>
B. <query filter>
C. <cursor modifier>
D. <projection>

103. **Which part of the statement would reflect that of the WHERE statement in SQL as illustrated in the lecture?** B

A. <projection>
B. <query filter>
C. <cursor modifier>
D. <collection>

104. **A sample part of the data structure is as follows: { _id:1, userIndex: 10, email: "arealeamil@notreallu.asd", retainRate:2}**
What would be the most likely statement that we would need to grab email info for user indexes greater than 24? C

A. db.userIndex.find({email:{$lte:24}}, {_id:0})
B. db.userIndex.find({email:{$gt:24}}, {_id:0})
C. db.email.find({userIndex:{$gt:24}}, {email:1, _id:0})
D. db.email.find({userIndex:{$lte:24}}, {email:1, _id:0})

105. **What does it mean to have a _id:0 within our query statement?** D

A. Grab as many objects as possible.
B. Grab the first object in the results.
C. Does not have an effect, simple convention left for compatibility issues.
D. Tell MongoDB not to return a document id.

106. **This quiz encompasses data and content from Week 1 and 2, so we recommend reviewing that material from last week for this quiz as well. What is the highest level that the team has reached in gameclicks? (Hint: use the MAX operation in postgres).** C

A. 9
B. 10
C. 8
D. 7
E. 6

107. **How many user id's (repeats allowed) have reached the highest level as found in the previous question? (Hint: For postgres: you may either use two queries or use a sub-query).** E

A. 106436
B. 122757
C. 67271
D. 98823
E. 51294

108. **How many user id's (repeats allowed) reached the highest level in game-clicks and also clicked the highest costing price in buy-clicks? Hint: Refer to question 4 for ideas.** D

A. 66887
B. 73226
C. 23301
D. 32747

109. **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

A. 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.
B. Displays the users who have bought items worth $3 and have had a team with level 5.
C. This is an invalid line of code, the subquery is not formatted properly.
D. Counts the users who exists between both gameclicks and buyclicks files.

110. **In the MongoDB data set, what is the username of the twitter account who has a tweet_followers_count of exactly 8973882?** A

A. FIFAcom
B. SasSpear
C. Autocenterit
D. CreateImga

111. **What is the main problem with big data information integration?** D

A. Mediated Schema
B. Pay-as-you-go model
C. Probabilistic Schema Mapping
D. Many sources

112. **What would be the two possible solutions associated with "big data" information integration as mentioned in lecture? (Choose 2)** AC

A. Probabilistic Schema Mapping
B. Customer Transactions
C. Pay-as-you-go Model
D. Attribute Grouping
E. Mediated Schema

113. C

What are mediated schemas?

A. Schemas created from customer info.
B. Schemas created entirely from attribute grouping.
C. Schema created from integrating two or more schemas.
D. A type of probabilistic schema mapping.

114. **In attribute grouping, how would one evaluate if two attributes should go together? (Choose 2)** AC

A. Similarity of Attributes
B. Customer Interaction
C. Probability of Two Attributes Co-occurring
D. Integrated Views
E. Candidate Designs

115. **What is a data item?** C

A. The real worth of a data value.
B. Data found in a mediated schema.
C. Data that represents an aspect of a real-world entity.
D. Data found in a customer transaction.

116. **What is data fusion?** D

A. Another term for customer analytics.
B. Extracting true sources from a data source.
C. Extracting a global value from a data source.
D. Extracting the true value of a data item.

117. **What is a potential problem of having too many data sources as mentioned in lecture?** B

A. None, the problem is not a problem when using big data methodologies.
B. Too many data values.
C. Too much data processing required for compres-

sion.
D. Schema mapping becomes impossible.

118. **What do we mean when we say "the true value of a data item"?** A

A. Extrapolated data from a data item that represents the worth of that item.
B. Another term for data fusion.
C. Data created from statistical estimations.

119. **What is a potential method to deal with too many data sources as mentioned in lecture?** A

A. Compare and weigh each source by their trustworthiness.
B. None, the more the better.
C. Randomly select a sample of sources to represent the various data sources.
D. Take less samples per tick.

120. **Which of the queries below will return the average population of the counties in Georgia (be careful not to include the population of the state of Georgia itself)?** A

A. source="census.csv" CTYNAME != "Georgia" STNAME="Georgia" | stats mean(CENSUS2010POP)
B. source="census.csv" STNAME="Georgia" | stats mean(CENSUS2010POP)
C. source="census.csv" CTYNAME != "Georgia" STNAME="Georgia" | stats sum(CENSUS2010POP)
D. None of the above

121. **What is the average population of the counties in the state of Georgia (be careful not to include the population of the state of Georgia itself)?** D

A. 243767.4564
B. 45373.454788

C. 394383.53786
D. 60928.635220

122. **Of the options below, which query allows you to find the state with the most counties?** C

A. source="census.csv" | stats count by CTYNAME | sort num(count)
B. stats count by STNAME | sort -count
C. source="census.csv" | stats count by STNAME | sort count desc
D. source="census.csv" | stats count by CEN-SUS2010POP | sort count

123. **What state contains the most counties?** A

A. Texas
B. California
C. Georgia
D. Alaska

124. **Of the options below, which query allows you to find the most populated counties in the state of Texas?** C

A. STNAME="Texas" CENSUS2010POP > 100000 | sort -CENSUS2010POP | table CEN-SUS2010POP,CTYNAME
B. STNAME="Texas" CENSUS2010POP > 100000 | sort CENSUS2010POP desc | table CEN-SUS2010POP,CTYNAME
C. Both
D. Neither

125. **What is the most populated county in the state of Texas?** A

A. Harris
B. Dallas
C. Travis
D. Bexar

126. **What is data-parallelism as defined in lecture?** A

A. Running the same function simultaneously for the partitions of a data set on multiple cores.
B. At each step of the data pipeline, process values simultaneously by using multiple cores.
C. Having multiple multiple data pipelines at the same time.
D. Simultaneously processing input data from multiple cores.

127. **Of the following, which procedure best generalizes big data procedures such as (but not limited to) the map reduce process?** A

A. split->do->merge
B. split->map->shuffle and sort->reduce
C. split->sort->merge
D. split ->shuffle and sort->map->reduce

128. **What are the three layers for the Hadoop Ecosystem? (Choose 3)** BCD

A. Data Creation and Storage
B. Data Management and Storage
C. Data Integration and Processing
D. Coordination and Workflow Management
E. Data Manipulation and Integration

129. **What are the 5 key points in order to categorize big data systems?** D

A. Coordination, Latency, Productivity, Flexibility, Fault Tolerance
B. Coordination, Latency, Productivity, Speed, Fault Tolerance
C. Execution model, Speed, Scalability, Flexibility, Fault Tolerance

D. Execution model, Latency, Scalability, Programming Language, Fault Tolerance

130. **What is the lambda architecture as shown in lecture?** B

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

131. **Which of the following scenarios is NOT an aggregation operation?** B

A. Averaging the total number of data per type.
B. Removing undefined values.
C. Counting the total number of data per type.
D. Counting the total number of data.

132. **What usually happens to data when aggregated as mentioned in lecture?** D

A. Data becomes personalized.
B. Data becomes faster to process.
C. Data become organized.
D. Data becomes smaller.

133. **What is K-means clustering?** D

A. Classify data by k decisions.
B. Divide samples using k lines.
C. Classify data by k actions.
D. Group samples into k clusters.

134. **Why is Hadoop not a good platform for machine learning as mentioned in lecture? (Choose 4)** ABDE

A. Bottleneck using HDFS.
B. Java support only.

C. Too massive.
D. Map and Reduce Based Computation.
E. No interactive shell and streaming.
F. Requires nodes and multiple machines.
G. Unable to support machine learning.

135. **What are the layers (parts) of Spark? (Choose 5)** BDFGH

A. Spark RDD
B. Spark Core
C. Worker Node
D. MLlib
E. Spark Graph
F. Graphx
G. Spark Streaming
H. SparkSQL

136. **What is in-memory processing?** D

A. Having the pipeline completely in memory.
B. Having the input completely in memory.
C. Writing data to disk between pipeline steps.
D. Writing data to memory between pipeline steps.
E. Having the input completely in disk.
F. Having the pipeline completely in disk.

137. **What does the following line of code do?** A

words = lines.flatMap(lambda line: line.split(" "))

A. Each line in the document is split up into words.
B. Each word is merged into lines to be counted later.
C. Each word in each line is counted.
D. Each line in the document is split into various Spark partitions.

138. **What does the following line of code imply about the state of partitions before the action is performed?** B

words = lines.flatMap(lambda line: line.split(" "))

A. There is only one single partition containing the full document.
B. Each Spark partition corresponds to a line in the document.
C. Each Spark partition corresponds to a word in the document.

139. **When the following command is executed, where is the file written and how can it be accessed?** D

counts.coalesce(1).saveAsTextFile('hdfs:/user/cloud-era/wordcount/outputDir')

A. The local file system and through the directory with the "cd" terminal command.
B. HDFS and through the system directory with the "cd" terminal command.
C. The local file system and through the "hadoop fs" command.
D. HDFS and through the "hadoop fs" command.

140. **What does the number one (1) allow us to do in the following line of code?** C

tuples = words.map(lambda word: (word,1))

A. The number represents the number of partitions in charge of keeping track of each word.
B. None, completely arbitrary in order to apply an algorithm that requires a tuple.
C. Treat each word with a weight of one during the counting process.
D. The number represents the number of partitions in charge of counting each line.

141. **Which part of SPARK is in charge of creating RDDs?** D

A. Spark Executor

B. Local CPU
C. Worker Node
D. Driver Program
E. Storage

142. **How does lazy evaluation work in Spark?** A

A. Transformations are not executed until the action stage.
B. Transformations are queued and executed at a certain threshold.
C. Actions are queued and executed at a certain threshold.
D. Actions are not executed until the transformation stage.

143. **What are the consequences of lazy evaluation as mentioned in lecture?** A

A. Errors sometimes do not show up until the action stage.
B. Hiccups within the system during queue execution.
C. There are no consequences.

144. **What is a wide transformation?** B

A. Transformations that take a lot of nodes to complete.
B. A transformation that requires data shuffling across node partitions.
C. A longer time-taking transformation compared to narrow transformations.
D. The name for the most used transformations.

145. **Where does the data for each worker node get sent to after a collect function is called?** E

A. Other Worker Nodes
B. None; Stays in the Same Node

C. Spark SQL
D. Spark Streaming
E. Spark Context

146. **What are DataFrames?** B

A. A special type of data node that contains framework to manipulate SQL.
B. A column like data format that can be read by Spark SQL.
C. A type of narrow transformation.

147. **Can RDD's be converted into DataFrames directly without manipulation?** A

A. No: lines have to be converted into row.
B. Yes
C. No: RDD's needed to be made relational first.
D. No: RDD's cannot be converted into DataFrames.

148. **What is the function of Spark SQL as mentioned in lecture? (Choose 3)** BEF

A. Better worker node interpolation.
B. Deploy business intelligence tools over Spark.
C. Efficient data manipulation using SQL like structure.
D. Better ability to manipulate big data.
E. Connect to variety of databases.
F. Enables relational queries on Spark.

149. **What is a triplet in GraphX?** A

A. A type of data to contain the information on connections between vertices and edges.
B. A type of data to contain vertex info.
C. A type of data to contain edge info.
D. A type of data to contain both edge and vertex info.

150. B

What does the following filter line of code do?

df.filter(df["teamlevel"] > 1)

A. Select the first two columns of the data and displays only team levels greater than 1.
B. Filter each row to show only team levels larger than 1.
C. Filter each column to show only team levels larger than 1.
D. Select the first two columns of the data and filter each column to show only team levels larger than 1.

151. **What does the following do?** A

df.select("userid", "teamlevel").show(5)

A. Select the columns named "userid" and "teamlevel" and display first 5 rows.
B. Display all columns except "userid" and "teamlevel".
C. Select the rows named "userid" and "teamlevel" and display first 5 rows.
D. Display all rows except "userid" and "teamlevel".

152. **What does the 1 represent in the following line of code?** B

ssc = StreamingContext(sc,1)

A. To specific debug output.
B. A batch interval of 1 second.
C. To create only one partition to manage the stream.
D. To create one single context.

153. **What does the following code do?** A

window = vals.window(10, 5)

A. Creates a window that combines 10 seconds worth of data and moves by 5 seconds.
B. Creates 10 windows with 5 seconds worth of data in them.
C. Creates a batch interval between 10 seconds and 5 seconds.
D. Creates 10 windows with 5 batch intervals inbe-tween.

154. **How many tweets have location not null?** A

A. 6937
B. 6945
C. 6973
D. 5957
E. No option applicable.

155. **How many people have more followers than friends? (Hint : use this.user instead of user).** C

A. 5206
B. 6673
C. 5809
D. 5590
E. 6238

156. **Perform a query that returns the text of tweets which have the string "http://". Which of the following sub-strings do NOT occur in the results? (Choose all that apply)** AC

A. @DundalkFC
B. @Ass0Star
C. @Infosmessi_
D. @TerraceImages
E. @espn

157. **Query: Return all the tweets which contain text "Eng-land" but not "UEFA". In these results the string "Euro 2016" appears in...** C

A. 0 tweets
B. 5 tweets
C. 2 tweets
D. More than 6 tweets.
E. 3 tweets

158. **Query: Get all the tweets from the location "Ireland" which also contain the string "UEFA". In this result the user with the highest friends count is...** B

A. Pauldonaghue
B. ProfitwatchInfo
C. irishexaminer
D. Insight4News4
E. DerekRantsGames

159. **How many different countries are mentioned in at least one tweet?** D

A. 211
B. 112
C. 64
D. 44

160. **How many times is any country mentioned in a tweet?** C

A. 26634
B. 52
C. 397
D. 211

161. **What are the three countries with the highest mentioned count** C

A. Thailand, Mexico, Denmark
B. Thailand, Iceland, Mexico
C. Norway, Nigeria, France
D. Nigeria, Slovakia, Germany

162. **How many times was France mentioned in a tweet?** D

A. 8
B. 25
C. 30
D. 42

163. **Which country was mentioned most: Kenya, Wales, or Netherlands?** B

A. Kenya
B. Wales
C. Netherlands

164. **What is the average number of times a country is mentioned? (Round to the nearest integer)** C

A. 15
B. 3
C. 9
D. 44

165. **What is NOT machine learning?** D

A. Learning from data
B. Data-driven decisions
C. Discover hidden patterns
D. Explicit, step-by-step programming

166. **Which of the following is NOT a category of machine learning?** D

A. Cluster Analysis
B. Regression
C. Classification
D. Algorithm Prediction
E. Association Analysis

167. **Which categories of machine learning techniques are supervised?** B

A. cluster analysis and association analysis
B. classification and regression
C. classification and cluster analysis
D. regression and association analysis

168. **In unsupervised approaches,** C

A. the target is unlabeled.
B. the target is provided.
C. the target is unknown or unavailable.
D. the target is what is being predicted.

169. **What is the sequence of the steps in the machine learning process?** A

A. Acquire -> Prepare -> Analyze -> Report -> Act
B. Prepare -> Acquire -> Analyze -> Report -> Act
C. Acquire -> Prepare -> Analyze -> Act -> Report
D. Prepare -> Acquire -> Analyze -> Act -> Report

170. **Are the steps in the machine learning process apply-once or iterative?** A

A. Iterative
B. The first two steps, Acquire and Prepare, are apply-once, and the other steps are iterative.
C. Apply-once

171. **Phase 2 of CRISP-DM is Data Understanding. In this phase,** C

A. we define the problem or opportunity to be addressed.
B. we prepare the data for analysis.
C. we acquire as well as explore the data that is related to the problem.

172. **What is the main difference between KNIME and Spark MLlib?** D

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

173. **Which of these statements is true about samples and variables?** B

A. A variable describes a specific characteristic of an entity in your data.
B. All of these statements are true.
C. A sample can have many variables to describe it.
D. A sample is an instance or example of an entity in your data.

174. **Other names for 'variable' are** C

A. sample, row, observation
B. numerical, quantitative
C. feature, column, attribute
D. categorical, nominal

175. **What is the purpose of exploring data?** D

A. To digitize your data.
B. To generate labels for your data.
C. To gather your data into one repository.
D. To gain a better understanding of your data.

176. **What are the two main categories of techniques for exploring data? Choose two.** CE

A. Correlations

B. Outliers
C. Visualization
D. Histogram
E. Summary statistics
F. Trends

177. **Which of the following are NOT examples of summary statistics?** C

A. skewness, kurtosis
B. mean, median, mode
C. data sources, data locations
D. standard deviation, range, variation

178. **What are the two measures for measuring shape as mentioned in the lecture? Choose two.** AC

A. Skewness
B. Mode
C. Kurtosis
D. Range
E. Contingency Table

179. **Which of the following would NOT be a good reason to use a box plot?** A

A. To show correlations between two variables.
B. To show and compare distribution values
C. To show data distribution shapes such as asymmetry and skewness.

180. **All of the following are true about data visualization EXCEPT** B

A. Should be used with summary statistics for data exploration.
B. Is more important than summary statistics for data exploration
C. Provides an intuitive way to look at data.
D. Is useful for communicating results.

181. **What is the maximum of the average wind speed measurements at 9am (to 2 decimal places)?** A

A. 23.56
B. 29.84
C. 5.50
D. 4.55

182. **How many rows containing rain accumulation at 9am measurements have missing values?** A

A. 6
B. 4
C. 3
D. 2

183. **What is the correlation between the relative humidity at 9am and at 3pm (to 2 decimal places, and without removing or imputing missing values)?** A

A. 0.88
B. 1.00
C. -0.45
D. 0.19

184. **If the histogram for air temperature at 9am has 50 bins, what is the number of elements in the bin with the most elements (without removing or imputing missing values)?** A

A. 57
B. 224
C. 49
D. 166

185. **What is the approximate maximum max_wind_direction_9am when the maximum max_wind_speed_9am occurs?** A

A. 70
B. 30
C. 312

186. **Which of the following is NOT a data quality issue?** A

A. Scaled data
B. Missing values
C. Inconsistent data
D. Duplicate data

187. **Imputing missing data means to** D

A. drop samples with missing values.
B. replace missing values with outliers.
C. merge samples with missing values.
D. replace missing values with something reasonable.

188. **A data sample with values that are considerably different than the rest of the other data samples in the dataset is called an/a _____.** C

A. Noise
B. Inconsistent data
C. Outlier
D. Invalid data

189. **Which one of the following examples illustrates the use of domain knowledge to address a data quality issue?** C

A. Drop samples with missing values
B. None of these
C. Merge duplicate records while retaining relevant data
D. Simply discard the samples that lie significantly outside the distribution of your data

190. B

Which of the following is NOT an example of feature selection?

A. Re-formatting an address field into separate street address, city, state, and zip code fields.
B. Replacing a missing value with the variable mean.
C. Adding an in-state feature based on an applicant's home state.
D. Removing a feature with a lot of missing values.

191. **Which one of the following is the best feature set for your analysis?** D

A. Feature set with the smallest number of features
B. Feature set that contains exclusively re-coded features
C. Feature set with the largest number of features
D. Feature set with the smallest set of features that best capture the characteristics of the data for the intended application

192. **The mean value and the standard deviation of a zero-normalized feature are** C

A. mean = 1 and standard deviation = 1
B. mean = 1 and standard deviation = 0
C. mean = 0 and standard deviation = 1
D. mean = 0 and standard deviation = 0

193. **Which of the following is NOT true about PCA?** A

A. PCA is a dimensionality reduction technique that removes a feature that is very correlated with another feature.
B. PCA stands for principal component analysis
C. PC1, the first principal component , captures the largest amount of variance in the data along a single dimension.
D. PC1 and PC2, the first and second principal com-

ponents, respectively, are always orthogonal to each other.

194. **If we remove all missing values from the data, how many air pressure at 9am measurements have values between 911.736 and 914.67?** A

A. 77
B. 287
C. 80

195. **If we impute the missing values with the minimum value, how many air temperature at 9am measurements are less than 42.292?** A

A. 28
B. 23
C. 1
D. 5

196. **How many samples have missing values for air_pressure_9am?** A

A. 3
B. 5
C. 1092
D. 0

197. **Which column in the weather dataset has the most number of missing values?** A

A. rain_accumulation_9am
B. number
C. They are all the same
D. air_temp_9am

198. **When we remove all the missing values from the dataset, the number of rows is 1064, yet the variable with most missing values has 1089 rows. Why did the number of rows decrease so much?** A

A. Because the missing values in each column are not necessarily in the same row
B. Because rows with missing values as well as rows with 0s are removed
C. Because rows with missing values as well as rows with duplicate values are removed

199. **Which of the following is a TRUE statement about classification?** C

A. In a classification problem, the target variable has only two possible outcomes.
B. Classification is an unsupervised task.
C. Classification is a supervised task.

200. **In which phase are model parameters adjusted?** A

A. Training phase
B. Testing phase
C. Data preparation phase
D. Model parameters are constant throughout the modeling process.

201. **Which classification algorithm uses a probabilistic approach?** C

A. k-nearest-neighbors
B. decision tree
C. naive bayes
D. none of the above

202. **What does the 'k' stand for in k-nearest-neighbors?** A

A. the number of nearest neighbors to consider in classifying a sample
B. the number of training datasets
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 samples in the dataset

203. **During construction of a decision tree, there are several criteria that can be used to determine when a node should no longer be split into subsets. Which one of the following is NOT applicable?** D

A. All (or X% of) samples have the same class label.
B. The tree depth reaches a maximum threshold.
C. The number of samples in the node reaches a minimum threshold.
D. The value of the Gini index reaches a maximum threshold.

204. **Which statement is true of tree induction?** D

A. An impurity measure is used to determine the best split for a node.
B. You want to split the data in a node into subsets that are as homogeneous as possible
C. For each node, splits on all variables are tested to determine the best split for the node.
D. All of these statements are true of tree induction.

205. **What does 'naive' mean in Naive Bayes?** A

A. The model assumes that the input features are statistically independent of one another. The 'naïve' in the name of classifier comes from this naïve assumption.
B. The full Bayes' Theorem is not used. The 'naive' in naive bayes specifies that a simplified version of Bayes' Theorem is used.
C. The Bayes' Theorem makes estimating the probabilities easier. The 'naïve' in the name of classifier comes from this ease of probability calculation.

206. **The feature independence assumption in Naive Bayes simplifies the classification problem by** D

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.

207. **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

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

208. **KNIME: Considering the Numeric Binner node again, what would happen if the "Append new column" box is not checked?**

A

A. The relative_humidity_3pm variable will become a categorical variable
B. The relaltive_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

209. **KNIME: How many samples had a missing value for air_temp_9am before missing values were addressed?**

A

A. 5
B. 3
C. 0

210. **KNIME: How many samples were placed in the test set after the dataset was partitioned into training and test sets?** A

A. 213
B. 851
C. 20

211. **KNIME: What are the target and predicted class labels for the first sample in the test set?** A

A. Both are humidity_not_low
B. Target class label is humidity_not_low, and predicted class label is humidity_low
C. Target class label is humidity_low, and predicted class label is humidity_not_low

212. **Spark: What values are in the number column?** A

A. Integer values starting at 0
B. Time and date values
C. Random integer values

213. **Spark: With the original dataset split into 80% for training and 20% for test, how many of the first 20 samples from the test set were correctly classified?** A

A. 19
B. 10
C. 1

214. **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

A. 730
B. 334
C. 70

215. **A model that generalizes well means that** D

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

216. **What indicates that the model is overfitting?** A

A. Low training error and high generalization error
B. Low training error and low generalization error
C. High training error and high generalization error
D. High training error and low generalization error

217. **Which method is used to avoid overfitting in decision trees?** A

A. Pre-pruning and post-pruning
B. None of these
C. Pre-pruning
D. Post-pruning

218. **Which of the following best describes a way to create and use a validation set to avoid overfitting?** B

A. random sub-sampling
B. All of these
C. k-fold cross-validation
D. leave-one-out cross-validation

219. **Which of the following statements is NOT correct?** B

A. The training set is used to adjust the parameters of

the model.
B. The test set is used for model selection to avoid overfitting.
C. The validation set is used to determine when to stop training the model.
D. The test set is used to evaluate model performance on new data.

220. **How is the accuracy rate calculated?** D

A. Add the number of true positives and the number of false negatives.
B. Divide the number of true positives by the number of true negatives.
C. Subtract the number of correct predictions from the total number of predictions.
D. Divide the number of correct predictions by the total number of predictions

221. **Which evaluation metrics are commonly used for evaluating the performance of a classification model when there is a class imbalance problem?** C

A. precision and error
B. accuracy and error
C. precision and recall
D. precision and accuracy

222. **How do you determine the classifier accuracy from the confusion matrix?** B

A. Divide the sum of the off-diagonal values in the confusion matrix by the total number of samples.
B. Divide the sum of the diagonal values in the confusion matrix by the total number of samples.
C. Divide the sum of the diagonal values in the confusion matrix by the sum of the off-diagonal values.
D. Divide the sum of all the values in the confusion matrix by the total number of samples.

223. **KNIME: In the confusion matrix as viewed in the Scorer node, low_humidity_day is:**

A. the target class label
B. the predicted class label
C. the only input variable that is categorical

A

224. **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

A

225. **KNIME: In the Table View of the Interactive Table, each row is color-coded. Blue specifies:**

A. that the target class label for the sample is humidity_not_low
B. that the target class label for the sample is humidity_low
C. that the predicted class label for the sample is humidity_not_low
D. that the predicted class label for the sample is humidity_low

A

226. **KNIME: To change the colors used to color-code each sample in the Table View of the Interactive Table node:**

A. change the color settings in the Color Manager node
B. change the color settings in the Interactive Table dialog
C. It is not possible to change these colors

A

227. **KNIME: In the Table View of the Interactive Table, the values in RowID are not consecutive because:**

A. the RowID values are from the original dataset, and only the test samples are displayed here
B. the samples are randomly ordered in the table
C. only a few samples from the test set are randomly selected and displayed here

A

228. **Spark: To get the error rate for the decision tree model, use the following code:**

A. print ("Error = %g " % (1.0 - accuracy))
B. evaluator = MuticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="error")
C. error = evaluator.evaluate(1 - predictions)

A

229. **Spark: To print out the accuracy as a percentage, use the following code:**

A. print ("Accuracy = %.2g" % (accuracy * 100))
B. print ("Accuracy = %100g" % (accuracy))
C. print ("Accuracy = %100.2g" % (accuracy))

A

230. **Spark: In the last line of code in Step 4, the confusion matrix is printed out. If the "transpose()" is removed, the confusion matrix will be displayed as:**

A. array([[87., 14.], [26., 83.]])
B. array([[83., 26.], [14., 87.]])
C. array([[83., 87.], [14., 26.]])

A

231. **What is the main difference between classification and regression?**

C

A. In classification, you're predicting a categorical variable, and in regression, you're predicting a nominal variable.
B. In classification, you're predicting a number, and in regression, you're predicting a category.
C. In classification, you're predicting a category, and in regression, you're predicting a number.
D. There is no difference since you're predicting a numeric value from the input variables in both tasks.

232. **Which of the following is NOT an example of regression?** D

A. Predicting the price of a stock
B. Predicting the demand for a product
C. Estimating the amount of rain
D. Determining whether power usage will rise or fall

233. **In linear regression, the least squares method is used to** A

A. Determine the regression line that best fits the samples.
B. Determine how to partition the data into training and test sets.
C. Determine whether the target is categorical or numerical.
D. Determine the distance between two pairs of samples.

234. **How does simple linear regression differ from multiple linear regression?** B

A. In simple linear regression, the input has only categorical variables. In multiple linear regression, the input can be a mix of categorical and numerical variables.
B. In simple linear regression, the input has only one variable. In multiple linear regression, the input has

more than one variables.
C. In simple linear regression, the input has only categorical variables. In multiple linear regression, the input has only numerical variables.
D. They are the just different terms for linear regression with one input variable.

235. **The goal of cluster analysis is** A

A. To segment data so that differences between samples in the same cluster are minimized and differences between samples of different clusters are maximized.
B. To segment data so that all categorical variables are in one cluster, and all numerical variables are in another cluster.
C. To segment data so that differences between samples in the same cluster are maximized and differences between samples of different clusters are minimized.
D. To segment data so that all samples are evenly divided among the clusters.

236. **Cluster results can be used to** D

A. Determine anomalous samples
B. Classify new samples
C. Create labeled samples for a classification task
D. All of these choices are valid uses of the resulting clusters.
E. Segment the data into groups so that each group can be analyzed further

237. **A cluster centroid is** B

A. The mean of all the samples in the two closest clusters.
B. The mean of all the samples in the cluster
C. The mean of all the samples in all clusters

D. The mean of all the samples in the two farthest clusters.

238. **The main steps in the k-means clustering algorithm are** B

A. Calculate the distances between the cluster centroids, then find the two closest centroids.
B. Assign each sample to the closest centroid, then calculate the new centroid.
C. Count the number of samples, then determine the initial centroids.
D. Calculate the centroids, then determine the appropriate stopping criterion depending on the number of centroids.

239. **The goal of association analysis is** D

A. To find the number of outliers in the data
B. To find the most complex rules to explain associations between as many items as possible in the data.
C. To find the number of clusters for cluster analysis
D. To find rules to capture associations between items or events

240. **In association analysis, an item set is** B

A. A set of items that infrequently occur together
B. A transaction or set of items that occur together
C. A set of items that two rules have in common
D. A set of transactions that occur a certain number of times in the data

241. **The support of an item set** B

A. Captures the correlation between the items in that item set
B. Captures the frequency of that item set
C. Captures how many times that item set is used in

a rule
D. Captures the number of items in that item set

242. **Rule confidence is used to** B

A. Measure the intuitiveness of a rule
B. Prune rules by eliminating rules with low confidence
C. Determine the rule with the most items
D. Identify frequent item sets

243. **What percentage of samples have 0 for rain_accumulation?** A

A. 157812 / 158726 = 99.4%
B. 157237 / 158726 = 99.1%
C. There is not enough information to determine this

244. **Why is it necessary to scale the data (Step 4)?** A

A. 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.
B. Since the values of the features are on different scales, all features need to be scaled so that all values will be positive.
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.

245. **If we wanted to create a data subset by taking every 5th sample instead of every 10th sample, how many samples would be in that subset?** A

A. 317,452
B. 1,587,257
C. 158,726

246. A

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

247. **Just by looking at the values for the cluster centers, which cluster contains samples with the lowest relative humidity?** A

A. Cluster 4
B. Cluster 3
C. Cluster 9

248. **What do clusters 7, 8, and 11 have in common?** A

A. They capture weather patterns associated with warm and dry days
B. They capture weather patterns associated with high air pressure
C. They capture weather patterns associated with very strong winds

249. **If we perform clustering with 20 clusters (and seed = 1), which cluster appears to identify Santa Ana conditions (lowest humidity and highest wind speeds)?** A

A. Cluster 12
B. Cluster 1
C. Cluster 16

250. **We did not include the minimum wind measurements in the analysis since they are highly correlated with the average wind measurements. What is the correlation between min_wind_speed and avg_wind_speed (to two decimals)? (Compute this using one-tenth of the original dataset, and dropping all rows with missing values.)**

A. 0.97
B. -0.12
C. 0.62

A

251. **Which of the following are graphs? (check all that apply)**

A.
https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/DAGEx.png
B.
https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/PieChartSmaller.png

A

252. **Which of the following is the correct adjacency matrix for this graph?**

https://blogger.googleusercontent.com/img/a/AVvXsEiUXhTJr0i7HDlgyKfwcdYWV5b5sAPXs80MBlETSkTc6EANBDCmnRb8BDVvfvJR5tguTFY5q7h8xUyVkXZ_a-1e2eDA1BYl2SGByjPQtOHrgiPpvsH84V-J7jR6mgVKa0pfOppSOjJUeZyEPGcQ6nLEEpRJ8PbVU-nW7isqOgESaoYJtI48eKBMEVrcHw=w145-h112

A.
https://blogger.googleusercontent.com/img/a/AVvX-

C

sEi7fi_ZD_l4zF1k7Z6l9DyOEicKbHjWDwHN2NiWrG-WwzJKalbu_FP4xFRv-lpRKhayKyY7cDqKL2S0zGdR263FmDz0JQWhu53gI ZqAKmRjIGlaLhbsZOd0peNIlWuNY-HupY9cf5lY0G4fFjQW0-NJVe9ygmbDvHX8D3vPFu-maE1KvbA=w167-h169
B. Neither option is correct.
C. https://blogger.googleusercontent.com/img/a/AVvX-sEih8UVbm76FFiOfBrB-wg1i2-40bqQ91WtEesa2O3YwCTj5ujJhjBhU-EyD-WaRCCLb55HhCfFg6AsUwmJAWNuMxCK7uCfSf-PO5yXm9uiM1JB2fXPiiRQxX07vpuU3M1nkY-GOsn_uq_cceRZ0fpwOcj1Nrh3llSSZGay655xWJir-JhReUYvZh9bHWkrivDw=w194-h180

253. **Which of the following content would be objects (or nodes) in a graph that represents the activity in a facebook page?** BCD

A. Created_post (the action of creating a post)
B. comment text
C. location
D. post text
E. friends (the action of making someone your friend)

254. **Based on the videos, which kinds of analysis might one be able to perform on a tweet graph?** ABC

A. find interacting groups of users
B. extract conversation threads
C. find influencers in a twitter community

255. **The key reason mentioned in the video that biology applications need Big Data analytics is...** B

A. The complexity of interactions that correlate to inform phenotypes.
B. The integration of multiple data sources from different researchers and of different sources of infor-

mation.
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.

256. **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?** D

A. Volume
B. Variety
C. Valence
D. Velocity

257. **Which of the Vs results in increased algorithmic complexity (which can cause analyses to not be able to finish running in reasonable amounts of time)?** A

A. Volume
B. Variety
C. Valence
D. Velocity

258. **Which of the Vs results in challenges due to graphs created from varying kinds, formats, sources, and meanings of data?** C

A. Valence
B. Volume
C. Variety
D. Velocity

259. **Which of the Vs causes increased interconnectivity of a graph -- which can cause problems in analysis due to density?** B

A. Variety
B. Valence

C. Volume
D. Velocity

260. **Updating a graph with a stream of posting information on facebook is an example of which of the Vs?** A

A. Velocity
B. Volume
C. Variety
D. Valence

261. **Studying Amarnath's gmail interactions over time (as gmail started to be used by more and more people) is BEST defined as an impact of which of the Vs?** A

A. Valence
B. Velocity
C. Variety
D. Volume

262. **A graph representing tweets would have only "one type" (e.g. label) of node.** B

A. True
B. False

263. **In a network representing the world wide web nodes would likely represent:** B

A. Hyperlinks
B. Webpages
C. Google search terms
D. Individual computers

264. **In a network representing the world wide web edges (or links) would likely represent** A

A. Hyperlinks
B. Webpages

C. Google search terms
D. Individual computers

265. **In an email network, which might reasonably be represented by weight on edges?** B

A. the total number of people who sent an email in a week
B. average number of emails sent from one user to another in a week
C. the total number of emails sent by one user in a week

266. **A loop in a graph is where:** C

A. when there is a edge from A->B, there is also an edge from B->A.
B. where there is a path in some way from a node, through 1 or more other nodes, back to the original node.
C. where there is an edge from a node to itself.

267. **An example of a loop in a graph could occur when:** A

A. Someone emails themself
B. Someone emails a friend who replies
C. Someone emails a friend, who emails another friend, who then replies to you

268. **When trying to represent a relationship between Maria and Julio who have more than one relationship to each other (e.g., tennis partner, co-worker, emergency contact) which of the following would be needed in a graph representing those relationships** B

A. Separate graphs for each kind of relationship
B. Multiple edges between Maria and Julio
C. Multiple nodes for each of Maria and Julio, to capture the various relationships

269. **In many applications paths (where we go from one node to another without repeating nodes) are more useful than walks (where we can repeat a node when going from one node to another).** A

A. True
B. False

270. **Trails (paths without repeated edges) can be interesting in which of the following problem applications?** C

A. An email network tracing frequency of emails from one person to another.
B. An email network tracing email replies.
C. Routing to avoid using the same bridge or road.
D. Routing to avoid visiting the same city.

271. **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

A. Julio's node was reachable from Maria node
B. Maria's node was reachable from Julio's node

272. **If I want to find the diameter of a graph, I should start by finding the shortest path between each set of nodes.** A

A. True
B. False

273. **What is the diameter of this graph?** B

A. 1
B. 2
C. 3

274. **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

A. Maximize
B. Minimize

275. **Which are the two kinds of constraints on paths discussed in the video on basic path analytics? (check 2) Hint: remember the example of Amarnath needing to get to work by taking his son to school.** AB

A. Exclusion of nodes and/or edges
B. Inclusion of nodes and/or edges
C. Directionality

276. **What are examples of preference constraints in the Google Maps application?** AB

A. Avoid roads under construction
B. Avoid highways
C. Include son's school

277. **Which of the statements below is true?** B

A. Dijsktra's algorithm is computationally efficient (has low computational complexity).
B. Dijsktra's algorithm is computationally inefficient (has high computational complexity).

278. **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)** AB

A. Reduction of the size of the graph
B. Splitting the task into 2 independent shortest path

problems
C. Changing the weights on the edges of the graph and/or subgraphs

279. **The example given in the lectures of when a power network loses power in large portions of its service area was an example of what?** B

A. a problem that can occur when centrality is too high
B. an attack which causes disconnection of the graph
C. high levels of connectivity which make it easy to bring a network down

280. **Is the following graph strongly connected, weakly connected or neither?** A

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/ABCD--Counter-Clockwise.png

A. strongly connected
B. neither
C. weakly connected

281. **Is the following graph strongly connected, weakly connected or neither?** A

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/ABCD-Pointing-to-C.png

A. weakly connected
B. neither
C. strongly connected

282. **If you were going to look for a node which would be most likely to be the target of an attack to disconnect a network, what would be the best characteristic to look for?**

A. low degree nodes
B. high degree nodes
C. nodes that, if they were removed, would cause the graph to go from strongly connected to weakly connected

B

283. **What is the out-degree of node B?**

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q5.png

A. 0
B. 1
C. 2
D. 3

A

284. **What is the in-degree of node B?**

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q5.png

A. 0
B. 1
C. 2
D. 3

D

285. **In the graph below, which node is the greatest listener?**

https://github.com/AlessandroCorradini/Universi-

B

ty-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q6.png

A. A
B. B
C. C
D. D

286. **In the graph below, which node is the greatest talker?** D

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q6.png

A. A
B. B
C. C
D. D

287. **In the graph below, which nodes are the greatest communicators? (Hint: there's a tie)** AC

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q7.png

A. A
B. B
C. C
D. D

288. **What would we be looking for if we followed the steps below? Note: we have 2 graphs.**
Create a table for each graph where, for each node, you list the degree of the node.
For each graph, create a histogram indicating how A

**many nodes in that graph have a specific degree (e.g., how many nodes have degree 1? 2? etc.).
Use advanced approaches (e.g. Euclidean distances) to compare these two histograms.**

A. Similarity
B. Centrality
C. Community
D. Connectivity

289. **Which of the following are the three type of analytics questions asked about communities?** ABC

A. Static
B. Evolution
C. Prediction
D. Connection

290. **What type of community analytics question is the following?
Did a community form on twitter around the 2014 World Cup in Brazil?** C

A. Static
B. Prediction
C. Evolution
D. Connection

291. **Which type of community analytics question is the following?
How tightly knit was the 2014 World Cup twitter community on July 13, 2014 (the day of the finals)?** D

A. Connection
B. Prediction
C. Evolution
D. Static

292. **What is the internal degree of the node indicated in the graph below?** C

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q12.png

A. 1
B. 2
C. 3
D. 4

293. **What is the external degree of the node indicated in the graph below?** A

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q12.png

A. 1
B. 2
C. 3
D. 4

294. **Which of the two graphs below is more modular?** B

https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization/raw/master/05%20-%20Graph%20Analytics%20for%20Big%20Data/img/Q13.png

A. A
B. B

295. **Which of the following community tracking phases best describes what (usually) happens a few months after two companies merge?** D

A. Birth

B. Split
C. Grow
D. Contract
E. Merge
F. Death

296. **Which of the following community tracking phases usually occurs when a company spins off a start-up?** D

A. Contract
B. Grow
C. Birth
D. Split
E. Death
F. Merge

297. **An influencer in a network is defined as:** A

A. a node which can reach all other nodes quickly
B. a node which has heavy weight edges to at least 1/2 of the nodes in the network
C. the biggest gossip in the network

298. **Which of the following are the 2 core "key player" problems that centrality analytics can address?** BD

A. What is the shortest path through a network
B. Which nodes' removal will maximally disrupt the network
C. Which nodes have the highest ratio of out-degree nodes to in-degree nodes
D. A set of nodes which can reach (almost) all other nodes

299. **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

A. Closeness

B. Group
C. Between-ness
D. Degree

300. **What kind of centrality would you want to analyze in a graph if you wanted maximize commodity flow in a network?** B

A. Group
B. Between-ness
C. Closeness
D. Degree

301. **What kind of centrality identifies "hubness"?** B

A. Between-ness
B. Degree
C. Group
D. Closeness

302. **Which of the following is a Cypher command used to combine two or more query results?** A

A. union
B. combine
C. merge
D. return

303. **For a graph network whose nodes are all of type "MyNode", which has both incoming and outgoing edges, and which has both root and leaf nodes, what will the following Cypher code return in a Neo4j report?** B

match (n:MyNode)<-[r]-() return n

A. All nodes and edges except leaf nodes and their edges.
B. All nodes except root nodes.

C. Edges but no nodes.
D. The entire network, all nodes and edges

304. **The Cypher query language shares some commands in common with SQL.** A

A. True
B. False

305. **The following query will return a graph containing whatever loops might exist.** A

match (n)-[r]-(n) return n, r

A. True
B. False

306. **Which Cypher pattern is used to represent a node?** A

A. ()
B. []
C. {}
D. <>

307. **Neo4j is a ...** A

A. Graph database
B. Relational database
C. None of the above

308. **Which Cypher command launches a Neo4j database search?** A

A. MATCH
B. RETURN
C. CREATE
D. None of the above

309. **Cypher does not include a specific command to find the shortest path in a graph network.** A

A. False
B. True

310. **Cypher includes a 'diameter' command to find the longest path in a graph network.** A

A. False
B. True

311. **What is the number of nodes returned?** B

A. 50,000
B. 9656
C. 9756
D. 8673

312. **What's the number of edges?** C

A. 50,000
B. 49,834
C. 46,621
D. None of the above

313. **The number of loops in the graph is:** C

A. 1035
B. 1395
C. 1221
D. 1243

314. **The query match (n)-[r]->(m) where m <> n return distinct n, m, count(r) gives us** A

A. the count of all non loop edges between every adjacent node pair.
B. the count of all edges between every adjacent node pair.
C. the count of all edges.
D. None of the above

315. **The query match (n)-[r]->(m) where m <> n return distinct n, m, count(r) as myCount order by myCount desc limit 1 produces what?** D

A. a random edge
B. the node with the maximum number of looping edges
C. two neighboring nodes, each with a high outdegree
D. the pair of nodes with the maximum number of multi-edges between them

316. **The query match p=(n {Name:'BRCA1'})-[:AssociationType*..2]->(m) return p produces what?** A

A. The 2-neighborhood of the node whose name is 'BRCA1'
B. The neighbors whose distance is greater than 1 and less than 2 of the node whose name is 'BRCA1'
C. The neighbors' neighbors of the node whose name is 'BRCA1'
D. The neighbors of the node whose name is 'BRCA1'

317. **How many non-directed shortest paths are there between the node named 'BRCA1' and the node named 'NBR1'?** B

A. 8
B. 9
C. 10
D. None of the above

318. **The top 2 nodes with the highest outdegree are:** D

A. GRB2 and TP53
B. EP300 and BRCA1
C. MEPCE and EGFR
D. SNCA and BRCA1

319. B

Applying the example queries provided to you, create the degree histogram for the network. How many nodes in the graph have a degree of 3?

A. 1351
B. 821
C. 675
D. 512

320. **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)**

ABC

```
val countries: RDD[(VertexId, PlaceNode)] =
sc.textFile("./EOADATA/country.csv").
filter(! _.startsWith("#")).
map {line =>
val row = line split ','
(100L + row(0).toInt, Country(row(1)))
}
```

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.

321. **In the metro example, what is an in-degree in relation to a country? Hint: this was covered in the Building a Degree Histogram Hands On exercise.**

A

A. A metro area or metropolis.
B. Another city.
C. A street in a city.
D. A continent.

322. **In the Hands On exercise on network connectedness and clustering, Antarctica was easy to identify. Why?** C

A. It had many edges
B. It had a vertex ID of 205.
C. It is the green dot that that has no connections, or it is the least connected cluster.

323. **In the Facebook graph example, the visualization looked like broccoli. Why?** B

A. In a directed graph, the stalks are large.
B. Social networks have communities or pockets of people who interact densely.
C. The high centrality of some people nodes in facebook gives the graph its broccoli shape.

324. **The ad-click events are listed in the file ad-clicks.csv. Each advertisement that is clicked on by a user generates $0.50 of revenue. What is the total amount of revenue generated by the ad-click events?** A

A. 8162
B. 5478
C. 112
D. 6844
E. None of the above

325. **How many different categories of advertisements are there?** A

A. 9
B. 8
C. 7
D. 6

326. **Let's say electronics generates $0.75, and the other types of advertisements generate $0.40. What is the total amount of revenue?** B

A. 6487
B. 6913
C. 3
D. 4009

327. **The file buy-clicks.csv lists the in-app purchases and the price of each purchase. When a user purchases an item, the company gets 2% of the price. How much revenue does the company make from the purchases in buy-clicks.csv?** A

A. 428
B. 824
C. 284

328. **How many distinct items can be purchased?** A

A. 6
B. 5
C. 4
D. 3

329. **How much does the most expensive item cost?** A

A. 20
B. 40
C. 60
D. 80

330. **What is the buyId of the item that is purchased the most?** A

A. 2
B. 3
C. 5
D. 9