Sample Exam

Edition 202310

Content

Introduction

This is the EXIN Data Analytics Foundation (DAF.EN) sample exam. The Rules and Regulations for EXIN's examinations apply to this exam.

This exam consists of 40 multiple-choice questions. Each multiple-choice question has a number of possible answers, of which only one is correct.

The maximum number of points that can be obtained for this exam is 40. Each correct answer is worth 1 point. You need 26 points or more to pass the exam.

The time allowed for this exam is 60 minutes.

Good luck!

Sample exam

1 / 40
Data analytics is the science of analyzing data to draw conclusions about it.

What does data analytics **not** focus on?

A) The conversion of data to information
B) The creation of data
C) The presentation of data
D) The processing of data

2 / 40
What is an example of a data cleaning step?

A) A company collects the revenue of all its branches and adjusts the monetary unit to "thousand dollars" before aggregating the datasets into its data warehouse.
B) A company collects the revenue of all its branches and builds a data model to predict the revenue of the coming year by geographic locations.
C) A company collects the revenue of all its branches and removes duplicate values before aggregating the datasets into its data warehouse.
D) A company collects the revenue of all its branches, finds the median value and identifies the branch whose revenue is furthest away from the median.

3 / 40
Compliance with legislation is a major area of concern for organizations to avoid risk.

What is **not** a major area of compliance with legislation regarding data?

A) Data analytics
B) Data ownership
C) Data privacy
D) Intellectual property

4 / 40
What is **not** a characteristic of interactive dashboards?

A) They allow users to drill down into data in detail.
B) They can combine data from multiple sources.
C) They may be difficult to use for non-technical users.
D) They provide what-if analysis for management.

5 / 40
An organization wants to collect text from a lot of different web pages.

What is the **best** way to source public data?

A) Via alternative data
B) Via internal acquisition systems
C) Via interviews and experiments
D) Via web scraping

6 / 40
What is a channel for collecting data?

A) Continuous data
B) Intent abstraction
C) Internal acquisition systems
D) Key-value store
E) Relational database management system (RDMS)
F) Structured data

7 / 40
What is an example of web scraping?

A) Downloading a CSV-file from a government website
B) Downloading data by parsing contents from a website
C) Downloading data through a file transfer protocol (FTP) connection
D) Downloading Excel files that are found on the internet

8 / 40
What is essential to comply with data laws when collecting data?

A) Completing transactions with open data
B) Ensuring transparency in collecting data
C) Mining minimal amounts of public data
D) Procuring of personal data

9 / 40
Which type of data is characterized by volume, velocity, and variety?

A) Alternative data
B) Big data
C) Structured data
D) Unstructured data

10 / 40
Which data solution has the explicit goal of post-analysis decision-making?

A) Enterprise data warehouse (EDW)
B) Key-value store
C) Relational database management system (RDMS)
D) Unstructured data systems

11 / 40
What is a difference between how an enterprise data warehouse (EDW) and a relational database management system (RDMS) are used?

A) An EDW is a solution for everyday operations and transactions, and an RDMS is accessed for pre-defined reports and dashboards.
B) An EDW uses Online Analytical Processing (OLAP) and an RDMS uses Online Transaction Processing (OLTP).
C) An RDMS is a tool that is applied to complex reports and analysis, and an EDW to real-time high-speed transactions.
D) An RDMS is based on a key-value store model and an EDW is based on an object-store model.

12 / 40
What is **not** a characteristic of distributed file systems?

A) Adding storage and servers at comparatively low costs
B) Processing chunks of data across servers simultaneously
C) Saving data easily as a blob without a schema definition
D) Storing data with no explicit use in its original format

13 / 40
A car lease company has a cloud solution for its core business in which they can make quotations, procure vehicles, organize maintenance, and do credit control.

The company also wants to use a cloud solution hosted by the same cloud provider for data analytics purposes.

What is **not** an advantage of the proposed data analytics solution?

A) The cloud solutions are cheaper than owning the environment of hardware, software, and IT-employees.
B) The import or export of massive data sets to different servers is unnecessary.
C) The solution can be scaled up in the same pace as the operations of the car lease company.
D) The solution can be scaled down when there is no longer a need for data analytics.

14 / 40
What is the relation between an independent variable and a dependent variable?

A) A dependent variable is changed by a data scientist and an independent variable is affected by this change.
B) A dependent variable's values reflect the effects that are observed and recorded on the change of the independent variable.
C) An independent variable is expressed as a lowercase 'y' and a dependent variable as an uppercase 'X'.
D) An independent variable should always be described as a categorical variable and a dependent variable can take any form.

15 / 40
Which type of variables is a categorical variable?

A) Boolean variables
B) Discrete variables
C) Nominal variables
D) Numerical variables

16 / 40
What type of variable is the mileage of a vehicle at the start and end of a year?

A) Boolean
B) Categorical
C) Numeric
D) TimeDate

17 / 40
Why is it important to distinguish between discrete and continuous variables?

A) Because it helps select the algorithms suitable for the variables
B) Because it is important to define the conceptual data model
C) Because only discrete variables can be aggregated with other variables
D) Because relational databases can only store continuous variables

18 / 40
What data scrubbing technique involves transforming variables into an integer format?

A) One-hot encoding
B) Variable merging
C) Variable selection
D) Web scraping

19 / 40
A car lease company's data analytics is based on a dataset with columns such as contract number, number plate, date of admission, brand, type, color, and end date of contract.

The columns 'date of admission' and 'end date of contract' are reduced to an integer of six positions, in which the first four correspond to the year and the last two to the month. This way, the contracts can be divided into categories containing all contracts ending in a certain month.

What is this an example of?

A) Binning
B) Merging variables
C) One-hot encoding
D) Variable selection

20 / 40
It is important to preserve the original source data used for data scrubbing to allow possible revisions.

To ensure this, what should be used?

A) A data retention policy
B) A data warehouse
C) A key-value store
D) Artificial Intelligence (AI) algorithms

21 / 40
What is an attribute of inferential methods?

A) They are favorable to situations where the data is well-documented and standardized into a single pool of information.
B) They help compress complex information into a convenient and easy-to-read format by summarizing obvious trends.
C) They involve isolating and analyzing a portion of the data and testing the results against another subset of the data.
D) They provide a convenient and concise summary of historical data generated from a potentially vast number of stand-alone events.

22 / 40
In what case should descriptive analysis be used?

A) When a data set cannot be analyzed easily
B) When limited information is available
C) When the retrieved data is very detailed
D) When there are gaps in record keeping

23 / 40
What is needed for a human operator to start data mining?

A) A defined hypothesis as “likely” and “unlikely”
B) A large number of possible input combinations
C) A model containing which inputs produce a given output
D) A problem that is deemed to be worth solving

24 / 40
What machine learning method uses training data and test data?

A) Data mining method
B) Descriptive method
C) Inferential method
D) Split validation method

25 / 40
A car lease company has a large dataset collected in its sales process. The company uses machine learning for fine tuning the quotations. Only the data from the sales process is fed to the software without any additional rules. The software then looks for patterns and will provide a model for calculating the quotation without giving too much discount.

What type of machine learning is this?

A) Reinforcement learning
B) Supervised learning
C) Unsupervised learning

26 / 40
An important aspect of natural language processing (NLP) is syntax analysis.

What does syntax analysis analyze?

A) Meaning of a sentence
B) Named entities
C) Search queries’ text
D) Structure of a sentence

27 / 40
A task in natural language processing (NLP) involves identifying important entities mentioned in text, such as name, location, or an activity.

Which NLP task is this?

A) Class identification
B) Named entity recognition
C) Sentiment analysis
D) Stemming

28 / 40
What is the association between data and algorithms?

A) Algorithms are designed to process and analyze data.
B) Data is designed for transition into algorithm knowledge.
C) Different people use the same algorithm when processing the same data.
D) Input data never changes for algorithms resulting in uniqueness.

29 / 40
Data from a real estate broker shows that the average price of houses sold decreases in accordance with the increase of mortgages interest rates.

What kind of regression analysis should be used to describe this pattern in the data?

A) Exponential regression analysis
B) Linear regression analysis
C) Logistic regression analysis
D) Non-linear regression analysis

30 / 40
What is the goal of regression analysis?

A) To analyze organizational performance that is used in decision-making
B) To categorize data points into groupings when no pre-defined classes exist
C) To find a line or a curve that describes patterns in the data in the best way
D) To understand how data is collected, organized, and interpreted

31 / 40
Which type of model generates category predictions and clustering?

A) Algorithmic
B) Classification
C) Regression

32 / 40
A company has a large dataset of shirts sales. The staff wants to better understand the figures and the data analyst is considering whether cluster analysis could offer any insight.

How does k-means clustering give a better understanding of the sales figures?

A) By clustering data in predefined groupings
B) By describing unexpected relationships between clusters
C) By explaining the role of centroids
D) By providing a new way of grouping customers

33 / 40
What is a **key** difference between association analysis and sequence mining?

A) Association analysis applies the generalized sequential patterns (GSP) technique, while sequence mining applies Apriori.
B) Association analysis disregards the order in which items appear, while sequence mining considers it.
C) Association analysis uses a confidence level to aid decision-making, while sequence mining is supervised learning.
D) Association analysis works better on large datasets, while sequence mining is best suited for small datasets.

34 / 40
Sequence mining uses a confidence level to inform decision-making. It also uses a kind of criterion to focus on frequent patterns to inform this decision-making.

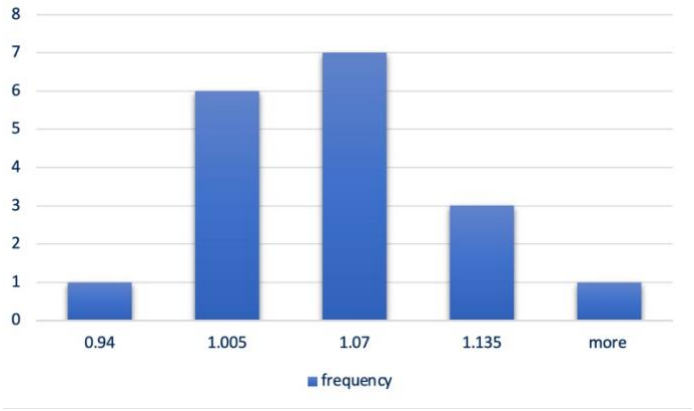
What criterion is this?

A) Apriori algorithm
B) Frequent itemset
C) Minimum support
D) Recursive elimination

35 / 40
Which graphic technique is typically used when delivering data to an external audience?

A) Expansionary
B) Explanatory
C) Explicatory
D) Exploratory

36 / 40
Please see the image below:



What is this an example of?

A) Bar chart
B) Box plot
C) Heatmap
D) Histogram

37 / 40
Which type of plot is used to display the spread and skewness of a dataset?

A) Box plot
B) Rug plot
C) Scatterplot
D) Violin plot

38 / 40
To get a better insight of the prices of houses sold in different quarters of a city, a real-estate broker wants to use a heatmap.

What does this heatmap show?

A) A plan of the city with levels of house prices represented in colors
B) A plan of the city with sold houses and the selling price per house
C) A table with average house prices in the past year per quarter sorted per average price
D) A table with prices of houses sold in the past year ordered by price

39 / 40
Why is aesthetic design important?

A) It enhances machine learning techniques.
B) It facilitates natural language processing (NLP).
C) It focuses on regression analysis models.
D) It improves the usability of data visualization.

40 / 40
What data visualization tool has the look and feel of Microsoft Excel?

A) Data Wrapper
B) Google Charts
C) Power BI
D) Tableau

Answer key

1 / 40
Data analytics is the science of analyzing data to draw conclusions about it.

What does data analytics **not** focus on?

A) The conversion of data to information
B) The creation of data
C) The presentation of data
D) The processing of data

A) Incorrect. This is the focus of the core activities of data analytics, which is about analyzing and processing data to make sense of it.
B) Correct. The creation of data precedes the need to make sense of it and is, therefore, not part of data analytics. (Literature: B, Chapters 2 and 3)
C) Incorrect. Data analytics focuses on the presentation of data to ensure that the results of the analysis are appropriate for the intended audience.
D) Incorrect. How data is processed and analyzed so that it makes sense is a central theme in data analytics.

2 / 40
What is an example of a data cleaning step?

A) A company collects the revenue of all its branches and adjusts the monetary unit to "thousand dollars" before aggregating the datasets into its data warehouse.
B) A company collects the revenue of all its branches and builds a data model to predict the revenue of the coming year by geographic locations.
C) A company collects the revenue of all its branches and removes duplicate values before aggregating the datasets into its data warehouse.
D) A company collects the revenue of all its branches, finds the median value and identifies the branch whose revenue is furthest away from the median.

A) Incorrect. This is not data cleaning, but data normalization/standardization.
B) Incorrect. This is not data cleaning, but predictive data analysis.
C) Correct. Removing duplicates is a typical data cleaning step to ensure data quality in terms of uniqueness. (Literature: B, Chapter 4)
D) Incorrect. This is not data cleaning, but data profiling, which refers to assessing data with statistical functions before applying machine learning techniques.

3 / 40
Compliance with legislation is a major area of concern for organizations to avoid risk.

What is **not** a major area of compliance with legislation regarding data?

A) Data analytics
B) Data ownership
C) Data privacy
D) Intellectual property

A) Correct. Legislation does not describe how data is analyzed, but rather which data can be collected and analyzed and for what purpose. (Literature: B, Chapter 7)
B) Incorrect. A growing number of countries currently have regulations that specify who owns the data collected and what are the owners' rights.
C) Incorrect. Data privacy regulations currently exist internationally and are an area of concern for organizations.
D) Incorrect. Intellectual property legislation is one of the first legislations that affects data analytics.

4 / 40
What is **not** a characteristic of interactive dashboards?

A) They allow users to drill down into data in detail.
B) They can combine data from multiple sources.
C) They may be difficult to use for non-technical users.
D) They provide what-if analysis for management.

A) Incorrect. Interactive dashboards allow users to drill down to more detailed data.
B) Incorrect. Interactive dashboards combine data from multiple sources to provide a comprehensive understanding.
C) Correct. Interactive dashboards are designed to be easy to use for non-technical users. (Literature: B, Chapter 5)
D) Incorrect. Interactive dashboards provide what-if insights to facilitate decision making.

5 / 40
An organization wants to collect text from a lot of different web pages.

What is the **best** way to source public data?

A) Via alternative data
B) Via internal acquisition systems
C) Via interviews and experiments
D) Via web scraping

A) Incorrect. Alternative data collates information from non-traditional sources, usually related to banking and finances. It is therefore not the best way.
B) Incorrect. Sourcing data from internal acquisition systems is the most common way to acquire internal data.
C) Incorrect. Interviews and experiments are not suitable for automation and code. Also, information retrieved from interviews is not public data.
D) Correct. Web scraping collects information from web using code and automation. (Literature: A, Chapter 1)

6 / 40
What is a channel for collecting data?

A) Continuous data
B) Intent abstraction
C) Internal acquisition systems
D) Key-value store
E) Relational database management system (RDMS)
F) Structured data

A) Incorrect. Continuous data is a type of variable, not a channel for collecting data.
B) Incorrect. Intent abstraction is a component of natural language processing (NLP), not a channel for collecting data.
C) Correct. Internal acquisition systems are a channel for collecting data. (Literature: A, Chapter 1)
D) Incorrect. Key-value store is a storage system for data, not a channel for collecting data.
E) Incorrect. RDMS is a software program used to store structured data. It is not a channel for collecting data.
F) Incorrect. Structured data is not a channel for collecting data. Instead, it is a format for storing data.

7 / 40
What is an example of web scraping?

A) Downloading a CSV-file from a government website
B) Downloading data by parsing contents from a website
C) Downloading data through a file transfer protocol (FTP) connection
D) Downloading Excel files that are found on the internet

A) Incorrect. Web scraping involves extracting valuable data by combining data, which is different from downloading a single file.
B) Correct. Web scraping involves searching through websites and extracting valuable data. (Literature: A, Chapter 1)
C) Incorrect. This does not include the web part of web scraping.
D) Incorrect. Although the files are found on the internet, scraping is not involved here.

8 / 40
What is essential to comply with data laws when collecting data?

A) Completing transactions with open data
B) Ensuring transparency in collecting data
C) Mining minimal amounts of public data
D) Procuring of personal data

A) Incorrect. The open data could contain personal information, in which case it would not necessarily comply with data laws.
B) Correct. Essential to reach compliance with contemporary data laws is transparency by sharing how data is collected. (Literature: A, Chapter 1)
C) Incorrect. Data laws do not mention that mining public information is essential for compliance.
D) Incorrect. Procuring personal data can only happen in accordance with the law. Hence, procuring it is not essential to reach compliance.

9 / 40
Which type of data is characterized by volume, velocity, and variety?

A) Alternative data
B) Big data
C) Structured data
D) Unstructured data

A) Incorrect. Alternative data is a channel of data collection.
B) Correct. Big data is characterized by volume, velocity, and variety. (Literature: A, Chapter 2)
C) Incorrect. Structured data describes information organized into a format that is well-defined and easy for algorithms to retrieve and process critical information.
D) Incorrect. Unstructured data consists of information that is not organized in a discernible way and does not fit into a standard structure like a tabular dataset.

10 / 40
Which data solution has the explicit goal of post-analysis decision-making?

A) Enterprise data warehouse (EDW)
B) Key-value store
C) Relational database management system (RDMS)
D) Unstructured data systems

A) Correct. An EDW has the explicit purpose of post-analysis for decision-making in organizations. (Literature: A, Chapter 2)
B) Incorrect. In a key-value store, data can be stored without a schema definition, as a blob. The result is that the value of the data is vague, and the result of a request cannot be controlled.
C) Incorrect. An RDMS aims at data storage. It is not used for post-analysis decision-making.
D) Incorrect. Unstructured data systems are used mainly for storage and processing of unstructured data. They are not used for post-analysis decision-making.

11 / 40
What is a difference between how an enterprise data warehouse (EDW) and a relational database management system (RDMS) are used?

A) An EDW is a solution for everyday operations and transactions, and an RDMS is accessed for pre-defined reports and dashboards.
B) An EDW uses Online Analytical Processing (OLAP) and an RDMS uses Online Transaction Processing (OLTP).
C) An RDMS is a tool that is applied to complex reports and analysis, and an EDW to real-time high-speed transactions.
D) An RDMS is based on a key-value store model and an EDW is based on an object-store model.

A) Incorrect. An EDW is not a solution for everyday operations because it needs to consolidate and integrate data from different sources. It focuses on analytical applications. An RDMS is a tool to manage transactional data.
B) Correct. An EDW is based on OLAP and an RDMS on OLTP. (Literature: A, Chapter 2)
C) Incorrect. An RDMS is used for real-time high-speed transactions and an EDW for complex reporting and analysis.
D) Incorrect. An RDMS is not based on a key-value store model. It is based on a relational model.

12 / 40
What is **not** a characteristic of distributed file systems?

A) Adding storage and servers at comparatively low costs
B) Processing chunks of data across servers simultaneously
C) Saving data easily as a blob without a schema definition
D) Storing data with no explicit use in its original format

A) Incorrect. The cost of incremental storage and servers is comparatively lower in distributed file systems.
B) Incorrect. With distributed file systems, chunks of data across servers can simultaneously be processed by splitting tasks across the network.
C) Correct. In a key-value store, data can be stored as a blob without a schema definition. This is not a utility of distributed file systems. (Literature: A, Chapter 2)
D) Incorrect. Distributed file systems allow data storage that does not have an explicit use in its original format.

13 / 40
A car lease company has a cloud solution for its core business in which they can make quotations, procure vehicles, organize maintenance, and do credit control.

The company also wants to use a cloud solution hosted by the same cloud provider for data analytics purposes.

What is **not** an advantage of the proposed data analytics solution?

A) The cloud solutions are cheaper than owning the environment of hardware, software, and IT-employees.
B) The import or export of massive data sets to different servers is unnecessary.
C) The solution can be scaled up in the same pace as the operations of the car lease company.
D) The solution can be scaled down when there is no longer a need for data analytics.

A) Correct. This cannot be said as a rule. The cloud solution may even be more expensive. (Literature: A, Chapter 2)
B) Incorrect. All data is already in the cloud ready to be used.
C) Incorrect. Cloud solutions, in general, can be scaled up without effort or investments of the customer.
D) Incorrect. Scaling up and scaling down are usually advantages of cloud solutions.

14 / 40
What is the relation between an independent variable and a dependent variable?

A) A dependent variable is changed by a data scientist and an independent variable is affected by this change.
B) A dependent variable's values reflect the effects that are observed and recorded on the change of the independent variable.
C) An independent variable is expressed as a lowercase 'y' and a dependent variable as an uppercase 'X'.
D) An independent variable should always be described as a categorical variable and a dependent variable can take any form.

A) Incorrect. The independent variable is the part that is modified. The dependent variable is affected by this change.
B) Correct. As the independent variable is modified, the effect is on the dependent variable. (Literature: A, Chapter 2)
C) Incorrect. The independent variable is expressed as 'X' in the equation.
D) Incorrect. An independent variable can also be described as numeric, Boolean and TimeDate, not only as a categorical variable.

15 / 40
Which type of variables is a categorical variable?

A) Boolean variables
B) Discrete variables
C) Nominal variables
D) Numerical variables

A) Incorrect. Boolean variables can either be true or false and are stored as 16-bit (2-byte) values. This is a different type of variable than categorical variables.
B) Incorrect. A discrete variable is a variable that takes on distinct, countable values. Categorical variables are uncountable.
C) Correct. A nominal variable is a type of categorical variable, which is a variable that can take on one of a limited, and usually fixed, number of possible values, that are qualitative in nature and cannot be aggregated. (Literature: A, Chapter 3)
D) Incorrect. Numerical variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Categorical variables are uncountable.

16 / 40
What type of variable is the mileage of a vehicle at the start and end of a year?

A) Boolean
B) Categorical
C) Numeric
D) TimeDate

A) Incorrect. The miles can be below or above the limit for the year, but the value itself is not Boolean. The value can be any number above 0.
B) Incorrect. The mileage variable can be compared with a categorical variable (the limits imposed by a certain type of lease contract), but it is not a categorical variable itself.
C) Correct. The variable can be expressed mathematically as an integer variable. (Literature: A, Chapter 3)
D) Incorrect. The miles are linked to an interval of time, but the number itself is not.

17 / 40
Why is it important to distinguish between discrete and continuous variables?

A) Because it helps select the algorithms suitable for the variables
B) Because it is important to define the conceptual data model
C) Because only discrete variables can be aggregated with other variables
D) Because relational databases can only store continuous variables

A) Correct. Recognizing whether a variable is discrete or continuous is important when analyzing data, as this will determine its compatibility with the selected algorithm. (Literature: A, Chapter 2)
B) Incorrect. Recognizing continuous or discrete variables does not interfere in the conceptual data model that captures higher abstraction contents (entities and relationships).
C) Incorrect. Discrete variables cannot be aggregated or mathematically manipulated with other variables.
D) Incorrect. Relational databases can store both discrete and continuous variables.

18 / 40
What data scrubbing technique involves transforming variables into an integer format?

A) One-hot encoding
B) Variable merging
C) Variable selection
D) Web scraping

A) Correct. The one-hot encoding technique involves transforming variables into an integer format for the algorithms to run successfully. (Literature: A, Chapter 4)
B) Incorrect. Merging variables involves combining related variables to preserve maximum information.
C) Incorrect. Variable selection involves selecting more valuable variables than other ones in a dataset.
D) Incorrect. Web scraping is not a technique for data scrubbing. It is a method of data collection.

19 / 40
A car lease company's data analytics is based on a dataset with columns such as contract number, number plate, date of admission, brand, type, color, and end date of contract.

The columns 'date of admission' and 'end date of contract' are reduced to an integer of six positions, in which the first four correspond to the year and the last two to the month. This way, the contracts can be divided into categories containing all contracts ending in a certain month.

What is this an example of?

A) Binning
B) Merging variables
C) One-hot encoding
D) Variable selection

A) Correct. Binning is converting numeric or timestamp values into a category using a discrete integer. (Literature: A, Chapter 4)
B) Incorrect. The two variables 'date of admission' and 'end date of contract' are not merged in any way.
C) Incorrect. One-hot encoding transforms discrete variables into a binary format. The resulting integers here can contain more than two values, which are not states.
D) Incorrect. Variable selection is about leaving out variables that give no insight. In the example nothing is left out.

20 / 40
It is important to preserve the original source data used for data scrubbing to allow possible revisions.

To ensure this, what should be used?

A) A data retention policy
B) A data warehouse
C) A key-value store
D) Artificial Intelligence (AI) algorithms

A) Correct. It is important to create and use a data retention policy to preserve the source data. (Literature: A, Chapter 4)
B) Incorrect. A data warehouse is a relational database that stores the data and does not ensure the preservation of the source data. Data governance and rules for data management should ensure this.
C) Incorrect. A key-value store is a database that stores data as a key-value pair.
D) Incorrect. AI algorithms do not ensure preserving the original source data. AI is applied to get insights, predictions, and inference.

21 / 40
What is an attribute of inferential methods?

A) They are favorable to situations where the data is well-documented and standardized into a single pool of information.
B) They help compress complex information into a convenient and easy-to-read format by summarizing obvious trends.
C) They involve isolating and analyzing a portion of the data and testing the results against another subset of the data.
D) They provide a convenient and concise summary of historical data generated from a potentially vast number of stand-alone events.

A) Incorrect. This refers to descriptive analysis, not to inferential methods.
B) Incorrect. This is a characteristic of descriptive analysis. Inferential methods are not used to compress information.
C) Correct. This is what split validation does, which is a technique used in inferential methods. (Literature: A, Chapter 5)
D) Incorrect. This refers to descriptive analysis. Inferential methods deal with the natural variance of relatively less concise data.

22 / 40
In what case should descriptive analysis be used?

A) When a data set cannot be analyzed easily
B) When limited information is available
C) When the retrieved data is very detailed
D) When there are gaps in record keeping

A) Incorrect. In this case, inferential statistical analysis should be used.
B) Incorrect. In this case, inferential statistical analysis should be used.
C) Correct. Descriptive analysis should be used when detailed data is available. (Literature: A, Chapter 5)
D) Incorrect. In this case, inferential statistical analysis should be used.

23 / 40
What is needed for a human operator to start data mining?

A) A defined hypothesis as “likely” and “unlikely”
B) A large number of possible input combinations
C) A model containing which inputs produce a given output
D) A problem that is deemed to be worth solving

A) Incorrect. An operator needs this for a standard statistical model.
B) Incorrect. An operator needs this for reinforcement learning.
C) Incorrect. An operator needs this for supervised learning.
D) Correct. An operator needs a goal, like a problem worth solving. (Literature: A, Chapter 5)

24 / 40
What machine learning method uses training data and test data?

A) Data mining method
B) Descriptive method
C) Inferential method
D) Split validation method

A) Incorrect. Data mining focuses on learning new things rather than understanding how what is learned can be applied in a specific context, as machine learning does.
B) Incorrect. Descriptive analysis favors instances where the data is well-documented and standardized into a single pool of information.
C) Correct. Inferential methods use 70 to 80% of a dataset to train an analytical agent to spot patterns and test them against the remaining 20 to 30%. (Literature: A, Chapter 5)
D) Incorrect. Inferential methods often use a technique called split validation, but split validation is not the entire method.

25 / 40
A car lease company has a large dataset collected in its sales process. The company uses machine learning for fine tuning the quotations. Only the data from the sales process is fed to the software without any additional rules. The software then looks for patterns and will provide a model for calculating the quotation without giving too much discount.

What type of machine learning is this?

A) Reinforcement learning
B) Supervised learning
C) Unsupervised learning

A) Incorrect. There is a large number of possible input combinations, but they are not randomly graded on their performance.
B) Incorrect. There is no specific or precise rule for calculating the correct price based on the independent input variables.
C) Correct. There are pre-labelled input and output combinations, but there are no known outputs to act as reference point. Unsupervised learning analyzes input values to find patterns in how the price can be set for the best possible performance. (Literature: A, Chapter 5)

26 / 40
An important aspect of natural language processing (NLP) is syntax analysis.

What does syntax analysis analyze?

A) Meaning of a sentence
B) Named entities
C) Search queries' text
D) Structure of a sentence

A) Incorrect. This is called semantic analysis.
B) Incorrect. This is called classification.
C) Incorrect. This is called text parsing.
D) Correct. The analysis of the structure of a sentence is called syntax analysis. (Literature: A, Chapter 10)

27 / 40
A task in natural language processing (NLP) involves identifying important entities mentioned in text, such as name, location, or an activity.

Which NLP task is this?

A) Class identification
B) Named entity recognition
C) Sentiment analysis
D) Stemming

A) Incorrect. Class identification labels a document with binary vectors of given tokens.
B) Correct. Named entity recognition picks out salient parts of text such as what, where and how. (Literature: A, Chapter 10)
C) Incorrect. Sentiment analysis is used for detecting the emotional intent of a document.
D) Incorrect. Stemming is used to parse a document's keywords into core meanings.

28 / 40
What is the association between data and algorithms?

A) Algorithms are designed to process and analyze data.
B) Data is designed for transition into algorithm knowledge.
C) Different people use the same algorithm when processing the same data.
D) Input data never changes for algorithms resulting in uniqueness.

A) Correct. Algorithm is a sequence of steps that react to cues and changing patterns that come from data. (Literature: A, Chapter 6)
B) Incorrect. Data has no design related to algorithm knowledge.
C) Incorrect. People use their own internal algorithm for evaluation of data.
D) Incorrect. Data always changes and is not unique for algorithms.

29 / 40
Data from a real estate broker shows that the average price of houses sold decreases in accordance with the increase of mortgages interest rates.

What kind of regression analysis should be used to describe this pattern in the data?

A) Exponential regression analysis
B) Linear regression analysis
C) Logistic regression analysis
D) Non-linear regression analysis

A) Incorrect. Exponential regression analysis is not suitable for linear relationships between variables.
B) Correct. The regression is linear and negative. (Literature: A, Chapter 7)
C) Incorrect. Logistic regression yields discrete variables as outputs.
D) Incorrect. The regression is linear, so a linear regression analysis is best suited in this case.

30 / 40
What is the goal of regression analysis?

A) To analyze organizational performance that is used in decision-making
B) To categorize data points into groupings when no pre-defined classes exist
C) To find a line or a curve that describes patterns in the data in the best way
D) To understand how data is collected, organized, and interpreted

A) Incorrect. This refers to business intelligence (BI), which can be defined as a set of tools for gathering, analyzing, and reporting information to decision-makers regarding the organization's performance.
B) Incorrect. Although logistic regression can be used to identify records with similar or close values, this cannot be said for all regression methods.
C) Correct. This is the objective of regression analysis. Although a single line or curve may oversimplify the data, it provides a useful reference point for making general predictions about future data. (Literature: A, Chapter 7)
D) Incorrect. This is related to statistics, which has the primary goal of determining the meaning of the data and its variations.

31 / 40
Which type of model generates category predictions and clustering?

A) Algorithmic
B) Classification
C) Regression

A) Incorrect. Algorithms are used in all data modeling techniques and are not a model as such.
B) Correct. Classification is the umbrella term for algorithms that generate category predictions and cluster analysis for topics as market research and customer profiling. (Literature: A, Chapter 8)
C) Incorrect. Regression analysis is a popular statistical technique used to model the relationship between one or more independent variables and the dependent variable.

32 / 40
A company has a large dataset of shirts sales. The staff wants to better understand the figures and the data analyst is considering whether cluster analysis could offer any insight.

How does k-means clustering give a better understanding of the sales figures?

A) By clustering data in predefined groupings
B) By describing unexpected relationships between clusters
C) By explaining the role of centroids
D) By providing a new way of grouping customers

A) Incorrect. Using k-means clustering, it is possible to find unidentified groupings.
B) Incorrect. Using k-means clustering, it is possible to find groups separated as far as possible. However, it is not possible to analyze the relations between the groups.
C) Incorrect. Centroids are a technical solution for making groups, but they do not play a role with a meaning.
D) Correct. K-means clustering is useful for cases where there is no knowledge of an existing category but a wish to find new unidentified groupings. It can provide a new insight using these groupings. (Literature: A, Chapter 8)

33 / 40
What is a **key** difference between association analysis and sequence mining?

A) Association analysis applies the generalized sequential patterns (GSP) technique, while sequence mining applies Apriori.
B) Association analysis disregards the order in which items appear, while sequence mining considers it.
C) Association analysis uses a confidence level to aid decision-making, while sequence mining is supervised learning.
D) Association analysis works better on large datasets, while sequence mining is best suited for small datasets.

A) Incorrect. Sequence mining uses GSP, whereas association analysis uses Apriori.
B) Correct. Association analysis does not consider the order of the events, which is important for sequence mining. (Literature: A, Chapter 9)
C) Incorrect. Sequence mining is not supervised learning, but a type of association analysis.
D) Incorrect. Association analysis and sequence mining do not differ in size of datasets.

34 / 40
Sequence mining uses a confidence level to inform decision-making. It also uses a kind of criterion to focus on frequent patterns to inform this decision-making.

What criterion is this?

A) Apriori algorithm
B) Frequent itemset
C) Minimum support
D) Recursive elimination

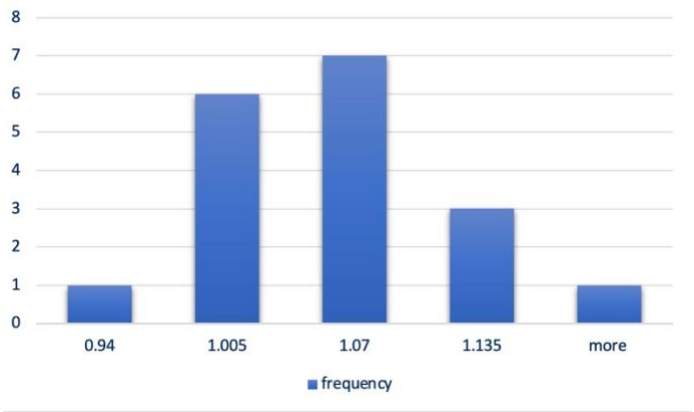
A) Incorrect. Apriori is an algorithm for frequent itemset mining and association rule learning over relational databases.
B) Incorrect. A frequent itemset appears in at least minimum support transactions from the transaction database, where minimum support is a threshold set by the user.
C) Correct. Like association analysis, sequence mining uses minimum support criteria to focus on frequent patterns as well as a confidence level to inform decision-making and prioritize implementation. (Literature: A, Chapter 10)
D) Incorrect. Recursive elimination is an algorithm for discovering frequent item sets in a transaction database.

35 / 40
Which graphic technique is typically used when delivering data to an external audience?

A) Expansionary
B) Explanatory
C) Explicatory
D) Exploratory

A) Incorrect. Expansionary is usually used in combination with the word 'policy'. It is a concept in economics relating to techniques used to prevent recession and unemployment.
B) Correct. Explanatory graphics aim to explain data and findings more simply and help the audience better understand complex data. (Literature: A, Chapter 11)
C) Incorrect. Explicatory is a philosophical and literary technique not used in data analytics.
D) Incorrect. Exploratory graphics are generated on-the-fly to aid internal understanding while analysis is in progress and the model is still in production mode.

36 / 40
Please see the image below:



What is this an example of?

A) Bar chart
B) Box plot
C) Heatmap
D) Histogram

A) Correct. A bar chart is the graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent. (Literature: A, Chapter 12)
B) Incorrect. A box plot displays the five-number summary of a set of data.
C) Incorrect. A heatmap is a two-dimensional representation of data in which values are represented by colors.
D) Incorrect. A histogram is the graphical representation of data grouped into continuous number ranges which correspond to vertical bars. A histogram does not have blank spaces between the bars.

37 / 40
Which type of plot is used to display the spread and skewness of a dataset?

A) Box plot
B) Rug plot
C) Scatterplot
D) Violin plot

A) Correct. The box plot describes the symmetry of the data and is used for summarizing and displaying the distribution of a set of continuous data. (Literature: A, Chapter 11)
B) Incorrect. The rug plot visualizes the distribution of a single variable, not the skewness of a dataset.
C) Incorrect. The scatterplot conveys information about the relationship between continuous variables, not a dataset's distribution and skewness.
D) Incorrect. The violin plot describes the distribution of the data and its density, but not the skewness.

38 / 40
To get a better insight of the prices of houses sold in different quarters of a city, a real-estate broker wants to use a heatmap.

What does this heatmap show?

A) A plan of the city with levels of house prices represented in colors
B) A plan of the city with sold houses and the selling price per house
C) A table with average house prices in the past year per quarter sorted per average price
D) A table with prices of houses sold in the past year ordered by price

A) Correct. A heatmap shows data in colors. (Literature: A, Chapter 12)
B) Incorrect. A heatmap shows data in colors. This plan does not indicate that any color is used.
C) Incorrect. A heatmap shows data in colors. This table, which is not a map, does not indicate that any color is used.
D) Incorrect. A heatmap shows data in colors. This table, which is not a map, does not indicate that any color is used. Ordering prices is also not necessary for a heatmap.

39 / 40
Why is aesthetic design important?

A) It enhances machine learning techniques.
B) It facilitates natural language processing (NLP).
C) It focuses on regression analysis models.
D) It improves the usability of data visualization.

A) Incorrect. Machine learning does not rely on data visualization. It is based on datasets.
B) Incorrect. NLP is related to language, syntax, and semantics. It does not rely on data visualization.
C) Incorrect. Aesthetic design improves all forms of data visualization, it does not focus on regression analysis.
D) Correct. Aesthetic design principles include careful design and color combinations, which improve the usability of data visualization. (Literature: A, Chapter 11)

40 / 40
What data visualization tool has the look and feel of Microsoft Excel?

A) Data Wrapper
B) Google Charts
C) Power BI
D) Tableau

A) Incorrect. Data Wrapper has an interface of its own.
B) Incorrect. Google Charts is compatible with Google products.
C) Correct. Power BI is compatible with other Microsoft products. (Literature: A, Chapter 12)
D) Incorrect. Tableau has an interface of its own.

Evaluation

The table below shows the correct answers to the questions in this sample exam.

Question	Answer	Question	Answer
1	B	**21**	C
2	C	**22**	C
3	A	**23**	D
4	C	**24**	C
5	D	**25**	C
6	C	**26**	D
7	B	**27**	B
8	B	**28**	A
9	B	**29**	B
10	A	**30**	C
11	B	**31**	B
12	C	**32**	D
13	A	**33**	B
14	B	**34**	C
15	C	**35**	B
16	C	**36**	A
17	A	**37**	A
18	A	**38**	A
19	A	**39**	D
20	A	**40**	C

Contact EXIN

www.exin.com