Question ID:	M01-Q03
Topic:	M1a – Data Visualization
Difficulty:	Medium
Blooms:	Apply

What role does visualization play in Data Science?

1. Helps humans understand patterns that exist in large datasets
2. It eliminates the need for data cleaning
3. It guarantees conclusions are correct
4. It replaces statistics entirely
5. It is only useful for small datasets

Question ID:	M01-Q04
Topic:	M1a – Statistics: Nonsense Protection
Difficulty:	Medium
Blooms:	Analyze

What does statistics help prevent in analysis?

1. False conclusions caused by random patterns
2. Collecting too much data
3. Good visualizations
4. The need for computers
5. Faster machine learning training

Question ID:	M01-Q08
Topic:	M1c – Arrays
Difficulty:	Medium
Blooms:	Understand

Which statement about arrays is TRUE?

1. Easy to insert in the middle
2. Always sorted
3. Fast access by index
4. Automatically distributed across different computers
5. Removing an element is always O(1)

Question ID:	M01-Q14
Topic:	M1c – Broadcasting
Difficulty:	Medium
Blooms:	Apply

In NumPy, what is "broadcasting"?

1. Automatically expanding arrays to compatible shapes during arithmetic
2. Sending arrays to a TV channel
3. Only used when adding identical shapes
4. A memory compression method
5. Copying data into SQL databases

Question ID:	M01-Q16
Topic:	M1c — Array memory organization
Difficulty:	Medium
Blooms:	Understand

One major reason arrays allow fast random access is:

1. The array keeps a hash table of all indexes
2. Each element stores the address of the next
3. Elements are stored contiguously so index lookup is constant time
4. The OS automatically accelerates access for arrays
5. Arrays are always stored in CPU cache

Question ID:	M01-Q24
Topic:	M1c — Hash table disadvantages
Difficulty:	Medium
Blooms:	Analyze

Which is a key drawback of hash tables?

1. They require keys to be sorted
2. They cannot delete items
3. They do not preserve meaningful order
4. Searching is always slower than a linked list
5. They can only store fixed-size elements

Question ID:	M01-Q25
Topic:	M1c — Broadcasting rules
Difficulty:	Medium
Blooms:	Apply

In NumPy, why does adding a 1×3 vector to a 3×3 matrix work?

1. The vector is broadcast across rows to match shape
2. The matrix is flattened automatically
3. NumPy guesses the user's intention
4. The vector overwrites the diagonal
5. Because 3 is a special broadcast-safe number

Question ID:	M03-Q1
Topic:	Exploratory Data Analysis
Difficulty:	Easy
Blooms:	Understand

Which part of the data science workflow primarily focuses on understanding the structure, patterns, and anomalies present in the data?

1. Data Collection
2. Exploratory Data Analysis (EDA)
3. Confirmatory Data Analysis
4. Feature Deployment
5. GPU Optimization

Question ID:	M03-Q4
Topic:	Nominal vs Nominal
Difficulty:	Medium
Blooms:	Apply

A bar chart comparing 'Plant Type' and 'Fruit Variety' represents what type of data?

1. Ordinal vs Quantitative
2. Nominal vs Nominal
3. Ordinal vs Ordinal
4. Quantitative vs Quantitative
5. Geospatial vs Quantitative

Question ID:	M03-Q5
Topic:	Scatter Plot
Difficulty:	Easy
Blooms:	Remember

Scatter plots are primarily used to visualize:

1. Relationships between two quantitative variables
2. Distribution of a single variable only
3. Hierarchical relationships
4. Statistical inference and p-values
5. Survey proportions

Question ID:	M03-Q8
Topic:	Matplotlib API
Difficulty:	Medium
Blooms:	Understand

In matplotlib, why is using ‘fig, ax = plt.subplots()‘ preferred over calling ‘plt.plot()‘ directly?

1. It uses more memory so it’s faster
2. It gives explicit references to figure and axes objects for better control
3. It enables automatic machine learning integration
4. It prevents adding labels and legends
5. It is required by NumPy

Question ID:	M03-Q9
Topic:	KDE
Difficulty:	Medium
Blooms:	Analyze

A KDE (Kernel Density Estimate) is used to:

1. Visualize category labels
2. Smooth the distribution of sampled data
3. Display geospatial movement
4. Simulate stock trading
5. Normalize missing values

Question ID:	M03-Q10
Topic:	Outliers
Difficulty:	Medium
Blooms:	Evaluate

The lecture suggests that before trusting an outlier, you should:

1. Remove all outliers automatically
2. Trace back to the original data and verify context
3. Replace with the mean
4. Convert to categorical encoding
5. Report it as a major scientific discovery

Question ID:	M03-Q15
Topic:	3-variable visualization
Difficulty:	Medium
Blooms:	Apply

Which chart type allows three variables to be shown simultaneously using two axes and color/size encoding?

1. Bar chart with sublevels
2. Scatter plot with a third encoding
3. Pie chart slices only
4. Table layout
5. Q-Q plot

Question ID:	M03-Q18
Topic:	Pairwise Relationships
Difficulty:	Medium
Blooms:	Understand

A matrix of scatter plots helps you:

1. Visualize pairwise relationships among many quantitative variables
2. Render 3D surfaces of a function
3. Encode hierarchical trees
4. Compute p-values for regression
5. Normalize geospatial coordinates

Question ID:	M03-Q19
Topic:	Correlation Heat Map
Difficulty:	Easy
Blooms:	Understand

What does a correlation heat map display?

1. The sign and magnitude of linear relationships between variables
2. Raw counts of categories
3. Geographic elevation
4. Financial OHLC patterns
5. Kernel bandwidths

Question ID:	M06-Q1
Topic:	ML Process: Problem Framing
Difficulty:	Medium
Blooms:	Analyze

Which step most directly ensures that the problem you're solving is actually a machine-learning problem and not a simple rules or query task?

1. Check whether a labeled target exists and if patterns must generalize to new data
2. Visualize predictions with a confusion matrix
3. Collect more data first
4. Tune hyperparameters with cross-validation
5. Train a baseline linear model

Question ID:	M06-Q2
Topic:	ML Process: Data Splits
Difficulty:	Easy
Blooms:	Understand

In a standard ML workflow, a hold-out test set is primarily used to...

1. Select the best model during tuning
2. Fit the feature scaler and imputer
3. Estimate the final generalization performance after all modeling choices are frozen
4. Balance class labels in the training data
5. Visualize learning curves

Question ID:	M06-Q3
Topic:	ML Process: Leakage Prevention
Difficulty:	Medium
Blooms:	Apply

Which practice best prevents **data leakage** when scaling features?

1. Fit the scaler only on the training data, then apply the fitted transform to validation/test
2. Use MinMax scaling instead of standardization
3. Use more features so leakage averages out
4. Shuffle the rows before scaling
5. Fit the scaler on all available data, then transform splits

Question ID:	M06-Q4
Topic:	ML Process: Baselines
Difficulty:	Easy
Blooms:	Understand

A **baseline** model in the ML process is best described as. . .

1. A model with zero variance predictions
2. The final, most complex model
3. A simple, often naive model used to set a minimum performance bar
4. Any linear model
5. A model trained without regularization

Question ID:	M06-Q10
Topic:	ML Process: Metric Fit
Difficulty:	Medium
Blooms:	Analyze

A confusion matrix is **not** the right tool to evaluate a model when. . .

1. You need class-wise errors
2. You want false positive rate
3. You must compare ranking quality at varying thresholds
4. You want recall per class
5. You are solving classification

Question ID:	M06-Q11
Topic:	Regression: OLS Objective
Difficulty:	Easy
Blooms:	Remember

In simple linear regression, the **least squares** estimator chooses coefficients that...

1. Equalize residuals across x
2. Minimize mean absolute error of residuals
3. Maximize R^2 directly
4. Minimize classification error
5. Minimize the sum of squared residuals

Question ID:	M06-Q14
Topic:	Regression: Overfitting
Difficulty:	Easy
Blooms:	Understand

You add a perfectly predictive but noisy feature. Training MSE drops sharply; validation MSE rises. The model most likely...

1. Underfit
2. Overfit due to high variance
3. Suffers from label leakage
4. Is unbiased
5. Has perfect calibration

Question ID:	M06-Q18
Topic:	Regression: Model Complexity
Difficulty:	Easy
Blooms:	Understand

Polynomial regression of degree 10 on small n most likely increases...

1. Sample size
2. Linearity
3. Bias
4. Variance
5. Homoskedasticity

Question ID:	M07-Q1
Topic:	Classification overview
Difficulty:	Easy
Blooms:	Understand

In supervised classification, what is the primary goal when some rows have known labels?

1. To delete rows without labels so training is cleaner
2. To reduce dimensionality for visualization only
3. To generate synthetic labels for all rows from noise
4. To label previously unlabeled rows using a learned model
5. To cluster the data into groups without labels

Question ID:	M07-Q2
Topic:	Applications of classification
Difficulty:	Easy
Blooms:	Remember

Which example task is most clearly a multi-class classification problem?

1. Predicting house prices in dollars
2. Grouping unlabeled news articles by topic
3. Estimating the mean of a Gaussian distribution
4. Finding a low-dimensional embedding of images
5. Assigning each handwritten digit image a class from 0–9

Question ID:	M07-Q4
Topic:	KNN and effect of k
Difficulty:	Medium
Blooms:	Understand

For k-Nearest Neighbors (kNN), increasing k typically has what effect on the decision boundary?

1. Has no effect on the boundary; only runtime changes
2. Converts the classifier into a linear separator
3. Smooths the boundary by averaging over more neighbors
4. Forces the classifier to overfit training data
5. Makes the boundary more jagged and sensitive to noise

Question ID:	M07-Q6
Topic:	Cross-validation
Difficulty:	Easy
Blooms:	Understand

Cross-validation is introduced primarily to combat which issue?

1. Overfitting and unreliable estimates from a single split
2. Class imbalance exclusively
3. Label noise only
4. GPU memory limits
5. Too many features

Question ID:	M07-Q7
Topic:	1-D decision threshold
Difficulty:	Medium
Blooms:	Apply

In one-dimensional two-class separation, the 'overlap region' between class distributions most directly corresponds to:

1. Training time
2. Regularization strength
3. Bayesian prior probability
4. Model bias
5. Classification error rate

Question ID:	M07-Q9
Topic:	LDA intuition
Difficulty:	Medium
Blooms:	Analyze

Linear Discriminant Analysis (LDA) can be viewed as finding:

1. A nonlinear kernel mapping to infinite dimensions
2. A decision tree that partitions space by axis-aligned cuts
3. A clustering of unlabeled data by k-means
4. A random forest that averages many trees
5. A projection that maximizes between-class separation relative to within-class scatter

Question ID:	M07-Q10
Topic:	LDA pros/cons
Difficulty:	Easy
Blooms:	Remember

Which is listed as a *pro* of LDA in the deck?

1. Nonlinear boundaries by default
2. Requires deep networks to perform well
3. Often overfits small datasets
4. Usually doesn't overfit and works with much less data
5. Slow classification at inference

Question ID:	M07-Q11
Topic:	Naïve Bayes assumptions
Difficulty:	Easy
Blooms:	Understand

Why is Naïve Bayes called 'naïve' in these slides?

1. It assumes features are independent given the class
2. It assumes k is chosen by cross-validation
3. It assumes infinite training data
4. It assumes a linear decision boundary
5. It assumes labels are uniformly distributed

Question ID:	M07-Q12
Topic:	Logistic regression
Difficulty:	Easy
Blooms:	Remember

Logistic regression models the log-odds (logit) as:

1. A sum of kernel functions over support vectors
2. A linear function of input features ($\mathrm{w}^T\mathrm{x}$)
3. A decision tree depth
4. A constant independent of input features
5. The reciprocal of a Gaussian density

Question ID:	M07-Q13
Topic:	Sigmoid link
Difficulty:	Easy
Blooms:	Remember

In the logistic regression derivation provided, which activation function maps $z=w^Tx$ to a probability p?

1. σ(z)=eˆz/(1+eˆz)
2. Hard threshold at z=0
3. tanh(z)
4. Softplus(z)=log(1+eˆz)
5. ReLU(z)=max(0,z)

Question ID:	M07-Q15
Topic:	Baselines to try first
Difficulty:	Medium
Blooms:	Remember

Which of the following is *not* one of the four 'brain-dead' baseline algorithms the slides recommend trying first?

1. Logistic Regression
2. Support Vector Machines with RBF kernel
3. Naïve Bayes
4. Linear Discriminant Analysis
5. k-Nearest Neighbors

Question ID:	M07-Q19
Topic:	Feature importance in logistic regression
Difficulty:	Medium
Blooms:	Apply

According to the M07B slides, what is a *critical* step before using logistic regression weights as feature importance?

1. Apply PCA to all features
2. Normalize or standardize features to comparable scales
3. Remove the intercept term
4. Use L1 regularization
5. Convert all features to binary indicators

Question ID:	M07-Q25
Topic:	Motivation for XAI
Difficulty:	Easy
Blooms:	Understand

Which of the following best captures the motivation slide 'Why do we need explainable models?'

1. To increase training speed only
2. To support trust, debugging, fairness, and accountability in decisions (e.g., 'Why didn't I get the loan?')
3. To reduce the need for labeled data
4. Because explainability always improves accuracy
5. So we never need test sets again

Question ID:	M07-Q30
Topic:	Fairness notions (context)
Difficulty:	Medium
Blooms:	Evaluate

Which statement about fairness and explainability is consistent with the surrounding course materials?

1. Choosing one fairness metric can create trade-offs with others
2. Fairness is solved by increasing model capacity
3. All statistical definitions of fairness are mutually compatible
4. Only demographic parity matters in practice
5. Explainability eliminates the need for fairness analysis

Question ID:	MT-Q1
Topic:	None
Difficulty:	Easy
Blooms:	N/A

Which of the following statements about Python dictionaries is TRUE?

1. Dictionaries preserve the order of key insertion in all Python versions
2. Keys in a dictionary must be immutable objects
3. Duplicate keys are allowed and the values are stored in a list
4. Dictionary keys and values must be of the same type

Question ID:	MT-Q3
Topic:	None
Difficulty:	Medium
Blooms:	N/A

Which of the following statements about the Central Limit Theorem is FALSE?

1. It applies to the distribution of sample means
2. It requires that the population distribution be normal
3. It becomes more accurate as sample size increases
4. It implies that the sampling distribution approaches a normal distribution

Question ID:	MT-Q6
Topic:	None
Difficulty:	Medium
Blooms:	N/A

Which of the following statements about hypothesis testing is TRUE?

1. A small p-value proves that the null hypothesis is false
2. Failing to reject the null hypothesis means it is true
3. A p-value is the probability, under the null hypothesis, of obtaining a result as extreme as the one observed
4. The null hypothesis always states that there is a difference

Question ID:	MT-Q7
Topic:	None
Difficulty:	Easy
Blooms:	N/A

Which of the following statements about cross-validation is TRUE?

1. In k-fold cross-validation, the value of k must equal the number of observations
2. Cross-validation provides a way to estimate model performance on unseen data
3. Cross-validation is only applicable to classification problems
4. Cross-validation always reduces overfitting

Question ID:	M04-Q1
Topic:	Story vs. Model
Difficulty:	Easy
Blooms:	Remember

In this course, a 'model' primarily serves to...

1. Guarantee that data will fit a normal distribution
2. Tell a compelling narrative that motivates the audience
3. Eliminate randomness from the world entirely
4. Automatically generate high-quality plots without code
5. Provide a mathematical abstraction that summarizes patterns and supports prediction

Question ID:	M04-Q4
Topic:	Density vs Distribution
Difficulty:	Hard
Blooms:	Analyze

For a continuous distribution with density p(x), a core property is that:

1. $\sum$ p(x) over the data equals 1
2. p(x) must always be less than 1
3. The derivative of p(x) is constant
4. p(x) must be symmetric around the mean
5. $\int$ p(x) dx over the support equals 1

Question ID:	M04-Q5
Topic:	Discrete vs Continuous
Difficulty:	Easy
Blooms:	Evaluate

In the slides, which notation pairing is used (abuse of notation aside)?

1. Discrete variables must be converted to z-scores
2. Only continuous variables can be modeled
3. Both must be written as integrals only
4. P(X) for discrete distributions, p(X) for continuous densities
5. p(X) for discrete, P(X) for continuous

Question ID:	M04-Q6
Topic:	Joint vs Conditional
Difficulty:	Medium
Blooms:	Create

Which expression corresponds to the conditional probability as defined in the slides?

1. P(C—X) = P(C) · P(X)
2. P(X—C) = P(C,X) / P(C)
3. P(C—X) = P(C,X) / P(X)
4. P(C—X) = P(C) + P(X)
5. P(C—X) = 1 - P(C,X)

Question ID:	M04-Q8
Topic:	MAP Estimate
Difficulty:	Hard
Blooms:	Understand

The Maximum A Posteriori (MAP) estimate selects:

1. Any class with probability above 0.1
2. The class C that maximizes P(C—X)
3. The class that maximizes P(X—C) regardless of priors
4. The class with the smallest label alphabetically
5. The parameter that minimizes variance only

Question ID:	M04-Q10
Topic:	MLE for Normal
Difficulty:	Medium
Blooms:	Analyze

For a Normal distribution fitted by Maximum Likelihood to i.i.d. data, the parameter estimates are:

1. Any pair that minimizes the max error
2. Both parameters are always zero
3. Sample median for μ and IQR for σ
4. Sample mean for μ and sample standard deviation for σ
5. Skewness for μ and kurtosis for σ

Question ID:	M04-Q11
Topic:	Frequentist simulation
Difficulty:	Medium
Blooms:	Evaluate

Observing 27 heads in 40 fair-coin flips is evaluated in lecture by:

1. Replacing missing flips with the mean
2. Assuming the coin is biased without testing
3. Comparing the observed statistic to a simulated null distribution
4. Counting only the last five flips
5. Using a geospatial plot of flips

Question ID:	M04-Q12
Topic:	Permutation Test
Difficulty:	Hard
Blooms:	Create

Permutation tests (label shuffling) assess:

1. How to order categories alphabetically
2. Whether variables are exactly deterministic
3. Which distribution family is true a priori
4. Whether the observed group difference could arise by chance under the null
5. The best color map for plots

Question ID:	M04-Q13
Topic:	Bootstrapping
Difficulty:	Easy
Blooms:	Remember

Bootstrapping, as presented, is primarily used to:

1. Approximate the sampling distribution of a statistic by resampling with replacement
2. Force data to follow a Normal distribution
3. Guarantee smaller variance than the original sample
4. Encrypt sensitive variables
5. Eliminate outliers by deletion

Question ID:	M04-Q14
Topic:	EDA
Difficulty:	Medium
Blooms:	Understand

A central purpose of EDA is to:

1. Guarantee future market trends
2. Reveal structure, anomalies, and relationships in raw data before modeling
3. Replace statistics with visuals
4. Finalize the causal model and deploy
5. Remove all missing data automatically

Question ID:	M04-Q16
Topic:	Narrative Components
Difficulty:	Hard
Blooms:	Analyze

Which is one of the narrative components highlighted for data stories?

1. Kernel bandwidth
2. Hash seed
3. ASCII table index
4. Conflict
5. Recursion depth

Question ID:	M04-Q21
Topic:	Prior
Difficulty:	Easy
Blooms:	Apply

In Bayes' theorem, the prior P(C) represents:

1. Belief about classes before seeing current data
2. Noise added to stabilize training
3. A probability that must always be 0.5
4. The maximum of the likelihood
5. A resampling weight in bootstrapping

Question ID:	M04-Q22
Topic:	Evidence
Difficulty:	Medium
Blooms:	Analyze

In the factorization P(C—X) = P(X—C)P(C)/P(X), P(X) (the 'evidence') mainly serves to:

1. Set the sampling rate for bootstrapping
2. Choose color bars for heatmaps
3. Encode class labels as integers
4. Maximize the likelihood for each class
5. Normalize the posterior over classes so probabilities sum to 1

Question ID:	M04-Q25
Topic:	Discrete Distributions
Difficulty:	Easy
Blooms:	Remember

Which is an example of a discrete distribution family discussed in the slides?

1. Fourier
2. Cauchy (continuous heavy-tailed)
3. Binomial
4. Beta-prime (advanced)
5. Gaussian mixture with infinite components

Question ID:	M04-Q31
Topic:	Analytics Types
Difficulty:	Medium
Blooms:	Remember

Diagnostic analytics asks primarily:

1. Why did this happen?
2. What should we do next?
3. Who is to blame regardless of data?
4. How to export to PDF?
5. What color should bars be?

Question ID:	M04-Q32
Topic:	Analytics Types
Difficulty:	Hard
Blooms:	Understand

Predictive analytics, per the slide taxonomy, addresses:

1. What happened, historically?
2. What may happen in the future?
3. Where to place legends
4. Which category is funniest?
5. How to set DPI for figures

Question ID:	M04-Q37
Topic:	Simulation Precision
Difficulty:	Easy
Blooms:	Remember

Repeating simulations many times primarily helps to:

1. Guarantee significance
2. Avoid computing summaries
3. Quantify variability of estimates (e.g., Monte Carlo error)
4. Reduce the true variance in the population
5. Change the historical data

Question ID:	M04-Q40
Topic:	Missing Data
Difficulty:	Hard
Blooms:	Analyze

Which Python package was explicitly recommended to visualize missingness patterns?

1. pytorch.missviz
2. matplotlib.auto_na
3. missingno
4. seaborn.naheat
5. cv2.missing

Question ID:	M04-Q44
Topic:	Discrete vs Continuous
Difficulty:	Hard
Blooms:	Understand

Which statement is TRUE regarding discrete vs. continuous modeling per lecture?

1. Discrete probabilities sum to 1; continuous densities integrate to 1
2. Continuous probabilities sum to 1 over observed bins only
3. Neither requires normalization
4. Discrete distributions cannot have expectations
5. Both integrate to 1 only

Question ID:	M04-Q45
Topic:	MAP Decision
Difficulty:	Easy
Blooms:	Apply

The slides define the 'best guess' class given X as:

1. argmax^i P(C^i—X)
2. Choose the smallest index i
3. argmin^i P(C^i—X)
4. Random selection
5. argmax^i P(X)

Question ID:	M04-Q46
Topic:	Deterministic Models
Difficulty:	Medium
Blooms:	Analyze

A finite-state machine (FSM) controlling an elevator is used to illustrate:

1. A continuous-time Markov chain only
2. A visualization library
3. An image compression algorithm
4. A stochastic bootstrap procedure
5. A deterministic system with defined transitions

Question ID:	M04-Q1
Topic:	Data Storytelling Components
Difficulty:	Easy
Blooms:	Remember

Which trio forms the core 'puzzle pieces' of data storytelling emphasized in the lecture?

1. Data, Narrative, Visualizations
2. Tables, Dashboards, Animation
3. Python, SQL, Excel
4. Fonts, Colors, Icons
5. Neural Nets, GPUs, TPUs

Question ID:	M05A-Q1
Topic:	Structured vs. Unstructured Data
Difficulty:	Easy
Blooms:	Remember

Which statement best captures why data is called 'structured' in this context?

1. It lives in a single file rather than multiple tables.
2. It is stored as images and audio files.
3. It contains only numeric values with no missing entries.
4. It comes from sensors in time order only.
5. It has a predefined schema with labeled columns and consistent record shape.

Question ID:	M06-Q1
Topic:	ML Process: Problem Framing
Difficulty:	Medium
Blooms:	Analyze

Which step most directly ensures that the problem you're solving is actually a machine-learning problem and not a simple rules or query task?

1. Check whether a labeled target exists and if patterns must generalize to new data
2. Visualize predictions with a confusion matrix
3. Collect more data first
4. Tune hyperparameters with cross-validation
5. Train a baseline linear model

Question ID:	MT-Q1
Topic:	None
Difficulty:	Easy
Blooms:	N/A

Which of the following statements about Python dictionaries is TRUE?

1. Dictionaries preserve the order of key insertion in all Python versions
2. Keys in a dictionary must be immutable objects
3. Duplicate keys are allowed and the values are stored in a list
4. Dictionary keys and values must be of the same type