100 Multiple-Choice Questions for Python Data Science

Section 1: Python Fundamentals & Data Structures

1. What is the output of `type(10/2)` in Python? a) `int` b) `float` c) `str` d) `bool`

   Answer: b) `float`

2. Which of the following is a mutable data type in Python? a) `tuple` b) `string` c) `list` d) `int`

   Answer: c) `list`

3. What is the correct way to define a function in Python? a) `def my_function:` b) `function my_function():` c) `define my_function():` d) `def my_function():`

   Answer: d) `def my_function():`

4. What will the following code print?

   Python

   ```
   x = [1, 2, 3]
   y = x
   y.append(4)
   print(x)
   ```

   a) `[1, 2, 3]` b) `[1, 2, 3, 4]` c) `[1, 2, 3, 4, 4]` d) An error

   Answer: b) `[1, 2, 3, 4]`

5. Which symbol is used for single-line comments in Python? a) `//` b) `#` c) `` d) `/* */`

   Answer: b) `#`

6. What does `len({"a": 1, "b": 2})` return? a) 1 b) 2 c) 3 d) Error

   Answer: b) 2

7. What is the purpose of the `__init__` method in a Python class? a) To destroy an object b) To define a class method c) To initialize the object's attributes d) To inherit from a parent class

   Answer: c) To initialize the object's attributes

8. How can you install an external Python library? a) `python install library_name` b) `install.package(library_name)` c) `pip install library_name` d) `setup install library_name`

   Answer: c) `pip install library_name`

9. What does the `range()` function return in Python 3? a) A list of numbers b) A tuple of numbers c) A generator object d) A string of numbers

   Answer: c) A generator object

10. Which of the following is an immutable data type? a) `list` b) `dict` c) `set` d) `tuple`

    Answer: d) `tuple`

Section 2: NumPy for Numerical Operations

11. Which Python library is the foundation for numerical computing and multi-dimensional arrays? a) Pandas b) Matplotlib c) Scikit-learn d) NumPy

    Answer: d) NumPy

12. What is the main advantage of NumPy arrays over standard Python lists? a) They are easier to create. b) They consume less memory and are faster for numerical operations. c) They can store different data types in the same array. d) They don't have indexes.

    Answer: b) They consume less memory and are faster for numerical operations.

13. How do you create a NumPy array from a Python list `[1, 2, 3]`? a) `np.array([1, 2, 3])` b) `numpy.list_to_array([1, 2, 3])` c) `np.to_array([1, 2, 3])` d) `array([1, 2, 3])`

    Answer: a) `np.array([1, 2, 3])`

14. Given a NumPy array `arr = np.array([[1, 2], [3, 4]])`, what does `arr.shape` return? a) `(2, 2)` b) `4` c) `[2, 2]` d) `(2, 4)`

    Answer: a) `(2, 2)`

15. What is the output of `np.zeros(3)`? a) `[0, 0, 0]` b) `array([0., 0., 0.])` c) `[[0, 0, 0]]` d) `(0, 0, 0)`

Answer: b) `array([0., 0., 0.])`

16. How do you select the element `4` from the NumPy array `arr = np.array([[1, 2, 3], [4, 5, 6]])`? a) `arr[1, 0]` b) `arr[0, 1]` c) `arr[1][0]` d) `arr.item(1, 0)`

    Answer: a) `arr[1, 0]`

17. What is broadcasting in NumPy? a) A method for saving arrays to a file. b) A way to perform operations on arrays of different shapes. c) A process for converting an array to a list. d) A function for printing array elements.

    Answer: b) A way to perform operations on arrays of different shapes.

18. What is the output of `np.arange(1, 5, 2)`? a) `array([1, 2, 3, 4])` b) `array([1, 3, 5])` c) `array([1, 3])` d) `array([1, 2, 3])`

    Answer: c) `array([1, 3])`

19. How do you find the mean of all elements in a NumPy array `arr`? a) `arr.mean()` b) `np.mean(arr)` c) `arr.average()` d) Both a and b

    Answer: d) Both a and b

20. What does `arr.reshape(2, 3)` do to a 1D array `arr` with 6 elements? a) It creates a new array with 2 rows and 3 columns. b) It changes the original array to have 2 rows and 3 columns. c) It raises an error because the number of elements is wrong. d) It flattens the array.

    Answer: a) It creates a new array with 2 rows and 3 columns.

Section 3: Pandas for Data Manipulation

21. Which Python library is essential for working with structured data, like spreadsheets or SQL tables? a) NumPy b) Matplotlib c) Pandas d) Scikit-learn

    Answer: c) Pandas

22. What are the two primary data structures in Pandas? a) Series and Dictionary b) List and DataFrame c) Series and DataFrame d) Array and List

    Answer: c) Series and DataFrame

23. What is the output of `df.head()` on a Pandas DataFrame `df`? a) The first 5 columns of the DataFrame. b) The last 5 rows of the DataFrame. c) The first 5 rows of the DataFrame. d) The column names of the DataFrame.

    Answer: c) The first 5 rows of the DataFrame.

24. How do you read a CSV file named `data.csv` into a Pandas DataFrame? a) `pd.read_csv('data.csv')` b) `pd.load_csv('data.csv')` c) `pd.read_file('data.csv')` d) `pd.from_csv('data.csv')`

    Answer: a) `pd.read_csv('data.csv')`

25. Which Pandas method is used to check for missing values in a DataFrame `df`? a) `df.isnull()` b) `df.isna()` c) `df.has_null()` d) Both a and b

    Answer: d) Both a and b

26. What does `df['column_name']` return? a) A Series object b) A DataFrame object c) A list d) A tuple

    Answer: a) A Series object

27. How do you remove rows with missing values from a DataFrame `df`? a) `df.dropna()` b) `df.fillna(0)` c) `df.drop_null()` d) `df.remove_na()`

    Answer: a) `df.dropna()`

28. Which of the following is used to remove duplicate rows from a DataFrame `df`? a) `df.drop_duplicates()` b) `df.unique()` c) `df.remove_duplicates()` d) `df.distinct()`

    Answer: a) `df.drop_duplicates()`

29. What is the purpose of the `groupby()` method in Pandas? a) To filter out rows based on a condition. b) To combine multiple DataFrames. c) To group rows based on one or more columns and apply an aggregation function. d) To sort the DataFrame by a specific column.

    Answer: c) To group rows based on one or more columns and apply an aggregation function.

30. How do you select rows from a DataFrame `df` where the value in the 'age' column is greater than 25? a) `df.loc[df['age'] > 25]` b) `df[df.age > 25]` c) `df.query('age > 25')` d) All of the above

    Answer: d) All of the above

Section 4: Matplotlib for Data Visualization

31. Which Python library is commonly used for creating static, animated, and interactive visualizations? a) NumPy b) Seaborn c) Matplotlib d) Pandas

    Answer: c) Matplotlib

32. What is the most common way to import Matplotlib's `pyplot` submodule? a) `import matplotlib.pyplot as mplt` b) `import pyplot as plt` c) `import matplotlib.pyplot as plt` d) `from matplotlib import pyplot`

    Answer: c) `import matplotlib.pyplot as plt`

33. Which type of plot is best for showing the distribution of a single numeric variable? a) Line plot b) Histogram c) Bar chart d) Scatter plot

    Answer: b) Histogram

34. Which function is used to create a scatter plot in Matplotlib? a) `plt.plot()` b) `plt.bar()` c) `plt.scatter()` d) `plt.pie()`

    Answer: c) `plt.scatter()`

35. How do you set the title of a plot in Matplotlib? a) `plt.set_title("My Title")` b) `plt.title("My Title")` c) `plt.add_title("My Title")` d) `plt.plot_title("My Title")`

    Answer: b) `plt.title("My Title")`

36. Which function is used to add a legend to a plot? a) `plt.add_legend()` b) `plt.show_legend()` c) `plt.legend()` d) `plt.legendary()`

    Answer: c) `plt.legend()`

37. A box plot is useful for visualizing: a) The relationship between two variables. b) The frequency of categorical data. c) The data distribution, median, quartiles, and outliers. d) Trends over time.

    Answer: c) The data distribution, median, quartiles, and outliers.

38. What is the purpose of `plt.xlabel()` and `plt.ylabel()`? a) To add a title to the plot. b) To label the x and y axes. c) To set the limits of the axes. d) To add a grid to the plot.

    Answer: b) To label the x and y axes.

39. What type of chart is ideal for showing proportions or percentages of a whole? a) Bar chart b) Line chart c) Pie chart d) Scatter plot

    Answer: c) Pie chart

40. How do you display a plot in Matplotlib? a) `plt.show()` b) `plt.plot()` c) `plt.display()` d) `plt.render()`

    Answer: a) `plt.show()`

Section 5: Statistics and Probability

41. What does the standard deviation measure? a) The central tendency of the data. b) The spread or dispersion of the data. c) The most frequent value. d) The correlation between two variables.

    Answer: b) The spread or dispersion of the data.

42. What is the range of a probability value? a) -1 to 1 b) 0 to 1 c) 0 to infinity d) -infinity to infinity

    Answer: b) 0 to 1

43. Which of the following is a measure of central tendency? a) Variance b) Standard deviation c) Mean d) Range

    Answer: c) Mean

44. The Central Limit Theorem states that the sampling distribution of the sample means will be approximately normal, regardless of the population distribution, as long as: a) The sample size is small. b) The population is normally distributed. c) The sample size is large enough. d) The population variance is known.

    Answer: c) The sample size is large enough.

45. What is the difference between a population and a sample? a) A population is a subset of a sample. b) A sample is a subset of a population. c) They are the same. d) A population is always larger than a sample.

    Answer: b) A sample is a subset of a population.

46. What does the term "correlation" refer to? a) A measure of central tendency. b) The relationship between two or more variables. c) The spread of data. d) The mode of a dataset.

Answer: b) The relationship between two or more variables.

47. What type of distribution is bell-shaped and symmetrical? a) Poisson distribution b) Binomial distribution c) Normal distribution d) Uniform distribution

    Answer: c) Normal distribution

48. What is the null hypothesis in a hypothesis test? a) The claim that there is an effect or a relationship. b) The claim that there is no effect or no relationship. c) The result of the statistical test. d) The confidence interval.

    Answer: b) The claim that there is no effect or no relationship.

49. What is a p-value? a) The probability of the null hypothesis being true. b) The probability of observing the data, or more extreme data, given that the null hypothesis is true. c) The probability of the alternative hypothesis being true. d) The confidence level.

    Answer: b) The probability of observing the data, or more extreme data, given that the null hypothesis is true.

50. What is a key characteristic of a `discrete` probability distribution? a) It can take any value within a given range. b) It describes outcomes that can be counted or are finite. c) It is always bell-shaped. d) It only has two possible outcomes.

    Answer: b) It describes outcomes that can be counted or are finite.

Section 6: Machine Learning Concepts

51. Which of the following is a **supervised** learning algorithm? a) K-Means b) Principal Component Analysis (PCA) c) Linear Regression d) DBSCAN

    Answer: c) Linear Regression

52. In supervised learning, the model is trained on: a) Unlabeled data. b) Data with features but no target variable. c) Labeled data (features and target variable). d) Only images and video data.

    Answer: c) Labeled data (features and target variable).

53. What is the primary goal of a regression model? a) To predict a category or class. b) To group similar data points. c) To predict a continuous value. d) To reduce the number of features.

    Answer: c) To predict a continuous value.

54. What is **overfitting**? a) When a model is too simple to capture the underlying pattern in the data. b) When a model performs well on training data but poorly on unseen test data. c) When a model performs poorly on both training and test data. d) When a model is trained on too little data.

    Answer: b) When a model performs well on training data but poorly on unseen test data.

55. What does a **confusion matrix** evaluate? a) The speed of a model. b) The accuracy and errors of a classification model. c) The memory usage of an algorithm. d) The quality of a dataset.

    Answer: b) The accuracy and errors of a classification model.

56. Which of the following is an **unsupervised** learning algorithm? a) Support Vector Machine (SVM) b) Decision Tree c) K-Means Clustering d) Naive Bayes

    Answer: c) K-Means Clustering

57. The **bias-variance tradeoff** refers to the balance between: a) Model training time and prediction time. b) Model complexity and interpretability. c) A model's ability to learn patterns (low bias) and its sensitivity to data fluctuations (low variance). d) Accuracy and recall.

    Answer: c) A model's ability to learn patterns (low bias) and its sensitivity to data fluctuations (low variance).

58. What is the purpose of **cross-validation**? a) To train a model on all available data. b) To evaluate a model's performance on different subsets of the data to ensure it generalizes well. c) To prevent a model from learning. d) To select the best features for a model.

    Answer: b) To evaluate a model's performance on different subsets of the data to ensure it generalizes well.

59. What is **feature scaling** used for? a) To increase the number of features. b) To convert categorical features into numerical features. c) To normalize the range of independent variables to prevent features with large values from dominating the model. d) To find the correlation between features.

    Answer: c) To normalize the range of independent variables to prevent features with large values from dominating the model.

60. What is a hyperparameter? a) A parameter learned by the model during training. b) A parameter that is set before the training process begins. c) A feature in the dataset. d) The output of the model.

    Answer: b) A parameter that is set before the training process begins.

Section 7: Machine Learning in Python (Scikit-learn)

61. Which Python library is a popular choice for machine learning tasks? a) Pandas b) Matplotlib c) Scikit-learn d) Seaborn

    Answer: c) Scikit-learn

62. What is the first step when using a `scikit-learn` model like `LinearRegression`? a) `model.fit(X_test, y_test)` b) `model.predict(X_train)` c) `model = LinearRegression()` d) `model.score(X_train, y_train)`

    Answer: c) `model = LinearRegression()`

63. Which `scikit-learn` function is used to split a dataset into training and testing sets? a) `train_test_split()` b) `split_data()` c) `train_split()` d) `data_split()`

    Answer: a) `train_test_split()`

64. After fitting a model with `model.fit(X_train, y_train)`, which method is used to make predictions on new data `X_test`? a) `model.predict(X_test)` b) `model.fit(X_test)` c) `model.score(X_test, y_test)` d) `model.transform(X_test)`

    Answer: a) `model.predict(X_test)`

65. What does the `accuracy_score()` function from `sklearn.metrics` measure? a) The model's performance on a regression task. b) The proportion of correct predictions in a classification task. c) The number of false positives. d) The mean squared error.

    Answer: b) The proportion of correct predictions in a classification task.

66. Which class in `scikit-learn` is used for K-Means clustering? a) `KMeans()` b) `K_Means()` c) `Cluster()` d) `KMeansClustering()`

    Answer: a) `KMeans()`

67. To improve a model's performance on an imbalanced classification dataset, which metric is often preferred over accuracy? a) Mean Absolute Error (MAE) b) R-squared c) F1-score d) Root Mean Squared Error (RMSE)

    Answer: c) F1-score

68. What is the purpose of the `StandardScaler` class? a) To convert strings to numbers. b) To normalize features by subtracting the mean and dividing by the standard deviation. c) To apply a logarithmic transformation to the data. d) To handle missing values.

    Answer: b) To normalize features by subtracting the mean and dividing by the standard deviation.

69. Which module in `scikit-learn` contains the classes for various classification, regression, and clustering algorithms? a) `sklearn.model_selection` b) `sklearn.preprocessing` c) `sklearn.metrics` d) `sklearn.tree`

    Answer: d) `sklearn.tree` (while other modules are important, tree-based models like `DecisionTreeClassifier` are a good example. The most general answer would be the top-level `sklearn` module, but `sklearn.tree` is a good example of a specific module.)

70. What is the output of `model.score(X_test, y_test)`? a) The predicted values for `X_test`. b) A single value representing the model's performance (e.g., accuracy for classification). c) The training loss. d) The model's hyperparameters.

    Answer: b) A single value representing the model's performance (e.g., accuracy for classification).

Section 8: Advanced Data Science & Machine Learning

71. What does **EDA** stand for in data science? a) Easy Data Analytics b) Exploratory Data Analysis c) External Data Automation d) Extended Data Arrangement

    Answer: b) Exploratory Data Analysis

72. The process of converting categorical data into a numerical format that can be used by machine learning models is called: a) Data scaling b) One-hot encoding c) Dimensionality reduction d) Data imputation

    Answer: b) One-hot encoding

73. What is **dimensionality reduction**? a) A technique to increase the number of features in a dataset. b) A method to convert numerical data to categorical. c) The process of reducing the number of random variables under consideration by obtaining a set of principal variables. d) The process of training a model.

    Answer: c) The process of reducing the number of random variables under consideration by obtaining a set of principal variables.

74. Which of the following is a popular algorithm for dimensionality reduction? a) Linear Regression b) K-Means c) Principal Component Analysis (PCA) d) Naive Bayes

    Answer: c) Principal Component Analysis (PCA)

75. What is an **ensemble learning** method? a) An algorithm that uses a single, complex model. b) A method that trains a model on a single dataset. c) A technique that combines multiple machine learning models to get better predictive performance. d) A way to preprocess data.

    Answer: c) A technique that combines multiple machine learning models to get better predictive performance.

76. Which of the following is an example of an ensemble learning algorithm? a) Linear Regression b) K-Nearest Neighbors (KNN) c) Random Forest d) Support Vector Machine (SVM)

    Answer: c) Random Forest

77. What is **regularization** used for in machine learning? a) To increase a model's complexity. b) To prevent a model from overfitting. c) To speed up the training process. d) To make the data normally distributed.

    Answer: b) To prevent a model from overfitting.

78. Which of the following is a type of regularization? a) K-Fold Cross-Validation b) L1 and L2 regularization c) Principal Component Analysis d) One-Hot Encoding

    Answer: b) L1 and L2 regularization

79. What is the main idea behind **Reinforcement Learning**? a) Learning from labeled data. b) Learning from unlabeled data. c) Learning by interacting with an environment and receiving rewards or penalties. d) Learning by combining multiple models.

    Answer: c) Learning by interacting with an environment and receiving rewards or penalties.

80. In a typical data science project lifecycle, what step comes after data cleaning and preprocessing? a) Deployment b) Model training c) Problem definition d) Data collection

    Answer: b) Model training

Section 9: Python Libraries for Specific Tasks

81. Which library is specifically designed for statistical data visualization and is built on top of Matplotlib? a) Pandas b) NumPy c) Seaborn d) Scikit-learn

    Answer: c) Seaborn

82. What is the purpose of Seaborn's `heatmap()` function? a) To visualize the distribution of a single variable. b) To show the correlation matrix of a DataFrame. c) To create a pie chart. d) To plot a line graph.

    Answer: b) To show the correlation matrix of a DataFrame.

83. Which library is often used for natural language processing (NLP) tasks in Python? a) OpenCV b) NLTK (Natural Language Toolkit) c) TensorFlow d) SciPy

    Answer: b) NLTK (Natural Language Toolkit)

84. What is the purpose of the `sklearn.metrics.roc_curve` function? a) To evaluate regression models. b) To find clusters in data. c) To evaluate the performance of a binary classification model at various thresholds. d) To calculate the F1-score.

    Answer: c) To evaluate the performance of a binary classification model at various thresholds.

85. Which library is a powerful and flexible tool for creating deep learning models? a) Pandas b) Matplotlib c) TensorFlow d) SciPy

    Answer: c) TensorFlow

86. What is the primary use of the `SciPy` library? a) Numerical computation. b) Scientific and technical computing (e.g., optimization, integration, signal processing). c) Data manipulation. d) Data visualization.

    Answer: b) Scientific and technical computing (e.g., optimization, integration, signal processing).

87. Which library is best suited for image and video analysis? a) Pandas b) OpenCV c) Seaborn d) SciPy

    Answer: b) OpenCV

88. What is the main purpose of the `datetime` module in Python? a) To perform numerical calculations. b) To work with dates and times. c) To handle network requests. d) To create random numbers.

    Answer: b) To work with dates and times.

89. The `re` module in Python is used for: a) Reading and writing files. b) Regular expressions. c) Random number generation. d) Object-oriented programming.

    Answer: b) Regular expressions.

90. Which `pandas` function is used to convert the data type of a column? a) `df.convert_dtype()` b) `df.astype()` c) `df.change_type()` d) `df.dtype_convert()`

    Answer: b) `df.astype()`

Section 10: General Data Science & Problem Solving

91. What is the first step in any data science project? a) Data cleaning b) Data collection c) Model training d) Defining the business problem

    Answer: d) Defining the business problem

92. What is a key step in data preprocessing to handle missing values? a) Dropping all rows with missing data. b) Imputing missing values with the mean, median, or a specific value. c) Converting missing values to a string like 'NA'. d) Ignoring the missing values.

    Answer: b) Imputing missing values with the mean, median, or a specific value.

93. What is the difference between a `training set` and a `test set`? a) The training set is used to evaluate the model, while the test set is used to build it. b) The training set is used to build the model, while the test set is used to evaluate its performance on unseen data. c) They are interchangeable and can be used for either purpose. d) The training set is always larger than the test set.

    Answer: b) The training set is used to build the model, while the test set is used to evaluate its performance on unseen data.

94. Which type of plot is best for visualizing time-series data? a) Bar chart b) Pie chart c) Line chart d) Box plot

    Answer: c) Line chart

95. What is the purpose of the **Anaconda** distribution? a) To provide an IDE for Python programming. b) To manage Python environments and packages, especially for data science. c) To speed up Python code execution. d) To convert Python code to other languages.

    Answer: b) To manage Python environments and packages, especially for data science.

96. What is a key characteristic of a `Jupyter Notebook`? a) It is a text-based editor for Python code. b) It allows for a mix of code, visualizations, and markdown text in one document. c) It is used to deploy machine learning models to production. d) It is a command-line interface.

    Answer: b) It allows for a mix of code, visualizations, and markdown text in one document.

97. What is **data wrangling**? a) The process of visualizing data. b) The process of cleaning, transforming, and structuring data for analysis. c) The process of deploying a machine learning model. d) The process of defining the business problem.

    Answer: b) The process of cleaning, transforming, and structuring data for analysis.

98. Which of the following is an example of a **classification** problem? a) Predicting a house price. b) Predicting whether an email is spam or not spam. c) Forecasting stock prices. d) Estimating the number of customers a business will have next month.

    Answer: b) Predicting whether an email is spam or not spam.

99. In **deep learning**, what is a **neural network** inspired by? a) The human brain b) The human digestive system c) Social networks d) A database

    Answer: a) The human brain

100. What is a common way to handle imbalanced datasets in classification? a) Using a simple accuracy metric. b) Oversampling the minority class or undersampling the majority class. c) Using a linear regression model. d) Removing all features from the dataset.

`**Answer:** b) Oversampling the minority class or undersampling the majority class`