Ace the upcoming Data Science Interview

You can't anticipate every question an interviewer will ask. However, there are many **critical questions** that you can prepare before the interview.

Our hiring partners have helped us curate a set of interview questions on key skills, which will help you prepare better for the data science job roles.

Filters

1. What are built-in function in python?

Basic Python

Hint?

Read about keywords, methods, functions (user-defined vs pre-defined), methods, difference between methods and functions.

2. Differentiate between Call by value and Call by reference

Basic Python

Hint?

Understanding difference between objects, function and methods in python

3. How to read any file (without using Pandas)

Intermediate Python

Hint?

https://www.analyticsvidhya.com/blog/2017/03/read-commonly-used-formats-using-python/

4. What is NaN in python?

Basic Python

Hint?

Read about missing values / null values and how that affects an analysis

5. What is the use of ID() function in python?

Basic Python

Hint?

https://www.geeksforgeeks.org/id-function-python/

6. How will you import multiple excel sheets in a data frame?

Basic Python

Hint?

https://www.caktusgroup.com/blog/2019/08/13/import-multiple-excel-sheets-pandas/

7. What are the different types of data types?

Basic Python

Hint?

Also understand all logical and arithmatic operator will work between 2 datatypes.

https://www.geeksforgeeks.org/python-data-types/

8. Difference between lists/ tuples/ dictionaries?

Basic Python

Hint?

https://www.edureka.co/blog/variables-and-data-types-in-python/

https://www.techbeamers.com/python-programming-questions-list-tuple-dictionary/

https://www.afternerd.com/blog/difference-between-list-tuple/

9. How would check a number is prime or not using Python?

Basic Python

```
# taking input from user

number = int(input("Enter any number: "))

# prime number is always greater than 1

if number > 1:

for i in range(2, number):

if (number % i) == 0:

print(number, "is not a prime number")

break

else:

print(number, "is a prime number")

# if the entered number is less than or equal to 1

# then it is not a prime number

else:

print(number, "is not a prime number")
```

10. How would check a number is armstrong number using Python?

Basic Python

```
# Python program to check if the number is an Armstrong number or not

# take input from the user

num = int(input("Enter a number: "))

# initialize sum

sum = 0

# find the sum of the cube of each digit

temp = num

while temp > 0:

digit = temp % 10

sum += digit ** 3

temp //= 10

# display the result

if num == sum:

print(num,"is an Armstrong number")

else:

print(num,"is not an Armstrong number")
```

11. What is an Append Function?

Basic Python

The append() method adds an item to the end of the list.

The syntax of the append() method is:

list.append(item)

12. For what Beautiful soup library is used for?

Basic Python

Hint?

Web scrapping

13. Which function is most useful to convert a multidimensional array into a one-dimensional

Basic Python

Hint?

ravel()

14. Python or R – Which one would you prefer for text analytics?

Intermediate Python

15. What is the lambda function in Python?

Intermediate Python

In Python, anonymous functions are defined using the lambda keyword

Syntax of Lambda Function in python

lambda arguments: expression

16. How negative indices are used in Python?

Intermediate Python

Python programming language supports negative indexing of arrays, something which is not available in arrays in most other programming languages. This means that the index value of -1 gives the last element, and -2 gives the second last element of an array. The negative indexing starts from where the array ends. This means that the last element of the array is the first element in the negative indexing which is -1.

17. How is the Python series different from a single column dataframe?

Intermediate Python

Python series is the data structure for a single column of a DataFrame, not only conceptually, but literally, i.e. the data in a DataFrame is actually stored in memory as a collection of Series

Series is a one-dimensional object that can hold any data type such as integers, floats and strings and it does not have any name/header whereas the dataframe has column names.

18. Which libraries in SciPy have you worked with in your project?

Intermediate Python

SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers etc

Subpackages include:

scipy.cluster

scipy.constants

scipy.fftpack

scipy.integrate

scipy.interpolation

scipy.linalg

scipy.io

scipy.ndimage

scipy.odr

scipy.optimize

scipy.signal

scipy.sparse

scipy.spatial

scipy.special

scipy.stats

scipy.weaves

19. How groupby function works in Python?

Intermediate Python

Pandas dataframe.groupby() function is used to split the data into groups based on some criteria. pandas objects can be split on any of their axes.

Syntax: DataFrame.groupby(by=None, axis=0, level=None, as_index=True, sort=True, group_keys=True, squeeze=False, **kwargs)

Parameters :

by: mapping, function, str, or iterable

axis: int, default 0

level: If the axis is a MultiIndex (hierarchical), group by a particular level or levels

as_index: For aggregated output, return object with group labels as the index. Only relevant for DataFrame input. as_index=False is effectively “SQL-style” grouped output

sort: Sort group keys. Get better performance by turning this off. Note this does not influence the order of observations within each group. groupby preserves the order of rows within each group.

group_keys: When calling apply, add group keys to index to identify pieces

squeeze: Reduce the dimensionality of the return type if possible, otherwise return a consistent type

Returns: GroupBy object

20. What does [::-1] do in python?

Intermediate Python

[::] just produces a copy of all the elements in order

[::-1] produces a copy of all the elements in reverse order

21. What are python packages?

Basic Python

Packages are namespaces which contain multiple packages and modules themselves. They are simply directories.

Each package in Python is a directory which MUST contain a special file called __init__.py. This file can be empty, and it indicates that the directory it contains is a Python package, so it can be imported the same way a module can be imported.

If we create a directory called foo, which marks the package name, we can then create a module inside that package called bar. We also must not forget to add the __init__.py file inside the foo directory.

22. How do you check missing values in a dataframe using python?

Intermediate Python

Pandas isnull() function detect missing values in the given object. It returns a boolean same-sized object indicating if the values are NA. Missing values get mapped to True and non-missing value gets mapped to False.

23. How do you get the frequency of a categorical column of a dataframe using python?

Basic Python

Using Series.value_counts()

24. Can you write a function using python to impute outliers?

Basic Python

```
import numpy as np
def removeOutliers(x, outlierConstant):
a = np.array(x)
upper_quartile = np.percentile(a, 75)
lower_quartile = np.percentile(a, 25)
IQR = (upper_quartile - lower_quartile) * outlierConstant
quartileSet = (lower_quartile - IQR, upper_quartile + IQR)
resultList = []
for y in a.tolist():
```

if y > = quartileSet[0] and y < = quartileSet[1]:

resultList.append(y)

return resultList

25. How can we convert a python series object into a dataframe?

Basic Python

Series.to_frame(name=None)

26. How can you change the index of a dataframe in python?

Basic Python

DataFrame.set_index(keys, drop=True, append=False, inplace=False, verify_integrity=False)

keys: label or array-like or list of labels/arrays

This parameter can be either a single column key, a single array of the same length as the calling DataFrame, or a list containing an arbitrary combination of column keys and arrays. Here, “array” encompasses Series, Index, np.ndarray, and instances of Iterator.

27. Is Python case sensitive?

Basic Python

Yes

28. What all ways have you used to convert categorical columns into numerical data using python?

Intermediate Python

One of the most used and popular ones are LabelEncoder and OneHotEncoder.

Both are provided as parts of sklearn library.

LabelEncoder can be used to transform categorical data into integers:

from sklearn.preprocessing import LabelEncoder

```
label_encoder = LabelEncoder()

x = ['Apple', 'Orange', 'Apple', 'Pear']

y = label_encoder.fit_transform(x)

print(y)

array([0, 1, 0, 2])
```

OneHotEncoder can be used to transform categorical data into one hot encoded array:

```
from sklearn.preprocessing import OneHotEncoder

onehot_encoder = OneHotEncoder(sparse=False)

y = y.reshape(len(y), 1)

onehot_encoded = onehot_encoder.fit_transform(y)

print(onehot_encoded)
```

29. How get_dummies() is different from one hot encoder?

Intermediate Python

OneHotEncoder cannot process string values directly. If your nominal features are strings, then you need to first map them into integers.

pandas.get_dummies is kind of the opposite. By default, it only converts string columns into one-hot representation, unless columns are specified.

30. How do you check the distribution of data in python?

Intermediate Python

A simple and commonly used plot to quickly check the distribution of a sample of data is the histogram.

```
from matplotlib import pyplot

pyplot.hist(data)
```

31. What is the difference between iloc and loc activity?

Basic Python

loc gets rows (or columns) with particular labels from the index.

iloc gets rows (or columns) at particular positions in the index (so it only takes integers).

32. Difference between univariate and bivariate analysis? What all different functions can be used in python?

Basic Python

Univariate statistics summarize only one variable at a time.

Bivariate statistics compare two variables.

Below are a few functions which can be used in the univariate and bivariate analysis:

1. To find the population proportions with different types of blood disorders.

df.Thal.value_counts()

2. To make a plot of the distribution :

sns.distplot(df.Variable.dropna())

3. Find the minimum, maximum, average, and standard deviation of data.

There is a function called 'describe'

4. Find the mean of the Variable

df.Variable.dropna().mean()

5. Boxplot to observe outliers

sns.boxplot(x = "", y = "", hue = "", data=df)

6. Correlation plot:

data.corr()

33. What all different methods can be used to standardize the data using python?

Intermediate Python

Min Max Scaler

Min Max Scaler.

Standard Scaler.

Max Abs Scaler.

Robust Scaler.

Quantile Transformer Scaler.

Power Transformer Scaler.

Unit Vector Scaler.

34. What is the apply function in Python? How does it work?

Basic Python

Pandas.apply allow the users to pass a function and apply it on every single value of the Pandas series.

Syntax:

s.apply(func, convert_dtype=True, args=())

35. How do you do upsampling of data? Name a python function or explain the code.

Intermediate Python

Up-sampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal.

There are several heuristics for doing so, but the most common way is to simply resample with replacement.

Module for resampling in Python:

from sklearn.utils import resample

36. Can you plot 3D plots using matplotlib? Name the function.

Intermediate Python

Yes

Function:

import numpy as np

```
import numpy as np
import matplotlib.pyplot as plt
fig = plt.figure()
ax = plt.axes(projection ='3d')
```

37. How can you drop a column in python?

Basic Python

DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')

38. What is the use of 'inplace' in python functions?

Basic Python

In-place operation is an operation that changes directly the content of a given linear algebra, vector, matrices(Tensor) with/without making a copy

When inplace = True is used, it performs an operation on data and nothing is returned.

When inplace=False is used, it performs an operation on data and returns a new copy of data.

39. How do you select a sample of dataframe?

Intermediate Python

1. Randomly select a single row: df = df.sample()
2. Randomly select a specified n number of rows: df = df.sample(n=3)
3. Allow a random selection of the same row more than once: df = df.sample(n=3,replace=True)
4. Randomly select a specified fraction of the total number of rows: df = df.sample(frac=0.50)

40. How would you define a block in Python?

Intermediate Python

A block is a group of statements in a program or script. Usually, it consists of at least one statement and declarations for the block, depending on the programming or scripting language. A language which allows grouping with blocks is called a block-structured language

41. How will you remove duplicate data from a dataframe?

Intermediate Python

DataFrame.drop_duplicates(subset=None, keep='first', inplace=False)

subset: Subset takes a column or list of column label. It's default value is none. After passing columns, it will consider them only for duplicates.

keep: keep is to control how to consider duplicate value. It has only three distinct value and default is 'first'.

42. Can you convert a string into an int? When and how?

Basic Python

Python offers the int() method that takes a String object as an argument and returns an integer. This can be done when the value is either of numeric object or floating-point.

But keep these special cases in mind:

A floating-point (an integer with a fractional part) as an argument will return the float rounded down to the nearest whole integer.

43. What does the function zip() do?

Intermediate Python

The zip() function takes iterables (can be zero or more), aggregates them in a tuple, and return it.

The syntax of the zip() function is:

zip(*iterables)

44. How many arguments can the range() function take?

Basic Python

It can take mainly three arguments.

start: integer starting from which the sequence of integers is to be returned

stop: integer before which the sequence of integers is to be returned.

The range of integers ends at stop – 1.

step: integer value which determines the increment between each integer in the sequence

45. What is the difference between list, array and tuple in Python?

Basic Python

List:

The list is an ordered collection of data types.

The list is mutable.

List are dynamic and can contain objects of different data types.

List elements can be accessed by index number

Array:

An array is an ordered collection of similar data types.

An array is mutable.

An array can be accessed by using its index number.

Tuple:

Tuples are immutable and can store any type of data type.

it is defined using ().

it cannot be changed or replaced as it is an immutable data type