1. Type I - Type II Errors? The Power of the test?

hypothesis testing → two types of Error.

A **Type I Error** is committed when we **reject the true null hypothesis!**

Type II error is made when we do **not reject the null false hypothesis**!

The probability of a **Type I error** is denoted by ***α*** and the probability of **Type II error** is denoted by ***β***.

For a given sample *n*, a decrease in *α* will increase *β* and vice versa. Both *α* and *β* decrease as *n* increases.

Decision	**H_0 is true**	**H_0 is false**
Reject the H_0	Type I error	Correct Decision
Fail to reject H_0	Correct Decision	Type II error

Type I error → False positive! Type II → False neg.!

Power of Test: Te probability of rejecting the null hypothesis when the null hypothesis is false. β is the probability of a Type II error, the power of the test is defined as 1- β.

2. Over-fitting and Under-fitting?

Overfitting ← A small amount of data and a large number of variables, and modelling has noise!

Underfitting ← Model is not modelling all the information! Ie, linear model for a non-linear data!

Overfitting → avoid it by **cross-validation techniques** (**K Folds**), **regularisation techniques** (**Lasso**).

3. When use Classification Tech over Regression?

It is used for **categorical variable (Discrete)**! Ie, Logistic Regression, naïve Bayes, Decision Trees & K nearest neighbours!

Regression Tech is for **Continuous variable**!

Linear regression, Support Vector Machine (SVM) and Regression trees. All of them Supervised ML Algos!

4. Data Cleansing?

A process of **removing or updating** the **incorrect, incomplete, duplicated** information, or **formatted improperly**. It is very important for the **accuracy** and **productivity** of modelling. Data cleansing **filters the usable data** from the **raw data**, avoid erroneous results.

5. Which are the important steps of Data Cleaning?

1. **Data Quality**
2. **Removing Duplicate** Data (**irrelevant** data)
3. **Structural errors**
4. **Outliers**
5. Treatment for **Missing Data**

Data Cleaning increases the accuracy of the model.

Data Scientists **spends 80%** of time **cleaning data**.

6. How is k-NN different from k-means clustering?

K-nearest neighbours → **classification (or regression) algo** (supervised)!

→ the number of **nearest neighbours classifies** or (**predicts** in continuous variable/regression) a test sample!

K-means → **Clustering** algorithm (unsupervised learning.)

K-means is **the number of clusters,** the algorithm is trying to learn from the data.

7. What is p-value?

p-value determines the strength of your results in a hypothesis test. (between 0 and 1). The claim → Null Hypothesis.

Lower p-values, ≤ 0.05 → we can reject the Null Hypothesis. A high p-value ≥ 0.05 → we can accept the Null Hypothesis. An exact p-value 0.05 → the Hypothesis can go either way.

P-value is the **measure of the probability** of events other than **suggested by the null hypothesis**.

8. Diff. of Data Science from Big Data and Data Analytics?

Data Science utilises algorithms and tools to draw meaningful and commercially useful insights from **raw data**. It **models**, **cleanses** data, **makes analysis**, **pre-processing** etc. → makes intelligent systems, they learn from data, make decisions!

Big Data → The enormous set of **structured**, **semi-struct.**, and **unstruct.** data in its raw form generated through various channels. → the process of handling large volumes of data! → manages data and processes and maintains the consistency of data!

Data Analytics provides **operational insights** into **complex business scenarios**, predicts upcoming **opportunities** and **threats** for an organisation

to exploit. → gains meaningful insights from the data through mathematical or non-math processes.

Statistics

9. What is the use of Statistics in Data Science?

Statistics provides tools and methods to identify patterns and structures in data to provide a deeper insight into it. Serves a great role in **data acquisition, exploration, analysis**, and **validation**. It plays a really powerful role in Data Science.

Data Science is derived from the **overlap of statistics probability and computer science**. To do estimations, statistics is required. Many algorithms are built on top of **statistical formulae and processes**.

10. Diff. between Supervised and Unsupervised Learning?

Supervised ML requires **labelled data** for **training**, **Unsupervised ML** does not! It can be trained on **unlabelled data**. It has no known results to learn and it has a state-based or adaptive mechanism to learn by itself.

Supervised ML involves **high computation costs**, **unsupervised learning** has low training cost. Supervised ML finds applications in classification and regression tasks whereas unsupervised ML finds applications in clustering and association rule mining.

11. What is a Linear Regression (LR)?

Linear regression is a one-degree equation (Y = mX + C) m is the **slope** of the line and C is **the standard error**. It is used for the continuous variables, like height, weight.

Simple LR → Dependent variable with one independent var! **Multiple LR** → Multiple independent variables!

Linear regression calculates the **best fit line** passing through the data points when plotted. The **best fit line** is the line that the distance of each data point is minimum from that! It **minimizes the overall error of the system**. The features in the data are linearly related to the target. It is often used in predictive analytics for calculating estimates.

12. What is Logistic Regression?

A technique which is used in **predictive analytics** on a **binary variable** like, yes/no - true/false. (Y = eX + e – X) . It is used for **classification based tasks**. It finds out probabilities for a data point to belong to a particular class for classification.

13. Explain Normal Distribution (Gaussian Distribution)

It is **a type of probability distribution** that most **values lie near the mean**:

1. The **mean, median, and mode** of the distribution **coincide**
2. The distribution has a **bell-shaped curve**
3. The **total area under the curve is 1**
4. **Symmetrical**

14. Mention some drawbacks of the Linear Model

Ans. Here a few drawbacks of the linear model:

1. The assumption regarding the **linearity** of the errors
2. It is not usable for **binary outcomes** or **count outcome**
3. It can't solve certain **overfitting** problems
4. It also assumes that there is no **multicollinearity** in the data.

15. Which is the best for text analysis, R or Python?

Python would be a better choice for text analysis as it has the **Pandas library** to **facilitate easy to use data structures** and **high-performance data analysis tools**. However, it depends on the complexity of data.

16. What steps do you follow while making a decision tree?

1. Determine the Root of the Tree Step
2. Calculate Entropy for The Classes Step
3. Calculate Entropy After Split for Each Attribute
4. Calculate Information Gain for each split
5. Perform the Split
6. Perform Further Splits Step
7. Complete the Decision Tree

17. What is correlation and covariance in statistics?

Correlation → the measure of the relationship between two variables. **Directly proportional** to each other → Positive correlation. **Indirectly proportional** to each other → Negative correlation. **Covariance** → Measure of how much two random variables vary together.

18. What is 'Naive' in a Naive Bayes?

Naive Bayes classifier assumes that the **presence or absence** of a particular feature of a class is **unrelated to the presence or absence of any other**

feature, given the class variable. Basically, it's "naive" **(lack of experience)** because it makes assumptions that may or may not turn out to be correct.

19. How can you select k for k-means?
Two methods to calculate the **optimal value of k in k-means**: **Elbow method, Silhouette score method**! **Silhouette score is the most prevalent** while determining the optimal value of k!

20. Native Data Structures in Python? Mutable?
Lists. tuples, sets, dictionary! **Tuples are immutable**.

21. Libraries to plot data in Python?
Matplotlib, seaborn, ggplot. And many open source tools.

22. How is Memory Managed in Python?
In Python, a **private heap** contains **all Python objects** and **data structures**. This heap is ensured internally by the Python memory manager.

23. What is a recall?
The rate of **true positives** with respect to **the sum of true positives and false negatives**. Known as **true positive rate!**

24. What is a lambda function?
Small anonymous function. It can take **any number of arguments**, but can only have **one expression**.

25. What is reinforcement learning?
Reinforcement learning is an **unsupervised** learning technique in ML. It is a **state-based learning** technique. The models have **predefined rules** for state change which enable the system to move from one state to another, while the training phase.

26. Entropy and Information Gain in decision tree algo?
Entropy is used to check the **homogeneity** of a sample. **Zero entrop**y means, the sample is completely **homogenous**. Entropy = 1, means, the sample is **equally divided**. Entropy controls how a Decision Tree decides to split the data. It actually affects how a Decision Tree draws its boundaries.

The **information gain** depends on the decrease in entropy after the dataset is split on an attribute. Constructing a decision tree is always about finding the attributes that return highest information gain.

27. What is Cross-Validation?
A **model validation technique** to asses how the outcomes of a statistical analysis will infer to an independent data set. It is majorly used where prediction is the goal and one needs to estimate the performance accuracy of a predictive model in practice. The goal here is to define a data-set for testing a model in its training phase and limit overfitting and underfitting issues. The validation and the training set is to be drawn from the same distribution to avoid making things worse. ***Read:*** *Why Data Science Jobs Are in Demand*

28. What is Bias-Variance tradeoff?
The **error** introduced in your model because of **over-simplification of the algorithm** is known as Bias. On the other hand, **Variance** is the error introduced to your model because of the **complex nature of machine learning algorithm**. In this case, the model also learns **noise** and perform poorly on the test dataset.
The bias-variance tradeoff is the **optimum balance between bias and variance** in a machine learning model. If **decrease bias**, **the variance will increase** and vice-versa.
Total Error= Square of bias+variance+irreducible error. Bias variance tradeoff is the process of finding the exact number of features while model creation such that the error is kept minimum, but also taking effective care such that the model does not overfit or underfit.

29. Types of biases that occur during sampling?
a. **Self-Selection Bias**. The participants of the analysis select themselves!
b. **Under coverage bias**. Few samples are selected from a **segment of the population**.
c. **Survivorship Bias.** The observations recorded at the end of the investigation are a non-random set of those present at the beginning of the investigation.

30. What is the Confusion Matrix?

2x2 table with four outputs provided by the **binary classifier** that **predicts** all data instances as either **positive** or **negative**.
True positive(TP) False-positive(FP)
True negative(TN) False-negative(FN)

error rate (FP+FN)/(P+N) specificity(TN/N)
accuracy(TP+TN)/(P+N) sensitivity (TP/P)
precision(TP/(TP+FP)).
It is essentially used to evaluate the **performance of a machine learning model** when the truth values of the experiments are already known and the target class has more than two categories of data. It helps in visualisation and evaluation of the results of the statistical process.

31. Explain selection bias
Selection bias (**selection effect**) occurs when the research does not have a random selection of participants. This type of sample collecting distorts statistical analysis. Take selection bias into account! Otherwise, the conclusions may be inaccurate. Some types of selection biases:
1. **Sampling Bias** – A systematic error that results due to a non-random sample
2. Data – Specific data subsets selection to support a conclusion or reject bad data
3. **Attrition** – The bias due to tests that didn't run to completion.

32. What are exploding gradients?
Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in NaN values. Hence the model becomes unstable and is unable to learn from the training data.

33. Explain the Law of Large Numbers
If an **experiment is repeated independently** a large number of times, **the average of the individual results** is close to the expected value. It also states that the sample variance and standard deviation also converge towards the expected value.

34. What is the importance of A/B testing
Its goal is to **pick the best variant among two hypotheses**, the use cases of this kind of testing could be a **web page** or **application responsiveness**, **landing page redesign**, **banner testing**, **marketing campaign performance**. The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

35. Explain Eigenvectors and Eigenvalues
Eigenvectors → Direction in which a **linear transformation** moves and acts by **compressing, flipping, or stretching**. Used to understand **linear transformations** and are generally calculated for a **correlation** or **covariance matrix**.
Eigenvalue → the strength of the transformation in the direction of the eigenvector. An eigenvector's direction remains unchanged when a linear transformation is applied to it. *Upskill with* Great Learning's DSBA Program *today!*

36. Why Is Re-sampling Done?
1. **Estimate the accuracy of sample statistics** with the subsets of accessible data at hand
2. **Substitute data point labels** while performing **significance tests**
3. Validate models by using random subsets

37. What is systematic sampling and cluster sampling
Systematic sampling is a type of probability sampling method. The sample members are selected from a larger population with a random starting point but a fixed periodic interval. This interval is known as the sampling interval. The sampling interval is calculated by dividing the population size by the desired sample size.
Cluster sampling involves dividing the sample population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. Analysis is conducted on data from the sampled clusters.

38.What are Autoencoders?
A kind of artificial neural network. **Used to learn efficient data coding**s in an unsupervised manner. It is utilised for learning
a representation (encoding) for a set of data, mostly for dimensionality

reduction, by training the network to ignore signal “noise”. Autoencoder also tries to generate a representation as close as possible to its original input from the reduced encoding.

39. What are the steps to build a Random Forest Model?
A Random Forest is essentially a build up of a number of decision trees. The steps to build it:
1. Select ‘k’ features from all ‘m’ features, randomly. **k << m**
2. Calculate **node D** by the **best split point** — on ‘k’ features!
3. Split the node into daughter nodes using best split
4. Repeat Steps 2 and 3 until the leaf nodes are finalised
5. Build a Random forest by repeating steps 1-4 for ‘n’ times to create ‘n’ number of trees.

40. How do you avoid the overfitting of your model?
Overfitting → Model setting only for a **small amount of data**. It ignores the **bigger picture**. Methods to avoid overfitting are:
1. Keeping the model simple—using fewer variables and removing major amount of the noise in the training data
2. Using **cross-validation** techniques. k folds cross-validation
3. Using **regularisation** techniques — (LASSO), to penalise model parameters that are more likely to cause overfitting.

41. Univariate, bivariate, and multivariate analysis.
Univariate data, contains only one variable. The univariate analysis deals with this data. **Bivariate data** contains two different variables. The bivariate analysis deals with causes, relationships and analysis between those two variables.
Multivariate data contains three or more variables. Multivariate analysis is like a bivariate, however, there exists more than one dependent variable.

42. How is random forest different from decision trees?
Ans. A Decision Tree is a single structure. Random forest is a collection of decision trees.

43. What is dimensionality reduction? What are its benefits?
Dimensionality reduction is defined as the process of converting a data set with vast dimensions into data with lesser dimensions — in order to convey similar information concisely.
This method is mainly beneficial in compressing data and reducing storage space. It is also useful in reducing computation time due to fewer dimensions. Finally, it helps remove redundant features — for instance, storing a value in two different units (meters and inches) is avoided.
In short, dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

44. For the given points, how will you calculate the Euclidean distance in Python? plot1 = [1,3] ; plot2 = [2,5]
Ans.
```
import math
# Example points in 2-dimensional space...
x = (1,3)
y = (2,5)
distance = math.sqrt(sum([(a - b) ** 2 for a, b in zip(x, y)]))
print("Euclidean distance from x to y: ",distance)
```

45. Feature selection methods to select the right variables.
Filter Methods: These methods involve
1. Linear discrimination analysis
2. ANOVA
3. Chi-Square
Wrapper Methods: These methods involve
1. **Forward Selection**: One feature at a time is tested and a good fit is obtained
2. **Backward Selection**: All features are reviewed to see what works better
3. **Recursive Feature Elimination**: Every different feature is looked at recursively and paired together accordingly.
Others are Forward Elimination, Backward Elimination for Regression, Cosine Similarity-Based Feature Selection for Clustering tasks, Correlation-based eliminations etc.

46. What are the different types of clustering algorithms?
Kmeans Clustering, **KNN** (K nearest neighbour), Hierarchial clustering, Fuzzy Clustering etc. Clustering algorithms.

47. How should you maintain a deployed model?
A deployed model needs to be **retrained** after a while so as **to improve the performance** of the model. Since deployment, a track should be kept of the predictions made by the model and the truth values. Later this can be used to retrain the model with the new data. Also, root cause analysis for wrong predictions should be done.

48. Which ML algorithm can be used to inpu missing values of both categorical and continuous variables? K-means clustering Linear regression K-NN (k-nearest neighbour) Decision trees
Ans. KNN and Kmeans

49. What is a ROC Curve? Explain how it works?
AUC – ROC curve is a performance measurement for the classification problem at various thresholds settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

50. Find RMSE and MSE in a linear regression model!
MSE is the squared sum of (actual value-predicted value) for all data points. It gives an estimate of the **total square sum of errors**. **RMSE** is the square root of the squared sum of errors.

51. False negative is more important than false positive?
Disease prediction on symptoms for diseases like **cancer**.

52. How can outlier values be treated?
Outlier treatment can be done by replacing the values with **mean**, **mode**, or **a cap off** value, or **removing** all rows with outliers if they make up a small proportion of the data. A **data transformation** can also be done on the outliers.

53. Calculate accuracy using a confusion matrix?
Accuracy score = (TP+TN)/(TP+TN+FP+FN)

54. Difference between "long" and "wide" format data?
Wide-format → Where we have a single row for every data point with multiple columns to hold the values of various attributes.
Long format → is where for each data point we have as many rows as the number of attributes and each row contains the value of a particular attribute for a given data point.

55. Explain the SVM machine learning algorithm in detail.
SVM is an ML algorithm which is used for **classification** and **regression**. For classification, it finds out a **multi dimensional hyperplane** to distinguish between classes. SVM uses **kernels** which are namely **linear**, **polynomial**, and **rbf**. There are few parameters which need to be passed to SVM in order to specify the points to consider while the calculation of the **hyperplane**.

56. The various steps involved in an analytics project?
1. Data collection
2. Data cleansing
3. Data pre-processing
4. Creation of train test and validation sets
5. Model creation
6. Hyperparameter tuning
7. Model deployment

57. Explain Star Schema.
→ A **data warehousing concept** in which all schema is connected to a **central schema**.

58. How Regularly Must an Algorithm be Updated?
It completely depends on the **accuracy** and **precision** being required at the point of delivery and also on **how much new data** we have to **train on**. For a model trained on 10 million rows its important to have new data with the same volume or close to the same volume. Training on 1 million new data points every alternate week, or fortnight won't add much value in terms of increasing the efficiency of the model.
59. What is Collaborative Filtering?

Ans. Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user.

60. How will you define the number of clusters in a clustering algorithm?
Ans. By determining the Silhouette score and elbow method, we determine the number of clusters in the algorithm.

61. What is Ensemble Learning? Define types.
Ans. Ensemble learning is clubbing of multiple weak learners (ml classifiers) and then using aggregation for result prediction. It is observed that even if the classifiers perform poorly individually, they do better when their results are aggregated. An example of ensemble learning is random forest classifier.

62. What are the support vectors in SVM?
Ans. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximise the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

63. What is pruning in Decision Tree?
Ans. Pruning is the process of reducing the size of a decision tree. The reason for pruning is that the trees prepared by the base algorithm can be prone to overfitting as they become incredibly large and complex.

64. What are the various classification algorithms?
Ans. Different types of classification algorithms include logistic regression, SVM, Naive Bayes, decision trees, and random forest.

65. What are Recommender Systems?
Ans. A recommendation engine is a system, which on the basis of data analysis of the history of users and behaviour of similar users, suggests products, services, information to users. A recommendation can take user-user relationship, product-product relationships, product-user relationship etc. for recommendations.

Data Analysis Interview Questions
66. List out the libraries in Python used for Data Analysis and Scientific Computations.
NumPy, Scipy, Pandas, sklearn, Matplotlib Seaborn. For deep learning Pytorch, Tensorflow is great tools to learn.

67. The difference between the expected and mean value?
Mathematical expectation, also known as the expected value, is the **summation** or **integration** of possible values from a random variable. **Mean value** is the average of all data points.

68. How are NumPy and SciPy related?
Ans. NumPy and SciPy are python libraries with support for arrays and mathematical functions. They are very handy tools for data science.

69. What will be the output of the below Python code?
```
def multipliers ():
return [lambda x: i * x for i in range (4)]
print [m (2) for m in multipliers ()]
```
Ans. Error

70. What do you mean by list comprehension?
An elegant way to define and create a list in Python. It often have the **qualities of sets** but are not all cases in set. Complete substitute for the lambda, map(), filter(), and reduce().

71. What is __init__ in Python?
A **reserved method** in python classes. **Constructor** in object-oriented concepts. This method is called when an object is created from the class and it allows the class to initialise the attributes of the class.

72. The difference between append() and extend()?
append() adds items to list. extend() iterates over its argument and adds each element in the argument to the list.

73. What is the output of the following? x = [‘ab’, ‘cd’] print(len(list(map(list, x))))
Ans. 2

74. Program to count the total number of lines in a text file.
```
count=0
with open ('filename.txt','rb') as f:
    for line in f:
        count+=1
print count
```

75. How will you read a random line in a file?
```
import random
def random_line(fname):
  lines = open(fname).read().splitlines()
  return random.choice(lines)
print(random_line('test.txt'))
```

76. Represent data effectively with 5 dimensions!
It can be represented in a np.array of dimensions (n*n*n*n*5)

77. Whenever you exit Python, is all memory de-allocated?
Objects having **circular references** are not always free when python exits. Hence when we exit python all memory doesn’t necessarily get deallocated.

78. How would you create an empty NumPy array?
```
import numpy as np
np.empty([2, 2]
```

79. Treating a categorical variable as a continuous variable would result in a better predictive model?
Ans. There is no substantial evidence for that, but in some cases, it might help. It’s totally a brute force approach. Also, it only works when the variables in question are ordinal in nature.

80. How and by what methods data visualisations can be effectively used?
Ans. Data visualisation is greatly helpful while creation of reports. There are quite a few reporting tools available such as tableau, Qlikview etc. which make use of plots, graphs etc for representing the overall idea and results for analysis. Data visualisations are also used in exploratory data analysis so that it gives us an overview of the data.

81. How deal with the features with > 30 % missing values.
In general, **we remove the column**. If it is too important to be removed **we may impute values** with convenient methods.

82. The skewed Distribution vs uniform distribution?
→ A distribution in which the majority of the data points lie to the right or left of the centre. A **uniform distribution** is a distribution in which all outcomes are **equally likely**.

83. The count of different categories in a column in pd!
Value_counts will show the count of different categories.

84. What is the default missing value marker in pandas, and how can you detect all missing values in a DataFrame?
Ans. NaN is the missing values marker in pandas. All rows with missing values can be detected by is_null() function in pandas.

85. What is root cause analysis (RCA)?
→ The process of **tracing back of occurrence of an event** and the factors which lead to it. It’s generally done when a **software malfunctions**. In data science, RCA helps businesses understand the **semantics** behind certain **outcomes**.

86. What is a Box-Cox Transformation?
A way to normalise variables. Normality is an important assumption for many statistical techniques!

87. What if instead of finding the best split, we randomly select a few splits and just select the best from them. Will it work?
The decision tree is based on a **greedy approach**. It selects the **best option** for each branching. If we randomly select the best split from average splits,

it would give us a locally best solution and not the best solution producing sub-par and sub-optimal results.

88. What is the result of the below lines of code?

```
def fast (items= []):
   items.append (1)
   return items
print fast ()
print fast ()
```

Ans. [1]

89. Produce list with unique elements from a list?

l=[1,1,2,2] → l=list(set(l))

90. How will you create a series from dict in Pandas?

```
dict = {'cat' : 10, 'Dog' : 20}
series = pd.Series(dict)
```

91. How will you create an empty DataFrame in Pandas?

```
column_names = ["a", "b", "c"]
df = pd.DataFrame(columns = column_names)
```

92. Get the items of series1 not present in series2, pandas?

[series1.isin(series2)].values

93. Get frequency counts of unique items of a series?

pandas.Series.value_counts

94. Convert a numpy array to a dataframe of given shape?

matrix np.array → df = pd.DataFrame(matrix)

95. What is Data Aggregation?

Aggregate functions are used to get the **necessary outcomes**, like **sum, count, avg, max, min**, after a **groupby**.

96. What is Pandas Index?

Ans. An index is a unique number by which rows in a pandas dataframe are numbered.

97. Describe Data Operations in Pandas?

Ans. Common data operations in pandas are data cleaning, data preprocessing, data transformation, data standardisation, data normalisation, data aggregation.

98. Define GroupBy in Pandas?

Ans. groupby is a special function in pandas which is used to group rows together given certain specific columns which have information for categories used for grouping data together.

99. How to convert the index of a series into a column of a dataframe?

Ans. df = df.reset_index() will convert index to a column in a pandas dataframe.

Advanced Data Science Interview Questions

100. How to keep only the top 2 most frequent values as it is and replace everything else as 'Other'?

Ans.

```
"s = pd.Series(np.random.randint(1, 5, [12]))
print(s.value_counts())
s[~s.isin(ser.value_counts().index[:2])] = 'Other'
s"
```

101. How to convert the first character of each element in a series to uppercase?

Ans. pd.Series([x.title() for x in s])

102. How to get the minimum, 25th percentile, median, 75th, and max of a numeric series?

Ans.

```
"randomness= np.random.RandomState(100)
s = pd.Series(randomness.normal(100, 55, 5))
np.percentile(ser, q=[0, 25, 50, 75, 100])"
```

103. What kind of data does Scatterplot matrices represent?

Ans. Scatterplot matrices are most commonly used to visualise multidimensional data. It is used in visualising bivariate relationships between a combination of variables.

104. What is the hyperbolic tree?
Ans. A hyperbolic tree or hypertree is an information visualisation and graph drawing method inspired by hyperbolic geometry.

105. What is scientific visualisation? How it is different from other visualisation techniques?
Ans. Scientific visualization is representing data graphically as a means of gaining insight from the data. It is also known as visual data analysis. This helps to understand the system that can be studied in ways previously impossible.

106. What are some of the downsides of Visualisation?
Ans. Few of the downsides of visualisation are: It gives estimation not accuracy, a different group of the audience may interpret it differently, Improper design can cause confusion.

107. What is the difference between a tree map and heat map?
Ans. A heat map is a type of visualisation tool that compares different categories with the help of colours and size. It can be used to compare two different measures. The 'tree map' is a chart type that illustrates hierarchical data or part-to-whole relationships.

108. What is disaggregation and aggregation of data?
Ans. Aggregation basically is combining multiple rows of data at a single place from low level to a higher level. Disaggregation, on the other hand, is the reverse process i.e breaking the aggregate data to a lower level.

109. What are some common data quality issues when dealing with Big Data?
Ans. Some of the major quality issues when dealing with big data are duplicate data, incomplete data, the inconsistent format of data, incorrect data, the volume of data(big data), no proper storage mechanism, etc.

110. What is a confusion matrix?
Ans. A confusion matrix is a table for visualising the performance of a classification algorithm on a set of test data for which the true values are known.

111. What is clustering?
Ans. Clustering means dividing data points into a number of groups. The division is done in a way that all the data points in the same group are more similar to each other than the data points in other groups. A few types of clustering are Hierarchical clustering, K means clustering, Density-based clustering, Fuzzy clustering etc.

112. What are the data mining packages in R?
Ans. A few popular data mining packages in R are Dplyr- data manipulation, Ggplot2- data visualisation, purrr- data wrangling, Hmisc- data analysis, datapasta- data import etc.

113.Techniques used for sampling? Advantage of sampling
There are two main Sampling techniques:
A. Probability sampling: Each individual of the population has an **equal possibility of presence** in the sample. Probability sampling methods include:
1. **Simple random sampling**: Each individual has an equivalent chance of being selected or included.
2. **Systematic sampling**: Every individual is **assigned a number** and are chosen at **regular intervals**.
3. **Stratified sampling**: The population is split into sub-populations to represent every sub-population in the sample It allows you to conclude more precise results.
4. **Cluster sampling**: It divides the population into sub-populations whose eeach should have analogous characteristics to that of the whole sample. Rather than sampling individuals from each subpopulation, you randomly select the entire subpopulation.
B. Non-probability sampling: Individuals are selected in non-random ways and they have no equal possibility!
1. **Convenience sampling**: Researcher collects data from an easily accessible group!
2. **Voluntary Response sampling**: Here, people or individuals volunteer themselves.
3. **Purposive** (judgmental) **sampling**: Researchers use their expertise to select a sample that is useful or relevant to the purpose of the research.

4. **Snowball sampling**: Used if the population is difficult to access. Recruit individuals via other individuals!

Advantages of Sampling

1. **Low cost** advantage
2. **Easy to analyze** by limited resources
3. **Less time** than other techniques
4. **Scope** is considered to be **considerably high**
5. **Sampled data is considered to be high**
6. **Organizational convenience**

114. What is imbalance data?

Imbalance data in simple words is a reference to different types of datasets where there is an uneven distribution of observations to the target class. Which means, one class label has higher observations than the other comparatively.

115. Define Lift, KPI, Robustness, Model fitting and DOE

Lift is used to understand the performance of a given targeting model in predicting performance, when compared against a randomly picked targeting model.

KPI (Key performance indicators) is a yardstick used to measure the performance of an organization or an employee based on organizational objectives.

Robustness is a property that identifies the effectiveness of an algorithm when tested with a new independent dataset.

Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained.

Design of Experiment (DOE) is a set of mathematical methods for process optimization and for quality by design (QbD).

116. Define Confounding Variable!

→ An **external influence** in an experiment. It changes the effect of a **dependent** and **independent** variable. A variable should satisfy below conditions to be a confounding variable :

1. It should be correlated to the independent variable.
2. It should be informally related to the dependent variable.

A study on whether a **lack of exercise** has an effect on **weight gain**! **Lack of exercise** is an independent variable and **weight gain** is a dependent variable. A confounder variable can be any other factor that has an effect on weight gain, like amount of food consumed, weather conditions etc.

117. Why are time series problems different from other regression problems?

Time series is **extrapolation**, Regression is **interpolation**. Time-series refers to an organized chain of data. Time-series forecasts what comes next in the sequence. Time-series could be assisted with other series which can occur together.

Regression can be applied to Time-series problems as well as to non-ordered sequences which are termed as Features. While making a projection, new values of Features are presented and Regression calculates results for the target variable.

118. The difference between the Test set and validation set?

Test set : A set of examples to evaluate the performance of a fully specified classifier. It is used to fit the parameters. Used to test the data which is passed as input to your model.

Validation set : A set of examples to tune the parameters of a classifier. It is used to tune the parameters. It is used to validate the output which is produced by your model.

Kernel Trick: A method where a **linear classifier** is used to solve **non-linear problems**. A **non-linear object** is projected to a **higher dimensional space** to make it easier to categorize where the data would be divided linearly by a **plane**.

$\mathbf{K(x_i,x_j) = x_i \cdot x_j = x^T_i x_j}$

Every data point is mapped into the high-dimensional space via some transformation $\mathbf{\Phi: x \rightarrow \Phi(x)}$

$\mathbf{K(x_i,x_j) = \Phi x^T_i \Phi x_j}$

Box Plot and Histograms: Types of charts that represent numerical data graphically. Easier way to visualize data! Easier to compare characteristics of data between categories.