1. How to define a distance measure between two values of a nominal or discrete attribute?
2. Why is it important to understand the nature of the missing values?
3. What are the different ways of handling missing values? Give examples of the situations when they are useful.
4. What are the advantages and disadvantages of imputations while handling missing values?
5. When is regression-based and knn based imputation are useful when handling missing values?
6. What are standard techniques for removing noise in the data?
7. What are the advantages of data normalization? Does it help in handling outliers?
8. What are some graphical methods to identify outliers?
9. What is one-hot encoding and when are they useful?
10. Choose a number uniformly at random from the range [1,1,000,000]. Using the inclusion-exclusion principle, determine the probability that the number chosen is divisible by one or more of 4, 10, and 25.
11. Why expectation of a random variable is always a constant?
12. Does linearity of expectations only follow when the random variables are independent? Why?
13. If the samples are expected to follow a normal distribution, then what is the required sample size to bound the margin or error by 0.04 for a 95% confidence interval?
14. In the above example, if the data collection is a costly task, will it help if you have access to the probability of success?
15. Explain what rejecting the null hypothesis and not rejecting the null hypothesis implies in terms of test parameters.
16. What are the types of errors? Explain them.
17. In your tests, you have replaced 95% confidence interval with a 100% confidence interval.
    a. Comment about your type-1 and type-2 errors.
    b. What will be your confidence interval around the empirical mean?
18. In case when data is unavailable, do we reject or we do not reject the null hypothesis?
19. In z-test, what is the ideal distribution of the test statistics for not to reject the null hypothesis?
20. A professor of the ‘Advanced Algorithms course thinks his students have a good grasp of ‘Graduate Algorithms, i.e., the average score is more than 70%. He randomly chose 10 students who had scored 62, 80, 75, 95, 59, 70, 63, 68, 90

and 84 in graduate algorithms out of 100. With a 5% level of significance, is the data have enough evidence to show that it is the other way around?

21. In the above question, let the professor sample ‘n’ students and their sum of scores in graduate algorithms is 700. Let the standard deviation of the population be 5. What is the smallest value ‘n’ such that the professor rejects the null hypothesis?
22. A manufacturer of hard safety hats for construction workers is concerned about the mean and the variation of the forces its helmets transmit to wearers when subjected to an external force. The manufacturer has designed the helmets so that the mean force transmitted by the helmets to the workers is 800 pounds (or less) with a standard deviation to be less than 40 pounds. Tests were run on a random sample of n = 40 helmets, and the sample mean and sample standard deviation were found to be 825 pounds and 48.5 pounds, respectively. Does the data provide sufficient evidence, at the 5% significance level, to conclude that the population standard deviation exceeds 40 pounds?
23. Suppose a fair coin is flipped 100 times. Find a bound on the probability that the number of times the coin lands on heads is at least 60 or at most 40.
24. If you have 50 people in a classroom whose birthday is assumed to be a uniformly random day, then how many pairwise duplicate birthdays do you expect?
25. In the above question, if you have 25 people, then with how much probability you hope to see at least one collision (use Markov)?
26. Sometimes, your friend forgets a few items when he leaves the house in the morning. The probabilities with which he forgets various pieces of footwear: left sock 0.2, right sock 0.1, left shoe 0.1, right shoe 0.3. Let X be the number of these items that he forgets. Give a tight bound on the probability that he forgets 3 or more items.
27. Suppose the coin is biased. It lands heads with a probability of 0.71 and tails with a probability of 0.29. How can you use this coin to mimic an unbiased coin?
28. Let a coin and two dies (4-faced & 6-faced) be unbiased. Now consider the game where you first toss the coin; if you get heads then roll the 4-faced die else, you roll the 6-faced die. If the face value is x, then you get x candies. What is the expected number of candies you hope to get?
29. Let a coin and two dies (4-faced & 6-faced) be unbiased. You are playing a game with your friend. Both of you toss the coin until there is a sequence of either HH or TH. If it is HH then you roll the 6-faced die and get the number of candies based on the face value. If it is TH then your friend gets to roll the 4-faced die and get the number of candies equal to the face value. What is the expected number of candies you hope to get?

30. Consider the following table. The row represents the salary category, and the column represents the happiness category.

Observed	Happy	Not Happy	Total
Below Average	**20**	**25**	**45**
Average	**15**	**25**	**40**
Above Average	**5**	**10**	**15**
Total	**40**	**60**	**100**

    a. With a 5% level of significance test if there is enough evidence to suggest that at least 50% of people are unhappy irrespective of their salary.
    b. Is there enough evidence in the data to suggest that with a 5% level of significance, the salary bracket is independent of an individual's unhappiness?
31. What are the advantages and disadvantages of using Bloom filters?
32. In terms of false positive and false negative what is the ideal setup to use a Bloom Filters to query a membership problem?
33. Does the False positive rate always decrease with an increasing number of hash functions in a bloom filter?
34. In what way is a Bloom filter a "probabilistic data structure"?
35. Consider a huge dataset of userid, shared between two machines M1 and M2. M1 uses one hash function, h1, to construct a bloom filter B1 for the dataset D of size m bits. M2 uses another hash function, k1 to construct a bloom filter B2 for the dataset D of size m bits.
    Let C = B1 bitwise-AND B2 and D = B1 bitwise-OR B2.
    a. Is C also a well defined bloom filter? If yes then what are (or is) its hash functions, else why this is not a bloom filter?
    b. Is D also a well defined bloom filter? If yes then what are (or is) its hash functions, else why this is not a bloom filter?
    c. If B1 uses two hash functions h1 and h2 and B2 uses two different hash functions k1 and k2, then answer a and b.

36. Let's say there are n students s1, s2, ..., sn in the course and they want to use a central server. We create a hash function, that operates on user-id i.e. student si is hashed to h(si). We plan to give access to the server in some ordering (say sorted) of the hash values returned by h(si). The hash value is of b bits and can be considered to be choosing values uniformly at random. Given two fixed users, what is the probability that they get the same hash value?
37. In the above question, what is the probability that at least one pair of students share the same hash value?
38. Let U be a set of ‘r’ unique elements, out of which we encounter ‘s’ elements such that s << r. If we construct a bloom filter of size ‘t’ bits using k hash functions h1, h2, …, hk such that its false positive is optimal, then answer the following.
    a. What is the size of t?
    b. What is the value of k?
    c. What is the optimum false positive rate?
39. Suppose that we wanted to extend Bloom filters to allO\v deletions as well as insertions of items into the underlying set. We could modify the Bloom filter to be an array of counters instead of an array of bits. Each time an item is inserted into a Bloom filter, the counters given by the hashes of the item are increased by one. To delete an item, one can simply decrement the counters. To keep space small, the counters should be a fixed length, such as 4 bits. Explain how errors can arise when using fixed-length counters. Assuming a setting where one has at most n elements in the set at any time, m counters, k hash functions, and counters with b bits, explain how to bound the probability that an error occurs over the course of t insertions or deletions.
40. What’s the TradeOff in using Bloom Filters?
41. How does Akamai prevent caching of One Hit Wonders?
42. What is load factor in hashing?
43. In hashing, what is the “power of two choices”?
44. Consider a universe U for which the goal is to design a universal hashing function that maps every element in U to a table of size m. For this, we pick a prime number p < |U| and design a family of hash functions H = $\{h_a(\ ) \mid a \in [1, 2, \ldots, p-1]\}$. A random sample of this family of hash functions is due to a random sample $a \in [1, 2, \ldots, p-1]$ such that for every element $x \in U$, $h_a(x) = (a \cdot x \bmod p) \bmod m$. Discuss the possible issues in such a hash function.
45. Define a non-negative random variable with appropriate distribution showing that Markov’s inequality is tight.
46. Chebyshev’s inequality always gives a better bound than Markov’s inequality. Prove the above statement if it is true else, give an example to counter it.

47. Using the moment generating function, define k^{th} moment of a random variable.
48. What is the Flajolet-Martin algorithm, and when is it useful?
49. If the stream of the dataset is skewed, that is, some elements appear much more frequently than others, then is FM still a good algorithm to estimate the number of unique elements. Why or why not?
50. Let U be a universe and let a variant of FM hashes h: U $\rightarrow$ [0, 1] such that if s is the largest hash value that FM algo maintains by the end of the stream, what do you return from this FM algo that well approximates the number of unique elements in the stream?
51. Let U be a universe with m unique elements. Recall that a variant of FM hashes h: U $\rightarrow$ [0, 1] such that if s is the smallest hash value that FM algo maintains by the end of the stream, then E[s] = 1/(m + 1). Then is E[1/s -1] = m?
52. Consider two separate streams s1 and s2, handled by two different machines. Let both machine uses the same hash functions in their respective FM algo where they locally maintain their bit array.
    a. Using the output of the local machines, how can we estimate the number of unique elements that are present in s1 or s2?
    b. Using the output of the local machines, how can we estimate the number of unique elements that are present in s1 and s2?
53. Each entry of a JL matrix of dimension k x d is a sample sample from N(0,1/k). If we sparsify the matrix such that 4/5 entries are 0, then from which distribution do we need to sample entries to ensure the desired properties from the sparse JL matrix?
54. What is the expected squared distance between two points generated from the surface of a unit d-dimensional sphere centered at the origin?
55. Explain how random projection can used to solve linear regression approximately but quickly.
56. For the previous question, explain the merits and demerits of your solution.
57. Let X be matrix of size 2^m x d representing 2^m points in d dimensional space. Let H_m be an 2^m x 2^m be a Hadamard matrix then such that H_1 = [1] and $H_m = \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix}$.
    a. For a fixed m, define H_m with appropriate scaling such that H_m is an orthonormal matrix.
    b. Let D be an 2^m x 2^m size diagonal matrix, such that every diagonal value is either +1 or -1 with an equal probability. Then what is the effect of DH_m on X (i.e., the 2^m points in $DH_m X$ relates with the points in X in terms of their square of L2 norm)?
58. Define the pseudo inverse of a matrix X of size n x m?

59. Let X be a dataset consisting of n points in d dimensional space and y be a vector of in R^n representing the targets for every points in the dataset. Let, the data is closely approximated by k-dimensional subspace in R^n. Using a JL matrix M of size d x klog(k) design an algorithm to quickly train a model v in R^d.
60. Let x be a random vector on the surface of a unit sphere in R^d.
    a. Prove that the matrix $E[xx^T] = I$.
    b. What is the rank of $E[xx^T]$ and xx^T and why?
61. Let $p = \{p_1, \ldots, p_n\}$ be a distribution on n points that lie on R^d. Consider X be the data matrix. Write a pseudo code to generate a matrix S that randomly selects m points out of n. Let Y be the sampled points from X using S such that for any vector v in R^d we get, $E[\|Yv\|_2^2] = \|Xv\|_2^2$
62. Write a pseudo code for CUR decomposition of a matrix A of size n x m where C has k columns and R has 2k columns.
63. When a sampling based algorithm for fast matrix multiplication is not useful.
64. Given an intuitive proof that the volume of a sphere is concentrated near the surface.
65. Recall that eigenvalue decomposition is used to compute the eigenvalues and eigenvectors of a square matrix. It takes $O(d^3)$ for a d x d square matrix. Let A be a rectangular matrix of size n x m. Using the eigenvalue subroutine design, an algorithm to compute the singular value decomposition of A in O(nm min(m,n))?
66. Let A and B be two matrices of size m x n and n x p, respectively. Let C = AB.
    a. Design an algorithm using a random projection that returns D such that E[D] = C.
    b. Propose a variant of the algorithm such that the $Var(D_{ij})$ is small, where D_{ij} is the entry corresponding to the i^{th} row and j^{th} column of D.
    c. Design an algorithm using random sampling that returns D such that E[D] = C.
    d. Derive the sampling probability such that the following holds and analyze its running time.
$$E[\|D - C\|_F^2] \leq \frac{\|A\|_F^2 \|B\|_F^2}{m}$$
    e. Design a sampling probability that gives a better bound than the above.

67. Why is the property of coreset useful for training a model?
68. To ensure the property of corset why do we rely on epsilon net?
69. Consider a data matrix X of size n x d (#points x #features) and a vector v in R^d. Define a sampling probability over n that is used to sample m points from n (say $Y \in R^{m \times d}$ be the sampled matrix) such that $E[\|Yv\|_2^2] = \|Xv\|_2^2$ and $var(\|Yv\|_2^2) = 0$
70. What is the expected squared distance between two points generated at random inside a d-dimensional cube centered at the origin having a side length of 2?