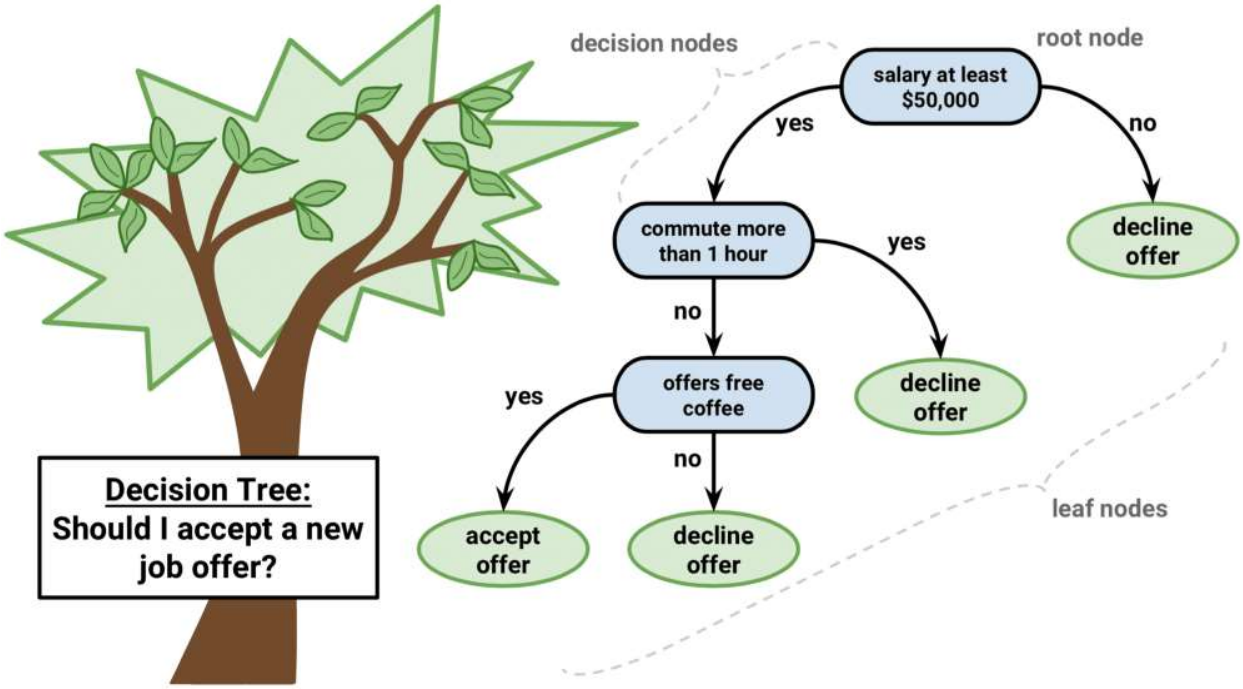
Miscellaneous

Q1. Explain the steps in making a decision tree.

1. Take the entire data set as input
2. Calculate entropy of the target variable, as well as the predictor attributes
3. Calculate your information gain of all attributes (we gain information on sorting different objects from each other)
4. Choose the attribute with the highest information gain as the root node
5. Repeat the same procedure on every branch until the decision node of each branch is finalized

For example, let's say you want to build a decision tree to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:



It is clear from the decision tree that an offer is accepted if:

- Salary is greater than $50,000
- The commute is less than an hour
- Coffee is offered

Q2. How do you build a random forest model?

A random forest is built up of a number of decision trees. If you split the data into different packages and make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

1. Randomly select k features from a total of m features where $k << m$
2. Among the k features, calculate the node D using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat steps two and three until leaf nodes are finalized
5. Build forest by repeating steps one to four for n times to create n number of trees

Q3. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Example: height of students

Height (in cm)
164
167.3
170
174.2
178
180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celsius)	**Sales (in K $)**
20	2.0
25	2.1
26	2.3
28	2.7
30	3.1

Here, the relationship is visible from the table that temperature and sales are directly proportional to each other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.

Example: data for house price prediction

The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range, minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

Q4. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection.

Filter Methods

This involves:

- Linear discrimination analysis
- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit
- Backward Selection: We test all the features and start removing them to see what works better
- Recursive Feature Elimination: Recursively looks through all the different features and how they pair together

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

Q5. In your choice of language, write a program that prints the numbers ranging from one to 50. But for multiples of three, print "Fizz" instead of the number and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz."

The code is shown below:

```
for x in range(51):

      if x % 3 == 0 and x % 5 == 0:
            print('fizzbuzz')

      elif x % 3 == 0:
            print('fizz')

      elif x % 5 == 0:
            print('buzz')

      else:
            print('fizzbuzz')
```

Q6. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way; we use the rest of the data to predict the values.

For smaller data sets, we can impute missing values with the mean, median, or average of the rest of the data using pandas data frame in python. There are different ways to do so, such as:

```
df.mean(), df.fillna(mean)
```

Other option of imputation is using KNN for numeric or classification values (as KNN just uses k closest values to impute the missing value).

Q7. For the given points, how will you calculate the Euclidean distance in Python?

```
plot1 = [1,3]
plot2 = [2,5]
```

The Euclidean distance can be calculated as follows:

```
euclidean_distance = sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)
```

Q8. What are dimensionality reduction and its benefits?

Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

Q9. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

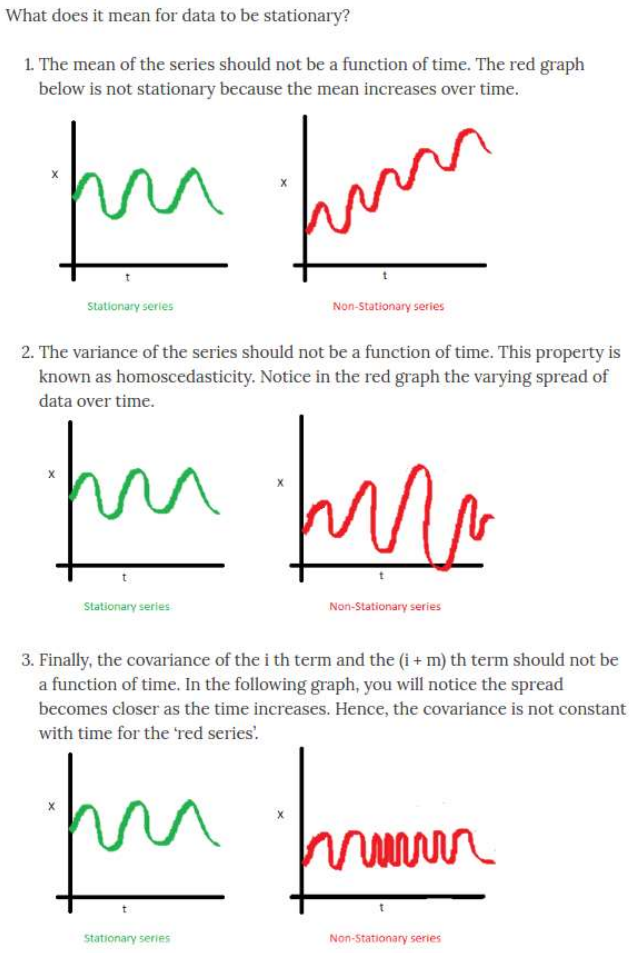
Determinant of $A - \lambda 1$ and solve to find λ.

Q10. How should you maintain a deployed model?

The steps to maintain a deployed model are (CREM):

1. **Monitor:** constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.
2. **Evaluate:** evaluation metrics of the current model are calculated to determine if a new algorithm is needed.
3. **Compare:** the new models are compared to each other to determine which model performs the best.
4. **Rebuild:** the best performing model is re-built on the current state of data.

Q11. How can a time-series data be declared as stationery?



Q12. 'People who bought this also bought...' recommendations seen on Amazon are a result of which algorithm?

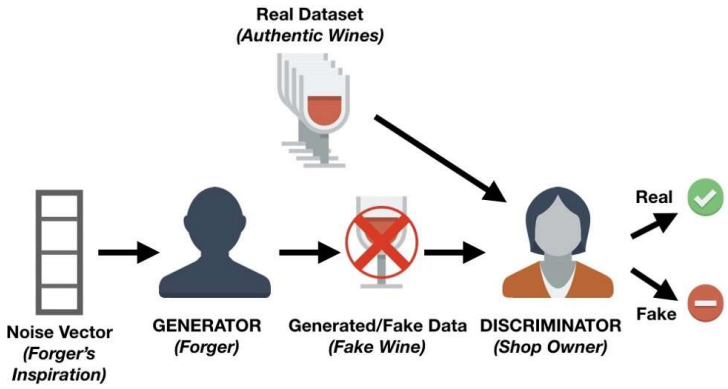
The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.
The engine makes predictions on what might interest a person based on the preferences of other users.

In this algorithm, item features are unknown.

For example, a sales page shows that a certain number of people buy a new phone and also buy tempered glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to buy tempered glass as well.

Q13. What is a Generative Adversarial Network?

Suppose there is a wine shop purchasing wine from dealers, which they resell later. But some dealers sell fake wine. In this case, the shop owner should be able to distinguish between fake and authentic wine. The forger will try different techniques to sell fake wine and make sure specific techniques go past the shop owner's check. The shop owner would probably get some feedback from wine experts that some of the wine is not original. The owner would have to improve how he determines whether a wine is fake or authentic.
The forger's goal is to create wines that are indistinguishable from the authentic ones while the shop owner intends to tell if the wine is real or not accurately.



- *There is a noise vector coming into the forger who is generating fake wine.*
- *Here the forger acts as a Generator.*
- *The shop owner acts as a Discriminator.*
- *The Discriminator gets two inputs; one is the fake wine, while the other is the real authentic wine. The shop owner has to figure out whether it is real or fake.*

So, there are two primary components of Generative Adversarial Network (GAN) named:

1. Generator
2. Discriminator

The generator is a CNN that keeps keys producing images and is closer in appearance to the real images while the discriminator tries to determine the difference between real and fake images. The ultimate aim is to make the discriminator learn to identify real and fake images.

Q14. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection and can greatly improve a patient's prognosis.

Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

Q15. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable? [0, 0, 0, 1, 1, 1, 1, 1]

The target variable, in this case, is 1 (the last)

The formula for calculating the entropy is, putting $p = 5$ and $n = 8$, we get:

$$Entropy = -\left(\frac{5}{8}\, log\left(\frac{5}{8}\right) + \frac{3}{8}\log\left(\frac{3}{8}\right)\right)$$

Q16. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case? Choose the correct option:

The most appropriate algorithm for this case is logistic regression.

Q17. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?

As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering is the most appropriate algorithm for this study.

Q18. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true? Choose the right answer:

The answer is A: {grape, apple} must be a frequent itemset.

Q19. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

One-way ANOVA: in statistics, one-way analysis of variance is a technique that can be used to compare means of two or more samples. This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or categorical input data, the "X", always one variable, hence "one-way".
The ANOVA tests the null hypothesis, which states that samples in all groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values.

Q20. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

Q21. What is root cause analysis?

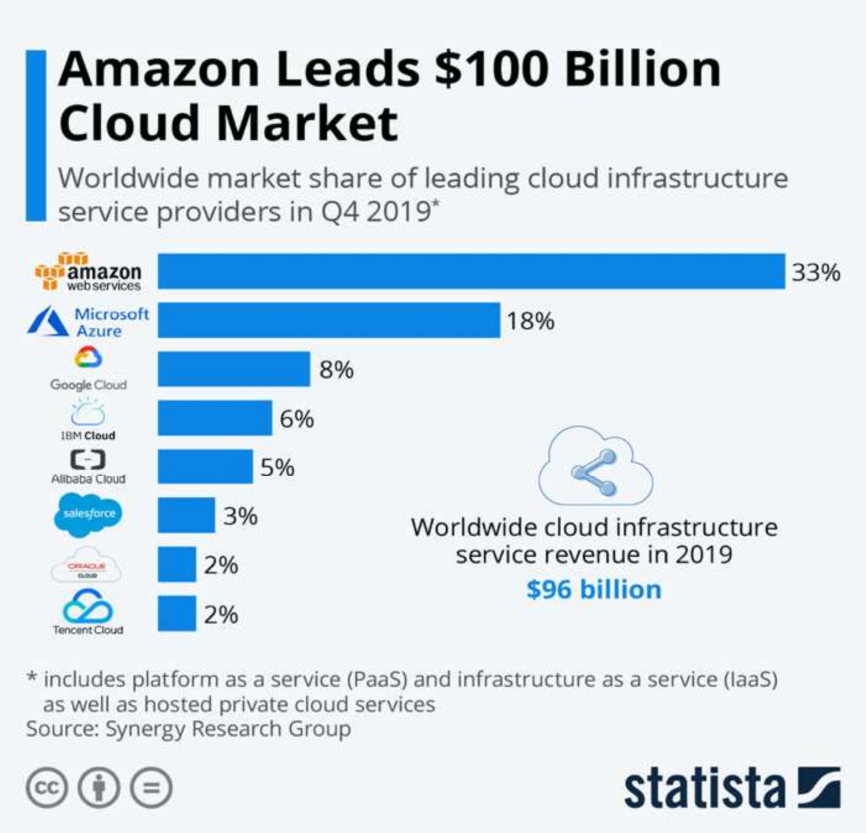
Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

Q22. Do gradient descent methods always converge to similar points?

They do not, because in some cases, they reach a local minimum or a local optimum point. You would not reach the global optimum point. This is governed by the data and the starting conditions.

Q23. What are the most popular Cloud Services used in Data Science?

https://www.zdnet.com/article/the-top-cloud-providers-of-2020-aws-microsoft-azure-google-cloud-hybrid-saas/



Q24. What is a Canary Deployment?

https://www.split.io/glossary/canary-deployment/

A canary deployment, or canary release, allows you to rollout your features to only a subset of users as an initial test to make sure nothing else in your system broke.

The initial steps for implementing canary deployment are:

1. create two clones of the production environment,
2. have a load balancer that initially sends all traffic to one version,
3. create new functionality in the other version.

When you deploy the new software version, you shift some percentage – say, 10% – of your user base to the new version while maintaining 90% of users on the old version. If that 10% reports no errors, you can roll it out to gradually more users, until the new version is being used by everyone. If the 10% has problems, though, you can roll it right back, and 90% of your users will have never even seen the problem.

Canary deployment benefits include zero downtime, easy rollout and quick rollback – plus the added safety from the gradual rollout process. It also has some drawbacks – the expense of maintaining multiple server instances, the difficult clone-or-don't-clone database decision.

Typically, software development teams implement blue/green deployment when they're sure the new version will work properly and want a simple, fast strategy to deploy it. Conversely, canary deployment is most useful when the development team isn't as sure about the new version and they don't mind a slower rollout if it means they'll be able to catch the bugs.

Q25. What is a Blue Green Deployment?

https://docs.cloudfoundry.org/devguide/deploy-apps/blue-green.html

Blue-green deployment is a technique that reduces downtime and risk by running two identical production environments called Blue and Green.

At any time, only one of the environments is live, with the live environment serving all production traffic. For this example, Blue is currently live, and Green is idle.

As you prepare a new version of your model, deployment and the final stage of testing takes place in the environment that is not live: in this example, Green. Once you have deployed and fully tested the model in Green, you switch the router, so all incoming requests now go to Green instead of Blue. Green is now live, and Blue is idle.

This technique can eliminate downtime due to app deployment and reduces risk: if something unexpected happens with your new version on Green, you can immediately roll back to the last version by switching back to Blue.