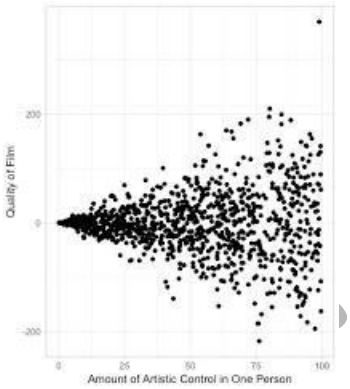
Day 3: Data Science Interview Questions

Q1. How do you treat heteroscedasticity in regression?

- Heteroscedasticity means unequal scattered distribution.
- In regression analysis, we generally talk about the heteroscedasticity in the context of the error term.
- Heteroscedasticity is the systematic change in the spread of the residuals or errors over the range of measured values.
- Heteroscedasticity is the problem because *Ordinary least squares (OLS)* regression assumes that all residuals are drawn from a random population that has a constant variance.



What causes Heteroscedasticity?

- Heteroscedasticity occurs more often in datasets, where we have a large range between the largest and the smallest observed values.
- There are many reasons why heteroscedasticity can exist, and a generic explanation is that the error variance changes proportionally with a factor.

We can **categorize Heteroscedasticity into two general types**: -

Pure heteroscedasticity: - It refers to cases where we specify the correct model and let us observe the non-constant variance in residual plots.

Impure heteroscedasticity: - It refers to cases where you incorrectly specify the model, and that causes the non-constant variance.

- When you leave an important variable out of a model, the omitted effect is absorbed into the error term.
- If the effect of the omitted variable varies throughout the observed range of data, it can produce the tell-tale signs of heteroscedasticity in the residual plots.

How to Fix Heteroscedasticity

Redefining the variables:

If your model is a cross-sectional model that includes large differences between the sizes of the observations, you can find different ways to specify the model that reduces the impact of the size differential.

To do this, change the model from using the raw measure to using rates and per capita values. Of course, this type of model answers a slightly different kind of question.

You will need to determine whether this approach is suitable for both your data and what you need to learn.

Weighted regression:

It is a method that assigns each data point to a weight based on the variance of its fitted value.
The idea is to give small weights to observations associated with higher variances to shrink their squared residuals.
Weighted regression minimizes the sum of the weighted squared residuals. When you use the correct weights, heteroscedasticity is replaced by homoscedasticity.

Q2. What is multicollinearity, and how do you treat it?

Multicollinearity means independent variables are highly correlated to each other.

In regression analysis, it is an important assumption that the regression model should not be faced with a problem of multicollinearity.

If two explanatory variables are highly correlated, it is hard to tell, which affects the dependent variable.

Let us say Y is regressed against X1 and X2 and where X1 and X2 are highly correlated. Then the effect of X1 on Y is hard to distinguish from the effect of X2 on Y because any increase in X1 tends to be associated with an increase in X2.

- Another way to look at the multicollinearity problem is: Individual t-test P values can be misleading.

- It means a P-value can be high, which means the variable is not important, even though the variable is important.

Correcting Multicollinearity:

1) Remove one of the highly correlated independent variables from the model.
2) If you have two or more factors with a high VIF, remove one from the model.

Principle Component Analysis (PCA) - It cut the number of interdependent variables to a smaller set of uncorrelated components.

Instead of using highly correlated variables, use components in the model that have eigenvalue greater than 1.

3) Ridge Regression - It is a technique for analysing multiple regression data that suffer from multicollinearity.

4) If you include an interaction term (the product of two independent variables), you can also reduce multicollinearity by "centering" the variables.

By "centering," it means subtracting the mean from the values of the independent variable before creating the products.

When is multicollinearity not a problem?

1) If your goal is to predict Y from a set of X variables, then multicollinearity is not a problem.

2) The predictions will still be accurate, and the overall R2 (or adjusted R2) quantifies how well the model predicts the Y values.

3) Multiple dummy (binary) variables that represent a categorical variable with three or more categories.

Q3. What is market basket analysis? How would you do it in Python?

Market basket analysis is the study of items that are purchased or grouped in a single transaction or multiple, sequential transactions.

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items.

It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

Understanding the relationships and the strength of those relationships is valuable information that can be used to make recommendations, cross-sell, up-sell, offer coupons, etc.

Q4. What is Association Analysis? Where is it used?

Association analysis uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction.

- The technique of association rules is widely used for retail basket analysis.
- It can also be used for classification by using rules with class labels on the righthand side.
- It is even used for outlier detection with rules indicating infrequent/abnormal association.

Association analysis also helps us to identify cross-selling opportunities, for example, we can use the rules resulting from the analysis to place associated products together in a catalog, in the supermarket, or the Webshop, or apply them when targeting a marketing campaign for product B at customers who have already purchased product A.

Association rules are given in the form as below:

A => B [Support, Confidence]

The part before => is referred to as if (**Antecedent**) and the part after => is referred to as then (**Consequent)**.

Where A and B are sets of items in the transaction data, a and B are disjoint sets.

RULE: Computer=>Anti–virus Software [Support=20%, confidence=60%]

Above rule says:

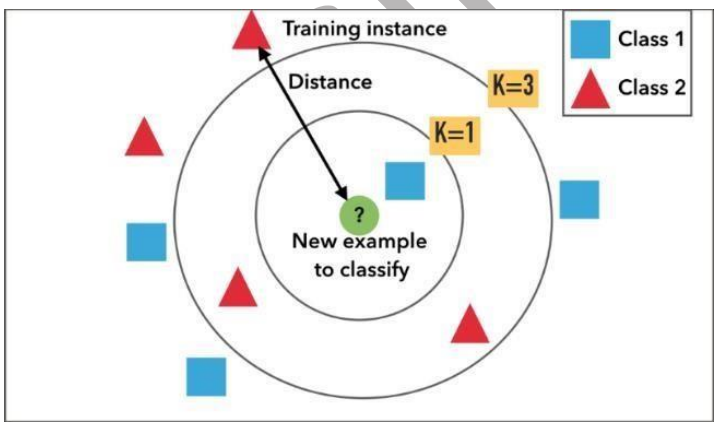
1. 20% transaction show Anti-virus software is bought with purchase of a Computer.
2. 60% of customers who purchase Anti-virus software is bought with purchase of a Computer.

An example of Association Rules * Assume there are 100 customers.

1. 10 of them bought milk, 8 bought butter and 6 bought both 2 bought milk => bought butter
2. support = P(Milk & Butter) = 6/100 = 0.06
3. confidence = support/P(Butter) = 0.06/0.08 = 0.75
4. lift = confidence/P(Milk) = 0.75/0.10 = 7.5

Q5. What is KNN Classifier?

KNN means **K-Nearest Neighbour** Algorithm. It can be used for both classification and regression.



- It is the simplest machine learning algorithm.
- Also known as **lazy learning** (why? Because it does not create a generalized model during the time of training, so the testing phase is especially important where it does the actual job.

- Hence Testing is very costly - in terms of time & money).
- Also called an instance-based or memory-based learning.

In k-NN classification, the output is a class membership.

An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).

If k = 1, then the object is assigned to the class of that single nearest neighbor.

In **k-NN regression**, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

All three distance measures are only valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used.

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

How to choose the value of K: K value is a hyperparameter which needs to choose during the time of model building.

Also, a small number of neighbors are most flexible fit, which will have a low bias, but the high variance and many neighbours will have a smoother decision boundary, which means lower variance but higher bias.

We should choose an odd number if the number of classes is even. It is said the most common values are to be 3 & 5.

Q6. What is Principal Component Analysis (PCA), and why we do?

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order.

So, in this way, the 1st principal component retains maximum variation that was present in the original components.

The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

Main important points to be considered:

1. Normalize the data.
2. Calculate the covariance matrix.
3. Calculate the eigenvalues and eigenvectors.

4. Choosing components and forming a feature vector.
5. Forming Principal Components.

Q8. What is t-SNE?

- (t-SNE) t-Distributed Stochastic Neighbor Embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data.
- It maps multi-dimensional data to two or more dimensions suitable for human observation.
- With the help of the t-SNE algorithms, you may have to plot fewer exploratory data analysis plots next time you work with high dimensional data.

Q9. VIF (Variation Inflation Factor), Weight of Evidence & Information Value. Why and when to use?

Variation Inflation Factor

It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

VIF = 1 / (1-R-Square of j-th variable), where R2 of jth variable is the coefficient of determination of the model that includes all independent variables except the jth predictor.

Where R-Square of j-th variable is the multiple R2 for the regression of Xj on the other independent variables (a regression that does not involve the dependent variable Y). If VIF > 5, then there is a problem with multicollinearity.

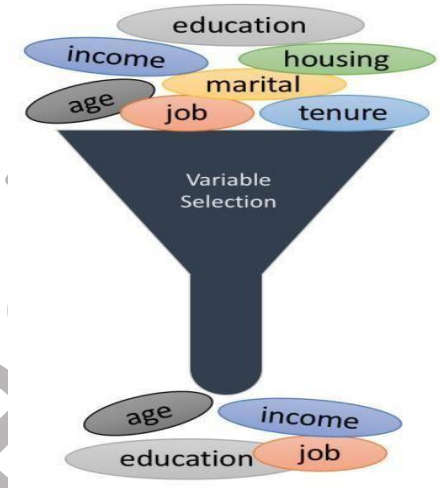
Understanding VIF

If the variance inflation factor of a predictor variable is 5 this means that variance for the coefficient of that predictor variable is 5 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

In other words, if the variance inflation factor of a predictor variable is 5 this means that the standard error for the coefficient of that predictor variable is

2.23 times ($\sqrt{5}$ = 2.23) as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

Weight of evidence (WOE) and information value (IV) are simple, yet powerful techniques to perform variable transformation and selection.



The formula to create WOE and IV is

$$WOE = \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

Variable Name	Min. Value	Max. Value	Count	# Event	# Non Event	Event%	Non event%	WOE	Event% - Non event%	IV
Age	10	20	1200	150	1050	28.3%	19.0%	0.3992	9.3%	0.03718
Age	21	30	900	120	780	22.6%	14.1%	0.4733	8.5%	0.04040
Age	31	40	1090	110	980	20.8%	17.7%	0.1580	3.0%	0.00479
Age	41	50	1460	100	1360	18.9%	24.6%	-0.2650	-5.7%	0.01517
Age	50	inf	1410	50	1360	9.4%	24.6%	-0.9582	-15.2%	0.14525
Total			**6060**	**530**	**5530**					**0.24279**

Here is a simple table that shows how to calculate these values.

The IV value can be used to select variables quickly.

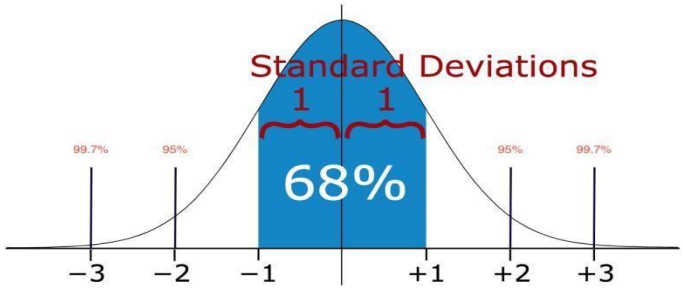
Q10: How to evaluate that data does not have any outliers?

In statistics, outliers are data points that don't belong to a certain population. It is an abnormal observation that lies far away from other values. An outlier is an observation that diverges from otherwise well-structured data.

Detection:

Method 1 — Standard Deviation:

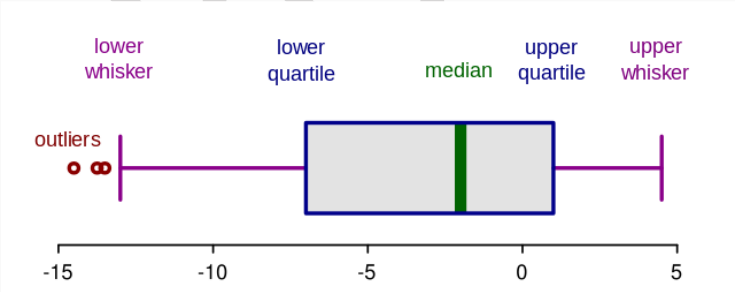
In statistics, If a data distribution is approximately normal, then about 68% of the data values lie within one standard deviation of the mean, and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations.



Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are highly likely to be anomalous or outliers.

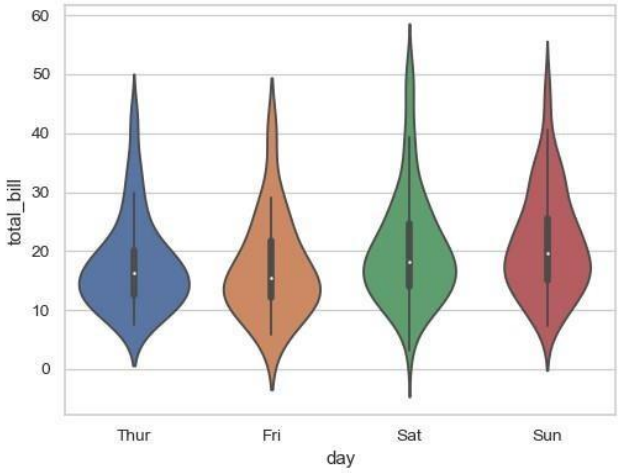
Method 2 — Boxplots: Box plots are a graphical depiction of numerical data through their quantiles. It is an amazingly simple but effective way to visualize outliers.

Think about the lower and upper whiskers as the boundaries of the data distribution. Any data points that show above or below the whiskers can be considered outliers or anomalous.

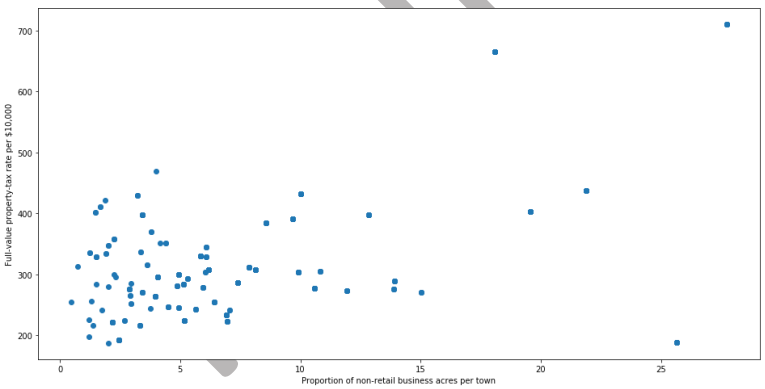


Method 3 - Violin Plots: Violin plots are similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.

Typically, a violin plot will include all the data that is in a box plot: a marker for the median of the data, a box or marker indicating the interquartile range, and possibly all sample points if the number of samples is not too high.



Method 4 - Scatter Plots: A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.



The points which are very far away from the general spread of data and have a very few neighbors are considered to be outliers.

Q11: What you do if there are outliers?

Following are the approaches to handle the outliers:

1. Drop the outlier records.
2. Assign a new value: If an outlier seems to be due to a mistake in your data, you try imputing a value.
3. If percentagewise the number of outliers is less, but when we see numbers, there are several, then, in that case, dropping them might cause a loss in insight. We should group them in that case and run our analysis separately on them.

Q12: What are the encoding techniques you have applied with Examples?

In many practical data science activities, the data set will contain categorical variables. These variables are typically stored as text values". Since machine learning is based on mathematical equations, it would cause a problem when we keep categorical variables as is.

Let us consider the following dataset of fruit names and their weights.
Some of the common encoding techniques are:

Label encoding: In label encoding, we map each category to a number or a label. The labels chosen for the categories have no relationship.

So categories that have some ties or are close to each other lose such information after encoding.

One-hot encoding: In this method, we map each category to a vector that contains 1 and 0 denoting the presence of the feature or not. The number of vectors depends on the categories which we want to keep.

For high cardinality features, this method produces a lot of columns that slows down the learning significantly.

Q13: Trade-off between bias and variances, the relationship between them.

Whenever we discuss model prediction, it is important to understand prediction errors (bias and variance). The prediction error for any machine learning algorithm can be broken down into three parts:

- Bias Error
- Variance Error
- Irreducible Error

The irreducible error cannot be reduced regardless of what algorithm is used.

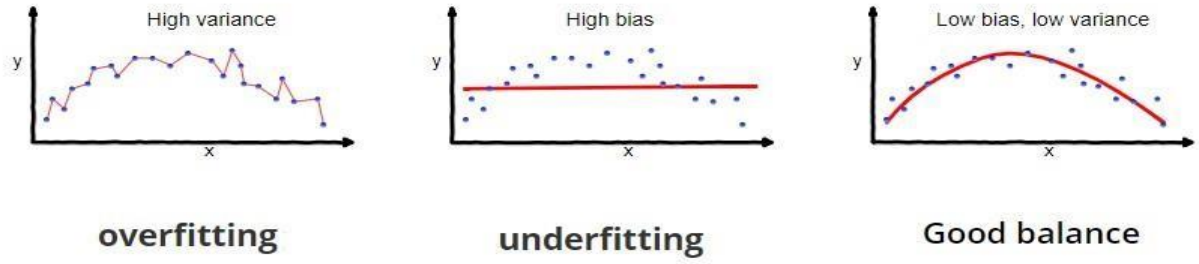
It is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable.

Bias: Bias means that the model favours one result more than the others. Bias is the simplifying assumptions made by a model to make the target function easier to learn.
The model with high bias pays extraordinarily little attention to the training data and oversimplifies the model. It always leads to a high error in training and test data.

Variance: Variance is the amount that the estimate of the target function will change if different training data was used.

The model with high variance pays a lot of attention to training data and does not generalize on the data which it has not seen before.

As a result, such models perform very well on training data but have high error rates on test data.



overfitting underfitting Good balance

So, the end goal is to come up with a model that balances both Bias and Variance. This is called *Bias Variance Trade-off*. To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Q14: What is the difference between Type 1 and Type 2 error and severity of the error?

Type I Error

A Type I error is often referred to as a “false positive" and is the incorrect rejection of the true null hypothesis in favour of the alternative.

In the example above, the null hypothesis refers to the natural state of things or the absence of the tested effect or phenomenon, i.e., stating that the patient is HIV negative.

The alternative hypothesis states that the patient is HIV positive. Many medical tests will have the disease they are testing for as the alternative hypothesis and the lack of that disease as the null hypothesis.

A Type I error would thus occur when the patient does not have the virus, but the test shows that they do. In other words, the test incorrectly rejects the true null hypothesis that the patient is HIV negative.

Type II Error

A Type II error is the inverse of a Type I error and is the false acceptance of a null hypothesis that is not true, i.e., a false negative.

A Type II error would entail the test telling the patient they are free of HIV when they are not.

Considering this HIV example, which error type do you think is more acceptable? In other words, would you rather have a test that was more prone to Type I or Types II error?

With HIV, the momentary stress of a false positive is likely better than feeling relieved at a false negative and then failing to take steps to treat the disease. Pregnancy tests, blood tests, and any diagnostic tool that has serious consequences for the health of a patient are usually overly sensitive for this reason – they should err on the side of a false positive.

But in most fields of science, Type II errors are seen as less serious than Type I errors. With the Type II error, a chance to reject the null hypothesis was lost, and no conclusion is inferred from a non-rejected null. But the Type I error is more serious because you have wrongly rejected the null hypothesis and ultimately made a claim that is not true. In science, finding a phenomenon where there is none is more egregious than failing to find a phenomenon where there is.

Q15: What is binomial distribution and polynomial distribution?

Binomial Distribution: A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.

The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Multinomial/Polynomial Distribution: Multi or Poly means many. In probability theory, the multinomial distribution is a generalization of the binomial distribution. For example, it models the probability of counts of each side for rolling a k-sided die n time.

For an independent trial each of which leads to success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

Q16: What is the Mean Median Mode standard deviation for the sample and population?

Mean It is an important technique in statistics.

Arithmetic Mean can also be called an average. It is the number of the quantity obtained by summing two or more numbers/variables and then dividing the sum by the number of numbers/variables.

Mode The mode is also one of the types for finding the average. A mode is a number that occurs most frequently in a group of numbers.

Some series might not have any mode; some might have two modes, which is called a bimodal series.
In the study of statistics, the three most common 'averages' in statistics are mean, median, and mode.

Median is also a way of finding the average of a group of data points.
It is the middle number of a set of numbers.

There are two possibilities, the data points can be an odd number group, or it can be an even number group.

If the group is odd, arrange the numbers in the group from smallest to largest. The median will be the one which is exactly sitting in the middle, with an equal number on either side of it.

If the group is even, arrange the numbers in order and pick the two middle numbers and add them then divide by 2. It will be the median number of that set.

Standard Deviation (Sigma) Standard Deviation is a measure of how much your data is spread out in statistics.

Q17: What is Mean Absolute Error?

What is Absolute Error? Absolute Error is the amount of error in your measurements. It is the difference between the measured value and the "true" value.

For example, if a scale states 90 pounds, but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.

This can be caused by your scale, not measuring the exact amount you are trying to measure. For example, your scale may be accurate to the nearest pound.

If you weigh 89.6 lbs, the scale may "round up" and give you 90 lbs. In this case the absolute error is 90 lbs – 89.6 lbs = .4 lbs.

Mean Absolute Error: The Mean Absolute Error (MAE) is the average of all absolute errors. The formula is: mean absolute error.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

Where, n = the number of errors, Σ = summation symbol (which means “add them all up”),
|xi – x| = the absolute errors.

The formula may look a little daunting, but the steps are easy:
Find all your absolute errors, xi – x.

Add them all up.

Divide by the number of errors.

For example, if you had 10 measurements, divide by 10.

Q18: What is the difference between long data and wide data?

There are many ways that you can present the same dataset to the world.

Let us look at one of the most important and fundamental distinctions, whether a dataset is wide or long.

The difference between wide and long datasets boils down to whether we prefer to have more columns in our dataset or more rows.

Wide Data A dataset that emphasizes putting additional data about a single subject in columns is called a wide dataset because, as we add more columns, the dataset becomes wider.

Long Data Similarly, a dataset that emphasizes including additional data about a subject in rows is called a long dataset because, as we add more rows, the dataset becomes longer.

It is important to point out that there's nothing inherently good or bad about wide or long data.

In the world of data wrangling, we sometimes need to make a long dataset wider, and we sometimes need to make a wide dataset longer.

However, it is true that, as a rule, data scientists who embrace the concept of tidy data usually prefer longer datasets over wider ones.

Q19: What are the data normalization method you have applied, and why?

Normalization is a technique often applied as part of data preparation for machine learning.

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

In simple words, when multiple attributes are there, but attributes have values on different scales, this may lead to poor data models while performing data mining operations.

so they are normalized to bring all the attributes on the same scale, usually something between (0,1).

It is not always a good idea to normalize the data since we might lose information about maximum and minimum values. Sometimes it is a good idea to do so.

For example, ML algorithms such as Linear Regression or Support Vector Machines typically converge faster on normalized data.

But on algorithms like K-means or K Nearest Neighbours, normalization could be a good choice or a bad depending on the use case since the distance between the points plays a key role here.

person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

Types of Normalisation:

1 Min-Max Normalization: In most cases, standardization is used feature-wise.

$$\hat{X}[:,i] = \frac{X[:,i] - \min(X[:,i])}{\max(X[:,i]) - \min(X[:,i])}$$

2 Z-score normalization In this technique, values are normalized based on a mean and standard deviation of the data

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

v', v is new and old of each entry in data respectively. σA, A is the standard deviation
and mean of A respectively.

standardization (or Z-score normalization) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with μ=0 and σ=1 where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows: $z=(x-\mu)/\sigma$

Q20: What is the difference between normalization and Standardization with example?

In ML, every practitioner knows that feature scaling is an important issue. The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically means it rescales the values into a range of [0,1].

It is an alternative approach to Z-score normalization (or standardization) is the socalled Min-Max scaling (often also called "normalization" - a common cause for ambiguities). In this approach, the data is scaled to a fixed range - **usually 0 to 1**. Scikit-Learn provides a transformer called **MinMaxScaler** for this. A Min-Max scaling is typically done via the following equation:

Xnorm = X-Xmin/Xmax-Xmin

Example with sample data: Before Normalization: Attribute Price in Dollars Storage Space Camera

- Attribute Price in Dollars Storage Space Camera
- Mobile 1 250 16 12
- Mobile 2 200 16 8
- Mobile 3 300 32 16
- Mobile 4 275 32 8
- Mobile 5 225 16 16

After Normalization: (Values ranges from 0-1 which is working as expected)

- Attribute Price in Dollars Storage Space Camera
- Mobile 1 0.5 0 0.5
- Mobile 2 0 0 0
- Mobile 3 1 1 1
- Mobile 4 0.75 1 0 • Mobile 5 0.25 0 1

Standardization (or Z-score normalization) typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance) Formula: **Z or X_new=(x−μ)/σ** where μ is the mean (average), and σ is the standard deviation from the mean; standard scores (also called z scores) Scikit-Learn provides a transformer called StandardScaler for standardization **Example:** Let's take an approximately normally distributed set of numbers: 1, 2, 2, 3, 3, 3, 4, 4, and 5. Its

mean is 3, and its standard deviation: 1.22. Now, let's subtract the mean from all data points. we get a new data set of: -2, -1, -1, 0, 0, 0, 1, 1, and 2. Now, let's divide each data point by 1.22. As you can see in the picture below, we get: -1.6, -0.82, -0.82, 0, 0, 0, 0.82, 0.82, and 1.63