**why do we need to fit ARIMA models on time series data?**

ARIMA (AutoRegressive Integrated Moving Average) models are used to forecast future values in a time series data by analyzing patterns and trends in historical data. They are particularly useful in analyzing and forecasting data with temporal dependencies, such as sales data, stock prices, and weather patterns. By fitting an ARIMA model to a time series, we can identify the underlying patterns and trends in the data, and use these to make more accurate predictions about future values. Additionally, ARIMA models can also be used to identify and account for seasonality and other temporal dependencies in the data, which can improve the accuracy of the forecasts.

**What are we trying to identify when we try to get the statistical summary of time series models such as ARIMA in python**

When obtaining the statistical summary of an ARIMA model in Python, the main things that are being identified are the coefficients of the model. Specifically, the summary will typically include the estimated values of the coefficients for the autoregressive (AR) component, the differencing (I) component, and the moving average (MA) component of the model. The summary may also include other statistics such as the standard errors of the coefficients, the t-values, and the p-values, which are used to determine the significance of the coefficients. Additionally, the summary may also include information about the goodness of fit of the model, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

In summary, statistical summary of ARIMA models gives information about the coefficients of the model and some measures of goodness of fit that allows to evaluate the model's performance.

**What characteristics of the coefficients or objective do we try to identify and achieve when we try to fit an ARIMA model to get the best fit**

When fitting an ARIMA model to a time series, there are several characteristics of the coefficients or objectives that we aim to identify and achieve in order to get the best fit:

1. Stationarity: We aim to identify and achieve a stationary time series, which means that the mean, variance, and autocovariance are constant over time. A stationary time series is a necessary assumption for an ARIMA model to work well.
2. Autocorrelation: We aim to identify and achieve a time series that has low autocorrelation. Autocorrelation is the correlation of a time series with its own lags. A time series with low autocorrelation is a desirable characteristic for an ARIMA model because it means that the past values of the time series are not highly correlated with the current value.
3. Model order: We aim to identify and achieve the appropriate model order (p,d,q) which is the number of lags in the autoregressive component (p), the number of times the time series is differenced (d), and the number of lags in the moving average component (q).
4. Coefficients: We aim to identify and achieve coefficients that are significant and have the smallest absolute values possible. The smaller the absolute value of the coefficient, the less weight the past values of the time series have on the current value.
5. Goodness-of-fit: We aim to achieve a model with a high goodness-of-fit, which means that the model fits the data well. This can be assessed using evaluation metrics such as AIC, BIC, and Mean Squared Error (MSE).

Ultimately, the goal of fitting an ARIMA model to time series data is to find a set of parameters that best explain the underlying patterns and trends of the data, and that can be used to make accurate predictions about future values.

**what is augmented dickey fuller test. How and why is it used**

The Augmented Dickey-Fuller test (ADF) is a statistical test used to determine whether a time series is stationary. Stationarity is an important assumption for many time series models, including ARIMA models, so it is necessary to check for stationarity before fitting such models.

The ADF test is based on the idea that a time series that is stationary will have a constant mean and a constant variance over time. The test statistic is calculated by regressing the time series on its own lags and a constant term, and then testing the null hypothesis that the coefficients on the lags of the time series are not different from zero. If the null hypothesis is rejected, it suggests that the time series is not stationary.

The ADF test is useful because it helps us determine whether a time series is stationary or not and it can be used to check the stationarity of time series data before fitting an ARIMA model. If the time series is not stationary, then differencing may be applied to make it stationary and then an ARIMA model can be fitted.

It's worth noting that the ADF test is a parametric test and makes certain assumptions about the underlying data distribution, so it might not be suitable or powerful for all types of data. Also, a weak or non-stationary time series might still be able to be modeled by some models or after some transformations or decomposition.

**weak vs non-stationary time series**

A weakly stationary time series is one where the mean and variance are constant over time, but the covariance between observations may change. A non-stationary time series, on the other hand, is one where the mean, variance, and/or covariance are not constant over time.

A weakly stationary time series can be transformed into a stationary one by applying a mathematical transformation such as a log or square root transformation. However, non-stationary time series cannot be made stationary using these types of transformations.

For non-stationary time series, differencing is often applied to make the time series stationary. Differencing involves subtracting the value of the time series at a previous point in time from the current value, creating a new time series that represents the changes in the original time series over time. This can be useful in removing trend or seasonality.

It's worth noting that a weakly stationary time series is often considered to be good enough for many models and purposes. For example, for modeling purposes, weak stationarity is sufficient for many linear models, and for forecasting, many models are able to handle weak stationarity by adjusting the initial values.

**ACF vs PACF when choosing the model order**

When fitting an ARIMA model to a time series, one important step is to determine the appropriate model order (p, d, q), which refers to the number of lags in the autoregressive (AR) component (p), the number of times the time series is differenced (d), and the number of lags in the moving average (MA) component (q) respectively.

Both the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are used to determine the appropriate model order.

The ACF is a plot of the correlation between a time series and lags of itself. It can be used to determine the order of the MA component (q) of the model. The MA component models the error term of the model, and the order of the MA component should be chosen to be the lag at which the ACF drops off to zero.

On the other hand, the PACF is a plot of the correlation between a time series and lags of itself, with the effect of the intermediate lags removed. It can be used to determine the order of the AR component (p) of the model. The order of the AR component should be chosen to be the lag at which the PACF drops off to zero.

In summary, ACF gives information about the MA component of the model, and PACF gives information about the AR component of the model. Therefore, the ACF and PACF plots can be used together to help determine the appropriate model order (p, d, q) for an ARIMA model.

**what is a time series with autocorrelation**

A time series with autocorrelation is a time series where the values at different time steps are correlated with one another. Autocorrelation, also known as serial correlation, is a measure of the similarity between a time series and a lagged version of itself. A time series with high autocorrelation means that the values at a given point in time are strongly related to the values at a previous point in time. This can make it difficult to model the time series because it suggests that past values of the time series are informative about future values.

Examples of time series with autocorrelation are stock prices, temperature, and sales data, where the value at a given point in time is likely to be similar to the value at a previous point in time. Autocorrelation can be positive or negative, positive autocorrelation indicates that a high value at time t is likely to be followed by a high value at time t+1, and negative autocorrelation indicates that a high value at time t is likely to be followed by a low value at time t+1.

Autocorrelation can be measured using the Autocorrelation Function (ACF) plot, which plots the correlation between a time series and lags of itself. If the plot shows high correlation at the first few lags, this indicates that the time series has high autocorrelation.

In summary, a time series with autocorrelation is a time series where the values at different time steps are correlated with one another, which can make it difficult to model. Autocorrelation can be measured using the Autocorrelation Function (ACF) plot.

**AIC and BIC in time series models**

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are both model selection criteria that are used to compare different time series models and select the best one.

AIC is a measure of the relative quality of a statistical model, relative to other models. It is based on the likelihood of the data given the model, and it also has a penalty term for the number of parameters in the model. The model with the smallest AIC value is considered to be the best one.

BIC is similar to AIC but it has a stronger penalty term for the number of parameters in the model, which makes it more likely to select simpler models. Like AIC, the model with the smallest BIC value is considered to be the best one.

In time series models, AIC and BIC are often used to compare different ARIMA models with different orders (p, d, q) and select the best one. The model with the lowest AIC or BIC value is considered to be the best one, as it balances the goodness of fit with the complexity of the model.

It's worth noting that AIC and BIC are asymptotic approximation, meaning that they work well when sample size is large, but in small sample sizes, these criteria might not be as reliable. Also, AIC tends to favor models with more parameters, while BIC tends to favor simpler models with fewer parameters.

In summary, AIC and BIC are both model selection criteria that are used to compare different time series models and select the best one, they balance the goodness of fit with the complexity of the model. AIC and BIC are asymptotic approximation, meaning that they work well when sample size is large, but in small sample sizes, these criteria might not be as reliable.

**Diagnostic summary statistics: Ljung-Box and Jarque-Bera (JB)**

The Ljung-Box and Jarque-Bera (JB) tests are diagnostic summary statistics that are used to check the residuals of a time series model for normality and independence.

The Ljung-Box test is a statistical test for checking the independence of a time series. It tests the null hypothesis that the autocorrelations of the residuals of a time series model are equal to zero up to a certain lag. If the p-value of the test is less than a significance level (e.g. 0.05), it suggests that the residuals are not independent and the model may be misspecified.

The Jarque-Bera test is a statistical test for checking the normality of a time series. It tests the null hypothesis that the residuals of a time series model are normally distributed. If the p-value of the test is less than a significance level, it suggests that the residuals are not normally distributed and the model may be misspecified.

Both Ljung-Box and Jarque-Bera tests are widely used in time series analysis. They are useful in identifying any patterns or issues in the residuals that may indicate that the model is not capturing all the important features of the data.

In summary, Ljung-Box and Jarque-Bera tests are diagnostic summary statistics that are used to check the residuals of a time series model for normality and independence, respectively. They are useful in identifying any patterns or issues in the residuals that may indicate that the model is not capturing all the important features of the data.

**The Box-Jenkins method**

The Box-Jenkins method, also known as the ARIMA (AutoRegressive Integrated Moving Average) method, is a systematic approach for identifying, fitting, and forecasting time series data. The method was developed by George Box and Gwilym Jenkins in the 1970s and is widely used for time series analysis and forecasting.

The Box-Jenkins method is a three-step process:

1. Identification: In this step, the goal is to identify the underlying patterns and trends in the time series data, such as trend, seasonality, and cyclical patterns. This is done by examining plots of the data and by using diagnostic statistics such as the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.
2. Estimation: In this step, the goal is to estimate the parameters of the model that best fit the data. This is done by fitting different ARIMA models with different orders (p, d, q) and selecting the best model based on the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).
3. Diagnostics: In this step, the goal is to diagnose and check the residuals of the model for any remaining patterns or issues that may indicate that the model is not capturing all the important features of the data. This is done by using diagnostic statistics such as the Ljung-Box test and Jarque-Bera test, and by examining plots of the residuals.

Once the best model is identified and any issues with the residuals are addressed, the model can be used to forecast future values of the time series data.

In summary, the Box-Jenkins method is a systematic approach for identifying, fitting, and forecasting time series data. It is a three-step process that involves identifying underlying patterns and trends in the data, estimating the parameters of the model that best fit the data, and diagnosing and checking the residuals of the model for any remaining issues.

**SARIMA**

SARIMA, or Seasonal AutoRegressive Integrated Moving Average, is a variant of the ARIMA model that is specifically designed to handle time series data with a strong seasonal component. A seasonal component refers to a repeating pattern within the data that occurs at regular, predictable intervals, such as daily, weekly, or yearly.

SARIMA models include additional parameters to account for the seasonal component of the time series. The seasonal component is modeled using a seasonal differencing term (D), a seasonal autoregressive term (P), and a seasonal moving average term (Q). These parameters are in addition to the non-seasonal parameters of an ARIMA model.

The process of fitting a SARIMA model is similar to that of fitting an ARIMA model: first, the model is identified using plots of the data and diagnostic statistics such as the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Then the model is estimated by fitting different SARIMA models with different orders (p, d, q, P, D, Q) and selecting the best model based on the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Finally, diagnostic statistics are used to check the residuals of the model for any remaining patterns or issues that may indicate that the model is not capturing all the important features of the data.

In summary, SARIMA is a variant of the ARIMA model that is specifically designed to handle time series data with a strong seasonal component. It includes additional parameters to account for the seasonal component of the time series, and the process of fitting a SARIMA model is similar to that of fitting an ARIMA model.

**SARIMA vs SARIMAX**

SARIMA and SARIMAX are both seasonal variants of the ARIMA model and are used for modeling and forecasting time series data with a strong seasonal component. The main difference between the two is that SARIMAX allows for the modeling of exogenous variables, which are variables that are not part of the time series being modeled but may have an effect on it.

SARIMA models include parameters for modeling the seasonal component of the time series, such as the seasonal differencing term (D), seasonal autoregressive term (P), and seasonal moving average term (Q). But these models do not include any exogenous variables.

SARIMAX models, on the other hand, include the same parameters as SARIMA models and also allow for the modeling of exogenous variables. These variables can be incorporated into the model as additional regressors, and they can be used to improve the forecasting accuracy of the model. Exogenous variables can be external factors such as weather, economic indicators, and other time series that are believed to have an effect on the time series being modeled.

In summary, the main difference between SARIMA and SARIMAX is that SARIMAX allows for the modeling of exogenous variables, which are variables that are not part of the time series being modeled but may have an effect on it. SARIMA models are useful in modeling the seasonal component of the time series, while SARIMAX models are useful in modeling both the seasonal component and the influence of exogenous variables on the time series, which can lead to more accurate forecasting. In practice, SARIMAX can be more appropriate in cases where there are other factors that might be influencing the time series data. However, it is worth noting that incorporating exogenous variables into the model can make it more complex and harder to interpret.

**Box Jenkins method in ARIMA and SARIMA**

The Box-Jenkins method is a systematic approach for identifying, fitting, and forecasting time series data using ARIMA or SARIMA models. The method is the same for both ARIMA and SARIMA, but the only difference is that SARIMA models include additional parameters to account for the seasonal component of the time series.

The Box-Jenkins method is a three-step process:

1. Identification: In this step, the goal is to identify the underlying patterns and trends in the time series data, such as trend, seasonality, and cyclical patterns. This is done by examining plots of the data and by using diagnostic statistics such as the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.
2. Estimation: In this step, the goal is to estimate the parameters of the model that best fit the data. This is done by fitting different ARIMA or SARIMA models with different orders (p, d, q and P, D, Q for SARIMA respectively) and selecting the best model based on the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).
3. Diagnostics: In this step, the goal is to diagnose and check the residuals of the model for any remaining patterns or issues that may indicate that the model is not capturing all the important features of the data. This is done by using diagnostic statistics such as the Ljung