

DAY 7: DATA SCIENCE INTERVIEW PREPARATION

Q1. What is the process to make data stationary from non-stationary in time series?

Ans: The two most common ways to make a non-stationary time series stationary are:

- Differencing
- Transforming

Let us look at some details for each of them:

Differencing:

To make your series stationary, you take a difference between the data points. So let us say, your original time series was:

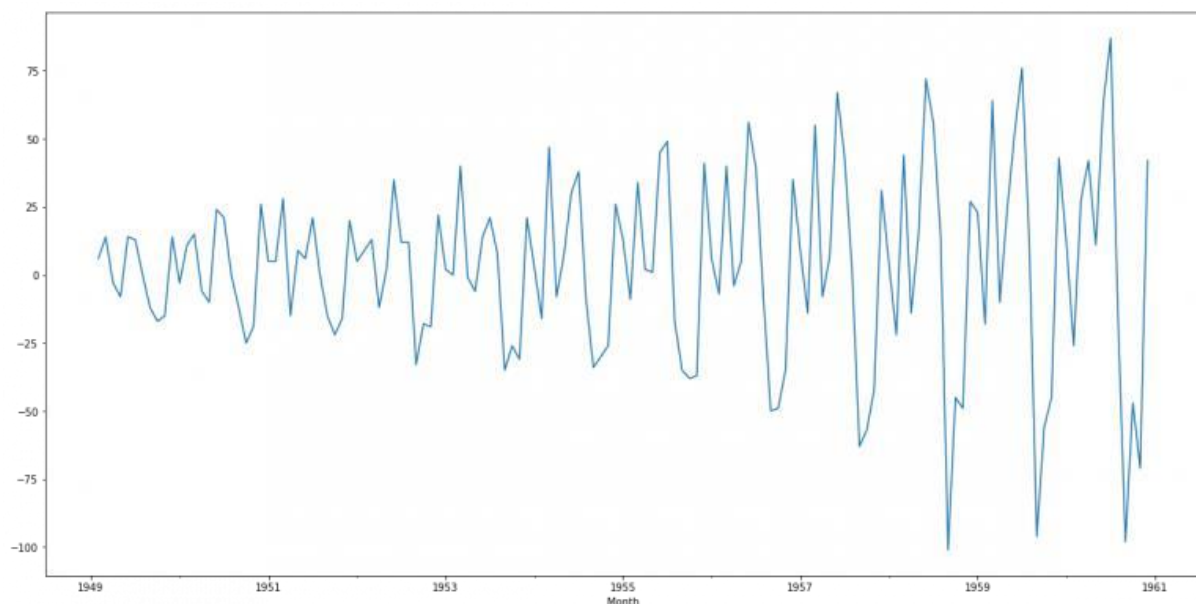
$X_1, X_2, X_3, \dots, X_n$

Your series with a difference of degree 1 becomes:

$(X_2 - X_1, X_3 - X_2, X_4 - X_3, \dots, X_n - X_{n-1})$

Once, you make the difference, plot the series, and see if there is any improvement in the ACF curve.

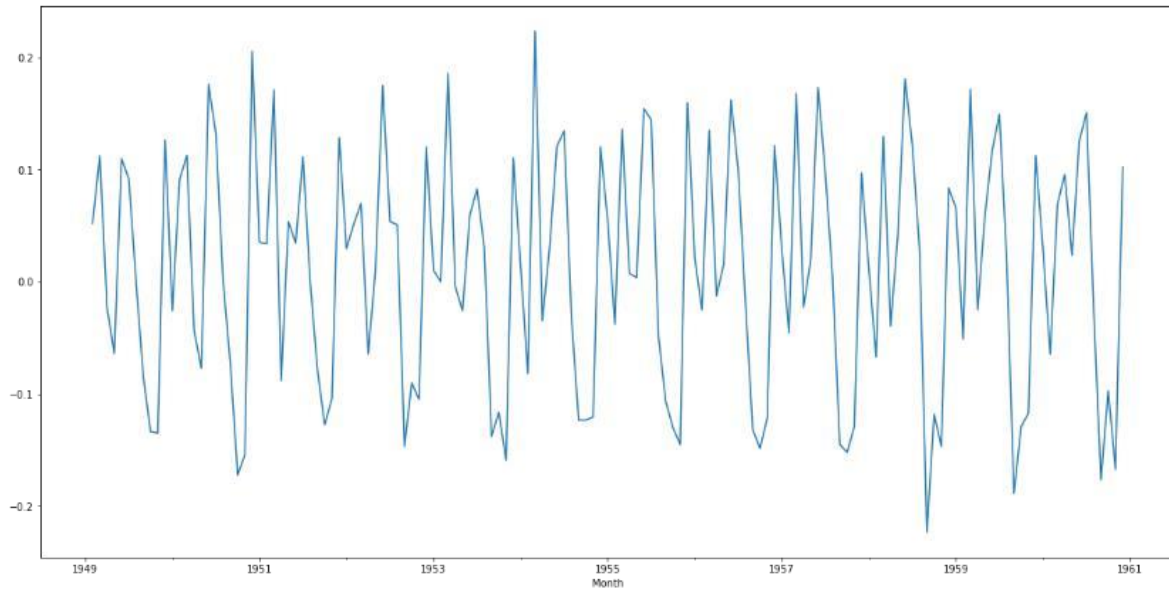
If not, you can try a second or even a third-order differencing. Remember, the more you difference, the more complicated your analysis is becoming.



Transforming:

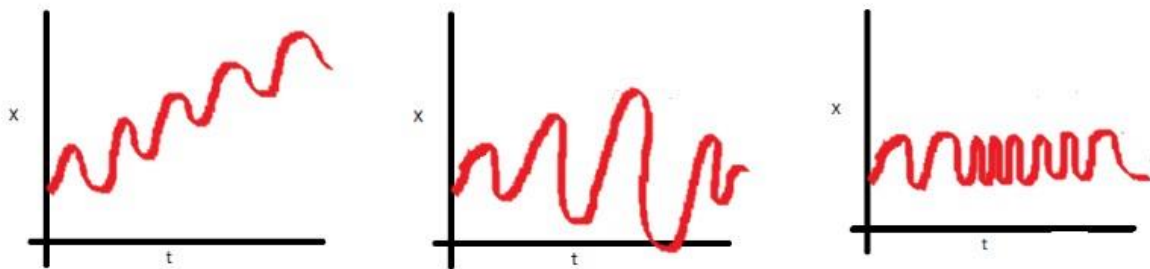
If we cannot make a time series stationary, you can try out transforming the variables. Log transform is probably the most used transformation if we see the diverging time series.

However, it is suggested that you use transformation only in case differencing is not working.



Q2. What is the process to check stationary data?

Ans: Stationary series: It is one in which the properties – mean, variance and covariance, do not vary with time.

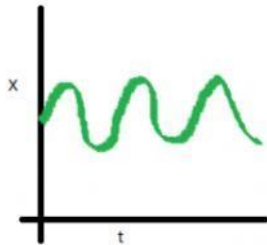


Let us get an idea with these three plots:

- In the first plot, we can see that the mean varies (increases) with time, which results in an upward trend. This is the non-stationary series.
For the series classification as stationary, it should not exhibit the trend.
- Moving on to the second plot, we do not see a trend in the series, but the variance of the series is a function of time. As mentioned previously, a stationary series must have a constant variance.

- If we look at the third plot, the spread becomes closer, as the time increases, which implies that covariance is a function of time.

These three plots refer to the non-stationary time series. Now give your attention to fourth:



In this case, Mean, Variance and Covariance are constant with time. This is how a stationary time series looks like.

Most of the statistical models require the series to be stationary to make an effective and precise prediction.

The various process you can use to find out your data is stationary or not by the following terms:

- Visual Test
- Statistical Test
- ADF (Augmented Dickey-Fuller) Test
- KPSS(Kwiatkowski-Phillips-Schmidt-Shin) Test

Q3. What are ACF and PACF?.

Ans: ACF is a (complete) auto-correlation function which gives us the values of the autocorrelation of any series with lagged values.

We plot these values along with a confidence band.

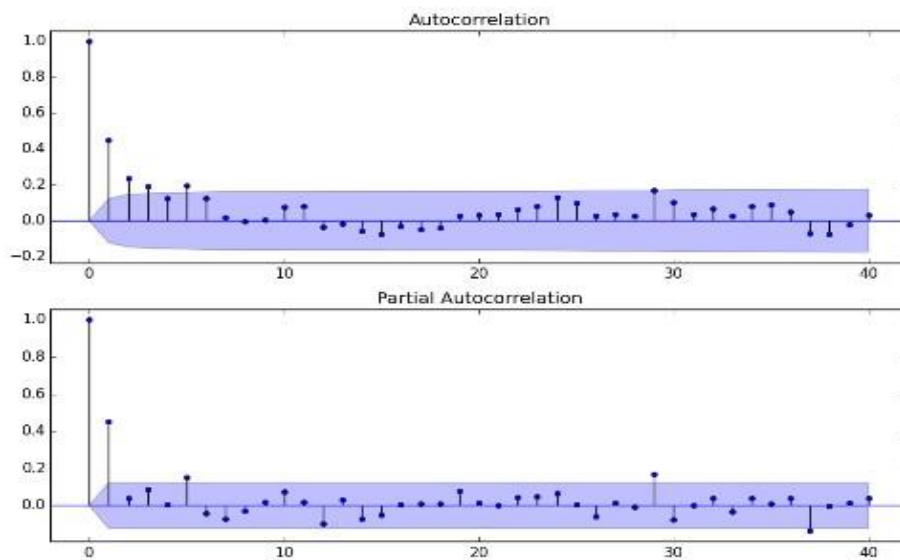
We have an ACF plot. In simple terms, it describes how well the present value of the series is related to its past values.

A time series can have components like the trend, seasonality, cyclic and residual.

ACF considers all the components while finding correlations; hence, it is a 'complete auto-correlation plot'.

PACF is a partial autocorrelation function. Instead of finding correlations of present with lags like

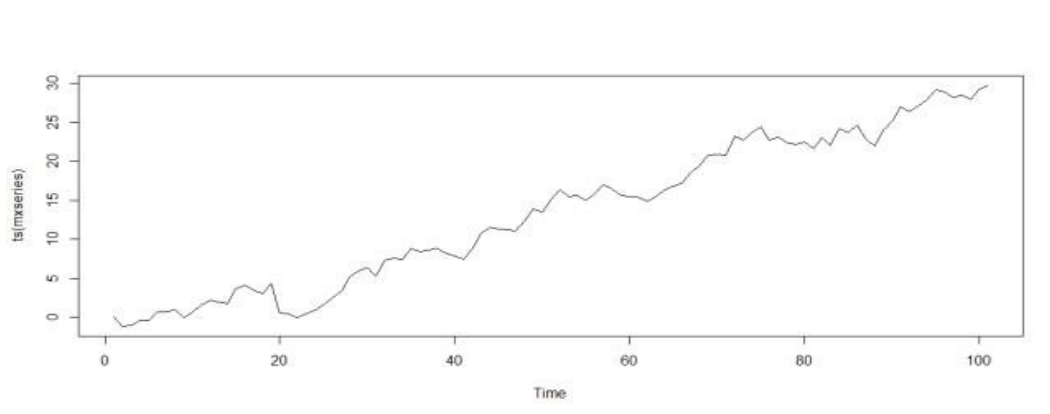
ACF, it finds the correlations of the residuals with the next lag value thus 'partial' and not 'complete' as we remove already found variations before we find next correlation. So, if there are any hidden pieces of information in the residual which can be modelled by next lag, we might get a good correlation, and we'll keep that next lag as a feature while modelling. Remember, while modelling we don't want to keep too many correlated features, as that it can create multicollinearity issues. Hence, we need to retain only relevant features.



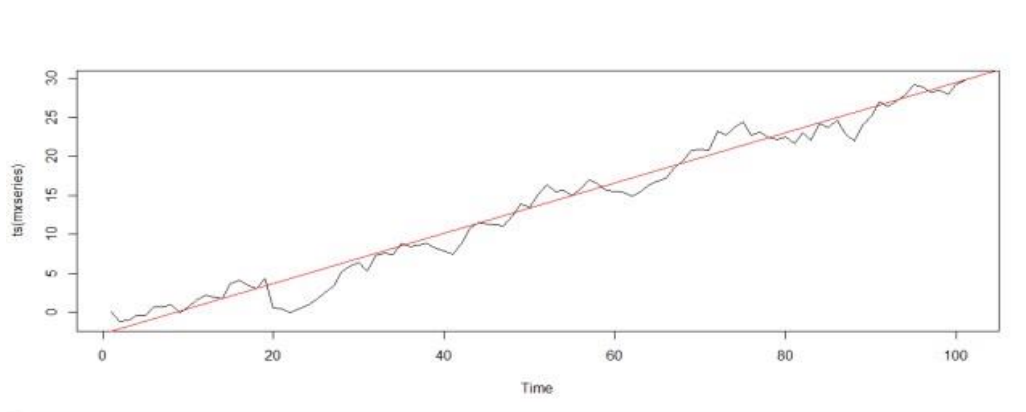
Q4. What do you understand by the trend of data?

Ans: A general systematic linear or (most often) nonlinear component that changes over time and does not repeat.

There are different approaches to understanding trend. A positive trend means it is likely that growth continues. Let us illustrate this with a simple example:



Hmm, this looks like there is a trend. To build up confidence, let's add a linear regression for this graph:



Great, now it is clear there's a trend in the graph by adding Linear Regression.

Q5. What is the Augmented Dickey-Fuller Test?

Ans: The Dickey-Fuller test: It is one of the most popular statistical tests. It is used to determine the presence of unit root in a series, and hence help us to understand if the series is stationary or not.

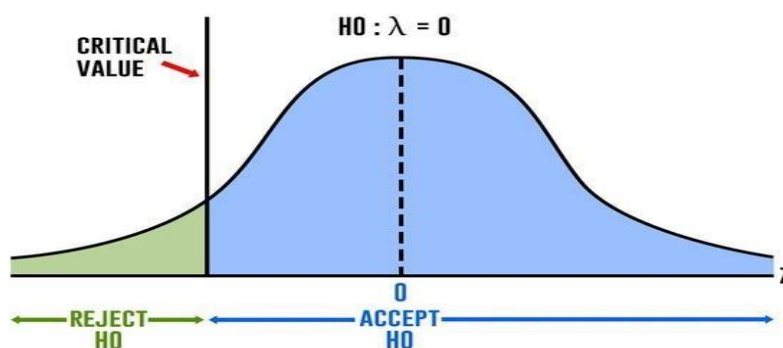
The null and alternate hypothesis for this test is:

Null Hypothesis: The series has a unit root.

(value of $A=1$)

Alternate Hypothesis: The series has no unit root.

If we fail to reject the null hypothesis, we can say that the series is non-stationary. This means that the series can be linear or difference stationary.



Q6. What is AIC and BIC into time series?

Ans:

Akaike's information criterion (AIC) compares the quality of a set of statistical models to each other.

For example, you might be interested in what variables contribute to low socioeconomic status and how the variables contribute to that status.

Let's say you create several regression models for various factors like education, family size, or disability status; The AIC will take each model and rank them from best to worst.

The “best” model will be the one that neither under-fits nor overfits.

- AIC
- K = number of estimated parameters in the model
- L = Maximized likelihood function for the estimated model

$$AIC = 2k - 2 \ln(L)$$

The Bayesian Information Criterion (BIC) can be

defined as: $k \log(n) - 2 \log(L(\theta))$.

Here n is the sample size.

K is the number of parameters which your model estimates.

θ is the set of all parameter.

$L(\theta)$ represents the likelihood of the model tested, when evaluated at maximum likelihood values of θ .

Q7. What are the components of the Time -Series?

Ans: Time series analysis: It provides a body of techniques to understand a dataset better. The most useful one is the decomposition of the time series into four constituent parts: -

- Level - The baseline value for the series if it were a straight line.
- Trend - The optional and linear, increasing or decreasing behaviour of series over time.
- Seasonality - Optional repeated patterns /cycles of behaviour over time.
- Noise - The optional variability in the observations that cannot be explained by the model.

Q8. What is Time Series Analysis?

Ans: Time series analysis: It involves developing models that best capture or describe an observed time series to understand the underlying cause.

This study seeks the “why” behind the time-series datasets. This involves making assumptions about the form of data and decomposing time-series into the constituent component.

Quality of descriptive model is determined by how well it describes all available data and the interpretation it provides to inform the problem domain better.

Q9. Give some examples of the Time-Series forecast?

Ans: There is almost an endless supply of the time series forecasting problems.

Below are ten examples from a range of industries to make the notions of time series analysis and forecasting more concrete.

1. Forecasting the corn yield in tons by the state each year.
2. Forecasting whether an EEG trace in seconds indicates a patient is having a seizure or not.
3. Forecasting the closing price of stocks every day.
4. Forecasting the birth rates at all hospitals in the city every year.
5. Forecasting product sales in the units sold each day for the store.
6. Forecasting the number of passengers through the train station each day.
7. Forecasting unemployment for a state each quarter.
8. Forecasting the utilisation demand on the server every hour.
9. Forecasting the size of the rabbit populations in the state each breeding season.
10. Forecasting the average price of gasoline in a city each day.

Q10. What are the techniques of Forecasting?

Ans: There are so many statistical techniques available for time series forecast however we have found a few effective ones which are listed below:

- **Simple Moving Average (SMA)**
- **Exponential Smoothing (SES)**
- **Autoregressive Integration Moving Average (ARIMA)**

Q11. What is the Moving Average?

Ans: The moving average model is probably the naivest approach to time series modelling.

This model states that the next observation is the mean of all past observations.

Although simple, this model might be surprisingly good, and it represents a good starting point.

Otherwise, the moving average can be used to identify interesting trends in the data. We can define a window to apply the moving average model to smooth the time series and highlight different trends.



Example of a moving average on a 24h window

In the plot above, we applied the moving average model to a 24h window. The green line smoothed the time series, and we can see that there are two peaks in the 24h period.

The longer the window, the smoother the trend will be.

Below is an example of moving average on a smaller window.



Example of a moving average on a 12h windows

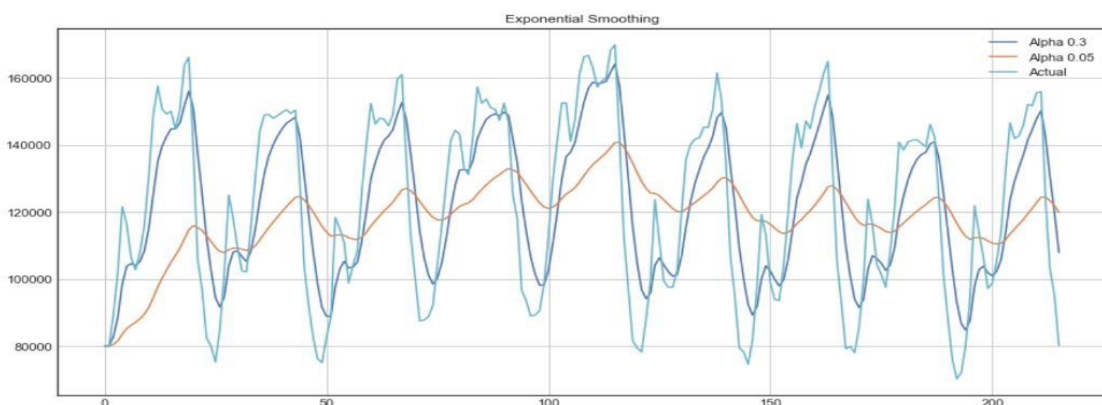
Q12. What is Exponential smoothing? Ans: Exponential smoothing uses similar logic to moving average, but this time, different decreasing weight is assigned to each observation.

We can also say less importance is given to the observations as we move further from the present.

Mathematically, exponential smoothing is expressed as:

$$y = \alpha x_t + (1 - \alpha)y_{t-1}, t > 0$$

Here, alpha is the smoothing factor which takes values between 0 to 1. It determines how fast the weight will decrease for the previous observations.



From the above plot, the dark blue line represents the exponential smoothing of the time series using a smoothing factor of 0.3, and the orange line uses a smoothing factor of 0.05. As we can see, the smaller the smoothing factor, the smoother the time series will be. Because as smoothing factor approaches 0, we approach to the moving average model.