TIME SERIES ANALYSIS NOTES

Q1. What is Time series analysis? Write its applications.

ANS:

It is a random sequence recorded in a time ordered fashion. Time series analysis is the endeavor of extracting meaningful summary & statistical information from points arranged in temporal order. It is to diagnose past behavior to make predictions about future behavior. The most common concerns of time series analysis are forecasting the future and classifying the past.

1. **Finance and Stock Market Analysis:** Time series analysis is extensively used in finance for forecasting stock prices, currency exchange rates, and commodity prices. It helps in making investment decisions and managing risk.
2. **Economic Forecasting:** Economists use time series analysis to forecast economic indicators such as GDP, inflation rates, and unemployment rates. This data is crucial for policy-making and business planning.
3. **Demand Forecasting:** Businesses use time series analysis to predict future demand for their products or services. This helps in inventory management, production planning, and resource allocation.
4. **Weather Forecasting:** Meteorologists use time series data for weather forecasting. Analyzing historical weather data allows them to make predictions about future weather conditions.
5. **Energy Consumption and Load Forecasting:** Utility companies use time series analysis to predict electricity and energy consumption patterns. This helps them in optimizing energy production and distribution.
6. **Healthcare and Epidemiology:** Time series analysis is used to track and predict the spread of diseases, monitor patient vital signs, and forecast healthcare resource requirements.
7. **Environmental Monitoring:** Time series data is crucial for tracking environmental changes, such as temperature trends, air quality, and water pollution levels.
8. **Quality Control:** Manufacturers use time series analysis to monitor the quality of their products over time. It helps in identifying defects and improving production processes.
9. **Marketing and Sales:** Businesses use time series analysis to analyze sales data, track marketing campaign effectiveness, and forecast sales trends.
10. **Traffic and Transportation:** Time series analysis is applied to traffic data to predict congestion patterns, optimize transportation routes, and plan infrastructure improvements.

11. **Social Sciences:** Time series analysis is used in social sciences to analyze trends in social, economic, and political data. It helps researchers understand and predict human behavior.

Q2: Define trend, seasonality, cyclicity, Irregularity.(factor affecting time series analysis)

Ans:

Trend is the increase or decrease in the series over a period of time, it persists over a long period of time.

Eg: population growth over the years can be seen as upward trend.

Seasonality is the regular pattern of up and down fluctuations. It is a short term variation occurring due to seasonal factors.

Eg: sales of icecream increases during summer season

Cyclicity is the mid term variation caused by circumstances, which repeat in irregular intervals

Eg: 5 years of economic growth , followed by 2 years of economic recession, followed by 7 years of economic growth followed by 1 year of economic recession.

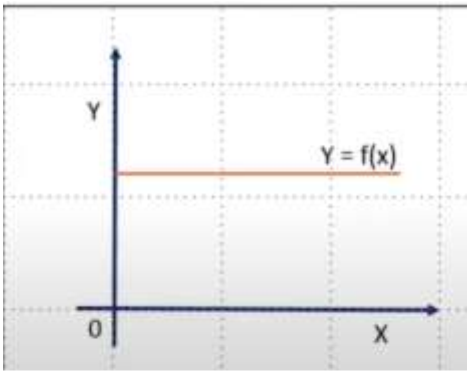
Irregularity refers to variations which due to predictable factors and also do not repeat in particular patterns.

Eg: variations caused by incidents like earthquakes, flood, war,etc.

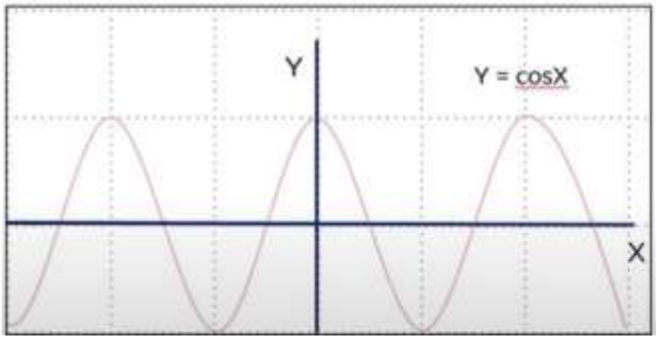
Q3: write conditions when to not use TSA?

Ans:

1) When values are constant over a period of time.



2) When values can be represented by known functions like cosx, sinx, etc.



Q4: stationary and non stationary time series.

Ans:

In time series analysis, one of the fundamental concepts is the distinction between stationary and non-stationary time series data. These terms refer to the statistical properties and characteristics of the data over time. Understanding whether a time series is stationary or non-stationary is crucial because it has implications for the choice of analytical methods and the reliability of forecasts.

Stationary Time Series:

A time series is considered stationary when its statistical properties do not change over time. In a stationary time series:

1. **Constant Mean (μ):** The mean (average) of the series remains constant over time.
2. **Constant Variance (σ^2):** The variance (spread or volatility) of the series remains constant over time.
3. **Constant Autocovariance:** The covariance between two observations at different time points (lag) remains constant.
4. **No Seasonal or Trend Components:** Stationary time series do not exhibit any systematic patterns or trends over time. There are no long-term upward or downward movements, and there are no repeating seasonal patterns.
5. **Statistical Tests Confirm Stationarity:** Formal statistical tests, such as the Augmented Dickey-Fuller (ADF) test, can be used to confirm the stationarity of a time series.

Stationary time series are easier to analyze and model because they exhibit stable statistical properties. Many traditional time series forecasting methods assume stationarity.

Non-Stationary Time Series:

Conversely, a time series is considered non-stationary when one or more of the properties mentioned above change over time. Non-stationary time series often exhibit trends, seasonality, and other time-dependent structures. Common characteristics of non-stationary time series include:

1. **Changing Mean:** The mean of the series exhibits a trend or systematic change over time. It may be increasing, decreasing, or following some other pattern.
2. **Changing Variance:** The variance of the series may change over time, leading to varying levels of volatility.
3. **Seasonal Patterns:** Non-stationary time series can have repeating seasonal patterns or cycles.
4. **Trends:** Non-stationary time series often exhibit long-term trends, which can be upward (growth) or downward (decay).
5. **Unit Roots:** Non-stationary series may exhibit unit roots, which are indicative of non-stationarity. The Augmented Dickey-Fuller (ADF) test is commonly used to test for unit roots.

Non-stationary time series can be challenging to model and forecast directly. Often, it is necessary to transform or differentiate the data to make it stationary before applying traditional time series forecasting techniques.

To summarize, the key difference between stationary and non-stationary time series lies in the stability of their statistical properties over time. Stationary time series exhibit stable means, variances, and covariances, while non-stationary time series display changing patterns and trends. Identifying whether a time series is stationary or non-stationary is a critical step in time series analysis to choose the appropriate modelling and forecasting methods.

Q5: how to convert data into stationary?

Ans:

•**Differencing:** Differencing is a method of removing trends from time series data. This is done by subtracting the previous value from the current value.

•**Logarithm transformation:** The logarithm transformation is a method of transforming the data to make it more stationary. This is done by taking the logarithm of the data.

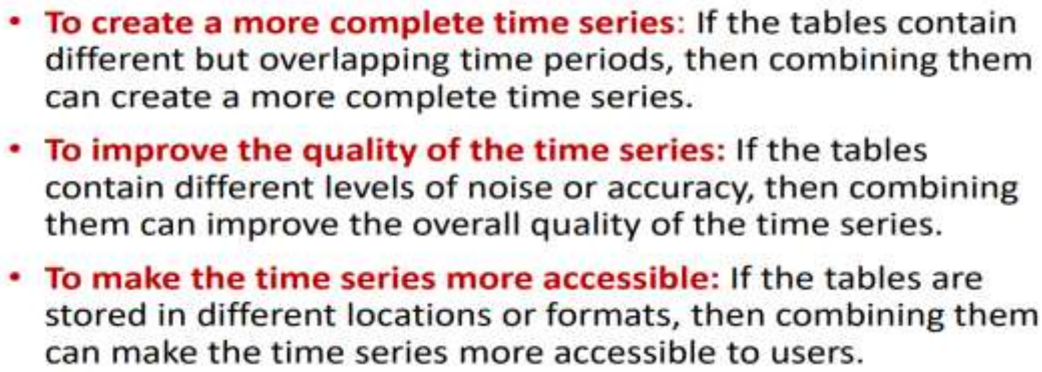
•**Seasonal adjustment**: Seasonal adjustment is a method of removing seasonal patterns from time series data. This is done by estimating the seasonal components of the data and then subtracting them from the original data.

•**Detrending:** Detrending is a method of removing trends from time series data. This is done by fitting a trend line to the data and then subtracting the trend line from the original data.

•**Combination of methods:** In some cases, it may be necessary to use a combination of methods to make time series data stationary.

Q6: Why retrofitting of time series data is performed from the collection of tables?

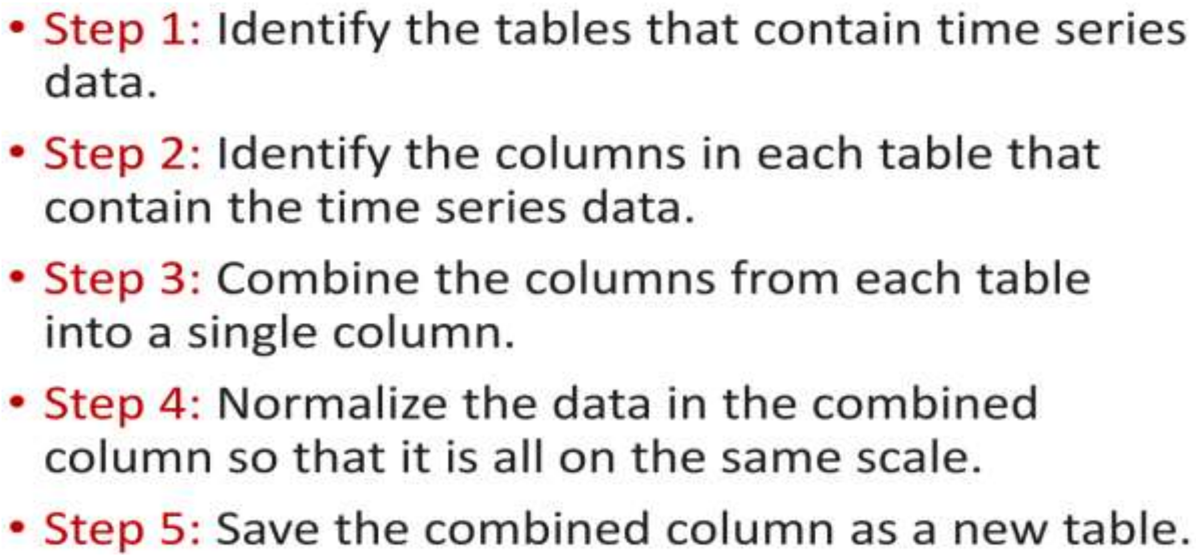
Ans:



- **To create a more complete time series:** If the tables contain different but overlapping time periods, then combining them can create a more complete time series.
- **To improve the quality of the time series:** If the tables contain different levels of noise or accuracy, then combining them can improve the overall quality of the time series.
- **To make the time series more accessible:** If the tables are stored in different locations or formats, then combining them can make the time series more accessible to users.

Q7: How to retrofit a time series data from a collection of tables?

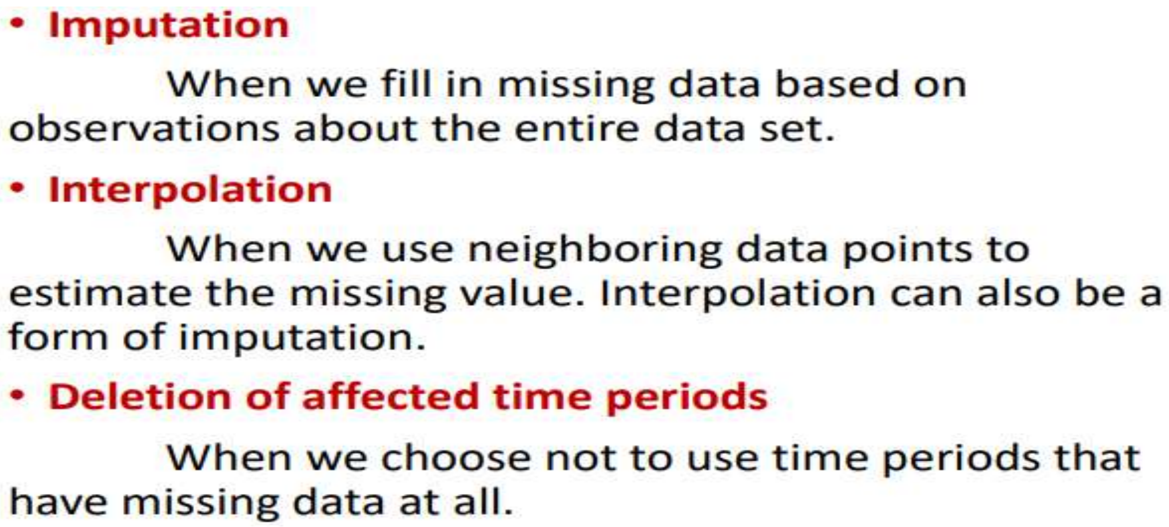
Ans:



- Step 1: Identify the tables that contain time series data.
- Step 2: Identify the columns in each table that contain the time series data.
- Step 3: Combine the columns from each table into a single column.
- Step 4: Normalize the data in the combined column so that it is all on the same scale.
- Step 5: Save the combined column as a new table.

Q8: How to handle missing data in time series?

Ans:



- **Imputation**

When we fill in missing data based on observations about the entire data set.

- **Interpolation**

When we use neighboring data points to estimate the missing value. Interpolation can also be a form of imputation.

- **Deletion of affected time periods**

When we choose not to use time periods that have missing data at all.

Q8: What is forward fill , backward fill, moving average and Interpolation?

Ans:

Forward Fill (or "ffill"):

Forward fill involves filling missing data points with the most recent preceding valid observation. In other words, it propagates the last known value forward in the dataset until a new valid value is encountered. This method is often used when missing data is assumed to carry forward the last observed value. It's particularly useful in cases where the data represents a sequence or time series, where continuity is important.

Date	Value
2023-01-01	10
2023-01-02	NaN
2023-01-03	15
2023-01-04	NaN
2023-01-05	NaN

After ffill

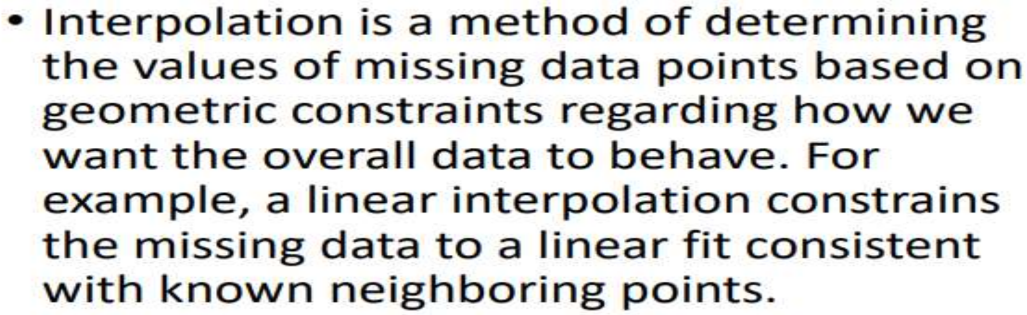
Date	Value
2023-01-01	10
2023-01-02	10 # Filled with the preceding value (10)
2023-01-03	15
2023-01-04	15 # Filled with the preceding value (15)
2023-01-05	15 # Filled with the preceding value (15)

Backward fill is the opposite of forward fill. It involves filling missing data points with the first valid observation encountered after the missing data point. In other words, it propagates the next known value backward in the dataset until a new valid value is encountered.

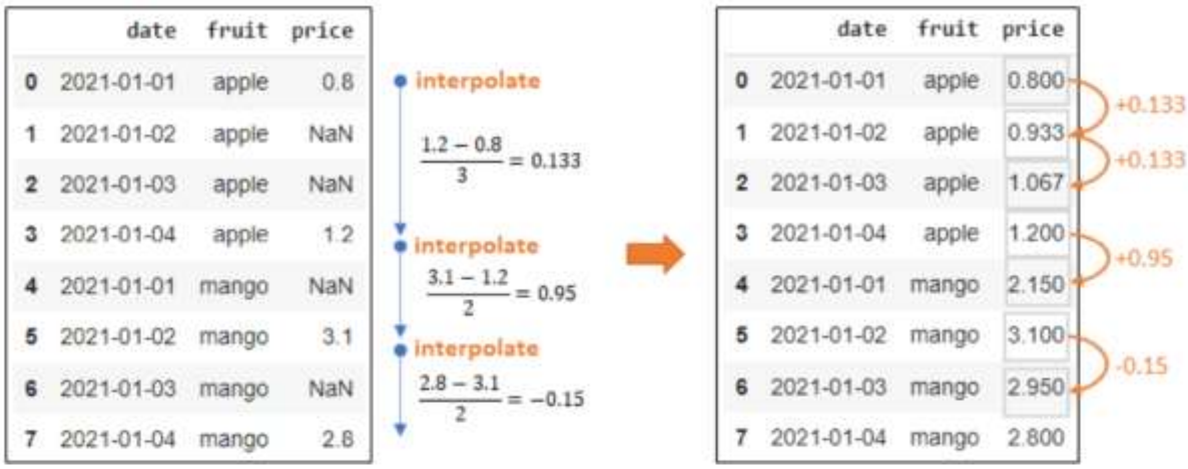
Using the same example data as above, but applying backward fill:

Date	Value
2023-01-01	10
2023-01-02	15 # Filled with the next valid value (15)
2023-01-03	15
2023-01-04	NaN
2023-01-05	NaN

A **moving average** is a commonly used statistical calculation applied to time series data to analyze and smooth out fluctuations in the data over time. It is particularly useful for identifying trends, patterns, and underlying patterns in noisy or volatile datasets. The moving average is calculated by taking the average of a set of data points within a sliding or "moving" window as it progresses through the dataset.



- Interpolation is a method of determining the values of missing data points based on geometric constraints regarding how we want the overall data to behave. For example, a linear interpolation constrains the missing data to a linear fit consistent with known neighboring points.

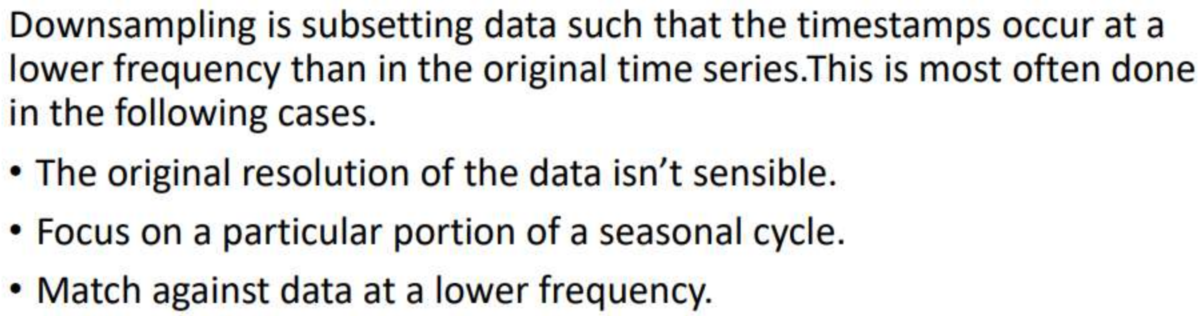


	date	fruit	price
0	2021-01-01	apple	0.8
1	2021-01-02	apple	NaN
2	2021-01-03	apple	NaN
3	2021-01-04	apple	1.2
4	2021-01-01	mango	NaN
5	2021-01-02	mango	3.1
6	2021-01-03	mango	NaN
7	2021-01-04	mango	2.8

	date	fruit	price
0	2021-01-01	apple	0.800
1	2021-01-02	apple	0.933
2	2021-01-03	apple	1.067
3	2021-01-04	apple	1.200
4	2021-01-01	mango	2.150
5	2021-01-02	mango	3.100
6	2021-01-03	mango	2.950
7	2021-01-04	mango	2.800

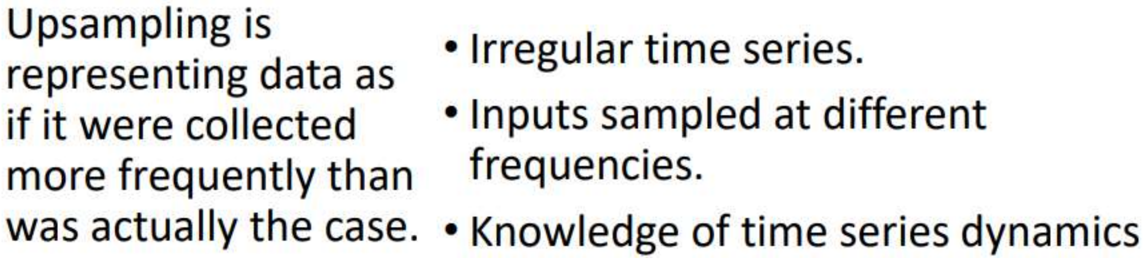
Q9: What is downsampling and upsampling? When to perform it?

Ans:



Downsampling is subsetting data such that the timestamps occur at a lower frequency than in the original time series.This is most often done in the following cases.

- The original resolution of the data isn’t sensible.
- Focus on a particular portion of a seasonal cycle.
- Match against data at a lower frequency.



Upsampling is representing data as if it were collected more frequently than was actually the case.

- Irregular time series.
- Inputs sampled at different frequencies.
- Knowledge of time series dynamics

Q10: what is smoothening of data? Why it is required? What are exponential smoothening methods and when to perfrom exponential smoothening?

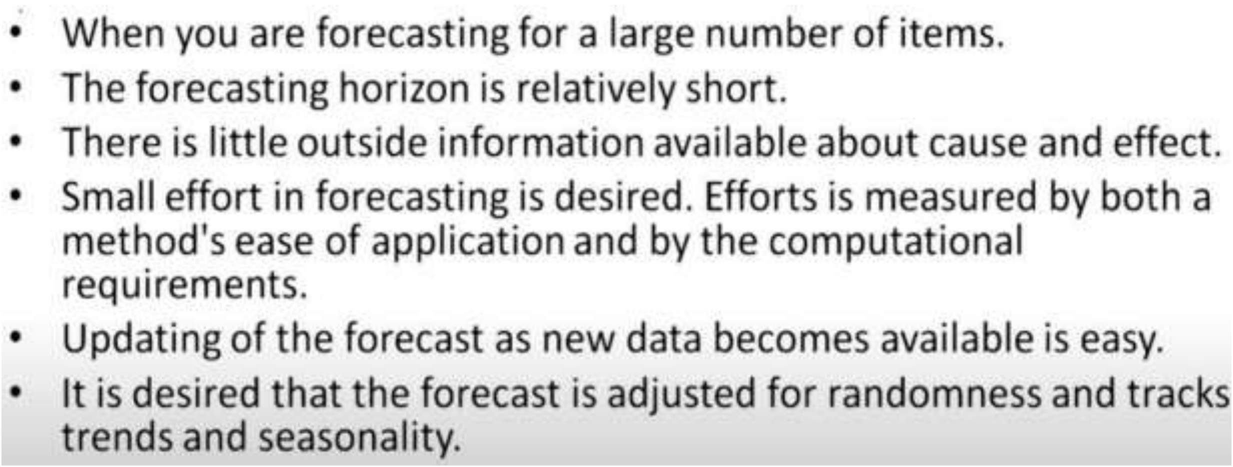
Ans:

Smoothing of data refers to the process of reducing noise or random variations in a dataset to extract underlying trends, patterns, or signals. It involves applying mathematical or statistical techniques to create a smoother representation of the data by averaging or aggregating values over a certain period or window. Smoothing is required for several reasons:

1. **Noise Reduction:** Many real-world datasets contain random fluctuations or noise that can make it difficult to identify meaningful patterns or trends. Smoothing helps to reduce the impact of this noise.
2. **Pattern Detection:** Smoothing can reveal hidden patterns or trends in the data that may not be immediately apparent in the raw data.
3. **Data Visualization:** Smoother data is often easier to visualize and interpret, making it valuable for data exploration and communication.
4. **Forecasting:** Smoothing is often used as a preprocessing step for time series forecasting, as it can help create a more stable and predictable dataset.

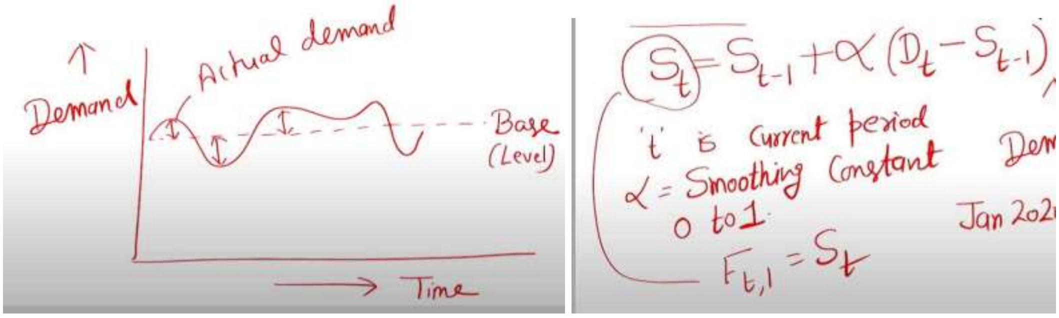
One common method of data smoothing is exponential smoothing. Exponential smoothing is a time series forecasting technique that assigns exponentially decreasing weights to past observations, giving more weight to recent observations.

When to use exponential smoothening:



- When you are forecasting for a large number of items.
- The forecasting horizon is relatively short.
- There is little outside information available about cause and effect.
- Small effort in forecasting is desired. Efforts is measured by both a method's ease of application and by the computational requirements.
- Updating of the forecast as new data becomes available is easy.
- It is desired that the forecast is adjusted for randomness and tracks trends and seasonality.

New Base = Previous Bases + α(New Demand- Previous Base)



$S_t = S_{t-1} + \alpha(D_t - S_{t-1})$

't' is Current period

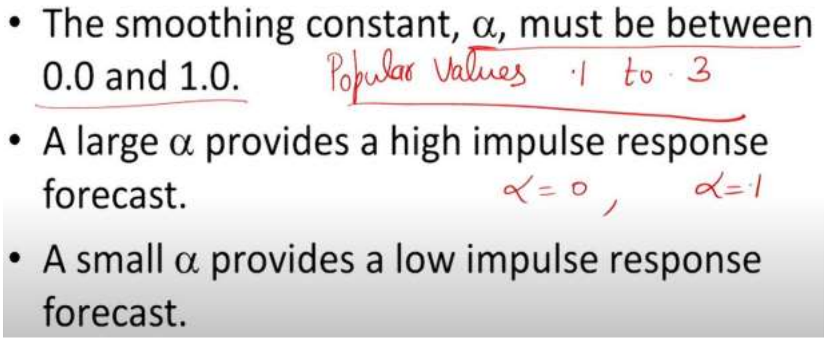
α = Smoothing Constant

0 to 1

$F_{t,1} = S_t$

Dem

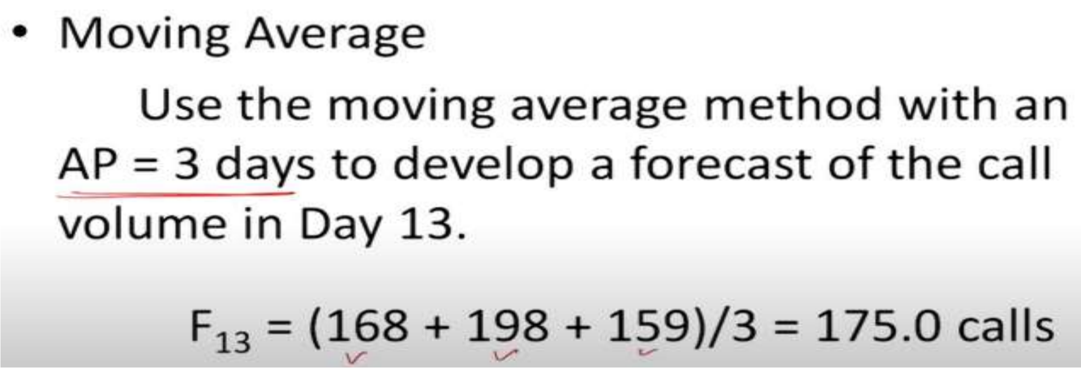
Jan 202



- The smoothing constant, α, must be between 0.0 and 1.0. Popular Values ·1 to ·3
- A large α provides a high impulse response forecast. $\alpha = 0$, $\alpha = \cdot 1$
- A small α provides a low impulse response forecast.

Q10: Numerical moving average and exponential smoothening

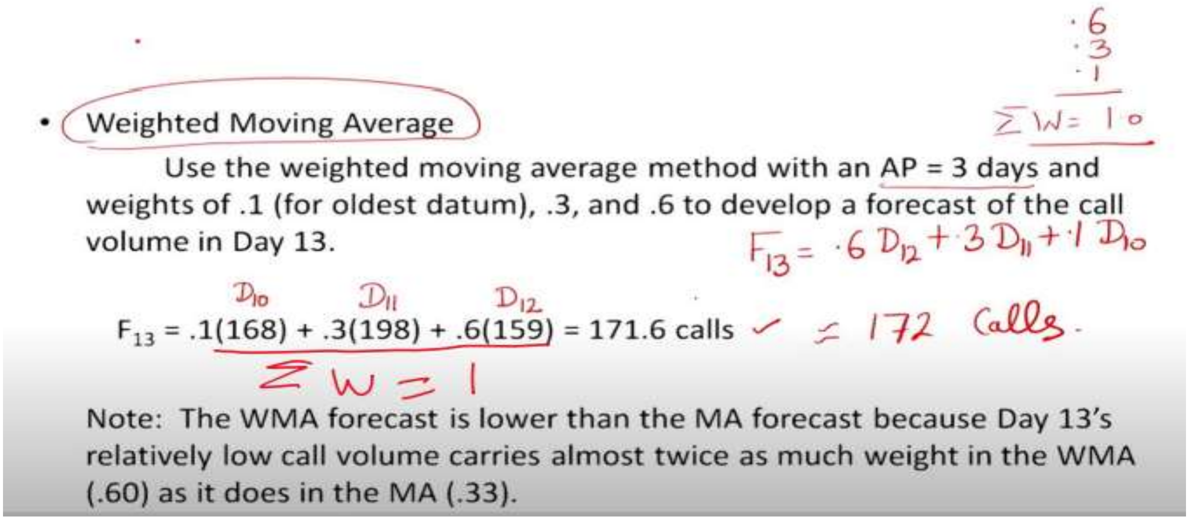
Ans:



- Moving Average

Use the moving average method with an AP = 3 days to develop a forecast of the call volume in Day 13.

$F_{13} = (168 + 198 + 159)/3 = 175.0$ calls



$\cdot 6$
$\cdot 3$
$\cdot 1$
$\Sigma W = 1 \cdot 0$

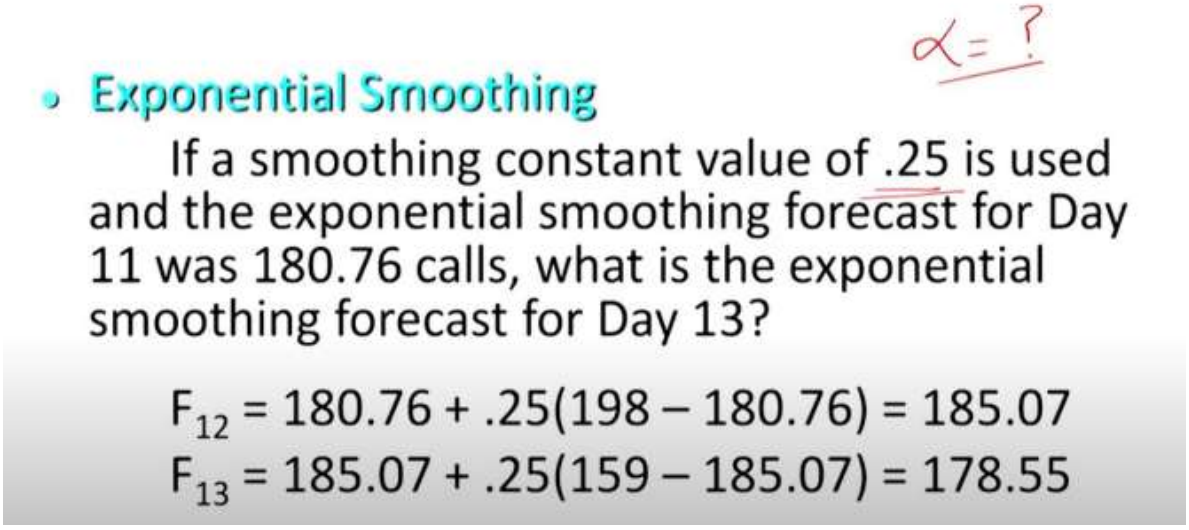
- Weighted Moving Average

Use the weighted moving average method with an AP = 3 days and weights of .1 (for oldest datum), .3, and .6 to develop a forecast of the call volume in Day 13.

$F_{13} = \cdot 6\, D_{12} + \cdot 3\, D_{11} + \cdot 1\, D_{10}$

$F_{13} = \underbrace{\overset{D_{10}}{.1(168)} + \overset{D_{11}}{.3(198)} + \overset{D_{12}}{.6(159)}}_{\Sigma W = 1} = 171.6$ calls ✓ $\simeq$ 172 Calls.

Note: The WMA forecast is lower than the MA forecast because Day 13's relatively low call volume carries almost twice as much weight in the WMA (.60) as it does in the MA (.33).



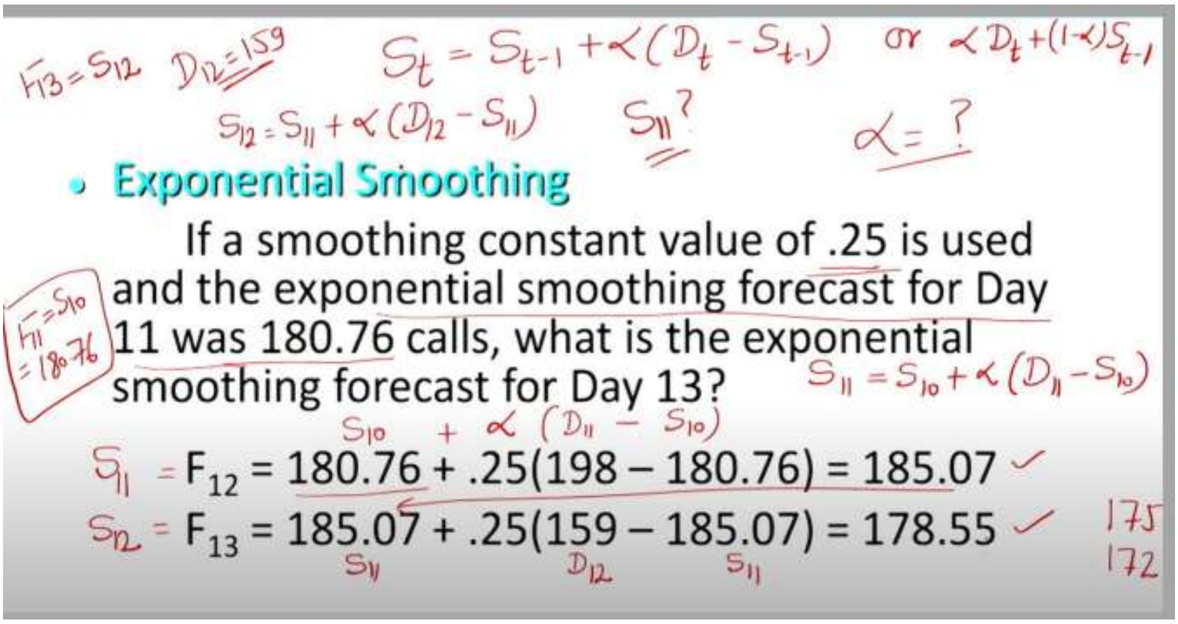
$\alpha = ?$

- Exponential Smoothing

If a smoothing constant value of .25 is used and the exponential smoothing forecast for Day 11 was 180.76 calls, what is the exponential smoothing forecast for Day 13?

$F_{12} = 180.76 + .25(198 - 180.76) = 185.07$

$F_{13} = 185.07 + .25(159 - 185.07) = 178.55$



$F_{13} = S_{12}$ $D_{12} = 159$ $S_t = S_{t-1} + \alpha(D_t - S_{t-1})$ or $\alpha D_t + (1-\alpha)S_{t-1}$

$S_{12} = S_{11} + \alpha(D_{12} - S_{11})$ $S_{11}?$ $\alpha = ?$

- Exponential Smoothing

If a smoothing constant value of .25 is used and the exponential smoothing forecast for Day 11 was 180.76 calls, what is the exponential smoothing forecast for Day 13?

$F_{11} = S_{10} = 180.76$

$S_{11} = S_{10} + \alpha(D_{11} - S_{10})$

$S_{10} + \alpha(D_{11} - S_{10})$

$S_{11} = F_{12} = 180.76 + .25(198 - 180.76) = 185.07$ ✓

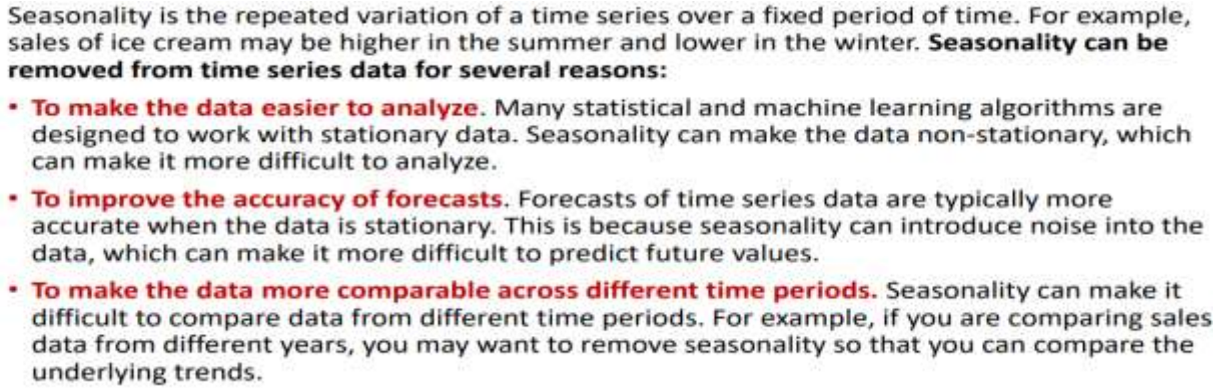
$S_{12} = F_{13} = \underset{S_{11}}{185.07} + .25(\underset{D_{12}}{159} - \underset{S_{11}}{185.07}) = 178.55$ ✓

175

172

Q11: what is seasonality and why it is removed?

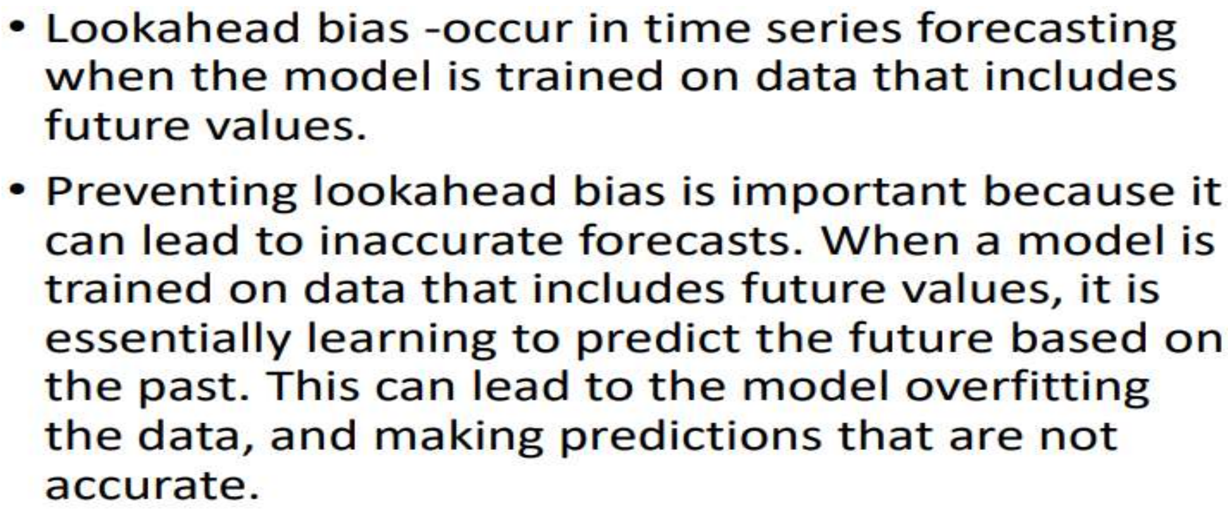
Ans:



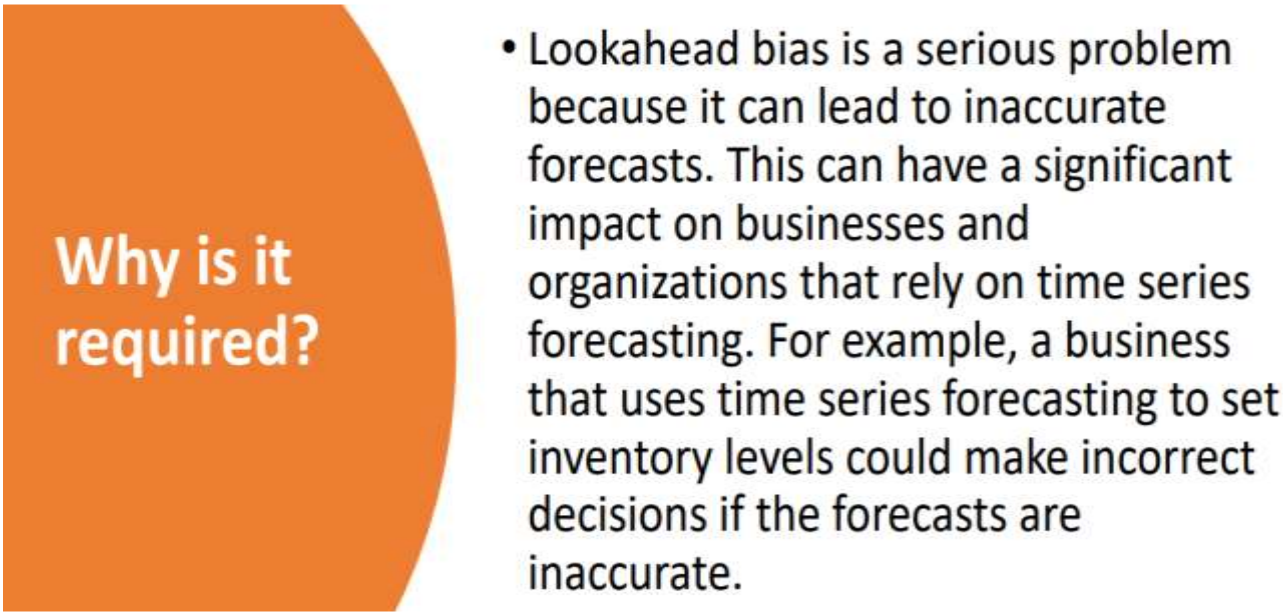
Seasonality is the repeated variation of a time series over a fixed period of time. For example, sales of ice cream may be higher in the summer and lower in the winter. **Seasonality can be removed from time series data for several reasons:**

- **To make the data easier to analyze**. Many statistical and machine learning algorithms are designed to work with stationary data. Seasonality can make the data non-stationary, which can make it more difficult to analyze.
- **To improve the accuracy of forecasts**. Forecasts of time series data are typically more accurate when the data is stationary. This is because seasonality can introduce noise into the data, which can make it more difficult to predict future values.
- **To make the data more comparable across different time periods.** Seasonality can make it difficult to compare data from different time periods. For example, if you are comparing sales data from different years, you may want to remove seasonality so that you can compare the underlying trends.

Q12: Lookahead bias

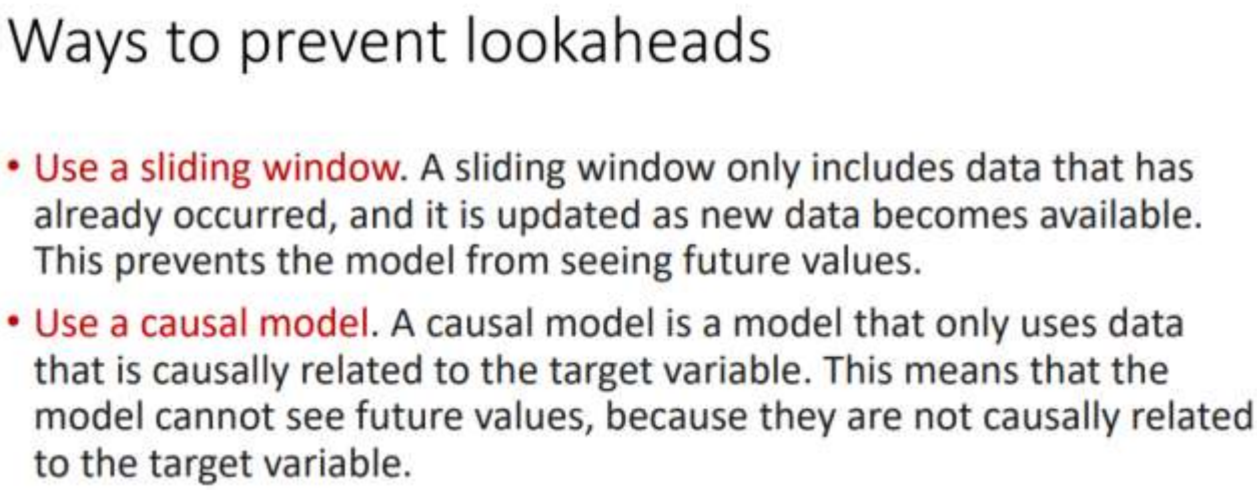


- Lookahead bias -occur in time series forecasting when the model is trained on data that includes future values.
- Preventing lookahead bias is important because it can lead to inaccurate forecasts. When a model is trained on data that includes future values, it is essentially learning to predict the future based on the past. This can lead to the model overfitting the data, and making predictions that are not accurate.



Why is it required?

- Lookahead bias is a serious problem because it can lead to inaccurate forecasts. This can have a significant impact on businesses and organizations that rely on time series forecasting. For example, a business that uses time series forecasting to set inventory levels could make incorrect decisions if the forecasts are inaccurate.



Ways to prevent lookaheads

- Use a sliding window. A sliding window only includes data that has already occurred, and it is updated as new data becomes available. This prevents the model from seeing future values.
- Use a causal model. A causal model is a model that only uses data that is causally related to the target variable. This means that the model cannot see future values, because they are not causally related to the target variable.