* Discuss the data sources used to test the hypothesis.
* Describe the process used to collect and clean the data, including any feature engineering techniques.
* Collect historical data of the stock prices over a period of time (e.g. daily, weekly, monthly).
* Split the data into train and test sets.
* Discuss the importance of collecting historical data from a reliable source, such as Yahoo Finance.

**II. Data Collection and Preparation**

* Check for missing values and outliers and handle them appropriately (e.g. impute, drop, replace).
* Discuss the need for handling missing values and outliers appropriately.

**III. Exploratory Data Analysis**

* Transform the data into a stationary series if needed (e.g. differencing, log transformation).
* Explain the need for transforming data into a stationary series if needed.
* Plot the data and observe the patterns and trends (e.g. seasonality, cyclicity, trend).
* Calculate descriptive statistics and summary measures (e.g. mean, median, standard deviation, skewness, kurtosis).
* Perform correlation analysis and test for autocorrelation and partial autocorrelation (e.g. ACF, PACF plots).
* Discuss the importance of collecting historical data from a reliable source, such as Yahoo Finance.
* Discuss the need for handling missing values and outliers appropriately.
* Explain the need for transforming data into a stationary series if needed.
* Discuss the importance of splitting the data into train and test sets.

o Plotting the data and observing the patterns and trends (e.g. seasonality, cyclicity, trend).

o Calculating descriptive statistics and summary measures (e.g. mean, median, standard deviation, skewness, kurtosis).

o Performing correlation analysis and testing for autocorrelation and partial autocorrelation (e.g. ACF, PACF plots).

o Plotting the time series and observing its shape, trend, seasonality and outliers (e.g. line plot).

o Testing for stationarity using statistical tests (e.g. Augmented Dickey-Fuller test) or visual methods (e.g. rolling mean and standard deviation).

o Decomposing the time series into trend, seasonal and residual components using additive or multiplicative models (e.g. seasonal\_decompose function in Python).

o Calculating and plotting autocorrelation and partial autocorrelation functions to measure the linear dependence of a time series with its own lagged values (e.g. acf and pacf functions in Python).

o Detecting change points or structural breaks in a time series using algorithms (e.g. PELT) or visual methods (e.g. CUSUM).

Seasonal subseries plots: This involves dividing the time series into seasonal periods and creating a subseries plot for each period to examine the patterns and trends within each season.

Box plots: This can be used to visualize the distribution of the time series and detect any outliers or extreme values.

Spectral analysis: This involves decomposing the time series into its frequency components using techniques such as Fourier analysis or wavelet analysis to identify any periodicity or cycles.

Plotting: In addition to plotting the raw stock price data, it may also be useful to plot the returns (percentage change) of the stock price over time, as this can help to identify patterns in the data that may not be immediately apparent when looking at the raw prices. Other useful plots might include histograms of the returns, or boxplots to compare the distribution of returns across different time periods.

Descriptive statistics: In addition to the basic summary statistics mentioned (mean, median, standard deviation, skewness, kurtosis), other useful measures might include the maximum and minimum values, the range of the data, and the coefficient of variation (CV), which is the ratio of the standard deviation to the mean.

Correlation analysis: In addition to calculating the Pearson correlation coefficient between the stock prices at different time points, it may also be useful to examine the cross-correlation function (CCF) between the stock prices and other variables that may be related, such as the prices of other stocks in the same industry, or macroeconomic indicators like interest rates or GDP growth.

Autocorrelation and partial autocorrelation: In addition to examining the ACF and PACF plots to identify potential AR and MA terms for a SARIMA model, it may also be useful to look at higher-order autocorrelations (e.g. ACF and PACF plots for lags > 12 for monthly data) to identify potential seasonality or longer-term dependencies in the data.

Determine stationarity: Stationarity is a key assumption for many time series models, and it means that the statistical properties of the time series (such as mean and variance) do not change over time. You can test for stationarity using statistical tests such as the Augmented Dickey-Fuller test.

1. Trend: Trend analysis is used to identify whether the time series data has a long-term upward or downward movement. Some common statistical tests used for trend analysis are:
   * Linear regression analysis: It is used to fit a linear trendline to the data and to test whether the slope of the trendline is significantly different from zero.
   * Moving average analysis: It is used to smooth out short-term fluctuations in the data and to identify the underlying trend.
   * Hodrick-Prescott filter: It is a mathematical method used to separate the trend and cyclical components of the time series data.
2. Seasonality: Seasonality analysis is used to identify whether the time series data has a repeating pattern at fixed intervals. Some common statistical tests used for seasonality analysis are:
   * Autocorrelation analysis: It is used to test whether the data exhibits a repeating pattern at a fixed interval.
   * Seasonal decomposition: It is used to separate the time series data into trend, seasonal, and residual components.
   * Spectral analysis: It is used to identify the dominant frequencies in the time series data and to test whether these frequencies correspond to seasonal patterns.
3. Cyclicity: Cyclicity analysis is used to identify whether the time series data has a periodic pattern that is not necessarily fixed. Some common statistical tests used for cyclicity analysis are:
   * Autocorrelation analysis: It is used to test whether the data exhibits a repeating pattern at any interval.
   * Spectral analysis: It is used to identify the dominant frequencies in the time series data and to test whether these frequencies correspond to cyclic patterns.
4. Stationarity: Stationarity analysis is used to identify whether the statistical properties of the time series data remain constant over time. Some common statistical tests used for stationarity analysis are:
   * Augmented Dickey-Fuller (ADF) test: It is used to test whether the data exhibits a unit root (i.e., a non-stationary trend).
   * Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test: It is used to test whether the data is stationary around a mean or a trend.
   * Phillips-Perron (PP) test: It is similar to the ADF test but allows for more flexibility in the specification of the model.
5. Serial correlation: Serial correlation analysis is used to identify whether the time series data exhibits a correlation between adjacent observations. Some common statistical tests used for serial correlation analysis are:
   * Autocorrelation function (ACF): It is used to test whether the data exhibits a correlation between adjacent observations at different lags.
   * Partial autocorrelation function (PACF): It is used to test whether the data exhibits a correlation between adjacent observations after accounting for the correlation at shorter lags.
6. White noise: White noise analysis is used to identify whether the time series data is random and uncorrelated. Some common statistical tests used for white noise analysis are:
   * Ljung-Box test: It is used to test whether the data exhibits no significant autocorrelation.
   * Breusch-Godfrey test: It is used to test whether the data exhibits no significant serial correlation.

**IV. Analysis**

SARIMA: In addition to specifying the appropriate order of the AR, MA, and seasonal components, it may also be necessary to include exogenous variables in the model, such as macroeconomic indicators or news events that could affect the stock price. To select the optimal model, one approach might be to use a grid search over a range of possible parameter values, and choose the model with the lowest AIC or BIC value.

Exponential smoothing: In addition to the basic exponential smoothing model, other variants that could be tested include Holt's linear exponential smoothing (which includes a trend component) and Holt-Winters' exponential smoothing (which includes both a trend and seasonal component).

LSTM: In addition to specifying the appropriate architecture and hyperparameters for the LSTM model, it may also be necessary to preprocess the data (e.g. scaling or normalization) and use techniques such as dropout or early stopping to prevent overfitting. To evaluate the model performance, metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE) could be used.

* Compare the effectiveness of different predictive models, such as ARIMA, LSTM, and Prophet.
* Choose a suitable model or models based on the characteristics of the data (e.g. ARIMA, SARIMA, ETS, LSTM).
* Specify the model parameters or hyperparameters and fit the model on the train set.
* Evaluate the model performance on the test set using appropriate metrics (e.g. MAE, MSE, RMSE).
* Compare different models or variations of models based on their performance metrics and choose the best one.
* Check for model assumptions and diagnostics (e.g. residuals analysis) and validate the model.

Select a model type: Based on the results of the previous steps, you can select a model type that is appropriate for your data. Some popular models for time series analysis include ARIMA, GARCH, and state-space models. You may need to try several different models and compare their performance using metrics such as mean squared error or Akaike Information Criterion (AIC) to determine the best model.

o Seasonal Autoregressive Integrated Moving Average (SARIMA): A model that captures both linear and seasonal dependencies in a time series1.

o Exponential Smoothing: A model that assigns weights to past observations that decay exponentially over time2.

o Long Short-Term Memory (LSTM): A type of recurrent neural network that can learn long-term dependencies in a time series2.

Split your data into training and testing sets .

• Train each model on the training set using appropriate functions such as fitlm, estimate, fitnet or trainNetwork. For example, you can train a linear regression model with one lagged variable as follows:

• Calculate the MSE of each model on the testing set using predict function and compare them. For example, you can calculate the MSE of the linear regression model as follows:

• Calculate the AIC or BIC of each model using aicbic function and compare them. For example, you can calculate the AIC of the linear regression model

• Choose the model that has the lowest MSE and AIC or BIC values.

o Fitting different models or variations of models to the train set using appropriate parameters (e.g. order for SARIMA) and hyperparameters (e.g. learning rate for LSTM).

o Comparing different models or variations of models based on their performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), etc.

o Checking for model assumptions and diagnostics such as normality, homoscedasticity and independence of residuals using statistical tests (e.g. Jarque-Bera test) or visual methods (e.g. QQ plot).

o Validating the model using cross-validation techniques such as walk-forward validation or rolling window validation.

Forecast evaluation: This involves comparing the actual forecasted values to the observed values and examining the accuracy and reliability of the forecasts, such as by using metrics such as mean absolute percentage error (MAPE) or symmetric mean absolute percentage error (SMAPE).

1. Forecast Evaluation:

• Use the trained model to generate forecasts on the test set and compare the actual forecasted values to the observed values.

• Evaluate the accuracy and reliability of the forecasts using metrics such as mean absolute percentage error (MAPE) or symmetric mean absolute percentage error (SMAPE).

• Adjust the model and repeat the evaluation process until satisfactory performance is achieved.

* Demonstrate a deep understanding of these models and explain why they were chosen.
* Discuss the evaluation metrics used to assess model performance.

Model testing

Discuss the different models that can be used, such as ARIMA, SARIMA, ETS, and LSTM.

Explain the process of specifying the model parameters or hyperparameters and fitting the model on the train set.

Discuss the importance of evaluating the model performance on the test set using appropriate metrics.

Model selection

Explain the process of comparing different models or variations of models based on their performance metrics and choosing the best one.

Discuss the need for checking for model assumptions and diagnostics, such as residuals analysis, to validate the model.

Model evaluation and prediction

Discuss the importance of applying the selected model to forecast future values of the stock prices using confidence intervals or prediction intervals.

Explain the need for updating or retraining the model periodically as new data becomes available.

1. To apply the Shapiro-Wilk test,. The function returns a p-value that indicates the level of significance for the null hypothesis that the data is normally distributed. Here's an example code:
2. To apply the Kolmogorov-Smirnov test,. The function returns a p-value that indicates the level of significance for the null hypothesis that the data is drawn from a standard normal distribution. Here's an example code:

Note that in both tests, a low p-value indicates that the null hypothesis can be rejected and the data is not normally distributed.

Also, keep in mind that these tests are sensitive to outliers. If your time series data contains many outliers, it might not be appropriate to use these tests for testing normality. In such cases, it might be better to use robust methods for testing normality, such as the Anderson-Darling test or the Lilliefors test.

What is the purpose and use of non-parametric tests that do not assume the normality of the data

Non-parametric tests are statistical tests that do not require specific assumptions about the underlying distribution of the data, such as normality. Instead, they rely on ranking or ordering the data to perform statistical inference. These tests are useful when the data violate the assumptions of parametric tests or when the sample size is small.

The use of non-parametric tests is common in fields such as environmental science, social science, and medical research where the data can be non-normal or skewed. They can be applied in various situations, including comparing two or more groups, testing for correlation or association, and assessing goodness-of-fit.

Non-parametric tests have several advantages, including their robustness to outliers, their ability to handle non-normal data, and their suitability for small sample sizes. They also do not require the estimation of parameters, making them more computationally efficient and requiring fewer assumptions.

Examples of non-parametric tests include the Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal-Wallis test, Friedman test, Spearman's rank correlation coefficient, Kendall's tau, and the runs test.

Stationarity and unit root are two concepts that are closely related to each other in time series analysis.

Stationarity refers to the statistical properties of a time series remaining constant over time. A stationary time series has a constant mean, variance, and autocovariance structure over time. In other words, the statistical properties of the time series do not depend on the time at which they are measured. Stationarity is an important concept in time series analysis because many statistical models assume that the time series is stationary.

A unit root, on the other hand, is a property of a time series that has a characteristic root of unity (i.e., a value of 1) in its autoregressive (AR) model. A time series with a unit root is non-stationary because its statistical properties change over time. A unit root indicates the presence of a trend or a random walk in the time series, which causes the mean and variance of the series to increase over time.

In summary, stationarity refers to the constancy of the statistical properties of a time series over time, while a unit root is a property of a non-stationary time series that indicates the presence of a trend or a random walk in the series. A stationary time series has no unit root, but a non-stationary time series may have a unit root.

Cointegration is another concept that is closely related to stationarity and unit root in time series analysis.

Cointegration refers to the long-run relationship between two or more non-stationary time series. Two or more non-stationary time series are said to be cointegrated if there exists a linear combination of them that is stationary. In other words, cointegration implies that even though the individual time series may not be stationary, there is a combination of them that is stationary, and this combination represents a long-term relationship between them.

The difference between cointegration and stationarity is that stationarity refers to the short-term statistical properties of a time series, while cointegration refers to the long-term statistical properties of two or more time series. Stationarity implies that the mean, variance, and covariance of a time series do not change over time, while cointegration implies that there is a long-term relationship between two or more non-stationary time series, even if their short-term statistical properties change.

The difference between cointegration and unit root is that unit root implies that a time series is non-stationary, while cointegration implies that two or more non-stationary time series are related in a long-term sense. Unit root indicates the presence of a trend or a random walk in the time series, which causes the mean and variance of the series to increase over time. Cointegration, on the other hand, implies that two or more non-stationary time series have a common trend or stochastic drift that can be removed by forming a linear combination of them.

In summary, cointegration is a long-term relationship between two or more non-stationary time series, while stationarity refers to the short-term statistical properties of a time series and unit root implies that a time series is non-stationary due to the presence of a trend or a random walk.