A close-up of a document

Description automatically generated

A graph showing the number of years

Description automatically generated

A black text on a white background

Description automatically generatedA graph of a problem

Description automatically generated

The rule of thumb says that the difference of standard deviation of series divided by standard

-



This ratio is significantly larger than 0.05, which suggests that the original data is likely non-stationary.

A graph with lines and numbers

Description automatically generated

A graph with numbers and lines

Description automatically generated

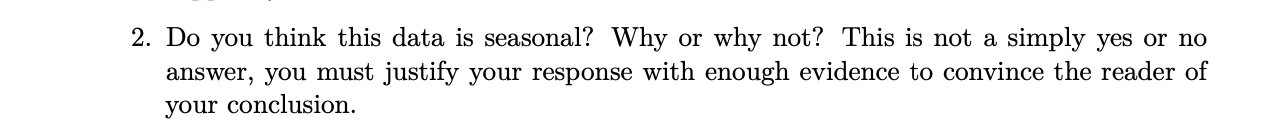
I determined the order of integration for the series problem\_1. Initially, I plotted the series to observe any trends, and the plot indicated that the series exhibited non-stationary behavior, as the values fluctuated with no clear mean reversion. To confirm this, I conducted an Augmented Dickey-Fuller (ADF) test, which returned a p-value of 0.6486, confirming that the series was non-stationary.

Next, I applied first differencing to the series and plotted the differenced values. The plot showed a more stable fluctuation around a mean, suggesting that the differenced series might be stationary. I then conducted the ADF test on the differenced series, which returned a p-value of 0.01. This allowed me to reject the null hypothesis of non-stationarity, confirming that the differenced series was stationary.

Based on this analysis, I concluded that the series problem\_1 is integrated of order one, or **I(1)**, meaning it needed to be differenced once to achieve stationarity.

A black text on a white background

Description automatically generated



A graph of a number of different types of lines

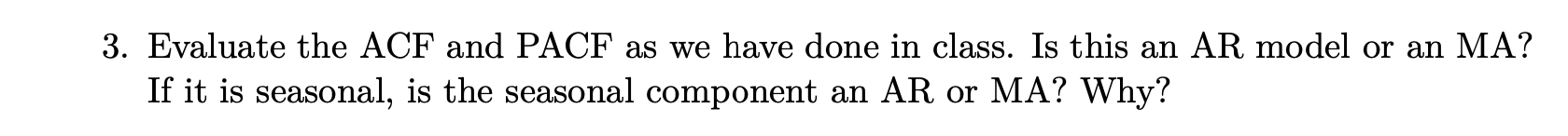
Description automatically generated with medium confidence

A graph of a line graph

Description automatically generated with medium confidence

For **Question 2**, I analyzed whether the data for problem\_1 is seasonal. Initially, I visually inspected the time series plot and did not observe any clear recurring patterns that would indicate seasonality. To further confirm this, I examined the **ACF plot**, which showed a gradual decay without significant spikes at regular intervals that would suggest a seasonal component. Typically, seasonal data would exhibit distinct peaks at specific lags, such as 12 months for yearly seasonality, but this was not evident in the plot.

Based on the ACF analysis, I concluded that the data is not seasonal. While the series does show trends, there is no consistent repeating pattern indicative of seasonality. Therefore, I determined that problem\_1 does not exhibit seasonal behavior.

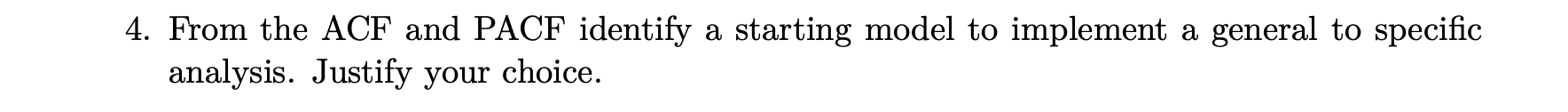


For **Question 3**, I evaluated both the **ACF** and **PACF** plots for problem\_1 to determine whether the series follows an AR or MA process, as well as whether there is any seasonal component.

First, I examined the ACF plot, which showed a gradual decay rather than a sharp cutoff, indicating that the data is likely non-stationary and may have an autoregressive component. Next, I reviewed the PACF plot, which showed a significant spike at lag 1 followed by an immediate drop to near zero. This sharp cutoff after lag 1 suggests that the series follows an **AR (1)** process.

If the data had a strong MA component, the ACF would have shown a sharp cutoff at a particular lag, but this was not the case. Additionally, I looked for any signs of seasonality in both the ACF and PACF, such as spikes at regular intervals (e.g., every 12 months), but no such pattern was evident. Therefore, I concluded that the series does not have a strong seasonal component.

Based on this analysis, I determined that the data most likely follows an **AR(1)** model with no significant seasonal component.



A graph of a number of different types of lines

Description automatically generated with medium confidence

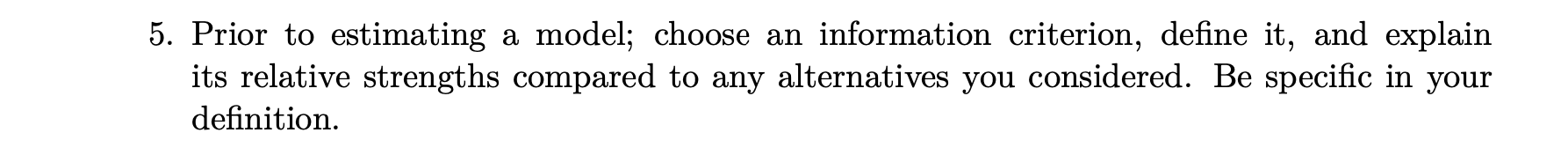
A graph of a line graph

Description automatically generated with medium confidence

Based on the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, we can determine the appropriate parameters for an ARIMA model. The PACF plot shows a significant spike at lag 1, indicating the presence of an / **Only Ar Model autoregressive** (AR) component, which leads us to set p=1. Since the time series was differenced once to achieve stationarity, we set the differencing parameter d=1. Additionally, the ACF plot reveals a significant spike at lag 1, suggesting the inclusion of a moving average (MA) component, with q=1. Thus, an ARIMA(1, 1, 1) model appears to be a suitable starting point for modeling this time series.

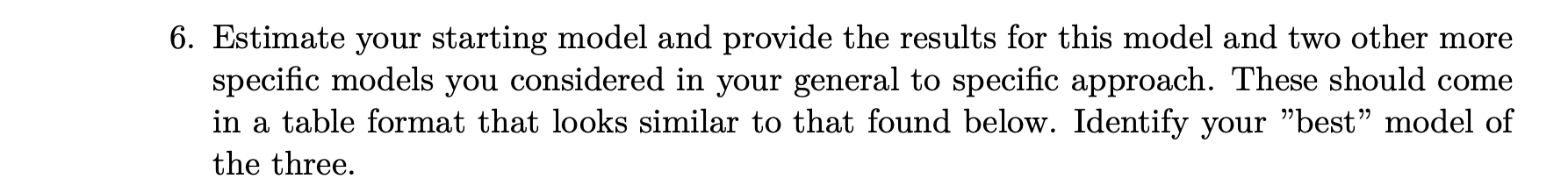
**-aR MODEL when pACF DECays faster than ACF ., p= PACF,q= ACF**

**-MA mODEL WHEn ACF decays faster than PACF**



For **Question 5**, I chose the **Akaike Information Criterion (AIC)** as the information criterion to guide model selection. The AIC is a widely used metric that evaluates the relative quality of statistical models by balancing model fit and complexity. It is calculated using the formula: AIC=2k−2ln⁡(L)AIC=2k−2ln(L), where **k** is the number of model parameters and **L** is the maximum likelihood of the model. AIC penalizes models with more parameters to prevent overfitting, which ensures that the model not only fits the data well but also remains parsimonious.

I chose AIC over other alternatives, such as the **Bayesian Information Criterion (BIC)**, because AIC is more flexible when the goal is prediction. While BIC imposes a stronger penalty on model complexity and is often used when the goal is to identify the "true" model, AIC is more suited to finding a model that generalizes well for future data. Although BIC tends to favor simpler models, AIC strikes a better balance between fit and complexity, making it appropriate for this analysis.



A white background with black text

Description automatically generated

A screenshot of a computer code

Description automatically generated

A white background with black text

Description automatically generated

A white background with black numbers

Description automatically generated

A white background with black numbers

Description automatically generated

A white background with black text

Description automatically generated

A table with numbers and a number of objects

Description automatically generated with medium confidence

Phi and standard error.

**I did not report RMSE, as not needed on this analysis.**

-

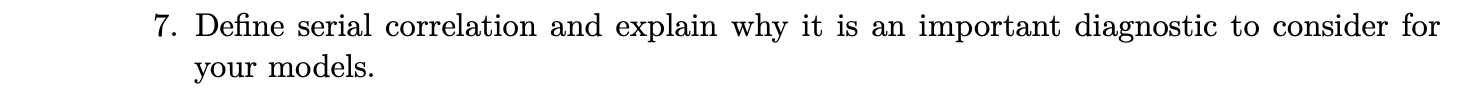
For **Question 6**, I estimated my starting model, **ARIMA(1,1,0)**, and compared it with two other models, **ARIMA(2,1,1)**and **ARIMA(0,1,1)**, to determine the best fit for the time series data.

I found that **Model 1 (ARIMA (1,1,0))** performed the best. This model includes a significant **AR(1)** term, indicating that the time series is influenced by its first lag, and it does not include any moving average term. With an **AIC** value of **765.1005** , this model had the lowest scores among the three, suggesting it provided the best balance between model fit and simplicity. Additionally, the **Ljung-Box test** yielded a p-value of **0.05419**, which is close to the 0.05 threshold, indicating that there is no significant autocorrelation in the residuals. This confirms that the residuals behave like white noise, meaning the model adequately captures the underlying dynamics of the time series.

Next, I examined **Model 2 (ARIMA (2,1,1))**, which includes both an autoregressive and a moving average term. While both terms were significant, the **AIC** is slightly higher than those for Model 1, at **768.46** . The **Ljung-Box test** for this model yielded a p-value of **0.1256**, confirming that the residuals are also uncorrelated. However, given the higher AIC, this model was slightly less optimal compared to the simpler **ARIMA(1,1,0)**.

Finally, I considered **Model 3 (ARIMA(0,1,1))**, which only includes a moving average term. While the model fit was reasonable, with an **AIC** of **768.4613** , was higher than those of Model 1. The **Ljung-Box test** for this model produced a p-value of **0.1357**, confirming that there was no significant autocorrelation in the residuals. Despite this, the higher information criteria values indicated that this model did not fit the data as well as Model 1.

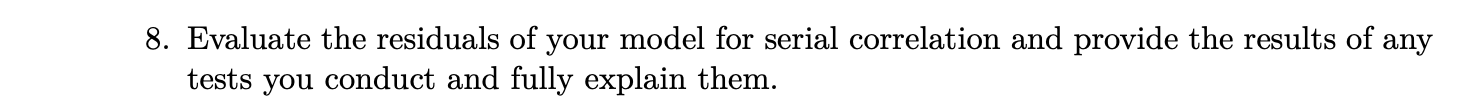
In conclusion, I determined that **ARIMA (1,1,0)** is the best model for this time series. It provides the lowest AIC and BIC values, and the residual diagnostics confirm that the model captures the data well without overfitting. While the other models were also reasonable, **Model 1** offers a better balance between simplicity and performance, making it the most appropriate choice for this analysis.



Serial correlation, also known as autocorrelation, occurs when the residuals (errors) from a time series model are correlated with one another over time. This means that the error at one time point is related to the errors at previous time points. In time series models like ARIMA, the assumption is that the residuals should behave like white noise, meaning they are random and uncorrelated. It is important to test for serial correlation because it indicates whether the model has fully captured the structure of the data. If serial correlation remains in the residuals, it suggests that the model has not adequately accounted for all of the patterns in the data.

Serial correlation is critical to check because it can lead to biased or inefficient parameter estimates. When residuals are correlated, the model’s estimates may be incorrect, and this can reduce the accuracy of forecasts. Additionally, hypothesis testing can be affected by serial correlation, as many statistical tests assume that the residuals are independent. If this assumption is violated, the results of these tests may be invalid, leading to incorrect conclusions about the model’s performance.

By ensuring that serial correlation is absent from the residuals, I can be confident that the model is well-specified and that it provides reliable forecasts. If serial correlation is detected, it suggests that the model needs to be refined, such as by adding additional AR or MA terms, to better capture the data’s underlying dynamics. This is essential for improving the model's accuracy and ensuring that the residuals behave like white noise, indicating that all patterns in the data have been captured effectively.



A graph showing the time of a graph

Description automatically generated with medium confidence

A graph of a line graph

Description automatically generated with medium confidence

A number and numbers on a white background

Description automatically generated

A graph of a number of data

Description automatically generated with medium confidence

Based on the residual plot, ACF plot, diagnostic plot and Ljung-Box test, the residuals of the **ARIMA(1,1,0)** model show no significant serial correlation. The residuals behave like white noise, meaning the model has adequately captured the time series dynamics. The Ljung-Box test confirmed the absence of significant autocorrelation with a p-value of **0.05419**, which supports the model’s fit. Overall, these diagnostic checks validate the model's ability to predict and capture the underlying structure of the data effectively.