Question-1

2024-04-11

```r
# Read the dataset
cet_temp <- read.csv("cet_temp.csv")

temperature_ts <- ts(cet_temp$avg_annual_temp_C, start=1900, frequency=1)
```

The dataset "cet_temp.csv" consists of two columns: year and avg_annual_temp_C. The year column ranges from 1900 to 2021, and the avg_annual_temp_C column represents the average annual temperature in degrees Celsius for each year in the Midlands region of England, as recorded by the UK Meteorological Office Hadley Climate Centre.

From my perspective, the dataset presents an opportunity to examine temporal patterns, trends, and seasonality in the temperature data over the 122-year period. Techniques such as time series decomposition can help in separating the data into trend, seasonal, and residual components.

Identifying any cyclic patterns or long-term trends is is very important for understanding climate change impacts on regional temperatures.

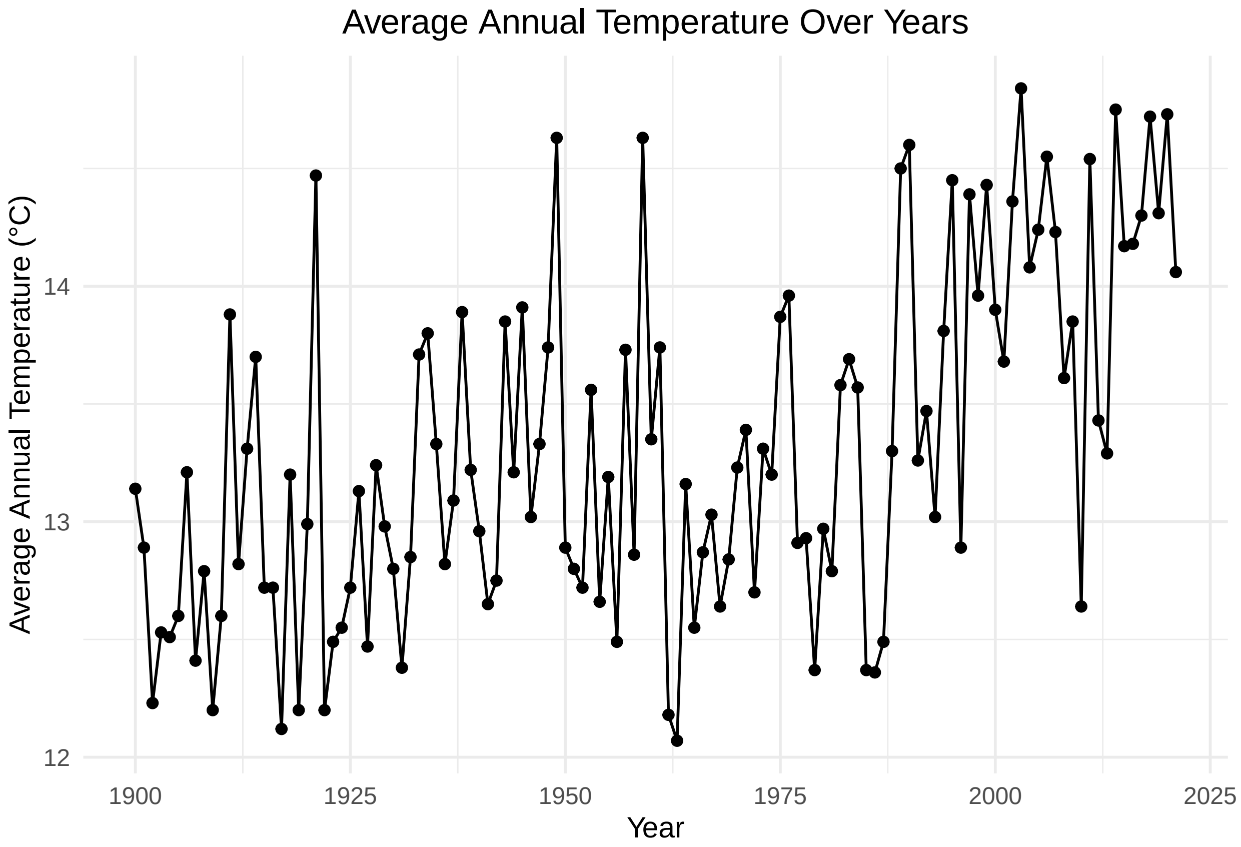
The dataset has no missing values, which means it's complete and ready for further analysis without the need for data imputation

Next, we will explore the time series aspect, focusing on trend analysis and potentially identifying any cyclical patterns or significant shifts in temperatures over time. This will involve visualizing the temperature data across the years and applying statistical tests or models to discern trends.

```r
# Load necessary library
library(ggplot2)



# Generate the time plot
ggplot(cet_temp, aes(x = year, y = avg_annual_temp_C)) +
  geom_line() + # Draw the line
  geom_point() + # Add points at each data entry
  theme_minimal() + # for a minimal/simple theme
  labs(title = "Average Annual Temperature Over Years",
       x = "Year",
       y = "Average Annual Temperature (°C)") +
  theme(plot.title = element_text(hjust = 0.5)) # Center the title
```

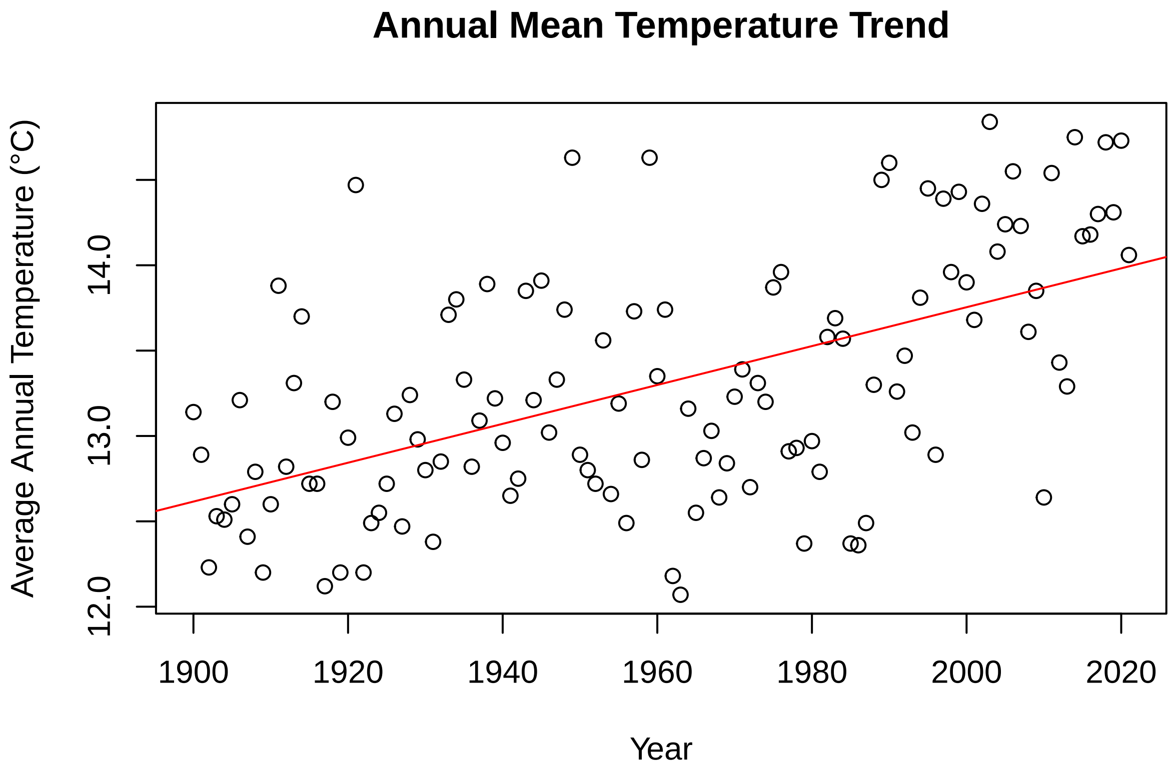


From the plot we can see, it appears that there are fluctuations in the average annual temperature over time. The mean of the series does not look constant as there seem to be periods where the temperature is generally higher than in other periods. Additionally, there's some visible seasonality: the fluctuations seem to follow a pattern over time. These observations suggest that the time series may not be stationary.

```r
# Linear model to analyze trend
model <- lm(avg_annual_temp_C ~ year, data=cet_temp)

# Plot the data
plot(cet_temp$year, cet_temp$avg_annual_temp_C, main="Annual Mean Temperature Trend",
     xlab="Year", ylab="Average Annual Temperature (°C)")

# Add the trend line
abline(model, col="red")
```

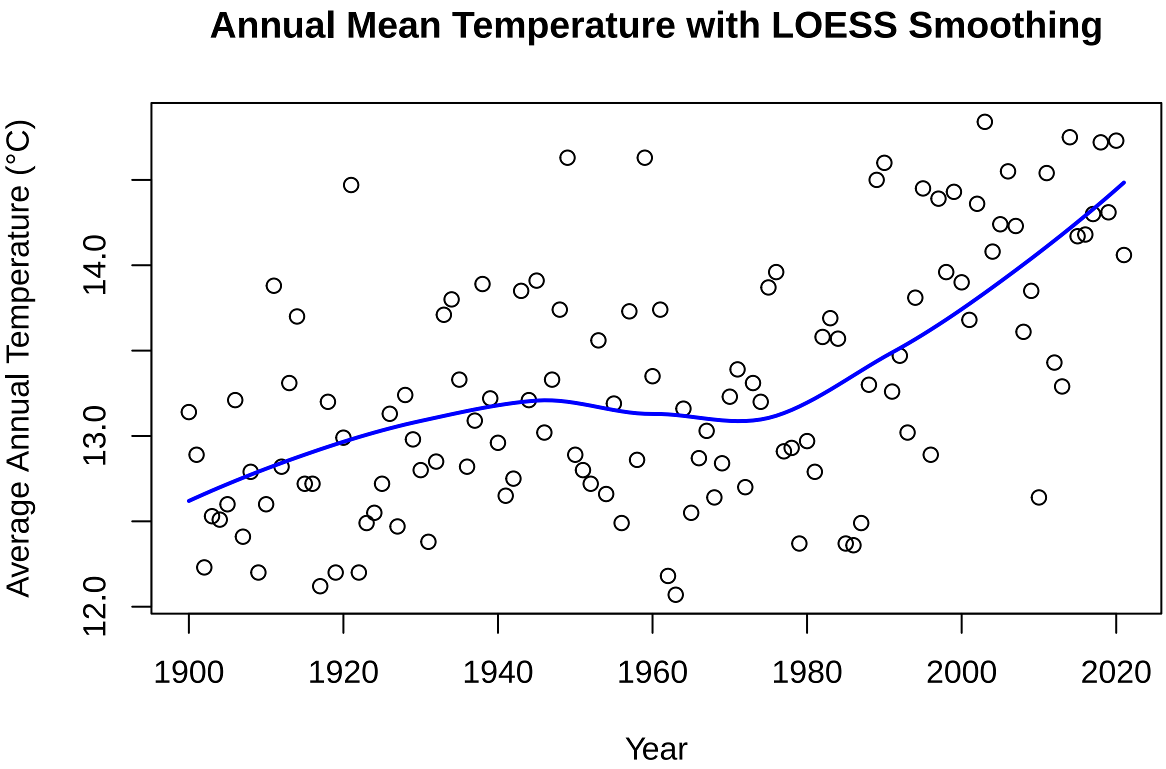


```r
# LOESS smoothing
loess_model <- loess(avg_annual_temp_C ~ year, data=cet_temp)

# Predict values
loess_pred <- predict(loess_model, data.frame(year=cet_temp$year))

# Plot the data
plot(cet_temp$year, cet_temp$avg_annual_temp_C, main="Annual Mean Temperature with LOESS Smoothing",
     xlab="Year", ylab="Average Annual Temperature (°C)")

# Add the LOESS smoothed line
lines(cet_temp$year, loess_pred, col="blue", lwd=2)
```



The seasonal decomposition of the dataset reveals the observed temperatures, the underlying trend, seasonal variations, and residual components:

Observed: The red graph shows the actual annual mean temperatures over the years. Trend: The green graph indicates a long-term increase in temperatures, suggesting a warming trend over the period of 122 years.

Seasonal: The blue graph shows the seasonal component, which, given the context of annual data with a frequency of 1, doesn't present a clear seasonal pattern as expected.

Residual: The yellow graph displays the residual or error component, which captures the fluctuations in the data that the trend and seasonal components do not.

To assess the stationarity of the time series, an Augmented Dickey-Fuller (ADF) test was performed {for technical and coding purposes, though we already have a idea about the stationarity of our time series}.

The ADF test results in a statistic of -2.15 and a p-value of 0.22. For the data to be considered stationary, the ADF statistic should be less than the critical values (for example, less than -3.48 for the 1% level) or the p-value should be below 0.05. Since neither condition is met, we can conclude that the time series is not stationary, indicating the presence of a trend.

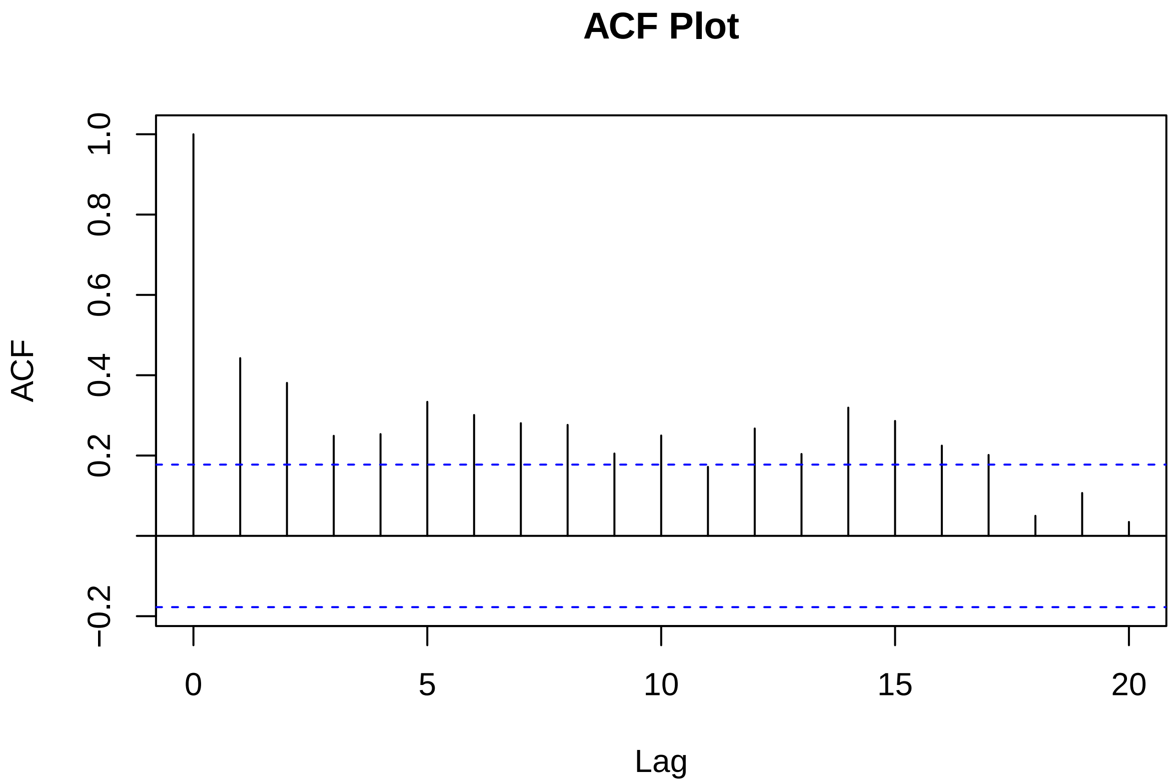
Given the non-stationarity of the series and the observed warming trend, fitting a time series model such as ARIMA (AutoRegressive Integrated Moving Average) could be suitable. The integration part of ARIMA can help to make the series stationary by differencing, while the AR and MA parts model the correlations in the data.

Before fitting an ARIMA model, we have to identify the order of differencing (d) needed to make the series stationary and to estimate the AR (p) and MA (q) terms.

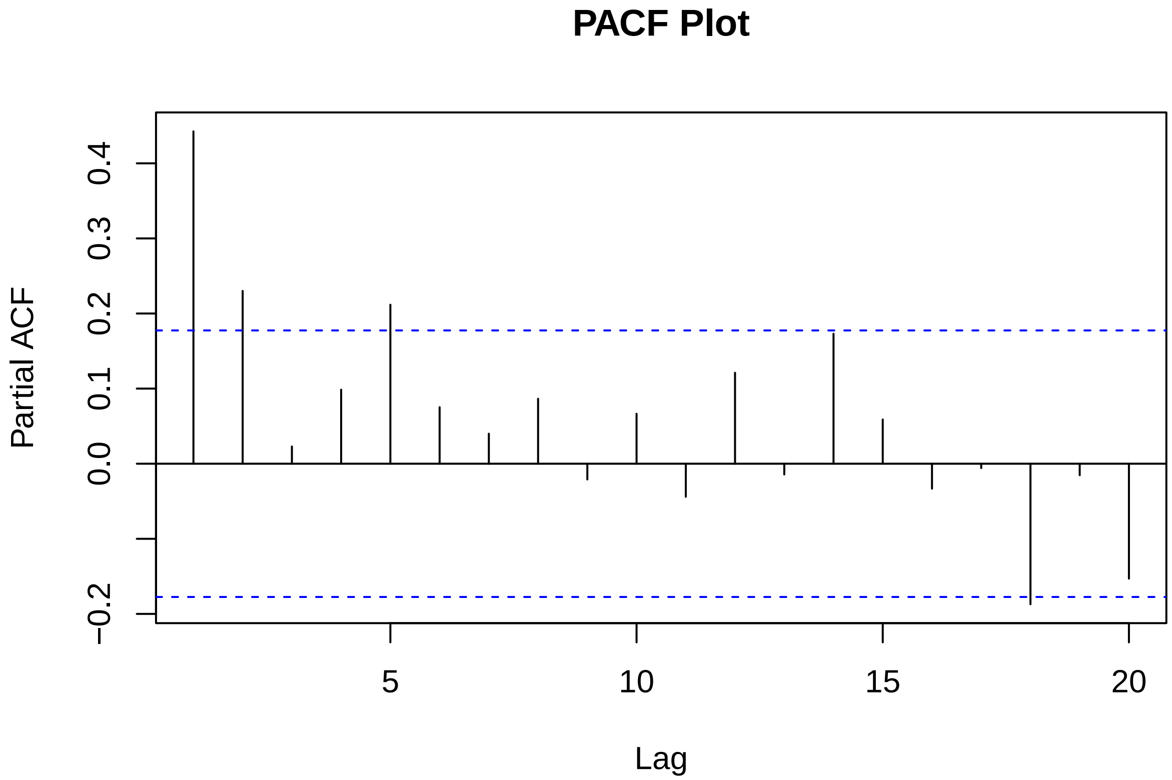
This is typically done by examining plots such as the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), along with experimenting with different values to find the best fitting model based on criteria such as the Akaike Information Criterion (AIC). Which we will see after this part.

```
# Defining the time series object
temperature_ts <- ts(cet_temp$avg_annual_temp_C, start=1900, frequency=1)

# Plot ACF
acf(temperature_ts, main="ACF Plot")
```



```
# Plot PACF
pacf(temperature_ts, main="PACF Plot")
```



The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are important in determining the ARIMA model parameters:

ACF Plot: Shows the correlation of the series with itself, lagged by x time units. The ACF plot gradually decreases, which is typical for a non-stationary time series. This indicates that differencing may be necessary to achieve stationarity.

PACF Plot: Indicates the partial correlation of the series with its own lagged values, controlling for the values of the time series at all shorter lags. The sharp cut-off after the first lag in the PACF plot suggests that an AR(1) model could be a good starting point, meaning the current value is significantly correlated with the immediate previous value. Given the patterns observed in these plots and the non-stationarity indicated by the Augmented Dickey-Fuller test, we might consider an ARIMA model with the following initial parameters:

p (AR term): The PACF plot suggests p=1 could be a good starting point, as there is a significant spike at lag 1 with a sharp decline afterwards. d (I term): Given the non-stationarity of the data, we might start with d=1, meaning we'll differentiate the series once to attempt to make it stationary. q (MA term): The ACF plot's gradual decline suggests some degree of differencing is needed, but the exact q value isn't as clear. We might start with q=1 based on the general guidance that for non-seasonal data, starting with lower values is preferable.

Let's fit an ARIMA(1,1,1) model to the data as a starting point and evaluate its performance. We will later adjust the parameters based on the model fit and diagnostics.

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
temperature_ts <- ts(cet_temp$avg_annual_temp_C, start=1900, frequency=1)

# Fit an ARIMA(1,1,1) model
arima_111_model <- Arima(temperature_ts, order=c(1,1,1))

# View the model summary
summary(arima_111_model)
```

```
## Series: temperature_ts
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.1137  -0.8749
## s.e.  0.1026   0.0454
##
## sigma^2 = 0.3679:  log likelihood = -110.81
## AIC=227.63   AICc=227.83   BIC=236.01
##
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.0777491 0.5990463 0.4767542 0.3909609 3.556565 0.7889396
##                     ACF1
## Training set -0.02525307
```

ARIMA(1,1,1) Summary:

AR1 Coefficient: The AR1 term is 0.1137 with a standard error of 0.1026, and is not statistically significant (p-value would likely be > 0.05). MA1 Coefficient: The MA1 term is -0.8749, with a standard error of 0.0454, and is statistically significant. Sigma^2: The estimated variance of the residuals is 0.3679. Log Likelihood: The model's log likelihood is slightly higher at -110.81 compared to the ARIMA(0,1,1) model. AIC/AICc/BIC: The AIC is 227.63, AICc is 227.83, and BIC is 236.01, which are slightly higher than those for the ARIMA(0,1,1) model.

```
# Fit an ARIMA(0,1,1) model for comparison
arima_011_model <- Arima(temperature_ts, order=c(0,1,1))

# View the model summary
summary(arima_011_model)
```

```
## Series: temperature_ts
## ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.8495
## s.e.   0.0480
##
## sigma^2 = 0.3685:  log likelihood = -111.43
## AIC=226.86   AICc=226.96   BIC=232.45
##
```

```
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.07490648 0.6020313 0.4764914 0.3678678 3.554959 0.7885047
##                    ACF1
## Training set 0.07073286
```

ARIMA(0,1,1) Summary:

MA1 Coefficient: The coefficient for the MA1 term is -0.8495 with a standard error of 0.0480, which is statistically significant. Sigmaˆ2: The estimated variance of the residuals is 0.3685. Log Likelihood: The model's log likelihood is -111.43, which indicates the probability of the data given the model. AIC/AICc/BIC: The Akaike Information Criterion (AIC) is 226.86, corrected AIC (AICc) is 226.96, and the Bayesian Information Criterion (BIC) is 232.45. These metrics are used for model comparison; lower values typically suggest a better model fit with a penalty for complexity.

Analysis:

1.Model Selection: Comparing both models, the ARIMA(0,1,1) has a slightly better (lower) AIC and BIC than the ARIMA(1,1,1), suggesting that it may be a better model due to its simplicity and similar explanatory power.

2.Significance of Coefficients: In the ARIMA(1,1,1) model, the AR1 coefficient is not significant, which implies that the additional AR term may not be providing valuable information to the model. In contrast, the MA1 term is significant in both models.

3.Error Metrics: Both models have similar error measures, which suggest that the predictive accuracy of the two models is quite close.

In the ARIMA(1,1,1) model, there are two parameters to consider:

AR1 (Autoregressive term of order 1): This term attempts to capture any autocorrelation in the data. In simpler terms, it checks if there's a linear relationship between the current year's temperature and the temperature of the previous year. However, in our ARIMA(1,1,1) summary, the AR1 term has a high standard error relative to its coefficient, suggesting that the estimated value of the AR1 coefficient is not reliably different from zero. If this were a significant coefficient, it would mean that last year's temperature (after differencing) has a strong influence on this year's temperature. But, since it's not statistically significant, it implies that the autoregressive part of the model does not provide additional value in explaining the variation in the data.

MA1 (Moving Average term of order 1): This coefficient is significant in both the ARIMA(0,1,1) and ARIMA(1,1,1) models. It accounts for the correlation between an observation and a residual error from a moving average model applied to lagged observations. The significance of this term suggests that the moving average component is useful in modeling the time series. Now, let's see and comment on the information criteria:

AIC (Akaike Information Criterion): It is a widely used criterion for model selection. It balances the model fit with the number of parameters used. The AIC penalizes complexity to avoid overfitting, so lower AIC values are preferred. BIC (Bayesian Information Criterion): It is similar to AIC but applies a larger penalty for models with more parameters. This makes BIC more stringent about complexity, favoring simpler models more than AIC when the sample size is large.

Both AIC and BIC are lower for the ARIMA(0,1,1) model compared to the ARIMA(1,1,1) model, indicating that it has a better trade-off between goodness of fit and simplicity. The ARIMA(0,1,1) model, despite being simpler (it has one less parameter), fits the data almost as well as the more complex ARIMA(1,1,1) model. In time series modeling, simplicity is an asset because it reduces the risk of overfitting (which happens when a model captures the noise in the data rather than the underlying process) and often leads to better out-of-sample prediction.

The lower AIC and BIC values, combined with the non-significant AR1 term in the ARIMA(1,1,1) model, suggest that the simpler ARIMA(0,1,1) model is preferable for forecasting purposes. It provides a similar level of explanation for the data while using fewer parameters, which is generally considered more robust for prediction.

Conclusion: Given the non-significant AR term in the ARIMA(1,1,1) model and the fact that both models have similar error measures and log-likelihoods, it's often recommended to choose the simpler model, which in this case is the ARIMA(0,1,1).

This model is uses fewer parameters, while still capturing the essential features of the data. The lower AIC and BIC values for the ARIMA(0,1,1) model further support this choice.

For performing diagnostics on the chosen model which is ARIMA(0,1,1) to ensure that the residuals (the differences between the observed values and the model's predictions) behave like white noise, which implies they are normally distributed, have constant variance, and are "uncorrelated".

1.Residual Analysis: Check if the residuals of the model are random, which would imply that the model has successfully captured the information in the data. This is usually done by plotting the residuals and looking for any patterns. Ideally, the residuals should show no patterns and look like white noise.

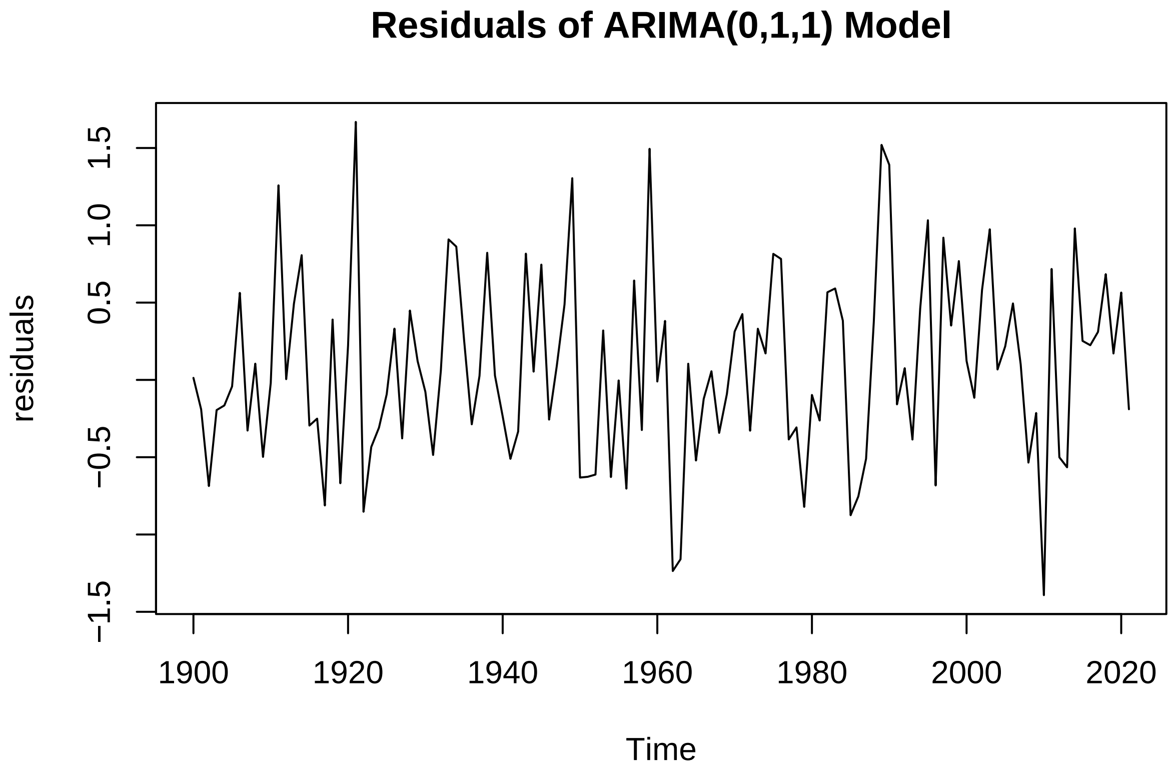
2.Ljung-Box Test: This test checks for autocorrelation in the residuals. If the p-value is high (typically above 0.05), it suggests that there is no significant autocorrelation at lag k.

3.Normality Test: The Jarque-Bera test is often used to test the normality of residuals. If the residuals are normally distributed, it suggests that the model has done a good job of explaining the variance of the data.

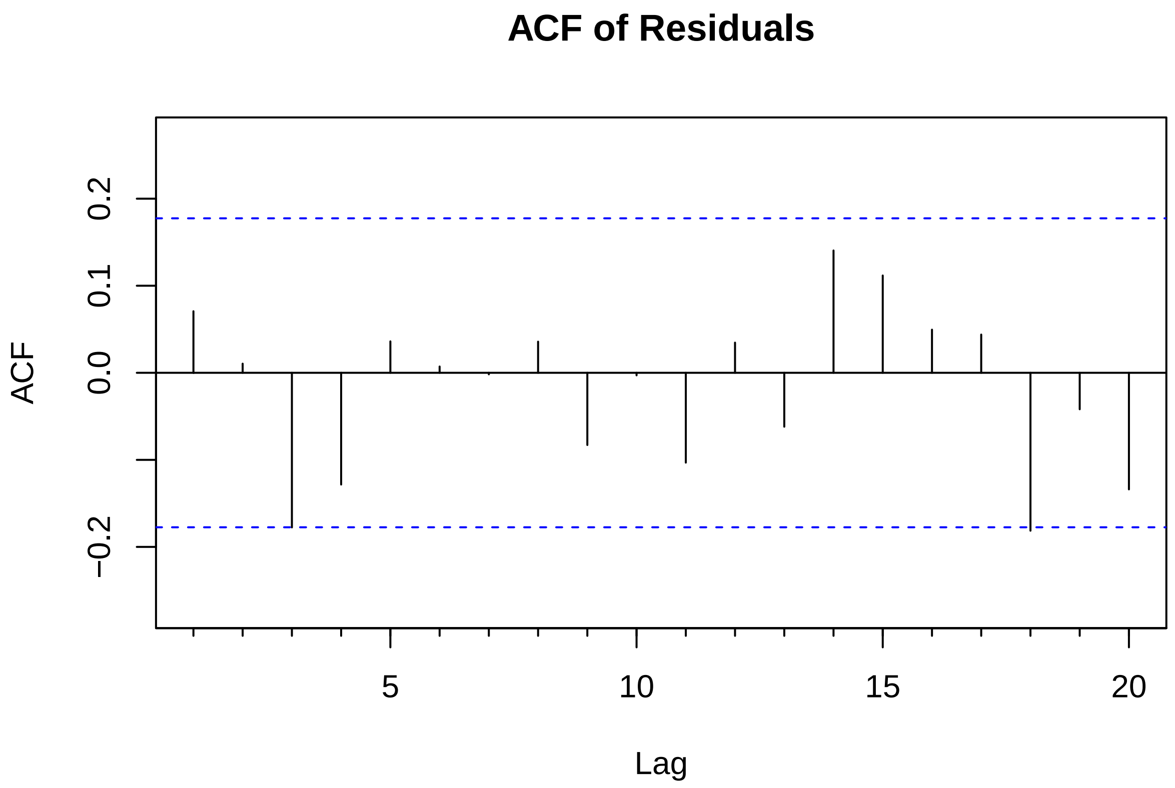
4.Forecasting: so now this model is validated, now we can proceed to use it to make forecasts for future values. The forecast will provide not only the expected value but also the confidence intervals for these predictions for our futher analysis.

```r
library(forecast)
library(tseries)


# 1. Plot residuals
residuals <- residuals(arima_011_model)
plot(residuals, main="Residuals of ARIMA(0,1,1) Model")
```



```
# 2. ACF plot of residuals to check for autocorrelation
Acf(residuals, main="ACF of Residuals")
```



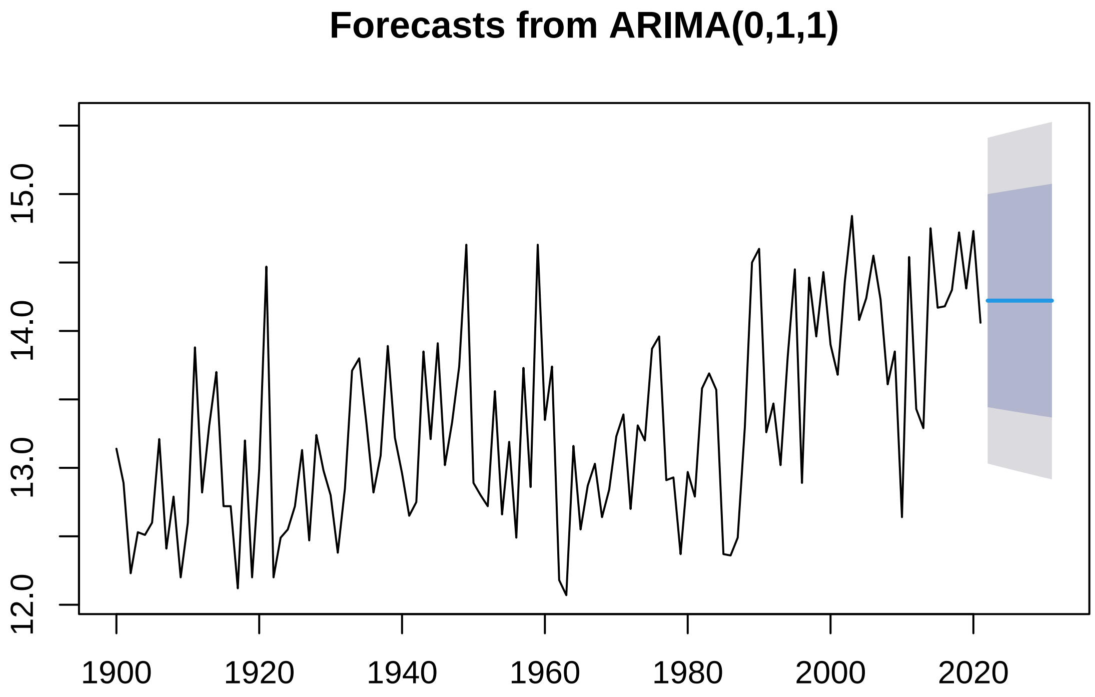
```
# 3. Ljung-Box Test
Box.test(residuals, lag=log(length(residuals)), type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals
## X-squared = 6.7621, df = 4.804, p-value = 0.2199
```

```
# 4. Normality Test of residuals
jarque.bera.test(residuals)
```

```
##
##  Jarque Bera Test
##
## data:  residuals
## X-squared = 1.6265, df = 2, p-value = 0.4434
```

```
# 5. Forecasting future values
forecast_arima011 <- forecast(arima_011_model, h=10) # h is the number of periods for forecasting
plot(forecast_arima011)
```



Analysing Diagnostics Checks.

ACF Plot of Residuals: The ACF plot shows that autocorrelations for all lags are within the confidence bounds (blue dashed lines), which indicates that there is no significant autocorrelation in the residuals. This is a good sign that the model is capturing the temporal structure of the data well.

Box-Ljung Test: The p-value from the Box-Ljung test is 0.2199. Since this value is greater than the common alpha level of 0.05, we fail to reject the null hypothesis that the residuals are independently distributed. This supports the conclusion from the ACF plot of residuals, indicating that there is no significant autocorrelation remaining in the residuals.

Jarque-Bera Test: The Jarque-Bera test has a p-value of 0.4434, which is well above the common significance level of 0.05. This means that we fail to reject the null hypothesis that the residuals are normally distributed. Therefore, we have no evidence to suggest that the residuals do not follow a normal distribution.

Residuals Plot: The residuals plot doesn't show any obvious patterns or systematic structure, which suggests that the residuals are random, supporting the assumption that the model has captured the underlying process adequately.

Forecast Plot: The forecast plot shows the forecasted values and the 80% and 95% prediction intervals. The prediction intervals are reasonably narrow, which indicates a level of precision in the forecasts. The forecast values are in line with the historical data, which suggests that the model has captured the central tendency of the series well.

Conclusion: Overall, the diagnostic checks indicate that the ARIMA(0,1,1) model is well-fitted to the data. The residuals appear to be random (white noise), which suggests that the model has captured the temporal structure in the data. The residuals are also normally distributed, which is an assumption of the ARIMA modeling process.

The forecast plot suggests that the model can be used for forecasting, and the prediction intervals give an indication of the uncertainty associated with these forecasts.

Therefore, based on these diagnostic results, the ARIMA(0,1,1) model seems adequate for forecasting purposes, and we can use this model for making future predictions.