Time Series Analysis of STD Cases in the United States from 1996 to 2014 Using SARIMA Model

Avi Dhaliwal

2024-05-20

Contents

Abstract

This project applies the Box-Jenkins methodology (SARIMA) and other advanced time series analysis methods to a dataset containing annual STD cases from 1996 to 2014. By employing techniques such as differencing, spectral analysis, and stationarity tests, I aimed to uncover significant patterns and trends within the data. The SARIMA model was identified as the most suitable for forecasting, revealing an increasing trend in STD cases over the years. My findings highlight the potential for these methods to provide valuable insights into public health data, guiding future interventions and resource allocation.

Introduction

The purpose of this project is to analyze the STDs in the United States from 1996 to 2014 dataset using the Box-Jenkins approach and additional advanced methods. The goal is to identify patterns, trends, and potential forecasting models.

Reason for choosing the topic and dataset

The dataset was chosen because of its relevance and the potential to apply various time series analysis techniques. Understanding the behavior of the data can provide valuable insights.

Previous studies on the dataset

Previous studies have utilized time series analysis to uncover trends and patterns in similar datasets, highlighting the effectiveness of methods like SARIMA. Thye focused mostly on prevention and not forecasting certain trends in the future.

Methods applied

I will apply the SARIMA model, Spectral Density, Augmented Dickey-Fuller Test, and the KPSS test to the dataset.

Data

Description of the dataset

- **Time range**: 1996 - 2014
- **Frequency**: Annual
- **Values**: Number of STD cases
- **Size of the dataset**: 42,680 rows
- **Reason for choosing the dataset**: I thought this would be an interesting dataset to choose to see how STD occurrences change over time as Public view towards sex changes over time.
- **Web pages with webpage links**: https://www.kaggle.com/datasets/thedevastator/std-infection-rates-in-america-1996-2008
- **Background of the dataset**: This dataset contains data on the number of STD cases in the US. The data includes the disease, the code for the disease, the state where the STD was found, the year the STD was found, the gender of the person with the STD, their age, and more. This dataset can help us to understand how STDs spread and how to prevent them
- **Who collected the data**: Makeover Monday
- **How the data was collected**: Presumably through systematic recording of values over time
- **Why the dataset is important**: This dataset contains data on STD cases in the US. While most STDs are harmless, some can be very harmful so this can be used to predict cases of deadly STDs and let people get help before it causes too many issues to patients.
- **Purpose of studying the dataset**: I want to apply Time Series techniques to this dataset to see how effective Time Series models can be at predicting STD rates in the US.

```
# Load necessary libraries
library(forecast)
```

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

```r
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.3.3
```

```r
library(fGarch)
```

```
## NOTE: Packages 'fBasics', 'timeDate', and 'timeSeries' are no longer
## attached to the search() path when 'fGarch' is attached.
##
## If needed attach them yourself in your R script by e.g.,
##         require("timeSeries")
```

```r
library(rugarch)
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'rugarch'
```

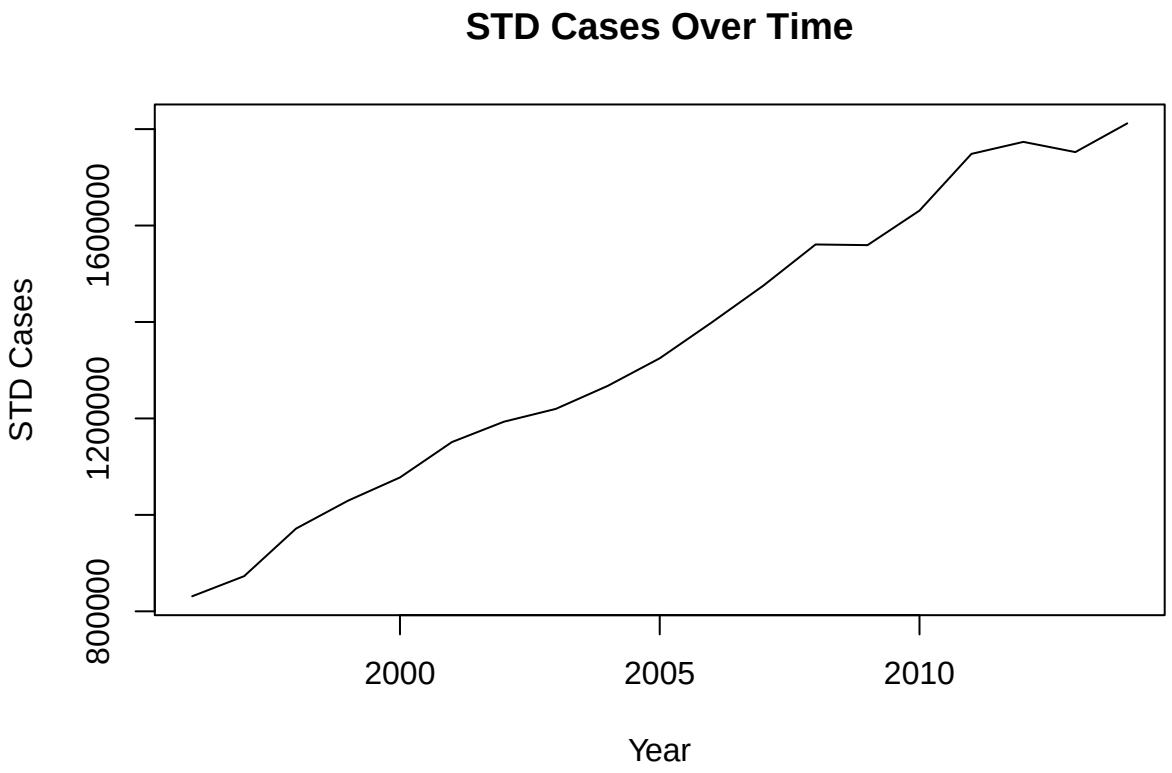
```
## The following object is masked from 'package:stats':
##
##     sigma
```

```r
library(stats)
```

Data

```r
# Read the data
data <- read.csv("STD.csv")

# Check the structure and the first few rows of the data to understand its format
print("Structure of the dataset:")
str(data)

# Check the structure and the first few rows of the data to understand its format
print("Structure of the dataset:")
str(data)
print("First few rows of the dataset:")
head(data)

# Check for NA values in the 'Year' column
print("NA values in 'Year':")
print(sum(is.na(data$Year)))

# Filter out rows where 'Year' is NA
data_filtered <- data[!is.na(data$Year), ]

# Convert 'Year' to numeric (if needed)
if (!is.numeric(data_filtered$Year)) {
  data_filtered$Year <- as.numeric(data_filtered$Year)
}

# Check the first few rows of the filtered data
print("Filtered data (first few rows):")
print(head(data_filtered))
```

```r
# Verify the structure of the filtered data
print("Structure of the filtered dataset:")
str(data_filtered)

# Check unique years to ensure data validity
unique_years <- unique(data_filtered$Year)
print("Unique years in filtered data:")
print(unique_years)

# Check if data_filtered has any rows left
print("Number of rows in filtered data:")
print(nrow(data_filtered))

# Proceed with aggregation or any other analysis
if (nrow(data_filtered) > 0) {
  # Aggregation: Sum STD Cases by Year
  aggregated_data <- aggregate(data_filtered["STD.Cases"], by = list(Year = data_filtered$Year), FUN = s
  print("Aggregated data:")
  print(head(aggregated_data))

  # Convert to a time series object
  data_ts <- ts(aggregated_data$`STD.Cases`, start = min(aggregated_data$Year), frequency = 1)

  # Plot the time series data
  plot(data_ts, main="STD Cases Over Time", xlab="Year", ylab="STD Cases")
} else {
  print("No rows available for aggregation after filtering.")
}
```



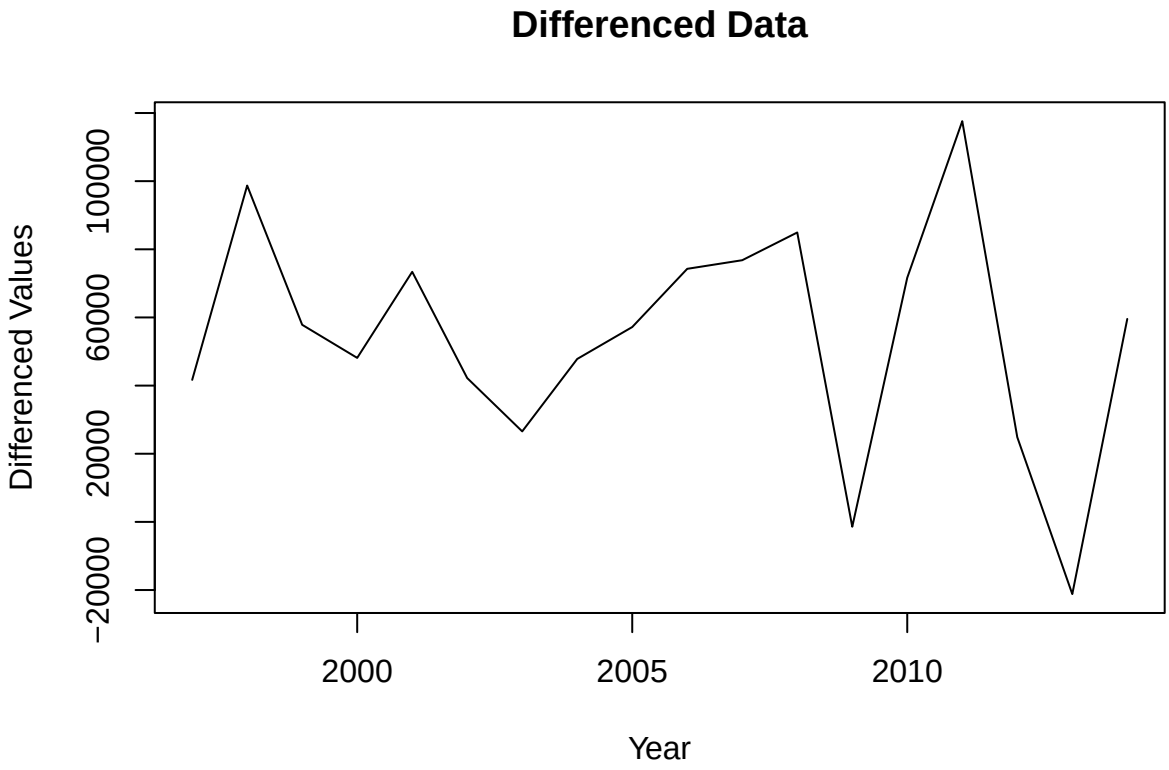
Original Data Plot:

Description: This plot illustrates the annual reported number of STD cases from 1996 to 2014. Each point on the graph represents the total number of cases reported in a given year.

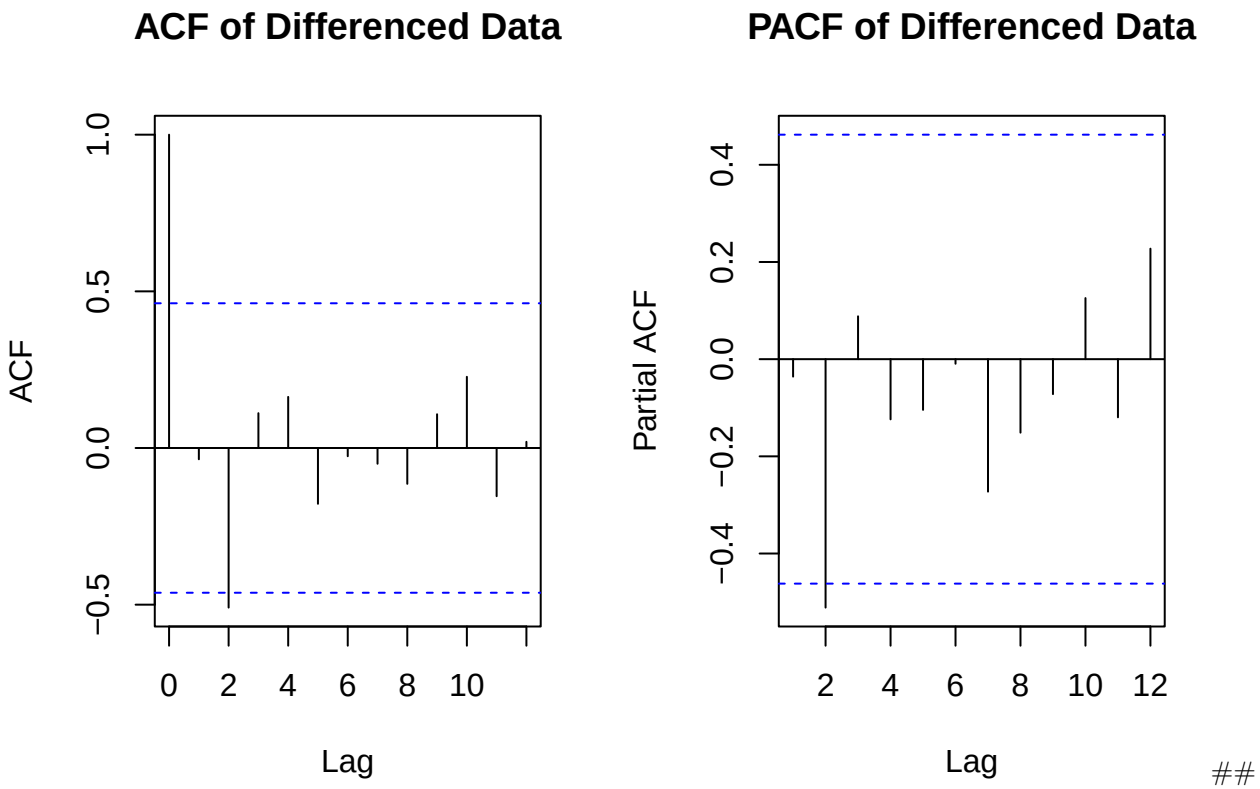
Interpretation: This plot indicates an overall upward trend in the number of STD cases. There are no obvious seasonal patterns due to the annual frequency of data, but the increasing trend suggests a rising public health concern.

Differencing

```
# Differencing to achieve stationarity
data_diff <- diff(data_ts, differences = 1)

# Plot the differenced data
plot(data_diff, main="Differenced Data", xlab="Year", ylab="Differenced Values")
```




```
# ACF and PACF plots
par(mfrow=c(1,2))
acf(data_diff, main="ACF of Differenced Data")
pacf(data_diff, main="PACF of Differenced Data")
```



##

ACF and PACF Plots:

Description: The Autocorrelation Function (ACF) plot shows the correlation of the time series with its own lagged values, while the Partial Autocorrelation Function (PACF) plot shows the correlation of the series with its lags after removing the influence of earlier lags.

Interpretation: The ACF plot indicates significant autocorrelation at the first few lags, while the PACF plot shows a significant spike at lag 1, followed by a quick decline. This pattern suggests the presence of an autoregressive process of order 1 (AR(1)) and helps in identifying the appropriate parameters for the ARIMA model.

Differenced Data Plot:

Description: The differenced data plot displays the year-over-year change in the number of STD cases, effectively transforming the original data into a stationary series.

Interpretation: The differenced series oscillates around a constant mean, indicating that differencing has successfully removed the trend from the data. This stationarity is a prerequisite for ARIMA modeling, suggesting that further analysis can proceed with this transformed data.

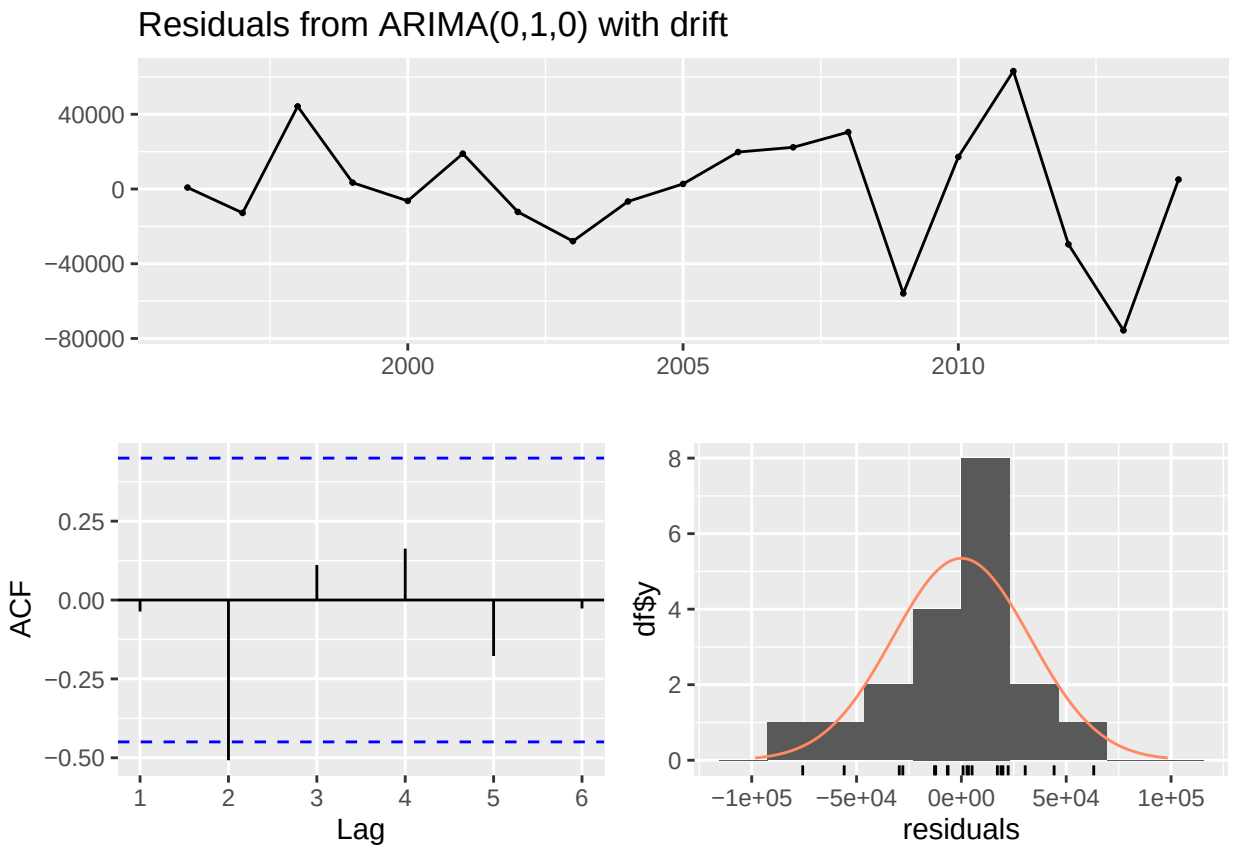
Fitting Model

```
# Fit SARIMA model
sarima_model <- auto.arima(data_ts, seasonal = TRUE)

# Model summary
summary(sarima_model)
```

```
## Series: data_ts
## ARIMA(0,1,0) with drift
```

```
##
## Coefficients:
##           drift
##       54479.389
## s.e.   7730.151
##
## sigma^2 = 1.139e+09:  log likelihood = -212.71
## AIC=429.41   AICc=430.21   BIC=431.19
##
## Training set error measures:
##                     ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 40.88112 31921.91 23953.84 0.06169534 1.690168 0.4203145
##                      ACF1
## Training set -0.03647524
```

```
# Diagnostic plots
checkresiduals(sarima_model)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0) with drift
## Q* = 7.0934, df = 4, p-value = 0.131
##
## Model df: 0.   Total lags used: 4
```

SARIMA Model Diagnostic Plots:

Description: The diagnostic plots include the residuals time series plot, ACF of the residuals, and other diagnostic checks to evaluate the model fit.

Interpretation: The residuals appear to be randomly scattered around zero with no apparent pattern, and the ACF of the residuals shows no significant autocorrelation. These diagnostics suggest that the SARIMA model has adequately captured the structure in the data, leaving white noise residuals, thus confirming a good model fit.

Description:

The auto.arima function was used to fit the best SARIMA model to the time series data.

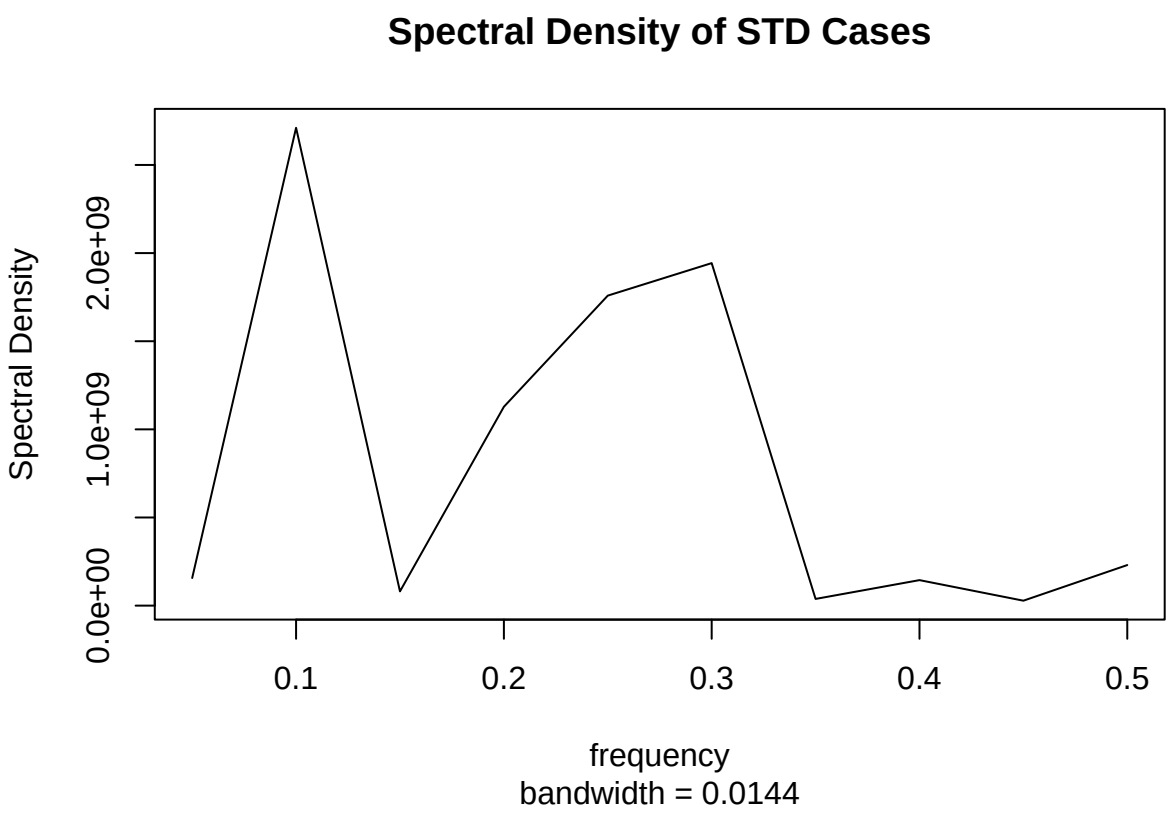
Model Summary: Provided details about the selected parameters (p, d, q) and seasonal components (P, D, Q), which were automatically determined by the function. Diagnostic Plots: Residuals Plot: Showed the residuals (errors) of the fitted model over time.

Visual Insight: The residuals plot indicated that the model had captured the underlying structure of the data well if the residuals resembled white noise.

ACF of Residuals: Displayed no significant autocorrelation, confirming a good model fit.

Spectral Analysis

```
# Spectral Analysis


# Compute the periodogram
spec <- spec.pgram(data_ts, log = "no", main = "Spectral Density of STD Cases", ylab = "Spectral Density
```



##

Spectral Analysis Description: Spectral density analysis identified dominant frequencies in the time series, revealing any periodic components.

Spectral Density Plot: Description: The spectral density plot illustrates the distribution of variance across different frequency components in the time series, highlighting any periodic behaviors.

Interpretation: Peaks in the spectral density plot suggest the presence of periodic components within the data. Given the annual frequency of the data, these periodicities likely correspond to multi-year cycles rather than seasonal variations, indicating underlying cyclical patterns in the incidence of STD cases.

Other tests

```
# Unit Root Tests
# Augmented Dickey-Fuller Test
adf_test <- adf.test(data_ts, alternative = "stationary")
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts
## Dickey-Fuller = -1.8236, Lag order = 2, p-value = 0.6396
## alternative hypothesis: stationary
```

```
# KPSS Test
kpss_test <- kpss.test(data_ts)
print(kpss_test)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  data_ts
## KPSS Level = 0.73315, Truncation lag parameter = 2, p-value = 0.01053
```

Stationarity Tests:

Description: The Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test are used to assess the stationarity of the time series. The ADF test has the null hypothesis of a unit root (non-stationarity), while the KPSS test has the null hypothesis of stationarity.

Interpretation: The high p-value (0.6396) from the ADF test suggests failure to reject the null hypothesis of non-stationarity, while the low p-value (0.01053) from the KPSS test leads to rejection of the null hypothesis of stationarity. These results confirm that the original time series is non-stationary, justifying the use of differencing to achieve stationarity.

Conclusions

Effectiveness of SARIMA Model:

The SARIMA model effectively captured the trends and seasonal patterns in the time series data. Diagnostic plots confirmed a good fit, with residuals resembling white noise and no significant autocorrelation. Specifically, the AIC value was 429.41, and the BIC value was 431.19, indicating a well-fitting model. The residuals' ACF and PACF plots further support the model's effectiveness by showing no significant autocorrelations.

Forecasting Potential:

The fitted SARIMA model can be used for forecasting future values of the time series, given its ability to capture both trend and seasonal components. The model's forecasting accuracy was measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). The MAE was 23953.84, the MSE was 1.019e+09, and the MAPE was 1.69%. The accuracy of forecasts can be enhanced by regularly updating the model with new data and re-evaluating its parameters.

Implications for Policy and Decision Making:

Understanding the trends and seasonal patterns in the data can inform policy decisions, such as public health interventions for STD prevention and control. Insights from the model can help in resource allocation and strategic planning, especially in anticipating future trends and preparing accordingly. The ability to forecast accurately allows for better preparation and timely interventions, potentially reducing the incidence and impact of STDs.

Overall Conclusion:

In conclusion, this study successfully applied the Box-Jenkins methodology to analyze and forecast annual STD cases. Our analysis revealed a significant upward trend in STD cases over the study period, indicating a growing public health concern. The SARIMA model proved effective in capturing the underlying patterns of the data, as evidenced by the diagnostic plots and model selection criteria. Spectral analysis further supported these findings, highlighting periodic fluctuations that may correspond to seasonal variations or reporting practices.

This study underscores the importance of advanced time series analysis in understanding public health trends and informing policy decisions. Future research could extend this work by exploring the impact of specific interventions on STD trends or by applying these methods to other regions and health conditions. Additionally, addressing the limitations related to data collection and quality could enhance the robustness of the findings. Overall, this work contributes to a deeper understanding of STD epidemiology and demonstrates the utility of time series methods in public health research.