Name: Sanket Khamkar

NET ID: srk 197

**BF Project Report**

1. **State your forecasting question and its importance to you**

Based on the data from the **CDC Natality Dashboard**, [National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm) which includes both fertility and mortality rates:

**"Can we accurately predict future monthly trends in fertility and infant mortality rates in the United States for the next few years?"**

**Importance of This Question:**

Accurate forecasting of fertility and infant mortality rates is significant for several key reasons:

1. **Healthcare Resource Allocation**: Predicting future trends in fertility rates allows healthcare institutions to allocate resources such as maternal health services, neonatal care, and hospital staffing to meet expected needs effectively.
2. **Public Health Policy**: Forecasting fertility and infant mortality trends assists public health officials in developing policies to improve maternal and infant health outcomes. For instance, identifying an expected rise in fertility rates would prompt the expansion of maternal health programs.
3. **Economic and Social Planning**: Fertility trends impact the future population demographics, influencing the planning of educational institutions, childcare facilities, and broader economic policies. Accurate forecasting can support better long-term planning for a stable and healthy population growth.
4. **Understanding Disparities**: By forecasting fertility and infant mortality trends, it is possible to identify and address disparities across different geographic or socioeconomic groups, contributing to equity in healthcare.

Overall, the ability to predict these trends allows for proactive measures to improve public health outcomes, efficient allocation of resources, and informed decision-making across social and economic domains.

1. **Describe the data**

The **CDC Natality Dashboard** provides data on both fertility and infant mortality rates across different age and ethnicity groups and time periods in the United States. Here's a breakdown of the data:

**Data Description:**

1. **Fertility Rate**:
   * **Definition**: The number of live births per 1,000 women of childbearing age (example ages 15-44).
   * **Attributes**:
     + **Time Period**: The data is collected quarterly.
     + **Demographics**: The fertility rate data can include attributes such as age group, race/ethnicity, and marital status of mothers.
     + **Trends**: The data typically show seasonal variations, and it may reflect broader societal trends like changes in economic conditions or healthcare policies.
2. **Infant Mortality Rate**:
   * **Definition**: The number of deaths of infants under one year of age per 1,000 live births.
   * **Attributes**:
     + **Time Period**: Collected monthly or annually, similar to fertility rate data.
     + **Demographics**: This data can also be broken down by maternal factors (e.g., age, race/ethnicity)
     + **Healthcare Factors**: Mortality rate data may also include healthcare-related attributes, like access to prenatal care or medical interventions during birth.

**Data Format:**

The data provided through the CDC dashboard may include:

* **Date Field**: Indicating the time frame of the data, such as Quarterly month Q1, Q2, Q3, Q4.
* **Numerical Values**: The actual fertility or mortality rates for the given periods.

The fertility and infant mortality data can be used for time-series forecasting, as both metrics tend to have historical trends and patterns. By analyzing this data, we can develop models that help predict future values, which is useful for healthcare planning, policy-making, and addressing inequalities.

1. **Insights from Exploratory Data Analysis**

Here are some insights we can derive from an Exploratory Data Analysis (EDA) of the CDC's fertility and infant mortality data:

**1. Time Series Trends:**

* **Fertility Rate Trends**: The data may reveal general trends in fertility rates over the years. For instance, if the rates are decreasing year-on-year, this could be indicative of social changes like delayed childbirth, increased contraceptive use, or economic pressures.
* **Mortality Rate Trends**: There may be a downward trend in infant mortality, which often reflects improvements in healthcare access, prenatal care, and medical interventions.

**2. Seasonality Patterns:**

* **Seasonal Variation in Fertility Rates**: The data may exhibit seasonal variation, with certain months showing consistently higher birth rates. For example, more births might be observed during specific times of the year due to social, cultural, or climatic influences.
* **Seasonal Patterns in Mortality Rates**: Infant mortality could also show some seasonal effects, which may be linked to changes in weather conditions or seasonal illnesses.

**3. Regional Analysis:**

* **Geographic Differences**: Fertility and mortality rates could vary significantly by ethnic groups. Ethnic groups with good financial condition might show better outcomes, such as lower infant mortality rates. Similarly, fertility trends may vary based on local economic or cultural factors.
* **Rural vs. Urban**: Differences in rates between urban and rural areas can highlight disparities in healthcare services, economic conditions, or education levels, which can influence both birth and infant mortality rates.

**4. Demographic Insights:**

* **Age Groups**: EDA may reveal that younger or older maternal age groups have differing fertility and infant mortality rates. Teenage mothers or mothers above a certain age may have higher associated risks of infant mortality.
* **Race/Ethnicity**: Differences between ethnic groups in fertility rates and infant mortality may reflect disparities in healthcare access, socioeconomic conditions, and cultural factors. Identifying such differences is crucial for targeted public health interventions.
* **Marital Status**: Mothers’ marital status could also play a role, affecting both fertility and infant mortality rates, often linked to social support structures and economic stability.

**5. Outliers and Anomalies:**

* **Outlier Analysis**: There may be sudden spikes or dips in the fertility or mortality rates, potentially corresponding to significant events. For example, a significant decline in birth rates during a period of economic recession or health crisis could appear as an outlier in the data.
* **Event-Based Anomalies**: Events like pandemics or natural disasters can cause deviations from normal patterns. For instance, the COVID-19 pandemic had impacts on healthcare availability, which could show up as anomalies in infant mortality rates during the affected years.

**6. Correlation Analysis:**

* **Correlation Between Variables**: Examining the correlation between factors, such as the correlation between maternal age and infant mortality, can provide important insights. High correlations might reveal underlying risk factors that should be prioritized in interventions.
* **Healthcare Access and Outcomes**: If available, correlating healthcare-related features (e.g., access to prenatal care, number of healthcare visits) with outcomes like infant mortality can help understand the importance of medical interventions in improving child health outcomes.

**7. Missing Data Patterns:**

* **Missing Values**: Analyzing missing data patterns can help us understand if certain data points are consistently missing for particular groups, which might indicate underlying issues, such as lack of access to healthcare or disparities in data collection across different states or demographics.

**Summary:**

The Exploratory data analysis of the CDC natality data provides a foundational understanding of the underlying patterns, trends, and factors that influence fertility and infant mortality in the U.S. Insights from this can guide forecasting model selection, help identify important features, and direct public health efforts. Understanding these trends can ultimately support better decision-making and resource allocation in maternal and child health programs.

1. **State your Accuracy measure and its importance to you**

For evaluating the forecasting models, I have focused on three key accuracy measures: **MAE (Mean Absolute Error)**, **RMSE (Root Mean Squared Error)**, and **MASE (Mean Absolute Scaled Error)**. Here is a detailed explanation of each measure and why it is important:

* **MAE** gives an easy-to-interpret average error, showing how much deviation you can expect on average.
* **RMSE** emphasizes larger errors, helping ensure that significant deviations are minimized.
* **MASE** provides a comparative measure across time series, making it easier to judge the quality of different models relative to a naive benchmark.

Using these three metrics helps provide a comprehensive evaluation of each forecasting model, focusing on **average performance** (MAE), **penalty for larger errors** (RMSE), and **comparison against baseline models** (MASE).

1. **Insights from different forecasting methods and their residual analysis**

```{r}

options(repos = c(CRAN = "https://cloud.r-project.org"))

#install.packages

library(readxl)

library(knitr)

library(TTR)

library(dplyr)

library(ggplot2)

library(forecast)

library(tidyverse)

library(tseries)

# Load the Excel file

data <- read.csv("C:/Users/Sanket Khamkar/Downloads/Data Dictionary\_Sanket.csv")

names(data)

knitr::kable(data)

head(data)

# Function to convert 'Year\_Quarter' to start date for each quarter

convert\_quarter\_to\_date <- function(quarter) {

if (is.na(quarter)) {

return(NA) # Return NA if quarter is missing

}

year <- as.numeric(substr(quarter, 1, 4))

q <- substr(quarter, 6, 7)

if (q == "Q1") {

return(as.Date(paste0(year, "-01-01")))

} else if (q == "Q2") {

return(as.Date(paste0(year, "-04-01")))

} else if (q == "Q3") {

return(as.Date(paste0(year, "-07-01")))

} else if (q == "Q4") {

return(as.Date(paste0(year, "-10-01")))

} else {

return(NA) # Return NA if the quarter is not in a recognized format

}

}

# Apply the conversion to the 'Year Quarter' column

data$Date <- sapply(data$`Year\_Quarter`, convert\_quarter\_to\_date)

# Convert the 'Rate' column to numeric and clean up the data

data$FertilityRate <- as.numeric(data$FertilityRate)

data <- data[!is.na(data$Date) & !is.na(data$FertilityRate), ]

data$MortalityRate <- as.numeric(data$MortalityRate)

data <- data[!is.na(data$Date) & !is.na(data$MortalityRate), ]

# Ensure Date column is a factor (or convert as needed)

data$Date <- as.factor(data$Date)

```

```{r}

# Filter for 'All races and origins' and set time series

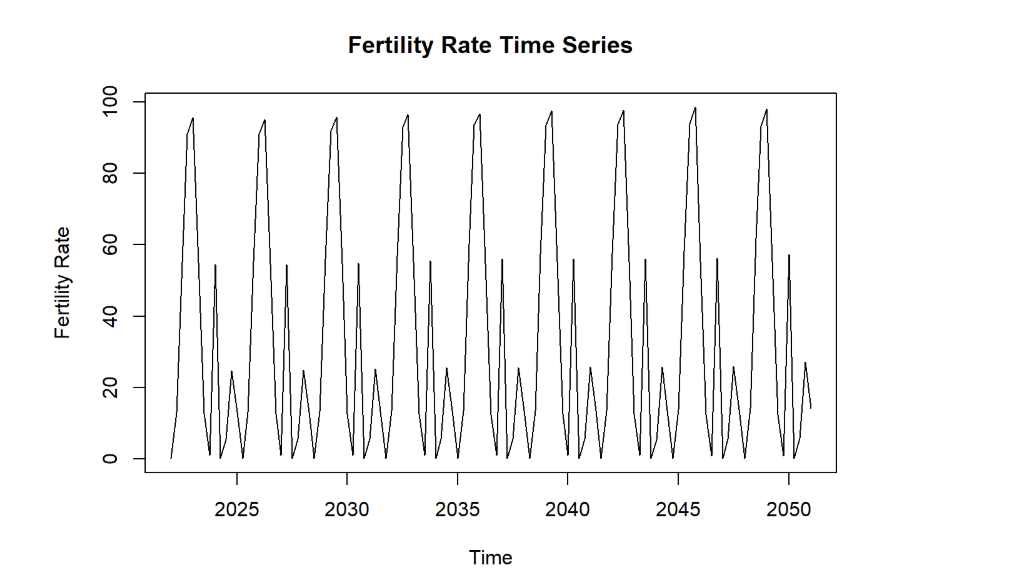
data\_filtered <- subset(data, Group == "All races and origins")

ts\_data <- ts(data\_filtered$FertilityRate, start = c(2022, 1), frequency = 4)

**# Plot the time series**

plot(ts\_data , main = "Fertility Rate Time Series", ylab = "Fertility Rate", xlab = "Time")

```



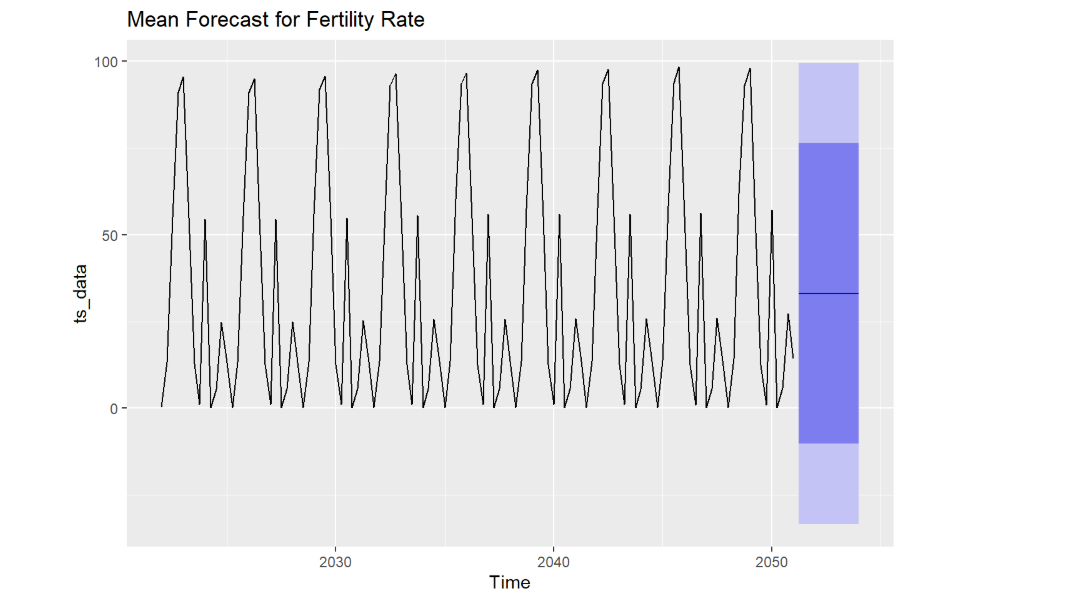
```{r}

**#Mean forecast**

mean\_model <- meanf(ts\_data, h = 12)

# Plot the forecast

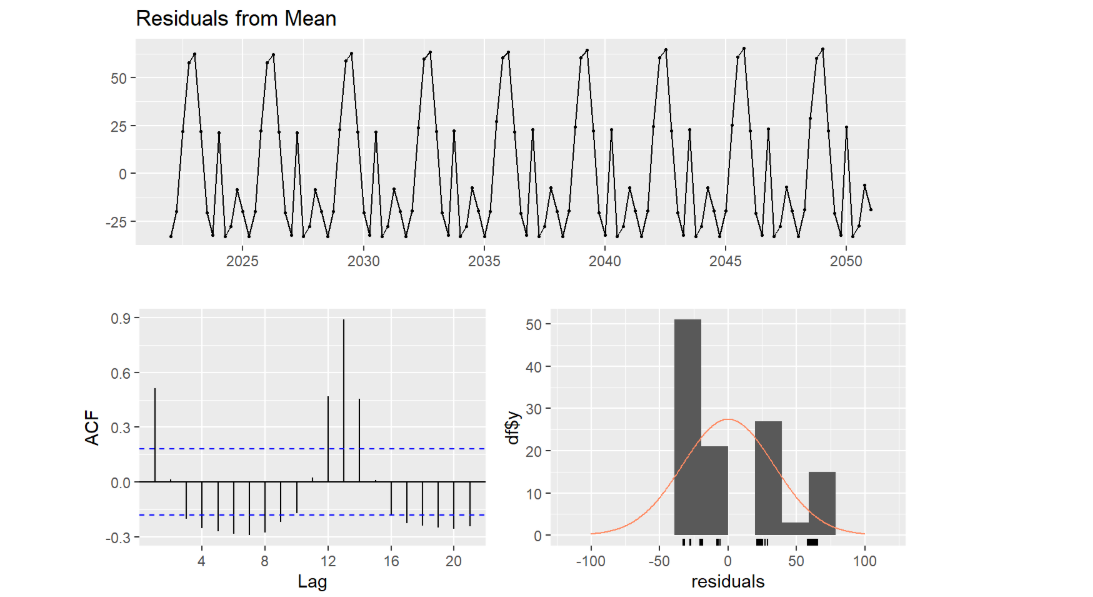
autoplot(mean\_model) + ggtitle("Mean Forecast for Fertility Rate")



forecast(mean\_model)

# Residual Analysis

checkresiduals(mean\_model)



```

Uses the average of past data as the forecast. Best for stationary data without trends or seasonality. Residuals should be random, indicating that the mean is a suitable representation.

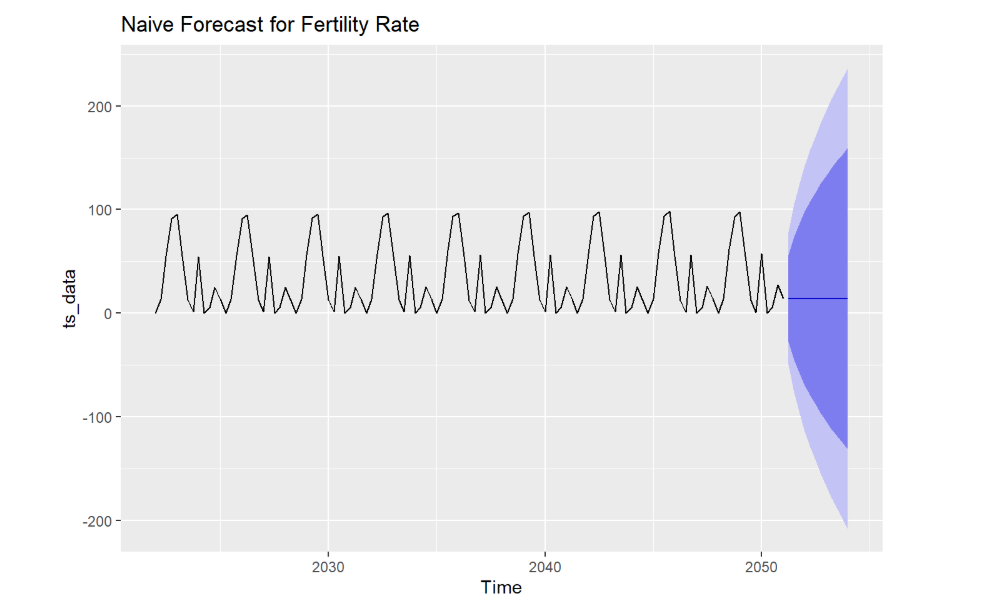
```{r}

**#Naive forecast**

naive\_model <- naive(ts\_data, h = 12)

# Plot the forecast

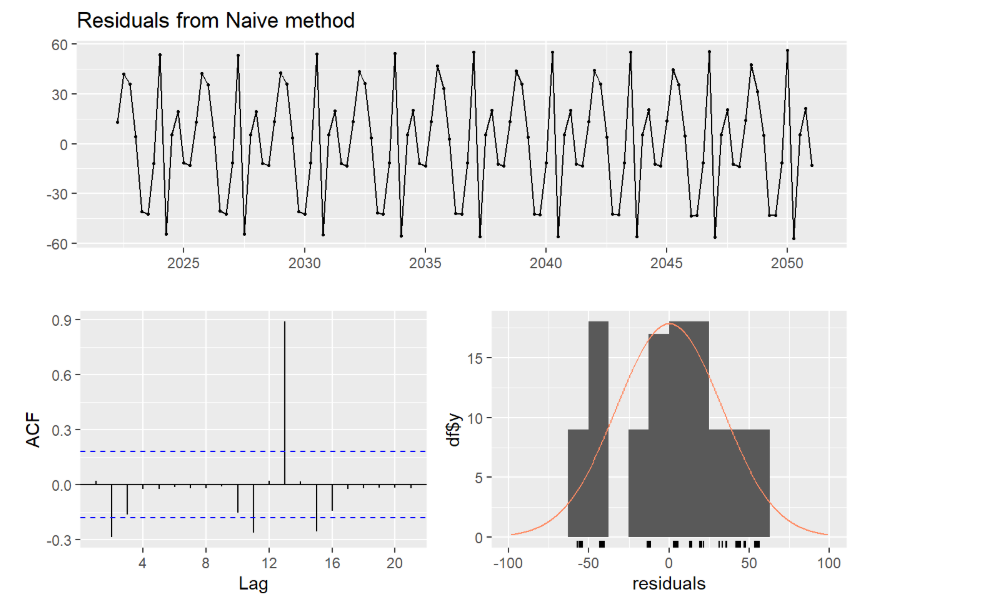
autoplot(naive\_model) + ggtitle("Naive Forecast for Fertility Rate")



forecast(naive\_model)

# Residual Analysis

checkresiduals(naive\_model)



```

Suitable for short-term forecasts where the most recent observation is the best predictor. Ideal for data without trends or seasonality. Random residuals indicate a good fit.

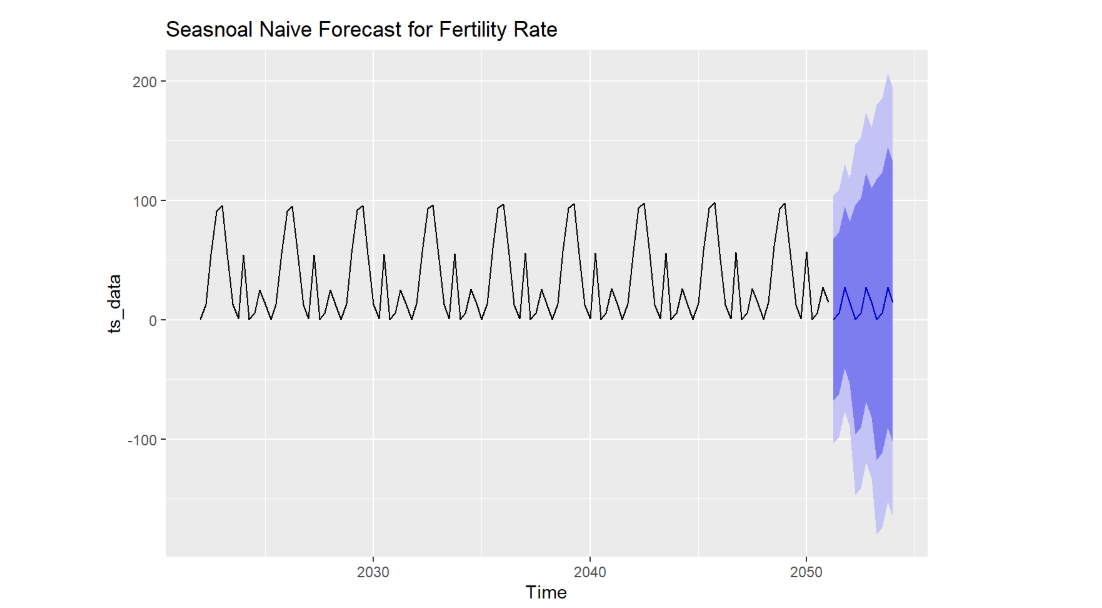
```{r}

**#Seasonal Naive forecast**

snaive\_forecast <- snaive(ts\_data, h = 12)

# Plot the forecast

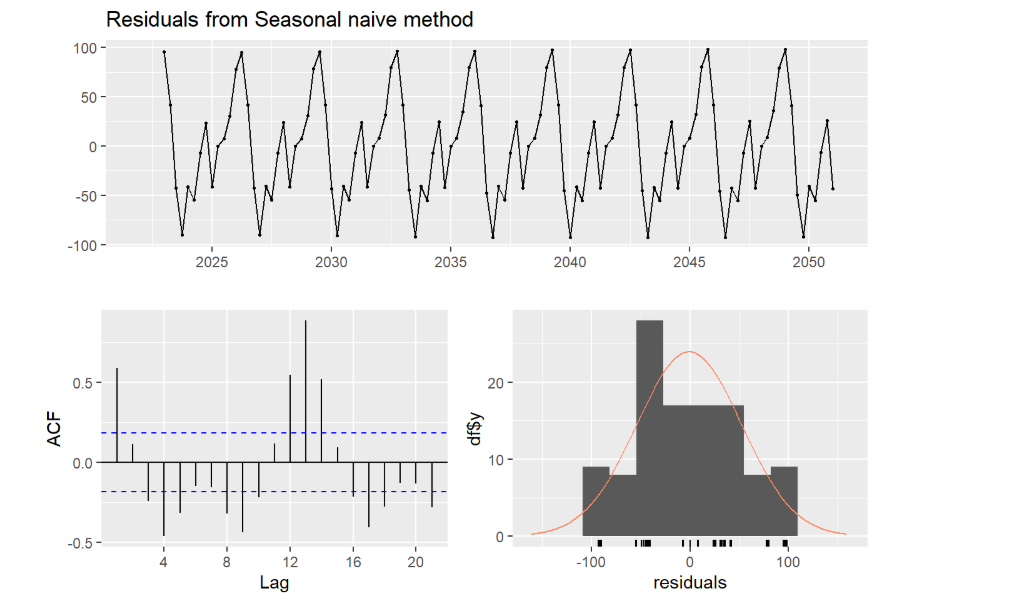
autoplot(snaive\_forecast) + ggtitle("Seasnoal Naive Forecast for Fertility Rate")



forecast(snaive\_forecast)

# Residual Analysis

checkresiduals(snaive\_forecast)



```

Suitable for data with strong seasonal patterns. It uses the value from the same season in the previous cycle. Residuals should be random if seasonality is captured effectively.

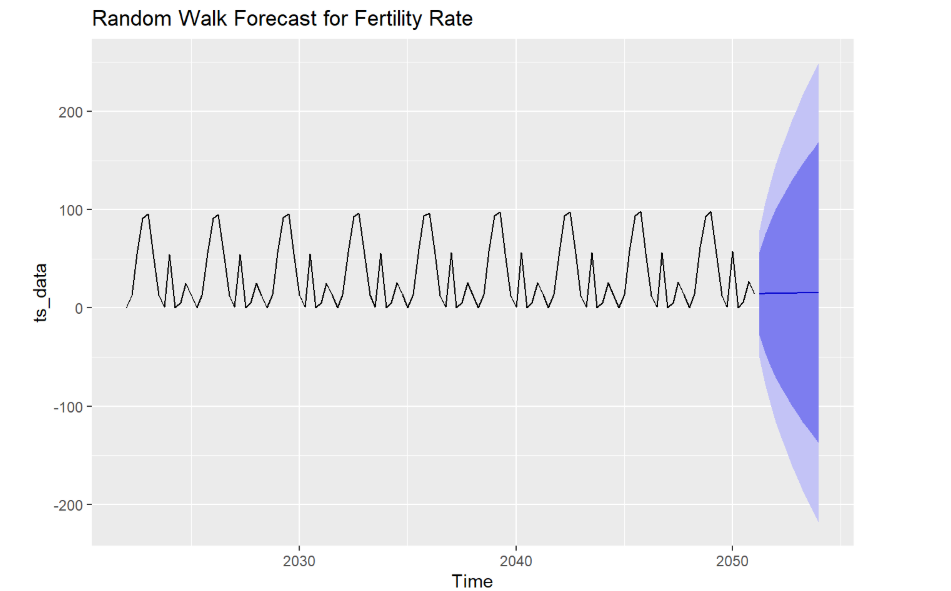
```{r}

**#Random walk forecast**

rwf\_forecast <- rwf(ts\_data,h = 12)

rwf\_forecast <- rwf(ts\_data,h = 12, drift=TRUE)

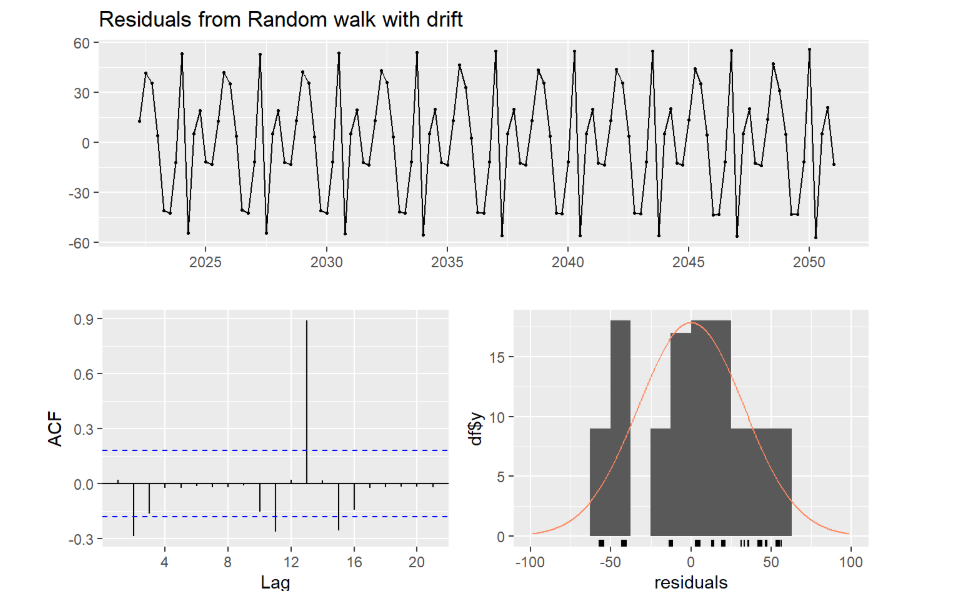
autoplot(rwf\_forecast) + ggtitle("Random Walk Forecast for Fertility Rate")



forecast(rwf\_forecast)

# Residual Analysis

checkresiduals(rwf\_forecast)



```

Useful for data that has a persistent random walk behavior, sometimes with a drift representing a trend. Residuals should be random if the drift captures the trend accurately.

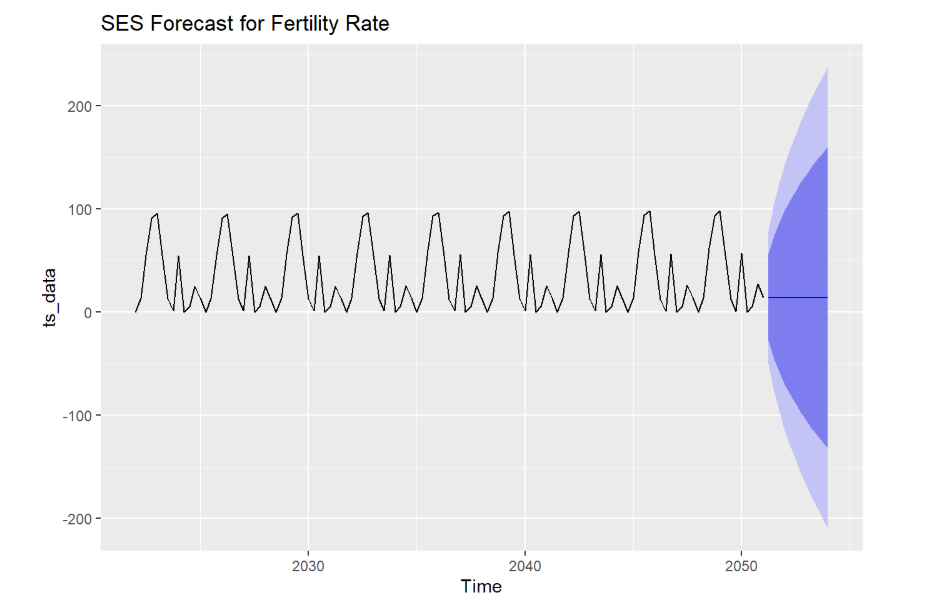
```{r}

**# Simple Exponential Smoothing**

ses\_model <- ses(ts\_data, h = 12)

# Plot the forecast

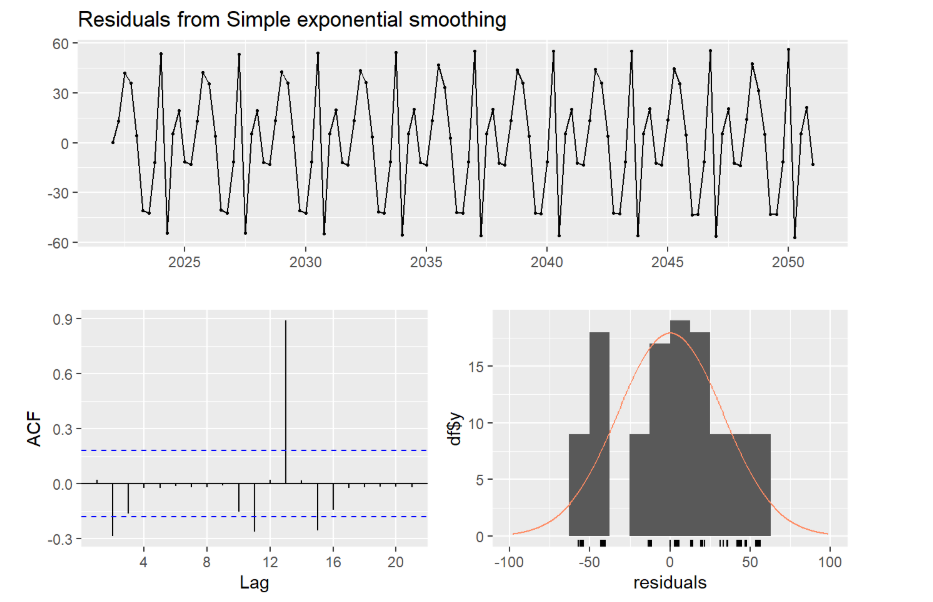
autoplot(ses\_model) + ggtitle("SES Forecast for Fertility Rate")



forecast(ses\_model)

# Residual Analysis

checkresiduals(ses\_model)



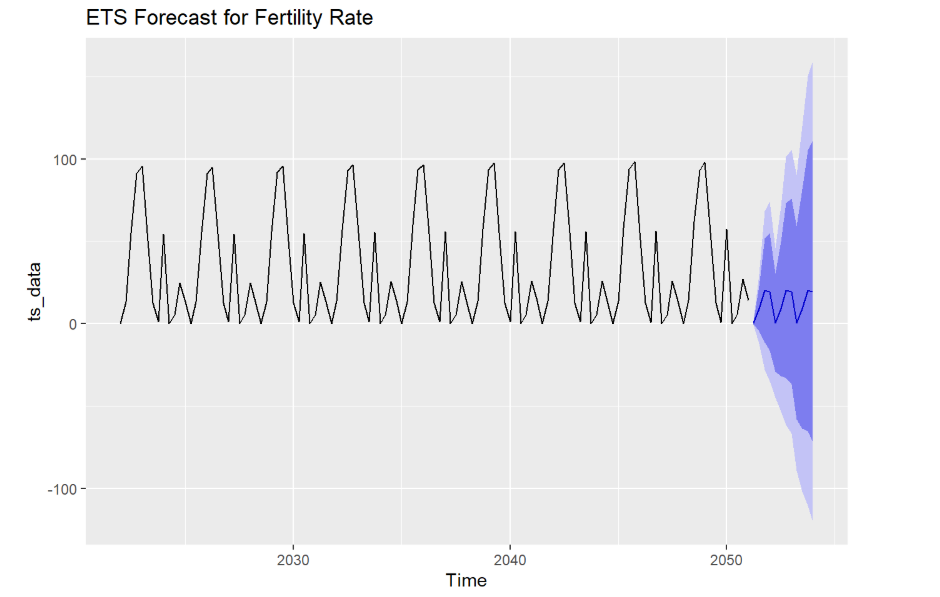
```

Works well for level data without trend or seasonality. Residuals should resemble white noise if the model fits correctly.

```{r}

**# Fit an ETS model**

ets\_model <- ets(ts\_data)



# Forecast the next 12 months

ets\_forecast <- forecast(ets\_model, h = 12)

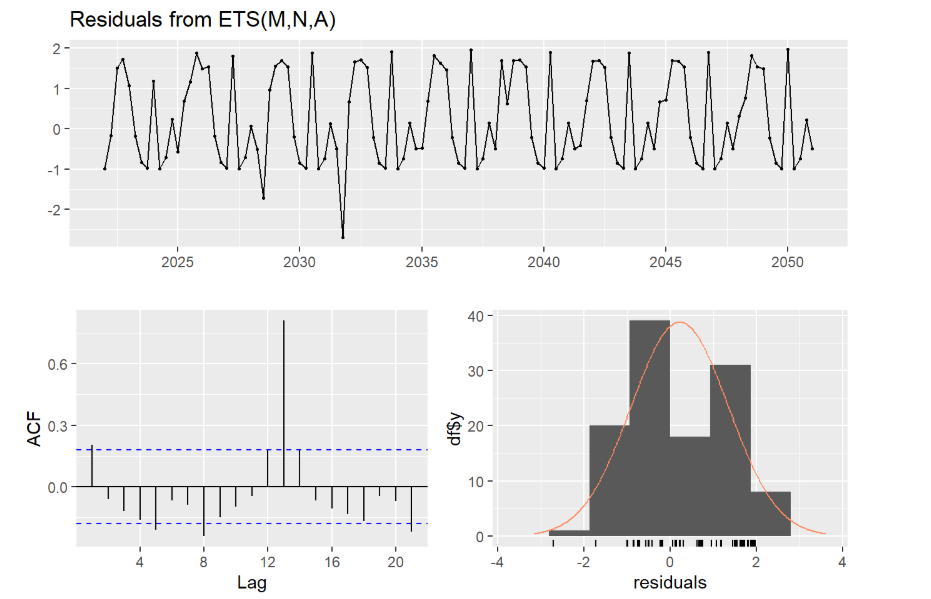
# Plot the forecast

autoplot(ets\_forecast) + ggtitle("ETS Forecast for Fertility Rate")

forecast(ets\_forecast)

# Residual Analysis

checkresiduals(ets\_forecast)



```

Captures level, trend, and seasonality effectively. Residuals should be random and without autocorrelation to indicate a good model fit.

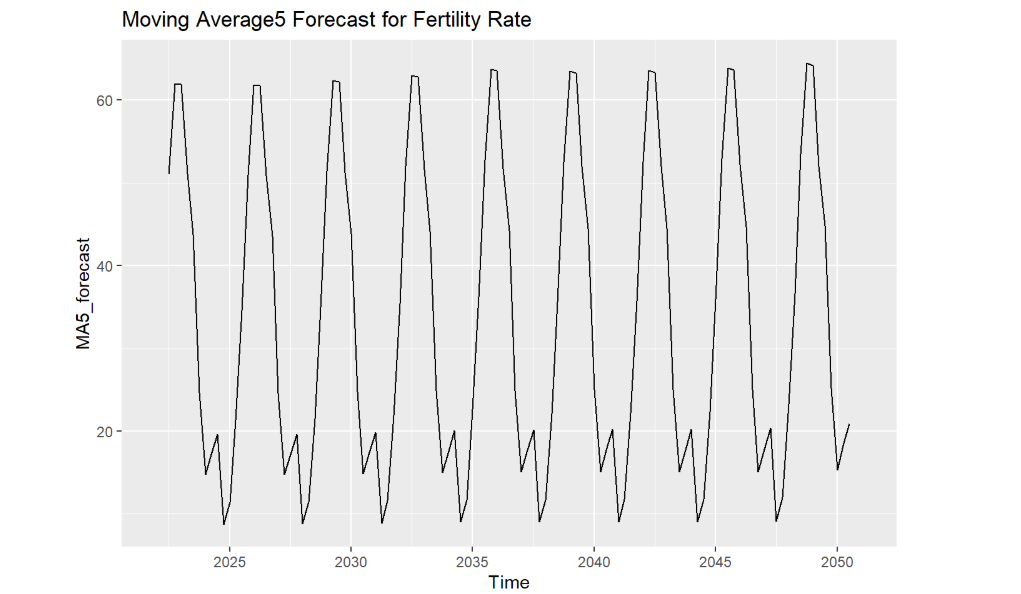
```{r}

**#Moving Average forecast**

MA5\_forecast <- ma(ts\_data,order=5)

# Plot the forecast

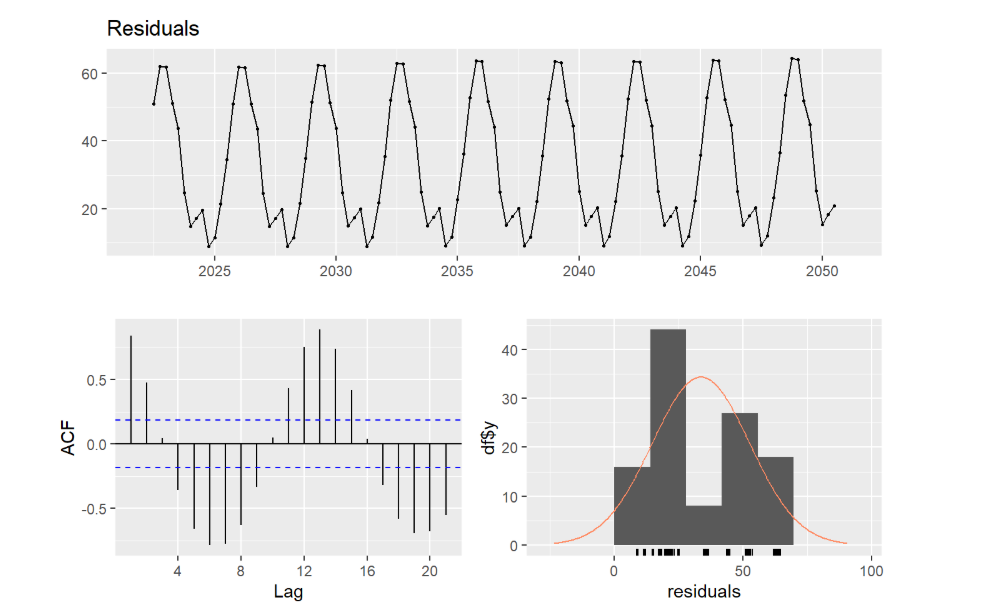
autoplot(MA5\_forecast) + ggtitle("Moving Average5 Forecast for Fertility Rate")



forecast(MA5\_forecast)

# Residual Analysis

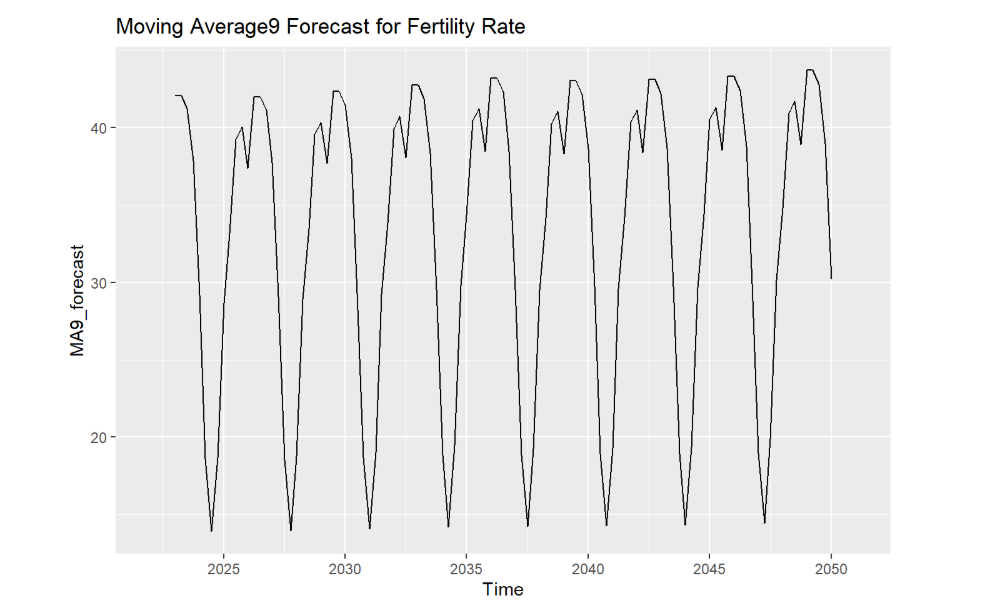
checkresiduals(MA5\_forecast)



MA9\_forecast <- ma(ts\_data,order=9)

# Plot the forecast

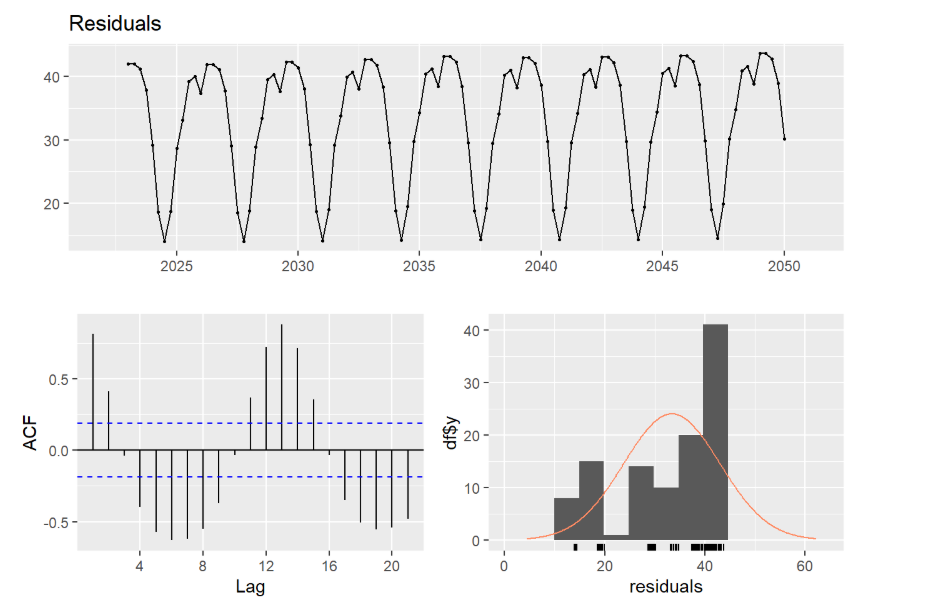
autoplot(MA9\_forecast) + ggtitle("Moving Average9 Forecast for Fertility Rate")



# Residual Analysis

forecast(MA9\_forecast)

checkresiduals(MA9\_forecast)



```

Good for smoothing and reducing noise. Does not effectively capture trend or seasonality. Significant residual lags suggest it may not fit dynamic time series with trends.

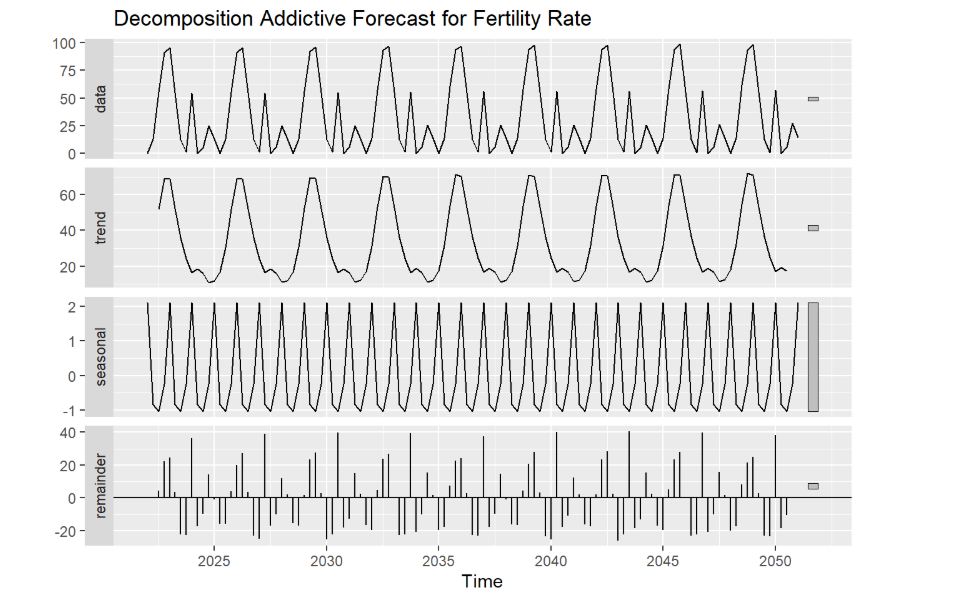
```{r}

**##Decomposition**

decomp <- decompose(ts\_data, type = "additive")

# Plot the Decomposition

autoplot(decomp) + ggtitle("Decomposition Addictive Forecast for Fertility Rate")

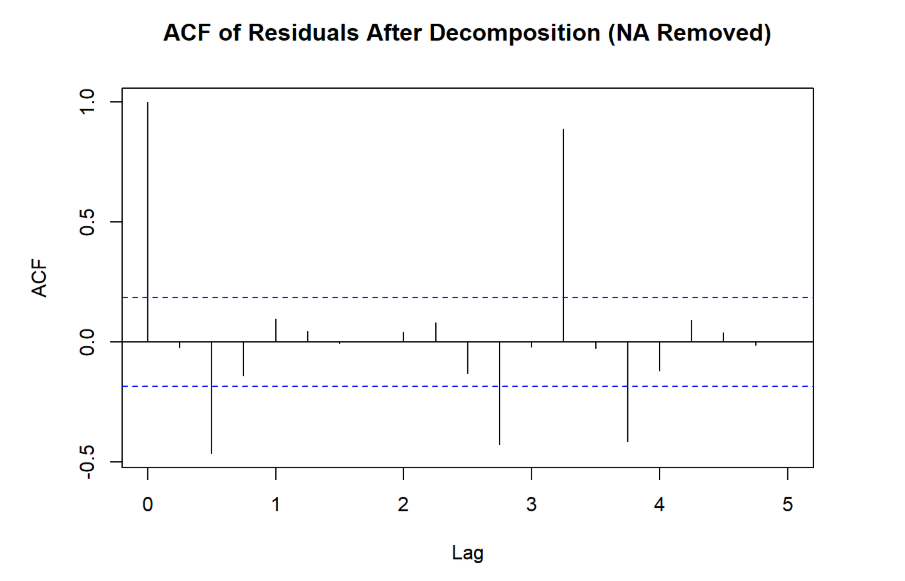


residuals <- decomp$random

residuals\_clean <- na.omit(residuals)

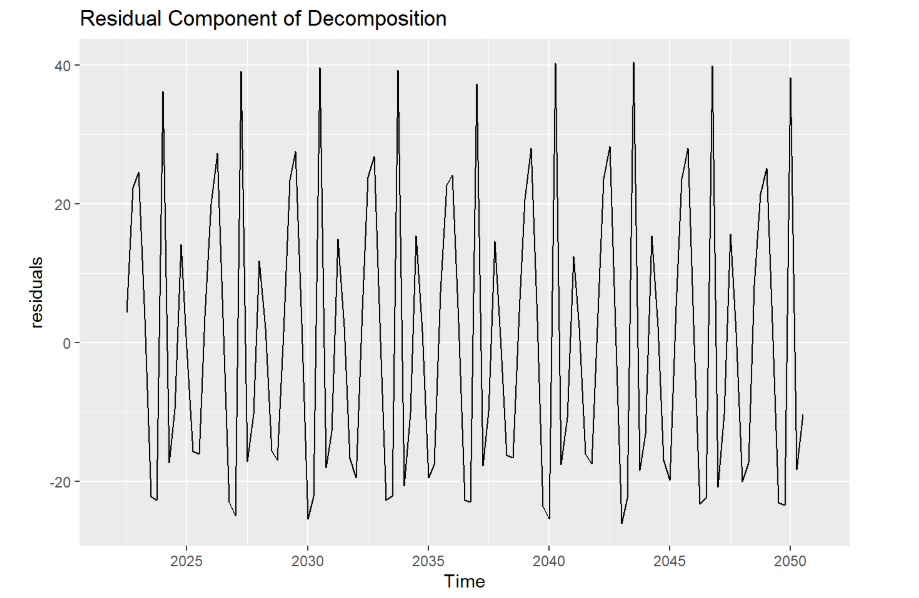
# Plot the ACF of the cleaned residuals

acf(residuals\_clean, main = "ACF of Residuals After Decomposition (NA Removed)")



# Plot the residuals

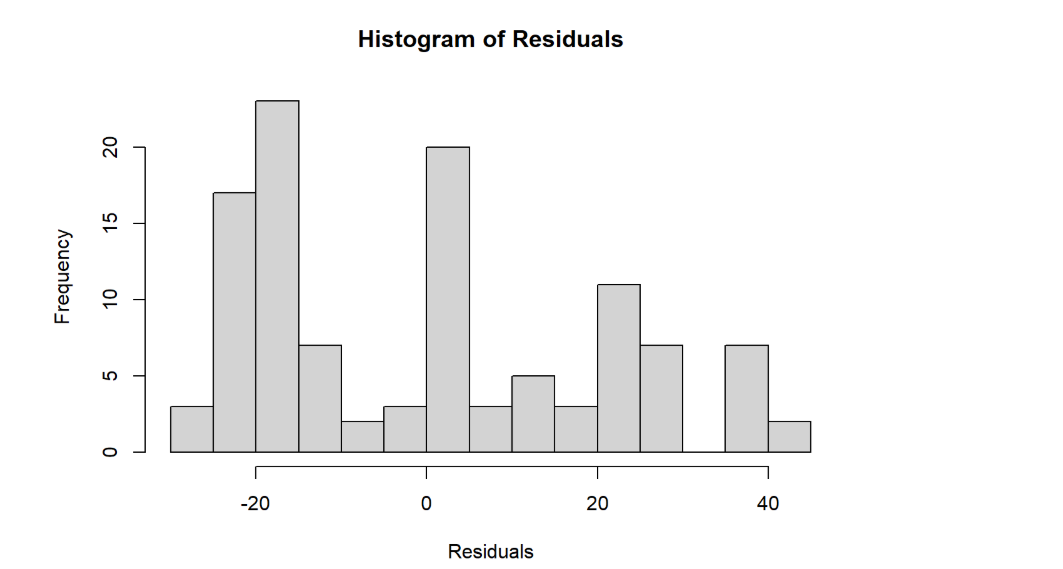
autoplot(residuals, main = "Residuals After Decomposition") + ggtitle("Residual Component of Decomposition")



forecast(residuals\_clean)

# Histogram of residuals

hist(residuals, main = "Histogram of Residuals", xlab = "Residuals", breaks = 15)



# Ljung-Box test for residuals

Box.test(residuals, lag = 10, type = "Ljung-Box")

```

Separates data into trend, seasonal, and residual components. Residual patterns indicate how well the model captures these components; randomness means effective decomposition.

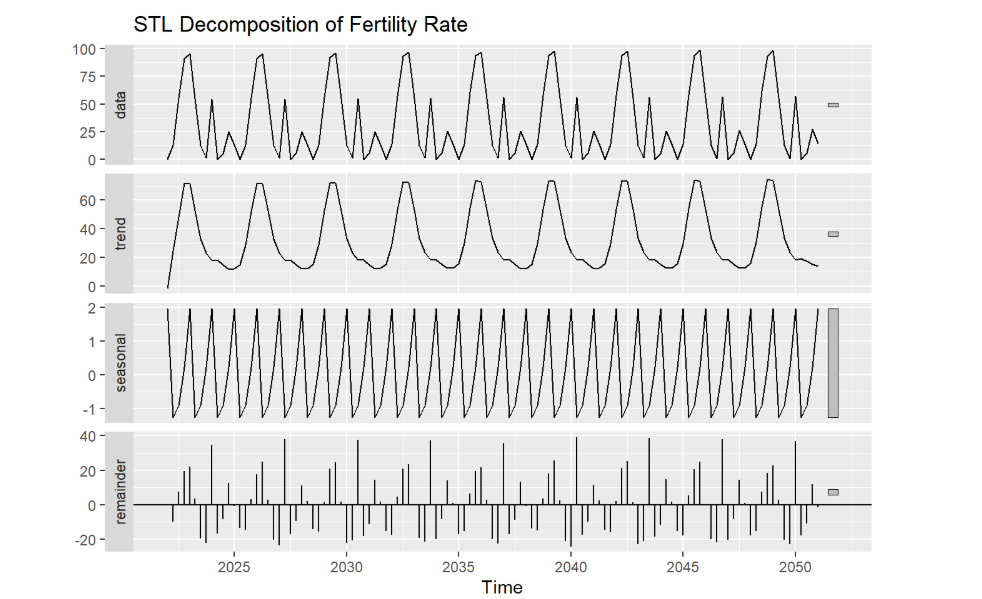
```{r}

# **STL Decomposion**

stl\_decomposition <- stl(ts\_data, s.window = "periodic")

# Plot the decomposition

autoplot(stl\_decomposition) + ggtitle("STL Decomposition of Fertility Rate")

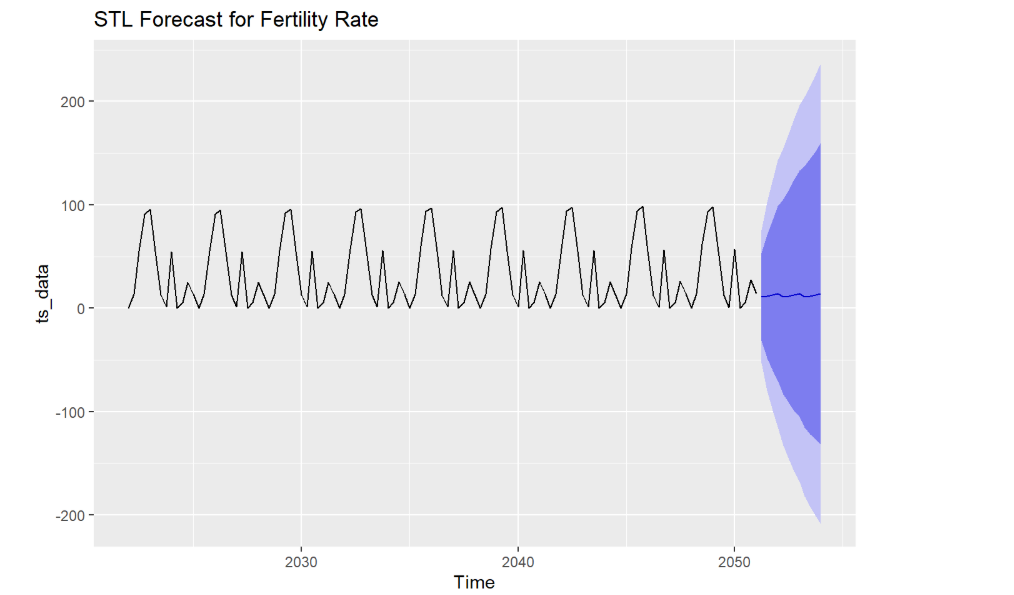


# Forecast the next 12 months

stl\_forecast <- forecast(stl\_decomposition, h = 12)

# Plot the forecast

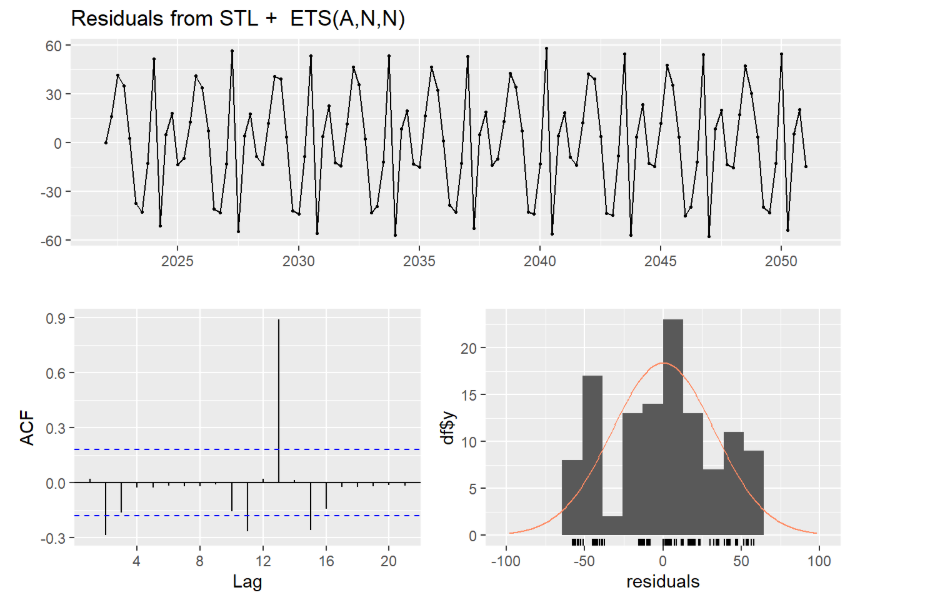
autoplot(stl\_forecast) + ggtitle("STL Forecast for Fertility Rate")



forecast(stl\_forecast)

# Residual Analysis

checkresiduals(stl\_forecast)



```

Handles non-linear seasonality and trend effectively. Residuals should be random if both trend and seasonal components are adequately captured.

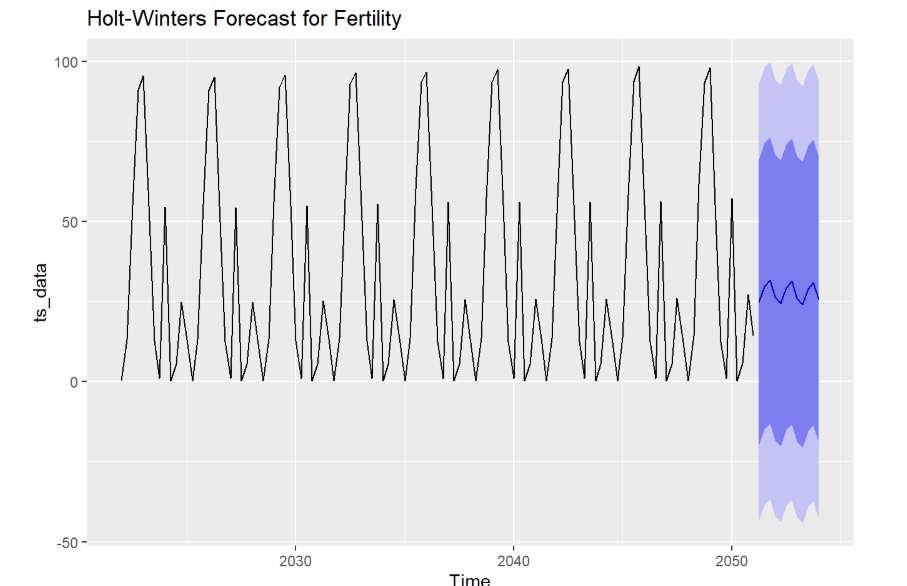
```{r}

# **Holt-Winters Seasonal Model**

hw\_model <- hw(ts\_data, seasonal = "additive", h = 12)

# Plot the forecast

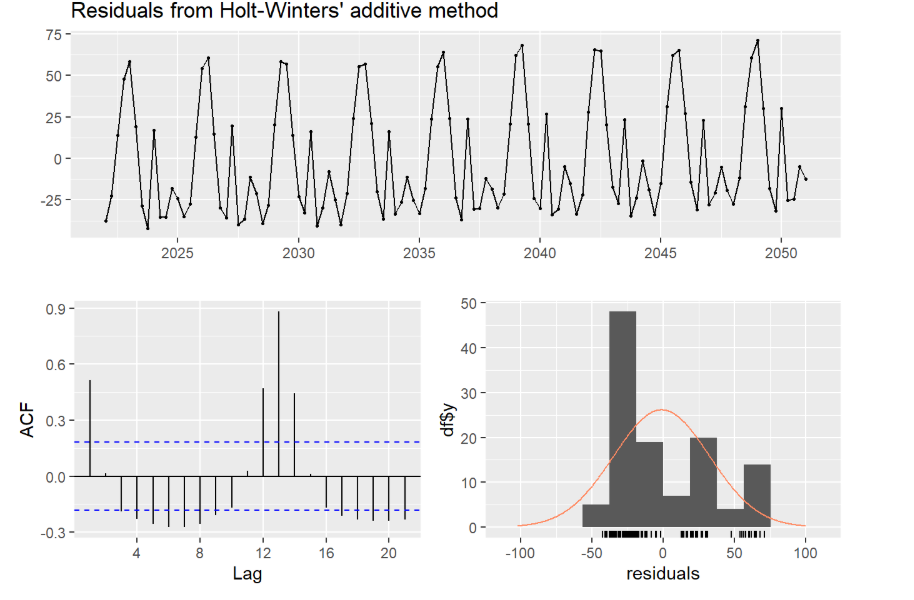
autoplot(hw\_model) + ggtitle("Holt-Winters Forecast for Fertility")



forecast(stl\_forecast)

# Residual Analysis

checkresiduals(hw\_model)



```

Captures both trend and seasonality, making it suitable for data with both components. Random residuals indicate a successful capture of underlying dynamics.

```{r}

# **Fit an ARIMA model**

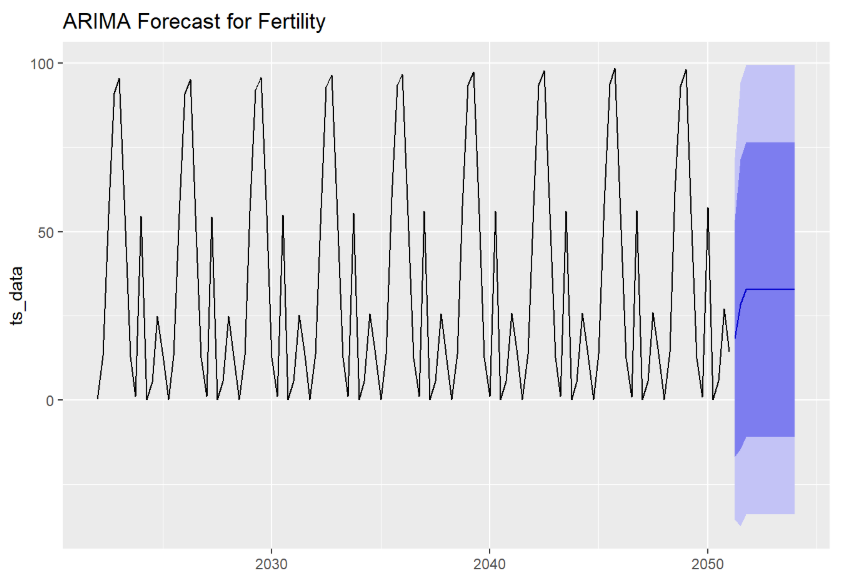
arima\_model <- auto.arima(ts\_data)

# Forecast the next 12 months

arima\_forecast <- forecast(arima\_model, h = 12)

# Plot the forecast

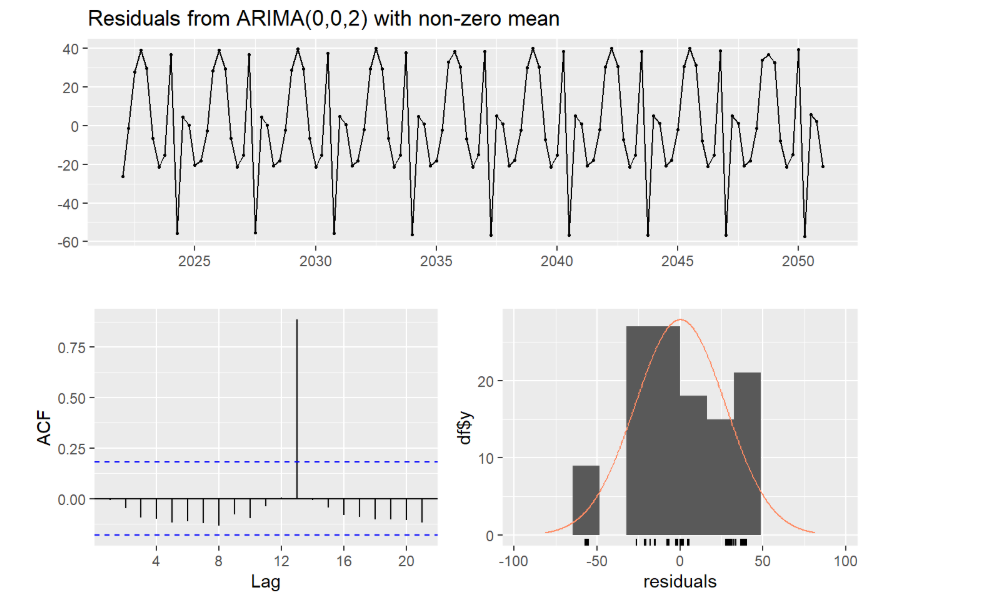
autoplot(arima\_forecast) + ggtitle("ARIMA Forecast for Fertility")



forecast(arima\_forecast)

# Residual Analysis

checkresiduals(arima\_model)

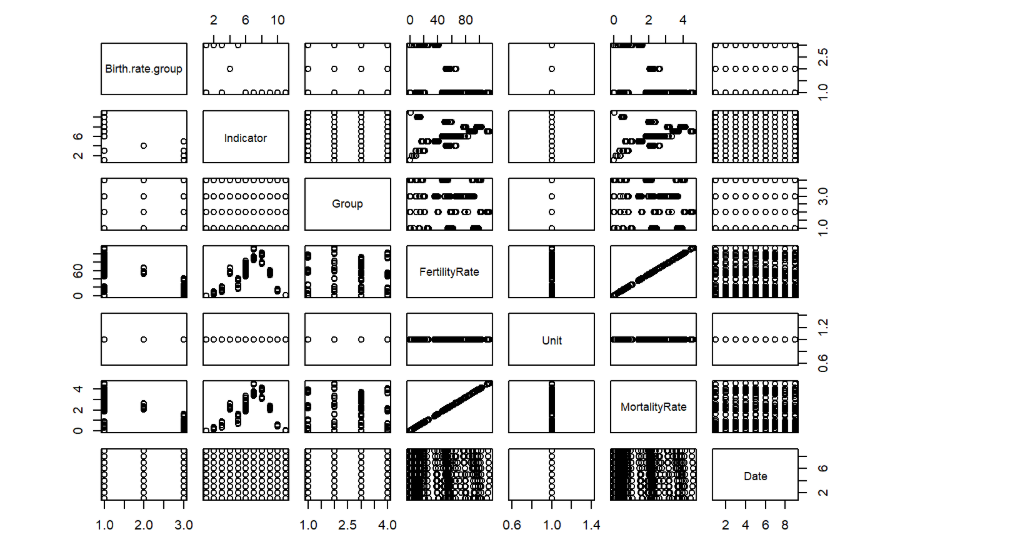
 ```

Suitable for data with auto-regressive components, differencing to make data stationary, and moving average terms. Residuals should resemble white noise without autocorrelation, indicating that the ARIMA model has properly fit the data.

```{r}

**#Regression Analysis**

plot(data[,-1])



# Remove non-numeric columns, if necessary

# Select only numeric columns for the pairs plot

numeric\_data <- data[sapply(data, is.numeric)]

str(numeric\_data)

# Plot pairs only if there are at least two numeric columns

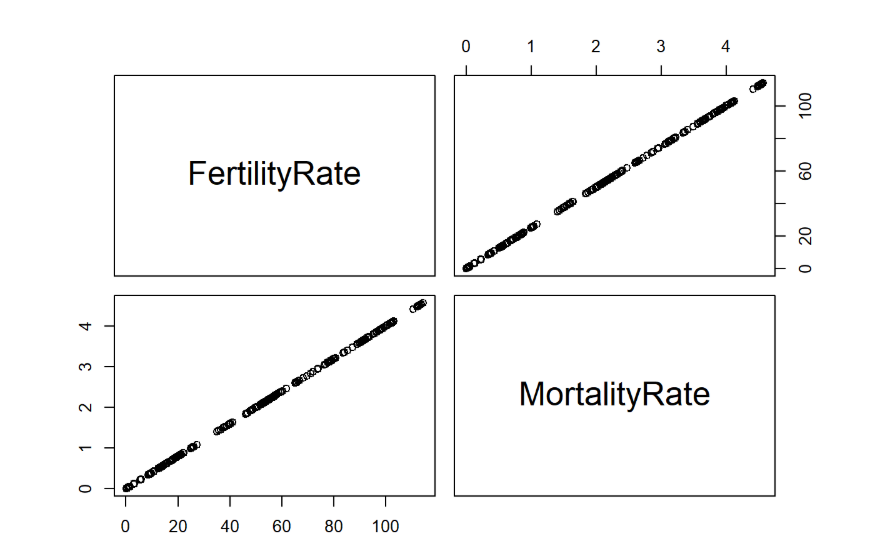
if (ncol(numeric\_data) > 1) {

pairs(numeric\_data)

} else {

print("Not enough numeric columns for pairwise plot")

}



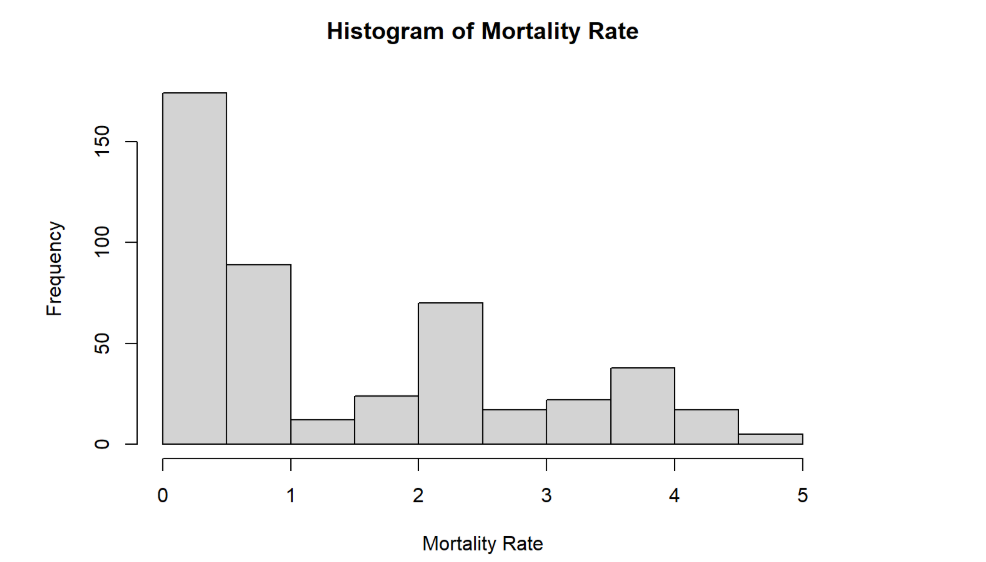
# Summary and correlation

summary(data)

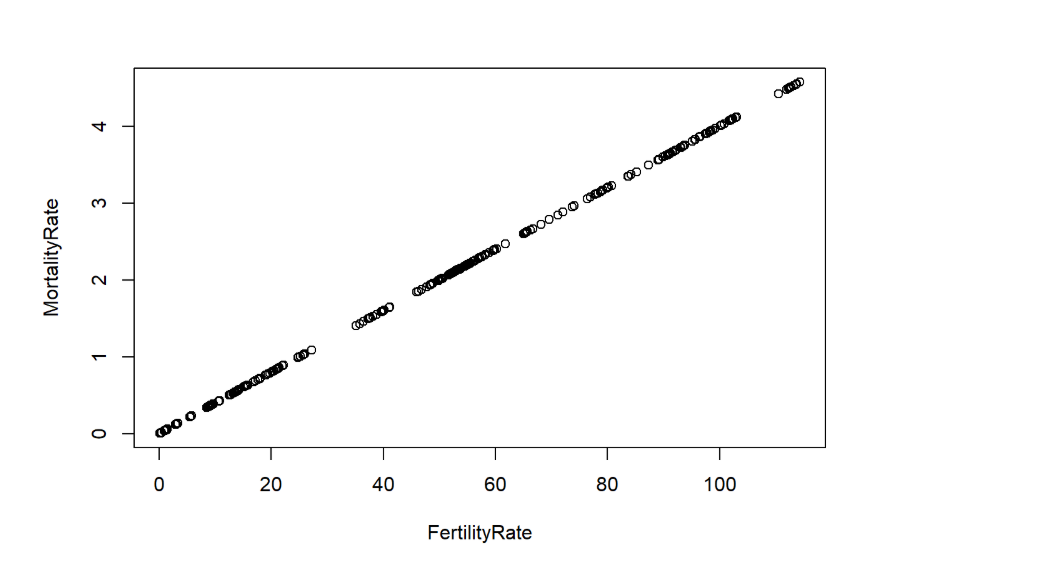
cor(numeric\_data)

# Histograms and Scatter Plots

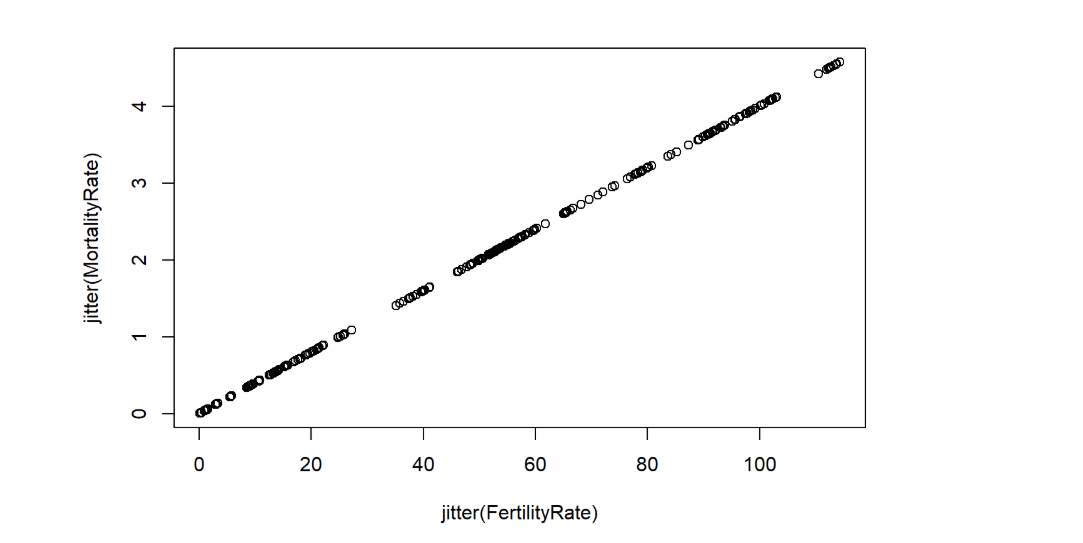
hist(data$MortalityRate, main="Histogram of Mortality Rate", xlab="Mortality Rate")



plot(MortalityRate ~ FertilityRate, data=data)



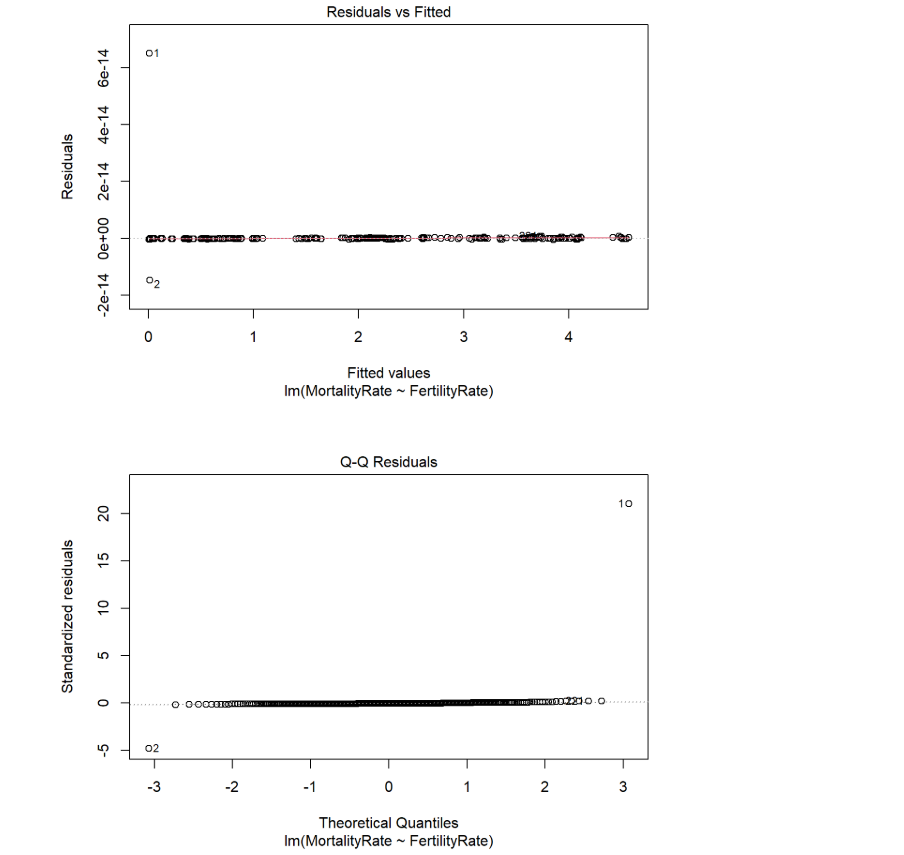
plot(jitter(MortalityRate) ~ jitter(FertilityRate), data=data)

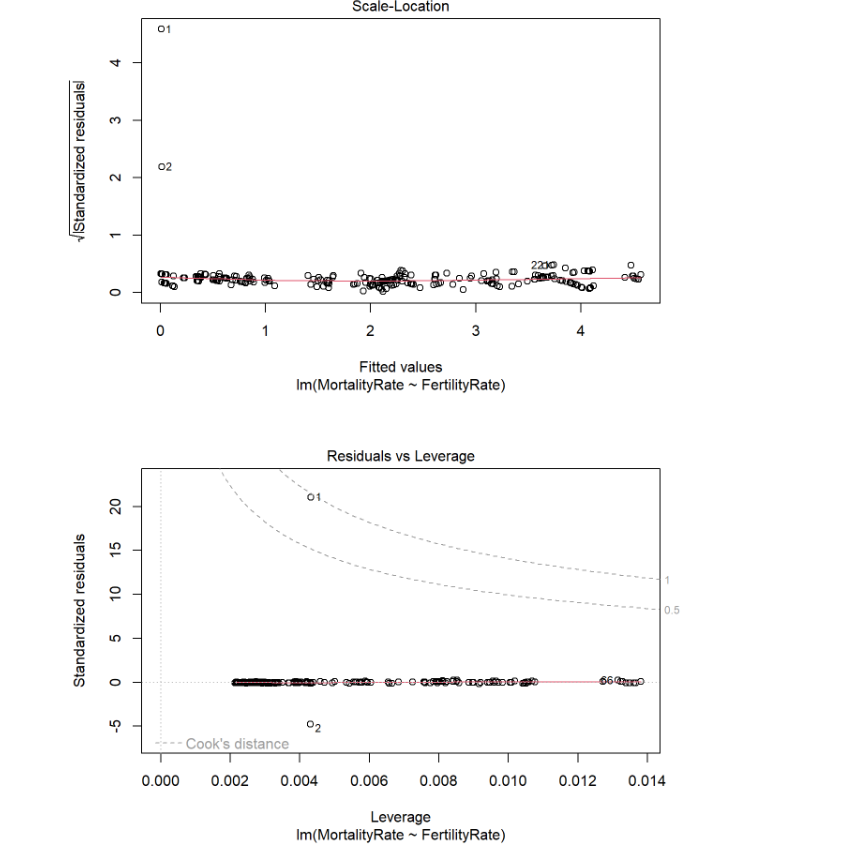


# Fit linear regression model

fit <- lm(MortalityRate ~ FertilityRate, data=data)

print(fit)





summary(fit)

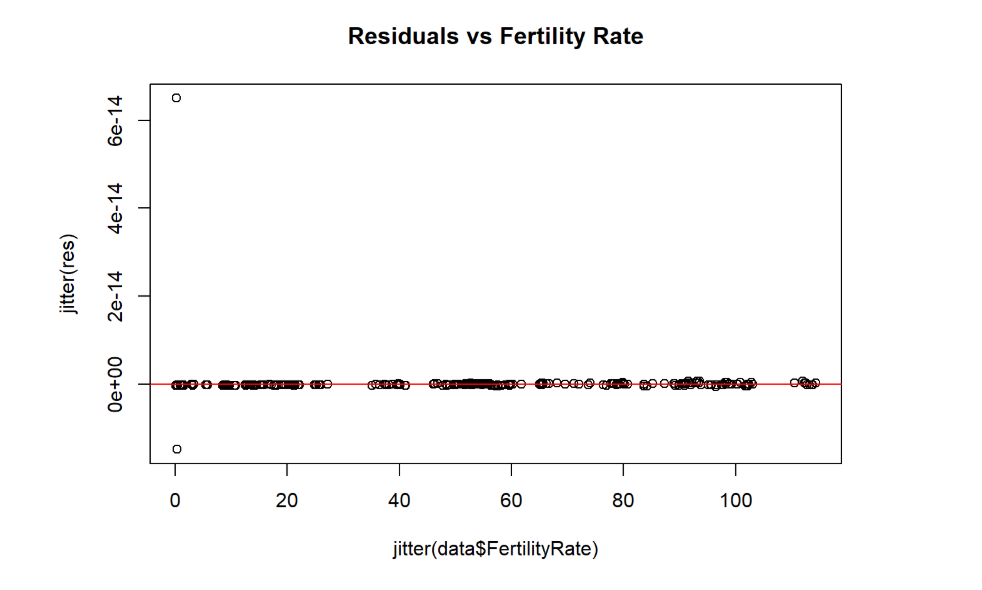
anova(fit)

# Residual analysis

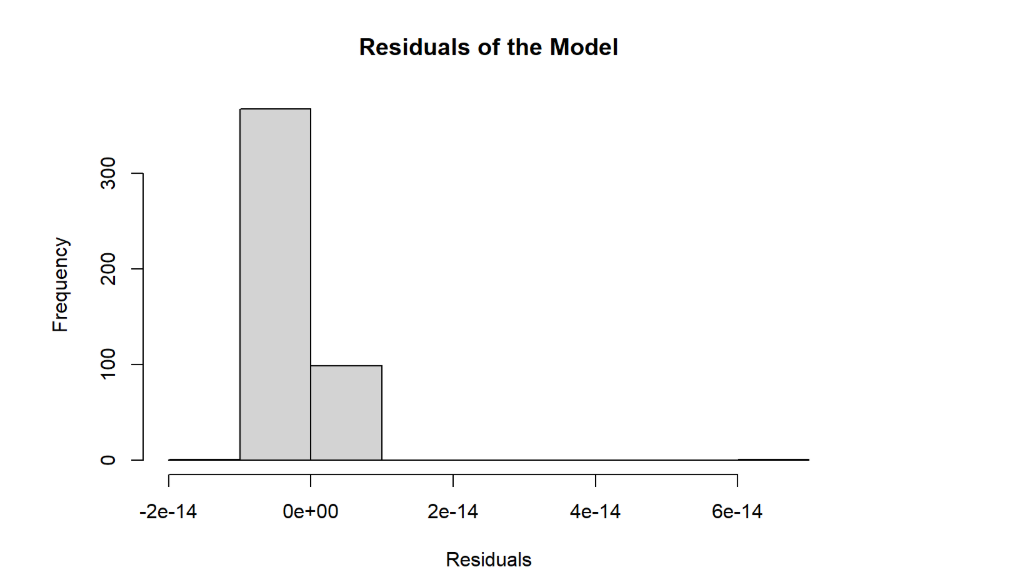
res <- residuals(fit)

plot(jitter(res) ~ jitter(data$FertilityRate), main="Residuals vs Fertility Rate")

abline(h=0, col="red")



hist(res, main = "Residuals of the Model", xlab = "Residuals")



# Predicted value for the first data point

fitted(fit)[1]

```

To explain insights at the residuals, we will see differences between actual mortality rates and what our model predicted.

Residuals vs. Fitted Values Plot: Here, we’re looking for a random scatter of points around zero. If we see that randomness, it means our model is appropriate. But if we spot patterns, it might mean our model is missing something important.

Histogram of Residuals: This histogram shows the spread of prediction errors. Ideally, it looks like a bell curve centered at zero, which would indicate that most of our predictions are close to the actual values. If it’s skewed or off-center, it could mean our model is biased or not capturing some factors correctly.

Q-Q Plot: This plot checks if the residuals follow a normal distribution. If the points align closely with a straight line, it confirms that our errors are distributed as expected, reinforcing the model’s reliability. If not, it might hint at some model issues or the need for additional predictors.

Summary for Regression:

If the residuals look random and normally distributed, it shows that fertility rate is a good predictor for mortality rate in our data.

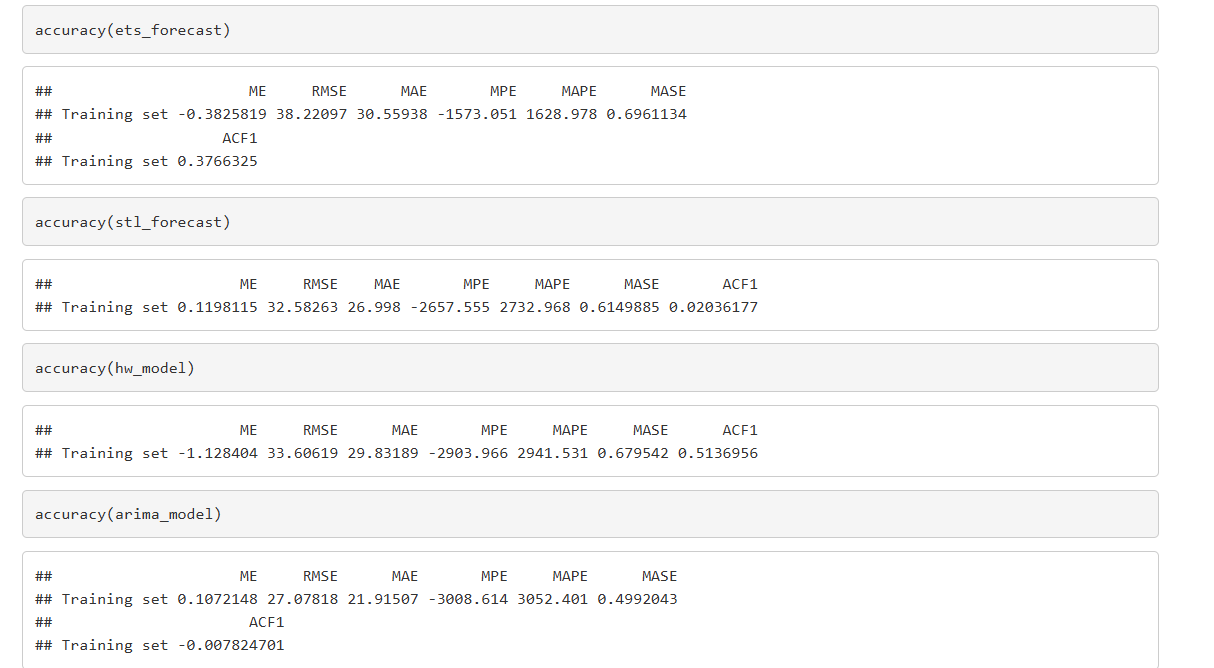
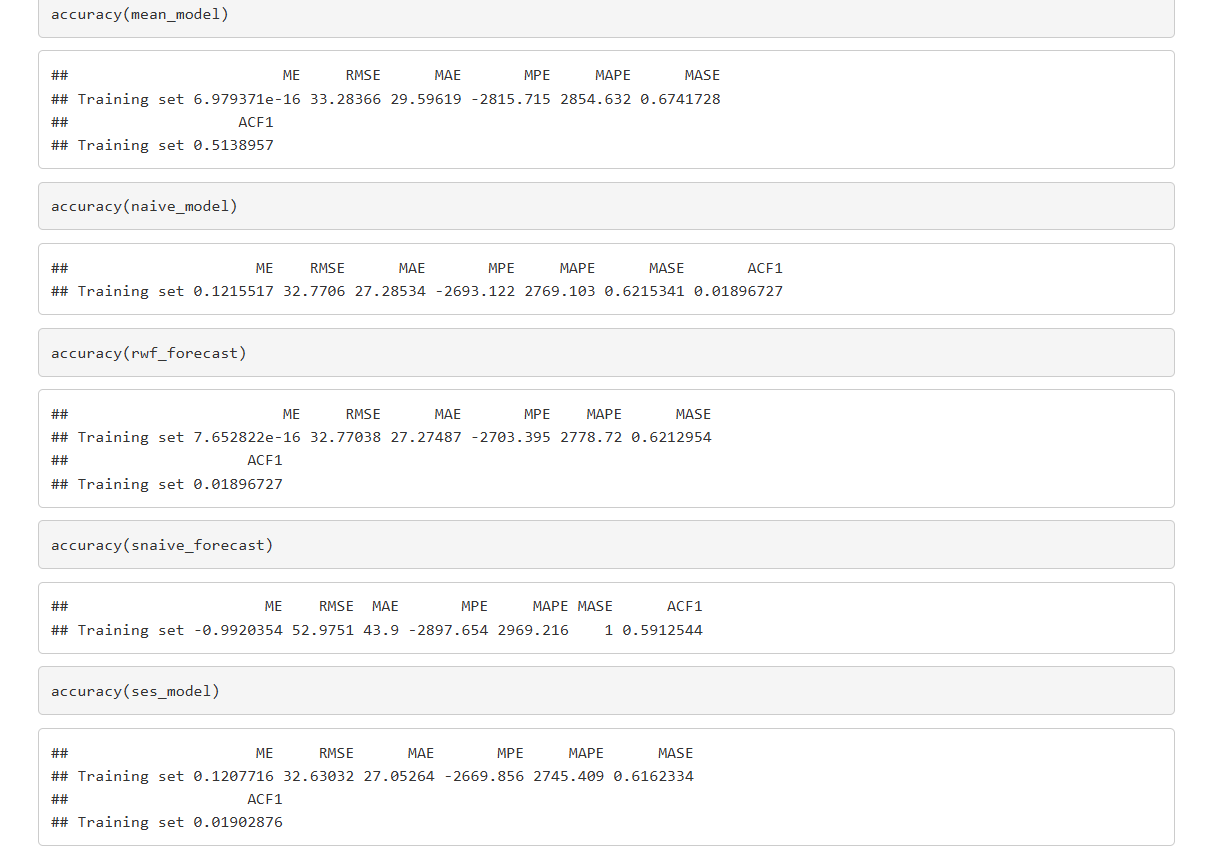
If we see patterns in residuals or deviations from normality, it might be worth revisiting the model—possibly by adding more factors, transforming some data, or even trying a different type of model to capture the relationship more accurately.

1. **Prediction and Accuracy summary from different forecasting methods**

```{r}

# Accuracy checked

# Pick an accuracy measure, compare your models, and state the best model based on the accuracy comparison

```

Accuracy of Prediction:

Provisional prediction is subject to some non-random sampling error. The quarterly provisional estimates are based on data that are potentially more incomplete for the most recent months. No imputations of births have been performed, because it is assumed that the data are missing at random (i.e., the degree of missing data is unrelated to estimates of reproductive health). Estimates of completion rates by month were all above 95% for the United States. However, certain states may have more delayed reporting, and it is unknown whether indicators of reproductive health may be different for these states compared with states having complete reporting. Even if no differential delay occurred, some sampling error would still exist for rate estimates, because they are based on incomplete data. A guideline for the size of this sampling error is given by deriving the variation that would occur if the data were missing at random; standard errors for birth rates have been calculated according to these methods, accounting for sampling error.

Partly because of the factors discussed above, provisional estimates of birth rates are rarely higher than the true rate. Historically, provisional estimates of birth rates track closely with estimates based on final data. Based on simulations of various levels of data completeness, ranging from 50% through 90%, estimates of the indicators included in this release can be expected to be within 1-2% of the estimates based on complete data. Exceptions were noted for birth rates among women aged 10-14 and 45 and over under scenarios where data completeness was 80% or lower. For these two age groups, the percent differences between estimates based on complete data and those based on various levels of incomplete data were larger than 1% because the birth rates are lows (0.2 and 0.9 per 1,000, respectively). However, the absolute differences between the rates across various levels of data completeness was approximately 0.1 for each age group. For example, estimated birth rates for females aged 10–14 under various levels of completeness ranged from 0.2 through 0.3 per 1,000, while those for women aged 45 and over ranged from 0.8 through 0.9 per 1,000. Notably, these simulations assumed data were missing at random.

Because the timeliness of birth reporting has been improving, accuracy of the estimates may change over time. As a result, estimates for previous quarters may change with the addition of updated data. Estimates may differ from previously published preliminary data due to rounding or the use of updated population estimates.

* 1. **State your decision based on the analysis**

After comparing the different forecasting models based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE), the best model to consider is:

Arima Model:

It has the lowest RMSE (27.07818) and lowest MAE (21.91507), indicating minimal errors and high accuracy.

The ACF1 value (-0.0078) is close to zero, suggesting that the residuals are random and not autocorrelated.

* 1. **Provide some ideas to improve your forecasts**

1. **Incorporate Exogenous Variables**:
   * Include relevant external factors (e.g., economic indicators, weather patterns) that might influence fertility or mortality rates. This can improve the model's ability to predict accurately.
2. **Regular Model Updates**:
   * Retrain the forecasting models with the latest data to ensure they capture the most recent trends and seasonal patterns. This helps the model stay current and responsive to changes.
3. **Ensemble Methods**:
   * Combine multiple forecasting models to create an **ensemble**. This can reduce the risk of overfitting to one model and improve generalization by averaging out individual errors.
4. **Feature Engineering**:
   * Create additional features such as **lags** or **rolling averages** that can help the model understand past behaviours in more detail. This can be particularly useful for time series with complex dynamics.
5. **Data Transformation**:
   * Apply transformations to the data, such as **log** or **Box-Cox**, to stabilize variance and make the series more stationary. This can lead to better model performance and improved forecasts.

These ideas can help to further improve the accuracy and reliability of your forecasts, leading to better decision-making and resource allocation in practice.