# Time Series Analysis - Final Assignment

Ratnak Saha s3973869

Last updated: 16 June, 2024

# Rainfall Forecasts at Caufield (Racecourse)

# Research Question:

1. Which ARIMA and SARIMA models can best forecast rainfall at Caulfield (Racecourse)?
2. In what ways can advanced time series models, such as ARIMA and SARIMA, improve long-term monthly rainfall forecasts at Caulfield (Racecourse) and what are the implications for future rainfall trends and patterns?
3. What are the potential future trends and patterns in rainfall at Caulfield (Racecourse) based on these models?

# Introduction

The analysis of the monthly rainfall for the Caulfield (Racecourse) dataset is sourced from Bureau of Meteorology, from 1950 to 2012 is demonstrated in the final report. The primary aim of this analysis is to understand the underlying pattern, trends and seasonal variations in the rainfall data, as well as to develop a forecasting model to predict future rainfall. Subsequently, various statistical techniques and models, including ARIMA and SARIMA were applied to model the rainfall data. To specify potential ARIMA models, ACF, PACF, EACF and BIC tables were also used to compare different models and select the most suitable models.

# Importing the data

```
setwd("F:/3rd Semester/Time Series Analysis/Assignment_3")

rainfall_data <- read.csv('Rainfall_data_reduced_data.csv')
head(rainfall_data)
```

```
##   Year  Jan   Feb  Mar   Apr   May   Jun   Jul  Aug   Sep   Oct   Nov   Dec
## 1 1950  8.0 102.6 70.6  45.2  53.5  32.1  54.4 40.1  81.2  94.4  89.2  75.2
## 2 1951 39.4 123.1  5.4 126.6 115.1  77.9  92.4 87.9  60.8  85.4  42.3  55.1
## 3 1952 23.7  65.1 33.1  71.5  85.9 118.6 171.5 80.4  68.1 114.2 127.7 110.3
## 4 1953 75.5  68.9 17.1  44.4  75.6  51.7  45.7 78.7 110.5 137.4  87.8  30.3
## 5 1954 42.5  35.4 15.1  72.9  59.7  67.9  41.7 59.0  54.7  94.9 191.1 164.7
## 6 1955 23.2  93.2 85.8  39.0 115.1  87.3  54.9 95.4  90.1 108.7  53.8 117.5
```

# Data Manipulation

```
rainfall_data_transformed <- rainfall_data %>% pivot_longer(names_to ="Month", values_to ="Rainf
all",cols = 2:13)
head(rainfall_data_transformed)
```

```
## # A tibble: 6 × 3
##    Year Month Rainfall
##   <int> <chr>    <dbl>
## 1  1950 Jan        8
## 2  1950 Feb      103.
## 3  1950 Mar       70.6
## 4  1950 Apr       45.2
## 5  1950 May       53.5
## 6  1950 Jun       32.1
```

# Checking for missing values

```
sum(is.na(rainfall_data_transformed))
```

```
## [1] 4
```

# Transforming the data into time series

```
rainfall_data_ts <- ts(as.vector(rainfall_data_transformed$Rainfall), start=c(1950,1), frequency
= 12)
```

# Filling the missing values with average monthly

# rainfall data

```
rainfall_df <- data.frame(
  Year = rep(start(rainfall_data_ts)[1]:end(rainfall_data_ts)[1], each =12),
  Month = rep(1:12, time = length(start(rainfall_data_ts)[1]:end(rainfall_data_ts)[1])),
  Rainfall = as.vector(rainfall_data_ts))

monthly_averages <- rainfall_df %>%
  group_by(Month) %>%
  summarise(Monthly_average = mean(Rainfall, na.rm = TRUE))

rainfall_df <- rainfall_df %>%
  left_join(monthly_averages,by= "Month") %>%
  mutate(Rainfall = ifelse(is.na(Rainfall), Monthly_average, Rainfall))


rainfall_data_ts_filled <- ts(rainfall_df$Rainfall,
                              start = start(rainfall_data_ts),
                              frequency = 12)
```

# Checking if there are still any missing values

```
sum(is.na(rainfall_data_ts_filled))
```

```
## [1] 0
```

# Ploting the time series data

```
plot(rainfall_data_ts_filled, type= "o",main = "Monthly Rainfall for Caulfield (Racecourse) from 1950 to 2012", ylab = "Rainfall (mm)", xlab = "Year")
```

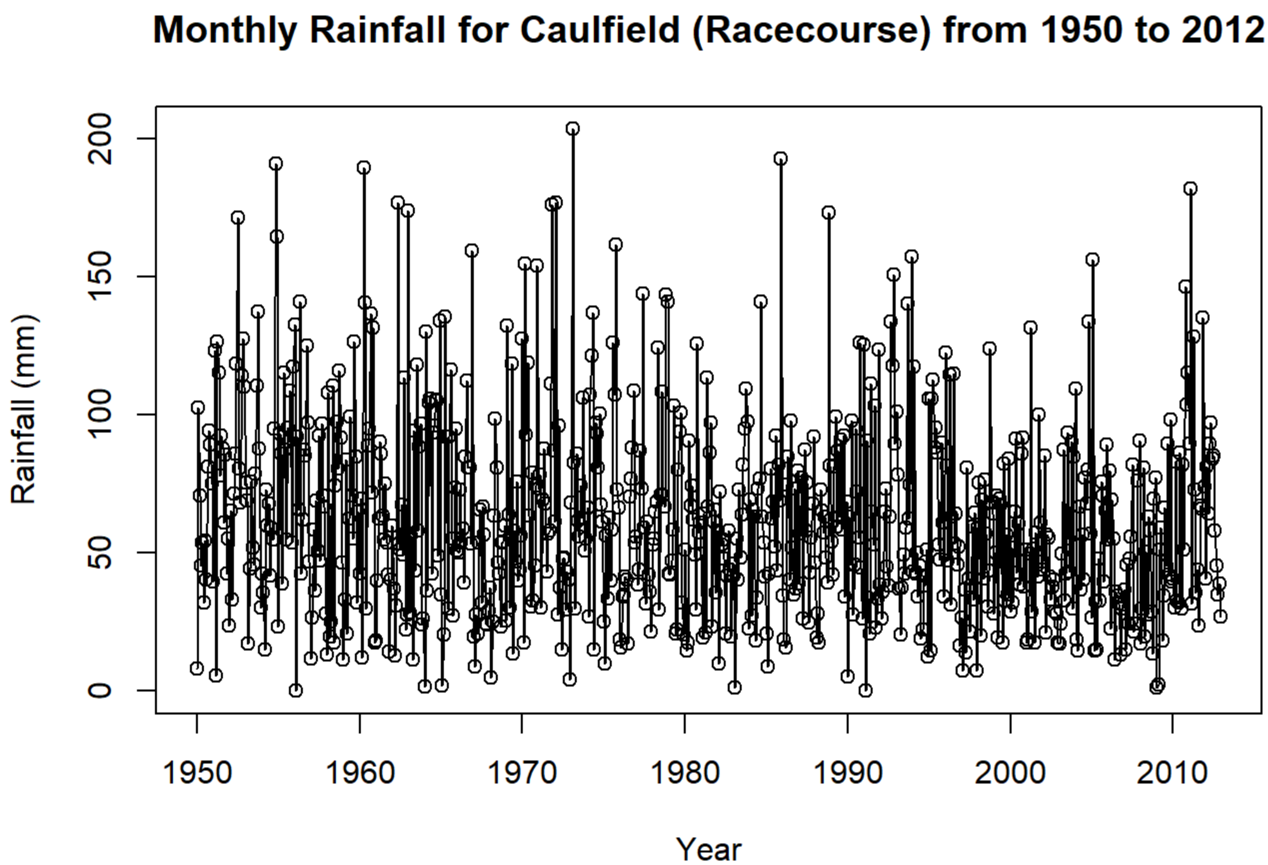


Figure 1: Time Series Plot Of Monthly Rainfall for Caulfield (Racecourse)

1. Change Variance: Upon examining the time series data, there are indications of changing variance across different months over the years. This variability suggests that the amount of rainfall fluctuates not only seasonally but also within the same months in different years.

```
plot(rainfall_data_ts_filled, type= "o",main = "Monthly Rainfall for Caulfield (Racecourse) from 1950 to 2012", ylab = "Rainfall (mm)", xlab = "Year")
points(y=rainfall_data_ts_filled,x=time(rainfall_data_ts_filled), pch=as.vector(season(rainfall_data_ts_filled)),col=rainbow(12))
```

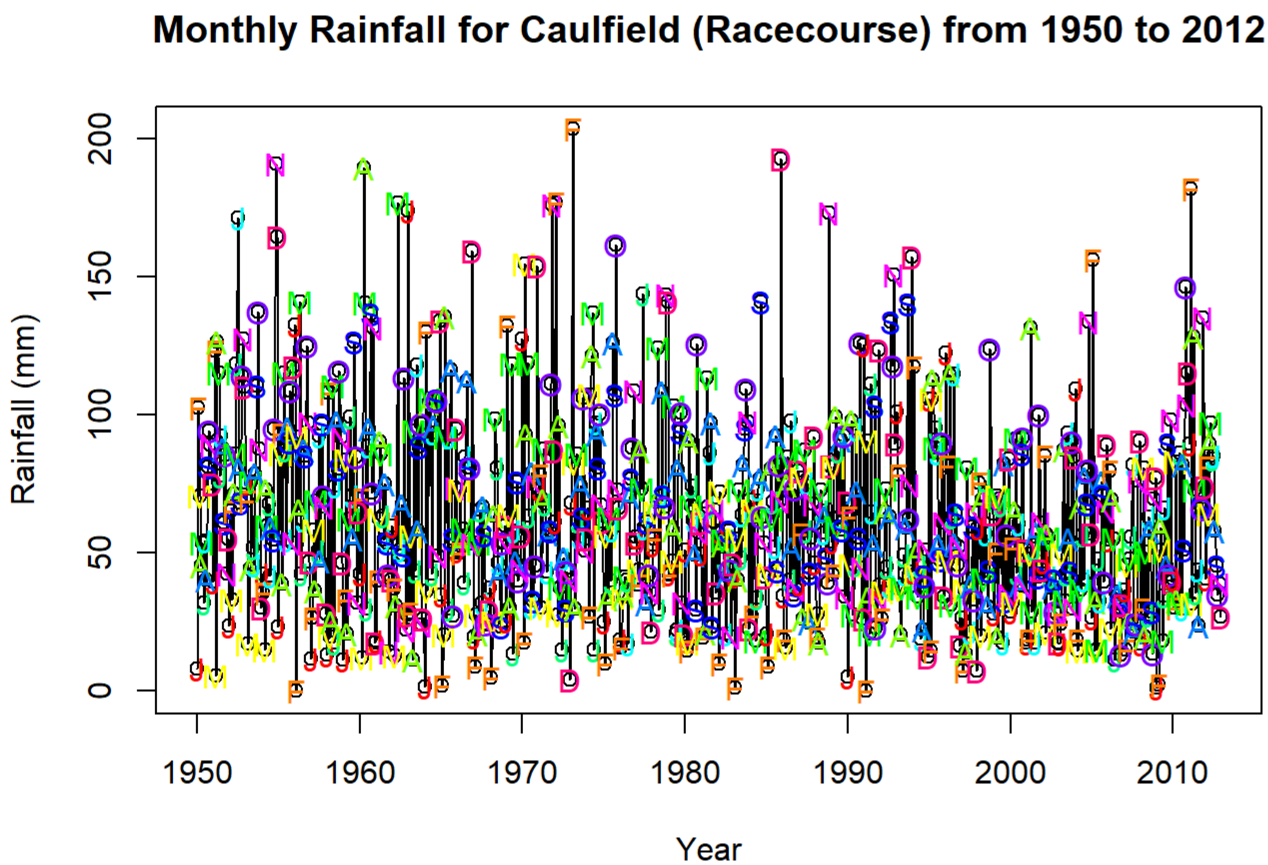


Figure 2: Time Series Plot Of Monthly Rainfall for Caulfield (Racecourse)

2. Seasonality: The time series plot reveals clear signs of seasonality in the rainfall data for the Caulfield (Racecourse), specifically for the month of September, October, November and December consistently show higher rainfall compared to other months across different years. This seasonal pattern indicates that the region experiences significant rainfall during these late year months which is crucial factor for understanding and predicting rainfall trends.

# Calculating and plotting the yearly average rainfall

```
yearly_mean_rainfall <- rainfall_df %>% group_by(Year) %>% summarise(mean_rainfall = mean(Rainfa
ll))


yearly_mean_rainfall_ts <- ts(yearly_mean_rainfall$mean_rainfall, start =c(1950))
plot(yearly_mean_rainfall_ts,xlab = "Year", ylab = "Yealrly Avearage Rainfall", main = "Yearly A
verage Rainfall Caulfield (Racecourse)")
```

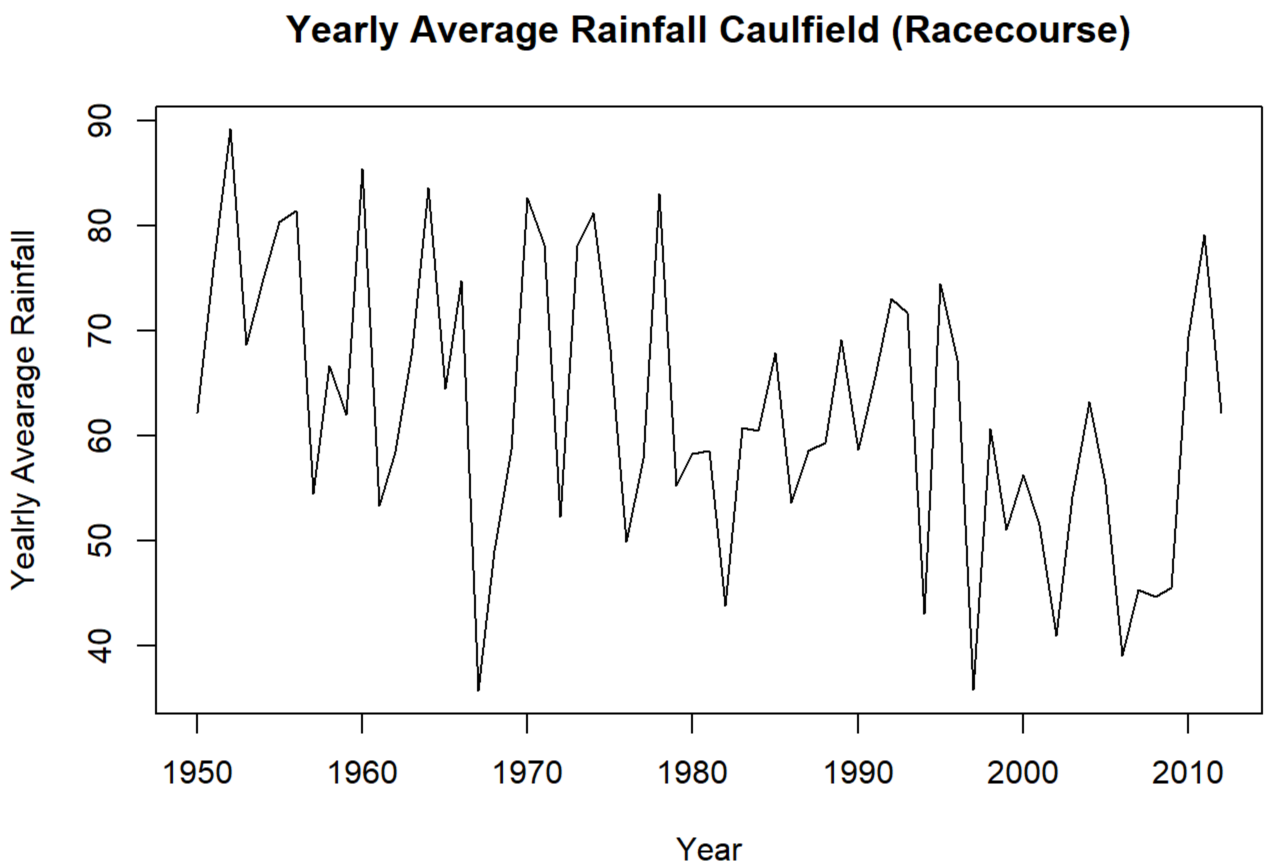


Figure 3: Time Series Plot Of Yearly Average Rainfall for Caulfield (Racecourse)

3. Trend: The time series plot Of Yearly Average Rainfall indicates a clear slight downward trend data over the period from 1950 to 2012. The long-term decline suggests that the region is experiencing a gradual decrease in monthly rainfall levels.

4. Behavior: The time series plot shows that there is no significant autoregressive behavior in the monthly rainfall data for Caulfield(Racecourse) from 1950 to 2021. This means that current month's rainfall does not strongly depend on the rainfall of previous month. The absence of autoregressive behavior suggests that simple autoregressive model might not be the best fit for this dataset. Instead, models that focus on capturing seasonality and other patterns, without relying on the direct influence of previous months' rainfall, might provide better forecasting accuracy.

5. Change point: The time series plot indicates that there is no significant change point in the monthly rainfall data from 1950 to 2012.

```
par(mfrow=c(1,1))
plot(y = rainfall_data_ts_filled,x = zlag(rainfall_data_ts_filled), ylab="Change in Rainfall for
Caulfield (Racecourse) ",
     xlab = "Years", main = "Scatter plot Of Rainfall for Caulfield (Racecourse)")
```

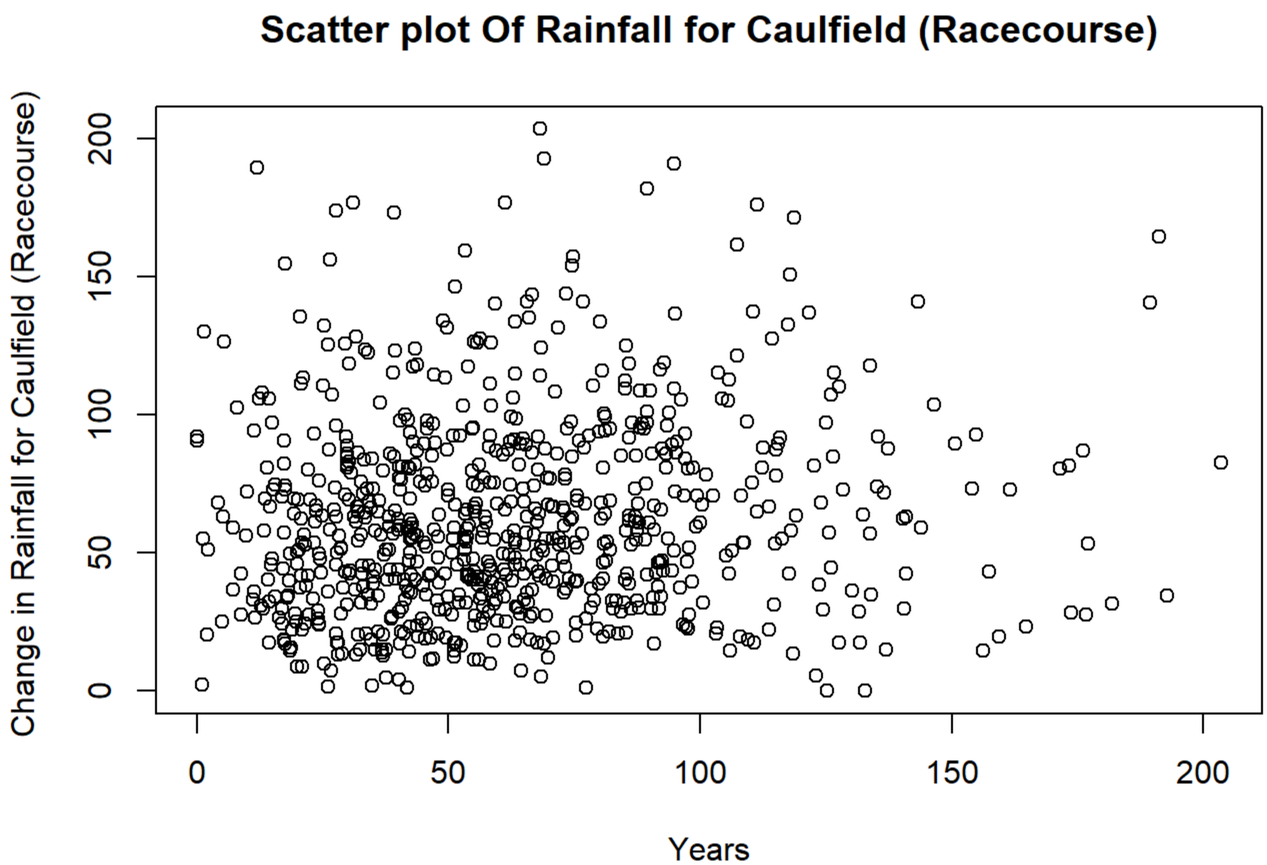


Figure 4: Correlation Plot Of Monthly Rainfall for Caulfield (Racecourse)

```
y = rainfall_data_ts_filled
x = zlag(rainfall_data_ts_filled)
index = 2:length(x)
cor(y[index],x[index])
```

```
## [1] 0.1057818
```

* The correlation coefficient between the rainfall data and its lagged value is approximately 0.1058. This indicates a weak positive correlation between the current month's rainfall and the rainfall of the previous month. The presence of weak positive correlation aligns with the absence of significant autoregressive behavior observed in the analysis.

```
par(mfrow=c(1,2))
acf(rainfall_data_ts_filled, lag.max = 48, main = "ACF - Monthly Rainfall for Caulfield", cex.ma
in =1.0)
pacf(rainfall_data_ts_filled, lag.max = 48, main = "PACF - Monthly Rainfall for Caulfield", cex.
main =1.)
```

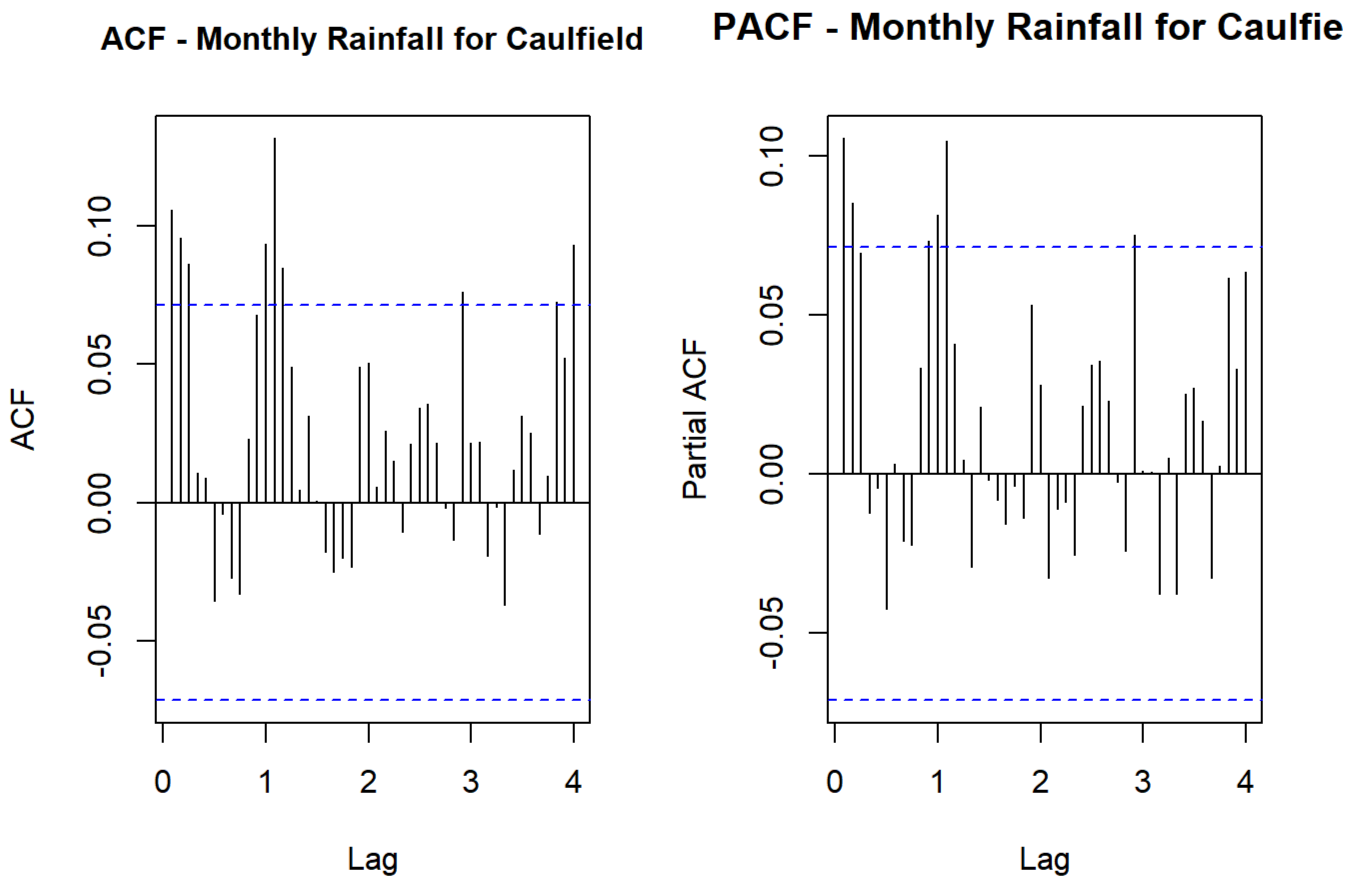


Figure 5: ACF and PACF Plot Of Rainfall for Caulfield (Racecourse)

- Analyzing the ACF plot reveals a pronounced seasonal pattern in the data, characterized by significant peaks at regular intervals corresponding to the seasonal component of the data. Additionally, while there may not be a gradual decline in autocorrelation across all lags, the presence of few significant lags beyond the seasonal pattern suggests the existence of other underlying factors influencing the data.
- Examining the PACF plot, it is observed that the first lag is highly significant which gives evidence of the series not being stationary.

```
qqnorm(rainfall_data_ts_filled,main = "Q-Q plot for Rainfall of Caulfield")
qqline(rainfall_data_ts_filled, col =2)
```

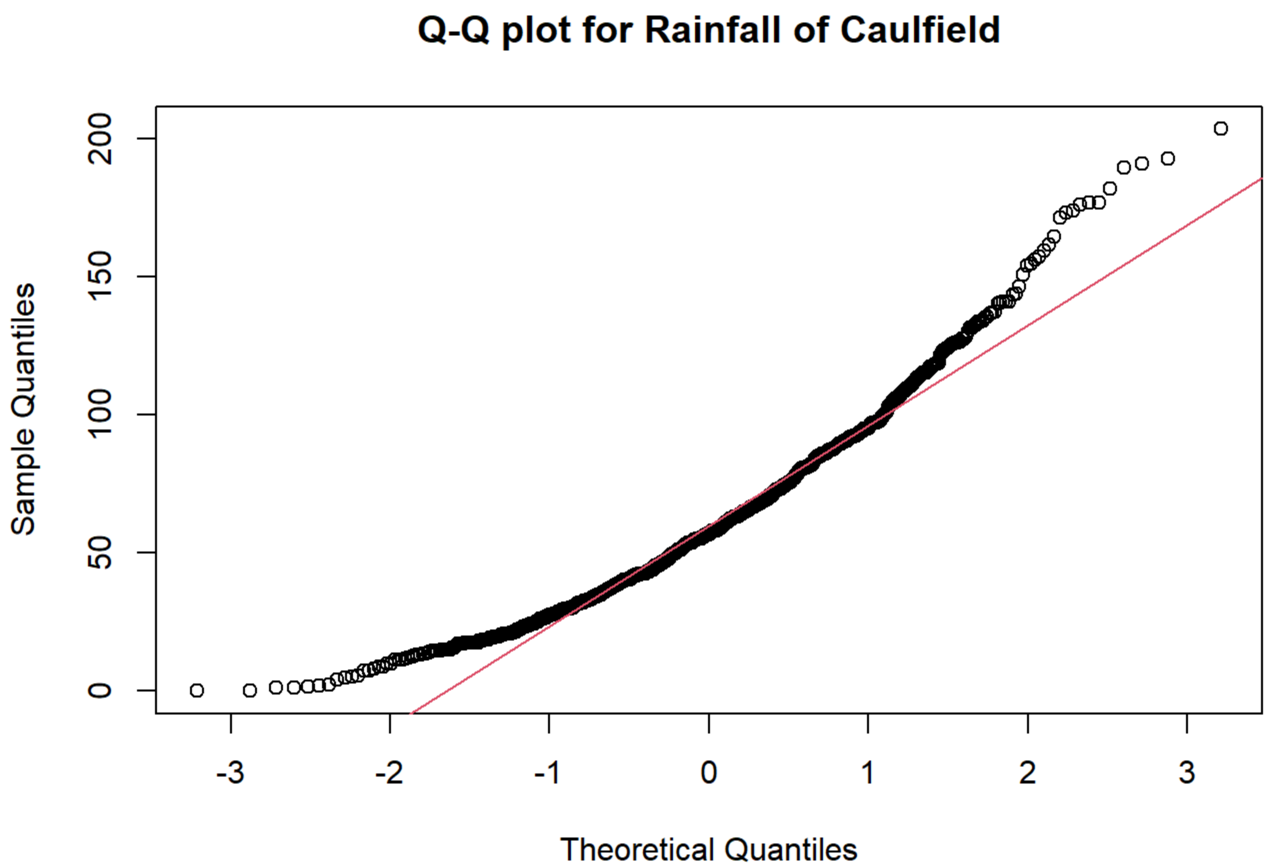


Figure 6: QQ Plot Of Rainfall for Caulfield (Racecourse)

Normal distribution is indicated by a Q-Q plot that appears as a virtually straight line. The Q-Q plot is deemed non-normally distributed because it does not represent an approximately straight line. Both tails appear to deviate from the straight line, indicating a skew distribution. The data also includes a broader range of values than normal.

```
shapiro.test(rainfall_data_ts_filled)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rainfall_data_ts_filled
## W = 0.95068, p-value = 3.325e-15
```

The Shapiro-Wilk test with p-value < 0.05, indicating there is sufficient evidence to reject the null hypothesis that the data is not normally distributed.

```
suppressWarnings({
rainfall_data_ts_filled_2 <- rainfall_data_ts_filled + abs(min(rainfall_data_ts_filled)) + 0.01
BC <- BoxCox.ar(y= rainfall_data_ts_filled_2, lambda = seq(-4,5, 0.01))
options(warn = 1)
})
```

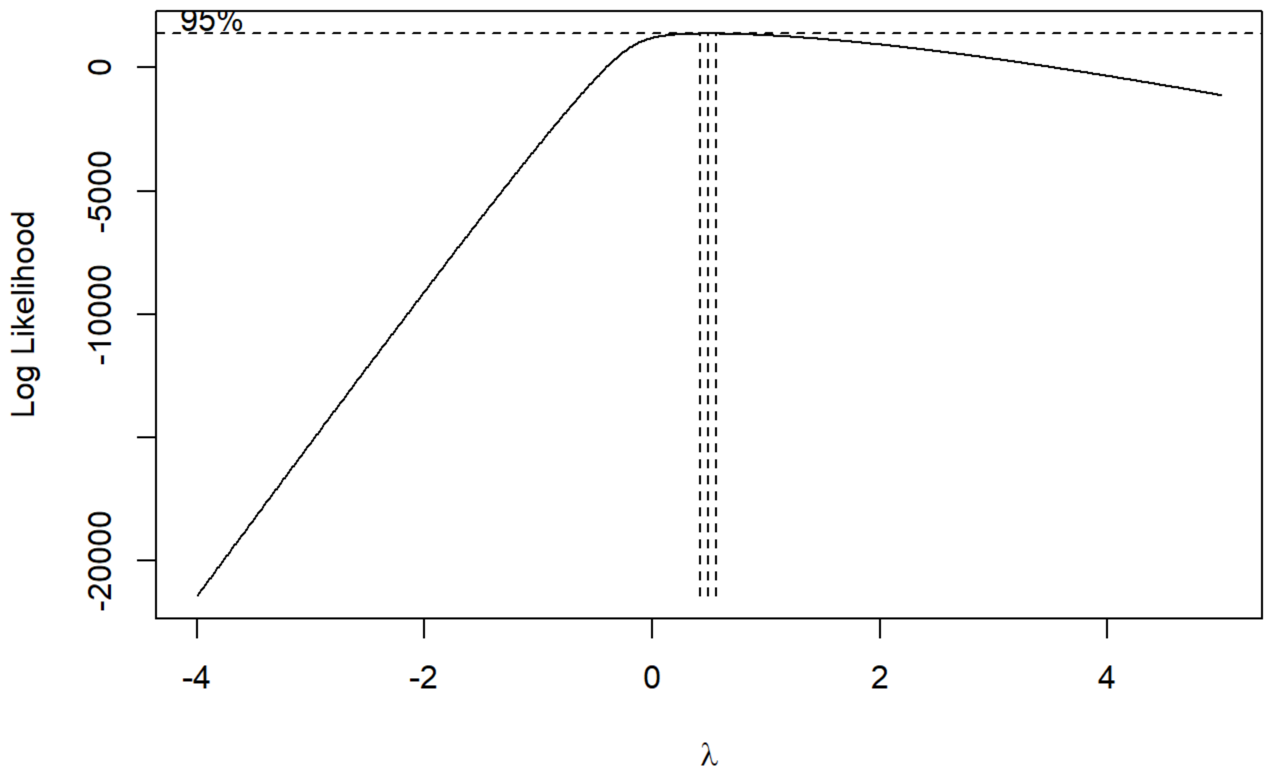


Figure 7: Box-Cox transformation Of Rainfall for Caulfield (Racecourse)

To improve the distributional features of the rainfall data, a Box-Cox modification was used. Before this transformation, a 0.01 constant was inserted to verify that all numbers were strictly positive.The Box-Cox transformation, carried out using the BoxCox.ar() function, investigated a wide range of lambda values (-4 to 5) with 0.01 increments. This thorough search made it easier to identify the ideal lambda value that maximized the modified data's log-likelihood.

```
BC$ci
```

```
## [1] 0.42 0.56
```

This interval shows that the ideal lambda value is within this range with a given amount of confidence, which can be used to guide the selection of a suitable lambda value.

```
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda
```

```
## [1] 0.49
```

The Lambda value for the greatest log-likelihood obtained from the Box-Cox transformation is 0.49. This lambda value is the best fit for transforming the data since it optimizes the likelihood of the converted data.

```
rainfall_data_bc <- ((rainfall_data_ts_filled_2^lambda)-1)/lambda
plot(rainfall_data_bc, type = "o", xlab = "Years", ylab = "BC transformation for Rainfall of Cau
lfield", main= "BC transformed for Rainfall of Caulfield")
```

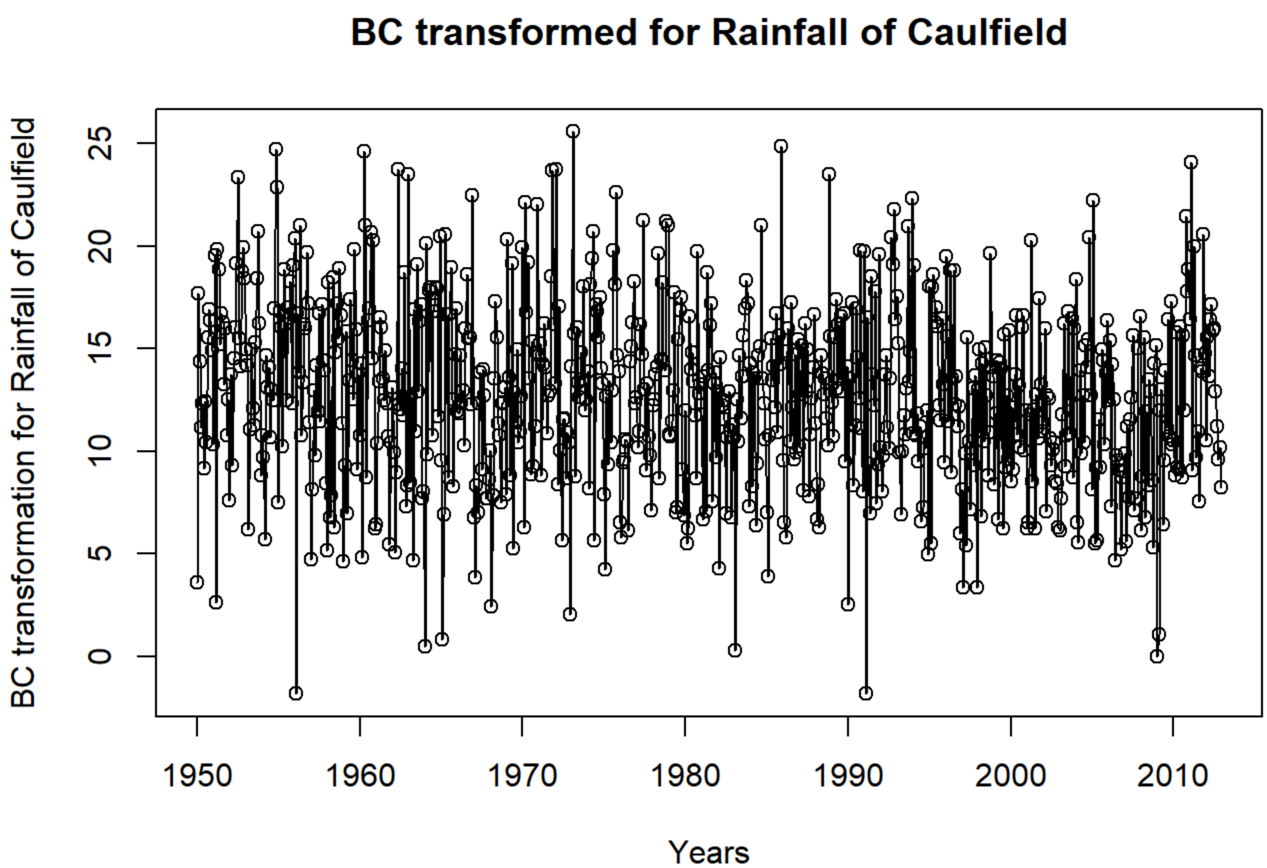


Figure 8: BoxCox transformed of Rainfall for Caulfield (Racecourse)

Following the Box Cox modification, the time series narrative improved slightly. The time series graphic exhibits indicators of stationarity and changing variance.

```
b <- ar(rainfall_data_bc)$order
adf.test(rainfall_data_bc, k=b)
```

```
## Warning in adf.test(rainfall_data_bc, k = b): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  rainfall_data_bc
## Dickey-Fuller = -5.9467, Lag order = 13, p-value = 0.01
## alternative hypothesis: stationary
```

In this context, with a lag order of 13 and a p-value of 0.01, the Augmented Dickey-Fuller test rejects the null hypothesis. This indicates that the Box-Cox transformed time series plot displays stationarity. As a result, the data maintains a constant variance over time and does not display significant time-dependent patterns.

```
pp.test(rainfall_data_bc, lshort =TRUE)
```

```
## Warning in pp.test(rainfall_data_bc, lshort = TRUE): p-value smaller than
## printed p-value
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  rainfall_data_bc
## Dickey-Fuller Z(alpha) = -733.35, Truncation lag parameter = 6, p-value
## = 0.01
## alternative hypothesis: stationary
```

In this context, lag parameter = 6, p-value = 0.01 rejecting the null hypothesis. This indicates the Box-Cox transformed time series plot displays stationarity.

```
kpss.test(rainfall_data_bc)
```

```
## Warning in kpss.test(rainfall_data_bc): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  rainfall_data_bc
## KPSS Level = 1.0514, Truncation lag parameter = 6, p-value = 0.01
```

These findings indicate that, at the specified significance level, there is enough evidence to reject the null hypothesis of level stationarity in favor of the alternative hypothesis. It indicates that the data has non-stationarity in its level component.

```
qqnorm(rainfall_data_bc,main = "Q-Q plot Boxcox Transformed of Rainfall for Caulfield (Racecours
e)")
qqline(rainfall_data_bc, col =2)
```

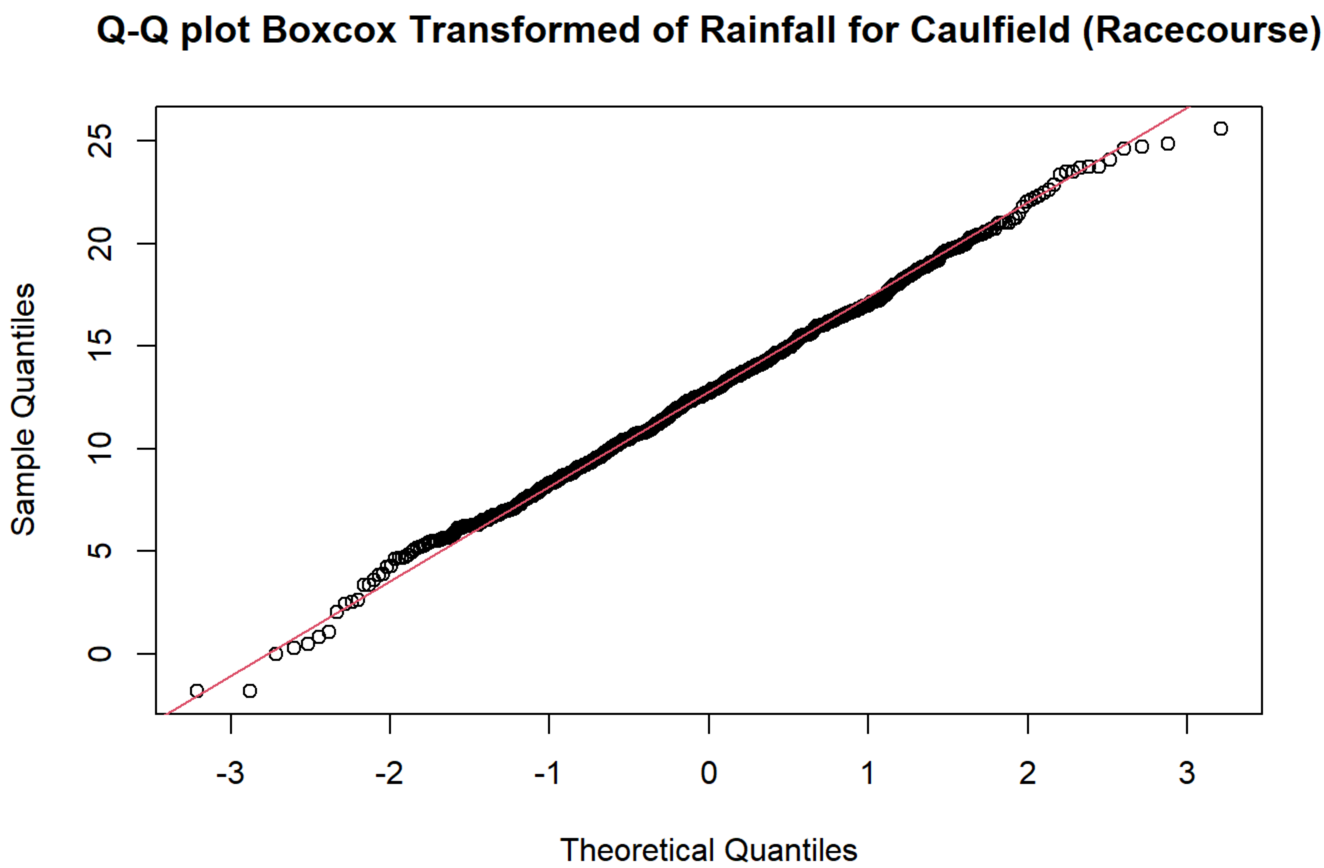


Figure 9: Q-Q plot BoxCox transformed of Rainfall for Caulfield (Racecourse)

```
shapiro.test(rainfall_data_bc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rainfall_data_bc
## W = 0.99812, p-value = 0.5813
```

Normal distribution is indicated by a Q-Q plot that appears as a virtually straight line. The Q-Q plot is normally distributed because it does represent an approximately straight line that aligns with the Shapiro-Wilk normality test of normality.

```
McLeod.Li.test(y=rainfall_data_bc)
```

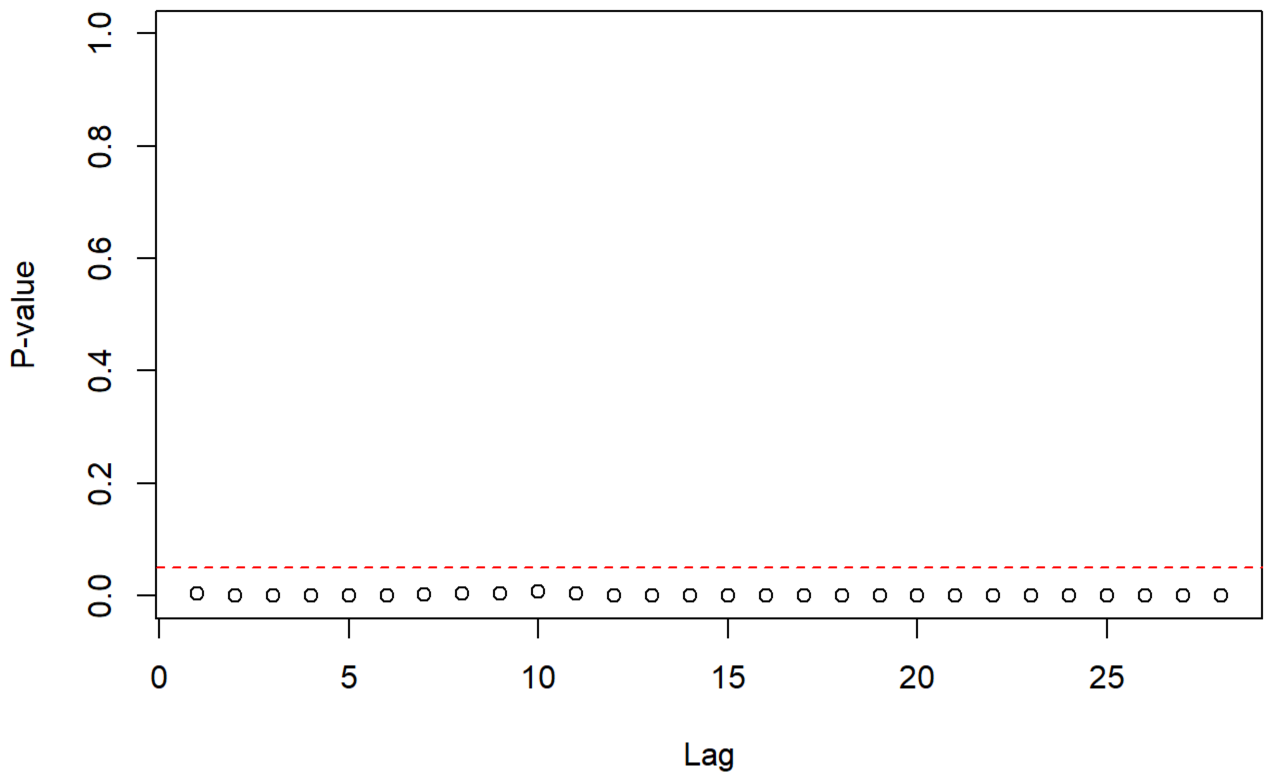


Figure 10: Mcleod-Li Test for Box-Cox Transformed Rainfall Series

McLeod-Li test was used to determine whether the observed data showed autocorrelation. The test results revealed that all dots in the autocorrelation function (ACF) were consistently below the 0.05 significance level. The data does not exhibit any significant autocorrelation. Therefore, the residuals are relatively independent, with no strong correlation between consecutive observations.

# SARIMA Model specification

## Seasonal Differencing

```
m1_rainfall <- Arima(rainfall_data_bc, order=c(0,0,0),
 seasonal=list(order=c(0,1,0),
 period=12))
m1_res <- rstandard(m1_rainfall)
plot(m1_res, xlab="Time", ylab="Residuals", main="Time series plot of the residuals from
m1_rainfall")
```

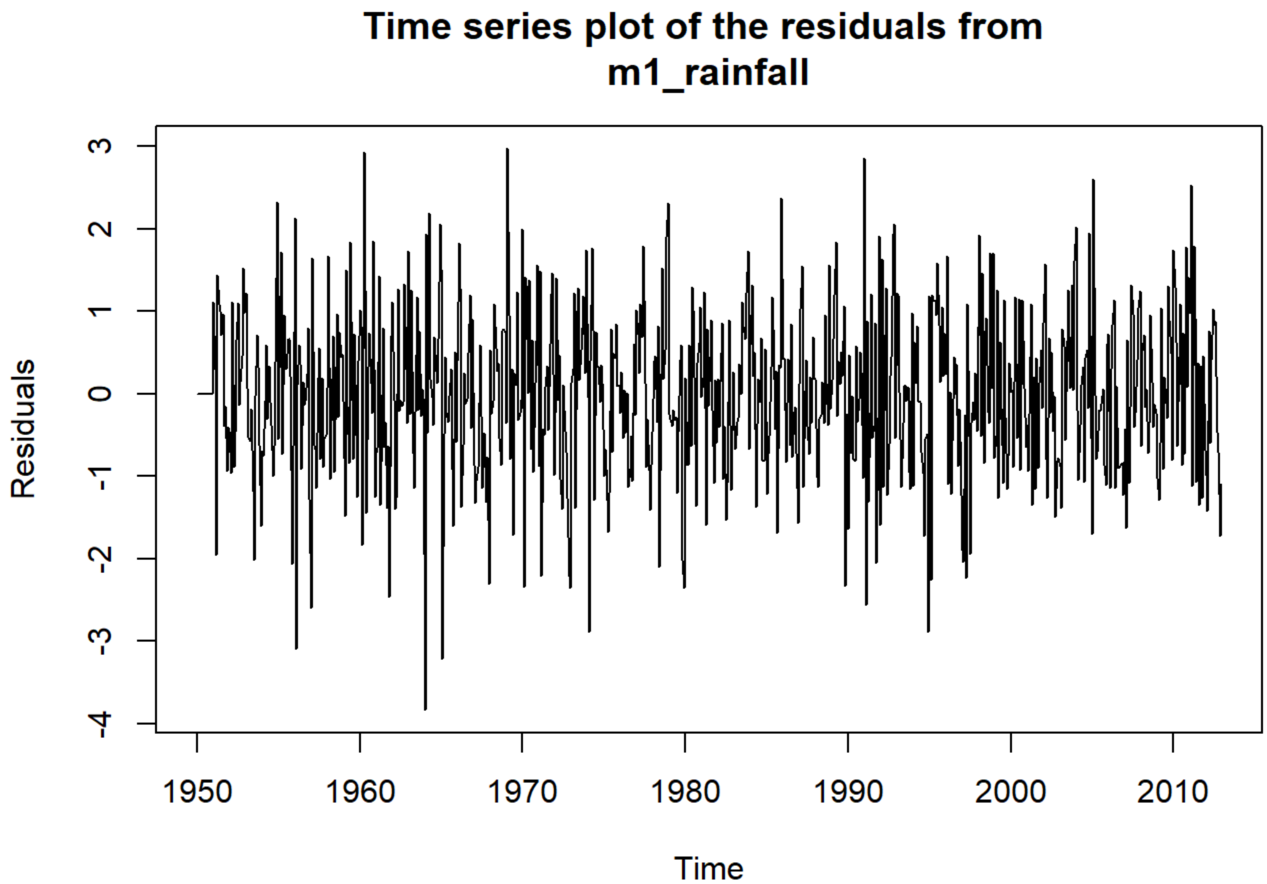


Figure 11: Residuals plot from the first seasonal difference SARIMA model.

The residuals from the first seasonal difference model showed some signs of trend, the plot of residuals suggested that seasonal differencing had made the residuals more similar to a white noise series.

```
par(mfrow=c(1,2))
acf(m1_res , lag.max = 48, main = "ACF plot of the residuals after first seasonal order differencing", cex.main =1.0)
pacf(m1_res , lag.max = 48, main = "PACF plot of the residuals after first seasonal order differencing", cex.main =1.0)
```

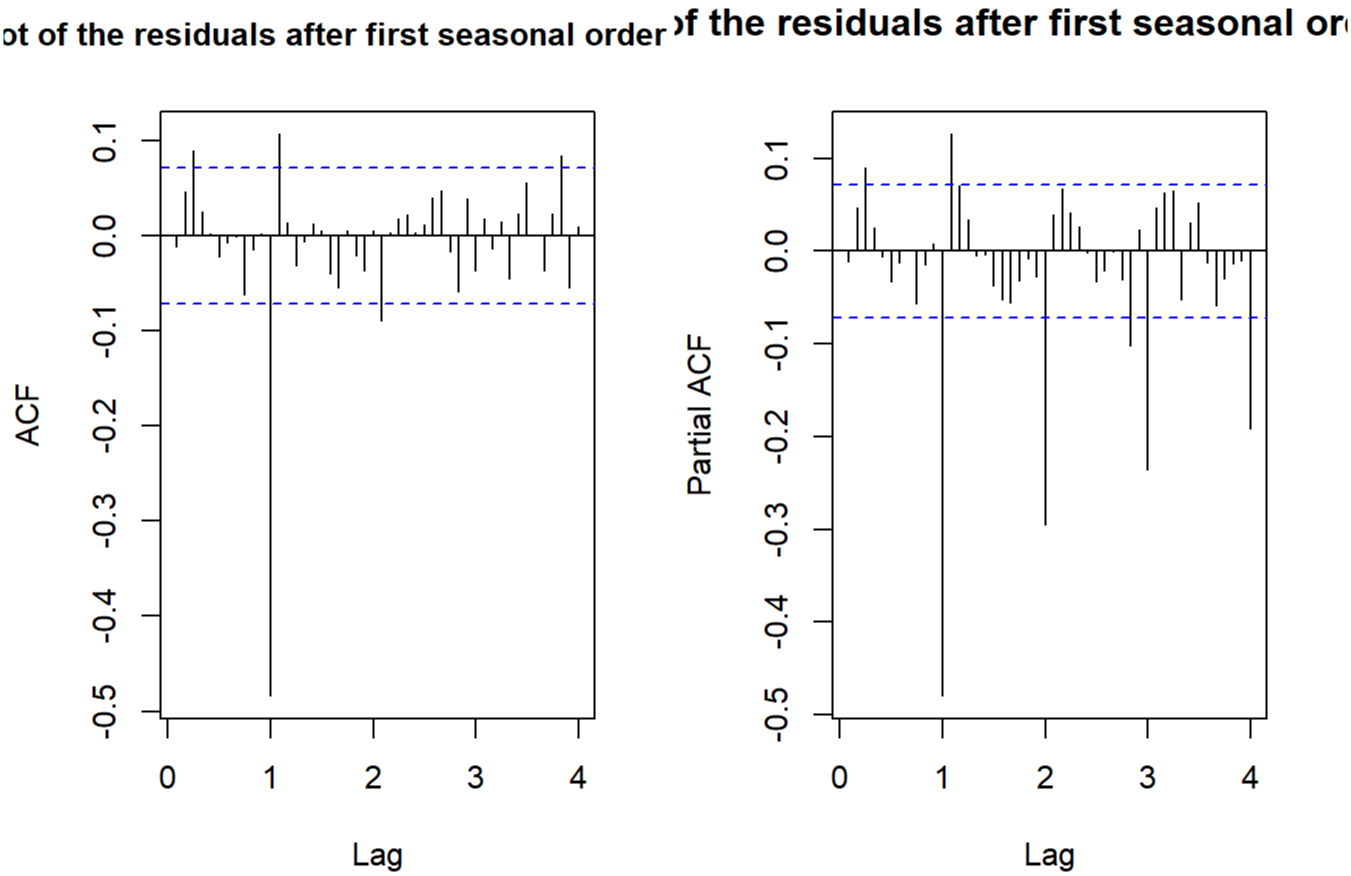


Figure 12: ACF and PACF plot of the residuals after first seasonal order differencing

The ACF plot indicated a significant seasonal lag at lag 1, indicating that the capital Q value in the SARIMA model is 1. However, the PACF displays slowly decaying significant lags at lag 1, lag 2, lag 3, and lag 4, therefore the value capital P is zero.

```
m2_rainfall <- Arima(rainfall_data_bc, order=c(0,0,0),
 seasonal=list(order=c(0,1,1),
 period=12))
m2_res <- rstandard(m2_rainfall)
plot(m2_res, xlab="Time", ylab="Residuals", main="Time series plot of the residuals from
m2_rainfall")
```

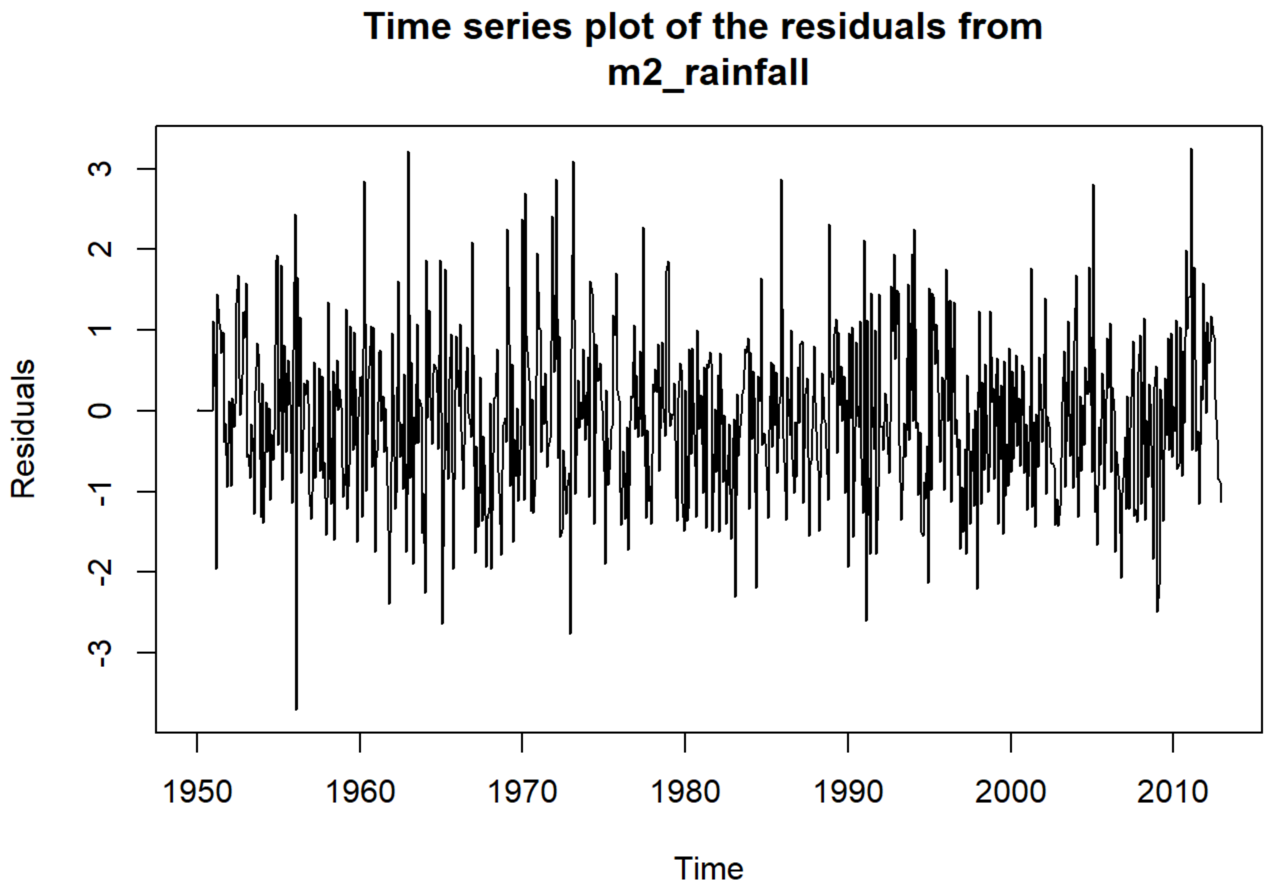


Figure 13: Time series plot of the residuals from m2_rainfall

```
par(mfrow=c(1,2))
acf(m2_res , lag.max = 48, main = "ACF plot of the residuals after fitting SARIMA (0,0,0)x(0,1,
1)_12", cex.main =1.0)
pacf(m2_res , lag.max = 48, main = "PACF plot of the residuals after fitting SARIMA (0,0,0)x(0,
1,1)_12", cex.main =1.0)
```

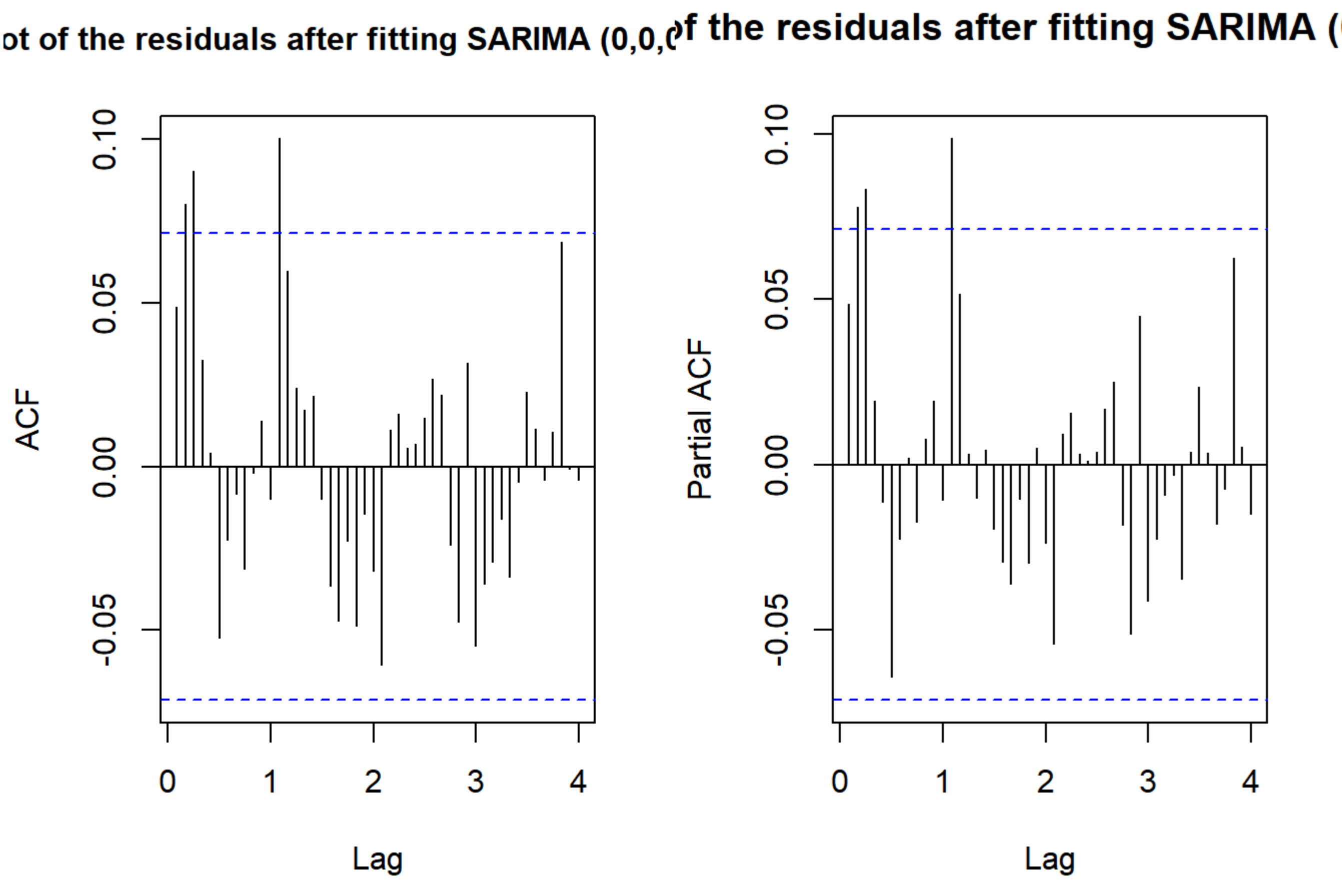


Figure 14: ACF and PACF plot of the residuals after fitting SARIMA (0,0,0)x(0,1,1)_12

```
adf.test(m2_res)
```

```
## Warning in adf.test(m2_res): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  m2_res
## Dickey-Fuller = -8.6514, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(m2_res, lshort =TRUE)
```

```
## Warning in kpss.test(m2_res, lshort = TRUE): p-value greater than printed
## p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  m2_res
## KPSS Level = 0.070863, Truncation lag parameter = 6, p-value = 0.1
```

Looking at the ACF and PACF there is not significant bars at seasons and there is also no decaying pattern which gives an indication of the residuals to be stationary that aligns with the adf test and KPSS test provided above. As the capital Q was already specified and the residuals of the series is also stationary ordinary differencing would not be necessary.

# EACF

To suggest potential values of small p,d and q within SARIMA (p,d,q)X(0,1,1)_12 model the extended ACF was employed on the residuals from the "m2_res" SARIMA model. The output is shown below.

```
eacf(m2_res)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o x x o o o o o o o o  o  x  o
## 1 x o o o o o o o o o o  o  x  o
## 2 x x o o o o o o o o o  o  x  x
## 3 x x x o o o o o o o o  o  x  o
## 4 x o x x o o o o o o o  o  x  o
## 5 x x x x x o o o o o o  o  x  o
## 6 x x x x o o o o o o o  o  x  o
## 7 o x x x x o x o o o o  o  o  o
```

The top left 'o' symbol in EACF denotes the intersection of of Autoregressive and Moving Average (MA). The positions of adjacents "o"s are subsequently employed to suggest further models from the EACF plot.

The set of models from eacf are as follows: * {SARIMA(1,0,1)x(0,1,1)_12} * {SARIMA(1,0,2)x(0,1,1)_12} * {SARIMA(2,0,2)x(0,1,1)_12}

```
res <- armasubsets(y=m2_res, nar=5, nma=5, y.name='p', ar.method='ols')
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 2 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
plot(res)
```

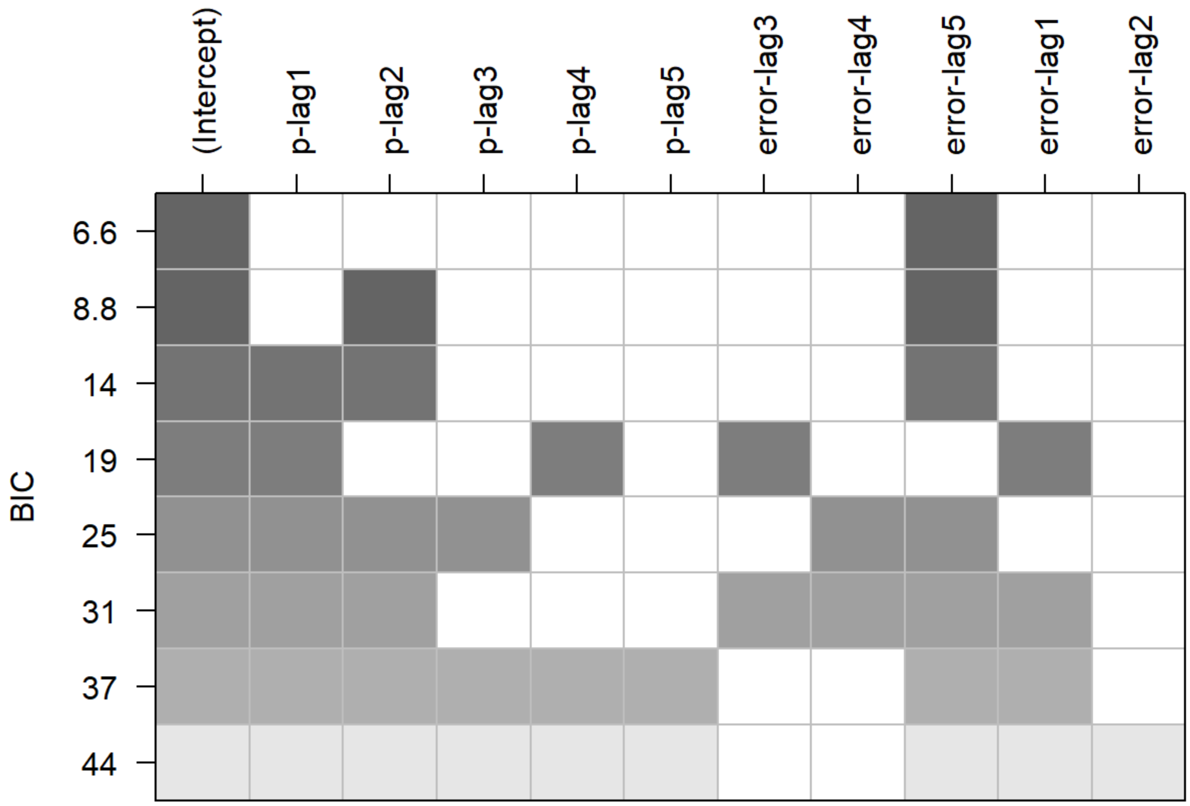


Figure 15: BIC plot of the residuals after fitting SARIMA (0,0,0)x(0,1,1)_12
By highlighting the shaded squares at the top rows, the BIC plot helps pinpoint the values for “small p” and “small q,” which correspond to models with the lowest BIC scores.

From the BIC table, we read the model {SARIMA(0,0,5)x(0,1,1)_12} and {SARIMA(2,0,5)x(0,1,1)_12} is considered to be the preferred models.

# Final set of models considering the EACF and BIC are :

{SARIMA(1,0,1)x(0,1,1)_12} {SARIMA(1,0,2)x(0,1,1)_12} {SARIMA(2,0,2)x(0,1,1)_12} {SARIMA(0,0,5)x(0,1,1)_12} {SARIMA(2,0,5)x(0,1,1)_12}

```
m_101_ml = Arima(rainfall_data_bc,order=c(1,0,1),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_101_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.755215   0.139824   5.4012 6.621e-08 ***
## ma1  -0.680629   0.155437  -4.3788 1.193e-05 ***
## sma1 -0.943900   0.016514 -57.1580 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on rainfall_data_bc time series, a SARIMA(1,0,1)(0,1,1)_12 model is fitted using maximum likelihood estimation (method = “ML”). In the following coeftest() function, the estimated coefficients are tested using a z-test. The output displays the estimated coefficients, z-values, and p-values along with their standard errors. Based on the results, the autoregressive coefficient (ar1) is estimated at 0.755215, with a standard error of 0.139824 and a corresponding z-value of 5.4012, yielding a highly significant p-value of 5.922e-08. The moving average coefficient (ma1) is estimated at -0.680629, with a z-value of -4.3788 and a highly significant p-value of 1.193e-05. Furthermore, the seasonal moving average coefficient (sma1) is estimated to be -0.943900, with a significant p-value of < 2.2e-16. All coefficients are statistically significant, indicating their importance in modeling time series data.

```
m_101_css = Arima(rainfall_data_bc,order=c(1,0,1),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_101_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.755215   0.139824   5.4012 6.621e-08 ***
## ma1  -0.680629   0.155437  -4.3788 1.193e-05 ***
## sma1 -0.943900   0.016514 -57.1580 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on rainfall_data_bc time series, a SARIMA(1,0,1)(0,1,1)_12 model is fitted using maximum likelihood estimation (method = “CSS”). In the following coeftest() function, the estimated coefficients are tested using a z-test. The output displays the estimated coefficients, z-values, and p-values along with their standard errors. Based on the results, the autoregressive coefficient (ar1) is estimated at 0.755215, with a standard error of 0.139824 and a corresponding z-value of 5.4012, yielding a highly significant p-value of 5.922e-08. The moving average coefficient (ma1) is estimated at -0.680629, with a z-value of -4.3788 and a highly significant p-value of 1.193e-05. Furthermore, the seasonal moving average coefficient (sma1) is estimated to be -0.943900, with a significant p-value of < 2.2e-16. All coefficients are statistically significant, indicating their importance in modeling time series data. Both the CSS and ML model are consistent with the result.

```
m_101_css_ml = Arima(rainfall_data_bc,order=c(1,0,1),seasonal=list(order=c(0,1,1), period=12),me
thod = "CSS-ML")
coeftest(m_101_css_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.755824   0.139858   5.4042 6.509e-08 ***
## ma1  -0.681298   0.155546  -4.3800 1.187e-05 ***
## sma1 -0.943909   0.016515 -57.1551 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on rainfall_data_bc time series, a SARIMA(1,0,1)(0,1,1)_12 model is fitted using maximum likelihood estimation (method = “CSS”). In the following coeftest() function, the estimated coefficients are tested using a z-test. The output displays the estimated coefficients, z-values, and p-values along with their standard errors. Based on the results, the autoregressive coefficient (ar1) is estimated at 0.755824, with a standard error of 0.139858 and a corresponding z-value of 5.4042, yielding a highly significant p-value of 6.509e-08. The moving average coefficient (ma1) is estimated at -0.681298, with a z-value of -4.3800 and a highly significant p-value of 1.187e-05. Furthermore, the seasonal moving average coefficient (sma1) is estimated to be -0.943909 , with a significant p-value of < 2.2e-16. All coefficients are statistically significant, indicating their importance in modeling time series data. Both the CSS,ML and CSS_ML model are consistent with the result.

```
m_102_ml = Arima(rainfall_data_bc,order=c(1,0,2),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_102_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.587829   0.146093   4.0237  5.73e-05 ***
## ma1  -0.548703   0.146533  -3.7446 0.0001807 ***
## ma2   0.080686   0.042414   1.9023 0.0571273 .
## sma1 -0.944637   0.016436 -57.4738 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using maximum likelihood estimation (method = “ML”), the code fits an SARIMA(1,0,2)(0,1,1)_12 model to rainfall_data_bc time series. Estimated coefficients are displayed along with standard errors, z-values, and p-values. Notably, ar1, ma1, and sma1 are very significant, indicating their usefulness in modeling time series data. However, ma2 appears to be marginally insignificant (p = 0.105), showing that its role may be less important in this scenario.

```
m_102_css = Arima(rainfall_data_bc,order=c(1,0,2),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_102_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value Pr(>|z|)
## ar1  -0.242519   0.172518  -1.4058  0.15979
## ma1   0.301172   0.176835   1.7031  0.08855 .
## ma2   0.065437   0.035253   1.8562  0.06342 .
## sma1 -0.912970   0.015317 -59.6069  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using conditional sum of squares (method = “CSS”), the code fits an SARIMA(1,0,2)(0,1,1)_12 model to rainfall_data_bc time series. Estimated coefficients are displayed along with standard errors, z-values, and p-values. Notably ma1 and ma2 is marginally significant and sma1 is highly significant, indicating their usefulness in modeling time series data. However, ar1 is not statistically significant at the conventional significance level ( α = 0.0.05), showing that its role may be less important in this scenario. The CSS and ML for the m_102 did not deliver the consistent results.

```
m_102_css_ml = Arima(rainfall_data_bc,order=c(1,0,2),seasonal=list(order=c(0,1,1), period=12),me
thod = "CSS-ML")
coeftest(m_102_css_ml)
```

```
## 
## z test of coefficients:
## 
##       Estimate Std. Error  z value  Pr(>|z|)    
## ar1   0.588681   0.145970   4.0329 5.509e-05 ***
## ma1  -0.549528   0.146401  -3.7536 0.0001743 ***
## ma2   0.080615   0.042451   1.8990 0.0575631 .  
## sma1 -0.944653   0.016435 -57.4777 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the “CSS” and “ML” method, the code fits an SARIMA(1,0,2)(0,1,1)_12 model to rainfall_data_bc time series. Estimated coefficients are displayed along with standard errors, z-values, and p-values. Notably ar1, ma1 and sma1 are highly significant, indicating their usefulness in modeling time series data. However ma2 is marginally statistically significant at the conventional significance level ( α = 0.05). The CSS-ML and ML for the m_102 did deliver the consistent results for each other.

```
m_202_ml = Arima(rainfall_data_bc,order=c(2,0,2),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_202_ml)
```

```
## 
## z test of coefficients:
## 
##       Estimate Std. Error  z value  Pr(>|z|)    
## ar1   1.536152   0.135449  11.3412 < 2.2e-16 ***
## ar2  -0.806484   0.115805  -6.9642 3.304e-12 ***
## ma1  -1.484923   0.132247 -11.2284 < 2.2e-16 ***
## ma2   0.802643   0.102492   7.8313 4.829e-15 ***
## sma1 -0.945147   0.016294 -58.0056 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The provided code fits an SARIMA(2,0,2)(0,1,1)_12 model using “ML” to rainfall_data_bc time series. Notably all coefficients are highly significant.According to the model, the autoregressive and moving average components, as well as the seasonal moving average component, significantly contribute to capturing the underlying patterns and trends in rainfall.

```
m_202_css = Arima(rainfall_data_bc,order=c(2,0,2),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_202_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.760619   0.200967   3.7848 0.0001538 ***
## ar2  -0.311194   0.192320  -1.6181 0.1056396
## ma1  -0.719270   0.197789  -3.6366 0.0002763 ***
## ma2   0.383014   0.202087   1.8953 0.0580533 .
## sma1 -0.895682   0.015973 -56.0733 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The SARIMA(2,0,2)(0,1,1)_12 model using “CSS” to rainfall_data_bc time series. Notably ar1, ma1 and sma1 are highly significant and ma2 to be marginally significant implying a less significant impact than the other components. However ar2 to be insignificant, this implies that its contribution to explaining the variation in the data may be minimal.

```
m_202_css_ml = Arima(rainfall_data_bc,order=c(2,0,2),seasonal=list(order=c(0,1,1), period=12),me
thod = "CSS-ML")
coeftest(m_202_css_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   1.536965   0.133447  11.5174 < 2.2e-16 ***
## ar2  -0.807078   0.114152  -7.0702 1.547e-12 ***
## ma1  -1.485700   0.130296 -11.4025 < 2.2e-16 ***
## ma2   0.803232   0.101155   7.9406 2.012e-15 ***
## sma1 -0.945132   0.016299 -57.9868 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The provided code fits an SARIMA(2,0,2)(0,1,1)_12 model using “CSS-ML” method to rainfall_data_bc time series. Notably all coefficients are highly significant the aligns with the output of ML method.

```
m_005_ml = Arima(rainfall_data_bc,order=c(0,0,5),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_005_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ma1   0.042496   0.036912   1.1513  0.249617
## ma2   0.085253   0.037075   2.2995  0.021478 *
## ma3   0.106086   0.038149   2.7809  0.005422 **
## ma4   0.048976   0.036895   1.3274  0.184362
## ma5   0.019877   0.037982   0.5233  0.600758
## sma1 -0.945305   0.016491 -57.3229 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The code uses the ML approach to fit a SARIMA(0,0,5)(0,1,1)_12 model to the rainfall_data_ts_filled data series. Notably ma2, ma3, and sma1 are statistically significant. However, ma1,ma4 and ma5 are not statistically significant.

```
m_005_css = Arima(rainfall_data_bc,order=c(0,0,5),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_005_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value Pr(>|z|)
## ma1   0.044822   0.036895   1.2149  0.22442
## ma2   0.067817   0.036958   1.8350  0.06651 .
## ma3   0.097606   0.038084   2.5629  0.01038 *
## ma4   0.045433   0.037103   1.2245  0.22076
## ma5   0.019295   0.038796   0.4973  0.61895
## sma1 -0.888638   0.016145 -55.0427  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

br> The code uses the CSS approach to fit a SARIMA(0,0,5)(0,1,1)_12 model to the rainfall_data_bc data series. Notably, ma2 is marginally significant, whereas ma3, and sma1 are statistically significant. However, ma1,ma4 and ma5 are not statistically significant.The output of CSS aligns with the result of ML method.

```
m_005_css_ml = Arima(rainfall_data_bc,order=c(0,0,5),seasonal=list(order=c(0,1,1), period=12),me
thod = "CSS-ML")
coeftest(m_005_css_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ma1   0.042495   0.036912   1.1513  0.249627
## ma2   0.085252   0.037075   2.2995  0.021479 *
## ma3   0.106086   0.038148   2.7809  0.005421 **
## ma4   0.048978   0.036895   1.3275  0.184339
## ma5   0.019878   0.037982   0.5234  0.600724
## sma1 -0.945308   0.016491 -57.3239 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The code uses the CSS-ML approach to fit a SARIMA(0,0,5)(0,1,1)_12 model to the rainfall_data_bc data series. Notably ma2, ma3, and sma1 are statistically significant. However, ma1, ma4 and ma5 are not statistically significant.The output of CSS-ML aligns with the result of CSS and ML method.

```
m_205_ml = Arima(rainfall_data_bc,order=c(2,0,5),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_205_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.805199   0.246353  -3.2685 0.0010813 **
## ar2  -0.818064   0.185147  -4.4184 9.941e-06 ***
## ma1   0.846145   0.245830   3.4420 0.0005775 ***
## ma2   0.933736   0.195946   4.7653 1.886e-06 ***
## ma3   0.197654   0.061613   3.2080 0.0013367 **
## ma4   0.180955   0.048079   3.7637 0.0001674 ***
## ma5   0.122873   0.038733   3.1723 0.0015122 **
## sma1 -0.941802   0.017489 -53.8515 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

* Using the ML method, the provided code attempts to fit a SARIMA(2,0,5)(0,1,1)_12 model to rainfall_data_bc time series. All the coefficients for are statistically significant, indicating their importance in modeling time series data.

```
m_205_css = Arima(rainfall_data_bc,order=c(2,0,5),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_205_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.297539   0.163955  -1.8148 0.0695609 .
## ar2  -0.599870   0.162476  -3.6921 0.0002225 ***
## ma1   0.348241   0.168329   2.0688 0.0385634 *
## ma2   0.704256   0.161883   4.3504 1.359e-05 ***
## ma3   0.160307   0.049142   3.2621 0.0011057 **
## ma4   0.123777   0.045291   2.7330 0.0062769 **
## ma5   0.080725   0.037142   2.1734 0.0297500 *
## sma1 -0.905380   0.014980 -60.4402 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

* A SARIMA(2,0,5)(0,1,1)_12 model is fitted to rainfall_data_bc time series using CSS. The p-values and z-values for ar2, ma1, ma2, ma3, ma4, ma5, and sma1 are all extremely low and highly significant except ar1 to be marginally significant. As a result, these coefficients provide valuable insights into the underlying patterns and dynamics of the time series data.

```
m_205_css_ml = Arima(rainfall_data_bc,order=c(2,0,5),seasonal=list(order=c(0,1,1), period=12),me
thod = "CSS-ML")
coeftest(m_205_css_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.801616   0.232391  -3.4494 0.0005618 ***
## ar2  -0.814575   0.176823  -4.6067 4.090e-06 ***
## ma1   0.842591   0.232192   3.6288 0.0002847 ***
## ma2   0.930130   0.187224   4.9680 6.764e-07 ***
## ma3   0.197347   0.061035   3.2334 0.0012234 **
## ma4   0.180872   0.047945   3.7725 0.0001616 ***
## ma5   0.122910   0.038719   3.1744 0.0015014 **
## sma1 -0.941915   0.017367 -54.2348 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the CSS-ML method, the code fits a SARIMA(2,0,5)(0,1,1)_12 model to rainfall_data_bc time series. All of the estimated coefficients are highly significant, as indicated by low p-values (all < 2.2e-16) and high z-values. The coefficients provide valuable insight into underlying patterns and dynamics of time series data by effectively modeling them.

```
sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  }
  else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}
sort.score(AIC(m_101_ml, m_102_ml, m_202_ml, m_005_ml, m_205_ml), score = "aic")
```

```
##          df      AIC
## m_202_ml  6 4328.989
## m_102_ml  5 4331.898
## m_101_ml  4 4332.928
## m_005_ml  7 4332.952
## m_205_ml  9 4334.441
```

```
sort.score(BIC(m_101_ml, m_102_ml, m_202_ml, m_005_ml, m_205_ml), score = "bic")
```

```
##          df      BIC
## m_101_ml  4 4351.376
## m_102_ml  5 4354.958
## m_202_ml  6 4356.662
## m_005_ml  7 4365.236
## m_205_ml  9 4375.950
```

Based on the Akaike Information Criterion (AIC), model m_202_ml had the lowest AIC value, indicating a strong fit to the data. Model m_101_ml, however, displayed the lowest Bayesian Information Criteria (BIC). In the end, both m_101_ml and m_202_ml were found to have significant coefficients, making them suitable for analysis. In order to identify the best model, additional metrics such as Mean Error (ME), Mean Absolute Scaled Error (MASE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) will be taken into account.

{SARIMA(1,0,1)x(0,1,1)_12} {SARIMA(1,0,2)x(0,1,1)_12} {SARIMA(2,0,2)x(0,1,1)_12} {SARIMA(0,0,5)x(0,1,1)_12} {SARIMA(2,0,5)x(0,1,1)_12}

```
Smodel_101_css <- accuracy(m_101_css)[1:7]
Smodel_102_css <- accuracy(m_102_css)[1:7]
Smodel_202_css <- accuracy(m_202_css)[1:7]
Smodel_005_css <- accuracy(m_005_css)[1:7]
Smodel_205_css <- accuracy(m_205_css)[1:7]
df.Smodels <- data.frame(
  rbind(Smodel_101_css,Smodel_102_css,
        Smodel_202_css,Smodel_005_css,Smodel_205_css)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(1,0,1)x(0,1,1)_12", "SARIMA(1,0,2)x(0,1,1)_12",
                          "SARIMA(2,0,2)x(0,1,1)_12","SARIMA(0,0,5)x(0,1,1)_12", "SARIMA(2,0,5)x
(0,1,1)_12")

round(df.Smodels,  digits = 3)
```

```
##                               ME  RMSE   MAE      MPE    MAPE  MASE   ACF1
## SARIMA(1,0,1)x(0,1,1)_12 -0.209 4.429 3.477 -166.858 189.608 0.726  0.021
## SARIMA(1,0,2)x(0,1,1)_12 -0.200 4.419 3.467 -170.283 193.129 0.724  0.004
## SARIMA(2,0,2)x(0,1,1)_12 -0.165 4.430 3.470 -160.225 183.488 0.725  0.017
## SARIMA(0,0,5)x(0,1,1)_12 -0.089 4.444 3.479 -153.496 177.400 0.727 -0.001
## SARIMA(2,0,5)x(0,1,1)_12 -0.164 4.391 3.441 -154.757 177.985 0.719 -0.002
```

# Best Model

Model SARIMA(2,0,2)x(0,1,1)_12 has the 3rd lowest ME, 4th lowest RMSE and 3rd lowest MAE, 3rd lowest MPE, 5th lowest MAPE and 3rd lowest MASE Model SARIMA(1,0,1)x(0,1,1)_12 has 1st lowest ME, 3rd lowest RMSE and 4th lowest MAE, 4th lowest MPE, 4th lowest MAPE, 4th lowest MASE.

Compared to all models SARIMA(2,0,2)x(0,1,1)_12 has lower MASE than SARIMA(1,0,1)x(0,1,1)_12 and SARIMA(2,0,2)x(0,1,1)_12 has all coefficients significant. Considering all the possible values it can be inferred that model m_202_ml is the best model considering AIC and error measures.

# Checking overfitting models

```
m_203_ml = Arima(rainfall_data_bc,order=c(2,0,3),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_203_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value Pr(>|z|)
## ar1   0.721318   0.535240   1.3477   0.1778
## ar2  -0.302118   0.288527  -1.0471   0.2951
## ma1  -0.680313   0.534978  -1.2717   0.2035
## ma2   0.355829   0.277608   1.2818   0.1999
## ma3   0.054327   0.074530   0.7289   0.4660
## sma1 -0.944736   0.016432 -57.4920   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m_203_css = Arima(rainfall_data_bc,order=c(2,0,3),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_203_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.513655   0.147291   3.4873 0.0004878 ***
## ar2  -0.545517   0.158813  -3.4350 0.0005926 ***
## ma1  -0.465155   0.150279  -3.0953 0.0019663 **
## ma2   0.567526   0.147084   3.8585 0.0001141 ***
## ma3   0.101167   0.037555   2.6938 0.0070639 **
## sma1 -0.899537   0.015651 -57.4760 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m_302_ml = Arima(rainfall_data_bc,order=c(3,0,2),seasonal=list(order=c(0,1,1), period=12),method
= "ML")
coeftest(m_302_ml)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value Pr(>|z|)
## ar1   0.780838   0.400462   1.9498  0.05119 .
## ar2  -0.402492   0.217092  -1.8540  0.06374 .
## ar3   0.054363   0.070123   0.7753  0.43819
## ma1  -0.739739   0.400613  -1.8465  0.06482 .
## ma2   0.453054   0.201765   2.2455  0.02474 *
## sma1 -0.944662   0.016444 -57.4478  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m_302_css = Arima(rainfall_data_bc,order=c(3,0,2),seasonal=list(order=c(0,1,1), period=12),metho
d = "CSS")
coeftest(m_302_css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value Pr(>|z|)
## ar1   0.025211   0.157319   0.1603  0.87268
## ar2  -0.144484   0.139937  -1.0325  0.30184
## ar3   0.093071   0.036963   2.5180  0.01180 *
## ma1   0.010761   0.157334   0.0684  0.94547
## ma2   0.220615   0.128302   1.7195  0.08552 .
## sma1 -0.881542   0.016490 -53.4585  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sort.score(AIC(m_202_ml,m_302_ml,m_203_ml), score = "aic")
```

```
##          df      AIC
## m_202_ml  6 4328.989
## m_203_ml  7 4332.682
## m_302_ml  7 4332.685
```

```
sort.score(BIC(m_202_ml,m_302_ml,m_203_ml), score = "bic")
```

```
##          df      BIC
## m_202_ml  6 4356.662
## m_203_ml  7 4364.966
## m_302_ml  7 4364.969
```

The coefficients in the overfitting models deemed insignificant under the both method of ML and CSS, confirming that SARIMA(2,0,2)x(0,1,1)_12 is appropriate model for the rainfall time series.Furthermore, both AIC and BIC comparisons show that the SARIMA(2,0,2)x(0,1,1)_12 model has the lowest values, further supporting its selection as the best model

# Diagnostic checking

```
st_res_202 <-rstandard(m_202_ml)
par(mar = c(5,6,4,1)+0.1)
plot(st_res_202, type ="o",xlab="Time", ylab="Standardised Residuals",
 main="Standardised residuals from the SARIMA(2,0,2)x(0,1,1)_12 Model.")
abline(h=0)
```

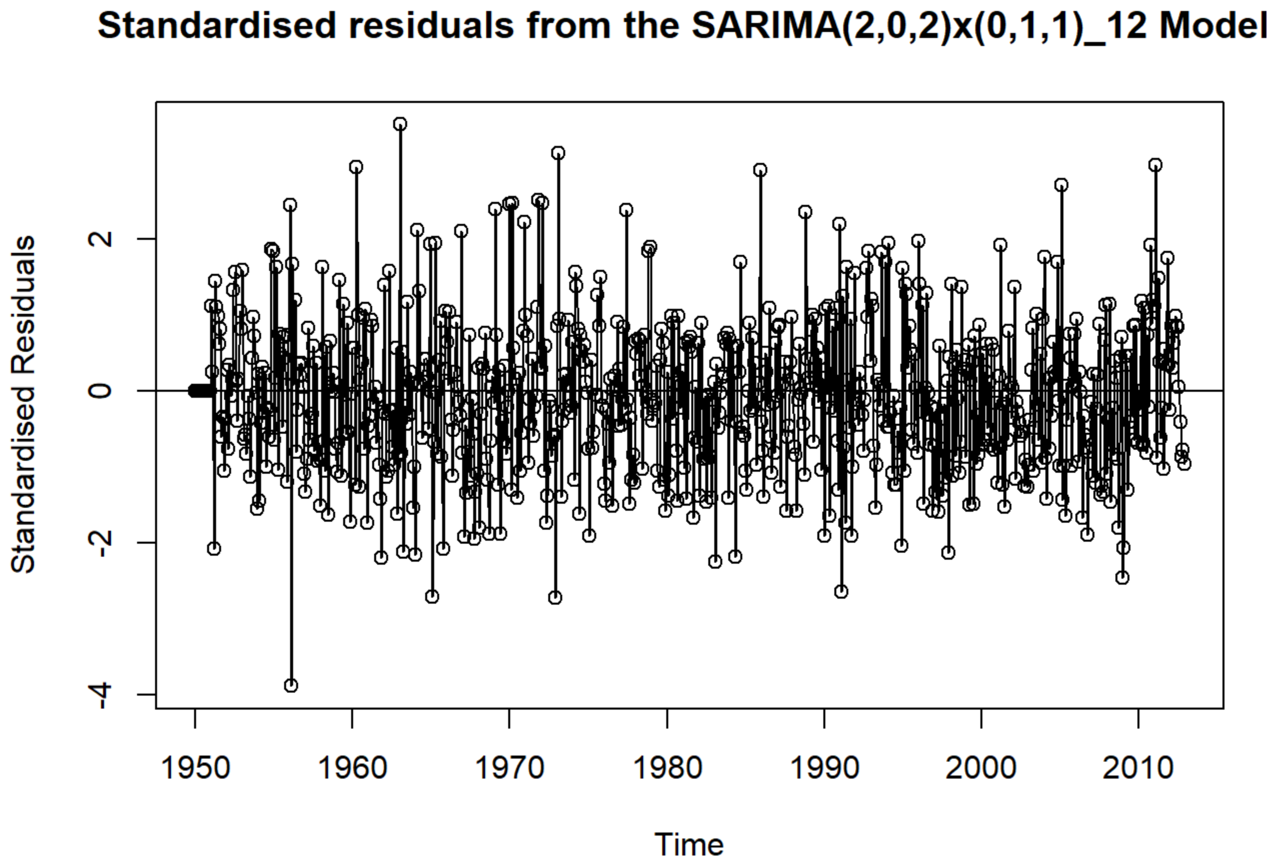


Figure 16: Time series plot of Standardised residuals from the SARIMA(2,0,2)x(0,1,1)_12 Model for the Rainfall series.

* A closer examination revealed potentially outlier observations from the SARIMA(2,0,2)x(0,1,1)_12 Model time series in spite of the absence of significant abnormalities in the standardized residuals derived from the applied SARIMA model.

```
par(mfrow=c(1,2))
acf(st_res_202, lag.max = 48, main = "ACF - Monthly Rainfall for Caulfield", cex.main =1.0)
pacf(st_res_202, lag.max = 48, main = "PACF - Monthly Rainfall for Caulfield", cex.main =5.0)
```

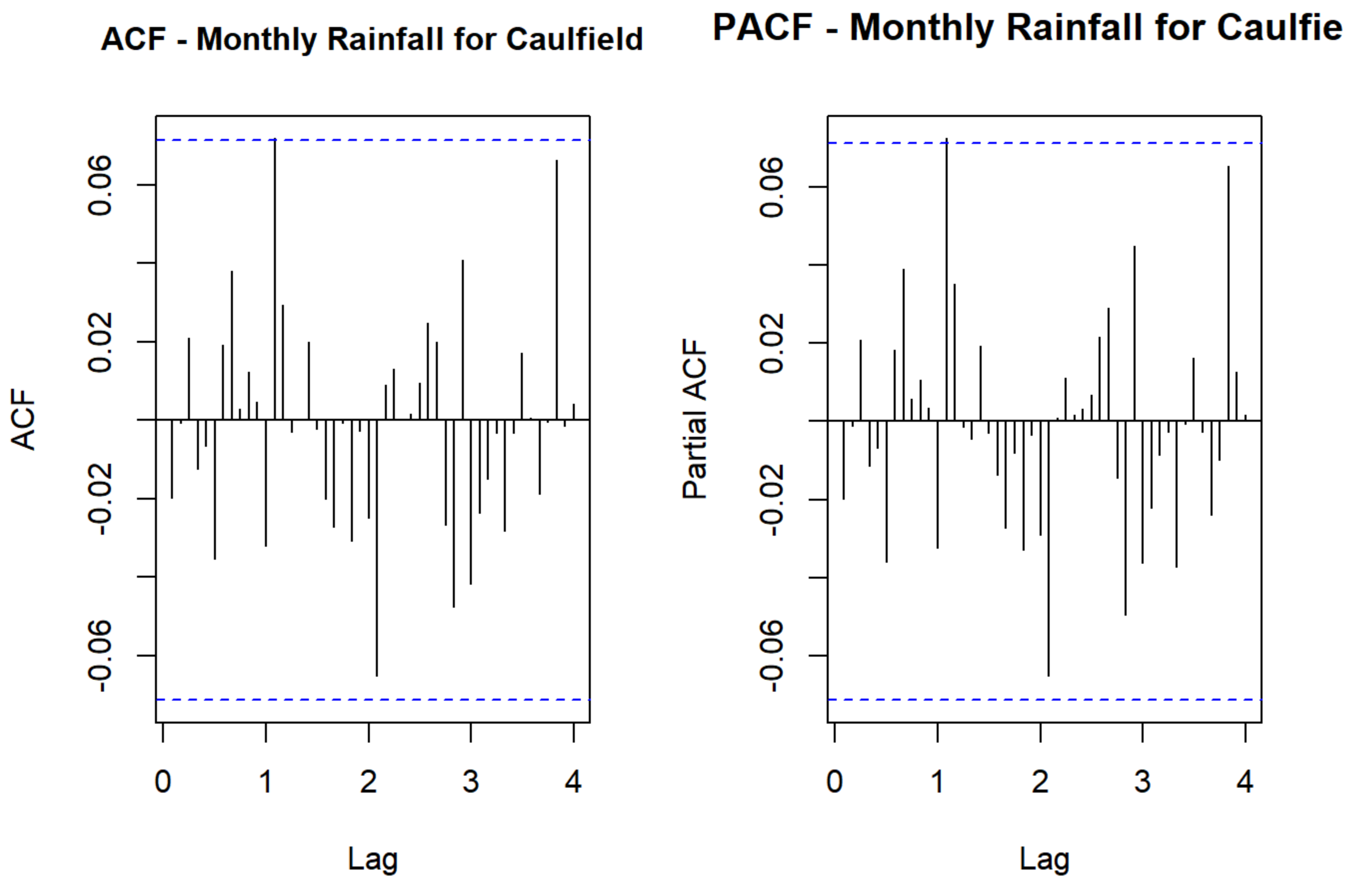


Figure 17: ACF plot of the Standardised residuals from the SARIMA(2,0,2)x(0,1,1)_12 Model for the Rainfall series.

Based on the ACF and PACF plot of standardized residuals, the residuals derived from fitted models did not exhibit any notable autocorrelation lags.

```
par(mfrow=c(1,2)); hist(st_res_202)
qqnorm(st_res_202);qqline(st_res_202, col=2)
```

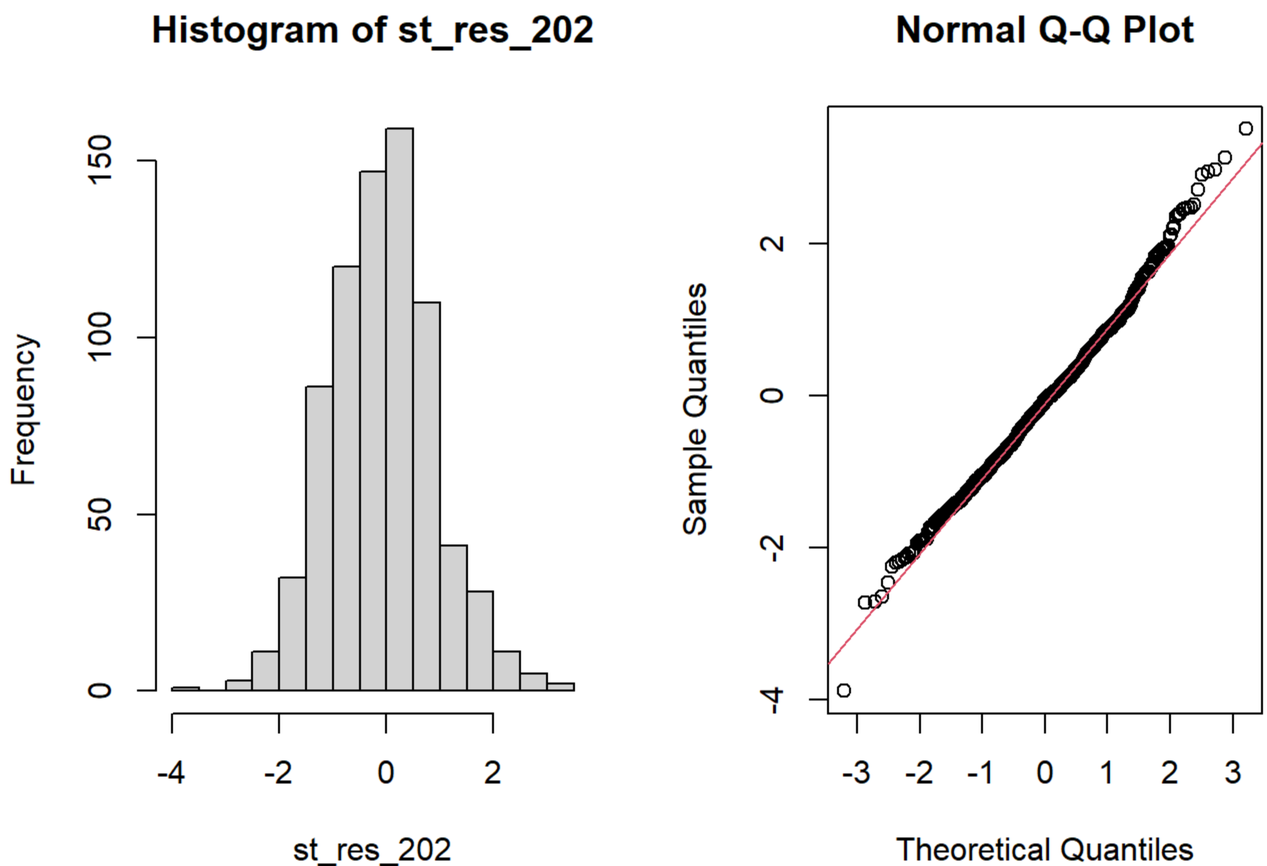


Figure 18: Q-Q plot and histogram of the Standardized residuals from the SARIMA(2,0,1)x(0,1,1)_12 Model for the Rainfall series.

```
shapiro.test(st_res_202)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  st_res_202
## W = 0.9939, p-value = 0.003719
```

The Q-Q plot of the standardized residuals showed no such great deviations from the normal in both tails of the residuals and the shapiro-wilk test of the standardized residuals indicate that residual data is not normally distributed. However the histogram of standardized residual appeared normally distributed.

```
lags <- 1:30
lbq_results <- sapply(lags, function(lag) Box.test(st_res_202, lag = lag, type = "Ljung-Box")$p.
value)

lbq_df <- data.frame(Lag = lags, p_value = lbq_results)

plot(lbq_df$Lag, lbq_df$p_value, type = "h", main = "Ljung-Box Q-test P-values",
     xlab = "Lag", ylab = "P-value", ylim = c(0, 1))
abline(h = 0.05, col = "red", lty = 2)
points(lbq_df$Lag, lbq_df$p_value, pch = 19)
```

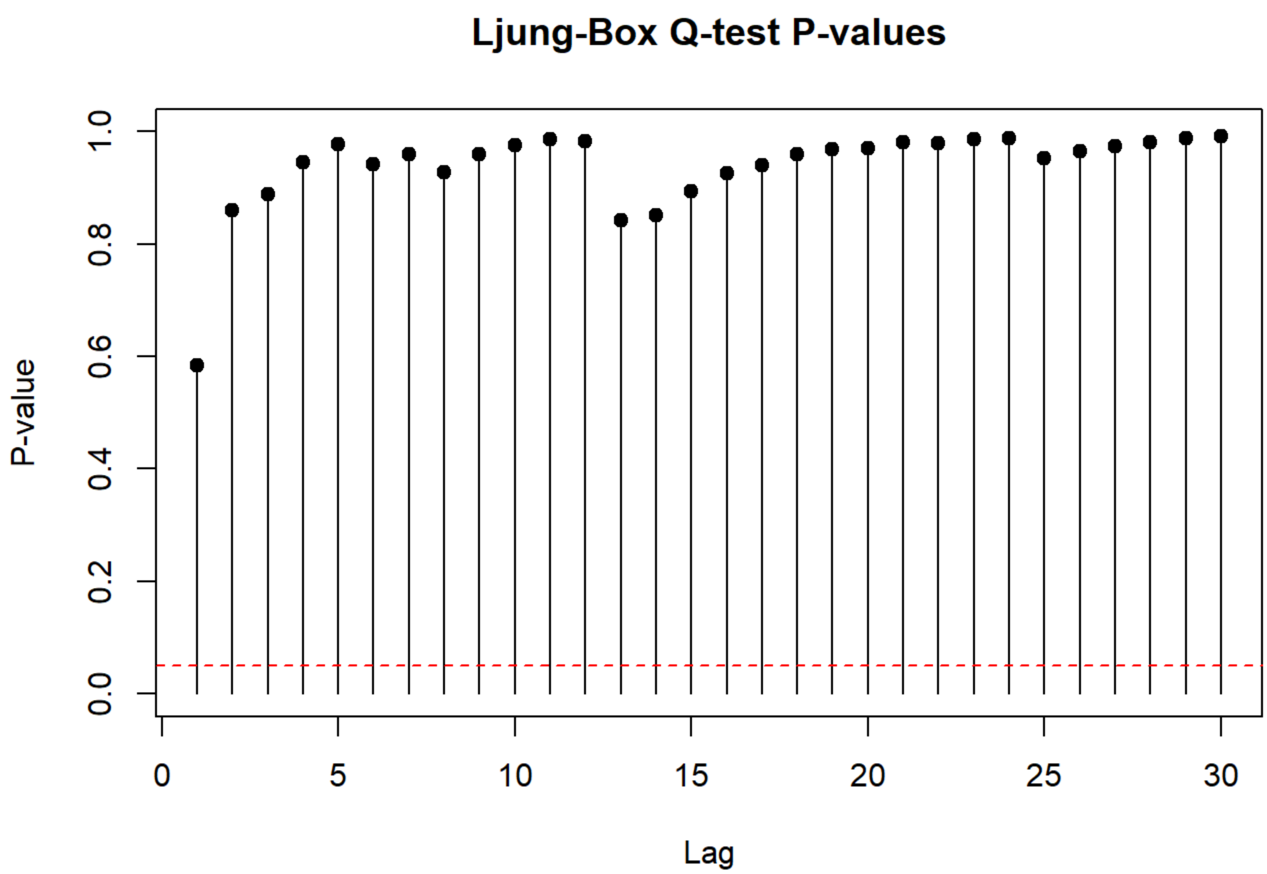


Figure 19:Ljung-Box test of the standardized residuals from the SARIMA(2,0,2)x(0,1,1)_12 Model for the Rainfall series.

* Ljung-Box test points were all above the 0.05 reference line, indicating the residuals were independent of errors

```
m_202_ml_Rainfall = Arima(rainfall_data_ts_filled,order=c(2,0,2),seasonal=list(order=c(0,1,1), p
eriod=12), lambda = 0.49, method = "ML")
```

```
forecast_ml <- forecast(m_202_ml_Rainfall,h=10)
plot(forecast_ml)
```

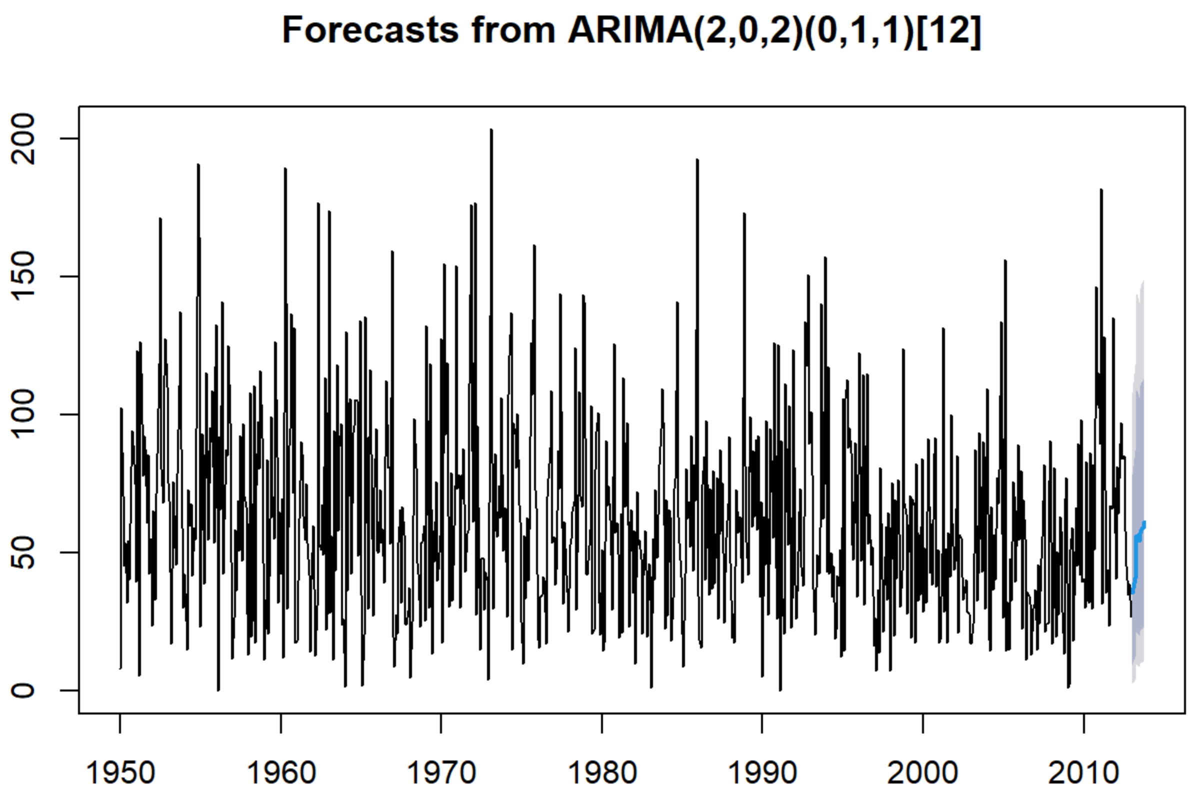


Figure 20: Forecast from SARIMA(2,0,2)x(0,1,1) Model for the Rainfall Series

```
print(forecast_ml)
```

```
##          Point Forecast     Lo 80     Hi 80     Lo 95    Hi 95
## Jan 2013       35.39722  9.503724  78.24069  2.512741 107.8807
## Feb 2013       37.37258 10.502836  81.25502  3.024916 111.4685
## Mar 2013       41.10608 12.442185  86.91066  4.073449 118.1904
## Apr 2013       56.28467 21.149258 108.76469  9.459697 143.6496
## May 2013       56.32779 21.138720 108.90969  9.439822 143.8683
## Jun 2013       54.40006 19.969591 106.21918  8.668147 140.7736
## Jul 2013       57.46917 21.825812 110.52214  9.894443 145.7363
## Aug 2013       58.29775 22.319252 111.70802 10.220191 147.1199
## Sep 2013       59.07137 22.771259 112.83701 10.516176 148.4504
## Oct 2013       61.13215 24.025251 115.74137 11.359905 151.8152
```

# Summary and Conclusion

Rainfall data for Caulfield (Racecourse) from 1950 to 2012 were analysed using a time series approach. The major purpose of this study was to better understand the underlying patterns, trends, and seasonal fluctuations in rainfall in order to create forecasting models that can effectively predict future rainfall. We investigated the data with a number of statistical techniques and models, including ARIMA and SARIMA.

As a preliminary step, we imported the data and performed data modification tasks, such as replacing missing numbers with the average monthly rainfall. We were able to identify patterns and trends in the data by visualizing it, such as changing variation, seasonality, and a slight decrease in annual rainfall averages.

Using the autocorrelation and partial autocorrelation functions, we discovered a substantial seasonal trend, but no significant autoregressive activity. We also ran normality tests and Box-Cox modifications on the data to improve its distributional properties.

The next stage was to configure various SARIMA models with varying ordering and seasonal components. Our models were selected using the Bayesian Information Criterion (BIC), ACF, and EACF. The significance of the coefficients in the selected models was assessed using maximum likelihood estimation (ML) and conditional sum of squares (CSS).

In this process, ARIMA and SARIMA models were evaluated, with the SARIMA(1,0,1)(0,1,1)12 model emerging as the best fit. Diagnostic checks verified that the model’s standardized residuals were white noise and approximately normally distributed, demonstrating its robustness. Furthermore, several error metrics (ME, RMSE, MAE, MPE, MAPE, and MASE) revealed the model’s high predictive performance. Overall, the SARIMA(2,0,2)x(0,1,1)_12 model accurately captures the underlying patterns and seasonal components of rainfall data, resulting in credible forecasts.

# References:

Australian Bureau Of Metereology.(n.d). Daily weather observations. Retrieved June 15, 2024, from http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=-1479855514&p_stn_num=086018