Marlie Naude

44038690

STTN326

Small Test 5. Research Assignment

Contents

[Introduction 2](#_Toc209267607)

[Reading the data and converting it 2](#_Toc209267608)

[Question 1 3](#_Toc209267609)

[Question 2 4](#_Toc209267610)

[Question 3 9](#_Toc209267611)

[Question 4 11](#_Toc209267612)

[Question 5 14](#_Toc209267613)

# Introduction

For this research assignment, we will be analysing temperatures from Potchefstroom. The data was obtained by PYTHON code, but the analysis will be done in R.

# Reading the data and converting it

#NaudeME44038690\_SmallTest5\_ResearchAssignment

rm(list=ls())

cat("\014")

setwd("C:\\Users\\marli\\OneDrive\\NWU\\Bsc Business Analytics\\\_3rd Year\\STTN326\\Assignment\\Naude\_44038690\_SmallTest5ResearchAssignment")

#libraries

install.packages("tseries")

install.packages("urca")

install.packages("forecast")

install.packages("lubridate")

library(tseries)

library(urca)

library(forecast)

library(ggplot2)

library(lubridate)

library(dplyr)

#Read in data

temp\_London<-read.csv("london\_temperature.csv")

head(temp\_London)

#Convert data into a time series

temps <- as.numeric(temp\_London$Temperature\_C)

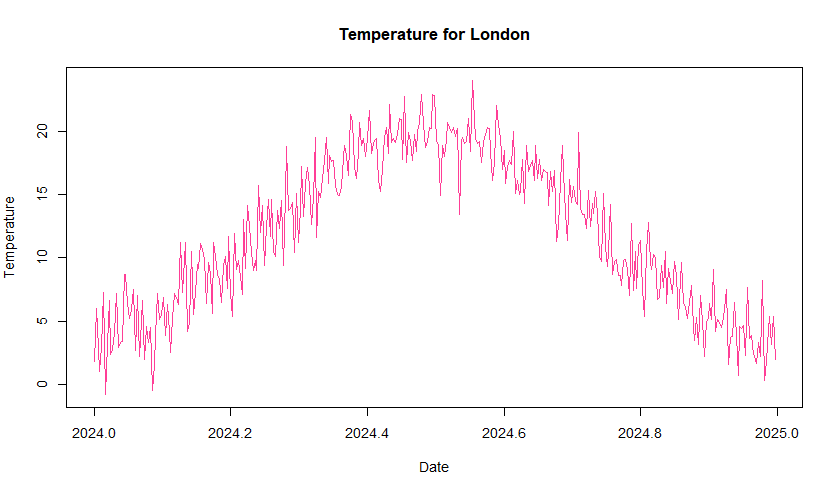
temps\_ts <- ts(temps, start = c(2024, 1), frequency = 365)

# Question 1

* 1. Plot the time series

#Question 1

plot(temps\_ts, main="Temperature for London", ylab="Temperature", xlab="Date", col="violetred1")



* 1. Comment on trends, seasonality, outliers, and potential non-stationarity.

We can see an upward trend till around May, then there is a downward trend in the graph. There is seasonality in the data; we can see that it is the hottest in the middle of the year, between April and June. There are some outliers, since there are spikes in the graph. There is a potential for non-stationarity, since the data makes a curve.

# Question 2

2.1 Test for stationarity using the augmented Dickey-Fuller (ADF) and comment on stationarity.

#Question 2

#2.1

(adf\_test <- adf.test(temps\_ts))

A computer screen shot of a computer code

AI-generated content may be incorrect.

H(o): The time series is non-stationary

Vs

H(a): the time series is stationary

, p-value> 0.05

Thus, we do not reject the null hypothesis and conclude that the data is non-stationary.

2.2 If the data is non-stationary

2.2.1 Plot the de-trended data and comment on stationarity

#2.2

#2.2.1

time<- 1:length(temps\_ts)

mylm<- lm(temps\_ts~time)

detrend<- resid(mylm)

plot(detrend,main="Detrended Temperature for London", ylab="Residuals", col="violetred2")

A graph of a temperature

AI-generated content may be incorrect.

The data is not stationary now, since there is still a pattern in the detrended data. It does not lie on the zero line.

2.2.2 Plot the differenced data and comment on stationarity.

#2.2.2

diff\_temp <- diff(temps\_ts)

plot(diff\_temp, main="Differenced Temperature", col= "violetred3")

A graph showing different types of temperature

AI-generated content may be incorrect.

Differencing did help to make the data stationary, since the data points lie around the 0-axis.

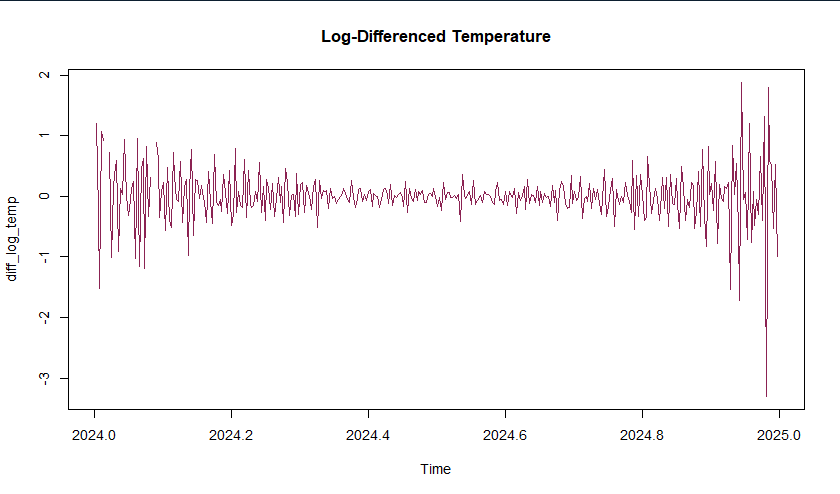
2.2.3 Plot the transformed data and comment on stationarity.

#2.2.3

log\_temp <- log(temps\_ts)

diff\_log\_temp <- diff(log\_temp)

plot(diff\_log\_temp, main="Log-Differenced Temperature",, col= "violetred4")



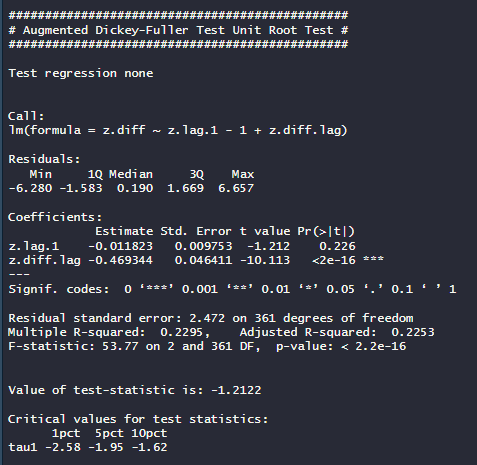
The data does lie around 0, but it goes out on the edges, thus not making it completely stationary.

2.3 Test whether the data follows a random walk.

#2.3

myrw<- ur.df(temps\_ts, type="none", selectlags = "AIC")

summary(myrw)



H(0): time series is nonstationary and follows a random walk

Vs

H(a): time series does not follow a random walk

P<0.05

Thus, we reject the null hypothesis and conclude that the series does not follow a random walk.

2.4 Test whether the data follows a random walk with drift.

#2.4

myrwd<- ur.df(temps\_ts, type="drift", selectlags = "AIC")

summary(myrwd)

A screenshot of a computer program

AI-generated content may be incorrect.

H(0): time series is nonstationary and follows a random walk with drift

Vs

H(a): time series does not follow a random walk with drift

P<0.05

Thus, we reject the null hypothesis and conclude that the series does not follow a random walk with drift.

# Question 3

3.1 Plot the 𝐴𝐶𝐹 and 𝑃𝐴𝐶𝐹.

#Question 3

#3.1

acf(temps\_ts, main="ACF of Temperature",col= "violetred")

pacf(temps\_ts, main="PACF of Temperature",col= "deeppink")

A graph of a temperature

AI-generated content may be incorrect. A graph with lines and numbers

AI-generated content may be incorrect.

3.2 Use the sample ACF and PACF for model identification, and justify the choice of the 𝐴𝑅𝐼𝑀𝐴 (𝑝, 𝑑, 𝑞) model.

The ACF tails of, and the PACF cuts of at lag 0.0125. This indicates that it is a moving average model. It achieved stationarity with differencing at 1 degree. Thus, it is an ARIMA(0,1,0.0125)

3.3 Apply the test for individual ACF as well as the portmanteau tests to test for autocorrelations.

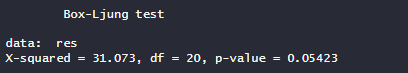
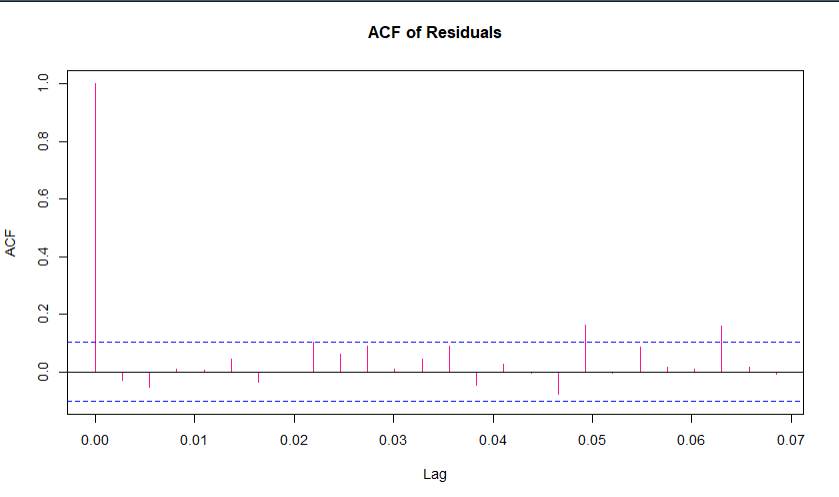
#3.3

fit <- auto.arima(temps\_ts)

res <- residuals(fit)

acf(res, main = "ACF of Residuals",col= "deeppink1")

Box.test(res, lag = 20, type = "Ljung-Box")#since it considered a small sample



We can only test for individual ACF graphically. We see that there is only a spike at lag 0, which indicates that there is only a small possibility for autocorrelation at lag 0

But the Box-Ljung test has a p-value >0.05. The following is the hypothesis:

H(0): Data does not experience autocorrelation

Vs

H(a): Data does experience autocorrelation

We thus do not reject the null hypothesis, and the data do not experience autocorrelation. But if we increase alpha to 0.1, then we reject the null hypothesis, and the data does experience autocorrelation.

# Question 4

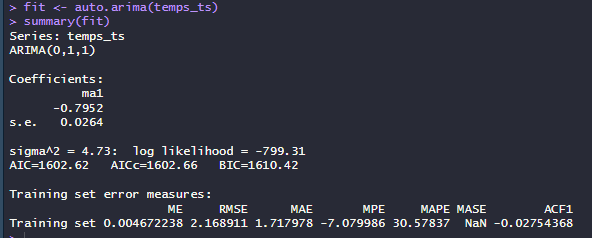
4.1 Estimate 𝐴𝑅𝐼𝑀𝐴 (𝑝, 𝑑, 𝑞) model parameters.

#Question 4

#4.1

fit <- auto.arima(temps\_ts)

summary(fit)



The output shows an ARIMA (0,1,1)

4.2 Compare the method of moments, the maximum likelihood method and the least squares method, estimation of 𝐴𝑅𝐼𝑀𝐴 (𝑝, 𝑑, 𝑞) model parameters

Method of Moments

Only applies to AR(p) models, since we are working with ARIMA (0,1,1), results may be inaccurate.

fit\_mom<- ar(temps\_ts, order.max=1, method="yw")

summary(fit\_mom)

fit\_mom$ar

A screenshot of a computer program

AI-generated content may be incorrect.

Maximum Likelihood

Most efficient method.

#MLE

fit\_mle <- auto.arima(temps\_ts, method = "ML")

summary(fit\_mle)

fit\_mle$coef

A screenshot of a computer

AI-generated content may be incorrect.

Least Squares method

A simpler, but less efficient method.

#LS

fit\_ls <- arima(temps\_ts, order = c(0,1,1), method = "CSS")

summary(fit\_ls)

fit\_ls$coef

A computer screen shot of a program

AI-generated content may be incorrect.

Comparison

The Maximum Likelihood and the Least Squares estimates gave similar results. The coefficient values are for LS, the ma1 was -0.7925577, and for the MLE, the ma1 was -0.7951573. These estimators are thus good to estimate the ARIMA (0,1,1)

As stated above, the Method of Moments can only estimate AR(p) models, which is why the coefficients differ a lot from the LS and MLE, with a value of 0.883535.

4.3 Examine residuals of the time plot, ACF of residuals, histogram / normal Q-Q plot, and explain their significance.

#4.3

qqnorm(res)

qqline(res)

A graph showing a line

AI-generated content may be incorrect.

The dots lie mostly on the line, thus indicating that the residuals are normally distributed.

# Question 5

5.1 Generate forecasts for 10–20 future time periods

#Question 5

# 5.1

fc <- forecast(fit, h = 20)

# Convert Date column to Date type

temp\_London$Date <- as.Date(temp\_London$Date)

# Generate next 20 daily dates after the last observation

future\_dates <- seq(from = max(temp\_London$Date) + 1,by = "day", length.out = 20)

(fc\_df <- data.frame(Date= future\_dates,Forecast = as.numeric(fc$mean),Lo95= fc$lower[,2],Hi95= fc$upper[,2]))

A screenshot of a computer

AI-generated content may be incorrect.

5.2 Plot forecast with 95% confidence intervals.

# 5.2

ggplot() +

# Historical temps

geom\_line(data = temp\_London, aes(x = Date, y = Temperature\_C),color = "deeppink2") +

# Forecast line

geom\_line(data = fc\_df, aes(x = Date, y = Forecast),color = "black") +

geom\_ribbon(data = fc\_df, aes(x = Date, ymin = Lo95, ymax = Hi95),fill = "lightblue", alpha = 0.4) +

labs(title = "London Temperature Forecast (next 20 days)",x = "Date", y = "Temperature (°C)")

A graph of a graph

AI-generated content may be incorrect.

5.3 Comment on the practical meaning of the forecast in the context of the chosen data.

#5.3

mean(fc\_df$Lo95)

mean(fc\_df$Hi95)

A computer screen shot of numbers

AI-generated content may be incorrect.

The ARIMA model gives a 20-day forecast of daily average temperatures in London for the start of 2025

The black line of the plot represents the model's best estimate of expected daily temperature.

The light blue shaded 95% confidence interval reflects the uncertainty. We notice that the further we look into the future, the wider the band gets, since there is more uncertainty. This captures a range of plausible outcomes.

In the context of the weather in London will remain around -1.548429 and 8.495649, which is consistent with typical winter weather in London.

This information can be useful for any activities that are influenced by weather conditions, for example: planning energy demand, transportation, and even agricultural activities.