**Group 2: Seasonal Time Series Analysis**

**MA 641 – Summer 2024**

## Introduction

Using a decade's worth of daily minimum temperature data from Melbourne, Australia, we've built predictive seasonal models to forecast the next points in a time series. This practical application aims to enhance weather predictions, aiding in resource planning, energy management, and climate adaptation.

## Data Description

Date range – Daily data collected from 1/1/1981 to 12/31/1990.

Raw data description – The dataset records ten years of minimum daily temperature values in Melbourne, Australia from January 1981 to December 1990. The dataset was imported from Kaggle. The data is structured in 3652 rows and 2 columns.

*Date* – Each date the minimum temperature was measured from 1981-1990. Date format was changed from (MM/DD/YYYY) to (YYYY-MM-DD) for easier formatting and usability.

*Temperature* – The recorded minimum temperature listed for each date listed in the dataset. Recorded in Celsius. Range of temperatures: 0 to 26.3 degrees Celsius.

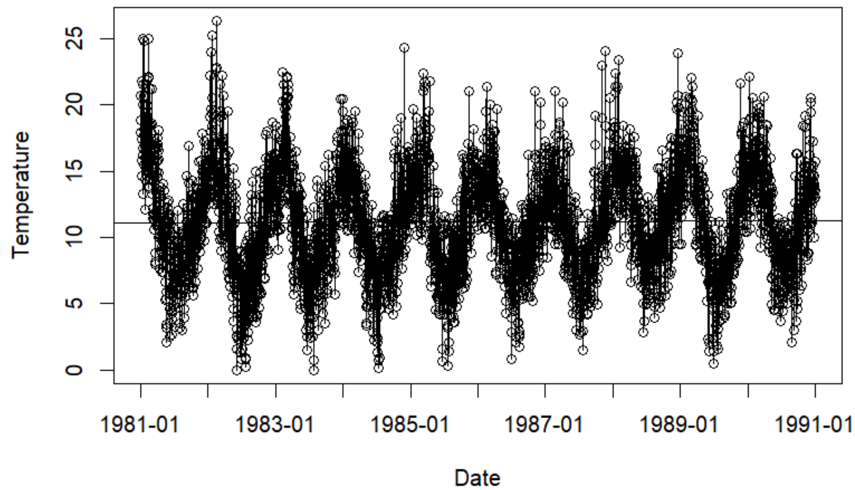
Link to data: [Daily Minimum Temperatures in Melbourne (kaggle.com)]

## Scope & Focus

The goal of this analysis is to use time series forecasting to predict the minimum temperature in Melbourne for the next 500 days within an 80% and 95% confidence interval. We will perform this time series forecasting using ACF and PACF plots to determine degrees of differencing needed to make forecasts. To make the final forecasts, we will be using seasonal autoregressive (AR) and moving average (MA) models. These models will be evaluated using maximum likelihood estimation and the Box-Ljung test.

**Initial Analysis**

We start with some initial exploratory analysis:




From plotting the raw time series data, we see there is a clear trend present. The highest minimum temperatures occur regularly during or around the month of January and lowest minimum temperatures occur regularly during or around the month of August. This leads us to interpret a likely seasonal pattern in the data, which would follow our intuition.

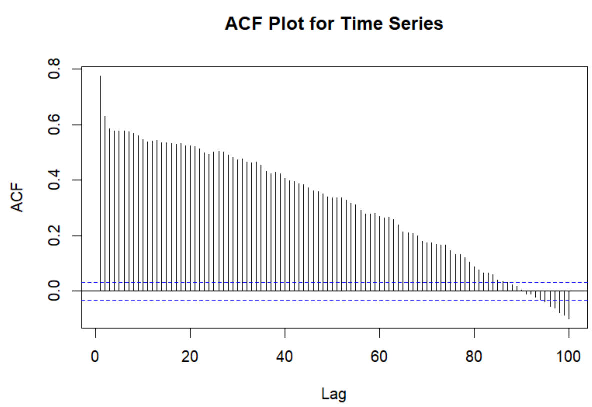
A linear model was also fitted to the data to produce a line-of-best-fit. The linear model produced a line that exactly bisects the raw time series. This visually suggests that the mean and variance are constant over time. We now perform proper stationarity testing to prepare the time series for forecasting.

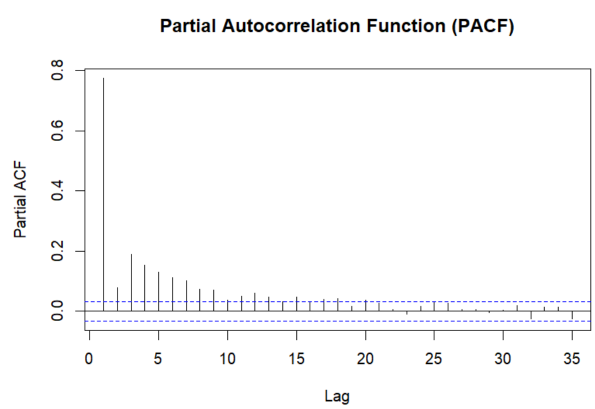
## Stationarity Testing

```
        Augmented Dickey-Fuller Test

data:  ts
Dickey-Fuller = -5.239, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary

        KPSS Test for Level Stationarity

data:  ts
KPSS Level = 0.17729, Truncation lag parameter = 9, p-value = 0.1

        KPSS Test for Trend Stationarity

data:  ts
KPSS Trend = 0.15625, Truncation lag parameter = 9, p-value = 0.04146
```

From these tests, the time series appears to be stationary. We will check the ACF and PACF plots to visualize for the dependence present in autocorrelation.

## ACF and PACF for Unadjusted Time Series

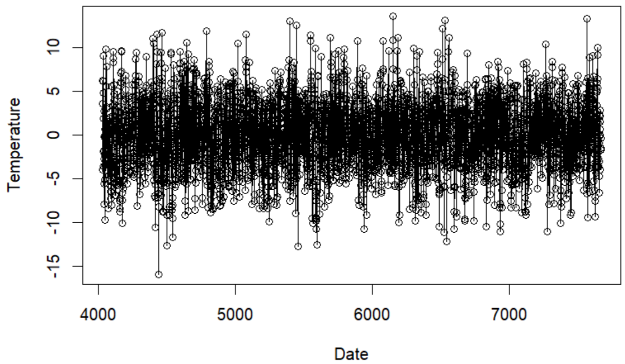


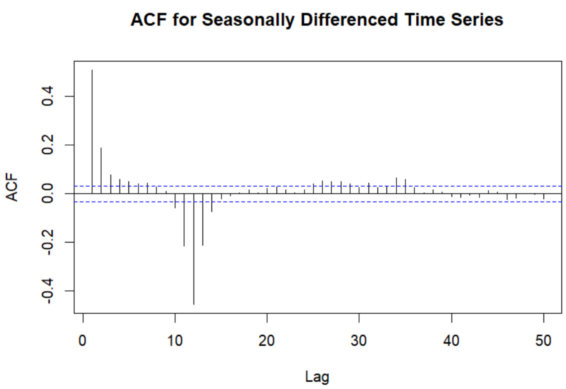



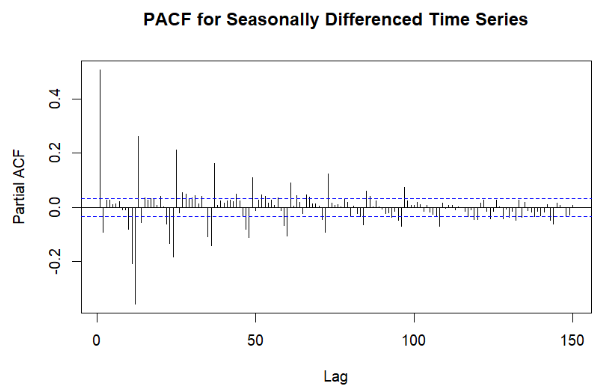

The ACF and PACF plots evidently indicate that there are more dependencies occurring within the time series. Due to the seasonal pattern observed in the raw data, our first method for addressing the dependencies in the time series is to use seasonal differencing.

## Seasonal Differencing

We took the seasonal difference at s = 12 then plotted the differenced time series along with its ACF and PACF plots:




From the plot of the differenced time series, we can determine the seasonal trend from the raw data has been normalized and the time series appears to be stationary. From the ACF plot we can see one clear season for the seasonal MA parameter. The ACF also illustrates significant values at lags 1 and 2 and possibly significant at lag 3. These lags will be used for the non-seasonal MA parameter.

The PACF plot also shows a clear pattern. After extended out to 150 lags, we can see that the pattern repeats 9 times leading us to interpret 9 seasons for the seasonal AR parameter. We also interpret 2 lags as significant for the non-seasonal AR parameter.

## Parameter Redundancy

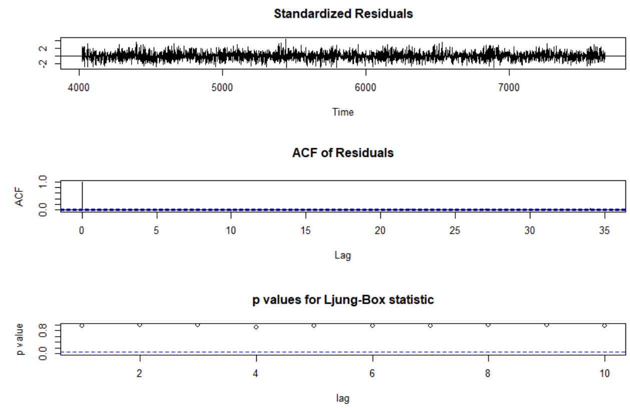
**SARMA(1, 0, 3) x (1, 1, 1)$_{s=12}$**

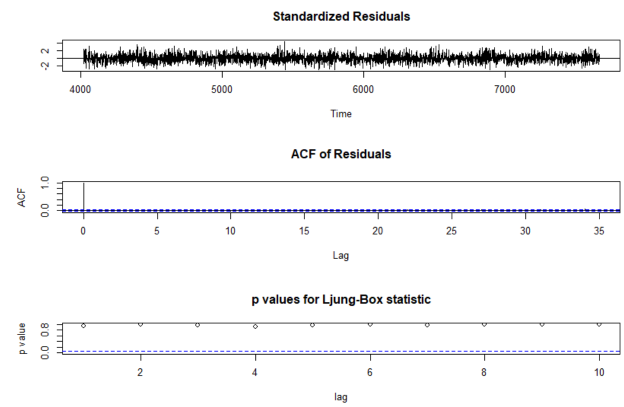
**SARMA(2, 0, 3) x (1, 1, 1)$_{s=12}$**

**SARMA(2, 0, 2) x (1, 1, 1)$_{s=12}$**

## Residual Diagnostics

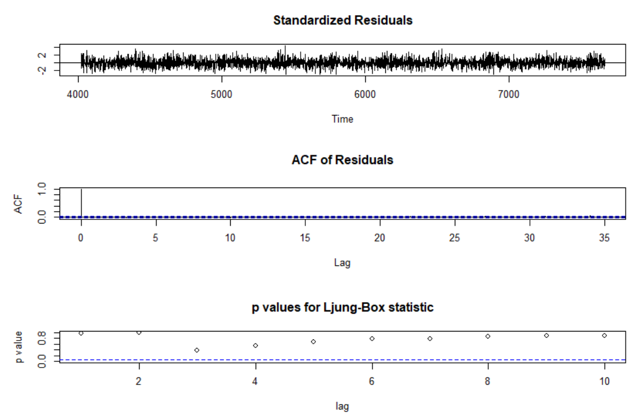
```
> fit52
Series: ts
ARIMA(1,0,3) with non-zero mean

Coefficients:
         ar1      ma1      ma2      ma3     mean
      0.9925  -0.4004  -0.3290  -0.1044  11.1836
s.e.  0.0023   0.0166   0.0171   0.0164   0.8517

sigma^2 = 5.789:  log likelihood = -8386.81
AIC=16785.61   AICc=16785.63   BIC=16822.83
```




```
> fit49
Series: ts
ARIMA(2,0,3) with non-zero mean

Coefficients:
         ar1      ar2      ma1      ma2      ma3     mean
      1.0280  -0.0353  -0.4355  -0.3135  -0.0911  11.1950
s.e.  0.1389   0.1377   0.1381   0.0625   0.0548   0.8461

sigma^2 = 5.791:  log likelihood = -8386.77
AIC=16787.54   AICc=16787.57   BIC=16830.96
```

```
> fit50
Series: ts
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1      ar2      ma1      ma2     mean
      1.2364  -0.2420  -0.6454  -0.2307  11.1832
s.e.  0.0377   0.0372   0.0370   0.0299   0.8617

sigma^2 = 5.793:  log likelihood = -8388.04
AIC=16788.08   AICc=16788.1   BIC=16825.3
```




From the Box-Ljung test on each model selected, we can see there is no significant autocorrelation present in any of the residual sets.

By looking at each model's BIC and following the principle of parsimony, we can interpret that SARMA(1, 0, 3) x $(1, 1, 1)_{s=12}$ would make the most sense to use for forecasting the next 500 values with an 80% and 95% confidence interval.

**Forecasting**