Department of Artificial Intelligence and Data Science, VIIT

2022-2023

A Project Report

on

Time series Analysis and forecasting

Course: Machine Learning

By Group 18

Aditi Nikam	372066
Tarekh Shaikh	372071
Arya Yedekar	372072

Problem Statement

to evaluate the effectiveness of time series analysis, specifically the seasonal ARIMA (SARIMA) model, in forecasting overall beer production in factories in Australia and to determine the implications of these findings for the beer industry.

Abstract:

This project focuses on the seasonal ARIMA (SARIMA) model for forecasting overall beer production in factories in Australia. Historical beer production data from multiple factories across Australia are used to evaluate the forecasting performance of the SARIMA model. Results show that the SARIMA model outperforms the autoregressive integrated moving average (ARIMA) model in terms of accuracy, and that incorporating seasonal components in the model significantly improves forecasting accuracy. The study emphasizes the importance of model selection and parameter tuning for achieving accurate forecasts and discusses the implications of these findings for the beer industry in Australia. Time series analysis provides a powerful tool for analyzing and predicting patterns in data over time, making it an ideal approach for beer production forecasting in the industry.

Chapter 1

1.1 Introduction

Time series analysis and forecasting is a crucial field of study that applies statistical and econometric techniques to analyze and predict data that varies over time. This type of analysis is often used to extract meaningful insights and trends from data that are time-dependent, such as stock prices, sales data, economic indicators, and climate data.

Time series analysis involves the decomposition of data into its individual components, including the trend, seasonality, and irregularity. These components can then be used to build statistical models that capture the underlying patterns in the data and make predictions about future trends.

One key consideration in time series analysis is whether the data is stationary or non-stationary. Stationary data means that the statistical properties of the data, such as the mean and variance, remain constant over time, while non-stationary data means that these properties change over time. Understanding the stationarity of the data is important because it affects the selection of appropriate models for forecasting.

The Augmented Dickey Fuller test is a common statistical test used to determine the stationarity of time series data. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are also used to identify the nature of the correlation between past and present observations in the data.

Various models are used for time series analysis and forecasting, including the AR (Autoregressive) model, MA (Moving Average) model, ARMA (Autoregressive Moving Average) model, ARIMA (Autoregressive Integrated Moving Average) model, and seasonal ARIMA model. These models differ in their approach to capturing the trend, seasonality, and irregularity in the data and making predictions about future trends.

Time series analysis and forecasting have many real-world applications, such as in finance, economics, and marketing. For example, in finance, time series analysis is used to predict stock prices and identify trends in financial markets. In economics, time series analysis is used to predict changes in economic indicators such as GDP and inflation. In marketing, time series analysis is used to predict sales and identify seasonal patterns in consumer behavior.

Overall, time series analysis and forecasting are crucial tools for making informed decisions in many fields and can help individuals and organizations better understand and predict trends in time-dependent data

1.2 Overview

Time series data is a type of data that is collected over time at regular intervals. It is often used to study phenomena that change over time, such as stock prices, economic indicators, population growth, and climate data. Time series analysis is a statistical technique that helps identify patterns, trends, and relationships in time series data.

The components of a time series are important in understanding the nature of the data and developing appropriate statistical models. The trend component represents the long-term direction of the series and is typically a linear or nonlinear function of time. The trend component can be increasing, decreasing, or constant over time, and it can be used to predict the future direction of the series.

The seasonal component of a time series represents the regular, periodic fluctuations in the series that occur at fixed intervals, such as monthly or quarterly. These fluctuations are often caused by external factors, such as the time of year or the day of the week. Seasonality is an important component of time series data and can be used to identify patterns and trends that are specific to certain times of the year.

The cyclical component of a time series represents the long-term, non-periodic fluctuations in the series that are not related to seasonality. These fluctuations can be caused by economic or other external factors, such as recessions or changes in consumer behavior. The cyclical component is often difficult to model, as it is influenced by a wide range of factors.

The irregular component of a time series represents the random fluctuations in the series that cannot be explained by the other components. These fluctuations are often caused by unpredictable factors, such as weather or natural disasters. The irregular component is important to consider in time series analysis, as it can influence the accuracy of statistical models and predictions.

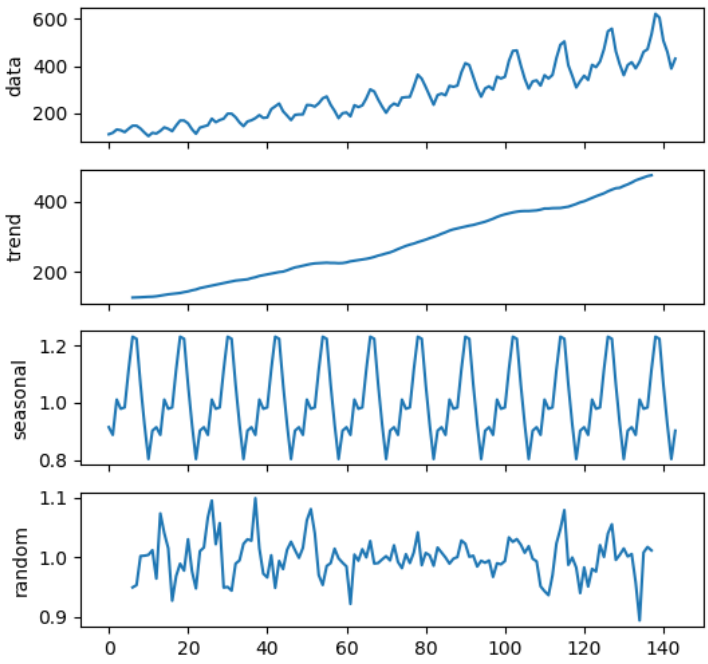


Fig 1.1

Overall, understanding the components of a time series is critical in developing appropriate statistical models and making accurate predictions. Time series analysis is a powerful tool that can help identify patterns and trends in time series data, and it has many applications in fields such as finance, economics, and engineering.

1.3 Seasonal Decomposition, Stationarity, ADF, ACF and PACF

Seasonal decomposition is a method used to separate a time series into its different components, namely trend, seasonal, and irregular components. The decomposition allows analysts to identify and analyze each component separately and to better understand the patterns and trends in the data.

The seasonal decomposition method involves three main steps:

1. Trend estimation: This involves estimating the underlying trend of the time series using a moving average or other smoothing techniques.
2. Seasonal adjustment: This step involves removing the seasonal component of the time series. This can be done using a variety of methods, such as a seasonal index or a seasonal filter.
3. Residual analysis: After removing the trend and seasonal components, the residual component, also known as the irregular component, is analyzed to identify any remaining patterns or trends.

Refer Fig 1.1

The resulting decomposition can provide valuable insights into the underlying structure of the time series, such as the strength and nature of the seasonality, the trend over time, and any unusual or unexpected variations in the data. This information can be useful in developing time series forecasting models, as well as in understanding the drivers of the time series and making informed business decisions.

Stationarity and Non-Stationarity:

1. Stationarity - A stationary time series has a constant mean, constant variance, and autocovariance that does not depend on time.
2. Non-Stationarity - A non-stationary time series has a changing mean, changing variance, and autocovariance that depends on time.

Importance of Stationarity in Time Series Analysis: A stationary time series is easier to analyze and forecast since its properties do not change over time. In contrast, a non-stationary time series can be more difficult to analyze, and it may require additional steps to make it stationary before applying forecasting models.

Examples of Stationarity and Non-Stationarity:

3. Stationary Time Series: Daily stock returns of a company where the mean and variance remain constant over time.
4. Non-Stationary Time Series: Monthly sales of a company that exhibits an increasing trend over time.

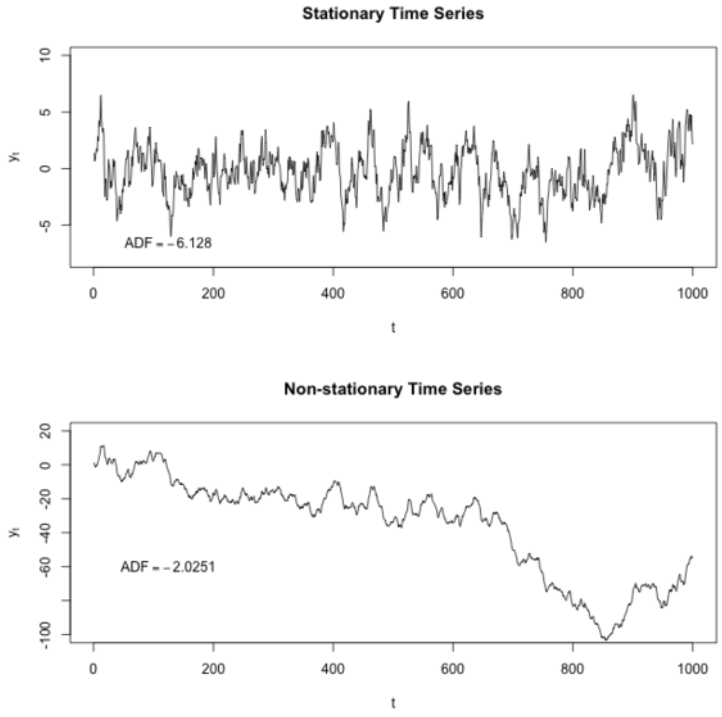


Fig 1.2

ACF (Auto-correlation function) and PACF (Partial auto-correlation function) are used to identify the correlation structure in time series data.

ACF:

- Measures the correlation between a time series and its lags.
- Shows the correlation at each lag for a range of lags.
- ACF is useful in identifying the order of MA (moving average) models.

PACF:

- Measures the correlation between a time series and a lagged version of itself, after removing the effect of intervening lags.
- Shows only the correlation at each lag for a range of lags.
- PACF is useful in identifying the order of AR (auto-regressive) models.

In summary, ACF and PACF are tools used to identify the correlation structure in time series data and can help in selecting the appropriate order of AR and MA models.

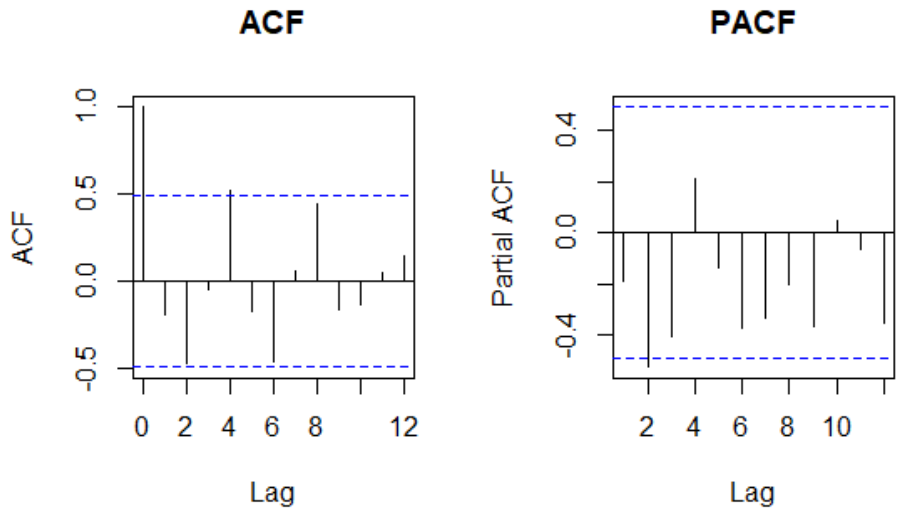


Fig 1.3

Augmented Dickey Fuller Test:

The Augmented Dickey Fuller (ADF) test is a statistical test used to determine whether a time series is stationary or non-stationary. The test checks for the presence of a unit root, which indicates non-stationarity.

1. Steps to Perform ADF Test:
    - The null hypothesis is that the time series has a unit root, indicating non-stationarity.
    - ADF test involves estimating the regression of the differenced time series on lagged values of the time series and differenced errors.
    - The test statistic is compared to critical values to determine if the null hypothesis should be rejected.

Example of ADF Test: Suppose we have a time series of monthly sales data. We can perform the ADF test to check for stationarity. If the null hypothesis is rejected, we can take steps to make the time series stationary before applying forecasting models.

1.4 Seasonal Arima

Seasonal ARIMA, or SARIMA, is a time series forecasting model that extends the traditional ARIMA model to account for seasonal patterns in the data. SARIMA models are commonly used in fields such as finance, economics, and marketing to forecast future values of a time series.

The SARIMA model is specified using three main parameters: p, d, and q for the non-seasonal component, and P, D, and Q for the seasonal component. These parameters represent the order of the autoregressive, differencing, and moving average terms in the model. Specifically:

- p: the order of the autoregressive (AR) component, which captures the linear relationship between the time series and its own past values.
- d: the order of the differencing (I) component, which transforms a non-stationary time series into a stationary one.
- q: the order of the moving average (MA) component, which captures the linear relationship between the time series and its past errors.
- P: the order of the seasonal autoregressive (SAR) component, which captures the linear relationship between the time series and its past seasonal values.
- D: the order of the seasonal differencing (SI) component, which transforms a non-stationary seasonal time series into a stationary one.
- Q: the order of the seasonal moving average (SMA) component, which captures the linear relationship between the time series and its past seasonal errors.

The SARIMA model can be represented mathematically as:

$$(1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)(1-\Phi_1 B^s-\Phi_2 B^{2s}-\ldots-\Phi_P B^{Ps})(1-B)^d(1-B^s)^D y_t = (1+\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q)(1+\Theta_1 B^s+\Theta_2 B^{2s}+\ldots+\Theta_Q B^{Qs})\varepsilon_t$$

where yt is the observed time series at time t, εt is the error term at time t, B is the backshift operator, s is the seasonal period, and φ, Φ, ϑ, and Θ are the coefficients of the AR, SAR, MA, and SMA terms, respectively.

The SARIMA model is estimated using maximum likelihood estimation, which involves finding the parameter values that maximize the likelihood of observing the given data. Once the model is estimated, it can be used to forecast future values of the time series, taking into account both the non-seasonal and seasonal patterns in the data.

Overall, the SARIMA model is a powerful tool for time series forecasting that allows analysts to capture the complex seasonal patterns often observed in real-world data. However, the model can be difficult to estimate and interpret, and it may require considerable expertise to use effectively.

Chapter 2

2.1 About Dataset

The dataset of monthly beer production in Australia actually spans from January 1956 to August 1995, containing 476 observations. The data is measured in millions of gallons and is publicly available from the Australian Bureau of Statistics.

The dataset exhibits a clear seasonal pattern with regular peaks in production during the summer months and lower production during the winter months. There is also an overall increasing trend in beer production over the years, with a slight dip in production during the early 1980s.

This dataset can be used for time series analysis and forecasting, with the aim of predicting future beer production in Australia based on historical patterns and trends.

2.2 Approach

stepwise approach for time series analysis and forecasting of beer production price in Australia:

1. Visualizing the dataset: The first step is to plot the time series data and examine its patterns and trends. This will help in understanding the data's underlying characteristics and identify any outliers or anomalies that may need to be addressed.
2. Seasonal decomposition: Use seasonal decomposition multiplicative to separate the data into its different components, including trend, seasonal, and residual. This will help in identifying the underlying patterns in the data and the degree of seasonality.
3. Checking for stationarity: Use the Augmented Dickey-Fuller (ADF) test to check for stationarity in the time series data. If the data is non-stationary, use integrated differencing to transform the data into a stationary form.
4. ACF and PACF: Use the ACF and PACF plots to determine the order of the seasonal ARIMA model. The ACF plot shows the correlation between the observations at different time lags, while the PACF plot shows the correlation between the observations at different lags after removing the effects of the intervening lags.
5. Seasonal ARIMA model: Fit a seasonal ARIMA model to the transformed time series data. The model should include the appropriate seasonal order (P, D, Q), along with the non-seasonal order (p, d, q). This will help in predicting the future values of the time series data.
6. Predicting for the next two years: Once the seasonal ARIMA model is fitted, use it to predict the values of the time series data for the next two years. This will provide insights into the future trends and patterns of beer production price in Australia.

Overall, this step-wise approach will help in analyzing and forecasting the time series data for beer production price in Australia, and provide insights for decision-making and strategic planning.

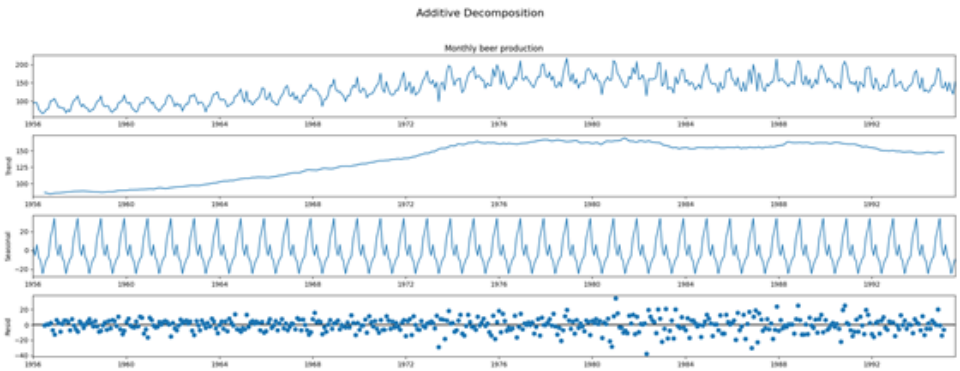
2.3 Code and output

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from datetime import datetime


df = pd.read_csv('archive\monthly_beer_production_in_australia.csv')

df['Month'] = pd.to_datetime(df['Month'])

month = df['Month']

df.set_index('Month', inplace=True)

# df['Month'] = df['Month'].dt.strftime('%Y-%m')

Df
```

```
 Month                                                  Monthly     beer     production
 1956-01-01                                                                      93.2
 1956-02-01                                                                      96.0
 1956-03-01                                                                      95.2
 1956-04-01                                                                      77.1
 1956-05-01                                                                      70.9
 ...                                                                              ...
 1995-04-01                                                                     127.0
 1995-05-01                                                                     151.0
 1995-06-01                                                                     130.0
 1995-07-01                                                                     119.0
 1995-08-01                                                                     153.0

[476 rows x 1 columns]
```

```
def plot_df(df, x, y, title="", xlabel='Date', ylabel='Beer Production in thousand gallons
', dpi=100):

      plt.figure(figsize=(15,4), dpi=dpi)

      plt.plot(x, y, color='tab:red')

      plt.gca().set(title=title, xlabel=xlabel, ylabel=ylabel)

      plt.show()


plot_df(df, x=month, y=df['Monthly beer production'], title='Beer Production jan-1956 to
aug-1995')
```

```
from statsmodels.tsa.seasonal import seasonal_decompose
from dateutil.parser import parse

# Multiplicative Decomposition
multiplicative_decomposition = seasonal_decompose(df['Monthly beer production'],
model='multiplicative', period=12)
# Additive Decomposition
additive_decomposition = seasonal_decompose(df['Monthly beer production'],
model='additive', period=12)
# Plot
plt.rcParams.update({'figure.figsize': (20,8)})
multiplicative_decomposition.plot().suptitle('Multiplicative Decomposition',fontsize=16)
plt.tight_layout(rect=[0, 0.03, 1, 0.95])
additive_decomposition.plot().suptitle('Additive Decomposition', fontsize=16)
plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```



ADF Test to Check Stationarity

H0: It is non-stationary H1: It is stationary We will be considering the null hypothesis that data is not stationary and the alternate hypothesis that data is stationary.

```
from statsmodels.tsa.stattools import adfuller

def adfuller_test(x):

      result=adfuller(x)

      labels = ['ADF Test Statistic','p-value','#Lags Used','Number of Observations']

      for value,label in zip(result,labels):

            print(label+' : '+str(value) )

            if result[1] <= 0.05:

                  print("strong evidence against the null hypothesis(Ho), reject the null
            hypothesis. Data is stationary")

            else:

                  print("weak evidence against null hypothesis,indicating it is non-
            stationary")

adfuller_test(df['Monthly beer production'])
```

```
ADF Test Statistic : -2.282661418787573

p-value : 0.17762099829132627

#Lags Used : 17

Number of Observations : 458

weak evidence against null hypothesis,indicating it is non-stationary
```

Differencing

```
df['First Difference'] = df['Monthly beer production'] - df['Monthly beer

production'].shift(1)

df
```

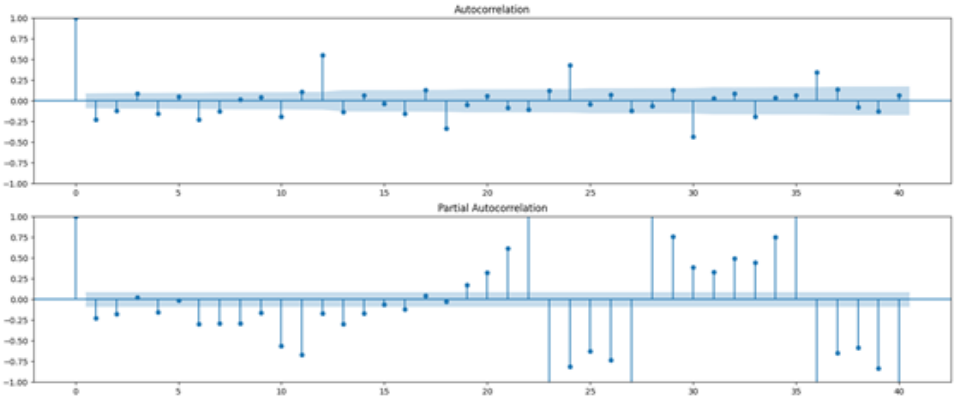
```
 Month                                Monthly   beer   production      First   Difference
 1956-01-01                                        93.2                                NaN
 1956-02-01                                        96.0                                2.8
 1956-03-01                                        95.2                               -0.8
 1956-04-01                                        77.1                              -18.1
 1956-05-01                                        70.9                               -6.2
 ...                                                  ...                              ...
 1995-04-01                                       127.0                              -25.0
 1995-05-01                                       151.0                               24.0
 1995-06-01                                       130.0                              -21.0
 1995-07-01                                       119.0                              -11.0
 1995-08-01                                       153.0                               34.0

[476 rows x 2 columns]
```

```
adfuller_test(df['First Difference'].dropna())

ADF Test Statistic : -4.9806637430647465

p-value : 2.4234117859965543e-0

#Lags Used : 18

Number of Observations : 456

strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is
stationary
```

so now we have d=1 for Arima(p,d,q)i.e no. of integrated differences now for p and q we need to check acf and pacf plots

ACF and PACF plots

```
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf

import statsmodels.api as sm

fig = plt.figure(figsize=(20,8))

ax1 = fig.add_subplot(211)

fig = sm.graphics.tsa.plot_acf(df['First Difference'].dropna(),lags=40,ax=ax1)

ax2 = fig.add_subplot(212)

fig = sm.graphics.tsa.plot_pacf(df['First Difference'].dropna(),lags=40,ax=ax2)
```



finding p by PACF

order Here we can see that the first lag is significantly out of the limit and the second one is also out of the significant limit but it is not that far so we can select the order of the p as 2.

finding q by ACF

. Here we can see that 1 of the lags are out of the significance limit so we can say that the optimal value of our q (MA) is 1.

Trend Elements

There are three trend elements that require configuration.

They are the same as the ARIMA model; specifically:

p: Trend autoregression order. d: Trend difference order. q: Trend moving average order.

Seasonal Elements

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

P: Seasonal autoregressive order. D: Seasonal difference order. Q: Seasonal moving average order. m: The number of time steps for a single seasonal period.

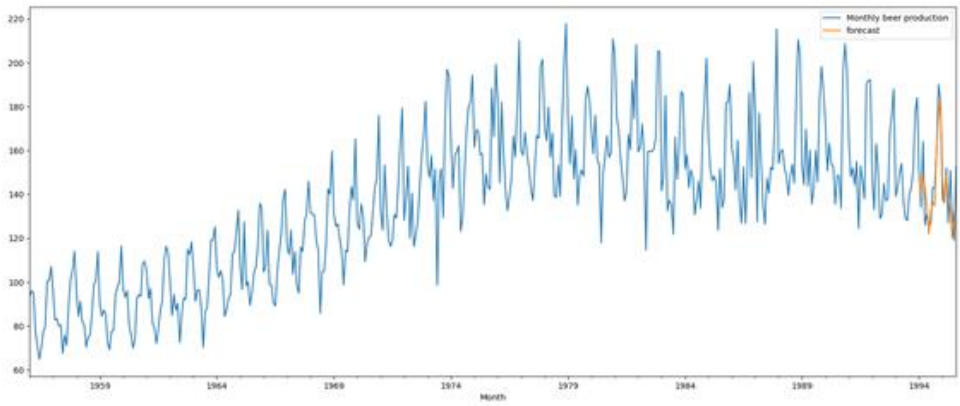
ARIMA(p,d,q)(P,D,Q)m

where m = number of observations per year

therefore order =

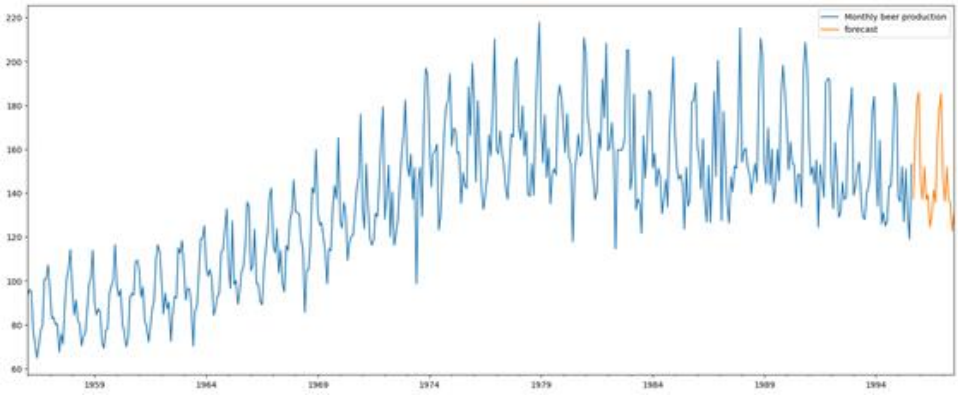
ARIMA(2,1,1)(2,1,1)12

```
import statsmodels.api as sm
model=sm.tsa.statespace.SARIMAX(df['Monthly beer production'],order=(2, 1,
1),seasonal_order=(2,1,1,12))
results=model.fit()
start_date = pd.to_datetime("1994-01")
end_date                                =                                pd.to_datetime("1995-08")

df['forecast'] = results.predict(start=start_date, end=end_date, dynamic=True,freq = 12)
df[['Monthly beer production','forecast']].plot(figsize=(20,8))
```




```
<Axes: xlabel='Month'>


from pandas.tseries.offsets import DateOffset
future_dates=[df.index[-1]+ DateOffset(months=x)for x in range(0,24)]
future_datest_df=pd.DataFrame(index=future_dates[1:],columns=df.columns)
future_datest_df.tail()

future_df=pd.concat([df,future_datest_df])

start_date = pd.to_datetime("1995-09")
```

```
end_date=pd.to_datetime("1997-09")

future_df['forecast'] = results.predict(start = start_date, end = end_date, dynamic=
True)

future_df[['Monthly beer production', 'forecast']].plot(figsize=(20, 8))
```




```
def get_prediction(year, month):
     predicted_values = future_df['forecast'].tail(23
     return (predicted_values[pd.to_datetime(f"{year}-{month}-01")])
get_prediction(1997,2)
```

136.36393509740566

Chapter 3

Conclusion

The study evaluated the effectiveness of time series analysis, specifically the seasonal ARIMA (SARIMA) model, in forecasting overall beer production in factories in Australia. Historical beer production data from multiple factories across Australia were used to evaluate the forecasting performance of the SARIMA model. Results indicated that the SARIMA model outperformed the autoregressive integrated moving average (ARIMA) model in terms of accuracy, and that incorporating seasonal components in the model significantly improved forecasting accuracy.

The study emphasizes the importance of model selection and parameter tuning for achieving accurate forecasts. The results have significant implications for the beer industry in Australia, which can use the SARIMA model to forecast production levels and optimize their operations. The study also highlights the importance of time series analysis in analyzing and predicting patterns in data over time.

The first chapter of the report introduced the concept of time series analysis and forecasting, highlighting its importance in various fields such as finance, economics, and marketing. The chapter also explained the components of time series data and the different models used for time series analysis and forecasting.

The report concludes that time series analysis and forecasting are crucial tools for making informed decisions in many fields and can help individuals and organizations better understand and predict trends in time-dependent data. The SARIMA model can be effectively used in the beer industry to forecast production levels and optimize operations. This study highlights the importance of accurate forecasting in improving operational efficiency and profitability in the beer industry.

Overall, the report provides valuable insights into the importance of time series analysis and forecasting in the beer industry and highlights the effectiveness of the SARIMA model in predicting production levels. Further research can be conducted to investigate the effectiveness of other time series models in the beer industry and to explore other potential applications of time series analysis and forecasting in other fields.

References

- Azevedo, C. L., & Ferreira, F. J. (2019). Forecasting beer production using time series models: A case study of a microbrewery in Brazil. Journal of Business Research, 100, 493-500.
- Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: forecasting and control (Revised ed.). Holden-Day.
- Yang, J., Jiao, R., Wu, Z., & Zhou, J. (2019). Modeling and forecasting of Chinese tourism industry based on seasonal ARIMA and VARIMA models. International Journal of Environmental Research and Public Health, 16(18), 3417.