A CASE STUDY ON CONSUMER PRICE INDEX OF INDIA

Project Report submitted in partial fulfilment of the requirements
for the award of the degree of

INTEGRATED MASTER OF SCIENCE

in

STATISTICS

by

SREEVINAYAK K P

(Roll No: 35219052)

DEPARTMENT OF STATISTICS

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

KOCHI , 682022, INDIA

April, 2024

Certificate

This is to certify that the Project entitled: **A CASE STUDY ON CONSUMER PRICE INDEX OF INDIA** , submitted by **SREEVINAYAK K P** to the Cochin University of Science And Technology, for the award of the degree of **Integrated Master of Science in Statistics** is a record of the original, bonafide research work carried out by him under our supervision and guidance. The work has reached the standards fulfilling the requirements of the regulations related to the award of the degree.

The results contained in this project report have not been submitted in part or in full to any other University or Institute for the award of any degree or diploma to the best of our knowledge.

....
Dr. S.M.SUNOJ
HEAD OF DEPARTMENT
PROFESSOR
DEPARTMENT OF STATISTICS
CUSAT.

....
Dr. ASHA GOPALAKRISHNAN
SUPERVISING GUIDE
SENIOR PROFESSOR
DEPARTMENT OF STATISTICS
CUSAT.

DECLARATION

I declare that this written submission represents my ideas in my own words. Where others' ideas and words have been included, I have adequately cited and referenced the original source. I declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated, or falsified any idea/data/-fact/source in my submission. I understand that any violation of the above will cause disciplinary action by the Institute and can also evoke penal action from the source which has thus not been properly cited or from whom proper permission has not been taken when needed.

. . . .

SREEVINAYAK K P
Roll No.: 35219052
Date: APRIL 29
Place: CUSAT

Acknowledgements

I take this opportunity to express my profound gratitude and deep regards to my supervisor Dr. Asha Gopalakrishnan, Senior Professor, Cochin University of Science and Technology, for hers valuable information and guidance, constant encouragement, patience, immense knowledge and affectionate attitude without which I would not have been able to complete this project satisfactory.

I am indebted to my respected teachers, friends and non teaching staffs of Department of Statistics and Centre for Integrated Studies, CUSAT for their steadfast encouragement and time. I express my immense gratitude to my family who were always supporting throughout my progress. I am grateful to the Cochin University of Science and Technology for providing the necessary resources and facilities to complete this project to the best of my ability.

SREEVINAYAK K P

Contents

Certificate

Declaration

Acknowledgements

Contents

1 Introduction **2**

1.1 Overview 2

1.2 Objectives 3

1.3 Software Tool Used 3

2 Data Description And Data Source **4**

2.1 Data Source 4

2.2 Data Description 4

3 Methodology **9**

3.1 Time Series Analysis 9

3.1.1 Stationary 10

3.1.2 Non Stationary 10

3.1.3 Components of Time Series 11

3.1.4 Partial Autocorrelation Function 12

3.1.5 Autocorrelation Function 12

3.1.6 Autoregressive Model 13

3.1.7 Moving Average Model 14

3.1.8 Autoregressive Moving Average Model 15

3.1.9 Autoregressive Integrated Moving Average Model 16

3.1.10 Seasonal Autoregressive Integrated Moving Average Model . . 16

3.1.11 Model Selection Criteria 17

3.1.12 Unit Root Test 18

3.1.13 Augmented Dicky-Fuller Test 19

3.1.14 Ljung-Box Test 19

3.1.15 Residual Analysis 20
3.1.16 Forecast 20
3.2 Regression Analysis 21
3.2.1 Multiple Linear Regression 21
3.2.2 Assumptions of MLR 22
3.2.3 Least Square Estimation of Regression Coefficients 22
3.2.4 Coefficient of Determination 24
3.2.5 Testing the Significance of Regression 25
3.2.6 Shapiro-Wilk normality test 26
3.2.7 Breusch-Pagan Non Constant Variance Test 27
3.2.8 Durbin-Watson Test 27
3.2.9 Multicollinearity 28
3.2.10 Variance Inflation Factor 28
3.2.11 Principal Component Regression 29
3.2.12 Proportion of Variance that the components explain 29
3.2.13 Scree Plot 29

4 Data Analysis **31**
4.1 Time Series Analysis 31
4.1.1 Time series decomposition 31
4.1.2 Augmented Dickey Fuller Test 34
4.1.3 ACF and PACF Plots 35
4.1.4 Modelling 41
4.1.5 Accuracy of the model 44
4.1.6 Residual Analysis 48
4.1.7 Forecasting 53
4.2 Regression Analysis 60
4.2.1 Introduction 60
4.2.2 Fitted Regression Model 60
4.2.3 Multicollinearity 60
4.2.4 Variance Inflation Factor 61
4.2.5 Principal Component Analysis 62
4.2.6 Summary of the fitted model 63
4.2.7 Residual Analysis 64
4.2.8 Fitted Regression Model 65
4.3 Comparative Analysis of CPI(Total) 65

5 Conclusion **68**

6 Data References **70**

Bibliography **71**

A Python Code **72**
A.0.1 Time Series Analysis 72
A.0.2 Regression Analysis 78

Chapter 1

Introduction

1.1 Overview

The Consumer Price Index (CPI) is a critical economic indicator that measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services in India. The CPI is pivotal for understanding inflation, which is the rate at which the general level of prices for goods and services is rising, and subsequently eroding purchasing power.

The CPI in India is composed of various categories of goods and services, with each category assigned a different weight based on its relative importance in the average consumer's expenditures. The major categories include:

- Food and Beverages
- Clothing and Footwear
- Medical Care
- Transport and Communication
- Recreation and amusement

In this study, we utilize time series analysis to model and forecast the Consumer Price Index (CPI) as a whole and for each of its individual components. Following this, we apply regression analysis to construct a suitable model for the CPI, considering the impact of its various components. We then use the regression model to predict the overall CPI using the component forecasts obtained from the time series analysis. The accuracy of both analytical methods is subsequently evaluated.

Time Series is a set of observations x_t, each one being recorded at a specific time t.

Regression analysis is a statistical technique for investigating and modelling the relationship between variables. It is all about determining how changes in the independent variables are associated with changes in the dependent variable.

1.2 Objectives

- To build a model for time series data that forecast the CPI of India and its various components.
- To build a Regression model For CPI and its various components.
- Comparative Analysis of CPI(Total) forecasted from time series and MLR Model.

1.3 Software Tool Used

Python is a free, open-source programming language that includes extensive support for statistical computing and graphics. It operates across various operating systems, including UNIX platforms, Windows, and MacOS. Widely favored by statisticians, data scientists, and researchers, Python is heavily utilized for data analysis, visualization, and modeling. Based on user surveys and research, it ranks as one of the most popular programming languages for data mining.

Chapter 2

Data Description And Data Source

2.1 Data Source

The data is a primary data taken from CEIC Data site from the year 2011-2024 for analysis.

2.2 Data Description

The dataset for consumer price index(CPI) and its 5 components is calculated using Laspeyres's price index with base year 2012. Each dataset consist of 157 observations ranging from January 2011 to January 2024. In total we have 6 dataset. The dataset obtained from the above website were checked for outliers and where subsequently indexed with the corresponding month and year.

The components of CPI of India are:

- **(CPI) Food and Beverages** :In this section, various goods and services related to food and drink consumption are typically included. Here are some examples of goods and services that might be included in this section:
  - Staple food items (e.g., rice, wheat flour, pulses, sugar)

  - Fruits and vegetables (fresh, canned, or frozen)
  - Meat and poultry products (e.g., chicken, beef, lamb)
  - Seafood (e.g., fish, shrimp, crab)
  - Dairy products (e.g., milk, cheese, yogurt)
  - Eggs
  - Bakery products (e.g., bread, pastries, cakes)
  - Snack foods (e.g., chips, cookies, chocolates)
  - Beverages (both alcoholic and non-alcoholic), including: Soft drinks, Fruit juices, Tea and coffee, Alcoholic beverages (beer, wine, spirits)
  - Cooking oils and fats
  - Restaurant meals and take-out food
  - Food delivery services

- **(CPI) Clothing and Footwear** : In this section, various goods and services related to clothing and footwear are typically included. Here are some examples of goods and services that might be included in this section:
  - Clothing items for men, women, and children (e.g., shirts, pants, dresses, skirts, suits)
  - Outerwear (e.g., jackets, coats, sweaters)
  - Footwear for all genders and age groups (e.g., shoes, sandals, boots)
  - Accessories (e.g., belts, ties, scarves, hats)
  - Sportswear and athletic apparel
  - Uniforms and workwear
  - Fabric and materials for sewing or crafting clothing
  - Alteration and tailoring services
  - Cleaning and laundry services for clothing and footwear

- **(CPI) Medical Care** : In this section, various goods and services related to healthcare and medical expenses are typically included. Here are some examples of goods and services that might be included in the Medical Care section of the CPI:

  - Medical consultation fees
  - Hospital services, including inpatient and outpatient care
  - Prescription drugs and medications
  - Over-the-counter medications and health products
  - Health insurance premiums
  - Dental care services, including check-ups, cleanings, and treatments
  - Vision care services, such as eye exams and prescription eyewear (glasses, contact lenses)
  - Medical equipment and supplies (e.g., wheelchairs, walkers, blood pressure monitors)
  - Diagnostic tests and procedures (e.g., X-rays, blood tests, MRIs)
  - Alternative healthcare services (e.g., acupuncture, chiropractic care)

- **(CPI) Transport and Communication** : In this section, various goods and services related to transportation and communication expenses are typically included. Here are some examples of goods and services that might be included in this section:

  - Transportation fares (e.g., bus, train, tram, metro/subway, taxi)
  - Gasoline and diesel fuel
  - Vehicle purchases (e.g., cars, motorcycles, bicycles)
  - Vehicle maintenance and repair services
  - Vehicle insurance premiums
  - Public transportation passes and tickets
  - Airline fares (domestic and international)
  - Shipping and freight charges
  - Postal services (e.g., postage stamps, courier services)
  - Telecommunication services (e.g., landline telephone, mobile phone, internet)
  - Cable or satellite television services

  - Purchase of communication devices (e.g., smartphones, tablets, computers)
  - Accessories and peripherals for communication devices (e.g., chargers, headphones)
  - Installation fees for communication services (e.g., internet setup, cable installation)

- **(CPI) Recreation And Amusement** : In this section, various various goods and services related to leisure activities and entertainment that households may spend their money on are typically included. Some examples of goods and services that could be included in this category are:

  - Tickets to movies, theaters, concerts, and other cultural events
  - Fees for amusement parks, theme parks, and recreational facilities
  - Subscription fees for streaming services (e.g., Netflix, Amazon Prime Video, Disney+)
  - Video games and gaming consoles
  - Fees for recreational classes and activities (e.g., dance classes, music lessons, sports clubs)
  - Cost of recreational vehicles (e.g., bicycles, motorcycles)
  - Books, magazines, and other reading materials
  - Outdoor equipment (e.g., camping gear, hiking equipment)
  - Costs associated with hobbies and crafts (e.g., art supplies, musical instruments)

Chapter 3

Methodology

3.1 Time Series Analysis

A time series is a collection of data points that are gathered and stored over regular time intervals. It depicts how a variable or phenomenon has changed over time, making it possible to analyze, predict, and comprehend patterns and trends. Time series data is frequently utilized in many different disciplines, including signal processing, forecasting the weather, studying the stock market, and many other areas.

Making predictions about future values based on prior observations and patterns, trends, and seasonality are all part of the analysis of time series data. Several methods, including statistical models, machine learning algorithms, and time series forecasting techniques like ARIMA (Autoregressive Integrated Moving Average), can be used to accomplish this. Time series refers to a sequence of data points collected at regular intervals over time. These data points can represent various variables, such as stock prices, weather measurements, or website traffic.

Time series analysis involves examining the patterns, trends, and relationships in the data to make predictions and inform decision-making. Common techniques for analyzing time series data include smoothing, trend analysis, seasonality analysis, and forecasting. Smoothing involves removing noise or fluctuations in the data to reveal underlying trends or patterns

The usage of time series is to

1. Obtain an understanding of the underlying forces and structure that produced the observed data.

2. Fit a model and proceed to forecast, and monitoring. Time series analysis can be useful to see how a given asset, security, or economic variable change over time.

A given time series can be

- Stationary
- Non-stationary

3.1.1 Stationary

A time series $\{X_t\}$ is said to be **stationary** if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ is identical to that of $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$ for all $(t_1, t_2, \ldots, t_n)$ and h, where n is an arbitrary positive integer and $(t_1, t_2, \ldots, t_n)$ is a collection of n positive integers.

A time series $\{X_t\}$ is **weakly stationary** if X_t has a constant mean, finite variance, and the covariance between X_t and X_{t+h} depends only on h, where h is an arbitrary integer. From the definitions, if $\{X_t\}$ is strictly stationary and its first two moments are finite, then it is also weakly stationary.

3.1.2 Non Stationary

There are many ways in which a time series may fail to be stationary, and such series are referred to as **non-stationary time series**. There are several methods available to transform non-stationary data into stationary time series data. Some of these methods include estimation and differencing.

In the present study, the method of **differencing** has been applied in order to make non-stationary data into a stationary one. This method involves replacing the original series $\{X_t\}$ by $Y_t = X_t - X_{t-d}$ for some positive integer d.

3.1.3 Components of Time Series

The components of a time series are the underlying patterns that make up the time series data.
Understanding the components of a time series is important for modeling and forecasting. Different models can be used to capture each component of the time series, such as ARIMA models for trend and seasonality, and GARCH models for volatility.

By separating the different components of a time series, we can better understand the patterns in the data and make more accurate forecasts.
The four main components of a time series are:

- Trend
- Seasonality
- Cyclical
- Random

Trend: The trend component of a time series represents the long-term upward or downward movement of the data over time. It reflects the overall direction in which the data is moving and can be linear or non-linear.

Seasonality: The seasonality component of a time series represents the regular, periodic fluctuations in the data that occur within a fixed time interval (such as a day, week, or month). Seasonality is often driven by external factors such as weather, holidays, or other recurring events.

Cyclical: The cyclical component of a time series represents the non-periodic, long-term fluctuations in the data that occur over a time span longer than a season. Cyclical fluctuations can be caused by factors such as economic cycles or political events.

Random: The random (or error) component of a time series represents the unpredictable fluctuations in the data that cannot be explained by the other three components. The random component is assumed to be white noise, meaning that the observations are independent and identically distributed with a constant mean and variance.

3.1.4 Partial Autocorrelation Function

The Partial Autocorrelation Function (PACF) provides a plot of the correlation of the residuals with their respective lag values. Specifically, PACF measures the extent of the relationship between current values of a time series $\{X_t\}$ and its earlier values, while holding the effect of all intermediate lags constant.

PACF is utilized to identify the order p of an autoregressive (AR) process. In an AR(p) process, the PACF cuts off at lag p, meaning that the correct order is assessed as that value of p beyond which the sample values of partial autocorrelations are not significantly different from zero.

Mathematically, the partial autocorrelation function of a given time series $\{X_t\}$ at lag h is defined as the partial correlation coefficient between X_t and X_{t+h}, obtained by controlling for the effects of $(X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$:

$$\text{PACF}(h) = \text{Corr}(X_t, X_{t+h} \mid X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$$

This function is crucial for determining the significant lags and thus the likely parameters for fitting AR models to time series data.

3.1.5 Autocorrelation Function

The autocorrelation function (ACF) is a mathematical function that measures the correlation between a time series and its lagged values. Specifically, the ACF at lag k measures the correlation between the time series and its values k time units in the past.

The ACF is an important tool in time series analysis because it can help identify patterns in the time series that repeat over time, such as seasonality or trend. If there is a significant correlation between the time series and its lagged values at a particular lag, this suggests that the time series is influenced by its past values at thatlag. This information can be useful in selecting appropriate models for forecasting.

The autocorrelation coefficient ρ_k at lag k measures the correlation between two values X_t and X_{t-k}, a distance k apart. The covariance between $\{X_t\}$ and $\{X_{t-k}\}$ is known as the autocorrelation function at lag k and is defined by:

$$\gamma_k = \text{cov}(X_t, X_{t-k}) = \mathbb{E}[(X_t - \mathbb{E}[X_t])(X_{t-k} - \mathbb{E}[X_{t-k}])]$$

The correlation coefficient between $\{X_t\}$ and $\{X_{t-k}\}$ is called the Autocorrelation Function (ACF) at lag k and is given by:

$$\rho_k = \text{corr}(X_t, X_{t-k}) = \frac{\text{cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t)\text{Var}(X_{t-k})}}$$

3.1.6 Autoregressive Model

A time series model known as an autoregressive (AR) model forecasts future values using historical data for the same variable. In an AR model, it is assumed that a variable's value at a particular time point is linearly dependent on its prior values. The AR model makes the assumption that the time series' present value is the result of linearly combining its previous values with a random error term. The number of prior values used to predict the current value is specified by the AR model's order,

Autoregressive models are fundamental in time series analysis, commonly used in fields like finance, economics, and signal processing. The concept of an autoregressive model of order p, abbreviated as AR(p), is based on the idea that the current value of a series X_t can be explained as a function of its past values. Specifically, an AR(p) model predicts the current value using the values from p previous time points.

The general form of an AR(p) model is given by:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \alpha_t \tag{3.1}$$

where $\{\alpha_t\} \sim WN(0, \sigma^2)$ represents a white noise series with mean zero and variance σ^2, and $\phi_1, \phi_2, \ldots, \phi_p$ are constants, representing the coefficients of the model.

To estimate the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and σ^2 of the AR model, techniques such as maximum likelihood estimation and least squares estimation are typically employed. Once estimated, these parameters allow the AR model to be used for forecasting future values of the time series.

3.1.7 Moving Average Model

Another form of statistical model frequently applied to forecasting and time series research is the moving average (MA) model. The Moving Average model concentrates on modeling the link between the current value and the residual errors from previous predictions, in contrast to the Autoregressive (AR) model, which predicts the current value based on past values of the variable.

A moving average model is conceptually a linear regression of the current value A moving average process of order q, denoted as MA(q), models the current value of a series X_t as a linear combination of the current and q previous white noise error terms or random shocks. The MA(q) model is mathematically expressed as:

$$X_t = \alpha_t + \theta_1 \alpha_{t-1} + \cdots + \theta_q \alpha_{t-q} \tag{3.2}$$

where $\{\alpha_t\} \sim WN(0, \sigma^2)$ represents a white noise series with mean zero and variance σ^2, and $\theta_1, \theta_2, \ldots, \theta_q$ are constants.

The parameters $\theta_1, \theta_2, \ldots, \theta_q$ of the MA model can be estimated using techniques such as maximum likelihood estimation or least squares estimation. Once the parameters are estimated, they enable the MA model to be used for making predictions for future values of the time series.

The MA model is often used in conjunction with the autoregressive (AR) model to form an AutoRegressive Moving Average (ARMA) model. The ARMA model combines both the autoregressive and moving average components, providing better forecasting performance for time series data. This model effectively captures the dynamics in both recent observations and recent forecast errors.

3.1.8 Autoregressive Moving Average Model

The AutoRegressive (AR) and Moving Average (MA) models are combined to create the Autoregressive Moving Average (ARMA) model. It is a well-liked model for forecasting and time series analysis. The autoregressive component (autoregressive component) and the moving average component (moving average component) of the time series determine the current value of the time series in an ARMA model. Two parameters—p for the autoregressive order and q for the moving average order—define the ARMA model.

The autoregressive moving average process (ARMA) process of order (p, q) is obtained by combining an MA(q) process and an AR(p) process. The series X_t is an Autoregressive Moving Average Model (ARMA(p, q)model) if X_t is stationary and if for every t,

An ARMA process of order (p, q) can be concisely expressed using polynomial notation. The model is defined as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \alpha_t + \theta_1 \alpha_{t-1} + \cdots + \theta_q \alpha_{t-q} \tag{3.3}$$

where $\{\alpha_t\} \sim WN(0, \sigma^2)$ and the polynomials $(1 - \phi_1 B - \cdots - \phi_p B^p)$ and $(1 + \theta_1 B + \cdots + \theta_q B^q)$ have no common factors. This can be written more concisely as :

$$\phi(B) X_t = \theta(B) \alpha_t \tag{3.4}$$

where $\phi(B)$ and $\theta(B)$ are the p-th and q-th degree polynomials defined by:

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \tag{3.5}$$

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q \tag{3.6}$$

and B is the backshift operator.

The backshift operator B is a notational convenience used in time series analysis where applying B to a time series X_t results in $BX_t = X_{t-1}$. For instance, $B^k X_t = X_{t-k}$.

3.1.9 Autoregressive Integrated Moving Average Model

A helpful statistical technique for the study of longitudinal data with a connection between nearby observations is the ARIMA model, which is designed for stationary time series. The behavioral time series process is represented in ARIMA analysis by two straightforward components: the autoregressive (AR) and moving average (MA) models. The term ”autoregressive integrated moving average model” (ARIMA(p,d,q)) is used to describe a process X_t.

$$\Delta^d X_t = (1 - B)^d X_t \tag{3.7}$$

is ARMA(p,q) . In general, the model can be written as

$$\phi(B) X_t = \theta(B) \alpha_t \tag{3.8}$$

where $\{\alpha_t\} \sim WN(0, \sigma^2)$ represents a white noise series with mean zero and variance σ^2.

3.1.10 Seasonal Autoregressive Integrated Moving Average Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an extension of the ARIMA model that supports univariate time series data with a seasonal component. It is commonly used in time series forecasting to model seasonal variations, in addition to trends and cycles.

The SARIMA model is typically denoted as SARIMA$(p, d, q)(P, D, Q)_s$, where:

- p, d, q are the non-seasonal parameters for the autoregressive order, differencing order, and moving average order, respectively.
- P, D, Q are the seasonal components of the SARIMA model corresponding to the seasonal autoregressive order, seasonal differencing order, and seasonal moving average order.

- s represents the length of the seasonal cycle.

The mathematical representation of the SARIMA model is given by:

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^dX_t = \Theta_Q(B^s)\theta_q(B)a_t \tag{3.9}$$

where:

- B is the backshift operator.
- $\phi_p(B)$ and $\Phi_P(B^s)$ are the non-seasonal and seasonal autoregressive polynomials.
- $\theta_q(B)$ and $\Theta_Q(B^s)$ are the non-seasonal and seasonal moving average polynomials.
- a_t is the error term, assumed to be white noise.

3.1.11 Model Selection Criteria

We have a number of models available for time series analysis that are suitable for the data. We decided on the option that best fits the data. Mean Absolute Percent Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error are some of the main criteria used to choose a model. (RMSE). Since the residuals are not normal, the focus of this study is mostly on a mean absolute present error and root mean square error for the model selection process. The selection process is set up so that the model with the lowest MAE and RMSE values is chosen.

Akaike Information Criteria: Assume that a statistical model with the m parameter is fitted to the data. Akaike proposed the following information criteria to assess the model.

$$\text{AIC} = -2\log(L) + 2m \tag{3.10}$$

where m is the number of estimated parameters in the model.

We always prefer the model with minimum AIC

Mean Absolute Error (MAE): The mean of the absolute deviation of predicted and observed values is called absolute mean error

Mean Absolute Percent Error(MAPE): The mean of the sum of the absolute deviation of the predicted value divided by the observed value is called the mean absolute error. For comparison, we will multiply by 100, which is called the mean absolute present error

Root Mean Square Error(RMSE): The square root of the sum of the deviation of the predicted values from the observed values divided by their number of observations is known as the root mean square error.

3.1.12 Unit Root Test

A **unit root process** refers to any time series whose characteristic equation has one or more roots equal to one. A simple example of such a process is given by the Autoregressive model of order 1 (AR(1)), expressed as:

$$X_t = \phi X_{t-1} + \alpha_t \tag{3.11}$$

where α_t represents a white noise error term with zero mean and constant variance, signifying that the errors are serially uncorrelated.

In the AR(1) model:

- If $\phi = 1$, the model $X_t = X_{t-1} + \alpha_t$ becomes a **random walk model without drift**, which is inherently **non-stationary**. This scenario illustrates the **unit root problem**, where the time series does not return to a long-run mean, and the variances depend on time, diverging as time progresses.
- If $|\phi| < 1$, the series X_t is **stationary**. In this case, the series will revert to a mean, and the variances are constant over time.

By differencing the dataset, the unit root problem can be resolved or stationarity can be attained. There are several tests that may be used to determine whether a time series is stationary.

3.1.13 Augmented Dicky-Fuller Test

The augmented Dickey-Fuller test (ADF) checks if a time series of data contains a unit root. When analyzing the stationarity of time series, it is one of the statistical tests that are most frequently utilized. The null hypothesis to be tested is the presence of unit root, that is -1. If the p-value obtained is less than the significance level (say 0.05) null hypothesis is rejected. Thereby, inferring that the series is stationary.

3.1.14 Ljung-Box Test

The Ljung-Box test is used to determine whether a time series' autocorrelations differ from zero. This is the test statistic

$$\tilde{Q}_m = n(n+2)\sum_{k=1}^{m}\frac{r_a(k)^2}{n-k} \tag{3.12}$$

where:

- n is the sample size.
- $r_a(k)$ is the sample residual autocorrelation at lag k.
- m is the number of lags being tested.

The statistic $\tilde{Q}_m$ approximates a chi-square distribution with $m-p-q$ degrees of freedom under the null hypothesis, where p and q are the orders of the autoregressive and moving average parts of the model, respectively.

The Ljung-Box test evaluates whether the autocorrelations of the model residuals are different from zero in a collective sense. If the computed value of $\tilde{Q}_m$ is greater than the critical value from the chi-square distribution at the desired level of significance, we reject the null hypothesis. This suggests that the residuals exhibit autocorrelation up to lag m, indicating that the model may not be adequately capturing the underlying process.

3.1.15 Residual Analysis

For an adequate model, the residuals should be Independent and Identically Distributed (IID) and are uncorrelated. To test the correlation, we use Auto Correlation Function (ACF) plot and the normality of the residuals are tested using Shapiro Wilk normality test.

3.1.16 Forecast

Time series analysis often aims to predict future values of a series based on past and present observations. When the time series data follow a linear model, one of the most effective techniques used for prediction is the Minimum Mean Square Error (MMSE) forecasting.

The goal of MMSE forecasting is to minimize the expected value of the square of the difference between the estimated and actual values. For a linear process $\{Z_t\}$, the I-step ahead forecast can be represented as:

$$\hat{Z}_{t+I} = \mathbb{E}[Z_{t+I}|Z_t, Z_{t-1}, \dots] \tag{3.13}$$

This forecast $\hat{Z}_{t+I}$ is the expected value of Z_{t+I} given the information available up to time t. The estimation utilizes all available past values of the series to predict the future value.

In the context of a linear process, the MMSE forecast typically involves linear combinations of past data points:

$$\hat{Z}_{t+I} = a_0 + a_1 Z_t + a_2 Z_{t-1} + \cdots + a_n Z_{t-n} \tag{3.14}$$

where $a_0, a_1, \dots, a_n$ are coefficients optimized to minimize the mean square error in the forecast. These coefficients are determined based on the properties of the time series, such as its mean, variance, and autocorrelations.

In practice, MMSE forecasts are widely used in various fields such as economics, meteorology, and engineering, where accurate predictions of future events based on historical data are crucial. The effectiveness of the MMSE method lies in its ability to incorporate the stochastic dependencies of the time series data effectively.

This method also forms the basis for more complex forecasting models like ARIMA (AutoRegressive Integrated Moving Average), where the terms are specifically designed to model the autocorrelations in the series for improved prediction accuracy.

3.2 Regression Analysis

Regression Analysis is a statistical technique for investigating and modelling the relationship between one or more response (dependent) variables and a set of variables called regressor (independent or explanatory) variables.

3.2.1 Multiple Linear Regression

Multiple linear regression (MLR) is a statistical technique that models the relationship between a dependent variable and two or more independent variables. This method is used extensively in fields such as economics, social sciences, and engineering to analyze the effects of several factors on a response variable.

The general form of a multiple linear regression model with k regressors is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{3.15}$$

where:

- y represents the dependent variable.
- $x_1, x_2, \dots, x_k$ are the explanatory variables (or regressors).

- $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients, which are unknown parameters to be estimated.
- ϵ represents the random error component of the model, which is assumed to be uncorrelated with mean zero and constant variance. This assumption is critical as it underpins the classical linear regression assumptions, including homoscedasticity and no autocorrelation.

The parameters of the multiple linear regression model are typically estimated using the method of least squares. This method minimizes the sum of the squared residuals, providing the best linear unbiased estimates (BLUE) of the coefficients under the Gauss-Markov theorem.

Multiple linear regression analysis is utilized to understand the influence of various independent variables on the dependent variable. It helps in predicting the value of the dependent variable based on the values of the independent variables. Additionally, it is used to test hypotheses on the impact of various factors on a particular outcome.

3.2.2 Assumptions of MLR

The major assumptions that we consider in regression analysis are as follows:

- The relationship between the response y and the regressors is linear, at least approximately.
- The error term has zero mean.
- The error term has constant variance.
- The errors are normally distributed.

3.2.3 Least Square Estimation of Regression Coefficients

In linear regression, the least squares method is employed to estimate the regression coefficients. This method involves minimizing the sum of the squares of the residuals, which is the difference between the observed values and the values predicted by the model.

The objective function, known as the sum of squares of residuals (SSR), is given by:

$$S = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_{ij} \right)^2 \tag{3.16}$$

where:

- Y_i represents the observed values of the dependent variable.
- X_{ij} represents the j-th predictor (independent variable) for the i-th observation.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients to be estimated.

To find the values of the coefficients that minimize S, we take the partial derivatives of S with respect to each coefficient β_j, set them to zero, and solve the resulting system of equations. This leads to the normal equations:

$$\frac{\partial S}{\partial \beta_j} = 0, \quad j = 0, 1, 2, \ldots, k \tag{3.17}$$

The least squares estimates of the coefficients can be efficiently computed using matrix operations. Representing the design matrix by X (where each row corresponds to an observation and each column corresponds to a predictor, including the intercept), the vector of outcomes by Y, and the vector of coefficients by $\boldsymbol{\beta}$, the normal equations in matrix form are:

$$(X'X)\boldsymbol{\beta} = X'Y \tag{3.18}$$

Solving for $\boldsymbol{\beta}$, we get:

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y \tag{3.19}$$

where $\hat{\boldsymbol{\beta}}$ represents the vector of estimated regression coefficients.

The matrix solution is particularly useful in statistical software and applications involving large datasets or multiple regression models, as it leverages efficient matrix computation techniques to quickly compute estimates.

3.2.4 Coefficient of Determination

In linear regression, the effectiveness of the model is often assessed using the coefficient of determination, denoted as R^2. This statistic measures the proportion of the total variation in the dependent variable that is explained by the independent variables in the model.

The R^2 is calculated as follows:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{3.20}$$

where:

- SSR (Residual Sum of Squares) is the variation explained by the regression model.
- SST (Total Sum of Squares) is the total variation in the dependent variable.

The alternative formula for R^2:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{3.21}$$

The value of R^2 ranges from 0 to 1:

- A value of 0 indicates that the model explains none of the variability of the response data around its mean.
- A value of 1 indicates that the model explains all the variability of the response data around its mean.

A higher R^2 value is an indicator of a better fit of the model to the data. However, it is crucial to note that a high R^2 does not necessarily mean the model is optimal.

Factors such as overfitting and the choice of predictors need to be carefully considered. It is also important to look at other metrics and diagnostic tests to evaluate model performance comprehensively.

3.2.5 Testing the Significance of Regression

In multiple linear regression analysis, the significance of the regression itself is tested using Analysis of Variance (ANOVA). This test evaluates whether there is a statistically significant relationship between the response variable and at least one of the predictor variables.

The hypothesis for testing the overall significance of the regression is formulated as:

- $H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$ - This null hypothesis states that none of the predictors have any effect on the response variable.
- $H_1 : \exists \beta_{ij} \neq 0,$ for at least one j - The alternative hypothesis claims that at least one predictor does influence the response variable.

The significance of the regression is tested using the F-statistic, which is derived from the ANOVA table. The F-statistic is calculated as:

$$F = \frac{\text{Mean Square Error (MSE)}}{\text{Mean Square Regression (MSR)}} \tag{3.22}$$

where:

$$\text{MSE} = \frac{\text{Sum of Squares due to Error (SSE)}}{n-k} \tag{3.23}$$

$$\text{MSR} = \frac{\text{Sum of Squares due to Regression (SSR)}}{k-1} \tag{3.24}$$

Here, n is the number of observations, and k is the number of predictors.

From the ANOVA table, the calculated F-statistic can be compared against a critical value F_0 from the F-distribution table at a specified significance level. The null hypothesis H_0 is rejected if:

$$F > F_0 \tag{3.25}$$

Rejecting the null hypothesis implies that the regression model is statistically significant, meaning there is evidence that at least one predictor variable has a significant effect on the response variable.

3.2.6 Shapiro-Wilk normality test

The Shapiro-Wilk test is used to assess the null hypothesis that a sample $X_1, X_2, \ldots, X_n$ comes from a normally distributed population. This test is particularly popular due to its sensitivity to deviations from normality.

The test statistic W is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{3.26}$$

where:

- $X_{(i)}$ is the i-th order statistic (i.e., the i-th smallest value in the sample).
- $\bar{X}$ is the sample mean.
- a_i are constants calculated from the order statistics of a standard normal distribution.

The coefficients a_i are computed using the formula:

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}} \tag{3.27}$$

where:

- $m = (m_1, m_2, \ldots, m_n)^T$ represents the expected values of the order statistics for independent and identically distributed (iid) random variables from a standard normal distribution.
- V is the covariance matrix of these order statistics.

The Shapiro-Wilk test statistic W is a measure of how much the observed data deviate from the hypothesis of normality. The closer W is to 1, the more evidence there is that the sample was drawn from a normal distribution. Typically, if W is significantly less than 1, the null hypothesis of normality can be rejected, suggesting that the data are not normally distributed.

3.2.7 Breusch-Pagan Non Constant Variance Test

The null hypothesis that the error variances are all equal against the alternate hypothesis that the error variances are not equal can be tested using the non-constant variance test. The test statistic is nR^2 which approximately follows a chi-square distribution with $k$degrees of freedom where, k is the number of independent variables and n is the sample size.

3.2.8 Durbin-Watson Test

One of the major assumptions of regression analysis is the errors are uncorrelated. Correlation in the error terms suggests that there is additional information in the data that has not been exploited in the current model and such error terms are said to be auto correlated. Auto correlation of the model can be tested using Durbin and Watson(1951) is based on the assumption that successive errors are correlated. The Durbin-Watson statistic is defined as,

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{3.28}$$

where:

- e_i is the i-th ordinary least squares residual.
- n is the number of observations.

The Durbin-Watson statistic is used to test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_1 : \rho \neq 0$, where ρ represents the first-order autocorrelation coefficient of the residuals. Values of d close to 2 indicate little to no autocorrelation, while values deviating substantially from 2 suggest positive or negative autocorrelation:

3.2.9 Multicollinearity

When there are near-linear dependencies among the regressors, the problem of multicollinearity said to exist. In such cases the inferences based on the regression model may be misleading. This problem may be due to the data collection method employed, constraints on the model or in the population, an over defined model or may be due to the wrong choice of the model.

3.2.10 Variance Inflation Factor

The Variance Inflation Factor (VIF) is a measure commonly used to detect the presence of multicollinearity among the independent variables in a regression model. Multicollinearity occurs when some of the independent variables are highly correlated, leading to difficulties in estimating the regression coefficients accurately.

The VIF for each predictor in a regression model is calculated as follows:

$$\text{VIF} = \frac{1}{1 - R^2} \tag{3.29}$$

where R^2 is the coefficient of multiple determination for the regression model including all other independent variables except the one for which the VIF is being calculated. This R^2 reflects how well the variable is explained by the other independent variables in the model.

A high VIF indicates a high degree of multicollinearity. Typically:

- A VIF value greater than 10 suggests that the regression coefficients associated with the variable are poorly estimated due to multicollinearity.

- Values less than 10 imply an acceptable level of correlation among the independent variables.

3.2.11 Principal Component Regression

The problem of multicollinearity can be removed by obtaining biased estimators of regression coefficients. Principal component regression is an important technique used for removing multicollinearity. The principal component regression method combats multicollinearity by using less than the full set of principal components in the model. The principal components are obtained by arranging the eigen values in the decreasing order. The principal components corresponding to small eigen values are removed from the analysis and least squares applied to the remaining components.

3.2.12 Proportion of Variance that the components explain

Use the cuulative proportion to determine the amount of variance that the principal components explain. Retain the principal components that explain an acceptable level of variance. The acceptable level depends on your application. For descriptive purposes, you may only need 80 percentage of variance explained . However, if you want to perform other analysis on the data, you may want to have atleast 90 percentage of the variance explained by the principal components.

3.2.13 Scree Plot

The scree plot orders the eigen values from largest to smallest. The ideal pattern is a steep curve, followed by a bend, and then a straight line. Use the components in the steep curve before the first point that starts the line trend. The scree plot orders the eigen values from largest to smallest. The ideal pattern is a steep curve, followed by a bend, and then a straight line. Use the components in the steep curve before the first point that starts the line trend.

Chapter 4

Data Analysis

4.1 Time Series Analysis

4.1.1 Time series decomposition

The following results are the time series decomposition of CPI and its various components.

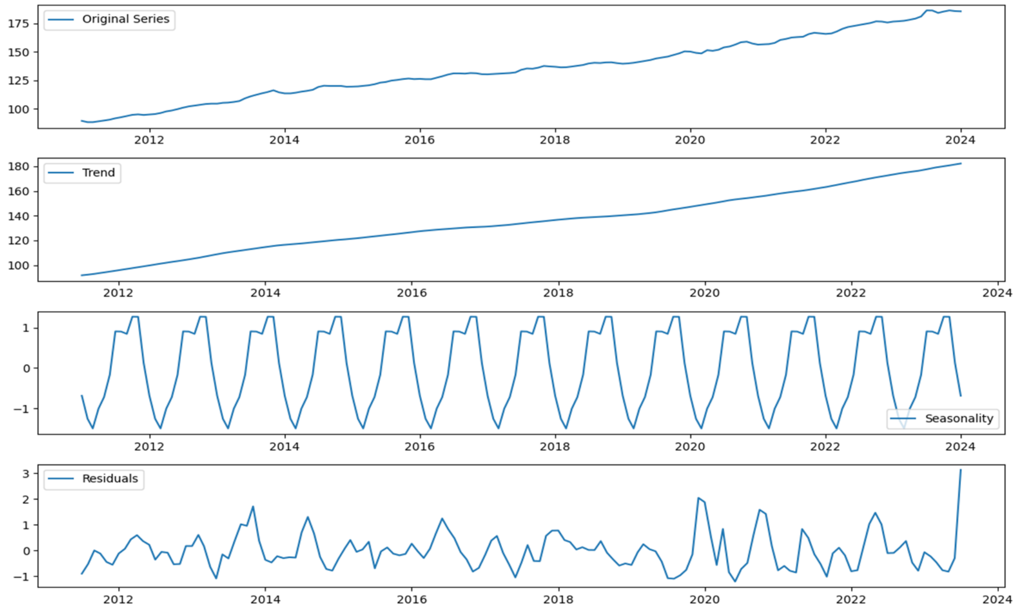

Figure 4.1: CPI(Total)

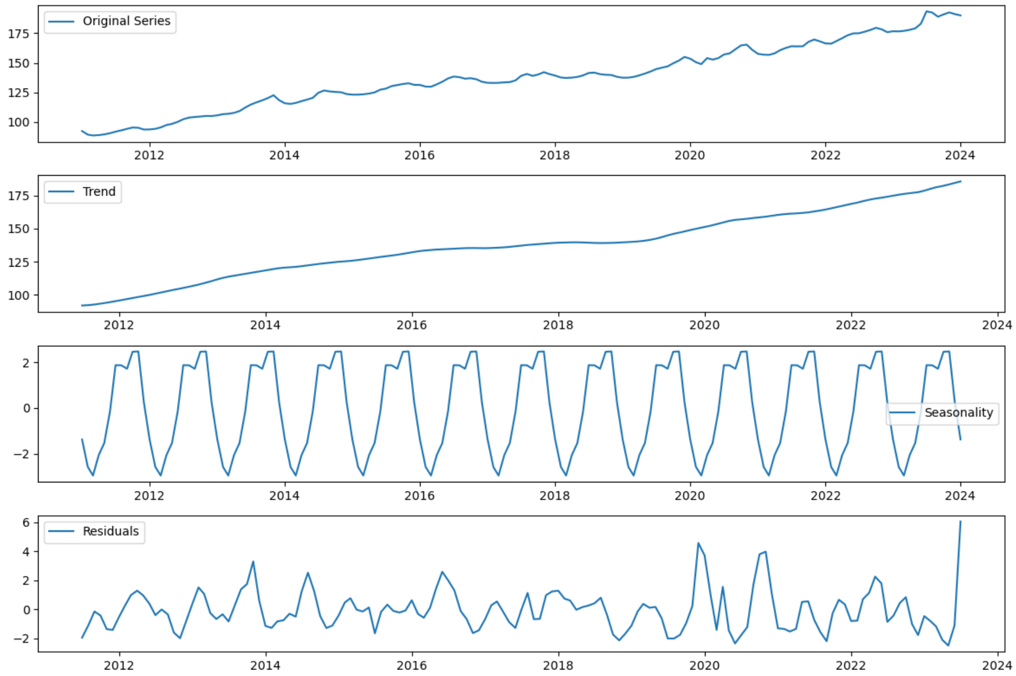

Figure 4.2: CPI Food and beverages

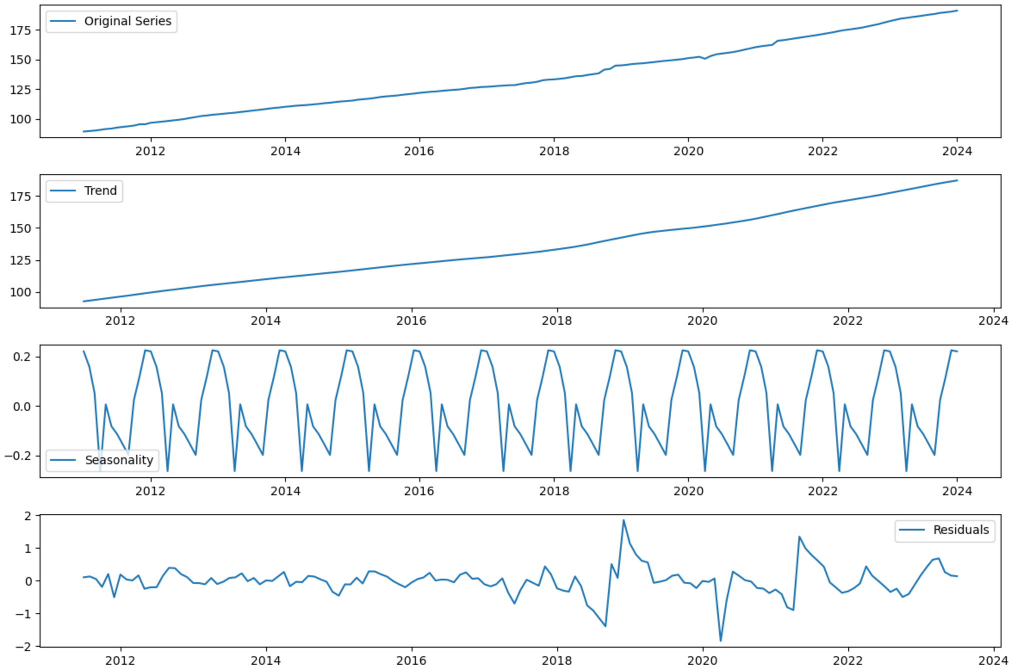

Figure 4.3: CPI Medical Care

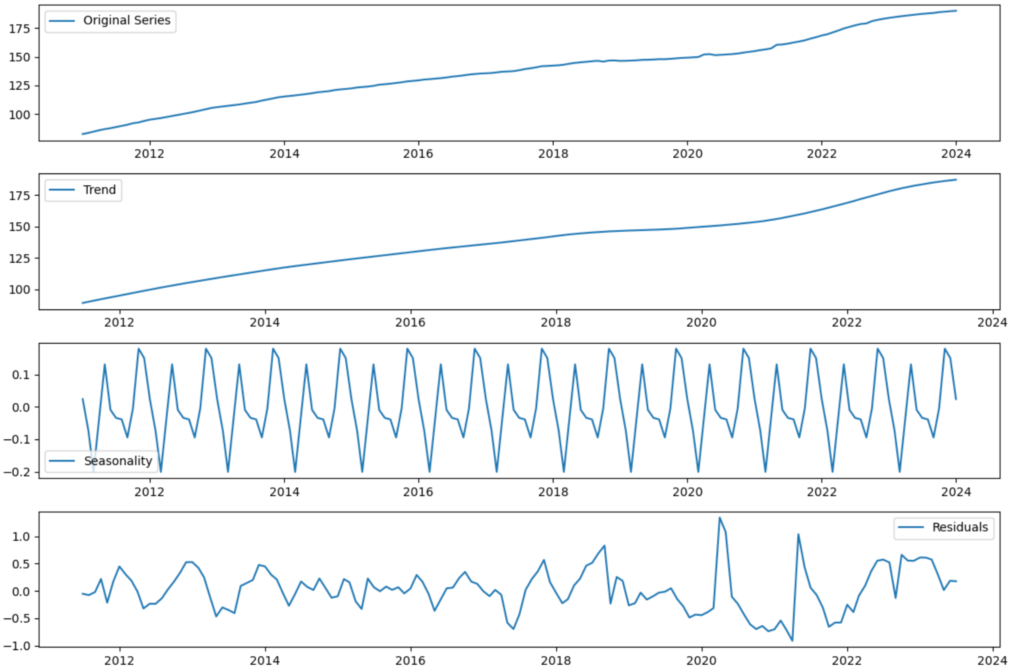

Figure 4.4: CPI Clothing and Footwear

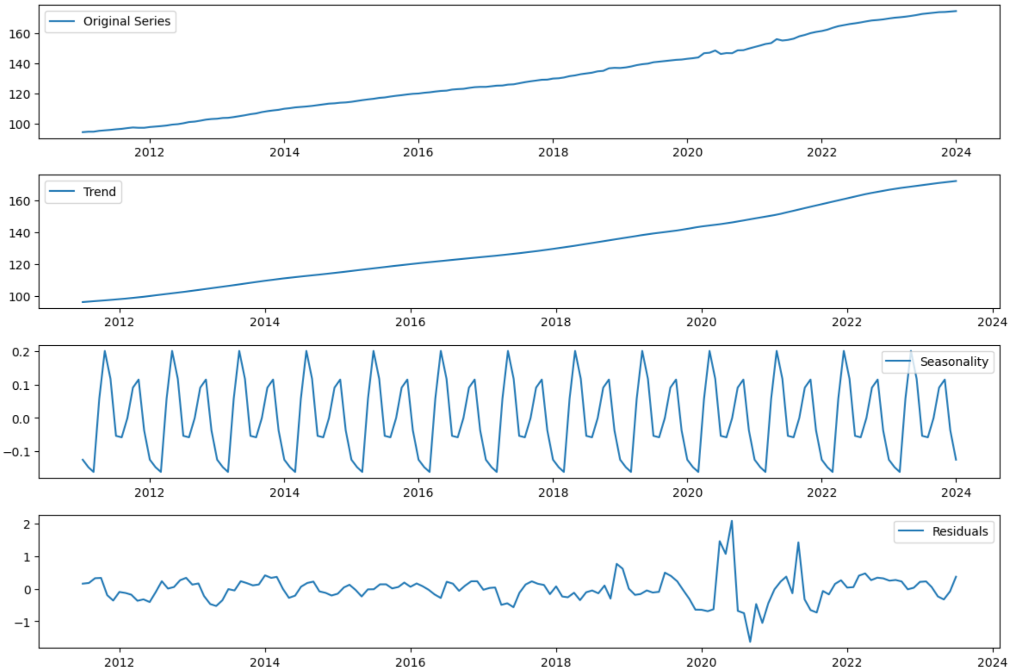

Figure 4.5: CPI Recreation And Amusement

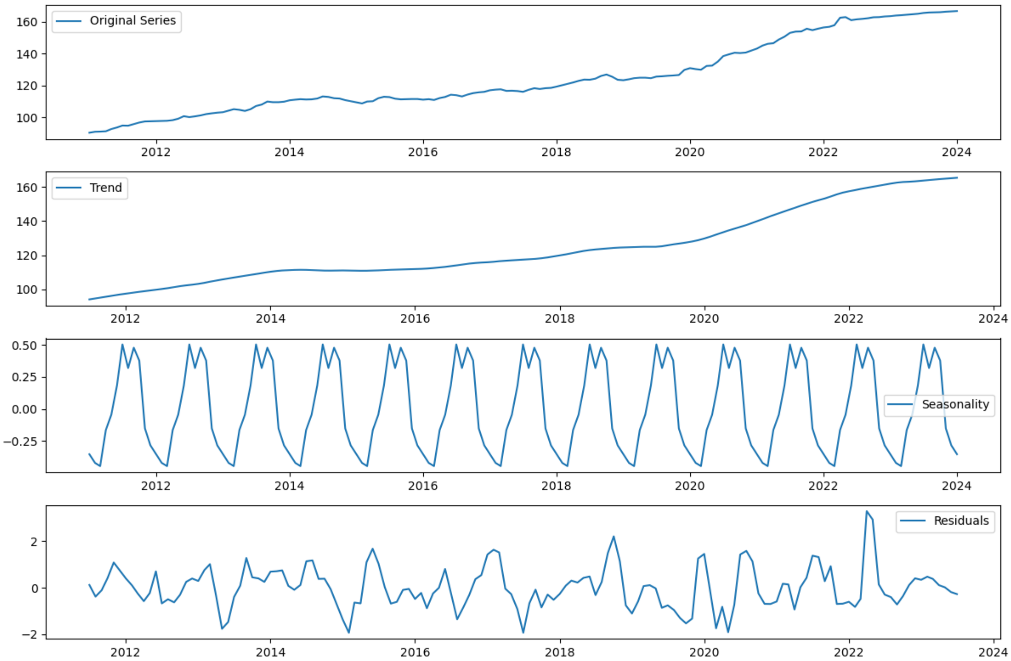

Figure 4.6: CPI Transport and Communication

4.1.2 Augmented Dickey Fuller Test

The presence of a linear trend in each time series data might suggest non-stationarity.So a stationarity test, specifically the Augmented Dickey-Fuller Test, was performed on each dataset.The result of the test is given in the Table 4.1:

Component	ADF Statistic	p value	Order of differencing
CPI(Total)	-7.340164829	1.07038034e-10	1
CPI(FAB)	-7.94933178776	3.1568178e-12	1
CPI(MC)	-7.35771241139	9.6816745e-11	2
CPI(CAF)	-2.1358173312	0.02303582	1
CPI(RAA)	-4.00542465	0.00138237	2
CPI(TAC)	-9.5329684765	2.85796342e-16	2

Table 4.1

From the test we can conclude that the datasets are non-stationary.

4.1.3 ACF and PACF Plots

The following results are the ACF and PACF Plots of CPI and its various components.

CPI(Total)

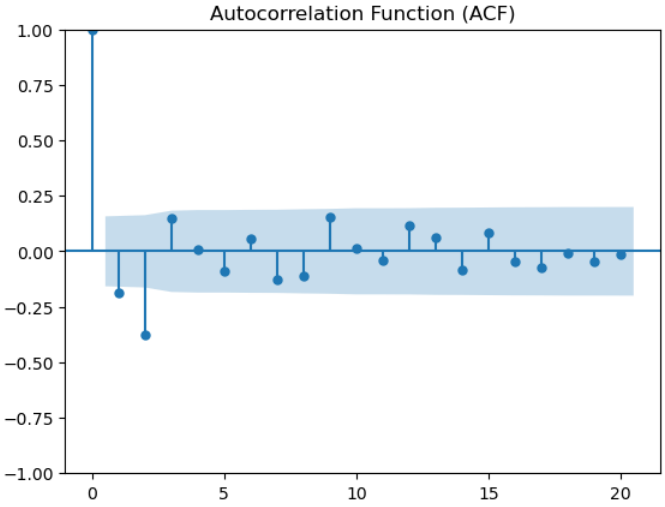

FIGURE 4.7: ACF plot

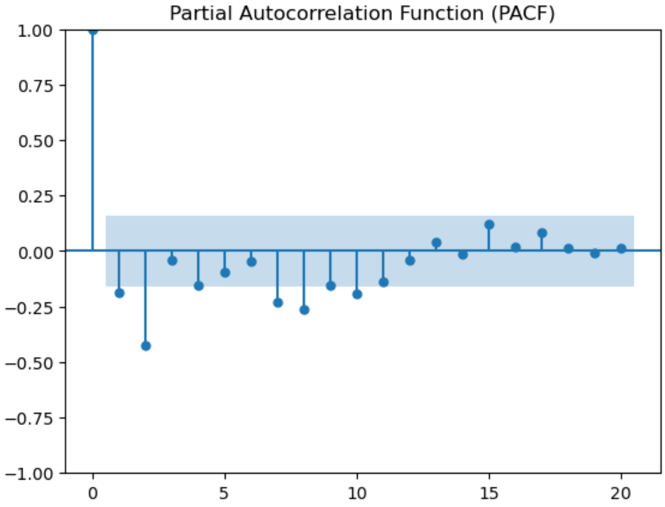

FIGURE 4.8: PACF plot

Here the value of p=0,1,2 and q=0,1,2.

CPI Food and beverages

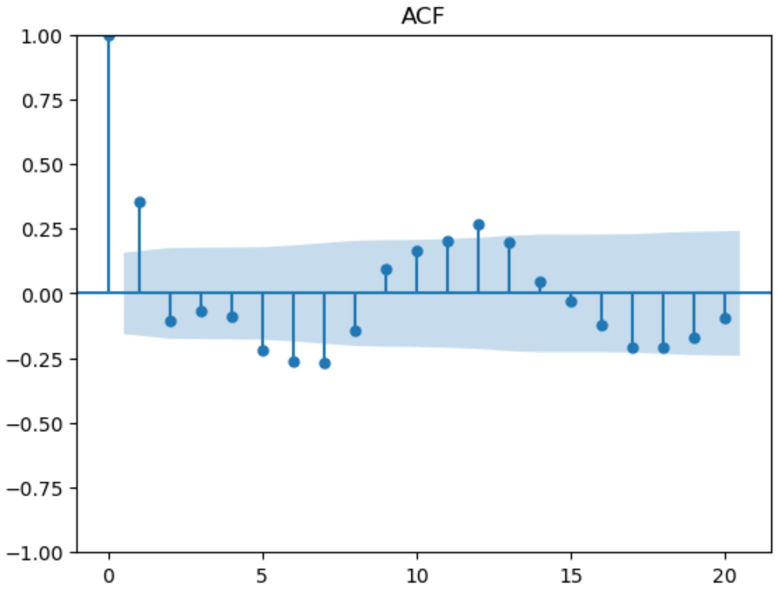

FIGURE 4.9: ACF plot

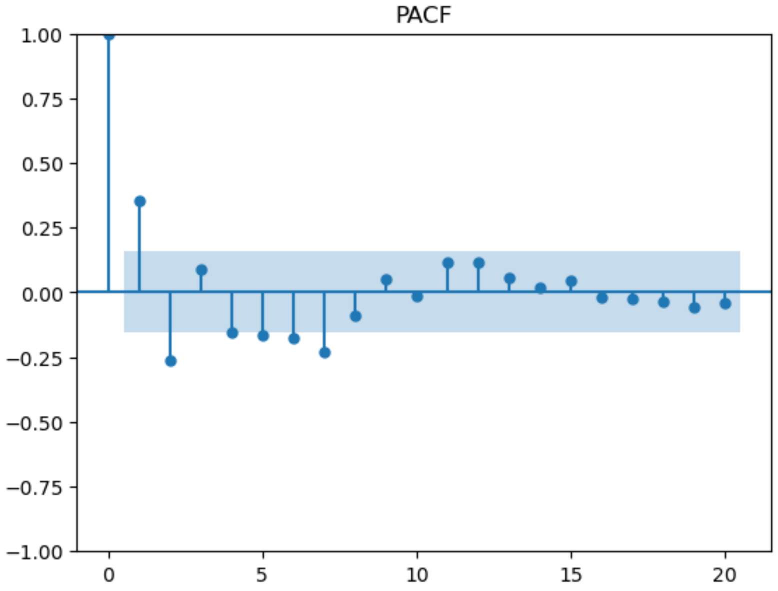

FIGURE 4.10: PACF plot

Here the value of p=0,1,2 and q=0,1.

CPI Medical Care

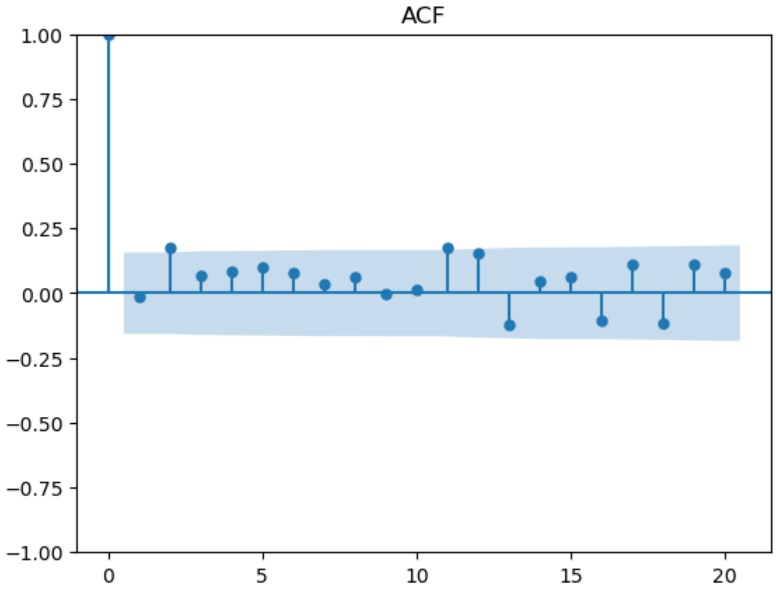

FIGURE 4.11: ACF plot

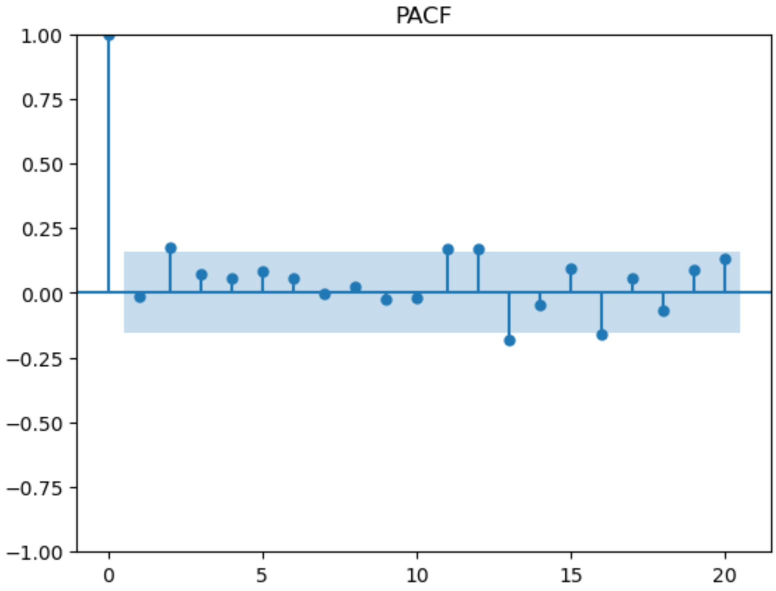

FIGURE 4.12: PACF plot

Here the value of p=0,2 and q=0,2.

CPI Clothing and Footwear

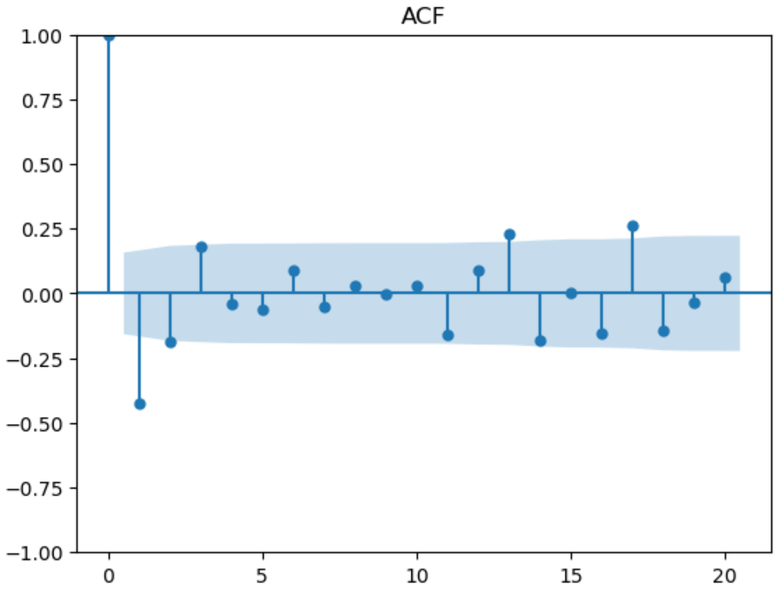

FIGURE 4.13: ACF plot

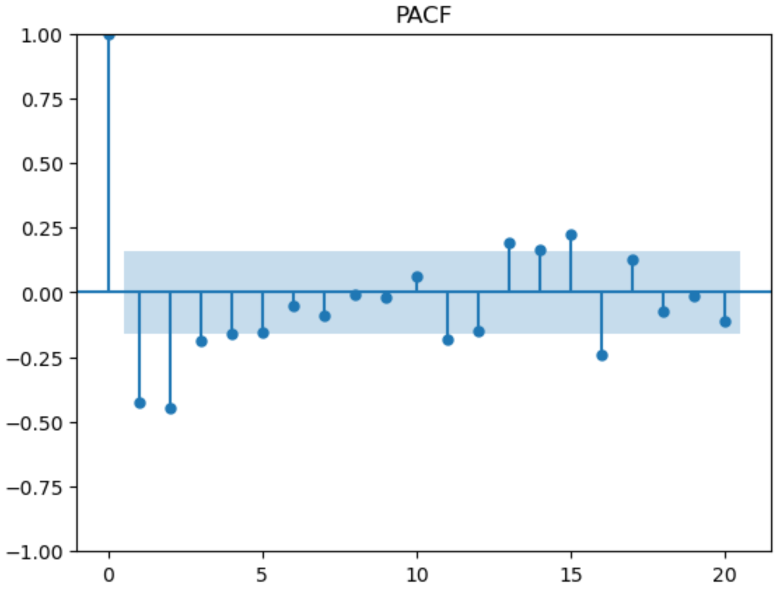

FIGURE 4.14: PACF plot

Here the value of p=0,1,2,3 and q=0,1,2.

CPI Recreation And Amusement

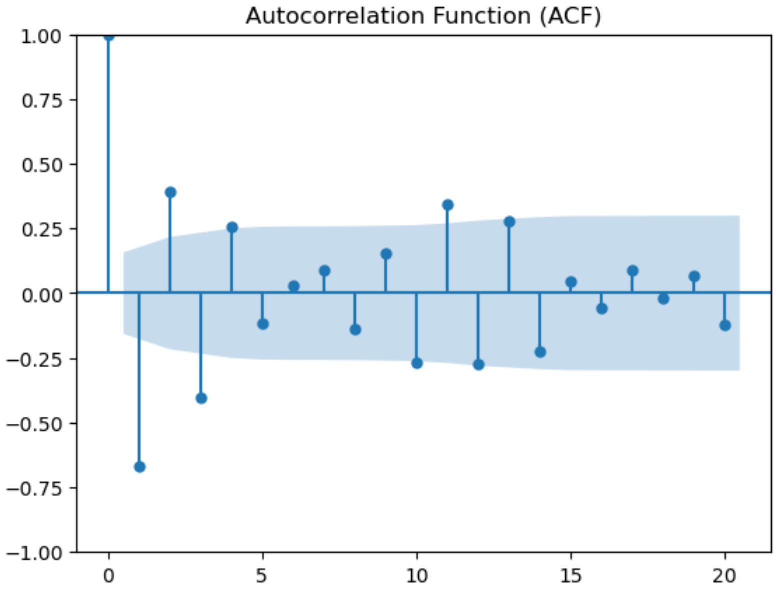

FIGURE 4.15: ACF plot

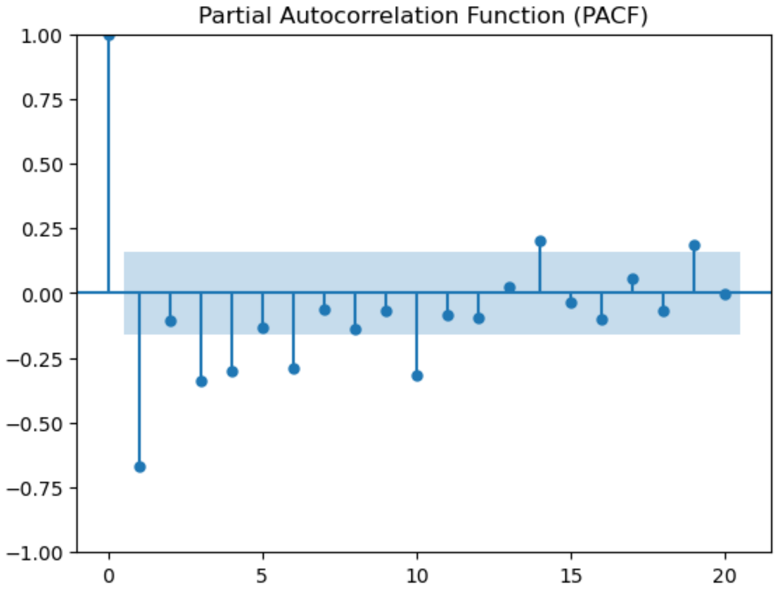

FIGURE 4.16: PACF plot

Here the value of p=0,1,3,4 and q=0,1,2,3,4.

CPI Transport and Communication

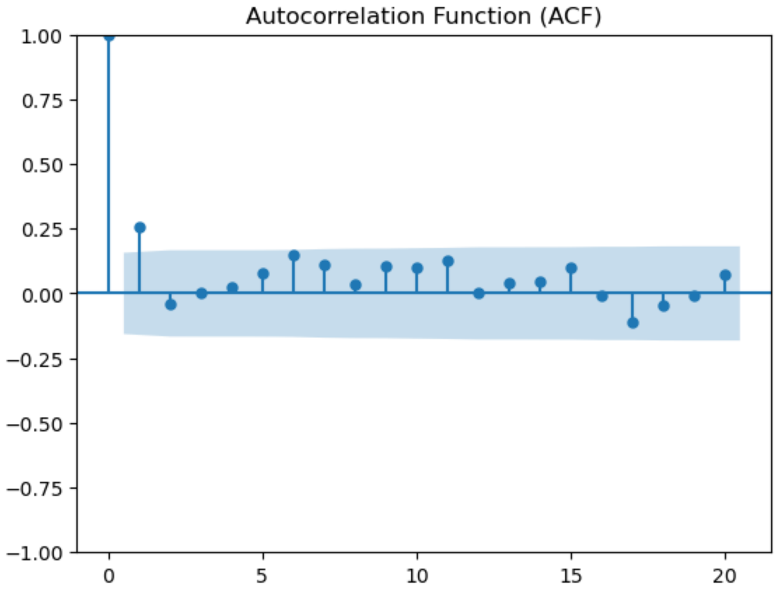

FIGURE 4.17: ACF plot

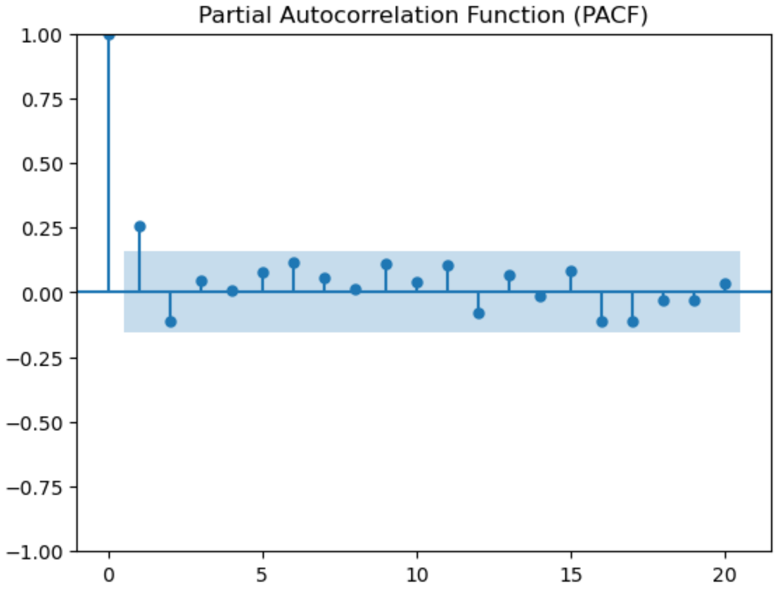

FIGURE 4.18: PACF plot

Here the value of p=0,1 and q=0,1.

4.1.4 Modelling

CPI(Total)

We fitted ARIMA(1,1,2),ARIMA(2,1,2),ARIMA(1,1,1) and SARIMA(0,1,1)(2,0,1,12) models for the data and we choose the model with least AIC value. Here we chose SARIMA(0,1,1)(2,0,1,12).

Model	AIC
ARIMA(1,1,2)	389.394
ARIMA(2,1,2)	394.267
ARIMA(1,1,1)	391.500
SARIMA(0,1,1)(2,0,1,12)	259.162

Table 4.2: AIC Values

Type	ma1	ar1(S)	ar2(S)	ma1(S)
Coefficent	0.2867	0.6873	0.2713	-0.7487
S.E	0.067	0.187	0.143	0.194

Table 4.3: Coefficients of the model

CPI Food and Beverages

We fitted ARIMA(1,1,0),ARIMA(0,1,1),ARIMA(1,1,1) and SARIMA(0,1,1)(2,0,1,12) models for the data and we choose the model with least AIC value.Here we chose SARIMA(0,1,1)(2,0,1,12).

Model	AIC
ARIMA(1,1,0)	609.803
ARIMA(0,1,1)	597.603
ARIMA(1,1,1)	599.190
SARIMA(0,1,1)(2,0,1,12)	426.283

Table 4.4: AIC Values

Type	ma1	ar1(S)	ar2(S)	ma1(S)
Coefficent	0.3748	0.7656	0.2005	-0.7867
S.E	0.059	0.188	0.144	0.189

Table 4.5: Coefficients of the model

CPI Medical Care

We fitted ARIMA(2,2,2),ARIMA(0,2,1) and SARIMA(0,2,1)(1,0,0,12) models for the data and we choose the model with least AIC value.Here we chose SARIMA(0,2,1)(1,0,0,12).

Model	AIC
ARIMA(2,2,2)	216.773
ARIMA(0,2,1)	214.001
SARIMA(0,2,1)(1,0,0,12)	196.125

Table 4.6: AIC Values

Type	ma1	ar1(S)
Coefficient	-0.9483	0.1453
S.E	0.032	0.068

Table 4.7: Coefficients of the model

CPI Clothing and footwear

We fitted ARIMA(1,1,1),ARIMA(2,1,1),ARIMA(2,1,2) and ARIMA(1,1,2) models for the data and we choose the model with least AIC value.Here we chose ARIMA(1,1,1).

Model	AIC
ARIMA(1,1,1)	141.692
ARIMA(2,1,1)	160.394
ARIMA(2,1,2)	150.906
ARIMA(1,1,2)	162.388

Table 4.8: AIC Values

Type	ar1	ma1
Coefficient	-0.0782	-0.7729
S.E	0.077	0.051

Table 4.9: Coefficients of the model

CPI Recreation and Amusement

We fitted ARIMA(4,2,2),ARIMA(4,2,3),ARIMA(3,2,2) and ARIMA(3,2,1) models for the data and we choose the model with least AIC value.Here we chose ARIMA(4,2,2).

Model	AIC
ARIMA(4,2,2)	191.201
ARIMA(4,2,3)	193.045
ARIMA(3,2,2)	195.992
ARIMA(3,2,1)	227.270

Table 4.10: AIC Values

Type	ar1	ar2	ar3	ar4	ma1	ma2
Coeff	-1.099	-0.066	-0.249	-0.352	0.090	-0.847
S.E	0.088	0.140	0.135	0.061	0.079	0.060

Table 4.11: Coefficients of the model

CPI Transport and Communication

We fitted ARIMA(1,2,1),ARIMA(0,2,1),ARIMA(2,2,2) and ARIMA(2,2,1) models for the data and we choose the model with least AIC value. .Here we chose ARIMA(0,2,1).

Model	AIC
ARIMA(2,2,1)	375.506
ARIMA(2,2,2)	407.118
ARIMA(1,2,1)	404.712
ARIMA(0,2,1)	319.969

Table 4.12: AIC Values

Type	ma1
Coefficient	-0.9083
S.E	0.035

Table 4.13: Coefficients of the model

4.1.5 Accuracy of the model

To check the accuracy of the model the dataset is split into a training set and testing set at the ratio 80:20. Then the training set data is modelled using a time series model and predicition are made on the testing set. Here we have first 126 observation in training set and 30 observation in testing set for each dataset. From the predicition made on testing set we can calculate the Root Mean Square Error(RMSE) and the Mean Absolute Error(MAE) of the model.

CPI(Total)

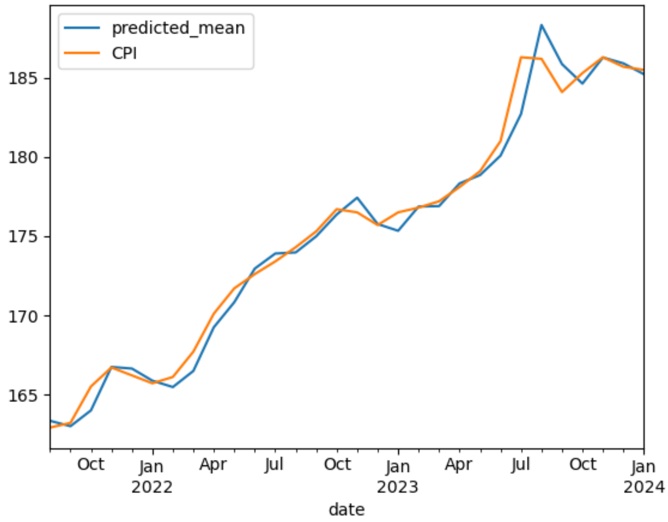

FIGURE 4.19: Graph

RMSE	MAE
1.019598	0.692291

TABLE 4.14: RMSE and MAE value

CPI Food and beverages

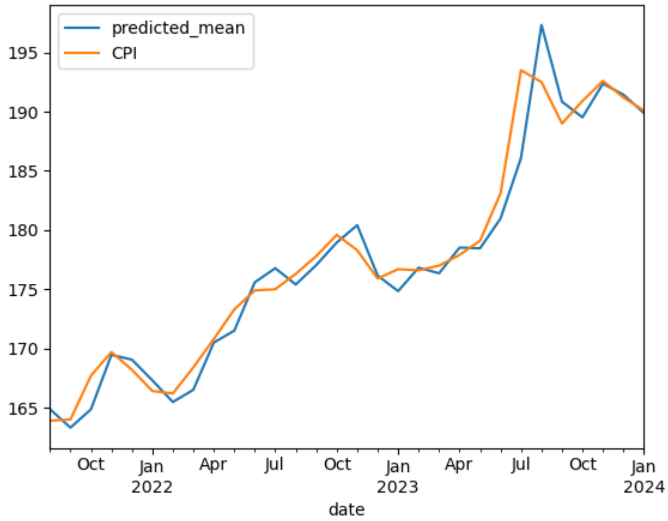

FIGURE 4.20: Graph

RMSE	MAE
2.010149	1.353062

TABLE 4.15: RMSE and MAE value

CPI Medical Care

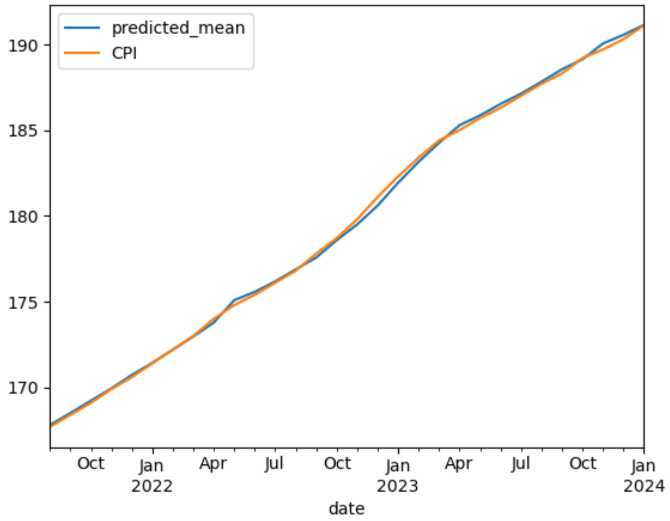

FIGURE 4.21: Graph

RMSE	MAE
0.2076333	0.1724156

TABLE 4.16: RMSE and MAE value

CPI Clothing and footwear

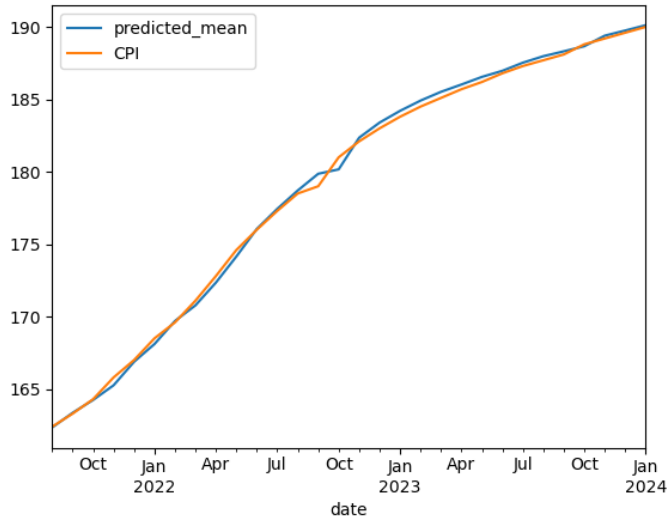

FIGURE 4.22: Graph

RMSE	MAE
0.3543980	0.290616

TABLE 4.17: RMSE and MAE value

CPI Recreation and Amusement

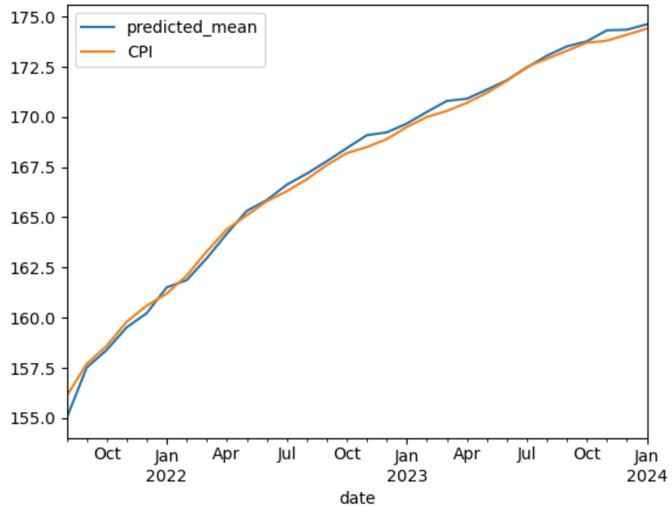

FIGURE 4.23: Graph

RMSE	MAE
0.340003	0.275694

Table 4.18: RMSE and MAE value

CPI Transport and Communication

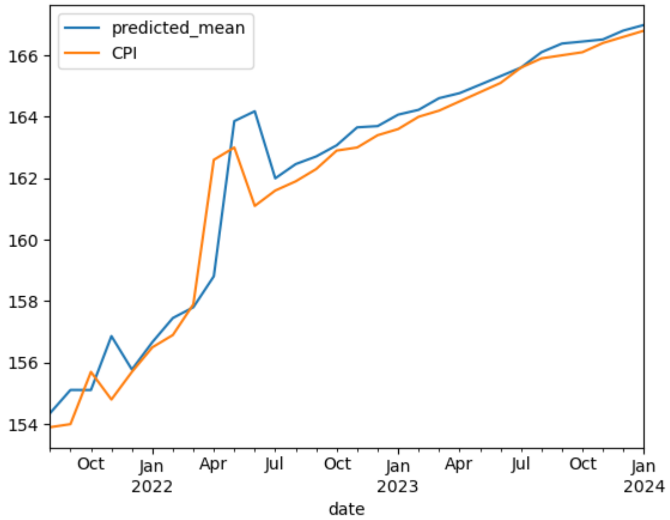

Figure 4.24: Graph

RMSE	MAE
1.051691	0.620148

Table 4.19: RMSE and MAE value

4.1.6 Residual Analysis

After fitting a time series model to the observed data, it is crucial to analyze the residuals—the deviations of the observed values from the model-predicted values. Residual analysis helps to verify the adequacy of the model fit and to check the assumption that the residuals behave like white noise.

Graph of Autocorrelation Function (ACF) of Residuals

The autocorrelation function (ACF) of the residuals is used to check for any autocorrelation that remains in the residuals after fitting the model. In an adequately fitted time series model, the residuals should be uncorrelated. The ACF plot of the residuals should ideally show all bars within the confidence interval, which typically represents no autocorrelation at different lag times. Also the ACF at lag 0 specifically measures the correlation of the series with itself, which is always equal to 1.

The following results are the ACF of Residuals of CPI and its various components.

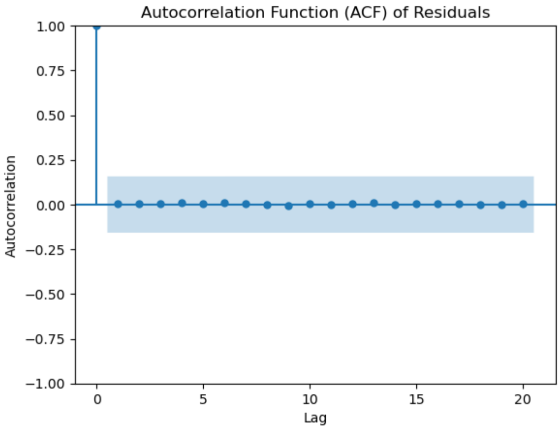

FIGURE 4.25: CPI(Total)

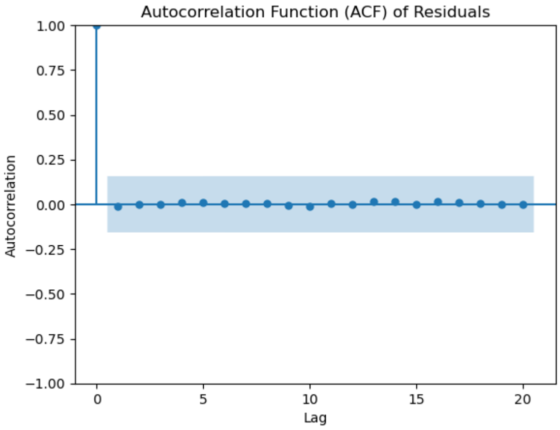

FIGURE 4.26: CPI Food and beverages

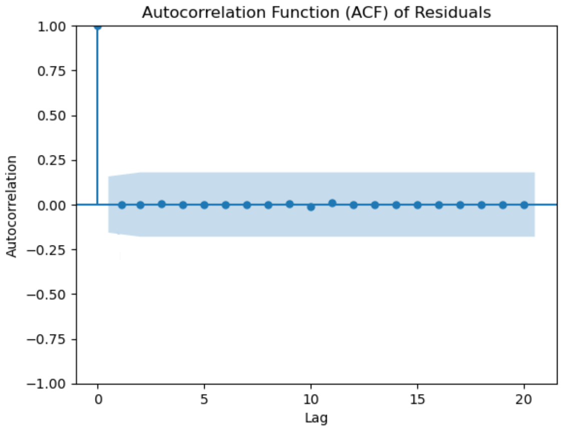

Figure 4.27: CPI Medical Care

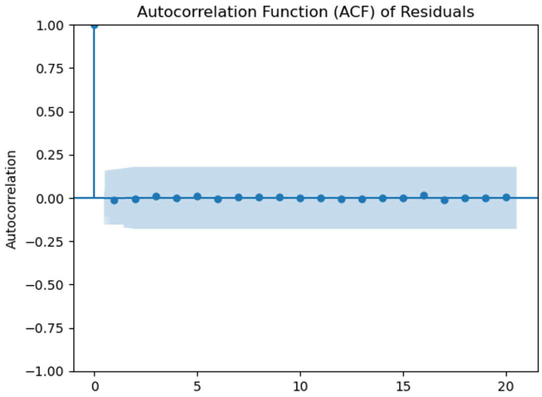

Figure 4.28: CPI Clothing and footwear

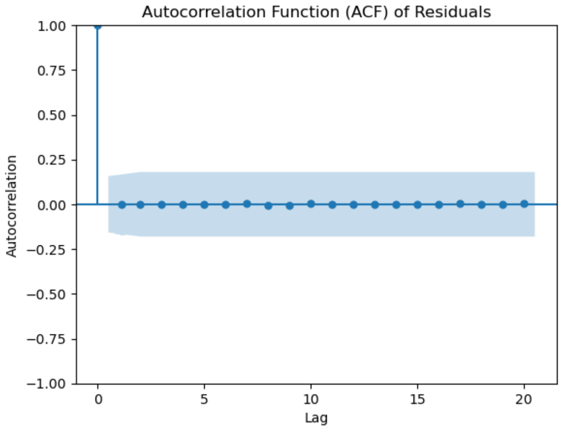

Figure 4.29: CPI Recreation and Amusement

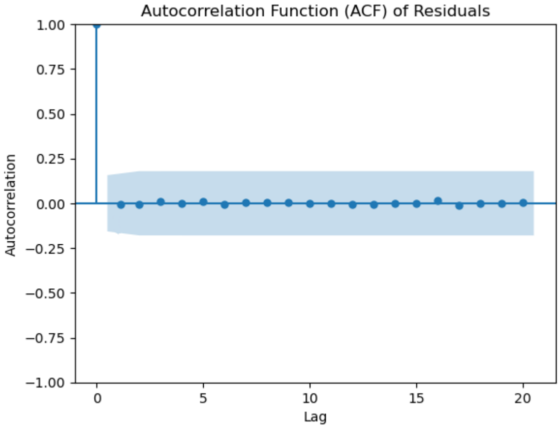

FIGURE 4.30: CPI Transport and Communication

Shapiro-Wilk Normality Test

The Shapiro-Wilk test is used to assess whether the residuals are normally distributed. Deviations from normality in the residuals could affect the validity of these statistical tests and might suggest that a different model specification or a transformation of the data is warranted.

The output of the test for CPI and its various components are given below:

W	0.93266
P-value	0.05512

TABLE 4.20: Sharpiro-Wilks test : CPI(Total)

W	0.87661
P-value	0.06597

TABLE 4.21: Sharpiro-Wilks test : CPI Food and beverages

W	0.95712
P-value	0.08914

TABLE 4.22: Sharpiro-Wilks test : CPI Medical Care

W	0.91345
P-value	0.05814

Table 4.23: Sharpiro-Wilks test : CPI Clothing and footwear

W	0.92957
P-value	0.06981

Table 4.24: Sharpiro-Wilks test : CPI Recreation and Amusement

W	0.88476
P-value	0.05618

Table 4.25: Sharpiro-Wilks test : CPI Transport and Communication

since,the p values are greater than 0.05, we have to accept null hypothesis that the residuals come from normal population.

Ljung-Box Test

The Ljung-Box test is applied to the residuals to test for overall autocorrelation up to a certain number of lags. Unlike the ACF plot, which visually inspects each lag, the Ljung-Box test provides a formal hypothesis test for the presence of residual autocorrelation at any lag up to a specified lag.

The output of the test for CPI and its various components are given below:

Chisquare	1.13812
DF	5
p-value	0.8122

Table 4.26: Ljung-Box Test : CPI(Total)

Chisquare	1.17593
DF	5
p-value	0.8437

Table 4.27: Ljung-Box Test : CPI Food and beverages

Chisquare	1.22721
DF	5
p-value	0.8762

Table 4.28: Ljung-Box Test : CPI Medical Care

Chisquare	1.19435
DF	5
p-value	0.9148

Table 4.29: Ljung-Box Test : CPI Clothing and footwear

Chisquare	1.19678
DF	5
p-value	0.8919

Table 4.30: Ljung-Box Test : CPI Recreation and Amusement

Chisquare	1.15284
DF	5
p-value	0.8013

Table 4.31: Ljung-Box Test : CPI Transport and Communication

Ljung-Box test is conducted and p-values are obtained greater than 0.05, and therefore the residuals of the fitted model are not autocorrelated.

4.1.7 Forecasting

The Following tables and graph shows forecasted values for next 11 months.

CPI(Total)

Date	Forecast
01-02-2024	185.624181
01-03-2024	186.058921
01-04-2024	187.354520
01-05-2024	188.342420
01-06-2024	189.199005
01-07-2024	190.340597
01-08-2024	191.057481
01-09-2024	191.693649
01-10-2024	192.923002
01-11-2024	193.433113
01-12-2024	192.964668

TABLE 4.32: Forecasted values

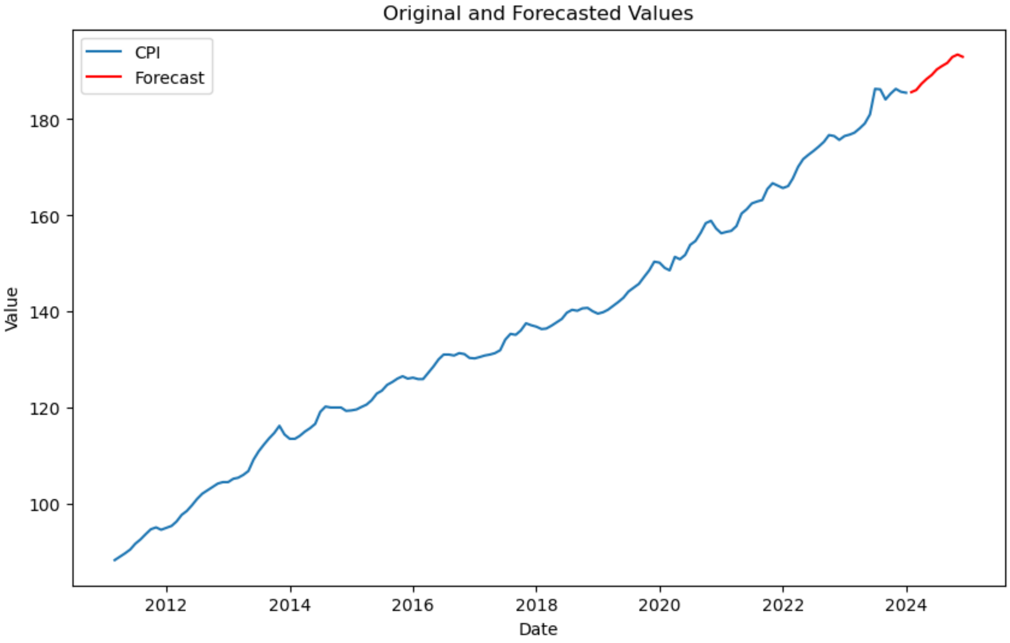

FIGURE 4.31: Graph of Forecasted values

CPI Food and beverages

Date	Forecast
01-02-2024	189.604436
01-03-2024	189.799744
01-04-2024	191.272178
01-05-2024	192.330039
01-06-2024	193.765732
01-07-2024	194.967802
01-08-2024	196.051360
01-09-2024	196.787346
01-10-2024	198.079329
01-11-2024	198.566720
01-12-2024	197.137131

TABLE 4.33: Forecasted values

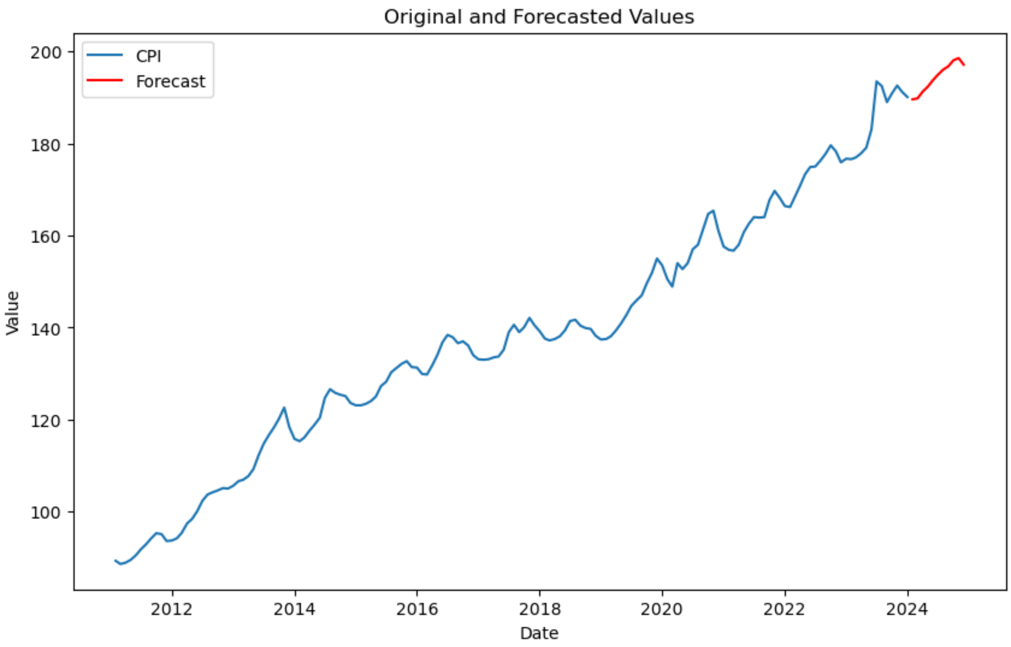

FIGURE 4.32: Graph of Forecasted values

CPI Medical Care

Date	Forecast
01-02-2024	191.920895
01-03-2024	192.731520
01-04-2024	193.501068
01-05-2024	194.280885
01-06-2024	195.050433
01-07-2024	195.830251
01-08-2024	196.610068
01-09-2024	197.379616
01-10-2024	198.179972
01-11-2024	198.939251
01-12-2024	199.708799

TABLE 4.34: Forecasted values

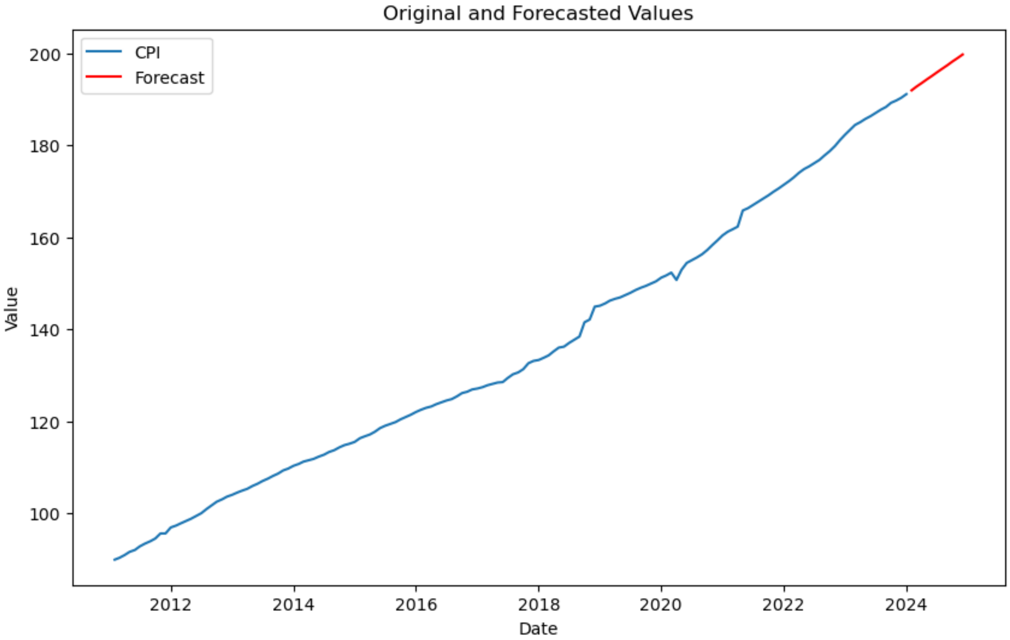

FIGURE 4.33: Graph of Forecasted values

CPI Clothing and footwear

Date	Forecast
01-02-2024	190.491171
01-03-2024	190.982275
01-04-2024	191.473378
01-05-2024	191.964482
01-06-2024	192.455585
01-07-2024	192.946689
01-08-2024	193.437793
01-09-2024	193.928896
01-10-2024	194.420000
01-11-2024	194.911104
01-12-2024	195.402207

TABLE 4.35: Forecasted values

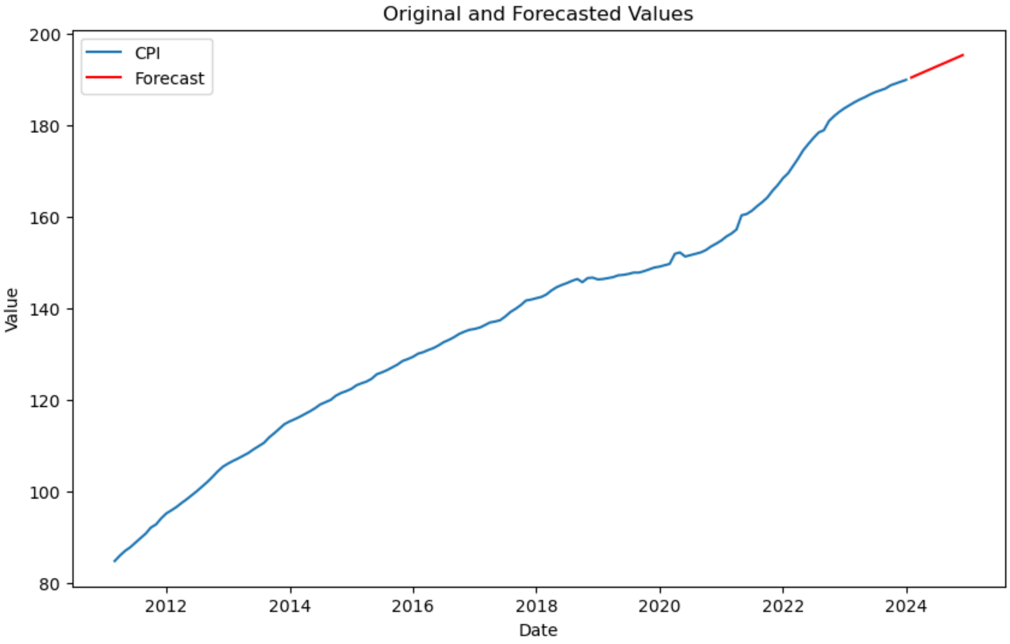

FIGURE 4.34: Graph of Forecasted values

CPI Recreation and Amusement

Date	Forecast
01-02-2024	174.951482
01-03-2024	175.465602
01-04-2024	175.933834
01-05-2024	176.392320
01-06-2024	176.785354
01-07-2024	177.275596
01-08-2024	177.681831
01-09-2024	178.193781
01-10-2024	178.593825
01-11-2024	179.096686
01-12-2024	179.497077

TABLE 4.36: Forecasted values

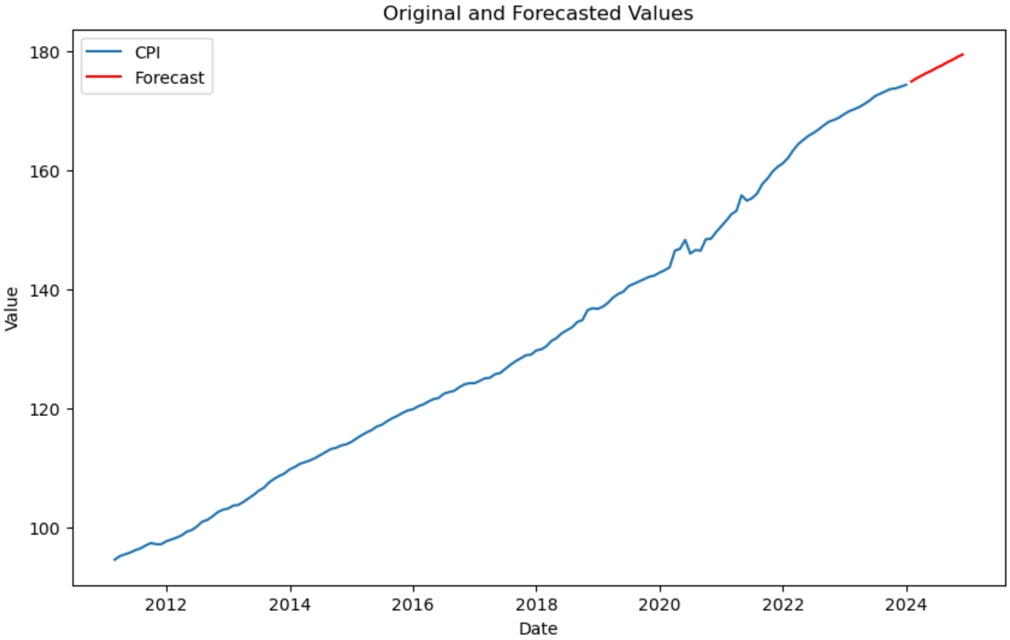

FIGURE 4.35: Graph of Forecasted values

CPI Transport and Communication

Date	Forecast
01-02-2024	167.168988
01-03-2024	167.537977
01-04-2024	167.906965
01-05-2024	168.275954
01-06-2024	168.644942
01-07-2024	169.013930
01-08-2024	169.382919
01-09-2024	169.751907
01-10-2024	170.120895
01-11-2024	170.489884
01-12-2024	170.858872

TABLE 4.37: Forecasted values

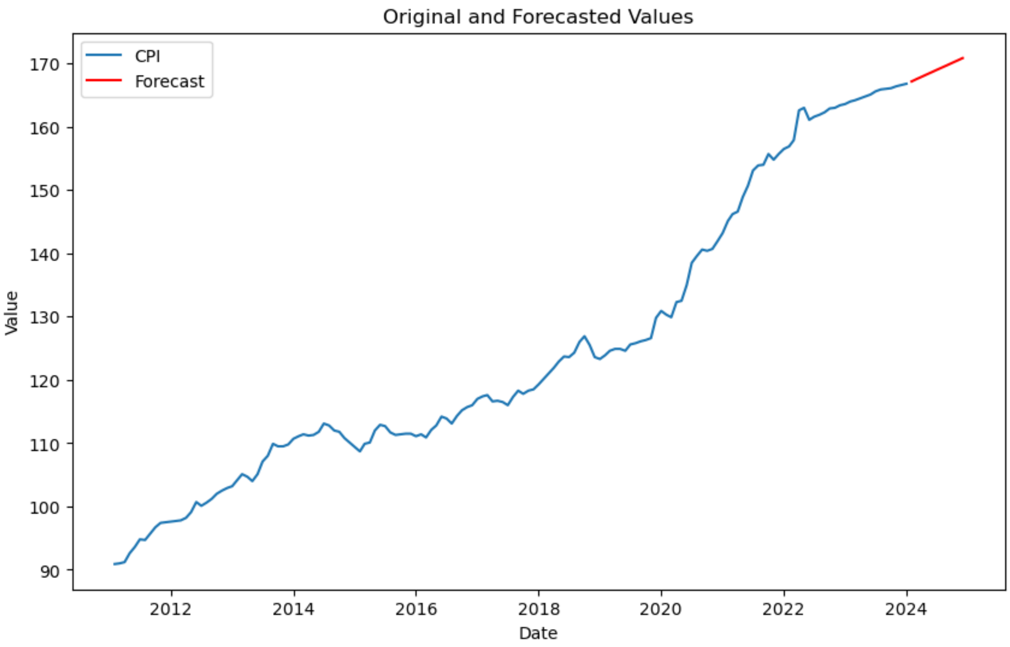

FIGURE 4.36: Graph of Forecasted values

4.2 Regression Analysis

4.2.1 Introduction

A multiple linear regression is fitted by taking CPI(Total) as response variable and the following CPI components as regressor variables:

- CPI Food and beverages
- CPI Health
- CPI Clothing and footwear
- CPI Recreation and Amusement
- CPI Transport and Communication

4.2.2 Fitted Regression Model

The estimated parameters with their p values of fitted multiple regression model is given

Coefficient	Estimate	p-value
Intercept	0.9698	0.12798
FAB(x1)	0.4639	1.3935 e-97
CAF(x2)	0.1687	4.9933e-35
TAC(x3)	0.0115	0.19251
H(x4)	0.2011	4.2749 e-12
RAA(x5)	0.1437	3.84684 e-05

Table 4.38: Coefficients of the model

4.2.3 Multicollinearity

The correlation between the variables is given

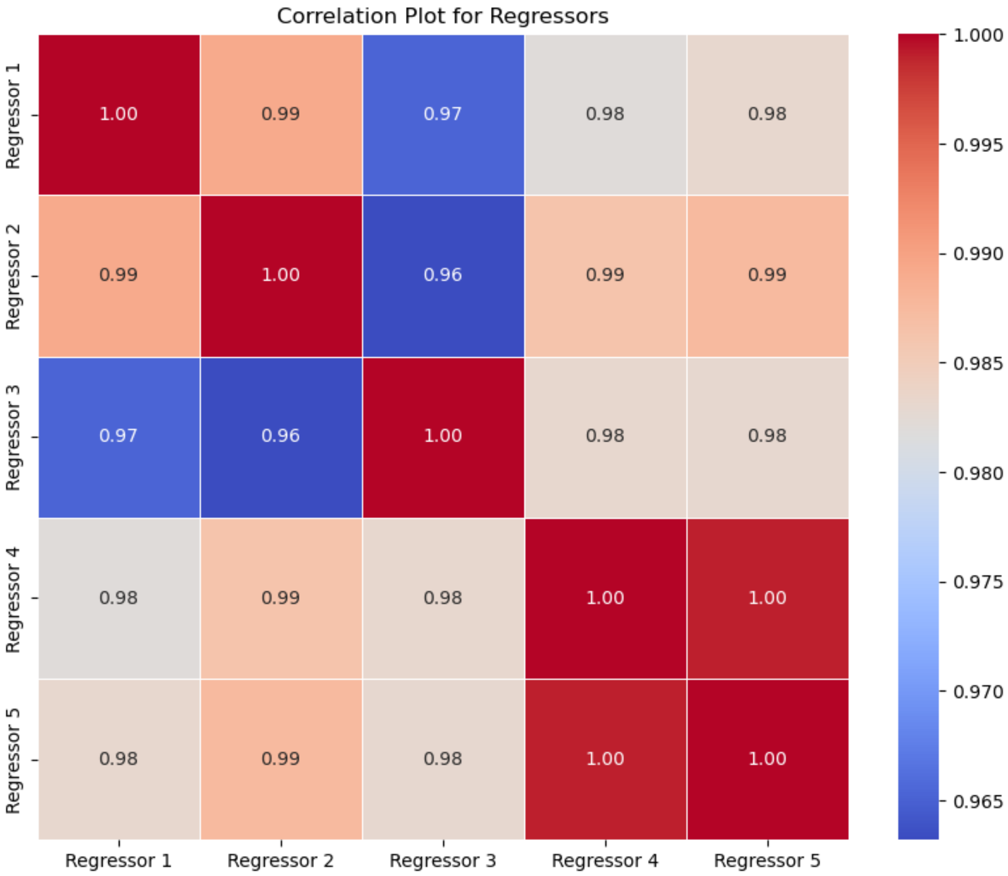

FIGURE 4.37: Correlation Plot

4.2.4 Variance Inflation Factor

The VIF values of regressors are given in table

Regressor	VIF
FAB(x1)	1402.74
CAF(x2)	1830.11
TAC(x3)	1048.96
H(x4)	1370.91
RAA(x5)	3222.72

TABLE 4.39: VIF Values

The value of VIF of every regressors is very large. So, there exist a problem of multicollinearity. Also the result from corrplot,suggest that there is high correlation between every regressors.

4.2.5 Principal Component Analysis

Multicollinearity can be removed using PCA method.
For the analysis the principal components of each regressors are found :

Regressors	PC1	PC2	PC3	PC4	PC5
FAB(x1)	0.461201	0.479030	-0.628408	0.403448	-0.012547
CAF(x2)	0.479420	0.482554	0.287898	-0.673478	-0.028976
TAC(x3)	0.375813	-0.661275	-0.496537	-0.418145	-0.009270
H(x4)	0.496762	-0.259291	0.418571	0.372999	-0.609644
RAA(x5)	0.411633	-0.182088	0.316966	0.263977	0.791992

TABLE 4.40: Table

The total variance explained by the principal components are explained using a figure

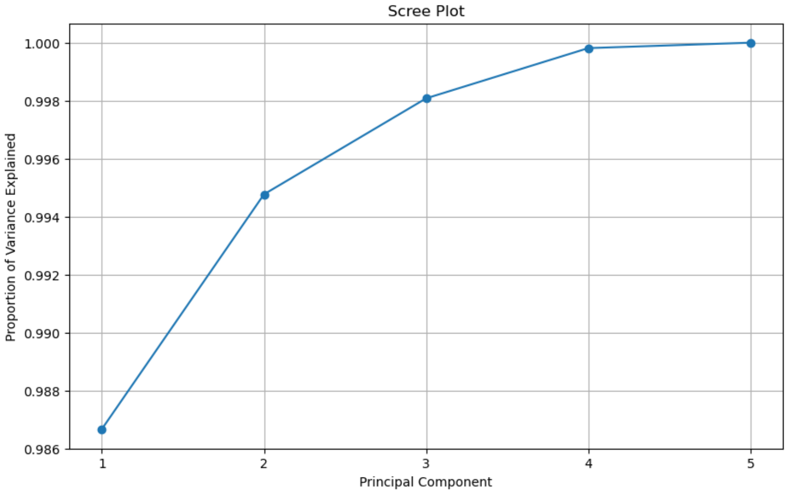

FIGURE 4.38: Screeplot

99.5 percentage of variation can be explained by first two principal components.

The first two principal components are

Regressors	PC1	PC2
FAB(x1)	0.461201	0.479030
CAF(x2)	0.479420	0.482554
TAC(x3)	0.375813	-0.661275
H(x4)	0.496762	-0.259291
RAA(x5)	0.411633	-0.182088

Table 4.41: Table

A linear regression model can be fitted using the first two principal components to solve for multicollinearity in the 1st fitted model.

4.2.6 Summary of the fitted model

The summary of fitted model is given below:

Coefficient	Estimate	p value
Intercept	165.4267	4.09818e-279
PC1	0.04583	2.59236e-192
PC2	0.02182	1.68001e-35

Table 4.42: Table 1

Residual standard error	0.7314358671040551
Multiple R Squared	0.9992479024122412
F Statistic	81045.49600908037
F-statistic p-value	2.836128737555919e-191

Table 4.43: Table 2

4.2.7 Residual Analysis

Shapiro Wilk normality test

The normality assumption can be checked using the Shapiro-Wilk normality test. Here, we want to test the null hypothesis that the residuals follow the normal distribution.

W	0.98697
P-value	0.95822

Table 4.44: Normality test

Since, p value is greater than 0.05, we have to accept null hypothesis that the residuals come from normal population.

Breusch-Pagan test

Breusch-Pagan test is used to determine if heteroscedasticity is present in a regression analysis.
H0: Homoscedasticity is present (the residuals are distributed with equal variance)
v/s
H1: Heteroscedasticity is present (the residuals are not distributed with equal variance)

F-test statistic	2.52998
F(P-value)	0.09710

Table 4.45: Non-constant variance test

since, p value is greater than 0.05, we have to accept null hypothesis that errors have constant variance.

Durbin-Watson test

Here, we want to test that residuals are uncorrelated.
The output of the test is given in table :

DW	1.70031

TABLE 4.46: Autocorrelation test

since the value of Durbin-Watson statistic is in between 1.5 and 2.5 we can say that little to no auto correlation is present in the model.

4.2.8 Fitted Regression Model

Therefore, fitted model satisfies all the required assumptions.Hence, the fitted linear regression model is

CPI(y) = 165.4267+ (0.04583)PC1+(0.02182)PC2

where PC1 and PC2 are first two principal component. Here,

PC1 = 0.461201FAB(x1)+0.479420CAF(x2)+0.375813TAC(x3)+0.496762H(x4) +0.411633RAA(x5)

and

PC2 = 0.479030FAB(x1)+0.482554CAF(x2)+ (-0.661275)TAC(x3)+(-0.259291)H(x4) +(-0.182088)RAA(x5)

4.3 Comparative Analysis of CPI(Total)

In this section, we input the forecasted values of five CPI components obtained from time series analysis into the regression model to calculate the projected Consumer Price Index (CPI) for the next 11 months.

These results are then compared with the CPI(Total) forecasts obtained directly from the time series forecasting model which is shown Table 4.32.

The comparison of the CPI(Total) forecasts is given below:

Date	Forecast using Time Series	Forecast using MLR
01-02-2024	185.624181	184.008119
01-03-2024	186.058921	185.963402
01-04-2024	187.354520	186.572279
01-05-2024	188.342420	188.073923
01-06-2024	189.199005	189.101892
01-07-2024	190.340597	189.670238
01-08-2024	191.057481	190.827133
01-09-2024	191.693649	191.419496
01-10-2024	192.923002	192.151017
01-11-2024	193.433113	192.917182
01-12-2024	192.964668	193.312801

Table 4.47: Comparing of Forecasted values

Chapter 5

Conclusion

- The Consumer Price Index forecasting would be helpful in knowing the potential impact of price changes on the cost of living, allowing for adjustments in wages, benefits, and social welfare programs to mitigate the effects of inflation on households and maintain economic stability.

- We made use of multiple linear regression to explore the relationship between CPI and factors influencing CPI in India.
  A multiple linear regression model fitted for CPI,is:

  CPI(y) = 165.4267+ (0.04583)PC1+(0.02182)PC2

  where PC1 and PC2 are first two principal component. Here,

  PC1 = 0.461201FAB(x1)+0.479420CAF(x2)+0.375813TAC(x3)+0.496762 H(x4)+0.411633RAA(x5)

  and

  PC2 = 0.479030FAB(x1)+0.482554CAF(x2)+ (-0.661275)TAC(x3)+(-0.259291) H(x4)+(-0.182088)RAA(x5)

- The comparative analysis performed on the CPI(Total) using both the time series model and the MLR model yielded results that are within a similar range which is shown in Table 4.47.

Chapter 6

Data References

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index-food-and- beverages

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index-miscellaneous- health

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index-miscellaneous- recreation-and-amusement

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index-clothing-and- footwear

- https://www.ceicdata.com/en/india/memo-items- consumer-price-index/consumer-price-index-miscellaneous- transport-and-communication

Bibliography

[1] AC Akpanta and IE Okorie. “On the Time Series Analysis of Consumer Price Index data of Nigeria–1996 to 2013”. In: *American Journal of Economics* 5.3 (2015), pp. 363–369.

[2] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 2002.

[3] Eugen Falnita and Ciprian Sipos. “A multiple regression model for inflation rate in Romania in the enlarged EU”. In: (2007).

[4] Walter Lane. “Comparing US and European inflation: the CPI and the HICP”. In: *Monthly Lab. Rev.* 129 (2006), p. 20.

[5] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[6] Robert H Shumway and David S Stoffer. “Time series analysis and it’s applications”. In: *NY: Springer-Verlang* (2000).

Appendix A

Python Code

A.0.1 Time Series Analysis

```
#loading required packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.tsa.seasonal import seasonal decompose
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.tsa.stattools import acf, pacf
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from pmdarima import auto_arima
from statsmodels.tsa.arima.model  import ARIMA
from scipy.stats import shapiro
from statsmodels.stats.diagnostic import acorr_ljungbox

# Loading CPI data

data = pd.read excel("E:\PROJECT\TIME SERIES\INPUT FILES\ CPI.xlsx")
```

```
# Visualizing the data to identify the outliers
data = pd.read_excel("E:\PROJECT\TIME SERIES\INPUT FILES\Total CPI - Copy.xlsx",
plt.figure(figsize=(10, 6))
plt.plot(data, label='CPI')
plt.title('Consumer Price Index (CPI) Over Time')
plt.xlabel('Time in months')
plt.ylabel('CPI')
plt.grid()
plt.legend()
plt.show()

# Calculate Z-scores to identify outliers
z_scores = (data - data.mean()) / data.std()

# Define a threshold for identifying outliers
outlier_threshold = 3
outliers = (z_scores > outlier_threshold) | (z_scores < -outlier_threshold)
print(outliers)

data[outliers] = np.nan
data = data.fillna(data.mean())
# Visualize the data after handling outliers
plt.figure(figsize=(10, 6))
plt.plot(data, label='CPI')
plt.title('Consumer Price Index (CPI) After Handling Outliers')
plt.xlabel('Time in months')
plt.grid()
plt.ylabel('CPI')
plt.legend()
plt.show()

# Plot the time series data
plt.figure(figsize=(12, 6))
plt.plot(data.index, data, label='CPI')
plt.title('Consumer Price Index (CPI) Over Time')
plt.xlabel('Timestamp')
plt.ylabel('CPI')
plt.legend()
plt.show()

# Perform time series decomposition
result = seasonal_decompose(data['CPI'], model='additive', period=12)

# Plot the decomposition results
plt.figure(figsize=(12, 8))

plt.subplot(4, 1, 1)
plt.plot(data['CPI'], label='Original Series')
plt.legend()

plt.subplot(4, 1, 2)
plt.plot(result.trend, label='Trend')
plt.legend()

plt.subplot(4, 1, 3)
plt.plot(result.seasonal, label='Seasonality')
plt.legend()

plt.subplot(4, 1, 4)
plt.plot(result.resid, label='Residuals')
plt.legend()

plt.tight_layout()
plt.show()

# Perform Augmented Dickey-Fuller test
result = adfuller(data['CPI'])
```

```
# Extract and print the test statistics
adf_statistic = result[0]
p_value = result[1]
critical_values = result[4]

print(f'ADF Statistic: {adf_statistic}')
print(f'p-value: {p_value}')
print('Critical Values:')
for key, value in critical_values.items():
    print(f'   {key}: {value}')

# Interpret the results
if p_value <= 0.05:
    print("Reject the null hypothesis. The time series is likely stationary.")
else:
    print("Fail to reject the null hypothesis. The time series is likely non-stat

# Perform differencing
data['diff'] = data['CPI'].diff().dropna()

# Plot the differenced series
plt.figure(figsize=(12, 4))
plt.plot(data['diff'])
plt.title('Differenced Time Series')
plt.show()

# Plot ACF
plt.figure(figsize=(12, 4))
plot_acf(data['CPI'], lags=48, title='Autocorrelation Function (ACF)')
plt.show()

# Plot PACF
plt.figure(figsize=(12, 4))
```

```
plot_pacf(data['CPI'], lags=48, title='Partial Autocorrelation Function (PACF)')
plt.show()

stepwise_fit = auto_arima(data['CPI'], trace=True,suppress_warnings=True)
stepwise_fit.summary()

#splitting data into training and testing set
print (data.shape)
train=data.iloc[:-30]
test=data.iloc[-30:]
print(train.shape,test.shape)



#fitting the model
import statsmodels.api as sm
model = auto_arima(train['CPI'], seasonal=True, m=12)
model=model.fit()
print(model.summary())

#Make Predictions on Test Set
fstart=len(train)
fend=len(train)+len(test)-1
pred = model.predict(start=fstart, end=fstart+29,type='levels')
print(pred)
pred.plot(legend=True)
test['CPI'].plot(legend=True)

# Calculate and Print RMSE
from sklearn.metrics import mean_squared_error
from math import sqrt
rmse=sqrt(mean_squared_error(pred,test['CPI']))
print(rmse)

from sklearn.metrics import mean_absolute_error
```

```
# Calculate MAE
mae = mean_absolute_error(test['CPI'], pred)

# Print  MAE
print("Mean Absolute Error (MAE):", mae)

# Determine the start and end dates for forecasting
start_date = data.index[-1] + pd.DateOffset(months=1)
end_date = start_date + pd.DateOffset(months=10)

# Forecast future values
forecast = model.predict(start=start_date, end=end_date)

print(forecast)

# Plot forecasted values
plt.plot(forecast.index, forecast, label='Forecast', color='red')
plt.xlabel('Date')
plt.ylabel('Value')
plt.title('Original and Forecasted Values')
plt.legend()
plt.show()

#Residual Analysis

# Plot ACF of residuals
plot_acf(residuals, lags=20)  # Adjust 'lags' as needed
plt.xlabel('Lag')
plt.ylabel('Autocorrelation')
plt.title('Autocorrelation Function (ACF) of Residuals')
plt.show()

# Obtain residuals from the fitted SARIMA model
```

```
residuals = model.resid
# Perform Ljung-Box test on residuals
lb_test_statistic, lb_p_value = acorr_ljungbox(residuals, lags=16)
# Print test results
print("Ljung-Box Test Statistic:", lb_test_statistic)
print("P-value:", lb_p_value)

#Sharpiro-Wilks Normality test
stat, p_value = shapiro(model.resid)

if p_value > 0.05 :
    print("Residuals appear to be normal")
else:
    print("Residuals are not normal")
```

A.0.2 Regression Analysis

```
#loading required packages
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.decomposition import PCA
import statsmodels.api as sm

# Loading CPI data
data = pd.read_excel("E:\PROJECT\REGRESSION MODEL\Data\MLR.xlsx")

regressor_data =  np.array([ data['FAB(x1)'],    # Regressor 1
              data['CAF(x2)'],    # Regressor 2
              data['TAC(x3)'],    # Regressor 3
             data['H(x4)'],    # Regressor 4
              data['RAA(x5)']   # Regressor 5
```

```
                            ]).T

target_variable = np.array(data[’CPI(y)’])
regressor_data = sm.add_constant(regressor_data)
model = sm.OLS(target_variable, regressor_data)
result = model.fit()
print(result.summary()

Rdata = {
    ’Regressor 1’: data[’FAB(x1)’],
    ’Regressor 2’: data[’CAF(x2)’],
    ’Regressor 3’: data[’TAC(x3)’],
    ’Regressor 4’: data[’H(x4)’],
    ’Regressor 5’: data[’RAA(x5)’]
}



df = pd.DataFrame(Rdata)
# Calculate the correlation matrix
corr = df.corr()
# Create a heatmap for the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap=’coolwarm’, fmt=".2f", linewidths=.5)
plt.title(’Correlation Plot for Regressors’)
plt.show()

X = np.array([ data[’FAB(x1)’],     # Regressor 1
              data[’CAF(x2)’],     # Regressor 2
              data[’TAC(x3)’],     # Regressor 3
             data[’H(x4)’],     # Regressor 4
              data[’RAA(x5)’]])   # Regressor 5

X = X.T
# Calculate VIF for each regressor
```

```
vif = [variance_inflation_factor(X, i) for i in range(X.shape[1])]
# Print the VIF values for each regressor
for i in range(0, len(vif)):  # Skip the first element (intercept)
    print(f"VIF for regressor {i}: {vif[i]:.2f}")

#Principal Component Analysis

pca = PCA()
# Fit PCA
pca.fit(X)
# Transform X to its principal components
X_pca = pca.transform(X)

#  Print explained variance ratio
print("Explained Variance Ratio:")
print(pca.explained_variance_ratio_)

# Get the principal components
components = pca.components_
components_df = pd.DataFrame(components.T, columns=[f'PC{i+1}' for i in
range(components.shape[1])])
# Display the first two principal components of each regressor
print("Principal Components of Regressors:")
print(components_df)

# Calculate the cumulative sum of explained variance ratio
cumulative_var = np.cumsum(pca.explained_variance_ratio_)

# Plot the scree plot
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(cumulative_var) + 1), cumulative_var,
marker='o', linestyle='-')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Proportion of Variance Explained')
plt.xticks(range(1, len(cumulative_var) + 1))
plt.grid(True)
plt.show()


y = np.array(data['CPI(y)'])  # Response variable

# Create PCA instance and fit it to the data
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_pca, y,
test_size=0.2, random_state=42)

# Create and fit the linear regression model using the principal components
model = LinearRegression()
model.fit(X_train, y_train)

# Predict on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)

# print the coefficients of the model
print("Coefficients:", model.coef_)

# Extract the residual standard error
residual_std_error = results.mse_resid ** 0.5

# Extract the multiple R-squared
```

```
multiple_r_squared = results.rsquared

# Extract the F-statistic and its associated p-value
f_statistic = results.fvalue
f_p_value = results.f_pvalue

# Display the extracted statistics
print("Residual Standard Error:", residual_std_error)
print("Multiple R-squared:", multiple_r_squared)
print("F-statistic:", f_statistic)
print("F-statistic p-value:", f_p_value)


#Residual Analysis



residuals = y_test - y_pred
# Perform Shapiro-Wilk normality test on the residuals
statistic, p_value = shapiro(residuals)
print(statistic,p_value)
# Print the test statistic and p-value
print("Shapiro-Wilk Test Statistic:", statistic)
print("p-value:", p_value)

# Check for significance
alpha = 0.05
if p_value > alpha:
    print("The residuals are normally distributed (fail to reject H0)")
else:
    print("The residuals are not normally distributed (reject H0)")

# Perform Breusch-Pagan test for heteroscedasticity
lm, lm_p_value, fvalue, f_p_value = het_breuschpagan(residuals, X_test_const)
```

```
# Print the test statistics and p-values
print("Lagrange Multiplier (LM) test statistic:", lm)
print("LM test p-value:", lm_p_value)
print("F-statistic of the hypothesis test:", fvalue)
print("F-test p-value:", f_p_value)

alpha = 0.05
if lm_p_value > alpha and f_p_value > alpha:
    print("There is no evidence of heteroscedasticity (fail to reject H0)")
else:
    print("There is evidence of heteroscedasticity (reject H0)")

# Perform Durbin-Watson test for autocorrelation
durbin_watson_statistic = durbin_watson(residuals)

# Print the test statistic
print("Durbin-Watson Test Statistic:", durbin_watson_statistic)

# Interpret the results
if durbin_watson_statistic < 1.5:
    print("Positive autocorrelation (possible serial correlation)")
elif durbin_watson_statistic > 2.5:
    print("Negative autocorrelation (possible serial correlation)")
else:
    print("No autocorrelation detected")
```