174 Final Project - Time Series Analysis of U.S. Inflation Data

Anoushka Menon - Email: anoushkamenon@ucsb.edu

2025-03-20

ABSTRACT

This report conducts a time series analysis on U.S. inflation rates, focusing on the post-war economic climate until present day. The study of inflation in the U.S. holds important to address recent economic instability, as well as to gain an understanding of the models that describe the progression of inflation rates over time. The methods applied for analysis include a differencing of the raw data followed by a SARIMA (p,d,q) x (P,D,Q) model, GARCH model, and forecasting. Conclusions from the SARIMA model include that the seasonal and non-seasonal components of inflation in the U.S. can both be modeled by a moving average process. A residual analysis highlighted unequal variance in inflation rates over time, which was further explored through a GARCH model. Lastly, forecasting was applied to visualize the next year's worth of inflation rates with two levels of confidence intervals applied to account for variation.

INTRODUCTION

Regarding the state of the United States economy, inflation has been one of the most studied and most relevant indicators of economic health. Patterns concerning inflation have previously been heavily correlated with unemployment, cost of common goods, wealth distribution, and financial legislation (Cogley and Sargent). Studying inflation through a statistical lens is crucial to understanding its dynamics over time, and to guiding effective economic policy. I chose to hone in on this topic for further analysis due to the fact that inflation is not only a measure of increasing prices of consumer goods, but also a reflection of labor markets, and macroeconomic trends moving forward (Cogley and Sargent).

Existing research on inflation has surrounded forecasting, applying statistical methods of non-linear machine learning, autoregressive, and structural models. Due to the volatile nature of inflation, the creation of valid generative models which can predict its future state is highly relevant today (Plakandaras et al 2016). Researchers through the process of attempting inflation forecasting have identified large datasets spanning multiple years with monthly data to be effective. Additionally, comparing structural to autoregressive methods, autoregression has proved more powerful in its forecasting abilities (Plakandaras et al 2016). The application of prior trends and related hypotheses to predict inflation are not ideal, since this may lead to a researcher's subjective choices in setting up a model, and may have a negative impact on the accuracy of future forecasted values. Contrastingly, a forecasting technique that minimizes the inclusion of subjective parameters, which relies primarily on the statistical relationship between numerical indicators of interest (such as CPI) is more beneficial (Fujiwara and Koga)

This report will apply time series analyses on U.S. inflation data, beginning with analyses of the raw data, and necessary differencing to ensure stationarity. The first set of methods applied on the transformed data include SARIMA (p, d, q) x (P, D, Q), diagnostic plots, Ljung-Box test. Follow-up analysis includes forecasting the next cycle of values (a year's worth of CPI indices) as well as a GARCH model to account for heteroscedasticity. The primary discovery of this report is a thorough analysis on the cyclic, long-term, and potential future trends which describe inflation in the United States overall.

DATA

The dataset used in this project was sourced from Kaggle, and is titled "US Inflation Dataset (1947-2023)". The dataset's time range from 1947 to mid-2023 (nearly present day) is highly significant as it follows the progression of inflation rates, through an examination of consumer price index (CPI), post World War II. The dataset focuses on collecting thorough data, with monthly data points for each year, considering the CPI on the final day of each month to ensure consistency. CPI as a measurement tool is an indicator of the average change over time in prices paid by consumers within the consumer goods and services market sector (Kaggle).

As alluded to previously, the frequency of this dataset is 12 data points per year, treating each year as a cycle, ending exactly halfway through the year 2023, and in total contains 918 data points. The two columns encompassed in the dataset are date in YYYY-MM-DD format and value in CPI, rounded to two decimal places.The high number of individual data points in this data set, paired with the extensive time range allows for a thorough investigation of not only cyclic patterns and how they adapt over time, but also any long-term upward or declining trend in CPI. This dataset was chosen due to the simplicity of measurement unit, consistency of data structure, and the relevancy of the topic at hand to current-day economic well-being. The webpage link corresponding to the datset is linked here: https://www.kaggle.com/datasets/pavankrishnanarne/us-inflation-dataset-1947-present

Further background on the source and data collection is also present on this webpage. This data was collected by the Federal Reserve Economic Data, Federal Researve Bank of St. Louis. It was collected via the CPI compiled and published by the U.S. Bureau of Labor Statistics (BLS), obtained through mass survey on items such as food, housing, and transportation. It holds importance in the study of inflation trends in the long-run which will affect the US economy on the side of suppliers and consumers.

METHODOLOGY

The original dataset adjusted to create a time series object (treating data with a frequency = 12) to proceed with thorough analyses. Once the original data was plotted to visualize any long term trends, **ACF and PACF plots** were created to identify the dependency structure of the data, and confirm whether it aligns to an AR(p) (autoregressive) process. An initial **Augmented Dickey-Fuller (ADF) test** was ran to check stationarity of the time series, which is an essential assumption to SARIMA and Box-Jenkins methodologies. Differencing was applied to the initial data, followed by a new set of ACF and PACF plots to re-evaluate the dependency structure. A follow-up ADF test was conducted to confirm stationarity before proceeding.

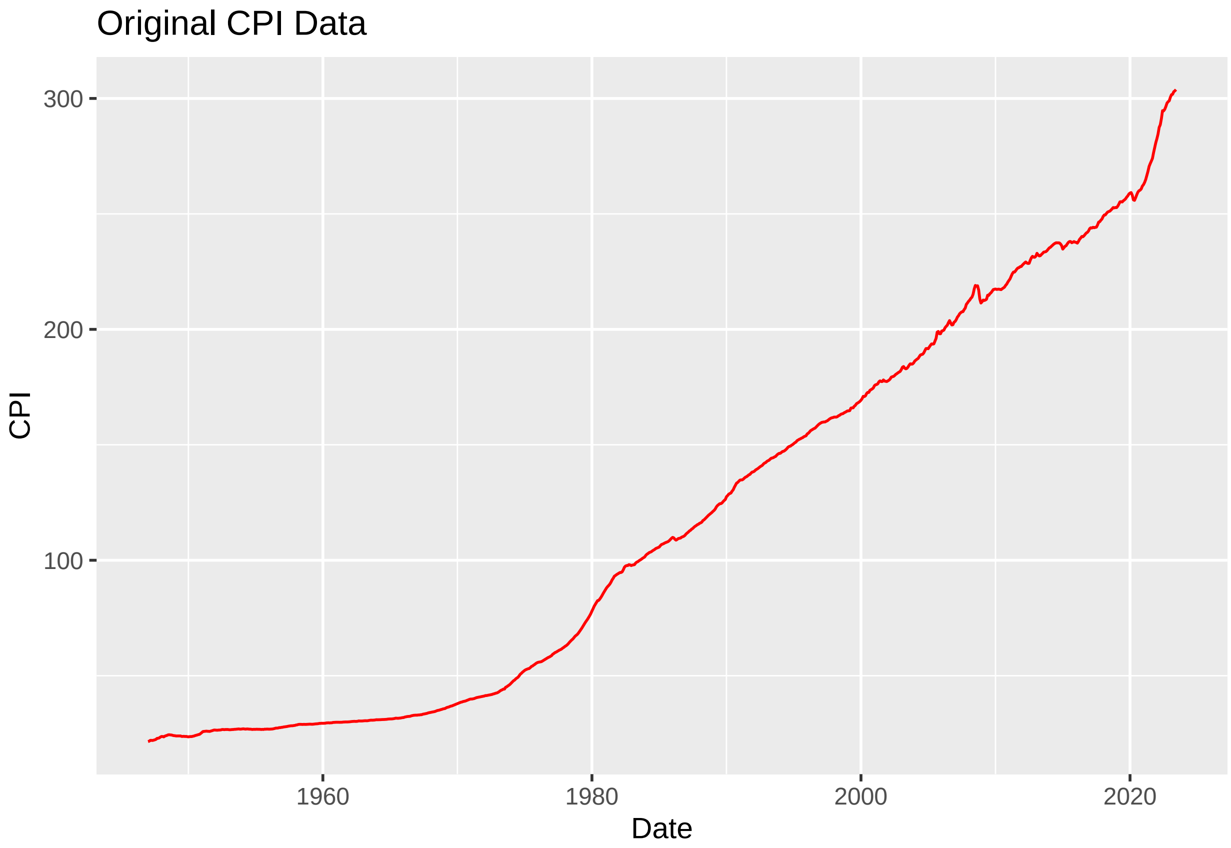
An SARIMA (Seasonal Autoregressive Integrated Moving Average) model, **SARIMA (p, d, q) x (P, D, Q)**, was conducted using the auto.arima() function in order to automatically select the best model parameters for a time series which follows seasonality. The model tests different combinations of (p, d, q) x (P, D, Q) using a stepwise procedure and applies the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to choose. **Model diagnostics** were conducted, including the following components: a *residual time series plot* to check for systematic trends in residuals, an *ACF of residuals* to check that residuals behave similar to white noise, and a *residual distribution* to assess the normality of residuals. Additionally an *Ljung-Box Test* was used to test whether residuals are independently distributed.

Follow-up analyses beyond the SARIMA and diagnostics included a **Forecast** and a **GARCH Model**. The first step in additional analysis was to forecast the next cycle (12 months) in the data to interpret the expected trend and any uncertainty in the future values, through the inclusion of confidence intervals. Lastly, the GARCH Model was added to model volatility in the residuals of the SARIMA model. This was vital for thorough analysis, as SARIMA models assume constant variance, but economic data such as CPI often contain periods of high volatility.

RESULTS

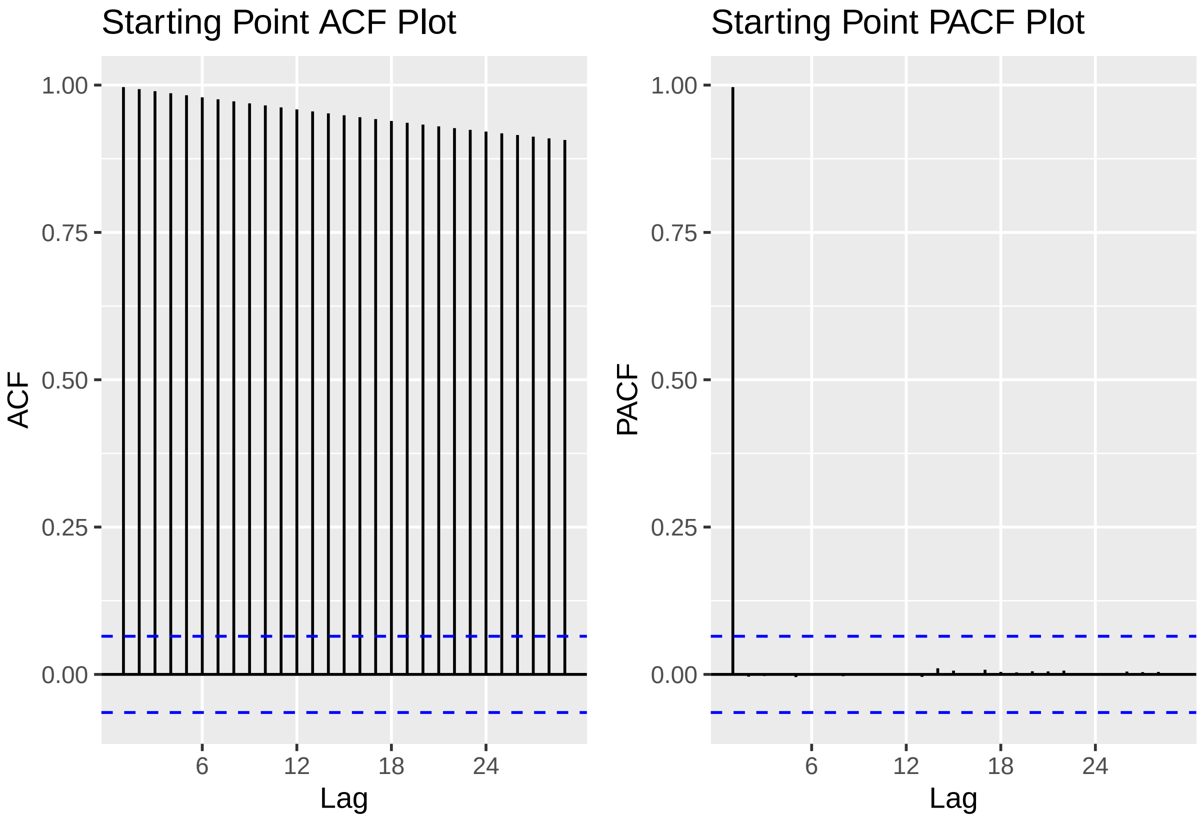
Run Analysis & Transformations on the Original Data

Plot Original Data



This plot shows the Consumer Prince Index (CPI) over time fro 1947 onwards, and illustrates a clear upward trend. The CPI, which measures changes in the price level of relevant subsets of consumer goods and services has steadily increased over this period, with very minor fluctuations within yearly cycles. The growth trend appears exponential in certain sections of time, particularly in the 1970s and 2020s, which may suggest inflation trends are not linear in some periods, with more rapid growth than others, but is always increasing.

Initial ACF, PACF Plots & Analysis



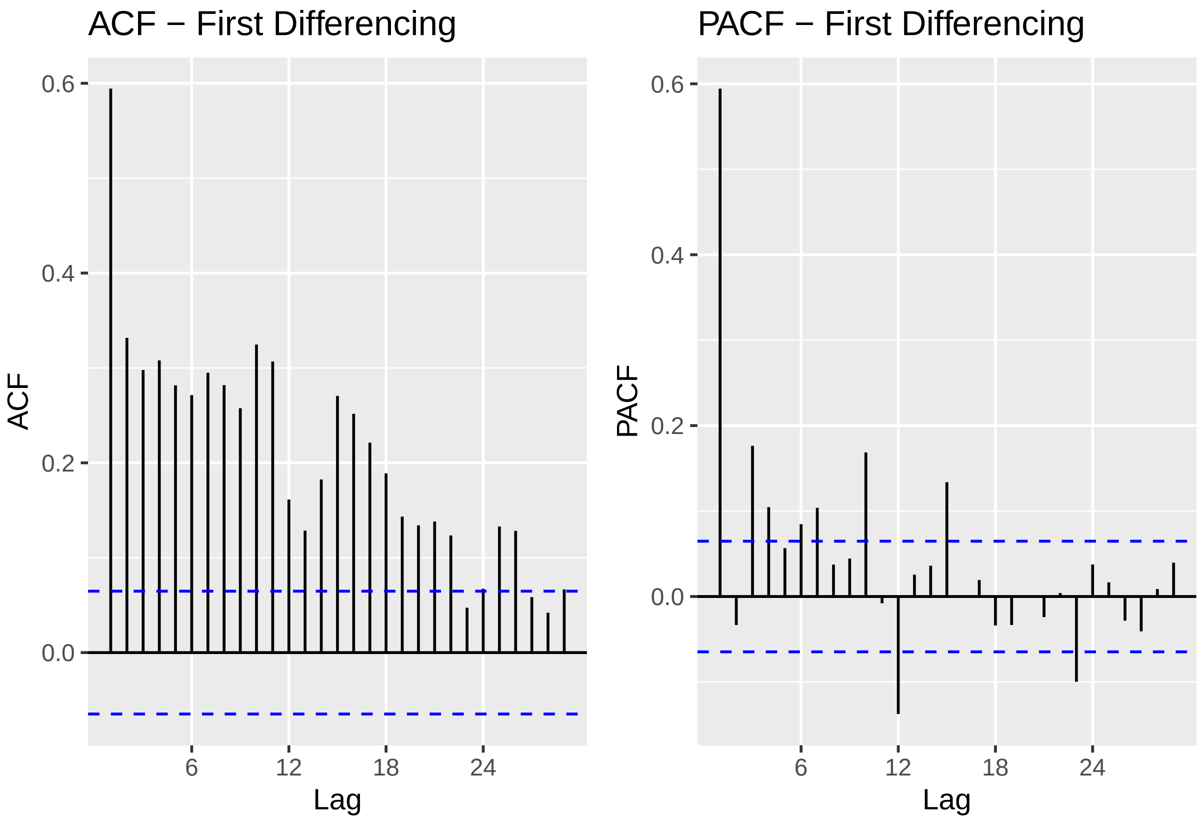
The ACF plot (on the left panel) shows strong positive autocorrelation at all lags with values near 1.0, with a slow decay which could suggest that the time series is non-stationary. The PACF plot on the right has a significant spike at lag 1 and spikes near-zero for the remaining lags, which suggests an AR(1) process and strong autoregressive behavior. It also confirms the conclusion of non-stationarity from the ACF plot.

Table 1: ADF Test Results

	Statistic	P_Value	Lag_Order	Alternative_Hypothesis
Dickey-Fuller	-1.07806	0.9261261	9	Stationary

The Augmented Dickey-Fuller (ADF) test is used to determine whether a time series is stationary or exhibits a unit root, indicating non-stationarity. Table 1 presents the results of the ADF test, with a test statistic of -1.0781, and a p-value pf 0.9261. This p-value fails to reject the null hypothesis H_0 that the data is non-stationary. This suggests that the time series exhibits strong autocorrelation and likely contains a unit root, which implies that it lacks stationarity. The next step is to apply first order differencing (d=1) before proceeding with modeling.

Apply Differencing & Re-Analyze



The ACF and PACF plots of data that has undergone first-order differencing help re-assess stationarity and a model strucure. The ACF plot on the left still shows a decay, but at a less gradual rate than previously, suggesting improvement, but some remaining level of autocorrelation. The PACF plot exhibits a cut-off after lag 1, and aligns less rigorously with the AR(1) component that was previously identified.

Table 2: ADF Test Results for Differenced Data

	Statistic	P_Value	Lag_Order	Alternative_Hypothesis
Dickey-Fuller	-5.521318	0.01	9	Stationary

Table 2 presents the results of the ADF test on differenced data, with a test statistic of -5.5213, and a p-value pf 0.01. This p-value is well below the critical value of 0.05, and provides statistical evidence to reject the null hypothesis H_0 that the data is non-stationary or has a unit root. This suggests that first order differencing was successful in making the series stationary, and that the series is appropriate for further modeling and forecasting.

Fit SARIMA (p, d, q) x (P, D, Q)

Table 3: SARIMA Model Order

Component	Order
Non-seasonal (p,d,q)	(0,1,3)
Seasonal (P,D,Q,s)	(0,0,2)[12]

Table 4: SARIMA Coefficients & SEs

	Estimate	Std_Error
ma1	-0.4299	0.0336
ma2	-0.3804	0.0329
ma3	-0.0640	0.0338
sma1	-0.1622	0.0351
sma2	-0.0596	0.0345

Table 5: Model Selection Criteria

Metric	Value
Log Likelihood	-357.1138
AIC	726.2275
AICc	726.3199
BIC	755.1476
Sigma^2	0.1281

Model Order Interpretation:

The selected SARIMA model has the following components. The non-seasonal (p,d,q) element is (0,1,3). This indicates that the model applies first order differencing to ensure stationarity (d=1). The MA term q=3 suggests the presence of three moving average terms, which imply the model has short term dependencies in residual terms. No AR terms (p=0) indicates that past values do not highly correlate with future values, and might not contribute to future-state predictions.

The seasonal (P,D,Q) component is (0,0,2), with a period of s=12, which aligns with the monthly progression of data points. The term d=0 suggests no seasonal differencing was applied. The term q=2 corresponds to 2 seasonal MA terms in the model, implying that variations in previous cycles of data may impact the current state. Lastly, the term p=0 remains, suggesting again that there are no AR terms, and that past seasonal trends likely do not influence future ones.

Coefficient Analysis:

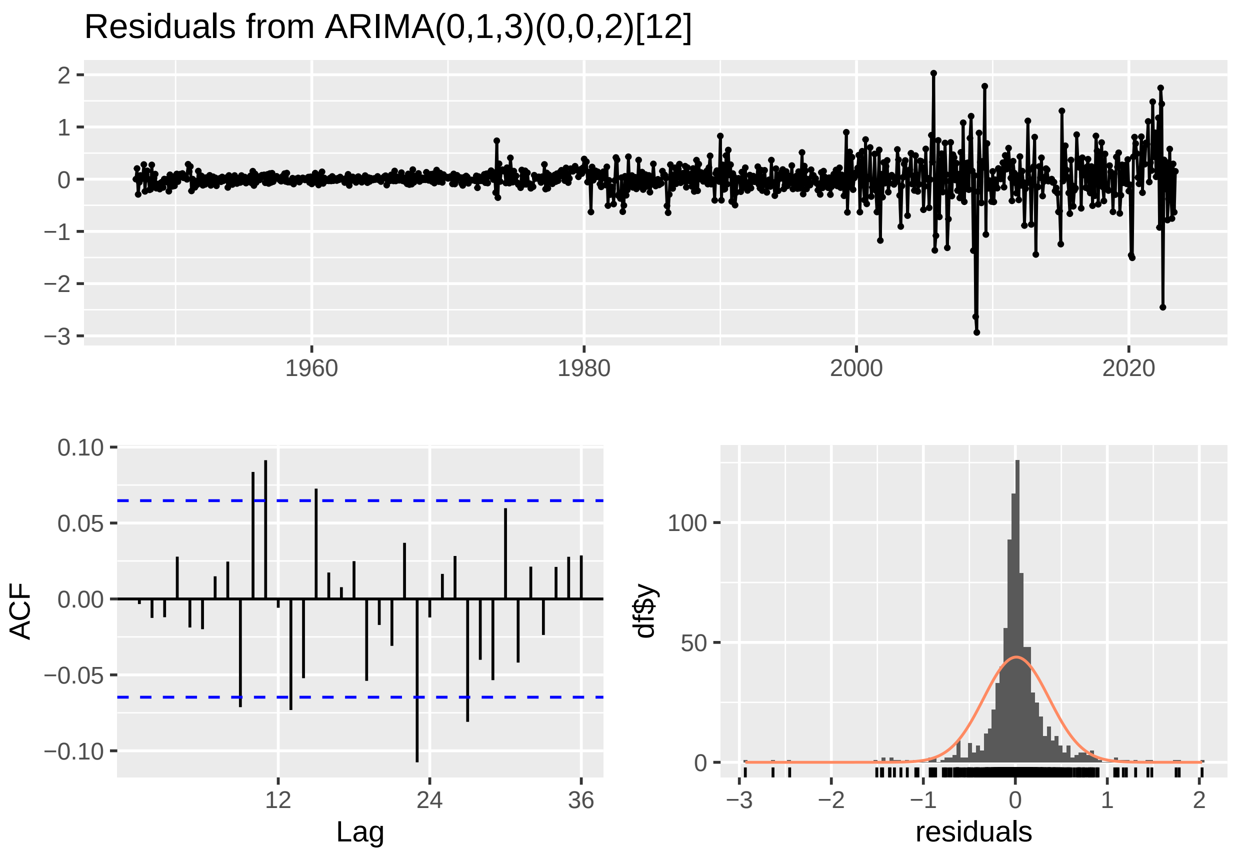
The moving average terms for the non-seasonal component of this SARIMA model are as follows: MA(1) = -0.4299 (SE = 0.0336), MA(2) = -0.3804 (SE = 0.0329), and MA(3) = -0.0640 (SE = 0.0338). The first two MA terms appear to be significant, as their values are large compared to their standard errors, while the third MA term is comparatively smaller. This indicates the more recent error (seasonal periods of increased variance) may influence future values more than build up of previous error.

The moving average terms for the seasonal components are as follows: seasonal MA(1) = -0.1622 (SE = 0.0351), and seasonal MA(2) = -0.0596 (SE = 0.0345). Since the first MA term is larger, seasonal fluctuations in recent years have a larger effect than previous years within a season.

Model Selection Criteria

The AIC and BIC are respectively 726.2275 and 755.1476. Given that the SARIMA model undergoes a stepwise iterative process to land on the optimal model, these AIC and BIC values are likely relatively small compared to alternate models. The σ^2 value of 0.1281 is also small, which shows that the model captures the remaining variation in inflation over time.

Diagnostic plots for SARIMA



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,3)(0,0,2)[12]
## Q* = 51.088, df = 19, p-value = 9.048e-05
##
## Model df: 5.   Total lags used: 24
```

Table 6: Ljung-Box Test

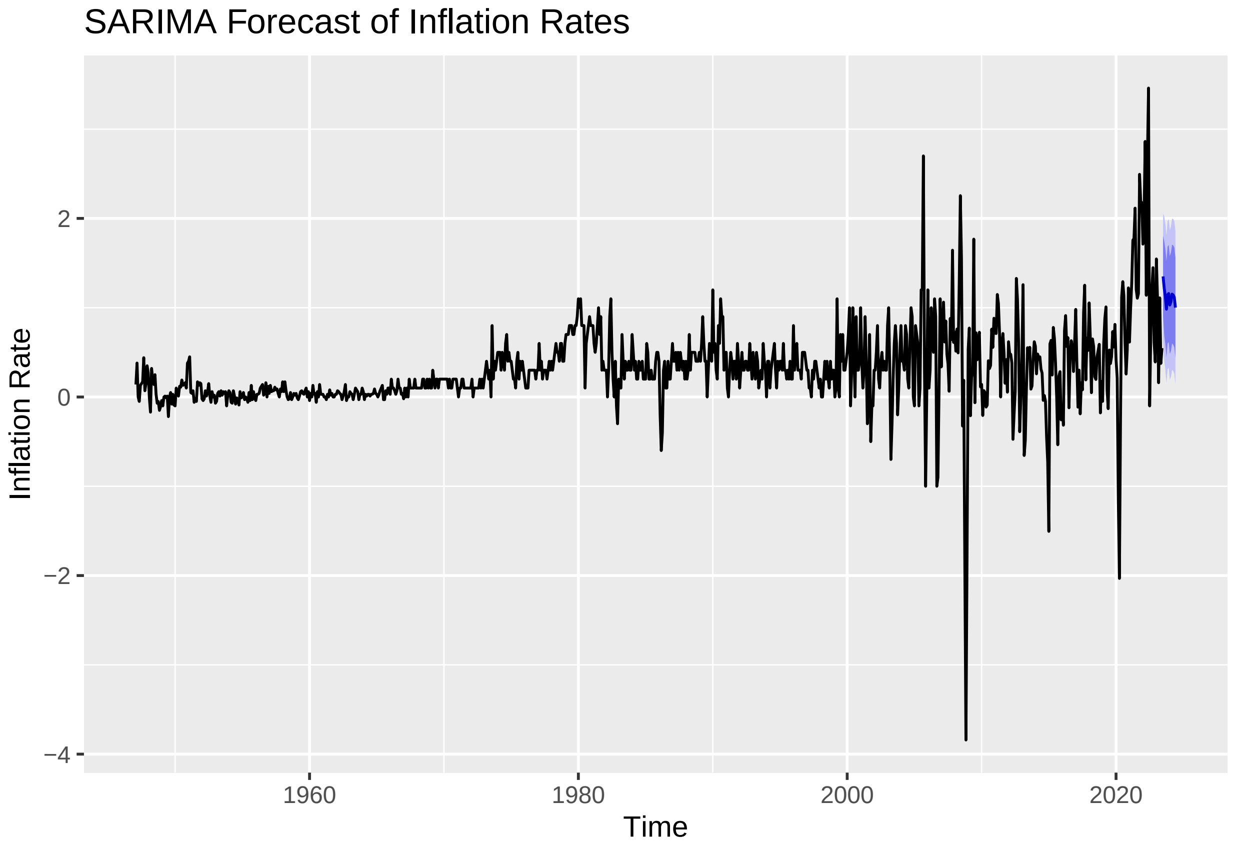
Metric	Value
Q*	51.088
Degrees of Freedom (df)	19
p-value	9.048e-05
Model df	5
Total Lags Used	24

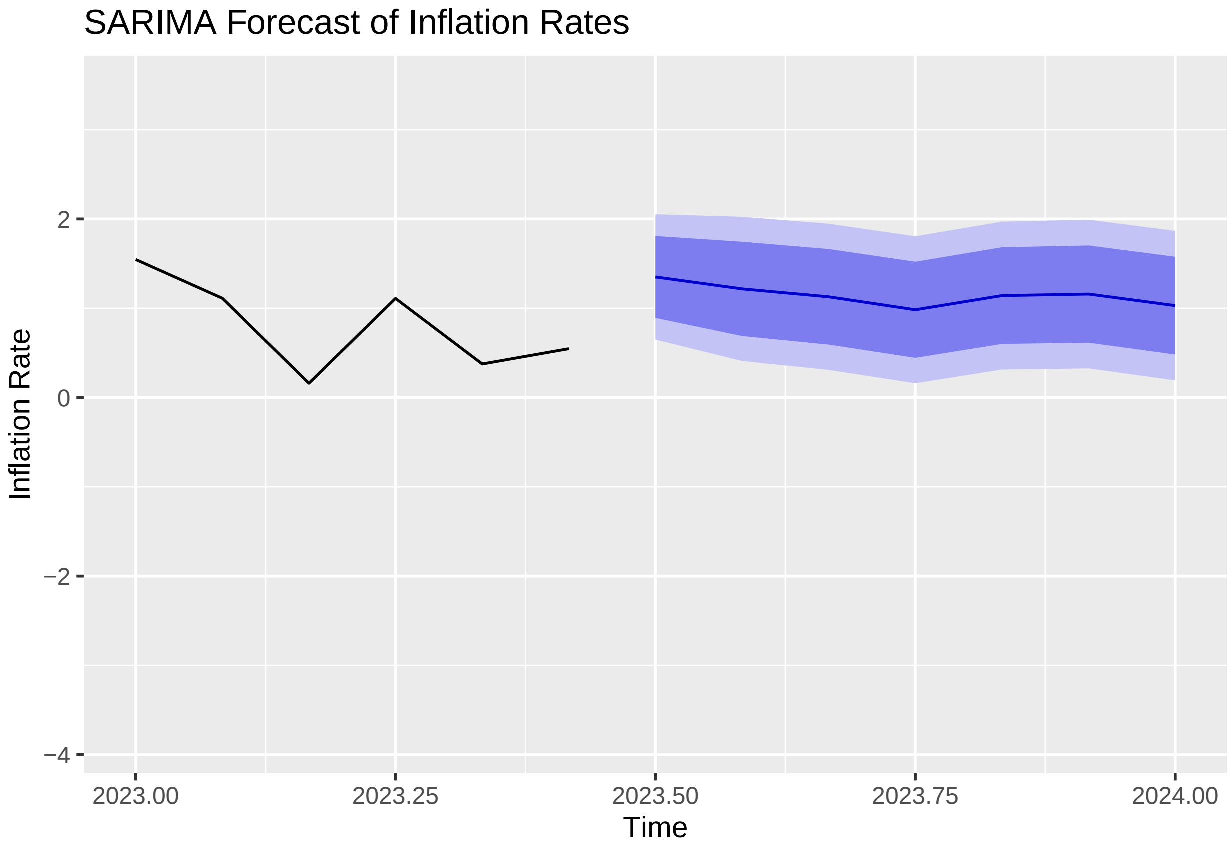
The diagnostic plots above analyze the residuals of the ARIMA(0,1,3)(0,0,2)[12] model, and contain a residual time series plot, ACF plot, and histogram of residuals. The **residual time series plot** displays that residuals fluctuate equally above and below zero for the most part, but variance is grows larger over time, appearing significantly larger after the year 2000, with the greatest fluctuation near 2008. This variance could be due to historical events such as recessions, and motivated by overall economic volatility after the year 2000. The **ACF plot** checks whether residuals show significant autocorrelation, based on the significance bounds in blue. Given that bars at a few lags exceed the bounds, this shows systematic error in the residuals. Lastly, the **histogram of residuals** represents the center of residuals around zero, but confirms that the variance of residuals extends well beyong the reference line (representing a normal distribution). This may mean that residuals are also not normally distributed.

The Ljung box-test for residuals is based on the null hypothesis H_0 that residuals are independently distributed, and the alternative hypothesis H_1 that residuals show autocorrelation. The small p-value of 9.048e-05 (< 0.05) provide sufficient statistical evidence to reject the null hypothesis. This indicates that the residuals are not random, or identically distributed, and they do exhibit autocorrelation. The next step to address this would be to conduct a GARCH model.

Additional Analyses

Forecast Next Cycle





The plots above were created to visualize a forecasting of the next 12 values in the cycle (the next year of CPI's). The first plot shows the historical data from the dataset from 1947 to mid-2023 alongside the predictions for the next 12 values, appearing in blue. According to the SARIMA model results above, the predictions have been made on the basis of non-seasonal and seasonal MA (moving average) terms.

The second plot zooms into the most recent data for a better visualization of the forecast with its point estimates, along with its 80% (dark blue) and 95% (light blue) confidence intervals. The estimates for the upper and lower bounds of each confidence interval, along with the point estimates for forecasted values are detailed in Table 7 below.

Table 7: Forecasted Inflation Rates

Month	Point_Forecast	Lo_80	Hi_80	Lo_95	Hi_95
Jul 2023	1.3503295	0.8915985	1.809060	0.6487610	2.051898
Aug 2023	1.2168431	0.6887946	1.744892	0.4092625	2.024424
Sep 2023	1.1281929	0.5930224	1.663363	0.3097203	1.946665
Oct 2023	0.9830364	0.4447669	1.521306	0.1598243	1.806248
Nov 2023	1.1420375	0.6006868	1.683388	0.3141131	1.969962
Dec 2023	1.1590879	0.6146733	1.703502	0.3264777	1.991698
Jan 2024	1.0295999	0.4821387	1.577061	0.1923302	1.866869
Feb 2024	1.0660355	0.5155445	1.616526	0.2241322	1.907939
Mar 2024	1.1523668	0.5988625	1.705871	0.3058551	1.998878
Apr 2024	1.1420167	0.5855156	1.698518	0.2909217	1.993112
May 2024	1.1169416	0.5574596	1.676424	0.2612877	1.972595
Jun 2024	1.0009436	0.4384966	1.563391	0.1407551	1.861132

GARCH Model

Table 8: GARCH(1,1) Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.002687	0.003539	0.759447	0.447586
omega	0.000216	0.000103	2.100799	0.035659
alpha1	0.155386	0.021060	7.378204	0.000000
beta1	0.843614	0.022563	37.389993	0.000000

The GARCH(1,1) mode was implemented to explain volatility within the time series, which is often present in financial and economic data such as this set. The mean parameter, $\mu = 0.002678$, with p-value of 0.4476 is not statistically significant, suggesting the average variance is not significantly different from zero. The omega parameter, $\omega = 0.000216$, with a p-value of 0.0357 is statistically significant, and represents that the baseline variance is non-zero. The alpha1 parameter, $\alpha 1 = 0.1554$, with a p-value < 0.0001, is highly statistically significant, and represents a significant effect of short-term volatility. Lastly, the beta1 parameter, $\beta 1 = 0.8436$, with a p-value of < 0.0001 is also highly significant, indicating that long-term volatility also affects future values. The $\alpha 1$ + $\beta 1$ sum is very close to 1.000, which indicates long-lasting effects of variance in inflation rates.

CONCLUSIONS & FUTURE STUDY

This report explores U.S. inflation through a time series analysis of existing data from 1947 to 2023. This is a key topic of relevance at present in the United States, and has key implications in economic policy, as well as public well-being, employment, and in the long-run, wealth distribution. The results from this experiment generally align with previous research which suggests that volatile markets will impact the near future inflation forecasts (Plakandaras et al 2016). The dataset applied in this experiment detailed inflation rates for 12 equidistant points for each year encompassed, and used CPI as the indicator on which subsequent analysis and models were built.

The original CPI data indicates an overall upward trend, with periods of exponential growth, and minor intermediate fluctuations, aligning with real-world expectations. An analysis consisting of ACF, PACF plots before and after a first differencing showed that once differencing was applied, the model still contained autocorrelation, indicating some systematic variance in the dataset. These plots, in addition to the output from a SARIMA model inndicated that both the seasonal and non-seasonal components of the inflation time series can be characterized by a moving average (MA) process. Specifically, the non-seasonal (p,d,q) was detailed to be (0,1,3), and the seasonal (P,D,Q) was (0,0,2) with a periodicity of 12.

Residual analysis further shows a high level of fluctuation of residuals, which although equally distributed above and below zero, show an exponential and systematic increase in variance post-2000. The autocorrelation present in residual led to the next step in analysis being a GARCH model, to further discuss the variance. The coefficients from this model aligned with the MA model suggested earlier, and confirmed that recent volatility has a more significant effect on future values, and previous volatility has a diminishing effect over time.

Future study to build on this analysis could involve refining the model to account for structural breaks, particularly in response to major economic events such as the 2008 financial crisis and the COVID-19 pandemic. Additionally, doing combinations of a multi-variable analysis with factors such as interest rates or unemployment rates could provide alternate models for comparison. Ideal next steps would be to incorporate machine learning approaches, which have the ability to capture complex patterns and provide a robust forecast.

REFERENCES

Cogley, T., & Sargent, T. J. (2001). Evolving post-world war II US inflation dynamics. NBER macroeconomics annual, 16, 331-373.

Fujiwara, I., & Koga, M. (2004). A statistical forecasting method for inflation forecasting: hitting every vector autoregression and forecasting under model uncertainty. Monetary and Economic Studies, 22(1), 123-142.

Gogas, P., Papadimitriou, T., Plakandaras, V., & Gupta, R. (2017). The informational content of the term-spread in forecasting the us inflation rate: A nonlinear approach. Available at SSRN 2990336.

"US Inflation Dataset (1947 - 2023)." Kaggle, 30 July 2023, www.kaggle.com/datasets/pavankrishnanarne/us-inflation-dataset-1947-present.

APPENDIX (R Code)

```r
# load necessary libraries
suppressMessages(library(forecast))
library(tseries)
library(ggplot2)
library(patchwork)
library(knitr)
suppressMessages(library(dplyr))
suppressMessages(library(zoo))
suppressMessages(library(rugarch))
```

```r
# Load the data
data <- read.csv("US_inflation_rates.csv")
data$date <- as.Date(data$date, format="%Y-%m-%d")

# Convert 'date' to time series object assuming frequency - 12
start_year <- as.numeric(format(min(data$date), "%Y"))
start_month <- as.numeric(format(min(data$date), "%m"))

# Create the time series object
ts_data <- ts(data$value, frequency=12, start=c(start_year, start_month))
```

```r
# Load the data
data <- read.csv("US_inflation_rates.csv")
data$date <- as.Date(data$date, format="%Y-%m-%d")

# Convert 'date' to time series object assuming frequency - 12
start_year <- as.numeric(format(min(data$date), "%Y"))
start_month <- as.numeric(format(min(data$date), "%m"))

# Create the time series object
ts_data <- ts(data$value, frequency=12, start=c(start_year, start_month))
```

```r
autoplot(ts_data, main = "Original CPI Data",
     xlab = "Date", ylab = "CPI", col = "red")
```

```r
# Create ACF and PACF plots
acf_plot1 <- ggAcf(ts_data) + ggtitle("Starting Point ACF Plot")
pacf_plot1 <- ggPacf(ts_data) + ggtitle("Starting Point PACF Plot")

(acf_plot1 | pacf_plot1)
```

```r
# Perform Augmented Dickey-Fuller test to check stationarity
adf_test1 <- adf.test(ts_data)

adf_table1 <- data.frame(
  Statistic = adf_test1$statistic,
  P_Value = adf_test1$p.value,
  Lag_Order = adf_test1$parameter,
  Alternative_Hypothesis = "Stationary")
kable(adf_table1, caption = "ADF Test Results")
```

```r
# First-order differencing (d=1)
ts_diff <- diff(ts_data, differences = 1)

# Plot ACF/PACF after first differencing
acf_plot2 <- ggAcf(ts_diff) + ggtitle("ACF - First Differencing")
pacf_plot2 <- ggPacf(ts_diff) + ggtitle("PACF - First Differencing")

(acf_plot2 | pacf_plot2)
```

```r
# Run ADF test after first differencing
suppressWarnings({
  adf_test2 <- adf.test(ts_diff)
})

adf_table2 <- data.frame(
  Statistic = adf_test2$statistic,
  P_Value = adf_test2$p.value,
  Lag_Order = adf_test2$parameter,
  Alternative_Hypothesis = "Stationary")
kable(adf_table2, caption = "ADF Test Results for Differenced Data")
```

```r
# Fit an SARIMA model using auto.arima()
sarima_model <- auto.arima(ts_diff, seasonal = TRUE,
                           stepwise = TRUE, approximation = FALSE)

# Create table for SARIMA model order
model_order <- data.frame(
  Component = c("Non-seasonal (p,d,q)", "Seasonal (P,D,Q,s)"),
  Order = c(
    paste0("(", sarima_model$arma[1], ",",
           sarima_model$arma[6], ",",
           sarima_model$arma[2], ")"),
    paste0("(", sarima_model$arma[3], ",",
           sarima_model$arma[7], ",",
           sarima_model$arma[4], ")[",
           sarima_model$arma[5], "]")))
```

```r
kable(model_order, align = "c", caption = "SARIMA Model Order")

# Create table for coefficients and standard errors
sarima_coeff <- coef(sarima_model)
sarima_se <- sqrt(diag(sarima_model$var.coef))
coeff_table <- data.frame(
  Estimate = sarima_coeff,
  Std_Error = sarima_se)
kable(coeff_table, digits = 4, align = "c", caption = "SARIMA Coefficients & SEs")

# Create table for model selection criteria
criteria_table <- data.frame(
  Metric = c("Log Likelihood", "AIC", "AICc", "BIC", "Sigma^2"),
  Value = c(logLik(sarima_model), sarima_model$aic,
            sarima_model$aicc, sarima_model$bic,
            sarima_model$sigma2))
kable(criteria_table, digits = 4, align = "c", caption = "Model Selection Criteria")
```

```r
# Run diagnostics
checkresiduals(sarima_model)
```

```r
# Create table for Ljung-Box test results
ljung_box_table <- data.frame(
  Metric = c("Q*", "Degrees of Freedom (df)",
             "p-value", "Model df", "Total Lags Used"),
  Value = c(51.088, 19, "9.048e-05", 5, 24))
kable(ljung_box_table, digits = 4, align = "c",
caption = "Ljung-Box Test")
```

```r
# Forecast the next 12 months
forecast_results <- forecast(sarima_model, h=12)

# Plot the forecasted values
autoplot(forecast_results) +
  ggtitle("SARIMA Forecast of Inflation Rates") +
  xlab("Time") +
  ylab("Inflation Rate")
```

```r
# Zoomed in plot of the forecasted values
suppressMessages(autoplot(forecast_results) +
    ggtitle("SARIMA Forecast of Inflation Rates") +
    xlab("Time") +
    ylab("Inflation Rate") +
    xlim(c(end(ts_data)[1], end(ts_data)[1] + 1)))
```

```r
# Print next 12 values
forecast_months <- as.yearmon(time(forecast_results$mean))
forecast_df <- data.frame(
  Month = format(forecast_months, "%b %Y"),
  Point_Forecast = as.numeric(forecast_results$mean),
  Lo_80 = as.numeric(forecast_results$lower[,1]),
  Hi_80 = as.numeric(forecast_results$upper[,1]),
  Lo_95 = as.numeric(forecast_results$lower[,2]),
  Hi_95 = as.numeric(forecast_results$upper[,2]))
kable(forecast_df, caption = "Forecasted Inflation Rates")
```

```r
# Define a GARCH(1,1) model
garch_spec <- ugarchspec(
  variance.model = list(model = "sGARCH", garchOrder = c(1,1)),
  mean.model = list(armaOrder = c(0,0), include.mean = TRUE),
  distribution.model = "norm")

# Print relevant criteria
garch_fit <- ugarchfit(spec = garch_spec, data = residuals(sarima_model))
results <- as.data.frame(garch_fit@fit$matcoef)
colnames(results) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")
kable(results, digits = 6, caption = "GARCH(1,1) Model Coefficients")
```