Time Series Analysis of Average Gasoline Prices in America

Kalp Shah (202103003),* Dhyey Patel (202103053),† and Jainil Patel (202101416)‡
Dhirubhai Ambani Institute of Information & Communication Technology,
Gandhinagar, Gujarat 382007, India
Time Series Analysis

In this project, we aim to analyse the data of Average Gasoline Prices in America based on the tools and models we learnt in the Time Series Analysis course (SC475) under Prof. Mukesh Tiwari.

I. INTRODUCTION

Gasoline prices affect everyone who drives a car or uses other vehicles, and they can also impact businesses and the economy. By studying how gas prices have changed over time and trying to predict future changes, we can better prepare for these fluctuations. This research can help us figure out if there are certain times of the year when gas prices are higher or lower, and why that happens. It can also help us see if there are any patterns in how gas prices change over many years. This information can be useful for making decisions about things like budgeting for gas expenses or planning for transportation needs.

We will examine the trend in the average fuel prices in the United States from the monthly data available from January 1967 to January 2024. We will try to analyse if the behavior of data and see if there are any predictable patterns, if there are any periodic recurring cycles present. We wish to predict future gasoline prices based on the past available data with high accuracy. Overall, studying gasoline prices can give us valuable insights into how the economy works and how it affects people's lives. It can also help us make better choices about things like when to buy gas and how to plan for future expenses.

Fig. 1 shows the last 10 data instances of the dataset.

	date	Avg. Monthly Price
676	2023-05-01	300.042
677	2023-06-01	302.412
678	2023-07-01	301.784
679	2023-08-01	326.825
680	2023-09-01	332.019
681	2023-10-01	317.678
682	2023-11-01	304.982
683	2023-12-01	303.242
684	2024-01-01	293.287
685	2024-02-01	304.302

FIG. 1: Data Instance

II. EXPLORATORY DATA ANALYSIS

The dataset of the Average Monthly Gasoline Prices used in this Time Series Project is taken from the Federal Reserve Economic Data (FRED) [1]. The dataset contains the average gasoline prices from January 1967 to January 2024. The dataset contains two columns, 'Date' and the 'Average Monthly Price' containing the average gasoline price for that particular date. The dataset contains 686 dataset instances and there is no missing data anywhere.

Fig. 2 gives the visual representation of the dataset from January 1967 to January 2024. There is a gradual increasing trend in the price over time. The increase becomes more rapid from 2000s onwards. There is a peak in price around 2008, which may be because of the 2008 Financial Crisis and again in 2022 which is because of the Russia-Ukraine conflict. It is clear that the prices are only going to increase in future.

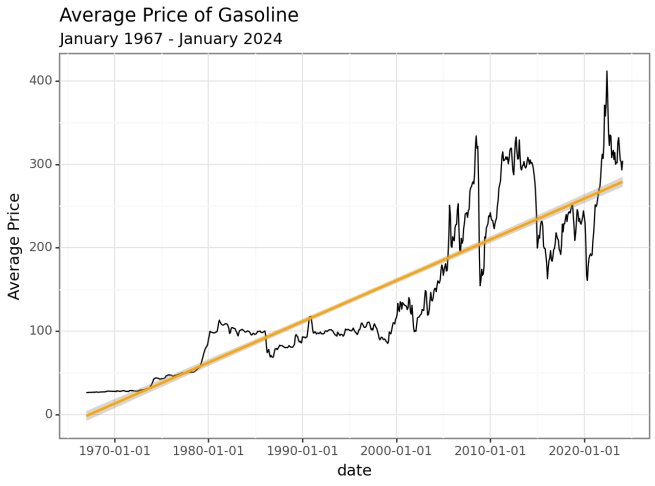

FIG. 2: Data Overview

*Electronic address: 202103003@daiict.ac.in
†Electronic address: 202103053@daiict.ac.in
‡Electronic address: 202101416@daiict.ac.in

Fig 3 shows the month wise price distribution. Prices are high around August - October. October, in particular, is renowned for its beautiful fall foliage in many regions of US.This natural spectacle attracts tourists who want to experience the stunning colors of the autumn. All these lead to a high demand of gasoline during this time and the prices increase. Prices are low during the winter months because of less demand during this period when most of America is covered with thick snow and most regions have extrememly low temperatures.

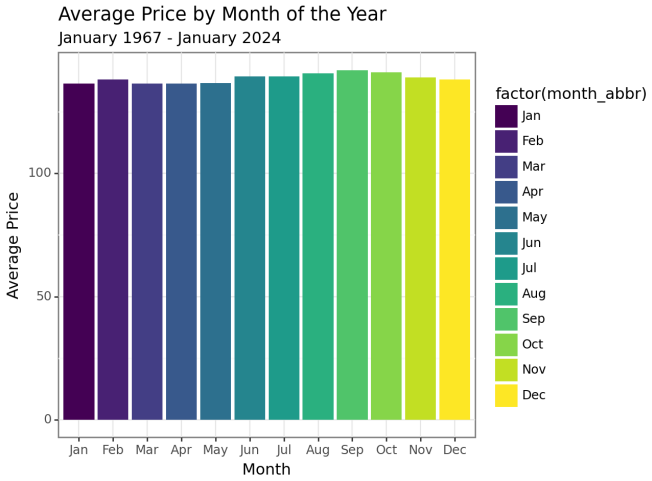

FIG. 3: Monthly Price Distribution

III. STATIONARITY CHECK AND NOISE

We perform the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to check the stationarity of the dataset.

The ADF test assumes that the data is non stationary (hypothesis) and so it checks for the presence of a unit root in the dataset. If the p-value output from the ADF test is less than the significance level 0.05, then we reject the null hypothesis of a unit root and so the time series is likely to be stationary.

The KPSS test is opposite to that of ADF test. The In this case, KPSS assumes that the data is stationary, that is, the hypothesis is that data is stationary. If the p-value is less than the chosen significance level 0.05, we reject the null hypothesis of the data being stationarity, and the series is non stationary.

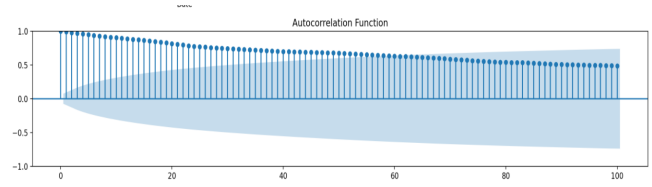

FIG. 4: ACF of Avg. Monthly Price

From Fig. 4, we can see that the ACF doesn't converge to zero faster as the lag increases. This shows that the average gasoline price is not stationary data.

Performing both the tests on our dataset, we get the following result -

- ADF test - p value - 0.773694 (Null hypothesis is true; non stationary)
- KPSS test - p value - 0.01000 (Null hypothesis rejected; non stationary)

The results of both the tests clearly state that the time series is not stationary, which is also evident from Fig. 2.

We need to convert this non stationary dataset to stationary before performing any forecasting.

A. First Order Differencing - We perform the first order differencing on the dataset (as shown in Fig. 5) and again perform the ADF and KPSS tests on the new dataset.

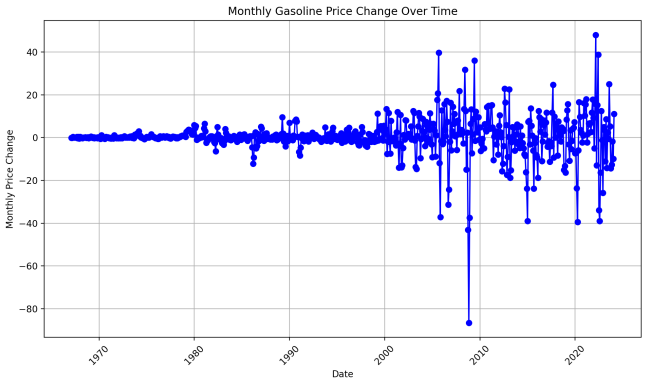

FIG. 5: Monthly Price Difference

The test results for First order differencing are -

- ADF test - p value - 2.682158e-09 (Null hypothesis rejected; stationary series)
- KPSS test - p value - 0.100000 (Null hypothesis stands true; stationary series)

The test results show that first order differencing has made the time series stationary. The autocorrelation function (Fig. 6) also converges to zero and rapidly (insignificant values after 2nd lag) which again shows that the time series is stationary.

Now we obtain the noise of the first order differencing time series. Remove the trend and seasonal components to find the noise of the time series. After obtaining noise, we check the distribution of the noise. Our job will become easier if the noise is white noise; ie, close to gaussian distribution. Fig. 7 represents the detrended time series,

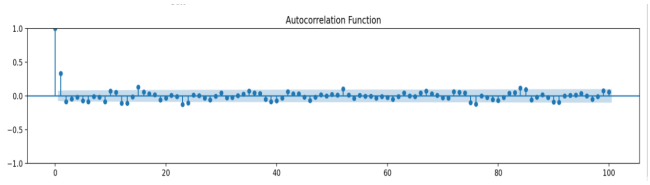

FIG. 6: ACF of first order differencing

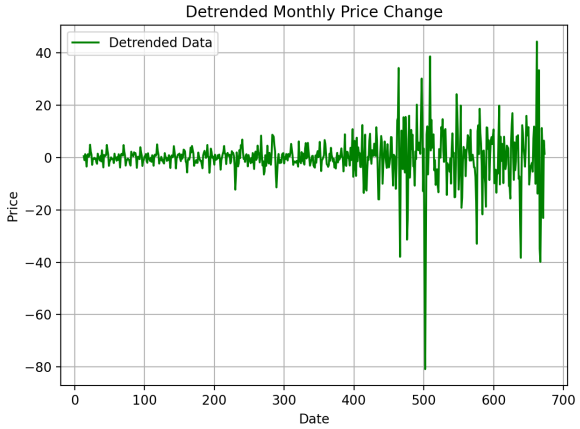

FIG. 7: Noise of first order differencing

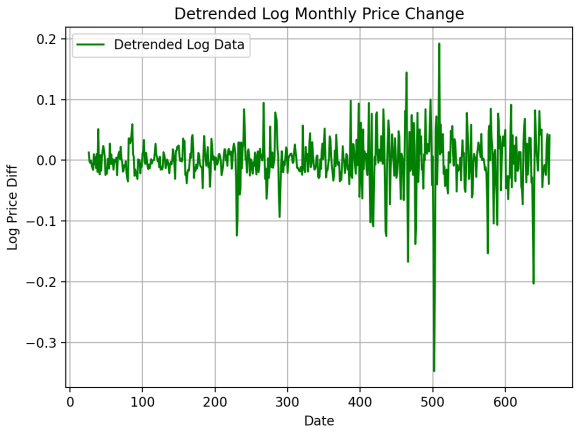

FIG. 8: Noise of first order log differencing

that is, the noise of the first order differenced time series. Fig. 8 represents the noise of the first order log differencing of the time series.

From Fig. 10, the Mean of the noise is -0.0082314688510436 and the Standard Deviation is 8.456986741118898.

From the ACF of the noise (Fig. 11), the ACF does not have significant values for lag greater than 1. No two values are correlated which have a lag of greater than 1. Ideally for white noise, no two values should be related for lag greater than 0.

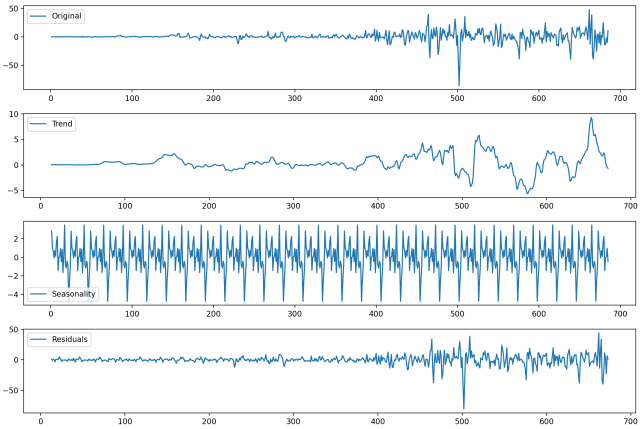

FIG. 9: Trend, Seasonality and Residual of Monthly Price change (first order differencing)

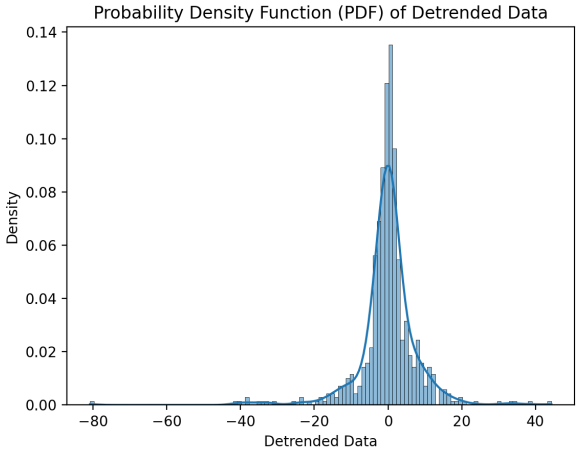

FIG. 10: pdf of Noise ; Mean - -0.0082, Standard Deviation - 8.45698

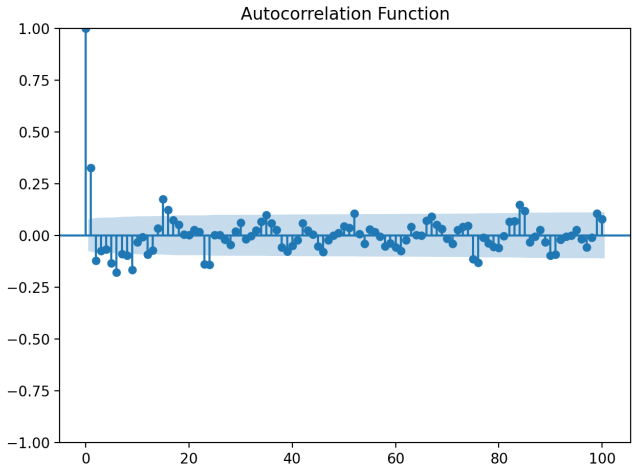

FIG. 11: ACF of Noise

B. First Order Log Differencing - First, we take the log of the entire original dataset. Then, we perform the first order differencing on the dataset (as shown in Fig. 12). Thus we obtain a new time series of the form [log(X_t) - log(X_{t-1})] and again perform the ADF and KPSS tests on the new dataset.

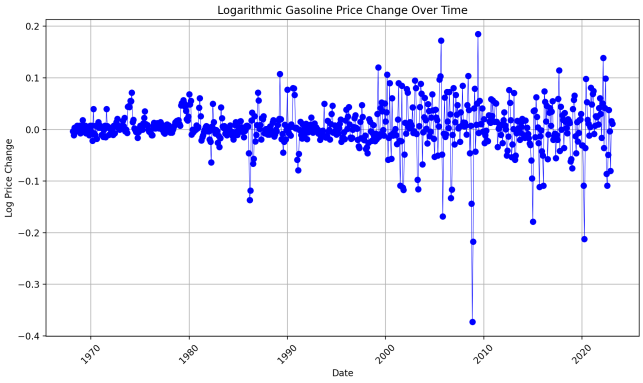

FIG. 12: Log Monthly Price Difference

The test results for First order differencing are -

- ADF test - p value - 9.326813e-13 (Null hypothesis rejected; stationary series)
- KPSS test - p value - 0.100000 (Null hypothesis stands true; stationary series)

The test results show that first order log differencing has made the time series stationary. The autocorrelation function (Fig. 13) also converges to zero and rapidly (insignificant values after 2nd lag) which again shows that the time series is stationary.

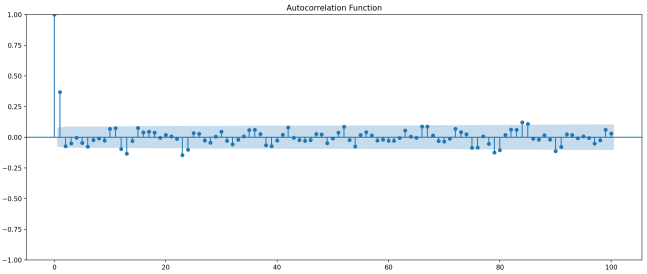

FIG. 13: ACF of first order log differencing

Now we obtain the noise of the first order log differencing time series. Remove the trend and seasonal components to find the noise of the time series just like we did in first order differencing. After obtaining noise, we check the distribution of the noise.

From Fig. 15, the Mean of the noise is 3.938717610751128e-05 and the Standard Deviation is 0.04069494777712663.

The mean of first order lag differencing is closer to 0 as compared to first order lag differencing and the standard

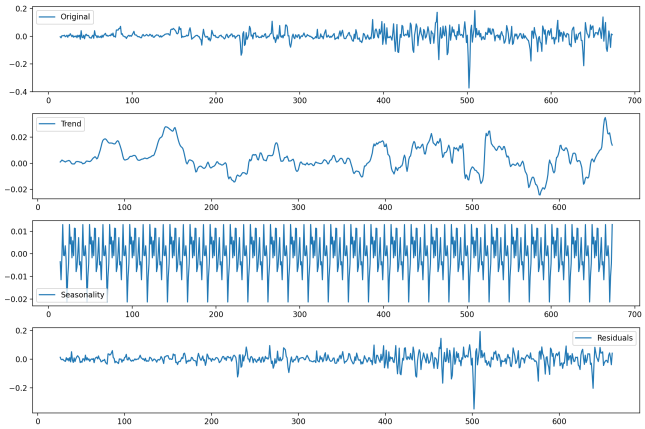

FIG. 14: Trend, Seasonality and Noise of first order log differencing

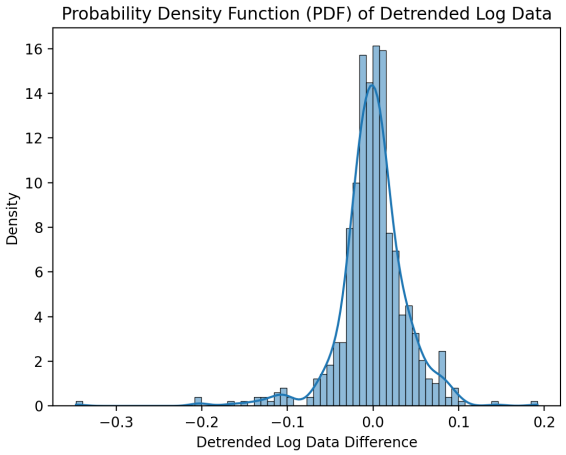

FIG. 15: pdf of Noise ; Mean - 3.938e-05, Standard Deviation - 0.0406

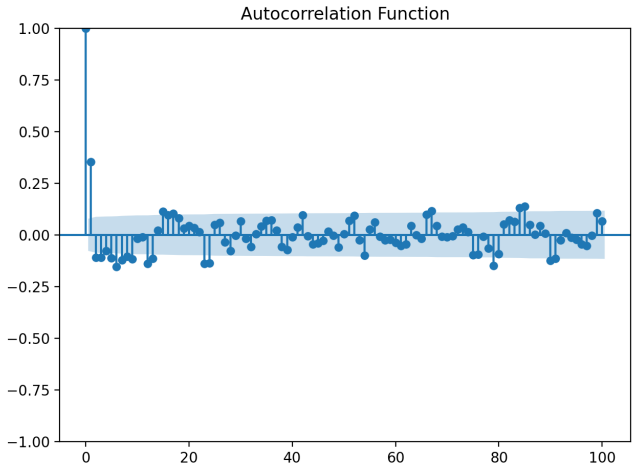

FIG. 16: ACF of Noise

deviation is also less as compared to first order differencing, that is, pdf converges to 0 faster and does not have a fat tail.

First order log differencing gives us the closest result of noise having Gaussian distribution.

IV. ARIMA

Auto Regressive Integrated Moving Average (ARIMA) is a widely used model used for time series forecasting. ARIMA model combines autoregression (AR), differencing (I), and moving average (MA) components to model and predict future time series values based on past time series observations. It is used for finding the best model for non stationary time series. The components of ARIMA are -

- **Auto Regressive (AR)** - This component captures the linear dependency of the current value of the time series with it's previous lagged observation values. 'p' is the AR parameter representing the order of AR component, that is, the current value depends on p lagged observations.
- **Integrated** - This component is used to check for the presence of non stationarity in the time series. It converts the non stationary data to stationary by differencing the consecutive observations which also decides the order of Integrated component. 'd' is the parameter representing the order of differencing the consecutive observations.
- **Moving Average (MA)** - This component represents the relationship between the current observation and the residual errors from moving average model applied to lagged observations. 'q' represents the order of MA component.

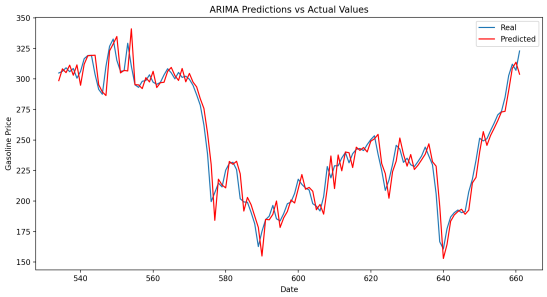

FIG. 17: ARIMA(0,1,1) Prediction

We tried to find the best possible parameters p(AR component), d(Integrated) and q(MA component) for the ARIMA model to best forecast the time series. We tested the model on different values of p,q,d parameters iterating from 0 to 3 each. The best possible fit for the ARIMA model is where we have the lowest value of AIC, accordingly, we get the parameter values -

$$p = 0, d = 1, q = 1$$

ARIMA(0,1,1) is the best fit model for forecasting time series. We took 80 percent of the time series for training the ARIMA model and the rest 20 percent for testing the model.

Fig. 17 shows the predicted time series values as compared to the actual time series values.

The general form of the ARIMA (0,1,1) model is -

$$\mathrm{x}_t = x_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

V. PRESENTATION LINK

Project Video Presentation - Video Link.
Code file - Code.

[1] U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: Gasoline (All Types) in U.S. City Average [CUSR0000SETB01], retrieved from FRED, Federal Reserve Bank of St. Louis;

https://fred.stlouisfed.org/series/CUSR0000SETB01,
May 3, 2024.