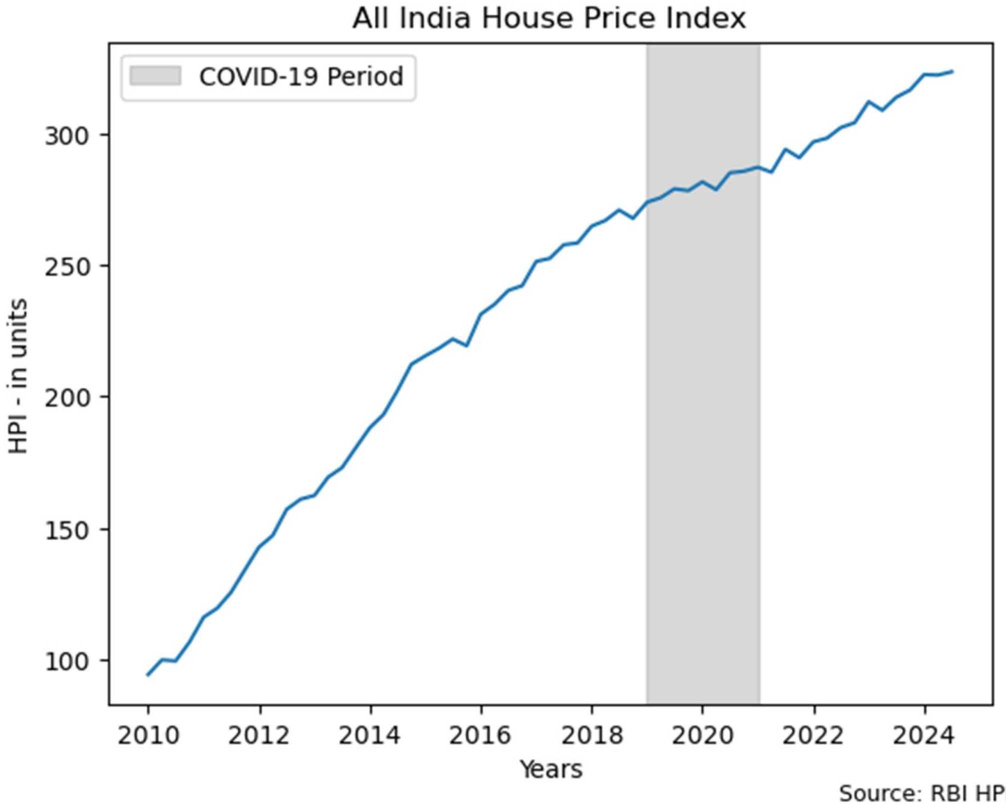
DILIPKUMAAR
II MA EE – Madras School of Economics

HPI CASE STUDY

NHB RESIDEX, India's first official housing price index, was an initiative of the National Housing Bank (NHB), a wholly owned subsidiary of the RBI undertaken at the behest of the Ministry of Finance, Government of India.

The HPI represents the price changes in residential housing properties. At present, the geographical coverage consists of 50 cities in India including 18 State/UT capitals and 37 smart cities.

This dataset is calculated by taking Q1 of 2010 – 11 as the base year. The data has records at 58 quarters with the minimum value recorded at Q1-2010 with 94.24 units and maximum value at Q3-2024 with 323.26 units. The values have been visualised below for easier understanding.



Source: RBI HPI Dataset

The index depicts a gradual upward trend over the years. When we look closer, a small flattening trend appears during the lockdown phase of Covid 19 Pandemic period. This maybe due to the health emergency situation causing a shift in people's mindset from investing to saving. Sellers decrease their prices in order to boost sale of houses.

In this case study, we will be forecasting the All-India House Price Index to the next 12 quarters or 3 years. We will be using Auto Regressive Integrated Moving Average (ARIMA) model using Python. Housing prices are not seasonal in nature and there are no seasonal patterns found in the graph above. Therefore, we will rule out Seasonal ARIMA. The dataset is small in size with only 58 data points and periods, ruling out the need for machine learning and other complex time series forecasting models.

Autocorrelation:

Autocorrelation helps in revealing repeating patterns or trends within a time series. The value of autocorrelation ranges from -1 to +1. The value derived for our data is **0.9986**. It confirms the presence of strong trends in the data.

Stationarity

One of the basic requirements to conduct a Time Series Analysis is to check for stationarity. Stationarity is observed when the statistics (like mean, variance and auto correlation) of a time series remains constant over time. Stationarity can be tested by using the *'Augmented Dicky Fuller Test'* which has:

Null Hypothesis (H0): Series is non-stationary

Alternative Hypothesis (H1): Series is stationary

When the p-value from the AD Fuller test is less than 0.05, we reject the null hypothesis and accept the alternate hypothesis that the 'series is stationary'. We need the dataset to be stationary to be able to perform time series analysis. If the data is non-stationary, meaning the data is dynamic in nature it leads to poor forecasting. In order to make the data stationary, we have to introduce differences between the values to reduce dynamic nature of the data.

The results of ADFuller test are as follows:

ADF Statistic: -3.1679098613025123
p-value: 0.021918958639846282

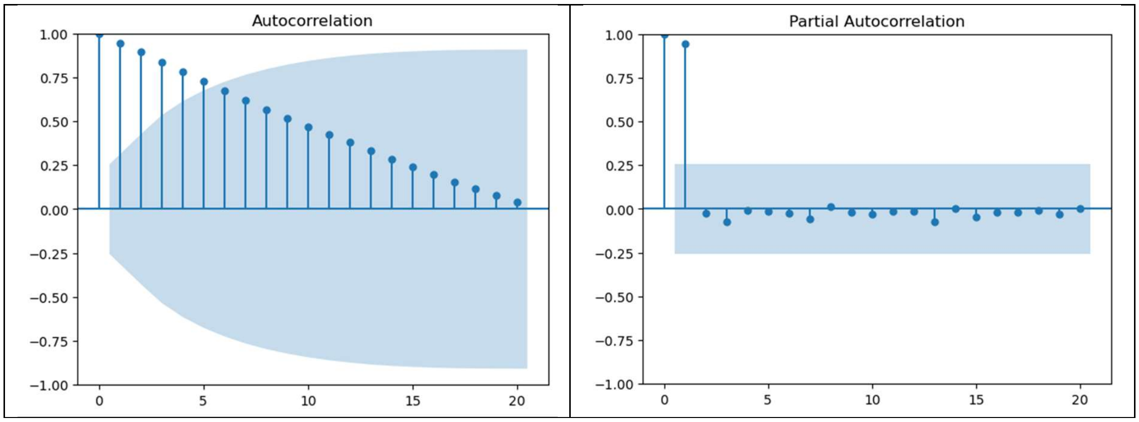
The p-value obtained from our AD Fuller test is **0.02** confirming stationarity in our dataset. Therefore, we can proceed with the analysis.

Autocorrelation and Partial Autocorrelation:

Autocorrelation measures the similarity between a time series and a lagged version of itself over the time intervals. The value of an autocorrelation ranges from -1 to +1. +1 stating that the past values have a strong influence on future values, mentioning a good possibility to conduct time series analysis.

The obtained autocorrelation value in our analysis is **0.99867** implying that the HPI values are highly correlated with their previous values, making it suitable for the study.

The difference between ***'autocorrelation function (ACF)'*** and ***'partial autocorrelation function (PACF)'*** is that ACF measures the correlation between the time series and its lagged values including both direct and indirect effects. Whereas, the PACF measures the direct correlation between the series and its lagged values after removing the effects of intermediate lags. ACF and PACF are plotted below:



Auto Correlation Function shows a gradual decay of lags whereas Partial Auto Correlation Function shows a spike at 1. These are necessary for determining the input parameters for the ARIMA model.

ARIMA Model:

The ARIMA model is short for AutoRegressive Integrated Moving Average and is a widely used forecasting method for time series data.

p (AR – AutoRegressive):
Number of past values (lags) used to predict the current value.

d (I – Integrated):
Number of times the data is differenced to make it stationary.

q (MA – Moving Average):
Number of past forecast errors used to predict current value.

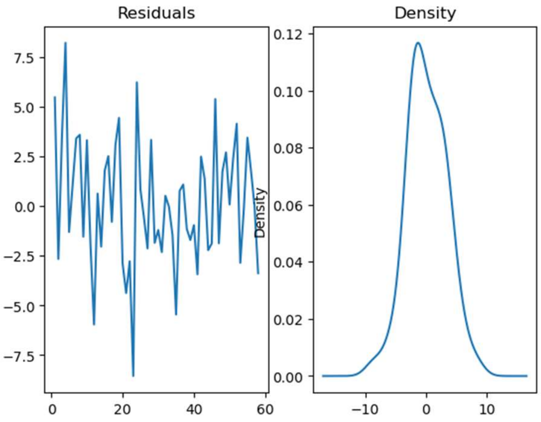
We use the model with parameters → **(p, d, q) = (3, 0, 3)**

This combination was obtained by running a loop of ARIMA models and selecting the model with the least value of **Akaike Information Criterion (AIC)**. Lower AIC value provides the best fit among all candidates without overfitting the data. This model was then used for further forecasting and analysis.

The least three AIC values appeared for following combinations:

Rank	ARIMA Order (p, d, q)	AIC Value
1	(3, 0, 3)	**325.05**
2	(4, 0, 3)	**326.05**
3	(3, 0, 2)	**327.45**

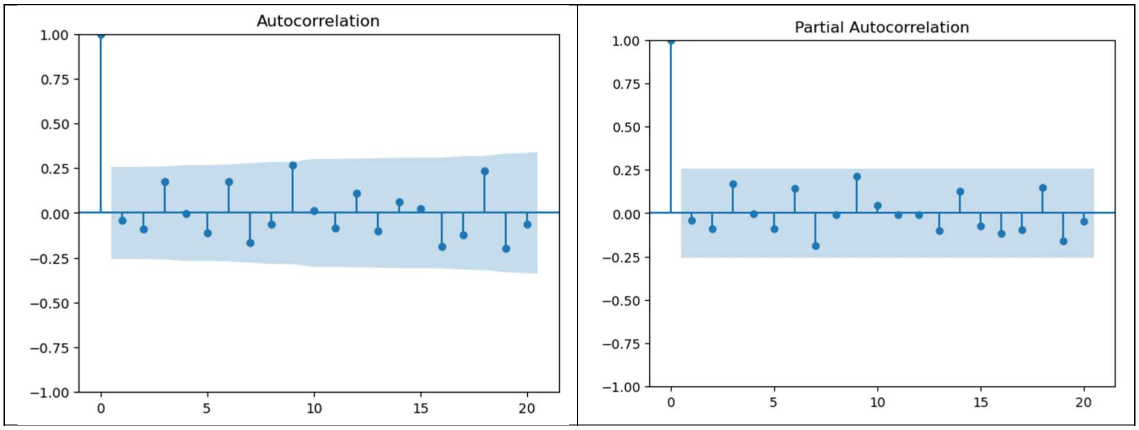
Residuals and Density Plots:



The Residuals plot shows fluctuation at Zero with randomness depicting no seasonality or trend, also suggesting that the model has captured the structure well.

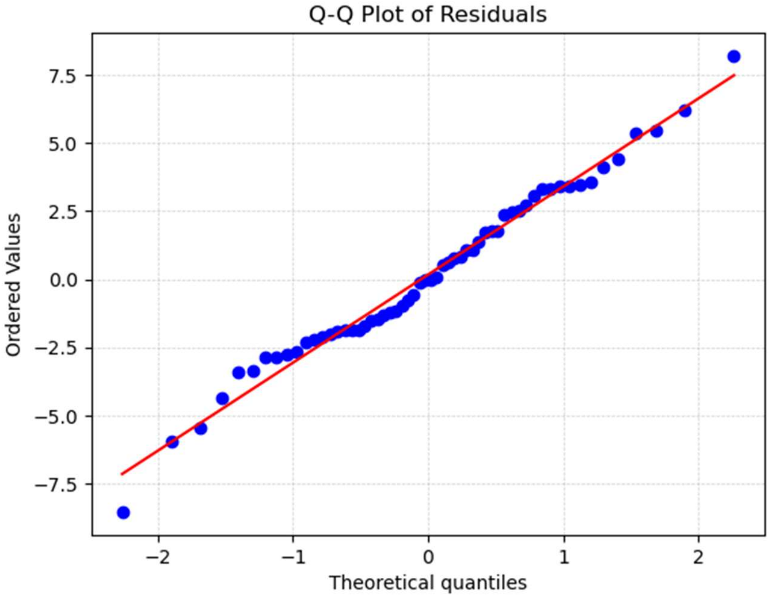
The Density plot shows that it is normally distributed with the bell-shaped curve centred at Zero, suggesting good model performance.

Autocorrelation and Partial Autocorrelation plots of the residuals:



Both the plots depict the spikes to be within the confidence intervals centred around Zero suggesting that the ARIMA model has captured the trends well.

Q-Q plot of Residuals:

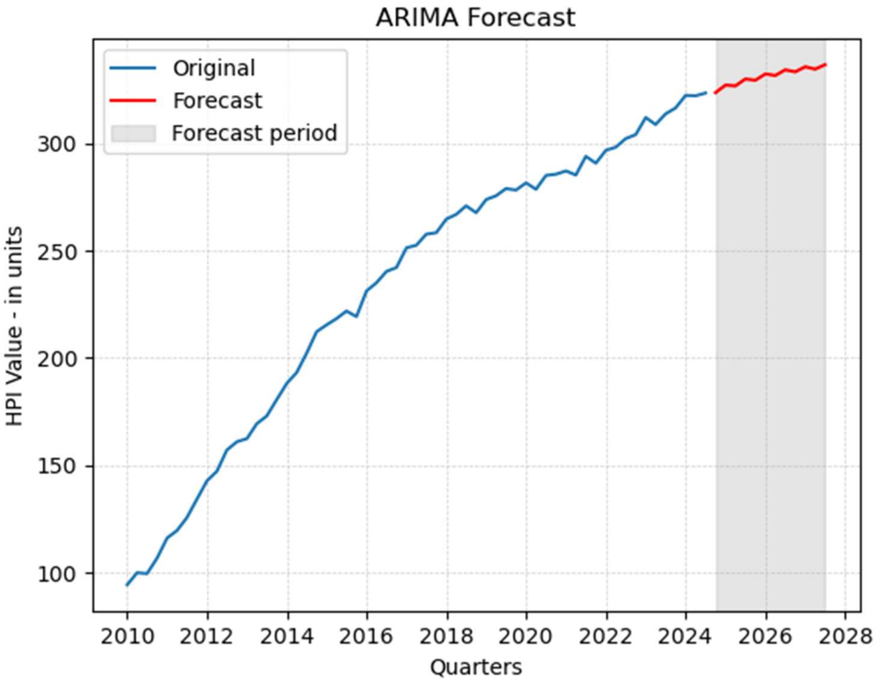


A **Q-Q plot** is a graphical tool used to assess whether a dataset follows a specific theoretical distribution—typically the normal distribution. If the residuals follow a normal distribution, the points should fall approximately along the 45-degree red reference line, which in our case it **satisfies** the key assumption for reliable statistical inference and forecasting.

FORECAST:

The forecast was done for the next 12 quarters. The dataset had HPI values for 58 quarters earlier. The forecasted values are as follows:

Index	Quarter	Units
59	2024-Q4	326.184
60	2025-Q1	328.3214
61	2025-Q2	330.8025
62	2025-Q3	333.106
63	2025-Q4	335.4725
64	2026-Q1	337.7912
65	2026-Q2	340.1133
66	2026-Q3	342.4152
67	2026-Q4	344.7079
68	2027-Q1	346.9864
69	2027-Q2	349.2532
70	2027-Q3	351.5072



The forecasted values show a gradual upward trend as such observed in the historical values reaching a value of 351.5072 units in Q3 of 2027. However, as this value is forecasted solely based on historical values of the Housing Price Index, there may be other factors such as income of the people, location of the property, resources available at location, distance from the city highly influencing their prices. Taking Chennai as an example, the housing prices are completely random and varies largely based on closer proximity to an IT park, type of housing (individual house, flat, gated community, etc), number of amenities provided by the builder, closeness to main road or beach, etc. Adding other control variables and larger datasets using complex models using Machine Learning or Deep Learning methods lead to different forecasts with higher accuracy.

As the index takes an aggregate of 50 cities to represent the levels of housing prices in India, this may try to represent what the housing market scenario maybe but may have biases and may at times not even be closer to the actual scenario.

When we look at the forecasted index values in particular, it appears to be raising. But taking the whole picture into account, even though when it seems to be raising, a flattening trend appears. This may be due to price ceiling issues as prices cannot be going in an upwards direction. A ceiling might hit and then prices may have to fall down for a bit at least.

One good indicator that we had observed is the auto-correlation value of 0.9986 observed in the dataset indicating an extremely strong correlation between current and past HPI values. For a macroeconomic indicator like the HPI, such high autocorrelations are not only expected but desired for. It reflects the inertial nature of the real estate prices, which gradually rises rather than fluctuating. In essence, the HPI is largely self-explanatory and external shocks seem to play a minimal role at least within the observed time frame.
