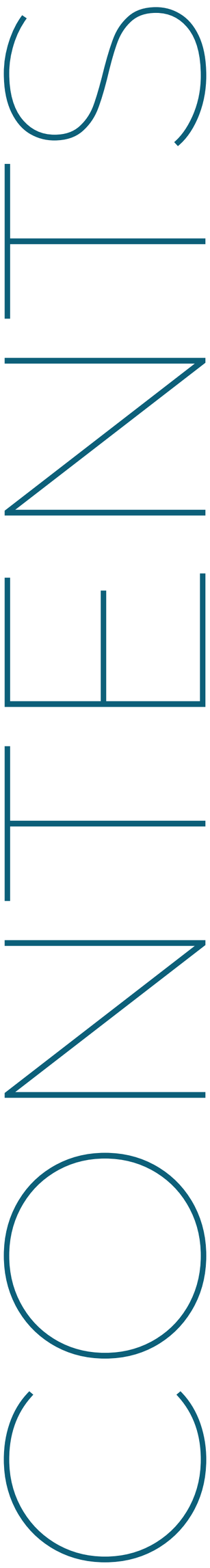
US MEDIAN HOUSE PRICE

Time series analysis

Prepared By:

Manjiri Gujar, Swati Mambilavil
Linh Cao, Colin Quinn

Table of Contents



1. DATA DESCRIPTION WITH EDA

The dataset comprises information on the Median Sales Price of Houses Sold for the United States, spanning from Q1 1963 to Q1 2023. It consists of 263 records and includes two columns: 'Date' and 'MSPUS,' representing the respective dates and median house prices. The original dataset's column 'Date' is of type 'Chr' (character), while the 'MSPUS' column is of type 'num' (numeric).

We obtained the dataset in a .CSV file format, which was then read and stored for further analysis. As for data quality, there are no null values or missing entries. The data source for this dataset is the U.S. Department of Housing and Urban Development, accessible via the FRED website: (Median Sales Price of Houses Sold for the United States (MSPUS) | FRED | St. Louis Fed). The values are measured in dollars and are not seasonally adjusted. The frequency of data collection is quarterly.

STATISTICS SUMMARY

The minimum price of US house is 17.800 USD in 1969 while the maximum price is 479.500 in 2022.

Most of the median house price of the US are from 17.800 USD to 317.800 USD, in which the bin 17.800 - 67.800 and 117.800 - 167.800 have the highest number of houses.

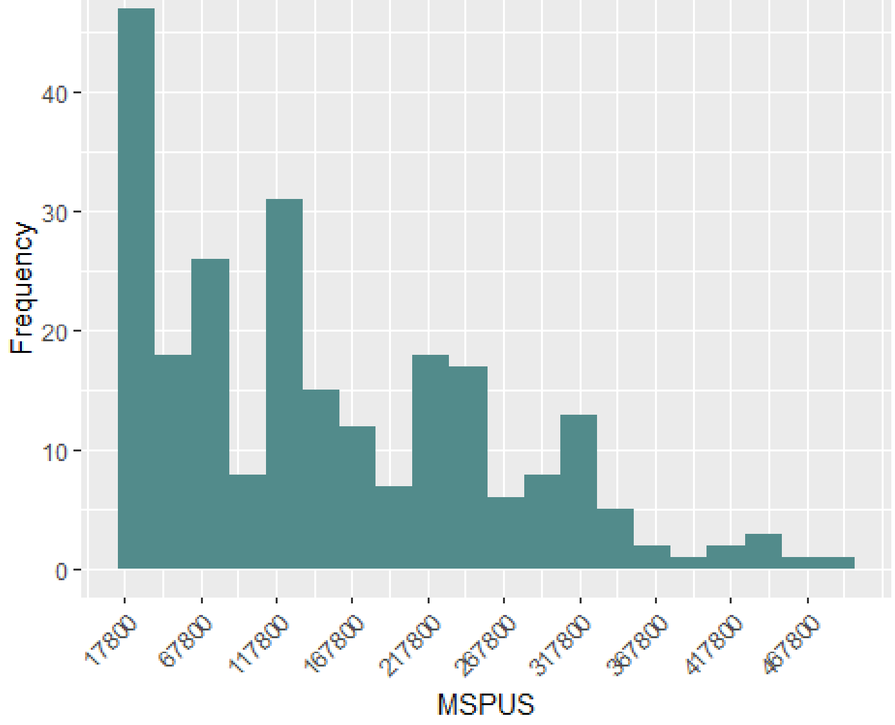


Figure 1.1. Median house price - Distribution

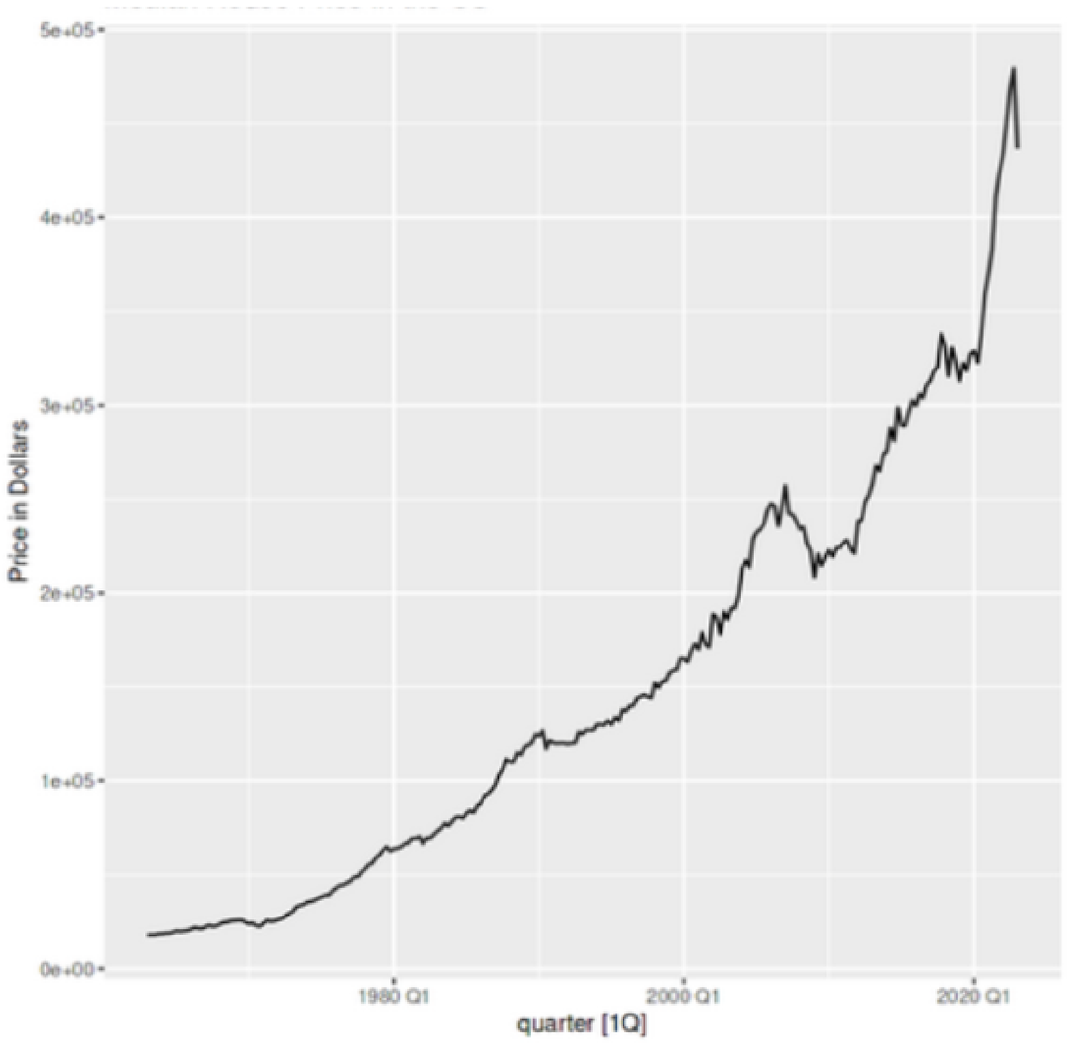


Figure 1.2. Time series on the US median house price

After this we plotted this data to get a view on the variance, trend and seasonality. So to get a detailed view of it we computed the STL decomposition, on performing the STL decomposition we understand that the Trend is upward and increases exponentially over the period of time, same is the case with variance and we can see there is seasonality in the data.

TIME SERIES ANALYSIS

After loading the data into the data frame we checked the summary of the data of MSPUS(House Price), we then assigned relevant column names for the previously known column names **'Date'** and **'MSPUS'** to now changed column names 'Quarter' and 'House Price'. Then, we converted the data frame into a tsibble , and converted the data type from **'chr'(Date)** and **'num'(House Price)** to **'<qtr>' (Quarter)** and **'<int>'(House_Price)**.

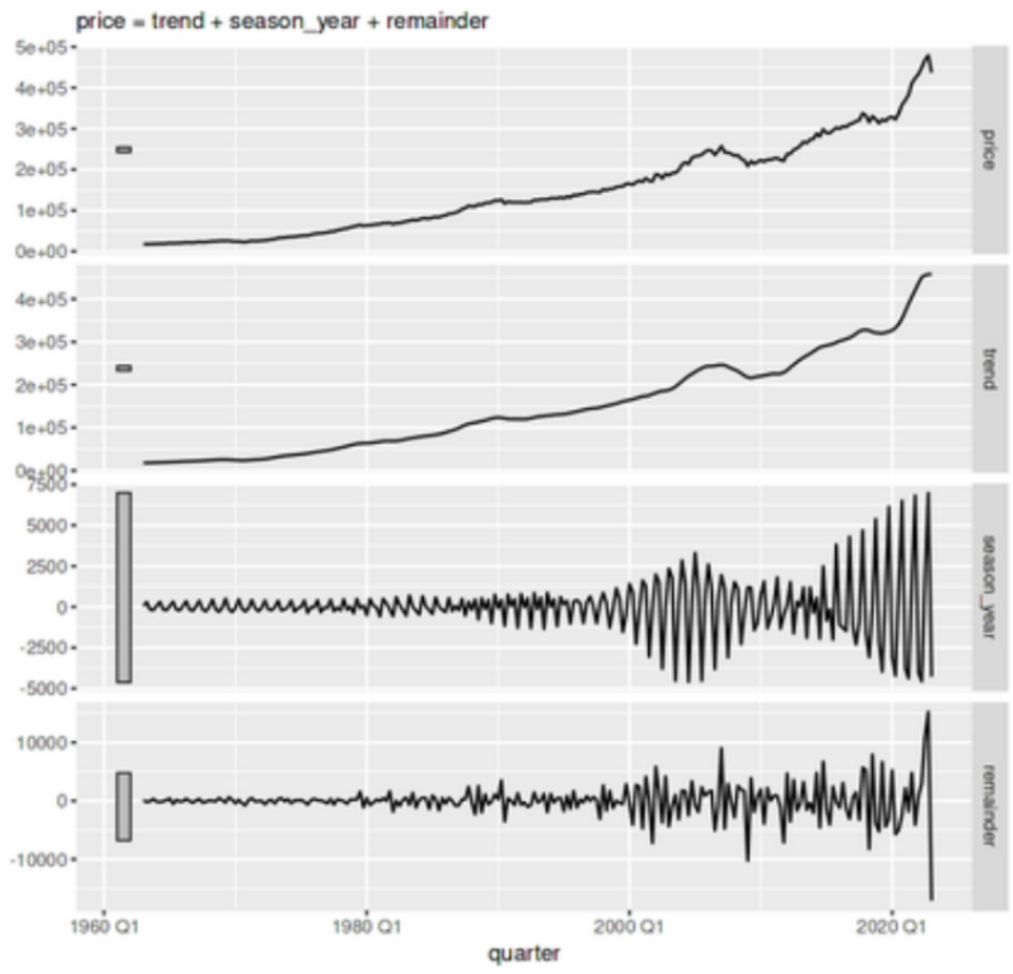


Figure 1.3. STL Decomposition of the Median Prices of the US Houses Sold

2. MODEL FITTING

We split the data into training and test sets with a 9:1 ratio. Then, we fit three models to the data: 1) Time Series Regression 2) Exponential Smoothing, and 3) ARIMA. In addition, we fit various benchmark models to use for determining the optimal approach. After plotting the data (Figure 1.2) and conducting STL decomposition (Figure 1.3), we observed that the data does not show variation that increases or decreases with the level of the series, hence we chose not to use any transformation on the data. We also fit the four benchmark models (Mean, Naive, Seasonal Naive, and Random walk with drift) to compare our models' accuracy to the accuracy of these benchmark models.

THE TIME SERIES REGRESSION

We used the TSLM() function on the data, generated the forecasts for 26 quarters in the test set, and calculated the accuracy on the test set using the accuracy() function from fabletools package. The TSLM model has a RMSE of 89263.16 and MAPE of 20.04.

We conducted residual diagnostics on the model and found that the residuals do have significant autocorrelation. This is also evident from the p-value of the Ljung-Box test, which was 0.

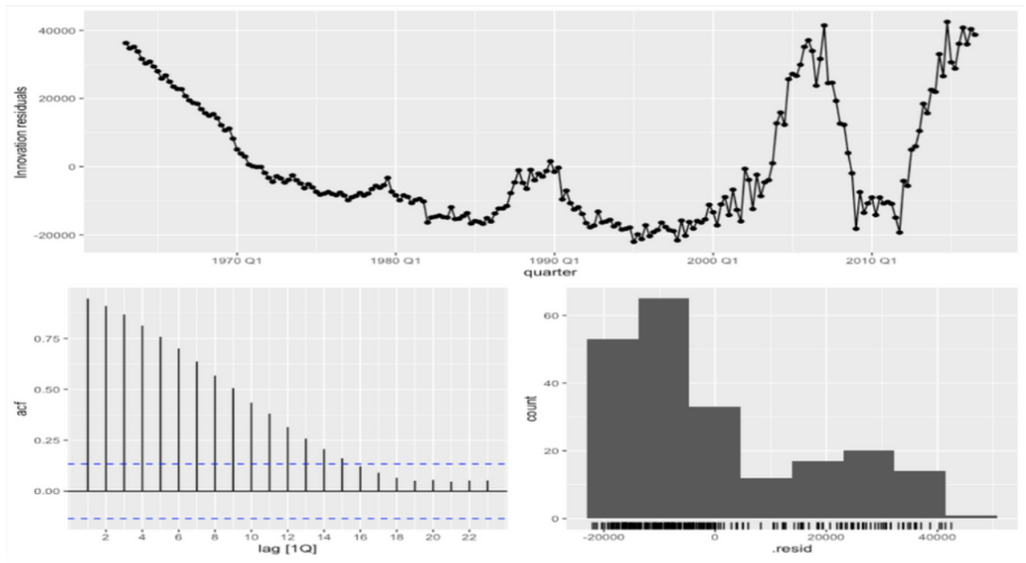


Figure 2.1. Residual Diagnostics of TSLM model

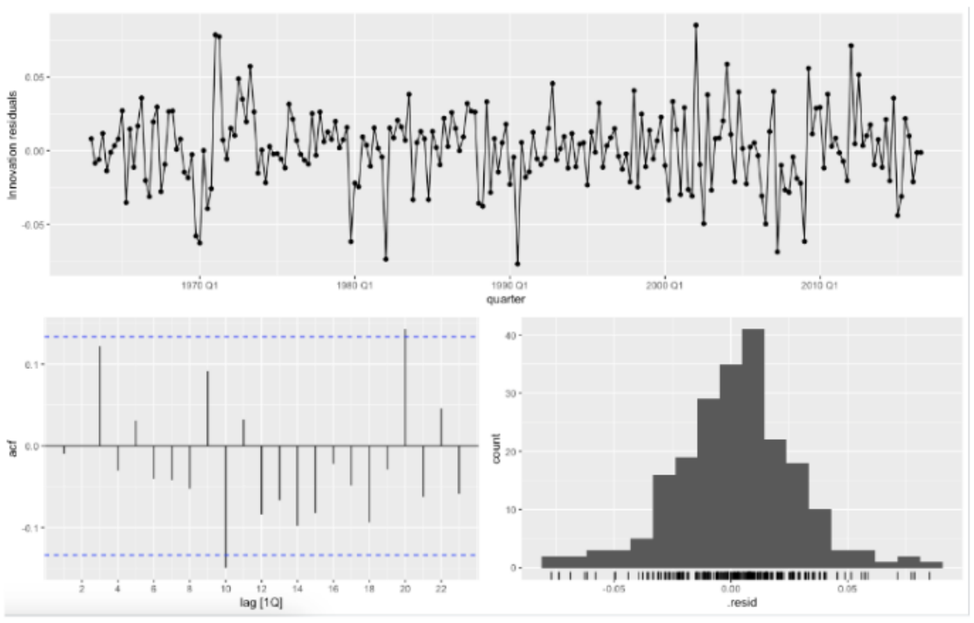


Figure 2.2. Residual Diagnostics of ETS model

EXPONENTIAL SMOOTHING

- We used the ETS() function on the data to obtain the best-fit model. Based on the lowest AICc of 4517.346, the best model was determined to be ETS(M,A,M).
- We predicted the forecasts for 26 quarters in the test set and calculated the accuracy on the test set using the accuracy() function from fabletools package. The best fitting ETS model has a RMSE of 40690.54 and MAPE of 7.02.
- We conducted residual diagnostics on the model and found that the residuals do not have significant autocorrelation. This is also evident from the p-value of the Ljung-Box test, which was 0.0949.

ARIMA

As part of our EDA we found out whether the data in the time series is stationary or not. As per Figure 1.3, we observed that the data is not stationary. We also conducted Unit Root Tests to find whether differencing is required for the time series or not, to make the data stationary. This process of using a sequence of KPSS tests to determine the appropriate number of first differences is carried out using the unitroot_ndiffs() feature. Through this process we determined it was appropriate to apply first differencing with no seasonal differencing required.

We used the ARIMA() function to find the best fitting model on the transformed data and we were provided with ARIMA(2,1,2)(1,0,1)[4] w/ drift as the best model with an AICc of 4173.37. This model provided us with an RMSE of 57378.2 and MAPE of 8.83. After assessing the residuals we were provided with a p-value of the Ljung-Box test that was greater than 0.05, hence we do not reject the null-hypothesis and conclude the residuals do not have autocorrelation.

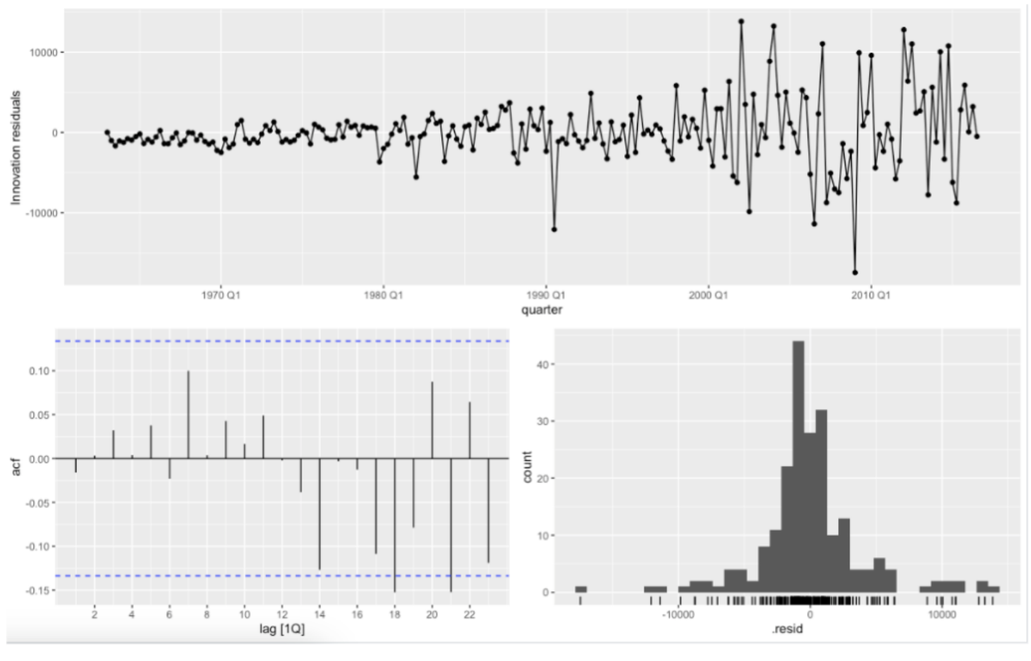


Figure 2.3. Residual Diagnostics of ARIMA model

BENCHMARK MODELS

It was determined (figure 2.4) that the Random Walk with Drift model provided us with the lowest RMSE of 59850, followed by the Naive Model with 78445.30, the Seasonal Naive Model with 79229.24, and the Mean Model with 243342.43. It is through these values that we are comparing the efficacy of the forecasts provided by our ARIMA and Exponential Smoothing models.

3. MODEL SELECTION & FORECASTING

In our time series analysis project on the Median prices of houses sold in the US, we explored the effectiveness of Time Series Regression, ETS and ARIMA models. By evaluating the residual diagnostics of all the models, we observed that the ETS and ARIMA both capture the underlying dynamics of the data reasonably well. The residual diagnostics for the TSLM showed that the residuals had autocorrelation and thus did not resemble white noise. However, a comprehensive comparison of the RMSE and MAPE values led us to conclude that the ETS model outperforms the TSLM and ARIMA models in terms of accurate forecasting.

We compared the RMSE and MAPE of all the models including the benchmark models and found that the ETS model is more accurate. Consequently, we selected the ETS model as our final choice and utilized it to generate forecasts for the next six years from our training data.

Figure 3.1 showcases the forecast plot derived from the ETS model, providing valuable insights into the trend and seasonality of the median house prices in the US. Notably, the ETS model demonstrates commendable accuracy in capturing the anticipated patterns. As we extend the forecast horizon, the prediction intervals naturally widen. This phenomenon is expected in multi-step predictions, as the wider intervals aim to encompass potential shifts in market conditions, changes in customer behavior, or the impact of unforeseen events.

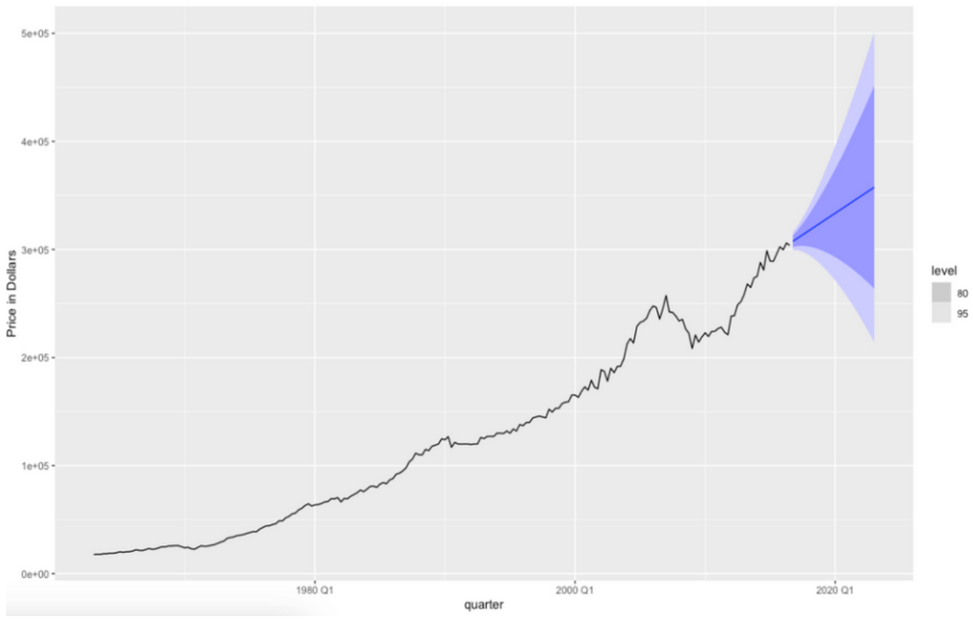


Figure 3.1. Forecasts from the ETS model

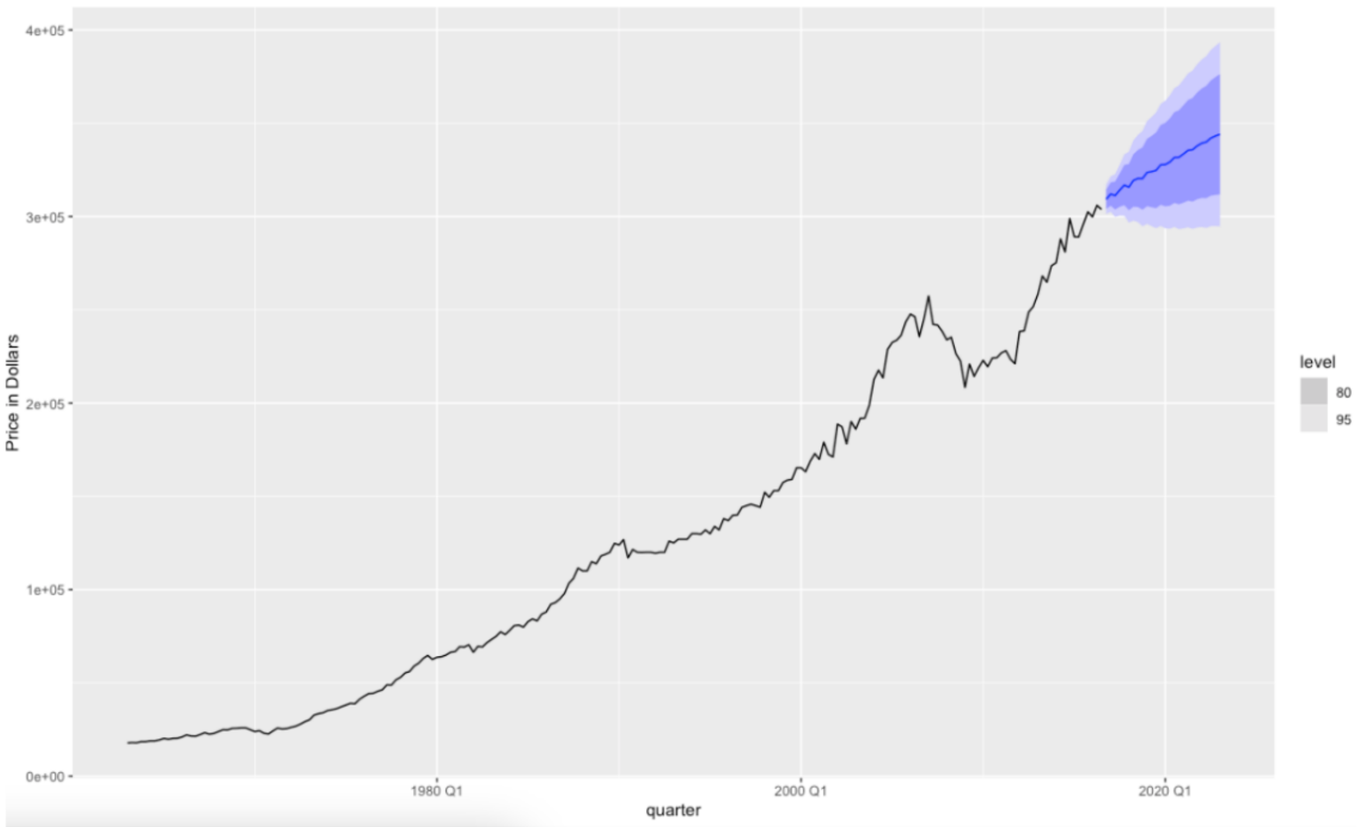


Figure 3.2. Forecasts from the ARIMA model

Conversely, the ARIMA model's forecasts (Figure 3.2) primarily exhibit a straight-line pattern with narrow prediction intervals. This is due to the fact that the model only accounts for the variation in the errors. Although the model may accurately capture the overall level of the data, it fails to account for the complexities arising from trend and seasonality.

The Time Series Linear Regression model's forecasts (Figure 3.3) captures the increasing trend and seasonality to some extent but it is not in continuation with the actual data. The prediction intervals are wide so they do account for the uncertainties affecting the median prices of houses sold in the US.

As a result, we favored the ETS model due to its superior performance in accurately predicting the future median house prices in the US.

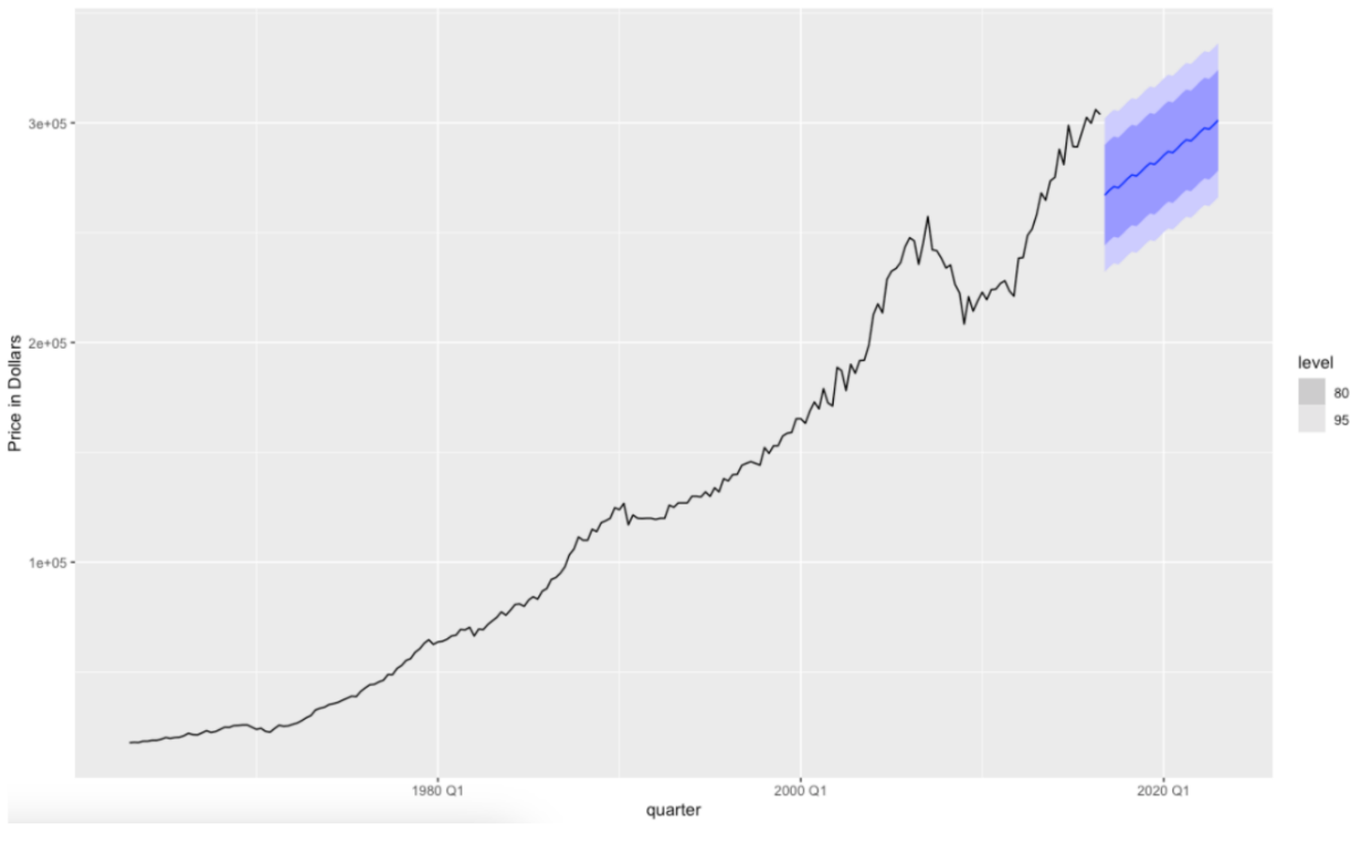


Figure 3.3. Forecasts from the Time Series Linear Regression model

4. CONCLUSION & LIMITATIONS

INSIGHTS

- The median prices of houses in the US have been influenced by socio-economic factors such as inflation, household income, rising interest rates, and Purchasing Power Parity.

- The time series plot indicates decrease in median house prices starting from 2005, with a significant drop during the 2008 recession. Prices started rising again after 2010.

- The forecasts obtained from fitting the ETS model align with the increasing trend observed in the time series data.

- External research and analysis of economic factors can provide additional insights into the fluctuations and trends observed in the median house prices.

LIMITATIONS

- The dataset's focus on median prices restricts the analysis to the impact of time alone on house prices, neglecting other influential factors such as location-specific variables.

- Without state or regional data, it becomes difficult to interpret localized fluctuations in house prices based on observed peaks and declines.

- Sole reliance on median values within the dataset fails to represent the entire range of house prices across different cities and states in the US from 1969 to 2023. Consequently, the insights provided may lack complete accuracy when applied to specific areas within the real estate industry. Moreover, accurate predictions would require the consideration of additional factors like socio-economic conditions, consumer behavior, and the correlation between house attributes and prices.

- The chosen model may not be fully optimized for predicting median house prices. The limitations of the dataset can impact the model's accuracy, consequently affecting the effectiveness of the generated forecasts. Since the dataset lacks comprehensive information regarding other influential factors, the model attempts to account for these variables during forecasting, leading to wider prediction intervals as demonstrated by the ETS model. Therefore, a conservative approach must be adopted when assessing the practicality of the forecasts.

CONCLUSION

In summary, the median house prices in the US have generally increased over time, with a notable decrease during the 2008 recession. As the model with lowest RMSE value (figure 4.1), we decided to use ETS model in our forecasting for the next six years. The prediction show that the median US house price has an upward trend in the next 6-year period..

REFERENCES

- St. Louis Fed. (n.d.). Median Sales Price of Houses Sold for the United States (MSPUS) | FRED | St. Louis Fed. Retrieved from https://fred.stlouisfed.org/series/MSPUS

- Hyndman, R. J., & Athanasopoulos, G.. Forecasting: Principles and Practice (3rd ed). Monash University, Australia. https://otexts.com/fpp3/

- Rothstein, Robin. “Housing Market Predictions for 2023.” Forbes, 8 June 2023, www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/

APPENDIX

```
> #Conducting unitroot tests to find the degree of differencing required for the data
> train %>%
+   features(price,unitroot_kpss)
# A tibble: 1 x 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1      4.27        0.01
>
> train%>%
+   mutate(diff_price = difference(price)) %>%
+   features(diff_price,unitroot_kpss)
# A tibble: 1 x 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1     0.463      0.0502
>
> train%>%
+   features(price,unitroot_ndiffs)
# A tibble: 1 x 1
  ndiffs
   <int>
1      1
>
> train %>%
+   mutate(diff_price = difference(difference(price))) %>%
+   features(diff_price,unitroot_kpss)
# A tibble: 1 x 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1    0.0165         0.1
>
> train%>%
+   mutate(diff_price = difference(log(price))) %>%
+   features(diff_price,unitroot_kpss)
# A tibble: 1 x 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1     0.442      0.0588
>
> train%>%
+   mutate(log_price = log(price)) %>%
+   features(log_price, unitroot_nsdiffs) #no differencing required for the seasonal component
# A tibble: 1 x 1
  nsdiffs
    <int>
1       0
```

Figure 2.5. Unitroot tests

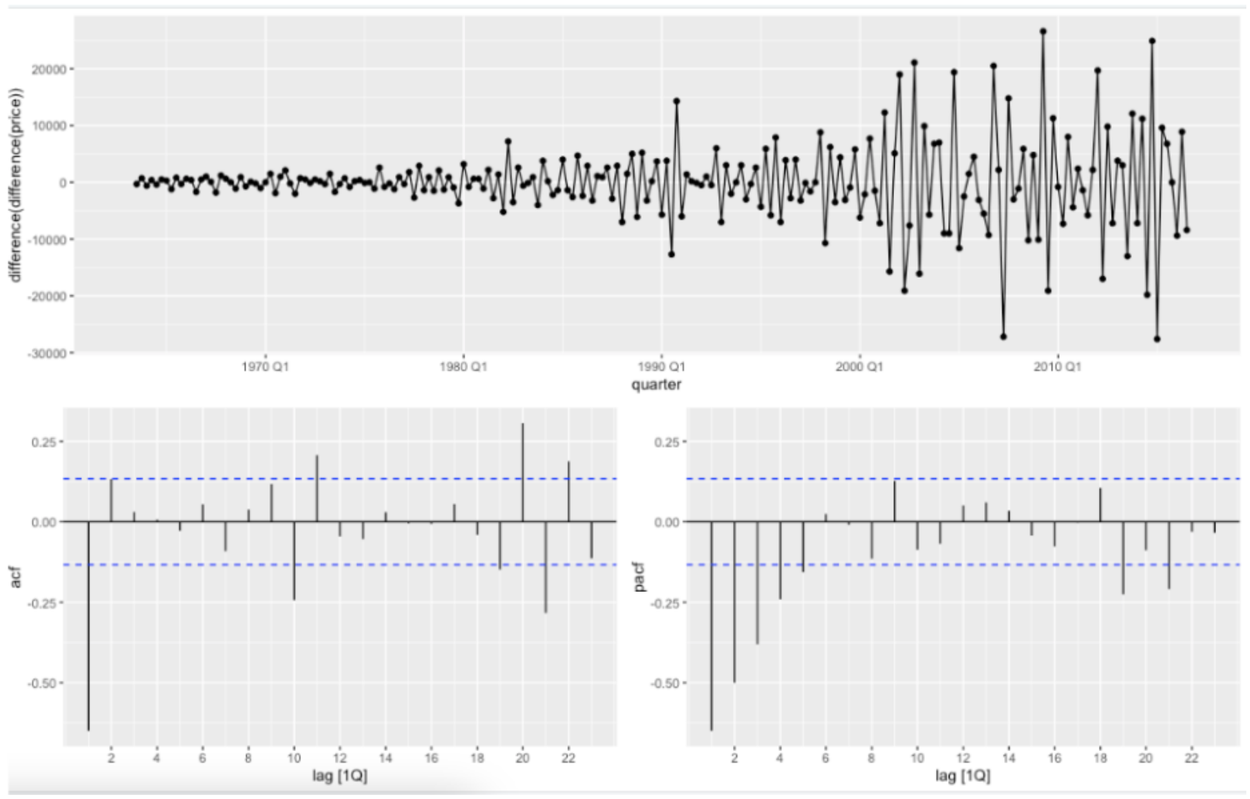


Figure 2.6. ACF and PACF plots

```
> # Print the resulting table
> data
                    RMSE      MAPE
ETS             40690.54  7.017065
ARIMA           57378.20  8.827264
Drift           59849.68  9.546365
Naive           78445.30 14.146831
Seasonal Naive  79229.24 14.401602
TSLM            89263.16 20.044445
Mean           243342.43 65.069221
```

Figure 4.1. Comparison of the accuracy measures of all models

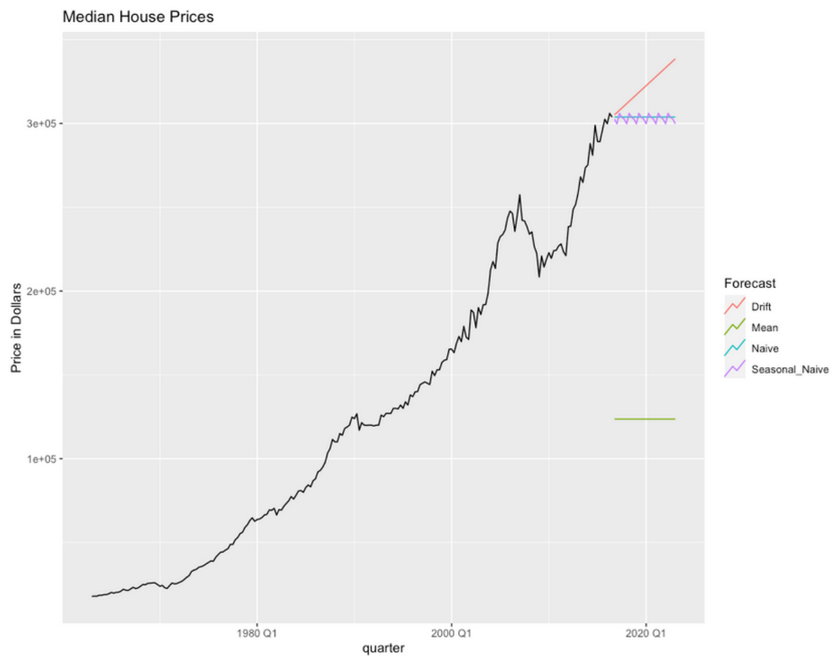


Figure 2.4. Forecasts from the benchmark models