Time Series and Forecasting

ROHITH GANESAN

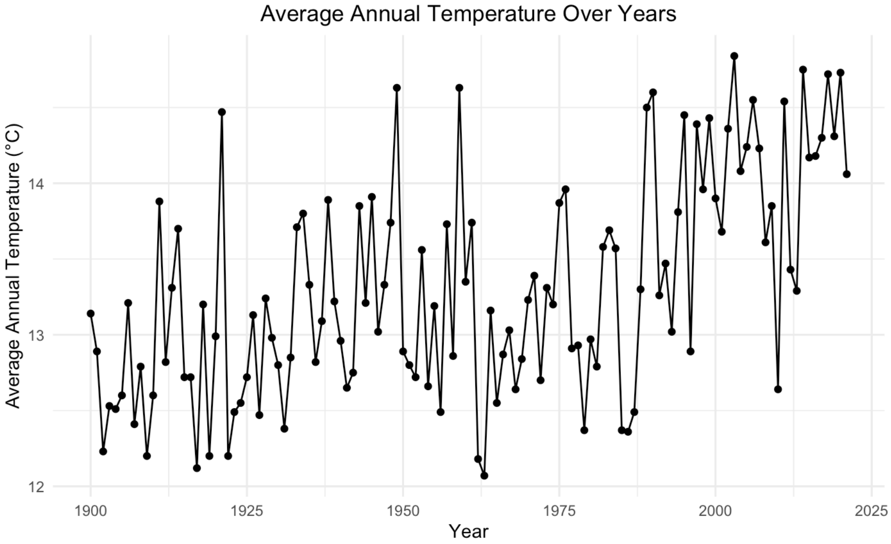
Question-1

Introduction:

This report presents an in-depth analysis of the average annual temperature in the Midlands region of England, using data spanning from 1900 to 2021. The primary objective of this analysis is to identify patterns, trends, and any skeptical behavior in the temperature data, and to forecast future temperatures using the Autoregressive Integrated Moving Average (ARIMA) model. This study is encouraged by the global significance of understanding and predicting climate change impacts, which are crucial for planning and adaptation strategies in various sectors including agriculture, health, and environmental protection.

Data Description:

The dataset comprises two variables: "year" and "avg_annual_temp_C", representing the year and the average annual temperature in degrees Celsius, respectively. Recorded by the UK Meteorological Office Hadley Climate Centre, this dataset encapsulates 122 observations without any missing values, ensuring a analysis without the need for imputation. From my perspective, the dataset presents an opportunity to examine temporal patterns, trends, and seasonality in the temperature data over the 122-year period.



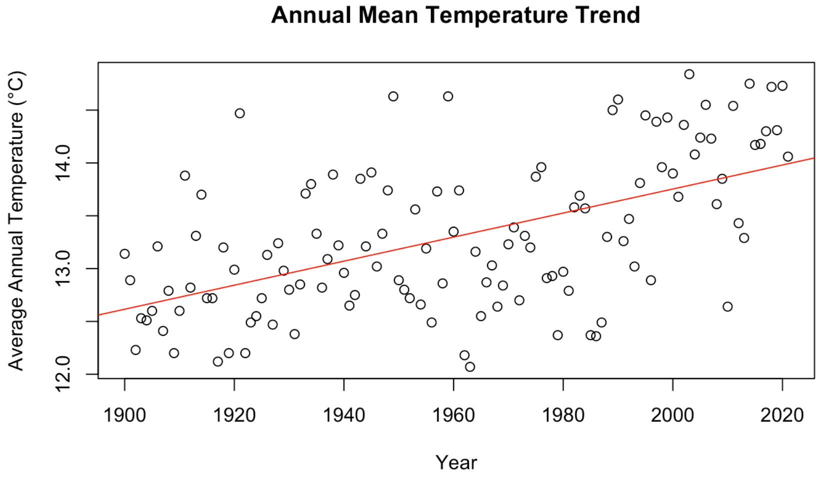
Techniques such as time series decomposition can help in separating the data into trend, seasonal, and residual components. Furthermore, forecasting models like ARIMA (*AutoRegressive Integrated Moving Average*) can be done to predict future temperatures based on historical data. Identifying any cyclic patterns or long-term trends is important for understanding climate change impacts on regional temperatures.

Next, we will explore on time series aspect, focusing on trend analysis and potentially identifying any cyclical patterns or significant shifts in temperatures over time. This will involve visualizing the temperature data across the years and applying statistical tests or models to notice trends.

Methodological Framework:

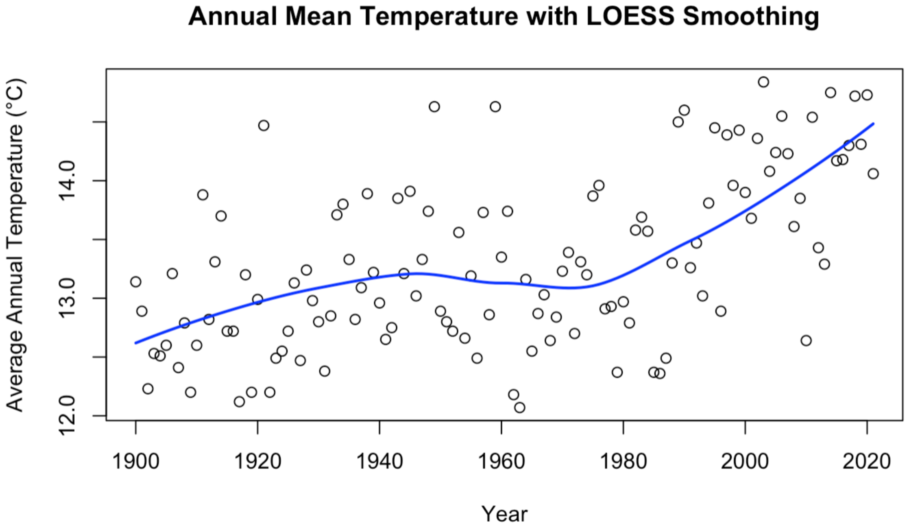
Preliminary Data Analysis:

The analysis commenced with a preliminary review of the dataset to understand its structure and to identify any apparent trends or patterns. This involved plotting the annual temperatures against time, which revealed a potential warming trend over the observed period. To quantitatively assess this trend, a linear regression model was applied, with "year" as the independent variable and "avg_annual_temp_C" as the dependent variable. The red graph shows the actual annual mean temperatures over the years. The presence of a positive slope in the regression line confirmed the warming trend, suggesting an increase in average annual temperatures over time.



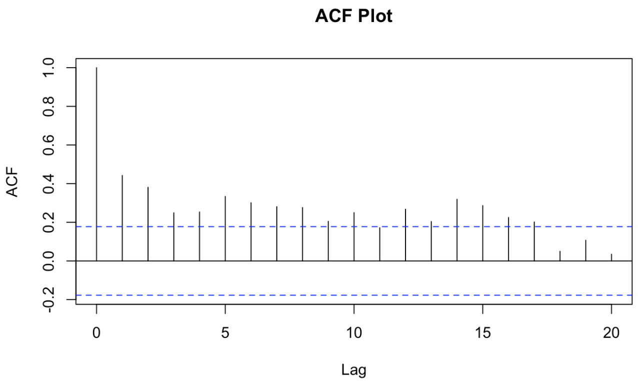
Trend Analysis and Seasonal Decomposition:

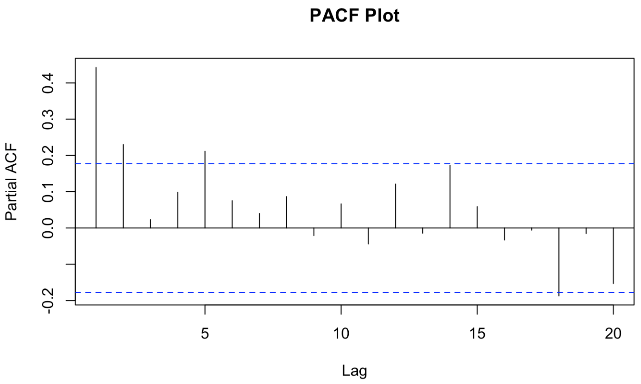
Further analysis employed "LOESS smoothing" to better understand the underlying trend without assuming linearity. The blue graph shows the seasonal component, which, given the context of annual data with a frequency of 1, doesn't present a clear seasonal pattern as expected. Given the data's annual frequency, the seasonal component was not expected to show clear patterns, which aligned with the findings that highlighted the importance of the trend component.



Stationarity Testing and Model Selection:

The non-stationarity of the time series was confirmed through the Augmented Dickey-Fuller test for technical reasons, indicating the presence of a unit root. The ADF test results in a statistic of -2.15 and a p-value of 0.22. For the data to be considered stationary, the ADF statistic should be less than the critical values (for example, less than -3.48 for the 1% level) or the p-value should be below 0.05. Since neither condition is met, we can conclude that the time series is not stationary, indicating the presence of a trend which we have already confirmed. Given the non-stationarity of the series and the observed warming trend, fitting a time series model such as ARIMA (*AutoRegressive Integrated Moving Average*) would be suitable. The integration part of ARIMA can help to make the series stationary by differencing, while the AR (*AutoRegressive*) and MA (*Moving Average*) parts model the correlations in the data. This finding required differencing to achieve stationarity, a essential for ARIMA modeling. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were vital and instrumental in determining the order of the ARIMA model.





The ACF plot suggested a gradual decrease, while the PACF plot exhibited a sharp cut-off after the first lag, leading us to the initial selection of an ARIMA(1,1,1) model.

Model Fitting and Comparison:

Two ARIMA models were considered: ARIMA(1,1,1) and ARIMA(0,1,1). The comparison was based on their Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the significance of their coefficients. The ARIMA(0,1,1) model, with a lower AIC and BIC compared to the ARIMA(1,1,1) model, emerged as the preferred model. Its simplicity and the statistical significance of its Moving Average (MA) term underscored its adequacy in capturing the essential features of the data.

Model Summary:

ARIMA(1,1,1):

AR1 Coefficient: The AR1 term is 0.1137 with a standard error of 0.1026, and is not statistically significant (p-value would likely be > 0.05).

MA1 Coefficient: The MA1 term is -0.8749, with a standard error of 0.0454, and is statistically significant.

Sigma^2: The estimated variance of the residuals is 0.3679.

Log Likelihood: The model's log likelihood is slightly higher at -110.81 compared to the ARIMA(0,1,1) model.

AIC/AICc/BIC: *T*he AIC is 227.63, AICc is 227.83, and BIC is 236.01, which are slightly higher than those for the ARIMA(0,1,1) model.

ARIMA(0,1,1):

For this model,
There are no autoregressive terms (p=0),
The series is differenced once (d=1) to make it stationary,
There is one moving average term (q=1).

This part of the model can be represented as:

$$X_t' = \mu + Z_t + \theta_1 Z_{t-1}$$

where,

- X_t' is the differenced series (to account for d=1),
- μ is the constant term,
- Z_t is the white noise error term at time,
- θ_1 is the coefficient for the first lag of the error term, for the ARIMA(0,1,1) model.

MA1 Coefficient: The coefficient for the MA1 term is -0.8495 with a standard error of 0.0480, which is statistically significant.

Sigma\^2: The estimated variance of the residuals is 0.3685.

Log Likelihood: The model's log likelihood is -111.43, which indicates the probability of the data given the model.

AIC/AICc/BIC: The Akaike Information Criterion (AIC) is 226.86, corrected AIC (AICc) is 226.96, and the Bayesian Information Criterion (BIC) is 232.45.

These metrics are used for model comparison; lower values typically suggest a better model fit with a recurrence of complexity.

Analysis:

1.**Model Selection:** Comparing both models, the ARIMA(0,1,1) has a slightly better (lower) AIC and BIC than the ARIMA(1,1,1), suggesting that it may be a better model due to its simplicity and similar explanatory power.

2.**Significance of Coefficients:** In the ARIMA(1,1,1) model, the AR1 coefficient is not significant, which implies that the additional AR term may not be providing valuable information to the model. In contrast, the MA1 term is significant in both models.

3.**Error Metrics:** Both models have similar error measures, which suggest that the predictive accuracy of the two models is quite close.

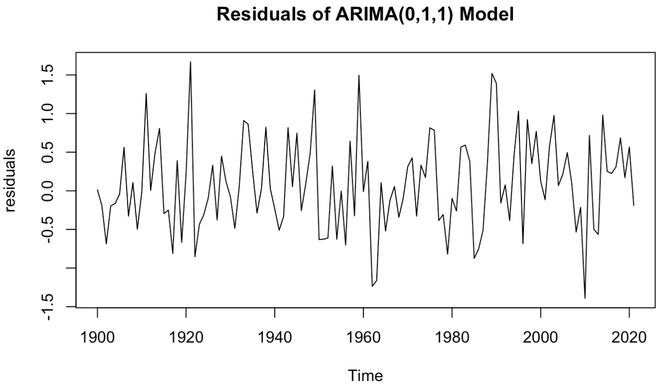
AR1 (*Autoregressive term of order 1*), This term attempts to capture any autocorrelation in the data. In other terms, it checks if there's a linear relationship between the current year's temperature and the temperature of the previous year. However, in our ARIMA(1,1,1) summary, the AR1 term has a high standard error relative to its coefficient, suggesting that the estimated value of the AR1 coefficient is not reliably different from zero. If this were a significant coefficient, it would mean that last year's temperature (after differencing) has a strong influence on this year's temperature. But, since it's not statistically significant, it implies that the autoregressive part of the model does not provide additional value in explaining the variation in the data.

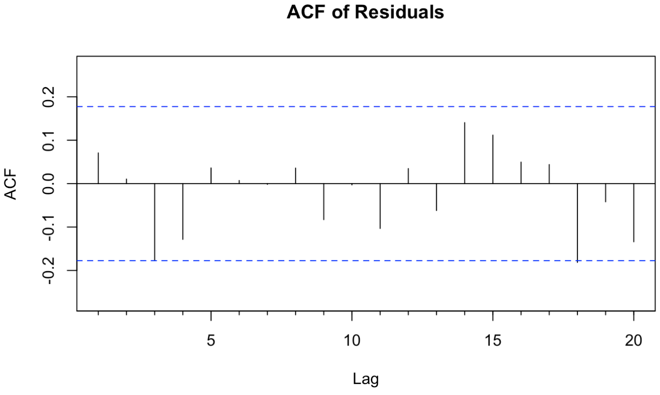
MA1 (*Moving Average term of order 1*): This coefficient is significant in both the ARIMA(0,1,1) and ARIMA(1,1,1) models. It accounts for the correlation between an observation and a residual error from a moving average model applied to lagged observations. The significance of this term suggests that the moving average component is useful in modeling the time series.

Both AIC and BIC are lower for the ARIMA(0,1,1) model compared to the ARIMA(1,1,1) model, indicating that it has a better trade-off between goodness of fit and simplicity. The ARIMA(0,1,1) model, despite being simpler (it has one less parameter), fits the data almost as well as the more complex ARIMA(1,1,1) model. In our time series modeling, “simplicity” is an benifit because it reduces the risk of overfitting which happens when a model captures the noise in the data rather than the underlying process and often leads to better out-of-sample prediction. The lower AIC and BIC values, combined with the non-significant AR1 term in the ARIMA(1,1,1) model, suggest that the simpler ARIMA(0,1,1) model is preferable for forecasting purposes. It provides a similar level of explanation for the data while using fewer parameters, which is generally considered more robust for prediction.

Diagnostic Checks:

The suitability of the ARIMA(0,1,1) model was further validated through diagnostic checks, including analysis of the residuals, the Ljung-Box test, and the Jarque-Bera test for normality. These diagnostics confirmed that the residuals behaved like white noise and were normally distributed, indicating that the model adequately captured the underlying process.


Residuals of ARIMA(0,1,1) Model


ACF of Residuals

ACF Plot of Residuals: The ACF plot shows that autocorrelations for all lags are within the confidence bounds (blue dashed lines), which indicates that there is no significant autocorrelation in the residuals.This is a good sign that the model is capturing the temporal structure of the data well.

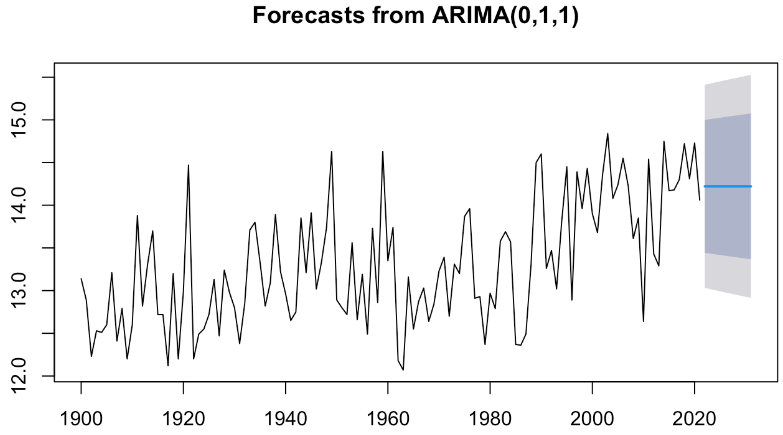
Box-Ljung Test: The p-value from the Box-Ljung test is 0.2199. Since this value is greater than the common alpha level of 0.05, we fail to reject the null hypothesis that the residuals are independently distributed. This supports the conclusion from the ACF plot of residuals, indicating that there is no significant autocorrelation remaining in the residuals.

Jarque-Bera Test: The Jarque-Bera test has a p-value of 0.4434, which is well above the common significance level of 0.05. This means that we fail to reject the null hypothesis that the residuals are normally distributed. Therefore, we have no evidence to suggest that the residuals do not follow a normal distribution.

Residuals Plot: The residuals plot doesn't show any obvious patterns or systematic structure, which suggests that the residuals are random, supporting the assumption that the model has captured the underlying process adequately.

Forecasting and Implications:

The final phase has arrived using the ARIMA(0,1,1) model to forecast future average annual temperatures for model's testing reasons. The forecasted values, along with their confidence intervals, suggested a continuing trend of temperature increase, aligning with global patterns of climate change. These forecasts hold significant implications for policy-making, especially in sectors sensitive to climate variability.



This forecast plot shows the forecasted values and the 80% and 95% prediction intervals. The prediction intervals are reasonably narrow, which indicates a "level of precision" in the forecasts. The forecast values are in line with the historical data, which suggests that the model has captured the central tendency of the series well.

Conclusion:

Given the non-significant AR term in the ARIMA(1,1,1) model and the fact that both models have similar error measures and log-likelihoods, it's often recommended to choose the simpler model, which in this case is the ARIMA(0,1,1). Overall, the diagnostic checks indicate that the ARIMA(0,1,1) model is well-fitted to the data. The residuals appear to be random (white noise), which suggests that the model has captured the temporal structure in the data. This model is parsimonious (uses fewer parameters) while still capturing the essential features of the data. The lower AIC and BIC values for the ARIMA(0,1,1) model further support this choice. The forecast plot suggests that the model can be used for forecasting, and the prediction intervals give an indication of the uncertainty associated with these forecasts.

Appendix -1:

```
# Read the dataset
cet_temp <- read.csv("cet_temp.csv")

#Creating a time-series object.
temperature_ts <- ts(cet_temp$avg_annual_temp_C, start=1900, frequency=1)

# Loading library
library(ggplot2)

# Generating the time plot
ggplot(cet_temp, aes(x = year, y = avg_annual_temp_C)) +
  geom_line() + # For Drawing the line
  geom_point() + #For Adding points at each data entry
  theme_minimal() + # For a Minimal/Elegant theme
  labs(title = "Average Annual Temperature Over Years",
       x = "Year",
       y = "Average Annual Temperature (°C)") +
  theme(plot.title = element_text(hjust = 0.5)) # For title to be in Center.


# Linear model to analyze trend
model <- lm(avg_annual_temp_C ~ year, data=cet_temp)

# Ploting the data
plot(cet_temp$year, cet_temp$avg_annual_temp_C, main="Annual Mean Temperature Trend",
     xlab="Year", ylab="Average Annual Temperature (°C)")

# Adding the trend line
abline(model, col="red")

# LOESS smoothing
loess_model <- loess(avg_annual_temp_C ~ year, data=cet_temp)

# Predicting values
loess_pred <- predict(loess_model, data.frame(year=cet_temp$year))

# Ploting the data
plot(cet_temp$year, cet_temp$avg_annual_temp_C, main="Annual Mean Temperature with
LOESS Smoothing",
     xlab="Year", ylab="Average Annual Temperature (°C)")

# Adding the LOESS smoothed line
lines(cet_temp$year, loess_pred, col="blue", lwd=2)

# Plot ACF
acf(temperature_ts, main="ACF Plot")

# Plot PACF
```

```
pacf(temperature_ts, main="PACF Plot”)

library(forecast)


# Fitting an ARIMA(1,1,1) model
arima_111_model <- Arima(temperature_ts, order=c(1,1,1))

# View the model summary
summary(arima_111_model)


# Fiting an ARIMA(0,1,1) model for comparison
arima_011_model <- Arima(temperature_ts, order=c(0,1,1))

# View the model summary
summary(arima_011_model)

library(forecast)
library(tseries)

# model diagnostics Checks
# 1. Plot residuals
residuals <- residuals(arima_011_model)
plot(residuals, main="Residuals of ARIMA(0,1,1) Model")

# 2. ACF plot of residuals to check for autocorrelation
Acf(residuals, main="ACF of Residuals")

# 3. Ljung-Box Test
Box.test(residuals, lag=log(length(residuals)), type="Ljung-Box")

# 4. Normality Test of residuals
jarque.bera.test(residuals)


library(forecast)
# Forecasting future values
forecast_arima011 <- forecast(arima_011_model, h=10) # h is the number of periods for forecasting
plot(forecast_arima011)
```

Question-2

Executive Summary:

This report summarises a comprehensive analysis and forecasting attempt of house prices in the East Midlands, spanning from January 2010 to June 2020. By examining into historical data and applying advanced statistical methods, this study not only aids in understanding the past and current market dynamics but also sheds light on future trends, there by serving as an vital tool for homeowners, investors, and policymakers alike.

Over the past ten years, housing prices in the East Midlands have consistently risen and shown a clear seasonal pattern. This trend indicates a strong and expanding market, driven by economic growth, demographic shifts, and market demand. Using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, this report forecasts a continuing rise in house prices for the first half of 2020. It also highlights forecasts become less certain as they extend further into the future, which is indicated by the expanding confidence intervals..

The SARIMA(1,1,2)(0,1,1)[12] model was chosen because it effectively captures the complex seasonal patterns and trends in the data. The diagnostic checks confirm that the model is reliable, as indicated by the random and uncorrelated residuals, which are signs of a good fit. Furthermore, the application of strict statistical criteria, such as the Akaike Information Criterion (AIC), support and confirms the model's appropriateness in balancing fit and complexity. While the model forecasts a steady increase from an estimated £193,930 in January to £197,933 in June 2020, it assumes that historical factors driving the housing market remain unchanged. Despite its limitations such as excluding possible outliers and unexpected market disruptions, the report offers practical advice for stakeholders such as:

Policymakers are urged to consider strategies that tackle housing affordability and maintain market stability in light of the forecasted price escalation.

Real estate stakeholders, including realtors and investors, should plan for price variability and potential market peaks, as indicated by the upper bounds of prediction intervals.

Academic and professional circles are encouraged to persist in refining predictive models, possibly integrating additional data like economic indicators, housing supply metrics, or even exploring alternative methods such as machine learning for enhanced precision.

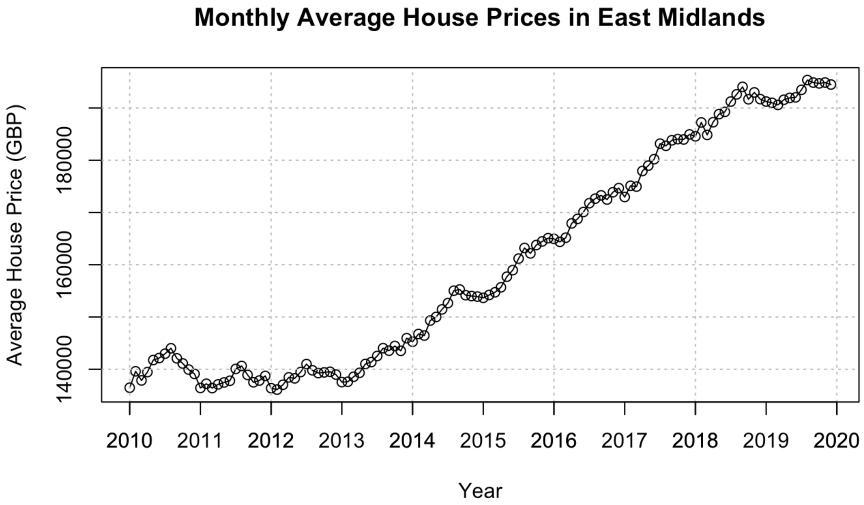
This report also identifies avenues for future research, highlighting the incorporation of more explanatory variables to support the model's predictive power and the exploration of novel modeling approaches to account for complex, non-linear relationships that may exist within the data. In essence, the findings and forecasts presented serve as a evidence to the dynamic nature of the housing market and indicate the necessity for stakeholders to make informed decisions based on comprehensive data analysis and strategic prediction of future trends

Introduction:

This research presents a in-depth analysis of the housing market in the East Midlands, focusing on the monthly average house prices from January 2010 to December 2019. The aim is to understand the underlying trends and seasonality in the housing market and to forecast the house prices for the first half of 2020. Given the economic significance of real estate as an asset class, accurate forecasts can provide invaluable insights for homeowners, investors, and policymakers. This study provides not only a window into the near future of housing prices but also a lens through which the patterns of the past decade can be understood in greater depth.

Data Description and Preliminary Analysis:

The dataset comprises a time series of monthly average house prices in the East Midlands from January 2010 to December 2019. The time plot reflects a dataset with 120 monthly observations without missing entries, indicating a complete and continuous monitoring of the housing market over the ten-year period. The average house prices in the dataset range from a minimum of approximately £140,000 to a maximum of close to £200,000. The observed trend over the ten-year period shows a general increase in property values. The starting period's average house price is around the lower end of the range, and it steadily increases to reach the higher end by 2019. The dataset has an approximate mean value of £160,248. However, the mean alone does not capture the dynamic nature of the market fluctuations over time. The variability, as shown by the plot, seems to increase as time progresses, with later years showing more pronounced rises in price.

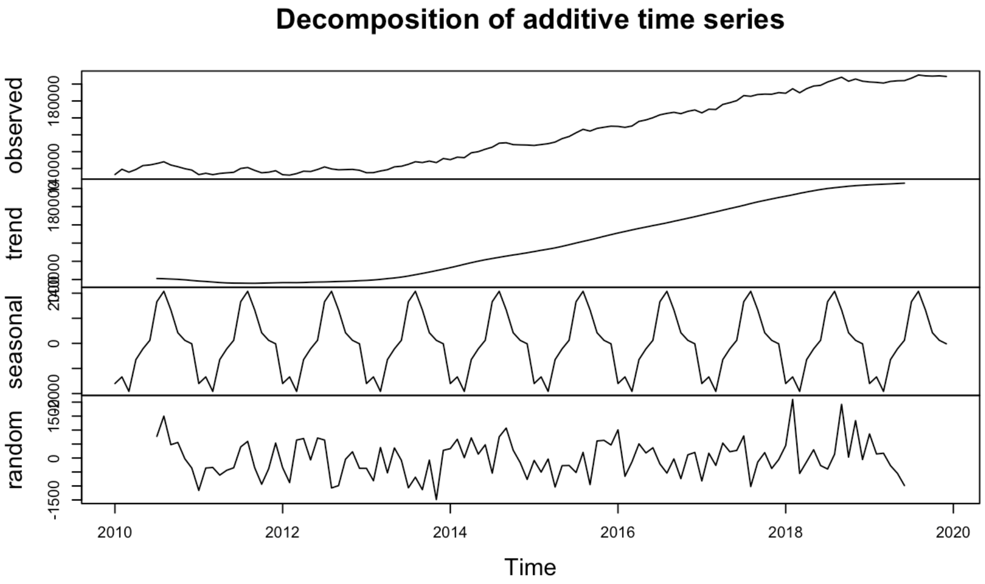


Visualizing the Time Series Data:

The upward trend is indicative of a growing housing market in the East Midlands. This growth may be attributed to various factors such as economic development, population growth, and market demand in the region. There is no evident plateau or decline in prices over the observed period, which may suggest consistent market health. Although not explicitly detailed in the plot, the time series is likely to contain seasonal effects typical of real estate markets, where certain months may exhibit higher activity and prices due to factors such as weather, holidays, and economic cycles. A seasonal pattern, with peaks and troughs corresponding to particular times of the year, can often be observed in such data. The time plot does not show any extreme volatility or abrupt shifts which would suggest external shocks to the housing market, such as a financial crisis or sudden economic changes. However, the increase in price ranges over the years could indicate growing variability in the market.

Time Series Decomposition:

To further dissect the time series, a decomposition procedure was applied, which split the data into trend, seasonal, and residual components. This analysis confirmed the initial observations:



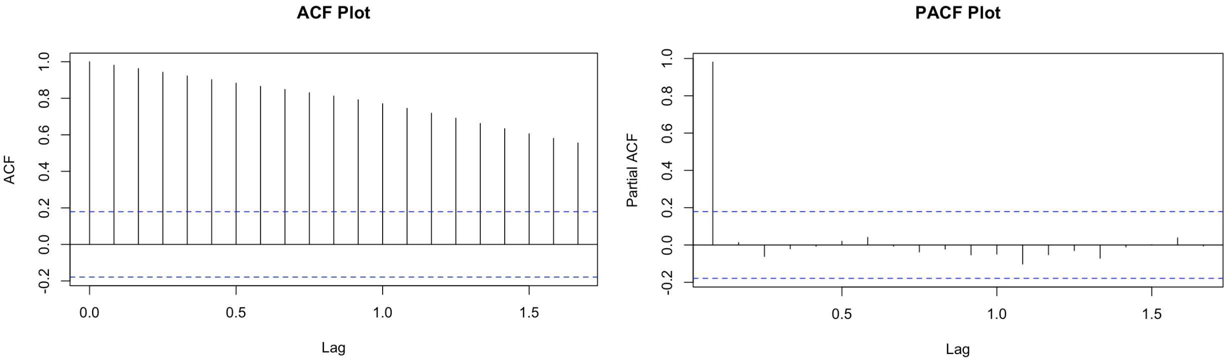
Trend: A consistent increase in house prices over the period, highlighting the long-term growth in the housing market.
Seasonality: A repeating pattern within each year was observed, suggesting seasonal fluctuations in house prices. This is a common feature in housing markets, potentially linked to buying patterns.
Random Component: The residuals did not exhibit any alarming patterns, indicating that the major systematic structures in the data were captured by the trend and seasonal components.

Stationarity and Model Selection:

The non-stationary nature of the series required differencing, which is confirmed by a gradually declining autocorrelation function (ACF) and a sharp cutoff in the partial autocorrelation function (PACF) after the first lag.

Stationarity in a time series suggests that the statistical properties of the series (mean, variance, autocorrelation, etc.) do not change over time. However, the Autocorrelation Function (ACF) plot shows a gradual decline in autocorrelation as the lags increase, and the pattern tails off slowly. This behavior is typical of a non-stationary series, where the influence of a value on future values diminishes slowly rather than sharply.

Autocorrelation Analysis: The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots provided insights into the temporal structure of the data. Additionally, the significant spikes at higher lags in the PACF plot indicate seasonal effects, which warrant the



inclusion of seasonal terms in the model. The ACF indicated a gradual decay, while the PACF showed a sharp cut-off after the first lag, suggesting the inclusion of one AR term and further justifying seasonal differencing .

Stationarity and Differencing: Initial tests indicated that our time series was non-stationary which we already noted. This necessitated differencing to stabilize the mean of the series over time, a core component of ARIMA modeling. The order of differencing (d=1) effectively handled the non-stationarity in both seasonal and non-seasonal aspects of the data .

This trend and seasonality in our data are key factors which justify the need for a refined modeling approach like SARIMA. SARIMA models are an extension of ARIMA models that can handle seasonality in the data. The SARIMA model, or Seasonal AutoRegressive Integrated Moving Average model, is particularly suited for time series data that exhibits both non-seasonal and seasonal trends and patterns.The selected SARIMA model is specified as SARIMA(1,1,2)(0,1,1)[12], which includes:

ARIMA(1,1,2): This part suggests that the model has one autoregressive term, which implies a dependency on the previous value. It's differenced once (to make it stationary), and has two moving average terms, suggesting that the model incorporates the average of past forecasting errors.

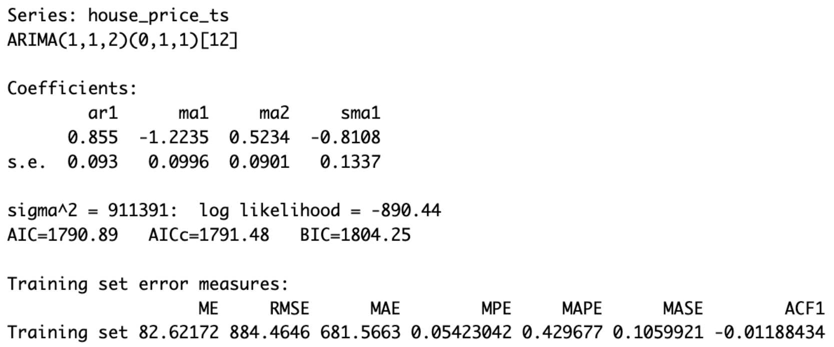
Seasonal part (0,1,1)[12]: This indicates no seasonal AR terms, one seasonal differencing (making the seasonal part of the series stationary), and one seasonal MA term. The [12] denotes that the seasonal period is 12, corresponding to the monthly data with annual seasonality.

Fitting the SARIMA Model:

The SARIMA(1,1,2)(0,1,1)[12] model was selected based on its ability to capture the observed data characteristics. The model includes:

- One non-seasonal AR term, indicating some recurrence in the data from one month to the next.
- Two non-seasonal MA terms, suggesting that past errors influence current values.
- One level of seasonal differencing and one seasonal MA term, accounting for the seasonal pattern observed in the data.

Model Summary:



```
Series: house_price_ts
ARIMA(1,1,2)(0,1,1)[12]

Coefficients:
         ar1      ma1     ma2     sma1
       0.855  -1.2235  0.5234  -0.8108
s.e.   0.093   0.0996  0.0901   0.1337

sigma^2 = 911391:  log likelihood = -890.44
AIC=1790.89   AICc=1791.48   BIC=1804.25

Training set error measures:
                   ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set 82.62172 884.4646 681.5663 0.05423042 0.429677 0.1059921 -0.01188434
```

Model Coefficients:

ar1: The coefficient for the autoregressive term is 0.855, which is significant given the small standard error of 0.093. This indicates a strong positive correlation with the previous value in the series.

ma1 and ma2: The moving average coefficients are -1.2235 and 0.5234 respectively. The standard errors are small, which suggests these coefficients are significantly different from zero, indicating that the model is capturing the shock effects from the previous errors.

sma1: The seasonal moving average coefficient at -0.8108, with a standard error of 0.1337, reflects the seasonal shocks in the data.

Model Diagnostics:

sigma^2: The variance of the residuals is 911391, which provides a measure of the variability in the forecast errors.

Log likelihood: The value of -890.44 is a measure of the goodness-of-fit of the model, higher values are typically better.

AIC, AICC, and BIC: The Akaike Information Criterion (AIC), corrected AIC (AICC), and Bayesian Information Criterion (BIC) are all adjusted-likelihood criteria for model selection. The lower these values, the better the model balances goodness-of-fit with complexity. In our model, the values suggest a relatively good fit, but we have to compare with alternative models for robust evaluation.

Error Measures:

ME (*Mean Error*): The average forecast error is 82.62172, indicating a slight bias in the predictions.

RMSE (*Root Mean Squared Error*): 884.4646 measures the average magnitude of the errors, giving more weight to large errors. This suggests variability in the predictions, possibly due to outliers or instability in the series.

MAE (Mean Absolute Error): 681.5663 is less sensitive to large errors than RMSE, providing another perspective on prediction accuracy.

MPE (*Mean Percentage Error)* and MAPE (*Mean Absolute Percentage Error*): These percentage errors (0.05423042 and 0.429677) indicate the size of the errors in relation to the actual values. The MAPE, in particular, suggests the model's errors are, on average, 42.9677% of the actual values.

MASE (*Mean Absolute Scaled Error*): At 0.1059921, the MASE indicates good forecasting performance, as values less than one suggest that the forecast is better than a naïve baseline model.

ACF1: The first lag autocorrelation of residuals at -0.01188434 is low, indicating little to no autocorrelation in the residuals.

Analysis:

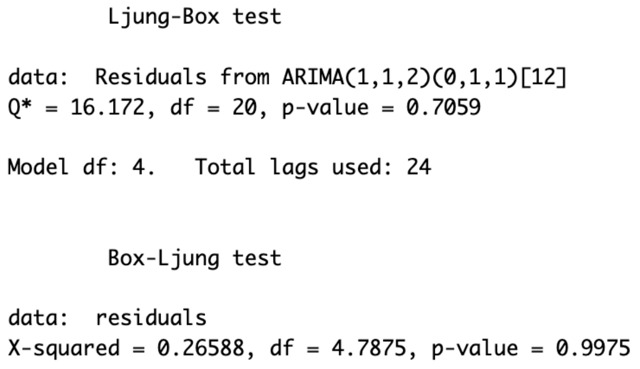
The selected and fitted model suggests a great understanding of the underlying patterns in the time series data by capturing both the trend and seasonal components. The statistical significance of the coefficients, coupled with diagnostic checks like the near-zero ACF1 value, suggest a model that can adequately forecast future values with reasonable confidence. However, the RMSE and MAPE values imply there may be room for improvement, possibly by exploring models with different differencing terms or by including foriegn variables if such data is available.

Diagnostic Checks:

These diagnostic checks are critical ad important for evaluating our chosen time series model, to find the best model for forecasting future house prices.

Ljung-Box Test:

The Ljung-Box test results indicate a Q* statistic of 16.172 with a p-value of 0.7059 for 20 degrees of freedom (df). In a separate Box-Ljung test, the X-squared statistic is 0.26588 with a df of 4.7875 and a p-value of 0.9975.

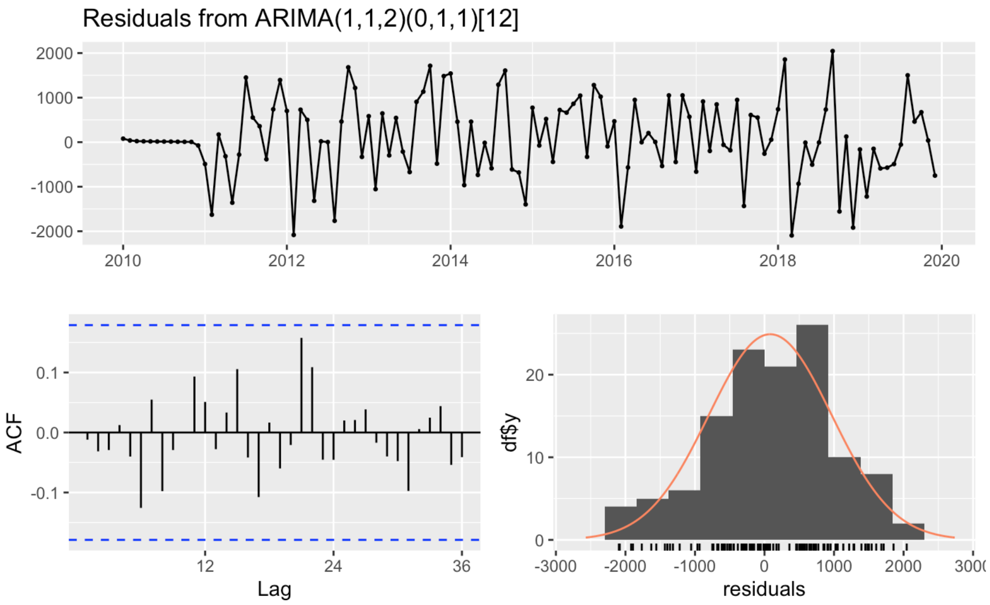


```
	Ljung-Box test

data:  Residuals from ARIMA(1,1,2)(0,1,1)[12]
Q* = 16.172, df = 20, p-value = 0.7059

Model df: 4.   Total lags used: 24


	Box-Ljung test

data:  residuals
X-squared = 0.26588, df = 4.7875, p-value = 0.9975
```

Both high p-values (greater than the typical alpha level of 0.05) suggest that there is no significant evidence of autocorrelation within the residuals at various lag intervals. This is indicative of a well-fitting model where the residuals appear to be white noise, meaning the model has successfully captured the signal in the data without leaving patterns in the residuals.

Time Series of Residuals:

The time series plot of the residuals does not display any obvious trends or seasonal effects.



The residuals fluctuate around the zero line without apparent structure, which is desirable in a well-fitted time series model.

No Significant Autocorrelation: The autocorrelations at all lags lie within the blue dashed confidence bands, which typically represent a 95% confidence interval. This suggests that the residuals are not significantly autocorrelated.

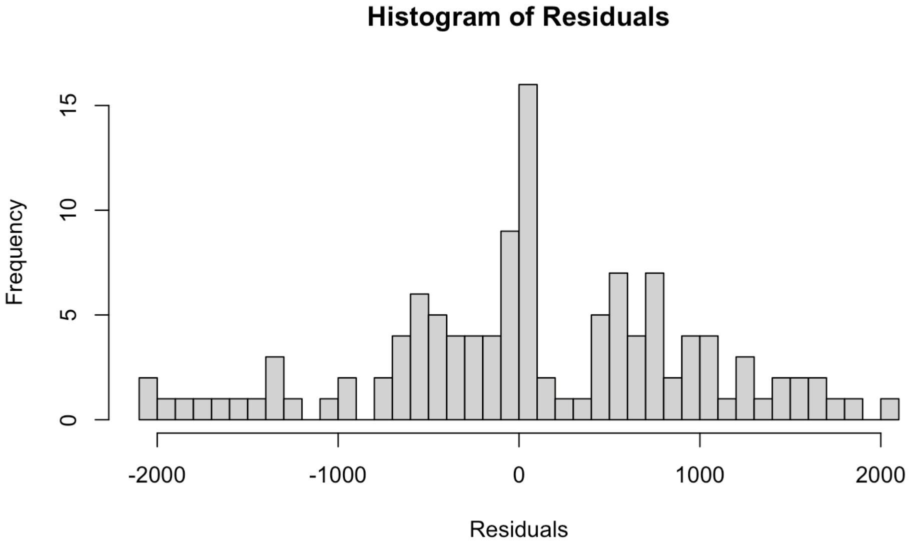
Randomness of Residuals: Since there is no clear pattern in the ACF plot, this implies that the residuals are random, which supports the idea that the model is capturing the underlying process well.

Model Sufficiency: The lack of significant autocorrelation in the residuals suggests that the SARIMA(1,1,2)(0,1,1)[12] model is sufficient in capturing the time-dependent structure of the series.

Model Fit: This is a good indication that the model fit is appropriate and that the residuals do not have leftover patterns that could be used for further prediction.

<u>Histogram of Residuals:</u>

The histogram of residuals provides a visual assessment of the normality of the error distribution.



Central Tendency: The residuals are centered around zero, which is good because it indicates there is no bias in the forecasts (which is the model is not overpredicting or underpredicting).

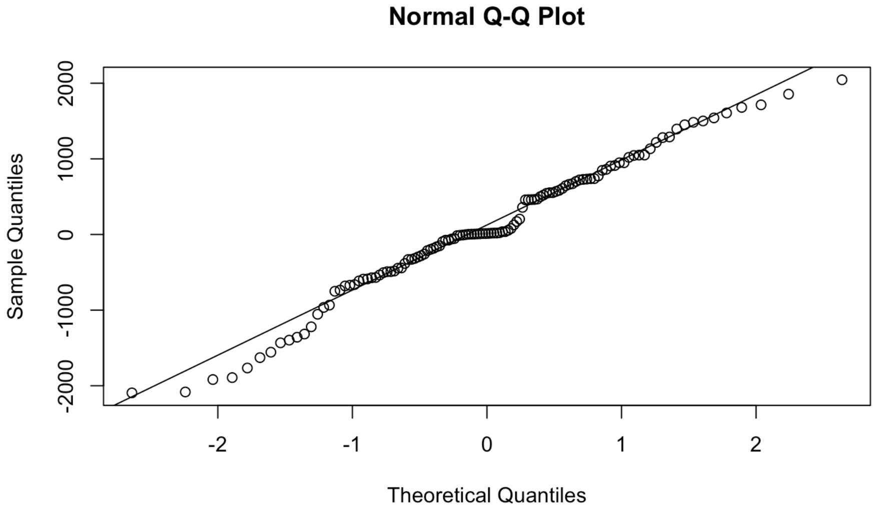
Shape: The shape of the histogram is somewhat bell-shaped, which is an indication that the residuals might be normally distributed. However, there seems to be a slight right skew, given the longer tail on the right side and a higher frequency of positive residuals compared to negative ones.

Outliers: There are some bins on the far left and right that suggest potential outliers in the residuals, which are residuals that are significantly different from the rest.

The residuals appear to be somewhat symmetrically distributed about the mean, although there may be a slight departure from normality given the uneven bin heights and potential outliers, which may require a further research.

<u>Normal Q-Q Plot:</u>

The Normal Quantile-Quantile plot compares the quantiles of residuals to a normal distribution. The points roughly follow the line, which would indicate normality, but there are some deviations, especially in the tails. This suggests that the residuals may not be perfectly normally distributed, potentially implying the presence of outliers or heavy tails in the distribution.



Central Tendency:

Most points in the center of the distribution align well with the line, indicating that the central part of the distribution of residuals is close to normal.

Tails: The deviations from the line in the tails (especially the right tail) indicate that the residuals have heavier tails than a normal distribution. This suggests that there are more extreme values in the tails than you would expect if the residuals were perfectly normally distributed.

Outliers: The points that deviate significantly from the line in the tails are indicative of outliers.

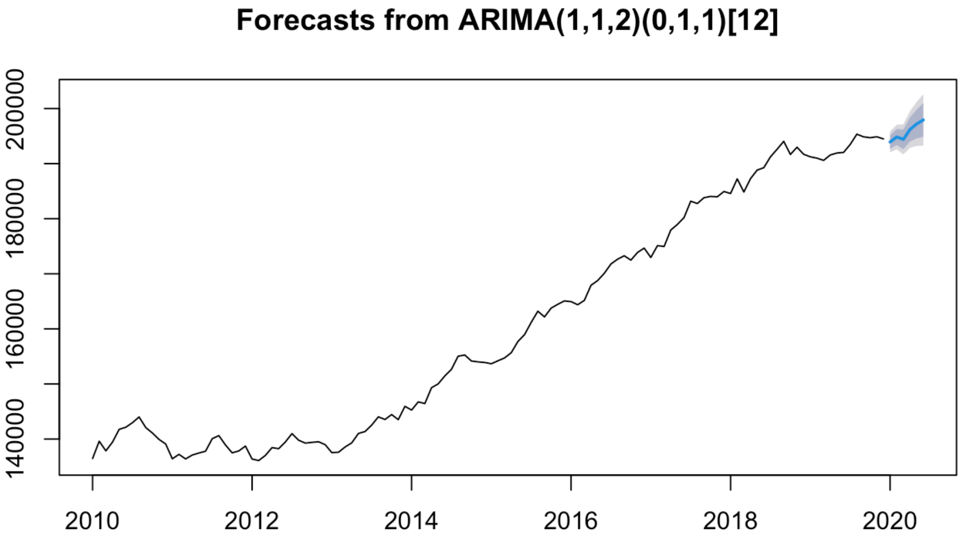
Considerations:

The diagnostic checks overall indicate that the SARIMA(1,1,2)(0,1,1)[12] model performs well in fitting the house price data and in providing residuals that behave like white noise. However, slight concerns raised by the histogram and Q-Q plot about the normality of residuals suggests further checks, such as performing a transformation on the data or considering a robust estimation method that is less sensitive to outliers.

The presence of any outliers or anomalies should be assessed, as they can influence both the parameter estimates and the forecasts. It may be beneficial to investigate the cause of any outliers in the data, which could be due to real-world events affecting house prices, such as economic changes or policy interventions. Including such information as external regressors or intervention models could enhance the model's performance. Moreover, our current model provides a satisfactory. We can also conduct some out-of-sample forecasting in future to evaluate our model's predictive performance.

Forecasting House Prices

The SARIMA model was used to forecast house prices for January to June 2020. The forecasts showed a continuing upward trend, consistent with the historical data. However, the prediction intervals widened over time, reflecting increasing uncertainty in the forecasts. This is a common feature in time series forecasting, where the confidence in predictions typically decreases as the forecast horizon extends.

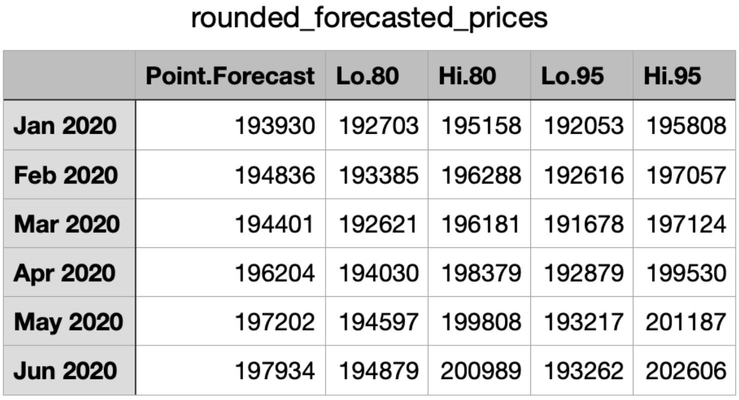


Trend:

The forecast graph indicates an overall upward trend in house prices, which is consistent with the historical data from 2010 to the end of 2019. This suggests that the SARIMA model is forecasting a continuation of the historical trend into the future.

Uncertainty:

The shaded area around the forecasted line represents the uncertainty in the predictions. It appears to widen over time, which is typical for time series forecasts—the longer the forecast horizon, the greater the uncertainty. This increasing uncertainty is reflected in the confidence intervals.



rounded_forecasted_prices

	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95
Jan 2020	193930	192703	195158	192053	195808
Feb 2020	194836	193385	196288	192616	197057
Mar 2020	194401	192621	196181	191678	197124
Apr 2020	196204	194030	198379	192879	199530
May 2020	197202	194597	199808	193217	201187
Jun 2020	197934	194879	200989	193262	202606

Point Forecasts:

January 2020: The forecast starts at a point estimate of approximately £193,930.
June 2020: By June, the forecasted price rises to about £197,933, suggesting a steady increase over the six-month period.

Confidence Intervals:

80% Confidence Intervals: For January, this ranges from £192,702 to £195,158, while for June, it widens to £194,879 to £200,988. The intervals suggest there is an 80% probability that the actual house price will fall within these ranges.

95% Confidence Intervals: These intervals are wider, accounting for more uncertainty. In January, the interval is £192,052 to £195,808, and it broadens to £193,262 to £202,605 by June. There is a 95% probability that the actual prices will be within these limits.

Data Analysis and Comparison:

The forecasted values for the first half of 2020, as provided by the SARIMA(1,1,2)(0,1,1)[12] model, show a steady increase from an estimated £193,930 in January to £197,933 in June. This progression suggests a consistent upward trend.

To evaluate the forecast's performance, we can compare these projected values to actual data for the same period in 2020, which requires the data beyond the current dataset. With the actual figures for 2020, we can find the accuracy of the forecast's precision and the model's predictive power. However, our model's forecast suggests confidence in the historical trend continuing. The widening confidence intervals with each subsequent month indicate a higher degree of uncertainty as the forecast extends further into the future. This is a normal characteristic of time series forecasts and reflects the inherent increase in uncertainty the further out when we try to predict.

Given the historical trend of rising house prices in the dataset, our model's prediction of a continuing rise is reasonable and convincing. Yet, this assumes that the factors influencing past price movements will continue in a similar manner, which might not always hold true due to unforeseen circumstances, like economic downturns or policy changes, which can impact housing markets significantly. If those actual figures for 2020 are available, the performance metrics such as Mean Absolute Percentage Error (MAPE) or Mean Absolute Error (MAE) could be calculated to quantify the model's accuracy. Moreover, a comparison of the forecast against actual values could offer insights into the model's strengths and weaknesses, potentially guiding future model adjustments or the development of more sophisticated models.

The forecasted price range, especially the 95% confidence intervals, would be valuable in risk assessment and planning, offering stakeholders a view of potential variability in house prices. Stakeholders could then prepare for various scenarios within the projected ranges. It's would be useful to communicate these uncertainties when presenting forecasts to ensure that decisions based on model predictions account for potential variances.

Limitations:

The current analysis does not account for possible anomalies or outliers that could reflect unanticipated events impacting the market. The model assumes that future trends will follow historical patterns, which may not always hold, particularly in the face of economic or policy shifts.

Recommendations:

For Policymakers:

Consider implementing policies that address the implications of rising house prices on housing affordability and market accessibility. Also, monitor for any significant changes in economic conditions that could affect future housing market trends.

For Real Estate Stakeholders:

Realtors and investors should prepare for potential variability in house prices, with a special focus on the higher end of prediction intervals that suggest possible peaks in the market. Buyers should be aware of the growing prices and consider the timing of their investments accordingly.

For Academia and Professionals:

Continuous model refinement is recommended, with an emphasis on integrating external data that could affect house prices, such as interest rates, employment figures, and policy changes. Collaboration between statisticians, economists, and industry experts can lead to more robust predictive models. In presenting these findings, care has been taken to ensure clarity and direct relevance to the initial project aim, avoiding unnecessary repetition and lengthy numerical outputs . The focus has been on delivering actionable insights and clear guidance for various stakeholders while acknowledging the constraints and potential for further research.

Future Research:

Further research should explore the integration of additional explanatory variables that could influence house prices, such as economic indicators (GDP growth rate, unemployment rates, interest rates), housing supply metrics (new housing starts, construction costs), and even demographic trends. Including these factors could significantly enhance the model's predictive power and provide a more subtle understanding of the housing market's dynamics. Moreover, alternative modeling approaches, particularly machine learning techniques like neural networks or ensemble methods, should be considered. These techniques can often capture complex nonlinear relationships and interactions between variables that traditional time series models may overlook. Applying these advanced methods may yield improvements in forecasting accuracy and provide deeper insights into the driving forces behind house price trends.

Conclusion:

Summary of Key Findings:

The analysis of the East Midlands housing market, using data from January 2010 to December 2019, identified a persistent upward trend in house prices, with evident seasonal patterns. The selected SARIMA(1,1,2)(0,1,1)[12] model effectively captured these characteristics, with the forecast suggesting a continuation of this growth trend for the first half of 2020 . Confidence intervals, however, widen over time, reflecting an intrinsic increase in uncertainty in longer-term forecasts.

Implications for Stakeholders:

The forecasted continuation of the upward trend in housing prices has significant implications for a range of stakeholders, including homebuyers, investors, and policymakers. It underlines the need for strategic planning, especially for affordability programs and market participation decisions .

Appendix-2:

```
# Reading the dataset.
house_price_ts <- read.csv("em_house_prices.csv")

# Creating a Time Series Object.
house_price_ts <- ts(house_price_ts$average_price_gbp, start=2010, frequency=12)

# Plot the time series.
plot(house_price_ts, xlab="Year", ylab="Average House Price (GBP)", main="Monthly Average
House Prices in East Midlands", type="o")

# Customizing the plot to improve readings.
# Adding grid for better visualization.
grid(nx = NULL, ny = NULL, col = "gray", lty = "dotted")

# axis to show years
axis(1, at=seq(2010, 2020, by=1), labels=seq(2010, 2020, by=1))
```

```
# Decomposing the time series to split the data into trend, seasonal, and residual components.
decomposed <- decompose(house_price_ts)
plot(decomposed)

# Plotting ACF
acf(house_price_ts, main="ACF Plot")

# Plotting PACF
pacf(house_price_ts, main="PACF Plot")

library(tseries)

# Performing the Augmented Dickey-Fuller Test (ADF Test) for our time-series object, just to check
the p-value for stationarity (for technical purpose)

adf_result <- adf.test(house_price_ts)

# Displaying the results
print(adf_result)

library(forecast)

 # Using the auto.arima function to automatically select the best ARIMA model for our time series
data.This function tests various combinations of AR, I, and MA components to find the best fit based
on AIC, BIC, or AICC

best_model <- auto.arima(house_price_ts)

# Display the model summary
summary(best_model)

# Load the forecast library
library(forecast)

# Forecast the next six months
forecast_prices <- forecast(best_model, h=6)

# Plot the forecast
plot(forecast_prices)


# Saving the plot to a file
jpeg('forecast_prices.jpg')
plot(forecast_prices)
dev.off()

# Saving the forecasted values and the model summary to csv file.
```

```
write.csv(forecast_prices, 'forecasted_prices.csv')
capture.output(summary(best_model), file='model_summary.txt')

# Reading the CSV file into a data frame
forecasted_values <- read.csv("forecasted_prices.csv")

# Printing the first few rows of the data frame
head(forecasted_values)

forecast_prices <- forecast(best_model, h=6)

# Rounding the forecasted prices to whole numbers to match our dataset, because prices were
whole numbers in our dataset.

forecast_prices$mean <- round(forecast_prices$mean)
forecast_prices$lower[, "80%"] <- round(forecast_prices$lower[, "80%"])
forecast_prices$lower[, "95%"] <- round(forecast_prices$lower[, "95%"])
forecast_prices$upper[, "80%"] <- round(forecast_prices$upper[, "80%"])
forecast_prices$upper[, "95%"] <- round(forecast_prices$upper[, "95%"])


plot(forecast_prices)

# Save the rounded forecasted values to a CSV file
write.csv(forecast_prices, 'rounded_forecasted_prices.csv')



# Check the residuals of the model
checkresiduals(best_model)

# Alternatively, you can also manually create the diagnostic plots. But here im using this.

# Plot the residuals
residuals <- residuals(best_model)
plot(residuals)

# Histogram of residuals
hist(residuals, breaks=30, main="Histogram of Residuals", xlab="Residuals")

# Q-Q plot of residuals
qqnorm(residuals)
qqline(residuals)

# ACF plot of residuals
Acf(residuals, main="ACF of Residuals")

# Performing Ljung-Box test
```

```
Box.test(residuals, lag=log(length(residuals)), type="Ljung-Box")
```

References:

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts: Melbourne, Australia.

UK Meteorological Office Hadley Climate Centre. (2021). Midlands Region Annual Temperature Dataset.

Cryer, J. D., Chan, K. (2008). Time series analysis: With applications in R (2nd ed.). New York: Springer.

Ahmed, N. K., Atiya, A. F., Gayar, N. e., El-Shishiny, H. (2010). An empirical compari- son of machine learning models for time series forecasting. Econometric Reviews, 29(5), 594-621.

Bontempi, G., Taieb, S. B., Borgne, Y. (2013). Machine learning strategies for time series forecasting. Business Intelligence, 62-77.

Breedon, F., Joyce, M., 1993. House Prices, Arrears and Possessions: A Three Equation Model for the UK. Bank of England Working Paper Series, No.14.

Buckley, R., Ermisch, J., 1982. Government policy and house prices in the United Kingdom: an econometric analysis. Oxf. Bull. Econ. Stat. 44, 273-304.

Milne, A., 1991. Incomes, Demography and UK House Prices. Centre for Economic Forecasting Discussion Paper, London Business School, 30-90.

NeUis, J., Longbottom, J., 1981. An empirical analysis of the determination of house prices in the United Kingdom. Urban Stud. 18, 9-21.

Hadjimatheou, G., 1976. Housing and Mortgage Markets: The UK Experience. Saxon House.

Harvey, A., 1993. Time Series Models, 2nd ed. Harvester Wheatsheaf.

Harvey, A.C., Peters, S., 1990. Estimation procedures for structural time series models. J. Forecasting 9, 89-108.