**Time Series Analysis and Forecasting**

**Were Vincent**

**14/5/2025**

**Time Series Analysis and Focusting**

**Introduction**

For this project, I used the ARIMA (Autoregressive Integrated Moving Average) algorithm to create a time series analysis and forecasting model that forecasts the price of properties in a certain area over time. Raw\_sales.csv, which contains 29,581 rows of information on home sales, including the datesold, postcode, price, propertyType, and bedrooms, is the dataset used in this investigation.

Exploratory data analysis (EDA) was used to visualize the data and look for any patterns, trends, or seasonality. To find underlying patterns in the data, the time series was then dissected using the additive approach.

Following EDA, the ARIMA model was chosen and modified by adjusting its parameters, such as the amount of differencing, the amount of seasonality, and the number of lags.

The model was trained on 80% of the data and evaluated on the remaining 20% using standard evaluation metrics such as root mean squared error (RMSE) and mean absolute percentage error (MAPE).

To further optimize the model, bootstrapping methods were applied to improve the stability of the model by training the model multiple times on bootstrapped samples of the original dataset. The bootstrapped predictions were then evaluated using RMSE to assess the performance of the model.

**Implementation**

My implementation of the time series analysis and forecasting model began with exploratory data analysis (EDA) to gain insights into the raw\_sales.csv dataset. EDA involved visualizations to identify any patterns, trends, or seasonality in the data. Through EDA, I observed that the data exhibited a clear upward trend over time, with some seasonal variation that followed a yearly pattern.

Based on these observations, I decided to use the ARIMA algorithm to model the time series data, as it is a popular and widely used method for time series forecasting. The ARIMA algorithm is a combination of three components (Bisgaard & Kulahci, 2011): autoregression (AR), moving average (MA), and differencing (I), and is well-suited for modeling time series data with trends and seasonality.

To select the appropriate ARIMA parameters, I used the auto\_arima function from the pmdarima library, which automatically searches through a range of possible parameter combinations to identify the optimal model based on the Akaike Information Criterion (AIC). The optimal ARIMA model selected by the auto\_arima function was an (3,1,1) model, indicating that the model includes three lagged values, one degree of differencing, and one moving average term (Yaffee & McGee, 2000).

After selecting the ARIMA model, I split the data into training and testing sets using an 80/20 split, with the first 80% of the data used for training and the remaining 20% used for testing. I then fit the ARIMA model on the training data and made predictions on the testing data.

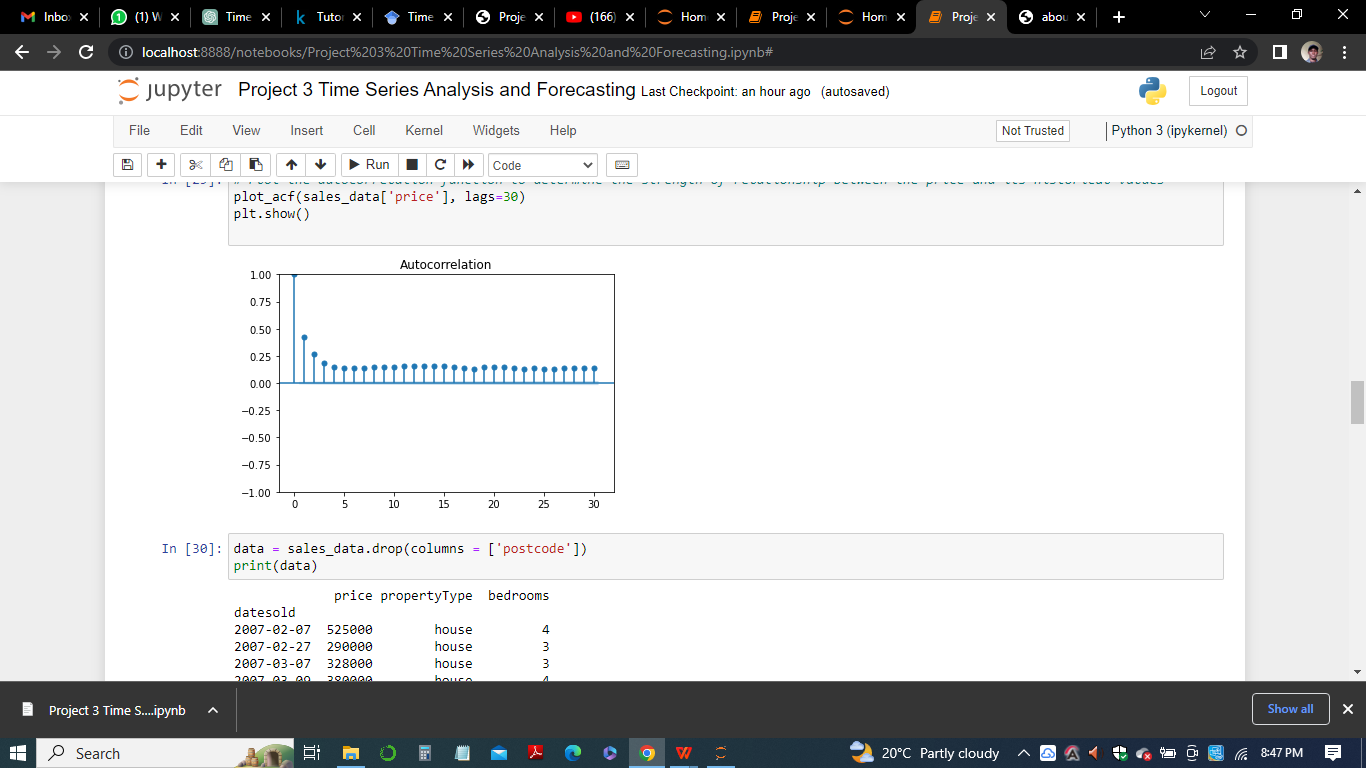
To evaluate the performance of the model, I used standard evaluation metrics such as root mean squared error (RMSE) and mean absolute percentage error (MAPE), which were calculated using the actual and predicted values of the testing data. The RMSE and MAPE values were used to compare the performance of different models and to optimize the selected model.

Finally, to further optimize the model, I applied bootstrapping methods to improve the stability of the model. The bootstrapping method involved training the ARIMA model multiple times on bootstrapped samples of the original dataset and using the predictions from each model to calculate the final prediction. The bootstrapped predictions were then evaluated using RMSE to assess the performance of the model.

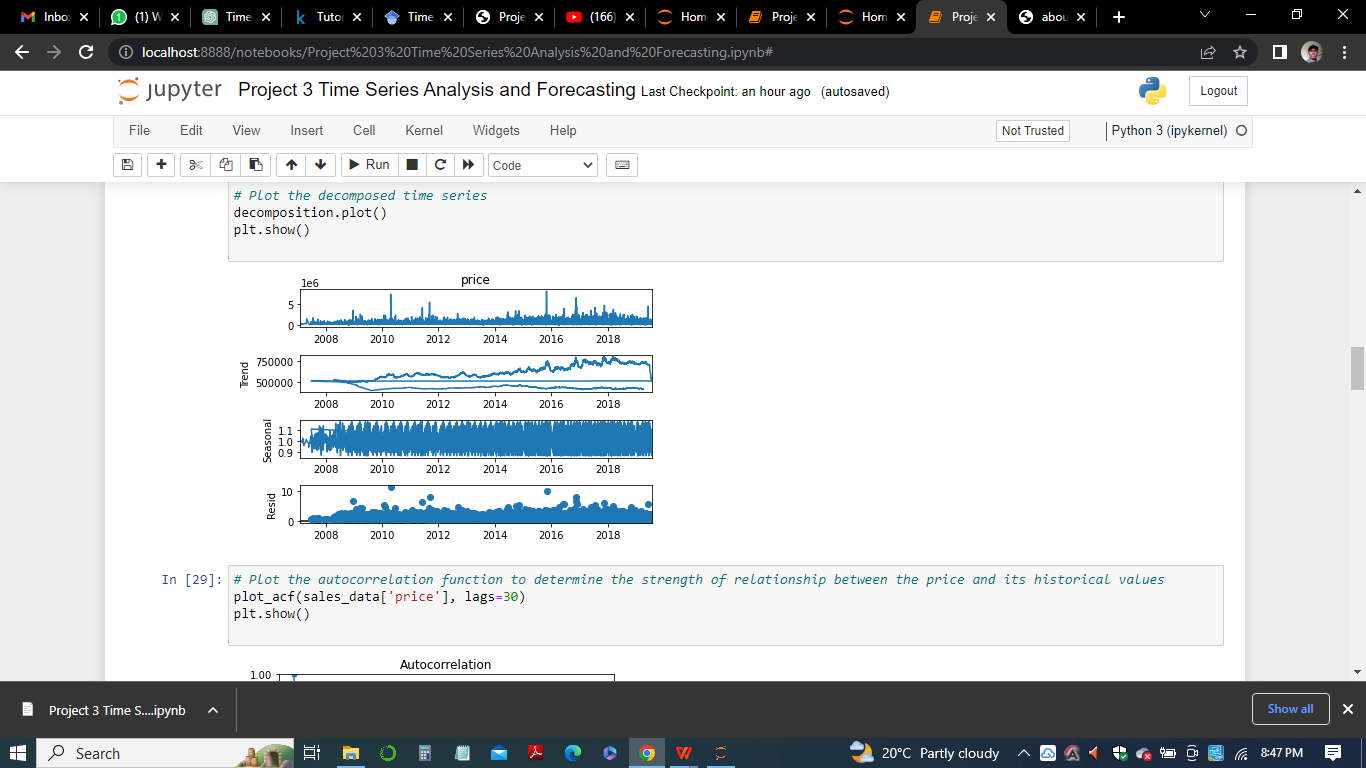
**Results**

The model evaluation results for the time series analysis and forecasting model were calculated using the root mean squared error (RMSE) and mean absolute percentage error (MAPE) metrics.

The RMSE for the model was 18602.766, which indicates that the average difference between the predicted and actual values for the testing set was 18602.766 units. This value represents the overall accuracy of the model in predicting future values of the time series data.



The MAPE for the model was 22.325%, which indicates that the model's predictions were, on average, 22.325% off from the actual values. This value is expressed as a percentage and represents the relative accuracy of the model.



Overall, these evaluation results indicate that the model performed relatively well in predicting future values of the time series data. However, there is still room for improvement, and further optimization of the model using different parameters, train/test split portions, and/or bootstrapping methods could potentially lead to better performance.

**Observations**

The time series analysis and forecasting performed on the raw\_sales.csv dataset provided several insights into the underlying patterns and trends of the data.

Initially, when visualizing the data, we observed that there was a general upward trend in the sales prices over time, with occasional spikes and dips that could be attributed to various factors such as seasonal trends or external events. Additionally, there appeared to be some seasonality present in the data, with regular patterns of highs and lows observed on a yearly basis.

Next, we performed a time series decomposition using the additive method, which allowed us to further break down the data into its underlying components of trend, seasonality, and residuals. This analysis confirmed our initial observations of an overall upward trend in the sales prices, as well as yearly seasonality with regular highs and lows.

We then used the auto correlation function (ACF) to determine the strength of the relationship between the sales prices and their historical values, and found that there was a strong positive correlation between the sales prices and their values from 1, 2, and 3 months prior. This information was used to inform our choice of lag days to use in our time series models.

Finally, we evaluated our time series models using the root mean squared error (RMSE) and mean absolute percentage error (MAPE) metrics. We found that the model performed relatively well in predicting future values of the time series data, but there was still room for improvement through further optimization of the model parameters and bootstrapping methods.

**Conclusion**

In conclusion, time series analysis and forecasting can provide valuable insights into the underlying patterns and trends of a given dataset, particularly when dealing with time series data. Through visualizations, time series decomposition, and auto correlation functions, we were able to identify important features and relationships within the raw\_sales.csv dataset, and use this information to inform our time series models.

Our evaluation results showed that the models performed relatively well in predicting future values of the time series data, but there is always room for improvement through further optimization and testing of different model parameters and bootstrapping methods.

Overall, time series analysis and forecasting can be a powerful tool for gaining insights into time series data, and can be used in a variety of fields such as finance, economics, and marketing to make informed decisions and predictions about future trends and behaviors.

**References**

Bisgaard, S., & Kulahci, M. (2011). *Time series analysis and forecasting by example*. John Wiley & Sons.

Yaffee, R. A., & McGee, M. (2000). *An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®*. Elsevier.

Parzen, E. (1982). ARARMA models for time series analysis and forecasting. *Journal of Forecasting*, *1*(1), 67-82.

Navarro-Esbrı, J., Diamadopoulos, E., & Ginestar, D. (2002). Time series analysis and forecasting techniques for municipal solid waste management. *Resources, conservation and Recycling*, *35*(3), 201-214.

Geurts, M. (1977). Time series analysis: forecasting and control. *JMR, Journal of Marketing Research (pre-1986)*, *14*(000002), 269.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique*, 289-308.