• Seasonal decomposition plots for Time Series 1 and Time Series 2

• Periodogram plots for both time series, highlighting significant lags

• Histograms of logarithmic returns for both time series, showing non-normal distributions

• Scatter plot of TS1 vs. TS2 returns, illustrating covariance normalised stock returns

I am writing the discussion section of my research. These were my research questions. “

investigate suitable approach for prediction when there is high volatility missing data and outliers due to extreme events

1. How does high volatility in stock price data impact the accuracy of traditional time series prediction models, and which models are most robust to high volatility, missing data, and outliers?
2. What are the most effective methods for handling missing data and outliers in the context of stock price prediction, considering high volatility and extreme events?
3. How do different preprocessing techniques, such as data imputation or outlier removal, affect the performance of prediction models in volatile stock price data?
4. What are the main challenges in predicting stock prices for the given dataset, and can these challenges be addressed using existing prediction methods or is there a need for new methods?
5. How do various data imputation techniques, such as linear interpolation and rolling mean, impact the accuracy of stock price predictions when dealing with missing data?
6. In what ways can outlier detection techniques be integrated into the stock price prediction process to enhance the model's robustness against extreme events and improve overall predictive performance?
7. How can the relationships between the two unnamed price series be leveraged to improve stock price predictions, and what is the potential for cross-series prediction, such as using one price series to predict the other or exploiting common trends and patterns between the two series?

“

This is the guideline for this section:

“Analyze how different prediction methods perform in the presence of high volatility and extreme events.

* Results and conclusions: 
  + a. Summarize the findings, highlighting the most effective techniques for handling missing data, outliers, and the challenges of stock price prediction in the presence of high volatility and extreme events.
  + b. Provide insights into the relationships between the two time series and their predictability.

“

Read my notes in light of the guideline and research questions and critically evaluate what analysis is missing, discussion is required based on the points I have so far. Make suggestions for what to include and rewrite these suggestions as dot points within my notes. These are my notes:

“The data sets prepared with the missing values removed preformed better. There was only improved in the case of hypothesis two were the outlier features added slightly improved the prediction.

Models

Ensemble forest methods proved more capable of handling the volatility in the price returns. While more simple decision tree methods were not

Experiment:

Prediction was not improved by imputing the missing data

Outliers as a feature itself did not significantly improve results.

Results suggest that even though interpolation has the least effect on the distribution of returns of the three imputation methods investigated, the train and test set prediction errors were

Improvements in error between train and test sets for the unimputed data with and without outliers, suggest that the model was able to generalise well to the training set. However in the case of the imputed data, the large increase in error suggests that the models likely overfit the training data, possibly due to the bias introduced by the interpolation method of the value of the missing data points. Suggests that artificially synthesising data in this case lead to lesser predictive ability of the model and likely was further propagated in the outlier features.

Hypotheses:

Results of comparison of h1 and h4 hypotheses and h2 and h3 show that using the history of the given stock to predict itself is more accurate than any correlations used for prediction that may be learned by the ML models between the two stocks.

Benchmarks:

Larger error than the benchmark random walk suggests that in most cases, the underlying returns behaviours was not captured well by the models.

It's essential to acknowledge the limitations of forecasting and focus on risk management

Extreme evens and heteroskedasticity:”

Highlighted plot of training and test set split

I am writing the discussion section of my research. These were my research questions. “

investigate suitable approach for prediction when there is high volatility missing data and outliers due to extreme events

1. How does high volatility in stock price data impact the accuracy of traditional time series prediction models, and which models are most robust to high volatility, missing data, and outliers?
2. What are the most effective methods for handling missing data and outliers in the context of stock price prediction, considering high volatility and extreme events?
3. How do different preprocessing techniques, such as data imputation or outlier removal, affect the performance of prediction models in volatile stock price data?
4. How do various data imputation techniques, such as linear interpolation and rolling mean, impact the accuracy of stock price predictions when dealing with missing data?
5. In what ways can outlier detection techniques be integrated into the stock price prediction process to enhance the model's robustness against extreme events and improve overall predictive performance?
6. How can the relationships between the two unnamed price series be leveraged to improve stock price predictions, and what is the potential for cross-series prediction, such as using one price series to predict the other or exploiting common trends and patterns between the two series?

“

Read my notes in light of the guideline and research questions and critically evaluate what analysis is missing, discussion is required based on the points I have so far. Make suggestions for what to include to make the analysis more insightful and critical and rewrite these suggestions as dot points within my notes. These are my notes:

* Addressing heteroskedasticity and extreme events:
  + Future research could investigate methods to model and account for heteroskedasticity in the prediction process, potentially improving the performance of the models in the presence of extreme events.
  + The main observations from the residual analyses are:

None of the models have normally distributed residuals, as indicated by the Jarque-Bera test (p-value < 0.05 in all cases).

Autocorrelation in residuals is present in some of the models, as indicated by the Ljung-Box test (p-value < 0.05 in some cases).

Example residual plot to show best model still did not account for heteroskedasticity entirely

* Performance of different prediction methods in the presence of high volatility and extreme events.
  + Ensemble forest methods showed better performance in handling high volatility in price returns compared to simpler decision tree methods.
  + Some models might have overfitted the training data, especially when using imputed data, resulting in decreased predictive ability.
* Most effective methods for handling missing data, outliers, and stock price prediction challenges:
  + Removing missing values led to better performance in most cases.
  + Imputing missing data did not improve prediction results significantly.
  + Including outliers as a feature did not significantly improve results either.
* The impact of pre-processing techniques on prediction model performance:
  + Interpolation, while having the least effect on the distribution of returns, led to increased error in train and test sets, suggesting that the models might have overfitted the training data due to bias introduced by the imputation method.
  + Using unimputed data without outliers resulted in better generalization to the test set.
* Comparing data imputation techniques:
  + Artificially synthesizing data (e.g., through interpolation) may lead to reduced predictive ability and further propagation of bias in outlier features.
  + **Since forecasting for daily accuracy, information lost or bias introduced by imputation has a significant affect on trend prediction, but will be evident in forecasting**
* Leveraging relationships between the two price series for prediction:
  + Results indicate that using a given stock's history to predict itself is more accurate than using correlations between the two stocks for prediction.
* Benchmarks and limitations:
  + Most models had larger errors than the benchmark random walk, suggesting that the underlying return behaviours were not captured well by the models. Best performing cases were H1 and H4 which correspond to predictions in stock 1. Suggesting that stock 2 is likely to behave more closely to a random walk, as such the predictions by the models did not fit well
* Integrating outlier detection techniques to enhance model robustness:
  + Future research could explore other outlier detection methods and their integration into the stock price prediction process.
  + The potential for cross-series prediction could be further investigated.

Performance of different prediction methods in the presence of high volatility and extreme events:

* Ensemble forest methods, such as random forests and gradient-boosted trees, demonstrated better robustness and performance in handling high volatility, missing data, and outliers compared to simpler decision tree methods.
* Overfitting, particularly when using imputed data, led to decreased predictive ability in some models, highlighting the importance of proper model validation and regularization techniques.

Addressing heteroskedasticity and extreme events:

* Analysis of residuals indicate the imputation and outlier features did not allow models to better handle the returns volatility and show robustness for heteroskedasticity.

Most effective methods for handling missing data, outliers, and stock price prediction challenges:

* Removing missing values generally resulted in better performance, while imputing missing data did not significantly improve prediction results.
* Including outliers as a feature did not noticeably improve results, suggesting that other methods for handling outliers may be more effective in the context of stock price prediction with high volatility and extreme events.

The impact of pre-processing techniques on prediction model performance:

* Interpolation, despite having the least effect on the distribution of returns, led to increased error in train and test sets, possibly due to overfitting and bias introduced by the imputation method.
* Using unimputed data without outliers resulted in better generalization to the test set, emphasizing the importance of appropriate pre-processing techniques in improving model performance.

Leveraging relationships between the two price series for prediction:

* The analysis showed that using a given stock's history to predict itself is more accurate than using correlations between the two stocks for prediction, indicating that further research is needed to identify the potential benefits and limitations of cross-series prediction.

Benchmarks and limitations:

* Most models had larger errors than the benchmark random walk, suggesting that the underlying return behaviors were not captured well by the models. This highlights the need for further investigation into alternative models or feature engineering techniques that can better capture these behaviors.

Comparing data imputation techniques:

* Artificially synthesizing data, such as through interpolation, can reduce predictive ability and propagate bias in outlier features

Integrating outlier detection techniques to enhance model robustness:

* Future research could explore other outlier detection methods, such as statistical tests, clustering, or machine learning-based techniques, and their integration into the stock price prediction process to improve the model's robustness against extreme events and enhance overall predictive performance.
* The potential for cross-series prediction and the benefits and challenges associated with it warrant further investigation to determine the viability of leveraging relationships between the two price series for improved stock price predictions.
* RF showed lower error, but GBR demonstrates greater robustness to volatility
* GBR predictions appear lagged, but closer the true price returns volatility than RF
* Consider more robust alternative to RMSE

# Discussion

## Performance of Prediction Methods in High Volatility and Extreme Events

### Ensemble forest methods (e.g., random forests, gradient-boosted trees) outperform simpler decision tree methods in handling high volatility, missing data, and outliers.

### Overfitting with imputed data can decrease predictive ability, emphasizing the need for proper model validation and regularization techniques.

### Daily resampling proved to categorically provide better predictions on training and test set compared to hourly suggesting that more significant down sampling (24hrs vs 2 mins) may be a suitable approach for minimising the influence of noise in the context of single step predictions

## Addressing Heteroskedasticity and Extreme Events

### Residual analysis shows that imputation and outlier features did not improve model robustness for heteroskedasticity.

### Consider alternative approaches to address heteroskedasticity in future research.

## Handling Missing Data, Outliers, and Prediction Challenges

### Removing missing values typically yields better performance than imputing missing data.

### Including outliers as a feature does not noticeably improve results, suggesting that alternative outlier handling methods may be more effective.

## Impact of Pre-processing Techniques on Model Performance

### Interpolation can lead to increased error in train and test sets due to overfitting and bias.

### Using unimputed data without outliers results in better generalization to the test set.

# Discussion

## Leveraging Relationships Between Price Series for Prediction

### Predicting a stock using its own history is more accurate than using correlations between two stocks.

### Further research is needed to explore the potential benefits and limitations of cross-series prediction.

## Benchmarks and Limitations

### Most models exhibited larger errors than the benchmark random walk, indicating the need for further investigation into alternative models or feature engineering techniques.

### Consider visualizing model performance compared to the benchmark for better illustration.

## Comparing Data Imputation Techniques

### Artificially synthesizing data (e.g., interpolation) can reduce predictive ability and propagate bias in outlier features.

### Consider evaluating alternative imputation techniques, such as rolling mean.

## Integrating Outlier Detection Techniques for Enhanced Robustness

### Future research should explore other outlier detection methods (e.g., statistical tests, clustering, machine learning-based techniques) for integration into the stock price prediction process.

### Investigate the potential for cross-series prediction and its associated benefits and challenges to improve stock price predictions.

# Further Research and Limitations

## Slide 12: Conclusion and Future Research

### The Random Forest model demonstrated the best performance in terms of RMSE, but it is crucial to consider other factors such as model complexity, interpretability, and generalization ability.

### Model performance should be validated using additional test datasets or cross-validation techniques to ensure robustness and reliability.

### More testing is needed to identify models and features that can account for correlation, volatility, and provide good forecasting ability with robustness to extreme events.

### Investigate the potential of other models, feature engineering techniques, and evaluation metrics in the context of stock price prediction with high volatility and extreme events.

Investigate different scaling more suitable for returns distribution with wide tails – RobustScaler

Consider alternative metrics that penalise outliers less, such MASE, AIC.

* + c. Discuss the limitations of the study and potential avenues for future research.

Given the RMSE values, the Random Forest model seems to be the best performing model among the tested models. However, it is important to consider other factors such as the complexity of the model, the ease of interpretability, and the generalization ability of the model. It is also essential to validate the model performance using additional test datasets or cross-validation techniques to ensure the model's robustness and reliability.

Despite the model's performance, the residual analyses show that there is room for improvement. To achieve a better performing model, one could try different algorithms, feature engineering, or hyperparameter tuning. Additionally, the assumptions of linear regression, such as homoscedasticity and normal distribution of residuals, should be further investigated and addressed, if possible.

Conclusion that more testing is required to find models and features that can account for correlation and volatility with good forecasting ability and robustness to extreme events.

Investigate different scaling more suitable for returns distribution with wide tails – RobustScaler

Consider alternative metrics that penalise outliers less, such MASE, AIC.

Squared residuals of a fitted model to a stock price returns series: If there is heteroskedasticity in the squared residuals of a fitted model, it indicates that the model has not adequately captured the changing volatility in the returns series. This can lead to inefficient parameter estimates and incorrect inferences.

# Feature Engineering for Supervised Learning

## Data preparation:

### • Converted to a supervised learning problem using lagged values

### • Single-step forecast: target variable is the next time step

## Lagged features:

### • Hourly data: 4 lags

### • Daily data: 3 lags

### • Features added: price and Boolean indicator for outliers

## Datetime features:

### • Day of the week

### • Month

### • Year