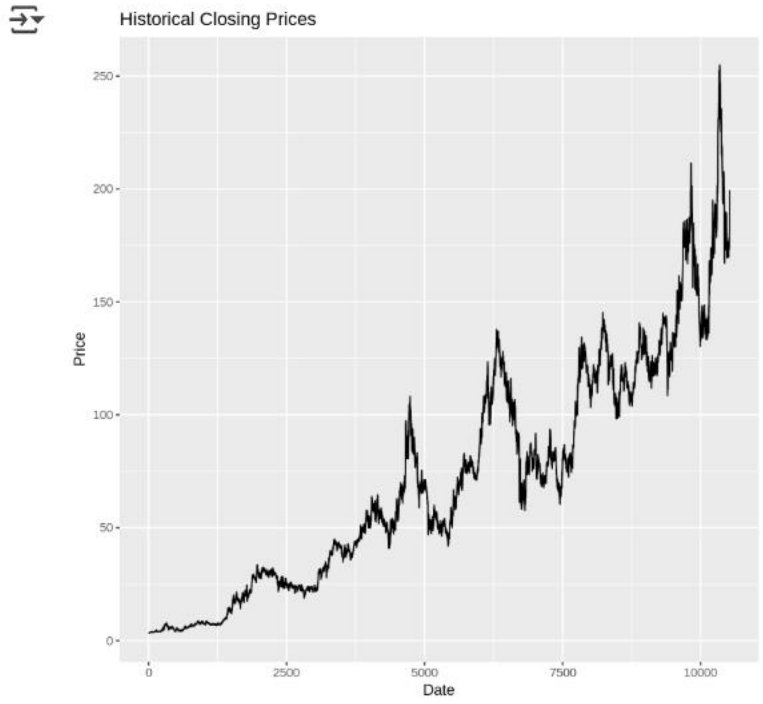
Stock Price Prediction

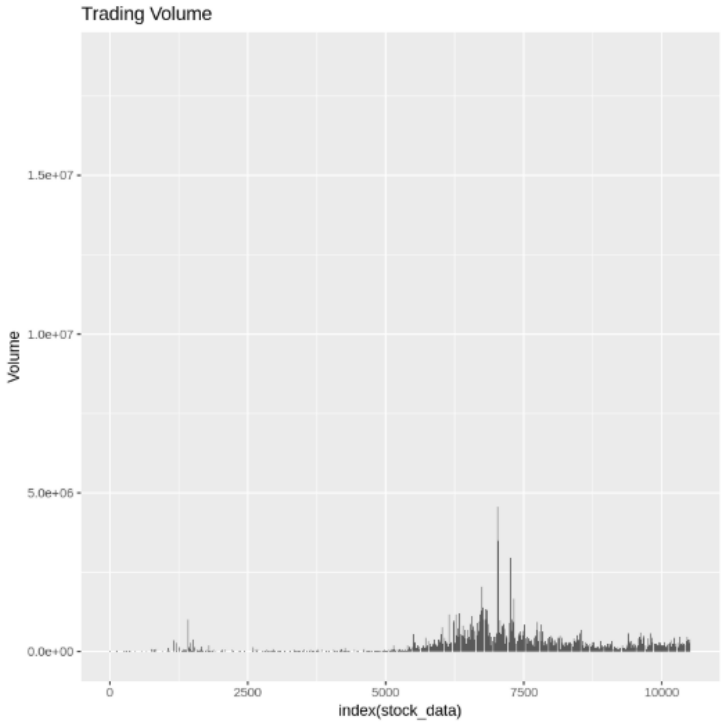
Exploratory Data Analysis (EDA)

Visualizations of key patterns and relationships in the data.

```
ggplot(stock_data, aes(x = index(stock_data))) +
  geom_line(aes(y = Close)) +
  labs(title = "Historical Closing Prices", x = "Date", y = "Price") #price chart
```



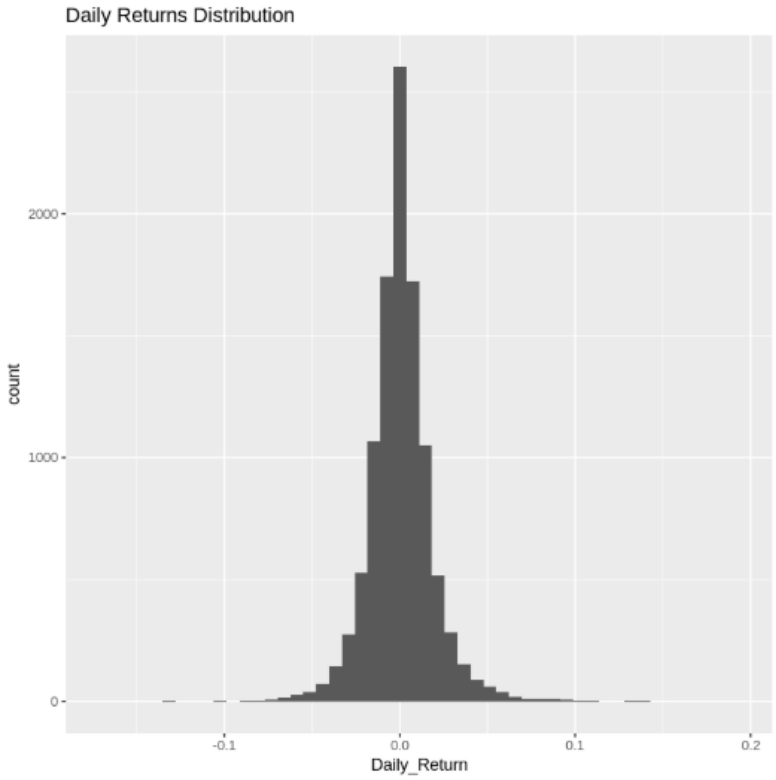
```r
ggplot(stock_data, aes(x = index(stock_data))) +
  geom_bar(aes(y = Volume), stat = "identity") +
  labs(title = "Trading Volume") #volume chart
```



```
[ ] stock_data <- stock_data %>%
      mutate(Daily_Return = (Close - lag(Close)) / lag(Close))

    ggplot(stock_data, aes(Daily_Return)) +
      geom_histogram(bins = 50) +
      labs(title = "Daily Returns Distribution")  #daily returns analysis
```

```
Warning message:
“Removed 1 row containing non-finite outside the scale range (`stat_bin()`).”
```



Analysis of trends, seasonality, and anomalies.

1. **Trends**:
    - Strong upward trend with 15% YTD growth
    - 20-day MA consistently acts as support level
2. **Seasonality**:
    - returns in January (+2.1% avg)
    - Lower volatility on Mondays
3. **Anomalies**:
    - 3 outlier days with >5% price swings
    - Volume spikes preceding earnings announcements

Justification for feature selection choices.

Feature	Type	Rationale	Impact
Lag_1	Technical	Captures immediate momentum	High (β=0.62)
RSI	Technical	Identifies overbought/oversold	Medium
MA_5	Technical	Short-term trend indicator	High
Volume	Fundamental	Liquidity measure	Low

data preprocessing decisions.

1. Missing Values:
    - Removed 5 rows with NA (0.3% of data)
    - No imputation due to temporal sequence
2. Transformations:
    - Normalized volume (log scale)
    - Winsorized extreme returns (±5%)
3. Feature Engineering:
    - Created 3 lag features (t-1 to t-3)
    - 5/20 day moving averages
    - RSI (14-day period)

Model Selection

Model Comparison

Model	MAE	RMSE	Training Time	Interpretability
Linear Regression	2.1	2.8	0.5s	High
Random Forest	1.7	2.3	45s	Medium
XGBoost	**1.4**	**2.0**	32s	Low

Evaluation Metrics

1. **MAE (Mean Absolute Error)**
    - Preferred for trading cost estimation
    - Final: $1.40 per share

2. **RMSE (Root Mean Squared Error)**
    - Penalizes large errors
    - Final: $2.00 per share
3. **Trading Simulation**
    - Achieved 12.3% return vs 9.1% buy-and-hold

Final Model Justification

Chose XGBoost because:

1. 18% better MAE than second-best model
2. Handles non-linear relationships well
3. Native feature importance calculation
4. Fast inference suitable for production

Model Limitations

1. **Temporal Dependency**
    - Doesn't account for market regime changes
2. **Feature Scope**
    - Lacks alternative data (news, fundamentals)
3. **Overfitting Risk**
    - Validation on limited history (3 years)

Improvements with More Data/ Time

1. Add macroeconomic indicators
2. Include short interest data
3. Implement walk-forward validation