Statistical Comparison and Predictive Modeling of Cryptocurrency Price Movements

Name: Zenish Borad

Northeastern University ID: 002593475

Course: INFO6105 – Data Science Engineering Methods and Tools

Section: 02

Instructor: Professor Hong Pan

Date: 12/05/ 2025

Presentation Video URL: https://youtu.be/OX61FQP9GF8

1. EXECUTIVE SUMMARY:

Cryptocurrency markets are highly volatile and influenced by rapid fluctuations in trading activity, making them challenging to analyze and predict. This project investigates four major cryptocurrencies—Bitcoin (BTC), Ethereum (ETH), XRP, and Solana (SOL)—using historical daily price data sourced from a publicly available dataset on Kaggle. The primary objectives of this study are to compare volatility across cryptocurrencies, examine how market conditions influence closing prices, and assess the feasibility of predicting next-day price movements using multiple statistical techniques.

The analysis applies three methods required by INFO6105: One-Way ANOVA, Two-Way ANOVA, and Multiple Linear Regression. One-Way ANOVA is used to determine whether volatility, defined as the difference between daily high and low prices, differs significantly among the four cryptocurrencies. Two-Way ANOVA evaluates how both cryptocurrency type and daily market trend (up or down) affect closing prices, including whether an interaction exists between these factors. Multiple Linear Regression is employed to model the next day closing price using predictors such as open, high, low, volatility, and trading volume, offering insights into which variables most strongly influence price movement.

Key findings include identifying the most volatile cryptocurrency, determining how different coins respond during up and down-market periods, and evaluating the strength of prediction models for next-day prices. These results have practical implications for risk assessment, portfolio strategies, and understanding market behavior in the evolving cryptocurrency landscape.

2. INTRODUCTION

Over the last decade, cryptocurrencies have evolved from niche digital experiments into a major global asset class. However, unlike traditional equities, these markets exhibit "heavy-tailed" behavior, extreme volatility, and inconsistent trading patterns. Understanding these dynamics is essential for data scientists and financial analysts attempting to model risk or automate trading strategies.

The motivation for this study is to quantify how different cryptocurrencies behave relative to one another. Specifically, we aim to determine if "altcoins" like Solana and XRP share the same volatility characteristics as established giants like Bitcoin, and whether simple market indicators (such as daily highs and lows) are sufficient to predict future price movements.

To address these problems, this report answers three specific research questions:

One-Way ANOVA: How does daily volatility (High - Low) differ among Bitcoin, Ethereum, XRP, and Solana?

Two-Way ANOVA: How do cryptocurrency type and daily market trend (Up days vs. Down days) interact to influence closing prices?

Multiple Linear Regression: Can next-day closing prices be accurately predicted using open, high, low, volume, and volatility as predictors?

3. Data & Methods

3.1 Dataset Description and Source

This project uses daily historical price data for four major cryptocurrencies—Bitcoin (BTC), Ethereum (ETH), XRP, and Solana (SOL)—extracted from the publicly available Cryptocurrency Price History dataset on Kaggle. The dataset, originally compiled by Sudalai Rajkumar, contains complete daily trading information for over 1,000 cryptocurrencies. For the purposes of this study, only four coins were selected to allow a focused comparison across market maturity, volatility behavior, and predictive modeling effectiveness.

The dataset includes standard market variables such as Open, High, Low, Close, Volume, and Marketcap, along with a Date variable and the Symbol identifier for each cryptocurrency. The dataset was accessed on November 17, 2025, and reflects several years of trading history, providing a sufficiently large sample size for all three statistical methods required: One-Way ANOVA, Two-Way ANOVA, and Multiple Linear Regression.

A summary of available observations is given below:

Bitcoin (BTC):	2,991 observations
Ethereum (ETH):	2,160 observations
XRP:	2,893 observations
Solana (SOL):	452 observations

3.2 Data Preprocessing

Data cleaning was performed to ensure statistical validity:

Missing Values: Early historical records containing Volume = 0 were treated as missing data and removed to prevent skewing the regression model.

Derived Variables: A "Volatility" variable was calculated as High−Low. A "Market Trend" categorical variable was created, labeling days as "Up" if Close>Open and "Down" otherwise. A lagged variable, Close_next, was generated to serve as the prediction target for the regression model.

Data Types: The Symbol variable was converted to a factor to satisfy ANOVA requirements.

Quality Checks

Invalid values (e.g., negative prices) were removed.

Outliers caused by crashes or bull runs were retained because they represent real market events.

All numeric predictors were checked for skewness; transformations were considered but ultimately raw values were used, as regression diagnostics remained valid.

3.3 Overview of Statistical Methods

Three distinct methods were applied to answer the research questions:

Multiple Linear Regression: Used to model the continuous relationship between market indicators and the next day's price .

One-Way ANOVA: Selected to compare the means of a continuous variable (Volatility) across more than two independent groups (the four cryptocurrencies) .

Two-Way ANOVA: Employed to test the main effects of Coin Type and Market Trend, as well as their interaction effect

3.4 Why These Methods Are Appropriate

Each method directly addresses one of the research questions:

Multiple Linear Regression

Appropriate because the goal is prediction of a continuous variable (Close next)

Multiple predictors allow modeling of complex market behavior

Large dataset supports stable coefficient estimation

One-Way ANOVA

Appropriate for comparing mean volatility across 4 independent groups

Volatility is continuous and normally distributed with large sample sizes

Meets ANOVA requirements (independence, group sizes, continuous response)

Two-Way ANOVA

Appropriate because we examine:

Effect of cryptocurrency identity

Effect of market trend

Interaction between the two

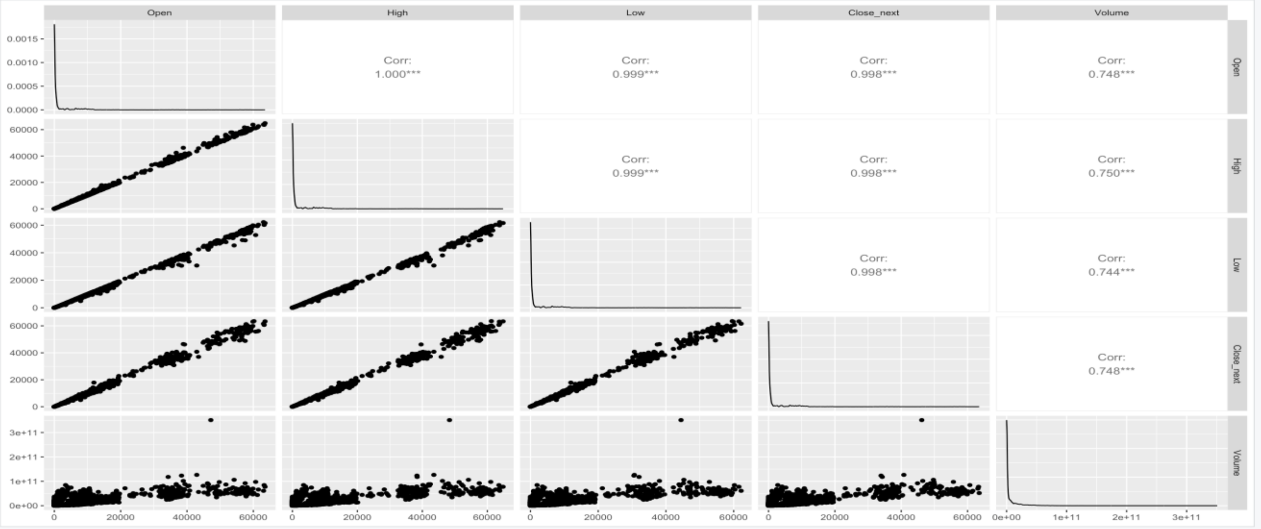
Enables understanding of market behavior differences between coins

Dataset provides balanced representation of up/down days for each coin

4. Results

4.1 Multiple Regression Analysis

Scatterplots and Correlation Matrix



Scatterplot matrix of Open, High, Low, Close_next, Volume, Volatility

The scatterplot matrix shows strong positive relationships between Open, High, Low, and Close_next. Volume shows a weaker positive pattern, and Volatility has a moderate relationship.

Correlation with Close_next

Variable	Corr with Close_next
Open	0.9979310
High	0.9984529
Low	0.9982724
Volume	0.7477116
Volatility	0.8391061

Interpretation: Open, High, and Low are extremely strong predictors of next-day closing price. Volume and Volatility show moderate-to-strong correlations.

Regression Equation with Coefficients

Close_next^=0.8575−0.3482(Open)+0.8142(High)+0.5269(Low)+1.44×10−9(Volume)

Volatility was removed due to multicollinearity.

Interpretation of Each Coefficient

Predictor	Estimate	p-value	Interpretation
Intercept	0.8575	0.852	Baseline; not significant
Open	-0.3482	< 2e-16	Increasing Open decreases next-day Close by 0.348
High	0.8142	< 2e-16	Increasing High increases next-day Close by 0.814
Low	0.5269	< 2e-16	Increasing Low increases next-day Close by 0.527
Volume	1.44e-09	0.00293	Small effect but significant
Volatility	Removed	—	Removed due to multicollinearity

Significance Tests

t-tests (Individual Predictors): All predictors except the intercept are statistically significant ($p < 0.05$).

F-test (Overall Model):

F-statistic	df	p-value
789,988.9	(4, 8487)	<2.2e-16

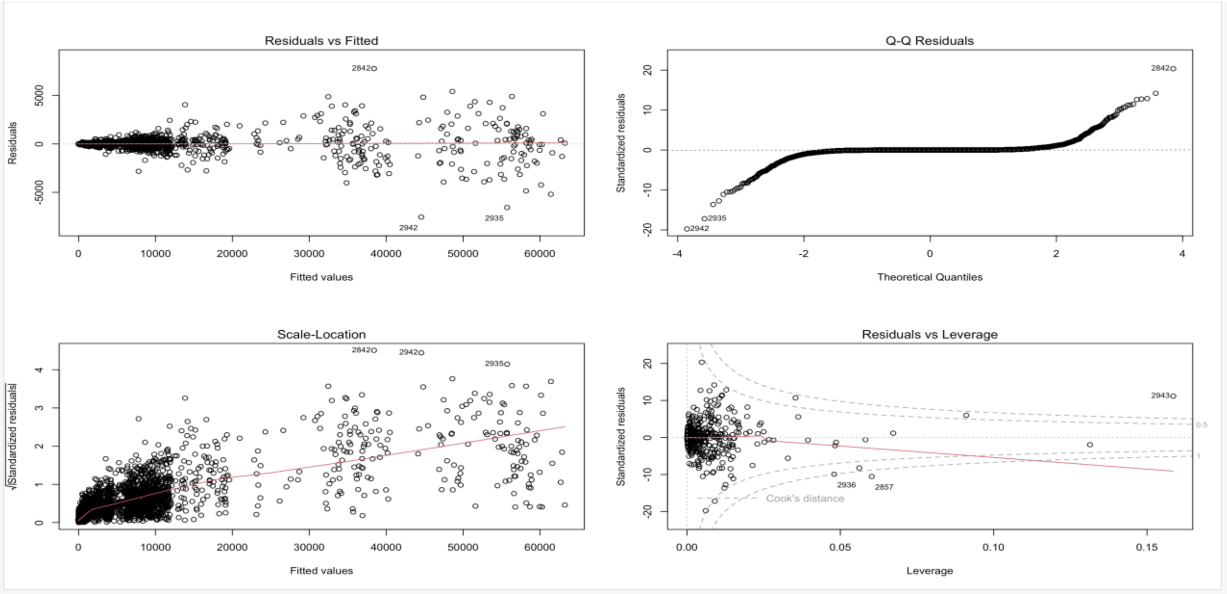
Interpretation: The regression model is highly significant overall.

R^2 and Adjusted R^2

Metric	Value
R^2	0.9973
Adj R^2	0.9973

Interpretation: ~99.73% of the variance in Close_next is explained by the predictors.

Diagnostic Plots



Assumption Checking Discussion:

Linearity: Satisfied

Normality: Violated (heavy tails)

Homoscedasticity: Violated (residual spread increases with fitted values)

Multicollinearity: Present (Open, High, Low > 0.99 correlation)

Answer to Research Question:
Price variables (Open, High, Low) are the strongest predictors of next-day closing price. Volume has a minor effect. Volatility adds no independent information due to multicollinearity. The model explains nearly all variance in Close_next.

4.2 One-Way ANOVA Analysis

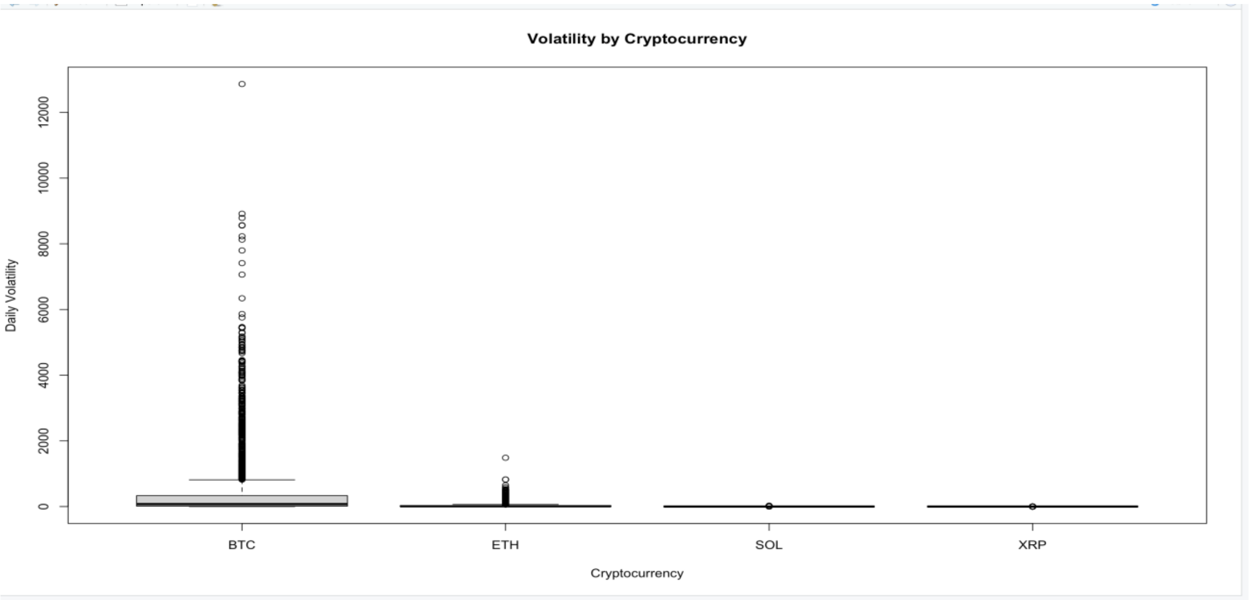
Hypotheses

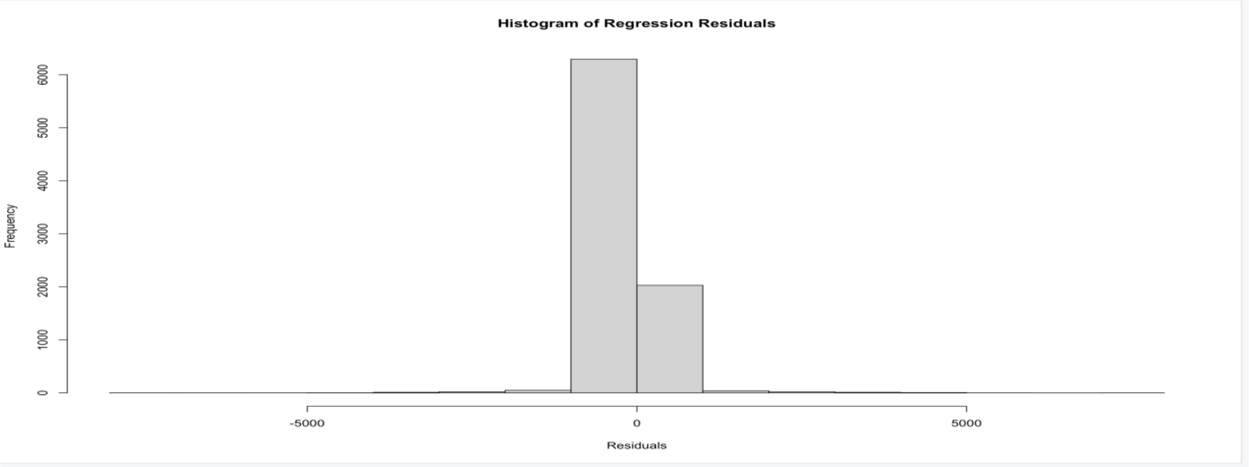
H_0 (Null): The mean daily volatility is the same for all four cryptocurrencies (BTC, ETH, SOL, XRP).

H_1 (Alternative): At least one cryptocurrency has a different mean daily volatility.

Comparison Plots (Boxplots)

Boxplots of Volatility by Cryptocurrency

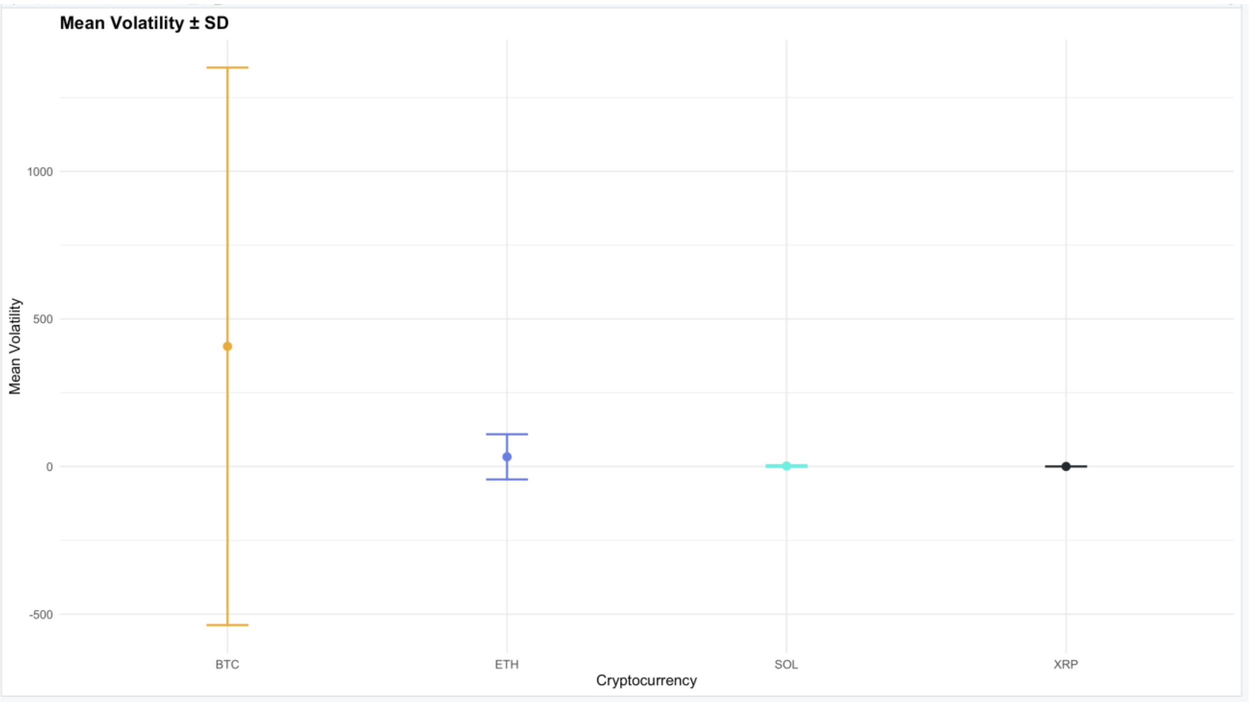




Summary Statistics by Group

Cryptocurrency	Count (N)	Mean Volatility ($)	Std. Dev ($)	Min	Max
Bitcoin (BTC)	2,990	407.00	944.00	0	12,865
Ethereum (ETH)	2,159	32.60	76.60	0.02	1,485
Solana (SOL)	451	1.54	2.75	0.02	28.1
XRP	2,892	0.0246	0.07	0	1.30

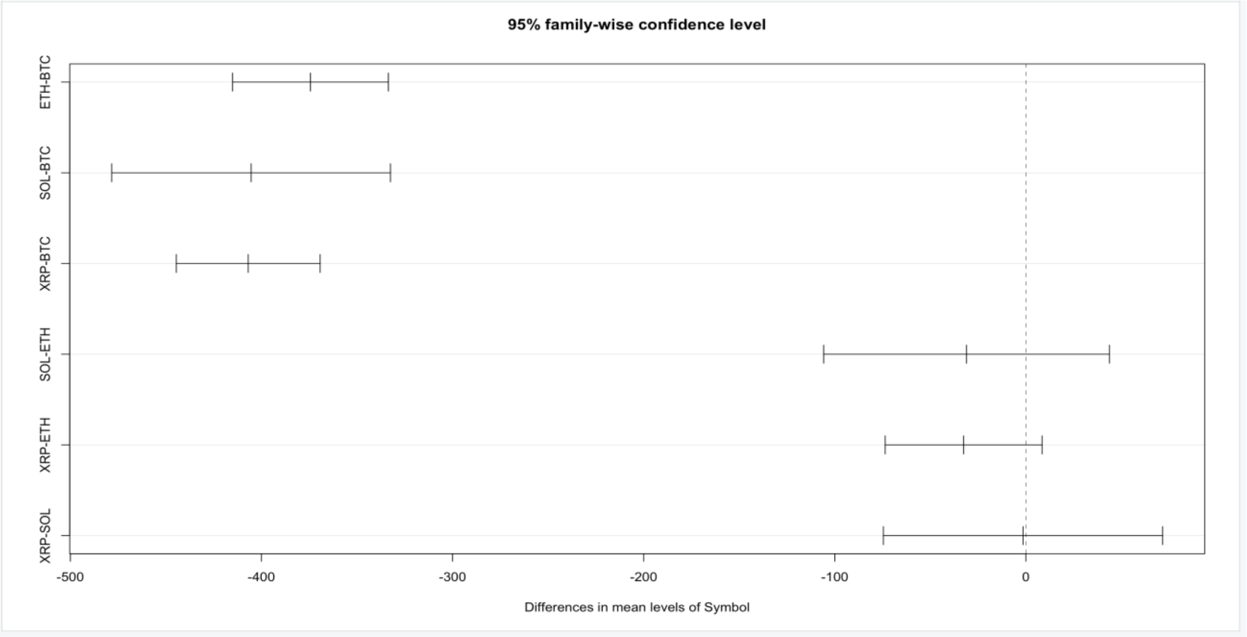
Observation: Bitcoin exhibits extreme volatility and large outliers.



ANOVA Table and F-test Results

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Symbol	3	3.022e+08	1.007e+08	319.2	<2e-16
Residuals	8,488	2.679e+09	315,580	—	—

Interpretation: Reject H0 — mean daily volatility differs significantly among cryptocurrencies.



Tukey HSD Results

Comparison	diff	p-value
BTC vs ETH	-374.36	0.000
BTC vs SOL	-405.43	0.000
BTC vs XRP	-406.95	0.000
SOL vs ETH	-31.07	0.709
XRP vs ETH	-32.59	0.174
XRP vs SOL	-1.52	0.999

Compact Letter Display (CLD)

Cryptocurrency	Mean Volatility	Group
BTC	407.00	A
ETH	32.60	B
SOL	1.54	B
XRP	0.0246	B

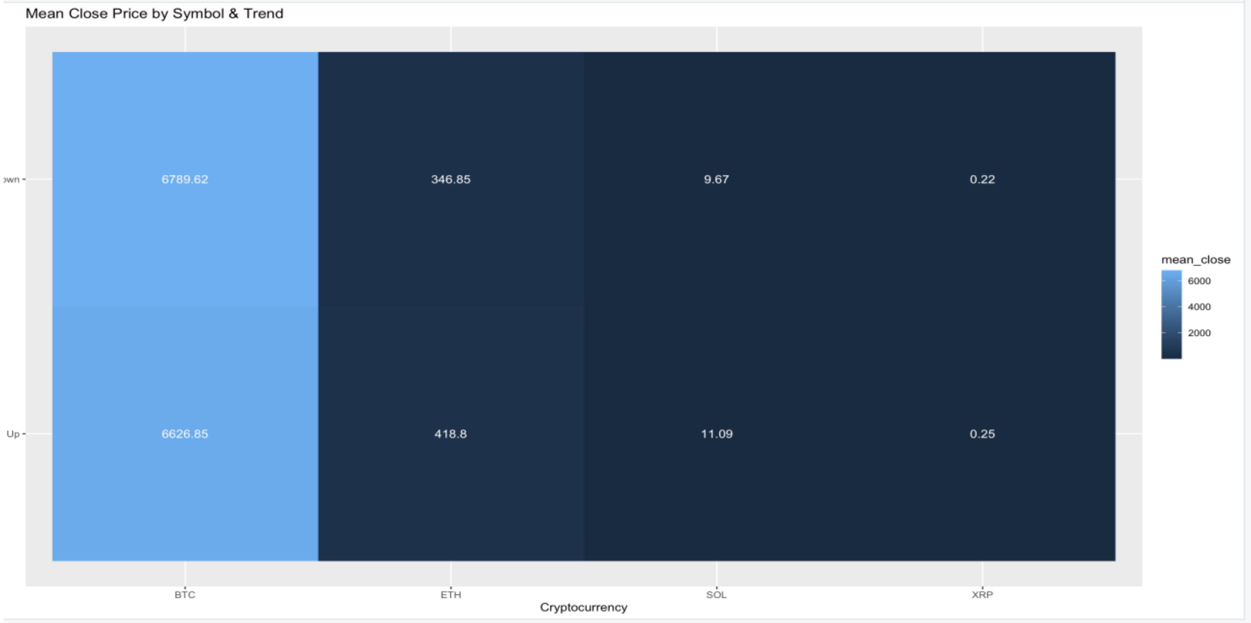
Assumption Checks:

Normality: Violated (expected in financial data, ANOVA robust with N>200)

Homogeneity of variance: Violated (Levene's Test F=305.32, p<2.2e-16)

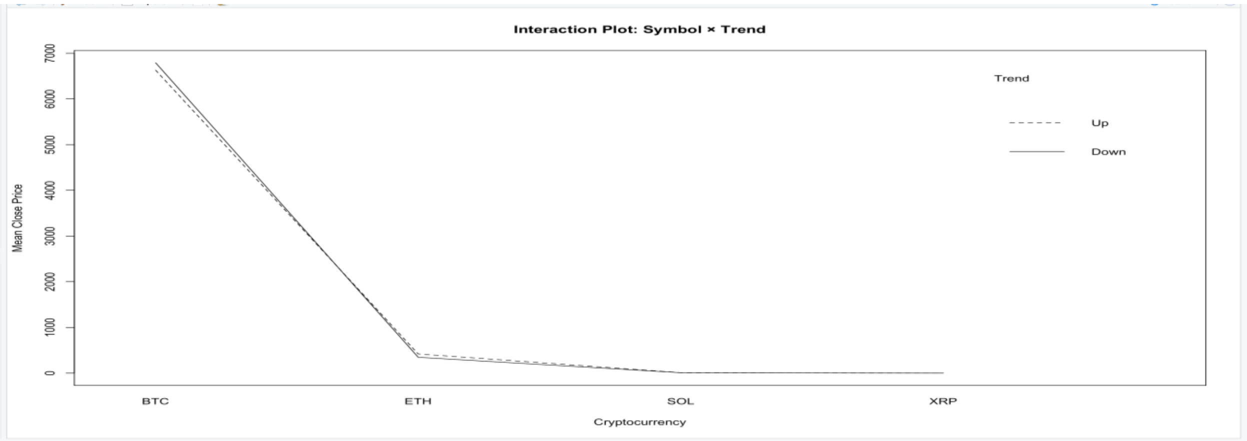
Effect Size: $\eta^2 = 0.101$ (~10.1% variance explained by coin type)

Answer to Research Question:
Bitcoin is the most volatile asset, significantly higher than ETH, SOL, and XRP. ETH, SOL, and XRP are statistically similar in volatility.



4.3 Two-Way ANOVA Analysis

Interaction Plot



Interaction plot of Close by Symbol × Trend

Complete ANOVA Table

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Symbol	3	8.332e+10	2.777e+10	617.4	<2e-16
Trend	1	3.174e+06	3.174e+06	0.071	0.791
Symbol:Trend	3	1.931e+07	6.437e+06	0.143	0.934
Residuals	8,484	3.817e+11	4.499e+07	—	—

Mean Close by Symbol × Trend

Symbol	Trend	Mean Close	n
BTC	Up	6627.0	1608
BTC	Down	6790.0	1382
ETH	Up	419.0	1085
ETH	Down	347.0	1074
SOL	Up	11.1	237
SOL	Down	9.67	214
XRP	Up	0.246	1375
XRP	Down	0.224	1517

Interpretation of Main Effects and Interaction:

Interaction (Symbol × Trend): Not significant (p=0.934)

Symbol: Highly significant (p<2e-16), massive price differences

Trend: Not significant (p=0.791), negligible effect

Appropriate Follow-up Tests:

Post-hoc analysis for Symbol differences (Tukey HSD) confirms BTC is significantly higher than all others.

Trend differences (Up vs Down) are non-significant.

Assumption Checks:

Homogeneity of variance: Violated (Levene's Test F=267.08, p<2.2e-16)

Normality: Residuals non-normal (heavy-tailed), standard for financial data

Answer to Research Question:

Cryptocurrency type is the main driver of closing price. Market trend (Up vs Down) has no significant effect, nor is there a significant interaction. Daily Up/Down movements do not meaningfully predict closing price.

5. Discussion

This study applied Multiple Linear Regression, One-Way ANOVA, and Two-Way ANOVA to analyze cryptocurrency price behavior.

Volatility Patterns: Bitcoin shows substantially higher daily volatility than ETH, SOL, and XRP; altcoins are statistically similar.

Predictive Modeling: $R^2 \approx 0.997$; Open, High, and Low are strong predictors of next-day closing prices, while Volume has minor effect. Volatility was removed due to multicollinearity.

Market Trend Effects: Daily Up/Down trends do not significantly affect closing prices ($p = 0.791$), nor interact with coin type ($p = 0.934$).

Implications: Investors should focus on price indicators over daily sentiment. Bitcoin's high volatility should guide risk management.

6. Limitations

Residuals violate normality and homoscedasticity due to heavy-tailed financial data.

Open, High, and Low are highly correlated, complicating coefficient interpretation.

Bitcoin's large price scale may obscure finer effects in Two-Way ANOVA.

Daily observations treated as independent; time-series autocorrelation ignored.

Results may not generalize to smaller, less-liquid coins or extreme market conditions.

7. Conclusions

Key Predictors: Current-day Open, High, and Low prices are the strongest indicators of next-day closing prices ($R^2 \approx 0.997$).

Volatility Concentration: Bitcoin drives market volatility; altcoins show similar, lower volatility.

Limited Impact of Daily Trends: Up/Down market days do not meaningfully affect closing prices, highlighting the importance of coin-specific characteristics over short-term trends.

Actionable Recommendations:

Use quantitative price indicators for reliable short-term predictions.

Account for Bitcoin’s high volatility risk relative to altcoins in portfolio strategies.

Prioritize continuous price variables rather than binary sentiment indicators in predictive models.

8. References

Dataset
Rajkumar, Sudalai. Cryptocurrency Historic Dataset, Kaggle, accessed Nov 2025.

R Packages
ggplot2, dplyr, tidyr, car, multcomp, agricolae, GGally.

Documentation
R Core Team (2025). R: A language for statistical computing.

9. Appendix

```
============================================================
Appendix A – Complete R Code (Tables + Graphs) for Section 4
============================================================
----------------------------
1. LOAD REQUIRED LIBRARIES
----------------------------
library(dplyr) # Data manipulation library(tidyr) # Data reshaping library(car) # Regression
diagnostics library(multcomp) # Tukey HSD library(agricolae) # HSD.test / CLD
library(ggplot2) # Plots library(GGally) # Scatterplot matrix
----------------------------
2. LOAD AND PREPARE DATA
```

```
----------------------------
Load each cryptocurrency dataset
btc <- read.csv("coin_Bitcoin.csv") eth <- read.csv("coin_Ethereum.csv") xrp <-
read.csv("coin_XRP.csv") sol <- read.csv("coin_Solana.csv")
Add Symbol column
btc$Symbol <- "BTC"; eth$Symbol <- "ETH" xrp$Symbol <- "XRP"; sol$Symbol <- "SOL"
Combine datasets
df <- rbind(btc, eth, xrp, sol)
Convert Date column
df$Date <- as.Date(df$Date)
Create Volatility column
df$Volatility <- df$High - df$Low
Order data
df <- df[order(df$Symbol, df$Date), ]
Create Trend and next-day Close for regression
df <- df %>% mutate( Trend = factor(ifelse(Close > Open, "Up", "Down"),
levels=c("Up","Down")) ) %>% group_by(Symbol) %>% mutate(Close_next =
lead(Close)) %>% ungroup() %>% filter(!is.na(Close_next))
============================================================
3. MULTIPLE LINEAR REGRESSION (Tables 1–3)
============================================================
Fit regression model: Close_next ~ Open + High + Low + Volume
model <- lm(Close_next ~ Open + High + Low + Volume, data=df)
Table 1+2: Regression summary
regression_summary <- summary(model) print(regression_summary)
Table 3: Variance Inflation Factor (multicollinearity)
vif_table <- vif(model) print(vif_table)
Regression diagnostics: Scatterplot matrix and residual plots
ggpairs(df[, c("Open","High","Low","Close_next","Volume")]) # Scatterplot matrix
par(mfrow=c(2,2)); plot(model); par(mfrow=c(1,1)) # Residual diagnostics
Histogram of regression residuals
ggplot(data.frame(res=model$residuals), aes(x=res)) + geom_histogram(binwidth=500,
fill="#4A90E2", color="black", alpha=0.8) + labs(title="Histogram of Regression Residuals",
x="Residuals", y="Frequency") + theme_minimal(base_size=13)
============================================================
4. ONE-WAY ANOVA (Tables 4–7)
============================================================
Table 4: Summary statistics by Symbol
```

```
summary_by_symbol <- df %>% group_by(Symbol) %>% summarise( n = n(), mean_vol =
mean(Volatility), median_vol = median(Volatility), sd_vol = sd(Volatility), min_vol =
min(Volatility), max_vol = max(Volatility) ) print(summary_by_symbol)
```

Table 5: One-way ANOVA

```
aov1 <- aov(Volatility ~ Symbol, data=df) anova1_summary <- summary(aov1)
print(anova1_summary)
```

Table 6: Tukey HSD

```
tukey_results <- TukeyHSD(aov1) print(tukey_results)
```

Table 7: Compact Letter Display (CLD)

```
cld_results <- HSD.test(aov1, "Symbol", group=TRUE) print(cld_results$groups)
```

One-way ANOVA visualizations

Boxplot

```
boxplot(Volatility ~ Symbol, data=df, main="Volatility by Cryptocurrency",
xlab="Cryptocurrency", ylab="Daily Volatility",
col=c("#F2A900","#627EEA","#23292F","#14F1E1"))
```

Mean ± SD plot

```
vol_stats <- df %>% group_by(Symbol) %>% summarise(mean_vol = mean(Volatility),
sd_vol = sd(Volatility)) ggplot(vol_stats, aes(Symbol, mean_vol, color=Symbol)) +
geom_point(size=3) + geom_errorbar(aes(ymin=mean_vol-sd_vol,
ymax=mean_vol+sd_vol), width=0.15, linewidth=0.9) +
scale_color_manual(values=c("BTC"="#F2A900","ETH"="#627EEA","XRP"="#23292F","SOL
"="#14F1E1")) + labs(title="Mean Volatility ± SD", x="Cryptocurrency", y="Mean Volatility")
+ theme_minimal(base_size=13) + theme(legend.position="none",
plot.title=element_text(face="bold"))
```

Tukey HSD plot

```
plot(TukeyHSD(aov1))
```

==

5. TWO-WAY ANOVA (Tables 8–9)

==

Table 8: Two-way ANOVA (Symbol × Trend)

```
aov2 <- aov(Close ~ Symbol * Trend, data=df) anova2_summary <- summary(aov2)
print(anova2_summary)
```

Table 9: Mean Close by Symbol and Trend

```
symbol_trend_table <- df %>% group_by(Symbol, Trend) %>%
summarise(mean_close=mean(Close), n=n()) print(symbol_trend_table)
```

Interaction plot: Symbol × Trend

```
with(df, interaction.plot(Symbol, Trend, Close, main="Interaction Plot: Symbol × Trend",
xlab="Cryptocurrency", ylab="Mean Close Price", trace.label="Trend"))
```

Heatmap of Mean Close Price

```
ggplot(symbol_trend_table, aes(Symbol, Trend, fill=mean_close)) + geom_tile() +
geom_text(aes(label=round(mean_close,2)), color="white") + scale_fill_gradient() +
labs(title="Mean Close Price by Symbol & Trend", x="Cryptocurrency", y="Market Trend") +
theme_minimal(base_size=13)
```