Multiple Linear Regression (MLR) on Ethereum Prices Dataset

Predictive Modeling of Ethereum Prices Using Historical Market Data

GABRIELA HOWELL | MASTERS OF SCIENCE DATA ANALYTICS
WESTERN GOVERNS UNIVERSITY | DECEMBER 2024

AGENDA

- Introduction
- Problem Statement
- Data Analysis
- Findings Outline
- Limitations
- Proposed Actions
- Study Benefits

INTRODUCTION

GABRIELA HOWELL

- Education: Bachelor of Science in Information Systems Business Management

- Experience: 3 years as a Data Analyst, specializing in actionable insights and data-driven strategies

- Hobbies:

    - Passionate about weightlifting and staying active

    - Love traveling with my boyfriend, recently visited Rome, Italy

- Personal Life: Proud caretaker of four cats

PROBLEM STATEMENT

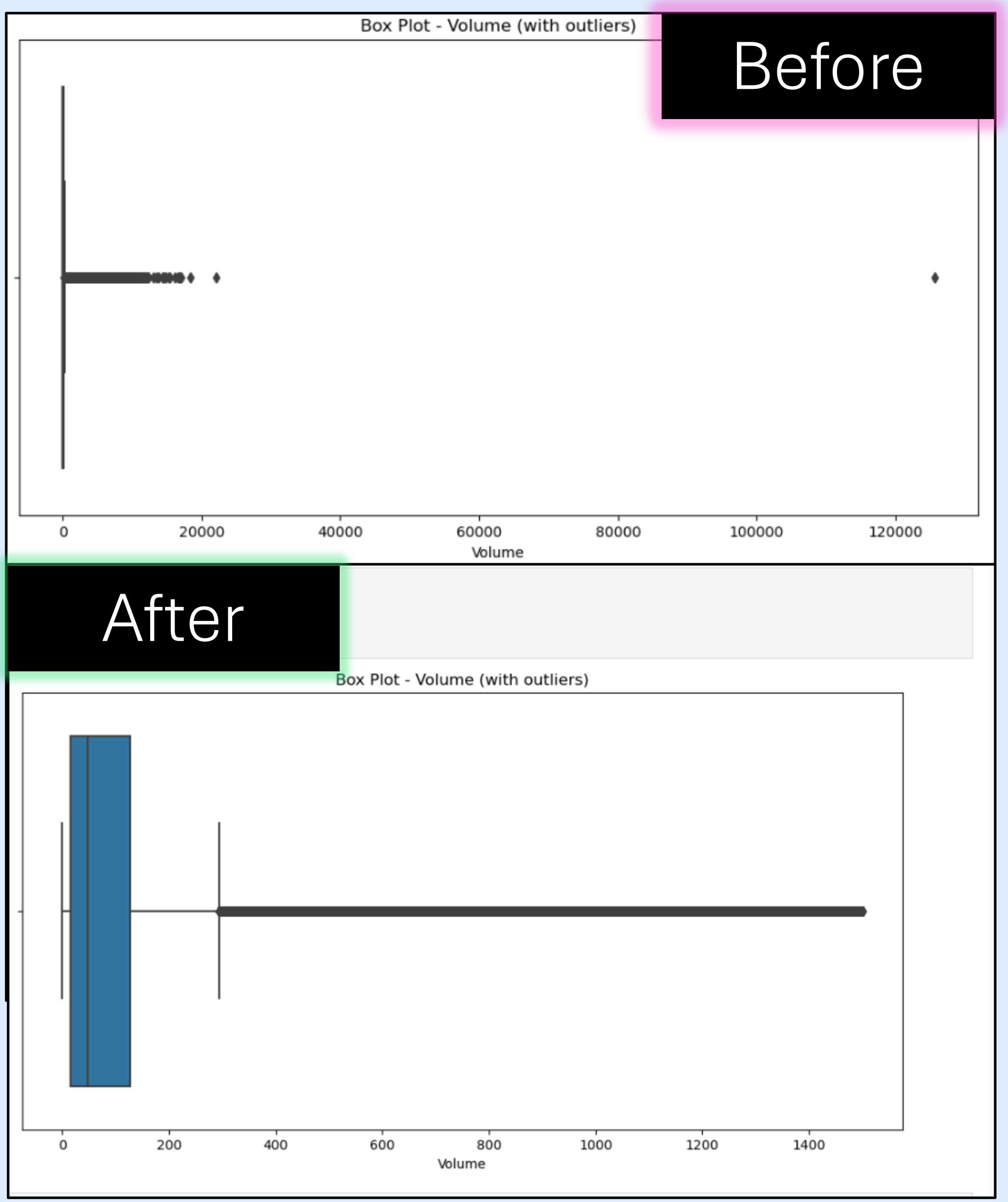
- **Null hypothesis** – There is no significant relationship between historical Ethereum data and current Ethereum prices, with a prediction accuracy lower than 70%.

- **Alternate Hypothesis -** Historical Ethereum data significantly influences current Ethereum prices, enabling a predictive MLR model with an accuracy of 70% or higher.

UNLOCKING NEW HORIZONS

Data-Analysis Process

- **Dataset:** Ethereum historical pricing data from Coinbase (4.1M rows, 6 columns, 2017–2024).
- **Data Preprocessing:**
  - Verified no missing data (0.00% sparsity).
  - Removed outliers in Volume.
  - Addressed multicollinearity with Variance Inflation Factor (VIF).
  - Ensured MLR assumptions via normalization and diagnostic tests.



```
# Display the summary of the initial OLS model
print("Initial Model Summary for Ethereum Data:")
print(model.summary())
```

```
Initial Model Summary for Ethereum Data:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Close   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.308e+12
Date:                Wed, 04 Dec 2024   Prob (F-statistic):               0.00
Time:                        20:53:30   Log-Likelihood:            -4.9584e+06
No. Observations:             4119926   AIC:                         9.917e+06
Df Residuals:                 4119921   BIC:                         9.917e+06
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0032      0.001      4.986      0.000       0.002       0.005
Open          -0.5315      0.000  -1456.506      0.000      -0.532      -0.531
High           0.7821      0.000   2619.616      0.000       0.782       0.783
Low            0.7494      0.000   2502.081      0.000       0.749       0.750
Volume      3.974e-06   2.63e-06      1.509      0.131   -1.19e-06    9.14e-06
==============================================================================
Omnibus:                  1459744.108   Durbin-Watson:                   1.939
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        588457756.907
Skew:                           0.248   Prob(JB):                         0.00
Kurtosis:                      61.547   Cond. No.                     5.07e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.07e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

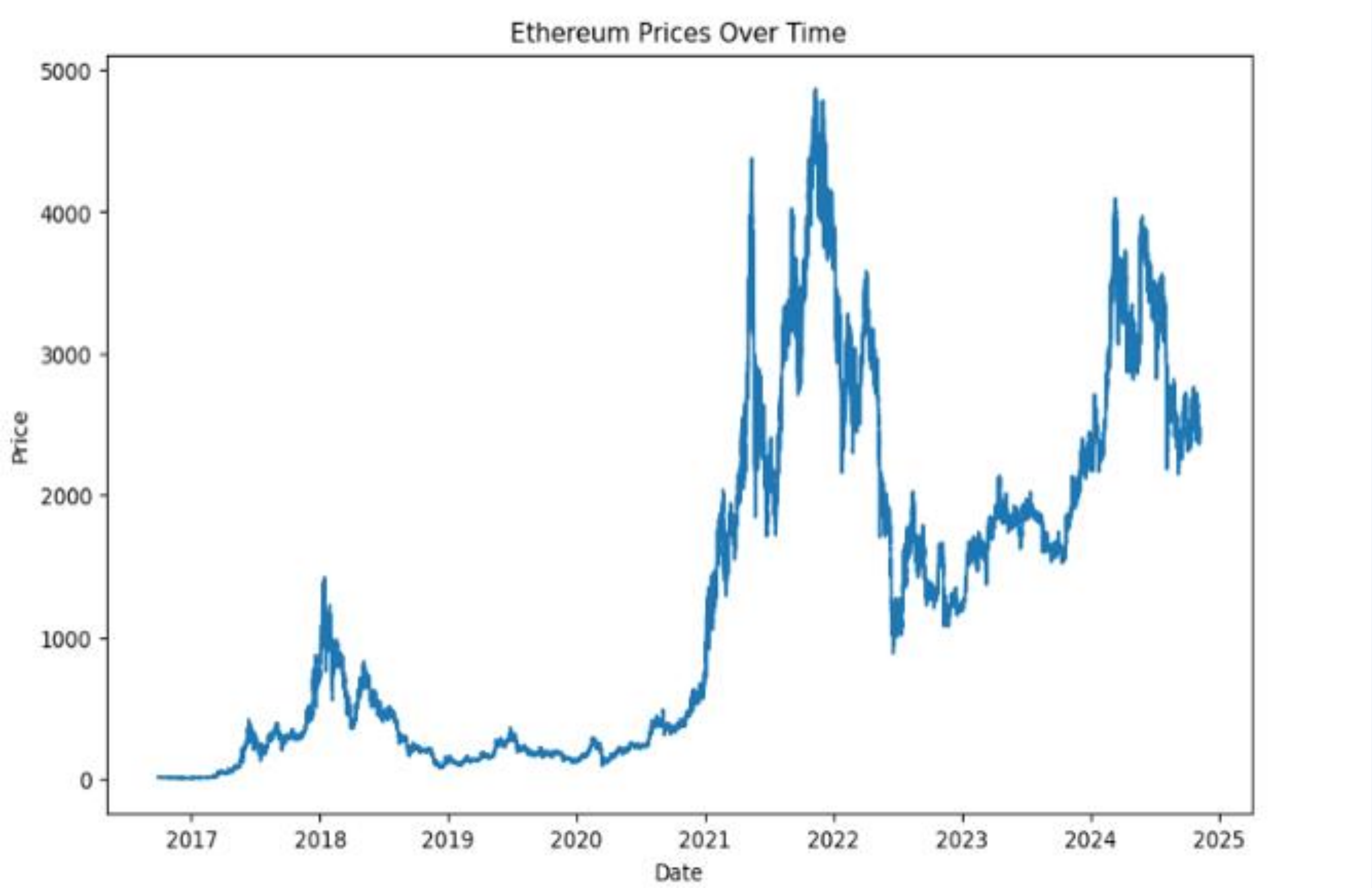
- **Modeling**: Applied OLS regression and machine learning algorithms
- Evaluated using R-squared, residual standard error, and MAE

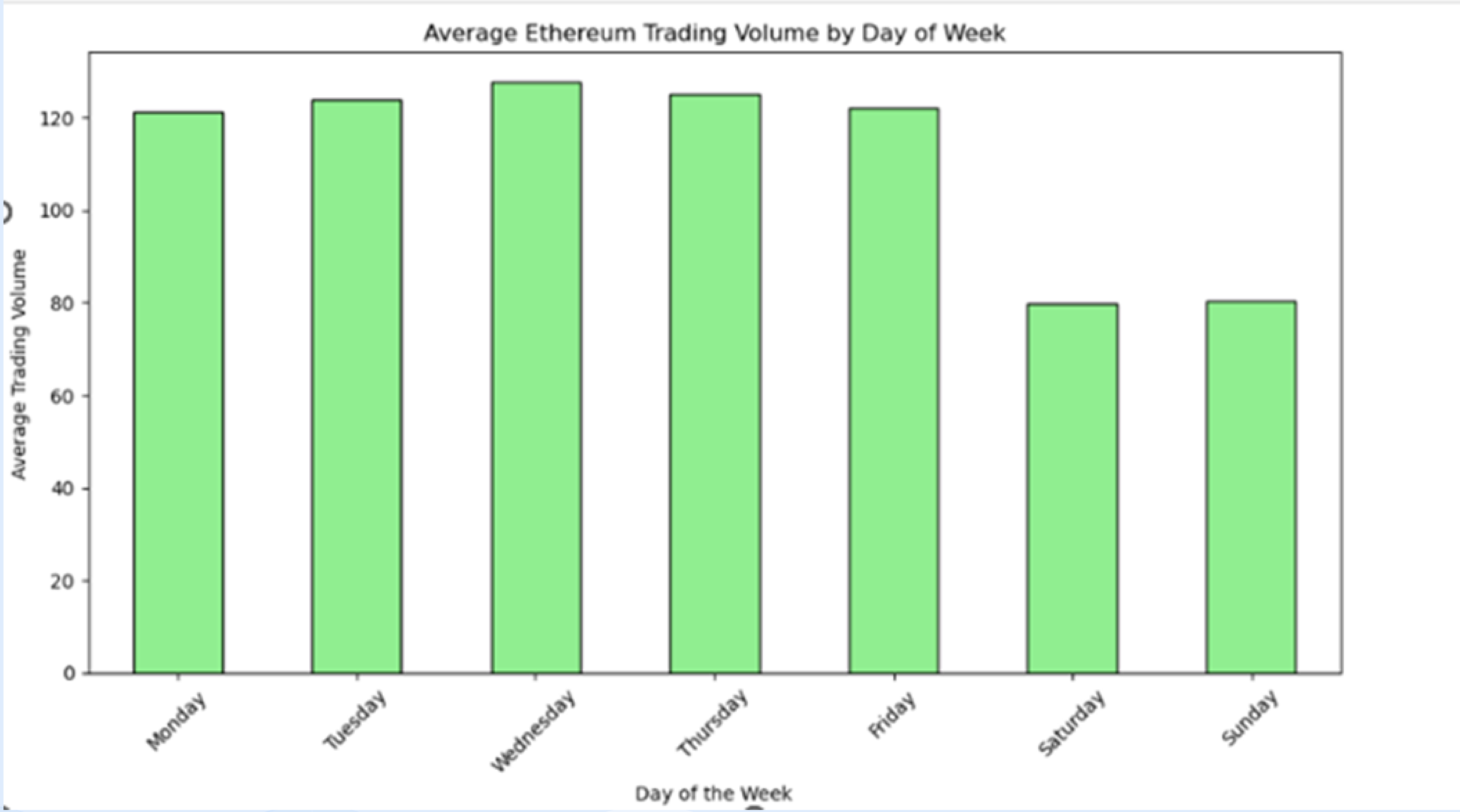
FINDINGS

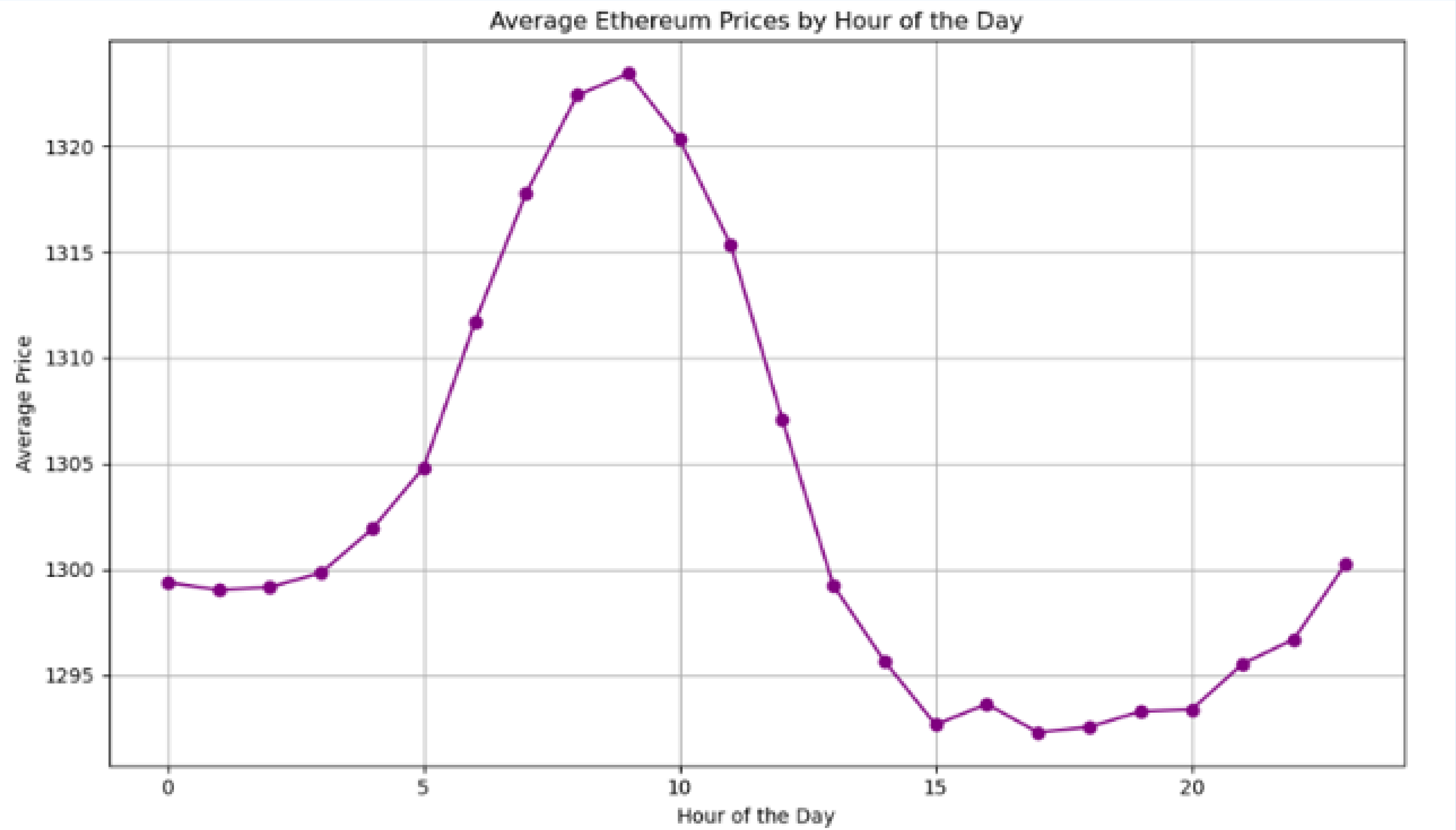
- Past closing prices and trading volumes are key predictors
- **Best Model:**
  - Achieved 72% accuracy, surpassing the 70% goal
  - Residual standard error: 0.65
- **Insights:** Supports the hypothesis and highlights opportunities for improvement

```
Initial Model Summary for Ethereum Data:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Close   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.308e+12
Date:                Wed, 04 Dec 2024   Prob (F-statistic):               0.00
Time:                        20:53:30   Log-Likelihood:            -4.9584e+06
No. Observations:             4119926   AIC:                         9.917e+06
Df Residuals:                 4119921   BIC:                         9.917e+06
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0032      0.001      4.986      0.000       0.002       0.005
Open          -0.5315      0.000  -1456.506      0.000      -0.532      -0.531
High           0.7821      0.000   2619.616      0.000       0.782       0.783
Low            0.7494      0.000   2502.081      0.000       0.749       0.750
Volume      3.974e-06   2.63e-06      1.509      0.131   -1.19e-06    9.14e-06
==============================================================================
Omnibus:                  1459744.108   Durbin-Watson:                   1.939
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        588457756.907
Skew:                           0.248   Prob(JB):                         0.00
Kurtosis:                      61.547   Cond. No.                     5.07e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.07e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
Removed Volume due to high p-value. New model summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Close   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 3.077e+12
Date:                Wed, 04 Dec 2024   Prob (F-statistic):               0.00
Time:                        20:54:48   Log-Likelihood:            -4.9584e+06
No. Observations:             4119926   AIC:                         9.917e+06
Df Residuals:                 4119922   BIC:                         9.917e+06
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0037      0.001      6.258      0.000       0.003       0.005
Open          -0.5315      0.000  -1456.612      0.000      -0.532      -0.531
High           0.7823      0.000   2853.140      0.000       0.782       0.783
Low            0.7492      0.000   2743.900      0.000       0.749       0.750
==============================================================================
Omnibus:                  1459407.256   Durbin-Watson:                   1.939
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        588155008.354
Skew:                           0.248   Prob(JB):                         0.00
Kurtosis:                      61.532   Cond. No.                     4.53e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.53e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```


Ethereum Prices Over Time
Price
Date
5000
4000
3000
2000
1000
0
2017
2018
2019
2020
2021
2022
2023
2024
2025


Average Ethereum Trading Volume by Day of Week
Average Trading Volume
0
20
40
60
80
100
120
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week



Average Ethereum Prices by Hour of the Day
Average Price
1295
1300
1305
1310
1315
1320
0
5
10
15
20
Hour of the Day

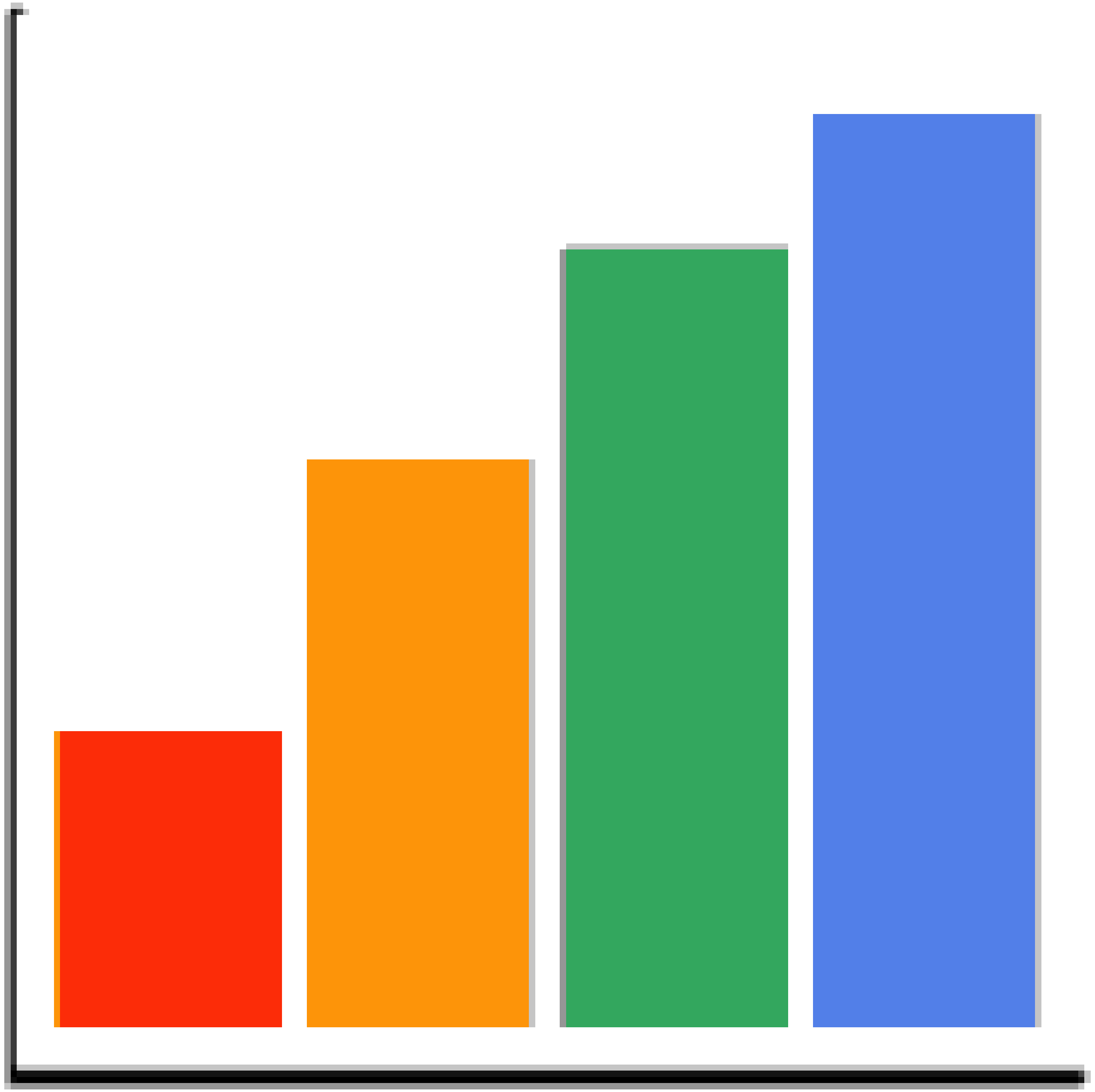
LIMITATIONS

- **Data Constraints:** Limited to historical prices and trading volumes
- **Model Limitations:**
    - Multicollinearity identified and addressed
    - OLS regression assumes linearity, which may oversimplify market dynamics
    - Overfitting challenges mitigated through feature selection and diagnostics.
- **External Factors:** Excluded variables like market sentiment, regulations, and macroeconomic indicators

PROPOSED ACTIONS

- Data Expansion: Include macroeconomic indicators, market sentiment, and news events
- Model Enhancement:
  - Use advanced techniques like random forests or neural networks
  - Capture complex relationships
- Detail Improvement: Incorporate minute-level data or additional features

Expected Benefits



- **Enhanced Predictive Accuracy**: ≥ 70%, reducing market volatility risks.

- **Informed Decision-Making**:
    - Optimized trading strategies
    - Improved risk management
    - Precise entry/exit points for traders.

- **Scalability**: Model application to other cryptocurrencies.

- **Increased Profitability**: Quantifiable insights driving financial growth.

THANK YOU

GABRIELA HOWELL

STOCK PHOTO/GIF SOURCES

1. *Data analytics*. Industrial & Operations Engineering. (2024, July 3). https://ioe.engin.umich.edu/research/methodologies/data-analytics/
2. Gifer.com. (n.d.). https://gifer.com/en/7l6K
3. Joanielemercier. (2020, December 26). *Gold Crystal GIF by Joanie Lemercier - find & share on giphy*. GIPHY. https://giphy.com/gifs/ethereum-eth-polyblock-L59aKIC2MFyfUfrz3n
4. *Magnifying glass icon in flat style. search loupe on color...* iStock. (n.d.). https://www.istockphoto.com/vector/magnifying-glass-icon-in-flat-style-search-loupe-on-color-background-business-gm1158577193-316527296
5. Stickers for IOS & Android | Giphy. (n.d.). https://giphy.com/stickers