**Introduction to Dataset:**

**Extracting the data:**

**A screenshot of a computer code

Description automatically generated**

**Creating the database:**

**A screen shot of a computer code

Description automatically generated**

**Save data to database:**

**A screenshot of a computer code

Description automatically generated**

**Scrape the data:**

**A screenshot of a computer code

Description automatically generated**

1. **Descriptive analysis**

A screenshot of a computer program

Description automatically generated

**Result:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## **Summary for Descriptive analysis:**

Mean Values: These indicate the average measurements across your dataset, helpful for understanding typical values.

Standard Deviation: This highlights the variability in your data, showing how much the values deviate from the average, which is crucial for assessing volatility.

Correlation Coefficients (r values): These reveal the strength and direction of relationships between pairs of variables, essential for identifying trends or dependencies.

Overall Distribution: Through minimum, maximum, and quartile values, you gain insights into the data distribution, which can assist in spotting outliers or abnormal values.

This analysis provides a foundation for deeper financial analysis, risk assessment, and investment decision-making.

1. **Regression Analysis:**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Results:**

A screenshot of a computer screen

Description automatically generated

**OLS Regression Results Overview**

* **Model Specifications**:
  + **Dependent Variable**: **price\_intraday**
  + **Model Type**: Ordinary Least Squares (OLS)
* **Fit Statistics**:
  + **R-squared**: 1.000 - Model explains all the variability in the dependent variable.
  + **Adjusted R-squared**: 1.000 - Adjusted for the number of predictors, still perfect.
  + **F-statistic**: 546,200 - Indicates a very strong overall model fit.
  + **Prob (F-statistic)**: Approximately 0.00 - Statistical test suggests the model fit is highly significant.
* **Coefficient Estimates**:
  + **Constant**: 0.4440 (p < 0.001) - Suggests a significant intercept term.
  + **50\_day\_average**: 0.8910 (p < 0.001) - Strong positive relationship with the intraday price.
  + **200\_day\_average**: 0.0936 (p < 0.001) - Positive but weaker relationship compared to the 50-day average.
* **Additional Notes**:
  + The covariance matrix of the errors is correctly specified.
  + High R² and significant coefficients suggest an excellent fit, but the high kurtosis and JB statistic warn of potential issues with normality and outlier influence.

**Summary for Regression analysis:**

Model Evaluation: The performance of the model is assessed on the test set, resulting in a Mean Squared Error (MSE) of 8.700650144491261 and a high R² score of 0.9947610906067818, indicating the model explains 99.47% of the variance in the target variable, which suggests a strong fit between the model and the data.

This analysis effectively captures the relationship between the moving averages and the

intraday price of ETFs, showing the model's robustness and predictive power.

1. **Visualization:**

A screenshot of a computer program

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

**Result:**

A graph of different types of data

Description automatically generated with medium confidence

A graph with a red line

Description automatically generated

A graph with a red line

Description automatically generated

**Summary for visualization:**

Loads Data: It connects to your SQLite database, retrieves the data, and ensures that relevant columns are numeric.

Data Cleaning: Converts string data to numeric where necessary and handles missing values by dropping rows with NaNs.

Histograms: Visualizes the distribution of 'price\_(intraday)', '50\_day\_average', and '200\_day\_average' to understand their individual characteristics.

Scatter Plot with Regression Line: Analyzes the relationship between '50\_day\_average' and 'price\_(intraday)'.

Correlation Heatmap: Provides a heatmap of correlations between the numerical variables, offering insights into their relationships.

This approach gives a thorough visual exploration of your ETF data, highlighting key patterns and correlations which can inform further analysis or investment decisions.

1. **Cluster Analysis**

**A computer screen shot of a code

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

**Results: A graph with colored dots

Description automatically generated**

**A graph of a graph

Description automatically generated with medium confidence**

**A screenshot of a computer code

Description automatically generated**

**3D cluster visualization of ETF Data:**

**A computer screen shot of a code

Description automatically generated**

**A screen shot of a computer code

Description automatically generated**

**Result:**

**A graph of data with numbers and lines

Description automatically generated with medium confidence**