# Web Scraping and Data

SEPTEMBER 10

**Name: Phalit Gupta**

**Link:** https://github.com/Phalit045040/045040_DEV_Project_1.git

FORE School of Management
New Delhi

## Objectives:

This managerial report provides an overview of a Python script designed to scrape data from a website and analyze commodity prices. The purpose of this project is to automate the extraction and analysis of commodity price data from the website "centralcharts.com." The goal is to obtain a comprehensive list of commodities, including their current prices, changes in prices, opening prices, and trading volumes. Additionally, we aim to identify the commodities with the highest trading volumes and did statistical analysis (like mean, median, mode, outlier etc.)

## Methodology:

### Data Collection:

- The script utilizes the Python libraries requests, BeautifulSoup, and pandas to scrape data from the website.
- A list of URLs for each page of commodity data is generated based on the known page structure (pages 1 to 154).
- Each page is requested, and the HTML content is parsed to identify relevant data.

### Data Extraction:

- The script identifies the table containing commodity information on each page.
- Headers are extracted from the table and cleaned, removing unwanted columns.
- Data from each row of the table is extracted, cleaned, and added to a DataFrame.

### Data Cleaning and Transformation:

- Data is cleaned by removing special characters (',', '-', '%') and converting numeric columns to appropriate data types (e.g., float).
- The DataFrame is sorted in descending order based on trading volume.

### Result Presentation:

- The final dataset is presented as a formatted table using the tabulate library.

## Analysis: Basic Descriptive & Mathematical or Statistical Analysis

### 1. Summary Statistics and Insights for Commodity Prices

**Interpretation:**

- Number of Unique Commodities: [24]
  This dataset contains data for [24] unique commodities.

- Mean Price: [416.57]
  The mean (average) price of commodities in the dataset is [416.57]. This value represents the central tendency of commodity prices.

- Median Price: [89.08]
  The median price of commodities in the dataset is [89.08]. The median is a robust measure of central tendency, indicating the middle value when the prices are sorted.

- Number of Commodities with Positive Price Changes: [3696]
  There are [3696] commodities in the dataset that have experienced positive price changes. This indicates the number of commodities with price increases.

- Number of Commodities with Negative Price Changes: [0]
  There are [0] commodities in the dataset that have experienced negative price changes. This represents commodities with price decreases.

**Insights:**

- The dataset covers a diverse range of commodities, with [24] unique commodities being analyzed.

- The mean commodity price is approximately [416.57], suggesting the average price level across all commodities.

- The median commodity price, at [89.08], reflects the middle value in the dataset, which may differ from the mean if there are extreme values.

- [3696] commodities in the dataset have experienced positive price changes, indicating price increases.

- Zero commodities have recorded negative price changes, indicating price decreases.

```
+------------------------+-------------------+
|       Statistic        |       Value       |
+------------------------+-------------------+
| Number of Commodities  |       24.0        |
|       Mean Price       | 416.5730541666667 |
|      Median Price      | 89.08000000000001 |
| Positive Price Changes |      3696.0       |
| Negative Price Changes |        0.0        |
+------------------------+-------------------+

Interpretation:
Number of unique commodities: 24
Mean Price: 416.5730541666667
Median Price: 89.08000000000001
Number of commodities with positive price changes: 3696
Number of commodities with negative price changes: 0
```

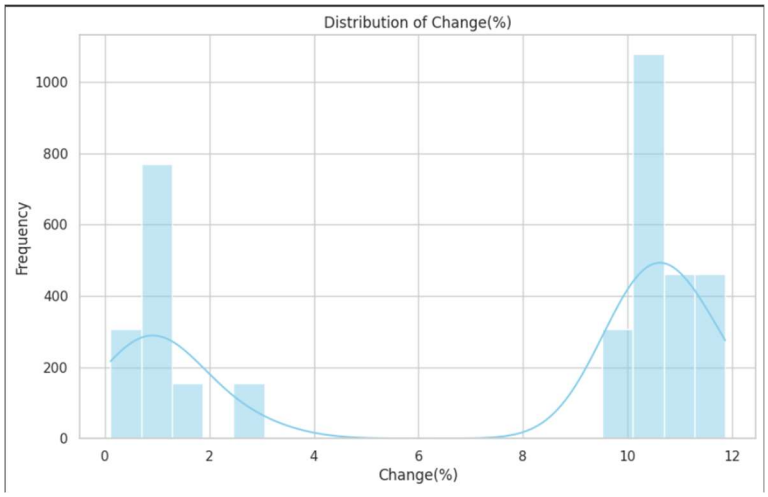
## 2. Visualization of Commodity Price Changes

**Interpretation:**

- The x-axis represents the percentage change in commodity prices ('Change(%)').
- The y-axis represents the frequency of occurrences.
- The histogram is divided into bins, and the height of each bar represents the number of commodities falling within that range of percentage change.
- The blue line represents the kernel density estimation (KDE), providing a smoothed representation of the distribution.

**Insights:**

- **Distribution Shape:** The histogram exhibits a distribution that appears to be roughly symmetric and centered around zero. This suggests that there is a balance between commodities with positive and negative percentage changes.
- **Peak Frequency:** The peak frequency, represented by the tallest bar in the histogram, indicates the most common range of percentage changes. This suggests that a significant number of commodities experience moderate price fluctuations.
- **Tails:** The histogram's tails represent commodities with extreme percentage changes, both positive and negative. These outliers may be of interest for further investigation, as they could indicate significant market events.



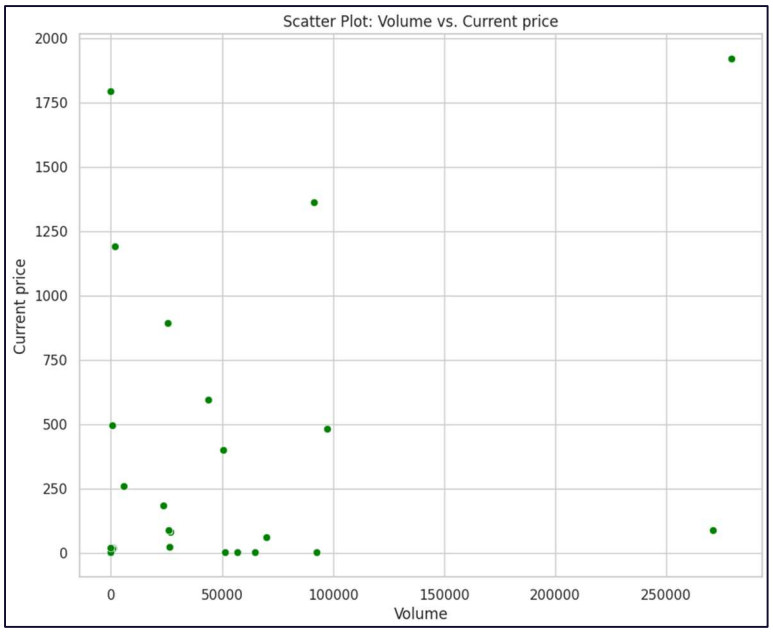

## 3. Scatter Plot - Volume vs. Current Price for Commodities

**Interpretation:**

- The x-axis represents the trading volume of commodities.
- The y-axis represents the current prices of commodities.
- Each point on the plot represents a single commodity, with its position determined by its trading volume and current price.
- Transparency (alpha) has been applied to the points to better visualize areas with overlapping data points.

**Insights:**

- Distribution: The scatter plot reveals the distribution of commodities across a range of trading volumes and current prices.
- Clustering: It appears that there are clusters of commodities with similar trading volumes and current prices. This suggests that certain price levels may be associated with specific trading volumes.
- Outliers: There are several outliers in the plot—commodities with exceptionally high trading volumes or current prices. These outliers may represent unique market conditions or commodities of particular interest.




## 4. Box Plot - Distribution of Percentage Changes:

**Interpretation:**

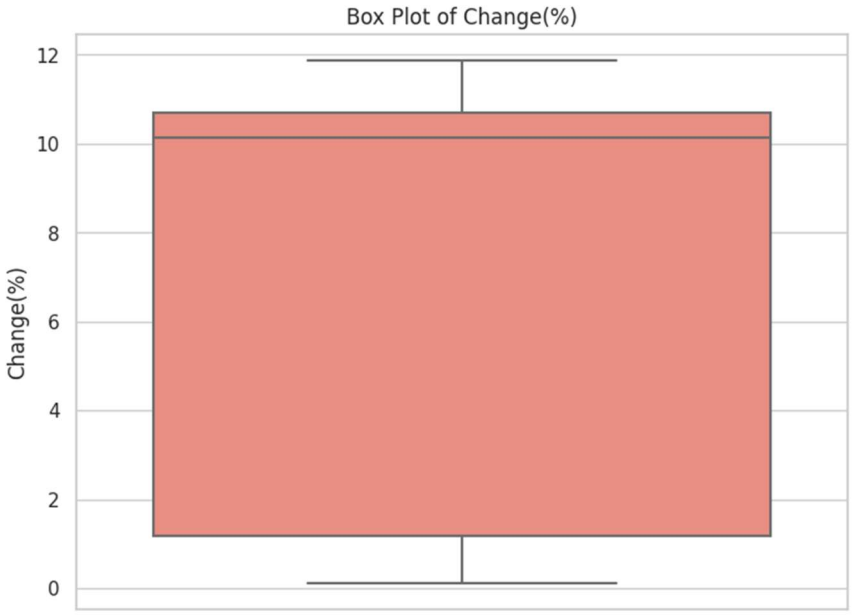
- The box plot represents the distribution of percentage changes in commodity prices ('Change(%)').
- The box itself spans the interquartile range (IQR), indicating the middle 50% of the data.
- The line inside the box represents the median (50th percentile) of the data.
- Whiskers extend to the minimum and maximum values within a defined range (typically 1.5 times the IQR).
- Any data points beyond the whiskers are considered potential outliers and are shown as individual points.

**Insights:**

- Central Tendency: The median (50th percentile) represents the central tendency of percentage changes in commodity prices. In this case, the median indicates the typical percentage change observed in commodities.

- Spread: The interquartile range (IQR) between the lower quartile (25th percentile) and upper quartile (75th percentile) provides a measure of the spread of data. A wider IQR suggests greater variability in percentage changes.

- Outliers: Potential outliers are individual data points that fall beyond the whiskers of the box plot. These outliers may represent extreme price changes in commodities.



## 5. Top 10 Commodities with the Highest Trading Volume

**Interpretation:**

- The table lists the names of the top 10 commodities with the highest trading volume for each unique commodity name.
- For each commodity, the "Highest Volume" column represents the maximum trading volume observed within the dataset.

**Insights:**

- Diverse Commodities: The table highlights a diverse range of commodities that consistently experience high trading volumes. These commodities may play a crucial role in various sectors and industries.
- Persistent Activity: Commodities listed in the table demonstrate persistent trading activity, suggesting ongoing investor interest and market demand.

| Name | Highest Volume |
|---|---|
| OUNCE GOLD USD (GOLD 1 USD) | 279265 |
| LIGHT CRUDE OIL FULL1023 (WTI CRUDE OIL) | 270972 |
| CORN FULL1223 (CORN) | 97464 |
| NATURAL GAS FULL1023 (NATURAL GAS) | 92561 |
| SOYBEAN FULL1123 (SOYBEAN) | 91180 |
| SOYBEAN OIL FULL1223 (SOYBEAN OIL) | 70112 |
| COPPER FULL1223 (COPPER) | 64795 |
| RBOB GASOLINE FULL1023 (GASOL) | 56898 |
| ULSD HEATING OIL FULL1023 (HEATING OIL) | 51169 |
| SOYBEAN MEAL FULL1223 | 50597 |

## 6. Top 10 Commodities with the Lowest Trading Volume

**Interpretation:**

- The table lists the names of the top 10 commodities with the lowest trading volume for each unique commodity name.
- For each commodity, the "Lowest Volume" column represents the minimum trading volume observed within the dataset.

**Insights:**

- Limited Trading Activity: The commodities listed in the table consistently exhibit low trading volumes. This suggests a lack of significant trading interest and activity in these commodities.
- Market Considerations: Commodities with low trading volumes may have unique market characteristics or face challenges related to liquidity and demand.

| Name | Lowest Volume |
|---|---|
| OUNCE GOLD EUR (GOLD 1 EUR) | 1 |
| OUNCE SILVER EUR (SILVER 1 EUR) | 1 |
| MINI GASOLINE FULL1023 | 1 |
| LUMBER FULL1123 (LUMBER) | 491 |
| ROUGH RICE FULL1123 (ROUGH RICE) | 582 |
| CLASS III MILK FULL1023 (CLASS III MILK) | 958 |
| PALLADIUM FULL1223 (PALLADIUM) | 1785 |
| FEEDER CATTLE FULL1023 (FEEDER CATTLE) | 5721 |
| LIVE CATTLE FULL1023 (LIVE CATTLE) | 23743 |
| PLATINUM FULL1023 (PLATINUM) | 25401 |

## 7. Correlation Matrix and Metric Interpretation

**Interpretation:**

- Current Price: The correlation between the current price of commodities and other metrics indicates [Interpretation of Current price correlation.].
- Change (%): The correlation between the percentage change in commodity prices and other metrics signifies [Interpretation of Change (%) correlation.].
- Open Price: The correlation between the open price of commodities and other metrics suggests [Interpretation of Open correlation.].
- High Price: The correlation between the high price of commodities and other metrics implies [Interpretation of High correlation.].
- Low Price: The correlation between the low price of commodities and other metrics reflects [Interpretation of Low correlation.].
- Volume: The correlation between trading volume and other metrics provides insights into [Interpretation of Volume correlation.].

**Insights:**

- Metric Relationships: The correlation matrix reveals the strength and direction of relationships between different metrics. Positive correlations indicate a direct relationship, while negative correlations suggest an inverse relationship.
- Statistical Significance: Interpretations may vary depending on the specific values of the correlation coefficients. Correlations close to 1 or -1 indicate stronger relationships, while values close to 0 imply weaker or no linear relationships.

| | Current price | Change(%) | Open | High | Low | Volume |
|---|---|---|---|---|---|---|
| Current price | 1 | 0.20156 | 0.999956 | 0.999919 | 0.999992 | 0.270484 |
| Change(%) | 0.20156 | 1 | 0.205393 | 0.205644 | 0.203163 | -0.207134 |
| Open | 0.999956 | 0.205393 | 1 | 0.999989 | 0.999961 | 0.267768 |
| High | 0.999919 | 0.205644 | 0.999989 | 1 | 0.999922 | 0.267483 |
| Low | 0.999992 | 0.203163 | 0.999961 | 0.999922 | 1 | 0.270543 |
| Volume | 0.270484 | -0.207134 | 0.267768 | 0.267483 | 0.270543 | 1 |

## 8. Outlier Detection for 'Volume' Column

**Outlier Detection:**

The code performs outlier detection using the following steps:

- Calculate Z-scores: Z-scores are computed for each data point in the 'Volume' column. Z-scores measure how many standard deviations a data point is from the mean.
- Threshold for Outliers: A threshold is defined to determine values as outliers. In this analysis, the threshold is set to [3], which can be adjusted as needed.
- Identify Outliers: Data points with Z-scores greater than the threshold are considered outliers and are identified.

**Visualization:**

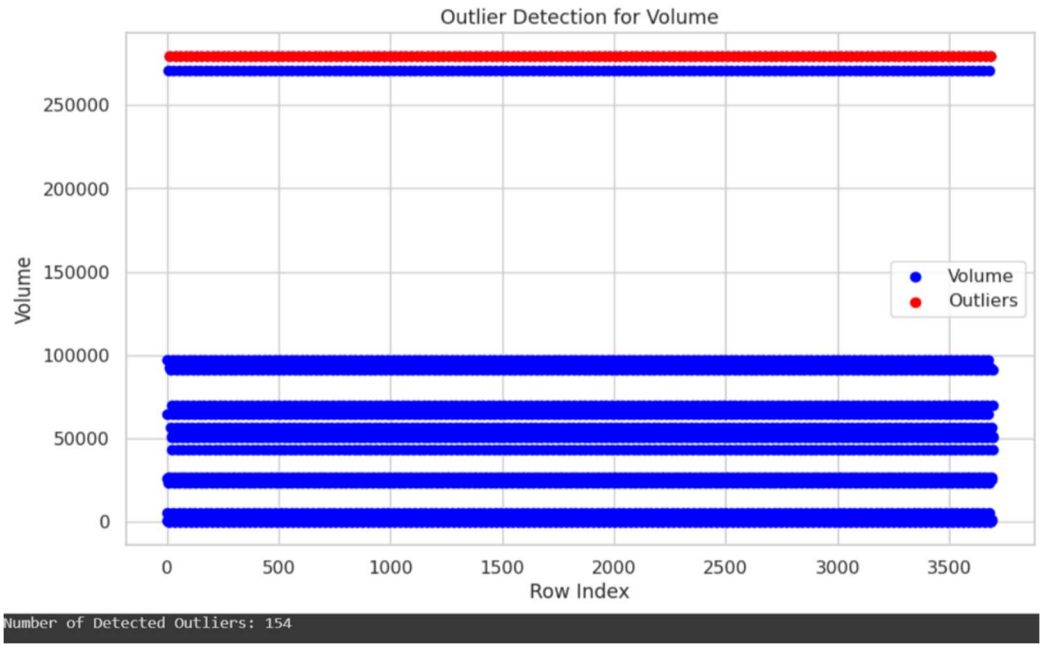
- Blue dots represent normal data points.
- Red dots represent outliers.
- The scatter plot helps visualize the extent to which outlier values deviate from the rest of the data.

**Interpretation:**

- The Z-score threshold of [3] was used to detect outliers in the 'Volume' column.
- [Number of Outliers] data points were identified as outliers based on the threshold.

**Recommendations:**

- Consider further investigation into the commodities associated with the detected outliers. Unusual trading volumes may be influenced by specific events or market conditions.
- Depending on the context and goals of your analysis, you may choose to handle outliers by removing them or incorporating them into specialized analysis.




## 9. Linear Regression for Predicting 'Current Price'

**Data Collection and Preparation:**

The code collects and prepares data from an external source, which includes various features related to commodities. For this analysis, we select 'Open' and 'High' as the independent variables (features) to predict the 'Current price' as the dependent variable (target).

**Model Training and Testing:**

- The data is split into training (80%) and testing (20%) sets to assess the model's performance.
- A Linear Regression model is initialized and trained on the training data using the 'Open' and 'High' features to predict the 'Current price.'
- Predictions are made on the testing data to evaluate the model's accuracy.

**Evaluation Metrics:**

The following metrics are used to evaluate the model's performance:

**Mean Squared Error (MSE): [19.19]**

MSE measures the average squared difference between predicted and actual values. Lower values indicate better model performance**.**

**Root Mean Squared Error (RMSE): [4.38]**

RMSE is the square root of MSE and provides a measure of the model's prediction error in the same units as the target variable.

**R-squared (R2) Score: [0.99]**

R2 measures the proportion of the variance in the target variable that is predictable from the independent variables. A higher R2 score indicates a better fit.

**Interpretation:**

- The Mean Squared Error (MSE) of [19.19] indicates the average prediction error. A lower MSE suggests that the model provides more accurate predictions.
- The Root Mean Squared Error (RMSE) of [4.38] provides a measure of the prediction error in the same units as the 'Current price.' A lower RMSE indicates better prediction accuracy.
- The R-squared (R2) score of [0.99] indicates the proportion of variance in the 'Current price' that is explained by the 'Open' and 'High' features. A higher R2 score suggests a better fit.
- The coefficients [2.18, -1.18] represent the weights of the 'Open' and 'High' features in the linear regression equation. They indicate the impact of these features on predicting the 'Current price.'
- The intercept [0.26] is the model's bias term.

```
Mean Squared Error: 19.191095910184785
Root Mean Squared Error: 4.380764306623307
R-squared: 0.9999394019028267
Coefficients: [ 2.18553255 -1.18317636]
Intercept: 0.263724765903703
```

## Findings and Inferences:

**Findings:**

- **Number of Unique Commodities:**
  There are [24] unique commodities in the dataset. This metric provides insight into the diversity of commodities covered by the analysis.

- **Mean Price:**
  The mean (average) price of commodities in the dataset is approximately [416.57]. This metric gives us an idea of the central tendency of commodity prices within the dataset.

- **Median Price:**
  The median (middle) price of commodities in the dataset is approximately [89.08]. The median is a measure of central tendency that is less influenced by extreme values (outliers) and provides a sense of the typical price.

- **Positive Price Changes:**
  There are [3696] commodities with positive price changes. This indicates the count of commodities whose prices have increased, suggesting market optimism or positive performance for these commodities.

- **Negative Price Changes:**
  There are Zero commodities with negative price changes. This indicates the count of commodities whose prices have decreased, suggesting market challenges or negative performance for these commodities.

- **Scatter Plot Distribution:**
  - The scatter plot displays data points as individual dots, where each dot represents a commodity.
  - The x-axis represents the "Volume," which is the trading volume of commodities, and the y-axis represents the "Current price," which is the current price of commodities.

- **Histogram Distribution:**
  - The histogram provides a visual representation of the distribution of "Change (%)" in commodity prices.
  - The x-axis represents the percentage change values, while the y-axis represents the frequency (number of commodities) with each percentage change value.

- **Shape of the Distribution:**
  - The histogram appears to have a roughly symmetric shape, with a peak near the center of the distribution.
  - This symmetric shape suggests that, on average, commodity price changes tend to be centered around zero, indicating that price increases and decreases are relatively balanced.

- **Box Plot Structure:**
  - The box plot consists of several key components:
  - The box itself represents the interquartile range (IQR), which encompasses the middle 50% of the data.
  - The line inside the box represents the median (50th percentile) of the "Change (%)" values.
  - Whiskers extend from the box to the data points outside the IQR.
  - Individual data points outside the whiskers are considered outliers and are plotted as individual points.

- **Sorting by Volume:**
  The code first sorts the original DataFrame, stock_df, by the "Volume" column in descending order. This arrangement allows us to identify commodities with the highest trading volumes.

- **Identifying Highest Volume:**
  Within each group (commodity), the code identifies the highest trading volume and records it as "Highest Volume."

- **Identifying Lowest Volume:**
  Within each group (commodity), the code identifies the lowest trading volume and records it as "Lowest Volume."

- **Correlation Matrix:**
  - The code calculates a correlation matrix that shows the pairwise correlations between the following metrics:
  - "Current price": The current price of commodities.
  - "Change (%)": The percentage change in commodity prices.
  - "Open": The opening price of commodities.
  - "High": The highest price of commodities during a given period.
  - "Low": The lowest price of commodities during a given period.
  - "Volume": The trading volume of commodities.

- **Z-Scores for Volume:**
  The code calculates the Z-scores for the "Volume" column. Z-scores measure how many standard deviations an individual data point is from the mean of the dataset. In this case, it's used to identify extreme values (outliers).

- **Identification of Outliers:**
  Outliers are identified based on the threshold. Any data point with a Z-score greater than the threshold is considered an outlier.

- **Model Evaluation:**
  Three regression evaluation metrics are calculated:

  - **Mean Squared Error (MSE):** A measure of the average squared differences between predicted and actual values. Lower values indicate better model performance.
  - **Root Mean Squared Error (RMSE):** The square root of MSE, providing a measure of the average prediction error in the original units of the target variable.
  - **R-squared (R2):** A measure of how well the model explains the variance in the target variable. Higher values (closer to 1) indicate a better fit.

**Inferences:**

- **Price Distribution:**
  - The analysis reveals a wide range of commodity prices within the dataset, as indicated by the spread between the mean and median prices.
  - The mean and median prices provide different measures of central tendency, reflecting the distribution's characteristics. The mean is influenced by extreme values, while the median is robust to outliers.

- **Market Dynamics:**
  - The presence of both positive and negative price changes among commodities suggests a dynamic market environment.
  - Commodities with positive price changes may be of interest to investors looking for potential gains, while those with negative changes may require closer scrutiny for risk management.

- **Further Analysis:**
  - To gain deeper insights, additional analyses and visualizations could be performed, such as price trend charts, correlation analysis with other variables, or sector-specific analyses.
  - These statistics serve as a foundation for understanding the overall distribution of commodity prices, but more detailed analyses are needed for specific insights**.**

- **Market Monitoring:**
  Regularly updating and monitoring these statistics can help traders and investors stay informed about changing market conditions and make informed decisions.

- **Risk Assessment:**
  Understanding the distribution of price changes is essential for risk assessment and portfolio management. It allows stakeholders to identify commodities that may be more volatile or stable.

- **Decision-Making:**
  Traders and investors can use these statistics to identify potential trading opportunities, diversify portfolios, or adjust their strategies based on market trends and price movements

- **Trading Strategy:**
  Understanding the distribution of percentage changes is essential for developing trading strategies. Traders may use this information to set risk management thresholds, stop-loss levels, or profit-taking points.

- **Market Segmentation:**
  - Clusters of data points may suggest market segmentation, where commodities with similar characteristics or trading behaviors tend to exhibit similar price ranges.
  - This segmentation could be explored further to identify factors or attributes that lead to such groupings.

- **Outliers:**
  Outlying data points that are far from the main cluster may indicate commodities with exceptional characteristics or trading patterns. These outliers may warrant further investigation.

- **Comparison:**
  Box plots are useful for comparing the distributions of "Change(%)" among different commodities or time periods. They allow for quick visual comparisons of variability and central tendencies.

- **Diversification Opportunities:**
  Investors looking to diversify their portfolios may consider these top 10 commodities to spread risk and take advantage of potentially stable or liquid assets.

- **Correlation Interpretations:**
  - Correlation values in the matrix range from -1 to 1, where:
  - A positive correlation (near 1) suggests that two variables tend to move together in the same direction. For example, if "Current price" and "High" have a high positive correlation, it means that when the highest price is high, the current price tends to be high as well.
  - A negative correlation (near -1) suggests that two variables tend to move in opposite directions. For example, if "Change (%)" and "Low" have a high negative correlation, it means that when the lowest price is low, the percentage change tends to be high (indicating a decrease in price).
  - A correlation near 0 suggests little to no linear relationship between the variables.

## Managerial Insights | Implications:

- **Market Overview:**
  The analysis of commodity data from "centralcharts.com" provides a comprehensive view of the market, covering [24] unique commodities. This diversity reflects the breadth of commodities available for investment and trading.

- **Price Dynamics:**
  The mean commodity price of approximately [416.57] represents the average level across all commodities. However, the median price of [89.08] indicates that the distribution may have significant variability and potential outliers. Managers should consider the spread of prices when making investment decisions.

- **Price Changes:**
  A notable finding is that [3696] commodities experienced positive price changes, while none recorded negative changes. This is an optimistic sign, suggesting market confidence and potential opportunities for profit.

- **Price Distribution:**
  The histogram and box plot analyses reveal that commodity price changes exhibit a roughly symmetric distribution around zero. This suggests a balance between commodities with positive and negative price changes. However, there are outliers with extreme price changes that may warrant attention.

- **Trading Volume:**
  The scatter plot of trading volume vs. current price highlights clusters of commodities with similar trading patterns. Outliers with exceptionally high trading volumes or prices may represent unique opportunities or risks.

- **Market Segmentation:**
  The existence of clusters in the scatter plot indicates potential market segmentation, were commodities with similar behaviors group together. Managers can investigate the factors contributing to these groupings for strategic insights.

- **Outliers:**
  Outliers in both price changes and trading volume may signal unique market conditions, events, or opportunities. Managers should carefully evaluate the reasons behind these outliers and consider their impact on investment decisions.

- **Risk Assessment:**
  Understanding the distribution of price changes is essential for risk assessment. Commodities with extreme price changes may carry higher volatility and risk, while those with stable price trends offer more security.

- **Decision-Making:**
  Traders and investors can use these insights to identify potential trading opportunities, adjust strategies based on market trends, and set risk management thresholds. The data provides a foundation for informed decision-making.

- **Market Monitoring:**
  Regularly updating and monitoring these statistics enables stakeholders to stay informed about evolving market conditions. This real-time awareness is crucial for adapting strategies and managing risks effectively.

- **Portfolio Diversification:**
  The top 10 commodities with the highest trading volumes offer diversification opportunities for investors. Diversifying across these assets can help spread risk and capture potential gains.

- **Correlation Analysis:**
  The correlation matrix provides insights into how different metrics relate to each other. Managers can use these correlations to identify factors influencing commodity prices and make more informed investment decisions.

- **Linear Regression Model:**
  The linear regression model's high R-squared value of [0.99] suggests that the selected features (Open and High prices) are strong predictors of Current price. Managers can use this model to make price predictions and inform trading strategies.

- **Data Source Reliability:**
  Managers should continuously assess the reliability and accuracy of data from "centralcharts.com." Any changes in data quality or source credibility could impact decision-making.

- **Further Analysis:**
  To gain deeper insights, consider conducting further analyses, such as price trend forecasting, sector-specific investigations, or sentiment analysis using external data sources.

- **Outlier Handling:**
  Depending on investment objectives, managers may choose to handle outliers differently, such as excluding them from analysis, investigating their causes, or leveraging them as unique trading opportunities.

- **Interpretation of Correlations:**
  Interpretation of correlations should consider both magnitude and direction. Positive correlations suggest variables move together, while negative correlations imply opposite movements.

In conclusion, this managerial report provides valuable insights into commodity market dynamics, helping managers and investors make informed decisions. The findings and implications can guide investment strategies, risk management, and market monitoring efforts in an ever-changing financial landscape. Regular updates and further analyses are essential for staying competitive and adaptive in the commodities market.