Name of Student: James Bobby Kiawu
Submitted To: Ratinder Rajpal, Lecturer.
Willis College: Introduction to Artificial Intelligence
Intro AI **Data Collection and Initial Analysis of Stock Market Data**
Date: November 20, 2025. GitHub: https://github.com/jkiawu7/colab-git-V2-Jkiawu7/tree/main

Technical Notes: Findings, Trends, and Initial Hypotheses on Stock Market Dataset

1. Overview of the Dataset: The dataset combines daily historical stock prices with company metadata such as sector, industry, and stock exchange. After cleaning and merging the datasets, the analysis focused on long-term patterns by segmenting data into decades (1970s through 2010s). Each decade was examined using summary statistics and a series of visualizations, including monthly average closing price trends, volume distributions, and boxplots for high and low prices.

2. Key Findings from Exploratory Data Analysis

2.1 Long-Term Price Trends: Across all decades, there is a noticeable upward trend in stock prices. Earlier decades such as the 1970s and 1980s show relatively low average price levels with minimal volatility. Beginning in the 1990s, the market exhibits steeper upward movement in monthly average closing prices, reflecting rapid technological growth, increased globalization, and higher participation in financial markets. The 2000s show more dramatic fluctuations, including a sharp decline during the 2008 global financial crisis. The 2010s continue the upward trajectory, with sustained growth supported by technology, consumer demand, and strong corporate earnings.

Interpretation: Stock prices generally rise over time, but later decades experience stronger fluctuations and steeper market cycles. This aligns with known macroeconomic events and higher sensitivity to global financial conditions.

2.2 Volume Patterns and Market Activity : Volume histograms reveal a significant increase in trading activity over the decades. The 1970s and 1980s show lower, tightly grouped volume levels, indicating limited participation and less electronic trading. From the 1990s onward, volumes become broader and significantly higher, with many right-tailed distributions showing extremely active trading days.

Interpretation:

Trading liquidity increased dramatically over time due to:

- Market digitization
- Growth of institutional investing
- Emergence of algorithmic and high-frequency trading
- Greater retail investor access

These shifts contributed to enhanced liquidity but also higher intraday volatility.

2.3 Price Range, Volatility, and Outliers :
Boxplots of high and low prices show that the spread between highs and lows widens considerably in the 1990s, 2000s, and 2010s. Earlier decades exhibit tighter and more stable fluctuation bands. The presence of outliers increases notably in later decades, especially during economic shocks. For example, research shows that the dot-com bubble peaked around 2000–2002 and erased vast valuations in the internet sector. Examples include:

- Tech bubble (late 1990s)[1]
- Housing/financial crisis (2008)[2]

- Short-term volatility during global events

2.4 Seasonal and Periodic Effects
While monthly averages smooth out much of the short-term noise, certain seasonal tendencies emerge:
- Prices often decline during recession periods
- Some decades show cyclical patterns corresponding to broader economic cycles
- Increased volatility tends to cluster during crisis or boom years
- Trading volume increases during periods of rapid price movement

3. Anomalies and Notable Observations :

Several anomalies were detected in the EDA:
- Sharp declines in monthly averages during crisis years (e.g., 2000–2002, 2008–2009).
- Sudden spikes in volume suggesting unusual trading activity, likely influenced by major news events or market corrections.
- Outliers in high/low price distributions particularly evident in the 2000s and 2010s.

These anomalies align with real-world events that disrupted financial markets.

4. Initial Hypotheses from the EDA

Based on the observed patterns, the following hypotheses can be proposed for further analysis or modeling:

Hypothesis 1: Decade and macroeconomic environment strongly influence stock prices.
Later decades with technological advancements and globalization display higher average prices and greater volatility.

Hypothesis 2: Trading volume is positively correlated with price volatility.
Periods of high-volume trading coincide with larger price swings, suggesting that market activity contributes to price instability.

Hypothesis 3: Certain sectors may respond differently across decades.
For example, technology and finance sectors might show stronger growth trends compared to utilities or consumer staples.

Hypothesis 4: Economic crises create identifiable patterns in price behavior.
Sharp declines during recessions suggest that macroeconomic shocks have a measurable and predictable impact on stock performance.

5. Conclusion
The exploratory data analysis highlights clear long-term growth in stock prices, coupled with rising volatility and trading activity in more recent decades. Visualizations and summary statistics reveal meaningful decade-by-decade differences, reflecting broader technological, economic, and structural changes in financial markets. These findings provide a strong foundation for deeper statistical modeling, sector analysis, or predictive modeling in future stages of the project.

References:
1. "Understanding the Dotcom Bubble: Causes, Impact, and …". Investopedia article. Investopedia https://www.investopedia.com/terms/d/dotcom-bubble.asp
2. McGill University Economics. (n.d.). *The U.S. Housing Collapse and the Financial Crisis of 2007-08*. Retrieved from https://www.mcgill.ca/economics/files/economics/the_u.s._housing_collapse_and_the_financial_crisis_of_2007-08.pdf