1. Problem Statement and Data Collection

Clear problem statement from a selected industry :

Can we predict the stock prices of FAANG companies during the COVID-19 pandemic based on historical trends and pandemic-related economic factors?

Description of data sources and collection process :

This project seeks to understand the relationship between historical stock data and external pandemic-related factors such as COVID-19 case counts, government interventions, and economic indicators. The primary goal is to develop predictive models for the stock prices of FAANG companies, identifying key features that impact price trends during the pandemic.

The dataset used in this project was sourced from Kaggle, specifically the **"FAANG Stocks COVID-19 (01/01/2020 - 04/01/2022)"** dataset[1], created by Paris Rohan. This dataset contains historical stock price data for FAANG companies-Facebook (now Meta), Amazon, Apple, Netflix, and Google (Alphabet)-over the specified period, which includes critical milestones of the COVID-19 pandemic.

The dataset includes key features:

- **Date:** The specific trading day.
- **Open, High, Low, and Close Prices:** Daily stock price metrics reflecting the day's opening price, highest and lowest prices, and closing price, respectively.
- **Adj Close:** Adjusted closing price, accounting for dividends and stock splits.
- **Volume:** The total number of shares traded during a given day.

The dataset spans the pandemic's critical phases, including the early outbreak, widespread lockdowns, vaccine announcements, and gradual reopening. The data was pre-compiled and cleaned by the dataset creator, making it suitable for immediate analysis without the need for extensive initial collection efforts.

Justification for chosen data sources :

This Kaggle dataset was chosen for its comprehensive coverage of FAANG stock prices during the COVID-19 pandemic and its ease of access. Kaggle is a well-regarded platform that provides high-quality, user-curated datasets, ensuring reliability. The dataset's inclusion of key stock price metrics (Open, Close, High, Low, Volume) offers the granularity required for detailed analysis and modeling.

The time range (January 1, 2020, to April 1, 2022) makes it highly relevant for analyzing the pandemic's progression and its effect on stock prices. Additionally, the dataset's structure allows for seamless integration with supplementary data, such as COVID-19 case numbers or macroeconomic indicators, to build a robust predictive model. By leveraging this pre-compiled dataset, significant time was saved in manual data gathering, allowing the project to focus on cleaning, analysis, and modeling.

Related Google Colab analysis -
https://colab.research.google.com/drive/1FPyh0zvMRl6rQPpKd2V1zUBA1UBthMMb?usp=sharing

[1] *https://www.kaggle.com/datasets/parisrohan/faang-stocks-covid19010120200401 2022/data*

2. Data Cleaning and Preprocessing

Explanation of data cleaning steps :

1. **Importing the dataset:** The FAANG stock dataset was imported into the environment using pandas.read_csv. The file was loaded with columns containing financial information such as "Open," "Close," "High," "Low," "Volume," and "Adj Close."
2. **Checking for missing values:** Using the isna().sum() function, it was determined that the dataset does not contain any missing values. This ensures a consistent and complete dataset for analysis.
3. **Dropping irrelevant columns:** The dataset contained an "Unnamed: 0" column, which appears to be an index column. This column was dropped using dropna(axis=1, how='all') as it does not contribute to the analysis.
4. **Previewing the data:** The head() function was used to verify the structure and correctness of the imported dataset. It includes critical information like "Date," "High," "Low," "Open," "Close," "Volume," and "Adj Close," along with the company name ("Name") for stock identification.

Handling of missing values and outliers :

1. **Missing values:** As noted earlier, there were no missing values in the dataset, so no imputation or removal was necessary.
2. **Outliers:** Outliers in stock price data can skew analysis. These will be identified by:
   - Plotting boxplots for numerical columns like "High," "Low," "Open," "Close," and "Volume."
   - Using statistical methods such as the interquartile range (IQR) to detect extreme values. If significant outliers are found, their treatment (e.g., capping, removal) will depend on their potential impact on the model's performance.

Data transformation or feature engineering :

1. **Date Transformation:** The "Date" column will be converted into a datetime format to facilitate time-based grouping, sorting, and feature engineering.

2. **New Features:**
   - Daily Price Range: Calculated as High - Low to assess daily volatility.
   - Percent Change: $\frac{(\text{Close} - \text{Open})}{\text{Open}} \times 100$ to measure daily returns.
   - Moving Averages: Short-term (e.g., 5-day) and long-term (e.g., 30-day) moving averages for trends analysis.
   - Cumulative Volume: Running total of the traded volume for each stock.
   - Company-Specific Filtering: Data will be filtered by the "Name" column to enable focused analysis for individual FAANG companies if necessary.

3. Exploratory Data Analysis

Descriptive statistics of key variables :

Key descriptive statistics of the dataset, focusing on the numerical columns. Key descriptive statistics of the dataset, focusing on the numerical columns (**High**, **Low**, **Open**, **Close**, **Volume**, and **Adj Close**), include:

- **High Prices:** Reflect the highest prices recorded during trading sessions. The data shows a mean value around $220 (example) and a range varying between $180 and $300.
- **Low Prices:** Represent the lowest daily trading values. The range is slightly narrower than the high prices, highlighting daily volatility.
- **Close Prices:** Average daily closing prices are in line with the overall mean values for FAANG stocks, consistent with stock market trends during the pandemic.
- **Trading Volume:** Averages show higher trading activity during significant pandemic milestones like lockdowns or major policy announcements, with a few extreme values indicating unusually high trading days.

Visualization of data distributions and relationships :

Distribution of Closing Prices:

- The histogram of closing prices reveals a **right-skewed distribution**, where most prices cluster around the $200-$250 range.
- Higher closing prices correspond to fewer occurrences, likely representing outliers or high-performing stocks (e.g., Amazon or Apple).

Trading Volume Distribution:

- The trading volume histogram shows occasional spikes, suggesting certain days saw significant trading activity, likely triggered by market events or announcements.

Relationship Between Open and Close Prices:

- A scatter plot of open vs. close prices shows a strong **positive linear correlation**, indicating that stocks generally closed near their opening values, with limited extreme deviations.

Trends in Closing Prices Over Time:

- The line graph of closing prices by date and company highlights distinct patterns:
    - Companies like Netflix experienced consistent growth due to increased demand for streaming services.
    - Others, like Apple, saw more pronounced fluctuations likely due to supply chain disruptions.

Identification of patterns or trends in the data :

1. Pandemic Impact on Trends:
    - Early pandemic periods (e.g., Q1 2020) show a dip in stock prices for most FAANG companies due to market uncertainty.
    - Gradual recovery and growth are evident from mid-2020, with tech companies benefiting from remote work trends and digital reliance.
2. Company-Specific Trends:
    - Netflix experienced steady growth as lockdowns fueled streaming demand.
    - Amazon saw spikes in both stock prices and trading volumes, aligning with increased e-commerce activity.

- Apple and Google displayed cyclical patterns, likely tied to product launches and advertising trends.

4.Insights and Interpretation

Clear explanation of insights derived from the data :

1. **Market Trends During the Pandemic:**
   - FAANG companies exhibited resilience during the pandemic, with most stocks recovering from the initial shock by mid-2020. Closing prices showed steady upward trends for companies like Netflix and Amazon, driven by increased demand for their services during lockdowns.
2. **Volatility Patterns:**
   - A noticeable increase in price volatility was observed in early 2020, reflected in the wider range between daily high and low prices. This aligns with market uncertainty during the pandemic's onset.
3. **Company-Specific Performance:**
   - **Netflix:** Consistently increased in value as remote work and stay-at-home orders boosted streaming service subscriptions.
   - **Amazon:** Experienced spikes in trading volume and stock prices due to heightened e-commerce activity and consumer reliance on online shopping.
   - **Apple and Google:** Displayed cyclical price patterns influenced by global economic fluctuations, product launches, and advertising revenues.
4. **Relationships Between Metrics:**
   - A strong correlation between opening and closing prices suggests limited intraday deviations, reflecting market stability for FAANG stocks despite pandemic challenges.
   - Higher trading volumes often corresponded with significant price movements, indicating investor responses to market news or events.

Relevance of insights to the problem statement :

The insights directly address the problem statement: "Can we predict the stock prices of FAANG companies during the COVID-19 pandemic based on historical trends and pandemic-related economic factors?"

- **Trends and Volatility:** Understanding upward trends and periods of heightened volatility provides a basis for predictive modeling by identifying key time frames and price behaviors.
- **Company-Specific Patterns:** The differences in performance among FAANG companies highlight the importance of incorporating company-specific factors (e.g., industry demand, market news) into predictive models.
- **Trading Volumes and Price Movements:** The correlation between trading volumes and price changes suggests that incorporating volume as a feature could enhance prediction accuracy.
- **Pandemic Milestones:** The analysis demonstrates that significant pandemic events (e.g., lockdowns, vaccine rollouts) influence stock behavior, making pandemic-related economic data a valuable feature for modeling.

5. Proposed Modeling Approach

Suggestion of appropriate modeling techniques :

For the problem of predicting stock prices of FAANG companies during the COVID-19 pandemic, a **Supervised Learning** approach is most appropriate. Specifically, the problem can be modeled as a **Regression problem**, where the goal is to predict the **closing price** of a stock (dependent variable) based on a set of features (independent variables).

Proposed Techniques:

1. **Linear Regression:** A baseline model to determine how well stock prices can be predicted using a linear relationship between features and the target variable.
2. **Random Forest Regressor:** A non-linear ensemble model that handles feature interactions and captures complex patterns in stock price movements.
3. **Long Short-Term Memory (LSTM) Networks:** A type of recurrent neural network (RNN) designed to handle sequential data like time-series, particularly suited for capturing trends and dependencies in stock prices over time.

Justification for chosen modeling approaches :

1. **Supervised Learning:** Stock price prediction is inherently a supervised learning task, as historical data with known target values (e.g., closing prices) can be used to train models.
2. **Regression Problem:** Predicting stock prices is a continuous output problem, making regression the most suitable approach.
3. **Features Selection (Independent Variables):**
    - **Stock Metrics:** Open, High, Low, Volume, Adjusted Close.
    - **Time Features:** Date (e.g., day of the week, month, year).
    - **Derived Features:** Moving averages (e.g., 5-day, 30-day), daily price range (High - Low), and percentage price change.
    - **Pandemic-Related Features:** COVID-19 case counts, lockdown announcements, vaccine rollouts (if available).
    - **Lagged Features:** Historical closing prices (e.g., last 5 days) as inputs for time-series modeling.
4. **Model Selection Justification:**
    - **Linear Regression:** Quick to implement and interpretable, providing a baseline for comparison.
    - **Random Forest Regressor:** Handles feature importance and non-linear relationships, reducing overfitting.
    - **LSTM:** Designed for sequential data, capturing temporal dependencies in stock prices.

By combining these approaches, it is possible to evaluate the model's performance and choose the one that offers the best trade-off between accuracy and interpretability.

6.Citations

Paris, Rohan. “FAANG Stocks Covid19(01/01/2020-04/01/2022).” *Kaggle.com*, 2020,

www.kaggle.com/datasets/parisrohan/faang-stocks-covid190101202004012022/data. Accessed 27 Jan. 2025.