Big Data and Capital Markets

Qianfan Cheng
Gavin
Hassan Mehmood
Sut Ring Ja
Shane Thomas
Gagan Tuli
Jeet Patel
Suraj Sinh Ramlavat
Jash Mayur Gosrani

3/9/24

Contents

Abstract:

This report outlines a detailed analysis of a comprehensive financial data project aimed at gathering, processing, analyzing, and visualizing various types of data related to stocks, financials, news, and other sources. The project involved a multi-step process, including data collection from APIs, storage in databases, ETL pipeline implementation, database migration, and dashboard creation using Power BI. The final product is to have the necessary data at a single point to make informed investment decisions.

1. Introduction:

The introduction provides an overview of the project's objectives, methodologies, and the significance of analyzing diverse datasets from multiple sources. It outlines the necessity of leveraging modern technologies and tools for effective data management and analysis in today's data-driven environment.

Data Sources

- Yahoo finance
- Financial Modelling Prep.
- Finn-Hub API.
- CSV data.

Databases

- MongoDB
- PostgreSQL
- MariaDB
- Hadoop
- SQL Express Server

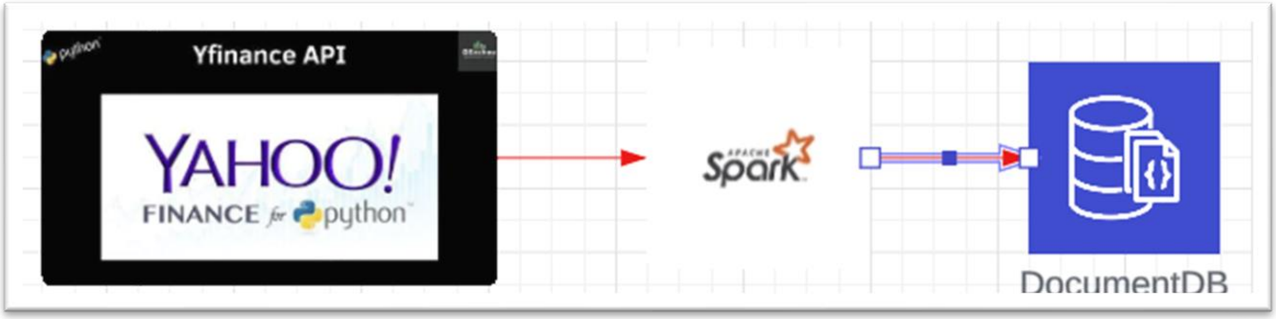
Programming Languages

- Spark
- SQL
- MongoDB Query Language

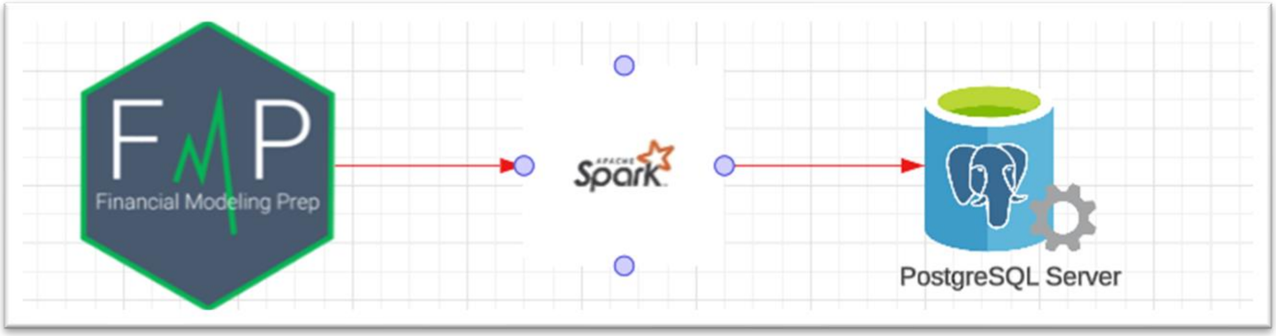
2. Data Collection and Storage:

This section delves into the specifics of data collection and storage. It discusses the sources of data, including yahoo finance, FMP API, Finn-hub API, and Snowflake for crime rate data. Each data set’s characteristics and relevance to the project are elaborated upon, along with the choice of databases for the data retrieved.

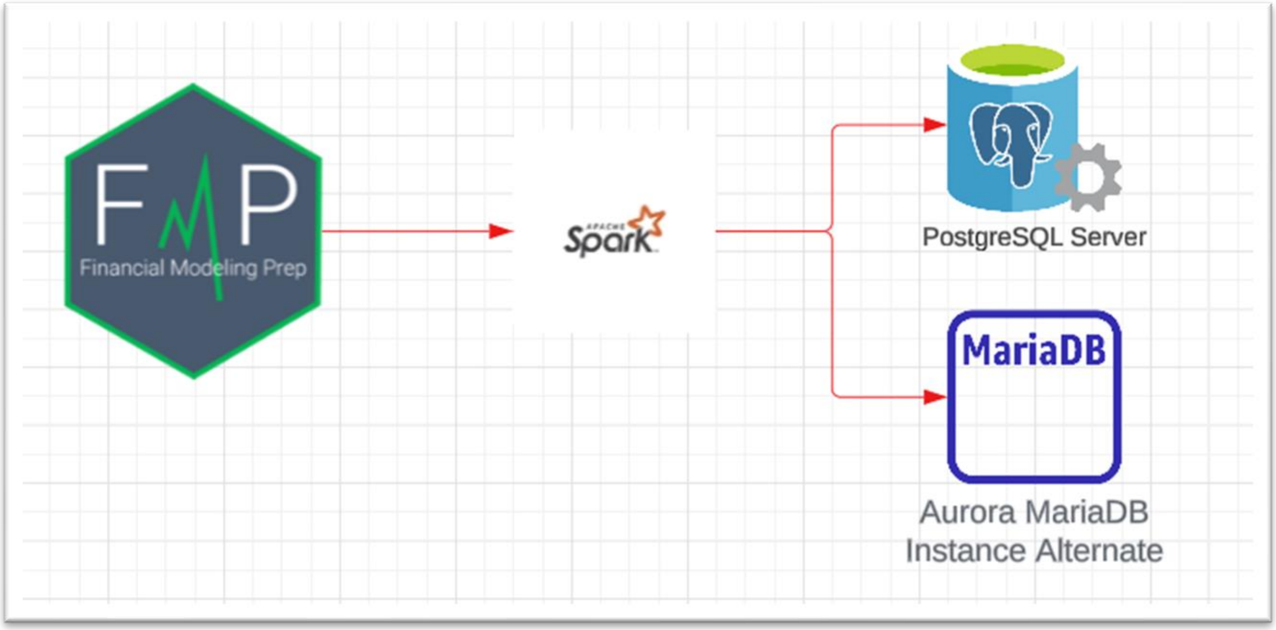
- **Stocks Market Data**: Fetched from Yahoo Finance (y-finance) and stored in MongoDB.



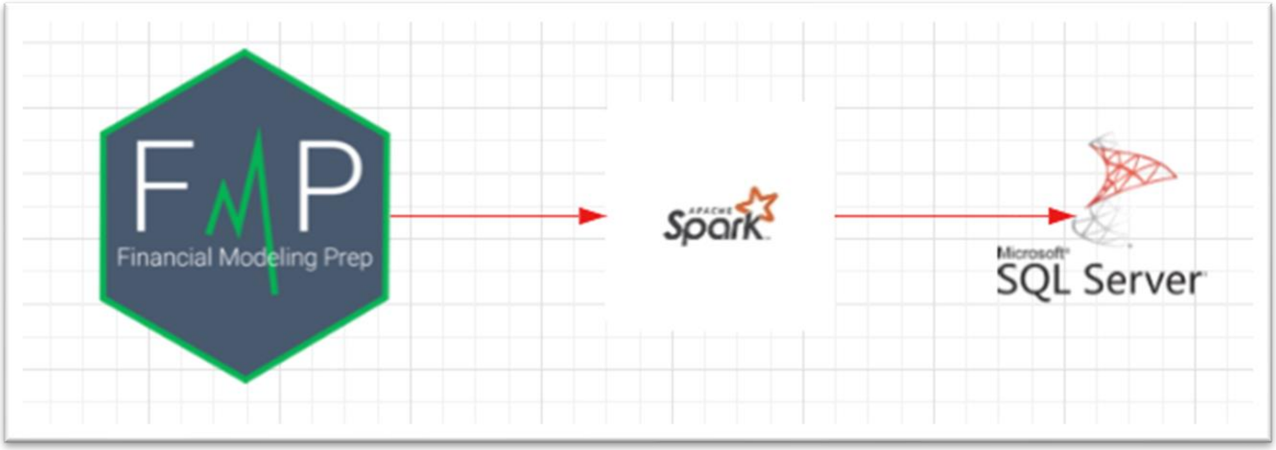
- **Income Statement**: Retrieved from the FMP API and stored in PostgreSQL.



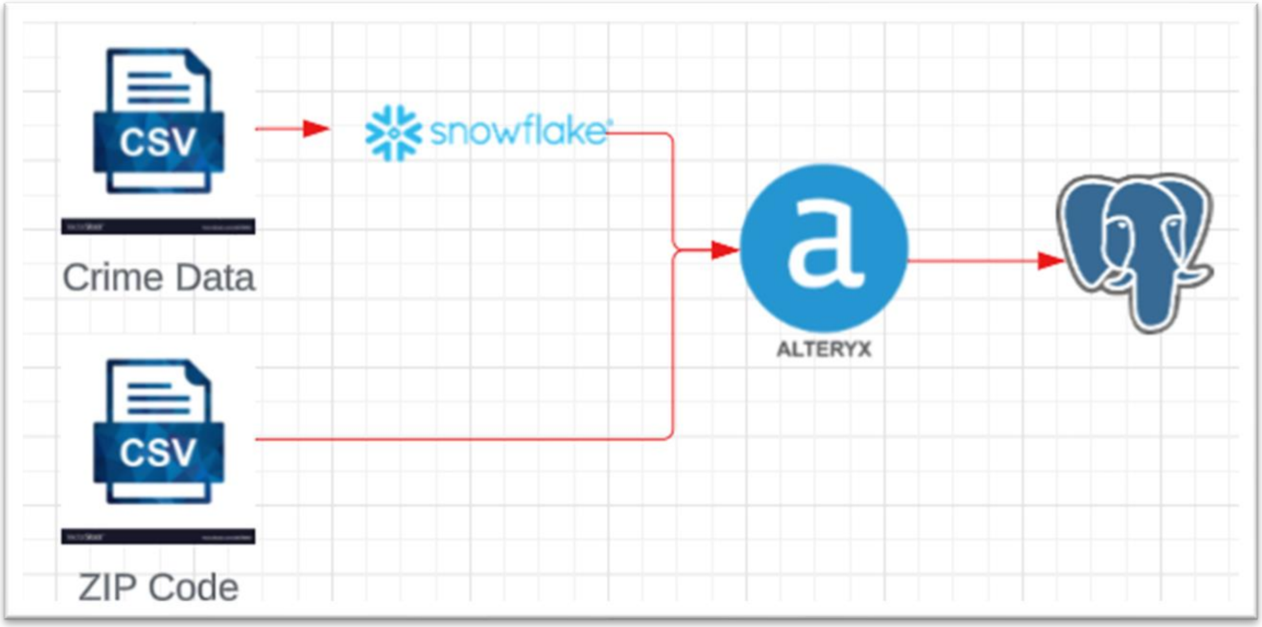
- **News Data**: Acquired from both the FMP API and Finn-hub API, stored in PostgreSQL and MariaDB respectively.



- **Company Profile**: Collected from the FMP API and stored in Hadoop.
- **Commitment of Traders Report**: Obtained from the FMP API and stored in SQL-Express Server.



- **Insider Trading**: Gathered from Finn-hub and saved in MongoDB.
- **Crime Rate CSV**: Stored in Snowflake.
- **ZIP Code CSV**: Additional geographical data.

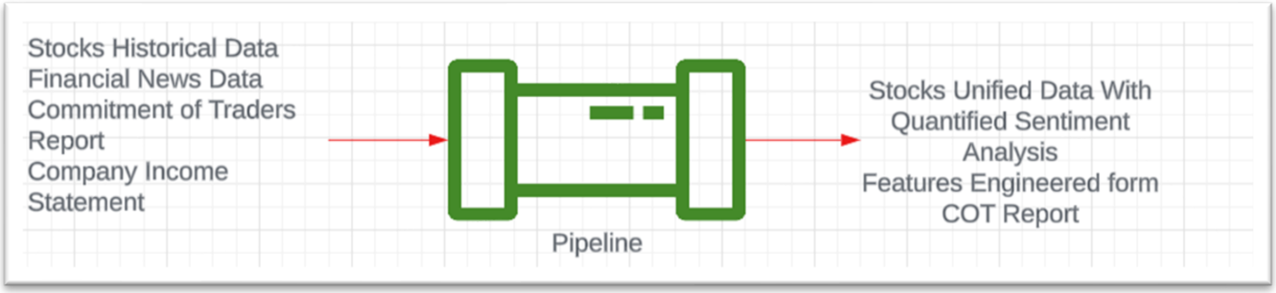


3. ETL Pipeline:

The ETL pipeline is described in detail, highlighting its role in extracting, transforming, and loading data from various sources into the staging database. The process of sentiment analysis using a pre-trained model from Hugging Face, feature engineering on the Commitment of Traders Report, and the integration of Crime Rate and ZIP Code data using Alteryx are explained step by step.

- **Sentiment Analysis**: Conducted using a pre-trained model from Hugging Face, joined with stock historical data.
- **Feature Engineering**: Applied to the Commitment of Traders Report.

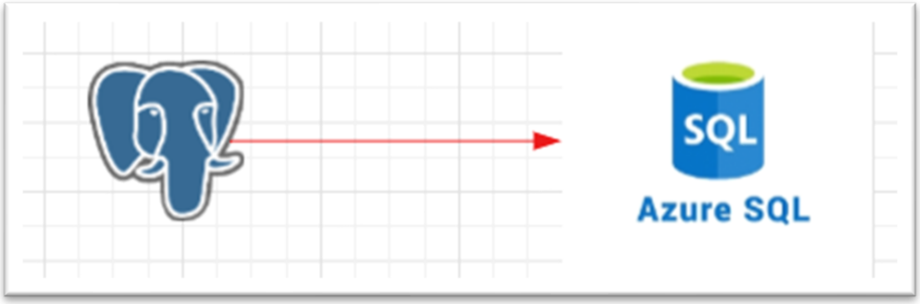
- **Data Transfer**: Income statement data, company profile data, and preprocessed data transferred to a local staging database (PostgreSQL).
- **Alteryx Integration**: Joined Crime Rate data and ZIP Code data, stored in PostgreSQL.


Stocks Historical Data
Financial News Data
Commitment of Traders Report
Company Income Statement
Pipeline
Stocks Unified Data With Quantified Sentiment Analysis
Features Engineered form COT Report

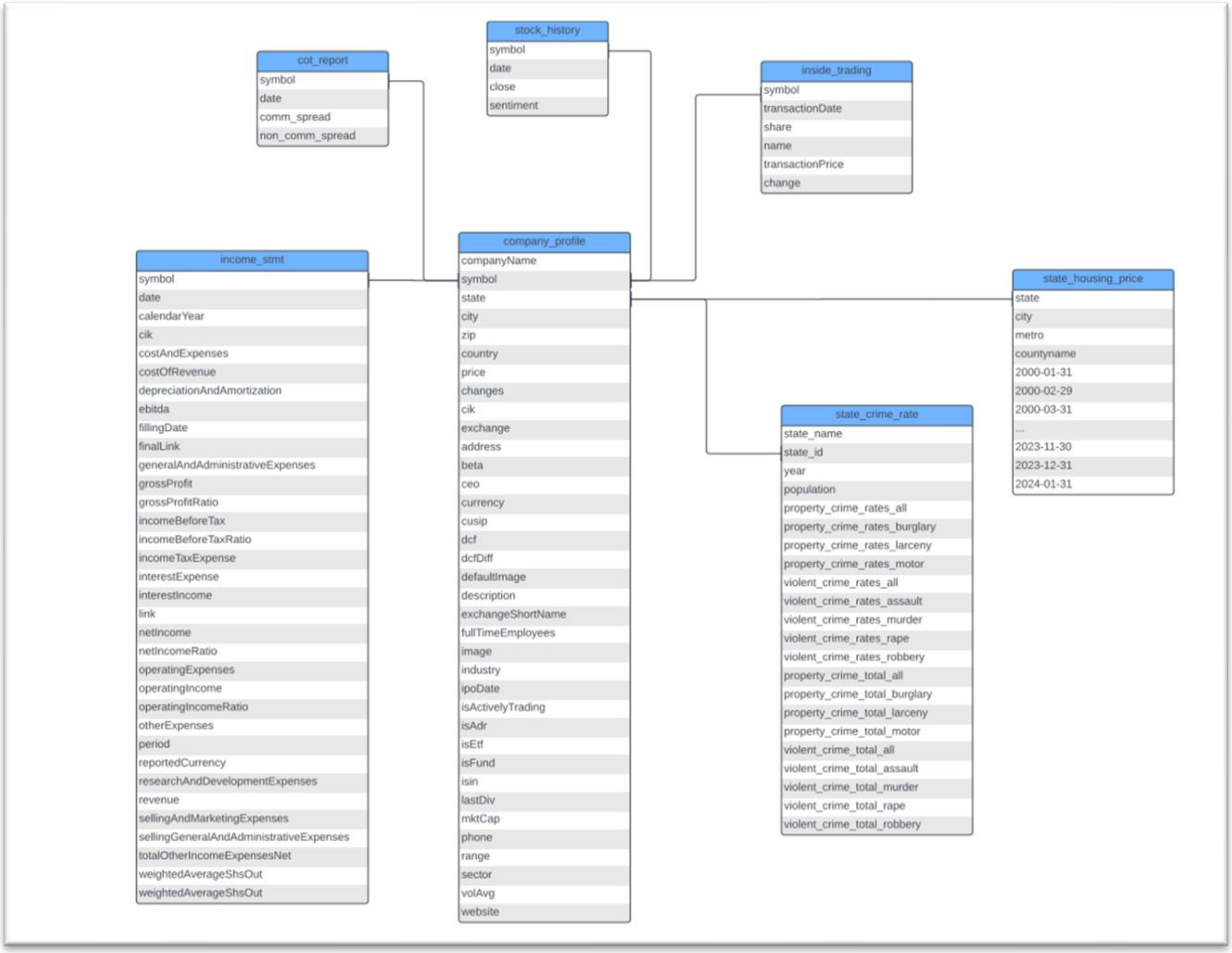
4. Database Migration:

This section discusses the migration of data from the local staging database (PostgreSQL) to Azure SQL for scalability and accessibility. The rationale behind choosing Azure SQL for hosting the final dataset is its advantages in terms of performance, security, and scalability.

- Data loaded from the local staging database (PostgreSQL) to Azure SQL for scalability and accessibility.
- We tried to ensure that the data stored in the final chosen database is not redundant yet contains information that might be useful for future analysis.



- Following is our final Entity Relation Diagram representing the star schema:



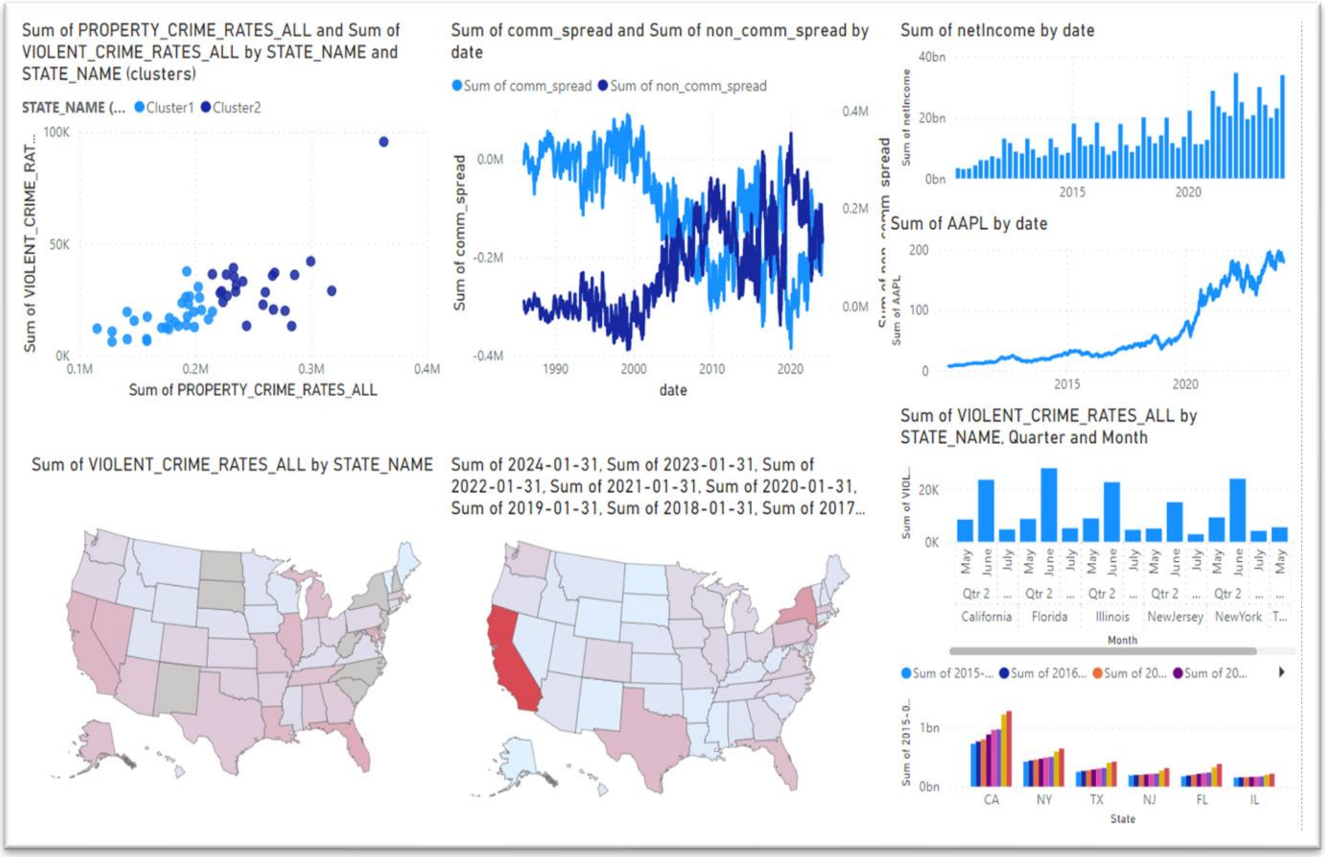
5. Data Analysis:

Data analysis techniques applied to the collected datasets, including sentiment analysis of news data, feature engineering on financial reports, and the integration of various datasets for deriving actionable insights. The methodologies used for data analysis and visualization are elaborated upon to provide a clear understanding of the analytical process.

- Correlation Analysis if the investor chooses to diversify investments.
- Exploratory data analysis to remove null values, duplicates.
- EDA was also useful in identifying data types of columns that were not representative of the data contained in that column.

6. Dashboarding with Power BI:

The creation of interactive dashboards using Power BI is detailed in this section. It covers the design principles, visualization techniques, and data storytelling aspects employed in developing insightful dashboards for stakeholders. Examples of key visualizations and their interpretations are provided to demonstrate the effectiveness of the dashboard in conveying actionable insights.



Sum of PROPERTY_CRIME_RATES_ALL and Sum of VIOLENT_CRIME_RATES_ALL by STATE_NAME and STATE_NAME (clusters)

Sum of comm_spread and Sum of non_comm_spread by date

Sum of netIncome by date

Sum of AAPL by date

Sum of VIOLENT_CRIME_RATES_ALL by STATE_NAME, Quarter and Month

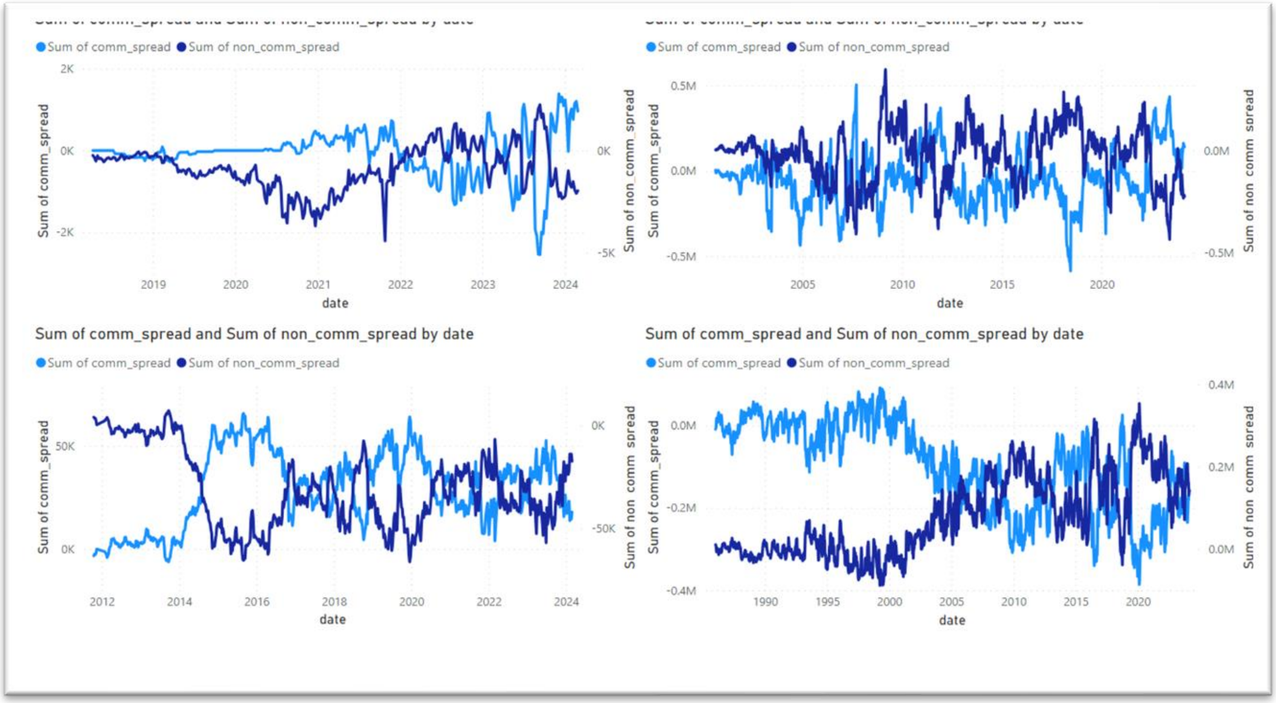
Sum of VIOLENT_CRIME_RATES_ALL by STATE_NAME

Sum of 2024-01-31, Sum of 2023-01-31, Sum of 2022-01-31, Sum of 2021-01-31, Sum of 2020-01-31, Sum of 2019-01-31, Sum of 2018-01-31, Sum of 2017...

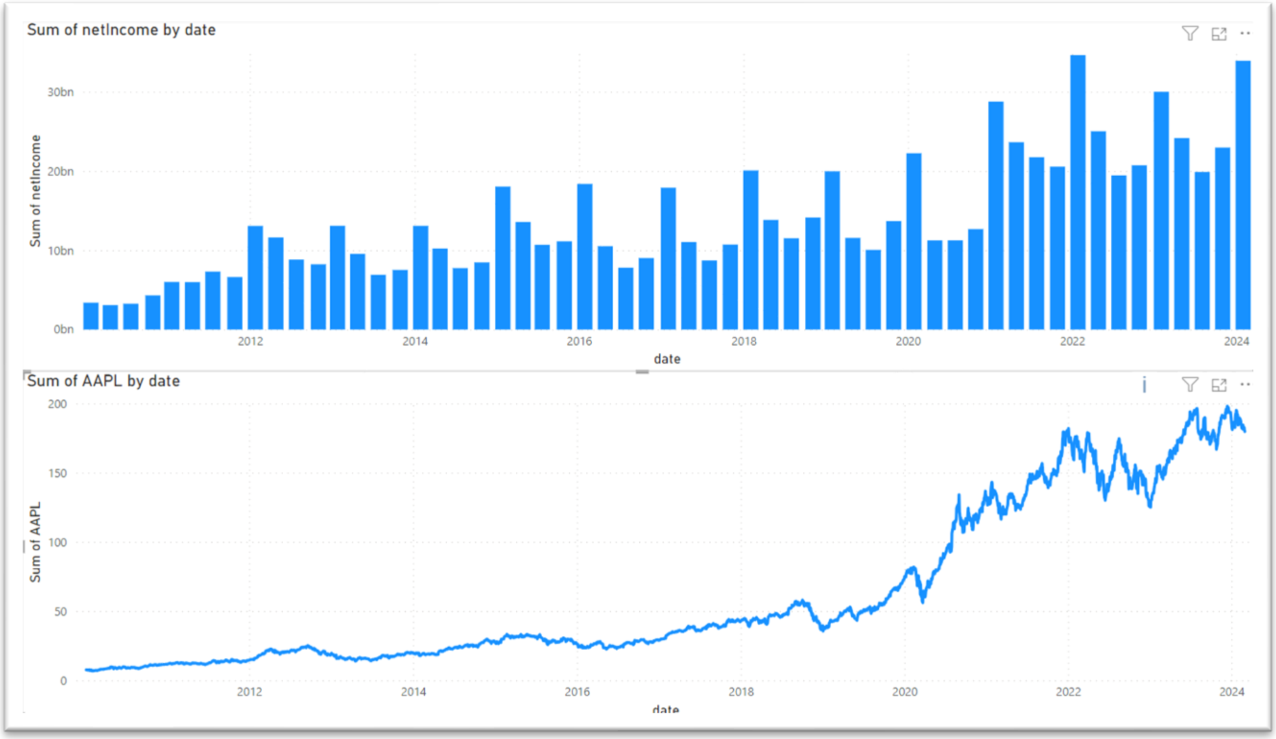
7. Results and Findings:

The results and findings of the data analysis are presented, highlighting significant trends, correlations, and insights derived from the analyzed datasets. Key findings related to stock market performance, financial indicators, news sentiment, are discussed, providing valuable insights for decision-making.

- Findings:
    - Commercial Hedgers and retail investors take positions that are generally opposite to each other, and market seems to agree with the commercial hedgers often. As evident from the samples taken from Bitcoin data, Oil, S&P 500, and gold data.

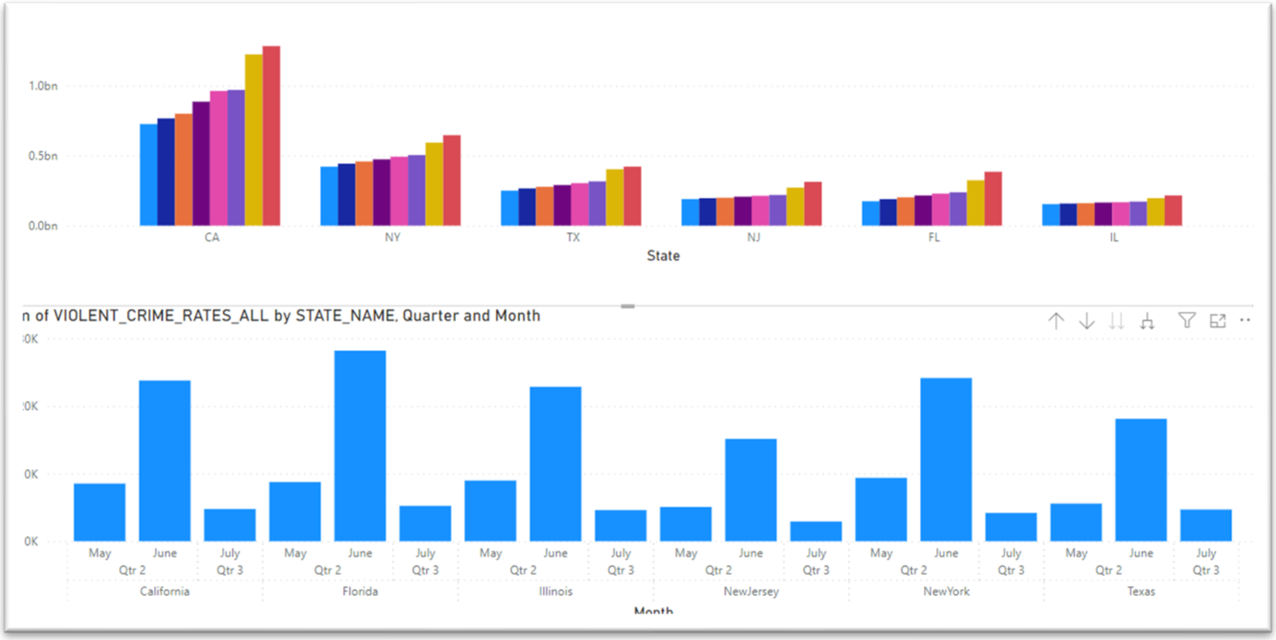


- Net quarterly income does have a say on stocks' performance for that quarter, as evident by our demonstration of AAPL's stock performance and quarterly net income.

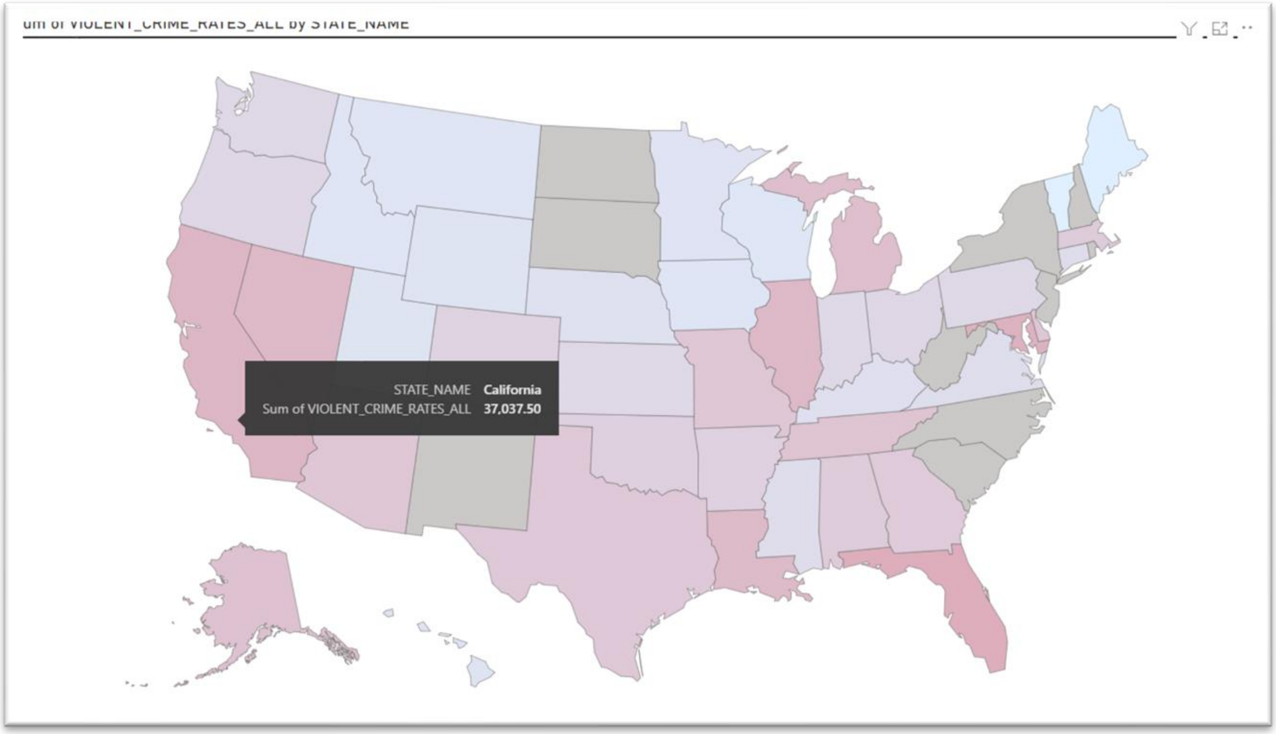


- Bitcoin and stocks of companies dealing with semi-conductors show strong correlation.

- Effect of Sentiment Analysis is in-conclusive, which aligns with “efficient market hypothesis.
- Most crimes are committed in the month of June, as displayed by the picture below:



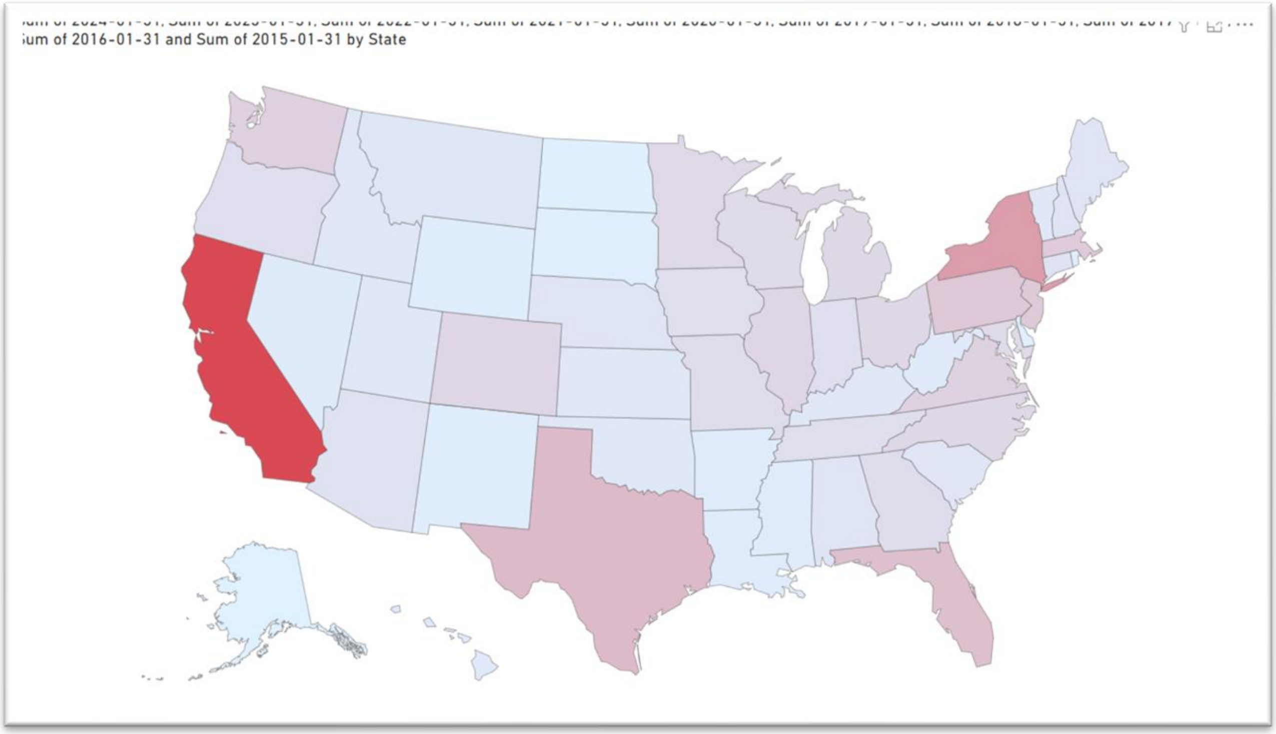
n of VIOLENT_CRIME_RATES_ALL by STATE_NAME, Quarter and Month

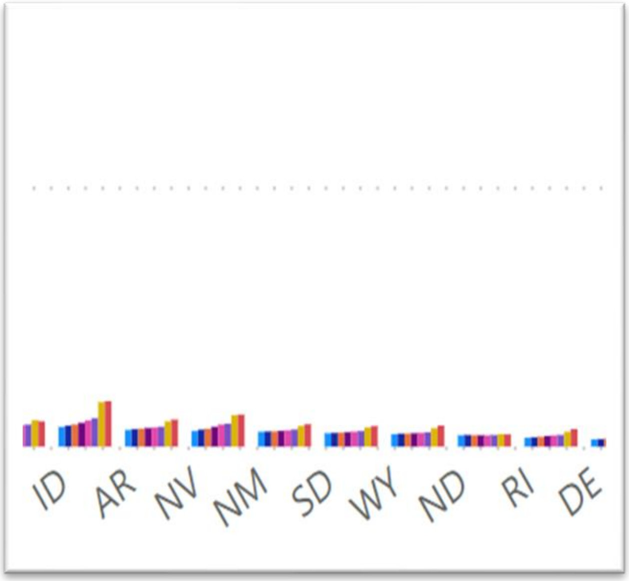


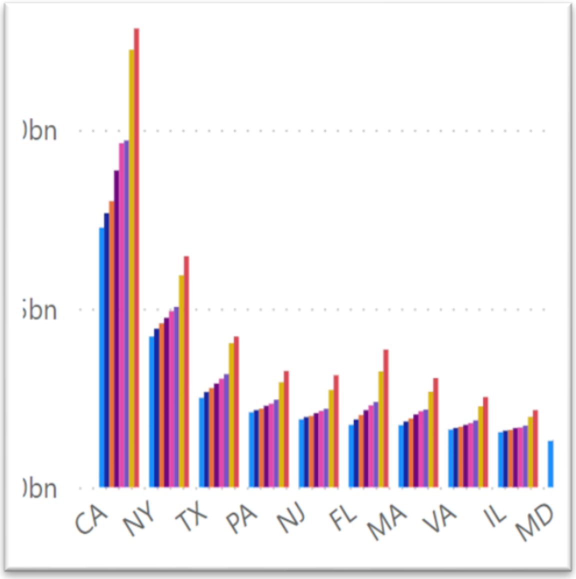
um of VIOLENT_CRIME_RATES_ALL by STATE_NAME

STATE_NAME California
Sum of VIOLENT_CRIME_RATES_ALL 37,037.50

- California has seen the highest appreciation in the housing market as opposed to “DC” where the market has stayed stable:


um of 2024-01-31, Sum of 2023-01-31, Sum of 2022-01-31, Sum of 2021-01-31, Sum of 2020-01-31, Sum of 2019-01-31, Sum of 2018-01-31, Sum of 2017
ium of 2016-01-31 and Sum of 2015-01-31 by State





8. Limitations and Challenges:

The limitations and challenges encountered during the project implementation are acknowledged and discussed. These may include data quality issues, technical constraints, and limitations of the analytical models used. Strategies for mitigating these challenges and areas for future improvement are also addressed. The issues we faced are mentioned below:

- Due to monetary constraints, we were unable to utilize the maximum benefits offered by spark.
- Our analysis was time bound because of our Azure Subscription expiration date.
- We took the data provided by the APIs at their face value with our assumption being that they are the source of truth.
- Finding the drivers for connecting spark to different databases was a challenge.
- Unable to have multiple clients working on the same spark server at the same time.

9. Conclusion:

To reiterate our workflow, the importance of data-driven decision-making in today's business environment and highlight the value added by the comprehensive data analysis undertaken. The importance of spark in distributed computing, azure with scalability and cloud computing cannot be overstated. We saw the importance of data analysis, and the hidden patterns discovered by it, firsthand. Recommendations for future research and development initiatives are provided to enhance the project's effectiveness and scalability:

- More monetary support.
- Integration of Deep Learning to forecast investment vehicles.
- Utilizing quantitative finance.

10. References:

- Yahoo Finance: https://ca.finance.yahoo.com/
- Financial Modelling Prep: https://site.financialmodelingprep.com/
- Finn-hub: https://finnhub.io/
- Investopedia: https://www.investopedia.com/