Project 2: ETL with Cryptocurrencies

Jacob Cortez, Julian Louden, Dan Tinarwo, Paola Mateo

SMU-Data Science Bootcamp

Overview

A brief look at what we will discuss on this presentation

- Purpose of Project

- Methods

- Potential Analysis

- Data Limitations

Purpose and Goals

Extract, Transform, and Load data

Data files were derived to observe trends of 3 cryptocurrency types (bitcoin, dogecoin, and ethereum) over the span of 3 years and imported into MongoDB

Could make future predictions

An analysis can easily be observed when data is imported into a database where all nine files may be accessed

View ongoing trends for a specific type of cryptocurrency

Keeping track of vital information can ensure the success of the company, and even the potential customer.

Guide on how Goal Was Reached

01

Collect Data Sources

Data sources were found on Kaggle.com with ~475,000 columns each dataset.

02

Data Cleaning

Files were imported into Jupyter by using the Pandas library. Null variables and unnecessary columns were deleted. Columns were renamed and reorganized.

03

Specific Data Transformation

Because the "timestamp" column was in epoch time, we changed it so it can have a readable date and time.

04

Load Data onto MongoDB

Once the CSV files were uploaded with the updated and relevant data, they were transferred to MongoDB in order to collectively view data at a time.
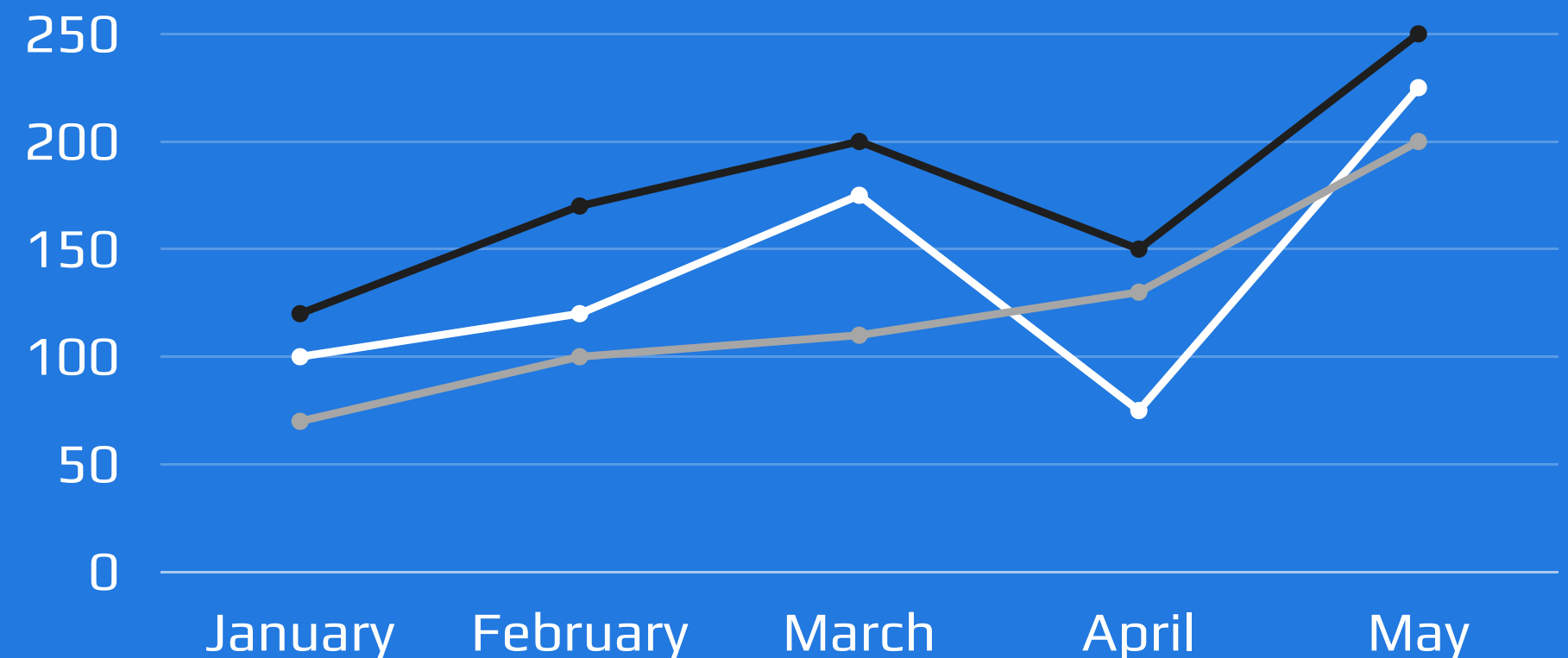
05

Filter Desired Information and Transfer Back to Jupyter

Specific data can be sent to Jupyter to view relationships between variables and make further analysis.

Data Limitations

Data Files Were Too Large

Once new data files were uploaded to MongoDB, attempting to push data back to Jupyter took roughly about 25-30 minutes, and sometimes even crashing the computer. Filtering is needed to further analyze data.

Inconsistencies in Timestamps

There are certain times where a specific crypto does not have trades. For example, Dogecoin may not have any trades for 10 minutes total one day, but trades for every minute the next day. This creates the data to be slightly inconsistent and difficult to create an accurate conclusion.

Conclusion

- 9 total CSV files were used for this project
- MongoDB was the database where CSV files were imported. We can load the data from this database to filter it out by the relationships that will be observed
- Potential analysis could include results regarding the best month for a trade in any of the 3 cryptocurrencies.
- Allowing room for error may be a solution to the listed data limitations.
- Overall, the purpose is to begin the process for a data analysis with different CSV files to create future predictions or data reports for ongoing trends.

Utilized Data Files:

Bitcoin Files:
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-bitcoin? select=full_data__1__2019.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-bitcoin? select=full_data__1__2020.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-bitcoin? select=full_data__1__2021.csv

Dogecoin Files:
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-dogecoin? select=full_data__4__2019.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-dogecoin? select=full_data__4__2020.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-dogecoin? select=full_data__4__2021.csv

Ethereum Files:
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-ethereum? select=full_data__6__2019.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-ethereum? select=full_data__6__2020.csv
- https://www.kaggle.com/yamqwe/cryptocurrency-extra-data-ethereum? select=full_data__6__2021.csv