Code files:

The manipulation was done step by step, some using pandas and some using spark
But it involved accessing the folders using a python script and copying the selected files to a new directory and hence this is not compiled into a single python file for execution on the cluster. They are present as ipynb files.

All the analysis tasks were each compiled into a python file, each named accordingly and were run on the Hadoop cluster to get the results.

Visualizations were done using matplotlib and cant be viewed while running them on the Hadoop cluster.
For this purpose, the whole code is also present as python notebooks.

Datasets:

Only the final selected data files have been attached in the zip file for submission.
The data required for the *Data Manipulation – Combining Datasets.py* and the three *Task1, Task2 and Task3* python files are added. But the complete 10 GB from which the files were selected in the *Data Manipulation – Stocks.ipynb* and *Data Manipulation-Crypto.ipynb* files is not included.

For the entire stock market and crypto dataset and all the files used in the is process, please visit
https://github.com/HareeshwarKarthikeyan/Correlation-of-Stock-Market-and-Cryptocurrency-Trading-Volumes-after-Covid19

I have set up my entire project with all the data there and by cloning it, you should be able to recreate everything done in the project.

The steps involved in this project are outlined in the workflow section of the report.

Note: The joined files for the indexes may already be present in the directory of that *particular stock index. Hence before running the code for joining files, make sure you delete the joined file. Else the newly joined file will contain the data of the already joined file as well as it is in the same directory.*