Data Collection & Pre-Processing

Group Assignment

Domain : Indian Stock Market

Team 2

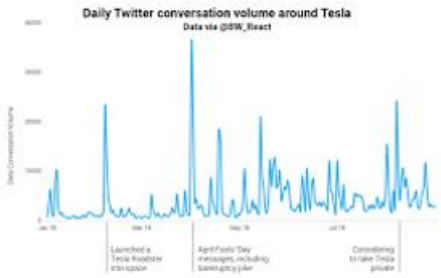
Abhishek Chintamani (PGID : 12120082)
Amit Shukla (PGID : 12120102)
Rupali Agarwal (PGID : 12120100)
Smaranika Sikdar (PGID : 12120092)
Varun Ananthula (PGID : 12120066)

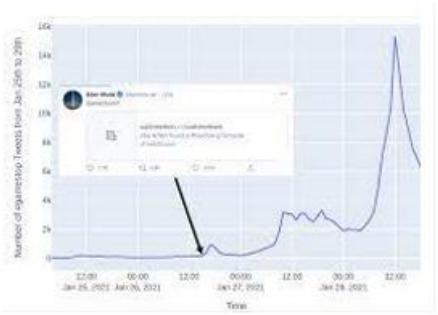
Project Github Link - https://github.com/ISB-Co-22/DCPP-Assignment

Table of Contents

1. Executive Summary

a. Problem Statement





No, our Data Collection project is NOT about Elon Musk!!!

But these articles about Musk are a good example of how social media is influencing the highly traditional stock markets globally.

We believe that with this increasing influence of social media, the traditional parameters for stock market analysis are no longer sufficient to efficiently track and predict stock market movements. We will have to update our data and analysis models to start including social media parameters to more closely correlate to the modern-day stock market. But just adding social media parameters is not sufficient - traditional parameters will still hold a big control over the stock market movement – we will need to find the correlation between the old parameters, new parameters and stock market overall.

To be able to do this, first we need to define a process for collecting traditional data from existing well defined sources – integrating key data from multiple sources. Then we will need to solve for defining social media parameters and setting up processes to efficiently collect data for these parameters. And finally we need to merge the traditional and the modern data parameters into a single data set for further analysis.

b. Proposed Solution

- Build a data collection model that efficiently and accurately handles traditional and new-age parameters for stock market analysis.
- For traditional parameters
    - Identify the seed source which has the most information available in an easily accessible format
    - Identify additional sources for enriching the traditional parameters
    - Setup the data collection process for these identified sources and the corresponding parameters from these sources
- For social media parameters
    - Identify the linkage between companies and their social media data – as part of this step finalize the platform(s) we want to track for this exercise
    - Define the data collection steps from the company data to the company's social media data
    - Finalize the key sources for each of the step and define the collection process for each source
    - Setup data cleanup and pre-processing steps to ensure minimal error rate in the data collected
    - Finally merge the social media and other new-age parameters into the traditional data

c. Challenges

- Traditional Data
    - One of the major challenges in collecting data through NSE or BSE is legal policies of exchanges. Automation or scraping or data mining directly through NSE or BSE official website is illegal
    - Due to stringent regulations, each report, if required from NSE or BSE directly, needs to be pulled out manually and this may increase work load of a daily trader.
    - If automation of report downloading is required, one may choose data feed vendors who are authorized by NSE or BSE. Some of them include Bloomberg, Ticketplant, Truedata, Ticker tape etc. where there is a nominal charge on monthly/yearly basis. Data from exchanges approved vendor is authenticated and reliable.
    - Part of the data used in this project is collected through these data vendors.
    - Also rules or definitions of certain attributes are updated regularly by exchanges, for instance promoter holding definition or change in lot size of index futures. Traders have to regularly monitor the notifications from exchanges so that they are aware of the current definitions of the key parameters. In case it is not monitored, attributes where definition is changed, may show 0 or blank, however in actual there would be some data.

- Social Media Data
    - Using social media parameters as part of stock market analysis is an emerging concept – especially in India. Most of the popular traditional sources of stock market data have not yet started tracking these new-age parameters. Even though social media data itself is abundantly available online, we had the challenge of finding the interlink between these data – i.e. mapping public listed companies to their social media data.
    - Companies too have not fully upgraded their online presence to easily identify and link their social media handles. For ex. Not all companies have included their Twitter handles info on their homepages and even if they updated it they have not done it in a consistent way.

2. Domain & Seed Sources

a. Chosen Domain

We chose the domain as “Indian Listed Companies” since we believed that this is one space which offers so much scope in terms of the sheer amount of data available and moreover so many kinds of analysis that can be done over that data.

India is a developing country, and the equity investment adoption is still it’s nascent stages and with the advent of automated trading and investments methodologies it becomes very important to have models being built on data backed principles to come up with Companies that can be trusted in terms of their return potential and quality of management. With such high volume of data being generated both on short term as well on a long-term basis for more than 5000 companies , it is not feasible to analyze manually and without any biases.

With this problem statement is mind, our group decided to select this domain and prepare a dataset of companies from both BSE and NSE with minimum of 10 Cr of Market Capital and being actively traded.

b. Seed Sources

Our seed sources can be broadly divided into two categories as explained in the subsequent sections

- **Structured** : These sources comprise of all the reliable or trusted sources from where we could gather Stock Related information in terms of all important Financial Ratios, Balance Sheet Data and Return data over the last few years.
  The following are the sources that helped us collect all required data that we needed to perform this exercise.
  We have described each of these in detail in the subsequent sections
    - Official NSE & BSE Web Portal
    - Value Research online , a Wealth advisory Portal

- Tickertape.in
- Bloomberg Quint

- **Unstructured :** These sources include sources from where we had to rely on automated ways of collecting the data through the established techniques of Web Crawling and scraping.
  We have described each of these in detail in the subsequent sections

  - MoneyControl.com
  - Twitter

Data from all these sources was gathered into separate worksheets within the excel file - Indian Stocks Dump.xlsx ,on which we performed the data cleaning and pre-processing.

3. Structured & Unstructured Sources

a. Structured Sources:

Structured data would form the very basis of this exercise that we undertook, and the idea was to build over this structured data. This would server as the main anchor of our analysis. Following are the main pre-requisites that had to be met before we finalize out structured data sources.

a) The data should be from a trusted and reliable source
b) The data should be easily retrievable
c) The data should have the required attributes and fields to drive some meaningful insights
d) The data should be consistent and streamlined in terms of data formats and types.

We divided the Data collection task amongst the Group members with the above-mentioned conditions and came up with the following sources

- Stock Selector | Value Research (valueresearchonline.com) VRO
  Value research online is more than a decade old wealth advisory portal and is one of the most trusted sources of historical Stock Data Information. We used the Stock Selector feature provided by VRO to select all stocks that have a market cap of more than 10 Cr with 35 different fields that helps us understand about the company and its performance in terms of Stock price historically.

Seed Data (4365 rows × 35 columns) [Indian Stocks Dump.xlsx -> Seed Stock Sheet]	
Column	**Description**
Company	Name of the Company
BSE code	Bombay Stock Exchange Code
NSE code	National Stock Exchange Code
ISIN	Unique Standard Code
Sector	Domain company works in
Industry	Industry company works in

Date	Data Collected on – 22nd April
Price	Day end price of the Share on 22nd April
52 Week Low	Lowest day end closing price in last 1 year
52 Week High	Highest day end closing price in last 1 year
3 Year Low	Lowest day end closing price in last 3 years
3 Year High	Highest day end closing price in last 3 years
5 Year Low	Lowest day end closing price in last 5 years
5 Year High	Highest day end closing price in last 5 years
All Time Low	Lowest day end closing price in since inception
All Time High	Highest day end closing price in since inception
Market Cap(Cr)	Current price of share x No. of free-floating shares
Enterprise Value(Cr)	Enterprise Valuation of the Company
1-Month Return	Absolute appreciation in % from the price a month back
3-Month Return(%)	Absolute appreciation in % from the price 3 month back
1-Year Return(%)	Absolute appreciation in % from the price 1 year back
3-Year Return	Absolute appreciation in % from the price 3 years back
5-Year Return(%)	Absolute appreciation in % from the price 5 years back
10-Year Return(%)	Absolute appreciation in % from the price 10 years back
Price to Earnings	Ratio of Current Price to last reported Earning
Price to Book	Ratio of Current Price to last reported Book Value
Price earnings to growth	Ratio of Current Price to Earnings Growth
Dividend Yield(%)	Dividend as a percentage of the current price
Price / Sales	Ratio of Current Price to last reported Sales
Price / Cash Flow	Ratio of Current Price to last reported Cash generated
Earning Per Share	Total Earnings / Total no.shares
Book Value Per Share	Total Book Value / Total no.shares
Cash Flow Per Share	Total Cash generated / Total no. of shares
Free Cash Flow Per Share	Total Free Cash generated / Total no. of shares
Dividend Per Share	Last declared Dividend per share

- Additional fields / attributes were sources from
  - Data related to company ISIN, Ticker name, Market capitalization & Total revenue (for last quarter) is available on the official NSE/BSE website:
    - BSE: All India Market Capitalization | BSE Listed stocks Market Capitalization (bseindia.com) ,
    - NSE: All Companies based on Market Capitalisation (nseindia.com)

  - Other data related to promoter holding, Total debt, free cash flow, 5Yr average return on equity, 5Yr historical revenue growth, Volatility and Stock Recommendations were taken from the website below.
    - Ticker Tape : Stock Analysis & Best Financial Tools for Indian Stock Market Evaluation | Tickertape
    - Bloomberg Quint: Stock Market News, Share Market Live Updates, BSE/NSE Live (bqprime.com)

Additional Stock Data (4626 rows × 11 columns) [Indian Stocks Dump.xlsx -> Seed Stock Sheet]	
Column	**Description**
Ticker	Display Name of the Company
ISIN	Unique Standard Code
5 Yrs. Avg ROE	Mean of 5 years Return on Equity
5 Yrs. Revenue Growth	Mean of 5 years Revenue Growth
Promoter Holding	% of Shares owned by Promoter
No. Of Shareholder	Total Number of Unique Shareholders

Pledged Promoter Holding	% of Promoter Shares pledged as Collateral to Banks / Creditors
Rating agency Buy Reco	No. of Buy Recommendations by Rating Agencies
Volatility	Measure of Price Movement Range
Total Debt	Total Debt on Company Balancesheet
Free Cash Flow	Total Free Cashflow generated by company

- Data source for the Nifty stocks is Investing.com.
  This source was chosen as the website that has the Nifty indices performance data of previous years for the comparison.We needed the Nifty 50 Index components historical data to see the trend of the stocks in Nifty 50 index.

Nifty50 (50 rows × 3 columns) [Indian Stocks Dump.xlsx -> Nifty50]	
Column	Description
Name	Display Name of the Company
ISIN	Unique Standard Code
Nifty 50 Stock	Flag for stock being part of Nifty Index

b. Unstructured Sources:

- DuckDuckGo Search
  - We found DuckDuckGo to have a good mix of returning **correct results** (hit rate) and supporting **bulk web scraping**
  - Needed to search for the company name on Moneycontrol website to get the company's Moneycontrol link
  - Tried the following options
    - Moneycontrol Website Search
      - Tried manually searching for company name on their website.
      - The hit rate was low (around 50%) and even for matched cases there was no simple url for replicating this search through code
    - Google Search
      - Google search had a very good hit rate (~98%) when we limit the searches using the 'site:' option to the specific Moneycontrol section / link.
      - But Google started blocking my web scraping requests after the first 50 – 70 calls and it was taking more than a day to unblock – hence we could not scale it for the 4.5K companies to be searched
      - Note that we did not explore rotating proxy solutions – as they were too costly and a bit advanced for this exercise.
    - Bing / Microsoft Search
      - Bing's hit rate was lower (~80%)
      - And even it was blocking web scraping requests within 50 – 70 calls
    - Less popular search engines
      - Tried various less popular search engines such as Dogpile, Ask.com and DuckDuckGo

        - Found a respectable hit rate with DuckDuckGo manual search (~95%)
    - DuckDuckGo
        - Using DuckDuckGo manual search we were able to get a decent hit rate of ~95% - had to do a bit of trial and error and tweak the search string
        - Also we were able to run higher number of requests without getting blocked during our testing phase.
- Moneycontrol Company page
    - Moneycontrol company pages had a good mix of **depth of info** and **ease of scraping** in terms of website structure and allowing bulk scraping.
    - We needed a source for online links for each company – specifically the company homepage (for links to their social handles) and their contact info (for future crowdsourcing steps)
    - Moneycontrol is one of the most popular consumer websites in India for information about public listed companies – so we started with Moneycontrol.
    - Since we were able to get the information we were looking for (and also there was lot more info available for future use) – we did not explore other options
    - Note that we are also storing each company's Moneycontrol page link in our data – so that in future if we need more details or updated stock market info we can directly scrape it from the respective Moneycontrol pages
- Company home pages
    - As mentioned in the Problem Statement – we wanted to include Twitter info in our data. This is to ensure that in addition to the traditional stock market data we also include some of the metrics that might be impacting the market in these modern times
    - Company home pages ended up being the ideal sources for the company's twitter handles due to **minimal errors**
    - Tried the following options
        - Twitter search
            - Direct twitter search using company name resulted in too many results for each search and there was no easy way to identify which of the results is the actual handle
            - Especially when most of the results were handles of aggregators and media companies who were reporting news about the company we were searching for.
            - Hence we could not use twitter search
        - Google search
            - Faced similar issues by searching for twitter handles through Google search.
            - Though the quality of the results improved (in terms of company name matching) – but here too there were too many results from aggregator and media company handles
        - Aggregators (Moneycontrol, Valueresearch, etc.)
            - We tried to identify if any of the aggregators provide social handles info for these companies

        - But we could not find this info on popular websites such as Moneycontrol and Valueresearch
        - Our sense is that the traditional sources for stock market info have not yet started tracking modern metrics such as # of Twitter Followers or recent tweets from the company regarding their performance, etc.
    - Company Homepages
        - During this exercise we realized that not all companies were using twitter to share their company updates – especially true in case of small and mid-sized companies in B2B sectors
        - We also noticed that companies that have a twitter account have a high chance of having a good homepage and mentioning their twitter handle on the homepage
        - With a pilot manual search we found a hit rate of ~20% and more importantly for the remaining 80% we could not find an easy way to identify their twitter handles
- Twitter
    - We identified that Twitter is a platform with a good mix of **business relevant info** and **high consumer (and investor) reach**
    - As part of identifying modern metrics that could be correlated to the stock market we wanted to track relevant social metrics such as presence and rate of activity on social media.
    - Tried the following options
        - LinkedIn
            - Very high relevance in terms of business content
            - But it is primarily targeted to working professionals and not really designed to reach a high number of consumers and investors
        - Facebook
            - Very high reach among consumers and investors
            - But low relevance in terms of business content
        - Twitter
            - This is where we found Twitter to be the ideal mix between business relevant info and good reach among consumers and investors
            - Companies worldwide are already using Twitter as a modern route for business communication (in addition to the traditional routes of press releases, etc.)
            - Once we identify the Twitter handle of a company, the info available on this handle is highly relevant for that company – there is no noise or incorrect data

4. Data Collection

a. Downloaded Data

Structured data was downloaded and collected from the following locations

All this structured data was finally organized into a spreadsheet “Indian Stocks Dump.xlsx”

Source	Excel Worksheet (Indian Stocks Dump.xlsx)
Stock Selector | Value Research (valueresearchonline.com) VRO	Seed Stock Sheet
All India Market Capitalization | BSE Listed stocks Market Capitalization (bseindia.com) , All Companies based on Market Capitalisation (nseindia.com) Ticker Tape : Stock Analysis & Best Financial Tools for Indian Stock Market Evaluation | Tickertape Bloomberg Quint: Stock Market News, Share Market Live Updates, BSE/NSE Live (bqprime.com)	Additional Stock Metrics
Investing.com.	Nifty50

All the sources mentioned above supported the provision of downloading the data in a structured form (csv, excel) but there were still some challenges faced across like

- Challenges
  - Not all the required fields or attributes were found in a single location or source. So all the Financial ratios, Return data over past 1,3,5,10 years were found from the Stock Selector feature of Value research Online but fields like Promoter holding, No. of shareholders & pledged promoter holding data is taken from NSE/BSE website directly.
  - If automation of report downloading is required, one may choose a data feed vendor who is authorized by NSE or BSE. Some of them include Bloomberg, Truedata, Ticker tape etc. where there is a nominal charge on a monthly/yearly basis. Data from exchanges approved vendor is authenticated & reliable.
  - Another major challenge in scrapping data through NSE or BSE is legal policies of exchanges. Automation or scraping or data mining directly through NSE or BSE official website is illegal. Due to stringent regulations, each report, if required from NSE or BSE directly, needs to be pulled out manually & this may increase work load of a daily trader.

b. Crawled Data

- DuckDuckGo Search
    - We first explored a python library for DuckDuckGo search – it had already integrated the technical aspects of making the search request and returning the required number of results.
    - But hit rate dropped to 70% vs 95% hit rate during manual testing. Same searches manually were giving correct results.
    - On deep dive understood that the library is using DuckDuckGo API and it is clearly mentioned by DuckDuckGo that their API does not yet search the full web (by design)
    - So we used DuckDuckGo search url directly for web scraping
    - After a bit of trial and error and tweaking the search string we were able to get to the hit rate of ~95%
    - A few challenges we faced and the corrective measures taken to overcome them were
        i. **Valid search header:** Got blocked in the first few attempts – realized that we had not setup the search header to ensure request is detected as that from a valid web browser. Had to wait for a day and re-run with the correct search header
        ii. **Sleep timer between searches:** We were getting blocked in Google and Bing searches after a few attempts – to ensure lower chance of blocking on DuckDuckGo we introduced a sleep timer of 30 secs after each search. Had to setup a dedicated system to run this for a few days for 4.5k companies
        iii. **Handle exceptions:** Since we were limiting the search to a section on Moneycontrol website in some cases it did not return any results and was resulting in the code crashing. Similarly we got errors when the webpage was not responding etc. Added the key code in try-except block to handle exceptions cleanly.
        iv. **Write to file periodically:** If we were not writing to file and saving it periodically, any error was resulting in data loss and we needed to run the code again.
        v. **Backup periodically:** After the first day's run the code covered 500 rows – but we were unable to open the data file - mostly got corrupted during a write operation or a system issue. We started taking manual backups every few hours.
    - The corresponding code for this can be referred to at –
        i. TwitterMoneyControlDataExtractor.py – searchOnDuckDuckGo() – Line 66
- Moneycontrol Company page
    - We used direct web scraping for collecting homepage and email info from Moneycontrol company page.
    - Two main reasons for this -
        i. Getting a direct link for the company's Moneycontrol page in the previous step with a 95% hit rate
        ii. Moneycontrol company pages are very cleanly and consistently structured

- One key challenge we faced at this step was errors due to info not being available for specific companies or when the link from the DuckDuckGo step is not a correct Moneycontrol company page.
- To handle such cases we used try-except blocks to handle exceptions cleanly.
- The corresponding code for this can be referred to at –
    i. TwitterMoneyControlDataExtractor.py – getWebsiteFromMoneycontrol() – Line 86

- Company home pages
    - We first did a manual review of page sources of 50 random company home pages to identify the pattern for getting Twitter handle from the home page.
    - Post that we wrote the code logic and did a test run on 200 companies and reviewed the twitter handles picked by code
    - Some of the challenges faced during this step were -
        i. **Fine tuning the scraping logic** - We had to find a fine balance between getting 'a' twitter handle vs. getting 'the' correct twitter handle.
            1. For ex. If we added a unique logic that would pick the correct handle for one company and ran it for remaining companies – it started picking incorrect handles for remaining companie
        ii. **Iterative and phase-wise process** – This was truly unstructured data – every company had their own way of mentioning their twitter handles (if mentioned at all). Hence we had to run this step in a phased manner and keep iterating to get the most efficient hit rate
        iii. **Handling Exceptions** – In this step there were too many unique exceptions due to the varied structures of company home pages. It was not sufficient to just catch the error and move on – during the iteration steps we reviewed the key errors to understand if we can fix the issue or if we had to skip this company
    - The corresponding code for this can be referred to at –
        i. TwitterMoneyControlDataExtractor.py – getTwitterHandleFromHomepage() – Line 116

- Twitter
    - We used Twitter API to get the Twitter info from company twitter handles. Two key reasons we went ahead with this approach –
        i. Twitter's free API supports upto 900 summary calls per day (i.e. calls to get summary info of a Twitter handle). Hence we were able to handle our volume easily with the free Twitter API
        ii. By using Twitter's approved method we did not have to worry about getting blocked and trying to figure out alternatives to
        iii. Clean format of data returned through the API
    - There was only one major challenge we faced at this step – setting up the Twitter API (requesting permissions etc.) and learning how to use the API in Python
        i. Twitter being a popular social platform, lot of info was available online to help with this step.
    - The corresponding code for this can be referred to at –
        i. TwitterMoneyControlDataExtractor.py – getTwitterUserInfoForAll() – Line 132

5. Data Conversion from Original Sources

a. Webpages

- There were broadly 3 processes or strategies we had to follow for data collection and conversion from webpages to structured data fields based on the nature of the webpage
- **Highly Structured Content**
  - When the webpage had a highly structured and consistent format, we used **BeautifulSoup parser** and extracted the data based on the 'class' info.
  - Ex. DuckDuckGo search results – every search query returned data in the same structure and the required data was returned as part of the class 'result__url'
- **Moderately Structured Content**
  - When the webpage had structured content but there were inconsistencies between returned content in terms of classes used and the level or depth of required data – in such cases we directly used **string manipulation on page source** to extract the relevant info
    - In these cases libraries such as BeautifulSoup were limited by their need for a consistent 'class' or 'attribute' structure
  - Ex. When scraping Moneycontrol company pages for company info (website link, email ID)
    - The 'class' structure varied slightly based on the company we were scraping (based on sector or age of the company)
    - But the visual and textual format of the webpage was consistent for all companies
    - So we directly used string manipulation on the page source to extract the company info using visual landmarks on the webpage
    - For company home page we searched for 'Internet' – which referred to the Internet label on the website.
- **Unstructured Content**
  - When there is no structure or consistency in the webpages that we are scraping we used string manipulation on the page source – we iterated this process on a few sample webpages to **fine tune the string we were searching for**
  - Ex, When we were scraping individual company home pages for their Twitter handle, each company home page had their own structure and and their own attributes or classes for listing the Twitter handle
  - We had to run multiple iterations to fine tune the string we were searching for and finally used https://twitter.com/ to identify Twitter handles mentioned on company home pages

6. Data Cleaning & Pre-Processing

Once we had all the data in the Structured Format, the following are the main spreadsheets available to us

- Seed Stock Sheet (Originally Structured)
- Web & Social Data (Transformed into Structured data from unstructured web sources)
- Additional Stock Metrics & Nifty 50(Originally Structured)

a. Structured Data

The following are the main steps that were executed using Python scripts (MergeHandler.py) to merge, clean and pre-process the data

i. **Merge** : Firstly, the core Seed Data had to be merged to the "Web & Social Data" and this was done using the common ISIN attribute. Since the Web and Social (twitter) data scrapping was done using the very same ISIN and Stock Names from the core sheet, this was a straightforward exercise.

ii. **Obsolete ISIN:** It was found that some stock records had old ISIN values and since we will be relying on ISIN value to do all merges with Additional Stock Metrics sheet, we will have to update such records with latest ISIN values. For this purpose we downloaded the latest Stock vs ISIN mapping sheet from NSE and did merge with the Seed Sheet to update those records (Refer to python code with comment – *# Map the obsolete ISIN with latest ISIN*")

iii. **Data Cleansing** : There were some Stocks that had NULL ISIN values and those had to be removed from the collection (Refer to python code with comment – *"# Drop NaN ISIN values from the Merged Sheet Data*")

iv. **Data type Casting** : The Additional Stock Metrics sheet had the 'Ticker' and 'ISIN' as object data type and had to be converted into 'String' type (Refer to python code with comment – *"#Cast Ticker & ISIN to String column types*")

v. **Clean Duplicate Data** : There were records with duplicate Ticker names and duplicate and Null ISIN values which had to be removed (Refer to python code with comment – *"# Drop duplicate Tickers / ISIN from the Additional Sheet"*)

vi. **Keep Stock Data only** : In the Additional Stock Metrics sheet there were some ETF (Exchange Traded Fund) , NCD and other tradeable entities which were in scope for our Stock only exercise and hence were removed as well. The naming format of ISIN (Stocks start with INE*****) helped us in getting rod of non-stock data (Refer to python code with comment – *# Remove the Funds & ETF data (starts with INF***** instead of INE***)from the Additional Data Sheet*)

vii. **Merge with Additonal Stock Metrics Sheet :** Post the clean up on both the Seed Sheet and Additional Stock Metrics, it was time to merge both (Refer to python code with comment – *# Merge the Additional metrics data with the merged data frame (seed and Web)*)

viii. **Non-trading Stocks** : Since we are interested only in the Actively traded stocks, post the merge from the Step 7 above, all the stocks which had 'Company' field as NULL (no matching record in Seed Sheet) were deleted (Refer to python code with comment – *# Cleanup Stocks that are not actively traded. These are the stocks that were not found in the Seed Sheet*)

ix. **Unreliable Stock Data** : For any stock to be evaluated for any kind of fundamental analysis, it is important to know at least the Earnings , Book Value and Cash Flow metrics. So for all the stock for which any of these value was missing had to be removed since those stock data can't be very credible (Refer to python code with comment – *# Remove the Stocks for whom the Earnings & Book Values is not available as it would be difficult to evaluate such companies*)

x. **Missing PE ratio** : For some of the stock the P/E ratio was found to be missing although the Current Price and Earnings per share were available. So we computed the PE ratio using those two fields and updated the missing values. (Refer to python code with comment – *# Fill in the missing values of Price to Earning given that the Price and EPS info is already present*)

xi. **Date Standardization** : The Datetime fields ('Date' and 'Twitter creation Date') were standardized to Datetime format (Refer to python code with comment – *# Convert Data Collection date and Twitter Handle creation date to standard datetime*)

xii. **Numeric Standardization:** All the numeric fields were rounded off to 2 decimal digits (Refer to python code with comment – *# Round off the Numeric values to a standard 2 decimal digit format*)

xiii. **JSON / Excel Creation** : Finally, the entire pre-processed dataset was transformed into a JSON and a Excel

a. **StockRefinedData.json**
b. **Final Processed Stocks Data.xlsx**

b. Unstructured Data

- As data from Web scraping is unstructured, it is prone to high number of errors and lot of noise. Hence it is important to have a clear and efficient process for Data Cleaning and Pre-Processing. For this exercise we have implemented the following key processes –
- **Clear process flow and strategy for data cleaning:**
  - We defined a clear process flow for our data collection process

    i. Web Search > Moneycontrol Company Page > Company Homepage > Twitter API
  - Identified the 'weak' areas in this process flow – i.e. steps with a high chance of data errors and noise
    i. Web Search - dependent on DuckDuckGo's results for a given company)
    ii. Company Homepage – high chance of error in identifying the right Twitter handle from a company homepage
  - Define a structured review and cleanup process for the 'weak' data sources
- **Algorithmic noise identification:**
  - We defined the following algo / logic based noise and data errors identification process that will cover the big chunk of errors
  - Number of repeated entries
    i. For Moneycontrol company page links and Twitter handles – since these should be mostly unique for each company (except subsidiaries). We reviewed and cleanup all duplicate entries starting from highest number of duplicates to lowest
    ii. We found some obvious errors in these cases – for ex. Whenever a company page was not found through web search it was returning a generic Moneycontrol link – this was repeated for 30+ companies
    iii. Similarly, few companies did not have their twitter handles on their homepage but included generic twitter links (share, tweet, etc.) and these were picked up multiple times
  - High values
    i. For Twitter info (# of followers, # of tweets, age of account), we reviewed the top 10 – 15 values to ensure correctness
    ii. For ex. Few of the companies with very high followers had been incorrectly tagged with Twitter's twitter handle, etc.
- **Visual dipstick process:**
  - Once the major data errors were cleaned up through the algo based cleanup process, we setup a visual dipstick process to identify any repetitive errors not caught in the previous step
  - Randomly selected 100 companies and did a quick visual review of the data – especially needed for the Twitter handles data as it was the most unstructured
  - Found issues in twitter handles where the actual handle is correct but we had also picked up special characters and additional links – based on the dipstick we ran a cleanup code on the Twitter handle data to remove special characters and additional links.
- **Decide on when to stop the cleanup process**
  - When we are dealing with bulk data (~5,000 rows X 30+ cols) – especially including unstructured data – it is not possible to ensure zero errors. The only way we can ensure this is to do a deep review of each and every cell in the dataset – hence it is very costly, sometime costlier than the data collection process itself.
  - We will have to decide at what stage the cost of reviewing and cleaning up further is too high
  - In our exercise we decided to stop the data cleanup process once

i. we reached an error rate of ~4% (on a random sample of 100 rows)
ii. and each error was specific to the company's home page and we could not extrapolate the same logic to remaining companies

7. Observation & Insights

Some of the key insights from our data collection exercise are –

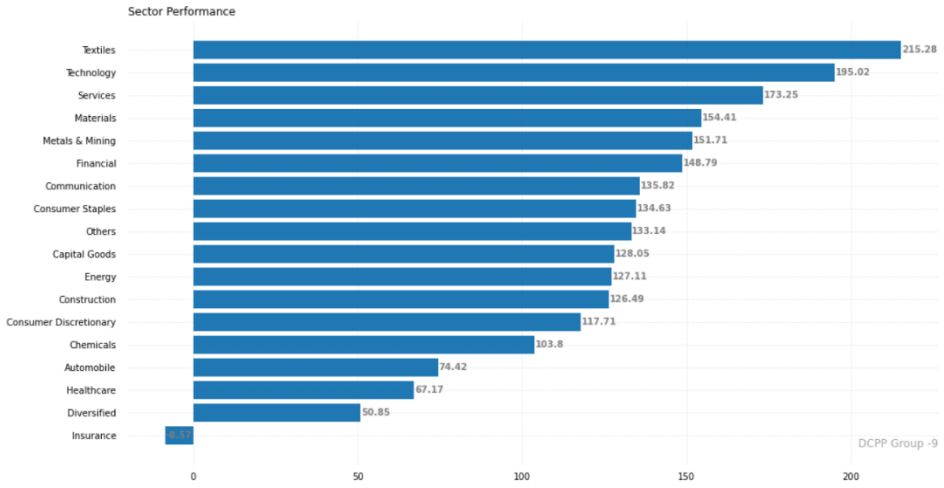
- **Identify the right set of sources**
  - During the initial planning stage lot of thought and research should go into identifying the ideal sources based on key points such as –
    - Depth and Width of data (Quantity of relevant data)
    - Quality of data
    - Ease of access
    - Long term availability of the source
- **Use efficient data collection processes** – while deciding the data collection process (download vs web scraping vs API, etc.) for each source, take care of the following parameters to ensure an efficient and scalable process –
  - Give preference to clean and legal methods of collection
    - Ex. API vs scraping – if API is cost competitive and provides all the required data we should use APIs.
    - While doing cost analysis also consider the cost of delays and data errors when using scraping.
  - Use scalable methods
    - Ensure the process can support bulk volumes and can be run periodically to keep the data up-to-date

- **Use algo based data cleanup processes**
  - Design the data cleanup process around the unique requirements of your data collection flow
  - Identify "weak" points in your process and setup review and cleanup processes around these points
  - Find a balance between the following methods for error identification –
    - Algo / logic based – ideally should handle the bulk of the errors in the data
    - Visual review – quick visual review of 10% random sample of the data to identify trends for error
    - Deep review – Detailed review of 3% - 5% of your data to get a final sense of data quality
- **Store intermediate data**
  - This will help with the review process and will also support future data enrichment.
  - For ex. When collecting Company homepage links from the Moneycontrol source we also stored the company's Moneycontrol link
  - This helped in reviewing if the correct Moneycontrol link was picked up

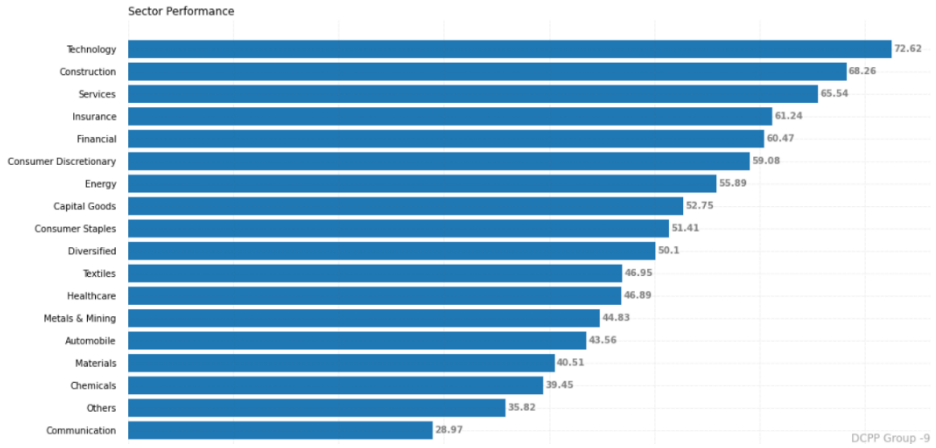
  - And also if in future we would like to pick more data from the Moneycontrol page of a company (including live stock price data) we can directly use the link in our data.
- **Ensure clean linkages between multiple data sources**
  - When taking data from multiple sources ensure there are clean linkages between these sources (i.e. primary keys) which will help merge the data among sources
  - For ex. When merging stock market data from two sources we ensured that we included ISIN in both the sources and used it to merge
  - Similarly when picking the list of company names for the Twitter Info search we used the company names directly from the Seed source.

Based on the data we collected, below is a sample analysis and key findings to help understand the importance of this data –

- The stock market is considered as a sentiment indicator & can impact gross domestic product (GDP). National stock exchange (NSE) recorded world's largest derivatives exchange by volume.
- For a regular stock market trader, it is necessary to analyze a lot of data to take a final decision.
- Hence it is necessary to have all data consolidated in one place & our project emphasizes this particular aspect.
- Below is sample of one such analysis:



- If we need to analyze sector-wise performance, we use our database & plot 1-yr avg equity returns against sectors.
- Above graph indicates Textile sector has out-performed all sectors in last one year & insurance sector is underperforming
- Below graph indicates average earnings of each sector. Even though price-wise textile sector is outperforming, same is not true in terms of earnings.



- Decision on buying a stock shouldn't be dependent on one criteria only. There should be a complete 360deg analysis, as shown in sample & then final decision should be taken.

8. Enhancing data with Crowd Sourcing Methods

- We plan to use crowdsourcing data strategies to enrich the Twitter info of companies, in terms of Twitter account details and usage details.
- We will use a two-pronged approach by reaching out to Companies and Investor community
- **Company Reachout –**
    - As part of our data collection exercise we have already collected the contact info (Email IDs) of all the companies.
    - We will design a survey form and email to be sent to the company email IDs.
    - Key info to be collected from the companies is –
        - Twitter Handle
        - Other Social media handles they use such as LinkedIn, Facebook
        - Do they use Twitter to connect with their investor community
        - Do their CXOs use Twitter to connect with their investor community
        - Share the twitter handle of your most active CXO
        - How often do they post their company updates on Twitter
            - More than Once a month
            - Once a month
            - Once a quarter
            - Less than once a quarter
        - How often do they review their mentions and respond
            - Daily
            - Once a Week

        - Once a Month
        - Less than once a Month
    - Responses to this survey will help cleanup the Twitter handles data in our data set and add additional qualitative data regarding Twitter usage by these companies
- **Investor Community –**
    - We will reachout to the investor community on popular stock market info portals (Moneycontrol, Value Research, etc.)
        - This will be a good sample set of tech savvy investors who are already using online info
            - There will be a higher chance of them responding to the survey
            - And also there is a higher chance of these users using Twitter and other social media for their investment process
        - The key info to be collected from them is –
            - Do they use social media to help with their stock market investment process
            - How often do they login to Twitter to read about stock market or company info
            - What are the top 5 companies that they closely track on Twitter
                a. Company Names in drop down and an option to fill the twitter handles for each
            - Any suggestions to companies to better connect with investors on Twitter ?
        - This info will help in two ways –
            - Validate some of the twitter handles that we already have in our data
            - Get qualitative info around investor behavior related to using social media for stock market investment decisions.

9. References & Sources

- Stock Selector | Value Research (valueresearchonline.com) VRO
- https://www.nseindia.com/all-reports
- https://www.nseindia.com/nse-terms-of-use
- https://www.bseindia.com/markets/equity/eqreports/AllIndiamktcap_Histori.aspx
- https://www.nseindia.com/regulations/listing-compliance/nse-market-capitalisation-all-companies
- https://www.tickertape.in/
- https://www.bqprime.com/quint
- Moneycontrol.com
- DuckDuckGo.com
- Twitter.com
- https://developer.twitter.com/en/docs/twitter-api
- https://realpython.com/beautiful-soup-web-scraper-python/

END OF DOCUMENT