NSE Fundamental Insights: Fundamentals (2019-2022)

Anant Kacholia

April 2024

https://github.com/ANANTKACHOLIA/NSE-Fundamental-Insights-Fundamentals-2019-2022-.git

1 Introduction

The dataset under examination, titled "NSE Fundamental Insights: Fundamentals 2019-2022," encapsulates a comprehensive collection of fundamental data pertaining to companies listed on the National Stock Exchange (NSE). Compiled over a span of four years, from 2019 to 2022, this dataset offers a rich repository of financial metrics and indicators essential for analyzing the performance and health of companies operating within various sectors of the economy.
With over 1800 individual stocks encompassing diverse sectors and segments, this dataset provides a granular understanding of the financial landscape within the NSE. Each stock entry includes a myriad of fundamental parameters, ranging from Price Earnings Ratio (P/E) and Earnings Per Share (EPS) to Return on Equity (ROE) and Debt-to-Equity Ratio, among others.
The dataset not only offers a snapshot of companies' financial standing at specific points in time but also facilitates the exploration of multi-year trends and patterns, enabling analysts and investors to gain insights into the evolution of companies' financial performance over time.
In the following sections of this report, we will delve deeper into the dataset, to uncover hidden patterns, relationships, and actionable insights that can inform investment decisions and strategic planning within the dynamic realm of financial markets.

2 Detailed Description

Motivation

PURPOSE OF CREATION:
The creation of the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset was driven by the need to provide comprehensive and granular insights into the financial performance and health of companies listed on the National Stock Exchange (NSE). Several key motivations underlie the creation of this dataset:

1. **Investment Analysis and Decision-Making:** One of the primary purposes of compiling this dataset was to facilitate investment analysis and decision-making for investors, analysts, and financial professionals. By offering a wide array of fundamental metrics and indicators for over 1800 stocks across various sectors, the dataset enables stakeholders to assess the financial standing and growth prospects of individual companies, identify investment opportunities, and mitigate risks.

2. **Predictive Modeling and Analytics:** Another motivation behind the dataset's creation was to support predictive modeling and analytics efforts within the financial domain. By providing multi-year data spanning from 2019 to 2022, the dataset offers a rich source of historical information that can be leveraged to develop predictive models for forecasting stock returns, estimating future financial performance, and assessing mar-

ket trends.

3. **Research and Academic Purposes:** The dataset serves as a valuable resource for academic researchers and scholars conducting studies in the fields of finance, economics, and data science. Researchers can utilize the dataset to investigate various phenomena, such as market dynamics, sectoral trends, and the impact of fundamental factors on stock prices, contributing to the advancement of knowledge in these domains.

4. **Market Transparency and Information Accessibility:** By consolidating fundamental data from publicly available sources and making it accessible in a structured format, the dataset promotes market transparency and enhances information accessibility for market participants. Investors and analysts can access standardized financial metrics for a wide range of companies, fostering a more informed and efficient market environment.

Overall, the creation of the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset aimed to address the need for comprehensive, reliable, and readily accessible financial data, empowering stakeholders with the tools and insights necessary to navigate the complexities of the financial markets effectively.

Composition

INSTANCES REPRESENTATION:
The instances in the dataset represent individual stocks listed on the National Stock Exchange (NSE). Each instance corresponds to a specific company within various sectors of the economy, such as automotive, technology, healthcare, etc. The dataset provides a comprehensive overview of fundamental metrics and financial indicators for these stocks, enabling analysis of their performance and health in the financial markets.

Furthermore, the dataset is organized into four folders, each corresponding to a specific year (2019-2022). Within each year's folder, there are multiple CSV files, each representing a sector or segment of the economy. These CSV files contain data for the respective stocks within each sector, along with their features.

TOTAL NUMBER OF INSTANCES:
The dataset comprises over 1800 instances, with each instance representing an individual stock listed on the NSE. These instances are distributed across different sectors and segments, providing a diverse representation of companies within the Indian stock market. There are 4 directories each for a year, each of them contains 18 CSVs, one for each sector.

DATA REPRESENTATION:
The dataset contains a comprehensive collection of instances representing individual stocks listed on the National Stock Exchange (NSE). While it may not include every single stock listed on the NSE, it covers a significant portion of the market, providing a representative sample of companies within the Indian stock market.

The sample is representative in terms of sectoral coverage, as it includes stocks from various sectors of the economy such as automotive, technology, healthcare, etc. This representativeness was ensured by including data from multiple sectors and segments within the NSE.

However, it's important to note that the dataset may not be entirely comprehensive, as certain stocks or sectors may be underrepresented or excluded due to various factors such as data availability, reporting practices, and dataset curation considerations.

DATA CONTENT:
Each instance in the dataset consists of features representing various fundamental metrics and financial indicators for individual stocks. These features include but are not limited to:

- Price Earnings Ratio (P/E) - Earnings Per Share (EPS) - Return on Equity (ROE) - Debt-to-Equity Ratio - Revenue from Operations - Profit Growth - Cash Earnings Retention Ratio - Market Capitalization - Enterprise Value - And many more

These features provide valuable insights into the financial performance and health of the companies represented in the dataset, enabling analysis and decision-making within the financial domain.

LABELS AND TARGETS:
Yes, there are labels or targets associated with each instance in the dataset. Two primary targets are provided:

1. Returns: Represents the returns generated by each stock in the respective year. 2. Next Year's Close: Represents the closing price of each stock in the following year, providing a forward-

looking target for predictive modeling and analysis.

These targets serve as key variables for assessing the performance and predicting the future behavior of the stocks within the dataset.

MISSING INFORMATION IN INSTANCES:
Yes, some details may be missing from individual instances in the dataset. This missing information could be due to various reasons, including but not limited to:

- **Unavailability of Data:** Certain fundamental metrics or financial indicators may be missing because the relevant data was unavailable or not reported by the respective companies. This could occur due to factors such as incomplete financial reporting or data collection limitations.

- **Legal Considerations:** In some cases, information may be intentionally removed or redacted from the dataset due to legal reasons or privacy concerns. This could involve sensitive financial information or proprietary data that cannot be disclosed publicly.

- **Data Curation:** The dataset may undergo a curation process where certain information deemed irrelevant or redundant is excluded from individual instances. While not technically missing, this information is intentionally removed as part of the dataset preparation process to streamline analysis and ensure data quality.

EXPLICIT RELATIONSHIPS:
In the dataset, relationships between individual instances are not explicitly made explicit. Each instance represents a standalone observation of a specific stock listed on the National Stock Exchange (NSE), and there are no explicit connections or links between different stocks within the dataset. However, implicit relationships can be inferred based on sectoral classifications and industry affiliations. Stocks within the same sector or industry may exhibit similar financial characteristics or performance metrics due to shared market dynamics and competitive factors.

RECOMMENDED DATA SPLITS:
The dataset is divided into four folders, each corresponding to a specific year (2019-2022). To facilitate predictive modeling and analytics, a common approach is to use the first three years (2019-2021) as the training set and the fourth year (2022) as the test set. This split allows for the development and validation of predictive models using historical data and then evaluates the model's performance on unseen data from the most recent year.

The rationale behind this data split is to simulate real-world scenarios where predictive models are trained on historical data and then applied to make predictions on future or unseen data. By evaluating model performance on a separate test set, one can assess the generalization ability and robustness of the model in predicting future outcomes.

ERRORS, NOISE, AND REDUNDANCIES:
The dataset does not inherently contain noise or errors. However, discrepancies or inaccuracies may arise from the data provided by the companies themselves. Financial reporting practices and data quality may vary among companies, leading to inconsistencies or inaccuracies in the reported fundamental metrics and financial indicators. These errors or sources of noise could include misreported financial figures, inconsistencies in accounting methods, or discrepancies in data formatting. Data cleaning and validation processes may be necessary to address such issues and ensure the reliability and accuracy of the dataset for analysis and modeling purposes.

EXTERNAL RESOURCES:
The "NSE Fundamental Insights: Fundamentals 2019-2022" dataset does not directly link to or rely on external resources such as websites, tweets, or other datasets. However, it is subject to government regulations, Securities and Exchange Board of India (SEBI) guidelines, and company policies, which may impact its existence and availability over time. Changes in regulatory requirements, reporting standards, or data disclosure policies by the government, SEBI, or individual companies could potentially affect the accessibility and continuity of the dataset.

As of the creation of this dataset, there are no official archival versions or guarantees regarding its permanence or stability over time. Future users should be aware of the dynamic nature of financial markets and the regulatory environment in which the dataset operates. While efforts may be made to maintain and update the dataset, there are no guarantees that it will remain unchanged

or accessible indefinitely.

CONFIDENTIALITY:
Currently, the dataset does not contain confidential data. The dataset primarily consists of publicly available financial metrics and indicators for stocks listed on the National Stock Exchange (NSE). It does not include sensitive or proprietary information that is protected by legal privilege, doctor-patient confidentiality, or non-public communications. The data provided in the dataset is derived from publicly accessible sources such as company financial reports, regulatory filings, and market data feeds, and does not breach any confidentiality agreements or privacy regulations.

POTENTIALLY OFFENSIVE CONTENT:
The dataset does not contain data that is offensive, insulting, threatening, or likely to cause anxiety. The dataset primarily consists of financial metrics and indicators related to stocks listed on the National Stock Exchange (NSE), and does not include any content that could be considered offensive or inappropriate. Users can analyze the dataset without encountering any offensive or distressing material.

SUBPOPULATIONS:
The dataset does not identify any subpopulations such as age or gender. Since the dataset primarily consists of financial metrics and indicators related to stocks listed on the National Stock Exchange (NSE), there is no demographic information associated with individual instances. Therefore, there are no subpopulations identified within the dataset based on demographic characteristics.

IDENTIFIABILITY:
It is not possible to identify individuals directly or indirectly from the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset. The dataset primarily consists of financial metrics and indicators related to stocks listed on the National Stock Exchange (NSE), and does not contain any personally identifiable information (PII) or individual-level data. Each instance in the dataset represents a specific stock, and there is no information that can be used to link these instances to individual natural persons.

SENSITIVE DATA:
The dataset does not contain data that might be considered sensitive in any way. The dataset primarily consists of financial metrics and indicators related to stocks listed on the National Stock Exchange (NSE), and does not include any information pertaining to racial or ethnic origins, sexual orientations, religious beliefs, political opinions, health data, biometric or genetic data, government identification, or criminal history. It solely focuses on publicly available financial information and does not breach any privacy regulations or confidentiality agreements.

By providing access to the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset, we do not intend to make any statement about any specific company or individual. The dataset is intended solely for informational and educational purposes, allowing users to analyze and explore financial metrics and indicators related to stocks listed on the National Stock Exchange (NSE). Users are encouraged to use the dataset responsibly and ethically, and to exercise caution when drawing conclusions or making decisions based on the information contained within the dataset. Any insights or findings derived from the dataset should be interpreted within the context of broader market dynamics and regulatory considerations. We aim to foster a collaborative and constructive environment for data analysis and exploration, and welcome feedback and contributions from users to enhance the utility and usability of the dataset.

Collection Process

DATA ACQUISITION PROCESS:
The data associated with each instance in the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset was acquired through a combination of direct observation and domain expertise. Approximately one-third of the raw financial data was gathered from various public sources, including company financial reports, regulatory filings, and market data feeds. The remaining data, including industry-relative metrics and growth-oriented features, was generated by domain experts based on their knowledge of the financial domain and market dynamics.

For the data reported by subjects or indirectly inferred/derived from other data, validation and

verification processes were employed to ensure accuracy and reliability. This validation process involved cross-referencing the derived metrics with multiple reputable sources, verifying calculations against industry standards, and conducting internal consistency checks to identify any discrepancies or outliers.

MECHANISMS AND PROCEDURES USED:
The data collection process for the dataset involved a combination of manual human curation and software-based techniques. Approximately one-third of the raw financial data was collected through web scraping techniques from various public domains, including company websites, financial news portals, and regulatory filings. The remaining two-thirds of the data, comprising industry-relative metrics and growth-oriented features, were calculated based on domain knowledge and expertise.

The mechanisms and procedures used for data collection were validated through several means. Firstly, web scraping techniques were implemented using reliable and established libraries and frameworks to ensure accurate extraction of data from public sources.Any discrepancies or inaccuracies are purely co-incidental.

PARENT DATASET
The dataset is not a sample from a larger set. Instead, it represents a comprehensive collection of financial metrics and indicators for stocks listed on the National Stock Exchange (NSE) over the specified time period (2019-2022). Therefore, there was no sampling strategy involved in the creation of this dataset, as it encompasses all available data within the defined scope.

ROLES IN COLLECTION PROCESS:
The data collection process for the dataset was primarily conducted by an individual researcher, with some involvement from undisclosed colleagues. However, the specific names of these colleagues cannot be disclosed at this time. No external parties, such as students, crowdworkers, or contractors, were involved in the data collection process, and no compensation arrangements or payments were made to external contributors.

DATA COLLECTION TIMEFRAME:
The data for the dataset was collected over a timeframe spanning from 2019 to 2022. This timeframe aligns with the creation timeframe of the data associated with the instances, as the dataset includes fundamental metrics and indicators for stocks listed on the National Stock Exchange (NSE) over this specified time period. Therefore, the data collection timeframe matches the creation timeframe of the dataset, ensuring that the collected data is relevant and up-to-date.

ETHICAL REVIEW PROCESSES:
Ethical review processes were not conducted for the dataset. However, the data collection and dissemination practices adhere to legal and regulatory guidelines governing financial data and market research. The dataset is compiled using publicly available information and does not include any personally identifiable information or sensitive data. While no formal ethical review was conducted, the data collection process is conducted within legal boundaries, and any necessary changes will be made in accordance with government policies and regulatory requirements.

DATA COLLECTION SOURCE:
The "NSE Fundamental Insights: Fundamentals 2019-2022" dataset does not directly relate to individuals.

IMPACT ANALYSIS:
No formal analysis of the potential impact of the dataset on data subjects has been conducted. However, as the dataset primarily consists of aggregated financial metrics and indicators for publicly listed companies, the potential impact on individual data subjects is minimal. The dataset does not contain personally identifiable information or sensitive data relating to individuals, thereby reducing the risk of privacy violations or negative impacts on data subjects. While no formal analysis has been conducted, the dataset's use and dissemination adhere to legal and regulatory requirements governing financial data and market research.

Preprocessing/cleaning/labeling

DATA PREPROCESSING:
Preprocessing and cleaning of the data were conducted as part of the data preparation process for the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset. The preprocessing steps included handling missing values, converting data types, removing unwanted characters, and ad-

dressing data integrity issues. Specifically, the dataset contained various forms of inconsistencies, such as NaN values, integers and floats represented as strings, and tampered data. These issues were addressed through a series of data cleaning techniques, including imputation for missing values, type conversion for numerical data, and data validation checks. Additionally, several preprocessing techniques were applied to prepare the dataset for modeling, such as feature scaling . Various machine learning models were also tested on the cleaned dataset to assess their performance and suitability for predictive analytics tasks.

RAW DATA:
The raw data used for the dataset was saved in addition to the preprocessed, cleaned, and labeled data. The raw data can be found in the "raw" folder within the dataset repository. Access to the raw data allows for transparency and reproducibility of the preprocessing and cleaning steps, facilitating future analyses or modifications if needed.

SOFTWARE USED FOR DATA PROCESSING:
The preprocessing, cleaning, and labeling of instances in the dataset were performed using Python programming language along with various libraries and tools. As such, the software used is readily available to the public as part of the Python ecosystem. The specific libraries and tools utilized may include pandas, NumPy, matplotlib, seaborn, and scikit-learn, among others. Access to these libraries and tools can be obtained through the official Python website (https://www.python.org/) and package repositories such as PyPI (Python Package Index).

Uses

UTILIZATION OF THE DATASET:
Yes, the dataset has been utilized for various tasks and analyses. However, the specific details and results of these tasks cannot be disclosed at this time due to confidentiality agreements and other reasons. The dataset holds immense value for predictive analytics, financial analysis, and research in the domain of stock market dynamics and investment strategies. Various machine learning models and analytical techniques have been applied to the dataset to derive insights, identify patterns, and make informed decisions in the financial domain.

REPOSITORY LINKING TO PAPERS OR SYSTEMS:
No specific repository linking to papers or systems that use the dataset has been created or identified at this time. However, the dataset remains publicly available for use, and researchers or practitioners are encouraged to cite the dataset directly in their publications or projects if utilized.

OTHER POTENTIAL TASKS:
The "NSE Fundamental Insights: Fundamentals 2019-2022" dataset can be utilized for a wide range of tasks beyond those already mentioned. Some potential tasks include:

1. Predictive Modeling: The dataset can be used to build predictive models for forecasting stock prices, predicting market trends, or identifying potential investment opportunities.
2. Financial Analysis: Researchers and analysts can leverage the dataset to conduct comprehensive financial analysis, including assessing company performance, evaluating industry trends, and identifying key drivers of financial success.
3. Portfolio Optimization: Investors can utilize the dataset to optimize their investment portfolios by analyzing the historical performance and fundamental characteristics of various stocks and sectors.
4. Risk Management: The dataset can aid in risk assessment and management by identifying potential risks associated with specific stocks, sectors, or market conditions.
5. Sectoral Analysis: Researchers can perform in-depth analysis of specific sectors or industries within the stock market, identifying key trends, challenges, and opportunities for growth.
6. Market Sentiment Analysis: Natural language processing techniques can be applied to analyze news articles, social media posts, and other textual data sources to gauge market sentiment and investor sentiment towards specific stocks or sectors.

These are just a few examples of the potential tasks that the dataset can be used for, highlighting its versatility and applicability in various domains within finance and investment.

POTENTIAL IMPACTS AND USAGE CONSIDERATIONS:
The composition of the "NSE Fundamental Insights: Fundamentals 2019-2022" dataset and the

process by which it was collected, preprocessed, cleaned, and labeled may have implications for future uses. Future users should be aware of the following considerations to avoid potential undesirable impacts:

1. Data Quality Issues: The dataset may contain missing values, inaccuracies, or inconsistencies due to factors such as data collection methods, data reporting practices, or errors in preprocessing. Future users should carefully evaluate the quality of the data and consider performing additional data cleaning and validation steps as necessary.

2. Legal and Regulatory Compliance: The dataset contains financial data and metrics related to publicly traded companies, which are subject to legal and regulatory requirements. Future users should ensure compliance with applicable laws, regulations, and industry standards governing the use and disclosure of financial data, including data privacy and securities regulations.

3. Ethical Considerations: Ethical considerations should guide the use of the dataset to prevent harm to individuals, groups, or society at large. Future users should adhere to ethical principles such as fairness, transparency, accountability, and respect for privacy when analyzing and utilizing the data.

By being cognizant of these considerations and taking proactive measures to address them, future users can mitigate potential undesirable impacts and promote responsible and ethical use of the dataset.

RESTRICTIONS ON USAGE:

The "NSE Fundamental Insights: Fundamentals 2019-2022" dataset should not be used for certain tasks or purposes due to ethical, legal, or practical considerations. Some tasks for which the dataset should not be used include:

1. Commercial Exploitation: The dataset is not intended for commercial use or profit-driven activities. Users should refrain from using the dataset for purposes that involve direct or indirect commercial exploitation, such as selling the data or using it for proprietary trading strategies without appropriate authorization.

2. Regulatory Compliance: Users should not rely solely on the dataset for making financial decisions or compliance-related activities without verifying the information with authoritative sources and ensuring compliance with relevant regulatory requirements. The dataset may not always reflect the most up-to-date or accurate information, and users should exercise caution and due diligence when using it for regulatory compliance purposes.

3. Misrepresentation or Misuse: The dataset should not be used to misrepresent or mislead others, manipulate financial markets, or engage in fraudulent activities. Users should refrain from misusing the data for purposes that could harm individuals, organizations, or the integrity of financial markets.

4. Violation of Terms of Use: Users should adhere to the terms of use and licensing agreements associated with the dataset. Any unauthorized or prohibited use of the dataset, including activities that violate intellectual property rights, data privacy laws, or contractual obligations, should be avoided.

5. Ethical Considerations: Users should consider the ethical implications of their use of the dataset and refrain from engaging in activities that could result in harm, discrimination, or unfair treatment of individuals or groups. This includes avoiding tasks that involve profiling, stereotyping, or discriminating against individuals based on sensitive attributes such as race, gender, or socioeconomic status.

By respecting these limitations and guidelines, users can ensure responsible and ethical use of the dataset while mitigating potential risks and harms associated with inappropriate or unauthorized usage.

Users should refer to SEBI (Securities and Exchange Board of India) guidelines and other relevant regulatory frameworks to ensure compliance with financial regulations and avoid any legal or regulatory trouble.

Distribution

DISTRIBUTION AND LICENSING:

The distribution of the dataset to third parties outside of the entity responsible for its creation is subject to certain considerations and limitations. While there are currently no plans for widespread distribution beyond the entity's control, access to the dataset may be granted to authorized third parties for specific purposes, such as academic research, educational use, or collaborative projects with appropriate permissions and agreements in

place.

DISTRIBUTION METHOD AND LICENSE:
The dataset is currently distributed via a GitHub repository. Users can access the dataset by downloading the relevant files from the repository's URL. As of now, the dataset does not have a digital object identifier (DOI). However, future considerations may include assigning a DOI to facilitate citation and tracking of dataset usage.

AVAILABILITY:
The dataset is currently available in the GitHub repository and can be accessed at any time.

LICENSE TERMS:
The dataset is distributed under the terms of the MIT License. A copy of the MIT License can be found in the GitHub repository containing the dataset. There are no fees associated with using the dataset under this license.

THIRD-PARTY RESTRICTIONS:
No third parties have imposed IP-based or other restrictions on the data associated with the instances.

EXPORT CONTROLS AND REGULATORY RESTRICTIONS:
As of now, there are no export controls or other regulatory restrictions that apply to the dataset or individual instances. Users are advised to comply with any relevant regulations or restrictions based on their jurisdiction.

Maintenance

SUPPORT AND MAINTENANCE:
The support, hosting, and maintenance of the dataset will depend on future availability and compliance with SEBI and other regulatory policies and guidelines.

CONTACT INFORMATION:
The owner/curator/manager of the dataset can be contacted via email at anant23k.kiran@gmail.com.

ERRATUM:
As of now, there are no known errata associated with the dataset.

UPDATES:
Updates to the dataset, such as correcting labeling errors, adding new instances, or deleting instances, may occur in the future. The frequency and process of updates will depend on various factors, including data availability, changes in regulatory requirements, and user feedback. Communication of updates to users may be facilitated through channels such as mailing lists or GitHub repositories.

DATA RETENTION:
As the dataset does not directly relate to people, there are no specific limits on the retention of individual data associated with instances. However, data retention policies will be aligned with relevant regulatory requirements and best practices to ensure compliance and safeguard privacy.

OLD VERSIONS:
At present, there are no plans to maintain older versions of the dataset. However, if there are significant updates or changes in the future, users will be informed through the dataset's repository on GitHub and any associated communication channels.

CONTRIBUTIONS:
Contributions to the dataset are welcome and can be submitted through the GitHub repository associated with the dataset. Contributions will be reviewed by maintainers to ensure quality and alignment with objectives. Once accepted, contributions will be incorporated into the dataset, and users will be notified through the repository's communication channels.

Disclaimer: The dataset is provided solely for educational and informational purposes. Users should exercise caution and refer to SEBI and other regulatory guidelines before utilizing the dataset for any commercial or investment-related activities.