

- Technical Documentation: Data Cleaning for Indian Startup Ecosystem Funding

Project Team: Team Name: Stockholm Team Lead: Faith Esther Njugu Mwai Data Analysts:

Dzeble Frank Kwame

Feisal Ali Hassan

Florence Mbabazi

Mark Wamache Kalerwa

Introduction: The Stockholm team undertook the Data Cleaning for Indian Startup Ecosystem Funding project, which involved cleaning and unifying four diverse datasets related to funding in India's burgeoning startup ecosystem. The datasets were sourced from various contributors, leading to inconsistencies and missing values. The objective of this project was to process and standardize the data to create a comprehensive and reliable dataset, named 'clean\_done,' ready for further analysis and insights derivation.

Data Cleaning Process: The data cleaning process was executed collaboratively by the Stockholm team, with each member contributing their expertise to different steps:

Data Collection and Merging:

Team members involved: Mark Wamache Kalerwa, Faith Esther Njugu Mwai

The team collected the datasets from multiple sources and merged them into a single dataset using Pandas. Care was taken to ensure a seamless merging process to maintain data integrity.

Handling Missing Location Data:

Team member involved: Faith Esther Njugu Mwai and Dzeble Frank Kwame

Missing location data for the company 'Sochcast' was addressed by filling in the missing value with 'Bangalore' using Pandas.

Unifying the 'Stage' Column:

Team member involved: Dzeble Frank Kwame

The 'Stage' column in the dataset was diverse, with different names for similar funding stages. To standardize the column, a mapping dictionary was created to group similar stages into common categories using Pandas.

Ensuring Valid Stage Categories:

Team member involved: Dzeble Frank Kwame

The validity of the 'Stage' column was checked to ensure that all stages were accurately categorized as 'Early Stage,' 'Mid Stage,' 'Late Stage,' or 'Other Stages' using Pandas.

Addressing Duplicate Entries:

Team member involved: Feisal Ali Hassan

Duplicate rows were identified and removed based on the 'Sector' column using Pandas to ensure data integrity and consistency.

Data Transformation:

Team member involved: Florence Mbabazi

The 'Amount' column was transformed to a float data type, and the 'Funding Year' column was transformed to an integer data type using Pandas, preparing the numerical data for future calculations and visualizations.

Conclusion: The collaborative efforts of the Stockholm team, led by Faith Esther Njugu Mwai, resulted in the successful completion of the Data Cleaning for Indian Startup Ecosystem Funding project. Each team member played a crucial role in different stages of the data cleaning process,

leveraging their expertise in Python libraries such as Pandas and NumPy. The comprehensive dataset 'clean\_done' is now ready for in-depth analysis, data visualization, and data-driven decision-making in the Indian startup ecosystem funding landscape. The team's commitment to data excellence and attention to detail ensures the dataset's reliability and accuracy for future research and exploration.

#### Visualization:

Below are some visualizations that provide a deeper understanding of the Indian startup ecosystem:

##### Histogram: Distribution of Funding Amounts

This histogram displays the distribution of funding amounts across different startups in India. It allows us to identify the common funding ranges and outliers.

##### Box Plot: Identifying Outliers in Funding Amounts

The box plot helps us visualize the presence of outliers in funding amounts. Outliers can significantly impact data analysis and identifying them is crucial for accurate insights.

##### Line Plot: Temporal Patterns of Funding

This line plot shows the funding amount trends over the years, providing insights into the growth and fluctuations of investments in the Indian startup ecosystem.

##### Bar Charts: Top Cities and Sectors by Startups and Funding

Bar charts highlight the top cities with the most startups and the highest amount of funding received. Additionally, they showcase the sectors that attract the most investment.

##### Pie Charts: Proportion of Startups and Funding by Sector

Pie charts visualize the proportion of startups and funding in different sectors. They allow us to compare the importance of each sector in the Indian startup landscape.

##### Scatter Plots: Correlation between Funding Amount and Sector/Location

Scatter plots examine the correlation between funding amounts and sectors or locations.

Understanding these relationships helps identify patterns and potential investment hotspots.

##### Kruskal-Wallis Test: Impact of Sector on Funding Amount

This statistical test examines whether the sector of a company has a significant impact on the amount of funding it receives. The results provide valuable insights into sector-specific investment patterns.

By incorporating data visualizations into our analysis, we gain a holistic view of India's startup ecosystem, empowering decision-makers and stakeholders to make informed and strategic choices.

●