INDIAN STARTUP FUNDING DATASET

Samiullah Shahid

Data Preprocessing

This report documents the data understanding, quality assessment, and data cleaning steps performed on the Startup Funding dataset. The purpose of these tasks was to prepare a clean and consistent dataset for further analysis.

1. Data Understanding & Loading- Dataset loaded:

startup_funding.csv- Columns inspected and renamed (e.g., "Date dd/mm/yyyy" → "date").- Extracted new columns: day, month, year from the date field.

2. Data Quality Assessment

Key issues identified in the dataset:

Issue	Description
Missing Values	Found in 'Amount in USD' and 'City'. Filled with 0 or 'unknown'.
Duplicates	Checked using duplicated() function.
Date Format	Converted from dd/mm/yyyy to standard datetime format.
Currency Formatting	Removed commas, 'undisclosed' entries handled.
City Names	Inconsistencies fixed (e.g., Bengaluru → Bangalore, Gurgaon → Gurugram).
Encoding	Whitespace stripped; some special characters may remain.
Industry/Investor	Converted to categorical but not fully standardized.

3. Data Cleaning

The raw dataset was assessed and underwent a thorough cleaning process to ensure accuracy, consistency, and readiness for analysis. The following steps were performed:

Handling Missing Values

- Missing values were identified in multiple columns.

- Appropriate strategies were applied based on the context:
    - Critical columns such as `Amount in USD` and `Date` were flagged or rows with missing values were removed.
    - Non-critical columns were imputed with placeholder values or left as nulls for analysis awareness.

Duplicate Removal

- Duplicate entries were detected and removed to prevent skewed analysis.

Date Standardization

- The `Date` column was converted to a standard datetime format (`YYYY-MM-DD`).
- Additional `Year` and `Month` columns were derived from the date for trend analysis.

Currency Column Cleaning

- The `Amount in USD` column was cleaned by removing commas and handling non-numeric entries.
- Values were converted to numeric format for accurate aggregation and computation.

City Name Standardization

- City names were standardized to consistent naming conventions (e.g., `Bangalore` → `Bengaluru`, `Gurgaon` → `Gurugram`).

Industry/Vertical Cleaning

- Inconsistent or misspelled industry/vertical categories were standardized for uniform classification.

Investor Name Formatting

- Investor names were cleaned to remove extra spaces, special characters, and inconsistent capitalization.

Encoding and Special Character Fixes

- Special characters and encoding issues in startup names (e.g., â€™) were corrected to standard UTF-8 representation.

Additional Preparations

- Columns were checked for consistency, correct data types were enforced, and the dataset was validated for readiness.

Data Analysis

Temporal Funding:

The temporal funding analysis highlights how startup funding in India evolved between 2015 and 2020. The results show a strong year-on-year growth in funding volumes, with notable peaks in 2017 and 2019. Monthly breakdowns indicate that January, June, and September consistently recorded higher funding activity, suggesting seasonal patterns tied to investment cycles. Overall, funding activity remained resilient, though slightly moderated in 2020, likely due to global macroeconomic conditions.

b) Geographic Analysis

Funding in India is highly **geographically concentrated**, with a few major hubs driving the majority of investment. The **Top 10 cities** by funding amount are dominated by **Bengaluru, Delhi NCR, and Mumbai**, together contributing the lion's share of deals. When analyzed by the **number of startups funded**, these same cities continue to lead, reaffirming their position as India's core innovation clusters. Secondary cities are emerging but still account for a relatively smaller proportion of overall funding.

c) Sector Analysis

Industry-level analysis reveals that **technology-driven sectors** attracted the bulk of funding. **E-commerce, FinTech, and Enterprise Tech** consistently rank among the top-funded verticals. FinTech shows a strong upward trend, reflecting investor confidence in digital payments and neobanking. E-commerce funding stabilized after its early boom years, while enterprise/B2B solutions show steady growth. When ranked by **average deal size**, capital-intensive sectors such as healthcare technology and consumer tech appear at the top.

d) Investment Type Analysis

The funding round distribution indicates that **Seed and Series A** dominate by frequency, reflecting India's dynamic early-stage startup ecosystem. However, when measured by **average deal size**, later rounds such as **Series C and beyond** command significantly larger investments. **Series B and C rounds** are critical inflection points for scaling startups, while **Seed funding** remains the most common entry path for new ventures.

e) Investor Analysis

The investor landscape highlights a group of highly active firms and individuals shaping India's funding ecosystem. The **Top 20 investors by deal count** include well-known venture capital firms, accelerators, and angel networks. When ranked by **total amount invested**, a smaller set of deep-pocketed investors dominate, often leading large rounds in unicorns. Investor-sector mapping reveals clear preferences: some investors concentrate heavily on FinTech, while others are diversified across e-commerce, enterprise tech, and consumer services.

f) Startup Analysis

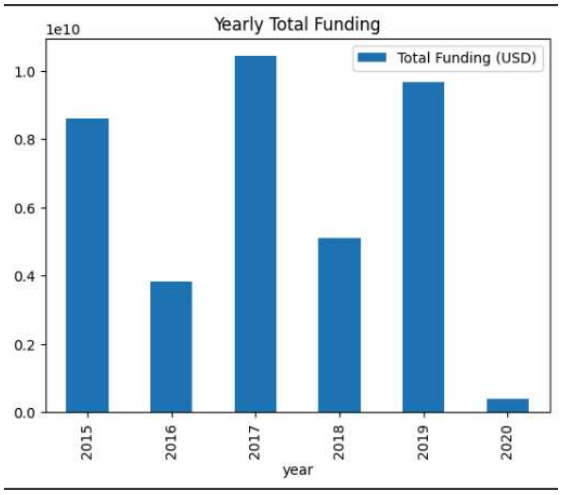
The startup-level breakdown highlights the **Top 10 most funded companies**, with household names such as **BYJU'S, Paytm, and OYO** leading the charts. Unicorn and mega-deal activity (>$100M) has accelerated over the years, signaling maturing investor confidence. For startups with multiple funding rounds, the **average time between rounds** shortens significantly once product-market fit is established, reflecting investor urgency to capture growth opportunities.

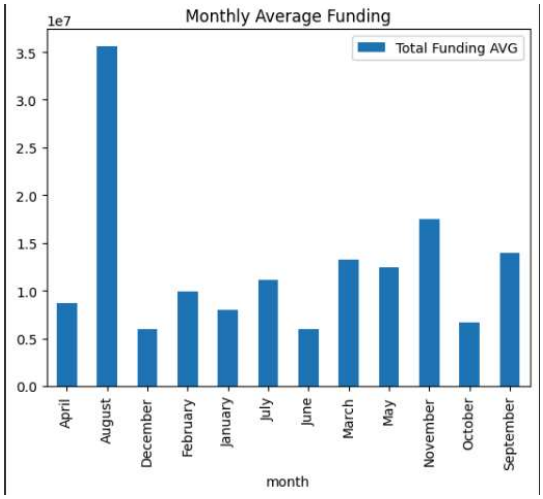
g) Deal Size Analysis

Deal size distribution shows a **heavily right-skewed pattern**, where most startups raise modest rounds under $5M, but a few very large deals dominate the aggregate funding amount. The **median funding amount** is significantly lower than the **mean**, indicating outlier effects from mega-deals. This underscores the polarized nature of startup

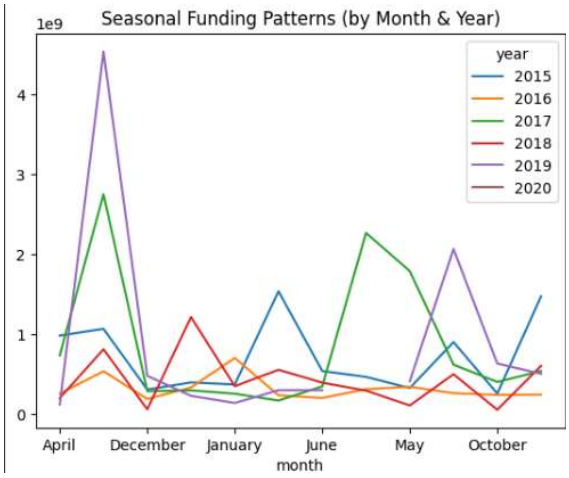
funding — while many companies raise smaller, incremental amounts, a select few capture blockbuster rounds that dramatically shape overall industry trends.

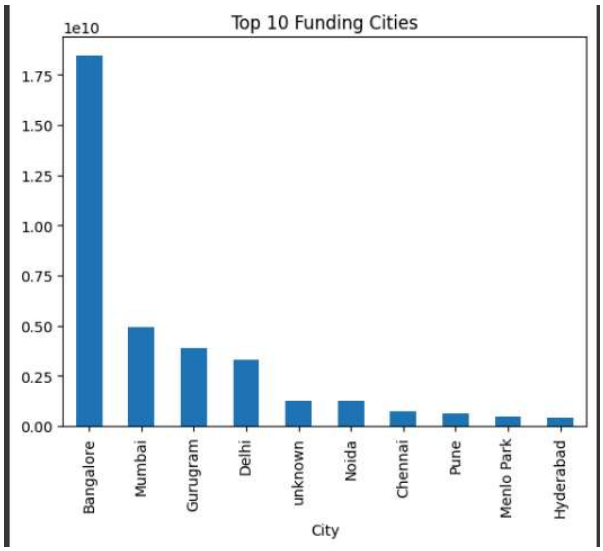
CHARTS AND GRAPHA

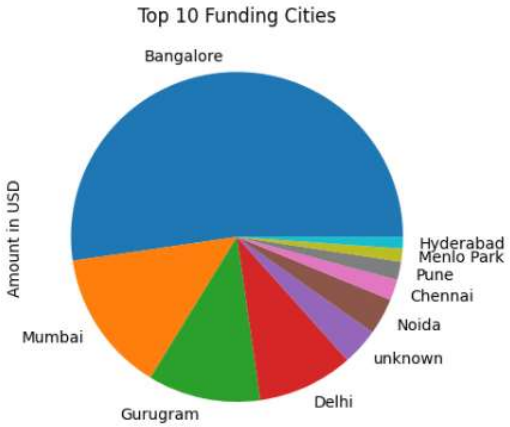


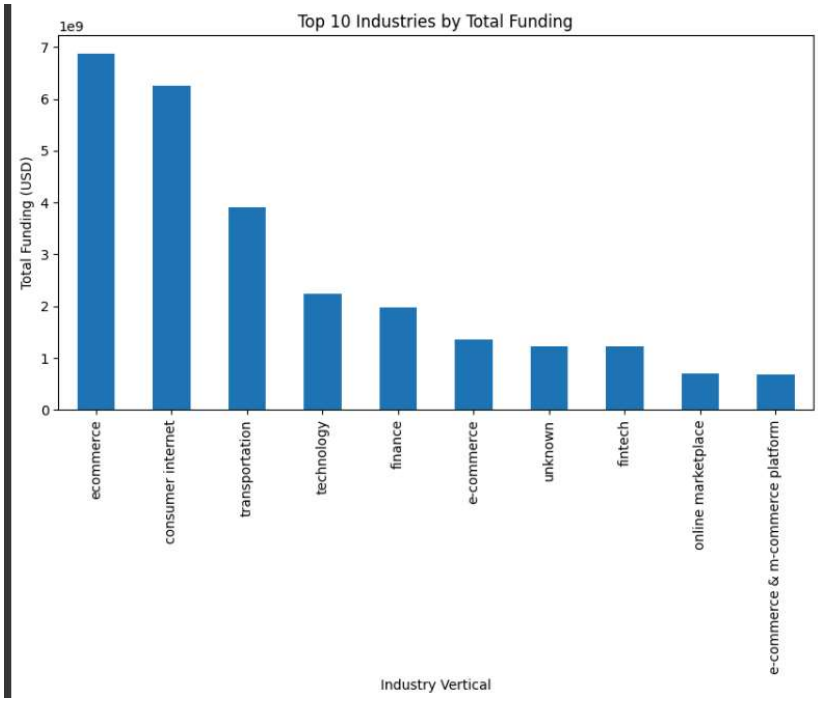

Monthly Average Funding
1e7
Total Funding AVG
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
April
August
December
February
January
July
June
March
May
November
October
September
month

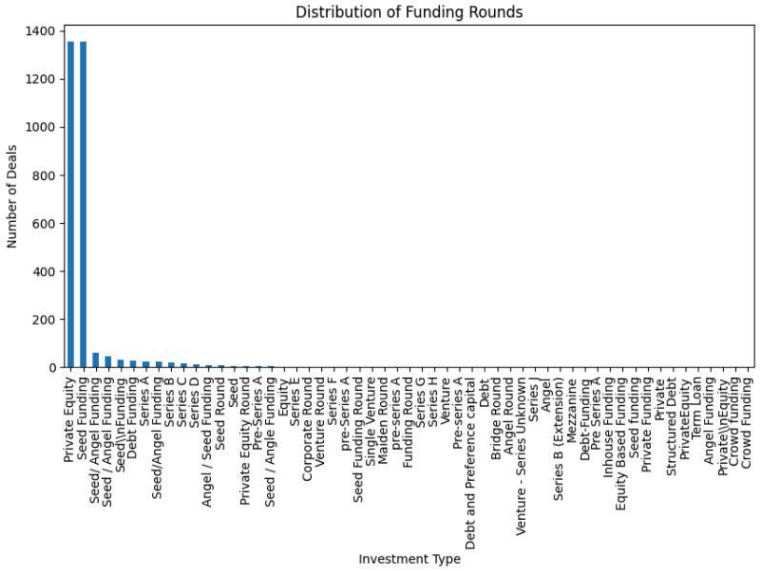

Montly Number of Deals
Number of Deals in month
290
280
270
260
250
240
230
220
April
December
January
June
May
October
month

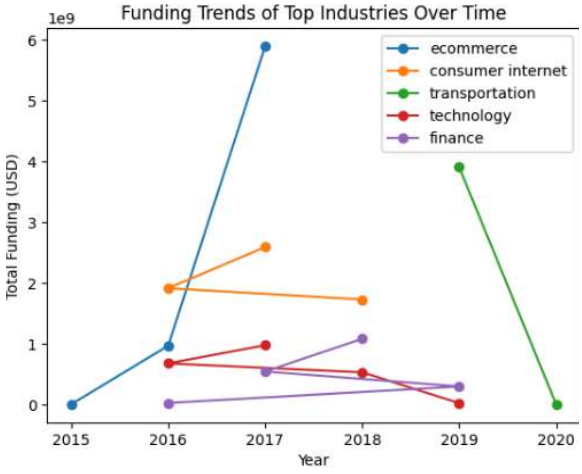

Seasonal Funding Patterns (by Month & Year)
1e9
year
2015
2016
2017
2018
2019
2020
4
3
2
1
0
April
December
January
June
May
October
month

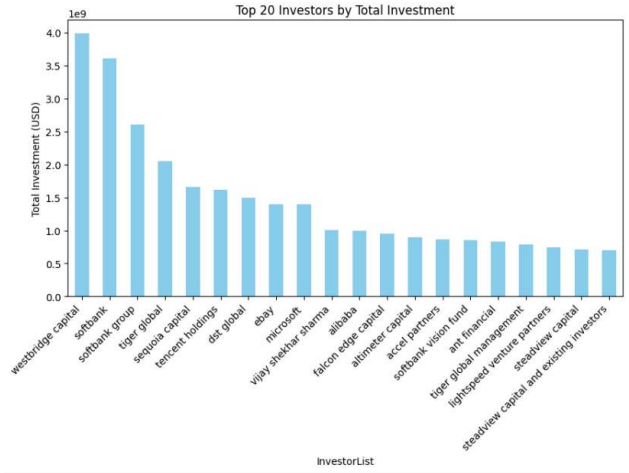

Top 10 Funding Cities
1e10
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Bangalore
Mumbai
Gurugram
Delhi
unknown
Noida
Chennai
Pune
Menlo Park
Hyderabad
City

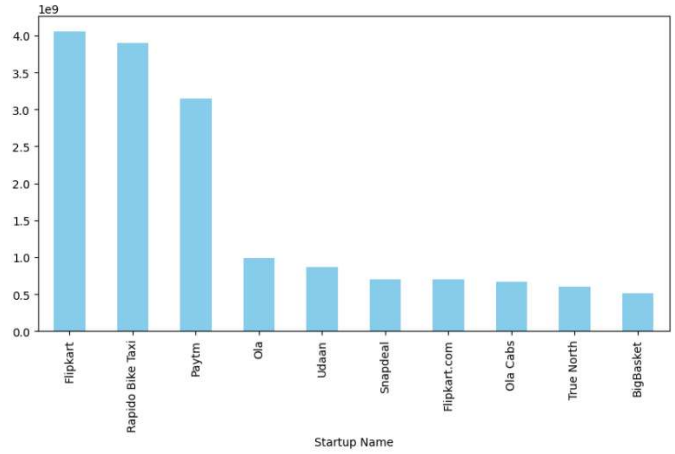

Top 10 Funding Cities
Bangalore
Amount in USD
Hyderabad
Menlo Park
Pune
Chennai
Noida
unknown
Delhi
Gurugram
Mumbai

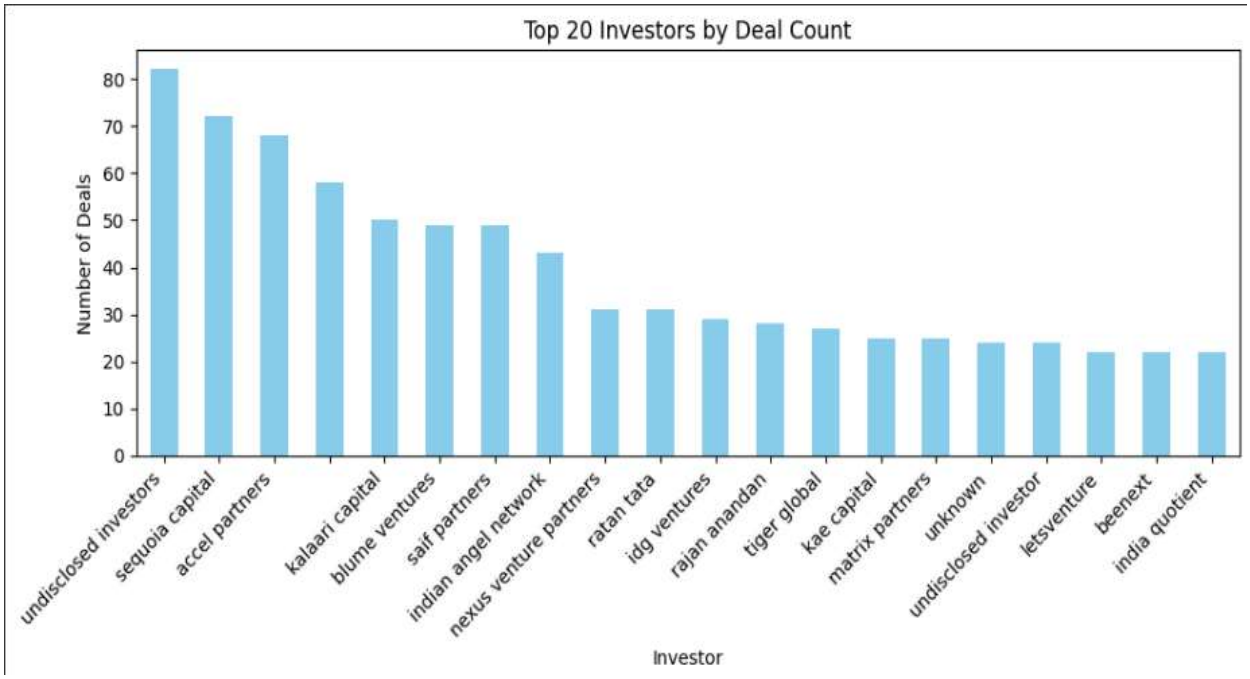

Top 10 Industries by Total Funding
1e9
Total Funding (USD)
0
1
2
3
4
5
6
7
ecommerce
consumer internet
transportation
technology
finance
e-commerce
unknown
fintech
online marketplace
e-commerce & m-commerce platform
Industry Vertical

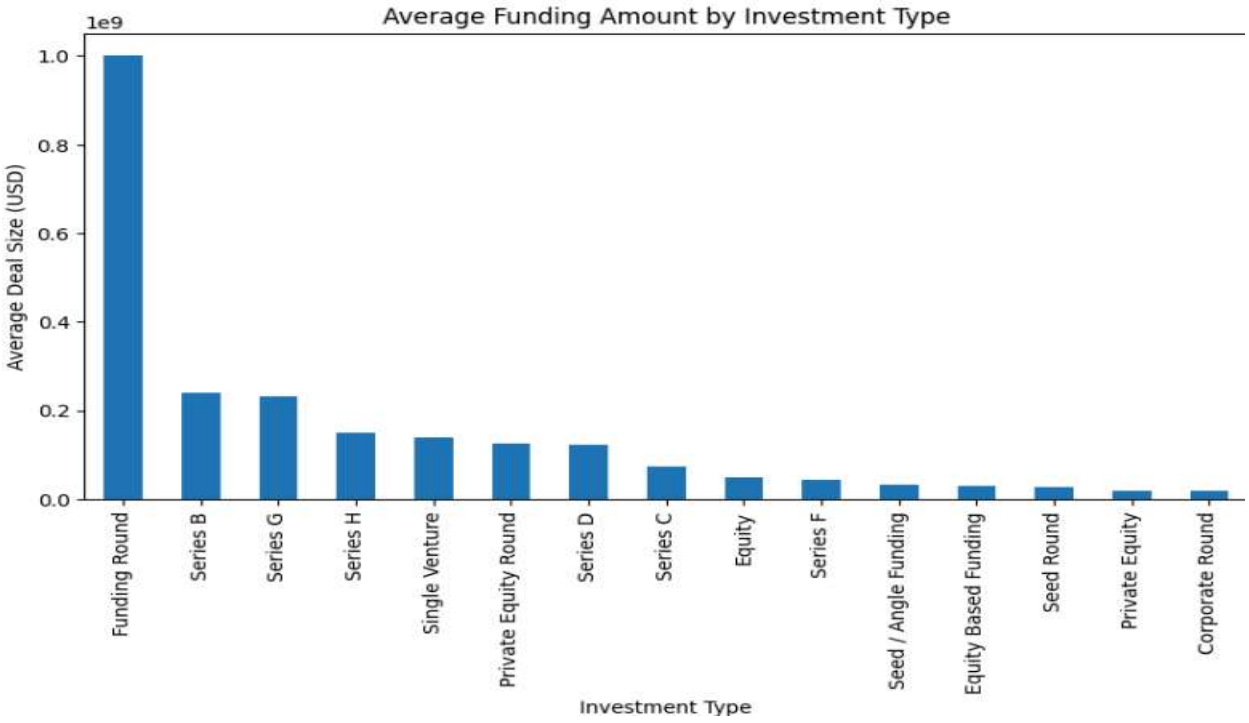

Distribution of Funding Rounds
Number of Deals
Investment Type
Private Equity
Seed Funding
Seed/ Angel Funding
Seed / Angel Funding
Seed\\nFunding
Debt Funding
Series A
Seed/Angel Funding
Series B
Series C
Series D
Angel / Seed Funding
Seed Round
Seed
Private Equity Round
Pre-Series A
Seed / Angle Funding
Equity
Series E
Corporate Round
Venture Round
Series F
pre-Series A
Seed Funding Round
Single Venture
Maiden Round
pre-series A
Funding Round
Series G
Series H
Venture
Pre-series A
Debt and Preference capital
Debt
Bridge Round
Angel Round
Venture - Series Unknown
Series J
Angel
Series B (Extension)
Mezzanine
Debt-Funding
Pre Series A
Inhouse Funding
Equity Based Funding
Seed funding
Private Funding
Private
Structured Debt
PrivateEquity
Term Loan
Angel Funding
Private\\nEquity
Crowd funding
Crowd Funding

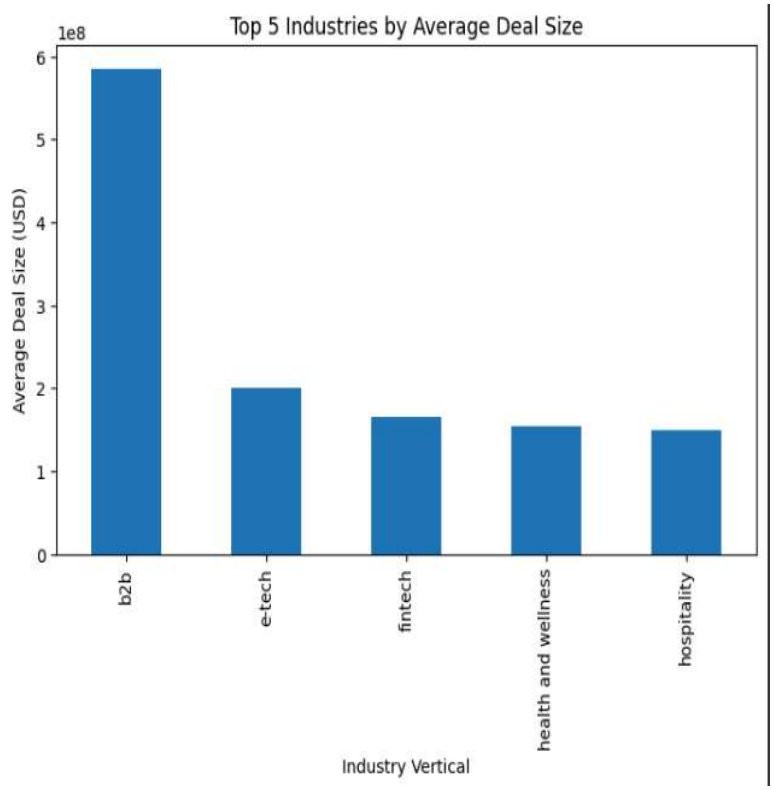

Funding Trends of Top Industries Over Time
1e9
Total Funding (USD)
Year
ecommerce
consumer internet
transportation
technology
finance

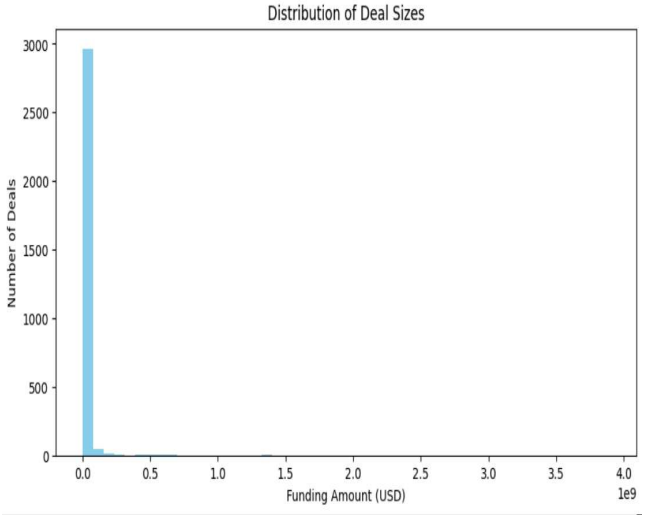

Top 20 Investors by Total Investment
1e9
Total Investment (USD)
InvestorList
westbridge capital
softbank
softbank group
tiger global
sequoia capital
tencent holdings
dst global
ebay
microsoft
vijay shekhar sharma
alibaba
falcon edge capital
altimeter capital
accel partners
softbank vision fund
ant financial
tiger global management
lightspeed venture partners
steadview capital
steadview capital and existing investors

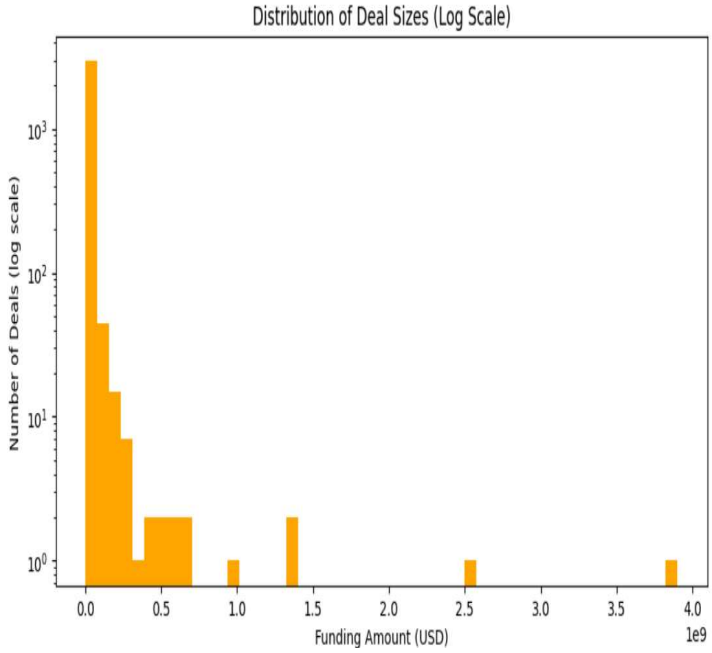

1e9
Startup Name
Flipkart
Rapido Bike Taxi
Paytm
Ola
Udaan
Snapdeal
Flipkart.com
Ola Cabs
True North
BigBasket


Top 20 Investors by Deal Count
Number of Deals
Investor
undisclosed investors
sequoia capital
accel partners
kalaari capital
blume ventures
saif partners
indian angel network
nexus venture partners
ratan tata
idg ventures
rajan anandan
tiger global
kae capital
matrix partners
unknown
undisclosed investor
letsventure
beenext
india quotient


Average Funding Amount by Investment Type
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Average Deal Size (USD)
Funding Round
Series B
Series G
Series H
Single Venture
Private Equity Round
Series D
Series C
Equity
Series F
Seed / Angle Funding
Equity Based Funding
Seed Round
Private Equity
Corporate Round
Investment Type


Top 5 Industries by Average Deal Size
1e8
6
5
4
3
2
1
0
Average Deal Size (USD)
b2b
e-tech
fintech
health and wellness
hospitality
Industry Vertical


Distribution of Deal Sizes
Number of Deals
Funding Amount (USD)
1e9


Distribution of Deal Sizes (Log Scale)
Number of Deals (log scale)
Funding Amount (USD)
1e9