# **📌Startup Investments Dataset📌 Documentation**

## **1. Overview of the Dataset**

This dataset contains information about **startup companies**, their **funding rounds**, and **investment details**. The data includes information like when a startup was founded, how much funding it received, and its location.

### **2.Understanding the Dataset**

The dataset includes several columns that represent different aspects of a startup’s journey. Here’s what each key column means:

1. **Company Information:**
   * **permalink:** A unique identifier for each startup.
   * **name**: The name of the startup.
   * **homepage\_url**: The official website of the startup (if available).
   * **category\_list**: Industries the startup operates in (e.g., software, healthcare).
   * **market**: A simplified industry classification for the startup.
2. **Funding Information:**
   * **funding\_total\_usd**: The total funding the startup has received in US dollars.
   * Different investment types such as **seed**, **venture**, **private\_equity**, **debt\_financing**, etc., represent specific funding amounts received in those categories.
   * **funding\_rounds**: The number of funding rounds the startup has gone through.
3. **Company Status and Location:**
   * **status**: Indicates if the startup is **operating, closed, or acquired**.
   * **country\_code**, **state\_code**, **region**, **city**: These columns define the startup’s location.
4. **Time-Based Data:**
   * **founded\_at**: The exact date the startup was founded.
   * **founded\_year**: The year the startup was established.
   * **founded\_month:** The month in which the startup was founded.
   * **founded\_quarter**: The quarter (Q1, Q2, etc.) in which the startup was founded.
   * **first\_funding\_at**: The date the startup received its first investment.
   * **last\_funding\_at**: The date of the most recent funding round.

**3. Understanding the Columns**

#### **Column Descriptions:**

|  |  |
| --- | --- |
| **Column name** | **Description** |
| permalink | Unique identifier for each startup |
| Name | The official name of the startup. |
| category\_list | Industries or sectors the startup belongs to. |
| market | A high-level classification of the business sector**.** |
| funding *total\_*UsD | The total funding received by the startup in US dollars. |
| Status | Indicates whether the startup is operating, acquired, or closed. |
| country\_code | The country where the startup is based. |
| state\_code | The state or province of the startup’s headquarters. |
| city | The specific city where the startup is headquartered. |
| funding\_rounds | The number of times the startup has received funding. |
| founded\_at | The exact date when the startup was founded. |
| Founded\_month | The month in which the startup was founded. |
| founded\_quarter | The quarter of the year in which the startup was founded. |
| founded\_year | The year the startup was established. |
| first\_funding\_at | The date when the startup received its first funding. |
| last\_funding\_at | The date of the most recent funding round. |
| seed | Amount of seed funding received. |
| venture | Amount of venture funding received. |
| Equity\_crowdfunding | Amount received from equity crowdfunding. |
| undisclosed | Amount of undisclosed funding received. |
| Convertible\_note | Amount received via convertible notes. |
| debt\_financing | Amount received via debt financing. |
| angel | Amount received from angel investors. |
| grant | Amount received as a grant. |
| private\_equity | Amount received from private equity investments. |
| post\_ipo\_equity | Amount raised through post-IPO equity. |
| post\_ipo\_debt | Amount raised through post-IPO debt. |
| secondary\_market | Amount received from secondary market transactions. |
| product\_crowdfuning | Amount raised through product crowdfunding. |
| funding\_rounds(A-H) | Amounts raised in Series A to Series H funding rounds. |

### **4.Data Cleaning Process**

To make the dataset suitable for analysis, we performed several **data cleaning** steps. Here’s a detailed breakdown:

#### **1. Standardizing Column Names**

* The column names were converted to lowercase, and spaces and special characters were replaced with underscores (\_).
* This ensures consistency and makes it easier to refer to columns in programming.

#### 

#### **2. Converting Funding Values to Numerical Format**

* Funding columns (e.g., **funding\_total\_usd, seed, venture,** etc.) contained values with **commas and special characters**.
* We removed these characters and converted the data into **numeric format** for calculations.
* Missing or invalid values (-) were replaced with **NaN** (Not a Number).

#### **3. Handling Missing Values**

* **For categorical columns** **(country\_code, state\_code, city, market, status**), missing values were replaced with "**Unknown**" to retain the record while marking incomplete data.
* **For numerical columns**, missing values were filled using the **median value** of that column to prevent bias in analysis.

#### **4. Converting Date Columns to Date Format**

* Date columns (**founded\_at, first\_funding\_at, last\_funding\_at**) were converted from text format to **datetime format**.
* This allows for easier date-based calculations, such as finding trends over time.
* If **founded\_at** was missing, it was estimated based on **founded\_year**, **founded\_month**, or **first\_funding\_at** (if available).

#### **5. Deriving Missing Date Information**

* If **founded\_year** was missing, it was extracted from founded\_at whenever possible.
* If **founded\_month** or **founded\_quarter** was missing, it was derived from **founded\_at** to enable time-based grouping.

#### **6. Optimizing Data Types**

* Categorical columns (e.g., **status, market, country\_code**) were converted to **category type** to improve memory efficiency.
* This reduces memory usage, making the dataset faster to process.

#### **7. Removing Duplicate Entries**

* Any duplicate records were dropped to ensure accuracy and prevent misleading insights.

#### **8.Creating new columns on basis of founded\_at and first\_funding\_at**

* Converts first\_funding\_at to datetime and fills missing founded\_year values with the year from first\_funding\_at, ensuring consistency and improving the completeness of the dataset for trend analysis.

#### 

#### **9. Saving the Cleaned Dataset**

* After applying all cleaning steps, the cleaned dataset was saved as **"cleaned\_dataset.csv**" for further analysis.

### **Why These Cleaning Steps Were Necessary?**

1. **Standardizing column names** ensures easy access and manipulation of data.
2. **Converting funding values to numerical format** allows mathematical operations like sum, average, and comparisons.
3. **Handling missing values** prevents incomplete data from affecting analysis.
4. **Ensuring correct date formats** enables time-series analysis and trend forecasting.
5. **Deriving missing values** improves dataset completeness.
6. **Optimizing data types** improves performance for large datasets.
7. **Removing duplicates** ensures accurate insights.

With this cleaned dataset, we can now perform **investment trend analysis, funding pattern detection, and startup success predictions** effectively.