📌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.