Nneka Maduike
VUG/CSC/24/12708
CSC 222 Assignment
Mr.Uloko, F.O.

Company Funding Data Analysis and Visualization Report

Introduction
This project analyzes a company funding dataset to extract insights on startup funding trends across different sectors, countries, regions, and cities. The analysis aims to identify leading startup hubs and sectors attracting significant investor activity, using real-world analytics techniques.
The report covers dataset exploration, data cleaning, exploratory data analysis (EDA), visual interpretations, and real-world implications of findings (Han, Kamber, & Pei, 2011).

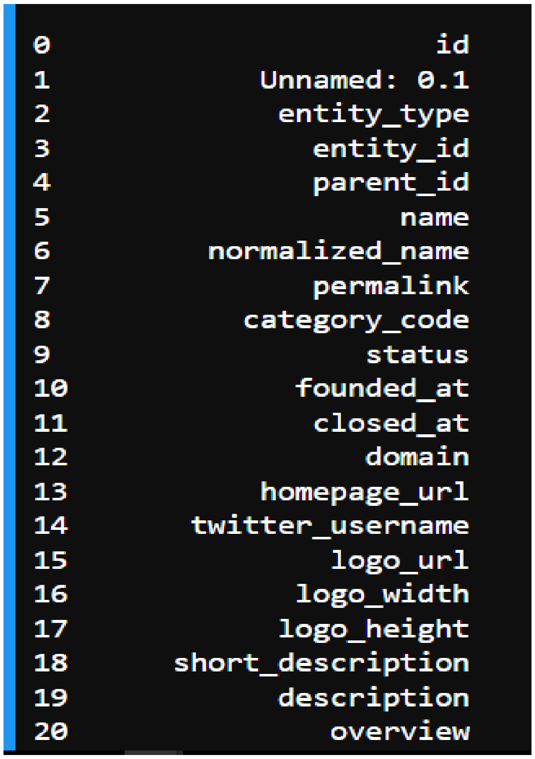
Dataset Overview
The dataset was sourced from Kaggle (Kaggle, n.d.) and tracks startup funding rounds, total funding amounts (in USD), startup establishment years, industry sectors, and geographic locations (region, city, and country).

Dataset summary
Shape

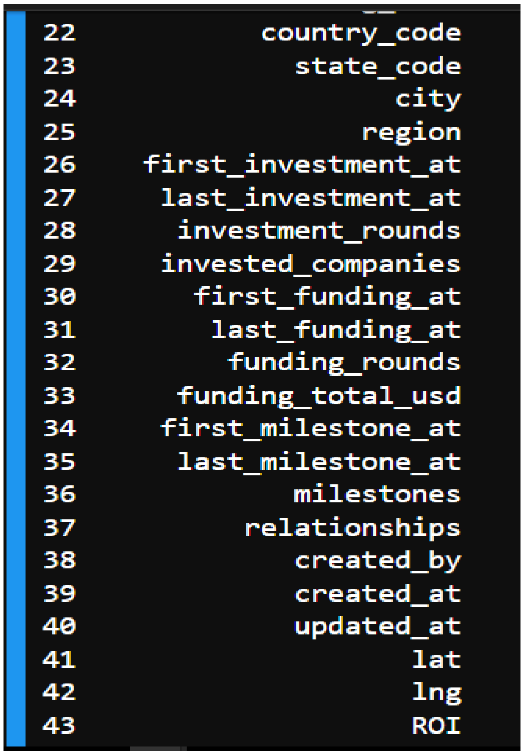
- Rows: 196,553
- Columns: 44

Features and Data Types
The dataset contains a mix of categorical, numeric, date, and text data:

- Numeric: funding_total_usd, funding_rounds
- Categorical: category_code, region, city, country
- Date: founded_at
- Text: Company descriptions and names



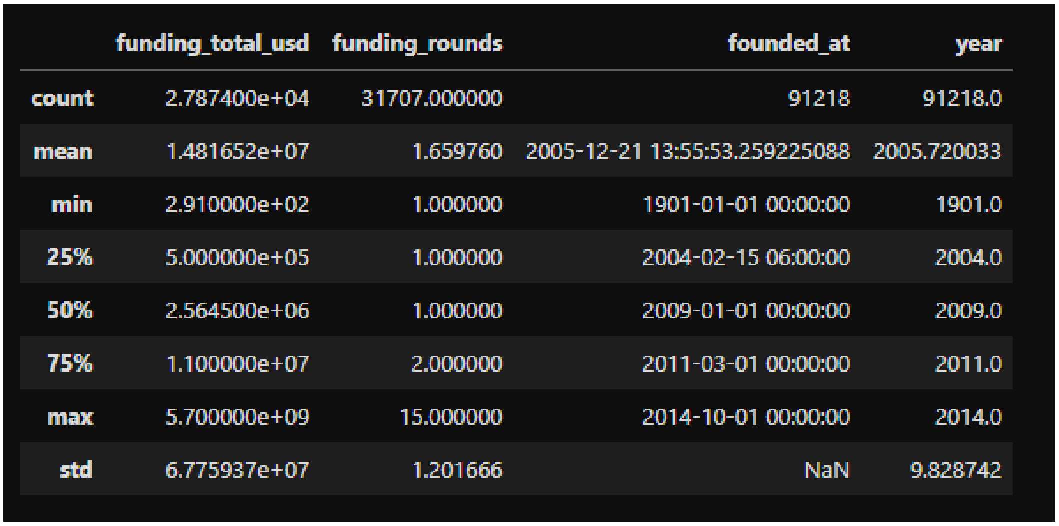
```
0                    id
1         Unnamed: 0.1
2          entity_type
3            entity_id
4            parent_id
5                 name
6      normalized_name
7            permalink
8        category_code
9               status
10          founded_at
11           closed_at
12              domain
13        homepage_url
14    twitter_username
15            logo_url
16          logo_width
17         logo_height
18   short_description
19         description
20            overview
```



```
22        country_code
23          state_code
24                city
25              region
26 first_investment_at
27  last_investment_at
28   investment_rounds
29  invested_companies
30    first_funding_at
31     last_funding_at
32      funding_rounds
33   funding_total_usd
34  first_milestone_at
35   last_milestone_at
36          milestones
37       relationships
38          created_by
39          created_at
40          updated_at
41                 lat
42                 lng
43                 ROI
```

Summary Statistics
For key numerical features:

- Funding Total (USD): Mean, median, minimum, maximum, and standard deviation calculated.
- Funding Rounds: Summary statistics provided.
- Founded At: Year extracted for temporal analysis.



	funding_total_usd	funding_rounds	founded_at	year
count	2.787400e+04	31707.000000	91218	91218.0
mean	1.481652e+07	1.659760	2005-12-21 13:55:53.259225088	2005.720033
min	2.910000e+02	1.000000	1901-01-01 00:00:00	1901.0
25%	5.000000e+05	1.000000	2004-02-15 06:00:00	2004.0
50%	2.564500e+06	1.000000	2009-01-01 00:00:00	2009.0
75%	1.100000e+07	2.000000	2011-03-01 00:00:00	2011.0
max	5.700000e+09	15.000000	2014-10-01 00:00:00	2014.0
std	6.775937e+07	1.201666	NaN	9.828742

Data Cleaning and Preprocessing

Several cleaning and preprocessing steps were undertaken (Dasu & Johnson, 2003):

- Feature Selection: Retained significant columns with sufficient data; dropped columns with over 70% missing values.
- Missing Values:
    - Numerical columns with missing data were dropped using .dropna().
    - Categorical missing data was filled with "Unknown" using .fillna().
- Duplicates:
    - 240 duplicate records were identified and removed using .drop_duplicates().
- Data Type Corrections:
    - founded_at was converted from object type to datetime.
- Feature Engineering:
    - Extracted year from founded_at.
    - Created new features: funding_per_round_per_startup, funding_per_round, and funding_range to enrich analysis.
- Column Renaming:
    - category_code renamed to sector for clarity.

Exploratory Data Analysis (EDA) Observations

1. Sectors with the Highest Funding

The Biotech sector attracted the highest total funding, outperforming other sectors by at least 25%, signaling strong investor confidence. Other highly funded sectors include Software, Cleantech, and Mobile.

2. Funding Trends Over the Last Five Years

All top-performing sectors experienced funding declines over the past five years, with E-commerce, Biotech, and Cleantech seeing the most significant drops.

3. Highest Funding Per Round and Per Startup

The Automotive, Nanotech, and Cleantech sectors recorded the highest funding per round. Notably, Nanotech startups also received the highest funding per startup, with the Government sector closely following.

4. Funding Trends (1902–2014)

The peak investment period occurred before 1910. Surprisingly, funding levels declined significantly during the early 2000s.

5. Funding Range Distribution

The majority of startups fell into the $1M–$5M funding range, followed by the $10M–$50M range, showing a skew toward early and mid-stage investment levels.

6. Top Startup Hubs by Funding

Bermuda recorded the highest average funding per startup, followed by Luxembourg, Malaysia, and China.

Visualizations and Interpretations

Seaborn is a powerful Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive, informative statistical graphics (Waskom, 2021). In this project, Seaborn facilitated the creation of clear and visually appealing charts that highlighted patterns, trends, and relationships within the startup funding data. Its capabilities, such as built-in themes, color palettes, and functions for complex visualizations (e.g., categorical plots, regression plots), made it an ideal tool for effective exploratory data analysis.

Each chart was generated using matplotlib and seaborn libraries (Hunter, 2007; Waskom, 2021), with clear titles and relevant commentary. Key charts included:

- Top Sectors by Total Funding: Highlighted the dominance of Biotech and Software sectors.
- Funding Trends by Sector: Illustrated sector-wise funding declines.
- Funding Per Round Analysis: Revealed sector competitiveness in funding efficiency.
- Historical Funding Trends: Exposed changes over more than a century.
- Funding Range Distribution: Showed predominant investment levels.
- Country-Level Analysis: Identified top funding hubs globally.

Conclusion

The analysis reveals significant sectoral and regional trends in startup funding. The Biotech sector stands out as a funding leader, while emerging sectors like Nanotech demonstrate high funding efficiency. Geographically, Bermuda, Luxembourg, Malaysia, and China are notable startup hubs.

Real-World Implications

- These findings have practical applications:
- Investors can target high-growth sectors and emerging hubs.

- Entrepreneurs can strategically position their startups in attractive regions.
- Policy Makers can design initiatives to boost startup ecosystems.
- Data Scientists can leverage such data for predictive modeling of funding trends.

Future studies could further enhance predictions using advanced machine learning models (Provost & Fawcett, 2013).

References

Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Elsevier.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

Kaggle. (n.d.). Startup Funding Dataset. Retrieved from https://www.kaggle.com

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Waskom, M. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021.