BACSE101 Problem Solving using Python

PROJECT REPORT

on

Startup Funding Analysis Project

Prepared by

Jeevitha.V.B -25BCE2345
Shangamithiraa.K.K-25BCE2338

Under the supervision of

Dr. GOUTAM MAJUMDER

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
Vellore Institute of Technology, Vellore.

Table of Contents

Abstract

This project focuses on analyzing startup funding data in India using Python and MySQL. It includes data cleaning, exploratory data analysis using Pandas, implementation of a menu-driven Python program, and integration with MySQL to store cleaned data. The analysis provides insights on top startups, active investors, cities, industries, and yearly funding trends, demonstrating skills in Python programming, data analytics, and database management.

In this project, we not only looked at funding trends in Indian startups but also focused on cleaning and organizing the data properly, which is a really important part of any real-world analysis. We handled missing values, fixed inconsistencies, and converted text into numbers to make the data usable. With this, we were able to find interesting patterns, like which industries got the most funding, which cities were leading in investments, and who the most active investors were. On top of that, by putting the cleaned data into a MySQL database, we made it easier to store, manage, and access the information whenever needed. This project shows how Python, data analysis, and databases can work together to give meaningful insights that could be useful for investors, entrepreneurs, or anyone interested in the startup ecosystem.

1. Introduction

Startups play a significant role in economic growth, innovation, and employment generation. Tracking funding patterns helps investors and entrepreneurs identify emerging sectors, successful startups, and active investors. This project performs a complete workflow from raw dataset cleaning to analysis and database integration.

1.1 Domain Information

The project uses a dataset of 3,000+ funding events for Indian startups. Each record contains information about startup name, investor name, funding amount, location, and industry vertical.

1.2 Software Libraries Used

- Python
- Pandas
- NumPy (if needed)
- MySQL Connector/Python

1.3 Contributions by Team Members

- Data cleaning and preprocessing: Jeevitha.V.B
- Exploratory Data Analysis: Jeevitha.V.B
- Menu-driven Python program implementation: Shangamithiraa.K.K
- MySQL database integration: Shangamithiraa.K.K

1.4 Challenges Faced

- Handling missing and inconsistent data in the funding amount column
- Converting dates and amounts into proper formats

2. Problem Statement and Objectives

Problem Statement:
The raw dataset contains inconsistent and missing data. Analysis without cleaning is unreliable. There is also no central database to store the dataset for queries.

Objectives:

1. Clean and preprocess the dataset
2. Analyze top startups, investors, cities, and industries by funding
3. Implement a menu-driven Python program for interactive queries
4. Store the cleaned dataset in a MySQL database

3. Implementation

3.1 Feature 1 – Data Cleaning & Preprocessing

Description:

This feature involves cleaning the raw startup funding dataset to make it suitable for analysis. It includes renaming columns, converting date columns to proper datetime format, and cleaning the funding amount column (Amount_USD) by handling missing values, removing commas, and converting values to numeric type.

Code:

```
import pandas as pd

data = pd.read_csv(r"C:\Users\balamurugan\Downloads\startup_fundings.csv")

data.head()
```

```
data.columns = [
    "Sr_No",
    "Date",
    "Startup_Name",
    "Industry_Vertical",
    "SubVertical",
    "City_Location",
    "Investors_Name",
    "Investment_Type",
    "Amount_USD",
    "Remarks"
]

data.head()
```

```
data["Date"] = pd.to_datetime(data["Date"], errors='coerce')
```

```
data["Amount_USD"] = data["Amount_USD"].astype(str).str.lower()


data["Amount_USD"] = data["Amount_USD"].replace(["undisclosed", "nan", "none"], None)


data["Amount_USD"] = data["Amount_USD"].str.replace(",", "", regex=True)
data["Amount_USD"] = pd.to_numeric(data["Amount_USD"], errors='coerce')
```

```
data.info()
```

```
data.to_csv(r"C:\Users\balamurugan\Downloads\startup_fundings_cleaned.csv", index=False)
```

Explanation:

- Columns renamed for consistency.
- Invalid dates are handled using errors='coerce'.
- Funding amounts cleaned to ensure only numeric values remain.

3.2 Feature 2 – Data Analysis & Insights

Description:
This feature performs various analytics on the cleaned dataset to generate insights such as top-funded startups, most active investors, funding by cities and industries, and year-wise funding trends.

Code:

```
import pandas as pd


df = pd.read_csv(r"C:\Users\balamurugan\Downloads\startup_fundings_cleaned.csv")

df['Date'] = pd.to_datetime(df['Date'], errors='coerce')


top_startups = df.nlargest(5, 'Amount_USD')[['Startup_Name', 'Amount_USD']]
print("\nTop 5 Startups by Funding:\n", top_startups)


active_investors = df['Investors_Name'].value_counts().head(5)
print("\nTop 5 Most Active Investors:\n", active_investors)

city_funding = df.groupby('City_Location')['Amount_USD'].sum().sort_values(ascending=False).head(5)
print("\nTop 5 Cities by Total Funding:\n", city_funding)

industry_funding = df.groupby('Industry_Vertical')['Amount_USD'].sum().sort_values(ascending=False).head(5)
print("\nTop 5 Industries by Total Funding:\n", industry_funding)

df['Year'] = df['Date'].dt.year
yearly_funding = df.groupby('Year')['Amount_USD'].sum().dropna().astype(int)
print("\nYear-wise Funding (in USD):\n", yearly_funding)
```

Explanation:

- Finds top-funded startups to identify leading companies.
- Determines the most active investors in the startup ecosystem.
- Aggregates funding by city and industry for trend analysis.
- Analyzes funding distribution year-wise to understand growth trends

3.2.1 Menu-Driven Analytics

Description:
Provides a simple menu for users to select which analysis they want to see. Each menu option corresponds to one insight.

Code:

```
def show_menu():
    print("\n--- Startup Funding Analysis Menu ---")
    print("1. Top 5 Startups by Funding")
    print("2. Top 5 Most Active Investors")
    print("3. Top 5 Cities by Total Funding")
    print("4. Top 5 Industries by Total Funding")
    print("5. Year-wise Funding Trend")
    print("6. Exit")

while True:
    show_menu()
    choice = input("Enter your choice (1-6): ")

    if choice == "1":
        top_startups = df.nlargest(5, 'Amount_USD')[['Startup_Name', 'Amount_USD']]
        print("\nTop 5 Startups by Funding:\n", top_startups)
    elif choice == "2":
        active_investors = df['Investors_Name'].value_counts().head(5)
        print("\nTop 5 Most Active Investors:\n", active_investors)
    elif choice == "3":
        city_funding = df.groupby('City_Location')['Amount_USD'].sum().sort_values(ascending=False).head(5)
        print("\nTop 5 Cities by Total Funding:\n", city_funding)
    elif choice == "4":
        industry_funding = df.groupby('Industry_Vertical')['Amount_USD'].sum().sort_values(ascending=False).head(5)
        print("\nTop 5 Industries by Total Funding:\n", industry_funding)
    elif choice == "5":
        df['Year'] = df['Date'].dt.year
        yearly_funding = df.groupby('Year')['Amount_USD'].sum().dropna().astype(int)
        print("\nYear-wise Funding (in USD):\n", yearly_funding)
    elif choice == "6":
        print("Exiting... Goodbye!")
        break
    else:
        print("Invalid choice. Please enter a number between 1 and 6.")
```

Explanation:

- Lets the user choose which insight they want to view.
- Makes the project interactive, satisfying the “menu-driven operations” requirement.

3.3 Feature 3 – MySQL Database Integration

Description:
This feature saves the cleaned dataset into a MySQL database. It creates a database, a table, and inserts all the cleaned records for future querying and analytics.

Code:

```
import mysql.connector
from mysql.connector import errorcode

db_name = "startup_funding_db"

try:
    conn = mysql.connector.connect(
        host="localhost",
        user="root",
        password="12345"   # Replace with your MySQL password
    )
    cursor = conn.cursor()
    print("Connected to MySQL!")
except mysql.connector.Error as err:
    print(f"Error: {err}")

try:
    cursor.execute(f"CREATE DATABASE IF NOT EXISTS {db_name}")
    print(f"Database '{db_name}' is ready!")
except mysql.connector.Error as err:
    print(f"Failed creating database: {err}")

conn.database = db_name

create_table_query = """
CREATE TABLE IF NOT EXISTS funding_data (
    Sr_No INT,
    Date DATE,
    Startup_Name VARCHAR(255),
    Industry_Vertical VARCHAR(255),
    SubVertical VARCHAR(255),
    City_Location VARCHAR(255),
    Investors_Name VARCHAR(255),
    Investment_Type VARCHAR(100),
    Amount_USD FLOAT
)
"""
cursor.execute(create_table_query)
print("Table 'funding_data' is ready!")
```

```
for i, row in df.iterrows():
    cursor.execute("""
        INSERT INTO funding_data (
            Sr_No, Date, Startup_Name, Industry_Vertical, SubVertical, City_Location,
            Investors_Name, Investment_Type, Amount_USD
        ) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)
    """, (
        int(row['Sr_No']),
        row['Date'].date() if pd.notnull(row['Date']) else None,
        row['Startup_Name'],
        row['Industry_Vertical'],
        row['SubVertical'],
        row['City_Location'],
        row['Investors_Name'],
        row['Investment_Type'],
        float(row['Amount_USD']) if pd.notnull(row['Amount_USD']) else None
    ))

conn.commit()
cursor.close()
conn.close()
print("Data inserted successfully and connection closed!")
```

4. Demo Screenshots

1.Data Cleaning Output

	Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount in USD	Remarks
0	1	09/01/2020	BYJU'S	E-Tech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	20,00,00,000	NaN
1	2	13/01/2020	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	80,48,394	NaN
2	3	09/01/2020	Mamaearth	E-commerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	1,83,58,860	NaN
3	4	02/01/2020	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	30,00,000	NaN
4	5	02/01/2020	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Round	18,00,000	NaN

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3044 entries, 0 to 3043
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Sr_No              3044 non-null   int64
 1   Date               1292 non-null   datetime64[ns]
 2   Startup_Name       3044 non-null   object
 3   Industry_Vertical  2873 non-null   object
 4   SubVertical        2108 non-null   object
 5   City_Location      2864 non-null   object
 6   Investors_Name     3020 non-null   object
 7   Investment_Type    3040 non-null   object
 8   Amount_USD         2065 non-null   float64
 9   Remarks            419 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(1), object(7)
memory usage: 237.9+ KB
```

2.Data Analysis Output

```
Top 5 Startups by Funding:
          Startup_Name    Amount_USD
60   Rapido Bike Taxi  3.900000e+09
651          Flipkart  2.500000e+09
830             Paytm  1.400000e+09
966          Flipkart  1.400000e+09
31              Paytm  1.000000e+09

Top 5 Most Active Investors:
 Investors_Name
Undisclosed Investors    39
Undisclosed investors    30
Ratan Tata               25
Indian Angel Network     23
Kalaari Capital          16
Name: count, dtype: int64

Top 5 Cities by Total Funding:
 City_Location
Bangalore    1.136159e+10
Bengaluru    7.098579e+09
Mumbai       4.921185e+09
New Delhi    3.017817e+09
Gurgaon      3.005296e+09
Name: Amount_USD, dtype: float64

Top 5 Industries by Total Funding:
 Industry_Vertical
Consumer Internet    6.253084e+09
eCommerce            5.002533e+09
Transportation       3.916632e+09
Technology           2.229708e+09
Finance              1.971438e+09
Name: Amount_USD, dtype: float64
```

```
Year-wise Funding (in USD):
 Year
2015.0    3587210368
2016.0    1527965000
2017.0    4702502000
2018.0    2301147824
2019.0    2854097895
2020.0     373158860
Name: Amount_USD, dtype: int64
```

3.Menu-Driven Analytics

```
--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  2

Top 5 Most Active Investors:
 Investors_Name
Undisclosed Investors    39
Undisclosed investors    30
Ratan Tata               25
Indian Angel Network     23
Kalaari Capital          16
Name: count, dtype: int64

--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  1

Top 5 Startups by Funding:
          Startup_Name    Amount_USD
60   Rapido Bike Taxi  3.900000e+09
651          Flipkart  2.500000e+09
830             Paytm  1.400000e+09
966          Flipkart  1.400000e+09
31              Paytm  1.000000e+09

--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  3

Top 5 Cities by Total Funding:
 City_Location
Bangalore    1.136159e+10
Bengaluru    7.098579e+09
Mumbai       4.921185e+09
New Delhi    3.017817e+09
Gurgaon      3.005296e+09
Name: Amount_USD, dtype: float64

--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  4

Top 5 Industries by Total Funding:
 Industry_Vertical
Consumer Internet    6.253084e+09
eCommerce            5.002533e+09
Transportation       3.916632e+09
Technology           2.229708e+09
Finance              1.971438e+09
Name: Amount_USD, dtype: float64
```

```
Top 5 Industries by Total Funding:
 Industry_Vertical
Consumer Internet    6.253084e+09
eCommerce            5.002533e+09
Transportation       3.916632e+09
Technology           2.229708e+09
Finance              1.971438e+09
Name: Amount_USD, dtype: float64

--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  5

Year-wise Funding (in USD):
 Year
2015.0    3587210368
2016.0    1527965000
2017.0    4702502000
2018.0    2301147824
2019.0    2854097895
2020.0     373158860
Name: Amount_USD, dtype: int64

--- Startup Funding Analysis Menu ---
1. Top 5 Startups by Funding
2. Top 5 Most Active Investors
3. Top 5 Cities by Total Funding
4. Top 5 Industries by Total Funding
5. Year-wise Funding Trend
6. Exit
Enter your choice (1-6):  6
Exiting... Goodbye!
```

5. Conclusion

The project successfully demonstrates:

- Cleaning and preprocessing real-world startup funding data
- Analyzing top startups, investors, cities, industries, and funding trends
- Implementing a menu-driven program using Python fundamentals
- Integrating data with MySQL database for query operations