**Exploratory Data Analysis (EDA) and Approach**

To develop the Founder-Investor Matching AI System, I began with **Exploratory Data Analysis (EDA)** on the investor data to understand its structure, distribution, and quality, which were crucial for preparing it for the matching algorithm.

**1. EDA Process:**

The EDA script is saved as **"vertex\_ai\_EDA.ipynb"**, which focuses on the following:

* **Data Inspection**: I loaded the dataset and checked its structure, columns, and data types.
* **Missing Values**: I identified missing values and summarized them for each column, then handled them by filling or removing values based on their importance.
* **Heatmap**: A **heatmap** was used to analyze correlations between numerical columns, helping to identify key relationships for the matching algorithm.

**2. Feature Engineering & Preprocessing:**

After EDA, I focused on preprocessing tasks:

* **Standardizing Categories**: I standardized inconsistent labels for industries and investment stages for uniformity.
* **Investment Range Conversion**: I transformed investment ranges (e.g., "$1M - $5M") into numerical values for better analysis.
* **Structured Data**: I parsed countries, industries, and stages into structured lists for better matching accuracy.

**3. Matching Algorithm:**

The **match score** is calculated based on several factors:

* **Industry Match (30%)**: Compares the founder’s industry with the investor’s focus.
* **Stage Match (20%)**: Compares the startup’s stage with the investor’s preferred investment stage.
* **Funding Match (25%)**: Checks if the required funding aligns with the investor’s range.
* **Geographic Match (10%)**: Considers the founder’s country and investor’s target geography.
* **Content Similarity (15%)**: Uses **TF-IDF** and **cosine similarity** to match the founder’s and investor’s descriptions.

**4.Gemini API Integration**

**The Gemini API enhances the matching accuracy between founders and investors by refining initial match scores with AI-driven insights.**

**How It Works:**

1. **Pattern Analysis: Gemini analyzes both the startup and investor profiles to identify deeper patterns, improving match accuracy.**
2. **Enhanced Match Score: The API refines the initial match score (0.0 to 1.0) by considering hidden compatibility factors not covered in the initial algorithm.**
3. **Reasoning Generation: It provides a brief explanation for the match, helping both founders and investors understand why they are compatible.**

**4. Evaluation and Saving Results:**

During the evaluation phase, I saved the matching results in a CSV file named "**startup.csv".** This file contains the top 10 investor matches for the startup, allowing me to easily review and analyze the best potential investors for my use case.

**5. Future Enhancements:**

Future improvements include:

* **Advanced NLP Models**: Integrating models like **BERT** to enhance semantic matching.
* **Historical Data**: Using past successful investments to improve accuracy.
* **Founder-Investor Personality Fit**: Adding personality analysis to refine matching.

.