Github Link: https://github.com/tkarim45/Data-Engineering/tree/main/Assignment%201

Group Number & Student IDs:

Group Number: Group 3Student 1 ID: 24280047Student 2 ID: 24280068

Contributions:

- 24280047: Implemented data collection scripts for Google Trends using Pytrends and Reddit API using PRAW.
- 24280068: Implemented data collection scripts for Reddit API using PRAW.
- Both: Conducted data preprocessing, initial analysis using pandas, and summarized key findings.

1. Overview of Our Topic

We chose to analyze trends in remote work, hybrid work, and work-from-home culture. Given the increasing shift in work environments post-pandemic, we aim to explore public interest trends and discussions regarding remote work. We expect to see:

- Increasing interest in work-from-home opportunities.
- High engagement in discussions about remote work challenges and benefits.
- Potential seasonal trends in job search activity.

2. Data Collection Process

Google Trends Data

- Used the pytrends library to extract search interest over time for keywords: "remote jobs," "hybrid work," and "work from home."
- Challenges: Google Trends has modified their bot detection bot which makes it difficult to make API calls through pytrends
- Addressed by selecting a verifies header of my own profile so google doesn't think it a request made by a library.

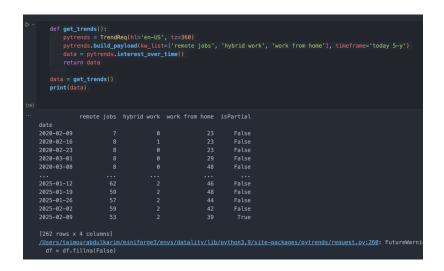
Reddit Data

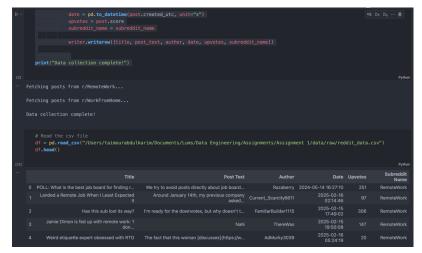
- Used the PRAW library to collect posts from RemoteWork and WorkFromHome.
- Extracted metadata: title, post text, author, date, upvotes, and subreddit name.
- Challenges: Some posts lacked text content, affecting analysis and Privacy concerns around personally identifiable information.

3. Initial Observations

We used pandas to generate basic summaries of our datasets:

- Google Trends: Search interest showed spikes around major events (e.g., pandemic lockdowns).
- **Reddit Data:** High engagement on topics related to productivity, isolation, and remote job scams.
- Kaggle Data: Increasing number of remote job postings over time.





4. Al Product Concept

Using this data, we aim to develop an Al-driven job trend analysis tool that:

- Identifies real-time shifts in remote work interest.
- Predicts future demand for remote jobs.

• Highlights key discussion themes from Reddit using NLP.

5. Terms of Service & Privacy Constraints

- Google Trends: Data can be used for analysis but should not be redistributed without proper attribution.
- **Reddit:** User-generated content cannot be stored indefinitely or republished without consent
- Kaggle: Public datasets may have their own licensing restrictions.

Mitigation:

- Store only aggregated insights rather than raw data.
- Follow API rate limits and adhere to platform-specific terms.

6. Data Quality & Challenges in Multi-Source Collection

- Benefits:
 - o Google Trends provides quantitative insights on search behavior.
 - o Reddit adds qualitative data from user discussions.
 - Kaggle offers structured datasets for validation.
- Challenges:
 - o Differences in update frequency (real-time vs. static datasets).
 - o Potential discrepancies between search interest and actual job postings.

7. Data Storage & Integration Strategy

- **Storage:** Use a relational database (PostgreSQL) or NoSQL (MongoDB) depending on data type.
- Integration:
 - o Google Trends: Time-series database.
 - o Reddit: Text-based storage with NLP processing.
 - o Kaggle: Standard relational format for structured analysis.
- ETL Pipeline: Automate data collection, cleaning, and merging.