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

Group Number & Student IDs:

- Group Number: Group 3
- Student 1 ID: 24280047
- Student 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.

```
def get_trends():
    pytrends = TrendReq(hl='en-US', tz=360)
    pytrends.build_payload(kw_list=['remote jobs', 'hybrid work', 'work from home'], timeframe='today 5-y')
    data = pytrends.interest_over_time()
    return data

data = get_trends()
print(data)
```

```
[16]
            remote jobs  hybrid work  work from home  isPartial
date
2020-02-09            7            0              23      False
2020-02-16            8            1              23      False
2020-02-23            8            0              23      False
2020-03-01            8            0              29      False
2020-03-08            8            0              48      False
...                 ...          ...             ...        ...
2025-01-12           62            2              46      False
2025-01-19           59            2              48      False
2025-01-26           57            2              44      False
2025-02-02           59            2              42      False
2025-02-09           53            2              39       True

[262 rows x 4 columns]
/Users/taimourabdulkarim/miniforge3/envs/datality/lib/python3.9/site-packages/pytrends/request.py:260: FutureWarni
  df = df.fillna(False)
```

```
            date = pd.to_datetime(post.created_utc, unit="s")
            upvotes = post.score
            subreddit_name = subreddit_name

            writer.writerow([title, post_text, author, date, upvotes, subreddit_name])

    print("Data collection complete!")
```

```
[3]                                                                                                    Python
Fetching posts from r/RemoteWork...

Fetching posts from r/WorkFromHome...

Data collection complete!
```

```
# Read the csv file
df = pd.read_csv("/Users/taimourabdulkarim/Documents/Lums/Data Engineering/Assignments/Assignment 1/data/raw/reddit_data.csv")
df.head()
```

```
[23]                                                                                                   Python
```

	Title	Post Text	Author	Date	Upvotes	Subreddit Name
0	POLL: What is the best job board for finding r...	We try to avoid posts directly about job board...	Razaberry	2024-05-14 16:27:10	251	RemoteWork
1	Landed a Remote Job When I Least Expected It	Around January 14th, my previous company asked...	Current_Scarcity6611	2025-02-16 02:14:46	97	RemoteWork
2	Has this sub lost its way?	I'm ready for the downvotes, but why doesn't t...	FamiliarBuilder1115	2025-02-15 17:49:02	306	RemoteWork
3	Jamie Dimon is fed up with remote work: 'I don...	NaN	ThereWas	2025-02-15 19:50:08	147	RemoteWork
4	Weird etiquette expert obsessed with RTO	The fact that this woman [discusses](https://w...	AdMurky3039	2025-02-16 05:24:19	20	RemoteWork

4. AI Product Concept

Using this data, we aim to develop an **AI-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:**
  - Google Trends provides quantitative insights on search behavior.
  - Reddit adds qualitative data from user discussions.
  - Kaggle offers structured datasets for validation.
- **Challenges:**
  - Differences in update frequency (real-time vs. static datasets).
  - 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:**
  - Google Trends: Time-series database.
  - Reddit: Text-based storage with NLP processing.
  - Kaggle: Standard relational format for structured analysis.
- **ETL Pipeline:** Automate data collection, cleaning, and merging.