Assignment 1 Data Engineering

Github link: https://github.com/Darab-Khan-1/DE-A1.git

1. Group Number, Student IDs, and Contributions

- Group Number: 30
- Student IDs: 24280036, 24280029
- Contributions:
    - Both of us worked almost on all the coding and analysis part but bellow is is a broad level breakdown
    - 24280029: Flowchart, Yahoo finance integration, analysis and public dataset from kaggle
    - 24280036:Reddit integration, Analysis of data and Report

2. Overview of the Topic

We chose the topic of Remote Work because we were curious to explore both the positive and negative aspects of remote work. Specifically we wanted to investigate how remote work affects organizational productivity compared to in-office work. By analyzing data from Reddit discussions, financial trends of remote work-related companies and public datasets on remote work trends, we aimed to gain an understanding of how it is affecting.

3. Data Collection Process

Reddit Data Collection

- API Used: PRAW (Python Reddit API Wrapper)
- Subreddit: r/RemoteWork
- Keywords: remote work, work from home, WFH, hybrid work
- Fields Collected: Title, post text, author, date, upvotes, subreddit
- Challenges:
    - We did not face API rate limits as we stayed within the free API quota.
    - Some posts had incomplete data, which we handled during the cleaning process.
- Storage: Data stored in datasets/raw/reddit_posts.csv.

Yahoo Finance Data Collection

- API Used: yfinance
- Stocks Analyzed: Zoom (ZM), Microsoft (MSFT), Salesforce (CRM), , Dropbox (DBX), Technology ETF (XLK)
- Data Collected: Closing stock prices over a two year period.
- Challenges:
    - Some stocks had missing values, which we cleaned during data processing.

- Storage: Data stored in datasets/raw/yfinance.csv.

Public Dataset

- Source: Kaggle (Remote Work Productivity Dataset)
- Link: https://www.kaggle.com/code/alaaabdelstar/remote-work-productivity
- Process:
    - We downloaded the dataset and cleaned it by removing missing values and irrelevant columns.
- Storage: Data stored in datasets/raw/public_data.csv.

---

4. Initial Observations

- Reddit Data:
    - We collected 100 posts for each keywork from the r/RemoteWork subreddit.
    - The dataset includes discussions on the benefits and challenges of remote work, such as productivity, work-life balance, and communication tools.

```
Collecting, cleaning and storing Reddit data
Searching for posts containing: remote work
Searching for posts containing: work from home
Searching for posts containing: WFH
Searching for posts containing: hybrid work
Reddit data collected and saved

First 5 rows:
                                        title                                          post_text            author         date  upvotes   subreddit
0  I'm going entirely insane trying to find a rea...  I've been applying on indeed and LinkedIn for ...  Sammy_Wants_Death  1.722877e+09      290  remotework
1           Where is everyone finding remote work?  I've applied to almost 100 remote positions wi...  Katieblahblahbloo  1.694366e+09      101  remotework
2  POLL: What is the best job board for finding r...  We try to avoid posts directly about job board...         Razaberry  1.715704e+09      250  remotework
3        Remote work is rarely what you think it is  I see so many say they're looking for entry le...     Deedle-Dee-Dee  1.693743e+09      174  remotework
4  Our employees aren't children. Spotify will co...                                                NaN   MarketsandMayhem  1.734997e+09    21129  remotework

Summary statistics for numeric columns:
              date     upvotes
count  4.000000e+02    400.0000
mean   1.722730e+09    505.1875
std    2.052666e+07   1440.7476
min    1.608139e+09      0.0000
25%    1.716474e+09      2.0000
50%    1.730315e+09     87.5000
75%    1.735968e+09    522.0000
max    1.739548e+09  21129.0000

Count of posts by keyword (title):
remote work: 101 posts
work from home: 52 posts
WFH: 97 posts
hybrid work: 35 posts
```

- Yahoo Finance Data:
    - We analyzed the stock performance of companies like Zoom and Microsoft over the past two years.

```
Collecting, cleaning and storing Yahoo Finance data
Yahoo Finance data collected, cleaned and stored

First 5 rows:
                                ZM        MSFT         CRM      GOOGL        DBX         XLK
Date
2023-02-15 00:00:00-05:00  80.769997  265.635223  170.149643  96.589851  24.170000  141.878128
2023-02-16 00:00:00-05:00  77.910004  258.563293  167.176254  95.165016  23.959999  139.416519
2023-02-17 00:00:00-05:00  76.110001  254.529266  164.252594  94.009209  21.219999  137.673706
2023-02-21 00:00:00-05:00  72.470001  249.212982  160.722290  91.458458  21.290001  134.404633
2023-02-22 00:00:00-05:00  73.389999  248.068863  162.452621  91.318962  21.219999  134.207718

Summary statistics for numeric columns:
              ZM        MSFT         CRM       GOOGL         DBX         XLK
count  502.000000  502.000000  502.000000  502.000000  502.000000  502.000000
mean    68.572450  375.588349  253.067385  146.870040   25.705139  193.730576
std      7.374509   55.938853   48.396058   27.608970    3.372604   30.817150
min     55.320000  242.900574  160.722290   88.808060   18.969999  133.439697
25%     63.289999  327.560425  209.208984  128.676044   23.139999  167.853516
50%     67.790001  400.661346  252.159111  142.194527   25.815000  199.666397
75%     71.309999  419.505447  287.726715  168.886848   27.887499  222.751789
max     89.029999  465.786438  367.450745  206.380005   33.160000  241.820007

Stock performance over time:
ZM:
  - Initial price: 80.77
  - Final price: 85.10
  - Percentage change: 5.36%
MSFT:
  - Initial price: 265.64
  - Final price: 408.43
  - Percentage change: 53.76%
CRM:
  - Initial price: 170.15
  - Final price: 326.54
  - Percentage change: 91.91%
GOOGL:
  - Initial price: 96.59
  - Final price: 185.23
  - Percentage change: 91.77%
DBX:
  - Initial price: 24.17
  - Final price: 32.78
  - Percentage change: 35.62%
XLK:
  - Initial price: 141.88
  - Final price: 239.97
  - Percentage change: 69.14%
```

- Public Dataset:
    - The dataset includes trends and statistics related to remote work productivity.

```
Collecting, cleaning and storing public data
Public data collected and saved

First 5 rows:
   Employee_ID Employment_Type  Hours_Worked_Per_Week  Productivity_Score  Well_Being_Score
0            1          Remote                     29                  75                78
1            2       In-Office                     45                  49                47
2            3          Remote                     34                  74                89
3            4          Remote                     25                  81                84
4            5          Remote                     50                  70                74

Summary statistics for numeric columns:
       Employee_ID  Hours_Worked_Per_Week  Productivity_Score  Well_Being_Score
count  1000.000000            1000.000000         1000.000000       1000.000000
mean    500.500000              39.720000           68.602000         63.975000
std     288.819436               8.042779           12.235494         13.870572
min       1.000000              16.000000           33.000000         14.000000
25%     250.750000              34.750000           60.000000         56.000000
50%     500.500000              40.000000           68.000000         65.000000
75%     750.250000              45.000000           76.000000         73.000000
max    1000.000000              64.000000          112.000000        104.000000

Count of unique values for categorical columns:
Employment_Type: 2 unique values
```

5. AI Product

A recommendation system can be developed to suggest remote work tools based on user preferences. For example the system could recommend tools like Zoom, Slack, or Microsoft Teams based on the user's needs and preferences. This system can utilize Reddit data to understand user preferences and the Yahoo Finance data to evaluate the financial performance of these tools.

6. Terms of Service Constraints & Privacy Issues

- Reddit:
    - Reddit's API has rate limits which we respected by staying within the free quota.
    - We ensured that we did not store or redistribute any personally identifiable information from Reddit posts.
- Yahoo Finance:
    - The yfinance library provides publicly available stock data, so there were no significant privacy concerns.

- We ensured that we complied with Yahoo Finance's terms of service by not using the data for commercial purposes.

---

7. Data Quality

Collecting data from multiple sources provides a more comprehensive view of the topic. However, it can also lead to discrepancies, such as differences in data formats or missing data. For example:

- The Reddit data was unstructured, while the Yahoo Finance data was structured, making it challenging to combine the two datasets.
- We addressed these challenges by cleaning the data and standardizing the formats before analysis.

8. Data Storage & Combination

- Storage:
  - We stored the data in CSV files within a structured folder as mentioned in the assignment instructions. There are several other ways for storing it as well such as relational, non relational databases or maybe storage spaces like s3 bucket.
- Combination:
  - We can first standardize data using pandas to clean and combine the datasets. For example, we can combine the Reddit and public datasets to analyze the relationship between remote work preferences and challenges.

Flow Chart Diagram