AI 601 - Assignment 1: Project Data Mosaic

Muhammad Huraira Anwer
Department of Computer Science
Lahore University of Management Sciences
25100314@lums.edu.pk

Muhammad Ahmad
Department of Electrical Engineering
Lahore University of Management Sciences
25100076@lums.edu.pk

GitHub Repository: https://github.com/muhmmad-ahmad-1/ai601-g11

1 Overview

The rise of remote work has been one of the most significant workplace transformations in recent years. While remote work existed before the COVID-19 pandemic, it saw an unprecedented surge as organizations worldwide were forced to adapt to lockdowns. Consequently, businesses rapidly embraced remote and hybrid work models. While initially a necessity, remote work soon became a preferred mode of work for many employees due to its flexibility, cost savings, and improved work-life balance and this preference is maintained after the pandemic is over.

As about-to-be fresh graduates, this shift in job trends (particularly in the tech industry) is very relevant to us. With many companies now offering remote or hybrid positions, understanding the long-term trajectory of remote work is crucial for shaping our career decisions and adapting to the evolving job market. Consequently, we chose the topic of `'remote work'`.

We expect to observe continued interest in remote job opportunities, reflected in online search trends, social media discussions, and employee sentiment surveys, and a consequent continued increase in remote job opportunities even after the end of the COVID-19 pandemic. Given the widespread adoption of remote work during the pandemic, we anticipate that companies will continue offering flexible work arrangements due to their benefits, such as cost savings, access to a global talent pool, and improved work-life balance for employees. We aim to examine whether this trend has sustained its momentum or if organizations have shifted back to traditional in-office roles. In the context of this assignment, a data pipeline was employed, which gathered historical Reddit posts, and Google trends data, as well as data from a public survey on employee well-being and health under remote work. Fig. 1 shows the broad data pipeline.

2 Data Collection and Initial Observations

2.1 Reddit

We consulted the PRAW documentation as a reference for extracting Reddit posts. Specifically, we used this API to retrieve posts related to remote work discussions from relevant subreddits.

To access the Reddit API, we first registered an `'app'` using an existing Reddit account, which provided us with a *client id* and *client secret*. These credentials were then used to initialize a `praw.Reddit` class object in Python. The data collection step involved selecting the most relevant subreddits—`'r/workfromhome'` and `'r/remotework'`—based on a preliminary search of trending discussions. Using the `reddit.subreddit` method, we extracted the latest 200 posts containing relevant keywords.

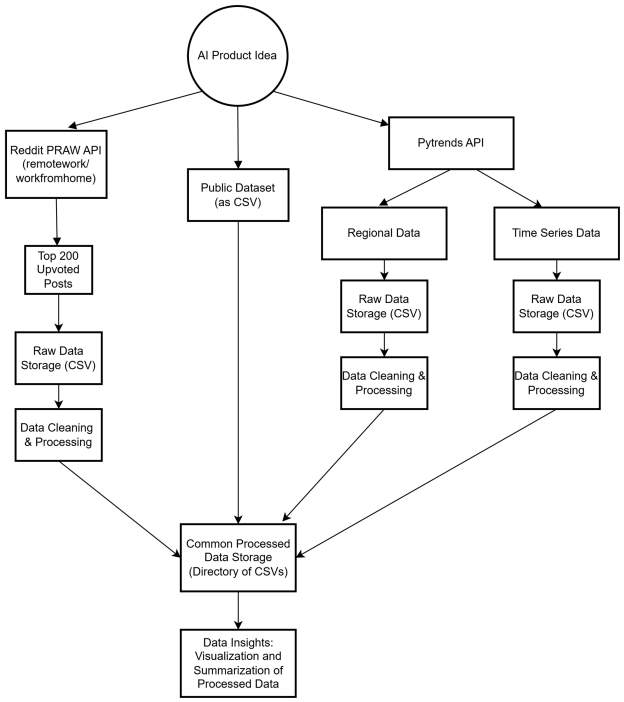

Figure 1: Broad Data Pipeline

The extracted data was stored in CSV format for further analysis. Since most fields in Reddit posts were non-empty, minimal preprocessing was required. The only major cleaning step involved handling missing post text by replacing empty strings with a 'No text' identifier.

The processing step was followed by summarization and visualization, which involved:

- Computing numerical summary statistics using `df.describe()` on the Pandas DataFrame.
- Generating line plots for word frequency in post titles and post texts.
- Creating a word cloud to analyze the most commonly used words in discussions.

Unlike other web scraping tools, PRAW did not impose API rate limits or trigger Terms-of-Service constraints, allowing smooth data retrieval, persumably since it is a well-maintained wrapper around Reddit's official API. Fig. 2 displays the first five rows of the dataset. Similarly, Table.1 provides its summary.

	Title	Post Text	Author	Date	Upvotes	Subreddit
0	Our employees aren't children. Spotify will co...	No text	MarketsandMayhem	2024-12-23 23:43:44	21118	remotework
1	You guys need to cherish your WFH life, I got ...	I'm angry all the time now, everyone is notici...	ZadarskiDrake	2024-12-05 12:14:09	10653	remotework
2	WFH is more efficient for the federal governme...	No text	Background-War9535	2025-01-29 15:49:44	10089	remotework
3	Long live Spotify	No text	vizzy_vizz	2025-01-19 16:46:40	8195	remotework
4	I interviewed, accepted, and signed contract f...	I've already reiterated to them that this was ...	bee_antlers	2025-01-06 23:33:45	6395	remotework

Figure 2: First 5 rows of Reddit dataset

2.2 Pytrends

We consulted the Pytrends documentation and repository as reference for generating the script. Specifically, we used the unofficial API for two types of data, 1) the weekly variation of the historical interest over time and 2) the interest by region, both over 12 months, with cumulative trends globally (rather than specific to a certain country) for the following keywords: *'remote work', 'work from home', 'remote job', and 'hybrid work'*. This was done by first building the payload (using the

Statistic	Upvotes
Count	200.000000
Mean	1510.640000
Std Dev	1964.266347
Min	590.000000
25%	724.500000
50% (Median)	942.500000
75%	1499.750000
Max	21118.000000

Table 1: Summary Statistics for Upvotes of Reddit Posts

`build_payload` method of the `TrendReq` class) and then invoking the `interest_over_time` and `interest_by_region` methods. The datasets provided by these methods are by default pandas Dataframes and were therefore easily stored in CSV format as two separate files (`pytrends` and `pytrends_regional`).

The collection step was followed by a processing step which involved the following:

- removal of columns not needed (**isPartial**)
- removal of columns with NaN (empty) entries
- removal of all zero rows (relevant mainly to regional analysis where such entries were prevalent)

This resulted in cleaned, final processed data frames which were stored in CSV formats (`cleaned_pytrends` and `cleaned_pytrends_regional`) like their preprocessed versions.

The final step was summarization and visualization of the data in each CSV. This was done by reporting the average interest of each keyword over the 1 year time window (as pure numbers) as well as the evolution (variation over time) of the interest for each (as line graphs). For regional analysis, the top regions with the highest interest in each keyword were reported in both (5) textual form as well as (20) as bar plots. This unofficial API was a substantial challenge (in getting it to work consistently) due to Google's security measures against such scraping and querying (anti-bot measures). This resulted in a TooManyRequests error when Google detected unusual use (which in normal browser use triggers a Captcha). Further details of this challenge will be given in section 4.

Fig.3,Table.2 and Fig.4, Table.3 provide the details of the two datasets after cleaning.

	date	remote work	work from home	remote job	hybrid work
0	2024-02-11	26	89	16	4
1	2024-02-18	26	90	19	4
2	2024-02-25	26	95	18	4
3	2024-03-03	26	90	17	4
4	2024-03-10	25	91	16	4

Figure 3: First 5 rows of Pytrend dataset

	geoName	remote work	work from home	remote job	hybrid work
0	Argentina	36	37	23	4
1	Australia	19	72	5	4
2	Austria	33	18	46	3
3	Bangladesh	14	43	42	1
4	Belgium	34	30	31	5

Figure 4: First 5 rows of Pytrends Regional dataset

Statistic	Remote Work	Work From Home	Remote Job	Hybrid Work
Count	53.000	53.000	53.000	53.000
Mean	26.906	89.245	17.472	4.038
Std	2.937	6.189	1.804	0.390
Min	22.000	69.000	12.000	3.000
25%	25.000	85.000	16.000	4.000
50%	26.000	90.000	17.000	4.000
75%	27.000	94.000	19.000	4.000
Max	41.000	100.000	21.000	5.000

Table 2: Summary Statistics for Pytrends Dataset

Statistic	Remote Work	Work From Home	Remote Job	Hybrid Work
Count	55.000	55.000	55.000	55.000
Mean	27.891	45.200	22.873	4.036
Std	9.354	17.317	12.456	1.885
Min	5.000	17.000	5.000	1.000
25%	21.500	34.000	13.500	3.000
50%	29.000	43.000	20.000	4.000
75%	35.000	58.500	30.500	5.000
Max	50.000	89.000	65.000	9.000

Table 3: Summary Statistics for Pytrends Regional dataset

2.3 Public Dataset

We searched Kaggle to find public datasets on our topic and decided on Remote Work and Mental Health dataset. The "dataset dives into how working remotely affects stress levels, work-life balance, and mental health conditions across various industries and regions" by surveying 5000 employees from different job sectors. These statistics together with gauging the interest in remote work globally and regionally, conceptualize a data pipeline that could help build an AI model that leverages this data to forecast interest across regions. Combined with the reviews of the employees currently engaged in remote/hybrid work, the model can make recommendations to a global company on increasing or decreasing remote jobs in different regions. We further discuss this in the next section. Table 4 and 5 show the first 5 rows of this public dataset on remote jobs and mental health. Table 6 shows the summary statistics of the dataset.

ID	Age	Gender	Job Role	Industry	Exp	Location	Hours	Meetings	Balance	Stress
EMP0001	32	Non-binary	HR	Healthcare	13	Hybrid	47	7	2	Medium
EMP0002	40	Female	Data Scientist	IT	3	Remote	52	4	1	Medium
EMP0003	59	Non-binary	Software Engineer	Education	22	Hybrid	46	11	5	Medium
EMP0004	27	Male	Software Engineer	Finance	20	Onsite	32	8	4	High
EMP0005	49	Male	Sales	Consulting	32	Onsite	35	12	2	High

Table 4: General Employee Data

No processing or cleaning of the public dataset was required since it was already well-structured and processed. For drawing insights from the data, there was an additional challenge since there were numerous fields, which might be categorical or numerical. Hence, we provide pie charts (for categorical) and bar charts (for numerical data or data with fewer categories) summarizing the key insights from the survey. We provide these figures in the section 6.

3 AI Product

As can be observed from the previous sections, we have four different kinds of information: a time series data showing the popularity of keywords relating to remote work globally (quantified by the relative number of searches of that keyword), global statistics on search hits over a period of an year across multiple countries globally, a text based dataset of the most popular reddit posts relating

ID	Mental Health	Resources	Productivity	Isolation	Satisfaction	Support	Activity	Sleep	Region
EMP0001	Depression	No	Decrease	1	Unsatisfied	1	Weekly	Good	Europe
EMP0002	Anxiety	No	Increase	3	Satisfied	2	Weekly	Good	Asia
EMP0003	Anxiety	No	No Change	4	Unsatisfied	5	NaN	Poor	North America
EMP0004	Depression	Yes	Increase	3	Unsatisfied	3	NaN	Poor	Europe
EMP0005	NaN	Yes	Decrease	3	Unsatisfied	3	Weekly	Average	North America

Table 5: Employee Mental Health & Well-being Data

Statistic	Age	Experience	Hours/Week	Meetings	Work-Life Balance	Company Support
Count	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00
Mean	40.995	17.810	39.615	7.559	2.984	3.008
Std	11.296	10.020	11.860	4.636	1.410	1.399
Min	22.000	1.000	20.000	0.000	1.000	1.000
25%	31.000	9.000	29.000	4.000	2.000	2.000
50%	41.000	18.000	40.000	8.000	3.000	3.000
75%	51.000	26.000	50.000	12.000	4.000	4.000
Max	60.000	35.000	60.000	15.000	5.000	5.000

Table 6: Summary Statistics for Public Dataset

to remote work, and a structured survey with numerous categorical and numerical data relating to the emotional well-being of employees doing remote work or hybrid work. So there is a mix of structured, and unstructured datasets, and within datasets themselves there is numeric and categoric information. Given the diverse datasets available, several AI-driven products can be developed.

The simplest utilization of this information is the popularity analysis of remote work in prospective employees, combined with sentiment analysis of current employees engaged in remote/hybrid work providing real-time and historical sentiment trends on remote work. This can be presented as a dashboard if the application is relevant to the corporate sector or as a web-based platform with reports and visual analytics to help a larger community (businesses, policymakers, and researchers) track evolution of remote work worldwide. A possible method of utilizing this information would be using the time series pytrends data to identify peaks and dips in remote work interest across regions and across time, applying natural language processing using LLMs e.g. on Reddit posts to extract key concerns and discussions, and correlating these insights with survey responses to detect shifts in employee sentiment globally.

These insights can further allow a recommendation system to be built suggesting remote work policies optimized for engagement and well-being as well as highlighting key regions or demographics which prefer remote work, suggesting consequent provision of such flexibility to accommodate current employees and attract further employees.

4 Challenges

4.1 Challenges faced during use

The praw API was easy to use since its a wrapper of the official reddit API. We faced no rate limit issues due to the smaller scale of our requests (a single query of 200 posts [with limit being of 1000 posts]) which did not trigger any ratelimit error as Reddit supports up to 10 queries per minute without OAuth authentication and 100 with authentication. This is very generous and allows further scaling. Even if there is immense scaling, the API will still remain functional by a built in abiding of the timeout notified by the Reddit server upon hitting a ratelimit. Pytrends, however, was a different story. Pytrends posed several challenges due to its status as an unofficial API. Since Google Trends does not offer a public API, Pytrends relies on web scraping techniques that are subject to Google's anti-bot measures. Due to Google's security mechanisms, repeated requests for the same data sometimes returned different results or missing values. Additionally, on suspected bot activity, Error 429 was easily (almost too easily) triggered claiming Too Many Requests being made even if the frequency of requests was very low (again due to the anti-bot detection). For this, we had to retry several times, and wait long periods for a few minutes of successful requesting and scraping. Timeouts or retries

didn't work nor did proxy based solutions so we had to resort to manual as well as loop-based retry methods which did eventual allow for somewhat consistent data fetching.

While PRAW was stable and scalable, Pytrends required constant monitoring and adjustments to ensure smooth data extraction. An unofficial alternative of pytrends is required to reliably fetch this trends data (or we might have to resort to the alternative of waiting for the long announced official Google Trends API release).

4.2 Potential service constraints and privacy issues

There are pertinent privacy and ethical issues on utilizing these data sources (particularly pytrends and Reddit PRAW API).Reddit's API prohibits the redistribution of raw post content and its API use does not give license to use the "user posts for other purposes, such as for training a machine learning or AI model, without the express permission of rightsholders in the applicable User Content" which amounts to seeking permission or verifying permission on a user to user basis. The API also does not permit modifying the User Content except to format it for such display. Consequently, transformations or augmentations and subsequent utilization for training an AI model are not supported and allowed legally. There is a further privacy issue due to user-generated posts containing personally identifiable information, requiring ethical considerations when storing and analyzing data. Preprocessing of data may involve masking of such personal information before use downstream.

In contrast, pytrends or the public dataset does not have such issues. Pytrends API reports coarse grained information on a regional level which does not contain personal information. The major issue is that it is an unofficial API which is not supported by Google itself, which raises questions on the legality of using it as a data source, since it already faces restrictions from Google, including rate limits and temporary IP bans. The public dataset is, however, sufficiently anonymized and therefore has little ethical and legal issues on its use. The dataset also has no license prohibiting its use in AI model training explictly.

4.3 Effect of Multiple Sources

4.3.1 Benefits

Combining multiple data sources improves the robustness of our analysis by allowing for cross-validation and contextual insights. Specifically, each dataset provides a different resolution for gauging popularity of remote work: Reddit provides real-world discussions, Google Trends reflects search interest, and Kaggle data offers structured survey responses. While each source has inherent biases — social media favors active keyboard users, search data shows intent but not sentiment, and surveys may reflect self-reporting bias - however, combining these sources mitigates these issues and allows a holistic view on the popularity of remote work, for analysis; drawing insights and providing recommendations.

4.3.2 Limitations

However integrating these sources is not straightforward. One of the major issues is the timeline and the resolution of each dataset. The public dataset is static over a single survey period and cannot be updated (update will mean starting a new survey), Reddit and Google Trends however provide real-time data, which is dynamic in nature. We also have the issue of different formats (categorical, numeric, or text-based, time-dependent, or time-invariant) as well as different structures (structured, semi-structured) of datasets. Reconciling these differences and combining to allow for a overall analysis requires significant pre-processing.

4.3.3 Potential Conflicts and Discrepancies

In the worst case, there is also a potential of conflicts between the different data sources due to the different in the resolution of the voices they capture as well as the presence or absence of sentiment from the different sources. For instance, Google Trends might show increasing search interest in remote jobs, but Reddit discussions could reveal growing dissatisfaction with remote work. Similarly, the public survey data might indicate strong support for remote work (which is specific to the current time the survey was held and the demographic surveyed), but online discussions could highlight concerns about isolation or burnout (relevant to a different demographic and time). Similarly, the

different kinds of data (qualitative for Reddit and survey and quantitative for Google trends) requiring different analytical approaches demands additional processing to reconcile potential conflicts. These issues might act as significant barriers in drawing insights as well as reliably training an AI model.

5 Combined Data Storage

To effectively store and integrate these datasets, we can leverage a combination of SQL, NoSQL, and Data Warehousing techniques. The structured public dataset can be efficiently stored in a relational database like MySQL, as it follows a well-defined schema with attributes like age, gender, job role, and experience. The semi-structured Reddit dataset, which contains textual content, is better suited for a NoSQL database (MongoDB, Elasticsearch) due to its flexible schema and the need for text indexing to support search and analysis. The aggregated Pytrends dataset, which consists of numeric values categorized by location, aligns well with a data warehouse or a relational database, making it easily joinable with the structured employee dataset through the common '*location*' attribute.

While structured and aggregated data can be seamlessly integrated using SQL-based relational joins, the semi-structured Reddit posts pose a challenge due to their unstructured text fields. A hybrid approach, such as storing structured data in SQL and linking it with NoSQL-stored textual data via a common attribute like location or date, would enable effective querying. However, directly merging all datasets into a single table is not feasible because Reddit posts lack structured attributes like '*id*', '*experience*', and '*weekly hours*', which exist in the employee dataset. Instead, a Data Lake can be used to store all data in raw form, allowing structured and unstructured datasets to coexist while enabling flexible transformations when needed. A data warehouse can then process and store cleaned, structured data for analytics, while the raw text data remains in a NoSQL system for specialized querying.

In summary, structured datasets can be combined through SQL-based joins, and semi-structured Reddit data can be linked through shared attributes (region, date, etc.) but is best stored separately in NoSQL for text processing. A Data Lake is an ideal overarching solution, ensuring that both structured and unstructured data remain accessible, while a data Warehouse can be used on top of that for analysis. This hybrid approach balances flexibility and efficiency, ensuring optimal data storage and integration across different formats.

6 Visualization

The following is the word frequency chart created to visualize the most spoken words from the Reddit posts.

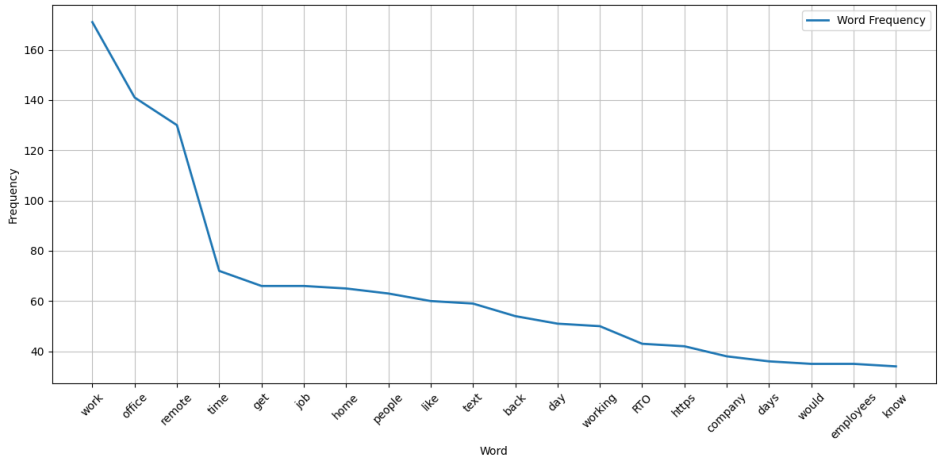

Figure 5: Word Frequency of Reddit posts

Next up is the line chart that depicts of how interest evolved overtime in different keywords related to remote work from the Pytrends dataset.

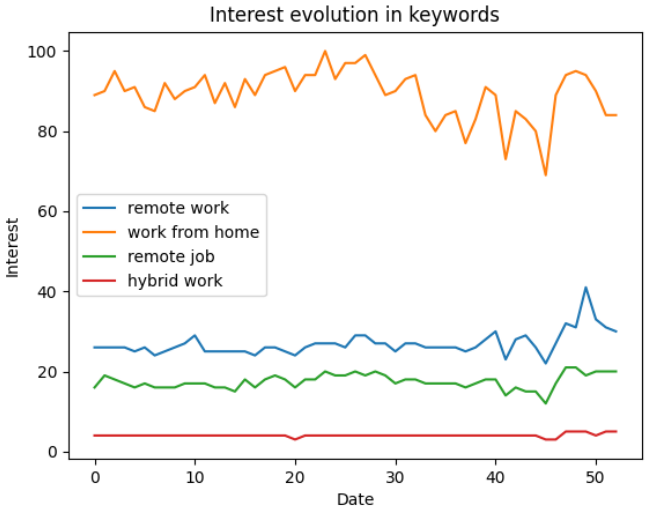

Figure 6: Interest evolution of Pytrends dataset

After that, is the bar chart that depicts the interest in different keywords region-wise from the Pytrends regional dataset.

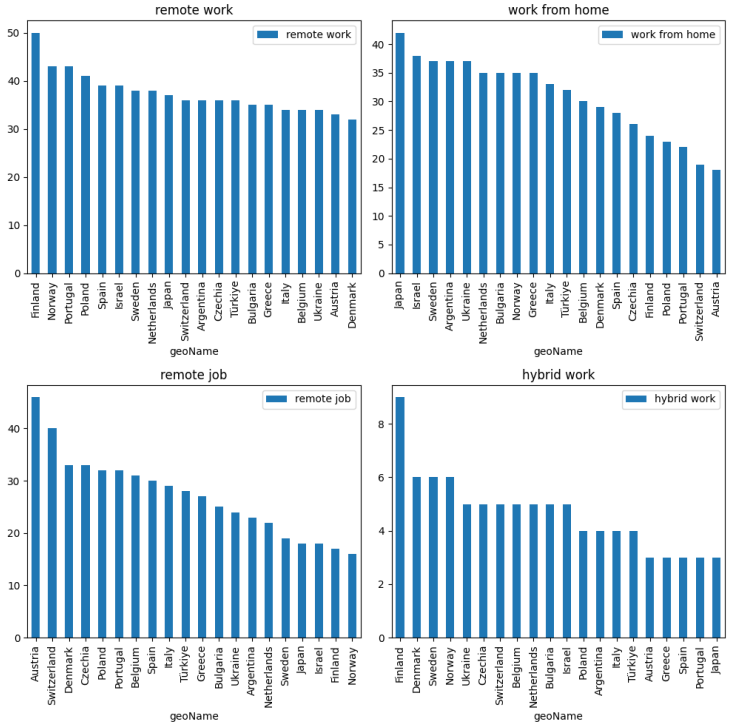

Figure 7: Region wise Interest of Pytrends (regional) dataset

Finally, we present multiple pie charts to depict the range of data points in each attribute from the public dataset.

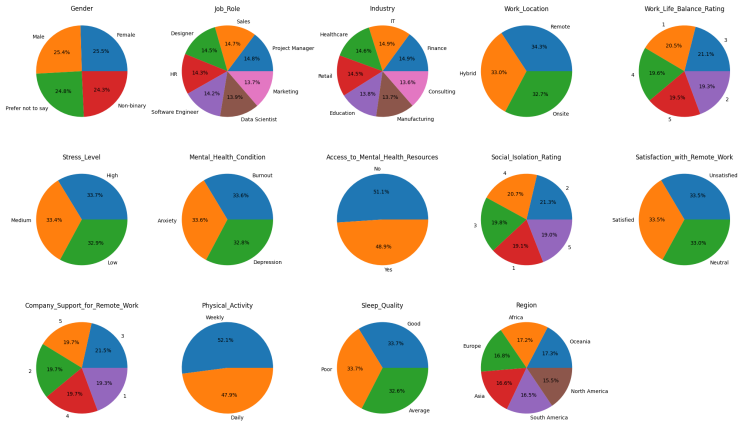

Figure 8: Pie charts of Public dataset

7 Contributions

Student	Task
Huraira Anwer	Reddit
Muhammad Ahmad	Pytrends
Both [Equal Contribution]	Public Dataset, Report

Table 7: Contributions