Syracuse University
School of Information Studies

IST652 Final Project Report

Analyzing present day Job Market Trends

Group 8

Sharif Bey
Archit Dukhande
Harika Gangu

Contributions:

Sharif Bey – *Conceived the original project idea and led the overall design of the project structure.*
Led the drafting and structuring of the final project report.
Troubleshot broken or malfunctioning code during development and helped debug errors.
Contributed to writing small parts of the notebook code, especially assisting with data processing and fixing issues.

Archit Dilip Dukhande – *Led the data collection and parsing of Hacker News "Who is Hiring?" threads using Python, building a pipeline to clean and structure semi-structured postings.*
Designed and implemented core NLP methods for job title extraction, skill identification, and location parsing.
Troubleshot and debugged Python scripts during development to ensure accurate data extraction and processing.

Harika Gangu – *Prepared project documentation, including organizing and formatting the markdown report.*
Contributed to structuring the report sections, proofreading content, and ensuring that the project outputs and analyses were clearly presented for submission.
Supported the overall project delivery by helping consolidate code outputs and visualizations into the final deliverables

Purpose

This project aims to analyze tech job market trends by collecting and examining both **current** and **historical** job postings data. The focus is on **remote Data/ML jobs**, aligning with our initial proposal to study in-demand skills, salary ranges, and job title trends in the tech industry. By leveraging two data sources (a live job API and historical community postings), we seek to address key questions:

- **In-demand skills and job titles:** What skills are most frequently requested, and which job titles are most common in data and tech roles?
- **Salary trends:** How do salaries for data/ML jobs vary, and what is the distribution of salaries in remote tech roles?
- **Trends over time:** How has the demand for remote tech jobs changed from 2020 to 2024 (e.g. growth of remote work, emergence of AI roles)?
- **Remote vs. on-site context:** (Originally proposed) How do remote job trends compare to on-site jobs? *(Our analysis primarily covers remote positions due to data availability.)*

By answering these questions, the project provides insights into the evolving job market, highlighting which skills are hot, how job roles have shifted, and how external factors (like the pandemic, Tech layoffs and AI boom) influenced hiring.

Data Sources

Our analysis draws from **two main data sources**, combining real-time data with historical records:

- **Remotive API (Live Remote Jobs, 2024):** We accessed Remotive.com’s public API to retrieve current remote job postings. This API returns a JSON feed of job listings worldwide. From this feed, we specifically filtered for jobs in data science, machine learning, AI, and analytics domains. Each job entry includes fields such as job title, company name, category, tags (skills/technologies), job type (full-time, etc.), salary (if provided), and publication date. Using the API ensured we have a **snapshot of the present** remote tech job market without needing to scrape web pages. For example, the Python snippet below shows how we fetched and loaded the Remotive data into a pandas DataFrame:

  ```
  url = "https://remotive.com/api/remote-jobs"
  response = requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})
  data = response.json()
  df = pd.DataFrame(data['jobs'])
  ```

- **Hacker News “Who is Hiring?” Threads (2020–2024):** To capture historical trends, we utilized monthly "Ask HN: Who is hiring?" discussion threads on Hacker News (a popular tech forum). In these threads, employers post job ads and often mention if roles are remote. We used the Algolia API for Hacker News to programmatically search for all "Who is hiring?" posts from 2020 through 2024 and collect their comments. Each comment typically contains a job description, which we treat as an unstructured job posting (often including the role, some required skills, location or remote indicator, etc.). This source provides a **community-driven dataset** of tech job postings over five years. We aggregated all comments from these threads, then filtered to those related to our focus (remote data/ML jobs) using keyword matching (e.g. comments containing "remote", "data", "machine learning", "AI", etc.).

Together, these data sources provide a rich foundation: the Remotive API gives a structured, up-to-date list of remote jobs, while the HN threads provide longitudinal data to observe how job postings and requirements evolved over time. All data collected was stored in structured form (pandas DataFrames, and intermediate CSV files for processed outputs).

Preprocessing

Before analysis, several **preprocessing steps** were necessary to clean and integrate the data from the sources:

- **Filtering relevant records:** From the Remotive dataset (df), we extracted only job postings relevant to data science and machine learning. For example, we filtered titles containing keywords like "data", "machine learning", "ML", "AI", or "analytics" (case-insensitive). This gave a subset df_data of remote jobs in our domain of interest. Similarly, for the HN data, after collecting all comments, we filtered the dataset to retain only comments that mention our keywords of interest (to find posts likely describing data/ML jobs, many of which also mention remote work explicitly).
- **Cleaning text data (HN comments):** Hacker News comments are unstructured HTML text (they can contain HTML tags, links, etc.). We wrote a cleaning function to strip HTML tags and non-alphanumeric characters, convert text to lowercase, and remove extraneous whitespace. We also removed URLs and punctuation to focus on actual words. This was done using Python's BeautifulSoup (to remove HTML tags) and regex substitutions for symbols and links. Below is a snippet of the text-cleaning function used on HN comment text:

```
from bs4 import BeautifulSoup
import re

def clean_html_and_symbols(text):
  if pd.isna(text):
    return ""
  text = BeautifulSoup(text, "html.parser").get_text()  # remove HTML tags
  text = re.sub(r"http\S+|www\S+", "", text)          # remove URLs
  text = re.sub(r"[^\w\s]", " ", text)               # remove special chars
  text = re.sub(r"\s+", " ", text)                   # collapse whitespace
  return text.strip().lower()
```

We applied this function to every comment in the filtered HN dataset, resulting in a clean text field ready for analysis (e.g. word frequency counting). This **cleansing** ensured that noise like HTML entities or punctuation would not skew our Natural Language Processing (NLP) of the job descriptions.

- **Parsing and formatting fields:** The Remotive data was already structured, but some fields needed transformation. In particular, **salary data** from Remotive came as strings (often ranges or formatted with currency symbols). To analyze salaries quantitatively, we created a parsing function to extract numeric values. For example, a salary string "$80,000-$100,000" was converted to a numeric average (90000). We used regular expressions to find numbers and took an average for ranges or left it as a single value if only one number was present. We then added a new column parsed_salary to the DataFrame for analysis. We also dropped entries without salary info or where parsing failed (to focus on meaningful salary data).
- **Date formatting:** For time-based analysis, we ensured date fields were proper datetime objects. In the Remotive data, publication_date (when the job was posted) was converted from string to Python datetime. In the HN data, the created_at timestamp of each comment was parsed to datetime and the year extracted. We had to be careful to handle any missing or malformed dates (dropping those if conversion failed).

- **Combining and saving datasets:** The HN comments from 2020–2024 were combined into one dataframe (df_hn_jobs before filtering, then df_hn_filtered after filtering/cleaning). This allowed us to easily group and analyze across years. We saved intermediate results (like the filtered HN dataset and top terms frequencies) to CSV files so that we could reuse them or inspect them outside the script if needed. The Remotive data and HN data remained separate datasets (since one is current snapshot and the other is historical), but we analyze them side-by-side in the Results section for comparison.

Overall, the preprocessing ensured that the data from both sources was clean, relevant, and in a structured format suitable for our analysis methods.

Methods of Analysis

With the data prepared, we employed several methods to analyze it and answer our research questions. Our analysis techniques included descriptive statistics, frequency analysis, simple natural language processing, and trend analysis, as outlined below:

Frequent Job Titles: To find the most common job titles, we counted occurrences of titles in each dataset. In the Remotive data, after filtering to df_data (data/ML jobs), we used pandas value_counts() on the job title field to identify the top 10 titles among current remote job postings. For the HN data, job titles were embedded in free-form text of comments. We devised a simple approach to infer a "normalized role" by searching for known role keywords in each comment (e.g., "Data Scientist", "Machine Learning Engineer"). We defined a mapping of keywords to standardized role names (e.g., any mention of "analytics engineer" would map to "Analytics Engineer"). Using this map, we categorized each HN job post comment into roles and then counted frequencies. This gave us the most frequent job roles mentioned in the community postings. The code snippet below (from the Remotive analysis) illustrates how we filtered and counted job titles in the current data:

```
# Filter remote jobs related to data/ML
df_data = df[df['title'].str.contains('data|machine learning|ml|ai|analytics', case=False)]
# Count top 10 job titles
top_titles = df_data['title'].value_counts().head(10)
print(top_titles)
```

- This method addresses the question of **what the most frequent job titles** are in our domain. It helps highlight, for example, whether "Data Scientist" or "Data Analyst" roles are more abundant, etc.
- **In-Demand Skills Extraction:**
- We took two approaches to identify key skills:
  - For the Remotive data (structured tags): Each job listing includes a list of tags (skills/technologies required for the role). We aggregated all tags from the filtered data jobs and counted their frequency using Python's Counter. This produced a ranked list of the most common tags (skills) in current remote data/ML job postings. These typically include programming languages, tools, or frameworks (e.g. Python, SQL, AWS).
  - For the HN data (unstructured text): We utilized NLP to extract frequent terms from the job descriptions. After cleaning the text, we tokenized each comment using spaCy (with a lightweight English tokenizer). We filtered out common stop words and very short words, then counted the remaining tokens to find which skills/terms appeared most often over 5 years of posts. This captures frequently mentioned technologies or keywords in job descriptions. The result is a list of top terms (which might include specific skills like "python" or general terms like "engineer", "remote"). By comparing the structured tag frequencies from Remotive with the unstructured term frequencies from HN, we can see overlap and differences in how skills are mentioned.

- **Salary Analysis:** To understand salary trends, we focused on the Remotive data (since HN posts rarely include standardized salary info). After parsing salary strings into numeric values (see Preprocessing), we computed summary statistics and plotted a distribution. We generated a histogram of the parsed_salary values for remote data jobs to see the range and common salary levels. We also attempted to correlate skills with higher salaries: by grouping the salary data by each skill tag, we calculated the average salary associated with postings that mentioned that skill. We identified the top 5 skills by average salary to see which expertise might command higher pay (for example, certain specialized skills or senior roles might show higher averages). This addresses the question of **salary variation** across different skill sets and roles in remote data jobs.

- **Temporal Trend Analysis:** Using the HN dataset (which spans 2020–2024), we analyzed how the volume of job postings changed over time. We grouped the filtered HN job posts by year and counted how many postings (comments) were in each year. This gave a year-over-year trend of remote/data job opportunities as reflected in the community. We visualized this as a line chart of job count by year. Additionally, within the Remotive data (which is a single snapshot but includes posting dates), we examined the **recent posting trend** by looking at the last 30 days of postings and counting jobs per day. This daily trend (plotted as a bar chart by date) provides insight into the current posting frequency and any spikes or dips in the past month.
- **Collating Results:** The results of these analyses were collated in the form of tables and charts. For instance, after counting top skills and titles, we created bar charts to visualize the top 10 titles and top 15 skills for easier interpretation. We compiled the top terms from HN into a dataframe and saved it as a reference (hn_top_terms.csv). Throughout, we printed intermediate summaries (like the top 15 terms from HN or the top 5 highest-paying skills from Remotive) to the console for documentation. By structuring the analysis around the key questions (titles, skills, salaries, trends), we ensured each question was addressed with the relevant fields from our data (titles, tags, salary field, dates, etc.), and the findings from each part could be compared or combined to draw higher-level insights.

In summary, our methods combined **data filtering**, **frequency analysis (counts)**, **NLP token extraction**, and **visualization** to answer the project questions. Next, we detail the program implementation and then present the outputs of these methods.

Program Overview

The project was implemented as a Python script that proceeds through the data retrieval, processing, analysis, and visualization steps in sequence. The program is organized into two main parts corresponding to the two data sources, followed by a summary section:

- **Part 1: Live Remote Jobs (Remotive API)** – We begin by importing required libraries (requests, pandas, matplotlib.pyplot, re, spacy, etc.). We load the **Remotive jobs** by making an HTTP GET request to the API as shown earlier. The JSON response is normalized into a pandas DataFrame. We then apply filters on the DataFrame to isolate data/ML jobs. Several analyses are performed on this subset:
    - Counting top job titles (using pandas value_counts).
    - Aggregating all skill tags and counting their frequencies (using Python Counter).
    - Parsing salary ranges and computing a histogram for salary distribution.
    - Calculating average salaries by skill and identifying top-paying skills.
    - Counting job postings in the last 30 days and plotting a daily trend chart. Each analysis in this section is accompanied by a visualization (bar charts for frequencies and histogram for salary). We used matplotlib to generate these plots. The code is well-documented with comments explaining each step.
- **Part 2: Historical Remote Jobs (HN "Who is Hiring" data)** – Next, the script fetches historical data. For each year 2020 through 2024, it uses the Hacker News search API (via requests calls) to find the monthly "Who is hiring?" posts and then retrieves all comments from those posts. This is done in loops with delays to avoid hitting API limits. The result is a DataFrame of thousands of job posting comments with fields for the posting text and timestamp. We filter this DataFrame using keywords to focus on posts likely about remote data/ML jobs. We then perform text preprocessing (using BeautifulSoup and regex as described earlier) on the comment text. After cleaning, we use **spaCy** to tokenize

the text and count term frequencies (excluding stopwords). We also apply our role extraction function to assign normalized job role categories to each post. With this processed data, we carry out analyses similar to Part 1:

  - Counting the most common roles mentioned (and plotting a bar chart of top roles).
  - Counting postings per year (and plotting the yearly trend line).
  - Listing the top 15 frequent terms in the job descriptions (output as a printed table). The code in this section also saves some results to CSV (e.g., the filtered dataset and top terms) for record-keeping. This part of the program showcases the use of an external API, text processing with NLP, and multi-year data aggregation.

- **Part 3: Summary of Insights** – Finally, the script includes a summary section (written in markdown comments within the code for clarity) that compares the findings from the two parts. This is not computational, but rather a narrative analysis that highlights differences and similarities between the current (Remotive) job market snapshot and the HN historical trends. Key insights about roles, skills, and temporal changes are noted here, effectively bridging the results of Part 1 and Part 2. This summary served as a basis for our Conclusions section in this report.

The program relies on standard data analysis libraries: pandas for data manipulation, requests for web/API calls, matplotlib for plotting, and spaCy for NLP. Additionally, BeautifulSoup from bs4 is used for HTML parsing in text cleaning, and Python's built-in collections.Counter helps with frequency counts. We structured the code to be modular (with helper functions like fetch_hn_who_is_hiring, fetch_comments_for_thread, clean_html_and_symbols, etc.) which makes it easier to follow the workflow and reuse parts if needed. The code is thoroughly commented to document each step and ensure clarity.

Overall, the program successfully implements the plan by gathering the data, performing the necessary preprocessing, and executing analysis steps to generate the outputs needed to answer our questions.

Output Documentation

The analysis produced several outputs in the form of charts and printed summaries. Each output is documented below, along with an explanation:

- **Top Job Titles (Current Remote Jobs):** *Output:* A horizontal bar chart titled **"Most Common Remote Data Job Titles"** showing the top 10 job titles from the Remotive data (2024). The x-axis represents the count of postings, and each bar corresponds to a job title. *Findings:* This chart reveals which data/ML roles are most frequently advertised in the current remote job market. For example, we found titles like **Data Scientist, Data Analyst,** and **Analytics Engineer** among the top occurrences. This indicates a strong demand for roles that specialize in data analysis and data science. The presence of multiple senior titles (e.g. *Senior Data Scientist*) also suggests companies are hiring at various experience levels for data roles.
- **Top Skills/Tags (Current Remote Jobs):** *Output:* A bar chart titled **"Top Required Tags (Skills) in Remote Data Jobs"** displaying the 15 most common skill tags from the Remotive job postings. Each bar corresponds to a specific skill or technology, and its height indicates how many job listings mention that skill. *Findings:* The chart showed that skills like **Python, SQL, JavaScript, AWS,** and **Machine Learning** are among the most frequently mentioned. This confirms that proficiency in programming (especially Python), data querying (SQL), and cloud or machine learning frameworks are in high demand. The tags provide a quick overview of the technical stack often expected for remote data/ML positions (for instance, **Python** being the top tag underscores its importance in data roles).
- **Salary Distribution (Current Remote Jobs):** *Output:* A histogram titled **"Salary Distribution for Remote Data Jobs"** illustrating the distribution of annual salaries (in USD) for data-related remote jobs from the Remotive dataset. The x-axis represents salary ranges (binned), and the y-axis shows the number of job postings offering salaries in those ranges. *Findings:* This plot showed a concentration of remote data job salaries in a certain range (for example, many salaries clustering around the mid-range of the spectrum, with fewer jobs at the

extreme low or high ends). Although not all jobs listed salary information, among those that did, we observed a central tendency (e.g., a common salary range might be around **$80k-$120k** for many data roles, with some higher-level or specialized roles offering more). This provides insight into how remote data/ML positions are compensated on average. There were some outliers at the high end (indicating a few roles with very high salaries, possibly senior or specialized jobs).

- **Top Paying Skills (Current Remote Jobs):** *Output:* A horizontal bar chart titled **"Top Paying Skills in Remote Data Jobs"** that ranks a selection of skills by the average salary of jobs requiring that skill. We calculated the average parsed_salary for each skill tag (considering skills mentioned in at least a few postings for reliability) and then took the top 5-10 skills with the highest average salaries. *Findings:* This analysis highlighted which skills are associated with higher paying roles. For instance, we found that certain skills (e.g. **Cloud Architecture**, **Leadership/Management**, or specific advanced technologies) tended to appear in job postings with above-average salaries. This suggests that roles requiring those skills (possibly more niche or senior skills) command higher compensation. It provides a nuanced view beyond just frequency: not only which skills are common, but which skills might boost earning potential.
- **Daily Posting Trend (Last 30 Days):** *Output:* A bar chart **"Job Postings Over the Last 30 Days"** showing the number of remote data job postings per day over the past month (based on Remotive data dates). The x-axis lists dates (covering roughly one month) and the y-axis is the count of jobs posted on that date. *Findings:* This gave a sense of the current activity in the job market. The chart typically shows minor fluctuations day-to-day with possible spikes on certain weekdays. We observed that postings were relatively steady with a few peak days (for example, possibly more jobs posted on Tuesdays and Wednesdays, and fewer on weekends). This short-term trend helps understand the volume of new opportunities appearing in the remote data/ML space on a daily basis.
- **Frequent Terms in Job Descriptions (Historical HN data):** *Output:* A printed table of the **top 15 terms** extracted from the HN "Who is Hiring" posts (2020–2024) relevant to remote/data jobs. The table lists the terms and their frequencies. *Findings:* The most frequent terms included words like **"engineer"**, **"remote"**, **"data"**, **"work"**, **"team"**, and **"python"**. Many of these reflect the

context of the posts: "engineer" suggests a lot of postings were for engineering roles (which aligns with HN's startup/engineering community), "remote" confirms that many posts mentioned remote work (as expected from our filter), and "data" indicates the focus on data-related positions. The presence of "python" in the top terms validates that Python has been a consistently mentioned skill in job posts over the years. Other terms like "developer", "product", or "cloud" might also appear, painting a picture of the common language used in job ads on HN. This output, derived via NLP, complements the structured tag analysis from Remotive by showing popular concepts in an open text setting.

- **Top Job Roles (Historical HN data):** *Output:* A horizontal bar chart titled **"HN Most Common Remote Data Job Titles"** illustrating the top normalized job roles from the HN postings. After mapping various phrases to broader role categories, we counted the occurrences of each role. *Findings:* This chart revealed roles like **Machine Learning Engineer, Data Scientist, Full Stack Developer, Data Analyst,** etc., as commonly mentioned in the HN threads. Interestingly, the HN data showed a mix of roles: not only pure data science roles but also general software roles (e.g., Full Stack Developer) that appeared frequently. This suggests that in the community postings, companies often look for a blend of skills (sometimes one person wearing multiple hats, which is common in startups). Compared to the Remotive top titles (which were more strictly data-focused titles), the HN top roles included more engineering and hybrid roles, highlighting a difference in how jobs are advertised in informal channels vs. formal job boards.
- **Yearly Trend of Remote/Data Job Posts (HN data 2020–2024):** *Output:* A line chart titled **"HN Remote Data/ML Jobs per Year (2020–2024)"** with years on the x-axis and number of job posting comments on the y-axis. Each point represents a year from 2020 through 2024, connected by a line to show the trend. *Findings:* The trend showed a **spike in 2021**, followed by a decline through 2022 and 2023, and a slight uptick or stabilization in 2024. The high point around 2021 aligns with the time when remote work opportunities dramatically grew (due to the COVID-19 pandemic pushing companies to hire remotely). The subsequent dip likely reflects market saturation and the tech industry slow-down (layoffs and hiring freezes in 2022–2023). By 2024, the curve suggests that hiring activity started to pick up again, potentially driven by new

needs (for example, an increase in AI and machine learning focus, as some HN posts in 2024 explicitly mention AI terms like GPT or LLM). This longitudinal view is valuable to understand how the demand for remote tech jobs changed over an unprecedented period for the job market.

Each figure and result above was generated directly by our Python code. They were examined to derive insights, as discussed in the next section. All outputs confirm that our code successfully gathered the intended information and that the results are meaningful in context. The combination of visualizations and text outputs allows us to cross-verify observations (e.g., seeing Python appear in both the tag chart and the HN terms list reinforces its importance).

Conclusions

Through this project, we have gained several **key insights** into the present-day job market trends for tech roles, especially focusing on data and machine learning positions in remote work settings. Below, we summarize our conclusions in relation to the initial research questions:

- **In-Demand Skills Across Roles:** Our analysis shows a strong convergence on certain skills as being in high demand. **Programming and data-related skills** are paramount – for instance, **Python** emerged as a top skill in both current job postings and historical posts. Alongside Python, skills like **SQL, machine learning frameworks, cloud technologies (AWS/Azure), and data visualization tools** were frequently mentioned. This indicates that employers consistently seek candidates with a blend of data handling, analysis, and software development skills. The overlap of terms between the structured Remotive tags and unstructured HN text (e.g., appearances of "python", "data", "ML") underscores that these skills have been and continue to be crucial. We also observed that newer **AI-specific skills** (like familiarity with LLMs or mention of "GPT") started appearing by 2024, suggesting that the rise of AI is influencing job requirements.

- **Job Title Trends:** There is a clear shift toward specialized roles in the data domain for remote jobs. In the Remotive data (current snapshot), we see a lot of **explicitly data-focused titles** such as *Data Scientist, Data Analyst, Machine Learning Engineer, Analytics Engineer,* etc. Meanwhile, the HN historical data — reflecting a more startup-oriented set of postings — often mentioned roles like *Machine Learning Engineer* and *Full Stack Developer*, indicating that earlier in the decade, companies (particularly startups) were looking for versatile engineers who could handle data along with broader development tasks. Over time, as the field matured, there's been a **shift from generalized engineering roles to more specialized data roles**. This matches our observation that postings on Remotive (likely from established companies) use more defined role titles. The prevalence of senior-level titles in current data (e.g. *Senior Data Scientist*) also suggests an increased stratification of roles as organizations build larger data teams.
- **Salary Variations:** We analyzed salary data for remote data jobs and found a wide range, but with notable patterns. Many remote data/ML jobs offer salaries in the mid-to-high five figures or low six figures (USD). The distribution suggests that while entry-level or junior data roles might start lower, experienced and specialized roles can command six-figure salaries. By linking skills to salary, we discovered that roles requiring certain **high-value skills**(for example, expertise in cloud architecture, leadership, or niche programming languages) tend to offer higher average pay. This implies that professionals who develop these in-demand and harder-to-find skills can leverage them for better compensation. However, since our data was primarily remote positions, we note that some extremely high-end tech salaries (often seen in specific locales like Silicon Valley) might not appear here; remote roles can sometimes average slightly lower than on-site roles at big tech hubs, but they also come with other benefits. A comprehensive location-based salary comparison wasn't possible given our data focus (most Remotive jobs did not specify a location beyond "remote"), but our findings give a general sense of remote tech salary levels and the influence of skills on pay.
- **Remote vs On-Site Trends:** Our project concentrated on remote job data, so a direct comparison with on-site roles is limited. However, the historical trend from

HN provides indirect insight. The significant spike in 2021 postings correlates with the shift to remote work during the pandemic, indicating that remote opportunities surged when on-site work was curtailed. The subsequent decline could suggest that after the initial rush to remote, the balance between remote and on-site hiring found an equilibrium or that overall hiring slowed. By 2024, with many companies resuming on-site operations but also embracing hybrid/remote for specialized talent, the data suggests that remote roles remain important, especially in the data/AI field, but are perhaps not as overwhelmingly dominant as in 2021. In other words, the market is adjusting: companies are more strategic about which roles can be remote. Our focus on remote jobs means all the trends we identified (skills, titles, salaries) apply to remote positions; on-site roles might have similar technical requirements, but remote roles may have broader geographic competition and sometimes slightly different skill emphases (e.g., communication across time zones). Fully addressing the remote vs on-site question would require additional data on on-site postings, but our findings confirm that remote data jobs are a significant and growing segment of the tech job market.

- **Impact of Macro Trends:** By examining 2020–2024, we can tie certain macro-level events to job market changes. The **COVID-19 pandemic** led to a boom in remote job postings (2020 into 2021), as reflected by the HN data surge. Following that, the tech industry faced **hiring freezes and layoffs (2022–2023)**, which likely contributed to the dip in our observed postings (fewer companies hiring, some caution in adding new roles). Towards 2024, the emergence of **generative AI and renewed tech innovation** corresponded with an uptick in specialized postings (we saw terms like "LLM" or "GPT" starting to appear), indicating new opportunities in AI-related jobs. These observations illustrate how external factors directly influence demand: companies rapidly adapted to remote work when needed, pulled back when the market tightened, and are now seeking talent in cutting-edge areas as new technology trends arise.

The project successfully met its objectives. We demonstrated the ability to gather data from web sources (API and online forum) and perform an end-to-end analysis of the tech job market. The findings provide a data-driven perspective on what skills and roles are currently sought after, how remote data job salaries are distributed, and how the job landscape has evolved over a turbulent few years. These insights can be valuable for professionals planning their career development (e.g., focusing on key skills like Python or cloud computing) or for companies understanding the talent market (e.g., recognizing competitive salary ranges and the continued interest in remote work).

Overall, the **present-day job market trends** in tech show a strong demand for data and machine learning expertise, an ongoing significance of remote work opportunities, and a landscape that responds to both technological advancements and broader economic events