

Remote Job Market Intelligence using Ethical Web Scraping

Presented by: Team G



Project Overview

This project focuses on collecting real remote job listings from Remote OK, building an end-to-end data science pipeline. It emphasizes **ethical web scraping** to ensure compliance with web standards, paving the way for actionable insights into skills demand and job roles.



Problem Statement & Challenges

Key Questions

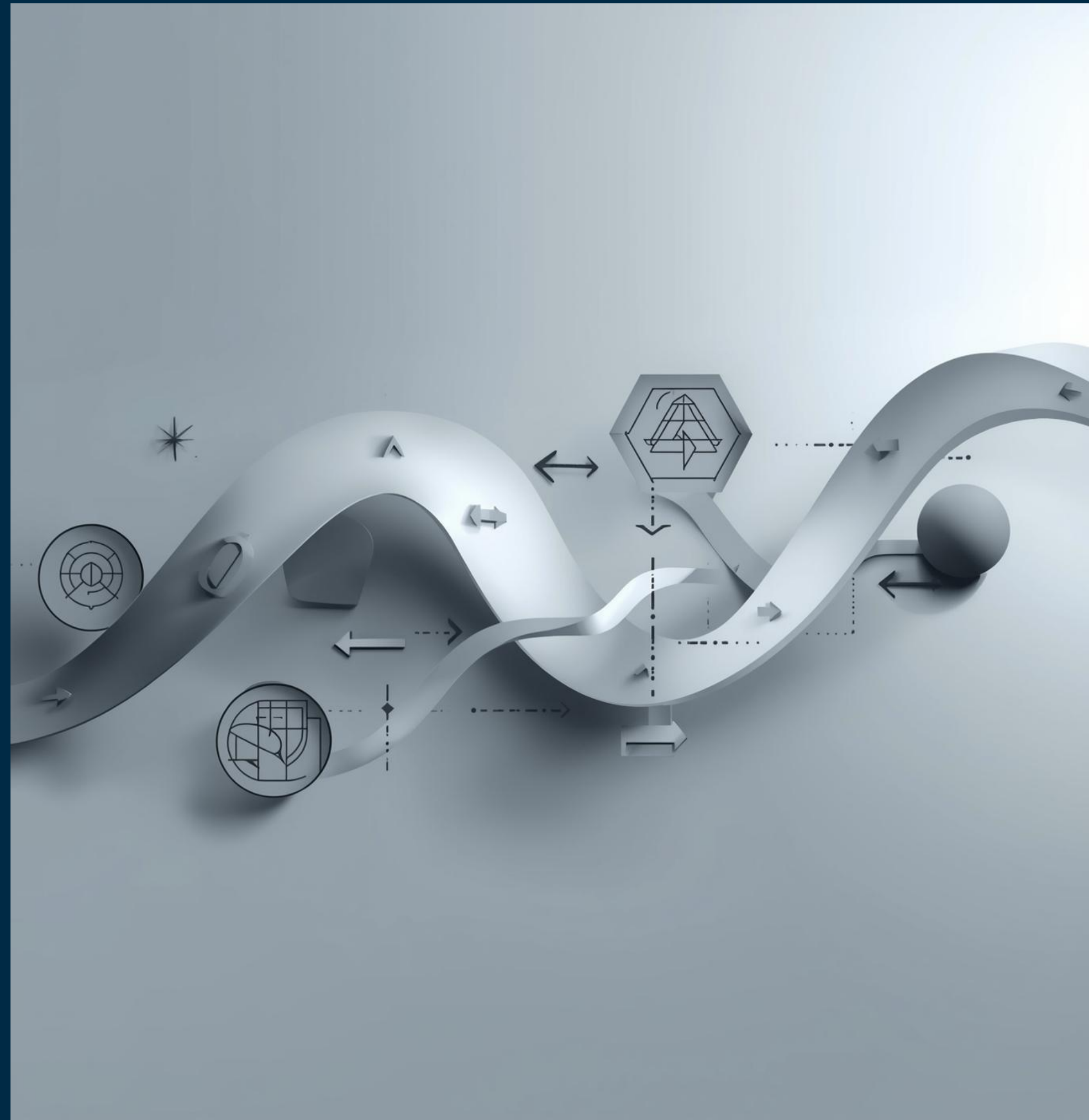
Addressing skills demand, job roles, job types, and geographic distribution in the remote job market.

Project Challenges

Managing unstructured raw data, inefficiencies of manual analysis, and adhering to ethical web scraping constraints.

Workflow Architecture

The project employs a comprehensive data science pipeline that includes stages such as website analysis, ethical web scraping, data cleaning, analysis, visualization, documentation, ethics review, and testing. Each stage is crucial and assigned to specific team members to ensure efficiency.



Website Analysis

This section outlines the **website structure mapping** conducted by **Sarang**. It details the exploration of the Remote OK site, focusing on critical HTML elements such as Job Title, Company, Skills, Location, and Job Type to ensure efficient scraping and data extraction.



Web Scraping Methodology

This section explains the ethical web scraping approach carried out by **Ansari and Rahul**. Using Selenium to handle dynamically rendered content, the team ensured compliance with web standards while efficiently gathering job listing data from Remote OK's platform.



Data Cleaning

In this phase, we focused on refining the raw data collected from web scraping. Key tasks included **removing duplicates**, addressing missing values, and ensuring text normalization to create a clean, structured dataset that facilitates accurate analysis and insights generation. Done by **Tejas**



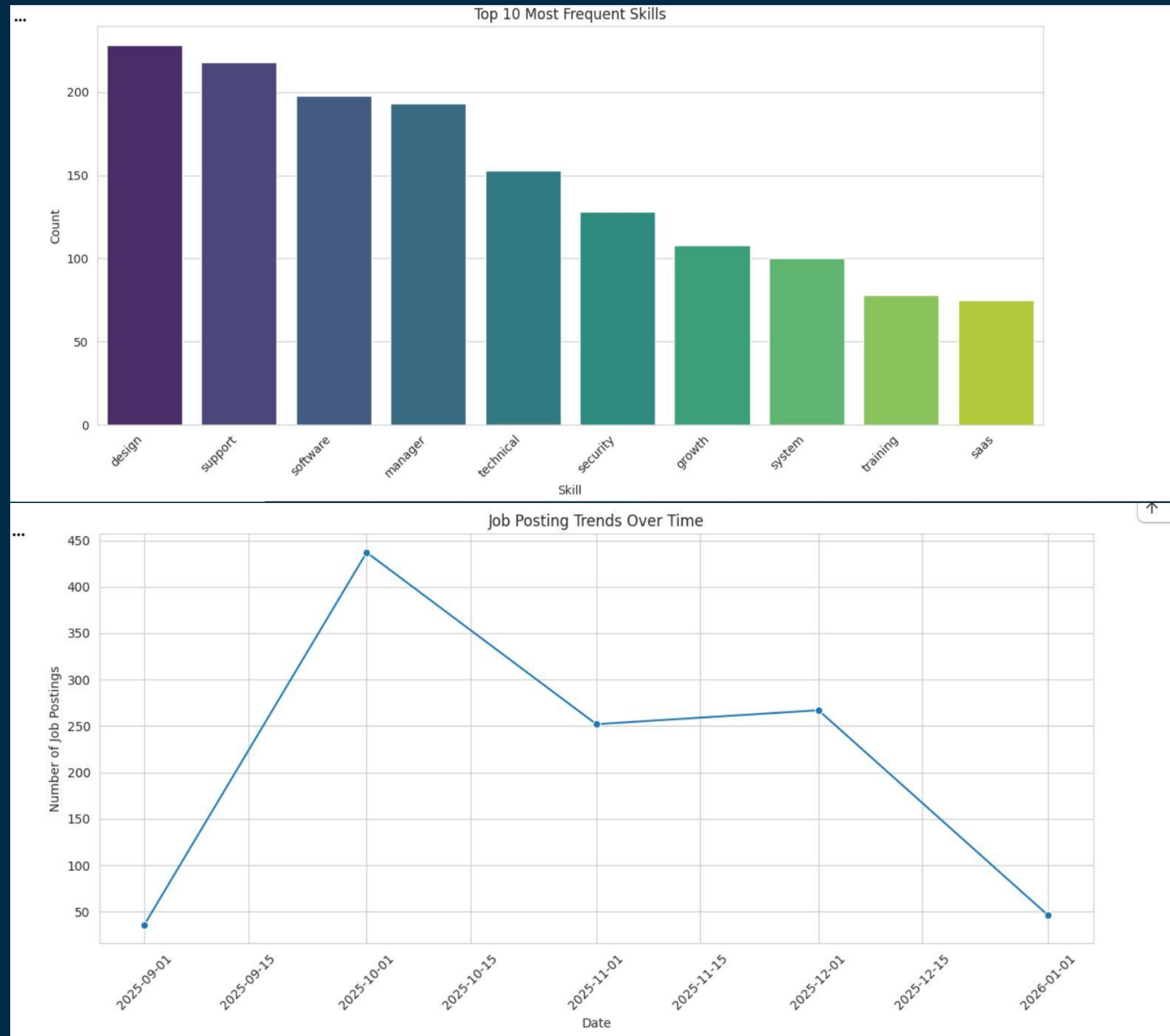
Data Analysis Insights

In this section, which is conducted by **Himanush**, we present the **key findings** from our data analysis, focusing on trends in job roles, skills demand, and geographical distribution. These actionable insights provide a comprehensive understanding of the remote job market landscape.



Data Visualization and Dashboard Creation

This section was handled by **Huda Saiyed** and focuses on transforming the analyzed data into clear and meaningful **visual representations**. It includes the creation of charts and visual summaries, such as skill demand, job type distribution, job trends over time, and top hiring locations. These visualizations effectively communicate key insights and support data-driven decision-making.



Documentation

This section was handled by **Priyanga** and focuses on compiling all source code contributions from the team. The collected scripts were carefully reviewed, refined, and integrated into a single, well-organized Jupyter Notebook (Team_G.ipynb). The **documentation** clearly presents the complete project workflow in a structured manner, making it easy to understand, review, and present.

```
226 fflush(stdin);
227 scanf("%[^\n]", R.note);
228 if(fwrite(&R, sizeof(R), 1, fp)){
229     gotoxy(5, 12);
230     puts("Note is saved successfully");
231     fclose(fp);
232 } else{
233     gotoxy(5, 12);
234     SetColor(12);
235     puts("\aFail to save!!\a");
236     ClearColor();
237     gotoxy(5, 15);
238     printf("Press any key.....");
239     getch();
240     fclose(fp);
}

printf("Error in opening the fi");
while(fread sizeof 1
if
gotoxy(10, 5+i);
printf("Note %d Day = %d: %
i++;
1++;
}
}
if(i == 0){
gotoxy(10, 5);
printf("This month contains no
}
gotoxy(10, 7+i);
printf("Press any key to back.....");
getch();
}

int main(){
ClearConsoleToColors(15, 1);
title("Project - Program")
```


Testing & Preparation

This section outlines the final **validation process** led by **Prajakta**, ensuring the pipeline's integrity. It also highlights the collaborative efforts of **Shifa and Vimalesh** in organizing the **presentation** flow, emphasizing key insights and ethical considerations for effective communication.



THANK YOU

