Brief Summary Report

Name: Sidra Zubair Shaikh | Roll Number: 25280035 | Course: AI-620 | Assignment #1

Theme: AI Labor Markets

Project Overview

This project implements a modern ELT pipeline to analyze trends in the AI labor market. The objective was to integrate heterogeneous data sources, store them in structured formats, perform data quality checks, and generate exploratory insights that could support downstream AI and machine learning applications.

The pipeline integrates three real-world data sources:

1. Kaggle AI Job Dataset (Structured CSV)
   This dataset contains over 15,000 AI job postings with salary, experience level, skills, remote ratio, and geographic information.

2. Remotive Jobs API (Semi-structured JSON)
   Live job postings were retrieved through a public API endpoint and normalized into tabular form for analysis.

3. Google Trends (Time-series Data)
   Search interest over time for AI-related job keywords such as "AI jobs," "machine learning jobs," and "remote AI jobs."

These datasets were extracted and stored in a layered structure using data/raw, data/processed, and data/cleaned directories. This separation ensures reproducibility and preserves raw data integrity while enabling structured transformations for analysis.

Key Findings

1. Experience Level Distribution

Analysis of the Kaggle dataset shows that AI hiring is concentrated at mid-level and senior-level positions. Entry-level roles are comparatively fewer, suggesting that organizations prioritize experienced candidates in AI-related positions.

2. Salary and Experience Relationship

A positive relationship exists between years of experience and salary in USD. The scatter plot indicates that compensation generally increases with experience. However, variability in salary levels suggests that other factors such as company size, industry, and location also influence compensation.

3. Remote Work Trends

The dataset includes a remote_ratio variable, which shows strong representation of partially and fully remote roles. This supports the idea that AI roles are well suited for distributed and remote work environments.

4. Search Demand Over Time

Google Trends data demonstrates sustained search interest in AI job-related keywords over the past five years. Although fluctuations occur, overall demand remains stable, indicating continued relevance of AI skills in the labor market.

Challenges Encountered

Several challenges arose during implementation:

- API data variability. The Remotive API returned semi-structured JSON with inconsistent fields across job postings, requiring normalization and careful handling of missing values.
- Missing salary data. Some job postings lacked salary information. Instead of imputing values that could distort analysis, rows with missing salaries were excluded only from salary-specific visualizations.
- Rate limiting from Google Trends. Occasional request limits required reducing keyword counts or retrying queries.
- Data type standardization. Date and numeric fields required explicit parsing to ensure accurate statistical summaries and time-series analysis.

Relevance to AI and Machine Learning Systems

The engineered dataset can support several AI applications, including salary prediction models, skill-demand forecasting, labor market trend analysis, and time-series demand forecasting using search interest signals.

By integrating structured job data with external search demand indicators, the pipeline demonstrates how modern data engineering practices support AI system development through reproducible, multi-source data workflows.