SECTION 1

Part 1 Questions: Extract and Load

(a) Data Heterogeneity

The three data sources used in this project represent different types of data structures.

The Kaggle AI job dataset is structured tabular data in CSV format. Each row represents a job posting, and the columns such as `salary_usd`, `experience_level`, `remote_ratio`, and `posting_date` follow a consistent schema. This makes it easy to load directly into a dataframe and perform statistical analysis.

The Remotive API provides semi structured JSON data. The response contains nested fields and optional attributes, meaning that not every job posting contains the same set of fields. The raw API response was saved as `remotive_raw.json` and then normalized into a tabular format for easier analysis.

Google Trends provides structured time series data. Each row corresponds to a date, and each column represents indexed search interest for specific AI job related keywords over time. Unlike the job datasets, this source captures demand trends rather than individual job postings.

(b) Extraction Challenges

A few practical challenges were encountered during extraction.

The Remotive API returned records with varying fields across job postings. Some entries were missing salary or location information, which required careful normalization and missing value handling.

Google Trends occasionally limits repeated requests. To manage this, the number of keywords was kept reasonable and queries were retried when necessary.

The Kaggle CSV required explicit type conversion. Numeric columns such as `salary_usd` and date columns such as `posting_date` had to be parsed properly to ensure consistency and accurate analysis later in the pipeline.

(c) Storage Justification

Storing data in multiple formats improves flexibility and reliability.

CSV files are ideal for structured tabular data and are easy to inspect and analyze using tools like Pandas or Excel.

JSON files preserve semi structured API responses and retain nested information that might be useful for future parsing or downstream machine learning workflows.

Separating raw and processed layers improves reproducibility. The raw layer keeps the original source data unchanged, while the processed layer provides analysis ready files. This separation makes debugging and validation much easier.

SECTION 2

Brief Documentation of Pipeline Structure

Pipeline Structure Overview

This project implements a modern ELT pipeline using a layered storage approach to integrate heterogeneous AI labor market data.

Data Sources Integrated

Public Dataset: Kaggle AI Job Postings CSV
This structured dataset contains job level attributes such as `job_title`, `salary_usd`, `experience_level`, `remote_ratio`, `required_skills`, and `posting_date`.

API Source: Remotive Jobs API
This source provides semi structured JSON data retrieved from a public endpoint. The raw response preserves nested and optional fields and is later normalized into a dataframe for processing.

Time Series Source: Google Trends
This dataset captures search interest over time for AI labor related keywords such as AI jobs, machine learning jobs, and remote AI jobs. It provides an external signal of labor market demand.

Directory Layout and Storage Strategy

data/raw
This folder stores the original extracts exactly as received. It includes the Kaggle CSV snapshot,

the raw Remotive JSON response, and the raw Trends CSV export. This layer preserves the source of truth and supports reproducibility.

data/processed
This folder stores standardized outputs for each source in two formats. CSV files support easy inspection and analysis, while JSON files preserve semi structured information. Each source produces corresponding processed CSV and JSON files.

data/cleaned
This folder contains transformed and quality checked datasets generated in Part 2. Cleaning includes duplicate removal, missing value handling, type standardization, and basic validation to prepare the data for exploratory analysis and potential modeling.

Orchestration and Modularity

The pipeline is implemented in a notebook based workflow with modular sections responsible for extracting each source independently and saving outputs consistently. This structure makes it easy to extend the pipeline with additional sources or transformations in the future.

Relevance to AI and Machine Learning

The final engineered dataset can support applications such as salary prediction, skill demand modeling, labor market forecasting, and remote work trend analysis. By combining structured job posting data with external time series demand signals, the pipeline demonstrates how data engineering supports real world AI system development.

SECTION 3

Part 2 Questions: Transform, Clean, and Analyze

(a) Cleaning Rationale

Duplicates were removed to avoid double counting job postings. The Kaggle dataset used `job_id` as the unique identifier, while the Remotive dataset used `id` where available.

Date columns were converted to datetime format to ensure consistent temporal analysis.

Numeric columns were coerced into numeric types to allow accurate summary statistics and plotting.

Missing salary values were not imputed because salary distributions are often skewed. Instead, rows with missing salaries were excluded only from salary based visualizations to prevent misleading conclusions.

(b) Visualization Insights

The Google Trends visualization shows how interest in AI job related keywords changes over time. This reflects shifting demand signals in the labor market.

The experience level distribution shows that mid level and senior level roles dominate AI hiring, suggesting a strong demand for experienced professionals.

The salary versus experience plot demonstrates a positive relationship. In general, higher years of experience correspond to higher compensation, although there is noticeable variation.

(c) Visualization Critique

Google Trends data is indexed between 0 and 100, meaning it represents relative interest rather than absolute search volume.

The Kaggle dataset may not fully represent the global AI labor market and may contain platform or collection bias.

The salary versus experience scatter plot does not control for other factors such as company size, industry, or location. A more advanced analysis using regression or grouped comparisons could provide deeper insight.