Yuvraj Singh

yuvraj.singh212@gmail.com 7354460006 LinkedIn GitHub

Education

- **B.tech in Computer Science** *(e-commerce technology)*, VIT, Bhopal (May, 2026) **8.09 CGPA**

Technical Skills

- **Languages & Databases:** Python, PostgreSQL, Bash, Azure PostgreSQL
- **Data Libraries & Tools:** Pandas, PySpark, NumPy, Streamlit, RegEx, Apache Airflow, Docker, Git
- **Technical Proficiencies:** Web Crawling, Web Scraping, ETL Pipelines, API Integration, Automation Scripts

Projects

Shopinion

[Python, Apache Airflow, Docker, Selenium-Stealth, Pandas, Azure PostgreSQL, ETL]

- **Situation:** A project to train a **context-aware BERT sentiment model** required a **large-scale dataset** of over 100,000+ product reviews, but manual collection was unfeasible due to Flipkart's robust **anti-scraping measures**.
- **Task:** To engineer a fully automated, **end-to-end ETL pipeline** that autonomously handled the entire data lifecycle—from stealthy web scraping against a protected site to the structured storage of only unique entries and provided a simple interface for others to use.
- **Action:** Orchestrated the entire workflow using **Apache Airflow** and **Docker**, designing DAGs for both sequential and parallel execution. Engineered a resilient scraper with **Selenium-Stealth** to bypass anti-bot measures, **automatically extracting over 100,000+ clean, unique reviews in under 8 hours**. Implemented an intelligent data loading module in **Pandas** that performed a **pre-emptive check** against **Azure PostgreSQL** to ensure idempotent writes and prevent data duplication.
- **Result:** Deployed a robust, **self-service data collection platform** that fully automated the complex scraping and cleaning process. The parallel processing mode, running in controlled batches of three, demonstrated a **20-22% improvement in execution time** over sequential scraping. The system enabled others to self-serve the creation of large-scale, clean datasets for their own analysis projects.

Outbreak Tracker

[Python, Pandas, PyFaker, Streamlit, Scikit-learn]

- **Situation:** Inspired by a local jaundice outbreak, recognized that standard **diagnostics failed** without awareness of real-time, **regional case counts**.
- **Task:** To overcome the complete **unavailability of a suitable public dataset** by **engineering a custom one** from scratch for a proof-of-concept.
- **Action:** Built the **100k+ record dataset** by **merging real state-wise statistics** with **synthetically generated patient profiles** from **PyFaker**, then developed a **Streamlit app** that applied **location-based rules** to predictions.
- **Result:** The deployed app cut **critical misclassifications** by **22%** and boosted overall **diagnostic accuracy** by **5-10%**.

Advanced Data Cleaning and Feature Engineering [GitHub]

[Python, Pandas, RegEx, Data Cleaning, ETL, Feature Engineering]

- **Transformed** a raw **9,999-entry dataset** into a validated set of **6,816 unique records** by scripting the removal of 3,183 duplicates and parsing inconsistent text into structured **start/end date columns** using Pandas and RegEx.
- **Engineered** a new, high-fidelity **Type** column by developing a **logical pipeline** that first classified content based on the presence of a **Director** tag, then refined categories using runtime data (runtime $\leq$ 40 min $\rightarrow$ **Short-movie**) and genre-based **keyword matching** (**Animation**, **Documentary**).
- **Eliminated** data gaps by imputing over 1,500 missing **Runtime** values (~22% of data) and over 1,000 missing **Rating** values (~13% of data) using **context-aware group means**—a **mathematically more robust** approach than using a single global average.

Extracurricular Activities

- **SGFI National-level Basketball Player**; represented at multiple regional and inter-school tournaments
- Member of **university sports council**, contributing to **planning and execution** of intra-college leagues

Certifications

- **Data Engineer Associate** (Data Camp, Feb 2025)
  Learned ETL workflows, SQL for data modeling, and database design using PostgreSQL and Snowflake.
- **Financial Modeling And Valuation** (Internshala, Sep 2021)
  Learned corporate finance fundamentals including DCF modeling, ratio analysis, valuation techniques, and financial forecasting using Excel.