Yujia (Sian) Jin

Data Scientist / Analyst with a focus on SQL, Python, and Statistical Models

Dallas, USA | Born 08/02/1999 | +1 2026296597
yujiajin2024@gmail.com | **LinkedIn** | **Github** | **Website**
I can provide my own visa/work permit for Germany within 5-8 weeks. My notice period is 14 days.

TECHNICAL SKILLS

Senior Data Scientist (5 years of project experience) **Programming Languages**: Python | R | SQL | Snowflake SQL **Data Processing & Engineering**: ETL Pipelines | Data Integration | Apache Spark | dbt | Apache Airflow **Machine Learning & AI**: AI/ML | LLMs | Regression Modeling | Logistic Regression | Random Forest | Gradient Boosting | Statistical Analysis | XGBoost | LightGBM | Scikit-learn | TensorFlow | PyTorch **Analytics & BI Tools**: Power BI | Tableau | KPI Development | Looker **Databases & Data Warehousing**: PostgreSQL | Snowflake SQL | Amazon Redshift | Google BigQuery **Cloud Platforms**: AWS | Azure | Google Cloud Platform (GCP) **DevOps & MLOps**: Docker | Git | MLflow | Kubeflow | CI/CD Pipelines

WORK EXPERIENCE

03/2025 – today

Senior Data Scientist
UT Southwestern Medical Center, Dallas, USA
Project: CPR Ventilation Analytics & Research Data Platform – US-based clinical research system integrating EMS, hospital EHR, and device-level data for CPR outcome studies.

- Led end-to-end analytics and data infrastructure, integrating 10+ heterogeneous data sources into a unified research database, improving data completeness by 35% and reducing processing time by 40%.
- Designed scalable ETL pipelines and optimized data capture workflows, cutting data latency by 45% and reducing manual data errors by 30%.
- Built Power BI dashboards to monitor pipeline health, data quality KPIs, and ventilation metrics, increasing operational visibility and accelerating study monitoring cycles by 50%.

Technologies used: Python | R | SQL | Machine Learnign | Power BI | Data Integration | Data Cleaning | Statistical Analysis

Project: LLM-Driven Clinical Information Extraction Platform

- Designed multi-step extraction workflows including entity identification, clinical inference, normalization, and relationship mapping using foundation models (GPT-4o, Llama).
- Established a novel human-in-the-loop error ontology framework to systematically classify model failure modes and iteratively refine prompts, reducing clinically significant error rates to <1% and achieving macro-F1 up to 0.99.

Technologies used: Technologies: Python | LLMs(GPT-4o, Llama), Prompt Engineering CI/CD | Clinical NLP | Human-in-the-Loop

10/2024 – 03/2025

Data Scientist / Analyst
APEXUS Tech, Remote, USA
Project: Automated Financial Analytics & Risk Monitoring Platform –

US-based system automating market data ingestion, sentiment analysis, visualization, and risk controls.

- Built Python automation pipelines and APIs, reducing manual analysis time by 60% and improving data accuracy by 25%.
- Implemented news sentiment analysis processing ~180 articles per run, improving signal precision by 30%.
- Developed Python and SQL workflows to monitor exposure and margin across 2,200+ futures transactions, ensuring 100% margin compliance.

Technologies used: Python | Pandas | SQL | PostgreSQL | NLTK (VADER) | APIs | Plotly | Grafana | Risk Modeling | Excel | AWS

07/2023 – 10/2024 **Information Analyst**
Briar Cliff University, Sioux City, USA
Project: Institutional Performance & Financial Analytics Platform – US-based reporting solution supporting admissions, enrollment, and financial planning.

- Delivered KPI-driven Tableau dashboards for 15+ departments, reducing reporting cycles by 60%.
- Automated ELT pipelines integrating academic and financial data, cutting manual processing by 70%.
- Applied regression analysis to multi-year financial data to support forecasting and cost optimization.

Technologies used: SQL | Tableau | Python | Excel | SSRS | ETL

02/2022 – 08/2022 **Data Scientist (Internship)**
Civilience, Remote, USA
Project: COVID-19 Social Media NLP Analytics Pipeline

- Built cloud-hosted pipelines for large-scale data ingestion and NLP topic modeling, increasing ingestion throughput by 50%.

Technologies used: Python | Twitter API | NLP | UMAP | SQL | GitHub

09/2020 – 06/2021 **Data Scientist**
Queen's University Belfast, Belfast, UK
Project: Bioinformatics & Cloud Healthcare Analytics Research

- Applied PCA and K-means clustering, achieving ~75% classification accuracy on high-dimensional biological data.
- Built Azure-based ETL pipelines processing ~10GB of healthcare JSON data with SCD Type I & II handling.

Technologies used: SQL | Python | Excel | Azure | JSON | SSRS

EDUCATION

08/2021 – 05/2023 Master of Science/ Data Science & Analytics
Georgetown University, Washington DC, USA

09/2017 – 07/2021 Master's in Pharmacy
Queen's University Belfast, Belfast, UK

LANGUAGE SKILLS

English (C2) | Mandarin (Mother tongue) | German (A1, currently learning)