Yuxing Liu

San Diego, CA | 503-380-5765 | yuxingliu0826@gmail.com
.linkedin.com/in/yuxing-liu26 | https://yuxing-liu-portfolio.streamlit.app/

SUMMARY

Data Analyst with 1+ year of research and industry experience, skilled in data analysis, statistical modeling, and building machine learning models using Python, SQL, and scikit-learn. Experienced in Agile teams applying data to real-world scenarios to improve strategic decisions. Proven ability to clean messy data and communicate findings through visualizations and dashboards.

EDUCATION

University of California, San Diego | *GPA 3.5*

Bachelor of Science Mathematics-Computer Science| Minor in Data Science — Expected Mar 2027

SKILLS

Programming Languages: Python, Java, JavaScript, HTML/CSS, C, Bash, R, STATA, Arduino

Data Analysis & ML: Microsoft Office, Pandas, Numpy, AWS (RDS, S3, Quicksight), Scikit-Learn, Matplotlib, PyArrow, Tableau

WORK EXPERIENCE

Center for Applied Internet Data Analysis

Data System Analyst Assistant | *MySQL, AWS RDS, Docker, Selenium, Linux* — Jun 2025 - Present

- Automated deduplication pipeline using decision trees and MySQL foreign-key cascades, improving data integrity and reducing manual cleaning by 75%.
- Engineered a resilient Selenium scraper with IP rotation to extract metadata from dynamic web pages; auto-filled 3000+ legacy data into an AWS RDS-hosted database with 86% parsing accuracy.
- Built a Python CLI tool to streamline CAIDA's publication workflow with MySQL integration via SSH tunneling, automated PDF analysis, and dynamic configuration management; cut curation time by 40% and decreased false positives by 80%.

RunBuggy

Data Analyst Intern | *Unsupervised machine learning, API scraping, MangoDB, React, Node.js* — Nov 2024 - Jun 2025

- Designed a composite site-rating metric by scrapping Yelp, Google, and BBB API data; cleaned and matched 1,300+ repossession site records to internal data, optimizing vehicle routing and improving accuracy.
- Built an unsupervised model (K-means & PCA) to evaluate 500+ repo sites using customer satisfaction and volume metrics, boosting vendor allocation efficiency and reducing low-performance site usage.
- Developed a full-stack web app (React + MongoDB) with geo-location filters to rank repo sites, reducing user search time.

China Data Lab

Research Assistant | *Text Classification, Statistical Method, Video Analysis, Excel, Mandarin* — Apr 2024 - Jun 2025

- Applied Bag-of-Words and TF-IDF to analyze 1,000+ social media posts on public policy topics, revealing how propaganda and censorship influence public discourse; compiled findings into a formal report for a postdoctoral researcher.

Project Experience

Intel Energy Usage Optimization Research | *Apache Arrow, DuckDB, PyArrow, PySpark* — Jun 2025 - Present

- Optimized DuckDB ingestion pipeline with parallel processing, reducing SQL query time on 300GB Parquet data to under 4 minutes. Analyzed 116+ tables with PySpark to detect anomalies and define custom energy inefficiency metrics.

Mergers and Acquisitions Prediction | *Matplotlib, Scikit-learn, Time series analysis, Tableau* — Nov 2024 - Jun 2025

- Built an M&A prediction pipeline using Crunchbase data, engineered financial indicators, and trained a GradientBoostingClassifier with automated preprocessing (imputation, encoding) to predict acquisition probability, achieving 0.843 ROC AUC and 91.7% test accuracy of acquisition likelihood; Used a synergy-adjusted DCF model to estimate company valuation post-acquisition and assess financial feasibility.