

# Jason Xu

[g7xu2024@gmail.com](mailto:g7xu2024@gmail.com) | (909) 760-9522 | La Jolla, San Diego, CA

---

## EDUCATION

### University of California San Diego, Bachelor of Science, Data Science

Jun 2022 – Jun 2026 (expected)

- GPA 4.0/4.0
- Provost Honors
- Course Work: Data Structure, Data Management, Data Visualization, Cloud Computing, Probability Theory, Modeling, Linear Algebra, Recommendation System, Optimization

### Walnut High School, Walnut, CA, High School Diploma

Jun 2018 – Jun 2022

- GPA 3.93/4.0
- English Excellence Award (sole recipient)

---

## SKILLS

- **Coding:** Python (Proficient), SQL (Proficient), Java (competent), R (competent)
- **Statistics Inference:** Hypothesis Testing, A/B Testing, Probability, Linear Regression, Logistic Regression, Time Series, Fairness Analysis
- **Machine Learning:** Data Cleaning, Feature Engineering, Fine-tuning, Scikit-learn, Pandas, NumPy, Random Forest
- **Web Scraping:** HTML, Regular Expression, BeautifulSoup
- **Data Visualization:** Tableau, Matplotlib
- **Big Data Skills:** Data Integration, ETL, ER diagrams, Dask

---

## EXPERIENCE

### Data Science Intern, LLM STRATEGIES

Dec 2024 – Present

- Developing product analyzing and tracking historical legal cases, helping eCommerce sellers address legal challenges
- Wrote Python scripts leveraging Tesseract OCR and Gemini Flash 2.0 LLM API to extract trademarks from **low-quality, multi-format** historical court documents
- Designed an image preprocessing schema that successfully identified all previously unrecognized trademarks
- Testing algorithm on extensive image datasets retrieved from MySQL database

### Student Research Intern, The Wu Lab, Scripps Research

Jun 2024 – Aug 2024, Nov 2024 – Present

- Contributed to the development of a biomedical API by integrating new data sources into the system
- Collaborating with domain experts and programmers to design data cleaning and mapping schemas for data integration
- Developed a multi-class logistic regression model to impute missing data for later analysis, achieving an **F1-score of 0.74**
- Presented findings via Microsoft PowerPoint to **50+** interdisciplinary biomedical researchers

### Data Analytical Intern, Tong Consulting Inc.

Aug 2024 – Nov 2024

- Facilitated meetings with airline customer clarifying requests, timeline, and objectives, focused on recommending profitable flight routes align with customer interests
- Designed data cleaning pipeline using Pandas to compute basic statistics, remove outliers, handle missing values
- Utilized visualizations, hypothesis testing, and linear regression in Jupyter Notebook to find relationships between attributes
- Planned and implemented rigorous merging schemas finding round-trip routes from one-way flights data
- Performed break-even analysis schema for the recommendation routes considering the evaluation metrics from customer
- Tested the algorithm with large datasets, reported abnormal issues back to the customer
- Delivered documentation with reviews on analytical logistics and results for customer

### Student Researcher, The DSTL Lab, UC San Diego

Aug 2024 – Nov 2024

- Contributed to the development of Pandas Tutor, a tool visualizing Pandas operations in data manipulation
- Designed and implemented python programs enhancing Pandas Tutor to handle and visualize large data set
- Published a research paper as **first co-author** at ACM SIGCSE, analyzing effective data analytical behaviors

### Instructional Assistant, UCSD Halıcıoğlu Data Science Institute

Sep 2023 – Jun 2024

- Taught **500+** students data science principle and application mainly using Pandas and NumPy
  - Maintained course grading pipeline in python that calculated and displayed students' grading summaries
-