Cheng Wu

(217) 607-4522 | cheng.w@columbia.edu | Linkedin | Github

EDUCATION

Columbia University, Engineering New York, NY
Master of Science in Data Science Dec 2026
Focus: probabilistic modeling, statistical inference, time-series & panel methods, machine learning, large-scale data processing (Python, SQL, AWS)

University of Illinois at Urbana-Champaign Champaign, IL
Bachelor of Science in Econometrics & Quantitative Economics, Minor in Statistics *GPA: 3.86/4.00* May 2025
Focus: econometric modeling, causal inference (DiD/IV), applied statistics, predictive modeling (Python, R)

WORK EXPERIENCE

Tarte Cosmetics New York, NY
Data Intern, Compliance Intern Oct 2025 - Present
- Built **statistical validation models** to detect discrepancies across multi-source import/export datasets; increased signal accuracy **8%** across ~300 weekly entries
- Developed **real-time monitoring pipelines** integrating SKU, shipment, and audit logs; automated anomaly-flagging reduced review latency **35%**
- Designed predictive **inconsistency-detection algorithms** identifying SKU/weight mismatches and missing attributes; captured **100% of errors** pre-submission
- Conducted **exploratory analyses** on failure patterns and formulated **hypotheses** on upstream data-generation issues, informing structural workflow adjustments

Donglai Natural BioTech Co. Ltd Remote
Data Scientist Intern, Strategies Team Jun 2024 - Sep 2024
- Applied **unsupervised learning (K-Means, DBSCAN)** to analyze **12k+ user sessions**; integrated cluster outputs into SQL pipelines, improving behavioral forecasting **10%**
- Estimated **demand elasticities** using price-response models and validated results via natural-variation tests; pricing updates increased order-completion **12%**
- Optimized **Spark/Dask pipelines** for large-scale log processing, reducing computational latency **40%** and enabling near-real-time analytics
- Performed exploratory research on drop-off drivers and converted findings into **testable hypotheses** used by PMs for iteration planning

Innovative Bloom Foundation Shanghai, China
Lead Data Analyst May 2019 - Jun 2024
- Developed **longitudinal data pipelines** (SQL + survey data) supporting **panel-style tracking** for ~1,200 participants across income, education, and engagement metrics
- Used **segmentation & interaction analysis** to identify underserved subgroups; insights informed program redesign and improved cohort participation
- Evaluated an e-commerce pilot using **ROI/payback modeling**; results guided product changes increased household income 80%
- Conducted **risk-factor analyses** and developed **performance indices** to support resource-allocation decisions

ZheShang Securities Co. Ltd Hangzhou, China
Data Scientist, Institutional Finance Department May 2023 - Sep 2023
- Assessed **20+ PE/M&A** opportunities using **multi-factor regression**, scenario testing, and sensitivity analysis; improved screening precision **25%**
- Consolidated accounting, portfolio systems into a **SQL/VBA pipeline**, reducing reconciliation time **30%**, improving data transparency
- Built **risk-return and exposure-tracking models** benchmarked to sector indices; enabled real-time attribution for investment committees

RESEARCH EXPERIENCE

Research Assistant, Professor Elizabeth T. Powers, UIUC Jan 2023 - Feb 2025
- Conducted causal-inference research using DiD and IV frameworks to estimate treatment effects and heterogeneous responses across income cohorts
- Processed & harmonized 50k+ multi-year survey/administrative records in R, reducing preparation time 30% and ensuring cross-wave consistency for panel analysis

Researcher, Quantitative Policy & Urban Risk Analytics Initiative April 2023 - Feb 2024
- Modeled 2M+ crime and rental observations using panel and lag-structure regressions to estimate rent elasticity to violent incidents (3–5% decline)
- Constructed spatial risk-index models integrating housing, crime, and demographic indicators to identify resilience patterns and redevelopment priorities

SKILLS

Quantitative Modeling: Statistical inference, hypothesis testing, panel & time-series regression, causal inference (DiD/IV), ML (XGBoost, LSTM), factor modeling, signal extraction, backtesting
Programming: Python (NumPy, pandas, scikit-learn, statsmodels), SQL (window functions, CTEs), R, C++
Data Systems: ETL pipelines, Spark, Dask, AWS (S3, Redshift), distributed data processing
Tools & Workflow: Git, Jupyter, reproducible research, large-scale data cleaning & validation