THANH VU

vulethanh07112002@gmail.com | (717) 869-9502 | linkedin.com/in/thanh25 | github.com/thanhvu911 | thanhvu911.github.io

EDUCATION

Franklin & Marshall College, Lancaster, PA *May 2026*

B.A. in Computer Science, Magna Cum Laude | GPA 3.72/4.0 | Honors List (5 semesters); Dean's List (Spring 2022)

- Computer Science: Algorithms, Artificial Intelligence, Machine Learning, Theory of Computer Science, Computer Organization
- Math, Stats & Logic: Symbolic Logic, Intro to Data Science, Statistical Modeling, Probability & Statistics, Applied Regression

RESEARCH INTERESTS

Machine learning, probabilistic reasoning, and applied statistics for noisy real-world data, plus reliable data systems for reproducible AI.

RESEARCH & APPLIED PROJECTS

Can Mail Data Predict Housing Vacancy? | Fabric, Power BI, DAX *Sep 2026; write-up AI-assisted*

- Built a Fabric pipeline and Power BI model testing if quarterly USPS vacancy proxies annual ACS (6 counties, 84 county-years).
- Found pooled r = 0.77, but county-level r from −0.01 to 0.72; same direction in 45 of 78 year-over-year changes (57.7%).
- Flagged a 2012 USPS jump consistent with a classification change; fixed two silent ingestion bugs.

Machine Learning Coursework (CPS 360) | Python, NumPy *Spring 2026; AI-assisted review Oct 2026*

- Wrote k-means and linear/logistic gradient descent from scratch.
- An AI assistant's review (disclosed) traced 90.7% accuracy to an unconverged bias; standardizing gave 99.3%. I now check convergence and scaling before trusting accuracy.

Grocery Transaction Data Warehouse | SQL Server, T-SQL *Jul–Aug 2025*

- Built a Medallion warehouse over 131M+ transaction lines with stored-procedure ETL, a star-schema Gold layer, and 21 data-quality checks; tuning cut processing time 25%.

Optimizing Decision Trees with Information Theory | Information Theory course project, co-written *Fall 2024; AI-run follow-up Oct 2026*

- Co-wrote an 18-page paper and slides with a classmate; wrote the entropy and information gain, feature selection, and pre-/post-pruning sections; we checked each other's parts.
- AI-run follow-up I commissioned (disclosed in repo): on 48,842 census records, pruning raised accuracy ~4 points but widened the women–men false-negative gap 6.7→9.9 points.

EXPERIENCE

Data Analyst Fellow | EDC Lancaster County, Center for Regional Analysis (via The Shalom Project) *Aug 2026 – Present*

- Replaced manual Excel downloads with an automated pipeline, validation checks and a star-schema Power BI model (25+ BLS, BEA, Census, USPS series; 20 years; daily refresh), saving 15+ hours/month.
- Gathered requirements from a business analyst and the director, then built a Google Apps Script web app with a Gemini anomaly-flagging assistant so 6+ non-technical colleagues update and explore data without code.

Data Science Intern | KCC Capital Partners, Sheridan, WY *Jan – May 2026*

- Cleaned data on 100+ mining companies in Excel, enriched contacts with Seamless.AI, screened with the firm's built-in AI tools and reviewed a 100+ buyer shortlist; automated weekly reports, saving 5+ hrs/week.

Data Engineer Intern | KPIM Joint Stock Company, Hanoi, Vietnam *Jan – Apr 2021*

- Built Python and SQL Server ETL pipelines that automated legacy client data workflows, cutting processing time 30%.

LEADERSHIP & COMMUNITY

Shalom Project | Lancaster, PA gap-year program *Aug 2026 – Present*

- Live with a cohort; volunteer with local non-profits, including serving meals to people experiencing homelessness.

Technology Committee | Diplomatic Congress, Franklin & Marshall College *Aug 2021 – Aug 2022*

- Led IT policy work and centralized license management across 15+ departments, cutting classroom tech costs 18%.

SKILLS & CERTIFICATIONS

- **Programming:** Python (pandas, NumPy, scikit-learn, PySpark), SQL (T-SQL, Spark SQL), R, C++, Java
- **ML:** regression, logistic regression, k-means clustering, model evaluation, Bayesian networks, probabilistic reasoning
- **Data:** Microsoft Fabric, SQL Server, Power BI, ETL, data modeling, Git
- **Certifications:** Microsoft Fabric Data Engineer Associate (DP-700), Oct 2026; HackerRank SQL (Advanced), Sep 2025