Joshua Allen PhD

Data Scientist

joshua.f.allen@gmail.com joshuafayallen.github.io @joshuafayallen Joshua Allen

Applied data scientist with 6+ years of experience in building end-to-end Bayesian models, causal inference pipelines, and production-ready data systems. Experienced with translating complex statistical models into business-facing products, including interactive web applications, and open-source datasets used by researchers worldwide. Proficient in Python, R, SQL, and modern ML tooling.

Experience

Applied Scientist

Georgia State University 2019 Aug. — 2025 Aug. Atlanta, GA

- Implemented and managed end-to-end data pipelines covering all phases of the data lifecycle using R and Quarto. Pipeline improved data deliverable time from 3 days to less than 2 hours.
- Applied NLP and LLMs to 3TB+ text corpora for named entity recognition and regression on embedding representations.
- Leveraged Double Machine Learning, Synthetic Control Method, and sensitivity tests to quantify treatment effect uncertainty for stakeholders.
- Built 18+ production-grade webscrapers and developed open-source packages in R and Python packages for the largest open-source political executive communication text dataset.
- Mentored junior data scientists and advised colleagues on statistical analysis, scientific programming, and computational reproducibility.

Instructor of Record

Research Methods 2023 Jan. — 2024 Aug. Atlanta, GA

- Instructed 100+ students annually in applied research design, including problem framing, causal inference, and research design.
- Developed curriculum translating abstract statistical concepts, DAG-based causal reasoning, linear regression, and hypothesis testing into structured frameworks for non-technical audiences.
- Developed original course materials, assessments, and programming exercises emphasizing reproducible analytical workflows and evidence-based decision making.

Projects

Bayesian Marketing Mix Model *Blog Post* - *Project Link*

- Adapted Marketing Mix Modeling (MMM) to quantify ROI for offensive play callers' use of personnel packages.
- Navigated practical identification challenges, including sparse panel data, endogenous channel usage, and left-truncated career histories.

The Football Factor Model *Blog Post* - *Project Link*

- Modeled player performance based on values over replacement outcomes using Bayesian hierarchical regression using PyMC and R
- Translated Bayesian high-density intervals into actionable roster recommendations, producing a 5-point per game improvement in fantasy football scoring.

Football Power Rankings *Full-Stack Web Application* - *Project Link*

- Built a Bayesian Bradley-Terry model to rank NFL teams and incorporate uncertainty in estimation.
- Designed a full-stack interactive web app using FastAPI, React, and D3.
- Improved application scalability by transitioning from live model execution to persistent cache, reducing load time from minutes to milliseconds.

Objective

Seeking applied data science roles where rigorous modeling and business impact intersect.

Education

Georgia State University
2019 - 2025 Atlanta, GA

PhD in Political Science

Georgia State University
2017 - 2019 Atlanta, GA

MA in Political Science

Sonoma State University
2013 - 2017 Rohnert Park, CA

BA in Political Science

Technical Expertise

• Causal Inference • Bayesian Statistics • Quasi-Experimental Design • Hierarchical Modeling

Languages

• Python • SQL • R • Stan

Libraries and Frameworks

• Polars • Pandas • Tidyverse • PySpark • PyMC • Brms • Scikit-learn • Tidymodels • FastAPI • React • D3.js

Infrastructure

• Docker • Git • Quarto • RShiny