

# KALEB JORDAN

kalebjordan888@gmail.com | [linkedin.com/in/kaleb-jordan-888k](https://www.linkedin.com/in/kaleb-jordan-888k) | [github.com/kjord3](https://github.com/kjord3)

## EDUCATION

### Northeastern University

M.S., Data Analytics Engineering

Boston, MA

May 2026 – Present

### University of Illinois Urbana-Champaign

B.S., Statistics; Data Science DISCOVERY Certificate

Champaign, IL

Aug 2021 – May 2025

**Relevant Coursework:** Statistical Modeling, Statistical Programming, Linear Algebra, Data Science Programming

**Certifications:** Google Data Analytics Professional Certificate (Oct 2025)

## SKILLS

**Languages:** SQL, Python, R, SAS, DAX

**Visualization:** Power BI, Tableau, Excel (Pivot Tables, Power Query)

**Analytics:** Regression, Time Series Analysis, A/B Testing, KPI Development

**Data & Cloud:** ETL, Data Modeling, AWS (S3, Redshift, SageMaker, Athena), Microsoft Fabric, Git

## EXPERIENCE

### Data Analytics Extern — Pfizer (via Extern) | Remote | Aug 2026 – Oct 2026

- Built AI pipelines to process pharmaceutical vendor documents using OCR (Tesseract, PaddleOCR) and RAG for structured data extraction and classification.
- Developed a RAG-based document retrieval system with LlamaIndex, optimizing chunking strategy and evaluating open-source LLMs (Mistral, Phi-2) for retrieval accuracy.
- Benchmarked OCR accuracy, RAG retrieval quality, and routing performance end-to-end; delivered a technical report on model trade-offs and deployment recommendations, plus a demo UI.

### Machine Learning & AI Intern — Applied Technologies, University of Illinois | Champaign, IL | Aug 2024 – Dec 2024

- Built predictive models (XGBoost, GBM) in Python to forecast operational outcomes, improving forecasting accuracy by 25% and supporting strategic planning decisions.
- Conducted exploratory data analysis on 10,000+ records, identifying key trends and performance indicators that informed process improvements.
- Designed interactive Power BI dashboards adopted by multiple stakeholder groups, cutting manual reporting effort by 20% and centralizing access to key metrics.

### Business Process Improvement Intern — Applied Technologies, University of Illinois | Champaign, IL | May 2024 – Aug 2024

- Analyzed cross-functional workflow and operational data across 3+ departments, identifying inefficiencies and driving process changes that increased efficiency by 15%.
- Built Power BI dashboards to track KPIs and operational performance, enabling leadership to monitor trends and reducing manual reporting time by 20%.
- Partnered with stakeholders across University Housing and Campus Police to redesign shared workflows, improving system integration and cutting data redundancy by 15%.

## PROJECTS

### Run Expectancy as a Markov Reward Process

Mar 2026 – Jul 2026

*Independent Research — Isolating Pitcher Quality from Situational Leverage*

- Reformulated the classical RE24 run-expectancy matrix as the value function of an absorbing Markov reward process ( $RE = (I - P)^{-1}R$ ), fit on ~1.27M plate appearances of Statcast data (2019–2026) via the `baseballr` package, validating the solved matrix against the classical estimator to within 0.037 runs.
- Designed an out-of-sample, cluster-robust framework (by pitcher) to adjudicate between two competing explanations for situational scoring shifts — game leverage vs. pitcher quality — using multinomial/Poisson state-transition models, pooled inverse-variance testing, and sign tests across 24 base-out states.
- Found a robust, statistically significant pitcher-quality effect (pooled +0.047 runs,  $p = 1.4 \times 10^{-11}$ , positive in 24/24 states) that replicated across 3 tier-construction methods, 3 train/test splits, 3 quality measures (including one built purely from pitch velocity/spin with no outcome data), and both pitch-clock and pre-clock eras — while the naive leverage effect failed to replicate, isolating pitcher selection and in-game fatigue as the true drivers.

GitHub: [github.com/kjord3](https://github.com/kjord3)

### Predictive Analysis of March Madness Seedings

Jan 2025 – Apr 2025

- Developed an ordinal logistic regression model to classify NCAA tournament seedings from team-level and contextual features, achieving 94% classification accuracy (within  $\pm 1$  seed) on held-out data.
- Delivered an analytical report and presentation translating model outputs into actionable insights for tournament forecasting and decision-making.