

Ishmeen Garewal

New York, NY | (929) 996-8705 | ig2548@columbia.edu | <https://www.linkedin.com/in/ishmeen-garewal/> | <https://ishmeen-11.github.io/Ishmeen-Portfolio/>

EDUCATION

Columbia University

Master of Science in Data Science, GPA: 3.83/4

Courses: Machine Learning, Statistical Inference, Causal Inference (DiD, IV, RDD, DAGs), Forecasting, Applied Deep Learning, Data Visualization

Fr. Conceicao Rodrigues Institute of Technology

Bachelor of Engineering in Computer Science, CGPA: 9.66/10

New York City, NY

Expected Dec 2026

Mumbai, India

Graduated Jun 2025

TECHNICAL SKILLS

Programming: Python, C++, SQL, R, Java, JavaScript

Machine Learning: PyTorch, TensorFlow, HuggingFace, Scikit-learn, Graph Neural Networks, CNNs, Transformers, NLP, Time Series

Data & ML Infrastructure: ETL Pipelines, Feature Engineering, Data Validation, Vector Databases, REST APIs (FastAPI, Django), BigQuery, MySQL, MongoDB

Systems & Tools: AWS (S3, EC2), Docker, Git, Unix/Linux, CI/CD

Data Processing: NumPy, Pandas

EXPERIENCE

Artie

New York City, NY

Data Science Intern

Jun 2026 - Present

- Built a YouTube reach-forecasting pipeline in Python; engineered per-video features and fitted view-accumulation curves to place videos on a common maturity basis, enabling a reliable read on final performance within the first week of publication.
- Built a walk-forward, point-in-time back-testing harness with bootstrap resampling to measure bias, spread, and interval coverage which caught and removed a data-leakage bug that had overstated interval coverage by $\sim 5\times$, then re-established the model's advantage over a naive baseline on corrected, leakage-free splits.
- Deployed the forecast on the company's first live sponsored campaign: the pre-publication range contained the realized outcome, which landed more than an order of magnitude below the creator's own expectation, replacing hype-based pricing with a defensible range.
- Built brand-to-inventory matching engine over 500+ scene-level placements and 30+ advertiser briefs, hard brand-safety gating, lexical retrieval, and sentence-embedding similarity for semantic category matching, expanding share of briefs returning viable matches from $\sim 40\%$ to $\sim 90\%$.
- Validated ranking quality with no labeled ground truth using weight-sensitivity sweeps and per-signal ablation; found a near-constant feature was driving roughly half of all top-ranked results, traced it to a mis-scaled ordinal encoding, and froze the weights after the fix

Northeast Big Data Innovation Hub

New York City, NY

Graduate Student Assistant

Jan 2026 - Present

- Developed a digital twin framework for simulating student sleep outcomes using multivariate physiological and behavioral time-series data.
- Lead the Digital Twins initiative and the NSDC data-science program: mentor 20+ students on projects ranging from ML, NLP, Data Science.
- Delivered Advanced Excel and Advanced SQL workshops to 25+ MS Data Science students; built time-series feature pipelines and open data infrastructure for predictive modeling.

Deloitte Touche Tohmatsu India LLP - Financial Advisory

Mumbai, India

Intern

Jun 2024 - Jul 2024

- Reduced manual valuation processing time by 75% by automating FX deal valuation pipeline using QuantLib with REST API integrations
- Presented automation framework and policy recommendations to 3 Partners and 15+ senior leaders
- Built Django-based visualization dashboard with CI/CD deployment adopted by 40+ consultants across practice, enhancing reporting efficiency
- Analyzed RBI vs. NBFC climate risk policies, identifying 3 major gaps with actionable mitigation proposals

PROJECTS

Fashion Trend Analysis & Prediction

Sep 2025 - Dec 2025

- Built multimodal AI engine analyzing 150K+ Instagram fashion images, leveraging Google Vertex AI multimodal embeddings for real-time style similarity search.
- Engineered automated multimodal embedding pipeline using PyTorch autoencoders and Chroma vector database for scalable similarity search.
- Implemented vector similarity search over BigQuery-partitioned datasets, optimizing indexing and achieving sub-second retrieval latency.
- Built decision-support pipeline enabling designers to validate new product concepts against real-time social signals, reducing trend validation from weeks of manual research to seconds of automated analysis.

Heat's Toll on Health

Nov 2025 - Dec 2025

- Integrated 4 CDC public government datasets (87K+ records, 1,806 counties, 50 states, 2000–2022), resolving schema mismatches and geographic resolution gaps to build a unified analytical dataset queried in SQL and analyzed in R
- Engineered a clean complete-case schema by handling 52% suppressed mortality records, splitting composite variables, and standardizing across datasets with inconsistent state reporting and mirroring real-world open data pipeline challenges
- Built geospatial and multivariate visualizations using ggplot2, tmap, mosaic plots, time series and alluvial diagrams in R to communicate county-level health trends across all 50 states

PUBLICATIONS

- Garewal, I.K., Mahamuni, C.V., & Jha, S. "Emerging Applications and Challenges in Quantum Computing: A Literature Survey." 2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Port Louis, Mauritius.
- Garewal, I.K., Jha, S., & Mahamuni, C.V. "Topic Modeling for Identifying Emerging Trends on Instagram Using Latent Dirichlet Allocation and Non-Negative Matrix Factorization." 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)