(202) 868-9664
kangruiliu@email.com
kangrui-liu
krliu67

Kangrui Liu

Master's degree in Data Science with a strong foundation in statistical modeling, machine learning, and data engineering. Experienced in analyzing large-scale datasets (e.g., All of US, NHANES) using Python, R, and SQL for advanced analytics, data visualization, and reporting. Skilled in building data pipelines, database development, and data quality assurance. Adept at collaborating on interdisciplinary research projects in public health, social sciences, and behavioral analytics.

Experience

May 2025 - Present **Research Intern**, *Ma Lab@UMD*
- Web scraping *All of US* datasets by searching, extracting, and filtering biomarkers in **Python**. Auto data cleaning, summary statistics, and visualizations (histograms, boxplots) across **30+** variables using **R**.
- Designed logic to compute Allostatic Load (AL) scores and developed quality-check criteria to ensure complete biomarker data coverage. Handled missingness of AL using multiple imputation via the **mice** package in R.

Dec 2023 - Present **Graduate Research Assistant**, *Joint Program in Survey Methodology*, ***Award*** *for poster at "The Past, Present and Future of Statistics in the Era of AI"* conference
- **Co-authored** a forthcoming paper: "*Gradient-Boosted Pseudo-Weighting: Methods for Population Inference from Nonprobability Samples*", where we developed two Boosted Propensity Score methods integrating Gradient Boosting Machines into pseudo-weighting frameworks for selection bias correction in nonprobability samples. In real-world data, the relative bias in incidence of diabetes mortality improved by **20%**.
- Managed **SLURM-based HPC** environment for simulations, hyperparameter tuning, and parallelized bootstrap variance estimation.

Feb 2023 - May 2023 **Intern**, *Jianxin Technology*, Guangzhou, China
- **Corporate Resource Library Optimization:** Designed and deployed a **MinIO**-based file storage and transfer system with Fast Transfer and Instant Upload in **Java**; reduced average upload latency by 30%, enhancing document accessibility. Developed a file prioritization algorithm based on metadata (e.g., file size, classification), ensuring high-priority files processed within 5 seconds under peak load.
- **Power Plant Gate Detection Project:** Labeled 1,200+ images using LabelMe for gate status (open/closed) classification to train ML models. Built image preprocessing pipelines in **Python** (e.g., grayscale conversion, bounding box filtering), improving feature extraction accuracy by 18%.

Education

Aug 2023 - May 2025 **M.S. in Survey and Data Science**, *University of Maryland College Park*
Key Courses: **Machine Learning**, **Statistical Modeling**, Data Collection, Applied Sampling, Multiple Imputation, **Web Scraping in R**, Inference, Modern workflow for Data Science

Aug 2019 - May 2023 **B.S. in Information and Computing Science**, *Ningbo University of Technology*
Key Courses: Data Structures and Algorithms, **Database in SQL**, Object-Oriented Programming, **Data Visualization with Python**, Mathematical Modelling, Big Data in **Spark**

Projects

Dec 2024 **Political Media Consumption and Voting Behavior in the UK**, Applications of Statistical Modeling
- Analyzed European Social Survey data (**1,500+ respondents, 30+ variables**) to examine how political media consumption influences voting likelihood. Applied **GLMs**, **GLMMs** with random intercepts, and **GEEs** to estimate marginal and conditional effects; models achieved **74% AUC** in classifying voters vs. non-voters.
- Identified media exposure, age, and income as statistically significant predictors of voting behavior, while controlling for region and political trust.

Apr 2024 **Childlessness Trends in New Jersey**, Machine Learning for Social Sciences
- Analyzed **1,200+** survey responses to identify behavioral and demographic predictors of childlessness using R and SQL. Built an **ETL pipeline** for data cleaning and feature engineering; trained interpretable ML models (LASSO, XGBoost), achieving **AUC = 0.82**.
- Applied **cross-validation and grid search** to optimize performance, improving F1-score by **12%** over baseline.

Dec 2023 **Content Analysis of Reddit User Responses to the SFFA Case**, Data Display
- Built a scalable Reddit data ingestion and analysis pipeline to examine public response to the **Students for Fair Admissions v. Harvard** case, processing over **2,000 posts** and **30,000 comments**.
- Automated data collection via API, followed by **topic modeling** (LDA) and **sentiment analysis** (NLTK, TextBlob), uncovering 5 major discussion themes (e.g., *meritocracy*, *racial equity*, *legal fairness*) and detecting a **40% increase in negative sentiment** after the court ruling.