MICHELLE KOH

Atlanta, GA | michelle.koh@emory.edu | 512-461-2910 | github.com/kohguma

Education

Rollins School of Public Health, Emory University Atlanta, Georgia
Master of Public Health, Biostatistics & Data Science May 2025
Relevant coursework: Statistical Inference, Machine Learning, Python Programming, Regression Analysis, Data Visualization

The University of Texas at Dallas Richardson, Texas
Bachelor of Science, Biology May 2023

Skills

Programming Languages: R, Python, SAS, LINUX Shell scripting, MySQL
Programming Libraries: Ggplot2, Dplyr, Tidyverse, Lubridate, Scikit-learn**,** NumPy, Pandas, Sci-Py, Matplotlib, Seaborn
Statistics: Statistical Inference, Regression Analysis, Bayesian Statistics, Multivariate Analysis, Likelihood Ratio Tests
Data Science & Analytics: Data Wrangling, Data Cleaning, Data Manipulation, Predictive Modeling, Hypothesis Testing
Software Tools: Excel, PowerPoint, Word, Tableau, Git, Docker

Experience

Emory School of Medicine Atlanta, Georgia
Graduate Research Assistant Sep 2024-Present

- Processed and analyzed 100,000+ genetic variants using statistical methods and bioinformatics tools (e.g., PLINK, Linux) to identify significant associations with orofacial clefts, contributing to genetic research insights
- Implemented data quality assurance pipelines to detect and resolve missing data, duplicates, and inconsistencies, improving dataset accuracy to 98%
- Optimized workflow efficiency by deploying customized scripts written in Bash and R that processed complex genetic datasets systematically
- Applied principal component analysis (PCA) algorithm to detect population stratification and outliers in genetic data

Centers for Disease Control and Prevention Atlanta, Georgia
ORISE Fellow, Primary Data Analyst Nov 2023-Present

- Performed advanced statistical modeling to quantify the impact of multivitamin use on birth defect prevention, providing evidence-based recommendations to enhance public health policies and maternal care initiatives
- Conducted exploratory data analysis (EDA) on complex survey datasets using R, creating visualizations to identify trends in folic acid supplementation and update national prevalence estimates for multivitamin use
- Developed and standardized reusable R scripts for data analysis, improving team efficiency and consistency by 30%
- Collaborated with cross-functional teams to interpret findings and present results to scientists, supporting data-driven decision-making in folic-acid research

Projects

NYT Games Connections Analysis Atlanta, Georgia
Python, R, Bash, Git, GitHub Jan 2025

- Developed an interactive treemap visualization and summary table using R to highlight common New York Times Connections group names, enabling users to explore data insights dynamically
- Trained a Random Forest machine learning model in Python to predict the most frequent group name given four words, leveraging TF-IDF vectorization for feature extraction
- Automated data processing and analysis workflows using Bash scripting for efficiency
- Published the project on GitHub with detailed documentation, ensuring reproducibility for future revisions

Chronic Kidney Disease Project Atlanta, Georgia
R, Bash, Git, GitHub, Docker Dec 2024

- Engineered data visualizations (e.g., Scatterplots, Heatmaps, and Boxplots) on the UCI Machine Learning Repository's Chronic Kidney Disease dataset, a multivariate dataset with 400+ patient records, to uncover key trends and relationships
- Built a Docker container to standardize the environment and ensure reproducibility across platforms
- Hosted the project on GitHub with clear documentation, enabling seamless collaboration and version control

Honors And Certifications

- PH125.1x: Data Science: R basics Certification, Academic Excellence Scholarship, Rollins Pathway Scholarship