Jake Haines

908 323 3690 | me@jakehaines.com | linkedin.com/in/jhainesnc

Education

North Carolina State University — Raleigh, NC
Bachelor of Science in Statistics, Minor in Computer Science — *August 2021 – May 2025*

Experience

North Carolina State University — Raleigh, NC
Student Researcher, Data Analytics — *January 2025 – May 2025*

- Built ETL pipelines with **Airflow**, **PostgreSQL**, and **Docker**, processing **100M+ subway ridership records** from API, climate data, **5GB** geospatial raster tiles, and economic data, reducing runtime **45%** with workflow optimizations
- Conducted statistical analysis using mixed-effects models in **R**, revealing insights on borough and weather specific subway ridership trends, published on Cornell ArXiV (**arXiv:2505.02990**)

WEX — Raleigh, NC
Analytics Engineer Intern — *September 2024 – December 2024*

- Enabled vehicle fleet managers to analyze fuel expenses with an interactive web application developed with **Flask** and **Javascript**
- Reduced dashboard query times to **under 2 seconds** by optimizing embed rendering and advanced filtering using ThoughtSpot SDK
- Improved visibility into fleet spending patterns by embedding live dashboards for **10,000+ transactions** using ThoughtSpot SDK

Tesla — Palo Alto, CA
Data Engineer Intern — *August 2022 – December 2022*

- Optimized big data pipelines for **100k+ vehicles**, reducing DAG execution time by **56%** using **Airflow**, **Spark**, and **Python**
- Developed ETL pipelines to process **10TB+** of vehicle sensor data daily, enhancing efficiency and scalability using **Pyspark**
- Built real-time sensor monitoring system using **PostgreSQL** and **Tableau API**, improving sensor reliability and deployment tracking

Tesla — Fremont, CA
Data Science Intern — *May 2022 – August 2022*

- Enhanced prototype sensor reliability by **25%**, analyzing time-series data relationships using **Python**, **Pandas**, and **Plotly**
- Built geospatial data workflows to identify optimal testing locations, visualizing insights with Mapbox API and **Python**
- Developed statistical validation models in **Scipy** and **Scikit-learn** for **six prototype technologies**, improving engineering decisions

Myndmap — Princeton, NJ
Product Management Intern, Research & Cloud Infrastructure — *May 2023 – August 2023*

- Built cloud infrastructure and data schemas for user event tracking with **AWS** and **Firebase**, enabling scalable backend
- Conducted **literature review** on therapeutic interventions and translated findings into modular app feature design supporting cognitive restructuring, time management, and emotional regulation
- Spearheaded project operations with **Jira** and **Confluence**, enabling seamless cross-functional development and documentation

Projects

Data Lake for EEG Brain Wave Data Storage

- Built system for ingestion and transformation of raw EEG brain wave data using AWS (S3, Athena), Pyspark, Snowflake and MNE-Python, facilitating smooth preprocessing and storage of scientific data accessible through REST API built with FastAPI
- Developed pipeline monitoring and data quality system using dbt, Grafana, Prometheus, and AWS Deequ, enabling system reliability

Data Workflow & Statistical Anomaly Detection System for NYC Rental Costs

- Designed data pipeline using Airflow, PostgreSQL, and Docker, cleaning and transforming unstructured data from multiple sources
- Modeled statistical distributions and anomalies using linear regression, residual thresholds, and variance inflation factors, identifying neighborhoods with unexpectedly low rent prices and powering interactive visualization tools

Skills

Programming & Scripting: Python, SQL, R, JavaScript, Bash
Data Tools & Frameworks: Airflow, Spark, Hadoop, Snowflake, PostgreSQL, REST, FastAPI, AWS Athena, Deequ
Cloud & Infrastructure: AWS (S3, RDS, Lambda, IAM), GCP (Firebase), Docker, Prometheus, Grafana
Libraries & Visualization:: Pandas, Scikit-learn, Statsmodels,, MNE-Python, Tableau, Seaborn, Matplotlib