Minh (Halley) Dao

0426051477 | halleydao2004@gmail.com | www.linkedin.com/in/halley-dao | https://github.com/halleyd2804 | Halley's Portfolio

EDUCATION

Bachelor of Science, Computing and Software Systems Jul, 2022 - Jul, 2025
The University of Melbourne

- Recipient of the Melbourne International Undergraduate Scholarship, awarded for academic excellence and potential
- Awarded Melbourne Plus: Innovation for contributions to innovation-focused extracurriculars and projects
- Relevant Coursework: Machine Learning, Elements of Data Processing, Probability for Statistics, Statistics, Econometrics 1/2, Database Systems, Foundations of Fintech

RELEVANT EXPERIENCE

Data Engineer Intern Jul, 2024 - Oct, 2024
Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC

- Implemented in-memory caching for datasets, resulting in 20% improved data retrieval speeds, reduced system crashes, and enhanced user experience
- Conducted benchmarking analysis of data storage formats (HDF5, Parquet, RDS), recommending RDS to optimize data access in R-based workflows and improve processing efficiency for large-scale machine learning tasks
- Evaluated the performance of R-based data analysis pipelines on Nectar VMs, benchmarking against Python workflows and presenting findings through interactive visualizations to support stakeholder decision-making

Data Analyst Intern Jul, 2023 - Jul, 2023
Viettel Digital, Hanoi, Vietnam

- Utilized Tableau (including Level of Detail (LOD) expressions, calculation fields, table calculations, parameters, and data blending) to design interactive, dynamic visualisations
- Cleaned and pre-processed over 100,000 records using Excel to ensure accuracy and consistency, streamlining analysis and improving overall quality of visualisations and insights

TECHNICAL SKILLS

- **Data Analysis & Visualization**: Tableau, MS Excel, R (data.table, dplyr, AER), Python (NumPy, Pandas, Matplotlib, NLTK, Beautiful Soup)
- **Database & Query Languages**: SQL (MySQL, PostgreSQL), MongoDB, Data Modeling
- **Statistical Analysis**: Exploratory Data Analysis (EDA), Statistical Modeling, Probability Theory
- **Machine Learning**: Scikit-learn, TensorFlow, Classification, Regression, Clustering techniques
- **Collaboration Tools**: Jira, Confluence, SharePoint, Git, GitHub

RELEVANT PROJECT WORK

Global COVID-19 Analysis (2020–2021)

- Conducted exploratory data analysis (EDA) on global COVID-19 datasets using SQL, identifying key patterns in confirmed cases, deaths, and recoveries across 2020–2021
- Designed and deployed interactive Tableau dashboards to communicate insights into pandemic hotspots, mortality rates, and recovery trends to non-technical audiences

Econometric Analysis of Education's Role in Malaria Prevention in Burkina Faso (2018)

- Analyzed data from 5,839 individuals in Burkina Faso to estimate the causal effect of primary education on malaria risk using multiple regression models (OLS, IV-2SLS) to evaluate robustness and address potential endogeneity and selection bias
- Cleaned, merged, and engineered features across multiple datasets using R (dplyr, data.table) to build a high-quality dataset for econometric analysis
- Validated instrument variables and delivered insights linking education to health outcomes to support policy decisions

Personalized Book Recommendation Systems

- Create recommendation systems using Surprise for collaborative filtering and Scikit-learn for content-based models, achieving 88% accuracy in generating relevant item recommendations
- Performed extensive data cleaning and preprocessing using Python (Pandas and NLTK for text analysis)
- Created visualizations to identify user behavior patterns and product relationships using Matplotlib

Multi-Model Traffic Sign Classification with Deep and Ensemble Learning

- Engineered image features to train classical ML models (Random Forest, SVM, XGBoost) and built a stacked ensemble with Logistic Regression, while also designing a high-performing CNN in TensorFlow using raw image data, achieving up to 97% accuracy.
- Conducted thorough error analysis to identify model limitations and improve classification performance across different traffic sign categories