HAIYU XIAO

86-18986189523 | xiaohaiyu2022@yeah.net | https://hx33.github.io/

EDUCATION

Georgetown University Washington D.C.

Master of Science in Data Science & Analytics GPA: 3.56/4.0 Sept. 2022--May 2024

Core Courses (All courses coded in Python and R): Time Series Analysis, Advanced Data Visualization, Statistical Learning for Analysis, Big Data and Cloud Computing

University of Washington Seattle, WA

Bachelor of Science in Applied Mathematics GPA: 3.41/4.0 Sept. 2018--May 2022

Core Courses: Scientific Computing (MatLab), Basic Programming (Java), Computational Methods for Data Analysis (Python & Machine Learning Methods), Statistical Methods (R)

RESEARCH EXPERIENCE

Traffic Safety Research: Injury Severity in Fatal Single-Vehicle Crashes Jul. 2025--Sept. 2025

- **Research Paper:** *Injury Severity in Fatal Single-Vehicle Crashes: The Role of Restraint Use and Vehicle Size*, (First Author: Haiyu Xiao). Accepted for publication in Conference on Mechatronics, Robotics and Automation (ICMRA 2025), forthcoming.
- Coded in R to analyze 3,483 front-seat passenger cases from the FARS dataset (2021–2023) in single-vehicle crashes where the driver died, focusing on restraint use and vehicle type
- Conducted EDA using frequency tables, ggplot2, and ggmosaic plots, demonstrating seat belt use is linked to significantly lower fatal and serious injury rates, especially in heavier vehicles
- Built linear regression models with interaction terms to quantify the effects of restraint use and car size on injury severity; unbelted status increased the predicted severity by 0.58 points on a 0–4 scale
- Concluded that seat belt use has a stronger protective effect in heavier vehicles, and that unbelted passengers face disproportionately higher risk regardless of car size

Big Data Project: Analysis of Recent Reddit Data about Soccer Sep. 2023--Dec. 2023

- Analyzed 5.6 million Reddit comments (Jan-Mar. 2023) from soccer subreddits using PySpark and AWS SageMaker, focusing on comment content, timestamp, controversiality, and score
- Employed EDA to explore the correlation between comment volume and match frequency, finding a moderate positive correlation ($r = 0.614$)
- Utilized NLP to score sentiment and assess fan reactions, noting that sentiment skewed negative on match days, regardless of team results
- Built machine learning models (logistic regression, SVM, neural networks) to predict match outcomes based on pre-match Reddit sentiment and discussion volume, achieving 88% accuracy using cross-validation

Time Series Analysis on Predicting Crops Production Jan. 2023--May. 2023

- Developed a project portfolio using Quarto, based on USDA data and leveraging R and Python
- Analyzed monthly US wheat yields since 2002, identifying an increasing trend, strong seasonality, and moderate autocorrelation
- Forecasted yields over 12 months using ARIMA/SARIMA, RNN, and GRU models
- Optimized model parameters via ACF/PACF plots, AIC scores, and K-fold cross-validation

- Compared forecasts with benchmark methods, concluding deep learning models outperformed others in accuracy (RMSE), while time series models surpassed benchmarks

Data Visualization Project: Refugee Crisis in Europe in the 21st Century Jan. 2023--May. 2023

- Created a visualization project using Quarto and flexdashboard with Python
- Analyzed the impact of the Russia-Ukraine invasion on refugees, using data from UNHCR since 2000
- Visualized refugee flows in Europe with box plots, choropleth maps, and linked bar/line charts, highlighting demographic correlations between factors like age and gender
- Concluded that the Middle East and Balkan countries have been key sources of refugees to Europe, with the Russia-Ukraine war significantly increasing nearby refugee numbers

Study of Student-Teacher Ratio in Public Schools Using Statistical Methods Oct. 2022--Dec. 2022

- Programmed in R to classify student-teacher ratios as "high" (worse) or "low" (better) based on the median
- Collected data from the NCES's Public School Characteristics 2020-2021 dataset, including parameters such as student enrollment (grades 1-12), school locale, and Title-I eligibility
- Applied NNET, logistic regression, and bagging with decision tree, compared model accuracy via confusion matrices, and calculated variable importance for each model
- Concluded that the bagging model achieved 81% accuracy, with Title-I eligibility being the most influential factor

INTERNSHIP EXPERIENCE

CPS Trading Limited Fremont, CA

System Software Test Engineering Intern Sep. 2024--Apr. 2025

- Analyzed data from the company's wireless charging system, including device model, charging time, and efficiency
- Developed an interactive data visualization tool in Python (Dash/Flask) with a color-coded dashboard, box plots, and dynamic charts
- Provided actionable insights to developers, such as optimizing charging efficiency by adjusting charging power and duration based on device model, and sending personalized user notifications for timely charging

Peltast Partners New York, NY

Data Analyst Intern Jun. 2020--July. 2020

- Developed a web scraper with Python and Octoparse to gather office demographic data for a client seeking office in San Diego, and presented results on an interactive map using Excel and PowerPoint
- Utilized Twitter API to conducted sentiment analysis for an investor on consumer views of leading car brands, using graphs for sentiment scores and word clouds for most associated opinions of each company

PROFESSIONAL SKILLS

Programming: Python, R, SQL, Matlab

Data Visualization Tools: Plotly, Matplotlib, Tableau

Technical Skills: HTML and Quarto, Data/web scraping and APIs, Excel