

# Yuqing (Summer) Xia

## EDUCATION

**Northwestern University**

*M.S. Data Science*

**University of California, Santa Barbara**

*Major: Statistics and Data Science B.S.*

*Dean's list: three quarters | Honors Program*

**Chicago, IL**

*Sep.2023-Aug.2024*

**Santa Barbara, CA**

*Sep.2019-June.2023*

## SUMMARY OF SKILLS

- Programming Languages: R (packages: shiny, keras, tensorflow, etc.), Python ( packages: pandas, matplotlib, seaborn, Pytorch, NumPy, SciPy, Keras, TensorFlow, PySpark, etc), SQL, C++, Javascript, Linux, HTML
- Technical Tools:
  - Databases: Alicloud, MongoDB, MySQL workbench, PostgreSQL, Dbeaver, Neo4j
  - Data visualization: PowerBI
  - Cloud based platform: AWS SageMaker, Databricks
  - Other: Excel (VLOOKUP), KNIME (KNIME Python integration), Godot, SAS, Celonis (process mining)

## PROFESSIONAL EXPERIENCE

**Northwestern University School of Professional Studies**

**Teaching Assistant (Sept 2024-Present)**

- In person: Applied Statistics with R, Remote: Database Systems, Practical Machine Learning
- Holding sessions to provide additional help and answer questions. Explaining course material and concepts to students. Organizing and overseeing group discussions or projects.

**Realix AI**

**Data Science Intern (March.2024-Oct.2024)**

- Prompt engineering: Fine tune training model (LLM), increased the BLEU score by 20%
- Led the data team. Text processing and data cleaning, version control, seed transcript generation. Designed and created automated ETL pipeline, transforming raw user conversation data into ready for use training data.

**Baker Tilly**

**Data Science Intern (June.2024-Aug.2024)**

- Identify and report data integrity issues and the potential origin of such issues
- Data storytelling with visualizations in PowerBI
- Develop machine learning model to identify underutilized offices, projected to improve the cost efficiency by 10%

**ARK.IO**

**Data Analyst Intern (Jul.2022-Sep.2022)**

- Monitored and managed AlibabaCloud database (memory and CPU utilizations, indexes, etc), retrieved and updated data using SQL queries
- Generated timely reports on user and post data with Tableau
- Determined the most active users and created banners for them to boost user interaction and stored the data to cloud for future analysis

## PROJECTS AND RESEARCH

**Capstone with CalCOFI**

**Jan 2023- May 2023**

*An eDNA window into larval fish habitat, ecosystem structure, and function using CalCOFI data*

- Conducted preliminary analysis, data cleaning, and model development for 18s sequence eDNA datasets
- Improved the overall interpretability of the data, pinpointed the issues within the data processing step and supervised the correction.
- Employed PCA to summarize the overall datasets, then used general linear model, and decision tree to build a predictive model for anchovy presence based on the eDNA.

**Survival Analysis Project**

**Fall 2022**

*Probability of world's political parties' leaders to stay in office for a certain time*

*Mentored by Professor Andrew Carter*

- Based on previous research by Horiuchi and Liang, this project tries to find the relationship between the time of election and the length of the political leader's term.
- Used R packages survival and survminer to plot the Kaplan Meier estimate probability, performed step AIC to select variables and created a Cox Proportional Hazard model, and finally explored the recurrent event model for the data.

**Time Series Project**

**Winter 2022**

*Ground level ozone in Los Angeles from 2000 - 2020*

- Visualized the time series with R, identified the time series as a seasonal ARMA model.
- performed spectral analysis on the model, and forecasted the ground level ozone up to March 2022 with 80% accuracy.