

# YEH CHAN YOO

Seattle, WA | yehchanyoo@gmail.com (personal) | yehchany@uw.edu (grad school) | linkedin.com/in/yehchan-yoo | yehchanyoo.github.io

## EDUCATION

**University of Washington - Seattle** *September 2024 - Present (Expected Completion in March 2026)*

Master of Science in Statistics - Advanced Methods and Data Analysis

- **Current Coursework:** Design and Analysis of Experiments, Statistical Inference, Statistics Seminars

**University of California - Berkeley**

*August 2017 - December 2019, August 2022 - December 2023*

Bachelor of Arts in Statistics and Political Economy with a Minor in Data Science

- **Note:** Paused undergraduate education from December 2019 to August 2022 due to mandatory military service in South Korea
- **Relevant Coursework:** Causal Inference; Concepts in Computing with Data; Concepts of Probability; Concepts of Statistics; Data, Inference, and Decisions; Econometrics (Math Intensive); Introduction to Artificial Intelligence; Linear Modelling: Theory and Applications; Modern Statistical Prediction and Machine Learning; Principles and Techniques of Data Science

## RELEVANT PROFESSIONAL EXPERIENCE

**University of Washington • Statistics Tutor**

*September 2024 - Present*

- Manages the environment of the Statistics Tutor & Study Center and provides tutoring for students in various undergraduate introductory statistics courses such as STAT 390: Statistical Methods in Engineering and Science and QSCI 381: Introduction to Probability and Statistics for 4.5+ hours a week
- Tracks study center usage and traffic to collect data on ways to improve the study center for future quarters

**Mindful Conversion • Data Scientist**

*January 2024 - August 2024*

- Served as the primary data scientist for the e-commerce marketing and data analytics agency Mindful Conversion, supporting several of the agency's product marketing and marketing analytics projects as part of a cross-functional team with both software engineers and professional marketers in a fast-paced environment
- Engineered and optimized the data pipeline for the company's predictive search engine optimization (SEO) product Kixelly to coordinate data mining, machine learning, and large language model (LLM) tasks using Python, Microsoft Azure, and Google Cloud Platform -- gathering and analyzing SEO data on ~100,000 keywords and URLs to improve Kixelly's SEO insights
- Led exploratory data analysis efforts on 10GB+ of marketing data to identify and resolve key data collection and content strategy issues, employing Python and SQL for robust data validation and business improvements
- Authored more than 5 data-driven reports and data visualizations that influenced the marketing strategies for the company and its clients

**UC Berkeley School of Education • Student Assistant/Undergraduate Research Apprentice**

*January 2023 - December 2023*

- Independently researched, designed, and engineered deep learning-based natural language processing (NLP) models for automatic short answer grading of 200+ student answers to mathematical and statistical questions to significantly improve grading efficiency; achieved 75% test accuracy with the use of a fine-tuned LLM (RoBERTa), Pytorch, and CUDA
- Graded and analyzed written answers to statistics assessment questions from more than 500 California high school and middle school students using Excel and R, contributing to refinements in assessment methodology and improvements in student learning outcomes
- Collaborated with a postdoctoral researcher on the design and writing of several of the project's statistical assessment questions for students in public secondary education

**UC Berkeley College of Computing, Data Science, and Society • DATA 102 Tutor**

*August 2023 - December 2023*

- Developed detailed assignment rubrics and provided individualized feedback for 250+ assignment submissions weekly in DATA 102: Data, Inference, and Decisions, enhancing students' mastery of advanced inference, modeling, and decision-making skills
- Graded labs, homework assignments, and exams for ~250 students under tight 7-day turnarounds
- Educated 20+ students weekly on complex data science concepts such as binary decision-making, neural networks, causal inference, differential privacy, and multi-armed bandits for 2+ office hours per week, ensuring a deeper understanding of these concepts

## RELEVANT PROJECTS

**Inference and Prediction on Crude Diabetes Prevalence in U.S. States Based on Vegetable Consumption**

*May 2023*

- Collaborated with 3 other teammates to examine the relationship between vegetable consumption and crude diabetes prevalence in U.S. states using inferential and predictive models for the final project in UC Berkeley's Data, Inference, and Decisions course; involved thorough analysis of CDC and U.S. Census data with 200+ data points using Python
- Found that increased vegetable consumption may have caused a decrease in diabetes prevalence in U.S. states with 95% confidence
- Demonstrated that random forest outperformed generalized linear models in forecasting diabetes prevalence in a U.S. state based on vegetable consumption, reducing train and test RSMEs by ~60% and ~25%, respectively

Link to paper: [https://yehchanyoo.github.io/DATA\\_CI02\\_Project\\_Final\\_Project\\_Submission\\_CD\\_TE\\_CM\\_YY\\_230512\\_.pdf](https://yehchanyoo.github.io/DATA_CI02_Project_Final_Project_Submission_CD_TE_CM_YY_230512_.pdf)

**Replication and Improvement on "How do 401(k)s Affect Saving? Evidence from Change in 401(k) Eligibility"** *December 2022*

- Worked together with a teammate to replicate, critique, and improve upon Professor Alexander M. Gelber's 2011 causal inference paper "How Do 401(k)s Affect Saving? Evidence from Changes in 401(k) Eligibility" using R for the final project in UC Berkeley's Causal Inference course based on an economic dataset with 4,000+ data points from National Bureau of Economic Research (NBER)
- Tested and replicated using doubly robust estimators the original paper's findings that 401(k) eligibility significantly increases 401(k) savings but not other types of savings such as individual retirement account (IRA) assets with 95% confidence
- Enhanced the original analysis by implementing additional econometric techniques such as propensity score matching

Link to paper: [https://yehchanyoo.github.io/group04\\_YY\\_XZ.pdf](https://yehchanyoo.github.io/group04_YY_XZ.pdf)

## SKILLS

- **Languages:** Python, R, SQL, Java, Git, Shell, LaTeX, Markdown, HTML, CSS, JavaScript
- **Programs/Platforms:** Jupyter Notebook, RStudio, Microsoft Excel, Microsoft Azure, Google Cloud Platform, Github
- **Packages:**
  - **Data Collection and Manipulation/Statistical Analysis:**
    - *Python:* NumPy, Pandas, SciPy, Xarray, Statsmodels, ArviZ, Requests, Beautiful Soup, Selenium
    - *R:* dplyr, tidyr, readr, stringr, data.table
  - **Machine Learning:** Scikit-learn (Python), PyMC3 (Python), PyTorch (Python), carat (R), e1071 (R), kernlab (R), mclust (R), nnet (R), Mice (R)
  - **Data Visualization:** Matplotlib (Python), Seaborn (Python), Plotly (Python), ggplot2 (R), igraph (R), shiny (R)