Yu Cai

10 River Road, Apt. 17A, New York, NY

(347)400-9068 | yuc4003@gmail.com | https://yuc4002.github.io/

Education

Cornell University | New York, NY — Dec 2019

- *Master of Science in Biostatistics and Data Science* — *GPA: 3.88*
- Coursework: Statistical learning, Data Management, Data Science, Statistical Programming

Hobart and William Smith Colleges | Geneva, NY — June 2017

- *Bachelor of Arts in Mathematics and Economics* — *GPA: 3.68*

Skills

- **Computer:** Java, R, SAS, MySQL, LaTeX, Tableau, Python, MS Office (Excel, PowerPoint, Word, Access)
- **Certification:** SAS Certified Base Programmer for SAS 9
- **Conference Talk:** MAA Seaway Section Conference and NCUWM Conference
- **Economics TA** for Principles of Econ., Econometrics, Macro & Micro, Statistics (SAS)
- **Mathematics TA** for 5 classes for a year, participate and offer suggestions in professors' weekly meeting

Professional Experience

Healthcare Data Analyst — 06/2019-09/2019

Weill Cornell Medicine — *New York, NY*

- Used Python to download online JOSON tables and used SQL quarry to join and clean a 310,000x200 table
- Applied multivariate logistic regression models with stratified analysis to explore the relationship between race and region in treatment for Low-Risk Prostate Cancer in the USA.
- Performed sensitivity analysis to deal with 20,000+ unknown data in the outcome variable

Phylogenetic Research Assistant — 05/2016-06/2018

Hobart and William Smith College — *Geneva, NY*

- Collaborated with Biology department to gather their requirements and provide the research status updates
- Proposed a quartet-based species tree structure algorithm with classification methods based on the literature review to improve the speed of estimation from one week to two days.
- Conducted literature review via Google Scholar and NCBI and conducted data analysis using programming languages such as Java, Perl, and R Studio in Linux System
- Presented the findings in figures created in Excel to show the extent to which the precision of the proposed algorithm has improved. Give presentations or poster talks for three conferences.

Data Analyst Intern — 08/2015-12/2015

Esperity — *Brussels, Belgium*

- Used SAS to collect data about gears that keep track on health-related information of the customers, such as Fitbit, Apple watch; presented the information in tables and graphs in Excel and write the results in reports
- Conducted literature reviews on Google Scholar on gears studies and user behavior to come out summary reports
- Caught the keywords by SQL from the unstructured data, like the users' chatting records, to construct a databased.

Selected Project

R packages with multiple hand functions(R)| *Biostatistics*

- 12 handy functions in 2 R packages, details can be found on my website. Here are some examples:
- Proved a good visualization of boxplots or bar chart for 1 or 2 predictors with p-values in one-line code
- Performed univariate or multivariate analysis with p-values in table and pretty plot in in one-line code

Database(SQL)| *Data Management*

- Built a structured database with form in Access for an unstructured data in Excel, with VBA code.
- Designed database schemas in MySQL with EER diagram by SQL queries and connect to Access by ODBC
- Pulled 100 JOSON tables from the web in R, union them and use SQL queries to analysis the dataset

Arrhythmia analysis prediction (R)| *Statistical Learning*

- Used R to clean up large raw data sets containing patient information from 4 databases, build statistical prediction models, use logistic regression, KNN, random forest, Lasso regression, Adaboost, support vector machine, and other methods to predict scale incidence rate
- The result comes shows the four ensemble methods did the best in prediction. The final model obtained AUC value of 0.8 accuracy in the 10-fold cross-validation, with 5 risk factors in advance.

Crime in Chicago (Python & R)| *Data Science*

- Used Hierarchical clustering and K-means clustered in Python to cluster the major 5 crimes in all community areas into 4 groups and plot them in the map with different colors
- Plot the first two principle components with all community areas and 4 crimes as arrows in the graph in R.