Sze Pui Tsang

szepui.tsang@caa.columbia.edu | (201) 496 1101| LinkedIn: Sallie Tsang | Personal Website: https://sallietsang.github.io

EDUCATION

Columbia University Mailman School of Public Health New York, USA
Master of Science, Biostatistics in Public Health Data Science 09/2021 - 07/2023
City University of Hong Kong Hong Kong
Bachelor of Science, Applied Biology with minor in Media Communication 09/2016 - 07/2020

SKILLS

- **Technical:** R (Shiny App, Markdown, ggplot), SQL, SAS, Python (Pandas, Numpy, NLTK), Tableau, QGIS, Geoda
- **Statical Analysis:** Survival Analysis, Natural Language Processing, Statistical Inference, Predictive Modelling, A/B Testing

EXPERIENCE

Columbia University

Data Analyst (Data Science Institute Scholar) 10/2022 - Present

- Engage in project HHEAR Data Repository to promote secondary analysis of pooled studies through data harmonization by achieving FAIR (findable, accessible, interoperable, reusable) and writing sharable script to GitHub
- Handle 26 environmental health datasets for data cleaning, descriptive statistics, and interactive dashboard via R
- Undergo analyses including summarizing exposure's biomarkers and comparing them to population from NHANES

Research Assistant 09/2022 - Present

- Preprocess time series data involving detrending data and checking stationarity and distribution of residues
- Measure transfer entropy value for directed pairs among biomarkers (e.g blood pressure, heartbeat)
- Brainstorm application of resulted transfer entropy value by considering its possibility applied to disease detection

Technical Service Division, HKSAR Government 07/2018 - 08/2018

Data Analysis Intern

- Participated in project ***Development of Image Processing Algorithm*** aiming to improve accuracy of optical character recognition (OCR) in recognizing serial number and character of Euro banknotes
- Trained OCR with 3500+ data characters and achieved 99.6% accuracy for character prediction
- Developed image preprocessing techniques to optimize banknote image and handled character labeling

PROJECTS

Natural Language Processing: Sentiment Analysis on Movie Review 10/2022 - Present

- Run sentiment analysis utilizing Naive Bayes and SVM classifiers on 400K+ reviews at Rotten Tomatoes using NLP
- Apply NLTK to perform extraction and tokenization of user comments to discern their sentiment preference
- Analyze relativity of features and predicted review impact with linear regression, aiming to achieve 85%+ accuracy

Machine Learning Project: Parkinson's Diseases Symptom Severity Prediction 03/2022 - 04/2022

- Conducted exploratory data analysis to visualize the relationship between 17 predictors and response in R
- Built 8 Supervised Machine Learning Models (Linear Regression, LASSO, Ridge, Elastic Net, GAM, MARS, PCR, Random Forest) to predict Parkinson's Disease severity based on patients' age, sex, and vocal features
- Tuned model parameters and selected MARS model with smallest median 10-fold cross-validated RMSE value

Data Science Project: Website and **Dashboard Building of the US Smoking Population** 11/2021 - 12/2021

- Cleaned 29482+ data and worked on team for descriptive statistics concluding 1.2% decrease on smoking population
- Utilized R(plotly, ggplot, gganimate) to visualize smoking geographical and demographic distribution by plot & map

Final Year Thesis, City University of Hong Kong 09/2019 - 05/2020

- Designed 8-month research experiment to investigate the negative impact of microplastic on marine animal mussels
- Performed statical analyses 2-way ANOVA & T-test comparing effect of plastic type and concentration on mussels
- Discovered high concentration nylon plastic; 7.2% more absorbable than low concentration natural cotton in mussels