# Wanchen Hong

(203)-768-1679 || wghong@bu.edu
LinkedIn: https://linkedin.com/in/wghong; GitHub: https://github.com/wghong02

## SUMMARY

Student seeking data scientist job opportunities. Strong background in statistics, machine learning, and proficient programming skills in **Python**, **R** and **SQL**. Previous project experience in **machine learning** and **data analytics**.

## EDUCATION

**Boston University** *Boston, MA* Expected Jan 2025
B.A./M.A. in Mathematics and Statistics; B.A in Physics and Computer Science *GPA: 3.85/4.0*
Coursework: Analysis of Variables, Probability, Computer Algorithms, Stochastic Processes, Applied Machine Learning, Linear Models, Causal Inference, Time Series, Generalized Linear Model, Deep Learning

## SKILLS

**Programming:** Python, R, SQL, Spark, C, MySQL, Java, C++, SAS, AWS, Tableau
**Machine Learning Techniques**

- Linear models, Decision Tree, Random Forest, Regression Tree, XGboost, LightGBM
- Principle Component Analysis, Regularization, Feature Engineering
- Neural Network, Q Learning, Markov Models

## EXPERIENCE

**Boston University** *Boston, MA*
***TCW Lab, Bioinformatics Research*** Jan 2022- Mar 2023

- Analyzed theimpact of gene's aging on Alzheimer's Disease (AD) through the mouse model using R.
- Performed RNA sequencing and single cell RNA sequencing with existing lab pipelines and R packages including scRNAseq, seurat, and deseq2.
- Identified key gene markers via gene enrichment analysis, such as GSEA and PEA.
- Developed a visualization pipeline that uses heatmaps, bar graphs, and upset graphs.

## PROJECTS

**San Francisco Crime Analysis in Apache Spark**

- Performed spatial and time series analysis on a 15-year dataset of reported incidents from SFPD.
- Built data processing pipeline based on Spark RDD, Dataframe, and Spark SQL for big data OLAP.
- Trained and fine-tuned an ARIMA model to forecast the number of theft incidents per month.
- Explored and visualized the variation of the spatial distribution of incidents over time.

**Automatic Liver Tumor CT Scan Segmentation**

- Collaborated with peers to improve upon UNet-based deep learning models for segmentation using PyTorch.
- Augmented data set with standardization and transformation using TorchIO.
- Implemented Atrous Spatial Pyramid Pooling (ASPP) and Attention layers into UNet++ and TransUNet architectures to improve feature extraction and segmentation accuracy.
- Achieved a 25% reduction in loss compared to the existing UNet architectures.

**Semantic Analysis for Youtube user Comments**

- Developed a Spark-based machine learning algorithm to categorize users according to their YouTube video comments.
- Cleaned the dataset by eliminating null entries and manually tagging a subset of users based on their commentary, and processed users comments via RegexTokenizer and Word2Vec in SparkML
- Tuned hyperparameters of Logistic Regression and Random Forest algorithms via k-fold cross-validation.
- Applied the TF-IDF methodology for feature extraction and implemented an unsupervised Latent Dirichlet Allocation (LDA) model to identify the top 5 topics among the target user group.

**Airbnb Rental Price Prediction**

- Developed machine learning algorithms to predict Airbnb rental prices with listing data in NY 2019 via Python programming.
- Preprocessed data set with data cleaning, feature standardization, categorical data encoding.
- Trained Linear Regression, Decision Tree, and Random Forest models, tuned parameters with k-fold cross-validation, and utlized boosting and ensembles to improve model performance.
- Reached an improvement of 40% in MSE compared to linear regression models.