CARY HTAN

(949) 9924422 | ch3756@columbia.edu | caryhtan.website | www.linkedin.com/in/cary-htan

Education

Columbia University, New York, NY **Expected graduation: May 2026**
Bachelor of Arts in Computer Science GPA: 3.54

Experience

Data Analyst Intern, **Fenics Market Data** **June 2024 – August 2024**

- Automated data validation in Quality Assurance using Pandas and Numpy which help reduced processing time by 80%
- Integrated MatplotLib and built a Streamlit app that sped up data anomaly and market bonds outlier detection process by 3x
- Developed a competitive analysis dashboard to extract revenue data from competitors via financial APIs and web scraping
- Implemented Power BI to visualize raw market data and assisted senior leadership in gaining insights/refining strategies

Bioinformatics Intern, Azul Bio **January 2024 – May 2024**

- Created a secure SSH key management system using Python's Click feature to streamline user access
- Designed custom CLI commands and built scripts to assist new users with secure server access
- Built a bead detection system using YOLOv7, Python, and Roboflow to automate bead classification

Technical Skills

- SQL (SQL Server, MySQL, PostgreSQL)
- Python (seaborn, matplotlib, scikit-learn)
- Tableau
- Microsoft Azure (DataBricks, Lake, Warehouse)
- PySpark
- Microsoft Power BI
- Excel
- Machine learning (PCA, KMeans)

Projects

Modeling Car Insurance Claim Outcomes **October 2024**

- Applied **Machine Learning** to build logistic regressions models and evaluated features for predicting insurance claims.
- Preprocessed customer data by replacing missing values with column means to ensure data completeness.
- Assessed model performance using confusion matrices and identified the best feature with an accuracy of 77.71%.

Reducing Traffic Mortality in the USA **August 2024**

- Analyzed road accident data with Python using **pandas**, **seaborn**, **matplotlib**, and **scikit-learn** for regression and clustering.
- Explored correlations and state-level patterns to identify variations in road accidents.
- Applied **PCA** and **KMeans** clustering to group states and provided target policy recommendations.
- Calculated fatal accident statistics and visualized cluster data to optimize resource allocation.

Word Frequency in Classic Novels **July 2024**

- Web scraped with **Python** and **BeautifulSoup** to extract text from classic literature, including *Moby Dick* and *Peter Pan*.
- Applied **Natural Language Processing** to identify the most frequent words in the extracted text.
- Analyzed word frequency distributions in selected novels and generated insights into linguistic patterns.

Analyzing Unicorn Companies **May 2024**

- Utilized **SQL** to extract data from multiple tables using **CTEs** and **JOIN** and analyzed unicorn trends.
- Aggregated data to identify the top three industries by unicorn count and average valuations in billions.
- Produced a final output table to deliver clear insights into industry growth and financial performance.