Sachin Otsuka Arjun

Data Scientist/Data Analyst

sangoten24@gmail.com
linkedin.com/in/sachin-otsuka-arjun
+1 240 618 6185
New York

Data Engineer with proven experience in designing scalable databases and optimizing data extraction pipelines. Skilled in implementing data quality frameworks, web development, and enhancing website responsiveness. Committed to improving platform performance and delivering clean, analysis-ready datasets.

Work Experience

Data Engineer

Aug 2023 - Jul 2024

PrivateBlok | Bengaluru, India

- Architected and maintained a scalable relational database by integrating datasets from MCA and external APIs, consolidating structured and semi-structured data from 1,000+ Indian Private Limited Companies to improve platform performance and accessibility.
- Developed and optimized automated web scraping pipelines, implementing multithreading to boost extraction speed by 400% and reduce ETL cycle time, enabling near real-time data availability for analytics teams.
- Implemented end-to-end data quality frameworks with anomaly detection and QA checks, ensuring 99%+ data accuracy; improved cross-team productivity by 20% through delivery of clean, analysis-ready datasets.

Web Development Intern

Sep 2022 - Oct 2022

CoachEd | Mysore

- Developed a dynamic Restaurant-Reservation System website using HTML, CSS, SASS, JavaScript, and Bootstrap
- Enhanced responsiveness of the website to ensure seamless adaptation across various devices and screen sizes
- Executed regular code reviews and integrated industry best practices to optimize website performance and security protocols

Projects

Predictive Modeling of Click-Through Rate(CTR) Using Advanced Machine Learning Techniques

Nov 2024 - Dec 2024

- Evaluated and analyzed complex datasets to identify factors influencing Click-Through Rate (CTR), applying rigorous data cleaning, exploratory data analysis (EDA), and statistical testing (ANOVA) to optimize model accuracy.
- Enhanced model performance by strategically handling missing values through techniques such as median/mode imputation and native handling capabilities of XGBoost, significantly reducing error rates.
- Implemented advanced modeling techniques, including XGBoost and Random Forest algorithms, significantly improving predictive accuracy (RMSE) by effectively managing missing data and categorical variables.

Geo-spatial analysis of New York Neighborhoods and Living Satisfaction

Sep 2024 - Dec 2024

- Conducted geo-spatial analysis to identify key determinants of neighborhood satisfaction for first-time residents in New York City, addressing decision paralysis among approximately 300,000 annual movers.
- Leveraged data extraction and ETL processes using Google Maps API, Zillow Research, NYC Open Data, and SimpleMaps to compile comprehensive cross-sectional datasets.
- Applied statistical modeling and linear regression analysis to identify factors strongly correlated with resident satisfaction and happiness, enabling targeted, personalized recommendations.
- Developed interactive Tableau dashboards and visualizations to present user-friendly, geo-spatial insights, enhancing data-driven decision-making and user engagement.

NYC-Crime Analytics – End to End Retrieval System

Jan 2025 - May 2025

- Built scalable ETL pipeline using PySpark to clean and standardize 8.9M+ crime records, correcting invalid dates and handling nulls.
- Managed MongoDB for raw storage and PostgreSQL for structured analytics data with scalable architecture (sharding, read replicas).
- Designed and queried Neo4j relationship models to uncover hidden crime-location-time patterns.
- Developed interactive Flask/HTML dashboard integrating Plotly charts and Folium maps with keyword and borough-level filtering.
- Implemented data search, filtering, and pagination features for crime records table.
- Applied data quality validation to ensure accuracy, completeness, and consistency.

MLS Soccer Analytics – Archetype & Anomaly Detection

Jan 2025 - May 2025

- Collected MLS player performance and demographic data via American Soccer Analysis API, integrating datasets using R.
- Applied K-means clustering to offensive players to identify three archetypes: Low Impact Players, Reliable Starters, and Elite Playmakers.
- Validated archetype and positional differences using Welch’s T-tests, ANOVA, and pairwise t-tests.
- Built linear regression and decision tree models to predict goals+assists and points added from physical traits, identifying age as the strongest predictor.
- Developed anomaly detection to flag players over- or underperforming relative to physical expectations, combining quantitative analysis with qualitative insights.

Core Skills

Languages:: HTML5, SQL, Javascript, CSS, Python, C Programming, R

Frameworks: Docker, Neo4j, Spark, MongoDB, PostgreSQL, MySQL, Tableau

Tools and Standards: NumPy, Scikit-learn, Git, Pandas, JSON, REST API, PySpark, Postman, Tensorflow, HTTP

Skills: Big Data Engineering, Data Integration, Data Structures, Operating Systems, Natural Language Processing, Algorithms, DevOps, Database Management, Micro-service Architecture, Data Warehouse, Data Visualization, A/B Testing, Data Lake, Technical competence, Performance metrics, Self-driven, Communication, Collaboration, Storytelling, Presentation Skills, Creativity

Cloud: Amazon Web Services(AWS), Microsoft Azure

Education

Columbia University in the City of New York

Sep 2024 - Present

Master of Science Applied Analytics
GPA: 3.75/4.0

Vidyavardhaka College of Engineering

Aug 2019 - Jun 2023

Bachelor of Engineering Computer Science & Engineering
GPA: 3.8/4.0