Jiayi Shi

New York, NY | (347) 483-4722 | js6857@columbia.edu | LinkedIn

EDUCATION

Columbia University — **New York, NY**
M.S. in Statistics | GPA: 4.00/4.00 — Expected: Dec 2026

The Chinese University of Hong Kong(ShenZhen) — **Shenzhen, China**
B.S. in Data Science & Big Data Technology (Honours) — Graduated: Jun 2025

PROFESSIONAL EXPERIENCE

KCC Capital Partners — Los Angeles, CA (Remote)
Data Science Intern — Dec 2025 – Present

- Defined funnel metrics (onboarding, match rate, meeting booking) and built SQL-based tracking using CTEs, LEFT JOINs, and window functions on clickstream and session tables, enabling full-funnel visibility
- Analyzed over 50k user interaction logs using Python (Pandas, NumPy, datetime), identified top failure patterns (e.g., unparseable time formats, repeated slot collisions), and reduced booking errors by 22% through process redesign
- Developed a rule-based internal matching engine using onboarding data (industry, region, ticket size); scored and ranked 2k+ investor-startup pairs with explainable logic, increasing match engagement by 18%
- Automated daily performance dashboards through Google API, surfacing key KPIs to inform weekly product iterations

Victoria Solutions — London, UK
Data Analyst Intern — Oct 2025 – Dec 2025

- Cleaned messy sales data using SQL and Python, resolving missing values and standardizing formats for analysis readiness
- Identified seasonal peaks and product trends by aggregating monthly revenue by category and region in SQL, then visualizing trends with Matplotlib; revealed Q4 spikes in Electronics and weak year-round performance in the West
- Built a sales prediction model and customer segmentation, uncovering high-value, low-frequency segments for retention
- Created Power BI dashboard to monitor KPIs, driving a bundling campaign projected to boost off-season sales by 20%

DingTalk Information Technology (Subsidiary of Alibaba, Fortune Global 500) — Hangzhou, China
Data Engineering Intern — Aug 2024 – Sep 2024

- Deployed ETL pipelines to ingest and standardize 200K+ records from Salesforce CRM and local ERP systems, unifying schemas and timestamp formats to support real-time sync across internal business platforms through DingTalk's iPaaS
- Engineered API integrations, building automated workflows reduced manual handoffs and cut data sync latency by 30%
- Implemented data validation rules and logging checkpoints (using Python & SQL) to detect nulls, schema drift, and timezone conflicts, improving pipeline stability and enabling downstream sales analytics and customer engagement tracking

Xiangyang City Big Data Center — Xiangyang, China
Data Analyst Intern — Dec 2023 – Jan 2024

- Forecasted application volumes using Hive-SQL and Python, cutting citizen wait times from 40 to 20 minutes
- Engineered time-aware features (lags, holidays), achieving 12% MAPE with interpretability for non-technical stakeholders
- Delivered trend dashboards and forecast reports to department managers, helping improve staffing efficiency by 8%

PROJECT EXPERIENCE

Global Sales Performance Dashboard with Power BI — Jan 2026 – Present

- Built interactive dashboards in Power BI to visualize 2.75B revenue and 830 orders across 15+ countries and 50+ products
- Cleaned and merged data from 9 tables (orders, products, employees) into star schema through DAX and Power Query
- Identified top SKUs and peak sales months by product category, supporting pricing and inventory decisions across regions

LLM Agent Evaluation for Interactive Analytics — Nov 2023 – May 2024

- Built TAPILOT-CROSSING with 1024 GPT-4 samples to benchmark multi-turn data analysis agents across 4 task modes
- Designed AIR prompting method, improving GPT-4's accuracy on feedback tasks by 44.5% through logic extraction
- Developed semantic evaluation metrics (Acc, AccR, CSE) to assess LLM performance on custom libraries and tool usage

Time-Series Demand Forecasting for Perishable Goods — Aug 2023 – Sep 2023

- Cleaned and merged 50K+ rows of sales, pricing, and spoilage data; analyzed seasonal trends across 20+ vegetable SKUs
- Built SARIMA models (MAPE ≈ 6.7%) to forecast daily demand, reducing overstock and spoilage costs by around 15%
- Proposed data-driven pricing tiers and inventory plans under display constraints, improving revenue by 12% in simulations

CORE COMPETENCIES

Statistical Modeling: SARIMA, Regression, Segmentation (RFM, Clustering), Rule-based Scoring, Prompt Engineering
Programming & Tools: Python (pandas, Matplotlib), SQL (CTE), Power BI, Tableau, Excel, Hive, DAX, R (glm, ggplot2)
Data Analytics: EDA, A/B Testing, Funnel Analysis, Time-series Analysis, KPI Design, Dashboard Automation, ETL Pipeline