KEUN YOUNG YOON

Healthcare Data Analyst with 5+ years managing large-scale user dataset

New York, Open to Relocation | (934) 500-4949 | yoon.keunyoung.yky@gmail.com |
| youngyoon.me | linkedin.com/in/keunyoung-yoon | github.com/youngyoony

EDUCATION

State University of New York at Stony Brook *Expected 2026*

Master of Science in Data Science, GPA: 3.52

Relevant Coursework: Statistical Computing, Data Management, Data Analysis, Introduction to Probability

Korea National Open University, Seoul, Korea *2021*

Bachelor of Science in Data Science and Statistics, GPA: 3.64

Relevant Coursework: Data Mining, Unstructured Data Analytics, Data Processing and Applications, Multivariate Data Analysis, Forecasting Methods and Its Applications, Introduction to Statistical Deep Learning

Sogang University, Seoul, Korea *2015*

Bachelor of Business Administration and Political Science (Dual Majors)

SKILL

- **Programming Languages & Tools** – SQL, Python, R, Pandas, NumPy, spaCy, NLTK
- **Data Visualization** – Tableau, Power BI (DAX), Matplotlib, Seaborn
- **Big Data & Cloud Platforms** – Databricks, Apache Spark, Apache Airflow, Apache Kafka, Elasticsearch
- **Analytics & Statistics** – Statistical Inference, Regression Analysis, Time Series Analysis, Survival Analysis, Causal Inference, A/B Testing, Predictive Modeling, Model Evaluation
- **Advanced Analytics & ML Tools** – LangChain, RAG, Vector DB, Docker
- **Project Management & Collaboration** – Jira, Confluence, Microsoft Visio, Data Governance Compliance, VBA, Microsoft Office

EXPERIENCE

GSK – ViiV Healthcare *Durham, North Carolina (Hybrid)*

Data Strategy and Operations Intern (Full-time, Summer) & Co-op (Part-time, Academic Year) *Jun 2025 - Present*

- **Built and optimized patient-level RWD and commercial data pipelines using PySpark and SQL, integrating 10+ specialty pharmacy datasets across 130+ territories to support real-world evidence generation and treatment access analytics.**
- Applied standardized mapping logic to harmonize multi-source datasets, enabling interoperable analytics across prescriber networks and payer systems.
- **Developed and refined Power BI dashboards using DAX to model patient journey data, perform ad hoc treatment pattern analyses, and improve leadership visibility into commercial and healthcare delivery performance.**
- Queried and validated tokenized datasets on Databricks (Spark SQL) to perform ingestion diagnostics and generate 7+ ad hoc business reports supporting sales decision-making.
- Documented 3+ end-to-end data flow diagrams in Visio to visualize new vendor ingestion logic, aligning with monitoring and compliance standards.

Nexon Korea *Gyeonggi-do, South Korea*

Senior Data Analyst, ELSWORD *2022 - 2024*

- Integrated Kafka and Spark to develop an advanced data streaming and real-time analytics pipeline, enabling faster detection of operational issues and improving system reliability.
- Used LangChain and Chroma to analyze community data and generate item recommendations beyond in-game logs, enhancing recommendation relevance.
- Analyzed monthly user attendance distribution, identifying two distinct user groups (short-term logins within 7 days and continuous logins for 28+ days), leading to a project aimed at growing the mid-range user segment to 20%.

Senior Data Analyst, FIFA Online 4 *2017 - 2022*

- **Led the design and implementation of 20+ data dashboards, enhancing real-time analytics capabilities and enabling faster decision-making processes across the team.**

- Analyzed complex probabilistic product purchase behaviors, discovering user patterns that improved user retention and drove a 10% year-over-year revenue growth.
- Managed the microtransaction strategy and execution of 25+ series of paid in-game items, achieving the highest-ever revenue for the series by leveraging advanced data analysis and market trends.
- Executed 8 long-term user satisfaction surveys, transitioning from basic trend analysis to segmented feedback, which provided actionable insights and led to a 25% improvement in tailored user experiences.

Data Analyst, Need for Speed: Edge *2015 - 2017*

- Led the planning and execution of in-game monetization strategies and user retention events, contributing to a 25% increase in revenue and a 15% boost in user engagement.
- Planned and executed large-scale testing events, including three major Closed Beta Tests (CBTs) for a racing game with over 200,000 participants and one Focus Group Test (FGT) and Interview (FGI). Analyzed feedback from core users, improving user experience and engagement by evaluating both qualitative and quantitative data.

Data Analyst, FIFA Online 3 *2014 - 2015*

- Analyzed KPIs (including Unique Users (UU), Churn Rate, and In-Game Play Metrics) to enhance performance reviews, boosting user retention by 12% and average purchase per user by 8%.

RESEARCH | PROJECT

Fairness-Aware Mortality Prediction Using MIMIC-IV and Machine Learning at Emory University School of Nursing (PI: Dr. Hyunjung Gloria Kwak)

- Investigated intersectional fairness in mortality prediction using MIMIC-IV, benchmarking interpretable (LR, EBM), tree-based (XGBoost), and transformer models (TabPFN, FT-Transformer, TabNet) under various class-balancing and debiasing strategies.
- Processed demographic and clinical features (e.g., age, race, gender, LOS), and resolved conflicting mortality labels using a majority-rule strategy.
- Applied oversampling, undersampling, class weighting, and fairness-aware techniques (GroupDRO, ADV/ADB); evaluated model fairness using subgroup AUC, Brier Scores, and Intersectional Fairness Gaps.
- Visualized calibration disparities with Reliability Diagrams; developed a structured data dictionary linking 80 MIMIC-IV tables for reproducibility.
- Currently drafting a peer-reviewed manuscript presenting cross-dimensional fairness benchmarking findings.

NLP-Based Detection of Neurocognitive Markers in Dementia Speech at The Rutgers Institute for Health, Health Care Policy, and Aging Research (PI: Dr. Michelle Chen)

- Parsed and engineered linguistic features from annotation-rich DementiaBank transcripts using pylangacq, focusing on syntax completeness, grammatical errors, and echolalia patterns.
- Cleaned and standardized CHAT-format clinical speech data with regex and rule-based preprocessing, developing logic for subject–verb detection, sentence type classification, and repetition collapse.
- Combined spaCy-based dependency parsing with annotation-derived features for downstream analysis of neurocognitive markers.
- Developed reproducible, Python-based pipelines for linguistic feature extraction using pandas, and am currently drafting a lab report summarizing methodological results for a planned manuscript submission.

Real-time Document AI Chatbot Development Using Solar LLM, Kafka, and Retrieval-Augmented Generation (RAG)

- Implemented a system to preprocess, split, and store PDF documents, ensuring efficient retrieval using Solar embeddings.
- Created a chatbot that can retrieve relevant document fragments and generate answers using LLM, providing real-time feedback to user queries.
- Used Kafka to manage real-time message flow, improving the scalability and response time of the chatbot.
- Reduced redundant document processing through caching, saving time and costs, especially for large files.