KEUN YOUNG YOON

Experienced Healthcare Informatics Specialist with a 5+ year track record of handling large user datasets (2.5M+ MAU)

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: 3.60

Relevant Coursework: Big Data Algorithms & Networks, Big Data Analysis, Statistical Computing, Data Management, Data Analysis

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

Sogang University, Seoul, Korea *2015*

Bachelor of Business Administration and Political Science (Dual Majors)

SKILLS

- **Programming Languages & Tools** – SQL, Python, R, Pandas, NumPy, VBA, 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 Excel, PowerPoint, Microsoft Visio, Power Automate

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*

- **Engineered scalable, automated data profiling pipelines using PySpark and SQL, integrating 10+ specialty pharmacy datasets across 130+ territories to support commercial analytics and real-world data quality governance.**
- Implemented standardized mapping logic to harmonize multi-source datasets, enabling interoperable analytics across specialty pharmacy, prescriber networks, and payer systems.
- **Built Power BI dashboards using DAX to analyze patient journey KPIs, ensuring data accuracy and improving leadership visibility into commercial and healthcare delivery performance.**
- Delivered 7+ ad hoc business reports by querying and validating tokenized datasets on Databricks (Spark SQL), performing ingestion diagnostics to support sales decision-making.
- Created 3+ end-to-end data flow diagrams in Visio to visualize complex multi-layer data architecture (landing, enriched, semantic), aligning with monitoring and compliance standards.
- Automated recurring reporting workflows using Power Automate, cutting report turnaround time by 95% (from 2 days to 1 hour).

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 — Emory University School of Nursing (Jan 2025 - Present; PI: Dr. Hyunjung Gloria Kwak)

- Evaluated bias mitigation strategies in ICU mortality prediction using MIMIC-IV EHR data (156K+ patients), benchmarking interpretable (LR, EBM), tree-based (XGBoost), and transformer models (TabPFN, FT-Transformer, TabNet).
- Engineered 40+ clinical features (vitals, labs, Charlson Comorbidity Index) using BigQuery SQL, applying outlier filtering and missing data indicators.
- Applied oversampling, undersampling, class weighting, and fairness-aware techniques (GroupDRO, ADV/ADB); evaluated model fairness using subgroup AUC, Brier Scores, and Fairness Gaps.
- Preparing first-author manuscript for peer-reviewed publication (target: Spring 2026).

NLP-Based Detection of Neurocognitive Markers in Dementia Speech — Rutgers Institute for Health, Health Care Policy, and Aging Research (May 2025 - Present; PI: Dr. Michelle Chen)

- Engineered linguistic features from DementiaBank's clinical speech transcripts using NLP libraries (spaCy, pylangacq), focusing on syntax completeness, grammatical errors, and disfluency patterns.
- Cleaned and standardized CHAT-format transcripts with regex and rule-based preprocessing for structured clinical speech analysis.
- Achieved 99.68% accuracy against manual validation for automated neurocognitive marker detection by combining dependency parsing with annotation-derived features.
- Built reproducible Python pipelines processing 236 files (~3,000 utterances) with pandas, generating structured outputs for downstream clinical analysis and manuscript preparation.

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.