KEUN YOUNG YOON

|New York, Willing to Relocate| (934) 500-4949 | yoon.keunyoung.yky@gmail.com |
|youngyoon.me|www.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
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

- **Analytics and Statistics** – Data Wrangling, ETL Processes, Regression Analysis, Data Mining, Multivariate Data Analysis, Probability and Statistical Inference, Statistical Computing, Fairness Evaluation, Forecasting Methods, Unstructured Data Analytics, Data Visualization
- **Big Data and ML/DevOps** – Databricks, Apache Kafka, Apache Spark, Apache Airflow, Elasticsearch, RAG, LLM, LangChain, Vector DB, Docker, XGBoost, MLP, Class Balancing
- **Data Visualization** – Microsoft Visio, Tableau, Power BI, Matplotlib, Seaborn
- **Project Management** – Jira, Confluence, Data Governance Compliance
- **Programming Language and other** – Python, R, SQL, Pandas, PyCaret, Regex, NLTK, Django, FastAPI, VBA, MS Office

EXPERIENCE

GSK *Durham, North Carolina*
Data Strategy and Operations Intern, ViiV Healthcare *Jun 2025 – Present*

- Standardized and integrated structured datasets from various vendors using domain-specific dictionaries and mapping logic to enable cross-vendor data integration.
- Engineered scalable data pipelines and documented end-to-end flow diagrams in Microsoft Visio, tightly aligning multi-vendor ingestion logic with operational monitoring and compliance requirements.
- Queried and validated tokenized datasets on Databricks (Spark SQL) to perform ingestion diagnostics, resolve schema drift, and optimize transformation logic across staging and production environments.
- Designed and deployed Power BI dashboards to monitor data latency, ingestion completeness, and key operational metrics across functional business domains, improving visibility for technical and business stakeholders.
- Led cross-functional onboarding efforts and contributed to data governance by aligning metadata definitions, schema mapping standards, and documentation practices using JIRA.

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 efficient collection and preprocessing of in-game activity data and user behavior metrics. This integration improved data processing efficiency and reduced latency, allowing for immediate data-driven decision-making and enhancing the overall player experience.
- Implemented a system that crawled external user community forums to analyze recommended specs for specific dungeons using Chroma and LangChain. This enabled us to provide item recommendation lists that matched user opinions, moving beyond simple log data to enhance the relevance of in-game recommendations.
- 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*

- Analyzed complex probabilistic product purchase behaviors, discovering user patterns that improved user retention and drove a 10% year-over-year revenue growth.

- Led the design and implementation of 20+ data dashboards, enhancing real-time analytics capabilities and enabling faster decision-making processes across the team.
- 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.
- Integrated various technical solutions such as re-integrating data protocols from TCP to UDP, developing an anomaly detection system, and creating a real-time notification service for probabilistic item data, resulting in a 20% reduction in system response times.

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, focusing on class balancing trade-offs in MLP and XGBoost models.
- Processed demographic and clinical features (e.g., age, race, gender, LOS), and resolved conflicting mortality labels using a majority-rule strategy.
- Applied oversampling, undersampling, and class weighting techniques; evaluated subgroup AUC and Brier Scores, and calculated Intersectional Fairness Gaps.
- Visualized calibration disparities with Reliability Diagrams; developed a structured data dictionary linking 10+ MIMIC-IV tables for reproducibility.
- Extended analysis with SHAP-based feature comparisons (with/without balancing), and tested GroupDRO and adversarial debiasing to mitigate subgroup-level model bias.

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-derived transcriptions in the DementiaBank Pitt dataset, focusing on syntax completeness, grammatical errors, and echolalia patterns.
- Cleaned and preprocessed clinical speech data using Python, developing logic to detect sentence structure violations (e.g., missing subjects or verbs) and repeated lexical units.
- Combined annotation-based features with token- and rule-based NLP features for downstream analysis, contributing to a pipeline aimed at detecting neurocognitive biomarkers.
- Authored modular Python notebooks for feature extraction and reproducibility, enabling integration of structured linguistic metrics into cognitive modeling pipelines.

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.