STAT GR5244 Final Project Proposal

Chuyang Su [cs4570], Kangyu Zhao [kz2537]
Columbia University
STAT GR5244 – Unsupervised Learning (Fall 2025)
Instructor: Prof. Genevera Allen

Abstract—We are currently uncertain which research direction best fits both the course scope and our available computing resources, so we have prepared three potential projects based on different Kaggle datasets. We hope to receive feedback on which option would be most appropriate to pursue. Our project materials and development progress are maintained in a GitHub repository at https://github.com/Schuyn/Unsupervised-Machine-Learning-Final-Project.git.

I. NBA Draft[link]

The NBA Draft dataset contains comprehensive career and performance statistics for all NBA rookies from 1989 to 2021, including demographic, draft, team, and box-score information.

Our goal is to analyze each player's first five seasons—combining box-score performance with draft position, college background, and team—to characterize a player's potential and career fulfillment path. We expect to obtain well-shaped clusters such as "rising superstars," "early peakers," and "under-the-radar contributors." Because players drafted after 2017 have not yet completed five years, we will train on 1989–2017 data and use later rookies to examine generalization and stability. We also plan to present representative templates for each cluster.

The novelty lies in three aspects: (1) the dataset itself is recent and rarely analyzed beyond simple EDA on Kaggle; (2) we will explore multiple clustering models and select the best one through both quantitative validation and subjective interpretability; (3) our conceptual focus differs from conventional basketball analytics—we seek to understand a player's developmental path and self-realization rather than isolated box-score efficiency. The dataset is small (212 KB) and thus highly feasible, yet large enough to yield meaningful structure.

II. NFL Big Data Bowl 2026[link]

The NFL Big Data Bowl 2026 dataset includes extensive play-by-play statistics, player-level tracking data, and even video-frame information. We do not intend to follow the original competition objective but instead to use the data to analyze team and player tendencies.

Two complementary perspectives are possible: from the team level, we could examine formations and play distributions to identify strategic styles; from the player level, we could analyze, for example, how different wide receivers run routes or how quarterbacks such as Patrick Mahomes tend to deliver passes. The dataset itself is novel—it was released in preparation for Super Bowl 2026—and almost no existing work has attempted unsupervised clustering on it. We expect to identify representative teams or players whose playing styles reveal interpretable clusters that might help fans and professionals alike understand football tactics from a new viewpoint.

However, the dataset's massive scale makes it difficult to control both computation and interpretability. If successful, it could yield the most insightful results, but we are uncertain whether team-level or player-level analysis is more feasible within one semester and would value your guidance.

III. Make Data Count[link]

The "Make Data Count" dataset contains basic metadata, citation-type labels, and full PDF/XML files of academic papers. We do not plan to use the "primary/secondary" labels directly; instead, we aim to analyze the content and citation relationships of papers to discover, in an unsupervised manner, the latent patterns and trends of academic research. Specifically, we will extract paper nodes, dataset nodes, and citation-context nodes, normalize DOIs, titles, and dataset names, and construct a "paper-citation-dataset" tripartite graph. Node features will be generated using high-dimensional embeddings from BERT, and edge weights defined by semantic similarity between text vectors. We plan to apply graph-based unsupervised methods, such as spectral clustering or GNN community detection, to identify latent topics and citation behaviors.

The novelty lies in exploring the potential of unsupervised learning on this new type of academic dataset while combining textual semantics and citation relationships to build a richer scholarly knowledge graph. Such a graph could reveal both explicit citation structures and hidden connections in data reuse. We acknowledge our limited experience with knowledge-graph construction and the possible computational challenges of large-scale NLP embedding, which may affect feasibility.

IV. Request for Feedback

Each project satisfies the learning objectives of STAT GR5244 but differs in complexity and interpretability. We kindly request your feedback on which dataset and framing offer the best balance between novelty, feasibility, and analytical depth for a one-semester project.