Ciara Fasullo, Tiffanie Kwakye, Lesly Veizaga

DS 2002

Final Project Reflection

5 December 2024

The journey of our group project, where we decided to analyze NBA Draft data and NBA player statistics from the 1989-2021 seasons, was both a challenging and rewarding experience. Working in a group of three, we wanted to draw meaningful insights into the relationship between collegiate basketball programs, draft positions, and subsequent NBA performance. This is something we became especially interested in exploring after witnessing the journey of Reece Beekman through UVA and his exceptional NBA career thus far. This reflection will cover the challenges encountered, lessons learned, and skills gained throughout the process, along with areas for future improvement.

The initial phase of identifying relevant datasets was more complex than we anticipated. While platforms like Kaggle have an abundance of data, ensuring both relevance and compatibility between two datasets was an unexpected hurdle. One challenge was determining how to align two datasets with different time ranges—draft data beginning in 1989 and player stats extending back to 1950. This discrepancy required filtering of the player stats data to focus only on individuals drafted within the relevant time frame. Additionally, understanding the structure of the datasets and identifying the key variables that could drive meaningful analysis was one of the more time-consuming aspects of the project.

Designing the ETL pipeline was another challenging aspect. Defining the extraction, transformation, and loading steps required us to think critically about how to clean and integrate data from two different sources. For example, inconsistencies in column names, missing values, and format mismatches required a significant amount of cleaning. Implementing the ETL

pipeline in Python and integrating MongoDB for data storage tested not our coding skills, but our ability to navigate the challenges of collaborating on code. As we were working in a group, cloud storage integration with Google Cloud added another layer of complexity, as we had to familiarize ourselves with credential management and access control across multiple devices.

The analysis phase posed its own challenges, primarily in ensuring that our visualizations effectively communicated the insights we uncovered. Balancing depth and clarity in our statistical analysis was difficult, especially when summarizing trends across large datasets. Additionally, identifying meaningful correlations required experimentation with various metrics and visualization techniques. As a group, we discussed the best ways to interpret the results, particularly when the data didn't immediately align with our hypotheses.

Further, working in a group brought its own set of challenges. Scheduling meetings around our individual commitments required flexibility, and dividing tasks equitably was not always straightforward. Miscommunication occasionally led to duplication of effort, such as when multiple members worked on similar aspects of the ETL script. Coordinating the presentation component added another layer of collaboration, as we needed to ensure that our insights were conveyed clearly and cohesively.

Although we encountered many challenges, this project reinforced the importance of data cleaning and preprocessing - something we have learned throughout this course and other data projects. Even seemingly small inconsistencies, such as minor name variations, can significantly disrupt analysis. We also learned the value of modular ETL design. By breaking the process into smaller, reusable functions, we improved both the pipeline's clarity and its maintainability. Additionally, integrating cloud storage highlighted the importance of securing credentials and managing access permissions effectively.

Clear communication and task delegation were also key lessons in this project. Early in the process, we recognized the need for check-ins with one another to ensure alignment on

progress and objectives. Using collaborative tools like shared documents and version control systems helped streamline our workflow and reduce redundancies.

In future projects, we would prioritize setting up a detailed project plan with defined milestones and responsibilities for each team member. This would reduce overlap and ensure that all aspects of the project receive adequate attention in the most efficient way possible.

We have definitely refined a variety of skills throughout this project. Notably, the project enhanced our proficiency in Python, particularly in using various libraries for data manipulation and database integration. Working with MongoDB strengthened our understanding of database management and query optimization. Additionally, we gained hands-on experience with Google Cloud storage, including setting up buckets and managing access permissions.

We also improved our ability to extract insights from large datasets, leveraging various data analysis techniques and statistical methods. Experimenting with a number of different visualizations using libraries like Matplotlib and Seaborn also refined our ability to communicate findings clearly and effectively.

Working as a team taught us the importance of active listening, constructive feedback, and compromise. Through this experience, we learned how to leverage each team member's strengths, resulting in a more comprehensive and well-rounded project.

While the project was successful, there are certainly areas we could develop further. Exploring more advanced analytical methods, such as machine learning models to predict player performance based on draft data, could provide deeper insights. Additionally, gaining proficiency in more sophisticated data storage solutions, such as cloud-based databases, would enhance the scalability of future projects. Finally, refining our time management and task allocation strategies would ensure smoother project execution.

Conclusively, this project was a valuable learning experience that tested our technical, analytical, and collaborative abilities. Despite the challenges, our group successfully explored the relationship between collegiate basketball programs, draft positions, and NBA performance, presenting insights that were both statistically and visually compelling. The skills we developed, including data cleaning, ETL implementation, and effective communication, will undoubtedly serve us well in future endeavors. With the lessons learned and areas for improvement identified, we feel better equipped to tackle more complex data analysis projects moving forward.