Anna Brown, Annie Edmonds, John Jordan, James Kulp, and Emory Wise

Professor Jason Williamson

DS 2002

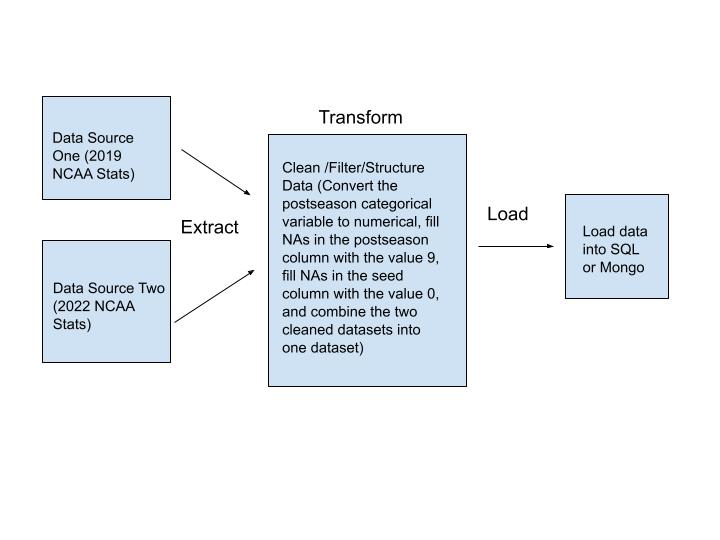
December 6, 2024

Final Project Reflection

For our final project, we authored a segment of an ETL pipeline that could ingest two Kaggle data sources — a CSV containing 2019 college basketball data and a CSV containing 2022 college basketball data. After fetching these remote data files from Github, we converted them into a usable format, modified the data frames to handle missing values and mapped the categorical values to numerical values, and merged the data frames into a combined 2019 and 2022 college basketball dataset to visualize cross-time trends in the data. Additionally, we developed functionality to load all three data sets (cbb\_19, cbb\_22, and cbb\_19\_and\_22) into either a Mongo or SQL storage destination for export as selected by the user. Finally, we stored the transformed data in Google Cloud and documented the setup process. While we did encounter some challenges — particularly with constructing our pipeline and configuring cloud storage — we successfully worked together as a team to design and implement our project. This effort not only provided valuable technical insights but also enhanced our ability to work effectively under time constraints.

Our objective was to analyze the effect of new transfer portal rules on regular and postseason success for each conference in the NCAA using statistics from before and after the changed rules. The first step in the ETL pipeline was to extract the two datasets: 2019 NCAA basketball statistics and 2022 NCAA basketball statistics. We then transformed this data by creating a dictionary to map the categorical values to new numerical values, developing a function to fill the NAs of a column with new values, and then combining the two cleaned datasets. Finally, we loaded the transformed data into either SQL or Mongo, depending on the user’s desired format and source. To uncover key insights about our data, we created visualizations depicting bar plots of total and average wins by conference for each team and scatterplots for the adjusted tempo of each team against their postseason success. Through these visualizations, we discovered discrepancies concerning parity among conferences and the success of Power 5 conferences in the postseason. Overall, we found that the transfer portal had little impact in its early years; however, we anticipate that future years will demonstrate a more significant effect of the portal as teams and players better understand how to use it.

Our last step was uploading our data to Google Cloud Storage by creating and configuring buckets. We named our bucket “college-basketball-insights,” assigned the region to the US, chose a standard storage class for frequently accessed data, and updated the access control to fine-grained for better security. The file structure, ‘transformed\_data/cbb\_19\_and\_22.csv’, processed and combined the 2019 and 2022 NCAA basketball statistics. Lastly, we addressed security considerations by creating a service account, developing a key within that account, and then editing the permissions to add team members as owners.



Our most significant challenge faced during data selection was setting up the cloud storage and the “export” section of our ETL pipeline. Specifically, developing a pipeline that could export cleaned data to *either* Mongo or SQL based on user preference was tricky. However, we were able to adapt the code from the Data Project and modify it to integrate flexible storage options to complete this task. Additionally, while most group members have experience in Python and ETL pipeline construction, none of us were proficient in Google Cloud services. Learning how to construct and manage Google Cloud storage buckets was an entirely new skill for us. The initial setup process was especially daunting, as we had to navigate unfamiliar documentation and troubleshoot connectivity issues between the local environment and the cloud. However, we ultimately succeeded in configuring the cloud storage, thanks to trial and error, as well as collaborative problem-solving.

Once the data was loaded and we began doing some basic EDA and creating visualizations to derive insights, we noticed some limitations, namely a lack of variation between the two years of data that hindered our ability to draw more meaningful conclusions about the effect the transfer portal might have had on team performance. With more data spanning multiple years, especially seasons in the future, any changes in team performance might be more pronounced as schools learn to take advantage of the transfer portal better, and its effects may become more pronounced. It is possible that the new rules surrounding the transfer portal may not have much impact on college basketball, and team success is more closely correlated to more permanent factors such as the team culture, historical success, coaching choices, or other changing factors in college basketball of which there are many.

Despite these technical challenges, we completed our project by leveraging strong teamwork and effective communication. While our team did struggle with coordinating over Thanksgiving Break, as team members due to differing priorities and schedules, we offset the impact of this challenge by conducting a thorough planning session before the break to identify each person’s strengths and weaknesses. Based on this assessment, we delegated tasks according to individual confidence levels and areas of expertise. For example, team members with stronger Python skills focused on coding the ETL pipeline, while others with analytical-oriented skills focused on developing data visualizations and drafting conclusions. This approach minimized delays and ensured that everyone could add meaningful contributions to the project. While there was some miscommunication early on, regular updates via text and shared Google files helped us stay aligned and resolve issues quickly.

In conclusion, this project allowed us to refine several technical and team-based skills. Technically, we deepened our understanding of Python-based ETL processes, including data transformation techniques and integrating multiple storage solutions. Working with Google Cloud significantly expanded our cloud computing knowledge, particularly in managing buckets and permissions. Additionally, adapting previous code for new uses honed our ability to reuse and modify existing solutions. As a team, we improved in task delegation, time management, and collaboration to meet deadlines under tricky circumstances. For future projects, we aim to develop our expertise further in cloud programs and other storage technologies to optimize performance.