### **Data Processing**

1. **Selection of Specific College Data**: Given the large volume of original data and the project’s focus on engineering, we filtered the dataset to include only subjects offered by the Grainger College of Engineering. These subjects include: ABE, AE, BIOE, CEE, CS, CSE, ECE, ENG, IE, ME, MSE, NPRE, PHYS, SE, TAM, and TE.
2. **Exclusion of Subjects with Limited Data**: To enhance the accuracy of predictions and prevent instability caused by low sample sizes, subjects with fewer than 5 records were removed. This ensured that every subject included in the training and prediction process had sufficient data to support reliable modeling, thereby mitigating the effects of sparse data.

### **Discussion of Data Quality Issues**

Throughout the data processing, we encountered and addressed the following data quality issues:

#### **1. Missing Data**

* **Problem**: Certain records may contain missing or null values, particularly for key predictive variables (e.g., GPA, grades) or important features (e.g., year, subject).
* **Solution**: For continuous variables (such as GPA), missing values can be filled using the median or mean to maintain consistency in data distribution. For categorical variables (like Subject), mode imputation is a viable option. When missing values are excessive in specific records, it may be more appropriate to remove those entries altogether to prevent potential biases.

#### **2. Imbalanced Data**

* **Problem**: Due to significant variations in student enrollment and course offerings across different subjects within the College of Engineering, some subjects may have substantially more data than others. For example, “CS” or “ME” may have a far higher number of records than “NPRE” or “BIOE.”
* **Impact**: Imbalanced data can lead the model to favor subjects with more data, potentially resulting in less accurate predictions for underrepresented subjects.
* **Solution**: Resampling techniques can help address this imbalance. Under-sampling larger classes or over-sampling smaller classes are two common methods. Alternatively, synthetic data generation methods, like SMOTE (Synthetic Minority Over-sampling Technique), can create virtual samples for minority classes, helping to balance the data distribution across subjects.

#### **3. Sparse Data**

* **Problem**: Sparse data refers to a lack of sufficient data for certain subjects across specific features (such as years), leading to incomplete time-series information for some subjects. For example, some subjects may only have data for a few specific years, making it difficult to identify a consistent trend over time.
* **Impact**: Sparse data makes it challenging for the model to learn changes in trends, which can negatively impact predictive performance.
* **Solution**: Aggregating data over broader time periods can alleviate sparsity. For example, combining adjacent years or merging certain attributes based on similarities across subjects can help fill gaps. Dimensionality reduction techniques may also be applied to focus the model on primary trends, reducing the impact of sparse features.