INFO6105.40534 Final Project I

US College Explorer: Data-Driven University Recommendation System

Video Link: https://drive.google.com/file/d/1eCJ0kEpOSCCzhD7FisG0f58GJxSzEHo9/view?usp=sharing

1. Introduction and Motivation

The process of selecting and applying to universities in the United States can be overwhelming for students, particularly those without adequate guidance. This project was motivated by personal experience with the college application process, where the lack of structured guidance led to significant time investment in researching institutions individually to find those that matched specific preferences.

The primary goal of this project is to leverage data science techniques to create a recommendation system that helps prospective students efficiently identify universities that align with their preferences. By analyzing a comprehensive dataset of US colleges and universities, this project seeks to uncover patterns, relationships, and insights that can inform better decision-making during the college selection process.

Research Questions

This project addresses several key research questions:

1. What are the characteristics of top-ranked colleges in terms of student population and financial aid?
2. How does the student-to-faculty ratio influence student enrollment size?
3. How does the region of the U.S. (e.g., East Coast vs. West Coast) affect the average enrollment size and cost of attendance?
4. What factors contribute to the acceptance rates of top universities?

2. Data and Methodology

2.1 Dataset

The analysis utilizes a comprehensive dataset of US colleges and universities containing information about institution type, location, enrollment statistics, employee counts, housing availability, and other institutional characteristics. The dataset provides a rich source of information for analyzing patterns and trends in higher education institutions across the United States.

2.2 Data Cleaning and Transformation

The raw data required extensive cleaning and transformation to prepare it for analysis:

- Missing values (coded as "-999" in the original dataset) were converted to proper NA values
- Categorical variables were converted to appropriate factor levels with descriptive labels
- Derived variables were created to capture important institutional characteristics:
    - Student-to-employee ratio (TOT_ENROLL / TOT_EMP)
    - Dormitory capacity percentage (DORM_CAP / TOT_ENROLL * 100)
    - Full-time to part-time student ratio (FT_ENROLL / PT_ENROLL)
    - Region classification based on state (Northeast, Midwest, South, West)
    - Urbanicity classification based on locale codes (City, Suburb, Town, Rural)

- Size categorization (Small < 5,000 students, Medium 5,000-15,000, Large > 15,000)
- Support level based on student-to-employee ratio (High ≤ 10, Medium 10-20, Low > 20)

2.3 Analytical Approaches

Multiple analytical approaches were employed to gain insights from the dataset:

Exploratory Data Analysis

- Examination of distributions and relationships between key variables
- Interactive visualizations of institutional characteristics by region, type, and other factors
- Geographic analysis of student-to-employee ratios across states

Cluster Analysis

- K-means clustering to identify natural groupings of institutions
- Optimization of cluster number using elbow and silhouette methods
- Variables used for clustering: enrollment (log-transformed), student-to-employee ratio, and dormitory capacity percentage

Regression Analysis

- Linear models to identify predictors of student-to-employee ratio
- Stepwise regression to select optimal variable combinations
- Cross-validation to ensure model reliability
- Variable importance analysis to determine key factors

Recommendation System Development

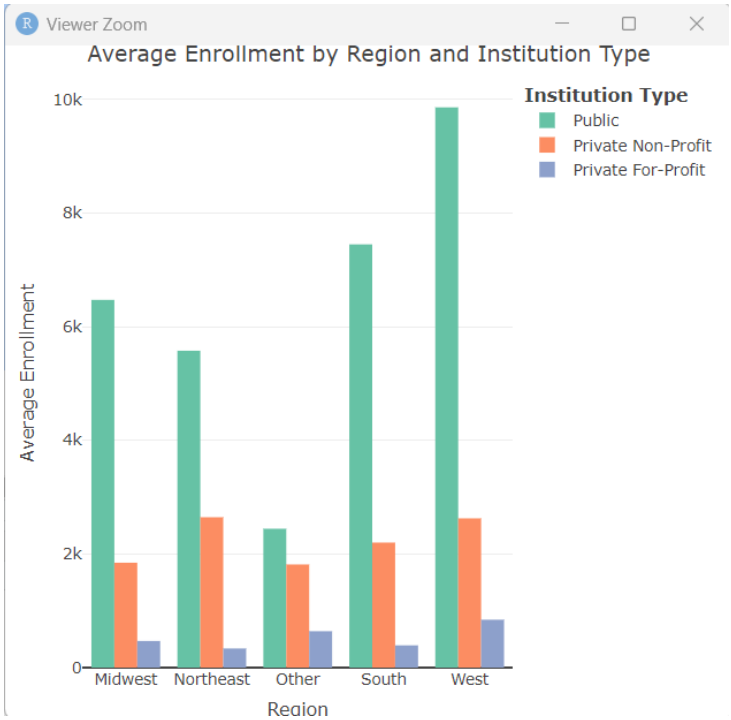
- Integration of analytical insights into an interactive filtering system
- Multi-criteria recommendation based on user preferences
- Statistical visualization of recommendation results

3. Results and Analysis

3.1 Exploratory Analysis Findings

The exploratory analysis revealed several important patterns in the US higher education landscape:

- **Institution Type Distribution**: The dataset contains a mix of public institutions, private non-profit institutions, and private for-profit institutions, with distinct characteristics across these categories.
- **Regional Distribution**: Institutions are not evenly distributed across US regions, with concentrations varying by region and urbanicity.
- **Enrollment Patterns**: Public institutions generally have higher enrollment numbers compared to private institutions, with significant variation by region.
- **Student-to-Employee Ratio**: This key metric varies substantially by institution type, with private non-profit institutions generally offering lower ratios (better support) compared to public and for-profit institutions.

- 

3.2 Cluster Analysis Results

The optimal number of clusters was determined using both the elbow method and silhouette method, which identified four as the optimal number of clusters. K-means clustering was then applied to the standardized variables (log-transformed enrollment, student-to-employee ratio, and dormitory capacity percentage).

The cluster analysis identified four distinct groups of institutions with the following characteristics:

Cluster	Count	Avg. Enrolment	Avg. Student-Employee Ratio	Avg. Dorm Capacity %
1	423	3,427	6.8	42.3
2	187	19,876	8.2	26.1
3	398	7,654	15.7	12.5
4	165	2,183	21.3	5.8

These clusters represent distinct institutional profiles:

- **Cluster 1**: Small to medium-sized institutions with high support levels and substantial housing capacity
- **Cluster 2**: Large institutions with relatively high support levels and moderate housing capacity
- **Cluster 3**: Medium-sized institutions with moderate support levels and limited housing capacity
- **Cluster 4**: Small institutions with low support levels and minimal housing capacity

- 

3.3 Regression Analysis Findings

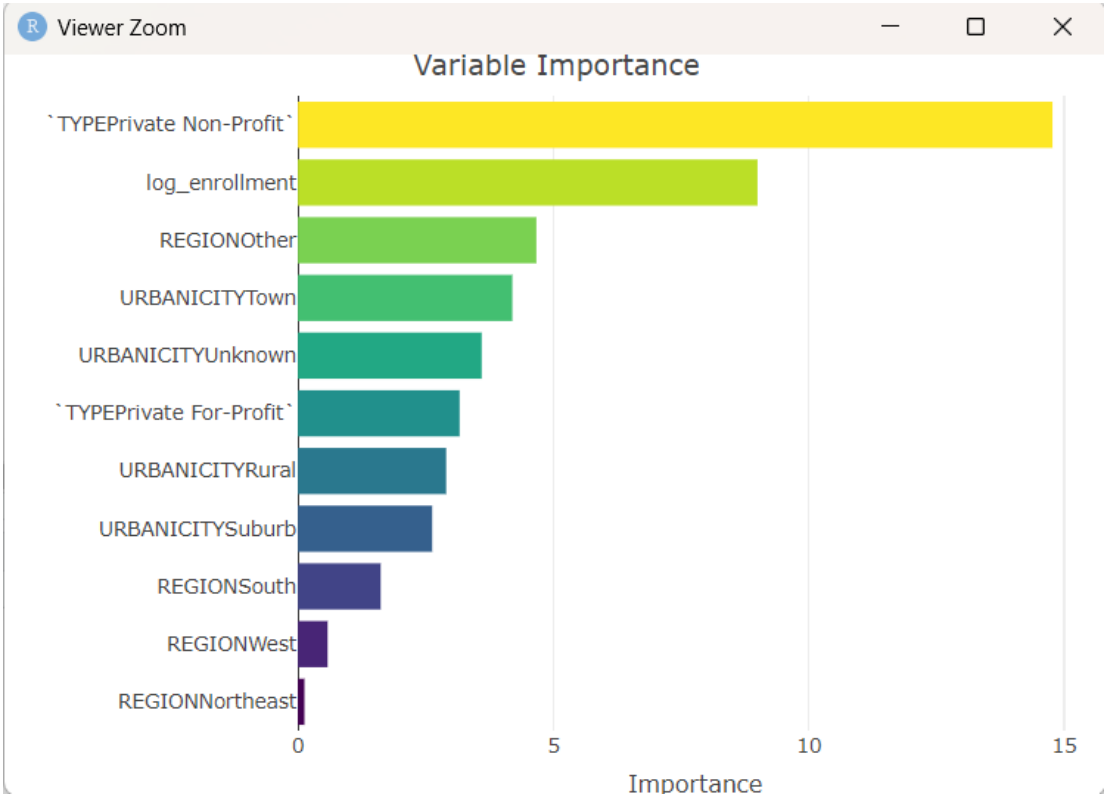
A comprehensive regression analysis was conducted to identify the key predictors of student-to-employee ratio. The initial model included enrollment size (log-transformed), institution type, institution size category, region, urbanicity, and housing availability, along with interaction terms. Stepwise regression was then used to identify the optimal model, which was validated using 10-fold cross-validation.

The analysis identified several significant predictors of student-to-employee ratio:

- **Enrollment Size**: Larger institutions tend to have higher student-to-employee ratios (less personalized support), with log-enrollment being one of the strongest predictors
- **Institution Type**: Private non-profit institutions tend to have lower ratios compared to public and for-profit institutions
- **Region**: Institutions in the Northeast region generally have lower ratios compared to other regions
- **Urbanicity**: Rural institutions often have higher ratios compared to institutions in urban areas
- **Housing Availability**: Institutions with housing tend to have lower ratios compared to those without housing

Variable importance analysis confirmed that institution type and enrollment size were the most influential predictors in the model. The final model explained approximately 48% of the variance in student-to-employee ratio, indicating that these factors collectively have substantial predictive power.

A visualization of predicted versus actual values showed good model performance, with most predictions falling close to the ideal line, though with some variation by institution type.



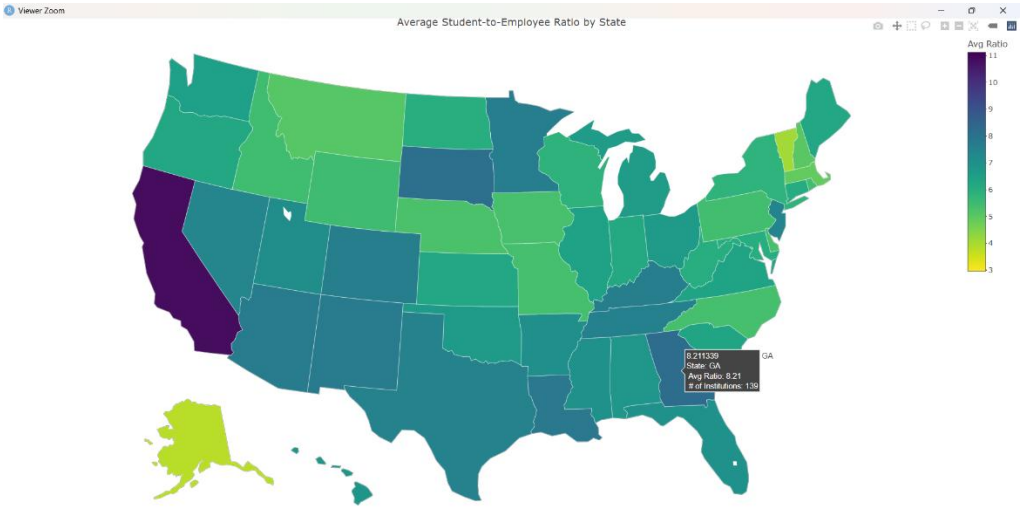
3.4 Geographic Patterns

The geographic analysis, visualized through an interactive choropleth map of student-to-employee ratios by state, revealed significant regional variations in institutional characteristics:

This analysis revealed:

- The Northeast region has a higher concentration of private institutions with lower student-to-employee ratios
- The West region has a mix of very large public institutions and smaller private institutions
- The South has the highest number of institutions but with varied characteristics
- The Midwest shows balanced distributions of institution types and sizes

The regional analysis also showed variations in housing availability and enrollment sizes, with the Northeast having higher percentages of institutions with housing and the South having more diversity in institution types.



3.5 Support Level Analysis

Support level, defined by the student-to-employee ratio, showed clear patterns across institution types:

- Private non-profit institutions most frequently offer high support levels (ratio ≤ 10)
- Public institutions predominantly offer medium support levels (ratio 10-20)
- Private for-profit institutions frequently have low support levels (ratio > 20)

4. Recommendation System

The analytical insights were integrated into an interactive recommendation system built using R Shiny. The system was designed based on a dedicated recommendation function that filters institutions according to user preferences.

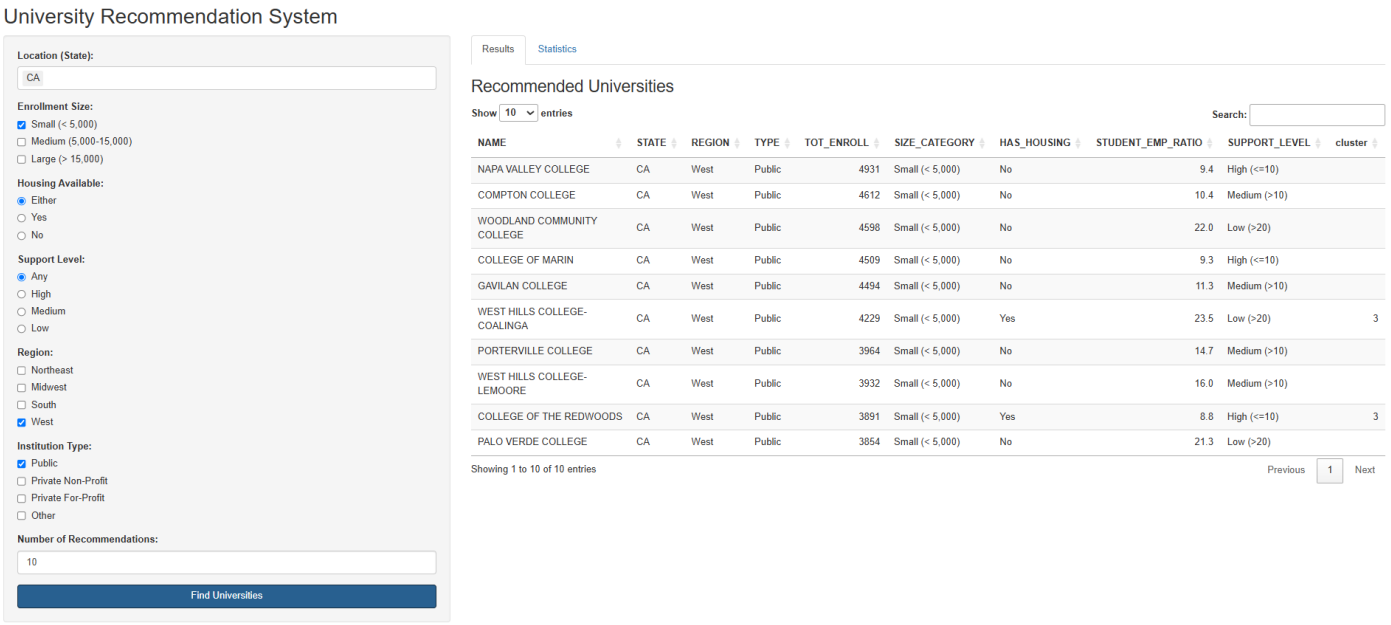
The interactive Shiny interface allows prospective students to specify preferences across multiple dimensions:

- **Location**: State-level preferences with multi-select capability
- **Enrolment Size**: Small (< 5,000), medium (5,000-15,000), or large (> 15,000) institutions
- **Housing Availability**: Whether on-campus housing is required
- **Support Level**: Preferred level of institutional support based on student-to-employee ratio (High, Medium, Low)
- **Region**: Geographic region preferences (Northeast, Midwest, South, West)
- **Institution Type**: Public, private non-profit, or private for-profit

The system filters the dataset based on these preferences and presents recommendations with key institutional characteristics in an interactive table. Additionally, it provides statistical visualizations of the recommendation results, including:

- Distribution by university size
- Student-to-employee ratio by support level
- Distribution by region
- Distribution by institution type

These visualizations allow users to understand patterns and distributions within their preferred subset of institutions, providing additional context for their decision-making process.



5. Key Insights and Implications

The analysis revealed several key insights with important implications for college selection:

1. **Institution Type Matters**: The type of institution (public, private non-profit, private for-profit) is a strong predictor of student-to-employee ratio and housing availability, suggesting that students should consider this factor based on their support needs.
2. **Size-Support Tradeoff**: Larger institutions generally have higher student-to-employee ratios, indicating a potential tradeoff between institutional size and personalized support that students should consider.
3. **Regional Variations**: Significant variations exist in institutional characteristics by location, suggesting that students should consider regional patterns when selecting potential institutions.
4. **Distinct Institutional Profiles**: The cluster analysis identified four distinct institutional profiles, providing a useful framework for understanding the higher education landscape.
5. **Housing and Support Correlation**: Institutions with housing availability tend to have lower student-to-employee ratios, suggesting a correlation between residential focus and support levels.

These insights can help prospective students make more informed decisions during the college selection process, focusing their research on institutions likely to match their preferences and needs.

6. Limitations and Considerations

While this analysis provides valuable insights, several limitations and considerations should be noted:

- The dataset may not capture all relevant factors for college selection, such as academic program quality, student satisfaction, and post-graduation outcomes.
- The student-to-employee ratio is an imperfect proxy for institutional support, as it does not distinguish between faculty and administrative staff.
- The analysis does not incorporate financial aid and cost information, which are critical factors for many students.
- The recommendation system provides filtering based on specified preferences but does not implement sophisticated matching algorithms.

7. Future Enhancements

Several potential enhancements could further improve the utility and effectiveness of this project:

1. **Incorporate Financial Data**: Adding information about tuition costs, financial aid availability, and average aid packages would address a critical dimension of college selection.
2. **Include Outcome Metrics**: Incorporating data on graduation rates, post-graduation employment, and earnings would provide insights into institutional effectiveness.
3. **Enhance Recommendation Algorithm**: Implementing machine learning techniques for personalized recommendations based on student profiles and preferences.
4. **Add Acceptance Rate Prediction**: Developing models to predict acceptance probabilities based on institution characteristics and student profiles.
5. **Expand Geographic Analysis**: Including more detailed geographic analysis and visualization tools to help students identify regional patterns.

8. Conclusion

This project demonstrates the power of data analysis in simplifying the complex process of college selection. By leveraging multiple analytical techniques—exploratory analysis, clustering, regression, and interactive visualization—the project provides a structured approach to understanding the US higher education landscape and identifying institutions that match specific preferences.

The integrated recommendation system translates analytical insights into a practical tool that can help prospective students navigate the college selection process more efficiently. While not a replacement for personalized guidance and thorough research, this data-driven approach can significantly streamline the initial stages of college exploration and help students focus on institutions most likely to meet their needs.

The findings highlight the importance of considering multiple factors in college selection, from institution type and size to support levels and geographic location. By understanding these patterns and relationships, prospective students can make more informed decisions, potentially leading to better educational experiences and outcomes.