Exploratory Data Analysis Report

Submitted by: Pragya Gupta
PGDM – Research & Business Analytics

Abstract

This project explores **Forbes America's Top Colleges 2019 dataset**, which contains rankings and key statistics of 650 U.S. colleges. The goal of the analysis is to perform exploratory data analysis (EDA) to understand trends, identify key insights about colleges across states, and highlight factors such as acceptance rates, financial aid, alumni salaries, and undergraduate population.

1. Introduction

Choosing a college is one of the most critical and expensive decisions for students and families. Forbes Magazine annually publishes a list of the top U.S. colleges based on academic quality, student experiences, career outcomes, and affordability.

This project applies EDA techniques to the Forbes dataset to:

- Understand the structure of the dataset.
- Clean and preprocess the data.
- Generate insights about the distribution of colleges, their financial aid, acceptance rates, and alumni outcomes.
- Highlight patterns useful for decision-making.

2. Dataset Description

- **Source:** Forbes America’s Top Colleges 2019
- **Size:** 650 rows × multiple columns
- **Key Features:**
    - College Name, City, State
    - Type: Public or Private

- Acceptance Rate, Student Population, Undergraduate Population
- Average Grant Aid, Alumni Salary
- Forbes Rank

Missing values were treated by dropping unnecessary or null-heavy columns (e.g., SAT/ACT ranges, Website) and handling NA rows. New fields such as **Percentage of Undergraduates** and **Alumni Salary Levels** were created for deeper analysis.

3. Methodology

The project followed these steps:

- **Data Loading & Inspection:** Read dataset, checked dimensions, datatypes, missing values.
- **Data Cleaning:**
  - Dropped unnecessary fields (SAT/ACT ranges, Website).
  - Removed rows with missing values.
  - Reset dataframe index.
- **Exploratory Analysis:**
  - Identified top-ranked colleges, state-level distributions, and counts of Public vs. Private.
  - Derived new variables (e.g., % of undergraduates, Alumni Salary categories).
  - Grouped and aggregated data (mean, counts, sorting).
- **Statistical Summaries & Insights:** Generated descriptive statistics and visual insights (head, tail, describe, groupby).

4. Results & Insights

1. **Top 5 Colleges (2019):** Princeton, Harvard, Columbia, MIT, and Yale.
2. **Distribution by State:** 49 states represented, with **NY and CA** having the highest number of colleges.

3. **Public vs Private:** The dataset shows more **private colleges** than public.
4. **Highest Colleges per State:** California and New York dominate in terms of number of listed colleges.
5. **Percentage of Undergraduates:** Created new metric – highlighted colleges with 100% undergraduate population.
6. **Grant Aid:** Top 5 colleges provide the highest average financial aid.
7. **State-Specific Analysis (CA):** Average undergraduate population in California colleges is high, reflecting larger institutions.
8. **Admission Chances:** Colleges with the highest acceptance rates (close to 100%) provide easier entry.
9. **Specialization Check:** Several "Art" colleges were identified from the dataset.
10. **Alumni Salary Buckets:** Divided salaries into 3 levels (L, M, H).
    - **High Salary Group (28 colleges):** Mostly prestigious Ivy League and private institutions.
11. **Mean Alumni Salary:** Strong positive correlation observed between high-ranking colleges and higher alumni earnings.

5. Conclusion

The analysis of Forbes' 2019 Top Colleges dataset revealed significant trends:

- Elite Ivy League colleges dominate the top rankings and offer high alumni returns.
- California and New York house the most institutions, reflecting their educational hubs.
- Private colleges dominate the dataset, though public universities often serve larger populations.
- Financial aid varies widely, with elite colleges providing substantial grants.
- Categorization of alumni salary levels allows identification of top-performing institutions in terms of career outcomes.

Future Work:

- Apply visualization techniques (bar plots, heatmaps, state-level maps).
- Perform clustering to group colleges with similar characteristics.
- Extend analysis across multiple years to see ranking trends.