

**Mapping the Educational Landscape: A**

**Comprehensive Data Analytics Study of Engineering Colleges in India**



**ENGINEERING COLLEGES IN INDIA**

Presented by: -

I.Teja Venkata Ram

**ACKNOWLEDGMENT**

I want to extend my sincere gratitude to the entire team at “Analytics Space” for granting me the opportunity to work on the “Data Analytics Study of Engineering Colleges in India” project. I would also like to express my heartfelt thanks to my academic team, 'Data Trained,' for providing me with this enriching experience. Working on this project not only allowed me to enhance my analytical skills but also exposed me to a plethora of new insights. The invaluable suggestions and guidance from both teams played a pivotal role in the successful completion of this project, and I am truly grateful for the learning and growth it has brought to me.

Chapter 1. INTRODUCTION

* 1. **Project Objective**

The objective of the “Data Analytics Study of Engineering Colleges in India” project is to conduct an in-depth analysis of the pricing variations among different colleges in India. The aim is to analyze college data across India to uncover student preferences, optimize fee structures, and provide actionable market insights. These insights will help enhance operations, marketing strategies, and overall competitiveness in the Indian education sector.

The data is taken from “Getmyuni” website by using some python-based web scrapping to extract the information from data in the form of datasets.

* **Overview of Project**

Discover the main objectives and scope of Project for web scraping and Data cleaning, Data Modelling, Data preprocessing In Python and data visualization in Tableau.

**1.2 Problem Statement**

Engineering education in India is marked by significant diversity in fee structures, student preferences, and institutional offerings. However, there is limited visibility into the underlying factors driving these variations. This lack of clarity poses challenges for students in making informed decisions, for institutions in optimizing their operations, and for policymakers in addressing affordability and accessibility.

Key challenges include:

* Fee Disparities: Wide-ranging fee structures across colleges with no standardized benchmarks.
* Lack of Insights: Limited availability of actionable data on student preferences and institutional performance.
* Accessibility: Difficulty for stakeholders (students and parents) to evaluate colleges based on pricing and value for money.
* Market Competition: Institutions struggle to position themselves effectively in a competitive landscape.

#### ****1.3 Project Objectives****

The primary objective of this project is to analyze a dataset comprising college name and fees and rank and its rating and its state and city of the college.

* Analyse Pricing Variations

Conduct an in-depth analysis of the fee structures of engineering colleges across India to identify disparities and trends.

* Uncover Student Preferences

Study student preferences and choices concerning college location, fees, and other parameters to understand their decision-making process.

* Optimize Fees Strategies

Provide insights to institutions for optimizing their fee structures while ensuring affordability and competitiveness.

* Provide Actionable Market Insights

Generate actionable insights to help colleges improve marketing strategies, operational efficiency, and overall market positioning.

* Support stakeholder Decision-Making

Facilitate informed decision-making for students, institutions, and policymakers by presenting a comprehensive analysis of the educational landscape.

#### ****1.4 Scope****

This project focuses on analysing pricing trends and student preferences among engineering colleges in India. By collecting data from the "Getmyuni" website using Python-based web scraping techniques, it aims to uncover insights into fee structures, student priorities, and market trends. The study provides actionable recommendations for optimizing fee strategies, improving institutional marketing, and enhancing decision-making for nm,.students, colleges, and policymakers. Covering colleges across India, the project aspires to promote affordability, accessibility, and competitiveness in the engineering education sector.

### **Chapter 2: Data Collection and Sources**

#### ****Data Source****

The data for this project is sourced from "Getmyuni," an online platform that provides comprehensive details about educational institutions in India. The website offers information on engineering colleges, including fee structures, course offerings, rankings, student reviews, and other vital details that influence student decision-making.

To gather this data, Python-based web scraping techniques are utilized. Tools such as Beautiful Soup, Selenium, and Scrapy are employed to extract structured information systematically. These tools ensure accurate retrieval of large volumes of data directly from the website.

The collected data includes:

* College Information: Names, locations, rankings, and specializations offered.
* Fee Structures: Detailed breakdown of tuition fees for various courses.
* Student Feedback: Reviews and ratings that provide insights into college performance.
* Other Details: Facilities, infrastructure, and placement statistics.

The extracted data is then stored in structured formats such as CSV or Excel, enabling easy accessibility and further analysis. To ensure reliability, the data undergoes preprocessing, including cleaning, removing duplicates, and handling missing values.

This comprehensive approach ensures the availability of high-quality datasets for analysing pricing trends and student preferences in engineering education across India.

#### ****Web Scraping Process****

The web scraping process for this project is designed to extract detailed information about engineering colleges from the "Getmyuni" website systematically.

The key Python libraries utilized in the web scraping process were:

1. **BeautifulSoup:** This library was used to parse the HTML content of the Getmyuni website. BeautifulSoup provides a simple way to navigate and search through the HTML structure, Enables easy extraction of specific data points, such as college names, fees, and reviews.
2. **Requests:** The Requests library was employed to send HTTP requests to the Getmyuni website and retrieve the HTML content of the pages. By using Requests, we could access multiple pages of the website programmatically, allowing for the extraction of data from many colleges.
3. **Pandas:** After extracting the data, Pandas was used to organize it into a structured format, typically a DataFrame. This allowed for easier manipulation and analysis of the data, including cleaning, filtering, and sorting.

The web scraping process involved several steps:

**Sending HTTP Requests**:

The initial step involves sending HTTP requests to the target website using the Requests library in Python.

The purpose is to retrieve the HTML content of the webpages containing college details, such as names, locations, courses, and fees.

**Parsing HTML Content:**

Once the HTML content is retrieved, the Beautiful Soup library is used for parsing.

This involves navigating through the HTML structure to locate specific tags and attributes containing the desired data.

For instance, college names, fee details, and course offerings are identified by examining HTML elements like <div>, <table>, or <span>.

**Extracting Data:**

After identifying the required elements, the data is extracted and stored in structured formats, such as lists or dictionaries in Python.

**Data Cleaning and Preprocessing:**

The extracted data is cleaned to remove duplicates, handle missing values, and correct formatting issues.

**Storing Data:**

The cleaned data is saved into structured file formats, such as CSV, Excel, or JSON, to facilitate further processing and analysis.

For this project, Python was the primary programming language used, due to its robust libraries and ease of use for data scraping tasks.

#### ****,****

#### ****Challenges****

While the web scraping process for the project was effective in collecting a comprehensive dataset on engineering colleges in India, several challenges were encountered:

1. **Handling Large Volumes of Data:** The vast number of colleges and related data (e.g., fee structures, course details, reviews, rankings), the data volume was immense. Processing such a large dataset required significant system resources and time. Data was collected in manageable batches to reduce memory overload. Optimized data structures were employed to store and process data efficiently.
2. **Website Restrictions:** Websites often implement measures to prevent automated scraping, such as CAPTCHAs, rate limiting, and dynamic content loading. These limitations posed significant challenges, particularly when trying to scrape data at scale. To mitigate these issues, strategies such as introducing delays between requests, randomizing user agents, and handling exceptions were employed.
3. **Data Accuracy and Consistency:** Another challenge was ensuring the accuracy and consistency of the data collected. Some Colleges listings had incomplete information or inconsistent formatting, which required additional data cleaning and preprocessing steps to ensure the dataset was usable for analysis.

Despite these challenges, the web scraping process successfully yielded a robust dataset that enabled detailed analysis of pricing structures, course preferences, and rankings across engineering colleges in India. This data formed the foundation for uncovering trends, optimizing pricing models, and offering actionable recommendations to stakeholders.

### **Chapter 3: Data Preprocessing**

#### Effective data preprocessing is a crucial step to ensure the quality and reliability of the analysis in Tableau. This phase involves cleaning the raw data, transforming it into a suitable format for analysis, and importing it into Tableau for visualization. Below is a detailed explanation of the steps undertaken during this phase.

#### ****Data Cleaning****

#### The raw data collected through web scraping contained inconsistencies and errors that needed to be addressed before analysis. Data cleaning ensures the dataset is accurate, consistent, and ready for visualization in Tableau. The following steps were performed.

1. **Removal of Duplicates:**
   * **Identification of Duplicates:** Duplicates were identified in the dataset, particularly when the same engineering college appeared multiple times due to minor variations in name formatting or multiple course offerings.
   * **Removal Process:** After importing the raw data into Tableau Prep, duplicate rows were identified using a combination of key fields such as college name, location, and affiliation. The Clean Step in Tableau Prep was used to remove these duplicates systematically, ensuring each college appeared only once in the dataset. This step helped avoid bias in visualizations, such as inflated averages or incorrect totals.
2. **Handling Missing Values:**
   * **Identifying Missing Data:** Missing values can significantly impact the results of the analysis if not handled properly. In this project, missing data was primarily found in columns related to customer reviews and ratings, where some entries lacked complete information.
3. **Correction of Errors:**
   * **Data Accuracy:** An essential part of preparing the dataset for analysis in Tableau involved correcting errors, ensuring consistency, and addressing anomalies. These steps were necessary to maintain the reliability and accuracy of insights derived from the data. Some tuition fees were recorded in varying formats, such as "₹1,50,000" and "150000," or displayed in incorrect units. Certain fields, such as course durations or accreditations, were mismatched across rows
   * **Standardization:**

**Currency Formatting:** All monetary fields were standardized to a consistent format (Indian Rupees) to ensure uniformity in analysis and comparisons.

* + Column Realignment: Misaligned data entries were corrected by ensuring that each value was properly categorized under its respective column.
  + Consistent Labels: Textual fields, such as course names or college types, were cleaned to ensure consistent capitalization and formatting for Tableau visualizations.

1. **Outlier Detection:**
   * **Identifying Outliers:**

Challenges: Outliers, such as unusually high or low tuition fees, student enrolment figures, or incorrect ratings, were detected during the preprocessing phase. The Interquartile Range (IQR) was employed to identify data points lying beyond 1.5 times the upper or lower quartiles.

* + **Handling Outliers:**

Retention: Outliers that represented meaningful data points (e.g., colleges offering premium, high-cost courses) were retained for specific analyses**.** Data points determined to be anomalies (e.g., fees listed as ₹0 or extremely high enrolments due to typographical errors) were corrected or removed. Tableau Prep's data profiling tools were used to visualize and manage these anomalies effectively.

### 

### **Chapter 4: Data Analysis and Visualization**

The Getmyuni platform, with its vast database of college-related information, offers a unique opportunity to analyse student preferences, college rankings, and other metrics critical for understanding the Indian higher education landscape. Tableau was utilized to process and visualize the data collected, uncovering actionable insights into areas such as student sentiment, fee trends, popular courses, and geographic distribution of colleges.

#### ****Overview of Analysis****

#### The primary focus of this project was to analyse data gathered from Getmyuni and present insights into college trends in India. The key metrics studied include:

1**.Student Sentiment**

* Analysis of reviews and ratings to understand the factors driving student satisfaction and dissatisfaction, as well as identifying the strengths and weaknesses of various colleges.

2**.Fee Structures and Affordability**

* Investigation of tuition fee trends across regions, colleges, and streams to understand affordability and competitiveness.

3**.Course Preferences**

* Analysis of the popularity of different courses and streams, such as engineering, management, arts, and sciences, to identify demand patterns.

4**.Geographic Distribution**

* Visualization of the geographic spread of colleges, including hotspots for higher education and regions with untapped potential for growth.

#### ****Visualizations****

The following visualizations were created in Tableau to illustrate the key findings from the data analysis.

1. **Fee Analysis**

**a. Distribution of Fees Across Colleges**

* **Visualization**: A histogram was used to show the distribution of tuition fees across colleges.
* **Findings**: Most colleges fell within a mid-range fee bracket, while only a few colleges were positioned as premium institutions with high fees.

**b. Regional Fee Variations**

* **Visualization**: A bar chart was employed to compare the average tuition fees across different states.
* **Findings**: States like Maharashtra and Karnataka showed higher average fees, likely due to their prominence as educational hubs.

**c. Fee Trends by Course**

* **Visualization**: A box-and-whisker plot was created to display the fee ranges for popular courses such as engineering, MBA, and medical studies.
* **Findings**: Engineering and MBA programs had wider fee ranges, reflecting both affordability and premium options within these courses.

2**. Student Sentiment Analysis**

**a. Sentiment Overview**

* **Visualization**: A pie chart summarized student sentiments (positive, neutral, and negative) from reviews.
* **Findings**: Positive reviews were dominant, emphasizing satisfaction with faculty quality and infrastructure. Negative reviews often mentioned placement opportunities and outdated curricula.

**b. Factors Influencing Satisfaction**

* **Visualization**: A stacked bar chart broke down factors driving student satisfaction and dissatisfaction, such as placements, faculty, campus infrastructure, and extracurricular activities.
* **Findings**: Placements and faculty were key drivers of satisfaction, while infrastructure and fees were often cited as areas of concern.

3**. Geographic Analysis**

**a. Distribution of Colleges Across India**

* **Visualization**: A map of India displayed the geographic distribution of colleges.
* **Findings**: States like Tamil Nadu, Maharashtra, and Uttar Pradesh had the highest concentrations of colleges, while northeastern states had comparatively fewer institutions.

**b. Popular Education Hotspots**

* **Visualization**: A heat map was used to identify regions with high demand for higher education.
* **Findings**: Cities like Bengaluru, Pune, and Hyderabad emerged as education hubs with a large number of colleges and students.

**4. Course Preferences**

**a. Popular Streams**

* **Visualization**: A bar chart was used to rank streams such as engineering, arts, management, and medical studies based on the number of colleges offering these courses.
* **Findings**: Engineering emerged as the most offered course, followed by arts and management.

**b. Student Preferences Across Streams**

* **Visualization**: A stacked bar chart analyzed student reviews and ratings for different streams.
* **Findings**: Management and medical colleges received higher average ratings compared to arts and engineering colleges, indicating varying levels of satisfaction across streams.

**c. Emerging Courses**

* **Visualization**: A line chart was created to show the year-over-year growth in the number of colleges offering emerging courses such as data science, AI, and digital marketing.
* **Findings**: There has been a noticeable uptick in colleges offering programs in technology-driven fields, reflecting industry trends.

#### ****Interpretation****

The visualizations provide a comprehensive view of the key trends and patterns in the Indian Colleges. The following interpretations can be drawn from the analysis:

1**.Student Satisfaction**

* Positive feedback highlighted faculty and placements as significant strengths of many colleges, while concerns over outdated facilities and curriculum need to be addressed to improve student experiences.

2.**Fee Structures**

* Affordable options dominate the landscape, but premium colleges continue to attract students by offering superior placements and modern facilities. Regional fee disparities suggest potential for better access in underserved areas.

3**.Regional Dynamics**

* States like Tamil Nadu and Maharashtra lead in terms of the number of colleges, but there is considerable scope to expand education infrastructure in northeastern states and rural areas.

4.**Demand for Emerging Courses**

* The rising popularity of data science and AI programs reflects growing student interest in industry-relevant skills. Colleges offering such courses are likely to attract more students in the coming years.

5**.Market Segmentation**

* Management and medical colleges show higher satisfaction ratings, suggesting a focus on quality over quantity in these streams. Engineering colleges, while numerous, need to focus more on placement quality to remain competitive.

Charts:-

This section details the functionality and insights derived from a Tableau dashboard designed to analyse college data on the Getmyuni platform. The dashboard incorporates filters and visualizations to explore metrics such as college fees, ratings, location-based insights, and student reviews.

**City Filter**

* **Function**: A slicer enabling users to filter the dashboard by a specific city.
* **Usage**: Selecting a city such as "Delhi" or "Mumbai" displays data for colleges within that location.
* **Example Insight**: Filtering by "Bangalore" shows 50+ colleges with insights into fees, ratings, and student preferences specific to Bangalore.

**Fee Range Filter**

* **Function**: A slider that allows users to set a minimum and maximum fee range.
* **Usage**: Users can narrow their search to display colleges with tuition fees falling within a specific budget, such as ₹50K to ₹1 lakh.
* **Example Insight**: Filtering colleges with fees less than ₹1 lakh reveals options suitable for cost-conscious students.

**Ratings Filter**

* **Function**: A slicer for filtering colleges based on their overall ratings (e.g., from 3.0 to 5.0).
* **Usage**: Users can explore colleges with higher ratings to find institutions with better reviews and student satisfaction.
* **Example Insight**: Filtering ratings between 4.0 and 5.0 highlights top-rated colleges in each city.

**Key Metrics and Visualizations**

**1. Average of Ratings (e.g., 4.2)**

* **Description**: A card displaying the average rating of colleges based on the selected filters.
* **Interpretation**: Provides a quick overview of the overall satisfaction level of students in the selected region or within a specific fee range.
* **Example Insight**: The average rating for colleges in Bangalore is 4.2, indicating high satisfaction levels.

**2. Average of Tuition Fees (e.g., ₹1.2 Lakh)**

* **Description**: A card showing the average tuition fees of colleges within the selected filters.
* **Interpretation**: Offers insight into the affordability of colleges in the selected city or course category.
* **Example Insight**: The average tuition fee for engineering colleges in Pune is ₹1.2 lakh.

**Dashboard Insights for College Data Analysis on Getmyuni Using Tableau**

This section details the functionality and insights derived from a Tableau dashboard designed to analyse college data on the Getmyuni platform. The dashboard incorporates filters and visualizations to explore metrics such as college fees, ratings, location-based insights, and student reviews.

**Filters**

**1. City Filter**

* **Function**: A slicer enabling users to filter the dashboard by a specific city.
* **Usage**: Selecting a city such as "Delhi" or "Mumbai" displays data for colleges within that location.
* **Example Insight**: Filtering by "Bangalore" shows 50+ colleges with insights into fees, ratings, and student preferences specific to Bangalore.

**2. Fee Range Filter**

* **Function**: A slider that allows users to set a minimum and maximum fee range.
* **Usage**: Users can narrow their search to display colleges with tuition fees falling within a specific budget, such as ₹50K to ₹1 lakh.
* **Example Insight**: Filtering colleges with fees less than ₹1 lakh reveals options suitable for cost-conscious students.

**3. Ratings Filter**

* **Function**: A slicer for filtering colleges based on their overall ratings (e.g., from 3.0 to 5.0).
* **Usage**: Users can explore colleges with higher ratings to find institutions with better reviews and student satisfaction.
* **Example Insight**: Filtering ratings between 4.0 and 5.0 highlights top-rated colleges in each city.

**Key Metrics and Visualizations**

**1. Average of Ratings (e.g., 4.2)**

* **Description**: A card displaying the average rating of colleges based on the selected filters.
* **Interpretation**: Provides a quick overview of the overall satisfaction level of students in the selected region or within a specific fee range.
* **Example Insight**: The average rating for colleges in Bangalore is 4.2, indicating high satisfaction levels.

**2. Average of Tuition Fees (e.g., ₹1.2 Lakh)**

* **Description**: A card showing the average tuition fees of colleges within the selected filters.
* **Interpretation**: Offers insight into the affordability of colleges in the selected city or course category.
* **Example Insight**: The average tuition fee for engineering colleges in Pune is ₹1.2 lakh.

**Bar Charts**

**3. Total Fees by College Name**

* **Visualization**: A bar chart displaying the total tuition fees for each college.
* **Key Insights**:
  + Premium colleges such as "IIT Delhi" and "IIM Bangalore" show higher fees.
  + Budget-friendly options like "Government Colleges" or "Regional Colleges" have lower fee sums.
* **Interpretation**: Helps users identify colleges within their affordability range and compare options.

**4. Minimum Fees by City**

* **Visualization**: A bar chart showing the minimum tuition fees across different cities.
* **Key Insights**:
  + Kolkata and Chennai have the lowest minimum fees for colleges.
  + Cities like Delhi and Mumbai have higher starting fees.
* **Interpretation**: Useful for identifying the least expensive college options across cities.

**Ratings Analysis**

**5. Sum of Ratings by College Name**

* **Visualization**: A bar chart aggregating ratings for each college based on student reviews.
* **Key Insights**:
  + Colleges such as "IIM Ahmedabad" and "BITS Pilani" receive the highest cumulative ratings.
  + Colleges with lower cumulative ratings may need to focus on improving areas like infrastructure or placements.
* **Interpretation**: Helps users identify colleges that consistently receive positive feedback.

**6. Ratings by College Location**

* **Visualization**: A bar chart aggregating ratings by college addresses.
* **Key Insights**:
  + Areas like "Whitefield, Bangalore" and "Sector 62, Noida" have colleges with high ratings.
  + Locations like "Outskirts of the City" tend to have lower ratings due to accessibility and infrastructure issues.
* **Interpretation**: Highlights how location impacts student satisfaction.

**7**. Geographic Visualizations

**. Map of Colleges by Location**

* **Visualization**: A map showcasing the geographic distribution of colleges across India.
* **Key Insights**:
  + Cities like Pune, Hyderabad, and Bangalore have a high density of colleges.
  + Northeastern states and rural areas show a lack of college infrastructure.
* **Interpretation**: Provides an overview of education hotspots and regions with potential for development.

**Chapter 5: Tools and Technologies**

Software Requirement:-

1)Python(Jupyter notebook)

2)Tableau

3)Excel

Main Software Libraries:-

1)Pandas

2)Numpy

3)Beautiful Soups

4)Requests

5)CSV

Chapter 6: Dashboard

A computer screen shot of a computer

Description automatically generated

Chapter 7: Conclusion

### This project aimed to leverage data analytics to deliver actionable insights and enhance decision-making for colleges and universities on the Getmyuni platform. By analysing data such as college names, locations, fees, ratings, reviews, and course offerings, the study explored trends in higher education, student preferences, and market demands. Key findings include factors driving student satisfaction, regional fee differences, and rising interest in courses like AI, Data Science, and Digital Marketing. The analysis also segmented colleges into premium and budget-friendly categories, revealing unique preferences among diverse student demographics. These insights provide a comprehensive understanding of India's higher education landscape, enabling institutions to refine strategies, improve student engagement, and align programs with current market trends. The findings establish a strong foundation for enhancing the educational ecosystem and addressing the diverse needs of prospective students effectively.