

**Mapping the Educational Landscape: A Comprehensive Data Analytics Study of Engineering Colleges in India**

Data analytics project

by

Basani Rakesh

**Acknowledgment**

We extend our heartfelt gratitude to everyone who contributed to the successful completion of this project, “*Mapping the Educational Landscape: A Comprehensive Data Analytics Study of Engineering Colleges in India”*

We sincerely thank our mentors and instructors for their invaluable guidance and constructive feedback, which were instrumental in overcoming challenges and achieving the project’s objectives.

We are grateful to the online platform careers360.com for providing the essential data that enabled an in-depth analysis of the hospitality industry in India.

Our appreciation also goes to the tools and technologies, including Python libraries like BeautifulSoup and Pandas, as well as visualization tools such as Tableau and Power BI, which made this project possible.

Finally, we thank our peers, family, and friends for their unwavering support and encouragement throughout this journey. This project is a testament to collaborative efforts, and we are humbled by the opportunity to contribute meaningful insights to the Education sector.

**Chapter 1: Introduction**

* 1. **Overview of the Engineering college in India**

Engineering education in India plays a crucial role in shaping the nation’s workforce and technological advancements. With thousands of institutions spread across various states, students have a plethora of choices when it comes to selecting an engineering college. However, the process of choosing the right institution can be overwhelming due to the vast number of options and the lack of centralized, easily accessible information.

The engineering colleges in India vary significantly in terms of quality, infrastructure, course offerings, and fee structures. Some colleges boast world-class research facilities and strong industry collaborations, while others struggle with outdated curricula and limited resources. This disparity makes it essential to have a structured and analytical approach to evaluating these institutions.

This project seeks to address these challenges by leveraging data analytics techniques to compile, analyze, and visualize data related to engineering colleges across India. By organizing and interpreting this information, we aim to provide a clear and detailed understanding of various aspects of engineering education, ultimately assisting students, parents, policymakers, and academic institutions in making informed decisions.

**1.2 Problem Statement**

Students, parents, and policymakers often struggle to make informed decisions due to a lack of centralized and structured information regarding engineering colleges. The scattered nature of data makes it difficult to evaluate colleges based on crucial aspects such as quality, affordability, and accessibility. Additionally, important factors like faculty expertise, placement records, research opportunities, and campus facilities remain inconsistent across different sources, leading to confusion and potential misinformation.

**Key Challenges:**

* **Lack of Centralized Data**: Information about engineering colleges is scattered across multiple sources, making it difficult to access comprehensive and reliable data.
* **Variation in Fee Structures**: There is a significant disparity between government and private institutions in terms of tuition fees, making it challenging for students from different socio-economic backgrounds to choose an affordable option.
* **Infrastructure and Facilities**: Many institutions do not provide clear information about the availability and quality of essential infrastructure like laboratories, libraries, and hostels, impacting students' learning experiences.
* **Employment and Placement Data**: Reliable placement records are often not publicly available, making it hard for students to assess career prospects post-graduation.
* **Course Availability and Specializations**: The variety of engineering specializations is growing, but information on course availability, quality, and interdisciplinary programs remains fragmented.
* **Policy and Decision-Making**: Policymakers lack structured data to formulate evidence-based education policies that support equitable access to quality engineering education.

This study aims to bridge these information gaps by leveraging data analytics to compile, analyze, and present a comprehensive dataset. By offering a structured and insightful analysis, the project seeks to empower students, parents, and policymakers with actionable information that facilitates better decision-making in the engineering education landscape.

delivering actionable recommendations to enhance decision-making and drive growth.

**1.3 Objectives of the Project**

The primary objectives of this project are:

1. **Engineering Colleges Across States**
2. **Course Availability**
3. **Fee Structure Analysis**
4. **Government and Private Colleges**
5. **Infrastructure Assessment**

1. **Engineering Colleges Across States**: Analyze the distribution of engineering colleges across different states. This will help identify regions with high and low concentrations of institutions, providing insights into accessibility and regional disparities in engineering education.
2. **Course Availability**: Categorize courses, including undergraduate and postgraduate programs. By understanding the range of available courses and specializations, students can make informed choices based on their career aspirations and interests.
3. **Fee Structure Analysis**: Examine annual and total course fees, differentiating between private and government institutions. This will provide clarity on affordability and assist students in selecting colleges that align with their financial capabilities.
4. **Government and Private Colleges**: Assess enrolment rates, infrastructure, and academic contributions of both types of institutions. This analysis will help students and policymakers understand the strengths and weaknesses of government-funded and private engineering colleges.
5. **Infrastructure Assessment**: Evaluate facilities such as laboratories, libraries, hostels, and sports amenities. Assessing the quality and availability of infrastructure will enable students to choose colleges that provide a conducive learning environment and support their academic growth.

.

**Chapter 2: Data Collection and Sources**

**2.1 Data Source**

To gain insights into the dynamics of the Indian Education sector, data was gathered from prominent online education platforms, specifically Careers360.com These platforms were selected for their vast databases that encompass a wide array of colleges across in India, Course availability, and fees structure. They offer a diverse selection of Facilities across various regions and fees ranges, thereby creating a rich dataset for thorough analysis.

**2.2 Web Scraping Methodology**

The data collection process involved web scraping, a technique for extracting information from websites. The following steps were undertaken:

**1. Identifying Key Data Points:**

The project focused on collecting key information, including college names, locations, reviews, ratings, courses and Fees. These data points were selected to offer a thorough understanding of Education landscape in India. By compiling this information, we were able to effectively analyse trends and patterns, establishing a solid foundation for generating valuable insights within the Education sector.

**2. Developing Scraping Scripts:**

To streamline the data extraction process, Python libraries such as BeautifulSoup were employed for efficient automation. BeautifulSoup was specifically utilized to parse static HTML content from web pages. For dynamic content, Selenium was incorporated into the workflow. Together, these tools facilitated accurate and seamless data collection from a variety of sources.

**3. Data Storage:**

The extracted data was organized into structured formats such as CSV files and SQLite databases to enhance accessibility. This systematic storage approach facilitated efficient data management and analysis. By using these formats, we ensured compatibility with various analytical tools.

**2.3 Challenges in Data Collection**

Several challenges were encountered during the data collection phase:

**• Dynamic Content:**

Many web pages used JavaScript to load data dynamically, requiring advanced techniques to capture the information.

**• Rate Limits:**

Websites imposed restrictions on the frequency of requests, necessitating the use of delays and proxies to avoid being blocked.

**• Inconsistent Formats:**

Variations in data formats across platforms required additional preprocessing to standardize the dataset. Despite these challenges, the data collection process yielded a rich dataset, forming the foundation for subsequent analysis.

**Chapter 3: Data Cleaning and Manipulation**

**3.1 Steps for Cleaning and Preprocessing**

The raw data collected through web scraping required extensive cleaning and preprocessing to ensure its usability. The following steps were undertaken:

**1. Standardizing Data Formats:**

Fields such as pricing, dates, and locations were standardized to maintain consistency across the dataset. This harmonization made comparisons easier and enhanced the overall usability of the data. The uniform formatting ensured compatibility with analytical tools and minimized errors during the analysis process.

**2. Removing Duplicates:**

Duplicate records in the dataset were identified and eliminated to ensure accuracy in the analysis. This step was critical to avoid redundant information skewing the results. By removing duplicates, the dataset became cleaner and more reliable. This enhanced the overall quality of insights derived from the data.

**3. Handling Missing Values:**

Missing data was managed using appropriate imputation techniques or exclusion based on its significance. Imputation helped fill gaps in the data without compromising analysis accuracy. Exclusion was used when the missing data was minimal or irrelevant. This ensured the dataset remained comprehensive and meaningful.

**4. Creating Derived Metrics:**

Additional metrics, such as average price per region and sentiment scores, were calculated to enhance analytical depth. Derived metrics provided new perspectives and improved decision making. These metrics allowed for more granular and actionable insights.

**5. Sentiment Labelling:**

Textual reviews were analysed to determine their sentiment as positive, negative, or neutral. Sentiment analysis provided valuable insights into customer opinions. Categorizing reviews helped identify trends in satisfaction or dissatisfaction. This added a qualitative dimension to the analysis.

**3.2 Tools Used**

**• Pandas:** For data manipulation and preprocessing.

• **NumPy:** For handling numerical data and calculations.

**• OpenRefine**: To assist with standardizing and cleaning data.

**3.3 Importance of Data Cleaning**

Data cleaning is a vital component of any analytics project, as the accuracy of the analysis relies heavily on the quality of the data used. By systematically addressing inconsistencies and filling in missing values, this project has ensured that the dataset is dependable and comprehensive, laying the groundwork for insightful findings. Clean data is crucial for informed decision-making, as it accurately captures the underlying characteristics of the phenomena being studied. By rectifying inconsistencies, managing gaps in the data, and standardizing formats, data cleaning creates a strong basis for effective analysis. Additionally, having clean data enhances algorithm performance and reduces computational strain, making it easier to extract meaningful insights. In industries like hospitality, where decisions often hinge on subtle trends and customer feedback, the significance of clean data is paramount. It empowers stakeholders to focus on actionable insights without being hindered by irrelevant noise or inaccuracies in the dataset.

**Chapter 4: Data Analysis and Insights**

The **Careers360** platform, with its extensive database of college-related information, provides a valuable opportunity to analyze student preferences, college rankings, and other key metrics that shape the Indian higher education landscape. By leveraging **Power BI** for data processing and visualization, this analysis uncovers critical insights into areas such as student sentiment, fee trends, popular courses, and the geographic distribution of colleges.

**Overview of Analysis**

The primary objective of this project was to analyze data gathered from **Careers360** and present key insights into trends shaping India's higher education sector. The study focuses on the following metrics:

**1. Student Sentiment Analysis**

* Examination of student reviews and ratings to identify the factors influencing satisfaction and dissatisfaction.
* Evaluation of qualitative feedback to determine common concerns, such as faculty quality, infrastructure, placement opportunities, and curriculum effectiveness.
* Comparative analysis of colleges based on student perception to highlight institutions excelling in specific areas.

**2. Fee Structures and Affordability**

* Investigation of tuition fee variations across different regions, college types (private, public, deemed universities), and disciplines.
* Identification of the most and least affordable institutions, helping prospective students make informed financial decisions.
* Trends in financial aid and scholarships offered by institutions, including their accessibility and impact.

**3. Course Preferences and Demand Trends**

* Analysis of student enrollment patterns across various disciplines such as **engineering, management, arts, sciences, law, and medicine**.
* Identification of emerging and high-demand courses, including new-age disciplines like **AI, Data Science, Cybersecurity, and FinTech**.
* Insights into specialization preferences within broad fields, such as mechanical vs. computer science in engineering or finance vs. marketing in management.

**4. Geographic Distribution of Colleges**

* Visualization of the concentration of colleges across states and major education hubs in India.
* Identification of **regions with a high density of institutions** (e.g., Maharashtra, Tamil Nadu, Karnataka, and Uttar Pradesh).
* Analysis of **education gaps** in underserved areas, indicating potential regions for new institutions and infrastructure development.

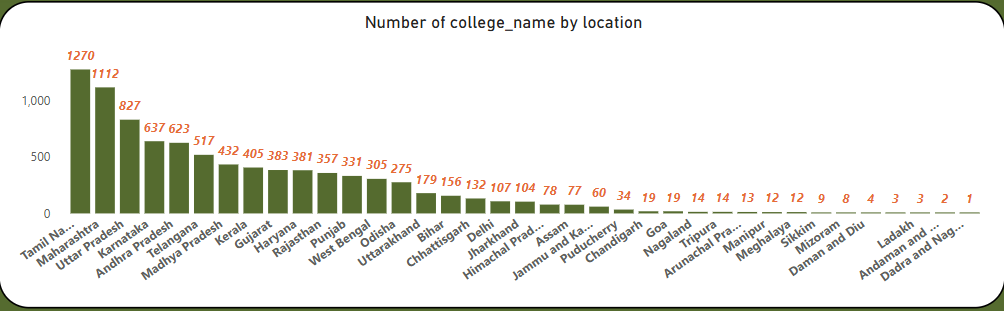
**Chapter 5: Visualization and Interpretation**

**5.1 Key Visualizations for Engineering college in India**

To illustrate the findings, several visualizations were created:

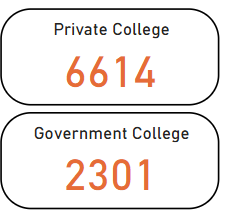
**1) College based on location:**

A column chart illustrates the total number of college present in particular states. In it the data values shown is the total colleges which combines private and government college.



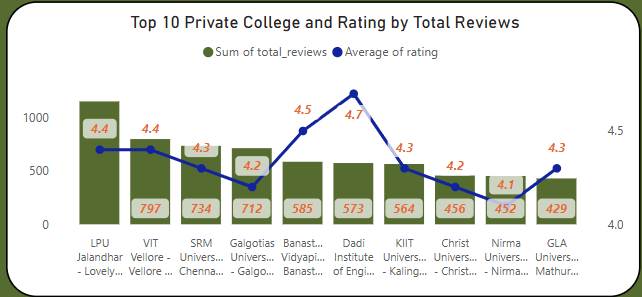
**2) College Type based on location:**

Two Cards illustrates the total number of private colleges and Government colleges. The card data will can based on location filter.



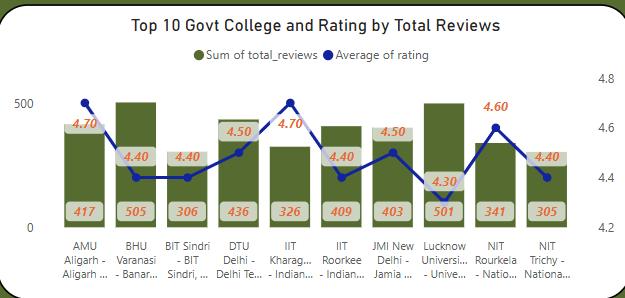
**3) Top 10 private college and rating by Total Reviews:**

A line and stacked column chart illustrates top 10 Private college and Ratings by total reviews. In It the college name is placed in x axis and average rating is placed in line y axis and sum of the reviews placed in the column y axis filter by total reviews.



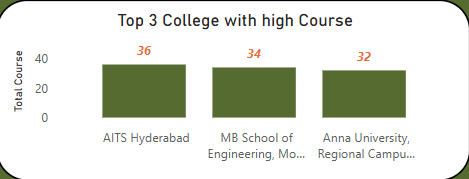
**4) Top 10 Government college and rating by Total Reviews:**

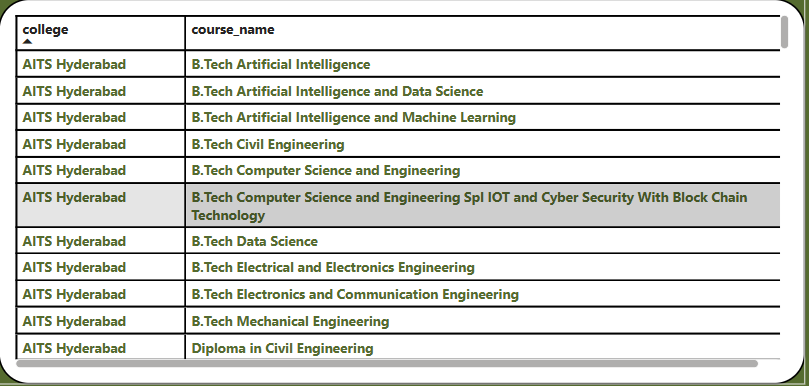
A line and stacked column chart illustrates top 10 Government college and Ratings by total reviews. In It the college name is placed in x axis and average rating is placed in line y axis and sum of the reviews placed in the column y axis filter by total reviews.



**5) Top 3 college based on courses available:**

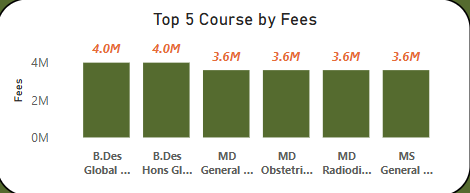
A stacked column chart illustrates top 3 college by total courses. In It the college name is placed in x axis and total courses is placed in y axis and table shows the courses name.





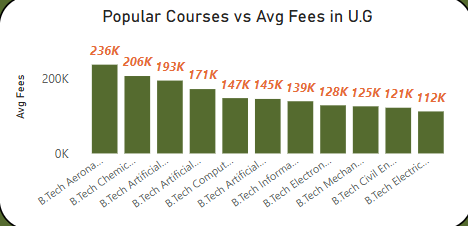
**6) Top 5 courses by Fees:**

A stacked column chart illustrates top 5 courses by Fees. In It the courses is placed in x axis and Fees is placed in y axis.



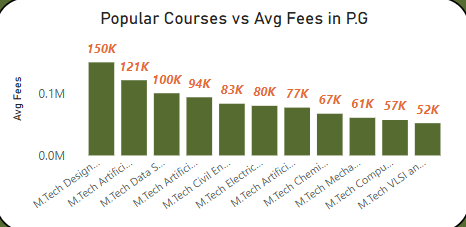
**7) popular courses by Average Fees in under graduate:**

A stacked column chart illustrates Popular Courses by Average Fees. In It the Courses is placed in x axis and Average fees is placed in y axis.

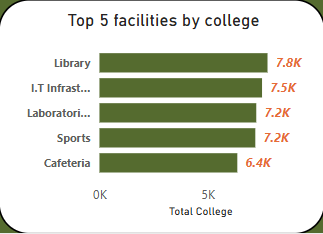


**8) popular courses by Average Fees in post graduate:**

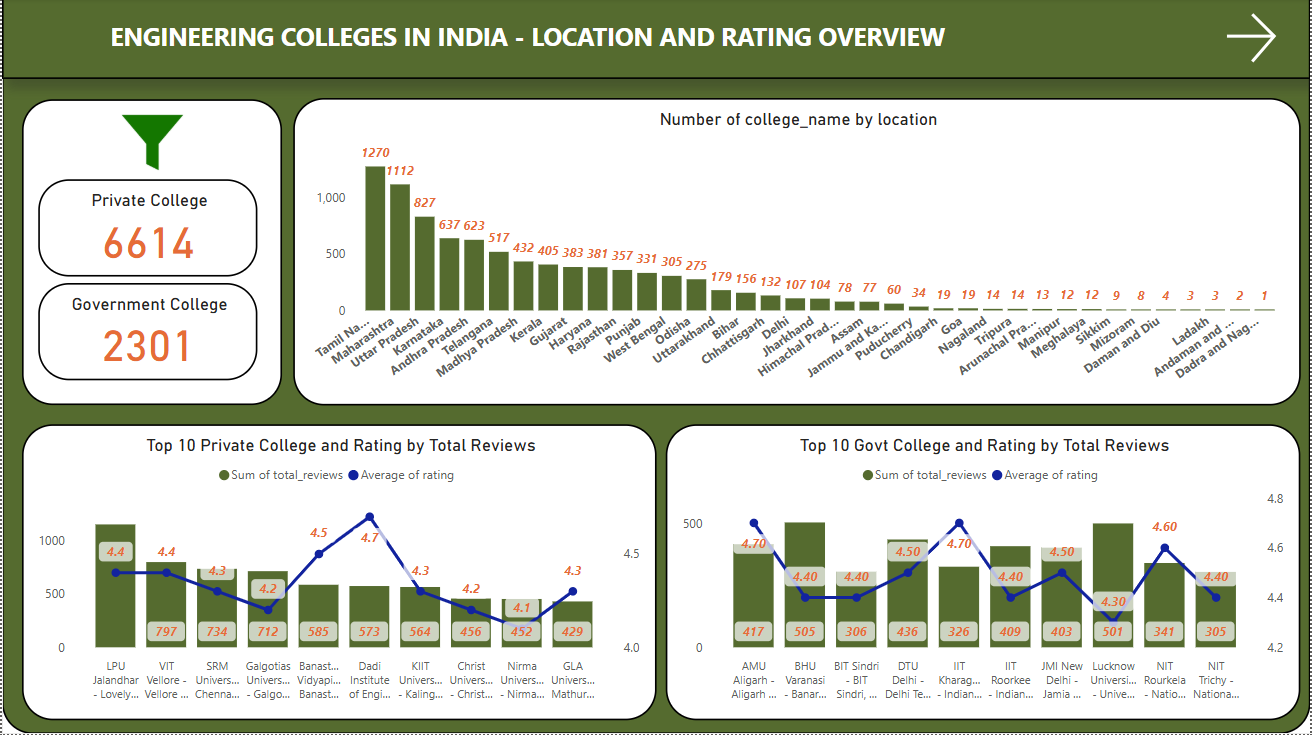
A stacked column chart illustrates Popular Courses by Average Fees. In It the Courses is placed in x axis and Average fees is placed in y axis.

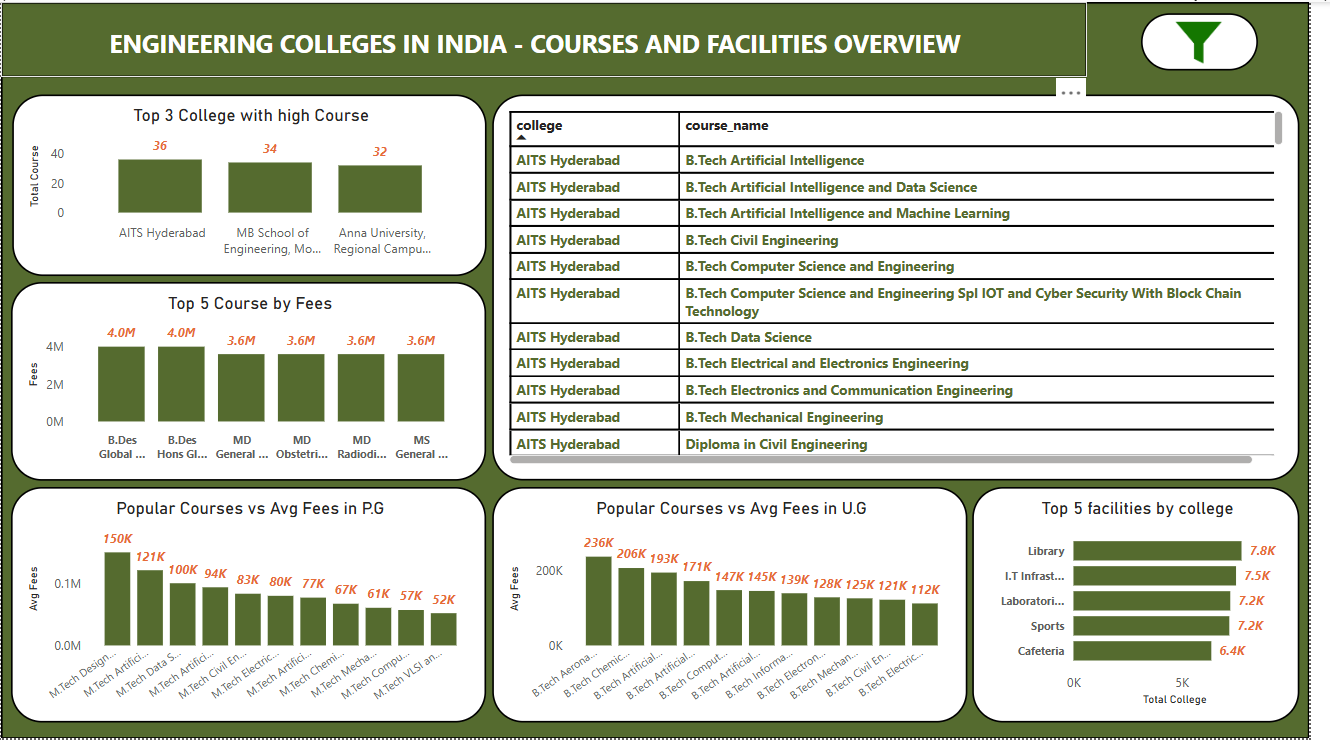


**8) Top 5 Facilities by college:**

A stacked bar chart illustrates Top 5 Facilities by college. In It the count of college is placed in x axis and Facilities is placed in y axis

**Chapter 6: DASHBOARD VISUALIZATION**

****

****

**Chapter 7: Tools and Technologies Used**

**9.1 Web Scraping**

**• Tools:** BeautifulSoup,

**• Purpose:** Extracting structured data from online platforms.

**9.2 Data Cleaning and Manipulation**

• **Tools:** Pandas, NumPy, OpenRefine.

• **Purpose:** Preparing data for analysis**.**

**9.3 Analysis and Visualization**

**• Tools:** Tableau, Power BI.

**• Purpose:** Generating insights and creating visualizations.

**Chapter 8: Challenges Faced**

**6.1 Technical Challenges**

The technical challenges encountered included:

**1. Large Data Volumes:** Handling large datasets necessitated the use of optimized algorithms to ensure efficient data processing. These algorithms helped manage the complexity and scale of the data without compromising performance. Additionally, efficient storage solutions were implemented to store vast amounts of data in a way that facilitated quick retrieval and analysis. This approach ensured smooth handling of large volumes of information.

**2. Unstructured Data:** Extracting meaningful insights from unstructured text data, such as customer reviews, required advanced preprocessing techniques. These techniques, such as text cleaning, tokenization, and sentiment analysis, helped convert raw data into a structured format for analysis. The goal was to transform the unstructured data into valuable insights that could inform business decisions and improve customer experience.

**6.2 Dataset Limitations**

**1. Incomplete Data:** Certain regions had sparse data, which limited the depth and comprehensiveness of the analysis. The lack of sufficient data from these areas made it challenging to draw meaningful conclusions about customer preferences and trends. Efforts were made to mitigate this issue by focusing on available data while acknowledging the limitations in the analysis.

**2. Data Bias:** Reviews were disproportionately concentrated on popular hotels, which potentially skewed the results. This bias may have led to overemphasis on well-reviewed hotels while underrepresenting smaller or less popular establishments. Recognizing this bias was important for ensuring that the findings provided a balanced view of the hospitality landscape.

**Chapter 9: Conclusion**

This study underscores the importance of data-driven decision-making in the realm of engineering education in India. By systematically analyzing data on colleges, courses, fees, infrastructure, and employment outcomes, we provide stakeholders with valuable insights that can guide their choices and policy decisions.

The findings of this project will help students and parents make informed college selection decisions, ensuring that they choose institutions that align with their academic and financial aspirations. Furthermore, policymakers will gain access to evidence-based insights that can inform strategic initiatives aimed at improving accessibility and quality in engineering education.

Additionally, institutions themselves can benefit from the study by identifying areas where they can enhance their offerings, infrastructure, and industry collaborations to remain competitive. The data analytics approach adopted in this study can serve as a model for future research in higher education planning and evaluation.

By bridging the existing information gap and promoting transparency, this project ultimately contributes to the continuous improvement and equitable development of the engineering education ecosystem in India.