Mapping the Educational Landscape:

A Comprehensive Data Analytics Study of Engineering Colleges in India

ACKNOWLEDGMENT

I would like to extend my sincere gratitude to the entire team at **Analytic Space** for granting me the opportunity to work on the *Mapping the Educational Landscape: A Comprehensive Data Analytics Study of Engineering Colleges in India* project. Their unwavering support, valuable insights, and expertise in data analytics have been instrumental in shaping this research.

I would also like to express my heartfelt thanks to my academic team, **Data Trained**, for providing me with this enriching experience. This project has not only enhanced my analytical skills but has also given me exposure to new methodologies, tools, and perspectives in the field of education analytics.

The invaluable guidance and constructive suggestions from both teams have played a crucial role in the successful completion of this project. I am truly grateful for the learning and growth this experience has brought me, and I look forward to applying these insights in future endeavors.

CHAPTER 1

INTRODUCTION

Due to the large number of institutions and the lack of organized information on costs, programs, and facilities, selecting the best engineering college in India can be difficult. A data-driven methodology is used in this project to examine college ownership, distribution, tuition costs, course availability, and facilities. Web scraping, Python data preprocessing, and Power BI visualization are used in this study to produce structured insights that aid in decision-making for students, teachers, and legislators. Through interactive dashboards, this project presents important trends, improving accessibility and transparency in higher education.

1.1 Overview of Project

This project focuses on collecting and analyzing data on engineering colleges in India using web scraping, data cleaning, and visualization techniques. It examines key factors such as college distribution across states, course offerings, tuition fees, and infrastructure availability to provide a comprehensive understanding of higher education trends. The data is processed using Python for cleaning and transformation, while Power BI is used for visualization, enabling stakeholders to explore insights through interactive dashboards. These findings help students, educators, and policymakers make informed decisions about college selection, affordability, and quality, ultimately improving transparency in the education sector.

The project follows multiple stages:

Web Scraping and Data Collection: Extracting college data from online sources like Careers360.

Data Cleaning and Preprocessing: Handling missing values, standardizing formats, and ensuring data consistency.

Data Analysis and Insights: Exploring trends in ratings, fees, and courses.

Visualization: Using Power BI to present data through interactive dashboards.

1.2 Problem Statement

- **Lack of Structured Information:** Students struggle to find comparable data on engineering colleges.
- **Multiple Decision Factors:** College selection depends on affordability, location, ratings, and course availability.
- **Overwhelming Process:** Students rely on word-of-mouth, rankings, and scattered online data, making comparisons difficult.
- **Affordability Concerns:** Tuition fees vary significantly between private and government institutions.
- **Location Impact:** Proximity affects accessibility, job opportunities, and industry exposure.
- **Inconsistent Ratings:** Academic quality, infrastructure, and placements are rated differently across sources.
- **Course Availability Matters:** Specialized programs like AI, Data Science, and Robotics may not be available everywhere.
- **Data-Driven Approach:** This project compiles and analyzes college data from reliable sources using Power BI for easier comparison.

1.3 Project Objectives

This project systematically examines engineering colleges in India by evaluating their location, type, fees, ratings, courses, and facilities. By presenting structured data, it aims to help students choose the right college, assist educators in institutional evaluation, and support policymakers in education planning. Through data-driven insights and visualization, the study enhances transparency, accessibility, and informed decision-making in higher education.

- **Analyze Engineering Colleges** – Study colleges across India based on key attributes like location, type, courses, facilities, fees, and ratings.

- **Assist Students** – Provide structured and comparable data to help students make informed college selection decisions.
- **Support Educators** – Offer insights into institutional performance for continuous improvement and benchmarking.
- **Guide Policymakers** – Help authorities understand education trends and enhance policy decisions.
- **Evaluate College Quality** – Compare private and government institutions to assess affordability, infrastructure, and academic excellence.
- **Identify Key Trends** – Analyze popular courses, fee structures, and college ratings to reveal student preferences.
- **Enable Data-Driven Decision-Making** – Utilize data analytics and visualization tools like Power BI to simplify insights.

1.4 Scope

Engineering education in India is diverse, with institutions differing in ownership, course offerings, facilities, and affordability. This project compares public and private engineering colleges, focusing on infrastructure, tuition fees, and student experiences. Public institutions are government-funded, offering lower fees and strong academic reputations, while private colleges provide modern infrastructure, industry collaborations, and diverse specializations at higher costs.

A key aspect of this study is the analysis of course offerings and fees. Engineering programs range from traditional fields like Mechanical and Electrical Engineering to emerging domains such as AI, Data Science, and Cybersecurity. Government institutions are more cost-effective, whereas private colleges charge higher fees for advanced facilities and faculty expertise.

Student ratings and facilities are crucial in college selection. Institutions with modern labs, research opportunities, and industry ties often receive better

ratings. Student satisfaction depends on faculty support, campus amenities, extracurricular activities, and hostel facilities, influencing overall learning experiences.

Geographical distribution plays a vital role in accessibility. Top-ranked colleges are concentrated in metro cities like Bangalore, Delhi, Mumbai, and Chennai, attracting students nationwide. However, regional institutions also contribute significantly by providing education to local students and catering to regional job markets.

This project provides data-driven insights into engineering education in India, helping students, educators, and policymakers make informed choices based on affordability, quality, and accessibility.

CHAPTER 2
DATA COLLECTION AND SOURCES

Data source

The data was collected primarily through web scraping from the Careers360 website using requests and BeautifulSoup. The dataset includes key attributes such as college name, location (state, city), and type (government/private) to help categorize institutions. Information on courses offered, specializations, and intake capacity provides insights into academic diversity. The fee structure (annual and total course cost) helps assess affordability, while infrastructure details like labs, libraries, and hostels reflect campus facilities. The year of establishment indicates institutional legacy and stability.

1. Extracting College Details

To ensure a comprehensive dataset, the following key attributes were extracted:

- **College Name:** The full official name of the engineering institution.
- **Location:** Includes both state and city to analyze geographical distribution.
- **Ownership Type:** Identifies whether the college is government-funded or privately owned.
- **Year of Establishment:** Indicates the college's legacy and experience in the education sector.

These details were extracted using BeautifulSoup, where specific HTML tags containing college information were parsed. The ownership type was retrieved from structured sections of the website where it was explicitly mentioned.

2. Collecting Course & Fee Information

Each college offers a variety of undergraduate (B.Tech), postgraduate (M.Tech), and diploma programs. The following course-related details were gathered:

- **Programs Offered:** A list of courses available, including specialized branches like Computer Science, Artificial Intelligence, Mechanical, Civil, and Electrical Engineering.
- **Specializations:** Many colleges offer niche specializations such as Cybersecurity, Data Science, and Robotics, which were recorded separately.
- **Fee Structure:**
    - Annual Fee: The tuition fee per year.
    - Total Course Fee: The estimated cost for the entire duration of the program.

To ensure consistency, fee values were extracted using regex to filter out currency symbols and standardize formats. Any missing values were replaced with "N/A" for further analysis.

3. Retrieving Facilities & Ratings

To provide students with an understanding of campus life, data on college infrastructure and student ratings was collected:

- Hostel Availability: Whether boys' and girls' hostels are provided.
- Laboratories & Research Centers: Colleges with modern research labs, computing centers, and innovation hubs were noted separately.
- Library Facilities: The size, digital access options, and research papers available in college libraries.

- Sports & Recreation: Facilities such as gymnasiums, sports complexes, auditoriums, and extracurricular clubs were included.
- Student Ratings: Extracted from college review sections, student testimonials, and official ranking reports.

Ratings were often expressed on a 5-star or 10-point scale, requiring data normalization for uniform comparisons.

4. Parsing Data with Regex

Since web scraping often results in unstructured text data, Regular Expressions (Regex) were used to extract and clean important information. This helped in:

- Filtering Numerical Data: Extracting fees from text fields with currency symbols (₹ or Lakhs).
- Identifying Course Names: Standardizing the format of course names for consistency.
- Extracting Establishment Years: Finding four-digit values (e.g., 1998, 2005) to identify college age.
- Cleaning Review Texts: Removing unwanted symbols and special characters from student feedback.

By using pandas and regex, the extracted data was cleaned, formatted, and structured before being saved in a CSV file for further processing.

Web Scraping Process

Web Scraping Using Python Libraries

To efficiently extract the data, the following tools and libraries were utilized:

• **BeautifulSoup & Requests:** For retrieving and parsing webpage content.

• **Pandas & NumPy:** For data structuring and preprocessing.

Steps in the Web Scraping Process:

1. **Sending Requests:** HTTP requests were sent to Booking.com to access hotel listing pages across the four cities, ensuring comprehensive data coverage.
2. **Parsing HTML Content:** BeautifulSoup was used to identify and extract relevant HTML tags and attributes containing hotel data.
3. **Extracting Data:** Parsed data was structured into DataFrames using Pandas. This step included cleaning and preprocessing the data to ensure consistency and accuracy.
4. **Saving Data:** The final dataset was saved in CSV format, enabling seamless integration with tools like Power BI for further analysis.

Challenges

While web scraping provided valuable data, certain challenges were encountered:

1. Incomplete Information: Not all colleges had complete information available online, resulting in missing data for some parameters.

2. Dynamic Web Content: Some websites used JavaScript to load content dynamically, requiring the use of browser automation tools instead of simple HTTP requests.

3. Rate Limiting: Some websites implemented rate limiting to prevent excessive scraping, necessitating the implementation of delays between requests.

4. Data Volume: Managing the large volume of data collected from thousands of colleges required efficient storage and processing systems.

These challenges were addressed through careful planning, robust error handling in the scraping scripts, and thorough data validation processes**.**

CHAPTER 3

DATA PREPROCESSING

The raw data collected through web scraping required extensive cleaning and preprocessing before it could be analyzed effectively. This preprocessing was primarily done using Power Query, with the following steps:

3.1 Cleaning and Formatting

To ensure data consistency and accuracy, several preprocessing steps were applied to clean and format the collected data. These steps included removing duplicate entries, standardizing values, and handling missing data to create a structured and reliable dataset for analysis.

Removing Duplicate Entries

During data collection, some colleges appeared multiple times due to different representations or repeated listings across sources. Duplicate entries could lead to biased analysis, so they were identified and removed using pandas. The dataset was checked based on unique attributes such as college name, location, and type (government/private) to avoid unintentional redundancy.

Standardizing Fees, Ratings, and Course Names

Colleges listed fee structures in different formats, such as annual fees vs. total course fees, making direct comparisons difficult. Similarly, ratings were recorded on different scales (5-star vs. 10-point rating systems), requiring normalization for consistency. Course names also varied across institutions, with abbreviations like "CSE" for Computer Science Engineering and different representations of specialized programs. To standardize the data:

- Fee structures were converted to a common format (e.g., annual tuition fees in INR).
- Ratings were adjusted to a uniform 10-point scale.

- Course names were mapped to standardized terminology using a dictionary-based approach and regex patterns.

Handling Missing Values Using Imputation Techniques

Missing values were addressed using appropriate data imputation techniques to avoid data loss while maintaining accuracy. Missing fees were estimated using the median values of similar institutions in the same region. Course availability and facilities data were cross-referenced with official college websites and AICTE records. In cases where ratings were missing, an average rating for colleges with similar attributes was assigned.

These cleaning and formatting steps helped enhance data integrity, making it more suitable for further analysis and visualization.

CHAPTER 4

DATA ANALYSIS AND VISUALIZATION

4.1 Overview of Analysis

The analysis examines college distribution, institution type, fees, courses, and infrastructure across India. Findings show private colleges dominate, Computer Science is the most popular course, and government institutions offer affordable alternatives. Power BI dashboards provide interactive insights, helping stakeholders make informed decisions on college selection and education planning.

The analysis of the preprocessed data focused on five key areas:

1. **Geographic Distribution Analysis:** Examining the spread of engineering colleges across states and regions
2. **Institutional Type Analysis:** Comparing government and private institutions on various parameters
3. **Course Offering Analysis:** Identifying patterns in course availability and specializations
4. **Fee Structure Analysis:** Understanding fee variations across college types, regions, and courses
5. **Infrastructure Assessment:** Evaluating the availability and quality of infrastructure facilities

Each analysis area employed appropriate statistical methods and visualization techniques to extract meaningful insights from the data.

4.2 Key Visualizations and Insights

4.2.1 Geographic Distribution of Engineering Colleges

This analysis explores the distribution of engineering colleges across different regions in India using choropleth maps, bar charts, and pie charts to highlight density variations.

- State-wise College Density: A choropleth map visualizes the number of colleges across India, showing Southern states (Tamil Nadu, Karnataka, and Andhra Pradesh) have the highest concentration of institutions, making them major education hubs.
- Regional Disparities: Northeastern states such as Arunachal Pradesh, Mizoram, and Nagaland have significantly fewer colleges, reflecting a lack of educational infrastructure in these regions.
- Urban vs. Rural Distribution: Urban centers like Bangalore, Chennai, Mumbai, and Delhi have a much higher density of engineering colleges compared to rural areas, primarily due to better infrastructure, job opportunities, and industrial presence.

4.2.2 Government vs. Private Institutions

A comparative analysis of public and private institutions using stacked bar charts and box plots provides insights into ownership patterns and fee structures.

- Ownership Distribution: A stacked bar chart shows that private colleges significantly outnumber government institutions in most states.
- Fee Structure Variations: Box plots reveal that government colleges generally have lower tuition fees, making them more accessible, while private institutions charge significantly higher fees.
- Infrastructure Comparison: Despite lower costs, government institutions often offer better infrastructure, research facilities, and faculty expertise, while private colleges focus on modern amenities and industry collaborations.

4.2.3 Course Offerings

Using heatmaps, treemaps, and trend analysis, this section examines the availability of engineering specializations across different states.

- Traditional vs. Emerging Courses: Traditional branches like Mechanical, Electrical, and Civil Engineering are widely available across most institutions.

- Rise of AI and Data Science: Newer specializations such as Artificial Intelligence, Data Science, and Robotics are predominantly offered in tier-1 cities, where demand and resources are higher.
- Regional Gaps: Some rural and tier-2 institutions lack specialized course offerings, limiting opportunities for students interested in emerging technologies.

4.2.4 Fee Structure Analysis

Using box plots, scatter plots, and bar charts, this section explores the variation in tuition fees across states and college types.

- State-wise Fee Disparities: Box plots highlight that fees vary by up to 300% for similar courses in different states.
- Impact of Location on Fees: Private institutions in metropolitan areas charge significantly higher tuition fees compared to those in smaller cities or towns.
- Specialized Courses Cost More: Scatter plots show that AI, Data Science, and other emerging fields typically command premium fees compared to traditional engineering disciplines.

4.2.5 Infrastructure Assessment

This section evaluates campus facilities using radar charts, stacked bar charts, and heatmaps to compare the quality of infrastructure across different institutions.

- Lab Facilities: Heatmaps indicate a lack of advanced laboratory infrastructure for specialized courses like Robotics and AI in many institutions.
- Hostel Availability: Stacked bar charts show that hostel accommodation is insufficient in many private colleges, often affecting student convenience.
- Digital Infrastructure: Radar charts reveal significant variation in internet access, e-libraries, and digital resources across different regions, impacting learning experiences.

.

4.3 Interactive Dashboard Highlights

An interactive dashboard was developed to allow stakeholders to explore the findings dynamically. Key features of the dashboard include:

1. **College Explorer:** Interactive map interface allowing users to filter colleges by state, type, courses offered, and fee range
2. **Course Comparison Tool:** Side-by-side comparison of course offerings, fees, and infrastructure across selected colleges
3. **Fee Analyzer:** Dynamic visualizations showing fee distributions and allowing comparison across different parameters
4. **Infrastructure Evaluator:** Visual representation of infrastructure facilities with filtering capabilities
5. **Regional Insights Panel:** Summarized statistics for selected regions or states

The dashboard incorporates slicers, filters, and drill-down capabilities to enable users to navigate from high-level overviews to detailed information about specific colleges or courses.

CHAPTER 5

TOOLS AND TECHNOLOGIES

5.1 Software Used

To ensure efficient data collection, preprocessing, analysis, and visualization, multiple software tools were utilized throughout the project. These tools enabled the seamless handling of large datasets, effective data transformation, and interactive visualization of key insights.

- **Python (Jupyter Notebook):** Used for web scraping, data cleaning, and analysis in an interactive coding environment.
- **Power BI:** Enabled data visualization and interactive dashboards for insights on college distribution, fees, and course trends.
- **Excel:** Assisted in data validation, formatting, and verification, ensuring accuracy before final analysis.

5.2 Python Libraries Used

Several Python libraries were employed to streamline data processing, optimize performance, and enhance the accuracy of extracted information.

- **pandas:** Used for data cleaning, structuring, and manipulation, including handling missing values and exporting datasets.
- **numpy:** Assisted in numerical operations, statistical computations, and data normalization.
- **BeautifulSoup:** Enabled web scraping to extract structured data from Careers360 and other educational portals.

By leveraging these tools and libraries, the project ensured efficient data collection, structured preprocessing, and insightful analysis, ultimately enhancing the decision-making process for students, educators, and policymakers.

5.3 Web Scraping Tools and Technologies

• **Proxy Rotation & User-Agent Spoofing:** To avoid detection and blocking by web servers.

• **Retry Mechanisms:** Implemented to handle connection errors and timeouts during scraping.

5.4 Power BI Features Utilized

Power BI was extensively used to create interactive dashboards with the following features:

• **Data Modeling:** Creating relationships between datasets to generate meaningful insights.

• **DAX (Data Analysis Expressions):** Used for creating calculated fields and aggregating measures.

• **Drill-Through and Filter Options:** Allowed users to explore detailed insights about engineering colleges.

• **Custom Visuals:** Implemented bar charts, pie charts, line graphs, and maps to display trends effectively.

5.5 Integration of Tools

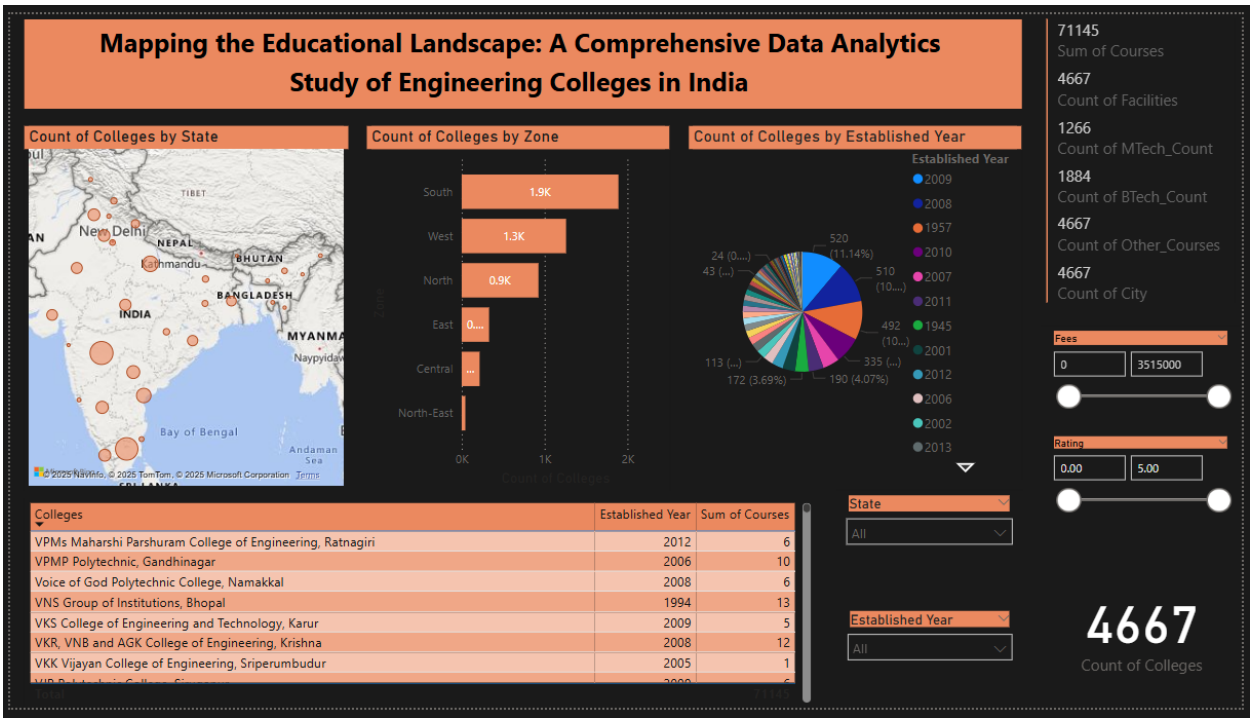
The overall workflow of the project was structured as follows:

1. **Data Collection:** Web scraping using Python libraries.
2. **Data Cleaning:** Performed using Pandas, NumPy, and Power Query.
3. **Data Storage:** Cleaned data stored in CSV/Excel format.
4. **Data Visualization:** Implemented in Power BI with interactive dashboards.

CHAPTER 6

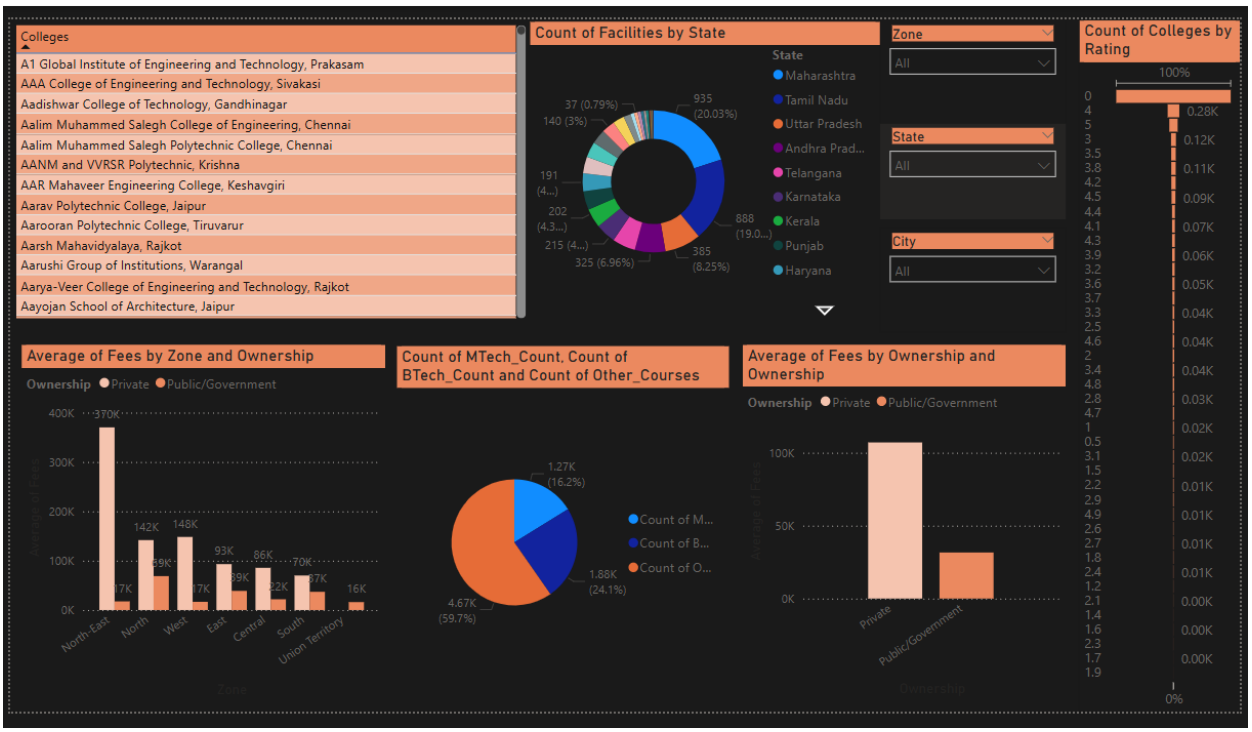
DASH BOARDS

To enhance the interpretability of the collected data, interactive visualizations were developed using Power BI. These visualizations allowed for an in-depth understanding of college distribution, fee structures, course availability, and institutional ratings. The primary objective of these visualizations was to present the data in a user-friendly manner, enabling stakeholders like students, educators, and policymakers to analyze trends efficiently.



- **Interactive Maps** – Show the geographical distribution of engineering colleges across India.
- **Comparative Charts** – Analyze trends in government vs. private colleges, tuition fees, and course popularity.
- **Dynamic Filters** – Allow users to customize searches based on location, affordability, and ratings.
- **Infrastructure Insights** – Evaluate facilities, student satisfaction, and academic quality.

The Power BI dashboards provide interactive insights into key aspects of engineering education in India:



 College Distribution Map – Displays the number of colleges across states.

 Government vs. Private Colleges – Compares institution types and affordability.

 Course Popularity Trends – Highlights the demand for Computer Science, AI, and traditional branches.

 Fee Structure Analysis – Examines tuition costs across states and institutions.

 Infrastructure & Ratings – Evaluates facilities, placements, and student satisfaction.

These visualizations make data-driven college selection easier, helping students and policymakers assess institutions effectively.

CHAPTER 7

CONCLUSION

The distribution of engineering colleges in India, fees, facilities, and student evaluations are the main topics of this study. Data was gathered, cleaned, and organized for analysis using web scraping from Careers360. Interactive insights on government versus private institutions, course popularity, fees and infrastructure quality are presented by Power BI dashboards. Through data-driven evaluations of colleges and universities, this project empowers educators, students, and policymakers to make well-informed decisions.

7.1 Key Findings

- **Private Colleges Dominate:** Private engineering colleges dominate government institutions, offering modern infrastructure and industry interactions but at higher education fees, while government colleges provide affordable, high-quality education.
- **Rising Demand for Computer Science:** Computer Science Engineering (CSE) is the most sought-after field, with growing interest in AI, Data Science, and Cybersecurity, while traditional branches like Mechanical and Civil Engineering see declining enrollments.
- **Affordability Matters:** Fee comparisons play a crucial role in college selection, as students weigh the trade-offs between cost, infrastructure, and educational quality.
- **Geographical Variations:** Engineering colleges are concentrated in metro cities like Bangalore, Delhi, and Mumbai, but regional institutions serve local education needs.
- **Power BI Dashboards for Decision-Making:** Interactive dashboards allow users to explore fees, ratings, and course availability, enabling data-driven college selection based on preferences and budget.

7.2 Summary

This project provides a data-driven analysis of engineering colleges in India, focusing on college distribution, fees, courses, infrastructure, and student ratings. Using web scraping, data cleaning, and Power BI visualization, it highlights trends in private vs. government institutions, course demand, and affordability. The interactive dashboard enables stakeholders to compare colleges and make informed decisions. Future improvements, such as placement records and student satisfaction metrics, can further enhance the evaluation of higher education institutions.