Exploratory Data Analysis on

National Institutional Ranking Framework

by

Group 3

Sobhan Behuria
ID: 202318040
Course: MSc(DS)

Taruna Sagar Mati
ID: 202318045
Course: MSc(DS)

Srinibas Masanta
ID: 202318054
Course: MSc(DS)

Course Code: IT 462
Semester: Winter 2023

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

May 2, 2024

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout my project titled “Exploratory Data Analysis on National Instituitional Ranking Framework.” Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am incredibly fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

Furthermore, I would like to thank the entire team at Dhirubhai Ambani Institute of Information and Communication Technology for fostering an environment of collaboration and innovation. The resources and facilities provided have been crucial in conducting comprehensive research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience. I am confident that the knowledge and skills acquired during this endeavor will be a solid foundation for my future endeavors.

Sincerely,

Sobhan Behuria, 202318040
Taruna Sagar Mati, 202318045
Srinibas Masanta, 202318054

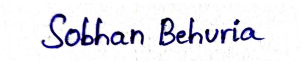
DECLARATION

We, [202318040, 202318045, 202318054] now declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data used in this project is obtained from the NIRF website. We also declare that we have adhered to the terms and conditions mentioned on the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

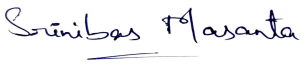
We acknowledge that we have received no external help or assistance in conducting this project except for the guidance provided by our mentor, Prof. Gopinath Panda. We declare no conflict of interest in conducting this EDA project.

We have now signed the declaration statement and confirmed the submission of this report on .... April 2024.



Sobhan Behuria
ID: 202318040
Course: MSc(DS)

Taruna Sagar Mati
ID: 202318045
Course: MSc(DS)



Srinibas Masanta
ID: 202318054
Course: MSc(DS)

CERTIFICATE

This is to certify that Group 3 comprising Sobhan Behuria, Taruna Sagar Mati, Srinibas Masanta has completed an exploratory data analysis (EDA) project on National Institutional Ranking Framework, which was obtained from NIRF website.

The EDA project presented by Group 3 is their original work. It was completed under the guidance of the course instructor, Prof. Gopinath Panda, who provided support and guidance throughout the project. The project is based on a thorough analysis of the PROJECT dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on , which demonstrates the analytical skills and knowledge of the students of Group 3 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

May 2, 2024

Contents

List of Figures 4

1 Introduction 1
- 1.1 Project idea . . . 1
- 1.2 Data Collection . . . 1
- 1.3 Dataset Description . . . 1
- 1.4 Packages required . . . 2

2 Data Pre-Processing 4
- 2.1 Data transformation: . . . 4
- 2.2 Data Characteristics: . . . 4
- 2.3 Challenges: . . . 4
- 2.4 Dataset Cleaning: . . . 5

3 Analysis and Visualization 6
- 3.1 Analysis of placement statistics in the institutes . . . 6
- 3.2 Analysis of the median salary for placed graduates among the top 5 institutes . . . 6
- 3.3 Analysis of Number of Faculties in Different Institutes . . . 7
- 3.4 Analysis of Total Amount Received by Institute Across Years . . . 7
- 3.5 Analysis of Total Students by Program . . . 7
- 3.6 Analysis on Distribution of male and female students within each institute . . . 8

4 Feature Engineering 9
- 4.1 Feature Scaling . . . 9

5 Model fitting 10
- 5.1 Regression . . . 10
- 5.2 Machine Learning Algorithms . . . 10
  - 5.2.1 Linear Regression . . . 10
  - 5.2.2 Random Forest Regression . . . 11

6 Conclusion & future scope 12
- 6.1 Findings . . . 12
- 6.2 Challenges . . . 12
- 6.3 Future plan . . . 12

List of Figures

1 Total Students (Institute of Chemical Technology) 6
2 Total Students (Jamia Millia Islamia) 6
3 Total Students (University of Delhi) 7
4 Placement Statistics 2019 7
5 Placement Statistics 2020 8
6 Placement Statistics 2021 8
7 Sponsorship Amount 9
8 Distribution of Students (Institute of Chemical Technology) 10
9 Distribution of Students (Jamia Millia Islamia) 10
10 Distribution of Students (University of Delhi) 11
11 Median Salary of Top 5 Institutes 11
12 Number of Faculty in Institutes 12
13 Heatmap of 2019 Features vs Rank 12
14 Heatmap of 2020 Features vs Rank 13
15 Heatmap of 2021 Features vs Rank 13
16 Relation Between Features 2019 14
17 Relation Between Features 2019 14
18 Relation Between Features 2020 15
19 Relation Between Features 2020 15
20 Relation Between Features 2021 16
21 Relation Between Features 2021 16
22 Actual vs Predicted 2019 17
23 Actual vs Predicted 2020 17
24 Actual vs Predicted 2021 17

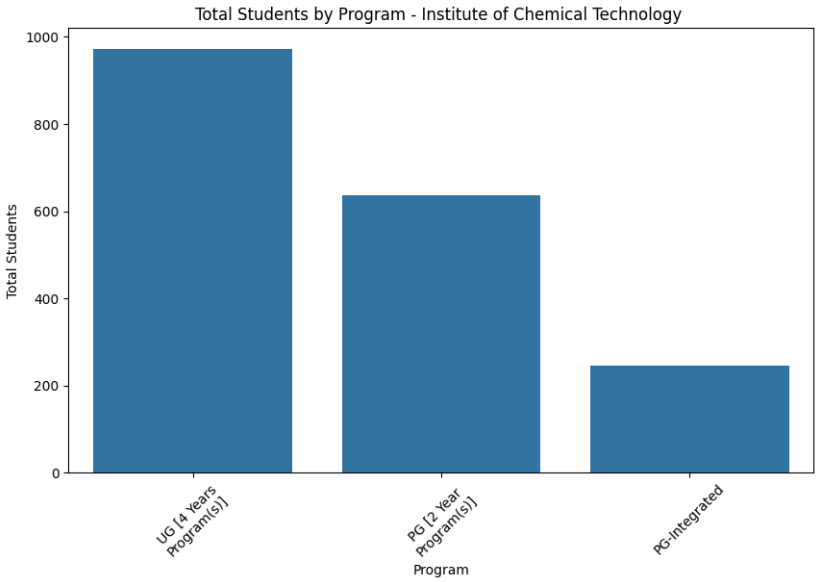


Figure 1: Total Students (Institute of Chemical Technology)

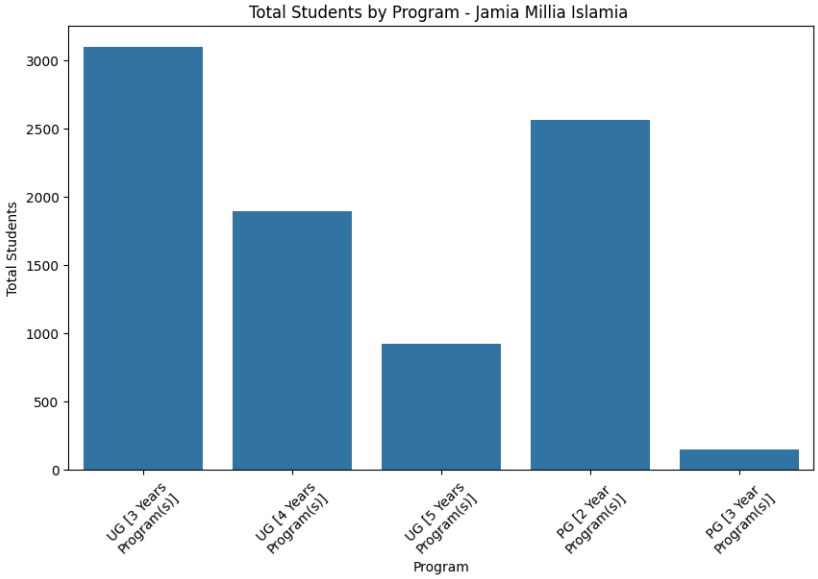


Figure 2: Total Students (Jamia Millia Islamia)

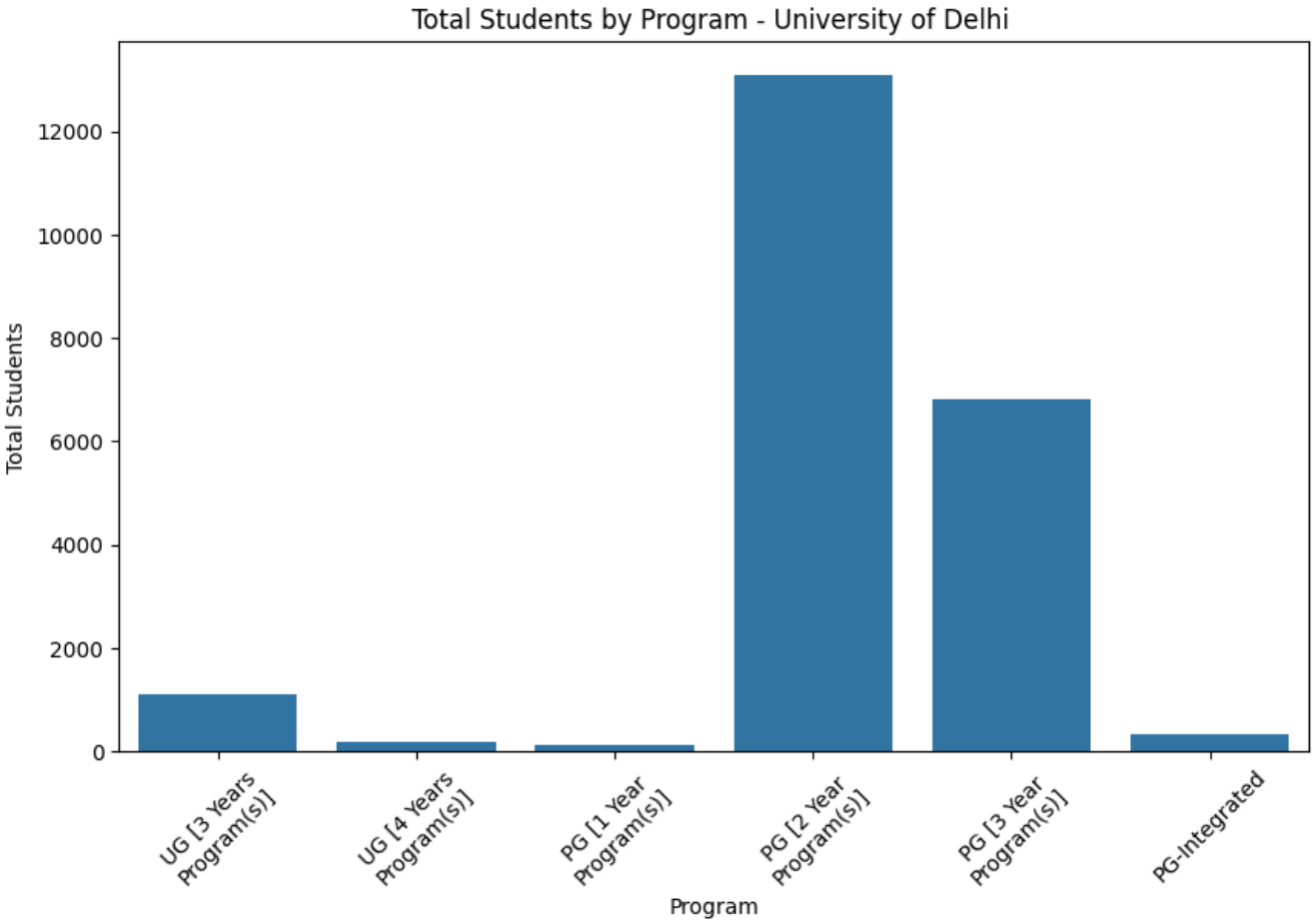


Figure 3: Total Students (University of Delhi)

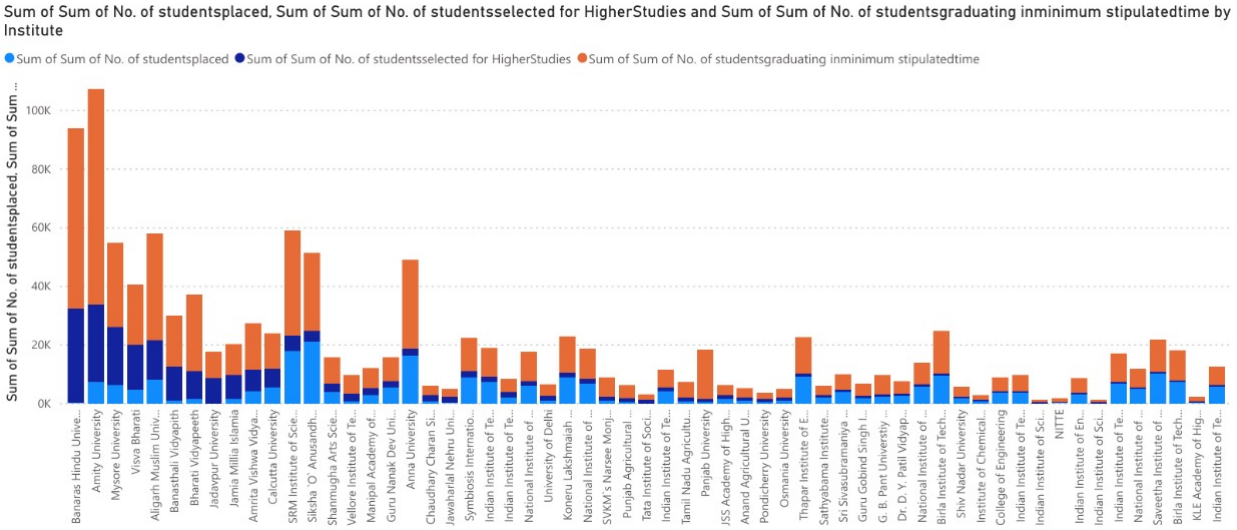


Figure 4: Placement Statistics 2019

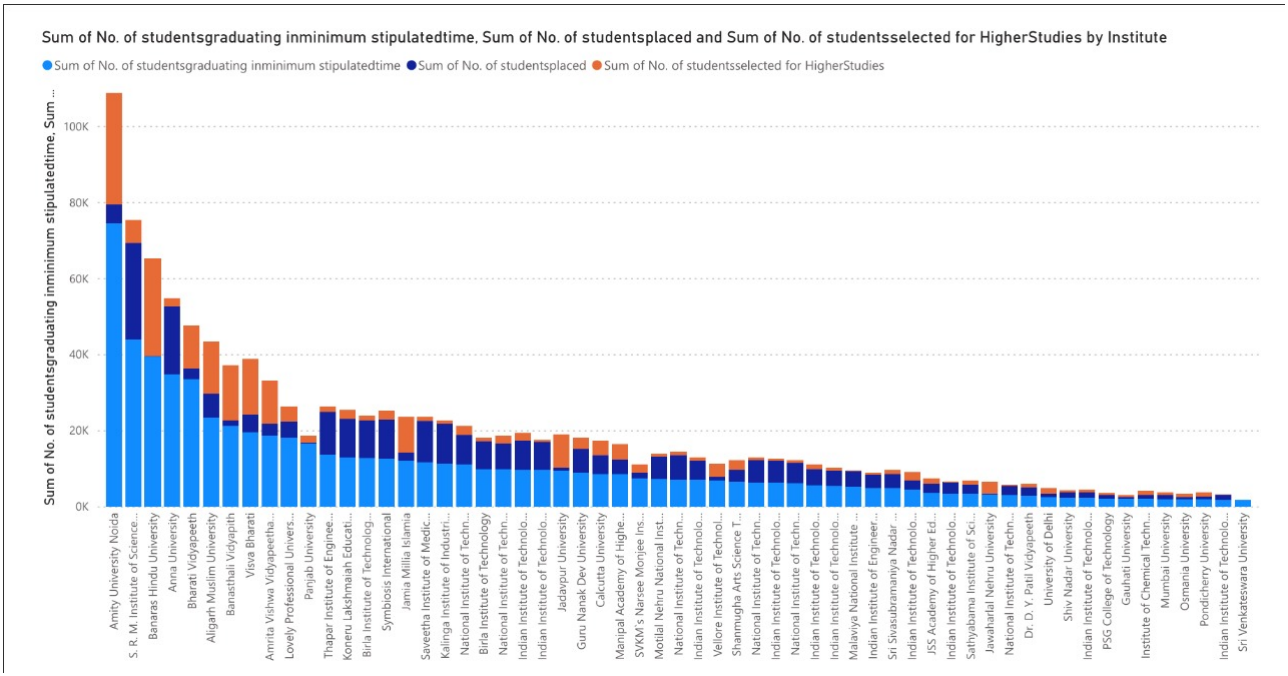


Figure 5: Placement Statistics 2020

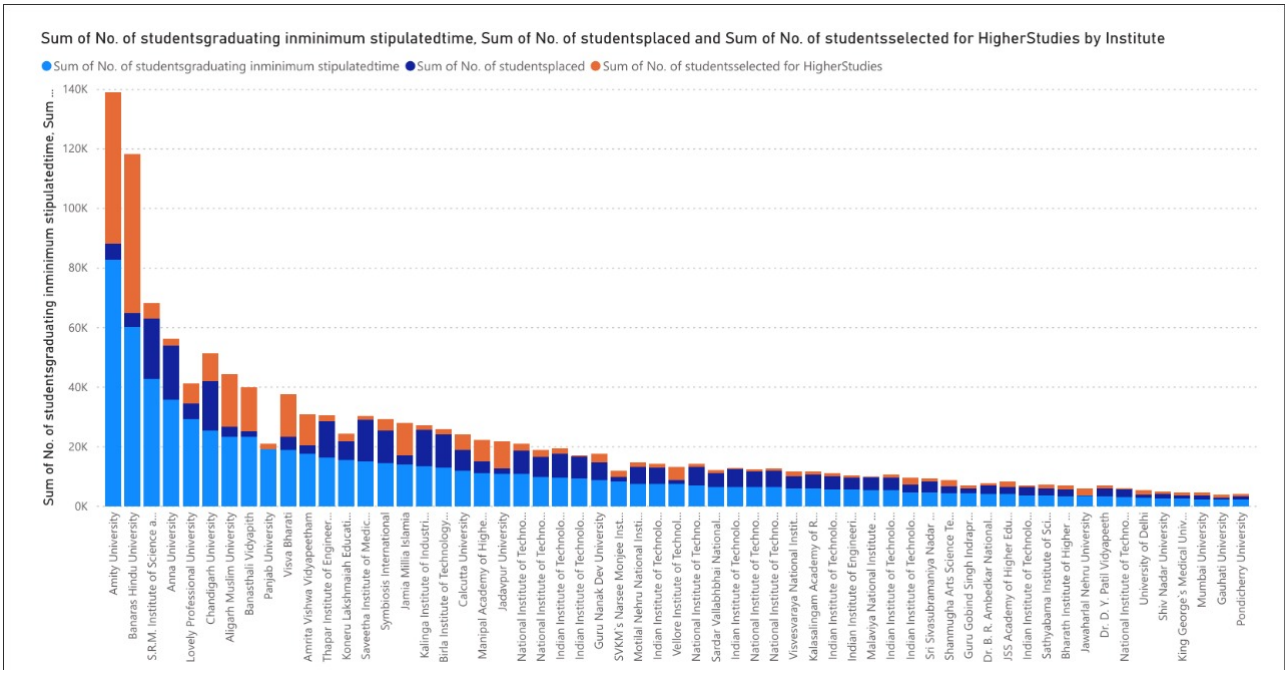


Figure 6: Placement Statistics 2021

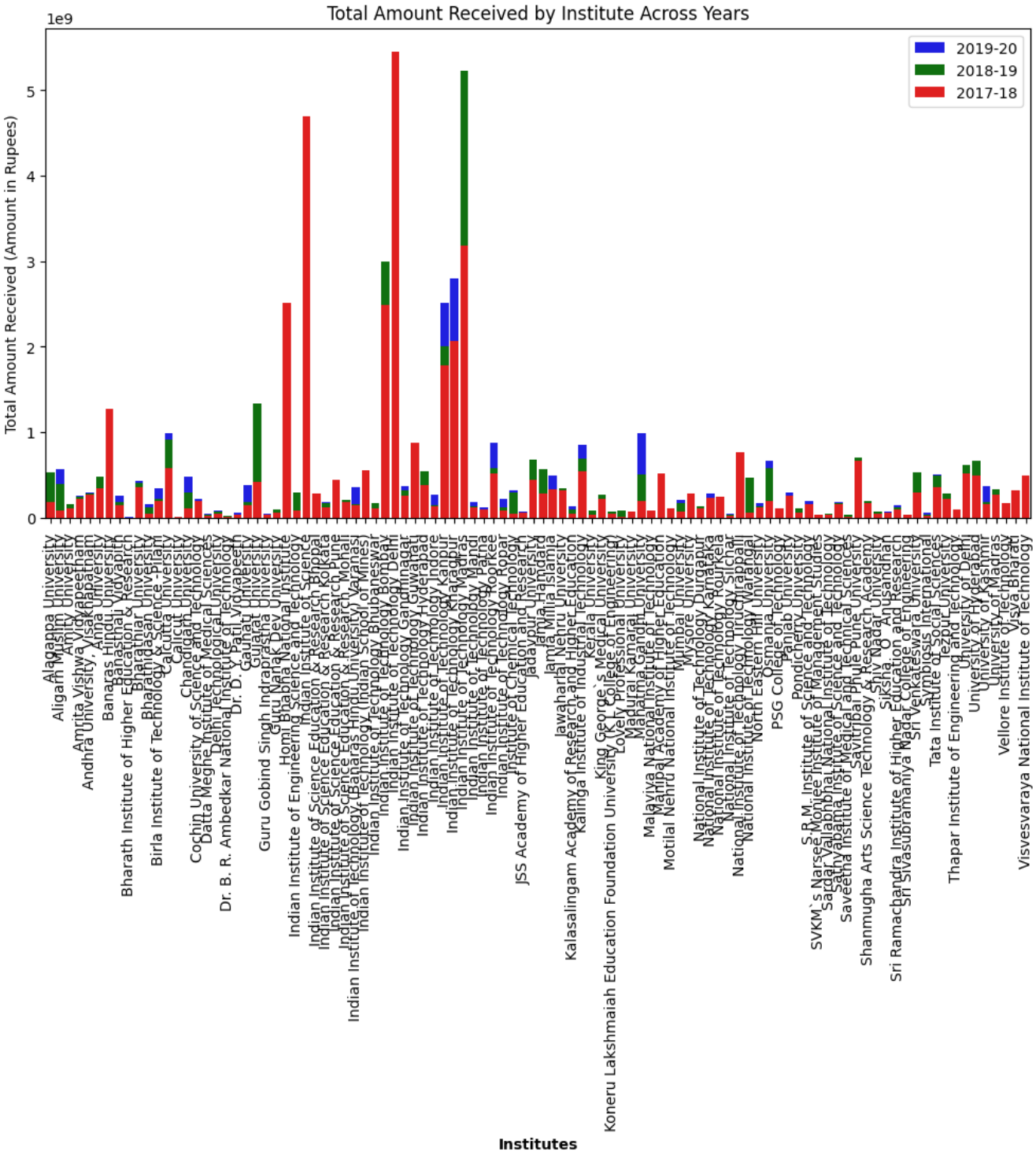


Figure 7: Sponsorship Amount

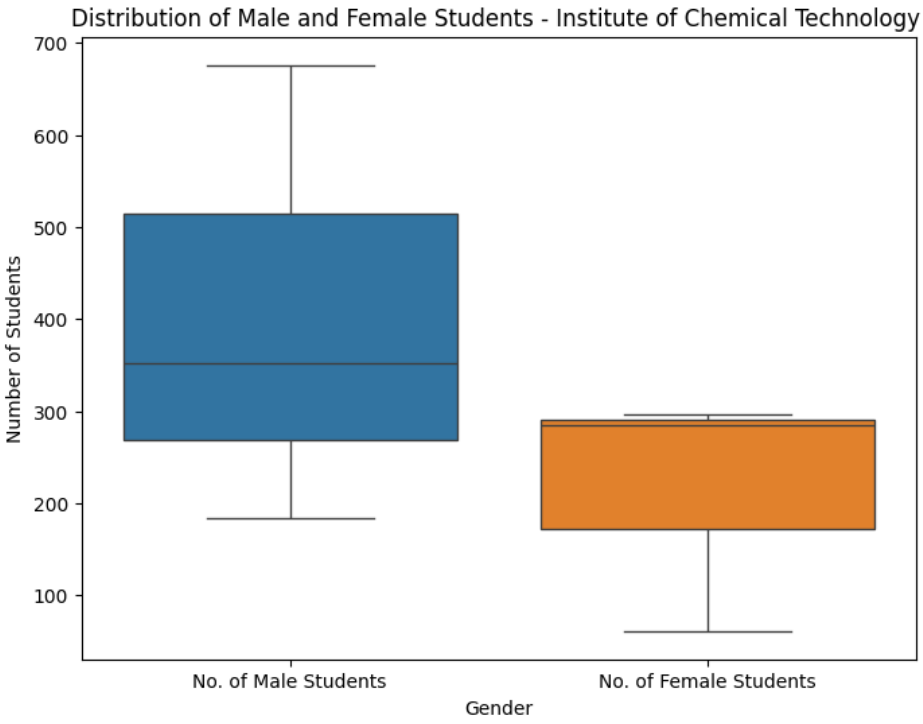


Figure 8: Distribution of Students (Institute of Chemical Technology)

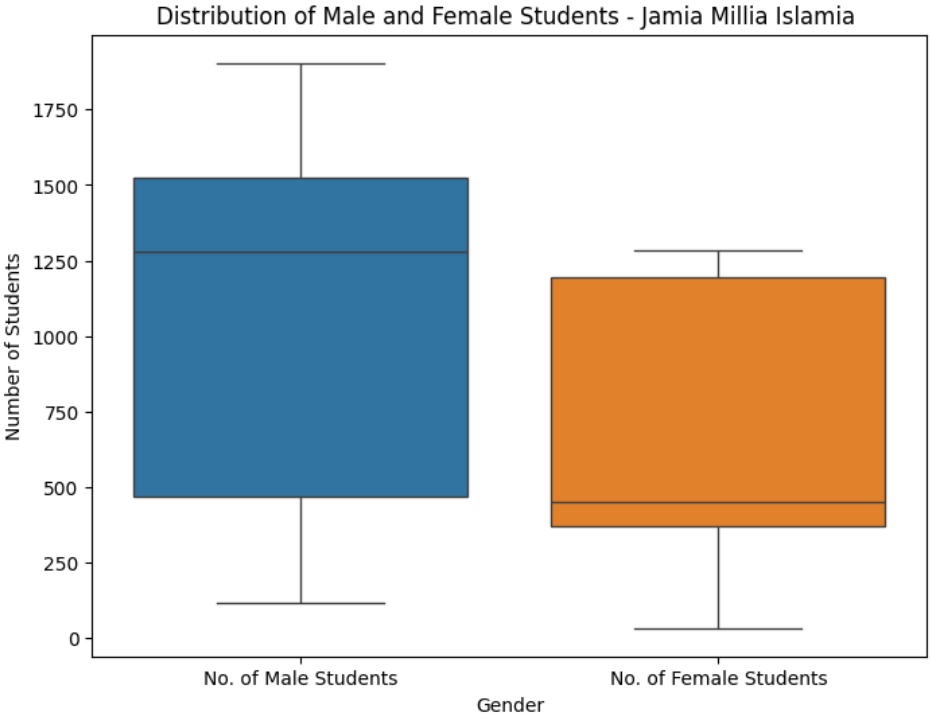


Figure 9: Distribution of Students (Jamia Millia Islamia)

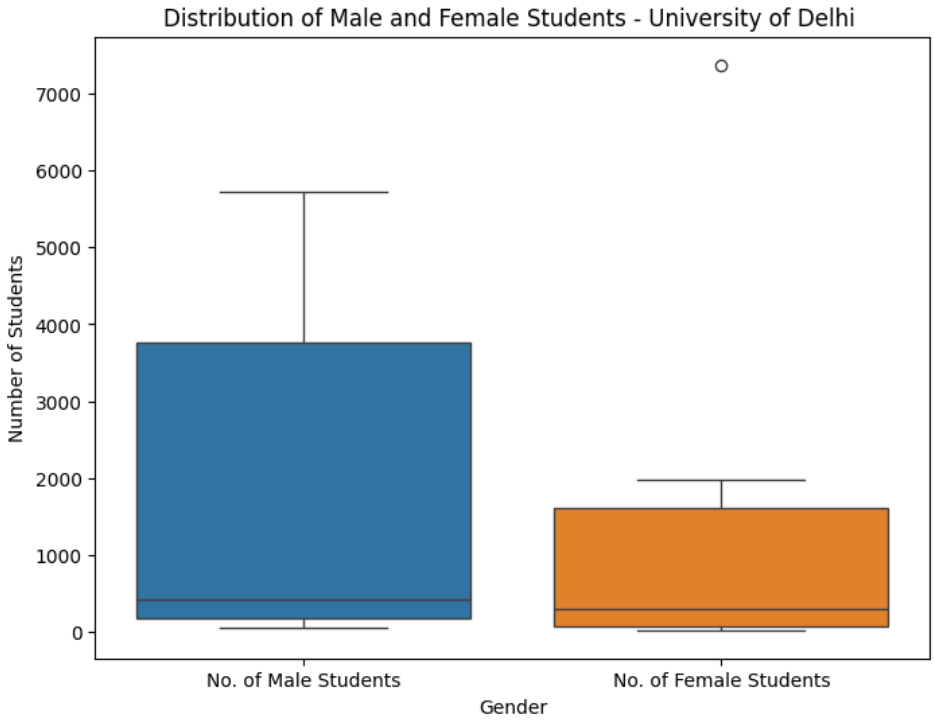


Figure 10: Distribution of Students (University of Delhi)

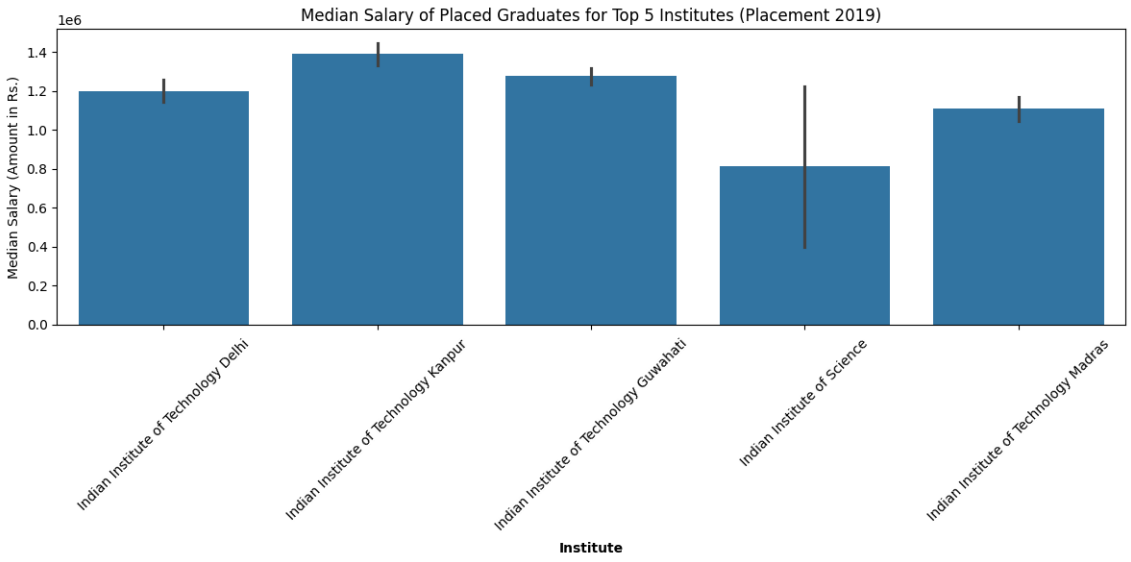


Figure 11: Median Salary of Top 5 Institutes

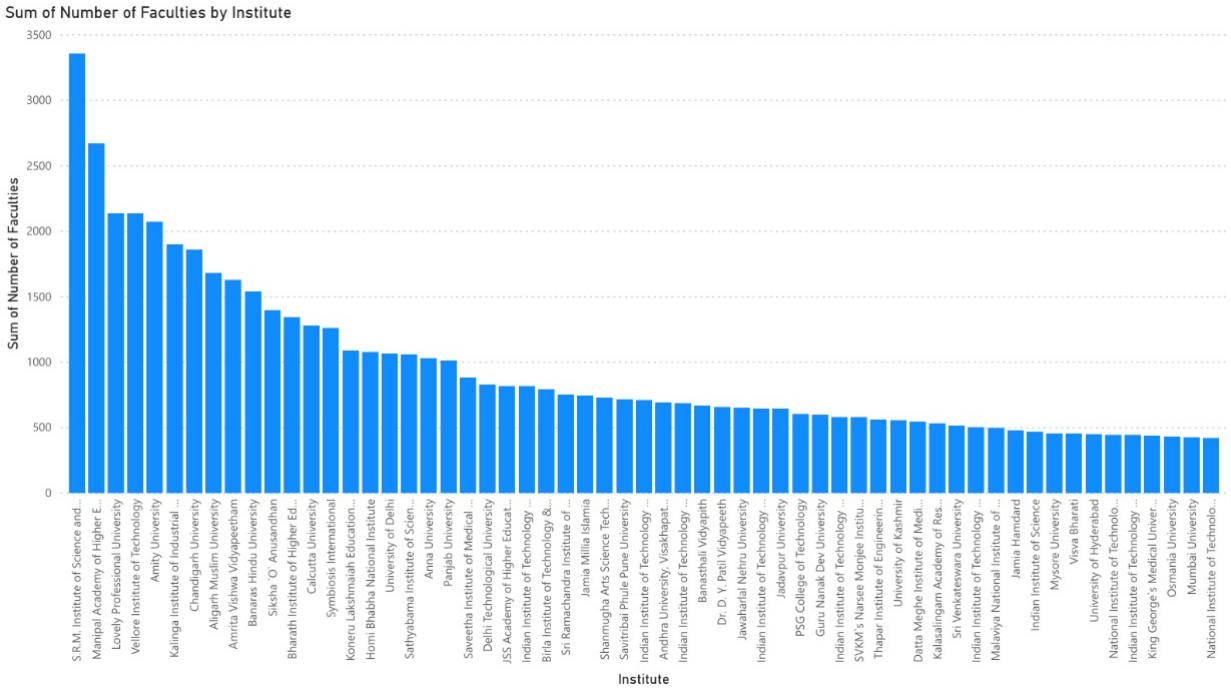


Figure 12: Number of Faculty in Institutes

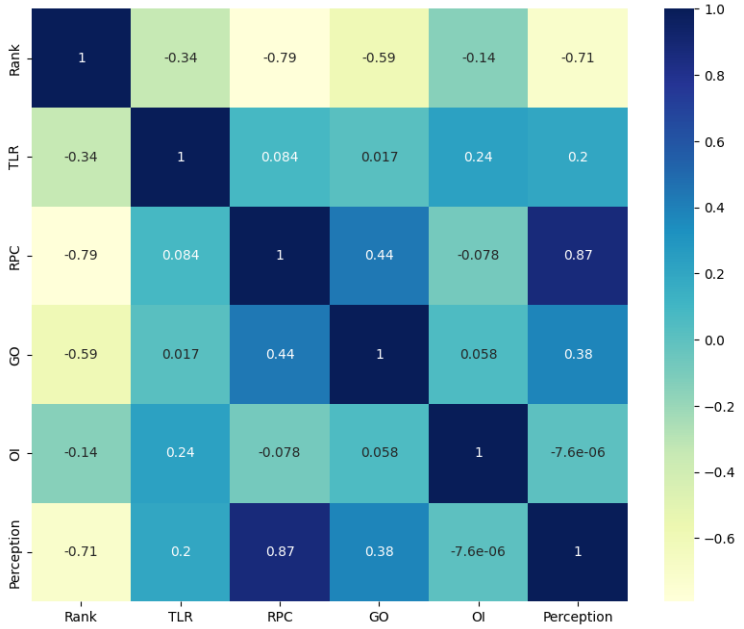


Figure 13: Heatmap of 2019 Features vs Rank

Figure 14: Heatmap of 2020 Features vs Rank

Figure 15: Heatmap of 2021 Features vs Rank

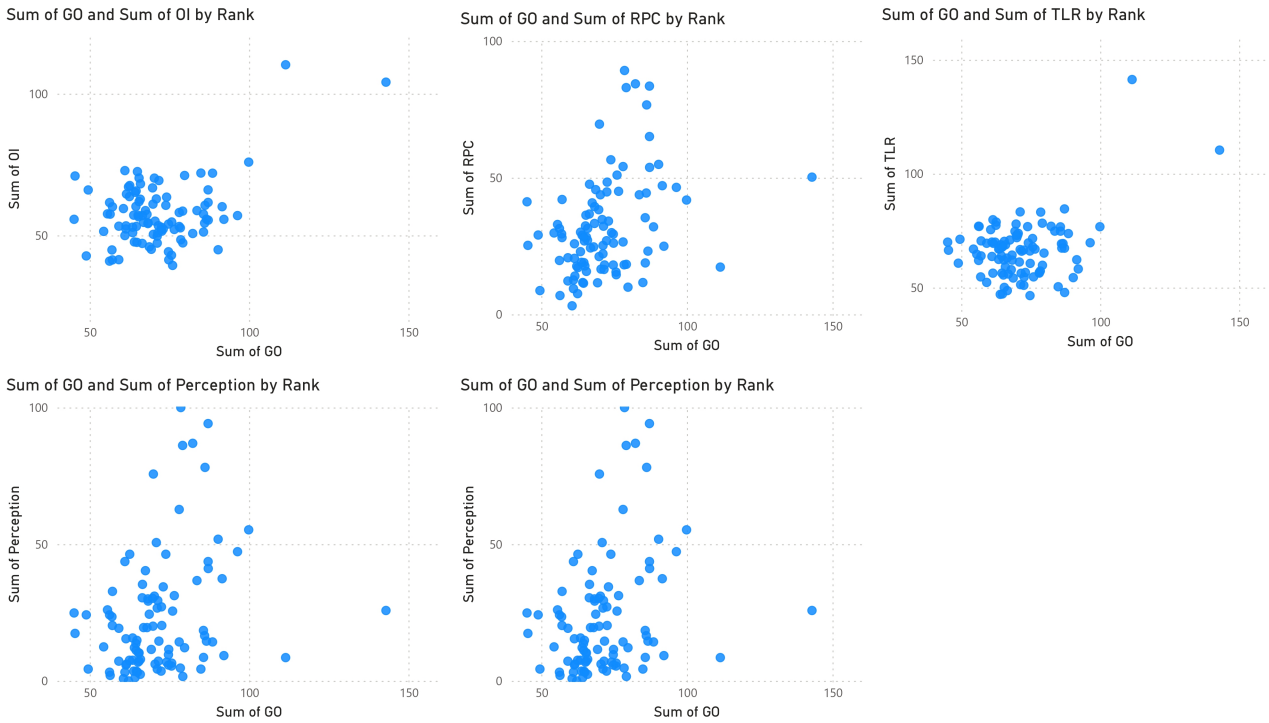


Figure 16: Relation Between Features 2019

Figure 17: Relation Between Features 2019

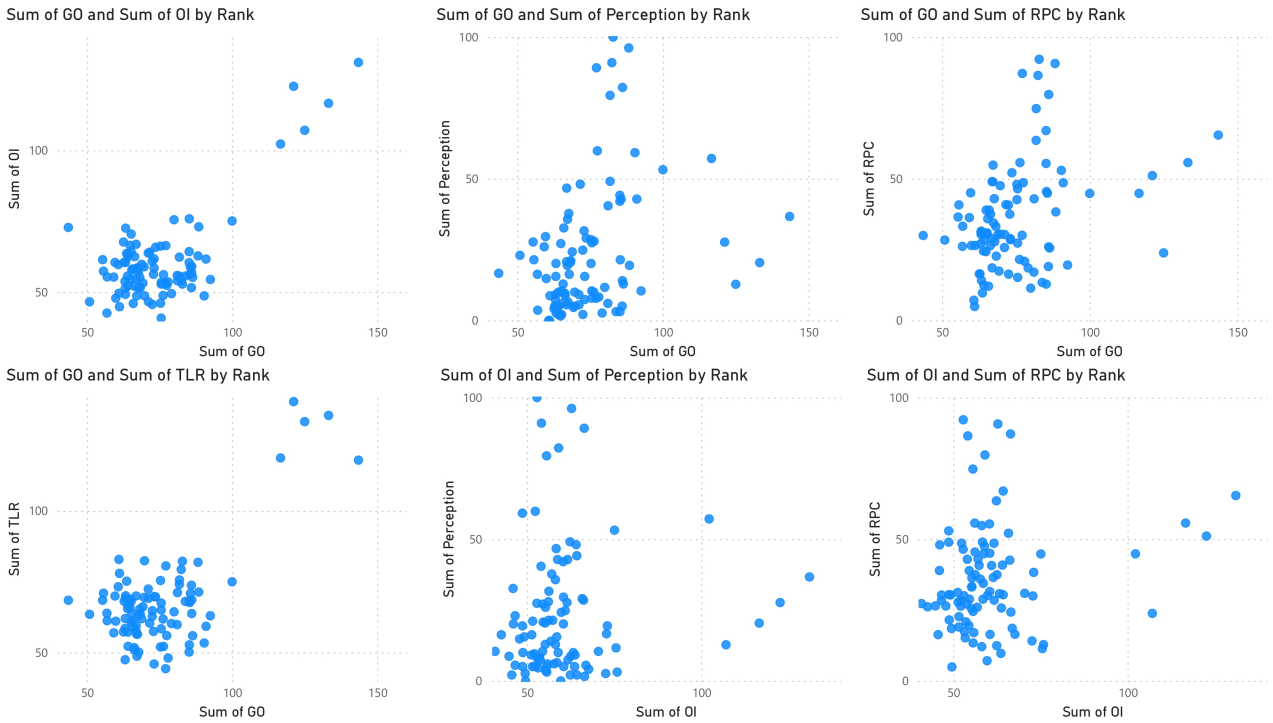


Figure 18: Relation Between Features 2020

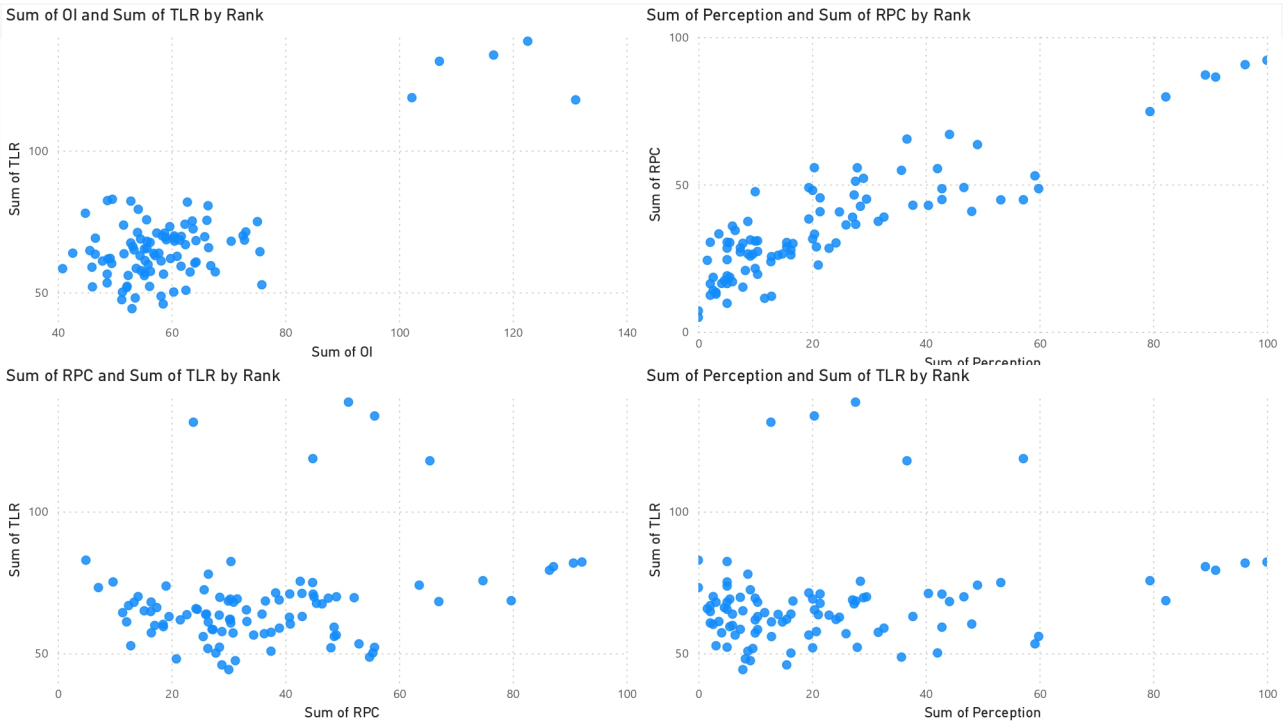


Figure 19: Relation Between Features 2020

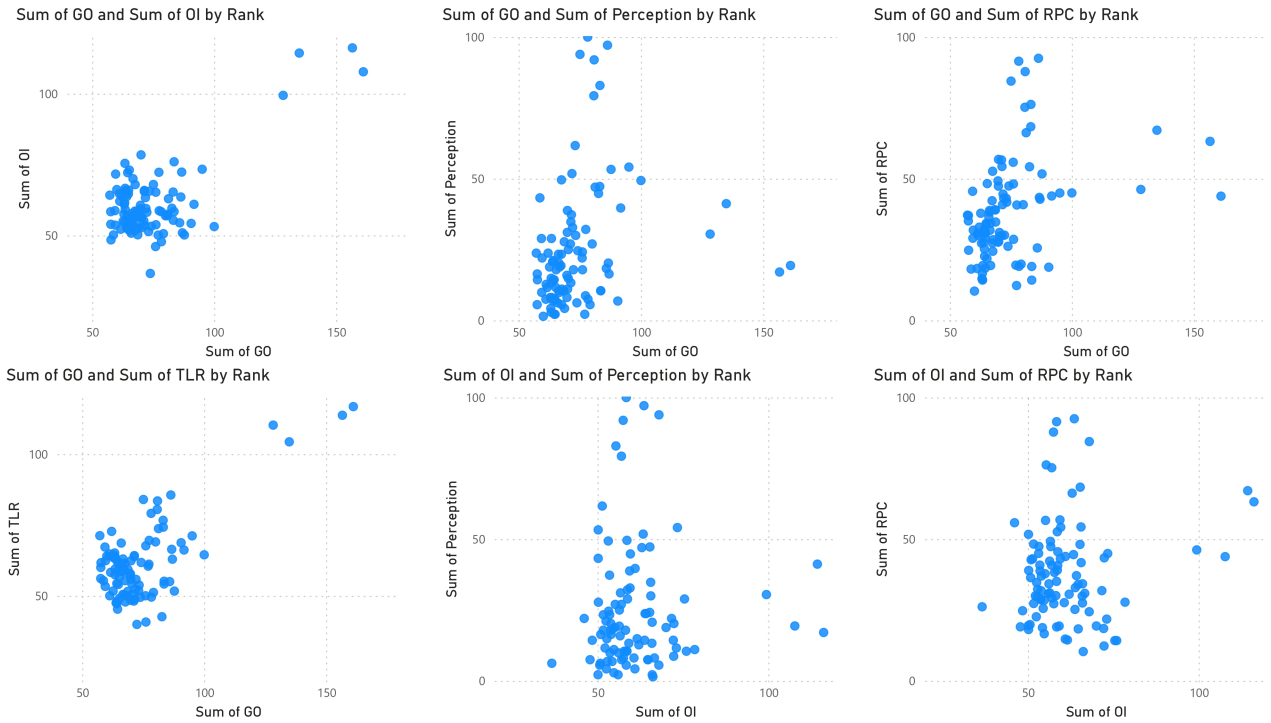


Figure 20: Relation Between Features 2021

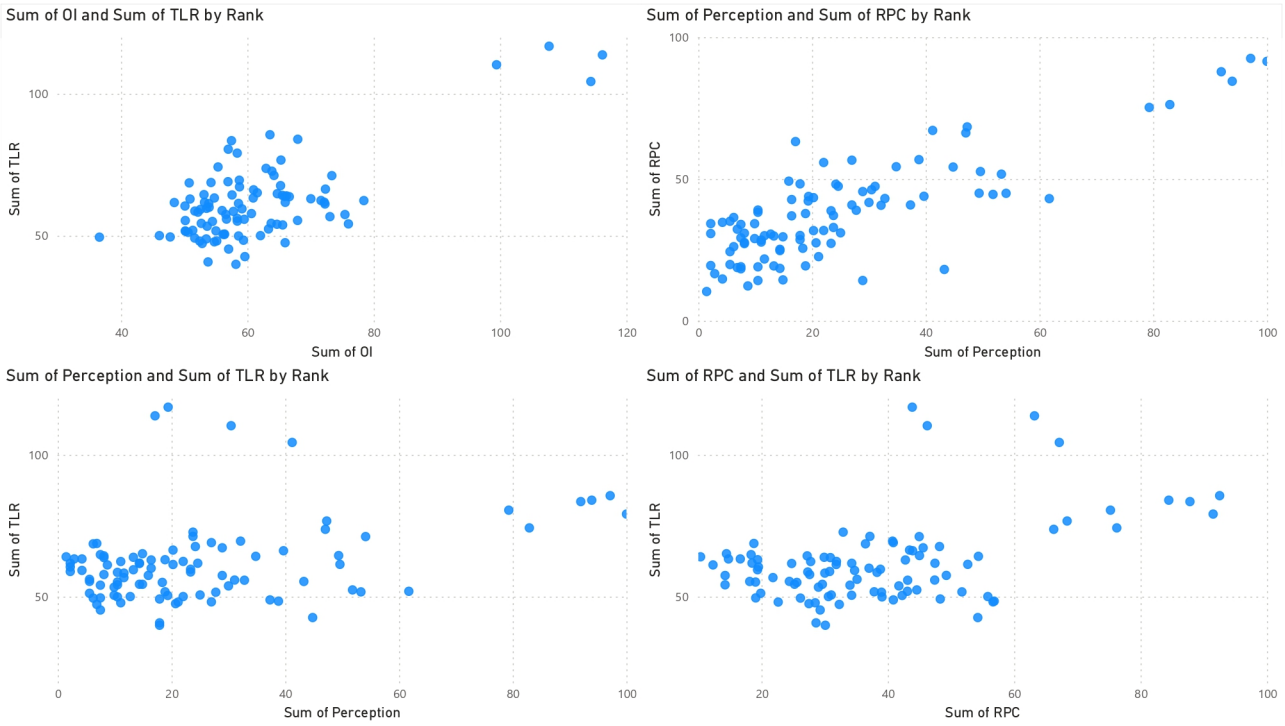


Figure 21: Relation Between Features 2021

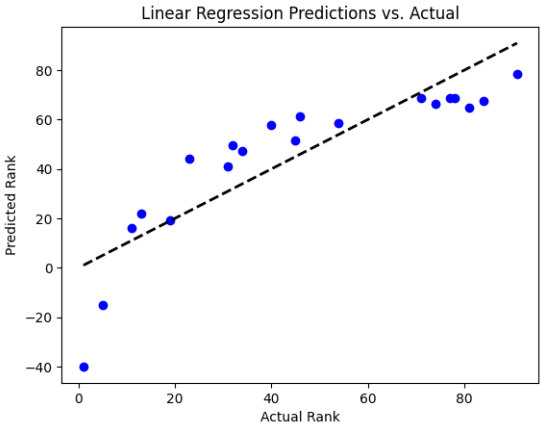


Figure 22: Actual vs Predicted 2019

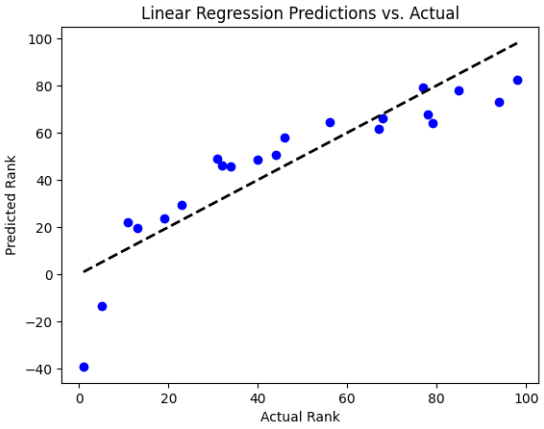


Figure 23: Actual vs Predicted 2020

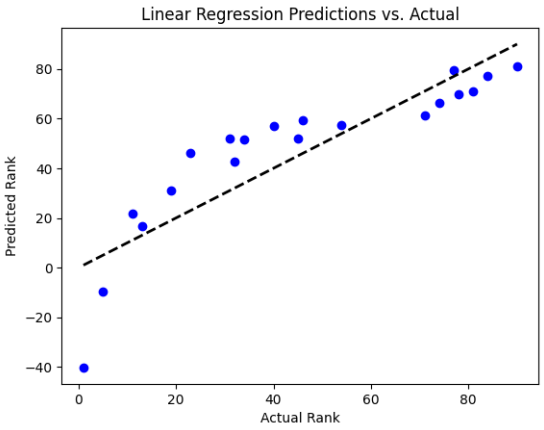


Figure 24: Actual vs Predicted 2021

List of Tables

1 Regression Results for 2019 18
2 Regression Results for 2020 18
3 Regression Results for 2021 18

Table 1: Regression Results for 2019

Metric	Linear Regression	Random Forest Regression
RMSE	15.4238	8.1627
R-squared	0.7031	0.9168

Table 2: Regression Results for 2020

Metric	Linear Regression	Random Forest Regression
RMSE	14.2601	8.1467
R-squared	0.7686	0.9245

Table 3: Regression Results for 2021

Metric	Linear Regression	Random Forest Regression
RMSE	15.1285	11.5530
R-squared	0.7128	0.8325

Abstract

The National Institutional Rankings Framework (NIRF) were established by the Ministry of Human Resources Development (MHRD) and were officially launched on September 29th, 2015, with the objectives of assess and rank higher education institutions across India. Although NIRF publishes much information regarding the rankings, it mainly exists in unstructured formats like PDF documents, limiting their usefulness for in-depth analysis and comparison across multiple institutes. This project aims to address this limitation by extracting data from NIRF's PDF documents and submitting it to various analytical methodologies. In addition to comparative analysis, this project strives to employ predictive model techniques to forecast the rankings of universities. By leveraging machine learn algorithms and historical data, the project seeks to develop predictive models that can estimate the future rankings of universities based on various performance metrics. By transforming unstructured data into analytically useful formats and predicting future rankings, this initiative aims to facilitate comprehensive and insightful comparisons among educational institutions, aiding stakeholders make informed decisions and fostering continuous improvements in the higher education sector.

Chapter 1. Introduction

1.1 Project idea

In our project, our primary objective is to perform comprehensive data analysis on information pertaining to various colleges listed on the National Institutional Ranking Framework (NIRF) over previous years. This involves extracting raw data from PDF documents and converting it into a CSV format as the initial step, streamlining the processing pipeline for ease of use. Furthermore, our intention is to leverage machine learning algorithms to forecast the future rankings of universities based on historical data trends. Through these efforts, we aim to gain deeper insights into the patterns within the data and present them visually through compelling graphical representations.

1.2 Data Collection

The dataset utilized in this project is sourced from the National Institutional Ranking Framework (NIRF), an initiative under the Ministry of Education, Government of India. This dataset, which includes NIRF rankings for the years 2019, 2020, and 2021, comprises PDF files corresponding to individual institutes. Leveraging web scraping techniques, we extracted and preprocessed PDF files for analysis. Each PDF file contains comprehensive information about the respective institute, encompassing various aspects such as sanctioned intake, student demographics, placement statistics, financial resources, research details, and faculty information. These PDF files serve as the primary data source for our analysis, enabling us to explore the performance and characteristics of higher education institutions within our country.

1.3 Dataset Description

- Sanctioned Intake: Each PDF file provides an extensive breakdown of the sanctioned intake for various academic programs offered by the institute. This encompasses the approved capacity for different programs over multiple years, serving as a foundational metric for understanding institutional capacity.
- Student Enrollment Statistics: The dataset offers comprehensive insights into student demographics, encompassing gender distribution, state-wise representation, and socio-economic backgrounds. It also delineates specific student categories, such as recipients of tuition fee reimbursements and individuals facing social challenges, fostering a nuanced understanding of inclusivity initiatives.

- Placement and Higher Studies Data: Detailed tables within the dataset elucidate the performance of graduating students, covering metrics like placement rates, pursuit of higher education, and median salary received. These metrics serve as indicators of an institute's effectiveness in preparing students for professional endeavors and academic pursuits.
- Research Endeavors: The dataset delves into the research ecosystem within institutes, encompassing parameters such as the number of full-time and part-time Ph.D. students, funding received for research endeavors, and collaborative engagements with external stakeholders.
- Executive Development Programs: Insights into professional development initiatives are provided through dedicated sections on executive development and management training programs. These segments offer details on program offerings, student enrollment, and financial resources allocated for such initiatives.
- Facilities for Physically Challenged Students: Institutional efforts towards fostering an inclusive learning environment are highlighted through sections detailing facilities for physically challenged students. Descriptions of facilities such as lifts, ramps, wheelchairs, and accessible toilets underscore institutional commitment to equitable access to educational opportunities.
- Faculty Profiles: The dataset includes comprehensive profiles of faculty members, encompassing information on their qualifications, areas of expertise, and institutional affiliations. This provides a holistic view of the academic talent pool within institutions, aiding in understanding institutional strengths and areas for improvement.

The dataset exhibits inherent variability across PDF files due to differences in the number of programs offered and reporting practices. Therefore, robust data preprocessing and analysis techniques are imperative to navigate the nuanced landscape of the NIRF dataset and derive meaningful insights for informed decision-making within the higher education sector.

1.4 Packages required

- BeautifulSoup: Utilized to perform web scraping and parse HTML and XML data structures, facilitating the extraction of text and metadata from the NIRF website's HTML framework.
- Mμ. : Used to navigate and extract text from PDF files efficiently, which enables streamline data extraction from the PDF documents containing NIRF rankings.
- Pandas: Acted as the primary library for data analysis and manipulation, providing robust tools for handling structured data, conducting exploratory data analysis (EDA), and preparing data for further processing.
- scikit-learn: Leveraged for implementing machine learning algorithms and predictive modeling techniques, allowing the development of models for ranking prediction based on historical NIRF data. Also, the 'MinMaxScaler' was employed for feature scaling, ensuring that all features were on a similar scale for optimal model performance.
- Matplotlib and Seaborn: Utilized for data visualization purposes, offering a diverse range of plotting functions and customization options to create informative and visually compelling graphs,

charts, and heatmaps. These aids in interpreting and effectively communicating insights derived from the NIRF dataset.

- Camelot: Employed to extract tabular data from PDF files, allowing for the extraction of structured information from the NIRF PDF documents. Camelot facilitates the conversion of PDF tables into usable data formats, aiding in the preprocessing and analysis of the NIRF dataset.

Chapter 2. Data Pre-Processing

2.1 Data transformation:

The process of data transformation presented a notable challenge due to the raw data's unstructured nature, originating from the official NIRF website, primarily in PDF format. To overcome this obstacle, we utilized a combination of Python packages, such as MuPDF and Camelot, to extract pertinent information from the PDF documents. Through the functionalities offered by these packages, we successfully extracted common yet significant data from the top colleges over the past three years and stored it in CSV files. These CSV files formed the basis for subsequent data processing and analysis, allowing us to uncover valuable insights and trends within the National Institutional Ranking Framework dataset.

2.2 Data Characteristics:

The data retrieved from the NIRF webpage presents several distinctive features that influenced our approach to data analysis. Firstly, each college is represented by an individual PDF file, resulting in a collection of PDF files for each year of NIRF rankings. This structure required us to handle and process each PDF document individually to extract relevant data for analysis. Secondly, the number of tables within these PDF files varied significantly, ranging from 9 to 19 tables per document. For example, in the 2021 dataset, some PDFs contained as few as 9 tables, while others included more than 15 tables. To streamline the data extraction process, we conducted a detailed analysis of the PDFs to identify consistent tables across all documents. Subsequently, we prioritized the extraction of Table 1 from each college's PDF and consolidated this information into a single CSV file, along with their respective college names. This standardized format enabled us to conduct comparative analysis across colleges on various key metrics, ensuring consistency and coherence in our analysis. By extracting and organizing the data uniformly, we were able to perform comprehensive assessments and derive meaningful insights from the NIRF dataset.

2.3 Challenges:

The process of extracting data from the NIRF PDF files presented various hurdles that demanded creative solutions. Firstly, in the year 2021, encountering single-lined tables posed a significant challenge, as existing packages struggled to extract such tables directly in tabular format. To tackle this issue, we devised a workaround by reading texts from the page to extract these tables. Additionally, some tables were split across two pages, which existing packages failed to detect. To address this, we developed a custom parser by meticulously analyzing the data files to accurately identify and extract split

tables. Furthermore, the PDF files for each year exhibited different formats, necessitating the creation of separate parsers for each year to ensure precise extraction. Moreover, minor discrepancies in institute names, such as variations like BITS Pilani and Birla Institute of Technology and Science Pilani, demanded special handling to maintain consistency in data processing. Additionally, some cross-page tables had mismatching numbers due to the type of program offered by the institute, requiring careful consideration during data extraction. Finally, numerous edge cases emerged, necessitating manual intervention due to the lack of uniformity in the data format. Despite these challenges, our iterative approach and tailored solutions empowered us to surmount these obstacles and extract valuable insights from the NIRF dataset.

2.4 Dataset Cleaning:

Ensuring the cleanliness of the NIRF dataset was a pivotal step in its preparation for analysis, involving diverse processes aimed at enhancing data quality and consistency. One notable challenge involved the variation in spellings for college names across different years, necessitating manual intervention to enforce uniformity. Additionally, the dataset contained NULL values for certain tables, which were addressed by either filling them with zeros or disregarding them, depending on the specific analytical requirements. To uphold dataset consistency, colleges that did not consistently appear in the top rankings for all three years were omitted from the analysis. Furthermore, some college names exhibited unnecessary spaces at the beginning or end, which were rectified using Python's strip function to ensure precision and uniformity in the data. These meticulous data cleaning procedures played a vital role in ensuring the accuracy, coherence, and readiness of the dataset for comprehensive analysis and interpretation.

Chapter 3. Analysis and Visualization

3.1 Analysis of placement statistics in the institutes

To analyze the placement statistics for 2019,2020 and 2021 we started by combining data from each institute into a detailed summary that included the total number of students placed, the number of students who graduated within the allotted time, and the number of students who were chosen to continue their education. This required grouping the data by institute using Pandas and adding the pertinent columns, which produced insightful information about how well each institute performed in terms of student placements and academic advancement.

To create a grouped bar chart for visual aids, we decided to use PowerBI. The chart's bars, each representing a different institute, are colored differently to indicate how many students were placed, graduated in the allotted time, and were chosen to pursue further education. We made it easier to compare these metrics across institutions by placing these bars next to each other. The chart's x-axis lists the institutes' names, and the y-axis shows the corresponding number of students. We also added labels and a brief title that summarized the chart's contents to guarantee clarity. All things considered, this visualization offers an understandable and straightforward representation of placement statistics for different institutes in 2021, promoting thoughtful decision-making and comparative analysis in the higher education sector. See 7,8,8 for visualizations.

3.2 Analysis of the median salary for placed graduates among the top 5 institutes

In our analysis of the median salary for placed graduates among the top 5 institutes in 2019,2020 and 2021 our process commenced with the conversion of the 'Median salary of placed graduates' column into numeric values. This transformation involved eliminating non-numeric characters and converting the values to floats using Pandas. Subsequently, we computed the median salary for each institute and identified the top 5 institutes based on their median salaries. We then filtered the dataset to exclusively include data for these top 5 institutes.

We decided to use Seaborn and Matplotlib, here each bar in the plot corresponds to a specific institute, with the height of the bar reflecting the median salary of placed graduates in that institute. The x-axis presents the names of the institutes, while the y-axis denotes the median salary in rupees. To enhance clarity, we included labels, including a concise title delineating the chart's content. Additionally, we rotated the x-axis labels for improved readability. In essence, this visualization provides a lucid and intuitive depiction of the median salary of placed graduates for the top 5 institutes in 2021, facilitating comparative analysis and informed decision-making within the higher education sector. See 11 for more information.

3.3 Analysis of Number of Faculties in Different Institutes

We started with extracting information from every PDF using the MuPDF and Camelot Libraries of Python. MuPDF was used in opening and reading PDFs, while to extract table contents we used Camelot. Algorithm was designed in such a way that we extracted faculty statistics from each PDFs' last table. We utilized Power BI for creating a bar graph visualizing the faculties distribution across different institutes. To ensure the visualization is at its best, we adjusted the figure size and increased the spacing between bars for clearer view. Each bar on the graph represents an institute, with its height indicates the number of faculties there.

For contextualizing the graph, we added labels to x-axis and y-axis, indicating the institute names and the number of faculties, respectively. Furthermore, a brief title was included to summarize the graph's content. To enhance readability, we turned the x-axis labels by 90 degrees.

Basically, this visualization provides an insight into how faculties are distributed among institutes, aiding comparative analysis and giving insights into academic resources available within each institute. See 12 for more information.

3.4 Analysis of Total Amount Received by Institute Across Years

In the graph, every bar corresponds to a specific institute, and the height of each bar reflects the number of faculties associated with that institute. For clarity, we labeled the x-axis with institute names and the y-axis with the corresponding number of faculties. Additionally, a descriptive title was provided to summarize the graph's content.

This visualization presents a clear and intuitive depiction of how faculties are distributed among different institutes. It enables comparative analysis and offers valuable insights into the academic resources available within each institute. See 9 for more information.

3.5 Analysis of Total Students by Program

The bar plot visualizes the distribution of students across various academic programs within different institutes. Each bar represents a specific program, with its height indicating the total number of students enrolled in that program.

This visualization offers insights into the enrollment patterns across different academic disciplines within each institute. For example, in Institute A, the Computer Science program appears to have the highest enrollment, followed by Engineering and Business Administration. This observation suggests a potential preference for the Computer Science program among students at Institute A.

Additionally, the plot facilitates comparisons between institutes, revealing disparities in program popularity and student preferences across different institutions. Institutes with a stronger emphasis on certain programs may indicate strengths or specializations in those areas. Conversely, differences in program enrollment between institutes may underscore variations in academic offerings or student demographics. See 6, 6, 8 for more information.

3.6 Analysis on Distribution of male and female students within each institute

The box plot visualizes the gender distribution of students within each institute.

For every institute, two side-by-side box plots are presented: one representing male students' distribution and the other female students'.

Each box plot encompasses key statistical measures like the median, quartiles, and any outliers.

By examining the box plots for male and female students, distinctions in the distribution of student populations based on gender within each institute can be discerned.

Furthermore, variations in the width, length, and position of the boxes provide insights into the dispersion and central tendency of the data for male and female students separately.

In essence, the box plot provides a visual synopsis of the gender distribution of students, facilitating comparisons and analyses of gender demographics within each institute. See 10, 10, 11 for more information.

Chapter 4. Feature Engineering

Feature engineering played a pivotal role in shaping the trajectory of our data analysis journey. While our dataset was relatively straightforward, boasting a limited number of features, the importance of feature engineering cannot be understated. Feature engineering involves the manipulation, transformation, and creation of features to enhance the performance of machine learning models. Despite the simplicity of our dataset, we recognized the potential for feature engineering to unlock hidden insights and improve predictive accuracy. However, given the inherently low complexity of our dataset, which already contained a concise set of features, we opted not to delve deeply into feature extraction or feature selection processes. Instead, we focused our efforts on refining existing features and ensuring their relevance to our predictive modeling objectives. By leveraging the inherent characteristics of our dataset and fine-tuning feature representations, we were able to streamline our analysis pipeline and achieve meaningful results without the need for extensive feature engineering.

4.1 Feature Scaling

It is a critical pre-processing step in machine learning that ensures all features contribute equally to the model training process. In our project, we employed min-max scaling to normalize the range of our feature values. This technique rescales each feature to a specific range, typically between 0 and 1, preserving the relative relationships between data points while mitigating the influence of outliers. By applying min-max scaling to our dataset, we effectively transformed our features to a uniform scale, thereby preventing features with larger magnitudes from dominating the model training process. This approach not only enhances the convergence of optimization algorithms but also improves the overall stability and performance of our machine learning models. Through meticulous attention to feature scaling techniques like min-max scaling, we ensured that our models could effectively leverage the information encoded within each feature, ultimately leading to more robust and reliable predictions.

Chapter 5. Model fitting

5.1 Regression

The dataset was scraped using BeautifulSoup from https://www.nirfindia.org/Home. Then this scraped information was stored in a dataframe according to respective academic years i.e. 2019, 2020 and 2021. The target variable was Rank and features were Teaching, Learning and Resources (TLR), Research and Professional Practice (RP), Graduation Outcomes (GO), Outreach and Inclusivity (OI), Peer Perception (Perception). Then correlation matrix was calculated and was visualized in form of Heatmaps. See 12, 13, 13. These Heatmaps show weak or strong correlations of features to target variable. See 14, 14, 15, 15, 16, 16 for better understanding. We employed two machine learning algorithms to predict rank according to given features, which we will discuss in next section.

5.2 Machine Learning Algorithms

5.2.1 Linear Regression

Linear regression serves as a statistical technique utilized to model the relationship between a dependent variable (target) and one or more independent variables (features). It operates under the assumption of a linear relationship between the variables. In the case of simple linear regression, which involves one independent variable, the model can be succinctly represented as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

where:
y is the dependent variable.
x_1 = TLR, x_2 = RPC, x_3 = GO, x_4 = OI, x_5 = Perception.
β_0 and β_1 are coefficients of the intercept and slope to be estimated.
ε is the error term.

The objective of linear regression lies in discovering the line that best fits the data, achieved by minimizing the disparity between actual and predicted values, typically accomplished through the least squares method. This line symbolizes the relationship between the dependent and independent variables within the most straightforward form of linear modeling. See 17, 17, 17 for better understanding.

Applications:

1. Prediction of house prices based on features such as area, number of bedrooms, etc.

2. Forecasting sales based on advertising expenditure.

3. Analyzing the impact of independent variables on a dependent variable.

5.2.2 Random Forest Regression

Random forest regression stands as an ensemble learning method rooted in decision trees. It constructs numerous decision trees and combines their predictions to enhance accuracy and mitigate over-fitting.

Random forest regression employs bootstrap sampling to generate subsets of both features and data points for each decision tree. Each tree grows independently without pruning, and the predictions from all trees are amalgamated for the final prediction. Its advantages encompass reducing over-fitting by amalgamating predictions, effectively handling missing data and outliers, and offering insights into feature importance, rendering it a versatile and potent regression approach.

Applications:

1. Prediction of stock prices based on historical data and market indicators.

2. Forecasting demand for products or services.

3. Medical diagnosis based on patient data.

4. Prediction of customer churn in business.

Chapter 6. Conclusion & future scope

6.1 Findings

- Leaping into the National Institutional Ranking Framework (NIRF) dataset revealed a plethora of insights and trends midst institutes and years. By careful data analysis, we came upon crucial performance indicators and spotted patterns pointing to institutional brilliance and zones for progression.
- Our analysis of the National Institutional Ranking Framework (NIRF) dataset spanning several years showcases the effectiveness of Random Forest Regression in forecasting institute scores. From 2018 to 2021, Random Forest Regression consistently surpassed Linear Regression, demonstrating lower Root Mean Squared Error (RMSE) values and higher R-squared values. This pattern underscores the superior predictive prowess of Random Forest Regression in evaluating educational performance. A comprehensive breakdown of model performance metrics can be found in Table 1,2,3

6.2 Challenges

Our path brimmed with challenges. Initially, pulling raw data from PDF documents raised hurdles demanding the utilization of specialized tools and techniques for efficient data extraction. Additionally, the assorted formats and structures of PDF files called for thorough data cleansing and pre- processing endeavors to ensure the precision and trustworthiness of our analyses. Moreover, the process of model selection and optimization raised hurdles, needing iterative experimentation and fine-tuning to reach optimum performance. Despite these obstacles, our persistence and devotion enabled us to conquer challenges and deliver meaningful results.

6.3 Future plan

Peering forward, our project sets the stage for future adventures in educational analytics and strategic planning. Capitalizing on our discoveries and insights, there exists significant potential for further research and development in predictive modeling and performance assessment. Future initiatives could involve sharpening existing models, exploring extra data reservoirs, and integrating advanced analytics techniques to boost predictive precision and granularity. Furthermore, our project highlights the importance of ongoing data collection and analysis in steering continuous improvement and innovation within the educational environment. By using data-driven insights, we can enable stakeholders to make informed decisions and drive positive modification for the education sector at large.

Group Contribution

All members actively participated in every phase of the project, including data collection, pre-processing, analysis, visualization, and prediction modeling. Each member contributed their skills and expertise to ensure the project's success, collaborating closely to tackle challenges and make informed decisions.

Member1(202318040)

Took the lead in handling all project tasks for the year 2019, including data collection, pre-processing, analysis, visualization, and prediction. Ensured thorough understanding and insights specific to the data from that year.

Member2(202318045)

Scraped data from NIRF website for the year 2020, and preprocessed PDF files into DataFrames and eventually into csv files. Then deriving meaningful data driven insights from csv files in PowerBI environment. Finally, analysed the performance of ML models and visualizing it.

Member3(202318054)

Spearheaded all project tasks for the year 2021, encompassing data collection, pre-processing, analysis, visualization, and prediction. Brought unique perspectives and predictions based on the dataset specific to that year.

Short Bio

Sobhan Behuria

I am a postgraduate student at DAIICT, currently pursuing Data Science. My academic background includes a bachelor's degree in Mathematics and Computing, providing me with a strong foundation in quantitative analysis and computational methods. Alongside my studies, I have freelanced as a tutor, focusing on Mathematics and Statistics which honed my ability to explain complex concepts clearly and assist others in their learning journey. My expertise includes machine learning, big data processing, statistics, and mathematical modeling. I am Proficient in Python, Oracle, PostgreSQL, R Studio, and Excel, which helps in extracting meaningful insights from data. I am passionate about leveraging data to solve real-world challenges, I continuously seek opportunities to contribute to the field of Data Science.

Taruna Sagar Mati

I am a Data Science graduating student. I have a Bachelor's degree in Maths and Computing. Specializing in machine learning algorithms, optimization techniques, big data handling, and data mining, I am also well versed in Python, SQL, PowerBI, Apache Spark, Matlab, Tensorflow, Excel. With a deep understanding of these concepts, I am passionate about leveraging data to solve any real-world problem and derive meaningful insights which are often. Through my academic journey, I have honed my expertise in developing innovative solutions and applying advanced analytical techniques across various domains other than mainstream, like Digital Signal Processing. My goal is to reach the peak of innovations in the field of Data Science and creating something that will live forever.

Srinibas Masanta

Currently pursuing a Master of Science in Data Science, I'm honing my skills and knowledge in this field, having completed a bachelor's degree in Mathematics and Computing. Proficient in math, statistics, and various technical tools like Python (including libraries like NumPy, Pandas, Matplotlib, and Seaborn), I navigate data complexities effectively. I'm also familiar with R Studio and PostgreSQL, broadening my skill set. With expertise in machine learning, SQL, Excel, and big data processing, I extract valuable insights from datasets. My goal is simple: to use data to make smart decisions, learn new things and drive positive change. With my knack for analysis, tech skills, and curiosity, I'm ready to do my part in the constantly changing world of data science.

References

[1] National Institutional Ranking Framework (NIRF) official website: https://www.nirfindia.org/

[2] Bhatia, A., Singh, S. P. (2021). Predicting NIRF Ranking using Machine Learning. In Proceedings of the 3rd International Conference on Computing Methodologies and Communication (pp. 547-553). Springer.

[3] Jha, P. C., Aggarwal, M. (2019). Predicting NIRF Ranking of Indian Universities and Institutes using Machine Learning Techniques. Journal of Data Science, 17(4), 611-626.

[4] Scikit-learn documentation: https://scikit-learn.org/stable/documentation.html

[5] Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

[6] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[7] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189-1232.2011.

[8] Chollet, F. (2018). Deep learning with Python. Manning Publications.

[9] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning. MIT press.

[10] Kingma, D. P., Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[11] Nigam, A., Singh, S. (2020). Predicting NIRF Ranking of Indian Engineering Institutions using Machine Learning Techniques. International Journal of Engineering Research and Technology, 13(2), 96-102

[12] Kumar, A., Kumar, M. (2021). NIRF Ranking Prediction using Ensemble Machine Learning Techniques. In 2021 4th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[13] Jain, A., Sood, S. K. (2020). NIRF Ranking Prediction of Indian Universities using Machine Learning Algorithms. International Journal of Computer Applications, 180(7), 1-5.