SURVEY AND DATA ANALYSIS OF EDUCATIONAL INSTITUTES

Project report submitted in partial fulfillment of the requirement for the degree of the

Bachelor of Technology

in

Information Technology

By

Shubham Rana (191508)

Under the supervision of

Dr. Shubham Goel, Assistant Professor CSE

To

Department of Computer Science & Engineering and Information Technology

Jaypee University Of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby affirm that the work presented in this report, " SURVEY AND DATA ANALYSIS OF EDUCATIONAL INSTITUTES" submitted to the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat, is an authentic representation of my own work completed between January and May(2023), and satisfies a requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology.

Additionally, I certify that I completed the aforementioned project work under the proficiency stream DATA SCIENCE.

No additional degree or certification has been proposed for the subject area of the report.

(Student Signature)

Shubham Rana(191508)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)

Dr. Shubham Goel

Assistant Professor

(Department of Computer Science & Engineering)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: …….

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: _____ __**Department:** _____ **Enrolment No** _____

Contact No. _____**E-mail.** _____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at ……..(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

<table>
<tr><th>Copy Received on</th><th>Excluded</th><th>Similarity Index (%)</th><th colspan="2">Generated Plagiarism Report Details (Title, Abstract & Chapters)</th></tr>
<tr><td rowspan="2"></td><td rowspan="4">• All Preliminary Pages
• Bibliography/Images/Quotes
• 14 Words String</td><td rowspan="2"></td><td>Word Counts</td><td></td></tr>
<tr><td rowspan="2">Report Generated on</td><td>Character Counts</td><td></td></tr>
<tr><td rowspan="2"></td><td>Submission ID</td><td>Total Pages Scanned</td><td></td></tr>
<tr><td></td><td>File Size</td><td></td></tr>
</table>

Checked by
Name & Signature **Librarian**

……

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

Acknowledgement

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

I really grateful and wish my profound my indebtedness to Supervisor Dr SHUBHAM GOEL, Assistant Professor (SG), Department of CSE Jaypee University of Information Technology. Waknaghat. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr SHUBHAM GOEL, Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

Signature of members

Shubham Rana (191508)

Table of Content

List of Figures

List of Abbreviations

Abbreviations	Full form
CSV	Comma Separated Values
Panda	Python data analysis
Numpy	Numerical Python
NIRF	National Institutional. Ranking Framework
ER	Entity Relationship

Abstract

The MHRD authorized the National Institutional Ranking Framework (NIRF), which was introduced on September 29, 2015, by the Honorable Minister of Human Resource Development.

This framework provides a mechanism for classifying educational institutions across the nation. In order to determine the broad criteria for rating different colleges and institutions, the process is based on the overall recommendations and broad understanding reached by a Core Committee established by MHRD. "Teaching, Learning and Resources," "Research and Professional Practices," "Graduation Outcomes," "Outreach and Inclusivity," and "Perception" are among the main categories covered by the parameters.

The National Institutional Ranking Framework (NIRF), which was started at the request of the Hon'ble Prime Minister, has released its initial round of rankings for engineering, management, pharmacy, and universities across the entire nation. Although the thorough ranking framework that described the methodology to rank institutions across the nation was well received, its eventual implementation covering the entire deck of institutions falling into different categories wasn't able to arouse the desired interest among observers and members of the general public. The entire exercise has laid the very foundations for DATA DRIVEN eduGOVERNANCE, though, when viewing the NIRF results from a different angle.

The board makes a lot of information available, but it does not provide it in an analytical format that would enable comparisons between different institutions. In order to perform various types of analysis on the data, we intend to extract the data that is currently available on the website in PDF format.

CHAPTER 1

1.1 Introduction

This project begins with data extraction, followed by data analysis and presentation in a comparable and understandable manner to address a variety of concerns for a variety of audiences.
The information provided on the NIRF website is in pdf format, and no audience analysis has been performed. There is a lot of material posted online, but it is not presented in an analytical manner that would enable a comparison of several institutes.

In our study, we attempted to do data analysis by removing information from the PDF files that were connected to the NIRF website.
Our project is sort of audience oriented, meaning we have tried to address questions of various types of audience. If any person/Institute/Organisation will approach institutes they can get answers to various types of questions, comparison between institutes, and more than what is present on NIRF's official website.

In order to address the 24 issue statements we generated, we extracted and analysed data from several institutes.
The following problem statement is given:

1. Which programme possesses the highest/lowest female to male ratio among the top (5) institutes?

2. What is the proportion of socially and economically disadvantaged persons (SC/ST, OBC, etc.) to all students at each institute?

3. Which institution provides a full fee waiver to the greatest number of applicants?

4. Which institution has the highest percentage of residents from outside the state (both overseas and from other states)? (Ratio of in-state to out-of-state students) Demonstrate the variety of each course's pupils.

5. Which universities have the highest over/underfill percentages?

6. Which institutions have elevators or ramps in fewer than 80% of their structures?

7. Which institutions forbid movements between buildings? Report the total proportion of these institutions as well.

8. Which institutions do not offer separate restrooms for individuals with disabilities in at least 80% of their buildings?

9. List the top five and bottom five institute names with the highest faculty-to-student ratio.

10. List the top 5 institutions that earned the most money annually from 2017–18 to 2019–20 (total) during the previous three years.

11. Which college benefits most from the executive development programme in terms of #Amount of Annual Earnings/#Total Students?

12. Identify the colleges where the majority of students pursue higher education.

13. Identify the universities with the lowest proportion of unplaced students. (year-wise)

14. Which college subject offers the best value overall (across all years)?

15. For each college, what is the average number of full-time and part-time PhD students during the past three years?

16. List the top five and lowest five institutions that received the most money from consulting projects during the last three years.

17. For each year, identify the top 5 institutions with the most projects.

18. The greatest ratio between the number of client projects and the number of client organisations should be found for each institute in the year in question.

19. For each institute over the course of three years, determine the percentage of supported research and consulting initiatives.

20. Which programme at each institute had the largest and lowest intake during the past two to three years?

21. Find the year out of the three stated above when each institute had the highest ratio of projects to agencies.

22. Find the top 5 institutions that got the most funding throughout the course of the previous three years, from 2017–18 to 2019–20, in decreasing order.

23. Determine each institution's total intake. the top five and the bottom five institutions.

24. Identify colleges offering specialised degrees (PG 1-year).

1.2 Objective:

The purpose of writing this kind of issue statement is to respond to various inquiries from various audiences/personas and to compare various institutions.
Students pick institutions depending on their interests because they come from diverse regions of our country.
We have developed these 24 problem statements while taking into consideration the preferences of various student sections.

The goal of the problem statements is:
1. Which programme has the largest or lowest female to male ratio among the top five institutes?
The goal of this issue is to provide information for female students in our nation.

2. What is the proportion of socially and economically disadvantaged persons (SC/ST, OBC, etc.) to all students at each institute?
The goal of this issue is to provide answers to the concerns of our nation's poorer regions.

3. Which institution provides a full fee waiver to the greatest number of applicants?
The goal of this issue is to provide information to students who are struggling financially.

4. Which institute has the percentage of people who do not belong to its state (foreign as well as different states)? show the diversity
present among the students of each course(ratio of instate to outside students)
This problem has the objective to show the comparison of institutes who have maximum diversity i.e, students from different states and different countries.

5. Which are the colleges that have the highest overfill/underfill percent?

Objective of this problem statement is to find out the institutes those are popular but are not managed well ,and also those institutes who are not popular but managed well.

6. Which institutions have elevators or ramps in fewer than 80% of their structures?

The goal of the issue statement is to learn more about the infrastructure of the institutions while also taking into consideration students and persons who are physically challenged.Which institutions have elevators or ramps in fewer than 80% of their structures?

The goal of the issue statement is to learn more about the infrastructure of the institutions while also taking into consideration students and persons who are physically challenged.

7.Which institutions forbid movements between buildings? Report the total proportion of these institutions as well.

Finding out a little about institutes' internal systems is the objective challenge.

8. Which institutions do not offer separate restrooms for individuals with disabilities in at least 80% of their buildings?

Finding information about the unique facilities for persons with special needs is the objective of the problem description.

9. List the top five and bottom five institute names with the highest faculty-to-student ratio.

The goal is to determine how much time, support, assistance, and mentoring a student may receive from faculty members at various institutions. Additionally, it conveys the notion of management and resources.

10. List the top 5 institutions that earned the most money annually from 2017–18 to 2019–20 (total) during the previous three years.

Which college benefits most from the Executive Development Programme? #Annual Salary / Total Number of Students

The goal of the aforementioned issue statements is to identify the institutions that are more economically stable and effectively run.

12. Identify the colleges where the majority of students pursue higher education.

Discovering institutions that prepare students for further education is the goal.

13. Identify the universities with the lowest proportion of unplaced students.(year-wise)

This issue statement's goal is to learn more about institutes that do a good job with placements. It also provides responses to inquiries from students that are focused on placement.

14. Which college subject offers the best value overall (across all years)?

The goal of this issue statement is to identify the institutions that provide excellent packages for various courses.

15. For each college, what is the average number of full-time and part-time PhD students during the past three years?

Finding reputable institutions for a PhD is the goal.

16. List the top five and lowest five institutions that received the most money from consulting projects during the last three years.

17. List the top five institutions for each year that had the most projects.

18. For each institute, identify the year with the largest ratio between the quantity of client projects and the quantity of client organisations.

19. What is the three-year proportion of supported research and consulting projects for each institute?

The goal of the aforementioned issue statements is to identify colleges that excel at student training and good training opportunities. institutions that large organisations rely on. .This also gives about the institutes who are innovative, students from these universities learn practically more. have high chances of a bright future.

20. For the past two to three years, which programme at each institute had both the greatest and lowest intake?

Finding out which institutions prioritise quality over number and which institutions don't give a damn about quality students is the goal of this study.

21. Which of the three years indicated above had the largest ratio of projects to agencies for each institute?

22. List the top 5 institutions that got the most money throughout the course of the previous three years, from 2017–18 to 2019–20, in decreasing order.

The goal of the two problem statements above is to identify the institutions that are successful at raising money from sources other than student fees and grants from other organisations.

23. Identify each institution's overall intake. the top five and the bottom five institutions.

The goal of this issue statement is to identify institutions that desire more admissions and those that seek less admissions.

24. Inform colleges about any uncommon courses (PG 1 year).

The goal of this issue statement is to identify institutions that provide uncommon courses. This is rather recent in the market.

Chapter-2

Literature Survey :

Literature Review 1 :

Impact of NIRF Ranking on Research Publications: A Study with Special Reference to North-East Indian Universities

1Pranjal Deka, 2Dr. Mukut Sarmah

1PhD Research Scholar, 2Associate Professor, Department of Library and Information Science, Assam

University, Silchar (Assam)

Email: pranjalpriya3@gmail.com

Abstract:

The aim of this study is to follow the effects of the NIRF rankings on research publications and to evaluate the research output, citation counts, and h-index of the relevant universities, with special reference to North-East Indian institutions.

Methodology: The top 100 universities in the NIRF rankings were used to choose the pertinent institutions. For information on publications, citations, and h-index in scientific research, consult the Web of the Science database. In this study, research papers from the eight institutions were analysed over a period of ten years, five years before and five years after the NIRF rankings (2011-2015 and 2016-2020), respectively.

Findings: The analysis shows that research articles play a significant role in how schools and institutions are ranked. NIRF rates publications according on their caliber.

According to the survey, just eight universities from NE India were ranked in the NIRF ranking. Tezpur institution is the most active institution in the region. The rankings of universities are always changing from one year to the next. Distribution of research publications, citations, and h-index during the previous 10 years, both before and after rankings, all institutions have made an attempt to boost their academic output.

Literature Review 2:

Impact of Research Output on NIRF Ranking: A Correlational Study

Vysakh C., Rajendra Babu H. & Spandana R.L

The primary goal of the study is to investigate how research production affects the academic standing of Indian Institutes of Management. Google Scholar and the official NIRF website were used to gather the study's data.

There are now 20 IIMs in the nation, however only 13 received high rankings in NIRF. The public's perception of key research completed in the past is humorous.

This report, the first of its type at the national level, aids Indian institutions in getting a comprehensive understanding of academic standing.

Keywords: NIRF Ranking, NIRF Citation, and IIM.

Literature Review 3:

Impact of NIRF Ranking on Research Publications: A Study with Special Reference

to North-Indian Universities

1Pranjal Deka, 2Dr. Mukut Sarmah

1PhD Research Scholar, 2Associate Professor, Department of Library and Information Science, Assam

University, Silchar (Assam)

Email: pranjalpriya3@gmail.com

Abstract:

With an emphasis on North-East Indian universities, the goals of this study are to evaluate the research output, citations, and h-index of the pertinent institutions and to track the impact of the NIRF ranking on research output.

The compilation of the top 100 institutions based to the NIRF rating was used to choose the pertinent universities. Data about research papers, citations, and h-index may be found using the internet version of the Science database. The 10 years of research papers from the eight institutions that were selected, from 2011 to 2015 and from 2016 to 2020, or a period of five years prior to and five years following the NIRF rankings, are examined in this study.

Findings: The study shows that research articles are a significant determinant in college and university rankings. The quality of publications is given a 70% weighting by NIRF.

According to the survey, just eight universities from NE India were ranked in the NIRF ranking. Tezpur institution is the most active institution in the region. The positions of universities fluctuate significantly from one year to the next.

According to the distribution of research publications, citations, and h-index during the previous 10 years, both before and after rankings, all universities have made an attempt to boost their academic output.

CHAPTER 3 SYSTEM DEVELOPMENT

This projects Hierarchical form is given below:

1.Selecting the institutes , so that based on their data we can answer our problem statements.
2.collecting pdf files of the selected institutes(first step of Data Extraction)
3.Data Extraction i.e, converting pdf file data into csv file , selecting the rows and columns which have required information from the pdf file and then exporting it to csv file.
4. Converting the extracted csv files into our required data set.
5. Performing different sorts of analysis on the data set we have created using python and its different libraries.
6. Using the Tableau for Data visualization, i.e presenting outputs in more visual/understable form

3.1 System Development:

Selecting the institutes:

Our project's first step is to choose the institutes, and then we may answer various issue statements using their data. This is all a theoretical process that is completed by consulting with our project supervisor.

In order to respond to the 24 various issue statements, we have chosen around 100 institutes depending on the information provided about them. To address various issue statements, we will be able to collect data and analyse it afterwards.

When choosing institutions, bear in consideration that

Institutes must have all the criteria necessary to address our issue statements and be registered on NIRF for the previous five years.

We conducted the study using the 2019–2021 data that was provided.

Some of the institutes which which are were selected are as follows:

- Indian Institute of Technology Roorkee
- Indian Institute of Technology Bombay
- Indian Institute of Technology Madras
- Homi Bhabha National Institute
- Indian Institute of Technology Guwahati
- Mahatma Gandhi University
- National Institute of Technology Tiruchirappalli
- Shanmugha Arts Science Technology & Research Academy
- Indian Institute of Technology Mandi
- Indian Institute of Science Education & Research Pune
- S.R.M. Institute of Science and Technology
- Vellore Institute of Technology
- Dr. D. Y. Patil Vidyapeeth

- Indian Institute of Technology Delhi…… similar type of 100 institute of all type which are registered on NIRF are selected and their data of 2019 ,2020, 2021 are studied manually so they can provide answer to our problem statements.

3.2 Collecting pdf files of the selected institutes:

After the selection process is complete, our next step is to obtain the pdf files from the NIRF official website. Since we have chosen around 100 institutes, we will need to download about 300 pdf files:

Downloaded pdf files look like this.

ubmitted by Institution for India Rankings "2021"

e Name: Indian Institute of Technology Madras [IR-O-U-0456]

ıned (Approved) Intake

Academic Year	2019-20	2018-19	2017-18	2016-17	2015-16	2014-15
ears Program(s)]	762	488	466	466	-	-
ears Program(s)]	157	358	372	372	372	-
ear Program(s)]	690	658	-	-	-	-
ear Program(s)]	245	245	245	-	-	-
grated	46	46	46	46	46	-

ctual Student Strength (Program(s) Offered by your Institution)

ograms years)	No. of Male Students	No. of Female Students	Total Students	Within State (Including male & female)	Outside State (Including male & female)	Outside Country (Including male & female)	Economically Backward (Including male & female)	Socially Challenged (SC+ST+OBC Including male & female)	No. of students receiving full tuition fee reimbursement from the State and Central Government	No. of students receiving full tuition fee reimbursement from Institution Funds	No. of students receiving full tuition fee reimbursement from the Private Bodies	No. of students who are not receiving full tuition fee reimbursement
ears ı(s)]	1874	332	2206	338	1848	20	215	1137	59	832	155	306
ears ı(s)]	1459	270	1729	313	1402	14	176	862	44	697	92	205
ear ı(s)]	1389	321	1710	282	1423	5	361	707	0	761	131	176
ear ı(s)]	626	123	749	175	574	0	275	232	6	374	96	31
grated	86	129	215	45	170	0	12	116	0	56	0	72

Figure 1 - data

PG [3 Years Program(s)]: Placement & higher studies for previous 3 years

Academic Year	No. of first year students intake in the year	No. of first year students admitted in the year	Academic Year	No. of students admitted through Lateral entry	Academic Year	No. of students graduating in minimum stipulated time	No. of students placed	Median salary of placed graduates(Amount in Rs.)	No. of students selected for Higher Studies
2016-17	245	212	2017-18	0	2018-19	130	81	1100000(Eleven Lakhs)	18
2017-18	245	221	2018-19	0	2019-20	170	131	1400000(Fourteen Laksh)	33
2018-19	245	193	2019-20	0	2020-21	148	107	1194000(Eleven Lakhs Ninety Four Thousand)	32

PG-Integrated [5 Years Program(s)]: Placement & higher studies for previous 3 years

Academic Year	No. of first year students intake in the year	No. of first year students admitted in the year	Academic Year	No. of students graduating in minimum stipulated time	No. of students placed	Median salary of placed graduates(Amount in Rs.)	No. of students selected for Higher Studies
2014-15	46	42	2018-19	38	16	1417176(Fourteen Lakhs Seventeen Thousand One Fourty six)	9
2015-16	46	42	2019-20	32	10	819000(Eight Lakhs Nineteen Thousand)	5
2016-17	46	44	2020-21	34	21	900000(Nine Lakhs)	9

Figure 2 - data

The technical phase now begins when we download the PDF files and use python libraries to extract the data from them.

A Jupyter notebook was initially installed on our machine.

Jupyter:

You may create and share documents with real-time code, mathematical equations, images, maps, graphs, and narrative prose using the web-based, open-source Jupyter Notebook interactive environment.

Jupyter offers the data analysis capabilities needed, particularly for our project. Jupyter notebook was used in this way.

We uploaded our PDF file to the Jupyter notebook after setting it up and pulled the data—a particular row and column—from it.

We used four separate Python libraries to carry out this operation, and they were as follows:

. **PYpdf**

.PYpdf2

. Camelot

.Tabulla.

These are the python libraries which i used to extract rows and columns from the pdf file:

.**PYPDF:**

a collection of Pure-Python-based PDF tools. It can trim pages, combine many pages into one, divide and merge documents page by page, extract document metadata (such the document's title and author), and encrypt and decode PDF files.

PYPDF2: Users may divide, combine, crop, and alter the pages of PDF documents with PyPDF2, an all-Python PDF library. It is both free and open-source. Additionally, it could offer password security, individualised information, and PDF file reading choices. Using PyPDF2, text and metadata may be retrieved from PDF files.

.**Camelot**

Extraction of data tables from PDF files is made simple by the Camelot Python programme. However, only text-based PDFs are compatible with this library; scanned PDFs are not.

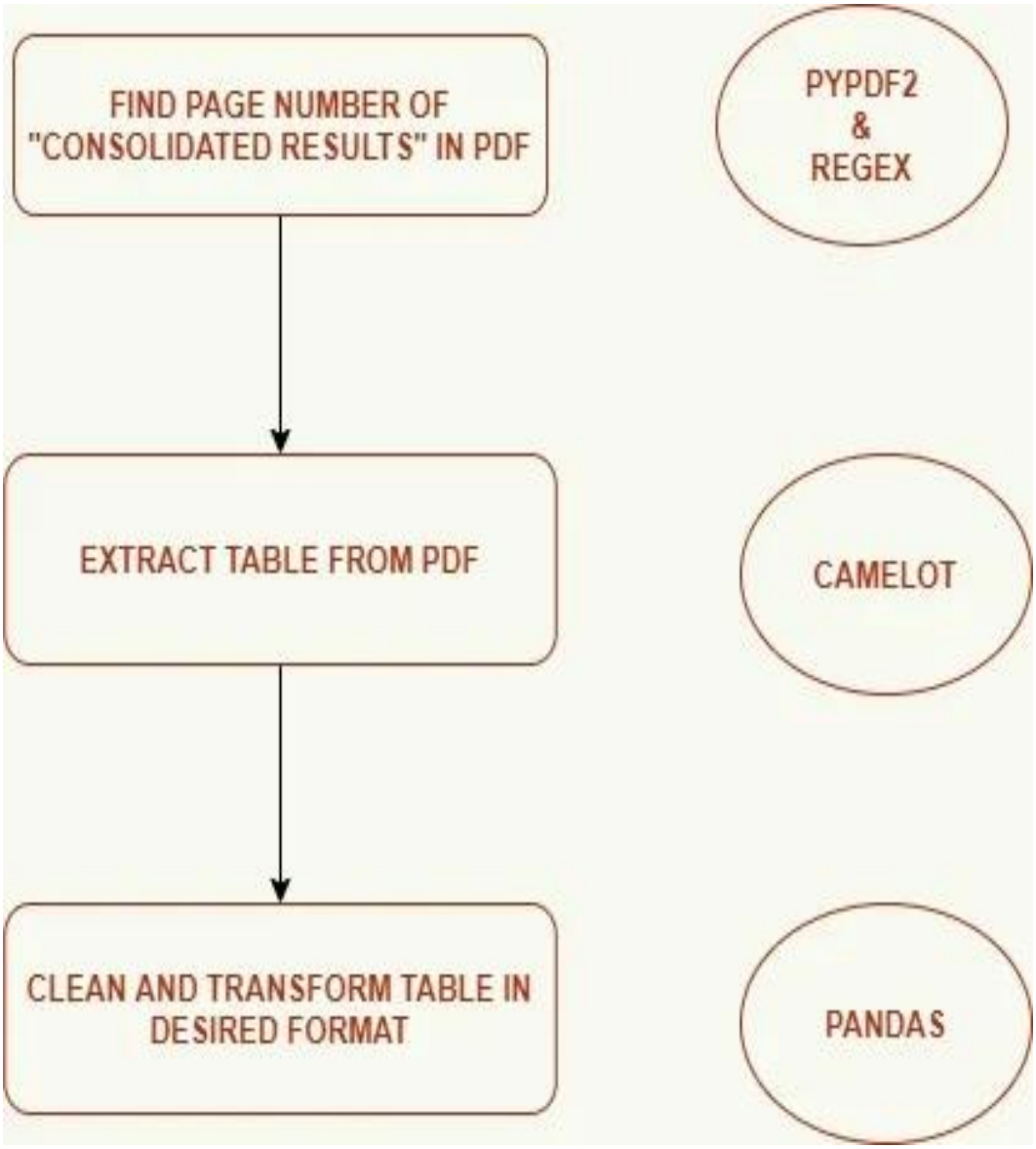
There are a number of parameters in Camelot that may be altered to enhance the extraction of data from tables.

Why was Camelot chosen?

You are capable of: Unlike other libraries and tools that either provide beautiful results or utterly fail (with no in-between), Camelot lets you customise table extraction. This is crucial since everything in the real world is hazy, including the extraction of PDF tables. On the basis of criteria like accuracy and whitespace, poor

tables may be removed without ever having to manually review each one. Every table is a pandas DataFrame, thus it can be used in processes with ease.

Flow chart of extracting tabular information from PDF using Python:



Flowchart

Why extracting tables from PDF is hard?

f you look at the PDF layout above, you will notice no concept of tables in it. A PDF

ontains instructions to place a character at an x,y coordinate on a 2-D plane,

etaining no knowledge of words, sentences, or tables.

ordinate Systems

coordinate system on a PDF page is called **User Space**. This is a flat 2- dimensional space, just like a
e of paper. And in fact that's a good way to think about it. The units of User Space are called "points" and
e are 72 points/inch. The origin, or 0,0 point is located in the bottom left hand corner of the page.
izontal, or X, coordinates increase to the rights and vertical, or Y, coordinates increase towards the top
ure 1)

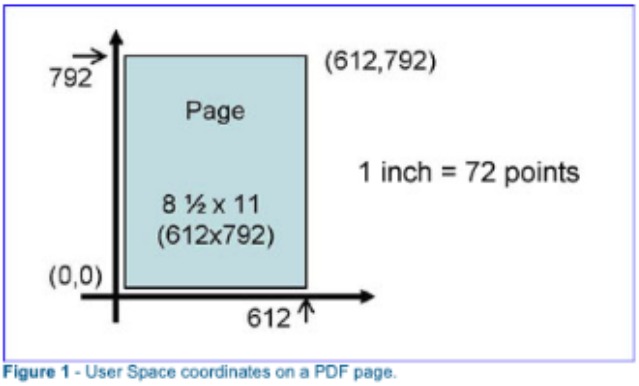


Figure 1 - User Space coordinates on a PDF page.

```
#conda (easiest way)
$ conda install -c conda-forge camelot-py

#pip after installing the tk and ghostscript dependencies
$ pip install "camelot-py[cv]"
```

Find page number of desired table in pdf document
Extract table using page number from pdf document
Clean extracted dataframe
Reformat dataframe
Extract desired information from each extracted and reformatted dataframe
If pdf is successfully processed, move to processed folder
If pdf encountered an error, move to error folder
Export final result to csv and move to output folder

Figure 3 - Psuedocode

Other liabraries used while extracting were pandas and os.

Pandas:

The main goal of the pandas library is data set manipulation, or the capacity to edit, swap out, and modify certain elements of a DataFrame type object.

Only two examples of the many services that pandas provides are the computation of descriptive statistics and the display of the columns and rows in a data collection.

OS: The OS module in Python has methods for a variety of tasks, including adding and deleting folders and accessing their contents, changing directories, determining the current directory, and more.Before you can interact with the underlying operating system, you must import the os module.We were able to respond to our issue statement by removing specific data from the rows and columns of many PDF files after employing all of these libraries.

3.3 Creating of Data set:

Now we are able to create 12 different data sets to answer our 24 problem statements .

Our data set are in the form of csv file some of our used data sets are are as follows:

- Consistent_Institutes.csv
- Total_Strength_three_years.csv
- cosultancy.csv
- executive-development-program.csv
- number-of-faculties.csv
- phd 2021 2020 2019.csv
- placement2019.csv
- placement2019_different_type_of_data.csv
- sanctioned-intake.csv
- sponsorship.csv
- total-actual-strength.csv

Institute	Programs	No. of Male Students	No. of Female Students	Total Students	Within State(Including male& female)
Institute of Chemical Technology	UG [4 Years Program(s)]	676	296	972	832
Institute of Chemical Technology	PG [2 Year Program(s)]	352	284	636	423
Institute of Chemical Technology	PG-Integrated	184	61	245	146
Jamia Millia Islamia	UG [3 Years Program(s)]	1904	1193	3097	1144
Jamia Millia Islamia	UG [4 Years Program(s)]	1523	371	1894	683
Jamia Millia Islamia	UG [5 Years Program(s)]	470	450	920	349
Jamia Millia Islamia	PG [2 Year Program(s)]	1280	1284	2564	847
Jamia Millia Islamia	PG [3 Year Program(s)]	115	33	148	52
Kalasalingam Academy of Research and Higher Education	UG [3 Years Program(s)]	759	627	1386	581

Figure 4 – csv-file

Institute	Number of Faculties
Institute of Chemical Technology	168
Jamia Millia Islamia	742
Kalasalingam Academy of Research and Higher Education	526
Sri Venkateswara University	510
Indian Institute of Technology Guwahati	436
Gujarat University	381
Osmania University	428
Indian Institute of Science Education & Research Kolkata	127
Anna University	1024
Indian Institute of Technology Roorkee	574

Academic Year	No. of first year students intake in the year	No. of first year students admitted in the year	Academic Year	No. of students admitted through Lateral entry	Academic Year	No. of students gra
2012-13	819	795	2013-14	125	2015-16	849
2013-14	819	783	2014-15	125	2016-17	835
2014-15	819	829	2015-16	101	2017-18	800
2012-13	819	795	2013-14	125	2015-16	849
2013-14	819	783	2014-15	125	2016-17	835
2014-15	819	829	2015-16	101	2017-18	800
2013-14	0	0	2014-15	0	2015-16	0
2014-15	300	230	2015-16	0	2016-17	180
2015-16	120	111	2016-17	0	2017-18	98
2013-14	0	0	2014-15	0	2015-16	0
2014-15	300	230	2015-16	0	2016-17	180

Figure 5-csv-file

3.4 Data Analysis part:

Now that a data set has been created, we can go on to our most crucial and nearly finished step, which is data analysis and displaying the results of our issue statements.

After the data set was constructed, it was once again uploaded to a Jupyter notebook, where data analysis was done on it to provide a solution or output for each of the issue statements we had formulated.

The programme remains the same, but we must interact with many additional Python modules to begin the data analysis portion.

i.e ,JUPYTER notebook.

- Various other libraries of python which were used as follows:
- .Pandas
- .Numpy
- .Matplot Library
- .Seaborn

.Pandas

The ability to edit, swap out, and modify particular parts of an object of the DataFrame class is the main goal of the pandas library.

Just two examples of the numerous capabilities that pandas provides are the computation of descriptive statistics and the visualisation of the columns and rows of a data collection.Pandas is a well-liked tool for managing data analysis. It facilitates data loading from external sources like as text files and databases and provides tools for data analysis and modification once the data has been added to your computer. With the use of pandas' capabilities, many regular tasks that once necessitated writing multiple lines of code in the original Python language may now be automated and simplified.

Numpy

The Python library NumPy provides the simple yet effective n-dimensional array data structure. Every Python data scientist's journey begins with learning NumPy since it is the building block from which almost all of the toolkit's capabilities are built.

Benefits of using NumPy for data analysis. Utilising NumPy is very beneficial when creating data objects with N dimensions.

Its framework functions quickly and easily while dealing with homogenous datasets.When doing numerical calculations, Python lists use more memory than NumPy arrays.

.Matplot Library

The Matplotlib toolbox for Python is a complete tool for producing static, animated, and interactive visualisations. Both straightforward and challenging projects are feasible using Matplotlib.

Publishable tales should be written. Create dynamic charts with zoom, pan, and update capabilities.

A basic Python tool for data visualisation is called Matplotlib. It replicates MATLAB-style graphs and presentations and is simple to use. The plots in this library, which include line charts, bar charts, histograms, and more, are built on NumPy arrays.

.Seaborn

The Seaborn library in python allows you to create statistical visuals.

Its foundation is Matplotlib, and Pandas data structures are intimately connected with it.

You may examine and comprehend your data with Seaborn.

Using these libraries and several kinds of data analysis, we were able to provide results for the issue statements we had defined.

We will now provide the findings we made for the several issue statements we created for this project.

CHAPTER 4 PERFORMANCE ANALYSIS

4.1 OUTPUTS:

Q.1

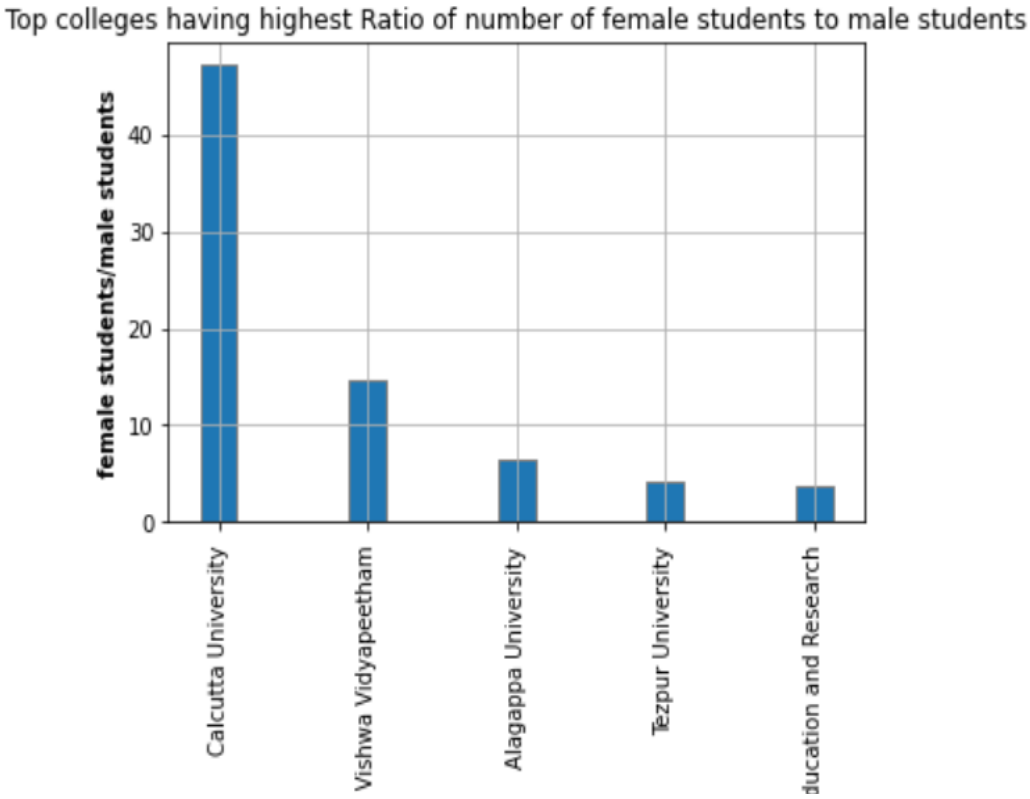


Figure 6 – Q1

Q.1

PART 2

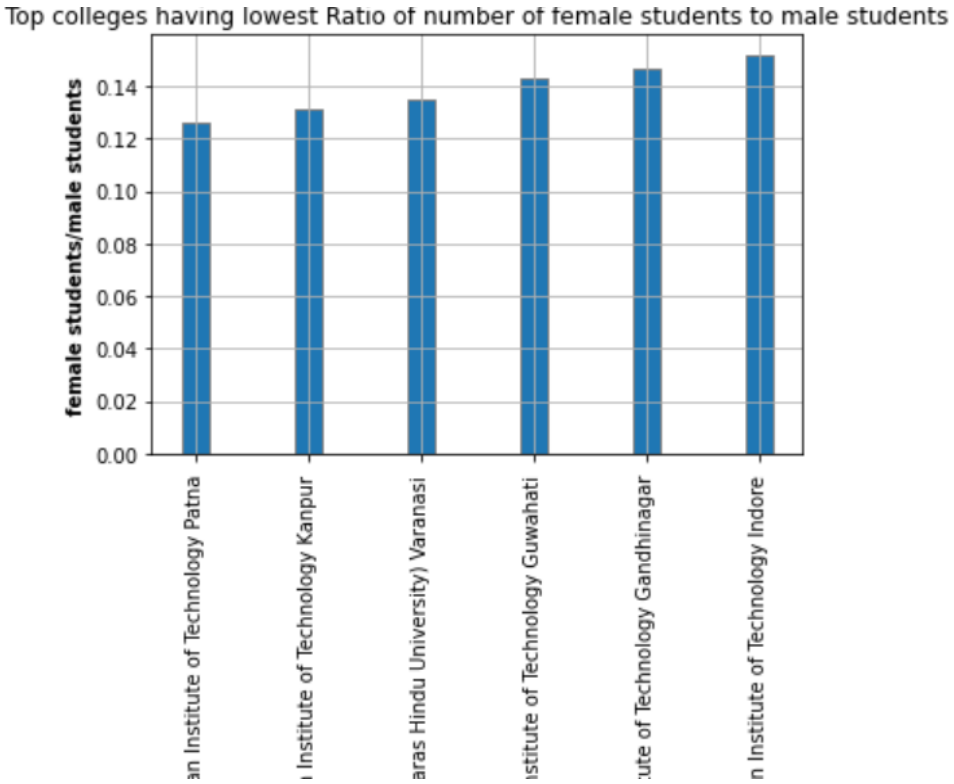


Figure 7 – Q1-Part-2

Q.2

Answer for the problem statement second , which is a csv file so, we are attaching a screenshot of the data.

Institute	ratio social
Bharathidasan University	0.9788279773156899
Bharath Institute of Higher Education & Research	0.884434239962322
PSG College of Technology	0.8698239222829387
North Eastern Hill University	0.796267087276551
Sri Venkateswara University	0.7516865221759724
University of Madras	0.7446054750402576
Anna University	0.7263845798402535
Kalasalingam Academy of Research and Higher Education	0.6994586233565352
Calicut University	0.6941451990632318
Pondicherry University	0.6521377474629845
Osmania University	0.6232775919732442

Figure 8 – Q2

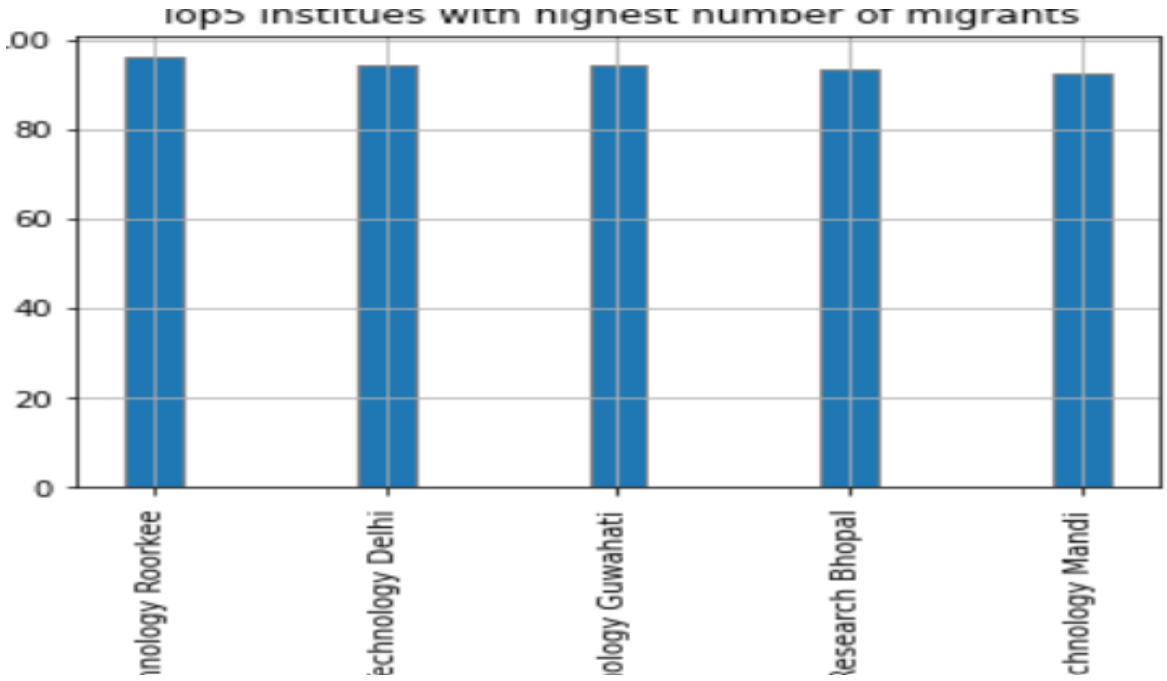
Q.3

Again solution to the problem statement three is csv file, i.e, are solution is in the form of row and column so again we are attaching screenshot.

	Institute	ratio economic
2	Jamia Millia Islamia	0.7463759712397078
3	University of Kashmir	0.6700876383763837
4	Calcutta University	0.6211902259590121
5	Bharathiar University	0.5233423545331529
6	Aligarh Muslim University	0.4940093078522626
7	Alagappa University	0.4762481089258699
8	Mysore University	0.46991078669910785
9	Dr. B. R. Ambedkar National Institute of Technology	0.441941539286672
10	Jadavpur University	0.380518889394629
11	Gauhati University	0.37229965156794426

Figure 9 -Q3

Q.4


Top5 institues with highest number of migrants
100
80
60
40
20
0
nology Roorkee
echnology Delhi
ology Guwahati
Research Bhopal
chnology Mandi

Q.4

Part 2

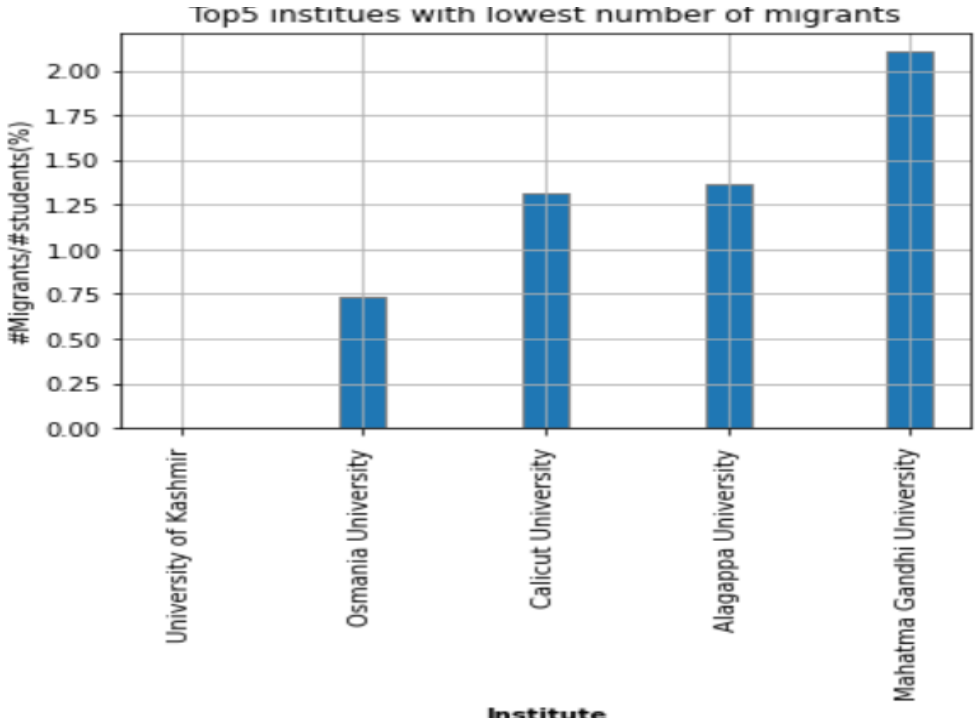

Top5 institues with lowest number of migrants
#Migrants/#students(%)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
University of Kashmir
Osmania University
Calicut University
Alagappa University
Mahatma Gandhi University
Institute

Figure 10 – Q4

Q.5

Part 2

Institute	Programs	Vacancies-Perc
Tezpur University	PG [1 Years Program(s)]	58.7302
Banasthali Vidyapith	PG [1 Years Program(s)]	53.3333
Visva Bharati	PG [1 Years Program(s)]	46.0526
Aligarh Muslim University	PG [1 Years Program(s)]	43.3333
Indian Institute of Technology Indore	PG [2 Years Program(s)]	34.0426

Figure 11 – Q5

Q.6

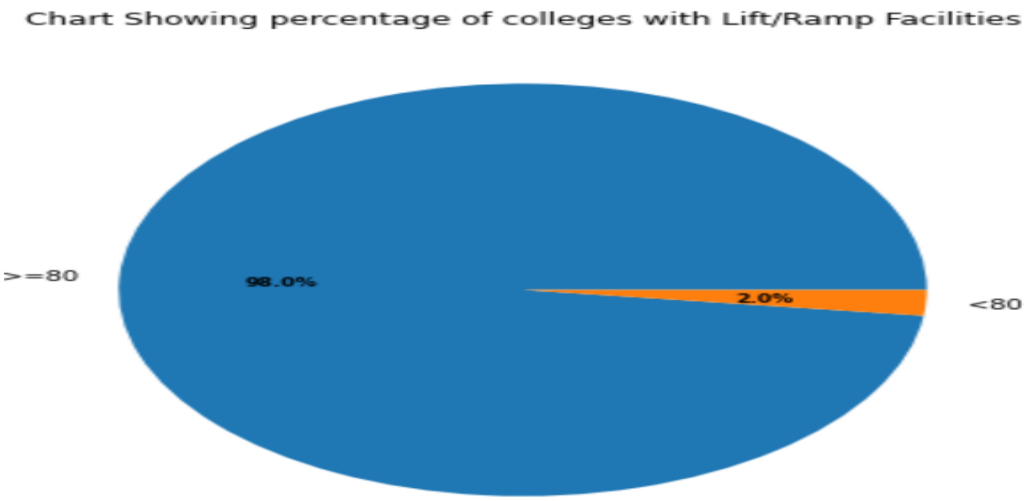


Figure 12 -Q6

Q.7

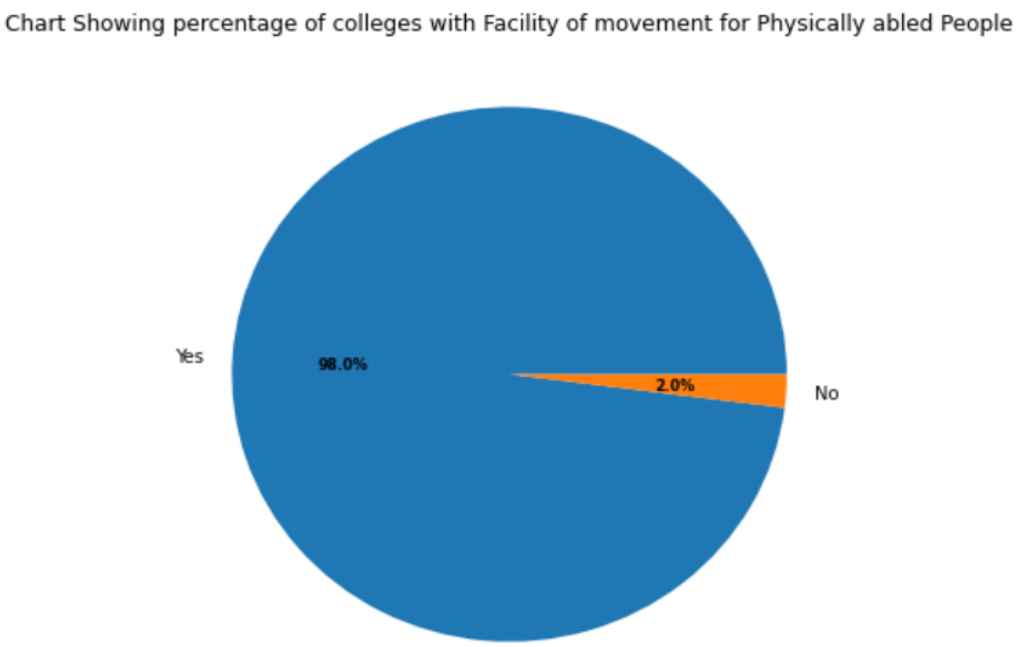


Figure – 13 – Q7

Q.8

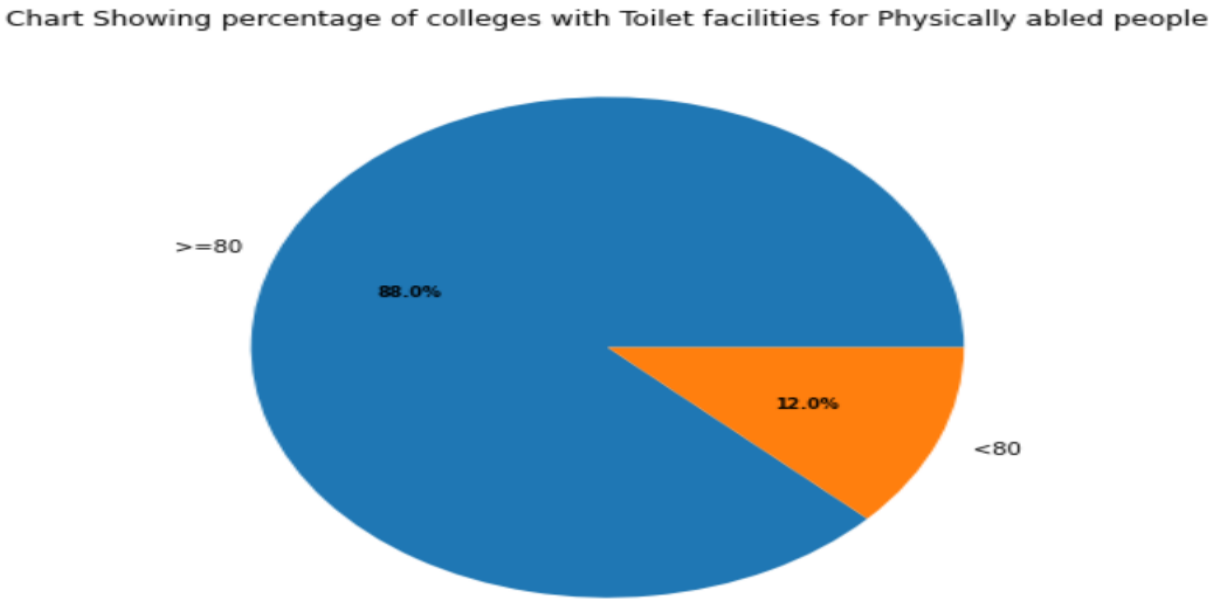


Figure 14 -Q8

Q.9

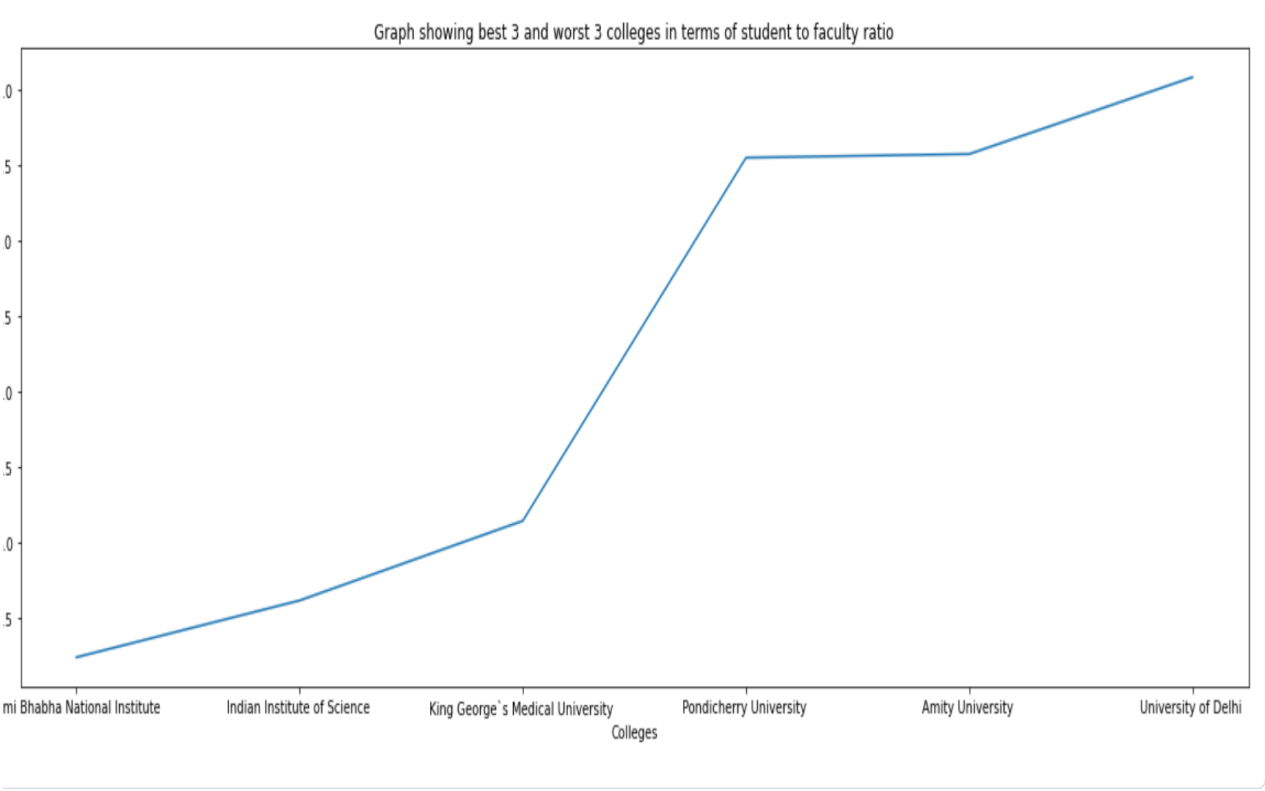


Figure 15 – Q9

Q.10

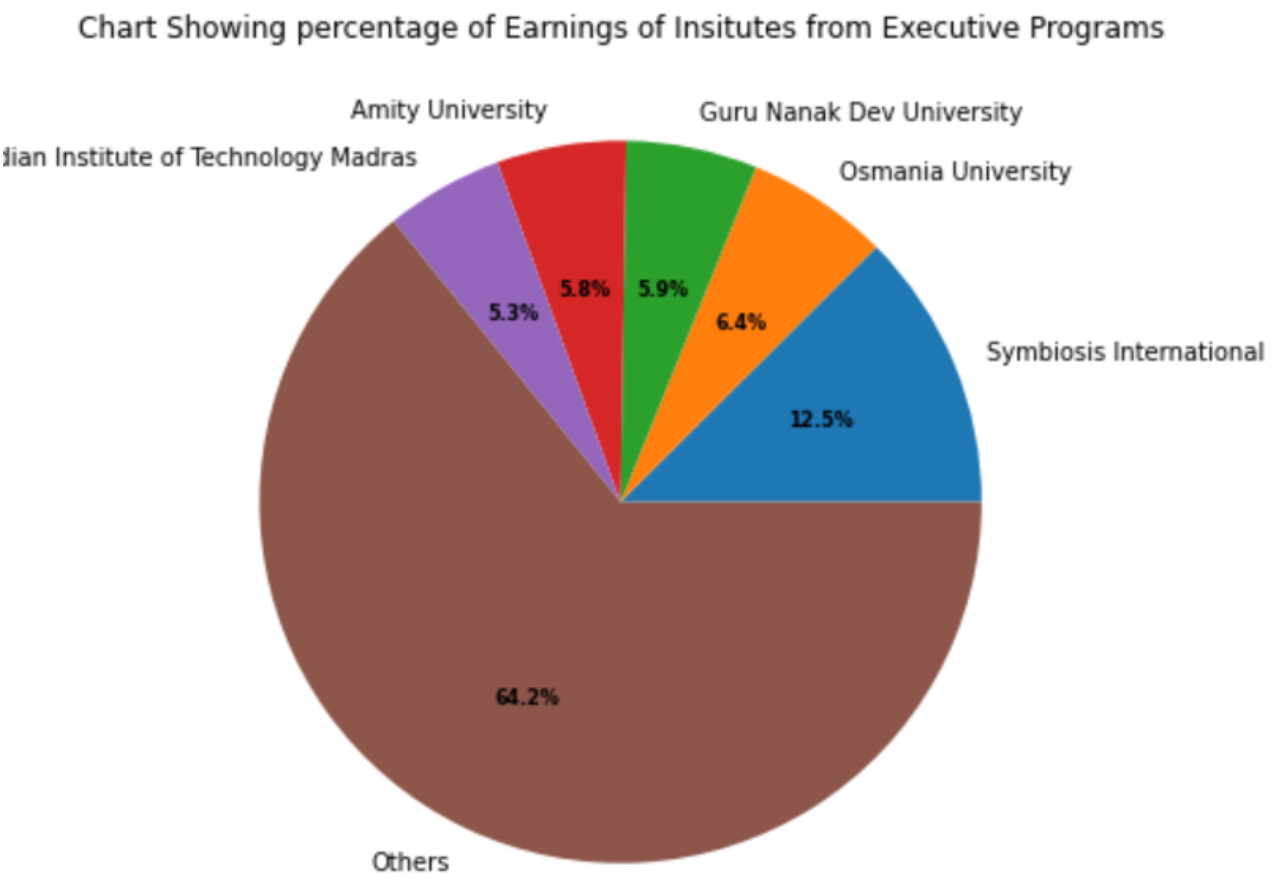


Figure 16 – Q10

Sponsored Research projects and consultancy projects.csv

1	Institute	Reasearch Project%(2017-18)	Reasearch Project%(2018-19)	Reasearch Project%(2019-20)	Consultancy Project%(2017-18)	Cons
2	Alagappa University	88.37209302325581	86.56716417910447	94.73684210526315	11.627906976744185	13.43
3	Aligarh Muslim University	72.0	70.96774193548387	53.289473684210535	28.000000000000004	29.03
4	Amity University	95.36423841059603	92.85714285714286	94.90445859872611	4.635761589403973	7.142
5	Amrita Vishwa Vidyapeetham	1.7744631606017742	1.6951147733961152	55.06329113924051	98.22553683939823	98.30
6	Andhra University, Visakhapatnam	30.0	21.57676348547718	21.544715447154474	70.0	78.42
7	Anna University	21.25	7.6541459957476965	5.309218203033839	78.75	92.34
8	Banaras Hindu University	99.47643979057592	99.46666666666667	99.47229551451187	0.5235602094240838	0.533
9	Banasthali Vidyapith	94.44444444444444	90.38461538461539	78.72340425531915	5.555555555555555	9.615
10	Bharath Institute of Higher Education & Research	8.695652173913043	3.816793893129771	7.415730337078652	91.30434782608695	96.18
11	Bharathiar University	92.3076923076923	80.0	94.44444444444444	7.6923076923076925	20.0
12	Bharathidasan University	88.37209302325581	89.04109589041096	79.71014492753623	11.627906976744185	10.95

Figure 17 – projects.csv

2017-18 EDP Earnings.csv

1	Institute	2017-18
2	Symbiosis International	159938589
3	SVKM`s Narsee Monjee Institute of Management Studies	90228000
4	Guru Nanak Dev University	80263540
5	Andhra University, Visakhapatnam	75481122
6	Indian Institute of Technology Madras	62869852
7	Indian Institute of Technology Delhi	60309589
8	Indian Institute of Technology Bombay	45885826
9	Indian Institute of Technology Kharagpur	41453100
10	University of Delhi	35300000
11	Indian Institute of Technology Kanpur	33226000
[illegible]	[illegible]	[illegible]

Figure 18 – earnings.csv

Maximum-project-to-client-ratio-year.csv

Institute	Year (Maximum)
Alagappa University	2018-19
Aligarh Muslim University	2019-20
Amity University	2019-20
Amrita Vishwa Vidyapeetham	2017-18
Andhra University, Visakhapatnam	2019-20
Anna University	2019-20
Banaras Hindu University	2017-18
Banasthali Vidyapith	2019-20
Bharath Institute of Higher Education & Research	2018-19

Figure 19 – clients.csv

Top 5 student to faculty ratio.csv

	Institute	Student_to_Faculty
2	Homi Bhabha National Institute	1.2216014897579144
3	Indian Institute of Science	3.0991379310344827
4	King George`s Medical University	5.7304147465437785
5	Savitribai Phule Pune University	6.217696629213483

Figure 20 – ratio.csv

Find the colleges with the highest percentage of students going for higher studies.

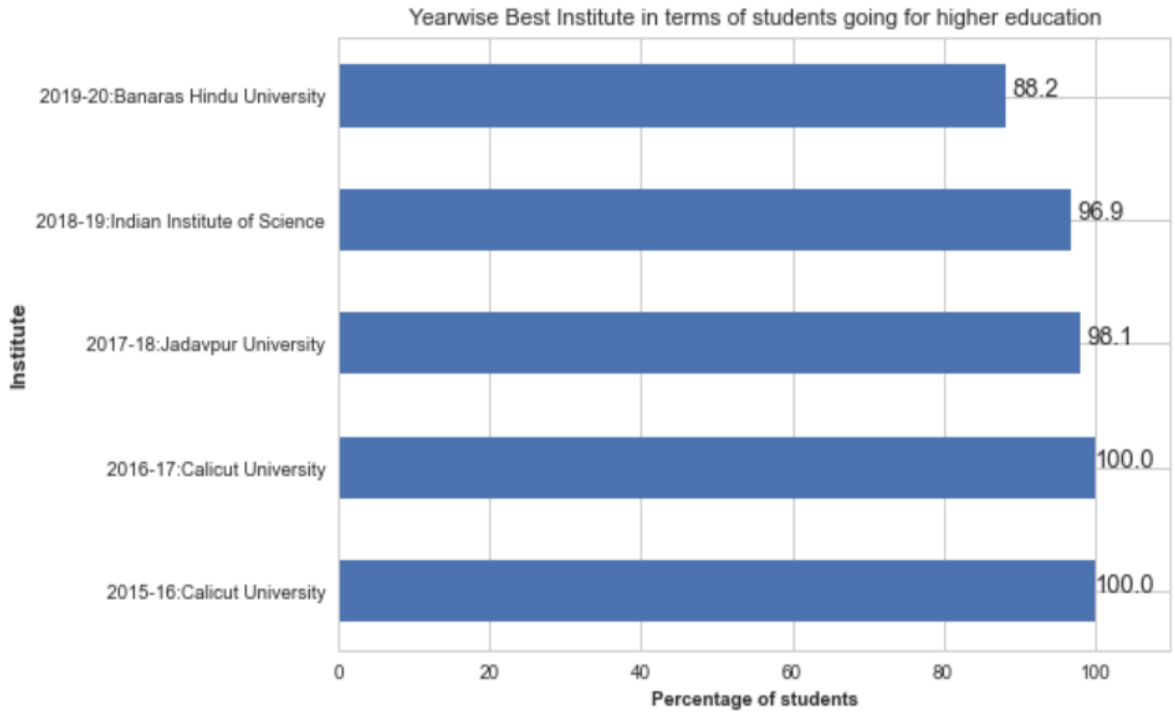


Figure 21 – list.csv

Worst institute in terms of students going for higher

Academic Year	Institute	% Students going for higher studies
2015-16	Alagappa University	0.0
2016-17	Bharath Institute of Higher Education & Research	0.0
2017-18	Bharath Institute of Higher Education & Research	0.0
2018-19	Bharath Institute of Higher Education & Research	0.0
2019-20	Cochin University of Science and Technology	0.0

Figure 22 – worstinstitute.csv

CHAPTER 5

5.1 Conclusions:

As a conclusion, we will claim that we have done something new in the field of data analysis and data extraction. First, we developed several issue statements while considering various student and professional groups.

After reviewing the information provided on the NIRF website, we used the website's PDF files to generate our own data collection. We next ran several types of analysis on the data set we had produced, using various Python packages, to see how each issue statement may be solved. We used Numpy, Pandas, Mataplotlib, and Seaborn in our data analysis section to do various forms of mathematical and statistical analysis, as well as to plot different types of graphs and to visualise results.

5.2 FUTURE WORK:

Data analysis of our nation's educational institutions is the focus of the entire project we are working on. In order to learn more about the existing educational system in our nation, particularly in regards to graduation and post graduation, we had a future plan to survey additional educational institutions using various methods, such as emails, linkedin, other social media platforms, etc.

Therefore, my future goal with this part—the survey part—is to learn more about other institutes and undertake data analysis on that data by coming up with various issue statements. so that our future generation would have access to more effective visual information when it comes to schooling.

5.3 Application:

Real life application of this project is too many.

outputs/ solutions created by this project will help different sections of students/professionals , if they aree going to join university in anyform. These projects will solve the major problems of students which they face while joining any institute.

References

[1] Amador-Perez and R. A. Rodriguez-Solis, “Analysis of a CPW-fed annular slot ring antenna using DOE” in Proc. IEEE Antennas Propag. Soc. Int. Symp., Jul. 2006, pp. 4301-4304, doi:10.1109/APS.2006.1711582.

[2] S. Ansolabehere et al., “Precinct-level election data,” Distributed· by Harvard Election Data Archive. Available at: http://hdl.handle.net/1902.1/21919UNF:5:5C9UfGjdLy2ONVPtgr45qA==, vol. 1.” January 20, 2014.

[3] S. P. Bingulac, “On the compatibility of adaptive controllers” in Proc. 4th Annu. Allerton Conf. Circuit Syst. Theor., New York, NY, USA, 1994, pp. 8-16.

[4] M. M. Chiampi and L. L. Zilberti, “Induction of electric field in human bodies moving near MRI: An· efficient BEM computational procedure,” IEEE Trans. Bio Med. Eng., vol. 58, no. 10, pp. 2787-2793, Oct. 2011, doi:10.1109/TBME.2011.2158315.

[5] H. Eriksson and P. E. Danielsson, “Two problems on Boolean memories,” Electron, IEEE, Trans., Devices, vol. Ed., vol. 11, no. 1, pp. 32-33, Jan. 1959.

[6] M. Ito et al., “Application of amorphous oxide TFT to electrophoretic display,” J. Non-Cryst. Solids, vol. 354, no. 19-25, pp. 2777-2782, Feb. 2008, doi:10.1016/j.jnoncrysol.2007.10.083.

[7] Klaus and P. Horn, Robot Vision. Cambridge, MA, USA: MIT Press, 1986.

[8] L. Stein, “Random patterns” in Computers and You, J. S. Brake, Ed. New York, NY, USA: Wiley, 1994, pp. 55-70.

APPENDICES

CODE –

```
from tabula import read_pdf
from tabulate import tabulate
import camelot.io as camelot
import os
import pandas as pd
import numpy as np
import PyPDF2
```

```
# Finding all file names in a folder 2020
file_names = []
for i in os.listdir('2019'):
    # print(i)
    if(i!='failed colleges'):
        file_names.append(i)
file_names


# In[24]:


## Dataframe containing placement data of all colleges
final = pd.DataFrame()
failed_colleges = [] #whose placement data has different shape than others
```

```
k = 0
for file in file_names:
    abc = camelot.read_pdf('2019/'+file, pages="all",flavor='lattice',line_scale=30) # file loation
    # line_space = 30 , this argument helps to read single row tables also. greater the value, higher the chances of detecting smaller tables.
    pdf_file = open('2019/'+file, 'rb')
    read_pdf = PyPDF2.PdfFileReader(pdf_file)
    number_of_pages = read_pdf.getNumPages()
    page = read_pdf.getPage(0)
    page_content = page.extractText()
    text=page_content.split('Name:')
    text = text[1].split('[')[0]
    print(text)
    name = text
    k = k+1
    df = []
    i = 0
    while(i<len(abc)):
        df1 = abc[i].df

        if(i==len(abc)-1):
            df.append(df1)
            break
        df2 = abc[i+1].df
        # merging conditions for placement tables
        condition = df1.loc[df1.shape[0]-1][0]=='Academic Year'
```

```
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2012-13') and (df2.loc[0][0]=='2013-14') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2013-14') and (df2.loc[0][0]=='2014-15') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2014-15') and (df2.loc[0][0]=='2015-16') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2015-16') and (df2.loc[0][0]=='2016-17') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2016-17') and (df2.loc[0][0]=='2017-18') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2017-18') and (df2.loc[0][0]=='2018-19') )
        condition = condition | ((df1.loc[df1.shape[0]-1][0]=='2018-19') and (df2.loc[0][0]=='2019-20') )

        if (condition):
            x = pd.DataFrame(np.vstack((np.array(df1),np.array(df2))))

            df.append(x)
            i = i+2
        else:
            df.append(df1)
            i =i+1
    x = df[1].copy()
    x.columns = x.loc[0]
    x = x[1:]
    x
    p = []
    for i in range(x.shape[0]):
        programme = x.iloc[i]['(All programs\nof all years)']
        p.append(programme)

    for i in p:
        p_name = i
        j = 2
        y = df[j].copy()
        y.columns = y.loc[0]
        j = j+1
        y = y[1:]
        y['Program_name'] = p_name
        y['Institute'] = name
        try:
            final = final.append(y, ignore_index=True)
        except:
            failed_colleges.append(name+p_name)
```

```
# Total-actual-strength
#
# 2. For all the insti find  the ratio of socially challenged (Sc/st, OBC etc) to
#  total number of children as well as economically backward people ?
#
# ================================================================================

import pandas as pd
import matplotlib.pyplot as plt
import warnings
import numpy as np
warnings.filterwarnings('ignore', category=FutureWarning)
import os
```

```
df = pd.read_csv(r"Datasets/total-actual-strength.csv")

df.columns
df = df.groupby('Institute').sum()
df['Institute'] = df.index
df.reset_index(inplace = True,drop = True)
df1 = df[['Institute', 'Total Students','SociallyChallenged (SC+ST+OBC Including male & female)']]
df = df[['Institute', 'Total Students','EconomicallyBackward(Including male & female)']]

df['ratio economic'] = df['EconomicallyBackward(Including male & female)'] / df['Total Students']
df1['ratio social'] = df1['SociallyChallenged (SC+ST+OBC Including male & female)'] / df1['Total Students']
```

```
df.sort_values("ratio economic", inplace=True, ascending=False)


df1.sort_values("ratio social", inplace=True, ascending=False)

df = df[["Institute", "ratio economic"]]
df1 = df1[["Institute", "ratio social"]]


df.to_csv('Outputs/Top institues with highest economically backward-total ratio.csv', index = False)
df1.to_csv('Outputs/Top institues with highest socaially backward-total ratio.csv', index = False)
```

```
# Total-actual-strength
#
# 5.  (Total allowed - actually admitted) percentage?
#
# ================================================================

import pandas as pd
import warnings
warnings.filterwarnings('ignore', category=FutureWarning)

df = pd.read_csv(r"Datasets/total-actual-strength.csv")
home = pd.read_csv(r"Datasets/sanctioned-intake.csv")

df = df[['Institute', 'Programs', 'Total Students']]
home = home[['Institute', 'Academic Year', '2019-20']]
```

```
#df.columns
#home.columns

for index, row in df.iterrows():
    st =  row[1].replace('\n',' ')
    st =  st.replace('Year ','Years ')
    df.at[index,'Programs'] = st

for index, row in home.iterrows():
    st =  row[1].replace('Year ','Years ')
    home.at[index,'Academic Year'] = st

home.columns = ['Institute', 'Programs', 'sanctioned-intake']
df.columns = ['Institute', 'Programs', 'Actual intake']

final = df.merge(home, how='inner', on=['Institute', 'Programs'])
final.columns

final['Actual/Sanctioned'] = ((final['Actual intake'] - final ['sanctioned-intake'])*100 )/final ['sanctioned-intake']
```

```
left = final[final['Actual/Sanctioned']<0]
left = left[['Institute', 'Programs', 'Actual/Sanctioned']]
left.columns = ['Institute', 'Programs', 'Vacancies-Percent']

excessive = final[final['Actual/Sanctioned']>=0]
excessive = excessive[['Institute', 'Programs', 'Actual/Sanctioned']]
excessive.columns = ['Institute', 'Programs', 'Overfill-percent']


left['Vacancies-Percent']=left['Vacancies-Percent'].apply(lambda x: x*-1)
```