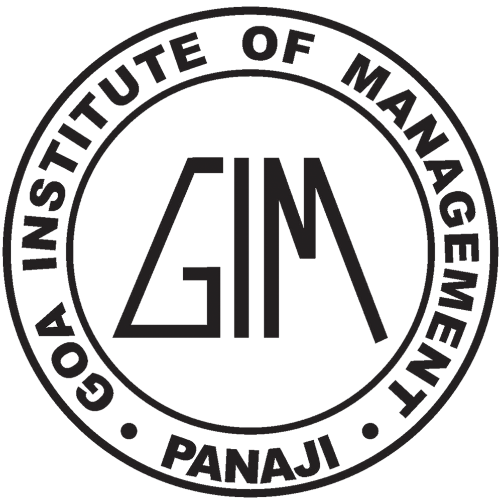
**Analytical Services to MOSP**

****

Faculty Guide: Dr. Alekh Gour

Academic Year: 2021-2022

Students Details:

|  |  |  |
| --- | --- | --- |
| Sr No | Roll No | Name |
| 01 | B2020071 | Darshan Dayanand Desai |
| 02 | B2021072 | Dayne Jose Maria Fernandes |
| 03 | B2021087 | Nihar Sameer Shah |
| 04 | B2021088 | Prakhar Parashar |
| 05 | B2021095 | Saurav Choudhury |
| 06 | B2021098 | Shubham K Jaiswal |

**ACKNOWLEDGEMENT**

We would like to express my deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we give to our Dr. Alekh Gour whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project and achieve our goals.

Furthermore, we would also like to acknowledge with much appreciation the crucial role of Dr Soumen Kumar Manna, who took sharp intrigued on our project work and guided us all along, till the completion of our project work by giving all the fundamental information.

Lastly, we would like to thank MOSPI Team, our client for this project, for their constant guidance and suggestions during this project work, without them this project wouldn’t have been possible.

**Contents**

[**Abstract** 4](#_Toc98429061)

[**1.** **Introduction** 5](#_Toc98429062)

[**Statement of Problem** 5](#_Toc98429063)

[**Significance of the Problem** 5](#_Toc98429064)

[**Objective** 6](#_Toc98429065)

[**2.** **Data and Methodology** 7](#_Toc98429066)

[**Discussions and conversations** 7](#_Toc98429067)

[**Deliverables and Tasks** 7](#_Toc98429068)

[**3.** **Results** 9](#_Toc98429069)

[**4.** **Future Directions** 11](#_Toc98429070)

[**5.** **Hurdles and Bottlenecks** 11](#_Toc98429071)

[**6.** **Conclusion** 12](#_Toc98429072)

[**7.** **Bibliography** 13](#_Toc98429073)

# **Abstract**

Ministry of Statistics and Programme Implementation (MoSPI) is our client for the Give Goa project.

The objective for our project was to research and create a process document that will clearly explain the process of automatic data extraction from different websites, including the technology/skills required, resources to put in place, etc.

We’ve converted the data to the required structured format, using libraries like pyPDF, tabula, beautiful soup, and regular expressions(re). We have extracted the data for six indices, Bank credit, GST collection, Index of industrial production – General, Index of industrial production by Use based, Receipts- Central Govt. operations, and Service.

In the future, we can automate the data extraction from the website and integrating it with the previous data.

# **Introduction**

The Ministry of Statistics and Programme Implementation (MoSPI) is an Indian government ministry that is responsible for the coverage and quality of statistics. The Ministry's surveys are conducted using scientific sampling procedures.

Our major duty consisted of data extraction and data structuring, with input data in PDF or DOC format being converted to the appropriate structured format in Excel.

The act or process of extracting data from typically unstructured or poorly structured data sources for subsequent processing or storage is known as data extraction.

Importing into the intermediate extracting system is hence followed by data transformation and a possible addition of metadata prior to exporting to next stage in the data workflow.

Data Structuring is the programmatic way of storing data so that data can be used efficiently. It is a way of arranging data on a computer so that it can be accessed and updated efficiently.

We used a PDF dataset of Bank Credit, GST Collection, Industrial indices, Revenue, Service from Ministry of Statistics and Programme Implementation.

Tabula-py, Py-PDF, Regular Expression, Selenium were used to convert the PDF datasets to excel format.

## **Statement of Problem**

Ministry of Statistics and Programme Implementation has an enormous repository of databases of the indices from multiple government sectors of India. It becomes very difficult to organize, manage and store such huge volume of data if done manually. Hence, our aim is to input data in PDF or DOC format and convert it into required structured format in excel using data extraction and data structuring so the manual labour gets reduced.

## **Significance of the Problem**

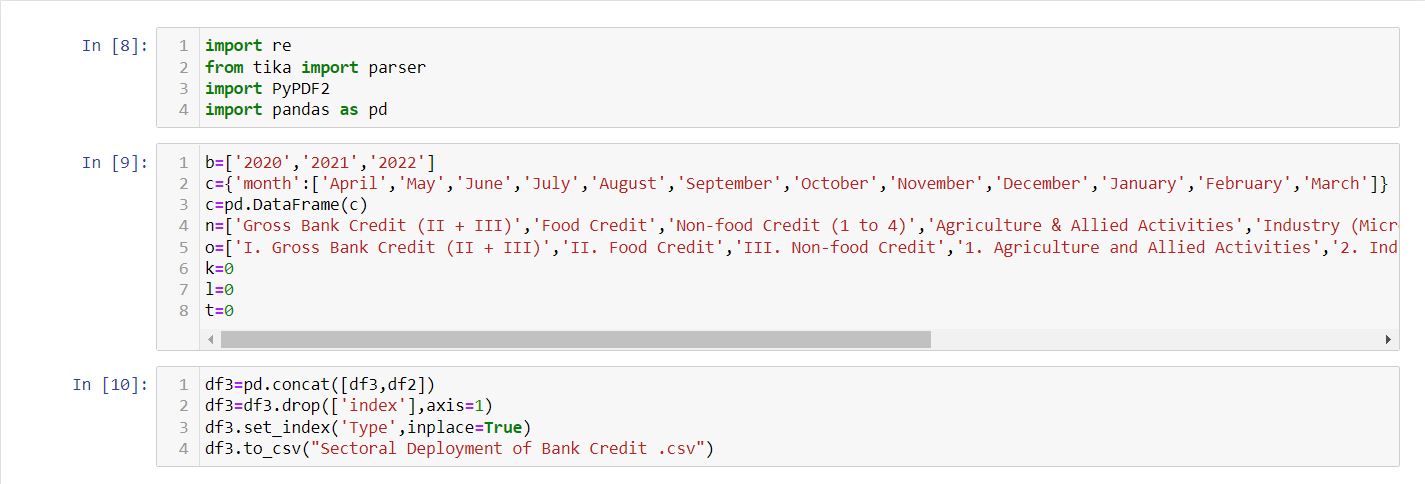
Ministry of Statistics and Programme Implementation (MOSPI) has a huge repository of database of the various indices from different governmental sectors of India.

The data extraction and structuring will be of immense significance for organizing, managing, and storing data efficiently. With the help of data structuring, the data items can be traversed easily. Data structure provides efficiency, reusability and abstraction.

## **Objective**

Data Extraction and Conversion

To do the research and create a process document that will explain clearly the process of automatic data extraction from different websites which includes the technology/skills required, resources to put in place, etc.



# **Data and Methodology**

## 

## **Discussions and conversations**

As discussed above over data was from the Ministry of Statistics and Programme Implementation website (https://www.mospi.gov.in/) in which wide variety of data for various indexes was present in the form of PDF format, and we needed to extract the same in structured format in excel, for this we strategized our steps in the whole process.

Below are the steps that we focused upon:

1. Downloading the data and checking whether the data was in standard PDF format. In some case we also tried to import the pdf file directly from the url link in the meta data excel file.
2. Following that, we concentrated on finding decent Python libraries for converting data into a well-structured format, for example PyPDF2, Tabula, Beautiful soup, Selenium and tried to get the desired result.
3. We tried extracting the table from the few pdfs given in the meta data.
4. We were able to extract table from few pdfs but in many pdf we faced a lot of difficulty and were not able to extract the table.

## **Deliverables and Tasks**

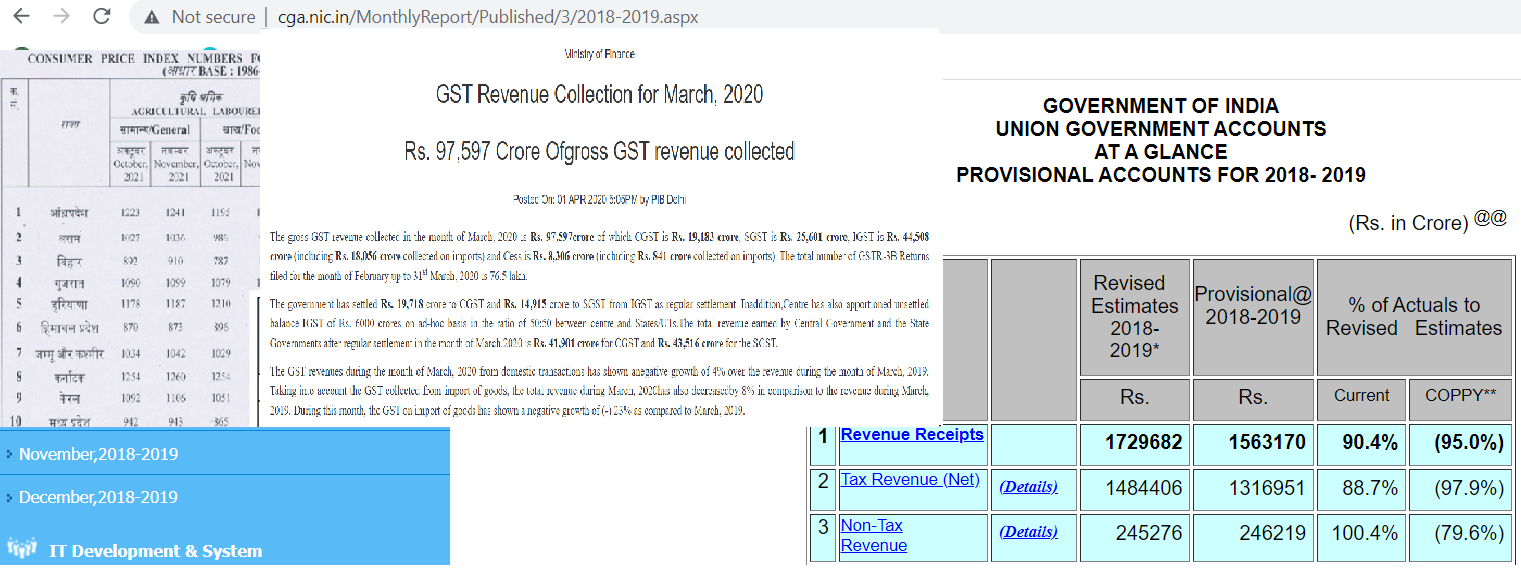
There were different tasks and deliverables that were divided upon different weeks

1. **Observation of data**

We discovered that different types of data were present in the PDFs that were provided to us, i.e., the tables were in various formats, some with different rows and columns, and the textual formats were also varied. They have both written and numerical data in them.

Some of the data that we found were in different languages some were in Hindi and some were in English. In some pdfs there were no table also we have to take out the information for the text. In some there was no pdf we had to extract from web directly.

Below are some of the different kinds of data we observed:



**2. Initial trials with different python libraries**

We came upon different Python libraries that can be used PyPDF, Regularexpression, Tabula Silenium are few of the libraries that we used.

Some pros that we found for PyPDF was that it can split documents page by page, it can merge documents page by page, it can also drop the pages and merge multiple pages into a single page, it can also encrypt and print media files and more.

Some cons that we found was that it cannot extract data from scanned PDF files, it does not have a way to extract images charts or other media from PDF documents and also it cannot write arbitrary text to a PDF.

We have used different libraries to extract the tables which made our work easy

1. **Finalising a PDFs that we can use it for extracting.**

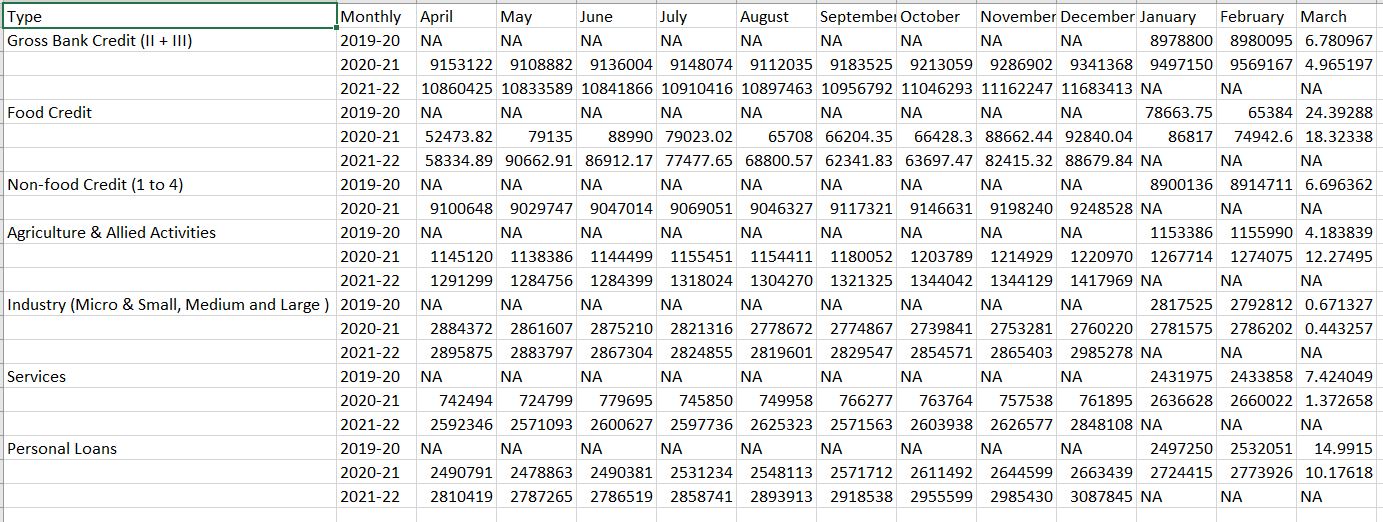
Finalising the PDF was a big issue. We tried extracting the data from all the pdfs but in most of the cases we were unable to extract the data. In many cases we had to extract the data using data cleaning techniques. In some case we had to import the file from website only and based on the import we extracted the table.

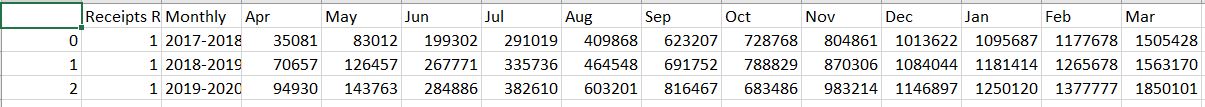
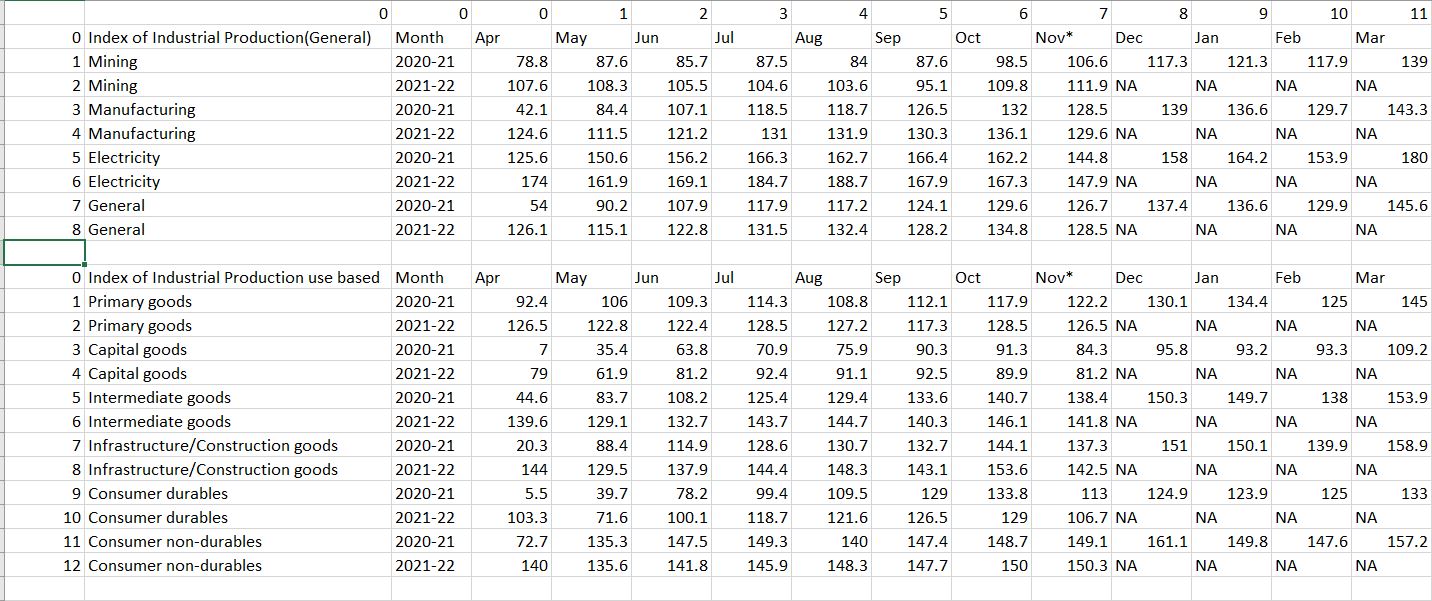
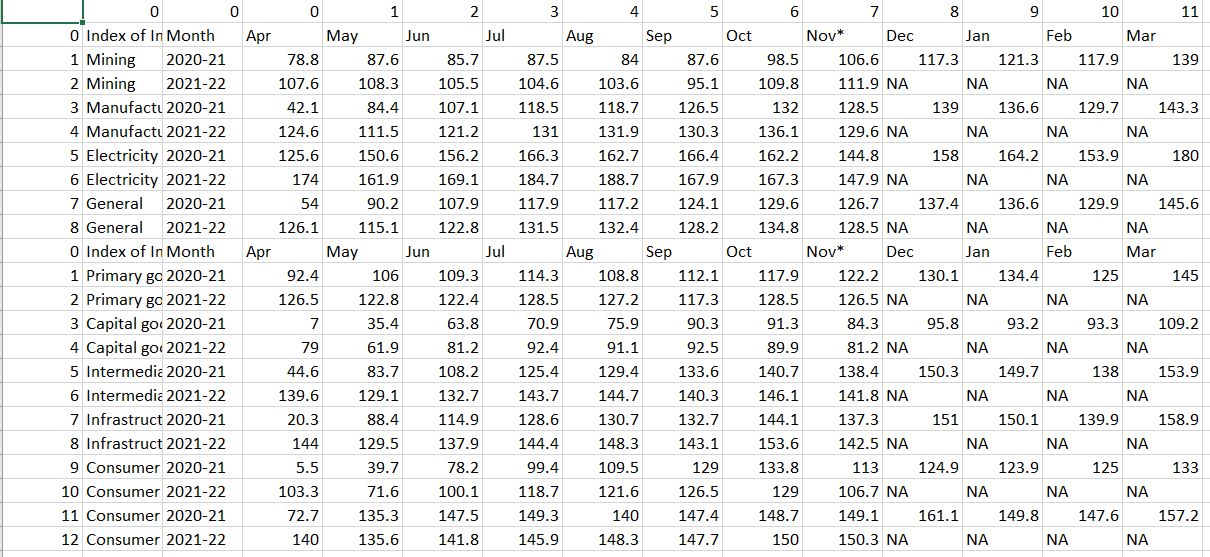
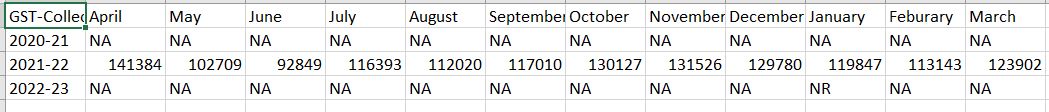
Finally we were able to successfully extract the table from six files bank credit, GST collection, Industrial Indices(General and Use Case) , Revenue and Service.

# **3.** **Results**

We used the below files to for the conversion process (left to right)

1. Bank Credit
2. GST Collection
3. Index of Industrial Production – General
4. Index of industrial production
5. Receipts – Central Govt. operations
6. Service





# **Future Directions**

1. In future we can automate the code and create a program in which the system picks the PDF from folder and convert the same into structure format.
2. Currently the data is in pdf format, in future if the data format changes we would try to incorporate different data format and extract it into csv format.
3. We can rewrite the algorithms for better restructuring of the data and fast processing and structured format of csv file.

# **Hurdles and Bottlenecks**

As we need to convert the data into a properly structured format, we have faced many problems during extraction.

• Lack of Consistency in information:

In many indices like in GST collection data, separate files were found for different periods, which was not the case with other files. Several files were not following proper consistency and it was hard to code for these files. We had to write separate code for each file and hard code for each period.

• Excessive cleaning:

We have gone through excessive cleaning as per the requirement of the code. Because when we had started extracting the data, data wasn’t coming in the desired format, data was cluttered, and different data was coming in the same tab. We have used techniques like text analytics, removed delimiters, and extracted desired numbers from the file.

• Continuous change in file format:

As the files were inconsistent and also formatting is changing every period, it was hard to automate codes. Even the file names were changing which should follow proper conventions to have consistency in the data. To ease the coding procedure, file names, table names, and columns should be in the proper format.

Web scraping wasn’t an easy task, as the files were in different formats like HTML, PDF, and XML. We also found difficulty in finding proper libraries which will give the desired output.

# **Conclusion**

Working with MoSPI was a wonderful experience, we have solved different hurdles and bottlenecks. We were able to successfully extract the data from different files. We have extracted the data for six indices, Bank credit, GST collection, Index of industrial production – General, Index of industrial production by Use based, Receipts- Central Govt. operations, and Service. We’ve used open-source libraries which are freely available and the same codes can be used for extracting the data from other files.

The mentioned codes will read the data and convert it into excel in a structured format. These codes will solve the manual data problem of the client and help them to automate the data entry procedure. It will reduce human intervention and will decrease human error

# **Bibliography**

* <https://www.geeksforgeeks.org/how-to-convert-pdf-file-to-excel-file-using-python/>
* <https://stackoverflow.com/questions/58690461/how-to-convert-pdf-file-to-excel-file-using-python>
* <https://pypi.org/project/tabula-py/>
* <https://www.youtube.com/watch?v=8ZHHnoaltTU>
* <https://blog.api.rakuten.net/top-scraping-apis/#:~:text=3.1%20Share%20this%3A-,What%20is%20a%20Scraper%20API%3F,time%2Dconsuming%20to%20collect%20manually>
* <https://youtu.be/GjKQ6V_ViQE>