Data Collection & Preprocessing Group Assignment

Submitted by :-

Group 10 (Section A)

Jaspreet Singh, PGID 12110053

Varun Mehta, PGID 12110115

Shivam Agarwal, PGID 12110020

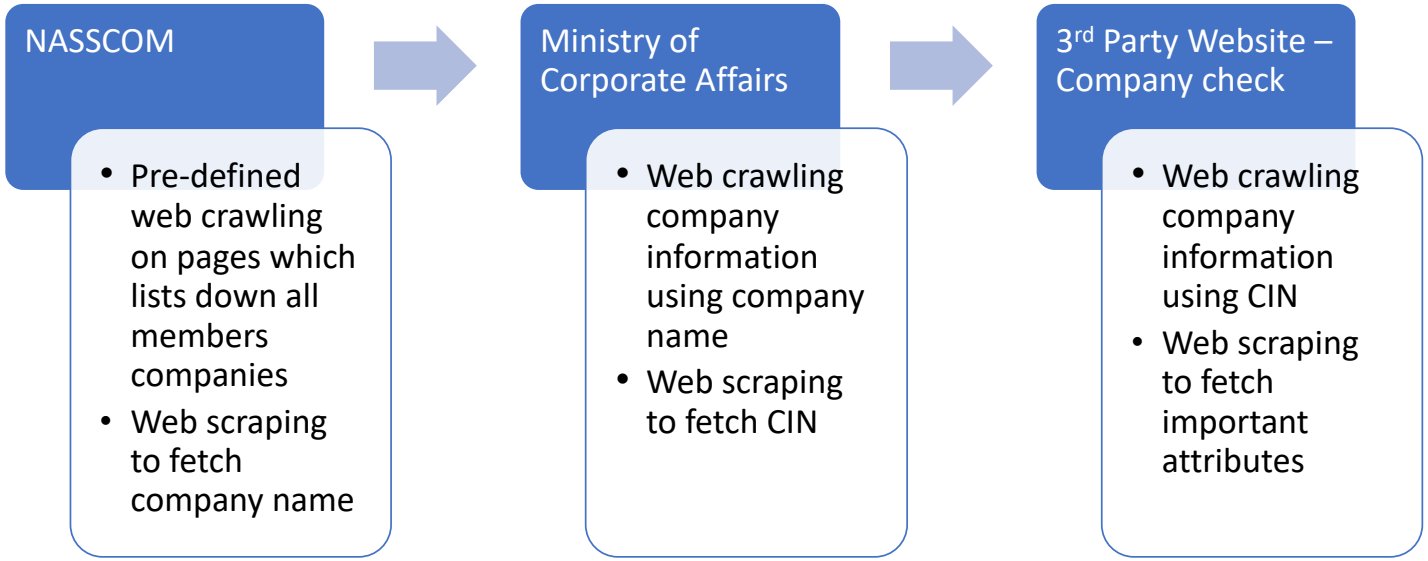
Nikhil Sharma, PGID 12110039

GitHub Reference: https://github.com/varunampba/Group_10_DCCP_Assignment.git

Executive Summary

- **Problem Statement**
  - Find qualitative and quantitative data of NASSCOM companies

- **Proposed Solution**



- **Challenges**
  - Limited and varying data of NASSCOM members if they are not registered in public space like Ministry of Corporate Affair, NSE, BSE.
  - Each company website stores information over webpages which follow different structure and layout. So, scrawling and scraping needs to be different for each company website. Hence, need to identify a series of websites which provide all company data at one place.

Domain & Seed Name

- Choosing Domain
  - Wanted to conduct an analysis of all Indian tech companies
  - Information of all Indian tech companies is not available at one place. Most business websites mostly cover publicly listed companies with silo publications on startups.
  - **NASSCOM** is credible non-profit organization est. in 1988
  - 3000+ tech. companies are listed under NASSCOM

- Choosing Seed Source
  - https://nasscom.in/members-listing
  - Provides list of all member companies across different webpages under a single domain **nasscom.in**

Data Sources
From open web

- NASSCOM website (https://nasscom.in/members-listing)
  - Provides list of all member companies across different webpages under a single domain nasscom.in

- Ministry of Corporate Affairs (https://www.mca.gov.in/mcafoportal/showCheckCompanyName.do)
  - Authentic information
  - Covers a large set of registered Indian companies, both private and public
  - Provides CIN no. which is unique and can be used to derive further insights

- CompanyCheck (https://www.thecompanycheck.com/company/)
  - Free and open source
  - Ethical crawling allowed with no authentication needed
  - Use API from Ministry of Corporate Affairs to derive the data

Data Downloading, Crawling & Collection

- **Methods**
  - Web Page Crawling using Selenium
  - Done across 3 websites – NASSCOM, Ministry of Corporate Affairs and thecompanycheck.com

- **Challenges & Solutions**
  - NASSCOM had stale element error
    - Used try and exception model to capture each element of the webpage
  - Crawling NASSCOM member data from Ministry of Corporate Affairs was difficult
    - Paid subscription : Rs. 100 per company
    - Unethical : Capcha check
    - Explored and identified another website which uses API from Ministry of Corporate Affairs, is open source and free
  - Synchronize the internet speed and crawling of web pages
    - Used library time.sleep

Conversion from Web Pages to Structured Data Fields

Methods

- Identify attributes of interest and collect data as a “list” for each attribute
- Create Data Frame using Pandas function with data captured under list of attributes

Challenges

- List length across attributes shall be equal to prepare a good quality flat file

Solution

- Used “try and exception method” to maintain the alignment between attributes in cases where specific attribute had missing value

Data Cleaning & Pre-processing

1. Removed **insignificant records** by evaluating value of CIN Number
   - Shall be 21 digit long
   - Shall start with letter U (for unlisted companies)
   - Shall start with letter L (for listed companies)
   - Records with any other CIN number are cleaned

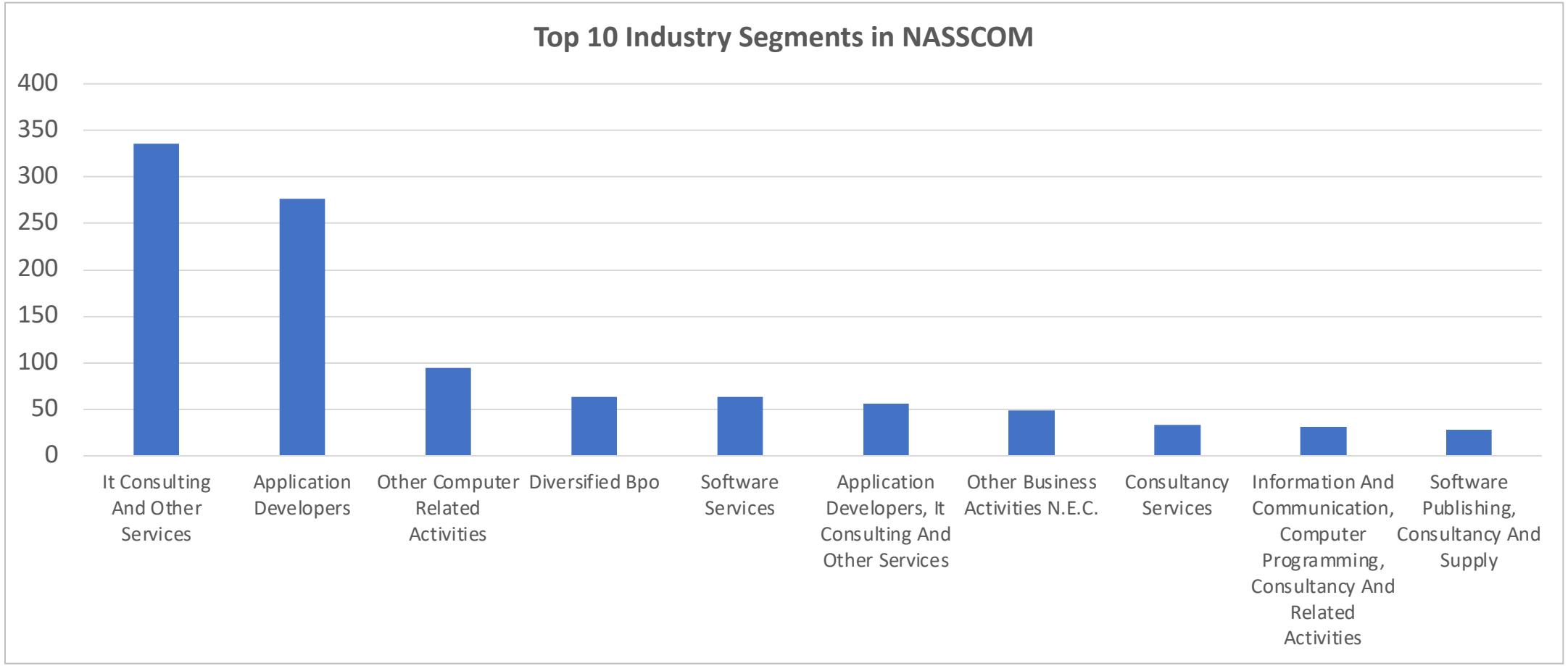
2. Removed **duplicate records**
   - Check if any two records have same CIN
   - Remove one of the record

3. Removed companies where **data is not available**
   - Check if CIN is not available. If CIN is not available in Ministry of Corporate Affairs, no other attributes will be available.
   - Remove records where CIN is not available

Collected quality data for 1780 NASSCOM members

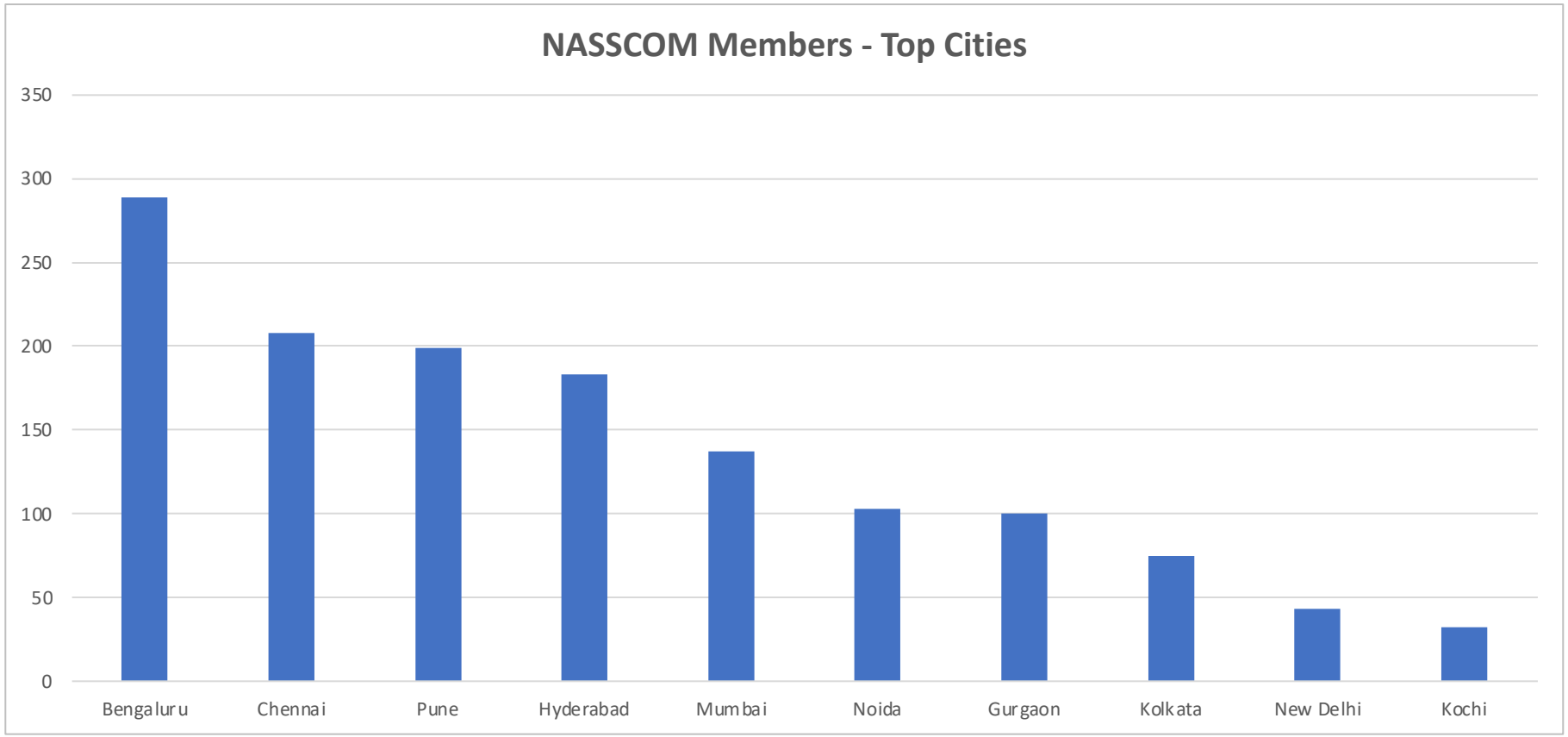
Data Analysis & Insights (1/3)

Active participation in NASSCOM from IT Consulting, Software Application and Computer related institutes



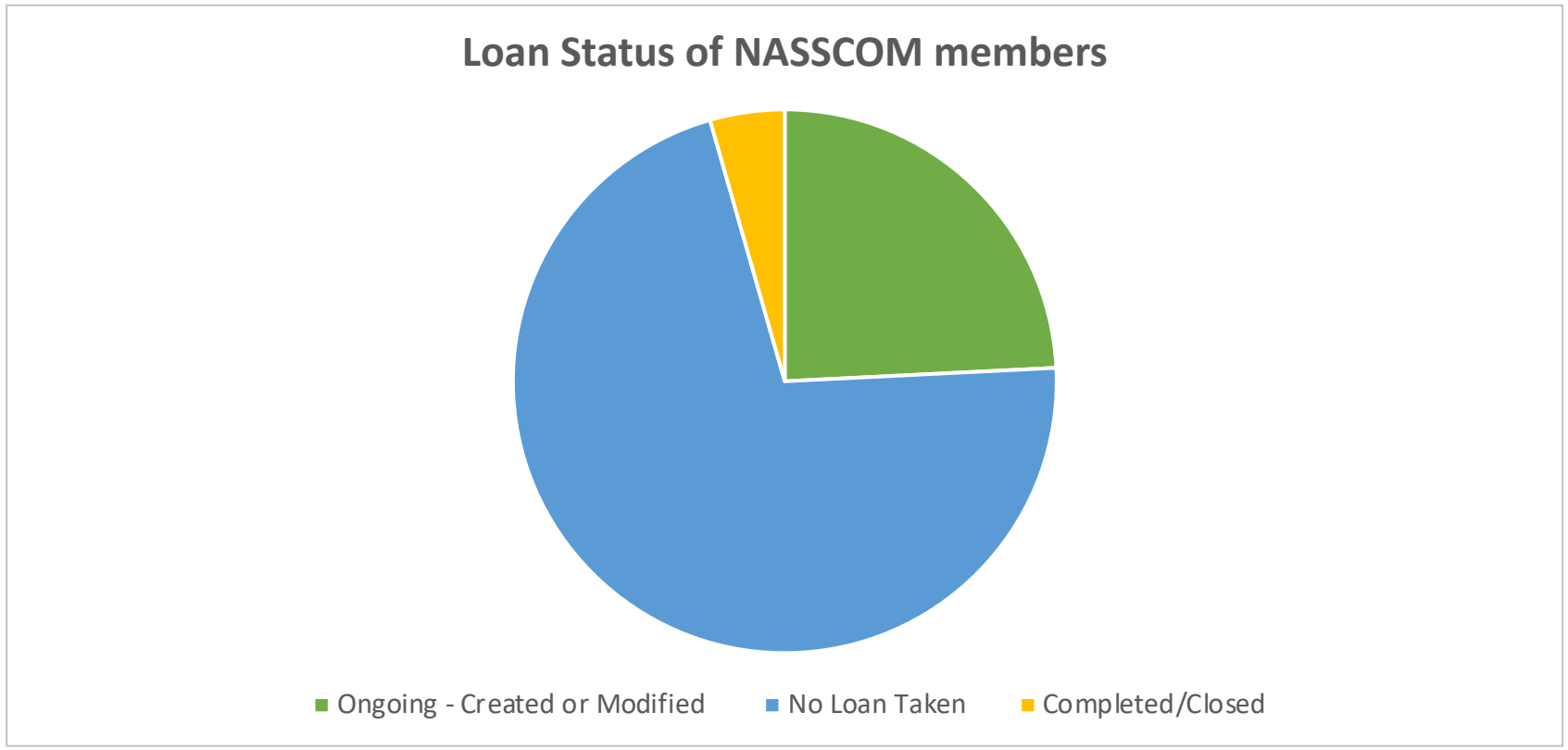
Data Analysis & Insights (2/3)

NASSCOM members are based in Southern cities of Bengaluru/Chennai/Hyderabad, followed by Western Cities of Pune/Mumbai, and northern cities of Delhi/NCR.



Data Analysis & Insights (3/3)

71% NASSCOM members have not taken any loan and 24% have ongoing loan.



Improve data coverage via crowd sourcing and other techniques

- **Increase depth** of collected data via crowd sourcing techniques
  - Do surveys with former employees of targeted companies
    - Capture data on employee satisfaction and happiness index
    - Capture data on HR processes, recognition and rewards
    - Capture data on active participation of NASSCOM in L&D of the members
    - Capture data on innovation index of the company

- **Increase length** of the collected data through paid channels (cannot be achieved via crowd sourcing)
  - Buy subscription-based APIs from Ministry of Corporate Affairs for NASSCOM members

References & Sources

- https://nasscom.in/members-listing
- https://www.mca.gov.in/mcafoportal/showCheckCompanyName.do
- https://www.thecompanycheck.com/
- https://www.indiafilings.com/learn/cin-number-meaning/


thank
you!