ASSIGNMENT SUBMISSION FORM

<u>This will be the first page of your assignment</u>

Course Name : Data Collection and Pre Processing
Assignment Title : Group Assignment
Submitted by : Group 6
(Student name or group names)

Student Name	PG ID
Unnati Khinvasara	12120097
Jhanvi Sharma	12120086
Chiraag Kumar	12120055
Divangana Bansal	12120037
Raktim Srivastava	12120044

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.

If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

(Please start writing your assignment below)

DATA CLEANING & PRE-PROCESSING (DCPP) - TERM 1

Report by Group 6

DOMAIN - UNLISTED COMPANIES

1. Executive Summary

1.1. Problem Statement

The problem statement is to compile a dataset using data collection techniques (like crawling & scraping) on the chosen domain by the group and pre-process it further to present a brief overview.

The domain chosen by the group is data pertaining to ***Companies not listed on the Indian stock exchanges***, *i.e., unlisted companies in India.*

1.2. Brief understanding of challenges

The primary challenge which would be faced is that generally, data of unlisted companies is not publicly available. The Ministry of Corporate Affairs ('MCA') has a repository of companies registered in India. We aim to leverage this source and populate our dataset. However, MCA website has a captcha code requirement for each company and hence it would get difficult to deploy a crawler on this website.

Further, unlisted companies (majorly private owned companies) are not legally obligated to share company data (eg - Annual Reports, Financial Reports, Directors Report, etc) to the public at large and hence only the data submitted by these companies via statutory compliances is available for data collection.

1.3. Proposed solution

We propose to find other similar data sources having unrestricted access, which mirror the MCA database and are accessible for crawling. This would ensure that we get accurate data available on the Unlisted Companies in India.

1.4. Final Dataset Output

We have submitted a final dataset of Unlisted Companies having 20,000+ rows with 26 attributes.

The following output deliverable has been submitted as a part of this project -

- PDF Report on end-to-end data collection pipeline
- GitHub Link - https://github.com/Raktim-Srivastava/DCPP-Project.git
- Code Files in .ipynb format (These are numbered sequentially)
- Json File of final cleaned dataset (Last number in the sequence submitted)

2. Chosen Domain & Seed Sources

2.1. Analysis Conducted

- *Domain Selection. -*
  The domain of Unlisted Companies' was chosen since it is a challenging domain to gather data for, as there is limited availability of data pertaining to these sets of companies.

- *Seed Source Selection -*
  We looked at various websites like Fundoodata, etrace, MCA, Zaubacorp for data collection and finally decided to go ahead with Zaubacorp as our Seed Source. This was because Zaubacorp is a more widely used website which has a high market credibility for providing data directly from the MCA source. It also presented data in a more systemic manner in comparison with other sites.

- *Sample Size Analysis -*
  We have conducted a sample size analysis for a confidence interval of **95%** with a margin of error of **0.01**. This results in a minimum number of sample size of ~ 9500 companies to meet desired statistical restraints. It should be noted that we have considered a higher number of observations **(20K +)** for the dataset.

2.2. Analysis within Seedsource and Process

- We have primarily taken only webpages from **Zaubacorp** as a source for the dataset. We noted that the Zaubacorp website has separate pages for each company and each company's web page was a mix for 'structured' and 'unstructured' data.

  We have filtered out a 'page list' of unlisted companies -

Unlisted companies located in India

Unlisted ✖

2,297,612 Companies Found

Page 1 of 13,333

CIN	Company	RoC	Status
U27106DL2012PTC246642	ASCON JOINTINGS PRIVATE LIMITED	Delhi	Strike Off
U27106DL2012PTC246726	JPC PIPES PRIVATE LIMITED	Delhi	Active
U27106DL2013PTC254154	FOREVER STEELS PRIVATE LIMITED	Delhi	Active
U27106DL2013PTC255946	PARAMHANS WIRES PRIVATE LIMITED	Delhi	Active
U27106DL2013PTC261257	PASCO STEEL & ALLOYS PRIVATE LIMITED	Delhi	Active
U27106DL2014PTC269319	REAL PLASMATECH PRIVATE LIMITED	Delhi	Active
U27106GA1985PTC000642	KAY PEE ROLLING MILLS PRIVATE LIMITED	Goa	Active
U27106GA1994PTC001582	SHIRDI STEEL RE-ROLLERS PRIVATE LIMITED	Goa	Active
U27106GA1994PTC001634	ELLENABAD STEEL PRIVATE LIMITED	Goa	Active

- Then, we realized that if we wanted to take out a sample for further statistical analysis, we would need a random set of companies. Hence out of the total page list of unlisted companies on Zaubacorp (over 13,333 pages), we randomised the pages and crawled a random series of pages. This ensured that we got a random set of pages on which the companies would not be in order.

- With respect to the company-wise details, we have scraped data available from the tables as well as the text. A snapshot of the available data is given below. It should be noted that financial data and other sensitive data is not available for public view and hence we have not considered that for our analysis.

PASCO STEEL & ALLOYS PRIVATE LIMITED

As on: March 29, 2022

Track this company

Basic Information | Documents | Trademarks | Directors | Map

Pasco Steel & Alloys Private Limited is a Private incorporated on 28 November 2013. It is classified as Non-govt company and is registered at Registrar of Companies, Delhi. Its authorized share capital is Rs. 100,000 and its paid up capital is Rs. 100,000. It is inolved in Manufacture of Basic Iron & Steel

Pasco Steel & Alloys Private Limited's Annual General Meeting (AGM) was last held on 30 November 2021 and as per records from Ministry of Corporate Affairs (MCA), its balance sheet was last filed on 31 March 2021.

Directors of Pasco Steel & Alloys Private Limited are Dinesh Mittha Lal Surana, Prema D Surana and .

Pasco Steel & Alloys Private Limited's Corporate Identification Number is (CIN) U27106DL2013PTC261257 and its registration number is 261257.Its Email address is primemetalloys@gmail.com and its registered address is 198, BHAGYA LAXMI APARTMENT SECTOR-9, ROHINI NEW DELHI North West DL 110085 IN , - , .

Current status of Pasco Steel & Alloys Private Limited is - Active.

Company Details

CIN	U27106DL2013PTC261257
Company Name	PASCO STEEL & ALLOYS PRIVATE LIMITED
Company Status	Active
RoC	RoC-Delhi
Registration Number	261257

Financial Report

Balance Sheet	
Paid-up Capital	🔒
Reserves & Surplus	🔒
Long Term Borrowings	🔒
Short Term Borrowings	🔒
Trade Payables	🔒

Contact Details

Email ID: primemetalloys@gmail.com

Website: Click here to add.

Address:

198, BHAGYA LAXMI APARTMENT SECTOR-9, ROHINI NEW DELHI North West DL 110085 IN

Director Details

DIN	Director Name	Designation	Appointment Date	
01917208	DINESH MITTHA LAL SURANA	Director	28 November 2013	View other directorships
02412048	PREMA D SURANA	Director	24 February 2018	View other directorships

3. Coding Procedure

- For all the code files, we have used Jupyter Notebook.
- We have used Python 3 as the programming language

3.1. Crawling Code :

a. Libraries used for running the crawl code are as follows:

Logging, time, json, urllib.parse, url_normalize, requests, bs4, random

b. We used a crawler class and crawled 25,000 links on the website and stored these **HTML links** in a .Json file in order to run it for scraping.
https://www.zaubacorp.com/company-list/listed-Unlisted-company.html

c. We have restricted the crawler to only crawl website hyperlinks within the selected domain in order to filter out any outgoing website links / advertisement pages.

d. In order to get a better clarity for the entire population of 20,00,000 + companies, we thought of randomizing the page numbers just to ensure we get an essence of the entire population in our sample dataset.

3.2. Scraping Code :

a. Libraries used for running the scrape code are as follows:

csv, json, urllib, urllib.request, requests, bs4, time, requests_html

b. Open CSV file to append data, generate dictionary key value pairs in order to append Columns from the website to be scraped using **<td>**, **<p>** and **<table>** tags as per the index numbers.

c. The reason we had to use the logic of index numbers is because of the lack of proper HTML tag in classes and sub categories done on the website.

d. We use Beautiful Soup to extract data from the web pages, we faced challenges majorly in scraping data for 3 no of columns (*Number of Directors*, *No of Prosecution* and *Charges/Borrowing Details*), as these column indexes kept varying between pages because of hidden tables on the webpage.

e. We overcame these problems by counting the number of hyperlinks for the Director Details column (View other Directorships), as every row which had the Director's name also had a link for 'View other Directorships', so in the code we tried to recognize this *string* for counting the number of directors in a company. We then multiplied the *Number of Directors* by 2 and then added it to the indexes of the other 2 columns as we realized for every Director there were 2 no of hidden tables created.

f. For *Charges/Borrowing details*, we went through the table as per the index number of the <td> tag and summed all the entries in the amount column making sure we replace all the commas (,) and empty spaces before converting it to float.

g. We then transformed this into a pandas data frame and converted it into an excel file as well as a .json file (uncleaned).

h. We faced an encoding error while we were trying to scrape the websites. However, we referred to stack overflow, in order to understand about the error and troubleshoot it using 'utf8' as an encoding.

4. Data Cleaning & Pre-processing

We have a rectangular & structured dataset for our domain. For Data Cleaning & Pre-Processing, we checked each attribute of the dataset and measured it against a common logic pertaining for that attribute as follows.

For eg - CIN should be a 21 digit unique number. We noticed that we had certain observations which did not follow this. On further inspection, it was realized that the dataset had 31 observations of LLPs (Limited Liability Partnerships) which is not a part of the domain we have selected. Hence these observations were deleted.

4.1. Preliminary Data cleaning & validation steps performed on the dataset are being consolidated and mentioned below -

Sr No	Column being referred	Observation/ Error / Challenge faced	Remarks / Possible Resolution
1	Number of Members	Should be absolute numbers and Should NOT have Rupee amounts	31 rows found with 'Rs.' Amount. This is caused due to misalignment of indexes for certain 'html' which are LLPs (Limited Liability Partnership) Rows to be removed since we do not want LLPs in the dataset
2	Name	Data set of Unlisted Companies shouldn't contain LLP observations Check for LLP in name	Found 33 Rows. 31 of which are LLPs (same as the rows mentioned in the above cell which are removed)
3	Website	Should contain an http:// link	More than 20k observations contain the text - "*Click here to add*". Further there are some observations with numeric values. We should delete this column as it provides no useful data **Note - We could leverage crowdsourcing though in order to populate this column.*
4	No. of Prosecution	Should contain an integer value	There are 6 rows with -1 value. Replace it 0. As the number of prosecutions cannot be negative.
5	Company Category	Only one row with "-" value.	To be removed as the company category has not been recorded.
6	CIN	Should be 21 digits and not an LLP No. which is triple letter - triple digit	We have already removed the LLP Data with 31 rows in the previous cleaning step (LLP's had alphanumeric values)
7	Activity	Every row contains the garbage string - 'Click here to see other companies involved in the same activity.'	This string to be removed as extra data is captured.

Sr No	Column being referred	Observation/ Error / Challenge faced	Remarks / Possible Resolution
8	Authorized and Paid up capital	- Remove the 'Rs' symbol and convert value to numeric - Create a column and check whether authorised capital is greater than paid up capital or not?	There are 275 observations where authorized capital is not greater than paid up. Since this is logically incorrect, we checked these entries further. After further probing these companies, the database on the MCA website also shows the same pattern. We concluded the data is a typographical error and excluded these observations.
9	ROC	Remove the "ROC-" before the city name to extract exact city	Clean the data and make it easier to filter basis location.
9	Address	To be used for making new attribute - Pincode	Extracted 5 numbers pincode out from the address string, in order to segregate companies based on their location.
10	Activity	Number of activities are very high. This can be further categorized	Categorized on broad activity scale
11	Age of Company		To be converted to number format and to be categorized as per age group in a newly generated variable
12	Paid up capital		Categorisation of Paid up capital

**Note - Since Zaubacorp is a widely used and well maintained database, there are minimal duplication errors / inconsistent identifiers. However, we have tried to check the same for completeness by using the approach of data validation for each attribute.*

4.2. Detailed Steps followed for Cleaning & Validation :

a. *Perform sanity checks for all the attributes of the dataset*.

Sanity checks are done to see whether data captured from the ZaubaCorp website has been extracted correctly and following the expected conventions.

- Sanity checks for Non-Numeric, Numeric, URL's, Email IDs, Date Values has been done on the dataset. Here, we have checked if a column is a Non-Numeric attribute, there should not be any numeric record. For Numeric attributes, all the records should be numbers, integers or floats. Email Column should have "@" sign and URL columns should have "https:" or "http:"

- On performing the above check, for Non-Numerical columns like 'Company Status', 'Category', 'Subcategory', only 0.15% of the records were being highlighted as issues. Thus, we have removed those records from the dataset.

- For all the Numerical Columns, almost all the records have been highlighted as an issue. which could essentially imply that the column has been stored as a string, even when we have *numeric values*. Hence, we need to perform further analysis in order for us to reach any conclusive remark.

- On further probing, we found the column 'Registration Number' which should have been in the number format was saved as *String/Object*. We have changed the same to *numeric values*.

- ‘Age of the Company’ was given as "xx years, xx months, xx days" . We need to convert it into a numeric value (*i.e we have converted it into No. of Years*). For Eg: if the Age of Company was 26 years 10 months 20 days, we have converted it to 26, which is now reflecting the Number of Years, since the Incorporation.

- For some records ‘Number of Members’ = "Login to view previous cins" &"-", which is causing issues with the sanity checks. As we do not have any information regarding the ‘Number of Members’ for more than 90% of the records, we should not consider this attribute for our analysis.

- Format of ‘Authorised Capital’ and “Paid up Capital' were not in the format that we expected it to be in. It was stored as “₹1,000” instead of 1000. Hence, we removed "₹" and "," and converted the attribute to numeric.

- For 252 Records, Paid Up Capital = “Login to view”, which essentially means, information was not scrapped correctly. Hence, we have removed all of the 252 Records/Companies.

- ‘Date of Incorporation’ , ‘Last Annual General’ and ‘Date of Latest Balance Sheet’ were not in date format, we have converted it to date format.

- In the “Activity” Column, we found that for each activity was ending with a garbage string “Click here to see other companies involved in the same activity”. We have removed that particular string from all the records in the Activity column.

- ‘Email ID’ and “Website” records are very few on the Zauba Corp website, so we plan to populate the column through crowd source techniques.

- ‘Address’ is missing just for 1 no of records, we plan to keep it as *blank*.

All *sanity checks* are resolved.

b. *Perform logical tests for the attributes of the dataset*.

- Checked if the ‘CIN’ number is not less than 18 digits for all records.

- Checked if RoC starts with "RoC -".

- Checked if ‘Company Status values’ have consistent Values.

- Checked if ‘Company Category values’ have consistent Values.

- Checked if ‘Company Sub Category values’ have consistent Values.

- Checked if ‘Class of Company’ has consistent Values.

- Checked if ‘Paid up Capital’ < ‘Authorized Capital’.
  (At any point, the paid-up capital of a company can never be more than its authorized capital; we are deleting 22 records where this does not stand true assuming it to be error in database)

- Checked if for every *Numerical column* we have no less than ‘0’ value.

- Checked if for every *Numerical column* we have no less than ‘0’ value.

- Number of Prosecutions cannot be less than ‘0’ logically. So we replace those values with ‘0’.

c. *Updating existing columns/ Creating Additional Attributes from existing Attributes*

- The Second Last element of the Address is the ‘Pincode’ and The Third Last element is the ‘State’, we can extract those and create new Attributes for ‘PINCODE’ and ‘State’.

- Validating if pincode is of 6 digits.

- Changed address attributes to ‘blank’, in case records are not correct.

- Creating age-groups basis age of company.

- Changed address attributes to ‘blank’, in case records are not correct.

- Changed address attributes to ‘blank’, in case records are not correct.

- Created ‘Paid up Capital group’
  (Categorized into buckets).

d. *Export the cleaned .XLS file to .Json file using the python library ‘excel2json-3’.*

.

5. Observations & Insights

Using the dataset compiled, we have tried to get inferences on the unlisted companies domain using Tableau as follows -

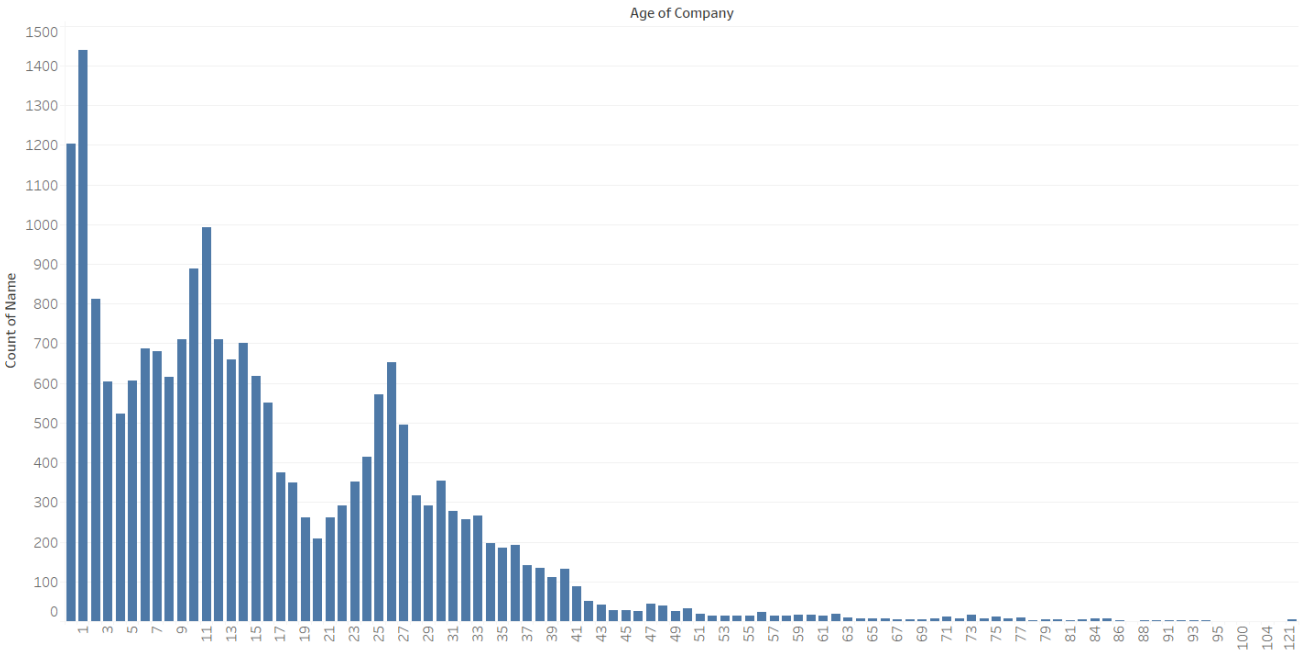
- **Total Companies in the dataset - 20,861**

  We have tried to get a count of companies based on the different attributes and categories made as follows -

  a) Class of Companies

Class of Company	
Private	19,027
Private(One Person Company)	231
Public	1,603
Grand Total	20,861

  b) Age of Companies



Categorisation of Age

Age Group	
1 to 3 Years	2,252
3 to 5 Years	1,126
5 to 10 Years	3,301
10 to 25 Years	7,638
Less than 1 Year	1,203
More than 25 Years	5,341
Grand Total	20,861

c) *Insight based on Company status*

Company Status	
Active	12,379
Strike Off	7,423
Under Process of Str..	364
Amalgamated	267
Converted to LLP	163
Under Liquidation	77
Converted to LLP an..	64
Not available for efil..	53
Dissolved	39
Dormant under secti..	26
Liquidated	6
Grand Total	20,861

The majority of the companies are either Active or Striked off. We can dig in further to gain a better understanding of these company types.

One of the areas we can look at are the number of directors in the active companies.

Number of Directors	Active
0	299
1	269
2	7,031
3	2,779
4	943
5	709
6	181
7	66
8	43
9	24
10	18
11	8
12	6
14	1
15	2
Grand Total	12,379

Based on this table of active companies, we can see that there are around 299 active companies which have nil directors.

We can infer that these companies need to be scrutinised further since active companies are required to have minimum of one / two directors (based on type of company)

d) ROC-wise List of Companies

Ro C	
Delhi	3,549
Mumbai	3,041
Kolkata	1,900
Ahmedabad	1,568
Hyderabad	1,314
Chennai	1,204
Pune	1,142
Kanpur	1,120
Jaipur	1,010
Bangalore	787
Ernakulam	693
Gwalior	534
Vijayawada	523
Chandigarh	485
Coimbatore	333
Cuttack	323
Patna	308
Shillong	237
Goa	157
Jharkhand	142
Chhattisgarh	118
HimachalPradesh	115
Jammu	101
Uttarakhand	100
Pondicherry	47
ANDAMAN	10
Grand Total	20,861

e) Based on Paid-Up Capital

Paid up Capital Group	
0 - 1 Lakh	3,818
1 Crore - 5 Crores	1,152
1 Lakh - 5 Lakhs	10,272
5 Crores - 25 Crores	387
5 Lakhs - 25 Lakhs	3,568
25 Crores - 100 Cror..	105
25 Lakhs - 1 Crore	1,512
100 Crores - 500 Cro..	35
500 Crores +	12
Grand Total	20,861

Other Possible Analysis

Another possible Analysis could be state-wise understanding of companies and their activities. Based on the concentration of activities in each state, the government / incentive policies can be formulated.

Other possible analyses include segregation of companies based on Activity Categorisation., No. of Prosecution, etc.

- **Taking out Proportions to comment on Population (using statistical models)**

Further, using this dataset as a sample, we can calculate proportions based on other attributes and use other statistical analysis tools (like confidence intervals) to comment on the complete population.

Eg - Segregation basis class of companies

Class of Company	
Private	91.21%
Private(One Person Company)	1.11%
Public	7.68%
Grand Total	100.00%

We can see that Private companies make up 91.21%, OPC are 1.11% and other public unlisted companies are 7.68%.

6. Strategy to enhance the data via Crowdsourcing

Crowdsourcing involves obtaining work, information, or opinions from a large group of people who submit their data via the Internet, social media, and smartphone apps. Crowdsourcing work allows companies to save time and money while tapping into people from all over the world. The advantages of crowdsourcing include cost savings, speed, and the ability to work with people who have skills that an in-house team may not have.

A few techniques we can use to enhance this dataset of Unlisted Companies are as follows -

a) Validation by Crowdsourcing

We can cross check the information we have gathered by pooling in the general crowd knowledge. This will ensure that any shortcomings or old data in our dataset is updated.
For eg - We can ask the crowd to select the activity of the company from given options on any crowdsourcing platform. This would ensure that the data we have collected would be cross verified and updated.

b) Creation of Data via Crowdsourcing

We can also take help from the crowd to generate new attributes. For eg, with respect to the activities of the company, we note that currently we have 100+ activities mentioned. To streamline this, we can make a category group of broad industries and then ask the crowd to select the most suited broad category for the company.

c) Targeted Crowdsourcing

We can target a pool from the crowd which would have accurate information of companies.
For eg - Business Consultants, Professionals like lawyers, accountants, advisors, etc. And ask them domain or location specific questions about the companies.

Platforms for Crowdsourcing

For our database of Unlisted Indian Companies, we feel Crowd Creation and Crowd voting can be used to get more information which is reliable and accurate. We can do that by hosting our data on business and financial platforms as follows -

i. Stock market apps (Upstox PRO, Sharekhan etc.) to have a separate section for unlisted companies. It would be beneficial for investors too if they want to invest in unlisted companies.

ii. E-magazine sites (Economic times, Financial Times etc.)

The required information can be collected while registration for seminars and workshops for these apps or e-paper sites. During the event and after the event in feedback/follow up for more seminars and interactive sessions, the companies would be allowed to come and update their details on these platforms thus increasing their reach, accessibility, and rating on various platforms.

The above practices will not only help and motivate the unlisted companies to update the data in our database from time to time but they will also be benefited from it and have ready access to an interested crowd for future reference.

7. References & Sources

- Zaubacorp Domain
- Code shared in Tutorial Material
- Stack Overflow for resolving generic errors in code