DC&PP GROUP ASSIGNMENT

CREATE AN END-TO-END DATA COLLECTION AND PREPROCESSING PIPELINE FOR COMPANY DATA

Group : 11, Section A
AMPBA 2022 S

Shruti Mantri PGID : 12110012
Vikrant Dhawan : PGID : 12110001
Karan Garg PGID : 12110070
Pankaj Madan PGID : 12110102

OUTLINE

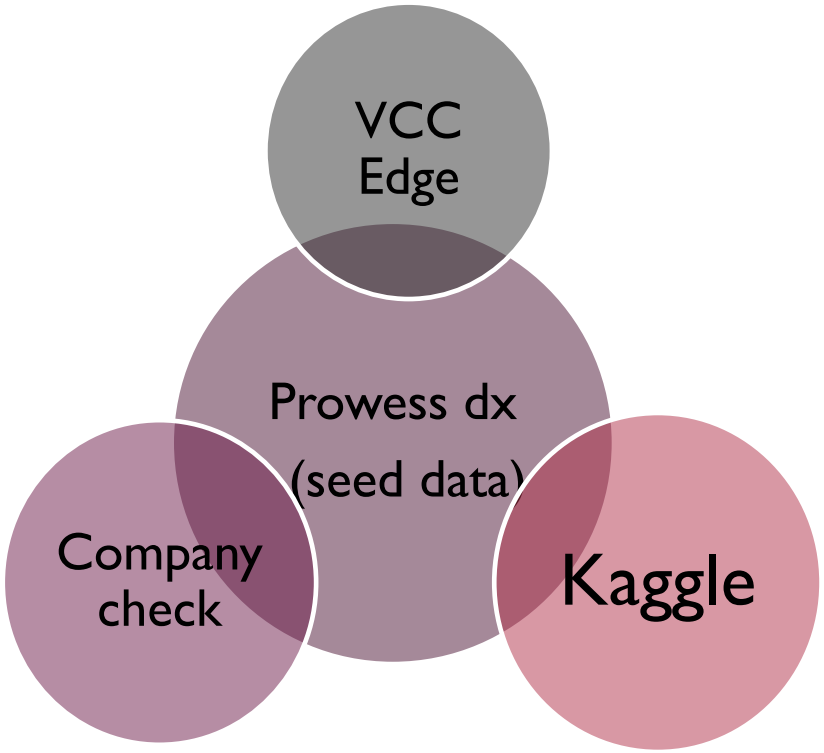
- Executive Summary
- Chosen domain & seed sources (structured and unstructured)
- Data collection approach
- Description of variables
- Download/crawl/collect data from all the sources
- Data cleaning/merge & pre-processing
- Exploratory Data Analysis
- Strategy to enhance the data with crowd sourcing methods
- Conclusions and way forward
- References

EXECUTIVE SUMMARY

Problem Statement	• Domain: “Indian Companies” & explore possible sources of seed set of structured and unstructured data • Use seed /similar sources to extract data using various techniques such as data scraping, web crawling etc. • To collect high quality data of companies with 50+ attributes.
Proposed Solution	• Use different Scraping libraries/ tools such as LXML, selenium, beautiful soup etc. to get the required data. • Merge the data to create a unified knowledge base of the data collected from Multiple sources based on “Corporate Identification Number” – primary key. • Store the data in excel sheets • Clean-up of data including removal of duplicates • Deriving meaningful variables from unstructured data. • Perform Exploratory Data Analysis by using libraries like Sweetiz (python), Excel • Define a Crowd sourcing strategy and methods to enhance the data.
Brief understanding of the challenges	• Multiple data formats within the same seed source (https://business.mapsofindia.com/india-company) • Most of the critical information was paid/ not easily available (e.g. Net Sales, PAT, EBITDA, Total debt etc.) • Many of the attributes were not relevant /null ,we had identified such attributes & removed from the dataset. • We could have used VCC Edge to extract Financials but owing to privacy issues and Contract violations only could only extract a sample set (~300 records), the Code can be extended to run for other companies. • Many of the companies were not common in different data sets and hence there are some null values in two of the attributes

CHOSEN DOMAIN AND DATA SOURCES (STRUCTURED & UNSTRUCTURED)

- The domain "Indian Companies" was primarily chosen to get meaningful insights that this domain has to offer.
- Further we wanted to explore varies dimensions such as regional distribution , commercially dense Industrial regions etc. and basic exploratory data analysis.



Chosen

Prowess dx — Seed data

The CompanyCheck

VCC EDGE

kaggle

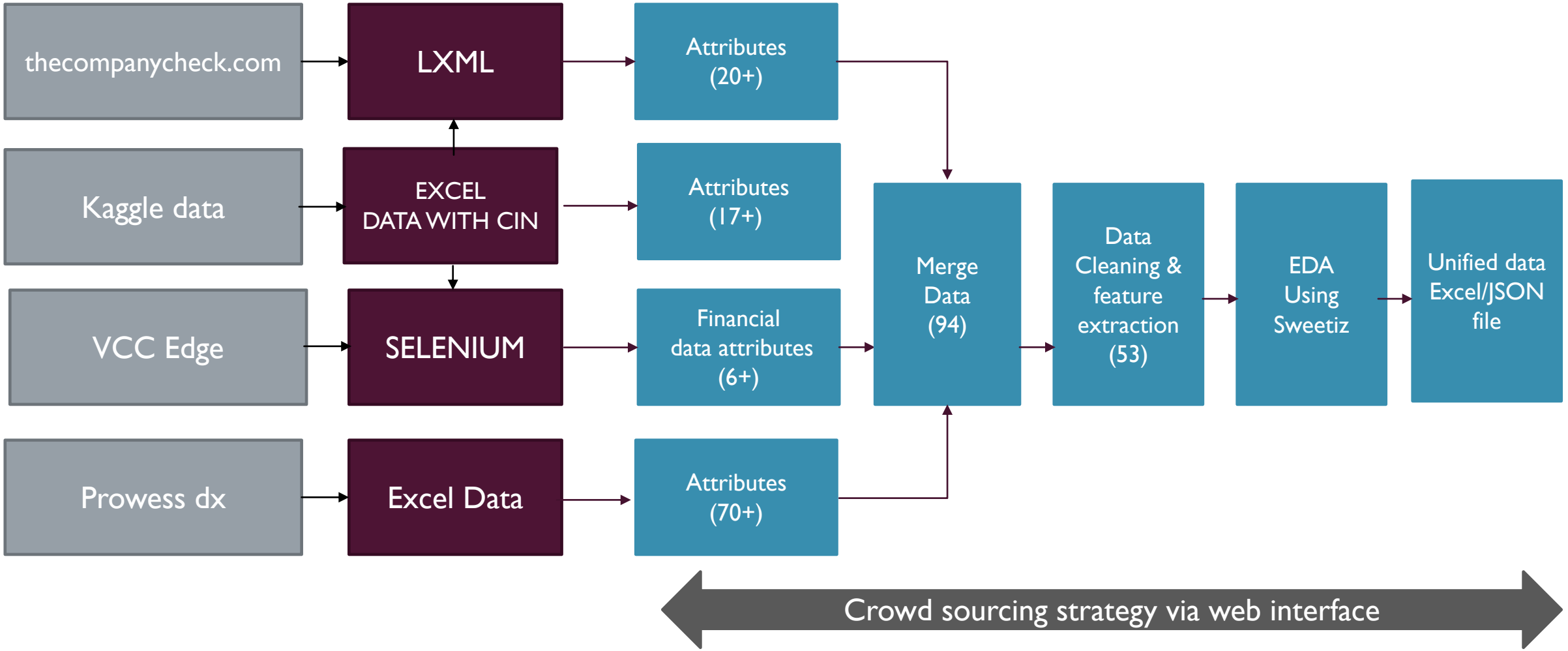
Other data sources (not chosen)

Zauba Corp — Multiple data formats within the same seed source

The CompanyCheck — Paid information could not be extracted

Business MapsofIndia.com — Multiple data formats within the same seed source

DATA COLLECTION APPROACH



DESCRIPTION OF VARAIBLES

Prowess$_{dx}$

Attribute Description	Attribute name	Data Source
Prowess company name	company_name	Prowess
MCA's CIN code	CIN	Prowess
State code	state_code	Prowess
ROC registration number	registration_no	Prowess
Entity type code	entity_type_code	Prowess
Entity type	mr_entity_type_name	Prowess
Ownership code	owner_code	Prowess
Ownership	owner_gp_name	Prowess
Industry type	co_industry_type	Prowess
Main product/service code	co_product_gp_code	Prowess
Main product/service name	product_name_mst	Prowess
Industry group code	co_industry_gp_code	Prowess
Industry name	co_industry_name	Prowess
NIC tree code	co_nic_code	Prowess
NIC code	nic_prod_code	Prowess
NIC name	nic_name	Prowess
Incorporation year	incorporation_year	Prowess
Age code by year of incorporation	age_code	Prowess
Age category	age_group	Prowess
Size code by deciles	decile_size	Prowess
Size by deciles	decile_size_group	Prowess
NSE symbol	nse_symbol	Prowess
BSE scrip code	bse_scrip_code	Prowess
BSE code	bse_code	Prowess
BSE scrip id	bse_scrip_id	Prowess
Registered office address	regdaddr	Prowess
Registered office district code	regddcode	Prowess
Registered office district	regddname	Prowess
Registered office state	regdstate	Prowess
Registered office pincode	regdpin	Prowess
Registered office telephone number/s	regdtele	Prowess
Registered office email address	regdemail	Prowess

The CompanyCheck

Attribute Description	Attribute name	Data Source
Paid Up Capital	Paid up Capital(in lakhs)	CompanyCheck
Registered State Name	State	CompanyCheck
Registrar of Companies Code	RocCode	CompanyCheck
Listed / Unlisted Company	ListingStatus	CompanyCheck
Industry	Industry	CompanyCheck
Company Details	CompanyDetails	CompanyCheck
Number of Open Loans	OpenLoans	CompanyCheck
Total Secured Amount	TotalSecuredAmt	CompanyCheck
WebSite URL	Website	CompanyCheck
Company Age	CompanyAge	CompanyCheck
Number of Directors	Directors	CompanyCheck
Status of Company	Status_N	CompanyCheck
Registered Company Address	Address_N	CompanyCheck

The CompanyCheck kaggle

Attribute Description	Attribute name	Data Source
Company Class	companyClass	CompanyCheck & Kaggle
Company Category	companyCategory	CompanyCheck & Kaggle
Company Sub Category	companySubCategory	CompanyCheck & Kaggle
Authorised Capital	AuthorizedCapital	CompanyCheck & Kaggle
Incorporation Date/ Establishment Date	IncorpDate	CompanyCheck & Kaggle

kaggle

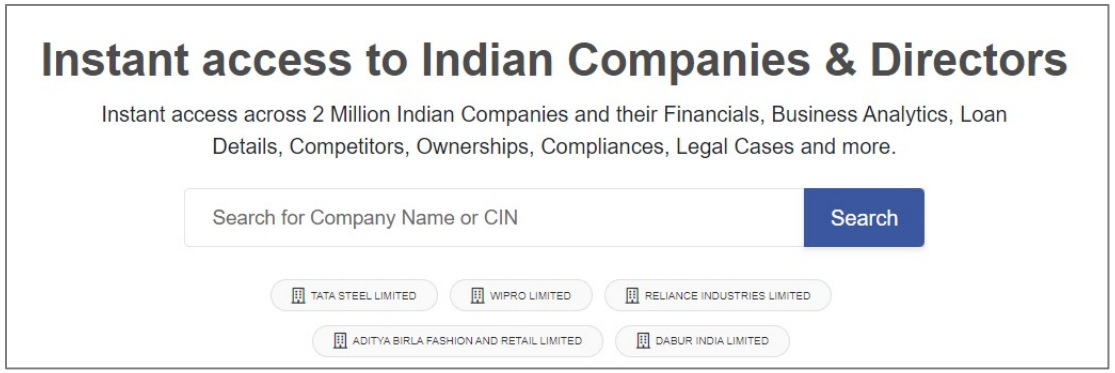
Attribute Description	Attribute name	Data Source
Company Status Code	companyStatus	Kaggle
Industrial Class	IndClass	Kaggle
Principle_business_activity_as_per_cin	ActivityCIN	Kaggle

WEB CRAWLING FROM COMPANY CHECK (LXML)

LXML , Python library which allows for easy handling of XML and HTML data
Xpath is used to extract attributes from theCompanyCheck website.
A Recursive “for” Loop extracts data for companies using CIN from “prowess dx” as a baseline.

```
dx=[]
for row in range(70, 500):
    CIN = sheet['A' + str(row)].value
    CNAME = sheet['B' + str(row)].value.strip().replace(' ','-')
    urlC='https://www.thecompanycheck.com/company/'+ CNAME+'/'+ CIN
    values = {'username': 'shrutijhawar13@gmail.com',
          'password': 'isb2022$'}

    page = requests.get(urlC, data=values)
    if (page.status_code==200):
        print("Data Extracted",CNAME+'/'+ CIN)
        tree = html.fromstring(page.content)

        #NSE
        StockSymbolNSEText=(tree.xpath('//*[@id="About"]/div[4]/table/tbody/tr[3]/td[2]/div/div[2]/span/text()'))
        if len(StockSymbolNSEText) == 0:
            StockSymbolNSE=""
        else:
            StockSymbolNSE=StockSymbolNSEText[0].strip()

        #BSE
        StockSymbolBSEText=(tree.xpath('//*[@id="About"]/div[4]/table/tbody/tr[3]/td[2]/div/div/span/text()'))
        if (len(StockSymbolBSEText) == 0):
            StockSymbolBSE=""
        else:
            StockSymbolBSE=StockSymbolBSEText[0].strip()

        #ROC
        RocCodeText=(tree.xpath('//*[@id="td_roc"]/text()'))
        if (len(RocCodeText) == 0):
            RocCode=""
        else:
            RocCode=RocCodeText[0].strip()

        #CompanyNumber
        CompanyNoText=(tree.xpath('//*[@id="td_regnumber"]/text()'))
        if (len(CompanyNoText) == 0):
```

Webpage



WEB CRAWLING FROM VCC EDGE (SELENIUM)

```
In [ ]: %%capture

        !pip install -U selenium
        !pip install -U webdriver-manager
        !pip install pandas
```

```
In [ ]: from selenium import webdriver
        import pandas as pd
        from lxml import html
        import requests
        from selenium.webdriver.common.by import By
        from selenium.webdriver.support.ui import WebDriverWait
        from selenium.webdriver.support import expected_conditions as EC
```

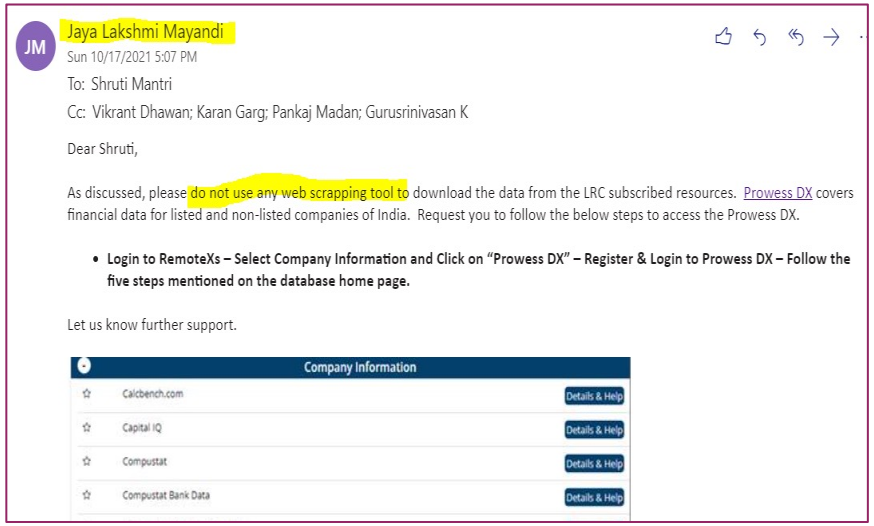
```
In [ ]: import openpyxl
        import os
        os.getcwd()
        file = 'ProwessSeed.xlsx'
        d1 = pd.ExcelFile(file)
        print(d1.sheet_names)
        df = d1.parse('Sheet1')
        df.info
        df.head(10)
```

```
dx=[]
n=0
while(n<300):

    driver = webdriver.Chrome('C:/Users/shrut/Downloads/CD/chromedriver.exe')
    driver.get("https://www.vccedge.com/login.php")
    driver.find_element(By.XPATH,"//*[@id='edit-name']").send_keys('Jaya_Lakshmi@isb.edu')

    driver.find_element(By.XPATH,"//*[@id='edit-pass']").send_keys('isblrc@21')
    driver.find_element(By.XPATH,"//*[@id='user-login']/div[3]/button").click()
    WebDriverWait(driver, 10)

    print("N",n)
    for row in range(n+1,n+101):
        print("row",row)
        # Get CIN Number from Excel
        CIN = sheet['A' + str(row)].value
        WebDriverWait(driver, 1).until(EC.element_to_be_clickable((By.XPATH, "/html/body/div[2]/header/nav/d
        driver.find_element(By.XPATH,"//*[@id='header_nav_toolbar']/div/div/div[3]/div/button").click()
        try:
            WebDriverWait(driver, 1).until(EC.element_to_be_clickable((By.XPATH, "/html/body/div[2]/div/div/
            driver.switch_to.window(driver.window_handles[1])
            # year
            try:
                objYear=driver.find_element(By.XPATH,"//*[@id='home']/div/table/tbody/tr[1]/td[1]")
                year=objYear.text
                print(year)
            except:
                year=""
```

Have used Selenium to extract data from VCC Edge. However due to privacy and Contract Violation issues did not extract Data for all 51809 records.

We had extracted sample set of 300 records to illustrate the process

Sample Output File is attached.



JM Jaya Lakshmi Mayandi
Sun 10/17/2021 5:07 PM

To: Shruti Mantri
Cc: Vikrant Dhawan; Karan Garg; Pankaj Madan; Gurusrinivasan K

Dear Shruti,

As discussed, please do not use any web scrapping tool to download the data from the LRC subscribed resources. Prowess DX covers financial data for listed and non-listed companies of India. Request you to follow the below steps to access the Prowess DX.

- **Login to RemoteXs – Select Company Information and Click on "Prowess DX" – Register & Login to Prowess DX – Follow the five steps mentioned on the database home page.**

Let us know further support.

DATA MERGE

Various Extracted Data files from Company Set and Kaggle were merged using left join with Prowess seed Dataset.

```
#Merging Datasets
f1 = pd.read_excel("ProwessSeed.xlsx")
f2 = pd.read_excel("ExDataCC.xlsx")

# merging the files
f3 = pd.merge(f1, f2[['CIN','companyClass','StockSymbolNSE','StockSymbolBSE','RocCode'

# creating a new file
f3.to_excel("FDSet.xlsx", index = False)
```

DATA CLEANING AND FEATURE EXTRACTION

The data cleaning and preprocessing was done to achieve two below mentioned objectives:

- Extract categorical information from variables with subjective data
- Wrap up the variables with incomplete information so that they can be stored directly in the final structured database
- Removing Spaces and Special Characters from the extracted data.

1.) In the extracted data from www.companycheck.com ,"CompanyDetails" variable had the data in form of the following sentences:

" It is an **Active company** established on 01 Apr 1971 with its office registered at **Godrej Coliseum, Office No.801,C-Wing,Behind Everard Nagar,Off Somaiya Hospital Road, Sion East Mumbai Mumbai City Mh 400022 In** and has been running since 50 years 6 months with a **paid up capital of 3.75 cr.** According to MCA records, 3 Directors are linked to this company as of 21 Jul 2021 "

The aim of the exercise was to extract the bold marked sections in the above sentence which would refer to the Status, Address and Paid up Capital of the company, respectively. The base snippet used for the procedure is the following:

```
data['Status'] = data['CompanyDetails'].astype(str).apply(lambda st: st[st.find("an")+3:st.find("established")])
```

2.) The variables storing the website information did not have "www." in front of the data so the information could not be used directly.

Variables: Website , Regdemail , Corpemail

"www." was added Infront of the data so that the information can be used directly. The base snippet used for the procedure is the following:

```
data['Website'] = "www." + data['Website']
```

DATA CLEANING AND FEATURE EXTRACTION(CONT.)

3.) The values extracted from www.companycheck.com , had black spaces, new line character appended at the leading and trailing sides and special characters like "₹ and , ".These were removed while extracting the variables and before storing them in the data frame.

Also, some variables like Number of directors were converted from String values to integer values so that analysis could be performed on them.

The base snippet used for the procedure is the following:

```
Directors=int(DirectorsText[0].replace('\n',""))

TotalSecuredAmt=TotalSecuredAmtText[0].replace('\n₹ ',"").replace('\n',"").replace(',',"").replace('₹ ',"")
```

EDA (PROCESS)

- **Eliminated attributes having large missing values**: Reduced attributes from 90 to 54
- **Used Sweetwiz in python for Exploratory Analysis**
- **Analysis of important attributes…**

Code:

pip install sweetviz

import pandas as pd
import sweetviz as sv
df=pd.read_csv('FinalDS_v2.csv')
df.info()
report=sv.analyze(df)
report.show_html('Final.html')

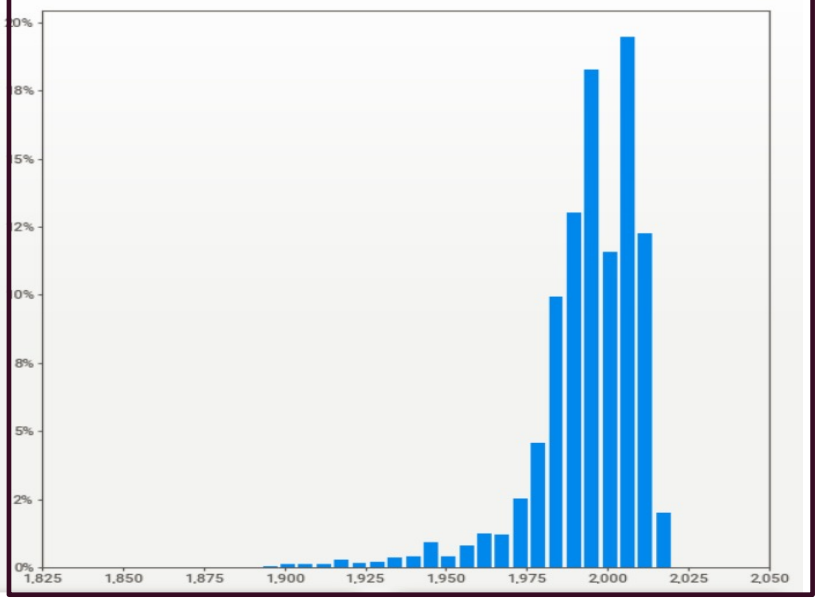
```
In [ ]: pip install sweetviz

In [ ]: import pandas as pd
        import sweetviz as sv

In [ ]: df=pd.read_csv('C:/Users/karan/ISB Study/Term 2/Data Collection/project/DC Final/FinalDS_v2.csv')

In [ ]: df.info()

In [ ]: report=sv.analyze(df)

In [ ]: report.show_html('C:/Users/karan/ISB Study/Term 2/Data Collection/project/DC Final/karan(Final).html')
```

ANALYSIS OF IMPORTANT ATTRIBUTES

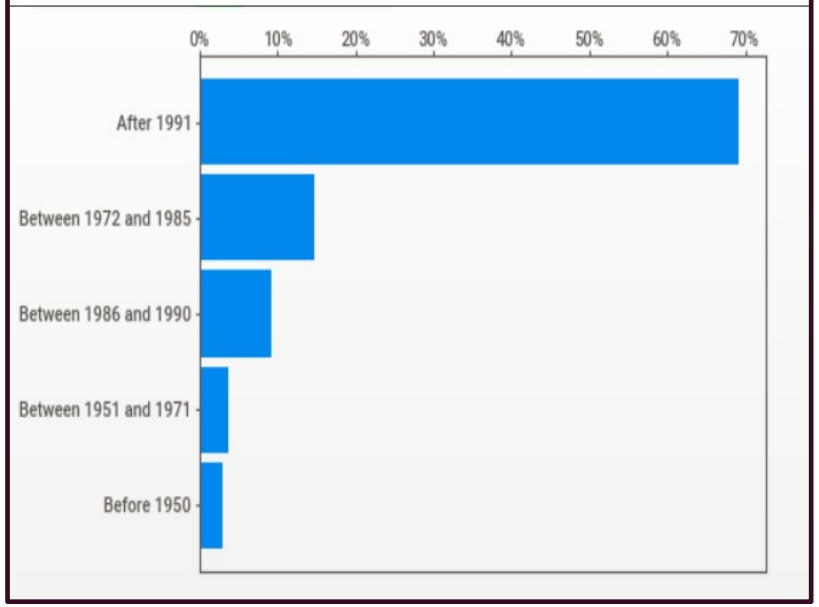
Most of the companies are incorporated b/w 1975 and 2010, 70% of them were registered after 1991 with 50% of them in Maharashtra and Delhi

Detailed EDA

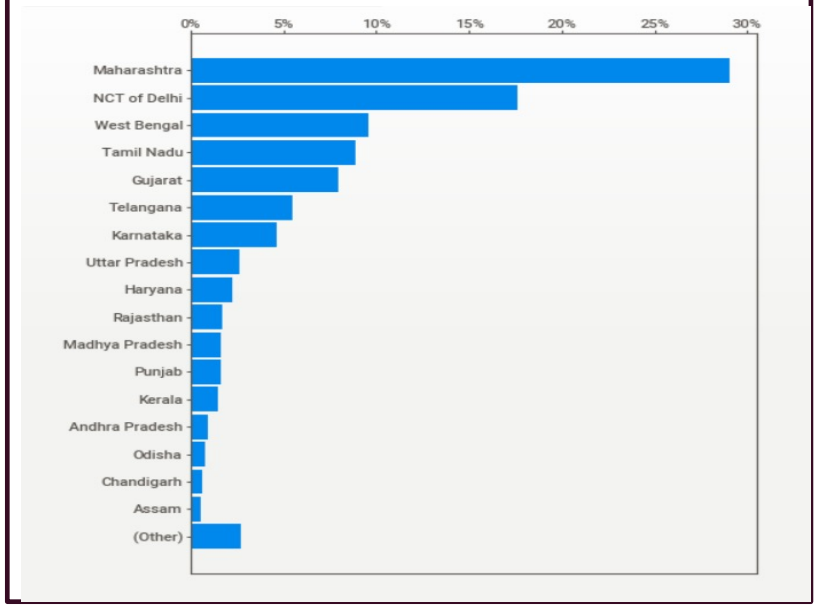
Incorporated Year:
Most of companies were incorporated b/w 1975 and 2010



Age Group:
70% companies started after year 1991



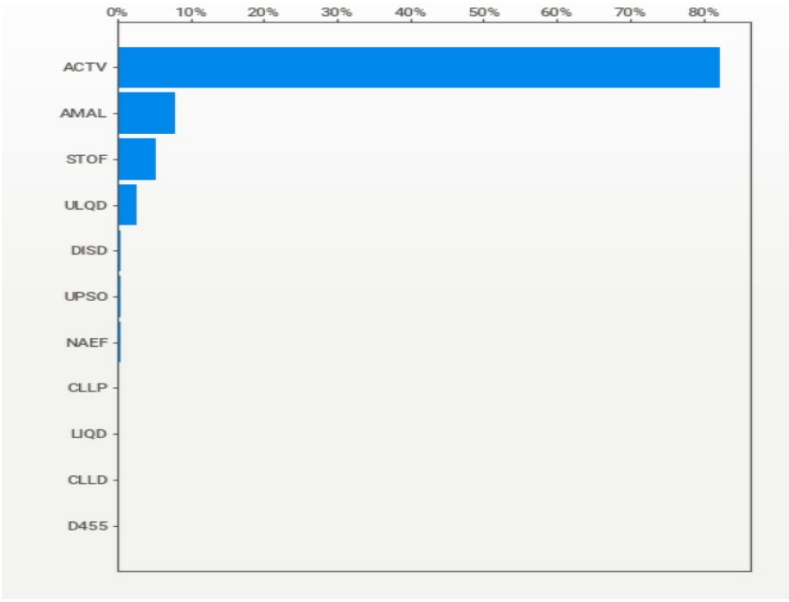
Registered State:
50% of companies are registered in Maharashtra and Delhi



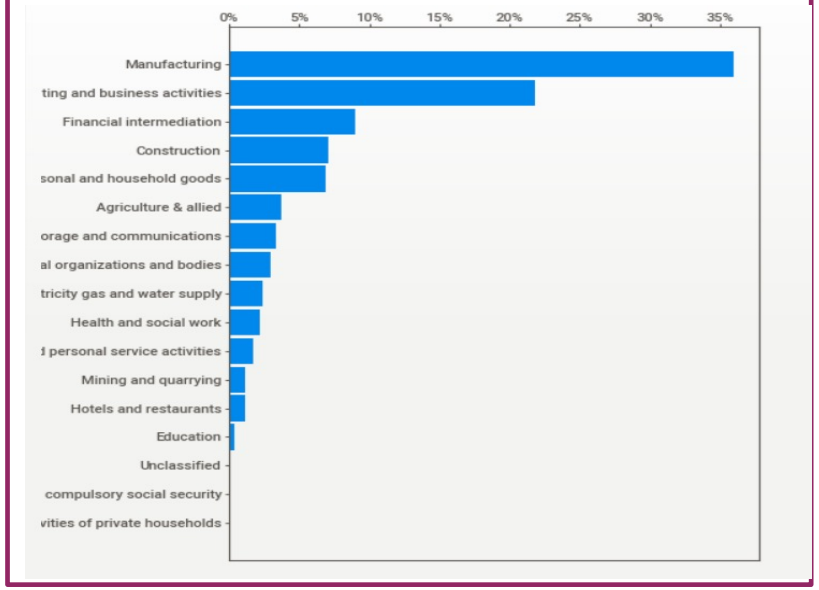
ANALYSIS OF IMPORTANT ATTRIBUTES

80% of the companies are in active status,70% of them belong to manf.,real state, const. and financial and only 12% of all the companies are listed on either BSE and NSE

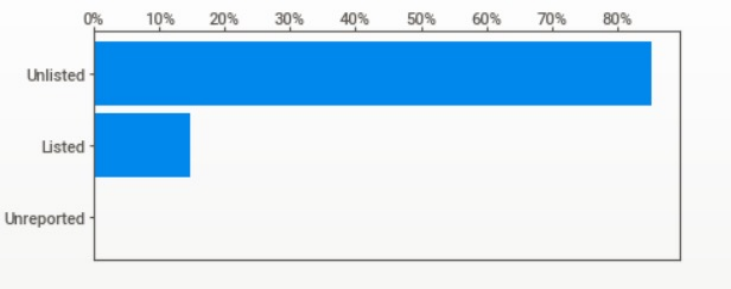
Company Status:
80% of the companies are in active status



Company Activity:
70% companies are from Manuf., Real state, Const and Financial



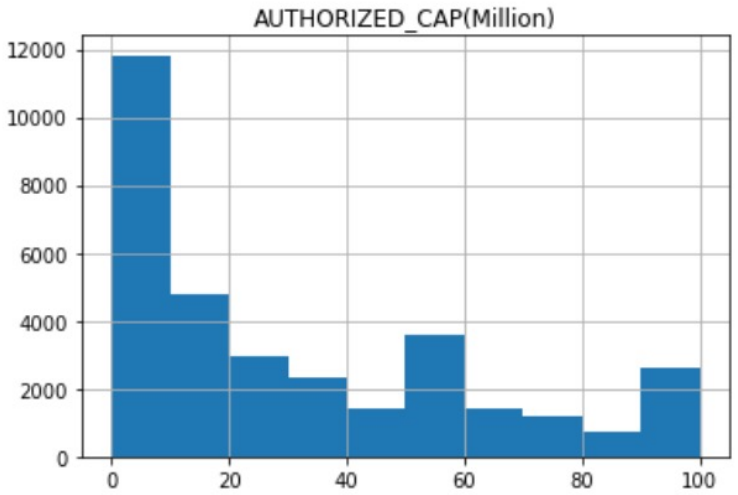
Listing status:
Only 12% companies are listed on either BSE or NSE



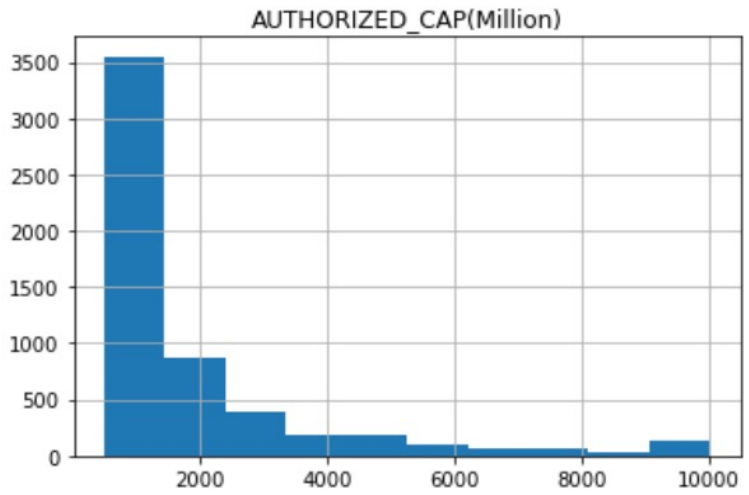
ANALYSIS OF IMPORTANT ATTRIBUTES

Authorized Capital defines a company's allowed capacity to withdraw capital by listing its shares to Public- 80% companies have Less than 100 Million INR out of which 20% have used almost all the capacity. Only 10% companies have not touched their authorized capital

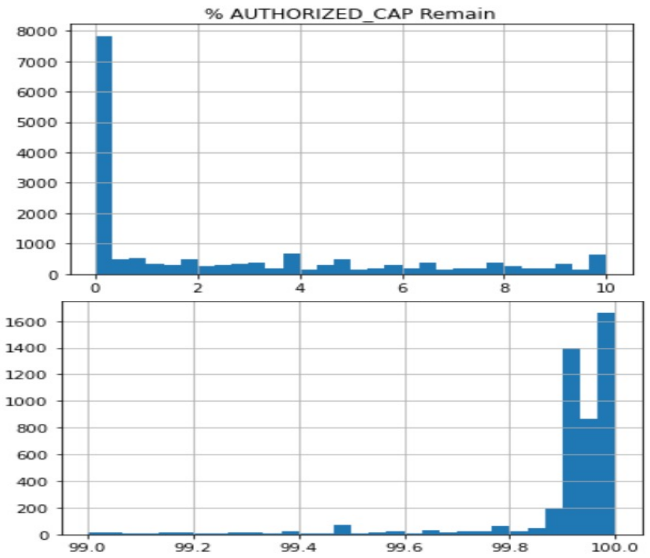
Authorized Capital(Million)
80% of the companies have authorized capital less than 100 Million INR



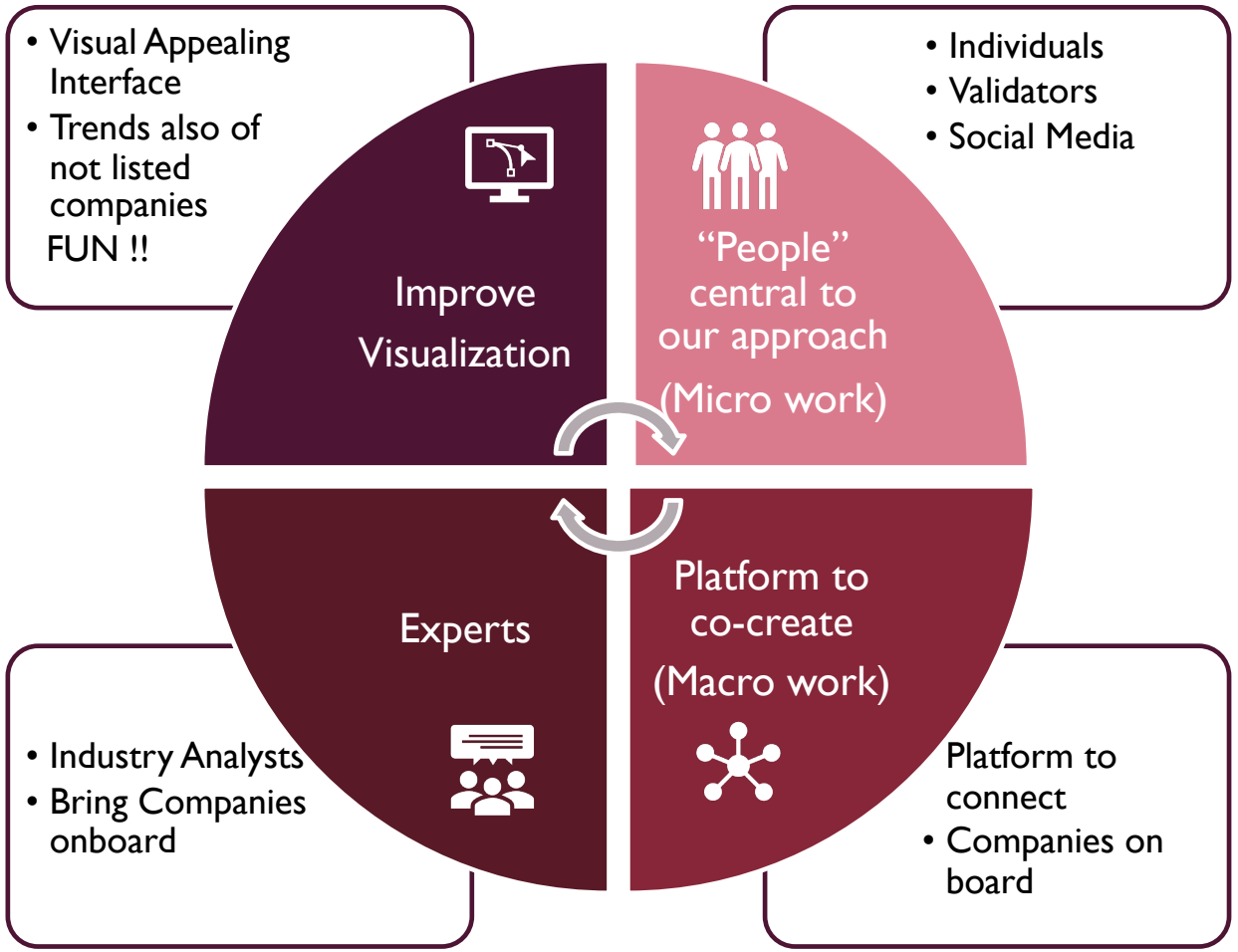
10% Companies have Authorized Capital greater than 500 Million INR



% Of Capital Remain:
20% Companies have used most of the authorized capital and 10% have almost 100% capital remain



CROWD SOURCING STRATEGY



- Presently the company data sites seems very boring we intend to create visualizations and geo plots with few crazy trends to make it more appealing and a common individual can compare different parameters of companies (Net worth, Profit, CSR, Employees) for FUN!

- A groups of individual contributors & validators Individuals' progress to validators over time
  - Incentive of $x (split between validators & contributors per update of 20+ parameters, new company info. etc.

- Create a platform of co-creation i.e. Graphic designs, Logo designs games , consumer voting games on best products

- Connect with Industry Analysts where there can present their view / rating for not just public but private, partnership & LLP companies'
  - Bring companies onboard to reinforce the model

CONCLUSION & WAY FORWARD

Conclusions

- Data Collection for any project requires a defined strategy to find the right resources(seeds) and extraction process since extraction from readily available resources is not free
- Powerful extraction techniques and tools like web crawling and scrapping helped in easy data retrieval from different web resources
- Tools in python like Sweetwiz helped in quick data analysis and decision making

Way Forward

- Build on the current extracted data using crowd sourcing
- Improve the final data to more structured format
- Convert the EDA into a business report
- Create a strategy to monetize the extracted data to potential customers

ATTACHMENTS & REFERENCES

GITHUB LINK : https://github.com/cgargk1/Data-Collection-Assignment-Group-11-Section-A-Karan-Garg-Pankaj-Madan-Shruti-Mantri-VikrantDhawan

S. No.	File name	description
1.	FinalCodeCompanyCheck.ipynb	Code to extract Data from CompanyCheck Website Through lxml
2	FinalCodeSeleniumVCC.ipynb	Code to extract Data from VCCEdge Website Through Selenium
3.	IndianCompanies_output.xlsx	Complete Extracted and Merged Output DataSet in excel
4.	IndianCompanies_output.json	Complete Extracted and Merged Output DataSet in json (In zip due to size restriction of Github (<100 MB)
5.	OutputSeleniumProwess.xlsx	Sample Data Set extracted from VCC Edge Website.
6.	EDA.html	Exploratory Data Analysis
7.	DCPPGroupAssignment_Group11.pdf	Report

References

https://www.kaggle.com/rowhitswami/all-indian-companies-registration-data-1900-2019

https://prowessdx.cmie.com/

https://business.mapsofindia.com/india-company/

https://www.zaubacorp.com/

https://www.thecompanycheck.com/

https://www.vccedge.com/login.php

THANK YOU !