Submitted by : Pranjal Upadhyay

Email: upranjal22@gmail.com

Phone: 8527702856

DATA EXTRACTION AND MINING FROM findinall.com

Link: https://www.findinall.com/

Language Used: Python

Following screenshots represent the steps in the process.

1. Importing the libraries and extracting data

Libraries such as urllib, urllib2, Beautiful Soup have been used to extract the data and for visualising the data, libraries like pandas, numpy, seaborn ,matplotlib.pyplot have been used. Besides, various string manipulation techniques have alse been used.

In [2]:
```
import pandas as pd
import numpy as np
import urllib2
import urllib
from bs4 import BeautifulSoup
import requests
import unicodedata
pd.set_option('display.max_colwidth', -1)
```

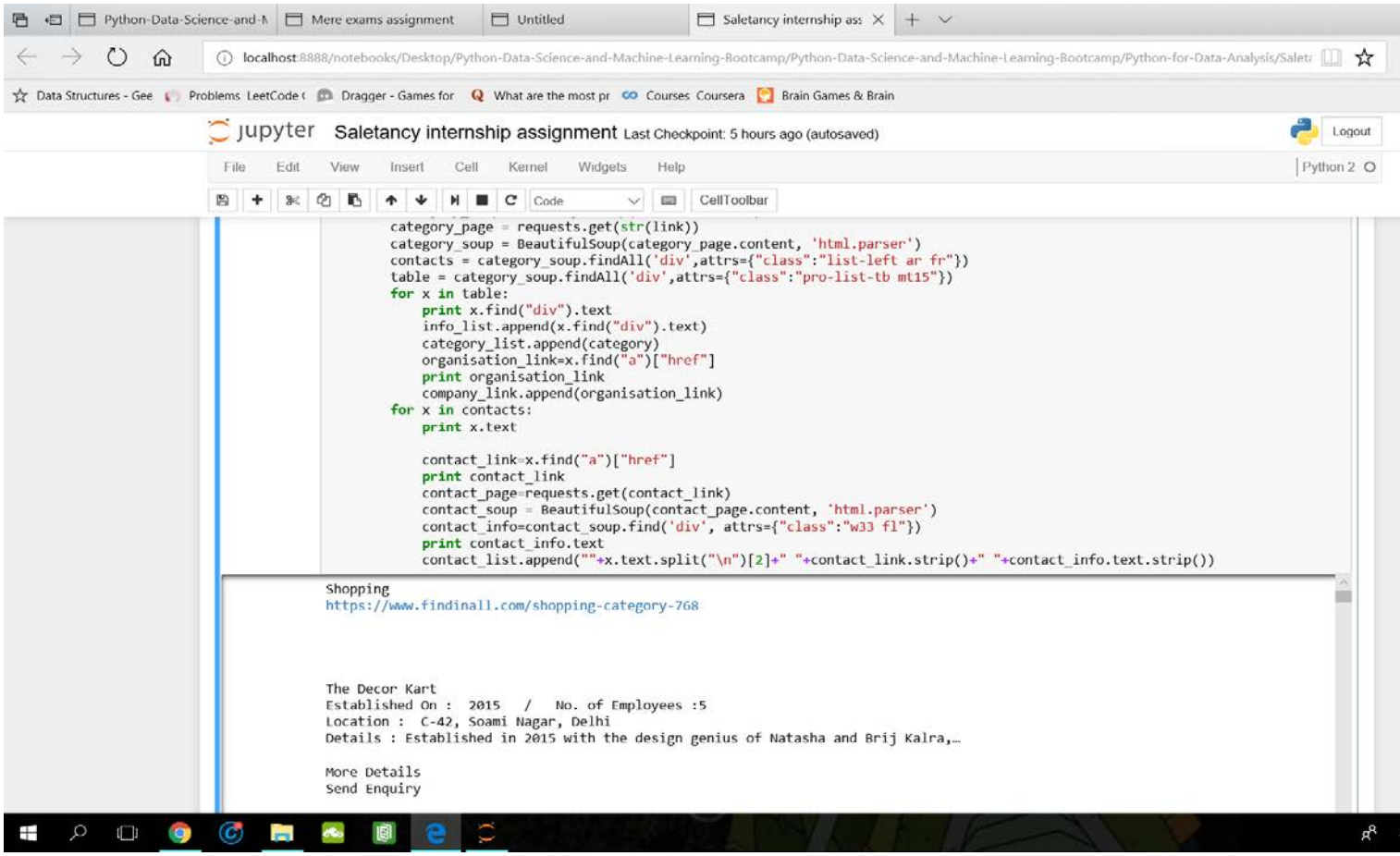
Extracting all the Category links

In [87]:
```
url = urllib.urlopen('https://www.findinall.com/')
content = url.read()
soup = BeautifulSoup(content, 'lxml')

table = soup.findAll('div',attrs={"class":"catlist-1 mt20"})
for div in table:
    links = div.findAll('a')
    for a in links:
        print a["title"]
        print  a['href']
```

```
Shopping
https://www.findinall.com/shopping-category-768
Hardware
```

Extracting all the information from all the categories

In [3]:
```
link_list=[]
info_list=[]
category_list=[]
company_link=[]
contact_list=[]
contact_list=[]
page = requests.get("https://www.findinall.com/")
#url = urllib.urlopen('https://www.findinall.com/')
#content = url.read()
#soup = BeautifulSoup(content, 'lxml')
soup = BeautifulSoup(page.content, 'html.parser')

table = soup.findAll('div',attrs={"class":"catlist-1 mt20"})
for div in table:
    links = div.findAll('a')
    for a in links:
        print a["title"]
        print  a['href']
        link=a["href"]
        category=a["title"]
        #url = urllib.urlopen(str(link))
        #content = url.read()
        #category_soup = BeautifulSoup(content, 'lxml')
        category_page = requests.get(str(link))
        category_soup = BeautifulSoup(category_page.content, 'html.parser')
        contacts = category_soup.findAll('div',attrs={"class":"list-left ar fr"})
        table = category_soup.findAll('div',attrs={"class":"pro-list-tb mt15"})
        for x in table:
            print x.find("div").text
            info_list.append(x.find("div").text)
            category_list.append(category)
```



Parsing the information into DataFrame

In [92]:
```
pd.set_option('display.max_colwidth', -1)
df=pd.DataFrame({"Category":category_list, "Company Info":info_list, "Company link":company_link })
df["Company Info"]=df["Company Info"].apply(lambda info: info.split("\n"))
df["Company Info"]=df["Company Info"].apply(lambda lst: remove_empty(lst))
```

In [93]:
```
name_list=[]
estabilished_list=[]
n_employees_list=[]
address_list=[]
state_list=[]
details_list=[]
```

In [94]:
```
for lst in df["Company Info"]:

    name_list.append(lst[0])
    line=lst[1].split("/")
    try:
        year=int(line[0].split(":")[1].strip())
        estabilished_list.append(year)
    except:
        estabilished_list.append(np.nan)

    try:
        n_emp=int(line[1].split(":")[1].strip())
        n_employees_list.append(n_emp)
    except:
        n_employees_list.append(np.nan)
```

Here is what the data extracted looks like:

```
In [30]: df.head()
```

Out[30]:

	Category	Company link	Contacts	Company Name	Year of Estabilishment	Number of employees	Address	S
0	Shopping	https://www.findinall.com/the-decor-kart-in-new-delhi-32513	Nihal : 9811331181	The Decor Kart	2015.0	5.0	C-42, Soami Nagar, Delhi	Delhi
1	Shopping	https://www.findinall.com/indian-concepts-online-in-delhi-32403	indianconceptsonline : 08527976973	Indian Concepts Online	2010.0	2020.0	Indian Concepts Online Mayur Vihar 1, Select State/Province	Select State/Prov
2	Shopping	https://www.findinall.com/awardsandtrophy-in-delhi-32401	Awardsandtrophy : 9911000035	Awardsandtrophy	1990.0	20.0	1734, Dariba Kalan Road,, Chandani Chowk, Delhi	Delhi
3	Shopping	https://www.findinall.com/indiangiftsadda-in-gurgaon-32381	Pankaj : 7011580516	IndianGiftsAdda	2017.0	10.0	Shop No.4 Pataudi Road Near Police Station, Gurgaon, Haryana	Haryana

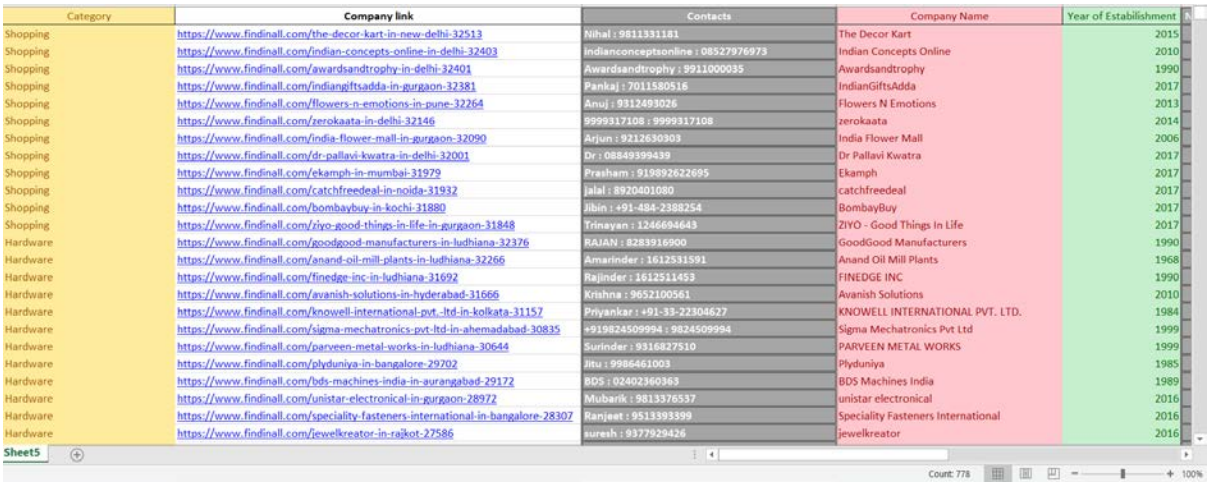
Saving the file to csv file:

Saving to file

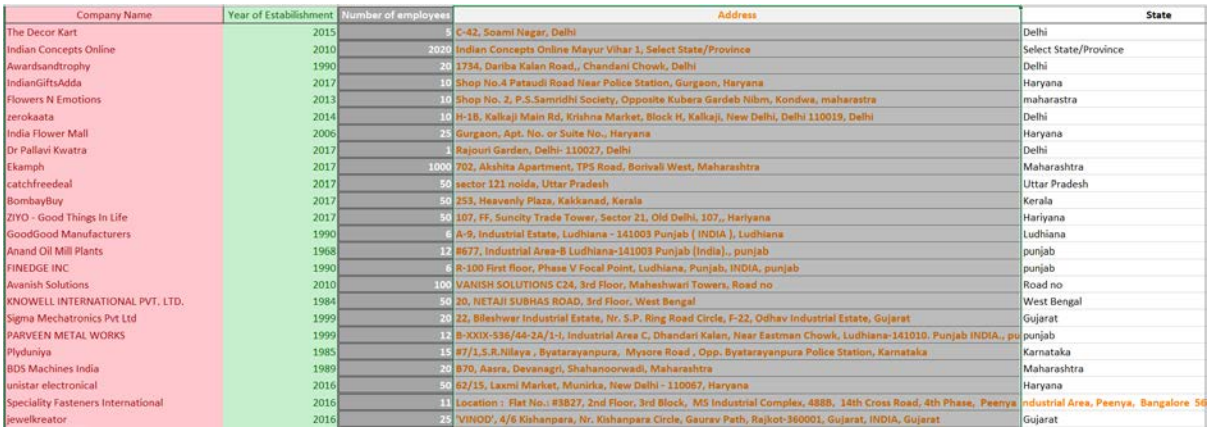
```
In [83]: df.to_csv("Saletancy_Data.csv", index=False)
```

```
In [84]: from pandas import ExcelWriter
```

```
In [85]: writer = ExcelWriter('PythonExport.xlsx')
         df.to_excel(writer,'Sheet5')
         writer.save()

         # DF TO CSV
         df.to_csv('PythonExport.csv', sep=',')
```



Category	Company link	Contacts	Company Name	Year of Estabilishment
Shopping	https://www.findinall.com/the-decor-kart-in-new-delhi-32513	Nihal : 9811331181	The Decor Kart	2015
Shopping	https://www.findinall.com/indian-concepts-online-in-delhi-32403	indianconceptsonline : 08527976973	Indian Concepts Online	2010
Shopping	https://www.findinall.com/awardsandtrophy-in-delhi-32401	Awardsandtrophy : 9911000035	Awardsandtrophy	1990
Shopping	https://www.findinall.com/indiangiftsadda-in-gurgaon-32381	Pankaj : 7011580516	IndianGiftsAdda	2017
Shopping	https://www.findinall.com/flowers-n-emotions-in-pune-32264	Anuj : 9312493026	Flowers N Emotions	2013
Shopping	https://www.findinall.com/zerokaata-in-delhi-32146	9999317108 : 9999317108	zerokaata	2014
Shopping	https://www.findinall.com/india-flower-mall-in-gurgaon-32090	Arjun : 9212630303	India Flower Mall	2006
Shopping	https://www.findinall.com/dr-pallavi-kwatra-in-delhi-32001	Dr : 08849399439	Dr Pallavi Kwatra	2017
Shopping	https://www.findinall.com/ekamph-in-mumbai-31979	Prasham : 919892622695	Ekamph	2017
Shopping	https://www.findinall.com/catchfreedeal-in-noida-31932	jalal : 8920401080	catchfreedeal	2017
Shopping	https://www.findinall.com/bombaybuy-in-kochi-31880	Jibin : +91-484-2388254	BombayBuy	2017
Shopping	https://www.findinall.com/ziyo-good-things-in-life-in-gurgaon-31848	Trinayan : 1246694643	ZIYO - Good Things In Life	2017
Hardware	https://www.findinall.com/goodgood-manufacturers-in-ludhiana-32376	RAJAN : 8283916900	GoodGood Manufacturers	1990
Hardware	https://www.findinall.com/anand-oil-mill-plants-in-ludhiana-32266	Amarinder : 1612531591	Anand Oil Mill Plants	1968
Hardware	https://www.findinall.com/finedge-inc-in-ludhiana-31692	Rajinder : 1612511453	FINEDGE INC	1990
Hardware	https://www.findinall.com/avanish-solutions-in-hyderabad-31666	Krishna : 9652100561	Avanish Solutions	2010
Hardware	https://www.findinall.com/knowell-international-pvt.-ltd-in-kolkata-31157	Priyankar : +91-33-22304627	KNOWELL INTERNATIONAL PVT. LTD.	1984
Hardware	https://www.findinall.com/sigma-mechatronics-pvt-ltd-in-ahemadabad-30835	+919824509994 : 9824509994	Sigma Mechatronics Pvt Ltd	1999
Hardware	https://www.findinall.com/parveen-metal-works-in-ludhiana-30644	Surinder : 9316827510	PARVEEN METAL WORKS	1999
Hardware	https://www.findinall.com/plyduniya-in-bangalore-29702	Jitu : 9986461003	Plyduniya	1985
Hardware	https://www.findinall.com/bds-machines-india-in-aurangabad-29172	BDS : 02402360363	BDS Machines India	1989
Hardware	https://www.findinall.com/unistar-electronical-in-gurgaon-28972	Mubarik : 9813376537	unistar electronical	2016
Hardware	https://www.findinall.com/speciality-fasteners-international-in-bangalore-28307	Ranjeet : 9513393399	Speciality Fasteners International	2016
Hardware	https://www.findinall.com/jewelkreator-in-rajkot-27586	suresh : 9377929426	jewelkreator	2016



Company Name	Year of Estabilishment	Number of employees	Address	State
The Decor Kart	2015	5	C-42, Soami Nagar, Delhi	Delhi
Indian Concepts Online	2010	2020	Indian Concepts Online Mayur Vihar 1, Select State/Province	Select State/Province
Awardsandtrophy	1990	20	1734, Dariba Kalan Road,, Chandani Chowk, Delhi	Delhi
IndianGiftsAdda	2017	10	Shop No.4 Pataudi Road Near Police Station, Gurgaon, Haryana	Haryana
Flowers N Emotions	2013	10	Shop No. 2, P.S.Samridhi Society, Opposite Kubera Gardeb Nibm, Kondwa, maharastra	maharastra
zerokaata	2014	10	H-1B, Kalkaji Main Rd, Krishna Market, Block H, Kalkaji, New Delhi, Delhi 110019, Delhi	Delhi
India Flower Mall	2006	25	Gurgaon, Apt. No. or Suite No., Haryana	Haryana
Dr Pallavi Kwatra	2017	1	Rajouri Garden, Delhi- 110027, Delhi	Delhi
Ekamph	2017	1000	702, Akshita Apartment, TPS Road, Borivali West, Maharashtra	Maharashtra
catchfreedeal	2017	50	sector 121 noida, Uttar Pradesh	Uttar Pradesh
BombayBuy	2017	50	253, Heavenly Plaza, Kakkanad, Kerala	Kerala
ZIYO - Good Things In Life	2017	50	107, FF, Suncity Trade Tower, Sector 21, Old Delhi, 107,, Hariyana	Hariyana
GoodGood Manufacturers	1990	6	A-9, Industrial Estate, Ludhiana - 141003 Punjab (INDIA), Ludhiana	Ludhiana
Anand Oil Mill Plants	1968	12	#677, Industrial Area-B Ludhiana-141003 Punjab (India)., punjab	punjab
FINEDGE INC	1990	6	R-100 First floor, Phase V Focal Point, Ludhiana, Punjab, INDIA, punjab	punjab
Avanish Solutions	2010	100	VANISH SOLUTIONS C24, 3rd Floor, Maheshwari Towers, Road no	Road no
KNOWELL INTERNATIONAL PVT. LTD.	1984	50	20, NETAJI SUBHAS ROAD, 3rd Floor, West Bengal	West Bengal
Sigma Mechatronics Pvt Ltd	1999	20	22, Bileshwar Industrial Estate, Nr. S.P. Ring Road Circle, F-22, Odhav Industrial Estate, Gujarat	Gujarat
PARVEEN METAL WORKS	1999	12	B-XXIX-536/44-2A/1-I, Industrial Area C, Dhandari Kalan, Near Eastman Chowk, Ludhiana-141010. Punjab INDIA., pu	punjab
Plyduniya	1985	15	#7/1,S.R.Nilaya , Byatarayanpura, Mysore Road , Opp. Byatarayanpura Police Station, Karnataka	Karnataka
BDS Machines India	1989	20	B70, Aasra, Devanagri, Shahanoorwadi, Maharashtra	Maharashtra
unistar electronical	2016	50	62/15, Laxmi Market, Munirka, New Delhi - 110067, Haryana	Haryana
Speciality Fasteners International	2016	11	Location : Flat No.: #3827, 2nd Floor, 3rd Block, M5 Industrial Complex, 488B, 14th Cross Road, 4th Phase, Peenya	ndustrial Area, Peenya, Bangalore 56
jewelkreator	2016	25	'VINOD', 4/6 Kishanpara, Nr. Kishanpara Circle, Gaurav Path, Rajkot-360001, Gujarat, INDIA, Gujarat	Gujarat

Describing the dataset:

Since only year of estabilishment and number of employee columns are numerical in nature, hence description will be based on these 2 columns.

Decription inclused variables like count, mean, standard deviation etc

```
df.describe()
```

	Year of Establishment	Number of employees
count	760.00000	758.000000
mean	1970.00000	561.174142
std	253.63602	4829.067349
min	1.00000	0.000000
25%	2000.00000	10.000000
50%	2009.00000	30.000000
75%	2015.00000	100.000000
max	2048.00000	100000.000000

Correlation between Year of estabilshment and Number of employees

```
d=df[df["Year of Establishment"]>1970][df["Number of employees"]<1000]
df.corr()
```

```
C:\Users\ABCD\Anaconda2\lib\site-packages\ipykernel\__main__.py:1: UserWarning: Bool
ataFrame index.
  if __name__ == '__main__':
```

	Year of Establishment	Number of employees
Year of Establishment	1.000000	0.000734
Number of employees	0.000734	1.000000

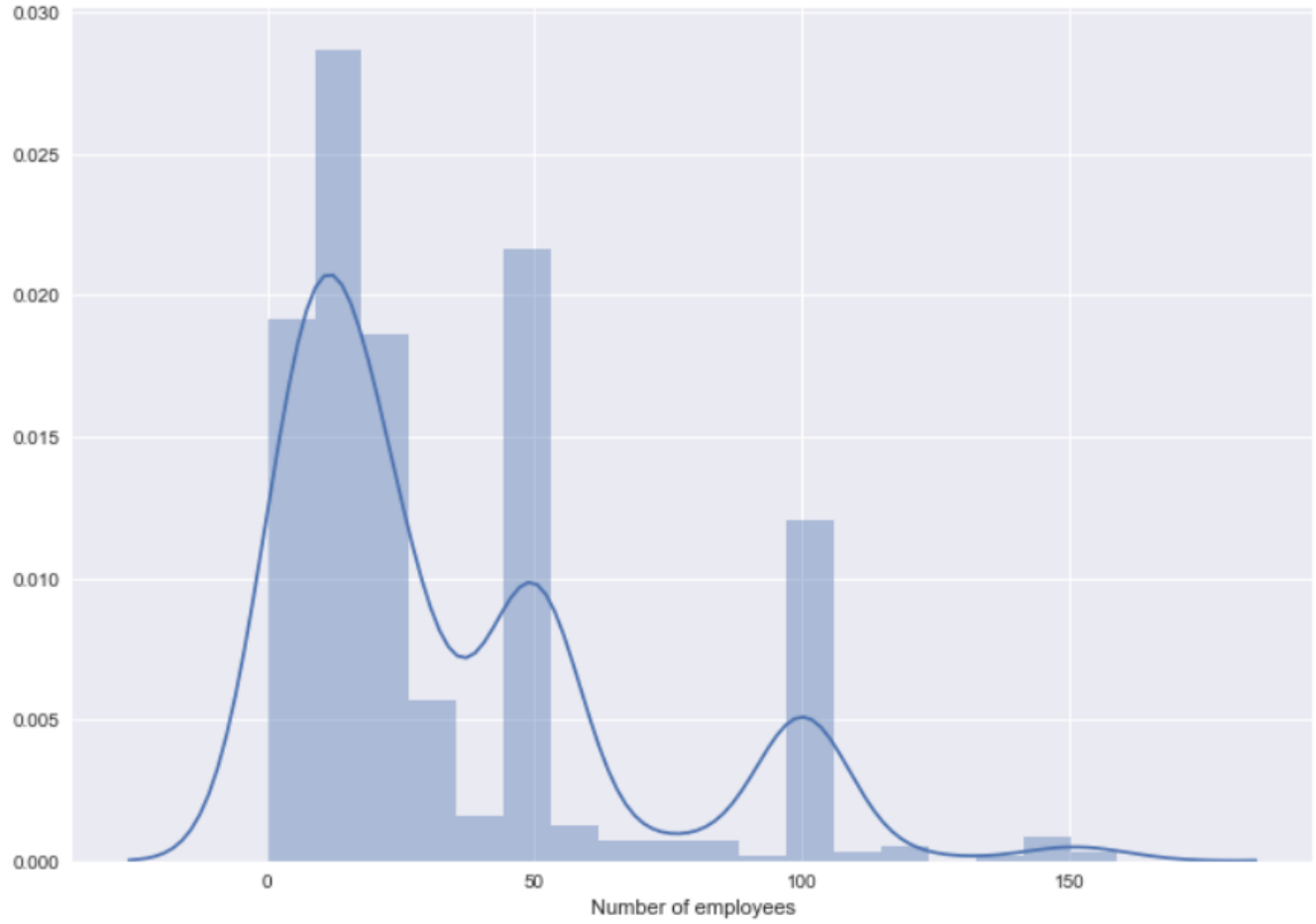
As we can see, there is a slight correlation between year of estabilishment and Number of employees. This correlation can further be illustrated by an various plots shown below.

Describing the data by visualisation techniques:

Distribution of number of employees Shows most of the companies have number of employees between 1 and 50

```
sns.distplot(df[df["Number of employees"]<200]["Number of employees"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x11d1da58>
```

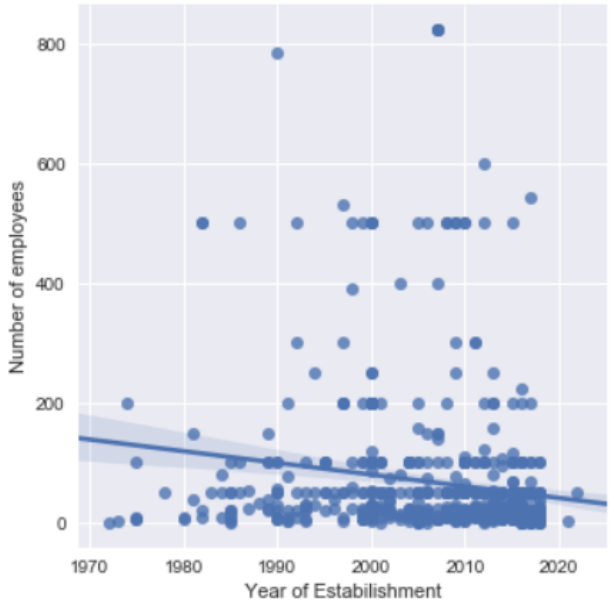


The above plot shows that most of the companies have number of employees between 0 to 50. Also, there are a few ouliers present.

Plot between Year of establishment and number of employees is shown below

	Year of Estabilishment	Number of employees
Year of Estabilishment	1.000000	0.000734
Number of employees	0.000734	1.000000

```
191]: sns.lmplot(x="Year of Estabilishment", y="Number of employees", data=d)
191]: <seaborn.axisgrid.FacetGrid at 0x14396128>
```



The above plot shows the regression line between the two quantities. It can be inferred that, the more recent companies are hiring lesser number of employees.

State-wise distribution of number of companies:

Statewise distribution of numbet of companies

```
city=pd.DataFrame(df.groupby("State").count()["Company Name"])
```

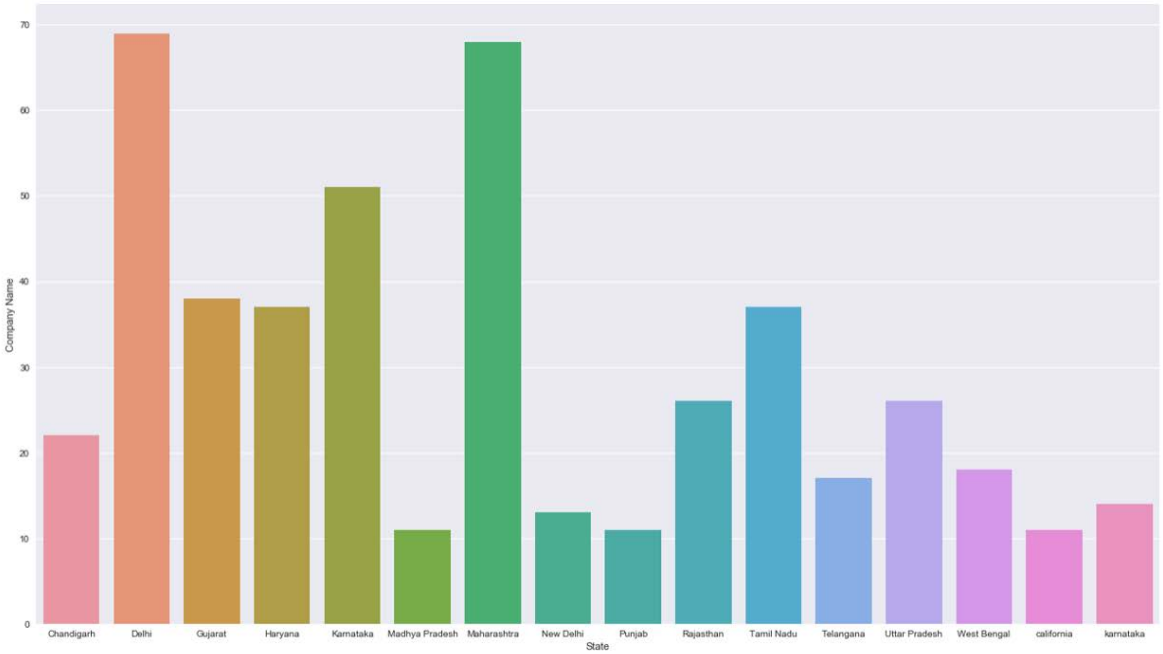
```
major_cities=city[city["Company Name"]>10]
major_cities["State"]=major_cities.index
major_cities["Number of companies"]=city["Company Name"]
```

```
C:\Users\ABCD\Anaconda2\lib\site-packages\ipykernel\__main__.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  from ipykernel import kernelapp as app
C:\Users\ABCD\Anaconda2\lib\site-packages\ipykernel\__main__.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  app.launch_new_instance()
```

```
sns.set(rc={'figure.figsize':(21.7,12.27)})
sns.barplot(x="State", y="Company Name", data=major_cities)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1414ddd8>
```

: <matplotlib.axes._subplots.AxesSubplot at 0x1414ddd8>



The above plot shows that Delhi and Maharahtra have most number of companies. Punjab and Madhya Pradesh have least.

Plot to show top 10 employing companies by number of employees working in them:

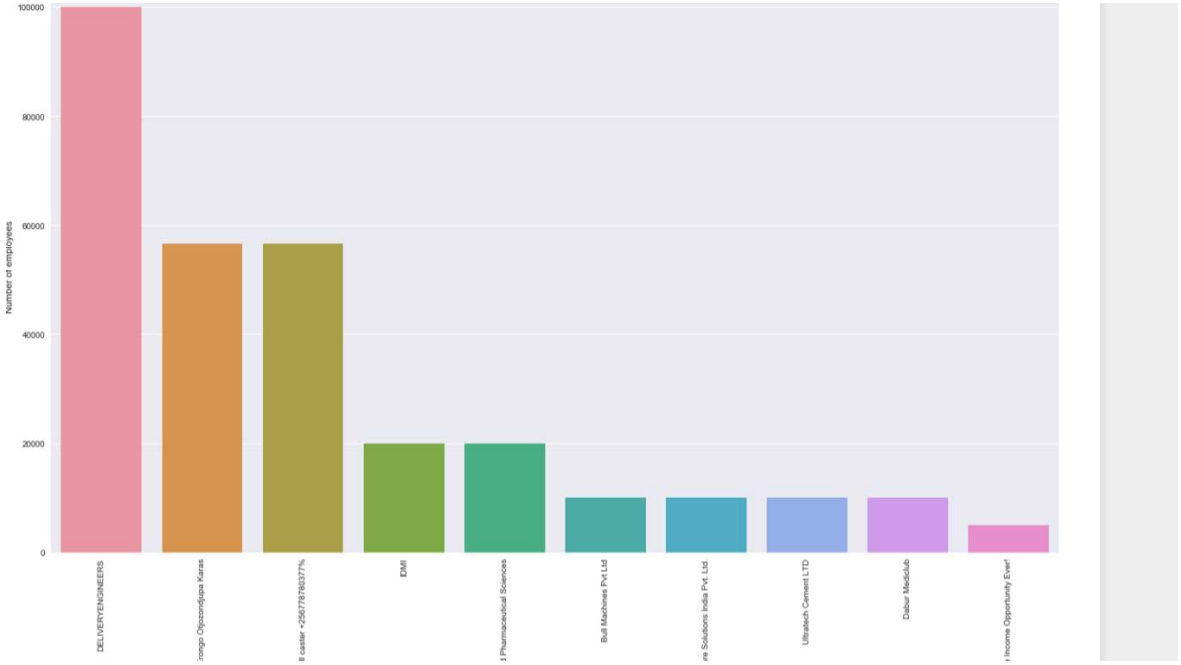
Plot Showing top 10 employers

In [217]:
```
top_10_employers=df.sort_values(by="Number of employees", ascending=False).head(10)
```

In [263]:
```
top_10_employers[["Company Name", "Number of employees"]]
```

Out[263]:

	Company Name	Number of employees
555	DELIVERYENGINEERS	100000.0
618	....@Spell caster lost lover 100% love spells caster make him love me +256778780377 Windhoek Rundu Walvis bay, Erongo Otjozondjupa Karas	56677.0
261	https://Powerfull% healer lost love% spell caster +256778780377%	56677.0
662	IDMI	20000.0
507	Gulf Congress on Pharmacy and Pharmaceutical Sciences	20000.0
605	Bull Machines Pvt Ltd	10000.0
447	Convexicon Software Solutions India Pvt. Ltd.	10000.0
481	Ultratech Cement LTD	10000.0
597	Dabur Mediclub	10000.0
720	Best Online Income Opportunity Ever!	5000.0



Plot showing top 10 industries by employee size

```
means=df.groupby("Category").describe()["Number of employees"]["mean"]
counts=df.groupby("Category").describe()["Number of employees"]["count"]
num_emp=means*counts
```

```
num_emp=num_emp.sort_values(ascending=False).head(10)
```

```
num_emp=pd.DataFrame(num_emp, columns=["Number of employees"])
num_emp
```

	Number of employees
Category	
Cosmetics	102108.0
Astrology	57596.0
Health & Beauty	57408.0
Advertising	20329.0
Event Management	20303.0
Jobs	17729.0
Building Materials	12338.0
Security & Administration	11061.0
Medical Shop	10765.0
Automobiles	10455.0

The plot downbelow shows that cosmetics and astrology have employed most number of people as visible from tha bar plot below.

Out[260]: <matplotlib.axes._subplots.AxesSubplot at 0x1921c630>