Mayank Singh

19,Nov,2022

Agenda - Which basic questions we should think of/ask when we see the Data for the first time?

Importing Liabrary

In [1]:
```
import pandas as pd
```

Reading Data

Sometimes reading Data from CSV files gets encoding errors, and to get over these errors we have many options to chnage the encoding (read documentation/Google) or we can ignore the encoding erros

In [2]:
```
## Ignoring the encoding errors in my case, using "encoding_erros = 'ignore' parc
df = pd.read_csv('Himachal_Pradesh_Political_party_candidates.csv', encoding_erro
```

In [3]:
```
# Reading first 5 rows of data using pd.head() function
```

In [4]:
```
df.head()
```

Out[4]:

	Sno	Candidate	Constituency	Party	Criminal Case	Education	Total Assets	L
0	1	Abhay Kumar Ashok	DHARAMSHALA	IND	0	Post Graduate	Rs97,40,093\r\n~ 97Lacs+	
1	2	Abhinay Bhardwaj	HAMIRPUR	Rashtriya Devbhumi Party	0	Graduate	Rs5,42,477\r\n~ 5Lacs+	Rs2,50
2	3	Abhishek Barowalia	SHIMLA	IND	0	Graduate Professional	Rs29,54,391\r\n~ 29Lacs+	Rs3,50
3	4	Abhishek Singh	SHAHPUR	AAP	0	12th Pass	Rs65,59,12,561\r\n~ 65Crore+	Rs1,07,88
4	5	Abhishek Thakur	SUNDERNAGAR	IND	0	Graduate Professional	Rs7,63,01,043\r\n~ 7Crore+	Rs4,16,16

Questions to ask - Begin

1. How big is the Data?

To get a baisc idea about the size of data should be first task, to understand that with which we are going to deal - is the data too big, small, average and so on....

In [5]:
```
df.shape
```

Out[5]:
```
(412, 8)
```

2. How does the Data look like?

We should have rough idea of our dataset. Which all coulmns are present, what are they depicting, how many int colmns, object/string colmns etc.

In [6]:
```
## Two methods are available for this task
# 1. pd.head()
df.head() # display/read the first 5 rows of the dataset
```

Out[6]:

	Sno	Candidate	Constituency	Party	Criminal Case	Education	Total Assets	L
0	1	Abhay Kumar Ashok	DHARAMSHALA	IND	0	Post Graduate	Rs97,40,093\r\n~ 97Lacs+	
1	2	Abhinay Bhardwaj	HAMIRPUR	Rashtriya Devbhumi Party	0	Graduate	Rs5,42,477\r\n~ 5Lacs+	Rs2,50
2	3	Abhishek Barowalia	SHIMLA	IND	0	Graduate Professional	Rs29,54,391\r\n~ 29Lacs+	Rs3,50
3	4	Abhishek Singh	SHAHPUR	AAP	0	12th Pass	Rs65,59,12,561\r\n~ 65Crore+	Rs1,07,88
4	5	Abhishek Thakur	SUNDERNAGAR	IND	0	Graduate Professional	Rs7,63,01,043\r\n~ 7Crore+	Rs4,16,16

In [7]:
```
# 2. pd. sample()
df.sample(10) # Randomly picks the data, so that user should not make a gernalize
```

Out[7]:

	Sno	Candidate	Constituency	Party	Criminal Case	Education	Total Assets	L
348	349	Sudhir Kumar	JHANDUTA (SC)	AAP	0	Post Graduate	Rs28,57,000\r\n~ 28Lacs+	Rs4,53
383	384	Tikender Singh Panwar	SHIMLA	CPI(M)	3	Graduate	Rs4,34,85,802\r\n~ 4Crore+	
169	170	Kushal Bhardwaj	JOGINDERNAGAR	CPI(M)	0	Post Graduate	Rs63,27,584\r\n~ 63Lacs+	
310	311	Reena	PACHHAD (SC)	BJP	0	Post Graduate	Rs1,21,21,682\r\n~ 1Crore+	Rs42,14
167	168	Kuldip Singh Tanwar	KASUMPTI	CPI(M)	20	Doctorate	Rs12,80,23,118\r\n~ 12Crore+	
86	87	Dhavinder Singh	DALHOUSIE	BJP	0	10th Pass	Rs15,67,49,714\r\n~ 15Crore+	Rs74,89
59	60	Bumber Thakur	BILASPUR	INC	2	Graduate Professional	Rs8,95,28,229\r\n~ 8Crore+	Rs55,61
226	227	Paras Ram	ANNI (SC)	IND	0	Graduate	Rs1,81,42,000\r\n~ 1Crore+	Rs20,00
197	198	Munish Sharma	SARKAGHAT	IND	1	Post Graduate	Rs41,94,392\r\n~ 41Lacs+	Rs30,50
375	376	Tara Chand	BALH (SC)	AAP	0	10th Pass	Rs73,78,500\r\n~ 73Lacs+	Rs3,78

3. What is the Data Types of columns?

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 412 entries, 0 to 411
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Sno            412 non-null    int64 
 1   Candidate      412 non-null    object
 2   Constituency   412 non-null    object
 3   Party          412 non-null    object
 4   Criminal Case  412 non-null    int64 
 5   Education      412 non-null    object
 6   Total Assets   412 non-null    object
 7   Liabilities    412 non-null    object
dtypes: int64(2), object(6)
memory usage: 25.9+ KB
```

4. Are there any missing values, if yes how many

Although pd.info() somehow/indirectly tells us anout the missing values, but to know exact numbers and coulumns we should find out missing values seperately

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Sno              0
Candidate        0
Constituency     0
Party            0
Criminal Case    0
Education        0
Total Assets     0
Liabilities      0
dtype: int64
```

5. How does the data look mathematically?

In [10]:
```
df.describe() # it provides the basic mathematical calculations ONLY ON INTEGER C
# Since the data is not cleaned, hece the results are of not much use at this par
# but if the data is somewhat in good shape, the function is of good use. At late
# convert some coulmns to int type and will try to perform this function again.
```

Out[10]:

	Sno	Criminal Case
count	412.000000	412.000000
mean	206.500000	0.601942
std	119.078406	2.323433
min	1.000000	0.000000
25%	103.750000	0.000000
50%	206.500000	0.000000
75%	309.250000	0.000000
max	412.000000	30.000000

6. Are there duplicate values?

In [11]:
```
df.duplicated().sum()
```

Out[11]:
```
0
```

7. How/What is the correlation between columns?

In [12]:
```
df.corr() # it provides the basic mathematical calculations ONLY ON INTEGER COLUM
# Since the data is not cleaned, hece the results are of not much use at this par
# but if the data is somewhat in good shape, the function is of good use. At late
# convert some coulmns to int type and will try to perform this function again.
```

Out[12]:

	Sno	Criminal Case
Sno	1.000000	0.005523
Criminal Case	0.005523	1.000000