Import a Data

```
from google.colab import files # for uploading data file
files.upload()
```

```
Choose Files TCPD_AE_Ut...-6-24.csv
```

- **TCPD_AE_Uttar_Pradesh_2021-6-24.csv**(application/vnd.ms-excel) - 6628954 bytes, last modified: 8/27/2021 - 100% done

```
Saving TCPD_AE_Uttar_Pradesh_2021-6-24.csv to TCPD_AE_Uttar_Pradesh_2021-6-24.csv
{'TCPD_AE_Uttar_Pradesh_2021-6-24.csv': b'\xef\xbb\xbf"Election_Type","State_Name","Asse
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Reading of Data

```
df = pd.read_csv('TCPD_AE_Uttar_Pradesh_2021-6-24.csv', index_col
df.head()
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarni
  interactivity=interactivity, compiler=compiler, result=result)
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

	Election_Type						Poll_No	DelimID
	AE	Uttar_Pradesh	14	1	2002	2	0	3
	AE	Uttar_Pradesh	14	1	2002	2	0	3
	AE	Uttar_Pradesh	14	1	2002	2	0	3
	AE	Uttar_Pradesh	14	1	2002	2	0	3
	AE	Uttar_Pradesh	14	1	2002	2	0	3

About Data

```
df.shape
```

```
(24744, 41)
```

This Data contain a 24744 Rows and 41 Columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 24744 entries, AE to AE
Data columns (total 41 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   State_Name              24744 non-null  object
 1   Assembly_No             24744 non-null  int64
 2   Constituency_No         24744 non-null  int64
 3   Year                    24744 non-null  int64
 4   month                   24744 non-null  int64
 5   Poll_No                 24744 non-null  int64
 6   DelimID                 24744 non-null  int64
 7   Position                24744 non-null  int64
 8   Candidate               24744 non-null  object
 9   Sex                     23888 non-null  object
 10  Party                   24744 non-null  object
 11  Votes                   24732 non-null  float64
 12  Candidate_Type          17777 non-null  object
 13  Valid_Votes             24732 non-null  float64
 14  Electors                24744 non-null  int64
 15  Constituency_Name       24744 non-null  object
 16  Constituency_Type       24744 non-null  object
```



```
 20  Vote_Share_Percentage   24732 non-null  float64
 21  Deposit_Lost            24744 non-null  object
 22  Margin                  24744 non-null  int64
 23  Margin_Percentage       24732 non-null  float64
 24  ENOP                    24732 non-null  float64
 25  pid                     24341 non-null  object
 26  Party_Type_TCPD         0 non-null      float64
 27  Party_ID                0 non-null      float64
 28  last_poll               24744 non-null  bool
 29  Contested               24341 non-null  float64
 30  Last_Party              5164 non-null   object
 31  Last_Party_ID           0 non-null      float64
 32  Last_Constituency_Name  5164 non-null   object
 33  Same_Constituency       5164 non-null   object
 34  Same_Party              5164 non-null   object
 35  No_Terms                24341 non-null  float64
 36  Turncoat                24341 non-null  object
 37  Incumbent               24744 non-null  bool
 38  Recontest               24744 non-null  bool
```

```
 39  Age                    17777 non-null  float64
 40  District_Name          12367 non-null  object
dtypes: bool(3), float64(12), int64(10), object(16)
memory usage: 7.4+ MB
```

df.describe()

	Assembly_No	Constituency_No	Year	month	Poll_No	Deli
count	24744.000000	24744.000000	24744.000000	24744.000000	24744.000000	24744.000
mean	15.472397	203.579696	2009.446330	3.280957	0.043000	3.499
std	1.074126	116.145798	5.345111	1.096573	0.207588	0.500
min	14.000000	1.000000	2002.000000	2.000000	0.000000	3.000
25%	15.000000	104.000000	2007.000000	3.000000	0.000000	3.000
50%	15.000000	205.000000	2011.000000	3.000000	0.000000	3.000
75%	16.000000	305.000000	2012.000000	5.000000	0.000000	4.000
max	17.000000	403.000000	2017.000000	5.000000	2.000000	4.000

type(df)

```
pandas.core.frame.DataFrame
```

df.count()


```
Constituency_No          24744
Year                     24744
month                    24744
Poll_No                  24744
DelimID                  24744
Position                 24744
Candidate                24744
Sex                      23888
Party                    24744
Votes                    24732
Candidate_Type           17777
Valid_Votes              24732
Electors                 24744
Constituency_Name        24744
Constituency_Type        24744
Sub_Region               12367
N_Cand                   24744
Turnout_Percentage       24732
Vote_Share_Percentage    24732
```

```
Deposit_Lost              24744
Margin                    24744
Margin_Percentage         24732
ENOP                      24732
pid                       24341
Party_Type_TCPD               0
Party_ID                      0
last_poll                 24744
Contested                 24341
Last_Party                 5164
Last_Party_ID                 0
Last_Constituency_Name     5164
Same_Constituency          5164
Same_Party                 5164
No_Terms                  24341
Turncoat                  24341
Incumbent                 24744
Recontest                 24744
Age                       17777
District_Name             12367
dtype: int64
```

```
df.columns
```

```
Index(['State_Name', 'Assembly_No', 'Constituency_No', 'Year', 'month',
       'Poll_No', 'DelimID', 'Position', 'Candidate', 'Sex', 'Party', 'Votes',
       'Candidate_Type', 'Valid_Votes', 'Electors', 'Constituency_Name',
       'Constituency_Type', 'Sub_Region', 'N_Cand', 'Turnout_Percentage',
       'Vote_Share_Percentage', 'Deposit_Lost', 'Margin', 'Margin_Percentage',
       'ENOP', 'pid', 'Party_Type_TCPD', 'Party_ID', 'last_poll', 'Contested',
       'Last_Party', 'Last_Party_ID', 'Last_Constituency_Name',
       'Same_Constituency', 'Same_Party', 'No_Terms', 'Turncoat', 'Incumbent',
       'Recontest', 'Age', 'District_Name'],
      dtype='object')
```



```
num = ['int16', 'int32', 'int64', 'float16', 'float32', 'float64']
num_df= df.select_dtypes(include=num)
print(num_df.columns)
```

```
Index(['Assembly_No', 'Constituency_No', 'Year', 'month', 'Poll_No', 'DelimID',
       'Position', 'Votes', 'Valid_Votes', 'Electors', 'N_Cand',
       'Turnout_Percentage', 'Vote_Share_Percentage', 'Margin',
       'Margin_Percentage', 'ENOP', 'Party_Type_TCPD', 'Party_ID', 'Contested',
       'Last_Party_ID', 'No_Terms', 'Age'],
      dtype='object')
```

```
len(num_df.columns)
```

```
22
```

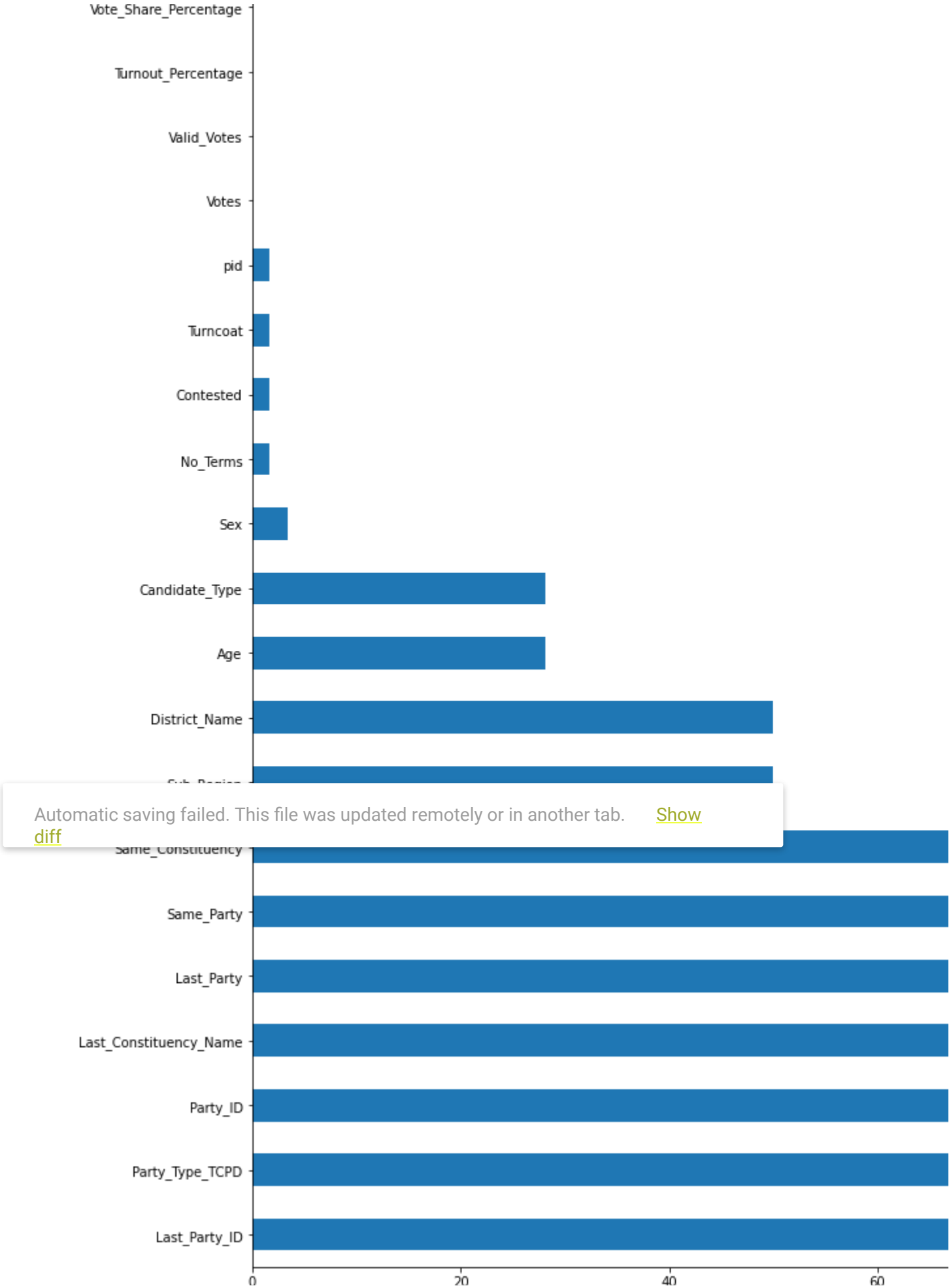
- Finding a presentage of missing value per column

```
missing_persentge = df.isnull().sum().sort_values(ascending=False
missing_persentge
```

```
Last_Party_ID            100.000000
Party_Type_TCPD          100.000000
Party_ID                 100.000000
Last_Constituency_Name    79.130294
Last_Party                79.130294
Same_Party                79.130294
Same_Constituency         79.130294
Sub_Region                50.020207
District_Name             50.020207
Age                       28.156321
Candidate_Type            28.156321
Sex                        3.459425
No_Terms                   1.628678
Contested                  1.628678
Turncoat                   1.628678
pid                        1.628678
Votes                      0.048497
Valid_Votes                0.048497
Turnout_Percentage         0.048497
Vote_Share_Percentage      0.048497
Margin_Percentage          0.048497
ENOP                       0.048497
Deposit_Lost               0.000000
Candidate                  0.000000
Assembly_No                0.000000
Constituency_No            0.000000
Year                       0.000000
month                      0.000000
Poll_No                    0.000000
DelimID                    0.000000
Recontest                  0.000000
Margin                     0.000000
Incumbent                  0.000000
last_poll                  0.000000
Electors                   0.000000
Constituency_Name          0.000000
Constituency_Type          0.000000
N_Cand                     0.000000
State_Name                 0.000000
dtype: float64
```



```
missing_persentge[missing_persentge!=0].plot(kind='barh',figsize=
```



Automatic saving failed. This file was updated remotely or in another tab. Show diff

Data Exploration and Visualisation

1. Distric_Name
2. sub_Region
3. Year wise Votes
4. Candidate_Type
5. Sex wise voters
6. Party
7. Age

1. District Wise votes

df.District_Name

```
Election_Type
AE          NaN
AE          NaN
AE          NaN
AE          NaN
AE          NaN
          ...
AE    SONBHADRA
AE    SONBHADRA
AE    SONBHADRA
Name: District_Name, Length: 24744, dtype: object
```


df.District_Name.unique()

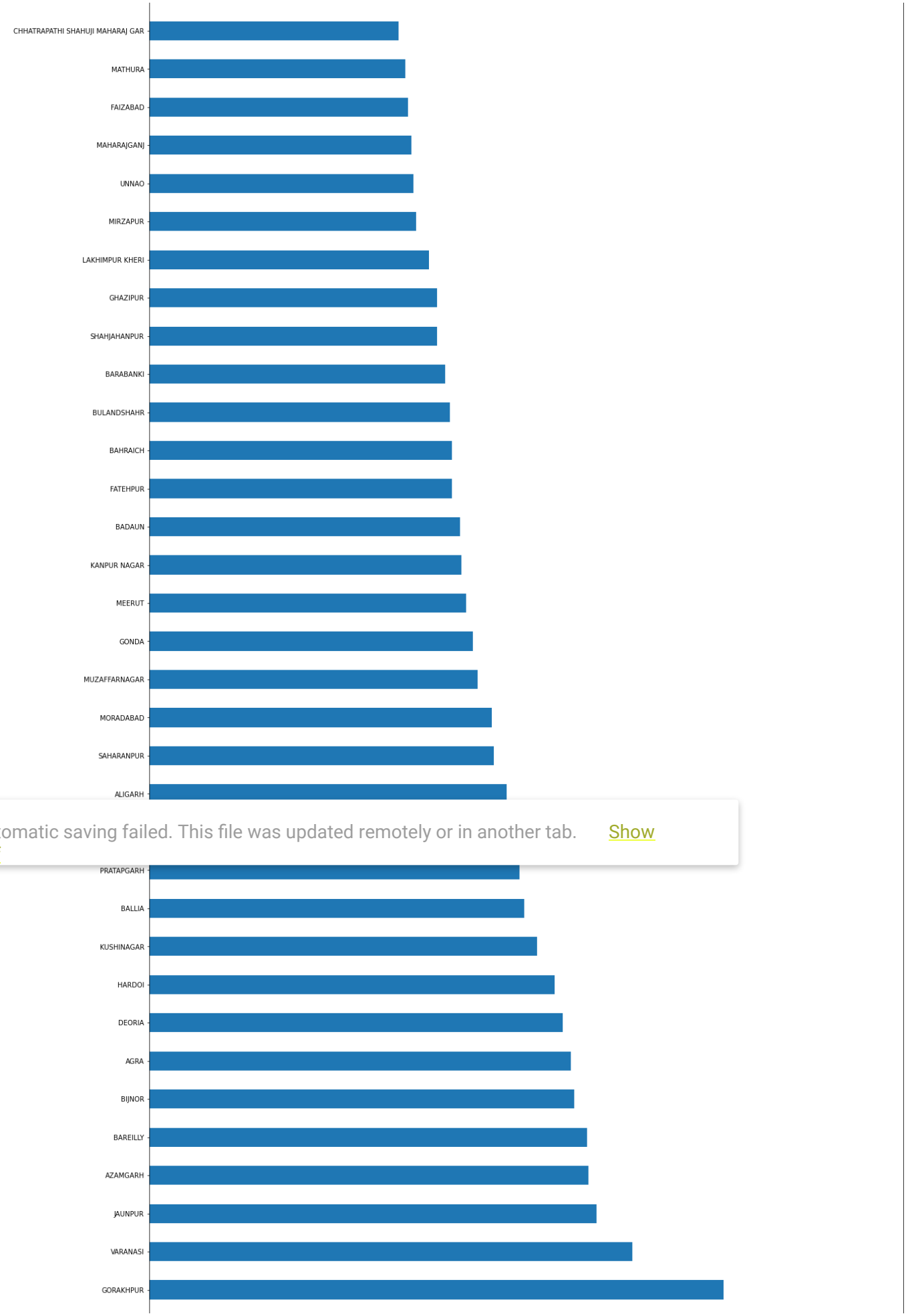
```
array([nan, 'SAHARANPUR', 'PRABUDDHA NAGAR', 'MUZAFFARNAGAR', 'BIJNOR',
       'MORADABAD', 'BHEEM NAGAR', 'RAMPUR', 'JYOTIBA PHULE NAGAR',
       'MEERUT', 'BAGHPAT', 'GHAZIABAD', 'PANCHSHEEL NAGAR',
       'GAUTAM BUDH NAGAR', 'BULANDSHAHR', 'ALIGARH', 'MAHAMAYA NAGAR',
       'MATHURA', 'AGRA', 'FIROZABAD', 'KANSHIRAM NAGAR', 'ETAH',
       'MAINPURI', 'BADAUN', 'BAREILLY', 'PILIBHIT', 'SHAHJAHANPUR',
       'LAKHIMPUR KHERI', 'KHERI', 'SITAPUR', 'HARDOI', 'UNNAO',
       'LUCKNOW', 'RAE BARELI', 'CHHATRAPATHI SHAHUJI MAHARAJ GAR',
       'SULTANPUR', 'FARRUKHABAD', 'KANNAUJ', 'ETAWAH', 'AURAIYA',
       'RAMABAI NAGAR', 'KANPUR GRAMEEN', 'KANPUR NAGAR', 'JALAUN',
       'JHANSI', 'LALITPUR', 'HAMIRPUR', 'MAHOBA', 'BANDA', 'CHITRAKOOT',
       'FATEHPUR', 'PRATAPGARH', 'KAUSHAMBI', 'ALLAHABAD', 'BARABANKI',
```

```
'FAIZABAD', 'AMBEDKAR NAGAR', 'BAHRAICH', 'SHRAVASTI', 'BALRAMPUR',
'GONDA', 'SIDDHARTH NAGAR', 'BASTI', 'SANT KABIR NAGAR',
'MAHARAJGANJ', 'GORAKHPUR', 'KUSHINAGAR', 'DEORIA', 'AZAMGARH',
'MAU', 'BALLIA', 'JAUNPUR', 'GHAZIPUR', 'CHANDAULI', 'VARANASI',
'SANT RAVIDAS NAGAR', 'MIRZAPUR', 'SONBHADRA'], dtype=object)
```

```
len(df.District_Name.unique())
```

```
78
```

```
dist_by_vote = df.District_Name.value_counts()
dist_by_vote
dist_by_vote.plot( kind= 'barh', figsize=(20,80));
```



Automatic saving failed. This file was updated remotely or in another tab. Show diff

Automatic saving failed. This file was updated remotely or in another tab. Show diff

Automatic saving failed. This file was updated remotely or in another tab. Show diff

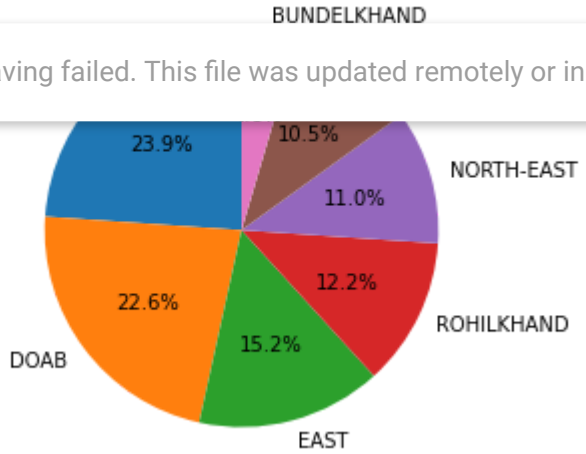
```
dist_by_vote[:10].plot( kind= 'barh', figsize=(10,10));
```

2. sub_Region

```
vote_by_subRegion = df.Sub_Region.value_counts()
vote_by_subRegion
```

```
AVADH          2956
DOAB           2800
EAST           1880
ROHILKHAND     1507
NORTH-EAST     1361
WEST           1299
BUNDELKHAND     564
Name: Sub_Region, dtype: int64
```


```
labels= ['AVADH','DOAB','EAST','ROHILKHAND','NORTH-EAST','WEST','
sizes= [2956, 2800, 1880, 1507,1361,1299,564]
plt.pie(sizes, labels=labels, startangle=90, autopct='%1.1f%%')
plt.axis('equal')
plt.show()
```



3. Year wise votes

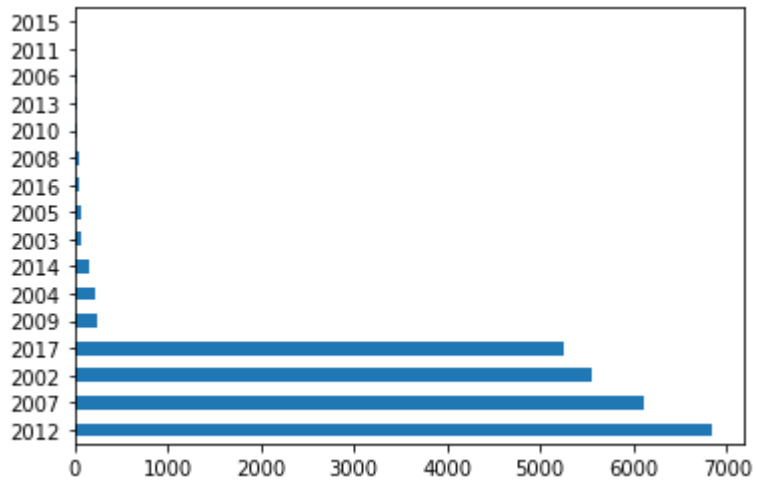
```
df.Year.unique()
```

```
array([2002, 2004, 2003, 2005, 2006, 2007, 2009, 2010, 2008, 2011, 2012,
```

```
       2014, 2016, 2013, 2015, 2017])
```

```
vote_by_Year = df.Year.value_counts()
vote_by_Year.plot(kind = 'barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8eab872510>
```



4.Candidate_Type

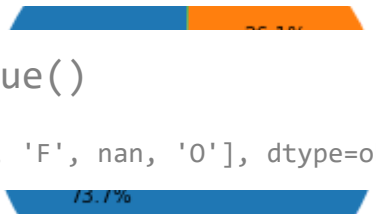
```
df.Candidate_Type.unique()
```

```
array([nan, 'GEN', 'SC', 'ST'], dtype=object)
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
plt.pie(c_type, labels=labels, startangle=90, autopct='%1.1f%%')
plt.axis('equal')
plt.show()
```

5.Sex wise voters

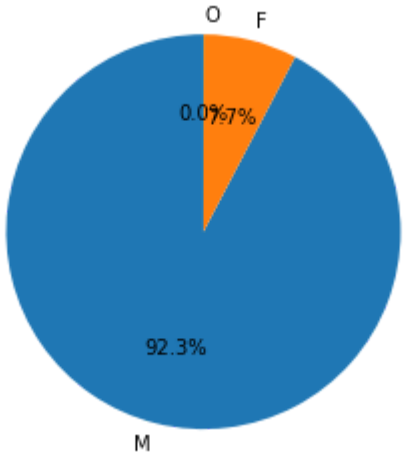


```
df.Sex.unique()
```

```
array(['M', 'F', nan, 'O'], dtype=object)
```



```
n = df.Sex.value_counts()
labels = ['M','F','O']
plt.pie(n, labels=labels, startangle=90, autopct='%1.1f%%')
plt.axis('equal')
plt.show()
```



Automatic saving failed. This file was updated remotely or in another tab. Show diff

6.Major Parties by votes

```
parties = df.Party.unique()
```

```
len(parties)
```

```
577
```

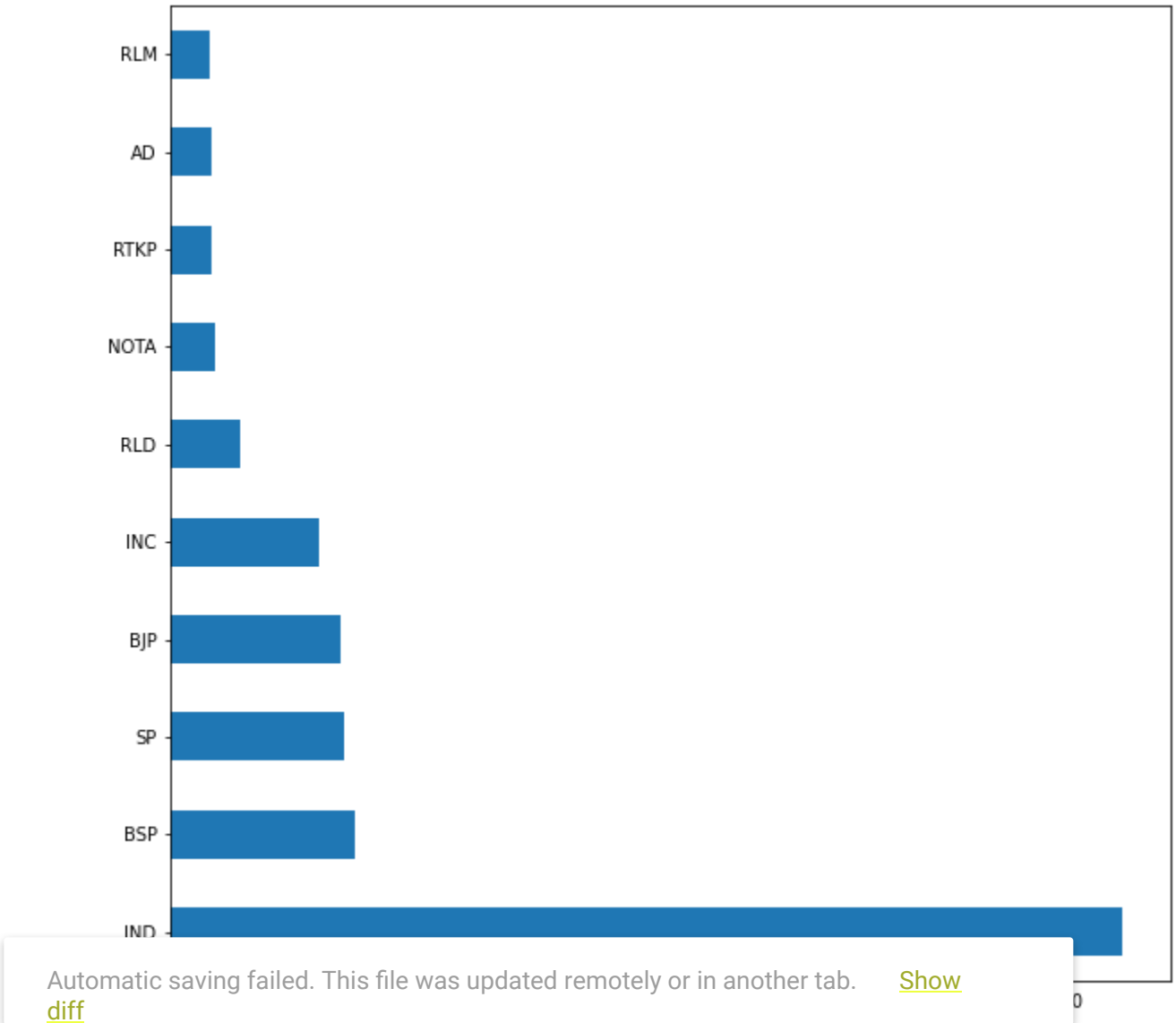
```
type(parties)
```

```
numpy.ndarray
```

```
parties = df.Party.value_counts()
```

```
parties[:10].plot(kind='barh',figsize=(10,10))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8eab30a690>
```



7. Age

28.16% Data is missing in Age column

hist plot before filling a data

```
df.Age.hist()
```