Merge data into a single DataFrame

In [103]:

```
import pandas as pd
import numpy as np
import os
import glob
```

In [104]:

```
path_dir =r'./OneDrive/Desktop/103/Project 1/Final_Dataset'
for Final_Dataset in os.listdir(path_dir):
    print(Final_Dataset)
```

```
generic_ballot_polls.csv
generic_ballot_polls_historical.csv
governor_polls.csv
governor_polls_historical.csv
house_polls.csv
house_polls_historical.csv
president_approval_polls.csv
president_polls.csv
president_polls_historical.csv
president_primary_polls.csv
senate_polls.csv
senate_polls_historical.csv
```

In [105]:

```

all_files = glob.glob(path_dir + "/*.csv")

li = []

for filename in all_files:
    df = pd.read_csv(filename, index_col=None, header=0,low_memory=False)
    li.append(df)

FinalData = pd.concat(li, axis=0, ignore_index=True)
```

In [106]:

```
FinalData.shape
```

Out[106]:

```
(32602, 33)
```

Display Dataset

```
In [107]: FinalData
```

Out[107]:

	Unnamed: 0	question_id	poll_id	race_id	cycle	pollster_id	pollster	sponsor_ids	spons
0	0	148863	77170	8990.0	2022.0	568	YouGov	352	Econo
1	1	148835	77166	8990.0	2022.0	1189	Morning Consult	0	
2	2	148899	77168	8990.0	2022.0	1508	Harris Insights & Analytics	763	Har
3	3	148624	77050	8990.0	2022.0	568	YouGov	352	Econo
4	4	148793	77149	8990.0	2022.0	736	NBC News/The Wall Street Journal	0	
...	...	...	...	...	...	...	...	...	
32597	163	139932	74407	8938.0	2022.0	525	University of New Hampshire	0	
32598	164	139357	74205	8936.0	2022.0	1515	Data for Progress	236	Mov
32599	165	139357	74205	8936.0	2022.0	1515	Data for Progress	236	Mov
32600	166	139460	69647	8949.0	2022.0	67	Braun Research	475,1094	Verr P Ra Verr
32601	167	139460	69647	8949.0	2022.0	67	Braun Research	475,1094	Verr P Ra Verr

32602 rows × 33 columns

Displaying all column names

In [108]:
```
print(FinalData.columns)
```

```
Index(['Unnamed: 0', 'question_id', 'poll_id', 'race_id', 'cycle',
       'pollster_id', 'pollster', 'sponsor_ids', 'sponsors', 'display_name',
       'pollster_rating_id', 'pollster_rating_name', 'fte_grade',
       'sample_size', 'population', 'population_full', 'methodology',
       'office_type', 'start_date', 'end_date', 'nationwide_batch',
       'created_at', 'stage', 'state', 'seat_number', 'election_date',
       'internal', 'ranked_choice_reallocated', 'answer', 'candidate_id',
       'candidate_name', 'candidate_party', 'pct'],
      dtype='object')
```

In [109]:
```
FinalData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32602 entries, 0 to 32601
Data columns (total 33 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Unnamed: 0                 32602 non-null  int64  
 1   question_id                32602 non-null  int64  
 2   poll_id                    32602 non-null  int64  
 3   race_id                    31495 non-null  float64
 4   cycle                      31495 non-null  float64
 5   pollster_id                32602 non-null  int64  
 6   pollster                   32602 non-null  object 
 7   sponsor_ids                32602 non-null  object 
 8   sponsors                   32602 non-null  object 
 9   display_name               32602 non-null  object 
 10  pollster_rating_id         32602 non-null  float64
 11  pollster_rating_name       32602 non-null  object 
 12  fte_grade                  32602 non-null  object 
 13  sample_size                32602 non-null  float64
 14  population                 32602 non-null  object 
 15  population_full            32602 non-null  object 
 16  methodology                32602 non-null  object 
 17  office_type                31495 non-null  object 
 18  start_date                 32602 non-null  object 
 19  end_date                   32602 non-null  object 
 20  nationwide_batch           31495 non-null  object 
 21  created_at                 32602 non-null  object 
 22  stage                      31495 non-null  object 
 23  state                      28650 non-null  object 
 24  seat_number                27456 non-null  float64
 25  election_date              27456 non-null  object 
 26  internal                   28650 non-null  object 
 27  ranked_choice_reallocated  27456 non-null  object 
 28  answer                     28650 non-null  object 
 29  candidate_id               28650 non-null  float64
 30  candidate_name             28650 non-null  object 
 31  candidate_party            27456 non-null  object 
 32  pct                        28650 non-null  float64
dtypes: float64(7), int64(4), object(22)
memory usage: 8.2+ MB
```

Rename all columns

In [110]:
```
FinalData.rename(columns={'Unnamed: 0':'Sr No.','question_id':'Question ID','pol
                     'pollster_id':'Pollster ID','pollster':'Pollster','sponsor_i
                     'display_name':'Display Name','pollster_rating_id':'Pollster
                     'population':'Population','population_full':'Population Full
                     'stage':'Stage','state':'State','seat_number':'Seat Number',
                     'pct':'Pct'
                       },inplace=True)
```

In [111]:
```
print(FinalData.columns)
```

```
Index(['Sr No.', 'Question ID', 'Poll ID', 'Race ID', 'Cycle', 'Pollster ID',
       'Pollster', 'Sponsor IDs', 'Sponsors', 'Display Name',
       'Pollster Rating ID', 'Pollster Rating Name', 'FTE Grade',
       'Sample Size', 'Population', 'Population Full', 'Methodology',
       'Office Type', 'Start Date', 'End Date', 'Nationwide Batch',
       'Created At', 'Stage', 'State', 'Seat Number', 'Election Date',
       'Internal', 'Ranked Choice Reallocated', 'Answer', 'Candidate ID',
       'Candidate Name', 'Candidate Party', 'Pct'],
      dtype='object')
```

In [112]:
```
FinalData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32602 entries, 0 to 32601
Data columns (total 33 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Sr No.                     32602 non-null  int64  
 1   Question ID                32602 non-null  int64  
 2   Poll ID                    32602 non-null  int64  
 3   Race ID                    31495 non-null  float64
 4   Cycle                      31495 non-null  float64
 5   Pollster ID                32602 non-null  int64  
 6   Pollster                   32602 non-null  object 
 7   Sponsor IDs                32602 non-null  object 
 8   Sponsors                   32602 non-null  object 
 9   Display Name               32602 non-null  object 
 10  Pollster Rating ID         32602 non-null  float64
 11  Pollster Rating Name       32602 non-null  object 
 12  FTE Grade                  32602 non-null  object 
 13  Sample Size                32602 non-null  float64
 14  Population                 32602 non-null  object 
 15  Population Full            32602 non-null  object 
 16  Methodology                32602 non-null  object 
 17  Office Type                31495 non-null  object 
 18  Start Date                 32602 non-null  object 
 19  End Date                   32602 non-null  object 
 20  Nationwide Batch           31495 non-null  object 
 21  Created At                 32602 non-null  object 
 22  Stage                      31495 non-null  object 
 23  State                      28650 non-null  object 
 24  Seat Number                27456 non-null  float64
 25  Election Date              27456 non-null  object 
 26  Internal                   28650 non-null  object 
 27  Ranked Choice Reallocated  27456 non-null  object 
 28  Answer                     28650 non-null  object 
 29  Candidate ID               28650 non-null  float64
 30  Candidate Name             28650 non-null  object 
 31  Candidate Party            27456 non-null  object 
 32  Pct                        28650 non-null  float64
dtypes: float64(7), int64(4), object(22)
memory usage: 8.2+ MB
```

Filling Categorical Variable

In [113]:

```
FinalData['State'] = FinalData['State'].fillna('Unknown')
FinalData['Office Type'] = FinalData['Office Type'].fillna('Unknown')
FinalData['Stage'] = FinalData['Stage'].fillna('Undefined')
FinalData['Candidate Name'] = FinalData['Candidate Name'].fillna('Unknown')
FinalData['Candidate Party'] = FinalData['Candidate Party'].fillna('Others')
FinalData['Answer'] = FinalData['Answer'].fillna('Undefined')
FinalData['Election Date'] = FinalData['Election Date'].fillna('Unknown')
FinalData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32602 entries, 0 to 32601
Data columns (total 33 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Sr No.                     32602 non-null  int64  
 1   Question ID                32602 non-null  int64  
 2   Poll ID                    32602 non-null  int64  
 3   Race ID                    31495 non-null  float64
 4   Cycle                      31495 non-null  float64
 5   Pollster ID                32602 non-null  int64  
 6   Pollster                   32602 non-null  object 
 7   Sponsor IDs                32602 non-null  object 
 8   Sponsors                   32602 non-null  object 
 9   Display Name               32602 non-null  object 
 10  Pollster Rating ID         32602 non-null  float64
 11  Pollster Rating Name       32602 non-null  object 
 12  FTE Grade                  32602 non-null  object 
 13  Sample Size                32602 non-null  float64
 14  Population                 32602 non-null  object 
 15  Population Full            32602 non-null  object 
 16  Methodology                32602 non-null  object 
 17  Office Type                32602 non-null  object 
 18  Start Date                 32602 non-null  object 
 19  End Date                   32602 non-null  object 
 20  Nationwide Batch           31495 non-null  object 
 21  Created At                 32602 non-null  object 
 22  Stage                      32602 non-null  object 
 23  State                      32602 non-null  object 
 24  Seat Number                27456 non-null  float64
 25  Election Date              32602 non-null  object 
 26  Internal                   28650 non-null  object 
 27  Ranked Choice Reallocated  27456 non-null  object 
 28  Answer                     32602 non-null  object 
 29  Candidate ID               28650 non-null  float64
 30  Candidate Name             32602 non-null  object 
 31  Candidate Party            32602 non-null  object 
 32  Pct                        28650 non-null  float64
dtypes: float64(7), int64(4), object(22)
memory usage: 8.2+ MB
```

Filling Boolean Variables

In [114]:

```
FinalData['Nationwide Batch'] = FinalData['Nationwide Batch'].fillna('TRUE')
FinalData['Internal'] = FinalData['Internal'].fillna('TRUE')
FinalData['Ranked Choice Reallocated'] = FinalData['Ranked Choice Reallocated'].f
```

Filling Numeric Variables

In [115]:

```
FinalData['Seat Number'] = FinalData['Seat Number'].fillna('0')
FinalData['Pct'] = FinalData['Pct'].fillna('0')
FinalData['Candidate ID'] = FinalData['Candidate ID'].fillna('0')
FinalData['Race ID'] = FinalData['Race ID'].fillna('0')
FinalData['Cycle'] = FinalData['Cycle'].fillna('0')
```

Converting Column Date from object to Datetime

In [116]:

```
FinalData['Start Date'] = pd.to_datetime(FinalData['Start Date'])
FinalData['Start Date'] = FinalData['Start Date'].astype('datetime64[ns]')
FinalData['End Date'] = pd.to_datetime(FinalData['End Date'])
FinalData['End Date'] = FinalData['End Date'].astype('datetime64[ns]')
FinalData['Created At'] = pd.to_datetime(FinalData['Created At'])
FinalData['Created At'] = FinalData['Created At'].astype('datetime64[ns]')
FinalData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32602 entries, 0 to 32601
Data columns (total 33 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Sr No.                    32602 non-null  int64
 1   Question ID               32602 non-null  int64
 2   Poll ID                   32602 non-null  int64
 3   Race ID                   32602 non-null  object
 4   Cycle                     32602 non-null  object
 5   Pollster ID               32602 non-null  int64
 6   Pollster                  32602 non-null  object
 7   Sponsor IDs               32602 non-null  object
 8   Sponsors                  32602 non-null  object
 9   Display Name              32602 non-null  object
 10  Pollster Rating ID        32602 non-null  float64
 11  Pollster Rating Name      32602 non-null  object
 12  FTE Grade                 32602 non-null  object
 13  Sample Size               32602 non-null  float64
 14  Population                32602 non-null  object
 15  Population Full           32602 non-null  object
 16  Methodology               32602 non-null  object
 17  Office Type               32602 non-null  object
 18  Start Date                32602 non-null  datetime64[ns]
 19  End Date                  32602 non-null  datetime64[ns]
 20  Nationwide Batch          32602 non-null  object
 21  Created At                32602 non-null  datetime64[ns]
 22  Stage                     32602 non-null  object
 23  State                     32602 non-null  object
 24  Seat Number               32602 non-null  object
 25  Election Date             32602 non-null  object
 26  Internal                  32602 non-null  object
 27  Ranked Choice Reallocated 32602 non-null  object
 28  Answer                    32602 non-null  object
 29  Candidate ID              32602 non-null  object
 30  Candidate Name            32602 non-null  object
 31  Candidate Party           32602 non-null  object
 32  Pct                       32602 non-null  object
dtypes: datetime64[ns](3), float64(2), int64(4), object(24)
memory usage: 8.2+ MB
```

Converting Dataframe to FinalCSV file

In [117]:

```
FinalData.to_csv(r'C:\Users\harsh\OneDrive\Desktop\103\Project 1\FinalCSV\Electi
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: