U.S. Presidential Elections, 1976-2020

Introduction to Data Science

Saiida Webb

Table of Contents

- Introduction of Dataset
- Objective of Project
- Method, Cleaning
- Storytelling, Visualization
- Conclusion

About the Data

year	state	state_po	state_fips	state_cen	state_ic	office	candidate	party_detailed	writein	candidatevotes	totalvotes	version	notes	party_simplified

- Kaggle, “US Elections Data Set (1976-2020)
- States, state identification, year, candidate information, party information, voting totals
- Amount of unique values, column names and data types

- Shape: 4287 rows, 15 columns
- Head and tail
  - Alphabetical
- Range of years = 44
  - 12 cycles

First Glance Goals

- State vs. Party
- State vs. Voter Turnout
  - Total votes, candidate votes
- Candidate vs. State
- Votes vs. Candidate

Cleansing the Data

Missing Values

```
print("Null Values by Column:")
print(df.isnull().sum())
```

```
Null Values by Column:
year                   0
state                  0
state_po               0
state_fips             0
state_cen              0
state_ic               0
office                 0
candidate            287
party_detailed       456
writein                3
candidatevotes         0
totalvotes             0
version                0
notes               4287
party_simplified       0
dtype: int64
```

Categorical Data

'state', 'state_po', 'office', 'candidate', 'party_detailed', 'writein', 'party_simplified', 'notes'

- Consider relevance
- Impact on the data
- Replace with filler
  - 'Unknown'
  - 'Non-Applicable'
  - Mode

Numerical Data

'year, 'state_fips', 'state_cen', 'state_ic', 'candidatevotes', 'totalvotes', 'version'

- When to drop columns
  - Too high or too low
- Replace with filler
  - Median
  - Mean

Cleansing the Data (cont.)

Replaced

- ‘Candidate’ : Unknown
- ‘Party_detailed’ : Non-Applicable
  - *Why?*

```
df_new.isnull().sum()

year                0
state               0
state_po            0
state_fips          0
state_cen           0
state_ic            0
office              0
candidate           0
party_detailed      0
writein             0
candidatevotes      0
totalvotes          0
version             0
party_simplified    0
dtype: int64
```

Dropped

- ‘Notes’ (100%)
- ‘Writein’ subset (0.07%)

*** No missing values in numerical columns*

Added Columns

Party_Numeric

```
df_clean['party_simplified'].unique()
```

```
array(['DEMOCRAT', 'REPUBLICAN', 'OTHER', 'LIBERTARIAN'], dtype=object)
```

```
party_numeric = {
    'DEMOCRAT' : 0,
    'REPUBLICAN': 1,
    'LIBERTARIAN': 2,
    'OTHER': 3,
}

# Create a new numeric column
df_clean['party_numeric'] = df_new['party_simplified'].map(party_numeric)
df_clean
```

Vote_Percentage

```
df_clean['vote_percentage'] = (df_clean['candidatevotes'] / df_clean['totalvotes']) * 100
df_clean
```

Analysis and Categorizations

Votes Over Time

```
votes_by_year = df_clean.groupby('year')['totalvotes'].sum().reset_index()
print('This is the total amount of votes per year:')
votes_by_year
```

This is the total amount of votes per year:

	year	totalvotes
0	1976-01-01	605944064
1	1980-01-01	663902096
2	1984-01-01	609936856
3	1988-01-01	537099170
4	1992-01-01	770486377
5	1996-01-01	728343795
6	2000-01-01	783441739
7	2004-01-01	768259747
8	2008-01-01	992684830
9	2012-01-01	879479158
10	2016-01-01	941573717
11	2020-01-01	1856741191

Votes By Candidate

```
votes_by_candidate = df_clean.groupby('candidate')['candidatevotes'].sum().reset_index()
votes_by_candidate_sorted = votes_by_candidate.sort_values(by='candidatevotes', ascending=False)
print('Top 10 Most Voted-For Candidates')
votes_by_candidate_sorted.head(10)
```

Top 10 Most Voted-For Candidates

	candidate	candidatevotes
240	TRUMP, DONALD J.	137201208
185	OBAMA, BARACK H.	135398119
33	BUSH, GEORGE W.	112484454
205	REAGAN, RONALD	98353843
46	CLINTON, BILL	92356201
32	BUSH, GEORGE H.W.	87989969
17	BIDEN, JOSEPH R. JR	81268908
40	CARTER, JIMMY	76306787
47	CLINTON, HILLARY	65853581
159	MCCAIN, JOHN	59948283

Votes By Party

```
party_votes = df_clean.groupby('party_simplified')['totalvotes'].sum().sort_values(ascending=False).reset_index()
print('Total Amount of Votes Gained By Party, Simplified (highest to lowest)')
party_votes
```

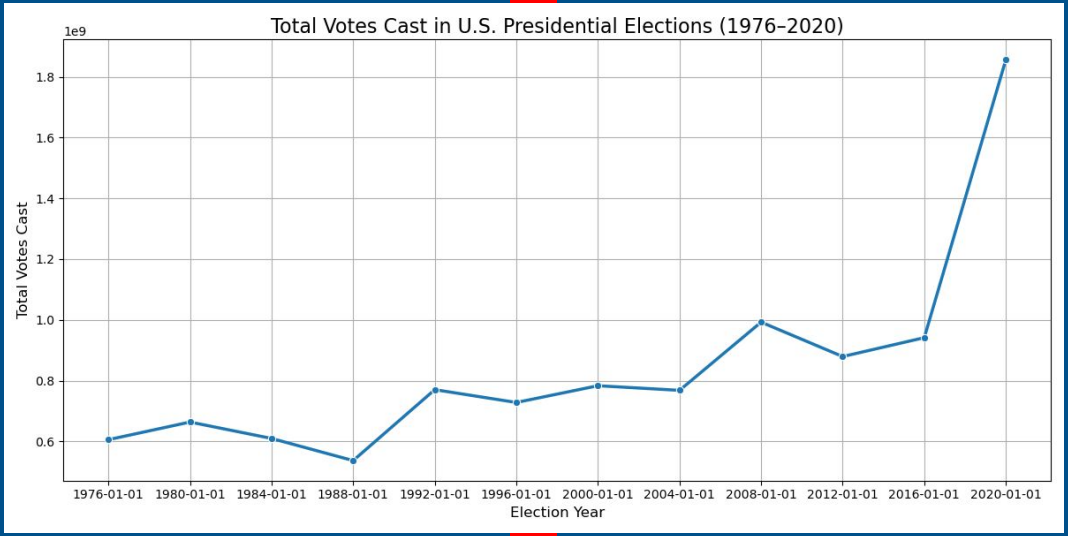
Total Amount of Votes Gained By Party, Simplified (highest to lowest)

	party_simplified	totalvotes
0	OTHER	6275824281
1	DEMOCRAT	1344886700
2	REPUBLICAN	1339929297
3	LIBERTARIAN	1177252462

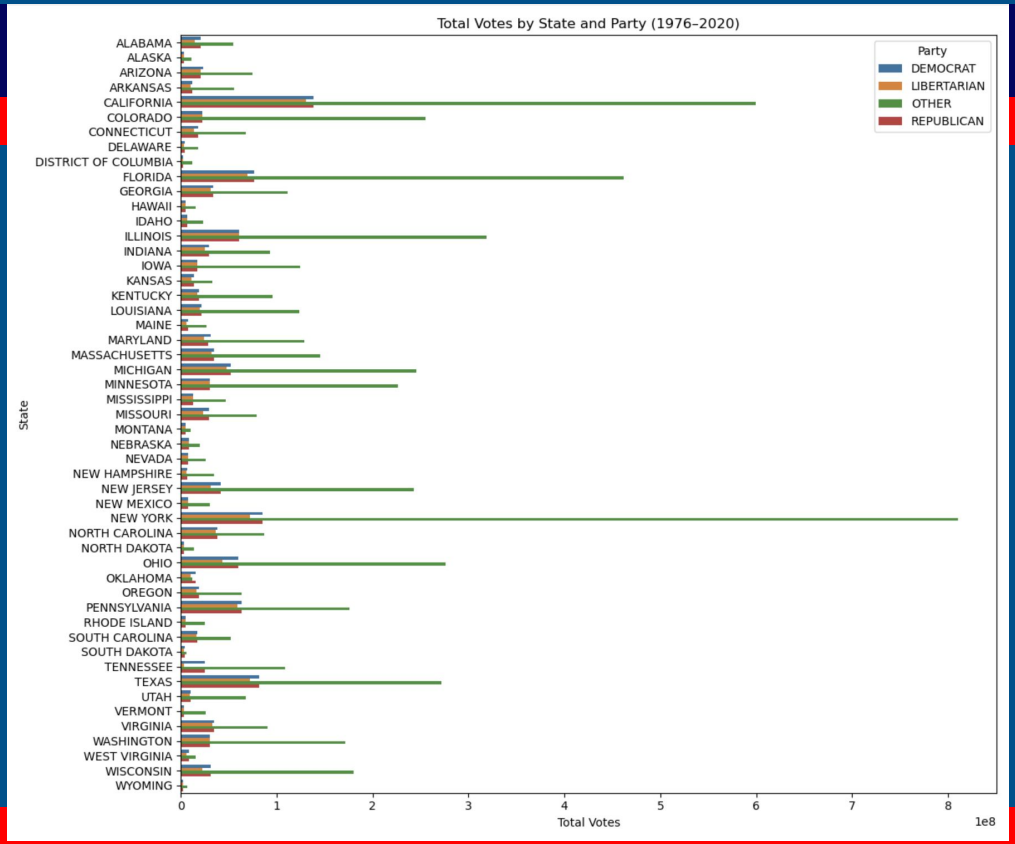
Visualizations

Top 4

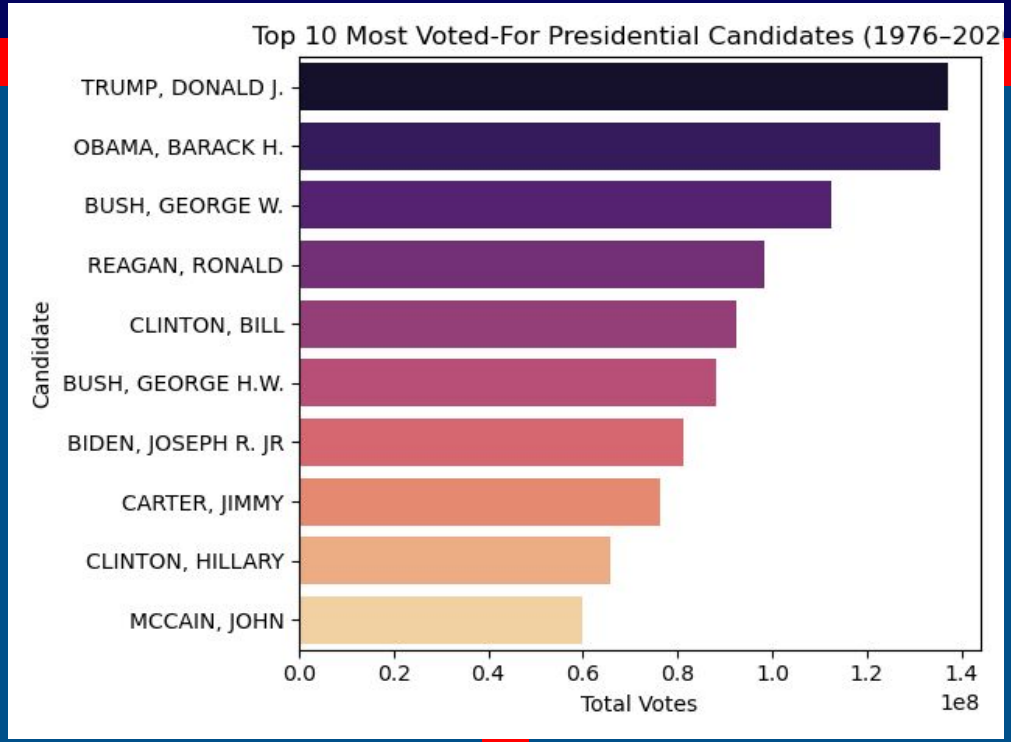
Votes Over Time



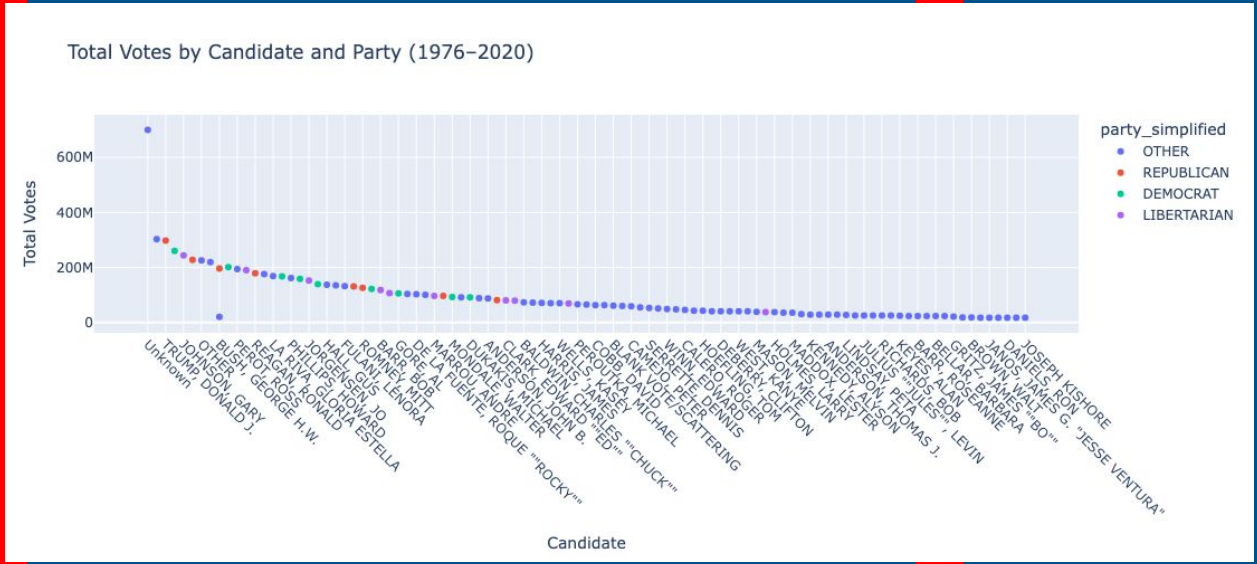
Votes, State, Party

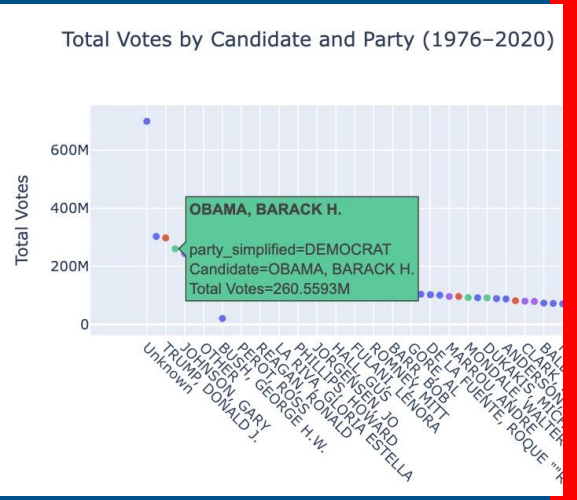


Votes, Candidate



Votes, Party, Candidate





Conclusions

Trends and Findings

- This project analyzed U.S. Presidential elections from 1976–2020 using a dataset covering all 50 states and D.C. Categorical data (like party and candidate) proved more valuable than numerical identifiers for visualizations and analysis. Key findings include:
    - Democrats narrowly led in total votes over Republicans (by 0.1%).
    - Donald Trump received the most votes, likely due to expanded access during the 2020 pandemic.
    - Larger states like California and Texas cast the most votes, reflecting their population size and electoral influence.

Thank You!