Dealing With Historical US Election Data - Part 3

A comprehensive guide to processing, cleaning, and analyzing complex data

Roee Diler

2025-03-24

Introduction

Welcome to the third and final part of the guide on validating historical US election data. In this section, we will focus on making some validation checks to ensure the data is complete and accurate. We will also address any remaining data quality issues and flag potential anomalies that require further investigation.

Required Packages

We load the necessary libraries using `pacman` for package management.

```
rm(list = ls())  # Clear workspace

library(pacman)
p_load(data.table, tidyverse, naniar, Hmisc, rlist, skimr, rstudioapi, haven, labelled, dplyr)
```

Data Setup and Directory Configuration

Define Paths

We set up directory paths dynamically based on the script location.

```
# Get the directory where this script is located
script_path <- dirname(getSourceEditorContext()$path)
basic_path <- script_path
data_path <- file.path(basic_path, "data")
icpsr_path <- file.path(data_path, "ICPSR/ICPSR_00001")
```

Data Import and Initial Examination

Loading Elections Data

We begin by importing the final clean and organized data from the previous steps. This data has been processed to address missing values, inconsistencies, and other data quality issues.

```
# Load the clean ICPSR data that was previously processed
icpsr_data <- read_rds(paste0(data_path, "/elections_returns/election_returns.rds"))
icpsr_data <- as.data.frame(icpsr_data)

# Display the first few rows of the dataset
icpsr_data
```

ICPSRST <int>	STATE <chr>	ICPSRCTY <int>	ICPSRNAM <chr>	YEAR <dbl>	ELECT_OFFICE_CODE <dbl+lbl>
11	Delaware	10	KENT	1823	2
11	Delaware	10	KENT	1823	2
11	Delaware	10	KENT	1823	2
11	Delaware	10	KENT	1823	2
11	Delaware	10	KENT	1823	2
11	Delaware	10	KENT	1823	2
11	Delaware	30	NEW CASTLE	1823	2
11	Delaware	30	NEW CASTLE	1823	2
11	Delaware	30	NEW CASTLE	1823	2
11	Delaware	30	NEW CASTLE	1823	2

1-10 of 10,000 rows | 1-6 of 22 columns Previous 1 2 3 4 5 6 … 1000 Next

Dataset dimensions: 2694714 rows and 22 columns

A quick summary of the final dataset:

```
skim(icpsr_data)
```

Data summary

Name	icpsr_data
Number of rows	2694714
Number of columns	22
_____	
Column type frequency:	
character	8
numeric	14
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
STATE	0	1.00	4	14	0	50	0
ICPSRNAM	0	1.00	3	17	0	1944	0
ELECT_OFFICE	0	1.00	3	4	0	5	0
ELECT_TYPE	39466	0.99	1	1	0	2	0
PARTY_NAME	176788	0.93	1	51	0	917	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
source	0	1.00	6	6	0	198	0
ELECT_OFFICE_ORIG	0	1.00	3	4	0	6	0
ELECT_TYPE_ORIG	39466	0.99	1	1	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ICPSRST	0	1.00	37.77	16.34	1	24	40.0	49	82	
ICPSRCTY	2118	1.00	987.03	958.10	10	350	790.0	1310	8400	
YEAR	0	1.00	1914.79	35.51	1823	1890	1920.0	1944	1968	
ELECT_OFFICE_CODE	0	1.00	2.55	0.93	1	2	3.0	3	5	
PARTY_CODE	0	1.00	1923.48	3438.44	1	200	361.0	646	9999	
VOTES	797	1.00	1616.83	13480.62	0	0	2.0	681	1692591	
TOTAL	3360	1.00	11127.30	56324.35	0	1302	3242.0	7379	5219597	
TOTAL2	2694183	0.00	4093.76	8427.16	213	1142	2065.0	3420	62823	
TOTAL_VOTES	0	1.00	11274.82	56794.75	0	1327	3281.0	7461	5219597	
MONTH	2694171	0.00	8.34	3.21	1	6	8.0	11	11	
CONG_NUM	2694690	0.00	70.50	0.51	70	70	70.5	71	71	
INDEX	0	1.00	1.13	0.86	1	1	1.0	1	25	
ELEC_ID	0	1.00	269255.64	170549.53	1	115420	263309.0	425685	564721	
ELECT_OFFICE_CODE_ORIG	31514	0.99	2.71	1.23	1	2	3.0	3	7	

Labels, variable types, and value labels:

```
look_for(icpsr_data) %>% print()
```

```
##  pos variable              label            col_type missing values
##  1   ICPSRST               ICPSR State Code int      0
##  2   STATE                 State Name       chr      0
##  3   ICPSRCTY              ICPSR County Co~ int      2118
##  4   ICPSRNAM              County Name      chr      0
##  5   YEAR                  Year of Election dbl      0
##  6   ELECT_OFFICE_CODE     Elected Office ~ dbl+lbl  0       [1] President
##                                                               [2] Governor
##                                                               [3] Congress
##                                                               [4] Senate
##                                                               [5] Attorney_Ge~
##  7   ELECT_OFFICE          Elected Office   chr+lbl  0       [PRES] President
##                                                               [GOV] Governor
##                                                               [CONG] Congress
##                                                               [SEN] Senate
##                                                               [ATGN] Attorney~
##  8   ELECT_TYPE            Election Type    chr+lbl  39466   [G] General
##                                                               [S] Special
##  9   PARTY_CODE            Party Code       int      0
##  10  PARTY_NAME            Party Name       chr      176788
##  11  VOTES                 Votes            int      797
##  12  TOTAL                 The Original TO~ int      3360
##  13  TOTAL2                Second TOTAL en~ int      2694183
##  14  TOTAL_VOTES           Sum of VOTES Pe~ int      0
##  15  source                Original ICPSR ~ chr      0
##  16  MONTH                 Month of Electi~ dbl      2694171
##  17  CONG_NUM              Congress Number  dbl      2694690
##  18  INDEX                 Index            int      0
##  19  ELEC_ID               Election ID      int      0
##  20  ELECT_OFFICE_CODE_ORIG Original Electe~ dbl+lbl  31514   [1] President
##                                                               [2] Governor
##                                                               [3] Congress
##                                                               [4] Senate
##                                                               [5] Senate
##                                                               [6] Senate
##                                                               [7] Attorney_Ge~
##  21  ELECT_OFFICE_ORIG     Original Electe~ chr+lbl  0       [PRES] President
##                                                               [GOV] Governor
##                                                               [CONG] Congress
##                                                               [SEN] Senate
##                                                               [HAL] House_of_~
##                                                               [ATGN] Attorney~
##  22  ELECT_TYPE_ORIG       Original Electi~ chr+lbl  39466   [G] General
##                                                               [S] Special
##                                                               [M] Multiple_G_~
##                                                               [W] Multiple_S_~
```

Data Quality Checks

Vote Count Distribution Analysis

Let's examine the distribution of vote counts to identify potential anomalies.

```
vote_dist <- table(icpsr_data$VOTES) %>% as.data.frame() %>% arrange(desc(Freq))
head(vote_dist)
```

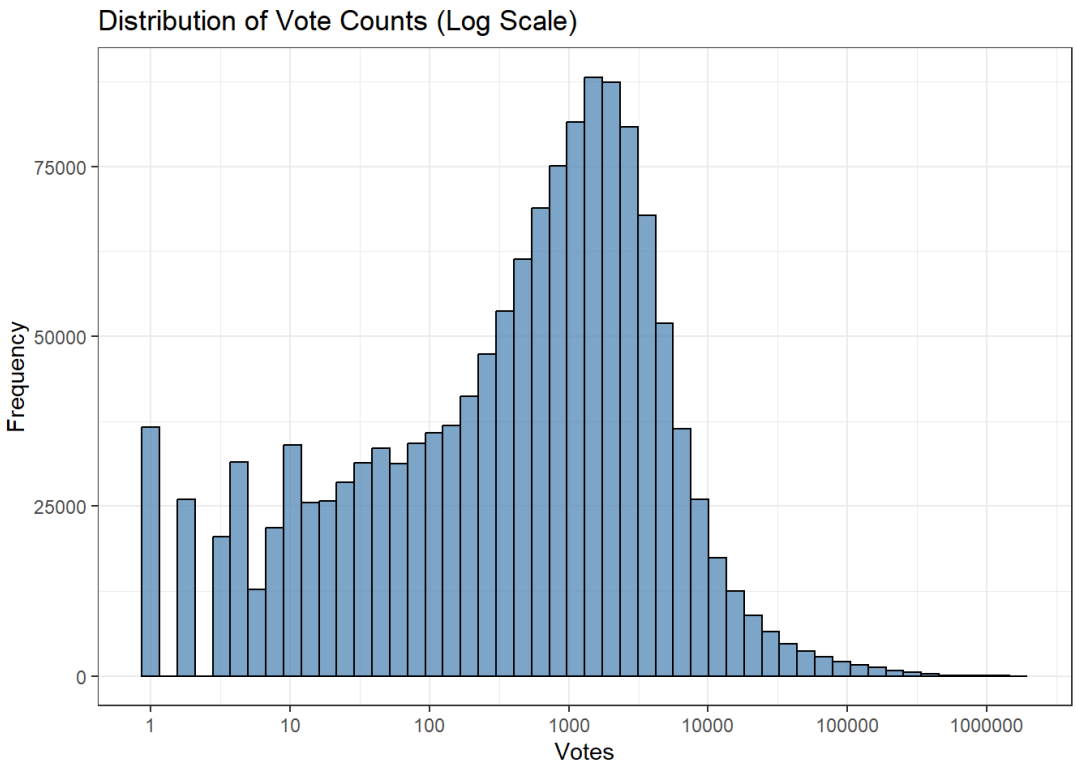
	Var1 <fct>	Freq <int>
1	0	1297632
2	1	36561

	Var1 <fct>	Freq <int>
3	2	26021
4	3	20532
5	4	17122
6	5	14403

6 rows

We observe that the vast majority of vote counts are low, with zero votes being the most common value. Let's visualize the distribution of vote counts on a log scale to better understand the data.

```
# create data for the plot using tidyverse syntax
icpsr_data %>%
  filter(!is.na(VOTES) & VOTES !=0) %>%
  ggplot(aes(x = VOTES)) +
  geom_histogram(bins = 50, color = "black", fill = "steelblue", alpha = 0.7) +
  scale_x_log10(breaks = c(1, 10, 100, 1000, 10000, 100000, 1000000)) +
  labs(title = "Distribution of Vote Counts (Log Scale)",
       x = "Votes",
       y = "Frequency") +
  theme_bw()
```



We can see that the distribution of vote counts looks like log-normal distribution.

Presedinential Election Over Time

We would like to check the validity of the data, the most reliable source of election data outside this dataset is about presidential elections. We can start by checking the number of presidential elections in the dataset and the years they cover. We know that presidential elections are held every four years, so we can use this information to validate the dataset.

```
# Expected presidential election years
ex_pres_years <- seq(1824, 1968, 4)

# Check the years covered in the dataset
pres_years <- icpsr_data %>% filter(ELECT_OFFICE == "PRES") %>% pull(YEAR) %>% unique()

diff1 <- setdiff(ex_pres_years, pres_years)
diff1
```

```
## numeric(0)
```

We can see that the dataset is not missing any presidential election years.

```
diff2 <- setdiff(pres_years, ex_pres_years)
diff2
```

```
## [1] 1878
```

However, there is an additional year in the dataset that is not a presidential election year and is labeled as such. We should investigate this further to understand the discrepancy.

```
icpsr_data %>% filter(YEAR %in% diff2 & ELECT_OFFICE == "PRES") %>%
  distinct() %>%
  select(STATE, YEAR, ELECT_OFFICE, ELECT_OFFICE_CODE, ELECT_TYPE, PARTY_CODE, PARTY_NAME) %>%
  apply(2, unique)
```

```
##             STATE              YEAR      ELECT_OFFICE ELECT_OFFICE_CODE
##        "Arkansas"            "1878"            "PRES"               "1"
##        ELECT_TYPE        PARTY_CODE        PARTY_NAME
##               "G"             "100"       "DEMOCRAT"
```

It seems like some kind of error in the data, maybe some of the data is mislabeled.

```
icpsr_data %>% filter(YEAR %in% diff2 & STATE == "Arkansas") %>%
  select(STATE, YEAR, ELECT_OFFICE, ELECT_OFFICE_CODE, ELECT_TYPE, PARTY_CODE, PARTY_NAME) %>%
  distinct()
```

STATE <chr>	YEAR <dbl>	ELECT_OFFICE <chr+lbl>	ELECT_OFFICE_CODE <dbl+lbl>	ELECT_TYPE <chr+lbl>	PARTY_CODE <int>
Arkansas	1878	PRES	1	G	100
Arkansas	1878	CONG	3	G	100
Arkansas	1878	CONG	3	G	320
Arkansas	1878	CONG	3	G	330

4 rows | 1-6 of 7 columns

We can see that we have data for the Democratic party both for the presidential and the congressional elections. Maybe the presidential data is mislabeled and should be congressional data. Let's see how many data entries we have for the each party in this election.

```
icpsr_data %>% filter(YEAR %in% diff2 & STATE == "Arkansas") %>%
  select(ELECT_OFFICE, PARTY_CODE) %>%
  table()
```

```
##             PARTY_CODE
## ELECT_OFFICE 100 320 330
##         CONG  50  50  50
##         PRES  74   0   0
```

Unfortunately, this it is probably not the case, since we have 50 entries for each party in the congressional election and 74 entries for the Democratic party in the presidential election.

We can assume that the data is mislabeled and the error is in the year of the election. We can try and check if the data is missing for the year 1876 or 1880.

```
icpsr_data %>% filter(YEAR %in% c(1876, 1880) &
                        STATE == "Arkansas" &
                        ELECT_OFFICE == "PRES" &
                        PARTY_CODE == 100) %>%
  select(YEAR) %>%
  table()
```

```
## YEAR
## 1876 1880
##   74   73
```

It seems like the data is not missing, but maybe the data is simply duplicated.

```
icpsr_data %>% filter(YEAR %in% c(1876, 1878, 1880) &
                        STATE == "Arkansas" &
                        ELECT_OFFICE == "PRES" &
                        PARTY_CODE == 100) %>%
  group_by(YEAR) %>%
  summarise(sum(VOTES))
```

YEAR <dbl>	sum(VOTES) <int>
1876	58086
1878	88726
1880	60489

3 rows

This data is not duplicated either, so unfortunately we cannot be sure what is the problem with this data. We saw in the previous part that this data set has some data quality issues in the columns' names data. Hence, it is only logical to assume that this is another data quality issue.

To fix this issue, we will need some deeper investigation and maybe some external data sources. For now, we will leave this issue as it is.

Vote Count Consistency Check

Let's try to make another check and see if the total vote counts in the presidential elections are consistent with the historical data.

First, we need to calculate the total votes in the dataset for each presidential election year (and exclude the year 1878, which is not a presidential election year).

```
pres_votes <- icpsr_data %>%
  filter(ELECT_OFFICE == "PRES" & YEAR != 1878) %>%
  group_by(ELEC_ID) %>%
  # leave only one row for each election
  slice(1) %>%
  group_by(YEAR) %>%
  summarise(TOTAL = sum(TOTAL, na.rm = T), TOTAL_VOTES = sum(TOTAL_VOTES)) %>%
  arrange(YEAR)
```

Next, we will compare the total votes in the dataset with historical data from the General Presidential Elections Turnout Data from 1824 to 1968 (retrieved from ChatGPT). This is obviously not the most reliable source of data, but it should give us a rough idea of the total votes in the presidential elections.

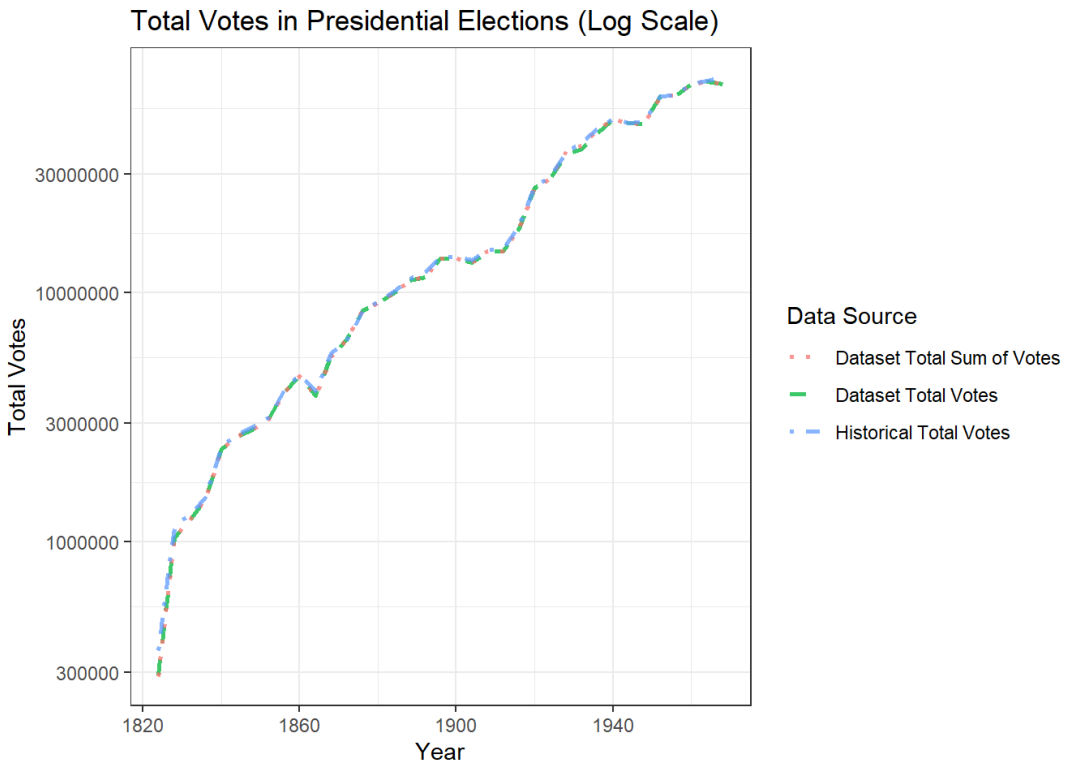
```
election_votes <- data.frame(
  YEAR = seq(1824, 1968, by = 4),
  GPT_VOTES = c(
    365833, 1155350, 1291728, 1501290, 2411808, 2694284, 2884746, 3144196,
    4053967, 4685561, 4034142, 5716082, 6430149, 8411618, 9217410, 10067610,
    11382242, 12059351, 13923102, 13970692, 13524463, 14887808, 15036869,
    18528728, 26750525, 29095023, 36805951, 39758759, 45646817, 49900706,
    47977296, 48793535, 61551118, 62026093, 68838204, 70645592, 73211875
  )
)
```

Now, we will merge the two datasets and plot the total votes in the dataset against the historical data.

```
pres_votes <- merge(pres_votes, election_votes, by = "YEAR")
```

And now we can plot the total votes in the dataset against the historical data.

```
# Line plot of total votes in the dataset vs. historical data
pres_votes %>%
  ggplot(aes(x = YEAR)) +
  geom_line(aes(y = TOTAL, color = "Dataset Total Votes"), size = 1, lty = 2, alpha = 0.75) +
  geom_line(aes(y = TOTAL_VOTES, color = "Dataset Total Sum of Votes"), size = 1, lty = 3, alpha = 0.75) +
  geom_line(aes(y = GPT_VOTES, color = "Historical Total Votes"), size = 1, lty = 4, alpha = 0.75) +
  scale_y_log10() +
  labs(title = "Total Votes in Presidential Elections (Log Scale)",
       x = "Year",
       y = "Total Votes",
       color = "Data Source") +
  theme_bw()
```



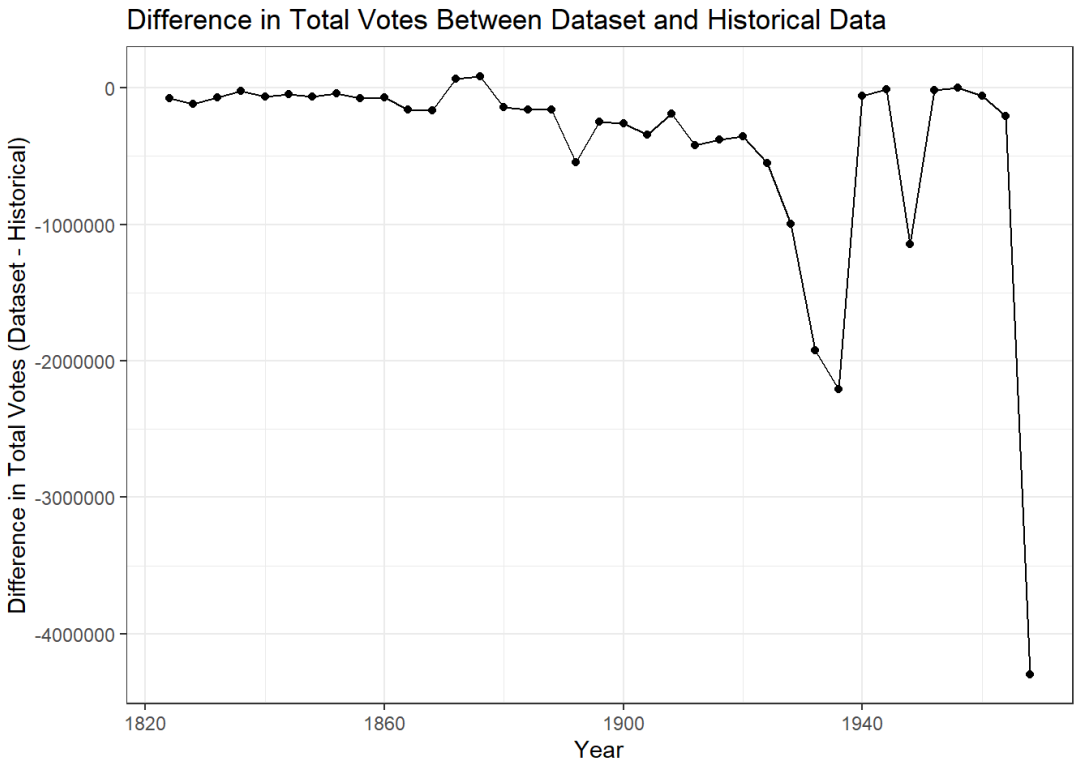
We can see that there is strong correlation between the total votes in the dataset and the historical data. Let's calculate the difference between the total votes in the dataset and the historical data and the percentage of this difference from the total votes.

These two metrics will help us understand the discrepancy between the dataset and the historical data.

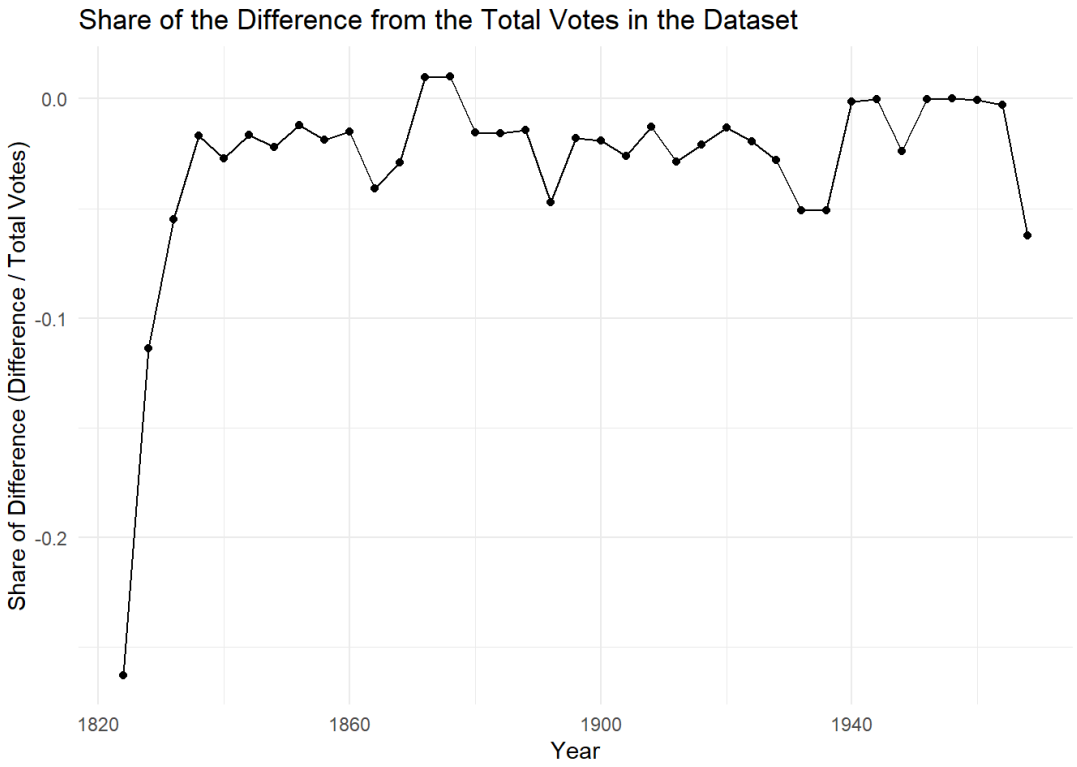
```
pres_votes <- pres_votes %>%
  mutate(DIFF = TOTAL - GPT_VOTES) %>%
  mutate(SHARE = DIFF / TOTAL)
```

Now, we can plot the difference in total votes between the dataset and the historical data.

```
pres_votes %>%
  ggplot(aes(x = YEAR, y = DIFF)) +
  geom_line() +
  geom_point() +
  labs(title = "Difference in Total Votes Between Dataset and Historical Data",
       x = "Year",
       y = "Difference in Total Votes (Dataset - Historical)") +
  theme_bw()
```




```
pres_votes %>%
  ggplot(aes(x = YEAR, y = SHARE)) +
  geom_line() +
  geom_point() +
  labs(title = "Share of the Difference from the Total Votes in the Dataset",
       x = "Year",
       y = "Share of Difference (Difference / Total Votes)") +
  theme_minimal()
```



This is very interesting, we can see that both the difference between the dataset and the historical data, and the share of the difference from the total votes are minimal in most of the years.

However, some discrepancies can be seen in later years, while the share of the difference from the total votes experiences disparity in some early years. This is probably due to the very low total votes in the dataset in the early years and the very high total votes in the later years.

Which indicator is more important is a context-dependent question, but it is good practice to check both of them.

Conclusions and Recommendations

We saw that the dataset has some data quality issues, but it is mostly reliable. We also saw that the dataset is mostly consistent with the historical data, but there are some discrepancies in some years.

Key Findings

1. **Data Quality Issues**: The ICPSR election dataset contains several types of data quality issues, including inconsistent missing value codes and potentially misclassified values.
2. **Missing Data Patterns**: Approximately 20% of vote count values were identified as missing or suspicious and converted to NA. This level of missingness should be considered when drawing conclusions from analyses.

Recommendations for Analysis

1. **Missing Data Handling**: Consider using multiple imputation techniques for analyses requiring complete data, but be cautious about imputing vote counts without strong theoretical justification.
2. **Temporal Comparisons**: When comparing election results across time periods, account for changes in reporting standards, state boundaries, and population distributions.
3. **Validation with Alternative Sources**: Where possible, validate key findings against alternative election data sources such as state election board records or newspaper archives.
4. **Transparency in Reporting**: When publishing analyses based on this data, clearly document the data cleaning steps and limitations to ensure reproducibility and proper interpretation.

Summary

This guide has provided a comprehensive overview of the process of validating historical US election data. By following the steps outlined in this guide, you can ensure that your analyses are based on high-quality, reliable data. Remember to document your data cleaning and validation steps thoroughly to maintain transparency and reproducibility in your research.

I hope you found this guide helpful and informative. If you have any questions or feedback, please feel free to reach out. Thank you for reading!