Name: Keerthi Devireddy

Student ID: 880173271

# COVID-19 Data Analysis Project Stage – I

## Report on Presidential Election Results (Enrichment Dataset)

## (Task 2 Part -2)

**Part 1:**

**Report describing the enrichment data and datatype – variable dictionary.**

**Table of contents**

**Introduction**

For a more comprehensive analysis of the COVID-19 dataset from usafacts.org, we have enriched our primary dataset with additional information from the "Presidential Election Results (Political leaning) dataset". This additional data provides insights into the political leanings of different regions in the United States during 2020 Presidential elections.

**Overview**

Source: Kaggle

URL: https://www.kaggle.com/unanimad/us-election-2020

Dataset overview:

Provides information like the state, county, candidate, party, total votes and the election result or candidate status (won or not).

**Dataset Description**

This dataset includes information related to the 2020 United States Presidential Election, specifically focusing on the vote counts and percentages for each major candidate at the county level. Additionally, it provides details about the political leaning of each county based on the candidate who received majority of votes in the county.

**Datatype – variable dictionary**

The data dictionary provides a clear understanding of the variables present in the Presidential Election Results – 2020 dataset, their definition, datatypes, possible values, and whether they include null data, or the field is required for analysis or not. It serves as a valuable reference for data analysis and interpretation.

**Geographic Information:** State, County

**Election Information:** Candidate, party, total votes, status (won or not)

**Presidential Election Results (Political leanings) Datasets**

**1. president_county_candidate.csv**

| Variable | Definition | Data type | Possible values | Any missing values? | Required? |
|---|---|---|---|---|---|
| state | The name of the US state. | String object | Names of US states. | No, the data does not contain any null or missing values. | Yes, this is a mandatory field. |
| county | The name of the county within US state. | String object | Names of US Sate counties. | No, the data does not contain any null or | Yes, this is a mandatory field. |

| | | | | | |
|---|---|---|---|---|---|
| | | | | missing values. | |
| candidate | A candidate running in the election for a county representing their specific party. | String | Person from list of people standing in elections. | No, the data does not contain any null or missing values. | Yes, this is a mandatory field. |
| party | A political organization that fields candidates. | String | Name of a party from list of political parties. | No, the data does not contain any null or missing values. | Yes, this is a mandatory field. |
| total_votes | Number of votes secured by won candidate. | Integer | Positive integers representing total votes count. | No, the data does not contain any null or missing values. | Yes, this is a mandatory field. |
| won | It is a Boolean value representing the status of the participated candidate (won or not) | Boolean | Boolean values (True- the candidate won, False- if lost) | No, the data does not contain any null or missing values. | Yes, this is a mandatory field. |

**2. president_county.csv**

| **Variable** | **Definition** | **Datatype** | **Possible values** | **Required?** |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| county | The name of the county within US state. | String | Names of US Sate counties. | Yes |
| current_votes | Total number of votes casted in 2020 presidential elections from the county. | Integer | Positive integers representing total polled votes count. | Yes |
| total_votes | Count of total votes in the county. | Integer | Positive integers representing total voter's count. | Yes |
| percent | Percentage of votes polled. | Float | Decimal values indicating polling %. | Yes |

**3. governors_county.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| county | The name of the county within US state. | String | Names of US Sate counties. | Yes |
| current_votes | Total number of votes casted in 2020 governor elections for the county. | Integer | Positive integers representing total polled votes count. | Yes |
| total_votes | Count of total votes in the county. | Integer | Positive integers representing total voter's count. | Yes |
| percent | Percentage of votes polled. | Float | Decimal values indicating polling %. | Yes |

**4.governors_county_candidate.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| county | The name of the county within US state. | String | Names of US Sate counties. | Yes |
| candidate | A candidate running in the election for a county for the governor's role. | String | Person from list of people standing in elections. | Yes |
| party | A political organization that fields candidates. | String | Name of a party from list of political parties. | Yes |
| votes | Number of votes secured by won candidate. | Integer | Positive integers representing total votes count. | Yes |
| won | It is a Boolean value representing the status of the participated candidate (won or not). | Boolean | Boolean values (True- the candidate won, False- if lost) | Yes |

**5. governors_state.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| votes | Count of total votes polled for governor in the state. | Integer | Positive integers representing total votes count. | Yes |

**6. house_candidate.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| district | The name of the US district. | String | Names of districts. | Yes |
| candidate | The name of the candidate contesting for house. | String | Names of candidates. | Yes |
| Party | Name of candidate's political organization. | String | Name of a party from list of political parties. | Yes |
| total_votes | Count of total votes secured. | Integer | Positive integers representing total votes count. | Yes |
| won | It is a Boolean value representing the status of the participated candidate (won or not). | Boolean | Boolean values (True- if the candidate won, False- if lost). | Yes |

**7.house_state.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| district | The name of the district within US state. | String | Names of US Sate districts. | Yes |
| current_votes | Total number of votes casted. | Integer | Positive integers representing total polled votes count. | Yes |
| total_votes | Count of total votes. | Integer | Positive integers representing total voter's count. | Yes |
| percent | Percentage of votes polled. | Float | Decimal values indicating polling %. | Yes |

**8.president_state.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| total_votes | Count of total votes polled for president election from the state. | Integer | Positive integers representing total votes count. | Yes |

**9. senate_county.csv**

| Variable | Definition | Datatype | Possible values | Required? |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| county | The name of the county within US state. | String | Names of US Sate counties. | Yes |

| | | | | |
|---|---|---|---|---|
| current_votes | Total number of votes casted for the senator. | Integer | Positive integers representing total polled votes count. | Yes |
| total_votes | Count of total votes in the county | Integer | Positive integers representing total voter's count. | Yes |
| percent | Percentage of votes polled. | Float | Decimal values indicating polling %. | Yes |

**10. senate_county_candidate.csv**

| **Variable** | **Definition** | **Datatype** | **Possible values** | **Required?** |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| county | The name of the county within US state. | String | Names of US Sate counties. | Yes |
| candidate | A candidate running in the election for a county for the senate elections. | String | Person from list of people standing in elections. | Yes |
| party | A political organization that fields candidates. | String | Name of a party from list of political parties. | Yes |
| Total_votes | Number of votes secured by won candidate. | Integer | Positive integers representing total votes count. | Yes |

**11. senate_state.csv**

| **Variable** | **Definition** | **Datatype** | **Possible values** | **Required?** |
|---|---|---|---|---|
| state | The name of the US state. | String | Names of US states. | Yes |
| total_votes | Count of total votes polled for senator from the state. | Integer | Positive integers representing total votes count. | Yes |

We will mainly use the president_county_candidate.csv dataset for future analysis and in the tasks that are needed to be performed further (task 3 part 2).

## Part 2:

## Identifying the individual variable which map between the datasets and merging the enrichment data with the primary COVID-19 dataset.

To merge the enrichment data with the primary covid-19 dataset from usafacts.org, we can make use of the common variables in both the datasets 'State' and 'County' to create a unified dataset. But the state name in the enrichment dataset and the state name in our primary covid19 dataset are not in the same format. In enrichment the state column has full name of the states, whereas the data present in the state column of the covid19 dataset do contain the short form notation of the state names. So, to resolve this issue we can take help of an additional dataset which includes state names and their respective short form notations. By, using this specific dataset we can either replace or add an additional column in our enrichment dataset to include the short form notations of the state names and finally we can perform merge operation on the datasets with the help of the newly obtained short form notations of the state names and the county column.

The query would be similar to:

Merge_data= covid_data.merge(election_data, on=['State', 'County'], how='left')

## Part 3:

## Describing how enrichment data can help in the analysis of COVID-19 spread.

The enrichment data from the Presidential Election Result dataset can significantly contribute to the analysis of COVID-19 spread in several ways:

**Political Leaning vs. COVID-19 Response**: We can examine whether political leaning at the county level correlates with differences in COVID-19 case counts. This may lead to hypotheses about how political factors influenced public health responses.

**Voting Patterns and COVID-19 Impact:** By analyzing voting patterns alongside COVID-19 data, we can find whether areas that voted for a particular candidate or party experienced different levels of COVID-19 impact, potentially due to varying policies or behaviors.

### Initial Hypothesis Questions:

Is there any correlation between the political leaning of a county and the number of covid cases reported?

Did counties that voted for a specific candidate in 2020 Presidential party experience a different trajectory of COVID-19 cases and deaths during the pandemic?

The enrichment data adds a political dimension to our analysis, allowing to explore potential relationships between political factors and the spread of COVID-19.