**CRISP DM METHODOLOGY**

**BUSINESS UNDERSTANDING**

The winner in elections is determined by the candidate with the grandest electors. Grand electors are attributed at the state level: in each of the 51 states, there is a given number of grand electors to win (roughly, but not exactly, proportional to the size of the state) and the presidential candidate receiving the most local votes wins ALL the Grand Electors in that state.

**BUSINESS OBJECTIVE**

* To help a candidate running for the presidency in the United States to win the elections.

**ACCESSING THE CURRENT SITUATION**

* Inventory of resources
  + Resources available for the project include:
    - Datasets:
      * Grand electors by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1AuPlNXKjmEdRTLGqKbP-OP1XZ5UQPIPw/view?usp=sharing)]
      * Population by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1VKt_hF2pRqPxcNb1DKotkVXWNd2HX_KL/view?usp=sharing)]
  + Software:
    - An online IDE is known as google collab edit.
    - SQLite

**BUSINESS SUCCESS CRITERIA**

Targeting all the states at the top of the list until the cumulative sum (also called running total) of grand electors won is larger than half the total number of Grand Electors in the country. These states will be prioritized more.

**PROCEDURE PLAN**

| **Plan** | **Duration** |
| --- | --- |
| Loading SQLite database into the environment | 12:50 PM |
| Loading the datasets into the environment | 12:51 PM |
| Cleaning the data | 12:53-1:00 PM |
| Merging the two tables and creating a new table (merged\_data) | 1:04-1:10 PM |
| Creating a new table (merged\_data\_sorted) | 2:00-2:10 PM |
| Creating a new table (final) | 3:00-4:00 PM |
| Data analysis | 4:00-4:30 PM |

**DATA UNDERSTANDING**

**Describing the data**

1. Grand electors by State dataset.

| **Column name** | **Datatype** |
| --- | --- |
| State | Char |
| GrandElectors | Int |

1. Population by State dataset.

| **Column name** | **Datatype** |
| --- | --- |
| State | Char |
| Population | Int |

**Data Previewing**

Previewing the first 5 rows in each dataset:

1. Grand electors by State dataset:

| State | GrandElectors |
| --- | --- |
| Alabama | 9 |
| Alaska | 3 |
| Arizona | 11 |
| Arkansas | 6 |

1. Grand electors by State dataset:

| State | Population |
| --- | --- |
| CALIFORNIA | 39144818 |
| TEXAS | 27469114 |
| FLORIDA | 20271272 |
| NEW YORK | 19795791 |

**Data exploration**

* The data in the Grand elector’s column is in uppercase while that for Population by State isn’t.

**Verifying Data Integrity**

1. Grand electors by State:
   1. The data is complete.
   2. The assumption that the data is correct and up to date
2. Population by State:
   1. The data is complete.
   2. The assumption that the data is correct and up to date

**DATA PREPARATION**

The datasets to be used for this project are stated below:

* + - * Grand electors by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1AuPlNXKjmEdRTLGqKbP-OP1XZ5UQPIPw/view?usp=sharing)]
      * Population by State. [[Link](https://drive.google.com/a/moringaschool.com/file/d/1VKt_hF2pRqPxcNb1DKotkVXWNd2HX_KL/view?usp=sharing)]

**Data cleaning**

Change the data in the state column of Grand electors by State dataset to march that one the one in the Grand electors by State dataset.

**Data Integrating**

After the data cleaning, it’s now possible to join the two tables via the state column. Through this, I’ll come up with a new table known as ‘merged\_table’.

**DATA ANALYSIS**

1. Merging the two tables

* **Preview of the merged\_table**

| State | GrandElectors | Population |
| --- | --- | --- |
| ALABAMA | 9 | 4858979 |
| ALASKA | 3 | 738432 |
| ARIZONA | 11 | 6828065 |
| ARKANSAS | 6 | 2978204 |
| CALIFORNIA | 55 | 39144818 |
| ALABAMA | 9 | 4858979 |

1. Ordering the states, by descending ratio of Grand Electors per capital to Population. This will make our priority list.

* Then create a new table of the sorted data know as **merged\_table\_sorted.**
  + Preview of the **merged\_table\_sorted table**

| State | GrandElectors | Population |
| --- | --- | --- |
| WYOMING | 3 | 586107 |
| ALASKA | 3 | 738432 |
| DC | 3 | 672228 |
| VERMONT | 3 | 626042 |
| DELAWARE | 3 | 945934 |
| WYOMING | 3 | 586107 |

1. Computing the cumulative frequency of the GrandElectors then saving it a new column, in a new table known as final.
   * Preview of the final table

| State | GrandElectors | Population | Cumulative\_frequency |
| --- | --- | --- | --- |
| WYOMING | 3 | 586107 | 5.0 |
| ALASKA | 3 | 738432 | 4.0 |
| DC | 3 | 672228 | 4.0 |
| VERMONT | 3 | 626042 | 4.0 |
| DELAWARE | 3 | 945934 | 3.0 |
| WYOMING | 3 | 586107 | 5.0 |

1. Computing the half total of GrandElectors overall(also known as threshold)
2. Selecting all the states for which the running total is below or equal to the threshold.

**Conclusion**

The number of states to prioritize the elections are, 21, which are less than the number of states which I began with these are:

| State | GrandElectors | Population | Cumulative\_frequency |
| --- | --- | --- | --- |
| WYOMING | 3 | 586107 | 5.0 |
| ALASKA | 3 | 738432 | 4.0 |
| DC | 3 | 672228 | 4.0 |
| VERMONT | 3 | 626042 | 4.0 |
| DELAWARE | 3 | 945934 | 3.0 |
| MAINE | 4 | 1329328 | 3.0 |
| NEW HAMPSHIRE | 4 | 1330608 | 3.0 |
| NORTH DAKOTA | 3 | 756927 | 3.0 |
| RHODE ISLAND | 4 | 1056298 | 3.0 |
| SOUTH DAKOTA | 3 | 858469 | 3.0 |
| ARKANSAS | 6 | 2978204 | 2.0 |
| HAWAII | 4 | 1431603 | 2.0 |
| IDAHO | 4 | 1654930 | 2.0 |
| KANSAS | 6 | 2911641 | 2.0 |
| MISSISSIPPI | 6 | 2992333 | 2.0 |
| MONTANA | 3 | 1032949 | 2.0 |
| NEBRASKA | 5 | 1896190 | 2.0 |
| NEVADA | 6 | 2890845 | 2.0 |
| NEW MEXICO | 5 | 2085109 | 2.0 |
| UTAH | 6 | 2995919 | 2.0 |
| WEST VIRGINIA | 5 | 1844128 | 2.0 |