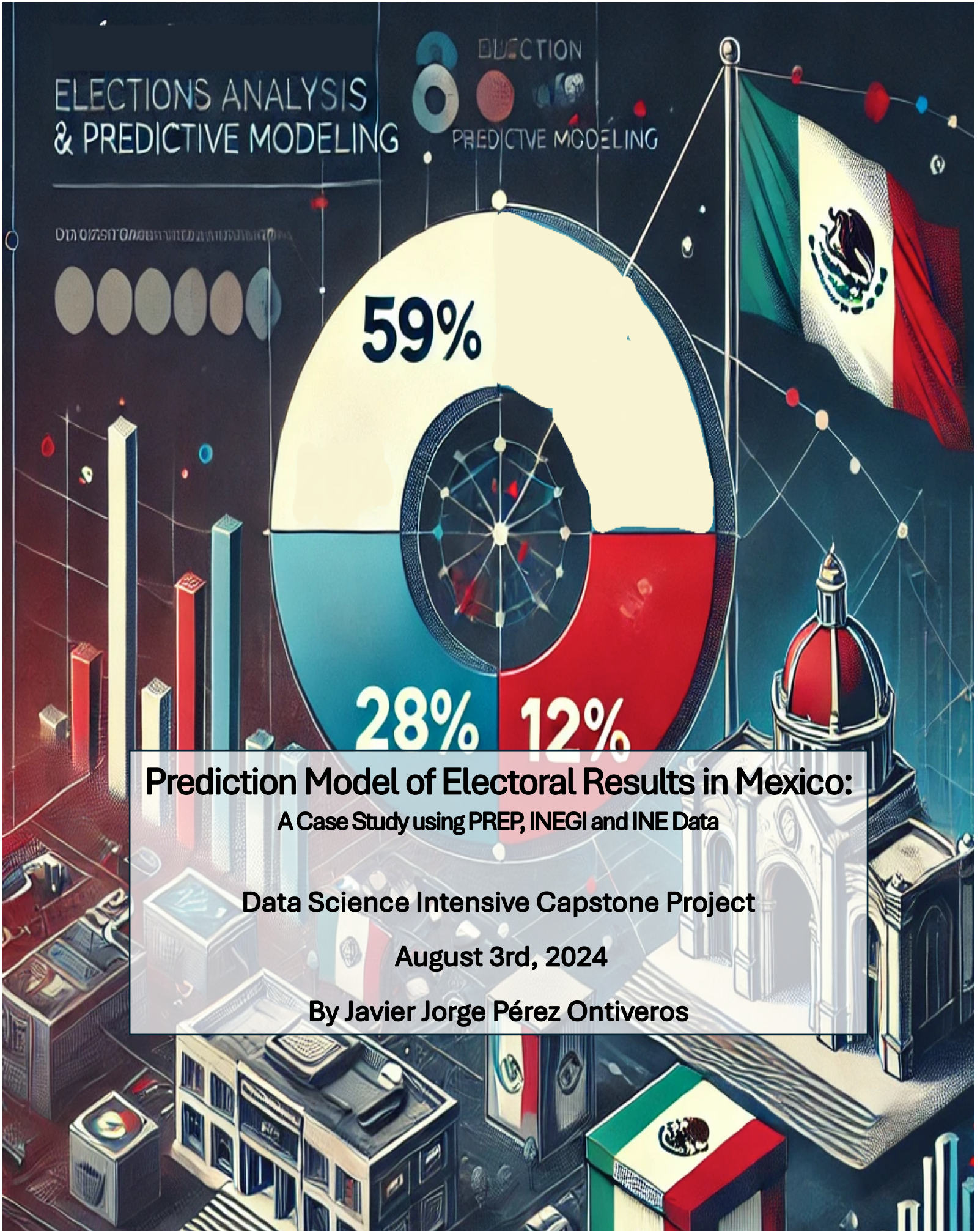



# Prediction Model of Electoral Results in Mexico:

## A Case Study using PREP, INEGI and INE Data

**Data Science Intensive Capstone Project**

**August 3rd, 2024**

**By Javier Jorge Pérez Ontiveros**

**Abstract**

This project uses the PREP, INEGI and INE databases, this provides detailed data on previous elections, including voter demographics, polling station information, economic variables, etc.

Analyzing this data helps identify hidden patterns and trends, which can enhance campaign strategies, policymaking, and voter engagement efforts.

The goal is to understand why the winning party (MORENA) achieved a 59% acceptance rate, while the opposition (PAN-PRI-PRD) secured only 28% of the votes.

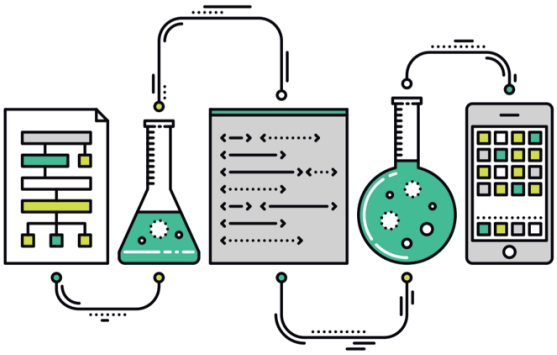
Objectives:

1) To generate a Model that predicts votes for MORENA

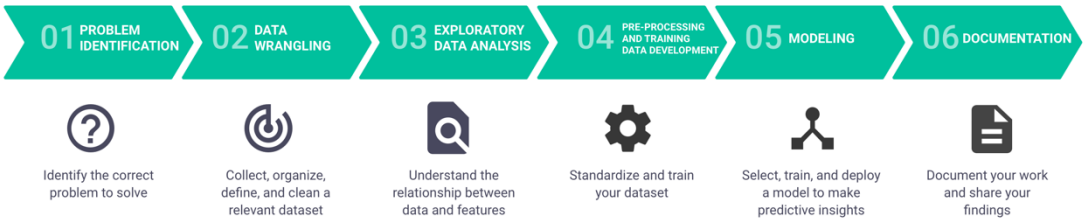
2) To provide actionable insights to support the Opposition Party in developing a strategy to improve their results in the 2030 elections.

Results:

- Geographic heatmap and interesting insights
- 2 Main Voters Profiles were discovered with unsupervised learning (k-Means Clustering).
- A final Machine Learning Model with 98% accuracy was developed.



**Methodology followed**




# • Problem definition

Elections are a critical component of democratic societies, and understanding the factors that influence voting behavior is essential for all stakeholders involved. The PREP (INE) database provides comprehensive data on past elections, including voter demographics, polling station information, and real-time results. Analyzing this data can uncover patterns and trends that are not immediately apparent, helping to improve campaign strategies, policymaking, and voter engagement efforts.

**Criteria for Success**

- Identify key factors that influence voter turnout, while generating valuable insights to optimize Opposition's campaign strategies.
- Develop an accurate predictive model with an R2 score greater than 70%.
- Deliver findings and actionable recommendations in a detailed final report and presentation.

**Scope of Solution Space**

- Data Analysis and Clustering: Use clustering techniques, such as K-Means, to segment voter groups based on demographics, voting history, and other relevant factors during the exploratory data analysis (EDA) phase.
- Data Quality and Validation: Ensure rigorous data cleansing and validation processes to maintain the accuracy and reliability of the models.
- Predictive Modeling: Develop a state-of-the-art predictive model to accurately forecast election results and voter turnout.

- Reporting and Visualization: Summarize key findings and recommendations in a final report and slide deck, including visualizations that highlight important trends and insights.

## Data Sources

- PREP Database: The primary data source, providing detailed election data, including voter demographics, polling station information, and real-time results.
- Census Data: Data from INEGI (Instituto Nacional de Estadística y Geografía) can provide detailed demographic information such as age, gender, education level, income, and employment status. This data can help identify demographic trends and correlations with voting patterns.
- INE Database: Data from INE (Instituto Nacional Electoral) to show the budget spent per political party, the public events and marketing strategies.

**Acquiring, Cleaning and Merging**

**PREP**
2024 and 2018 election results exported from database to CSV file.

55,976,881 votes
170,944 polling stations
1,580 municipalities
32 states

**INEGI**
Socioeconomic inputs for the 32 states in csv:

Hospitals, Schooling Years, vehicles purchased, catholic believers, murders per year, population ages, Poberty ratios, Average Income, etc.

**INE**
Performance and Behavior of Parties per state in csv:

# of Campaign events, Marketing Budget, Federal Welfare,

## Final Dataframe:

A 40 columns x 32 rows Data frame was created to predict the votes for MORENA Entity.

| ENTIDAD | Morena_Votes_2024 | Urban_Pollings_Stations | Non_Urban_Polling_Stations | AMLO_2018 | GDP_19 88 | GDP_19 94 | GDP_20 00 | GDP_20 06 | GDP_20 12 | GDP_20 18 | GDP_20 24 | Federal_Social_Welfare_Budget | Welfare_Recipients | MORENA_Events | Xochitl_Events | Mayn ez_Events | Avg_Scholoing_Years | Avg_Income_Yr | Avg_Expense_Yr | Extreme Poberty_habitants | Poberty_Habitants | Private_Hospital_Users | Public_Hospitals | Total_Vehicles | Catholic_Believers | Murders_Yr | Median_Age | Children | Teenagers | Adults | Elders | Total_Population |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AGUASCALIENTES | 270389 | 1279 | 563 | 190820 | 86482 | 121684 | 158300 | 200318 | 242758 | 327926 | 316500 | 495079 | 51394 | 3 | 3 | 4 | 10 | 313152 | 181260 | 26 | 326 | 34968 | 1153954 | 721372 | 1159832 | 88 | 27 | 256986 | 261684 | 761561 | 145376 | 1425607 |
| BAJA CALIFORNIA | 862661 | 4671 | 703 | 675810 | 417816 | 548270 | 790647 | 772687 | 726418 | 878817 | 929459 | 604660 | 61194 | 5 | 1 | 1 | 10 | 355648 | 201252 | 50 | 461 | 176457 | 2783913 | 2208801 | 2187369 | 2925 | 30 | 574174 | 627987 | 2187557 | 379302 | 3769020 |
| BAJA CALIFORNIA SUR | 175756 | 827 | 287 | 136806 | 62347 | 75145 | 95187 | 126685 | 147270 | 184618 | 177421 | 472056 | 49688 | 2 | 4 | 3 | 10 | 365668 | 192796 | 6 | 106 | 14851 | 666840 | 587090 | 544008 | 90 | 29 | 133230 | 133134 | 456475 | 75608 | 798447 |
| CAMPECHE | 240693 | 734 | 496 | 171328 | 588026 | 713799 | 826278 | 959641 | 691262 | 564591 | 473101 | 536081 | 55722 | 5 | 4 | 7 | 10 | 229832 | 143104 | 92 | 337 | 10355 | 721266 | 372668 | 515526 | 110 | 29 | 165244 | 156308 | 504195 | 102616 | 928363 |
| CHIAPAS | 2E+06 | 2519 | 4354 | 1E+06 | 230322 | 281577 | 309797 | 321114 | 370237 | 353600 | 368789 | 826240 | 83773 | 8 | 4 | 2 | 8 | 159380 | 103772 | 1608 | 2231 | 64941 | 3677747 | 6368520 | 2704411 | 503 | 24 | 1053437 | 1259351 | 5405537 | 1491619 | 9209944 |
| CHIHUAHUA | 744823 | 4601 | 1160 | 441965 | 326595 | 417237 | 581264 | 635598 | 711481 | 853323 | 919617 | 386882 | 38358 | 6 | 11 | 2 | 10 | 327716 | 160384 | 80 | 589 | 247512 | 2975346 | 1127781 | 2514110 | 2157 | 29 | 562295 | 537428 | 1698511 | 348537 | 3146771 |
| CIUDAD DE MEXICO | 3E+06 | 13166 | 265 | 3E+06 | 1897079 | 2448263 | 2903709 | 3134225 | 3332209 | 3694575 | 3840388 | 2551158 | 276864 | 398 | 78 | 29 | 11 | 357240 | 235592 | 159 | 2071 | 537244 | 6352039 | 397307 | 6634532 | 747 | 35 | 114318 | 121464 | 4040152 | 91557 | 731391 |
| COAHUILA | 813432 | 3307 | 836 | 515518 | 387781 | 479740 | 655491 | 798549 | 870635 | 986056 | 922287 | 1235326 | 132323 | 7 | 4 | 1 | 10 | 300504 | 176020 | 59 | 538 | 113714 | 2487607 | 1138170 | 2157764 | 170 | 29 | 1214788 | 1109709 | 2711574 | 507757 | 5543828 |
| COLIMA | 182063 | 667 | 362 | 178123 | 74614 | 90021 | 110551 | 114844 | 135600 | 153808 | 148130 | 9277541 | 924361 | 2 | 3 | 2 | 10 | 277284 | 172468 | 9 | 149 | 12360 | 606110 | 1931820 | 567548 | 887 | 30 | 616247 | 660048 | 2045753 | 419821 | 3741869 |
| DURANGO | 402566 | 1530 | 1088 | 247076 | 149136 | 181317 | 208654 | 241373 | 286058 | 308981 | 290963 | 1509159 | 158018 | 4 | 3 | 2 | 10 | 228880 | 142532 | 118 | 523 | 22732 | 1366822 | 694906 | 1383653 | 127 | 27 | 346808 | 333261 | 943103 | 209478 | 1832650 |
| GUANAJUATO | 1E+06 | 5044 | 3100 | 608766 | 410583 | 515054 | 653601 | 769496 | 892379 | 1120603 | 1126686 | 3228960 | 336389 | 9 | 4 | 3 | 9 | 240400 | 145836 | 203 | 1870 | 119487 | 4805125 | 2271471 | 5107664 | 4329 | 28 | 1098440 | 1101393 | 3285727 | 681374 | 6166934 |
| GUERRERO | 1E+06 | 2304 | 2800 | 844065 | 210191 | 244535 | 256194 | 286083 | 294551 | 317845 | 306831 | 3142321 | 340995 | 5 | 7 | 1 | 8 | 167016 | 122094 | 801 | 1373 | 23333 | 2632011 | 1447351 | 2576502 | 1404 | 27 | 700885 | 675033 | 1721233 | 443534 | 3540685 |
| HIDALGO | 1E+06 | 1518 | 2692 | 744219 | 235180 | 297707 | 337331 | 349080 | 379811 | 418630 | 426819 | 2108249 | 214434 | 5 | 4 | 1 | 9 | 212936 | 136344 | 214 | 1080 | 46182 | 2130675 | 695875 | 2285681 | 390 | 30 | 510948 | 556182 | 1632036 | 383675 | 3082841 |
| JALISCO | 2E+06 | 7782 | 3081 | 1E+06 | 843957 | 1031967 | 1259070 | 1340006 | 1488617 | 1754180 | 1783505 | 4558279 | 467720 | 13 | 9 | 14 | 10 | 286976 | 178776 | 181 | 1676 | 305198 | 5635979 | 4369650 | 6843249 | 1863 | 29 | 1440134 | 1431165 | 4477767 | 999085 | 8348151 |
| MEXICO | 5E+06 | 16527 | 4452 | 4E+06 | 1005116 | 1239028 | 1560109 | 1710986 | 1911402 | 2243798 | 2184863 | 18896673 | 1934530 | 20 | 17 | 6 | 10 | 228932 | 154596 | 1032 | 6395 | 565431 | 10827568 | 9421189 | 12369271 | 3257 | 30 | 2665745 | 2941214 | 9466005 | 1919454 | 16892418 |
| MICHOACAN | 1E+06 | 3494 | 2936 | 620449 | 304548 | 364985 | 451793 | 485162 | 531827 | 633757 | 646901 | 2934380 | 306141 | 9 | 5 | 2 | 9 | 227836 | 160544 | 372 | 1691 | 61986 | 2930541 | 1950503 | 3837269 | 2329 | 28 | 874985 | 834612 | 2437550 | 601699 | 4748846 |
| MORELOS | 578230 | 1956 | 621 | 521571 | 144824 | 176585 | 206507 | 232044 | 248100 | 273245 | 256740 | 1681672 | 170369 | 10 | 6 | 3 | 10 | 228956 | 152036 | 118 | 708 | 31301 | 1405667 | 1195466 | 1298610 | 1175 | 30 | 305322 | 330831 | 1061464 | 273903 | 1971520 |
| NAYARIT | 306423 | 1110 | 706 | 240273 | 87659 | 112713 | 115656 | 135897 | 142898 | 155758 | 161869 | 1466501 | 149317 | 3 | 1 | 1 | 10 | 261304 | 155448 | 81 | 289 | 18508 | 967852 | 542623 | 944500 | 196 | 29 | 218711 | 217591 | 641759 | 157395 | 1235456 |
| NUEVO LEON | 1E+06 | 6277 | 1164 | 551927 | 694488 | 916740 | 1212181 | 1469584 | 1621930 | 1909026 | 1945060 | 730338 | 73990 | 15 | 8 | 16 | 11 | 344072 | 193584 | 65 | 907 | 417040 | 4413414 | 2686334 | 4152646 | 1410 | 30 | 936860 | 934616 | 3258916 | 654050 | 5784442 |
| OAXACA | 1E+06 | 2232 | 3687 | 1E+06 | 268785 | 344896 | 355731 | 380271 | 386538 | 383160 | 407238 | 2066353 | 215982 | 7 | 3 | 1 | 8 | 173372 | 106388 | 860 | 1624 | 30777 | 2918560 | 1005139 | 2855785 | 805 | 28 | 758000 | 757070 | 2066501 | 550577 | 4132148 |
| PUEBLA | 2E+06 | 5076 | 3243 | 1E+06 | 344990 | 425378 | 563689 | 650283 | 747712 | 848556 | 820790 | 2460353 | 270154 | 7 | 3 | 4 | 9 | 197536 | 136108 | 766 | 2861 | 131689 | 4572862 | 1263461 | 5057571 | 1089 | 28 | 1183000 | 1210575 | 3444284 | 745419 | 6583278 |
| QUERETARO | 585662 | 1773 | 1365 | 350246 | 189848 | 266658 | 368561 | 427147 | 526761 | 611368 | 592139 | 841132 | 87247 | 4 | 3 | 1 | 10 | 299824 | 194912 | 43 | 494 | 97490 | 1821860 | 820112 | 1861516 | 192 | 29 | 391686 | 396742 | 1339817 | 240222 | 2368467 |
| QUINTANA ROO | 536134 | 1938 | 558 | 334458 | 102517 | 140495 | 180224 | 234333 | 280381 | 377341 | 357539 | 503768 | 49611 | 4 | 2 | 0 | 10 | 287608 | 184580 | 80 | 437 | 41972 | 1341934 | 988280 | 942844 | 647 | 28 | 324141 | 301830 | 1100962 | 131052 | 1857985 |
| SAN LUIS POTOSI | 773086 | 2031 | 1875 | 413328 | 201731 | 264259 | 310465 | 377360 | 437595 | 557796 | 537933 | 1876774 | 201188 | 7 | 1 | 2 | 10 | 240576 | 154280 | 213 | 807 | 75626 | 2300947 | 1357909 | 2218856 | 759 | 29 | 475600 | 506577 | 1478479 | 361599 | 2822255 |
| SINALOA | 782221 | 3538 | 1621 | 632646 | 283923 | 344144 | 391567 | 411251 | 466782 | 527714 | 516711 | 1174206 | 121748 | 5 | 2 | 2 | 10 | 287552 | 171644 | 56 | 612 | 60854 | 2448164 | 1341447 | 2196411 | 587 | 30 | 488957 | 521753 | 1622714 | 393519 | 3026943 |
| SONORA | 674746 | 2983 | 916 | 462120 | 329333 | 446593 | 591564 | 605739 | 681379 | 782098 | 804668 | 756531 | 77435 | 6 | 6 | 2 | 10 | 301080 | 169672 | 51 | 599 | 100270 | 2352956 | 1552856 | 2094915 | 1759 | 30 | 467781 | 514507 | 1604148 | 358404 | 2946840 |
| TABASCO | 811270 | 1531 | 1580 | 766515 | 306786 | 375022 | 425907 | 478125 | 595176 | 471565 | 635465 | 935361 | 98415 | 5 | 2 | 3 | 10 | 208380 | 128100 | 273 | 863 | 87994 | 1576238 | 737187 | 1372585 | 354 | 29 | 431811 | 431358 | 1281904 | 257525 | 2402598 |
| TAMAULIPAS | 910332 | 3795 | 1157 | 607278 | 336905 | 419240 | 565665 | 657147 | 676616 | 732094 | 746946 | 946850 | 89313 | 7 | 8 | 1 | 10 | 254084 | 158012 | 103 | 860 | 109007 | 2740917 | 1434974 | 2223753 | 482 | 30 | 571338 | 600803 | 1929367 | 426227 | 3527735 |
| TLAXCALA | 486946 | 1283 | 438 | 358631 | 69999 | 87365 | 112212 | 117236 | 128975 | 147753 | 143717 | 1222596 | 127898 | 1 | 2 | 1 | 10 | 185180 | 131812 | 93 | 627 | 24102 | 950161 | 585821 | 1045600 | 155 | 28 | 249123 | 240925 | 707043 | 145886 | 1342977 |
| VERACRUZ | 2E+06 | 5796 | 5189 | 2E+06 | 676136 | 837586 | 896917 | 992039 | 1084723 | 1029905 | 1032889 | 2540337 | 253349 | 17 | 9 | 2 | 9 | 178548 | 117224 | 1078 | 3166 | 100470 | 5804663 | 2319548 | 5615966 | 787 | 31 | 1278905 | 1367016 | 4258766 | 1157892 | 8062579 |
| YUCATAN | 696450 | 2195 | 769 | 330914 | 145505 | 187127 | 232449 | 270590 | 299649 | 364798 | 369864 | 819399 | 89747 | 15 | 5 | 1 | 10 | 249484 | 152308 | 133 | 787 | 69445 | 1775984 | 994656 | 1597707 | 55 | 30 | 365924 | 385500 | 1280439 | 289035 | 2320898 |
| ZACATECAS | 369235 | 1112 | 1522 | 307197 | 89469 | 112636 | 124587 | 159901 | 201298 | 228388 | 221026 | 1574047 | 169039 | 3 | 1 | 3 | 9 | 200544 | 139452 | 84 | 642 | 11459 | 1300692 | 639134 | 1356905 | 1459 | 28 | 309307 | 292563 | 820489 | 199779 | 1622138 |

# ● Data Engineering / Wrangling

This step involves gathering the data, organizing it, and ensuring it is well-defined.

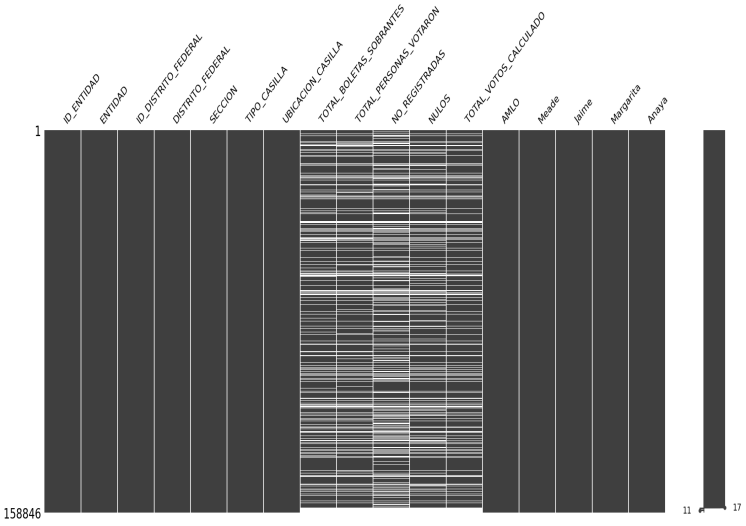
Careful attention to these tasks will yield significant benefits later.

Some initial data cleaning might be performed at this stage, while ensuring that the data frame does not have Null values. Also, during this first stage, we explore the data to gain a better understanding of it.

From Prep Data source up to 49,157 null values were identified and classified as “Ceros”, as basically this values represented the documents of population with null votes.

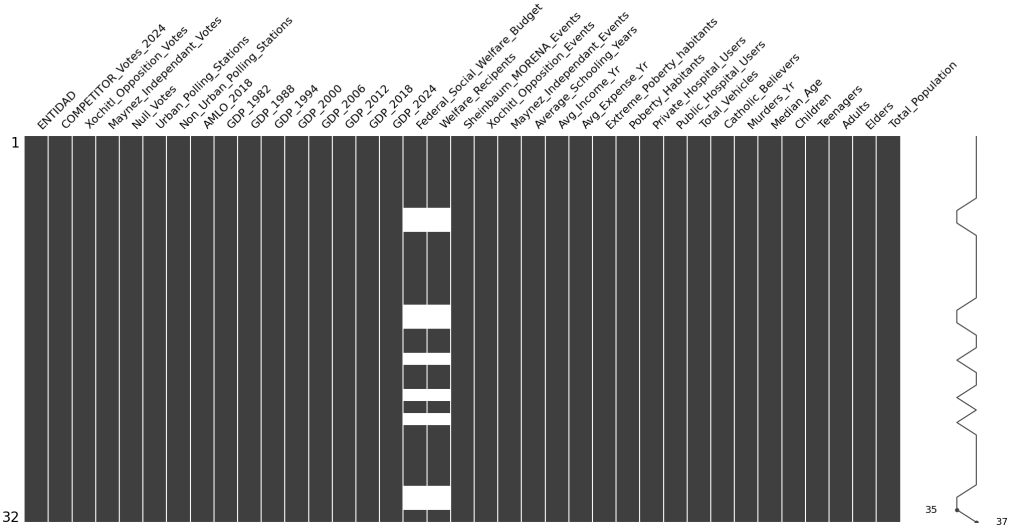
```
1  df.isnull().sum()
```

| | |
|---|---|
| ID_ENTIDAD | 0 |
| ENTIDAD | 0 |
| ID_DISTRITO_FEDERAL | 0 |
| DISTRITO_FEDERAL | 4 |
| SECCION | 0 |
| TIPO_CASILLA | 0 |
| UBICACION_CASILLA | 0 |
| TOTAL_BOLETAS_SOBRANTES | 35751 |
| TOTAL_PERSONAS_VOTARON | 35806 |
| NO_REGISTRADAS | 49157 |
| NULOS | 35316 |
| TOTAL_VOTOS_CALCULADO | 30297 |
| AMLO | 0 |
| Meade | 0 |
| Jaime | 0 |
| Margarita | 0 |
| Anaya | 0 |
| dtype: int64 | |




Once that a single data frame was merged from the 3 data sources, it was identified that Federal Social Welfare and Welfare Recipients Features had null values and in the exact same rows.

I realized that the null values of the merged data frame, were caused due to incompatibility of formats in the foreign keys of the left join due t o typographic differences (‘apostrophes’).




After all the cleansing and wrangling, the final data frame had 32 rows and 37 columns, all with numeric dtypes, except for the entities.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 37 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   ENTIDAD                               32 non-null     object
 1   COMPETITOR_Votes_2024                 32 non-null     int64
 2   Xochitl_Opposition_Votes              32 non-null     int64
 3   Maynez_Independant_Votes              32 non-null     int64
 4   Null_Votes                            32 non-null     int64
 5   Urban_Polling_Stations                32 non-null     int64
 6   Non_Urban_Polling_Stations            32 non-null     int64
 7   AMLO_2018                             32 non-null     int64
 8   GDP_1982                              32 non-null     int64
 9   GDP_1988                              32 non-null     int64
 10  GDP_1994                              32 non-null     int64
 11  GDP_2000                              32 non-null     int64
 12  GDP_2006                              32 non-null     int64
 13  GDP_2012                              32 non-null     int64
 14  GDP_2018                              32 non-null     int64
 15  GDP_2024                              32 non-null     int64
 16  Sheinbaum_MORENA_Events               32 non-null     int64
 17  Xochitl_Opposition_Events             32 non-null     int64
 18  Maynez_Independant_Events             32 non-null     int64
 19  Average_Schooling_Years               32 non-null     int64
 20  Avg_Income_Yr                         32 non-null     int64
 21  Avg_Expense_Yr                        32 non-null     int64
 22  Extreme_Poberty_habitants             32 non-null     int64
 23  Poberty_Habitants                     32 non-null     int64
 24  Private_Hospital_Users                32 non-null     int64
 25  Public_Hospital_Users                 32 non-null     int64
 26  Total_Vehicles                        32 non-null     int64
 27  Catholic_Believers                    32 non-null     int64
 28  Murders_Yr                            32 non-null     int64
 29  Median_Age                            32 non-null     int64
 30  Children                              32 non-null     int64
 31  Teenagers                             32 non-null     int64
 32  Adults                                32 non-null     int64
 33  Elders                                32 non-null     int64
 34  Total_Population                      32 non-null     int64
 35  Federal_Social_Welfare_Budget_2024    32 non-null     int64
 36  Welfare_Recipents                     32 non-null     int64
dtypes: int64(36), object(1)
```

# ● Exploratory Analysis (EDA) and Visualization

The Exploratory Data Analysis (EDA) phase is a crucial step in our Data Science project. In this phase, we delve into the dataset to uncover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

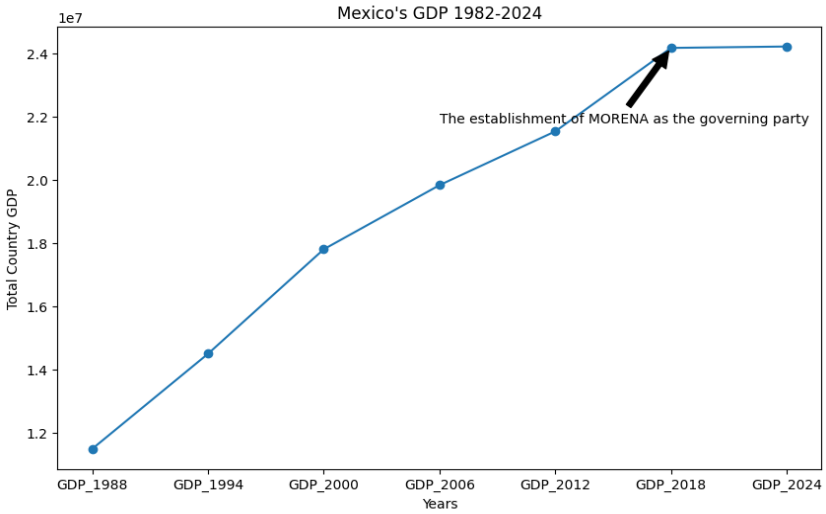
The primary goal of EDA is to gain a deeper understanding of the data, which will guide us in making informed decisions during the modeling phase. This involves identifying key features that influence voter decisions and understanding the underlying structure of the data. By doing so, we can develop strategies for the opposition (PAN-PRI-PRD) to enhance their chances in upcoming election cycles.

By thoroughly exploring the data, we aim to build a robust predictive model that can provide actionable insights for the opposition, enabling them to develop effective campaign strategies and allocate resources efficiently for future elections.

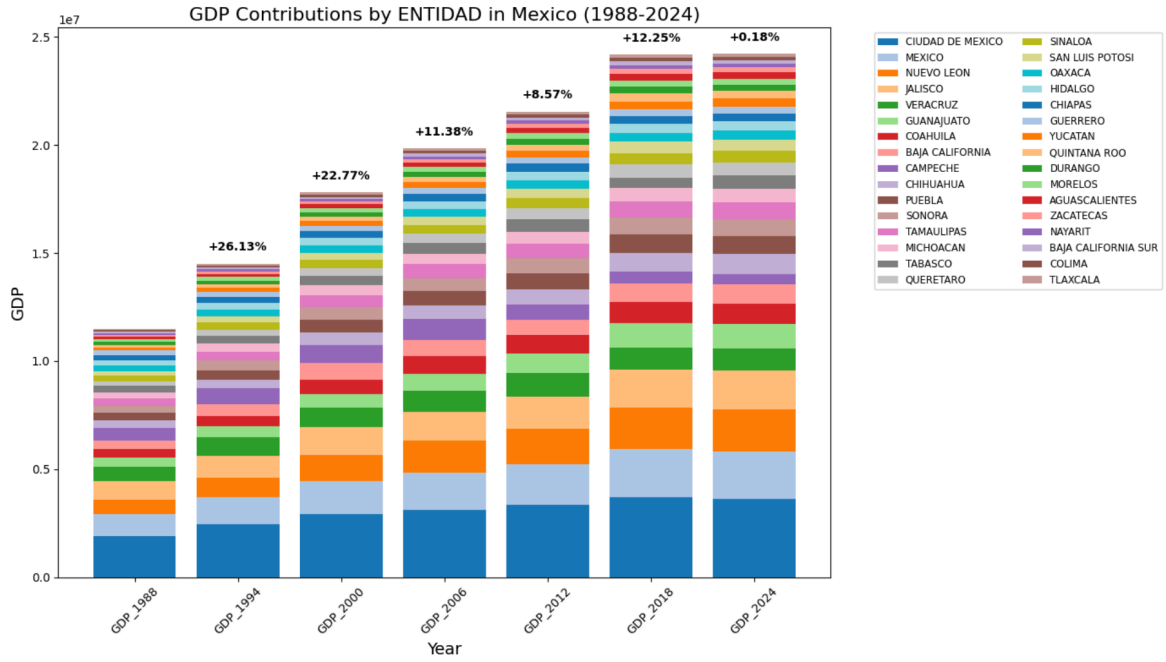
During EDA, I uncovered 4 interesting insights:

### 1) Economic Slowdown

The data shows an Economic Slowdown During the Period When MORENA Assumed Power



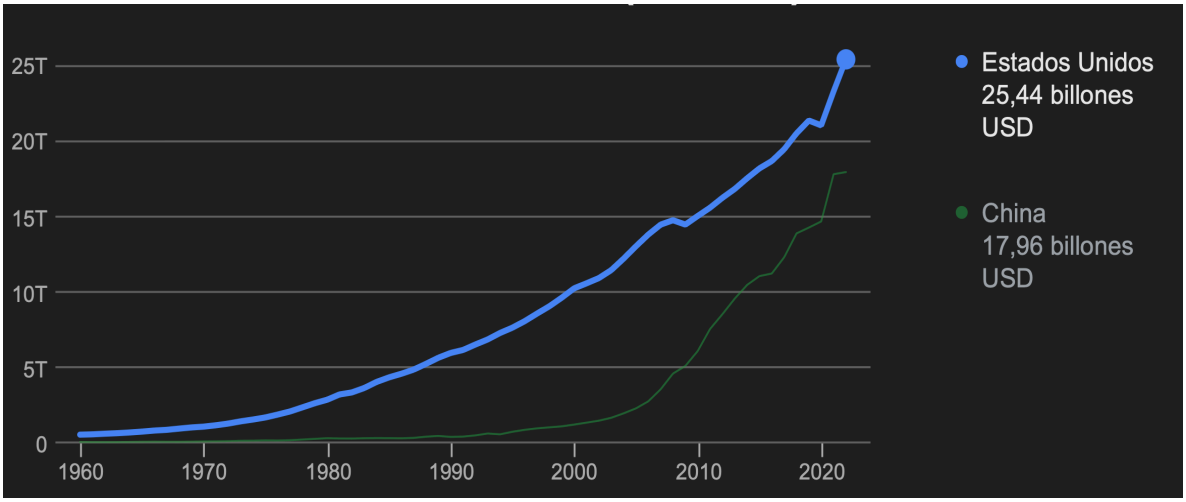
Same data can be split by Entity as follows:




As it can be observed, GDP Growtth from 2018 to 2024 was only 0.18%, while the Opposition parties had a 10%-20% growth  from 1988 to 2018.

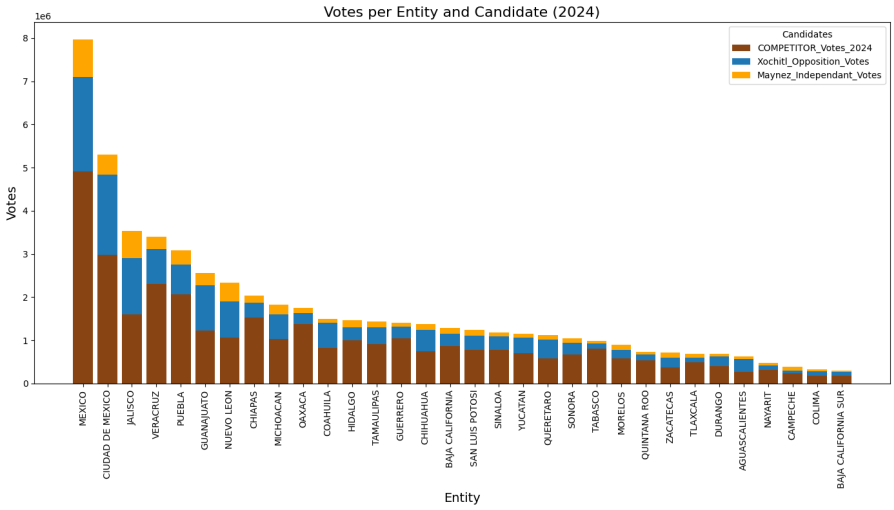
When I shared this insight with some polítical Experts, the first question that poped to their mind was: “Was this effect due to Covid 19?”.

For that reason, I compared the performance of the top 2 economic powers to review if those countries had a similar “slow down” or “stagnation”. But I observed that they didn’t have a similar effect from 2018-2024, so something else is causing the slowdown in México.



## 2) **Effect of Independent Party “Movimiento Naranja”.**

Contrary to what many expected, that the Independent Candidate Maynez would weaken the opposition, this effect can only be observed in states where the “Movimiento Naranja “party currently holds power, such as Nuevo León and Jalisco.”



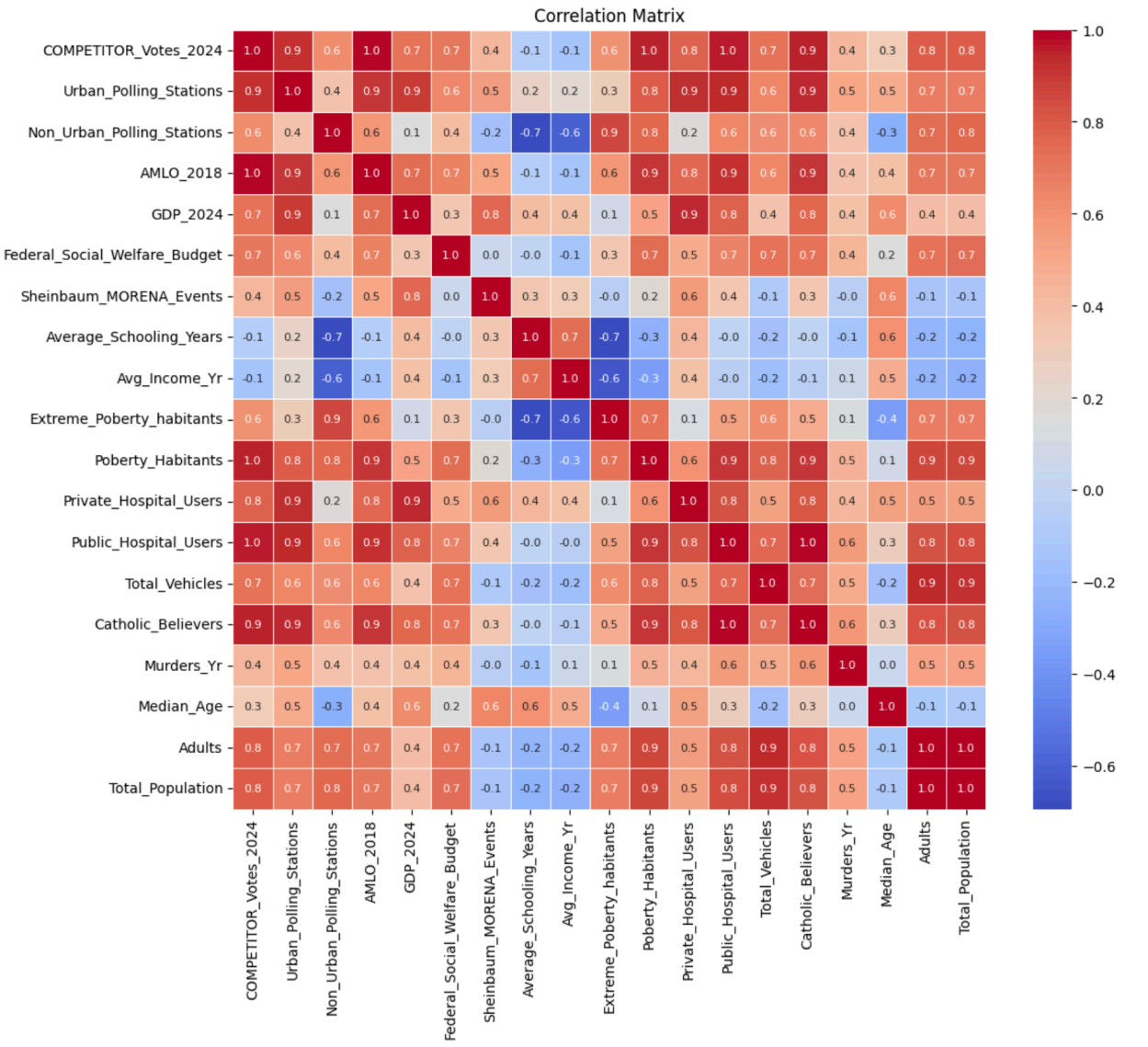

## 3) **Geographical Heatmap:**

The strongest resistance to MORENA is observed in the Bajío region, known for its industrial and economic strength, while the strongest support for MORENA is in the less industrially and economically developed Southeast region. To have a chance of winning from 2024-2030, the opposition must focus on helping the less privileged population.

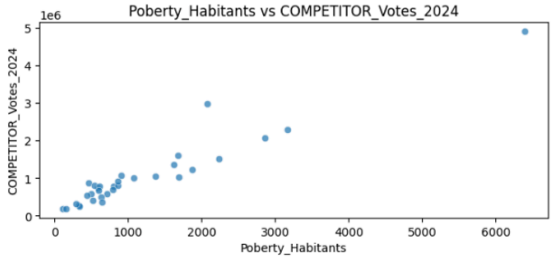
## 4) Features Correlated to Votes:

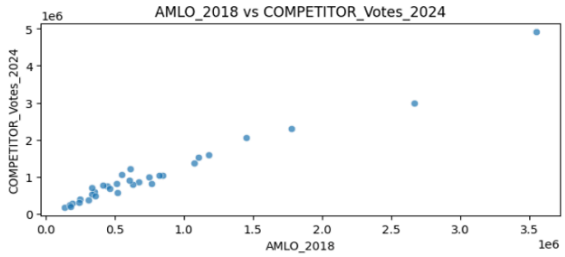
The following variables showed a positive correlation with MORENA votes:

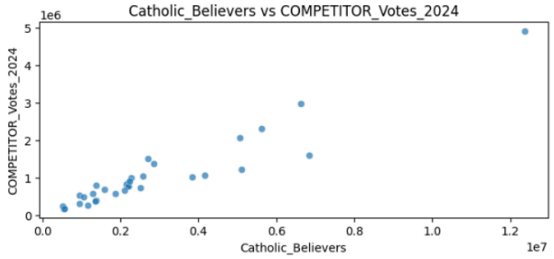
- Preference for AMLO in 2018
- Federal Warefare budget
- Poberty Habitants
- Public Hospitals
- Catholic Believers
- Adults Amount

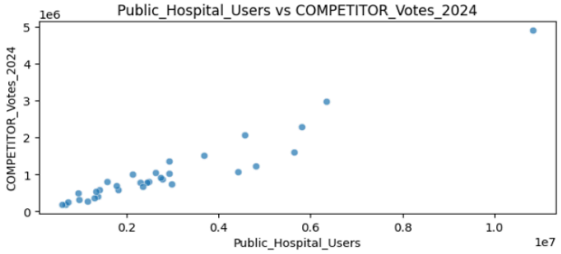


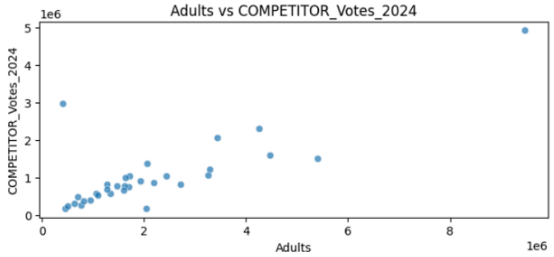
Their Linear correlations show that the votes for MORENA increase when these variables increase:

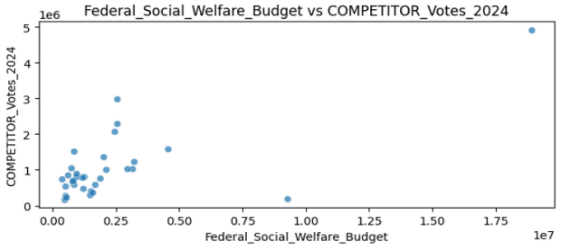


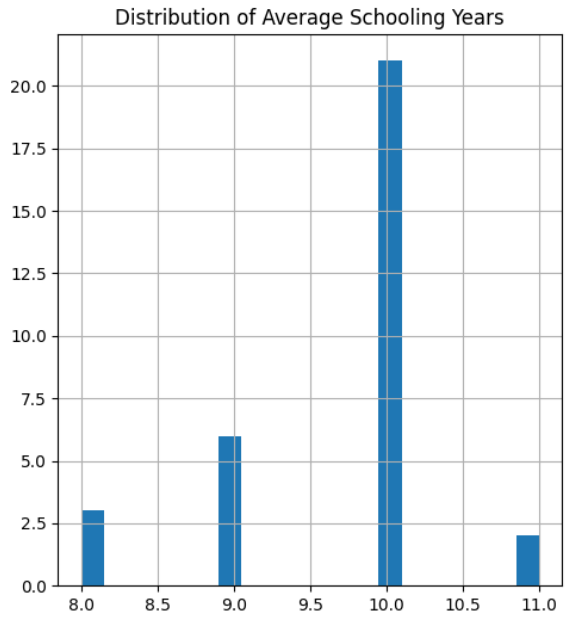



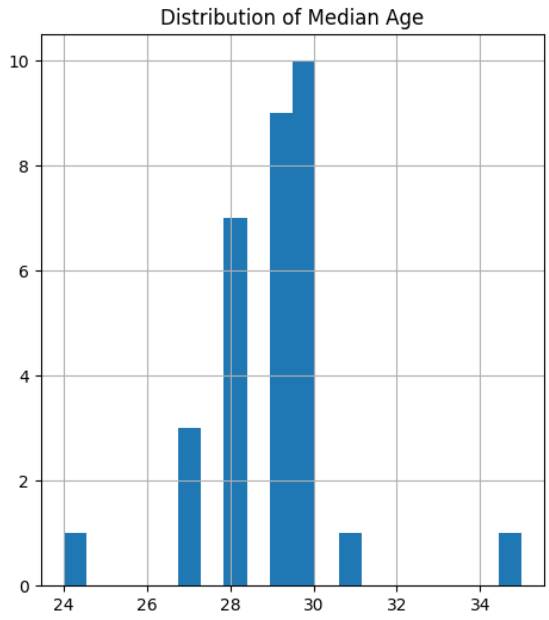
# ● Feature Engineering and Clusterization

2 features have a behavior of categorical variables (despite they are numerical):

* Average_schooling days

* Median Age




These 2 variables were encoded using “one-hot encoding” technique:

| Schooling_Bins_Low | Schooling_Bins_Medium | Schooling_Bins_High | Schooling_Bins_Very High | Age_Bins_Young Adult | Age_Bins_Adult | Age_Bins_Middle Age | Age_Bins_Senior | Age_Bins_Elder |
|---|---|---|---|---|---|---|---|---|
| False | True | False | False | True | False | False | False | False |
| False | True | False | False | False | False | True | False | False |
| False | True | False | False | False | True | False | False | False |
| False | True | False | False | False | True | False | False | False |
| False | False | False | False | False | False | False | False | False |
| False | True | False | False | False | True | False | False | False |
| False | False | True | False | False | False | False | False | True |
| False | True | False | False | False | True | False | False | False |
| False | True | False | False | False | False | True | False | False |
| False | True | False | False | True | False | False | False | False |

**Ratios:**

Some Features in my dataframe might seem to have very high levels or values de pending on the population of citizens living in each Entity. In

order to get a more realistic feature, some ratios where calculated from those type of features that depend on the size of the population.:

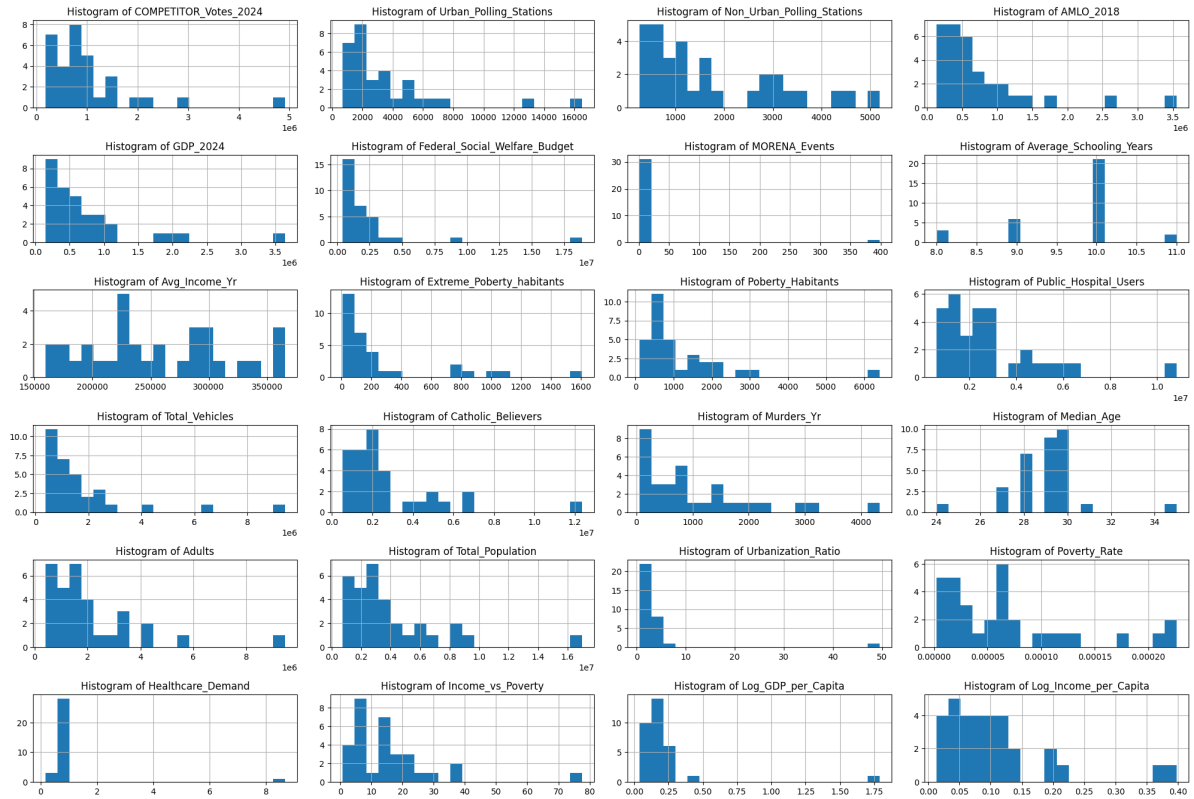
```
# Create new features with Ratios
df_encoded['Urbanization_Ratio'] = df_encoded['Urban_Polling_Stations'] / (df_encoded['Non_Urban_Polling_Stations'] + 1)
df_encoded['Poverty_Rate'] = df_encoded['Extreme_Poberty_habitants'] / df_encoded['Total_Population']
df_encoded['Healthcare_Demand'] = df_encoded['Public_Hospital_Users'] / df_encoded['Total_Population']
df_encoded['Income_vs_Poverty'] = df_encoded['Avg_Income_Yr'] * df_encoded['Poverty_Rate']
df_encoded['Vehicles_per_Capita'] = df_encoded['Total_Vehicles'] / df_encoded['Total_Population']
```

+ Código + Markdown

```
# Other proportions vs Total Population
df_encoded['GDP_per_Capita'] = df_encoded['GDP_2024'] / df_encoded['Total_Population']
df_encoded['Income_per_Capita'] = df_encoded['Avg_Income_Yr'] / df_encoded['Total_Population']
df_encoded['Vehicles_per_Capita'] = df_encoded['Total_Vehicles'] / df_encoded['Total_Population']
```

## Standardization

As the data still seemed to be very skewed, normalization techniques were used in this study:



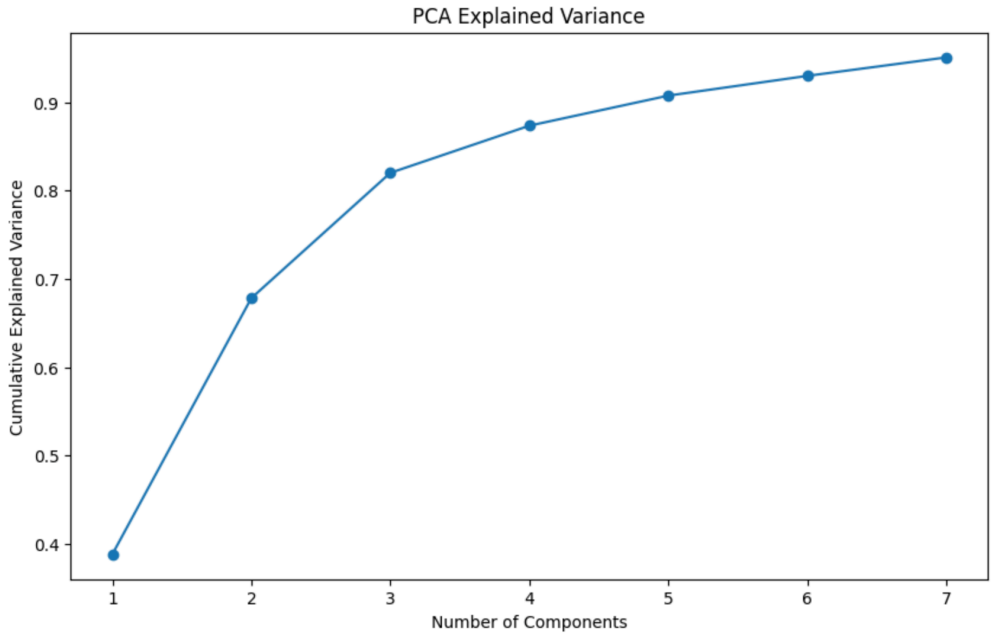

| Median_Age | Adults | Total_Population | Urbanization_Ratio | Poverty_Rate | Healthcare_Demand | Income_vs_Poverty | Log_GDP_per_Capita | Log_Income_per_Capita |
|---|---|---|---|---|---|---|---|---|
| -1.16692 | -0.760959 | -0.778683 | -0.179094 | -0.818624 | -0.107315 | -0.684212 | -0.054838 | 1.050425 |
| 0.58346 | 0.034264 | -0.052355 | 0.344803 | -0.898666 | -0.158091 | -0.752846 | 0.013539 | -0.172075 |
| 0.00000 | -0.931093 | -0.973068 | -0.106661 | -0.991260 | -0.088872 | -0.888999 | -0.054285 | 3.063130 |
| 0.00000 | -0.904481 | -0.932801 | -0.273967 | 0.483175 | -0.130636 | 0.495079 | 0.670421 | 1.305532 |
| -2.91730 | 1.828805 | 1.634029 | -0.381746 | 1.698576 | -0.401373 | 0.844107 | -0.608122 | -0.995469 |

**Principal Components Analysis (PCA)**

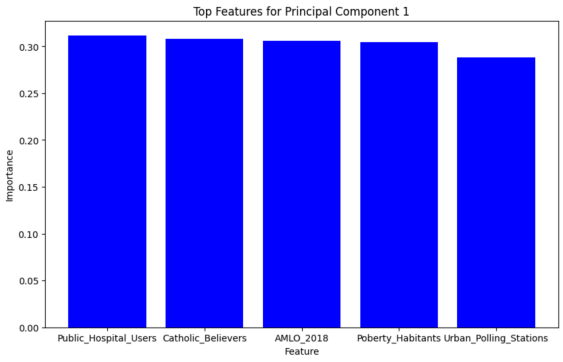
I conducted PCA to identify the most influential features in our dataset. The top features for the first principal component (PC1) included Public_Hospital_Users, Catholic_Believers, AMLO_2018, Poberty_Habitants, and Urban_Polling_Stations.

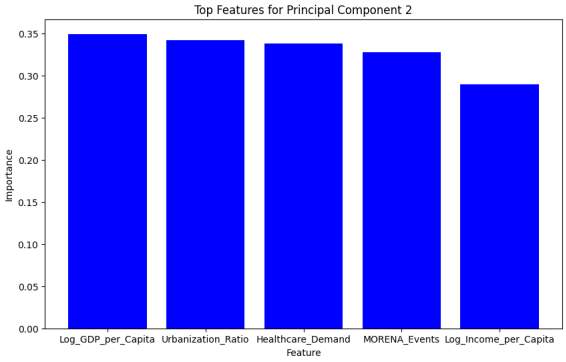
For the second principal component (PC2), the top features were Log_GDP_per_Capita, Urbanization_Ratio, Healthcare_Demand, Sheinbaum_MORENA_Events, and Log_Income_per_Capita.

PCA helped me to reduce dimensionality while retaining the most significant variance in the data.




- 2 Componenets explain 67% of the Total Variation
- 3 Components explain 87% of Total Variation



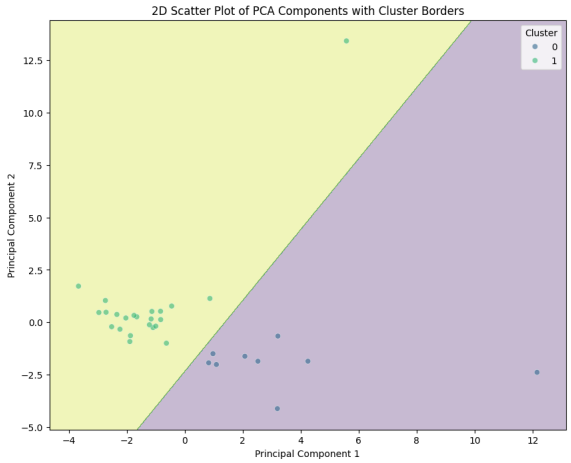



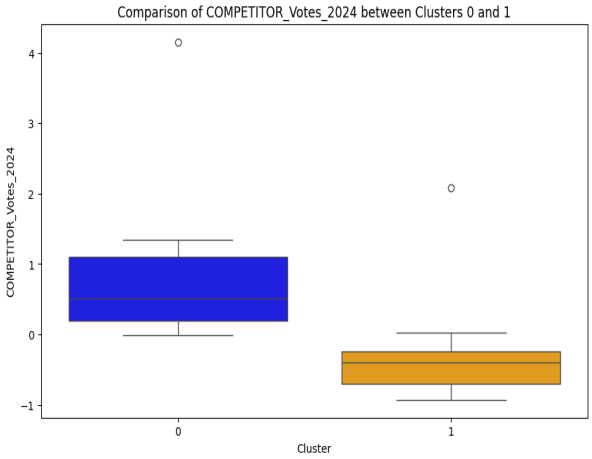
## Clustering by using K-Means

I applied K-means clustering on the PCA-transformed datato identify potential clusters within our dataset. 2 clusters were visualized, revealing distinct groups that indicate an underlying pattern in the data.

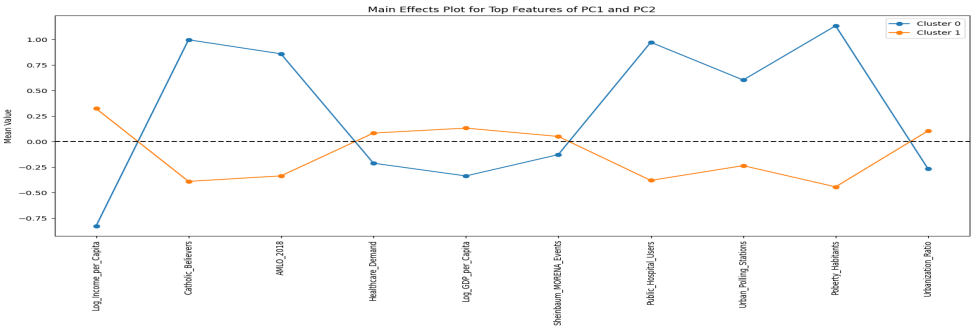
Orange Cluster #1 have these characteristics in general:

- LESS VOTES FOR MORENA
- Higher Income per Capita.
- Less chance of being catholic.
- Voted less for AMLO in 2018.
- Preference for Private Hospitals.
- Live in cities that have a lower Ratio of Poor Habitants




## Detailed behavior of the features of both clusters:




The graph shows the normalized Average of different characteristics of both clusters.

- **Cluster #0 (BLUE)** has a **higher** preference for **MORENA**
- **Cluster #1 (ORANGE)** has a **lower** voting preference for **MORENA**.

# ● Predictive Modeling

The modeling phase is one of the most crucial stages in a data science project. This phase involves selecting and applying appropriate modeling techniques to our data to build a model that can make accurate predictions or provide valuable insights.

Choosing the right model is fundamental. Different problems require different types of models. For example:

- **Classification Models:** When the goal is to categorize data (e.g., spam/not spam).
- **Regression Models:** When the goal is to predict a continuous value (e.g., house price).
- **Clustering Models:** To find natural groups within the data (e.g., customer segmentation).

In this specific project the purpose is to predict continuous values and the objective is to test at least 3 different models to measure their performance.

Once the model is selected and trained, it's essential to evaluate its performance using appropriate metrics. This helps to understand whether the model is accurate enough and if it will generalize well to unseen data. Common metrics include accuracy, recall, F1-score for classification, and RMSE, MAE for regression.

In this study as our output variable is continuous, so we used $R^2$ and MSE.

As a result of this phase, 5 **Models where developed:**

- **Decision Tree**
- **Random Forest**
- **Linear Regression**
- **Ridge**
- **Lasso.**

## The Lasso model had the best performance (98% accuracy in testing and training)

```
# Initialize models
models = {
    'Random Forest': RandomForestRegressor(random_state=42),
    'Lasso': Lasso(alpha=0.1),
    'Linear Regression': LinearRegression(),
    'Ridge': Ridge(alpha=1.0),
    'Decision Tree': DecisionTreeRegressor(random_state=42)
}

# Train and evaluate models
conclusions = []

for name, model in models.items():
    model.fit(X_train_scaled, y_train)
    y_train_pred = model.predict(X_train_scaled)
    y_test_pred = model.predict(X_test_scaled)

    mse_train = mean_squared_error(y_train, y_train_pred)
    r2_train = r2_score(y_train, y_train_pred)
    mse_test = mean_squared_error(y_test, y_test_pred)
    r2_test = r2_score(y_test, y_test_pred)

    conclusions.append({
        'Model': name,
        'Train MSE': mse_train,
        'Train R²': r2_train,
        'Test MSE': mse_test,
        'Test R²': r2_test
    })

# Convert conclusions to DataFrame
conclusions_df = pd.DataFrame(results)

# Print the results:

# Sort the results by Test R² in descending order
conclusions_df = conclusions_df.sort_values(by='Test R²', ascending=False)
```

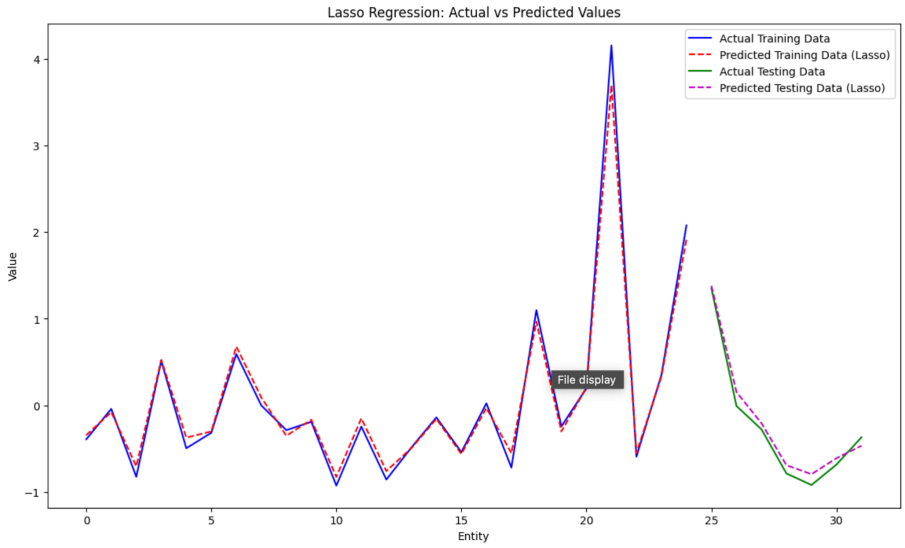
Metrics of the Model:

| | |
|---|---|
| Model | Lasso |
| Train MSE | 0.014998 |
| Train R² | 0.986565 |
| Test MSE | 0.010197 |
| Test R² | 0.97997 |
| Model | Ridge |
| Train MSE | 0.002201 |
| Train R² | 0.998029 |
| Test MSE | 0.047715 |
| Test R² | 0.906278 |
| Model | Linear Regression |
| Train MSE | 0.0 |
| Train R² | 1.0 |
| Test MSE | 0.054846 |
| Test R² | 0.892273 |
| Model | Random Forest |
| Train MSE | 0.100472 |
| Train R² | 0.910003 |
| Test MSE | 0.067189 |
| Test R² | 0.868028 |
| Model | Decision Tree |
| Train MSE | 0.0 |
| Train R² | 1.0 |
| Test MSE | 0.127489 |
| Test R² | 0.749587 |

## Testing Predictions:

```
Lasso Coefficients:
                   Feature  Coefficient
11                AMLO_2018     0.567453
18       Poberty_Habitants      0.202205
19   Public_Hospital_Users      0.209553
```

```
Lasso Equation:
Y =  + 0.57 * AMLO_2018 + 0.20 * Poberty_Habitants + 0.21 * Public_Hospital_Users
```




Lasso Equation ML Model developed had a great performance that can be graphically seen in the above equation.

I suggest to choose the Lasso Model due to 3 main points:

1) The simplicity of its equation:

**Lasso Equation: Y = + 0.57 * AMLO_2018 + 0.20 * Poberty_Habitants + 0.21 * Public_Hospital_Users**

2) The great acuracy it has in training and Testing Datasets (98% $R^2$)

3) With a low amount of features I can still predict the votes accurately.

## Model Metrics

An aditinal trial was run in a CSV file with all the dataframe features to confirm the predictive power of the Model, showing similar results, MSE 0.007 and R2 of 99%.

ModelMetricsfile

| AMLO_2018 | Poberty_Habitants | Public_Hospital_Users | COMPETITOR_Votes_2024 | Predicted | Difference |
|---|---|---|---|---|---|
| **-0.7847173190976150** | -0.7082027433568750 | -0.7970364622246030 | -0.8262555701128480 | -0.7563070776241820 | -0.06994849248866610 |
| **-0.1143552063371590** | -0.5959998662049440 | -0.0202714265554901 | -0.1910465079640970 | -0.18863944042982300 | -0.002407067534274330 |
| **-0.8593764649741140** | -0.8910518764933560 | -1.0291730519087100 | -0.9277501104378760 | -0.8841813012347460 | -0.04356880920313 |
| **-0.8116595205902790** | -0.6990602867000510 | -1.0032360720022700 | -0.8581043968505120 | -0.8131375591969460 | -0.044966837653565900 |
| **0.4811805121176670** | 0.8751045231203730 | 0.4056895938703400 | 0.5134736491721350 | 0.5344886112439160 | -0.021014962071780900 |
| **-0.4375800733621880** | -0.4896149160164470 | 0.0709569178833855 | -0.3174272341599740 | -0.3324426722642260 | 0.0150154381042516 |
| **2.642901014090320** | 0.7421233353847510 | 1.6801366939321400 | 2.077905693252200 | 2.0077069508341800 | 0.07019874241802350 |
| **-0.3359137617271650** | -0.5320026696071760 | -0.1614775186467470 | -0.243844389546412 | -0.33178165702173600 | 0.08793726747532420 |
| **-0.8022673464089400** | -0.8553131822894070 | -1.0581142342428600 | -0.9209848087453530 | -0.8505590131019790 | -0.07042579564337440 |
| **-0.7069592395177220** | -0.5444696559573910 | -0.6955931623545990 | -0.6844963307639760 | -0.6579352618110460 | -0.02656106895293010 |
| **-0.2070246582596980** | 0.5750657182918760 | 0.9429471643874960 | 0.1962412740405790 | 0.19502799297172200 | 0.0012132810688574300 |
| **0.1182099543517430** | 0.1619929038881000 | -0.0926610771222894 | -0.0016202055644616 | 0.0803194285624325 | -0.0819396341268941 |
| **-0.0197990262678175** | -0.0815288961527572 | -0.331575231356973 | -0.0416457380890304 | -0.09722202278817170 | 0.05557628469914130 |
| **0.5850652843209090** | 0.4138260281624340 | 1.3388947538335000 | 0.5921765873621260 | 0.6974203160004410 | -0.10524372863831500 |
| **3.862305164225730** | 4.335939933939930 | 3.8129722090114400 | 4.156590755587110 | 3.8694260942890600 | 0.28716466129805000 |
| **0.0855674840844155** | 0.4262930145126490 | 0.0496048731858473 | -0.0076036580284104 | 0.14444909219967500 | -0.15205275022808500 |
| **-0.3275471936153710** | -0.3907101576380780 | -0.6770813851807150 | -0.4960974790652660 | -0.40703102277632700 | -0.08906645628893870 |
| **-0.7163624714464410** | -0.738954643020738 | -0.8857242922056600 | -0.7876092668172880 | -0.7421196386918080 | -0.04548962812548040 |
| **-0.2855885711799470** | -0.2253148053918980 | 0.756275346944859 | 0.0227608349200244 | -0.049030623792529000 | 0.07179145871255340 |
| **0.4408957561919710** | 0.3706071421483570 | 0.0438952682863672 | 0.350638800508142 | 0.33465001579923200 | 0.015988784708910200 |
| **0.9569898530282180** | 1.3987179498293800 | 0.8322610811917750 | 1.0999570951993400 | 1.000002633242230 | 0.09995446195710380 |
| **-0.5643557648615520** | -0.5685724962344720 | -0.4787425493632280 | -0.4881266924167510 | -0.5359332205842570 | 0.047806528167506200 |
| **-0.5861782293170290** | -0.6159470443652880 | -0.7074536620406230 | -0.5412452501572880 | -0.6058762686122950 | 0.06463101845500760 |
| **-0.4771626623292210** | -0.3084280477266620 | -0.2504312662919310 | -0.2871152934916490 | -0.3862588929942940 | 0.09914359950264460 |
| **-0.1740172823288130** | -0.4704988702794510 | -0.1802742756829030 | -0.2773180469854340 | -0.23114722287672300 | -0.0461708241087109 |
| **-0.4097214610553630** | -0.4813035917829710 | -0.2256461197084190 | -0.3925845021858100 | -0.37718763629691900 | -0.015396865888891200 |
| **0.0110189156817218** | -0.2618846320191940 | -0.5957949301845360 | -0.2461631248443250 | -0.17121307980401000 | -0.07495004504031510 |
| **-0.2090813992724770** | -0.2643780292892370 | -0.0407613833419057 | -0.1399195743901760 | -0.18061189394495900 | 0.04069231955478300 |
| **-0.5527658633883330** | -0.4580318839292370 | -0.8941550258903860 | -0.5939991594235920 | -0.5944554743541780 | 0.0004563149305858440 |
| **1.4132352743857700** | 1.6522133389504100 | 1.419281949632180 | 1.350763388271900 | 1.434035983612730 | -0.0832725953408271 |
| **-0.5910768114066040** | -0.3250506961936150 | -0.5006049844161070 | -0.3693070595836330 | -0.5070509684678700 | 0.1377439088842370 |
| **-0.6238588857346110** | -0.4455648975790220 | -0.7271077414383660 | -0.7202436786893890 | -0.5974051700865890 | -0.1228385086027990 |
| **Mean Squared Error** | 0.007706573204951880 | | | | |
| **R^2 Score** | 0.9922934267950480 | | | | |

# Conclusion: Key Insights

- Target Population: MORENA focuses on impoverished populations, which are their major voters. However, these voters do not necessarily contribute to GDP growth or industrial/economic development, presenting an opportunity for the opposition to highlight long-term benefits.
- Campaign Efforts: MORENA held significantly more public events (617) compared to the opposition (223), despite spending 3x less on campaign budgets. The opposition should focus on engaging directly with impoverished populations rather than spending excessively on propaganda.
- Key Demographics: The opposition should prioritize engaging with populations that have a strong preference for AMLO, federal welfare recipients, impoverished communities, Catholic believers, and adults.
- Socioeconomic Clusters: There are two distinct clusters: one with a high socioeconomic profile (less likely to vote for MORENA) and another with a low socioeconomic profile (MORENA followers).
- Political Polarization: AMLO has polarized these clusters with rhetoric of “Fifis vs Chairos.” This polarization should be addressed as all Mexicans are equal, and unity should be emphasized.
- Future Projects: For future projects, incorporating demographic, economic, and social information per city could help create a more robust model that generalizes the data better.