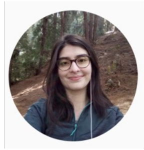
Prediction Model of Electoral Results in Mexico:
A Case Study using PREP, INEGI and INE Data.

Data Science Intensive Capstone Project, August 3rd, 2024

By Javier Jorge Pérez Ontiveros

Mentor



Ale Berbesi

Executive Summary:

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 u**nderstand why** the winning party (**MORENA) achieved a 59% acceptance rate**, while the opposition (PAN-PRI-PRD) secured only 28% of the votes.

Objectives:

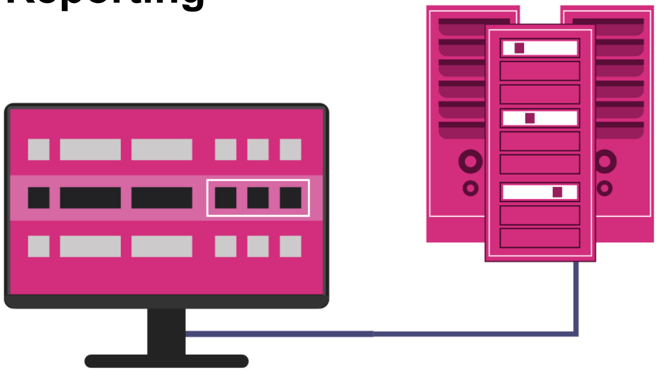
1) To generate a **Model that predicts votes for MORENA**
2) **To provide actionable insights** to support the Opposition Party in developing a strategy to improve their results in the 2030 elections.

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

- **Data Engineering**
- **Data Wrangling**
- **Exploratory Analysis (EDA) and Visualization**
- **Feature Engineering and Clusterization**
- **Predictive Modeling**
- **Reporting**



Data Sources

Data Engineering and Data Wrangling:
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:

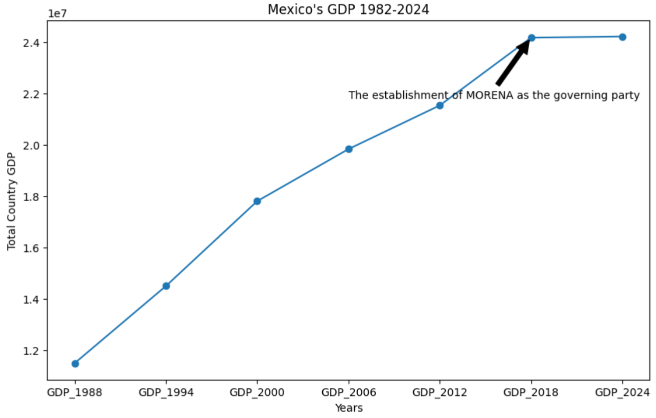
# of Campaign events, Marketing Budget, Federal Welfare,

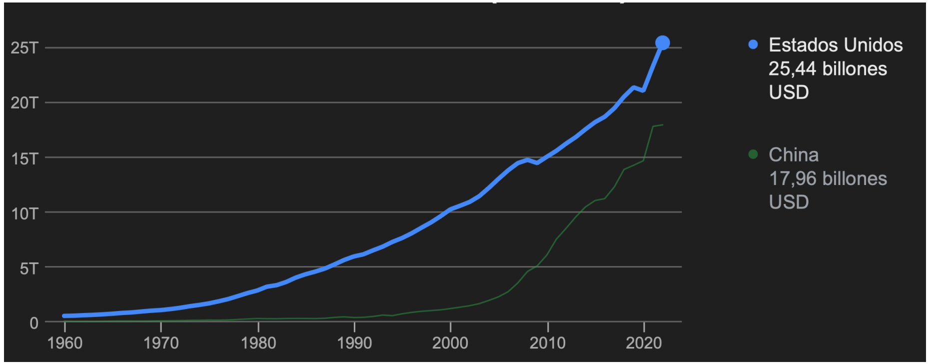
Final DataFrame

ENTIDAD	Morena_Votes_2024	Urban_Polling_Stations	Non_Urban_Polling_Stations	AMLO_2018	GDP_1988	GDP_1994	GDP_2000	GDP_2006	GDP_2012	GDP_2018	GDP_2024	Federal_Social_Welfare_Budget	Welfare_Recipents	MORENA_Events	Xochitl_Events	Maynez_Events	Avg_Schooling_Years	Avg_Income_Yr	Avg_Expense_Yr	Extreme_Poberty_habitants	Poberty_Habitants	Private_Hospital_Users	Public_Hospital_Users	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	175755	827	287	136806	62347	75145	95187	126685	147270	184618	177421	472056	46686	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	3640388	2551158	276864	398	78	29	11	357240	235592	159	2071	537244	6352039	397307	6634532	747	35	114318	121464	404052	91557	731391
COAHUILA	813432	3307	836	515518	387781	479740	655491	798549	870635	966055	922287	1235326	132323	7	4	1	10	300504	176020	59	538	113714	2487607	1136170	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	296963	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	1128686	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	280083	294551	317845	306831	3142321	340995	5	7	1	8	167016	122084	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	16992418
MICHOACAN	1E+06	3494	2936	820449	304548	364985	451793	485162	531827	633757	646901	2934380	308141	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	179369	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	2006353	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	650293	747712	849556	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	2944840
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	946950	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

A 40 columns x 32 rows Dataframe was created to predict the votes for MORENA Entity
This Dataframe was further modified and expanded with the use of some ratios and feature engineering.

Insight #1: GDP has stagnated since 2018





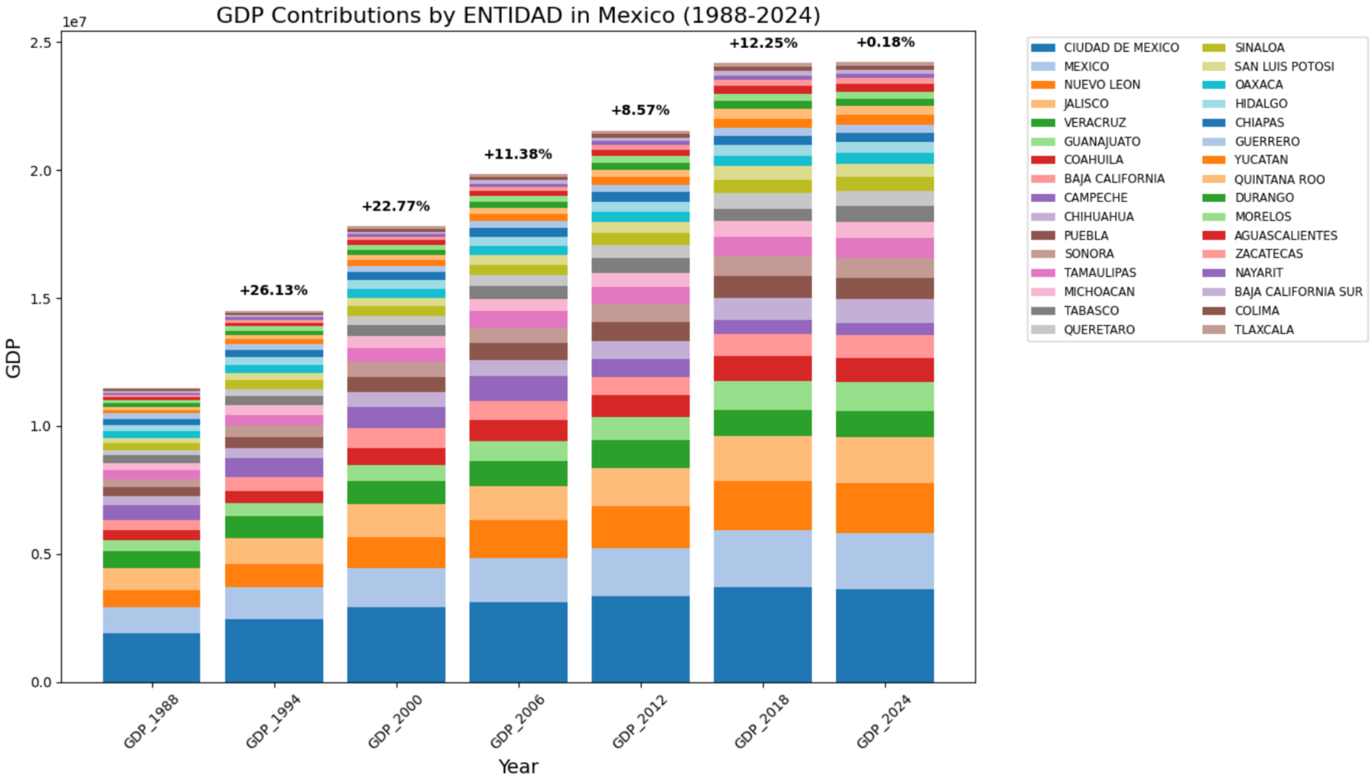
Since AMLO president won elections in 2018, the economy in Mexico has slowed down.

In comparisson, USA and China have a very positive trend (even after Covid)

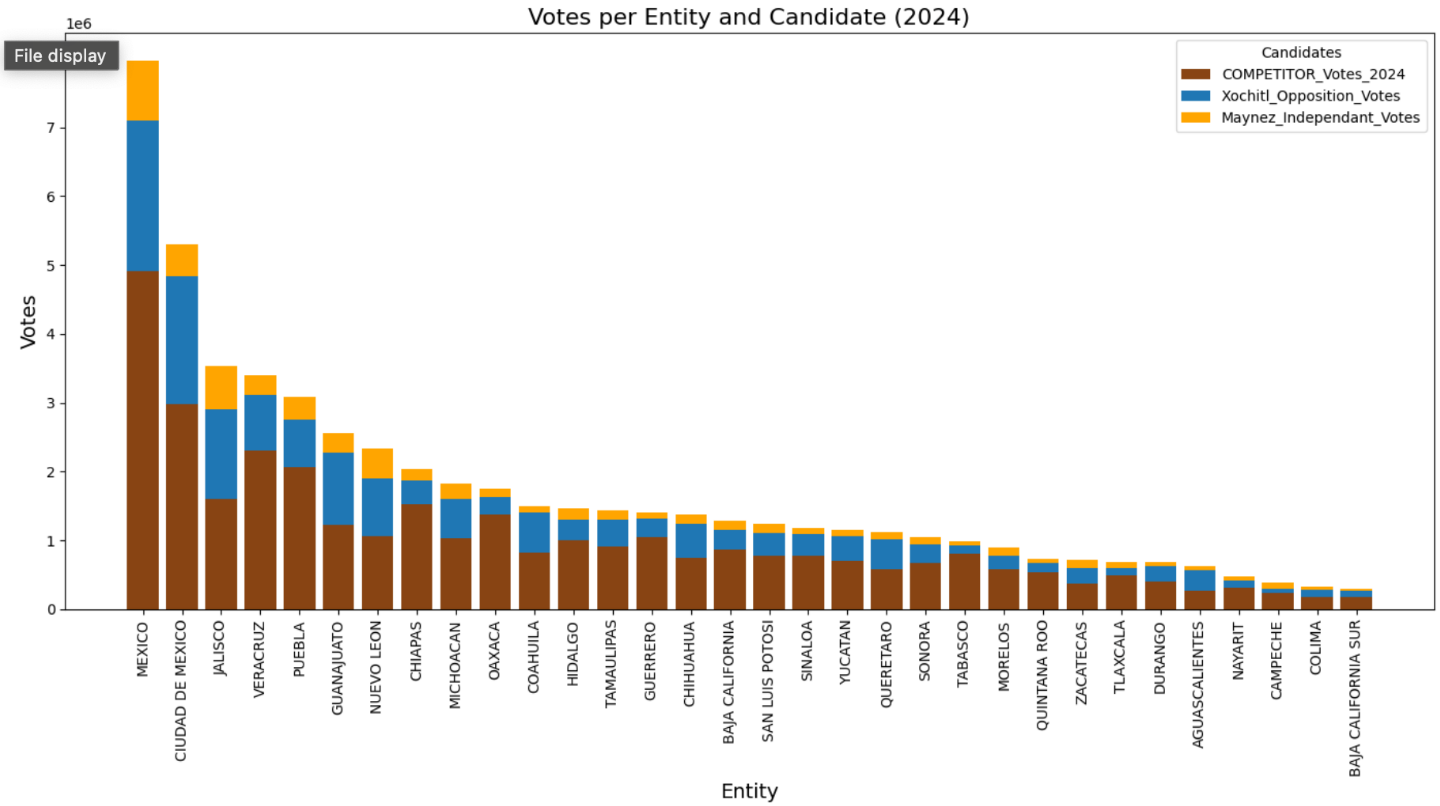
In the other hand, with Opposition parties, From 1988-2018, the GDP had a 10%-20% growth in average

This is a Key argument that needs to be “Sold” to the mexicans to gain more votes.

Crime Scene: Economy stagnation per State



Votes per Entity



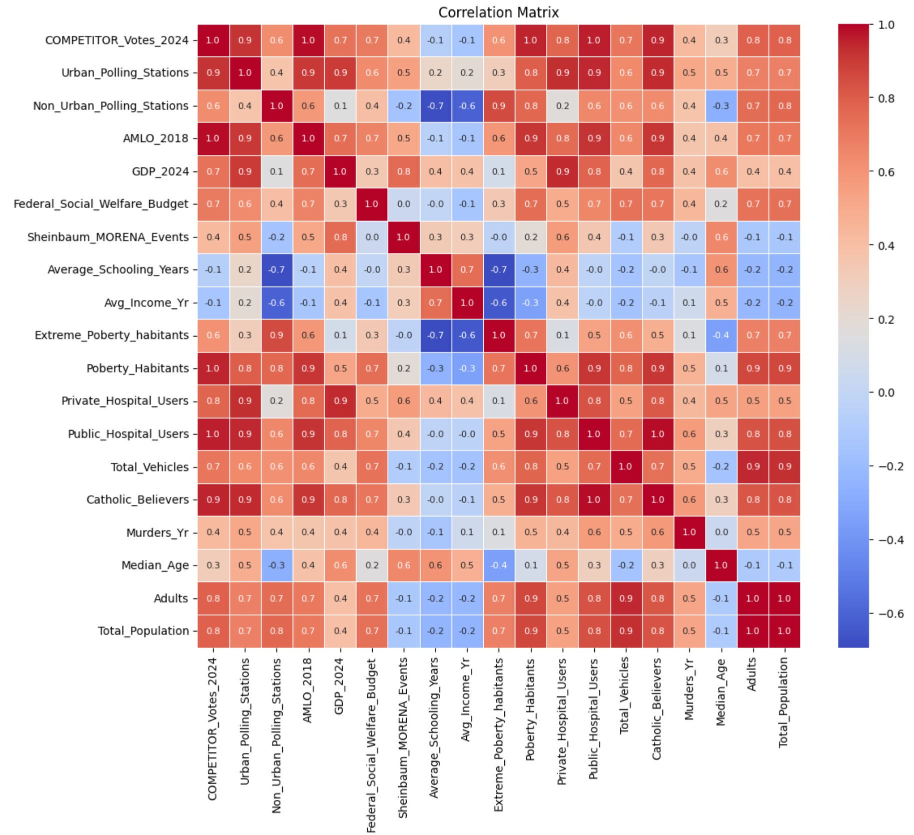
Geographical Heatmap Distribution

The map shows a significant trend where the highest resistance (blue color) is observed in the Bajío region, known for its industrial and economic strength. While the stronger MORENA states are in the South-East Region (Less Industiral and Economical development)

This is interesting as votes for MORENA, the competitor party, correlate highly with poverty ratios.

Opposition Party MUST focus on helping the less priviledged population from 2024-2030 if they want to have a chance of winning.

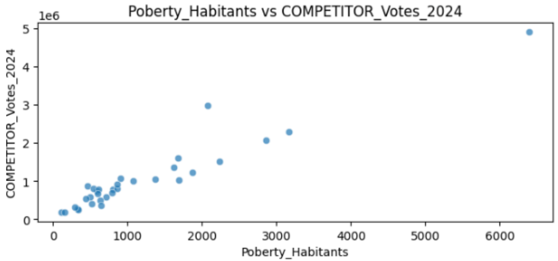
Features vs Response Heatmap

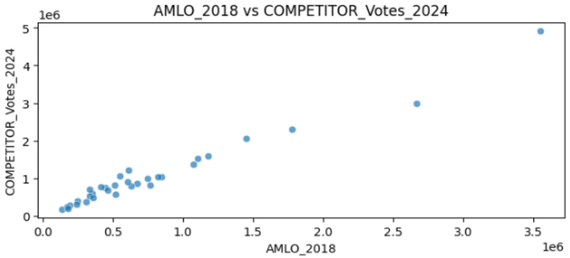


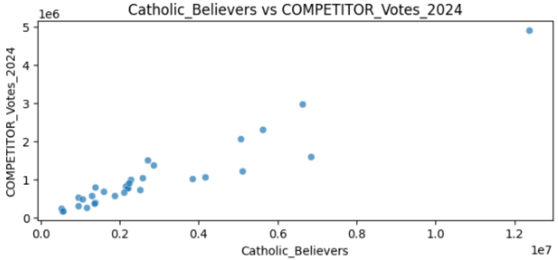
Since the begining of this study these variables showed a positive correlation with MORENA votes:

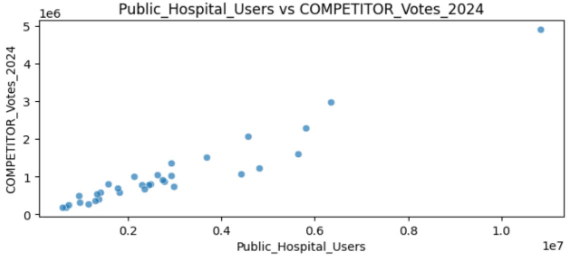
- Preference for AMLO in 2018
- Federal Warefare budget
- Poberty Habitants
- Public Hospitals
- Catholic Beliefs
- Adults Amount

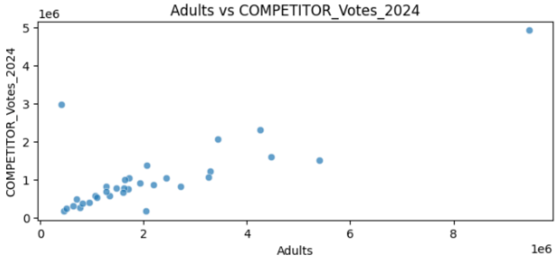
Linear Correlations

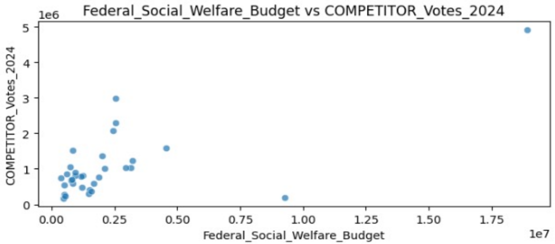








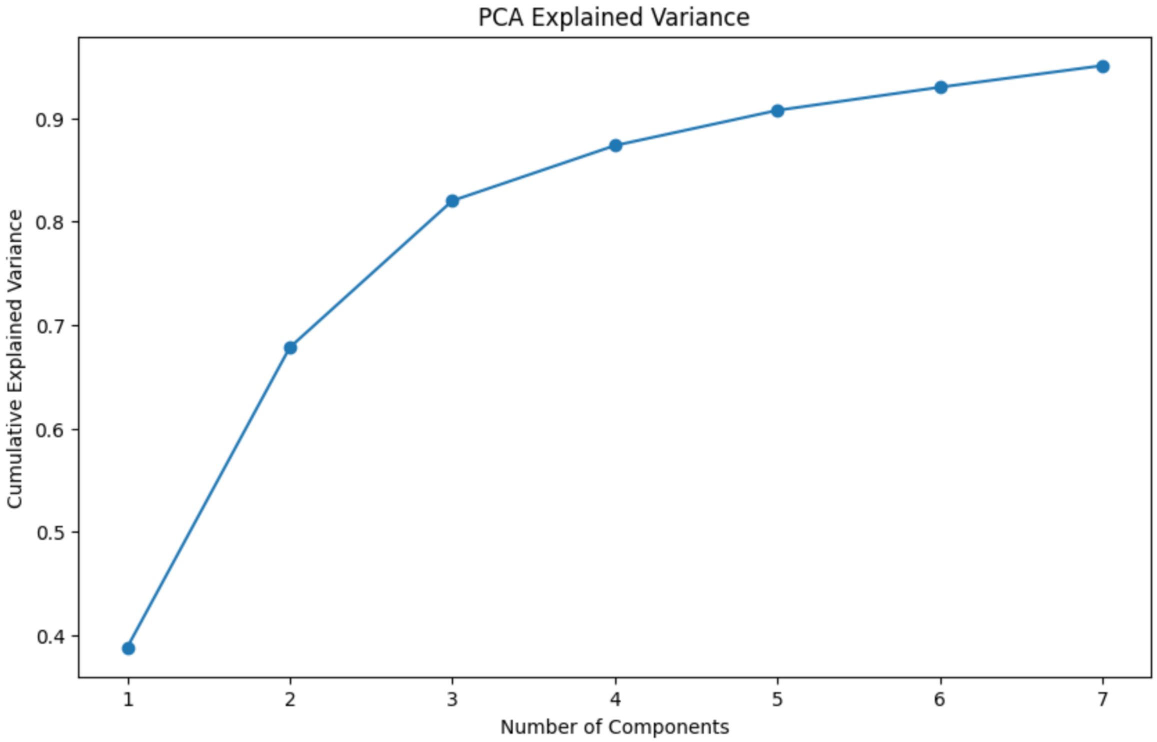


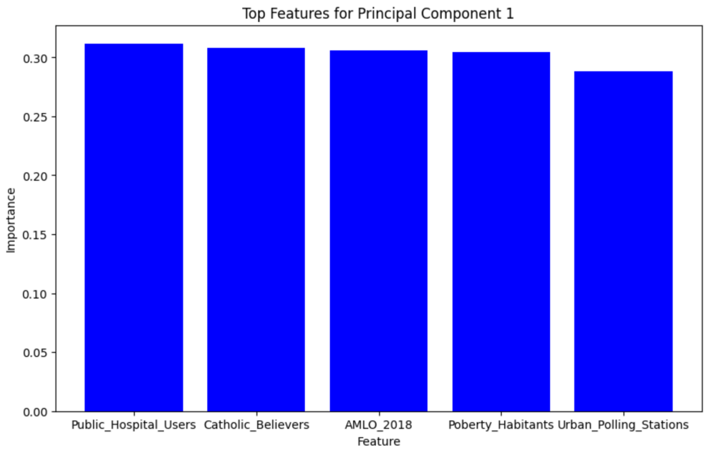


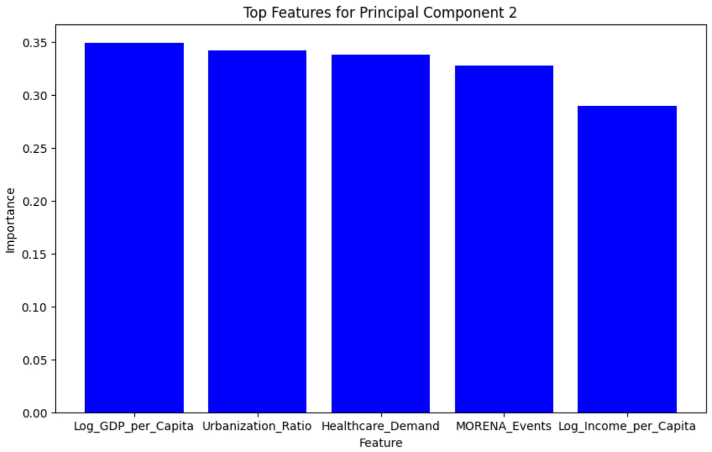
In General the votes for MORENA increas
when these variables increase:

- Poberty Habitants
- Catholic Believers
- Adults 40-50 yrs old
- AMLO followers since 2018
- Public Hospital Users
- Federal Wellfare Budget

Clustering (PCA)



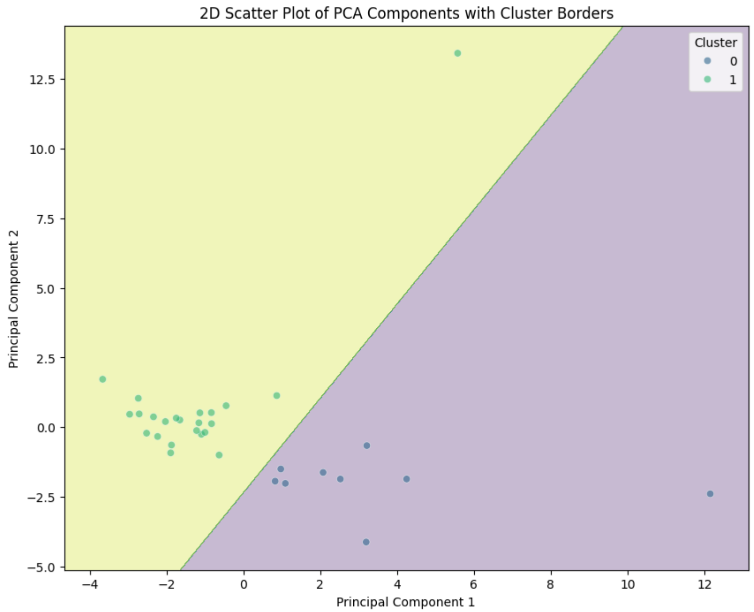


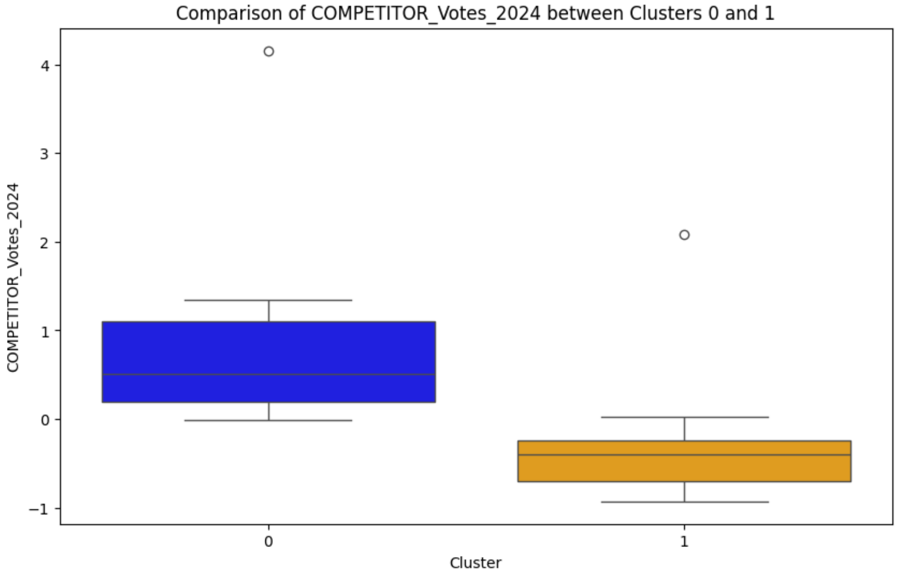


2 Componenets explain 67% of the Total Variation

3 Components explain 87% of Total Variation

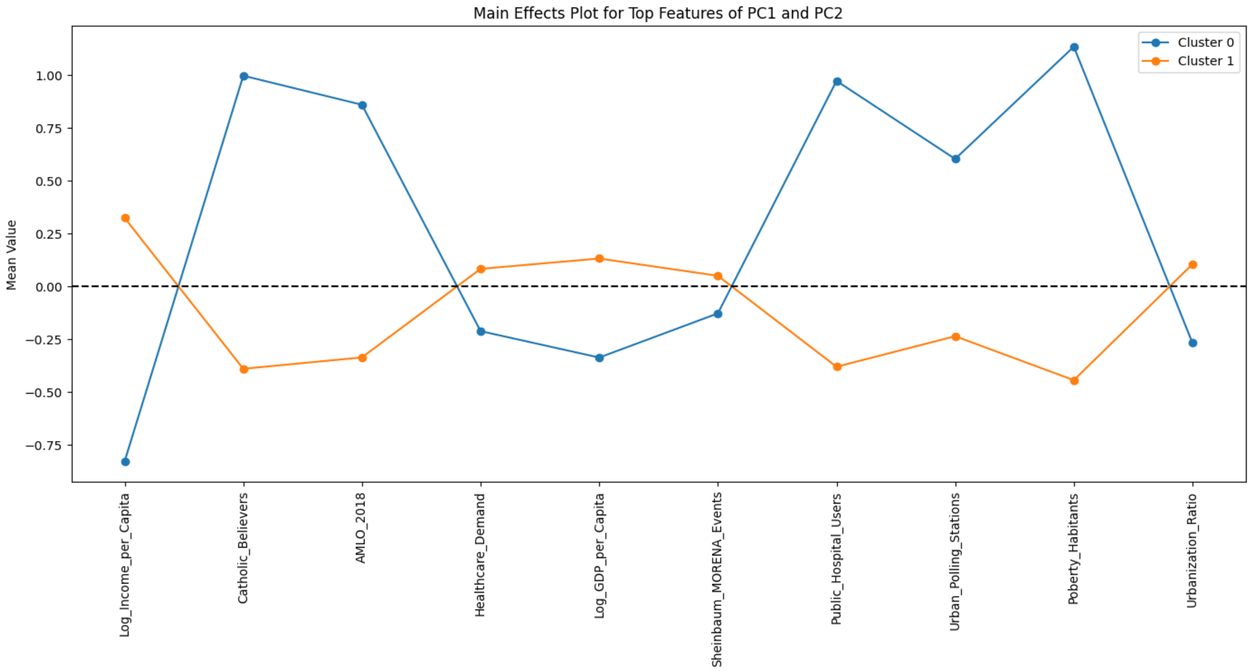
Clustering





Cluster #1 (PCA 1) has a lower voting preference for MORENA.
While has this features as key characteristics: High GDP, High Urbanization Ratio, Higher Healthcare demand and Higher Income per Capita.
Also is important to mention that in this cluster is where MORENA performed more campaing events!

Behavior of the 2 clusters identified (Morena and Opposition)



Cluster #0 has a **higher** preference for **MORENA** while,
Cluster #1 (PCA 1) has a **lower** voting preference for **MORENA**.

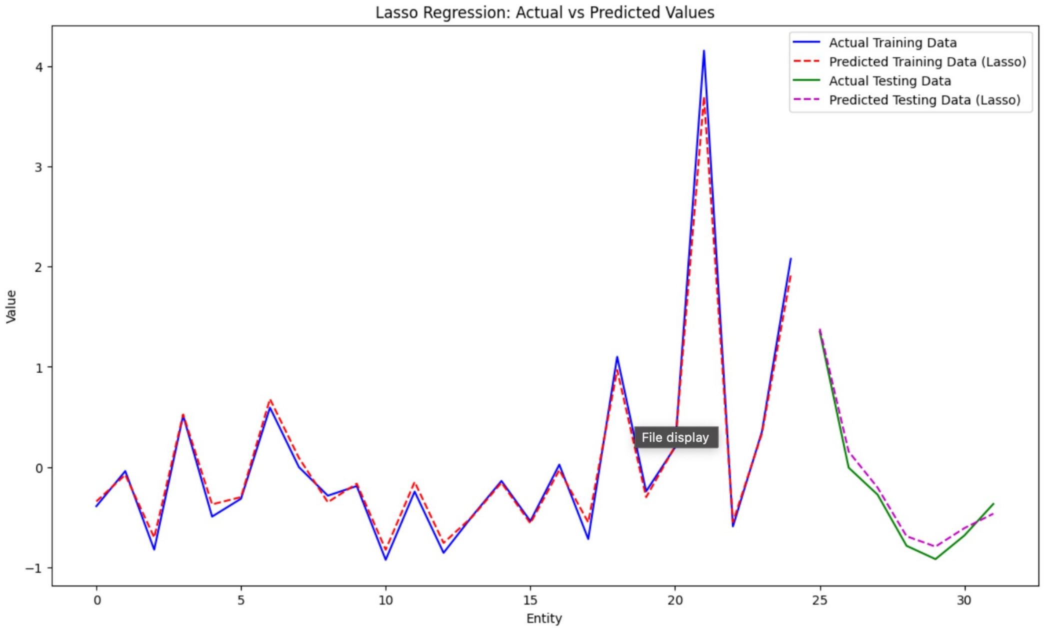
Modeling

```python
# 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)
```

Model	Lasso
Train MSE	0.014998
Train R^2	0.986565
Test MSE	0.010197
Test R^2	0.97997
Model	Ridge
Train MSE	0.002201
Train R^2	0.998029
Test MSE	0.047715
Test R^2	0.906278
Model	Linear Regression
Train MSE	0.0
Train R^2	1.0
Test MSE	0.054846
Test R^2	0.892273
Model	Random Forest
Train MSE	0.100472
Train R^2	0.910003
Test MSE	0.067189
Test R^2	0.868028
Model	Decision Tree
Train MSE	0.0
Train R^2	1.0
Test MSE	0.127489
Test R^2	0.749587

5 Models where developed: Decission Tree, Random Forest, Linear Regression, Ridge and Lasso. The last model had the best performance (98% accuracy in testing and training)

Results



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 representation

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.

Thank You!

https://www.linkedin.com/in/jjpo/

javierjorge77@gmail.com

https://github.com/javierjorge77/Springboard/tree/main/Capstones/Capstone2

(+521) 8711777903