Red Cross Donor Prediction Project

Exploratory Data Analysis & Machine Learning (2025)

Content

- **Business Case** :

  Donor Prediction
  This project focuses on predicting the donor indicator, enabling the organization to allocate limited resources more efficiently by targeting the most promising prospects for outreach.

- **Data Acquisition** – Donor data was obtained in CSV format.
- **Data Preparation** – Unnecessary, redundant, or large missing data was dropped and ‘PostalCode’ was mapped to show actual regions (details in the following page)
- **Data Visualization** – Graphs were created using Tableau and multiple data reports were created to understand the relation of features.
- **Machine Learning** – Multiple classification machine learning models were tested and evaluated and the best model was selected to predict the donor indicator.

Data Preparation

- **Total number of records** = 34,508
- **Following missing values were found**:
  - Marital Status (9,940 records), Wealth Rating(2,709 records), Academic Degree Level(7,606 records), Donor Date of Birth(13,318 records)
- **Missing values treatment:**
  - Among values with more than 20% missing, ‘Donor Date of Birth’ was dropped due to redundancy, but the rest were kept to gain insight from the limited values.
- **Donor Postal Code remapping:**
  - The value was remapped to show the actual region(State and City) of the donors.
  - 1,210 values were found to be non-valid Zipcodes and was labeled ‘None’.
- **Donation by fiscal Year modification:**
  - The ‘$’ sign and decimal points were dropped to better represent significant relationship.
- **Wealth Rating remapping:**
  - The data was remapped to show numbers 1 through 8 for each tier (i.e. '$1-$24,999' to “1”, '$25,000-$49,999' to “2”...)
- **No Duplicates were found**

Data Preparation

Summary table

	Variable class	# unique values	Missing observations	Any problems?
DonorUniqueId	numeric	34508	0.00 %	
DonorPostalCode	character	20992	0.00 %	×
DonorAge	numeric	102	0.00 %	×
MaritalStatus	character	7	71.20 %	×
GenderIdentity	character	6	1.43 %	×
IsMemberFlag	character	1	0.00 %	×
IsAlumnusFlag	character	2	0.00 %	
IsParentFlag	character	2	0.00 %	
HasInvolvementFlag	character	2	0.00 %	
WealthRating	numeric	9	92.15 %	×
AcademicDegreeLevel	character	8	77.96 %	
PreferredAddressType	character	5	11.72 %	
HasEmailFlag	character	2	0.00 %	
ConsecutiveDonorYears	numeric	33	0.00 %	×
LastFiscalYearDonation	numeric	121	0.00 %	×
Donation2FiscalYearsAgo	numeric	128	0.00 %	×
Donation3FiscalYearsAgo	numeric	133	0.00 %	×
Donation4FiscalYearsAgo	numeric	130	0.00 %	×
Donation5FiscalYearsAgo	numeric	146	0.00 %	×
CurrentFiscalYearDonation	numeric	113	0.00 %	×
CumulativeDonationAmount	character	1595	0.00 %	×
DonorIndicatorFlag.	character	2	0.00 %	
Region	character	46	0.00 %	

- The Summary Table for features were collected by ***DataReporter*** and several variables had red flags.
- ‘Donor Age’: The following possible outlier values were detected : Ages under 10 (65 cases).
- ‘Gender Identity’: ‘Blank’, ‘U’, ‘Uknown’, ‘Unknown’ accounts for 1,597 cases.
- ‘Is Member Flag’: ‘No’ was the only value, with no missing data.

Exploratory Data Analysis

The Features (i.e., variables) are segregated into three Categories namely:

- **Dependent Variable (Target) –** The donor indicator were to be predicted
- **Noise Features :** Variables which would not have a significant impact on the value of the Target

  Redundant variables(‘Donor Postal Code’, ‘Donor Date of Birth’)

  Feeder variables (‘Cumulative Donation Amount’)

  Variables irrelevant to prediction (‘Donor Unique ID’)
- **Predictor Variables** : Variables which were considered to have an impact on the Target

Exploratory Data Analysis (contd..)

Noise Variables

Donor Unique ID
Donor Postal Code
Donor Date of Birth
Cumulative Donation Amount

Predictor Variables

Consecutive Donor Years	Marital Status
Current Fiscal Year Donation	Gender Identity
Last Fiscal Year Donation	Is Member flag
Donation 2 Fiscal Years Ago	Is Alumnus flag
Donation 3 Fiscal Years Ago	Is Parent flag
Donation 4 Fiscal Years Ago	Has Involvement flag
Donation 5 Fiscal Years Ago	Wealth Rating
Donor Age	Academic Degree Level
Preferred Address Type	Has Email Flag
Region	

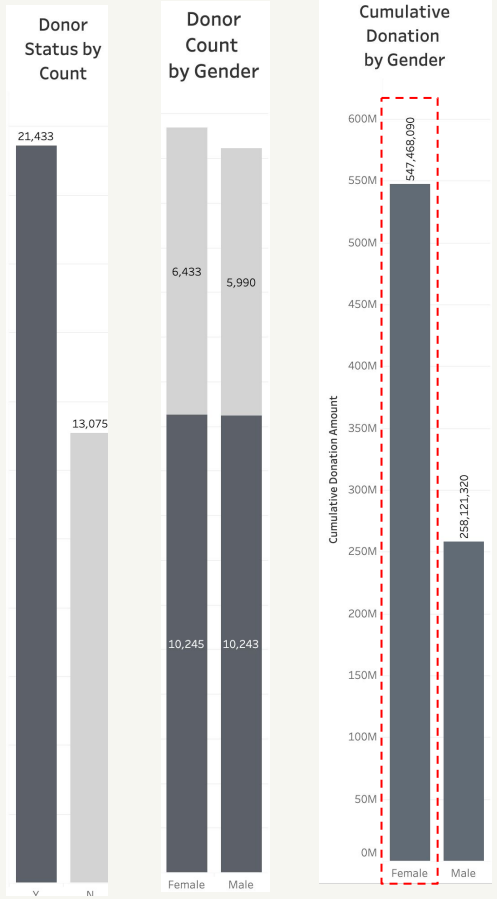
Data Visualization

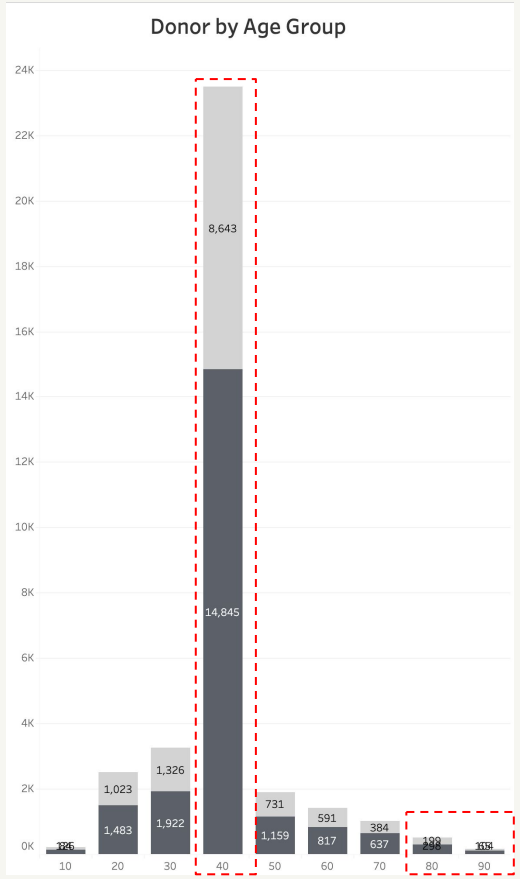
- Data visuals were created using ***Tableau*** where charts are plotted using the independent variables against the dependent variable (Donor Indicator flag).

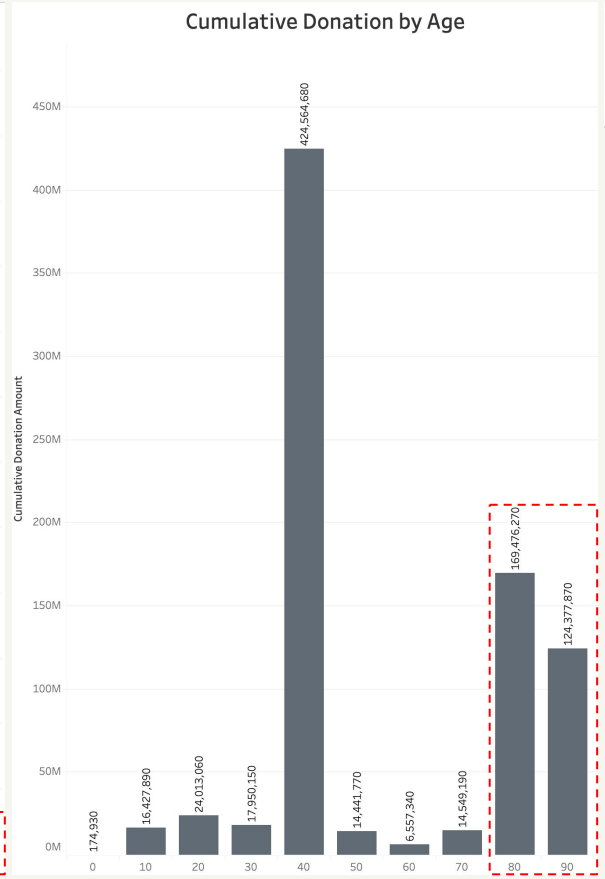
- Reports were created using ***DataReporter*** and ***DataExplorer*** to understand the correlation and aggregation of features.

- These charts help us in getting a preliminary idea about the **relationship** between the **independent variables** and the **dependent variable.**

Data Visualization



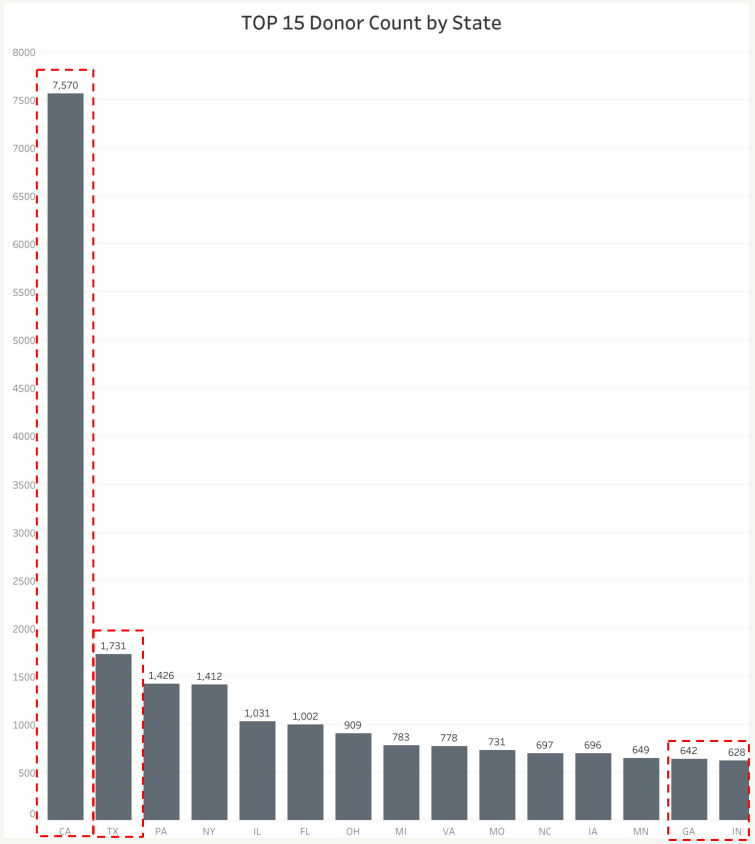


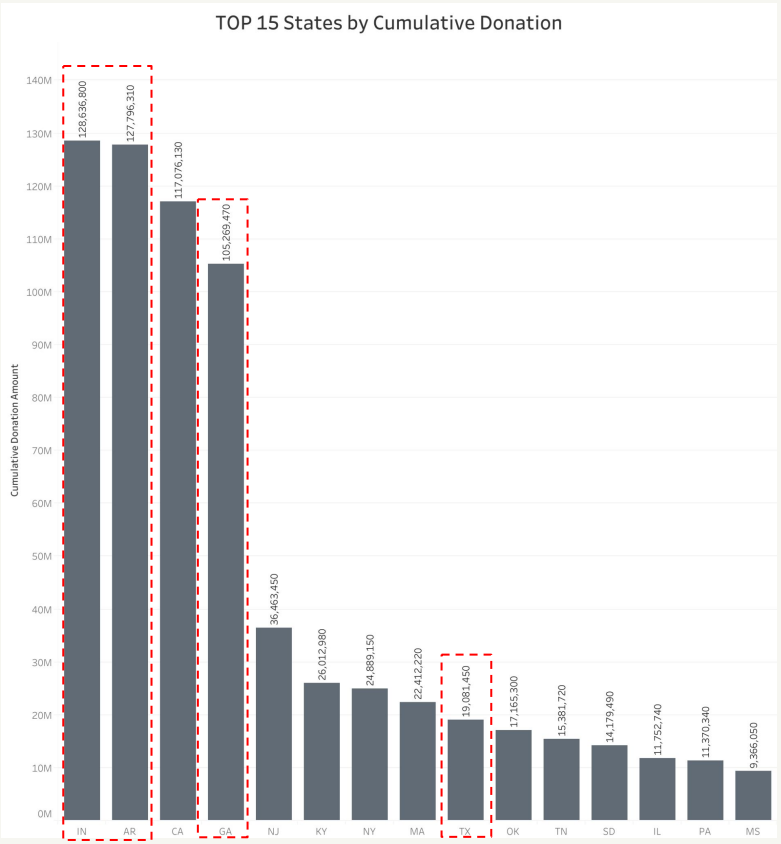


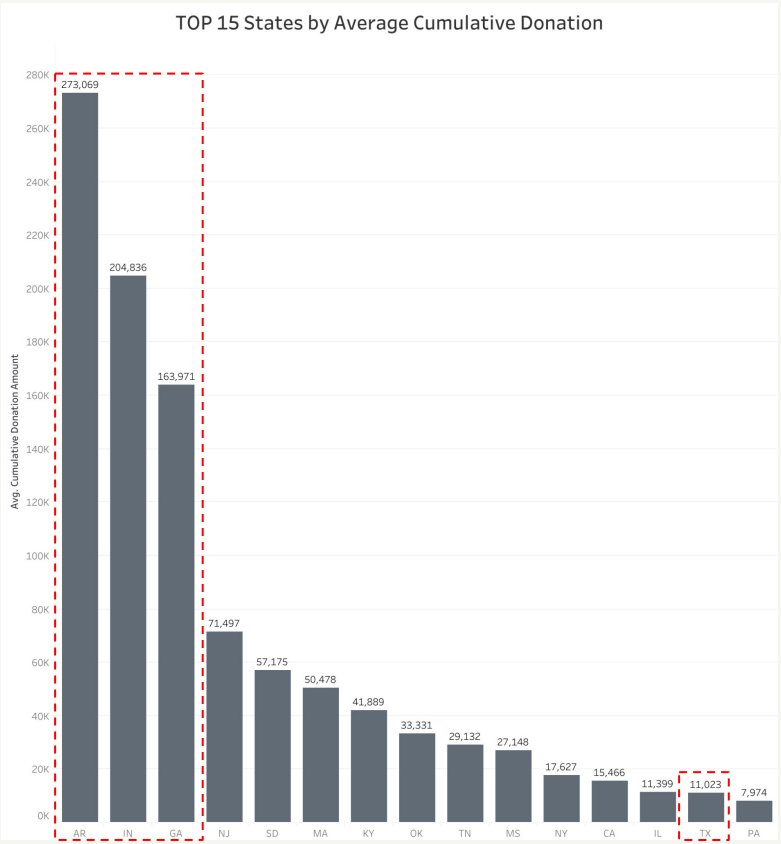
Key Highlights

- The data shows that **Donors** make up over **60%** of the total, outnumbering **Non-Donors**.
- While there was **no significant difference** in the **total number of Male and Female** members or donors, **Females** contributed significantly more in **cumulative donations**—nearly $290 million more than males.
- The **majority** of members and donors are in their **40's**.
- Although the **number** of donors in their **80's** and **90's** is **small**, their **cumulative donations** rank **second** and **third**, respectively.

Data Visualization



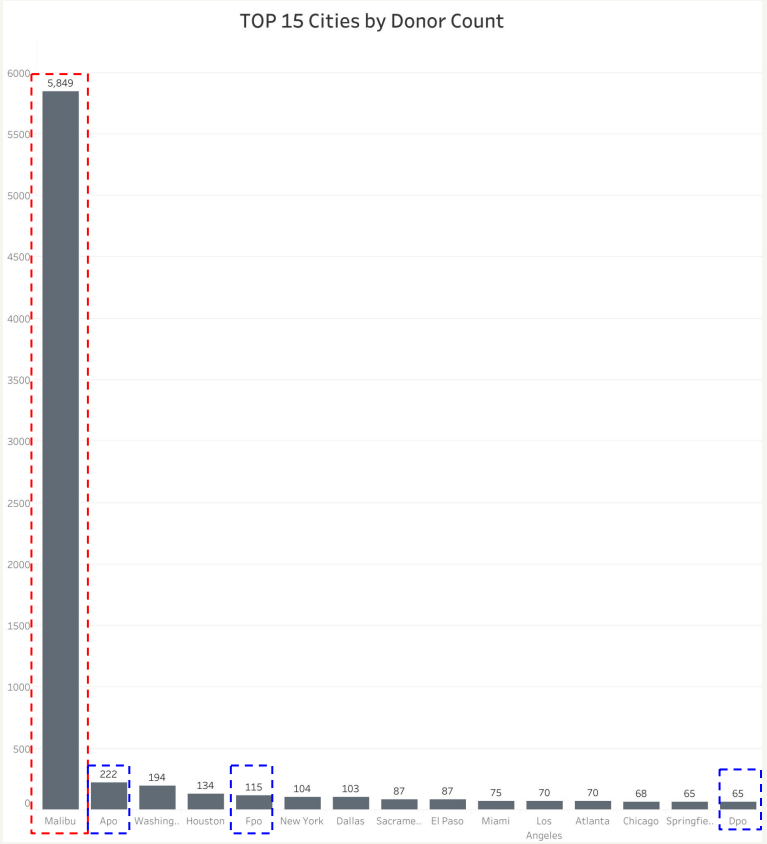


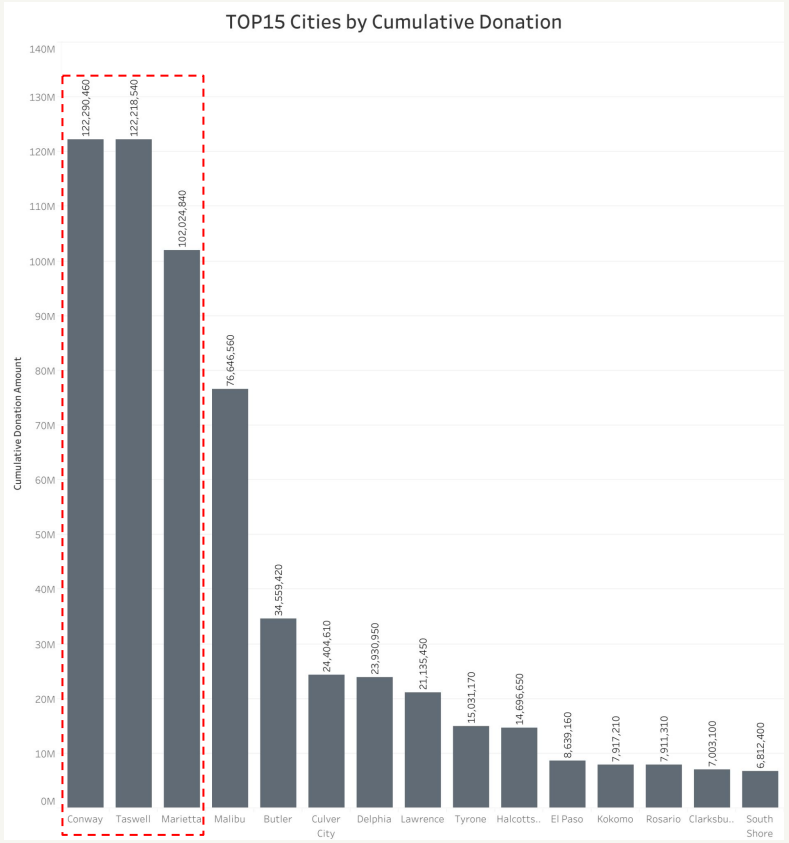


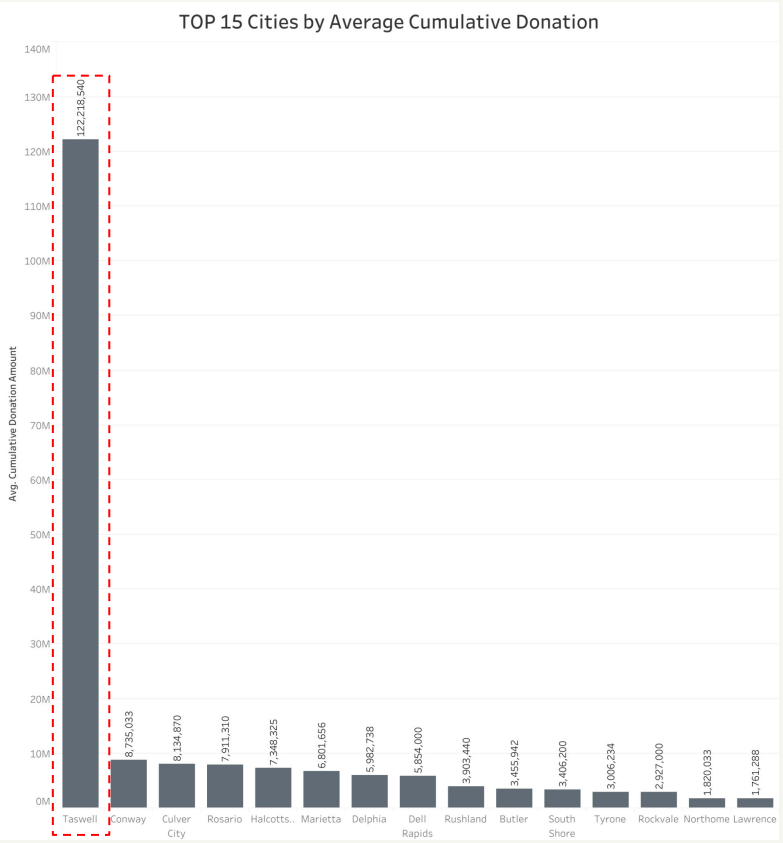
Key Highlights

- **California** leads in donor count. However, **Indiana** and **Arkansas** are the leaders in terms of **cumulative donation amount**, and **average cumulative donation**, despite **Arkansas** ranking **27th** and **Indiana 15th** by donor count.
- **Georgia**, which is only 14th in donor count, ranked **4th** in **cumulative donation** and **3rd** in **average cumulative donation**.
- **Texas**, which ranked **2nd** in **donor count**, only ranked **9th** in **cumulative donations** and **14th in average cumulative donation**.

Data Visualization



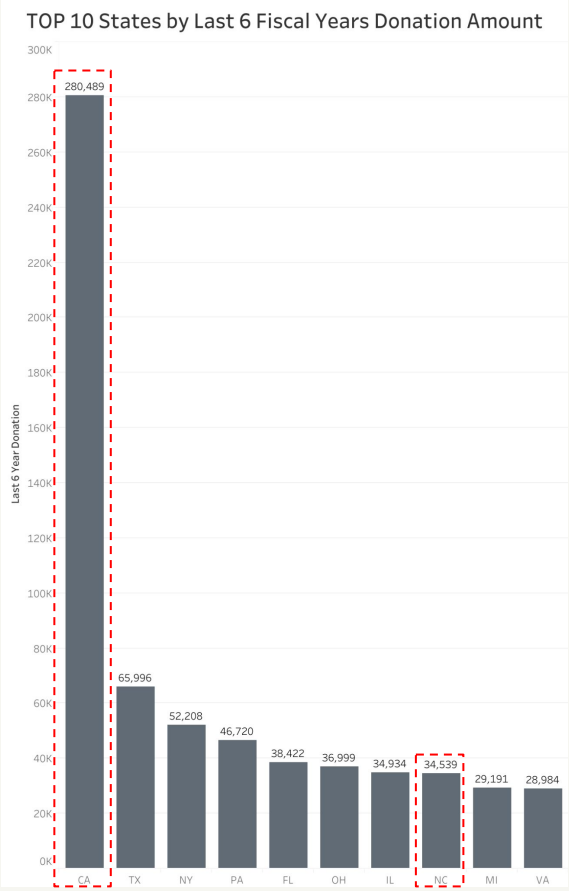


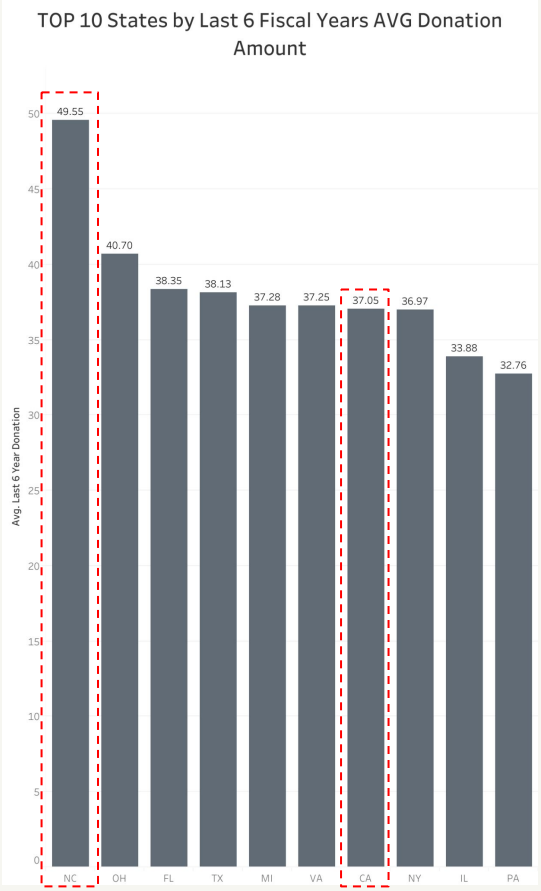


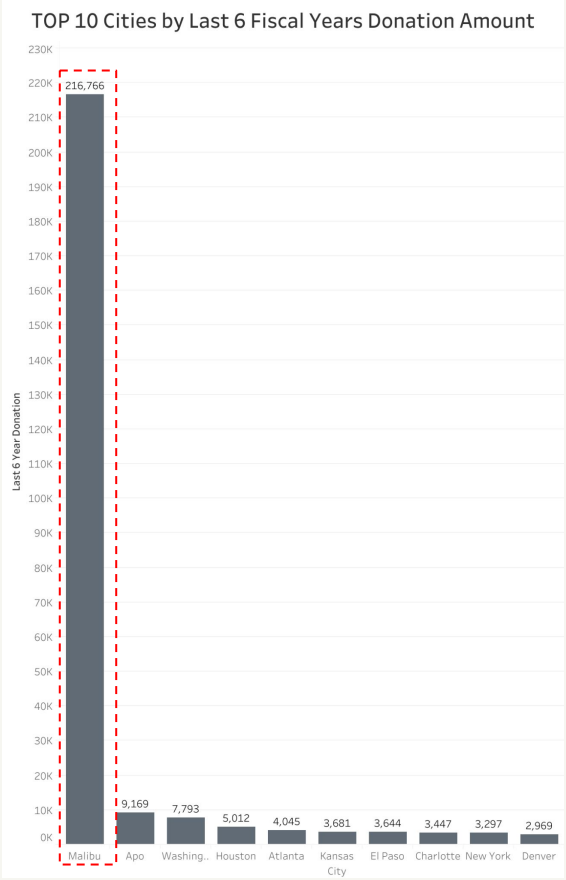
Key Highlights

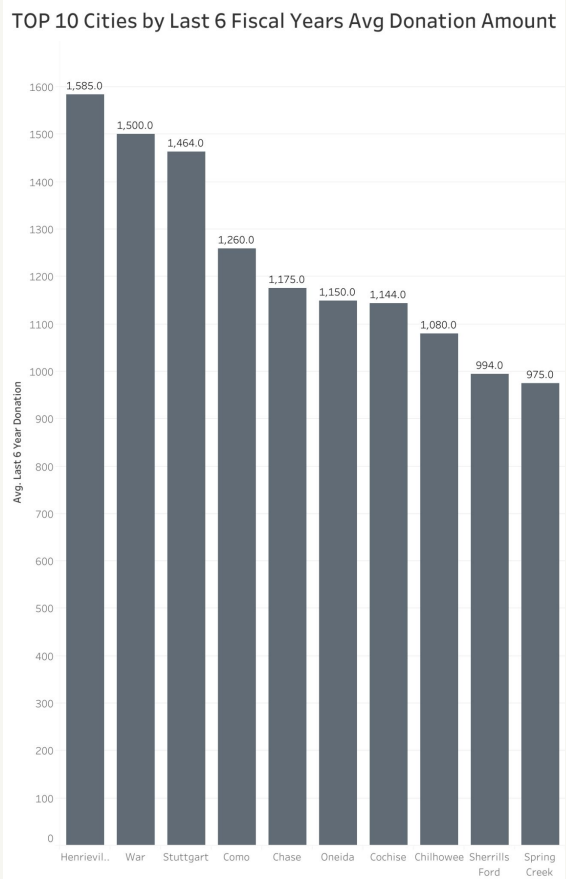
- **Malibu** is the dominant leader in donor count among cities. However, it only comes in staggering **4th** in **cumulative donation**, and is beyond TOP 50 cities when it comes to average cumulative donation.
- **APO**(Army Post Office), **FPO**(Fleet Post Office), and **DPO**(Diplomatic Post Office) are among the Top 15 postal codes.
- **Conway(South Carolina), Taswell(Indiana),Marietta(Georgia)** are the **top 3** cities by **cumulative donation**.
- **Taswell** is the ultimate contributor when it comes to **average cumulative donation**, contributing over **85%** of total average cumulative donation.

Data Visualization





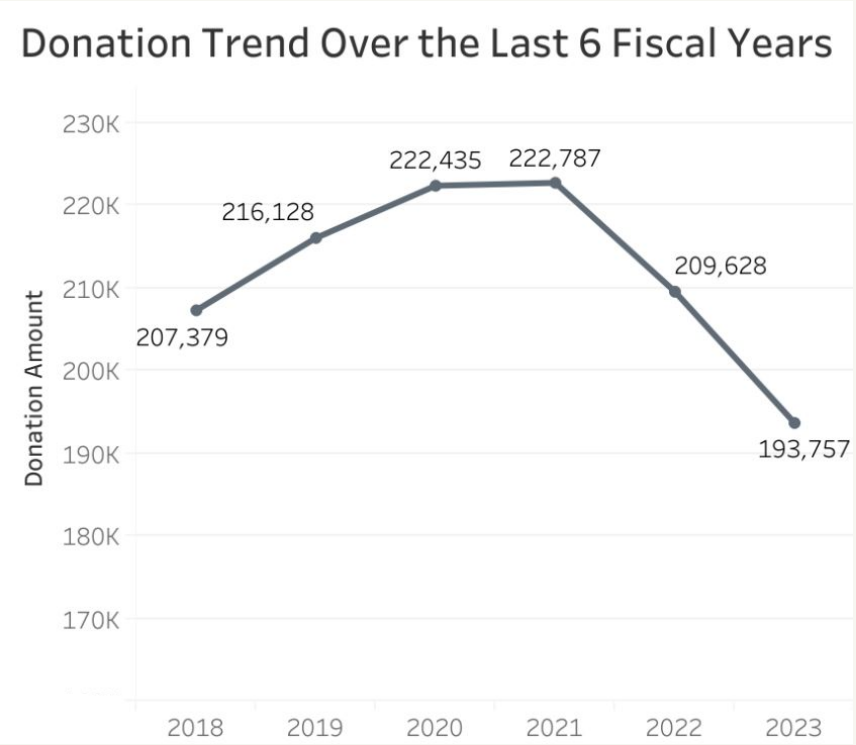




Key Highlights

- When analyzed by **last 6 fiscal years donation amount, California** takes the lead, but ranks only **7th** in average donation amount.
- **North Carolina** leads in **average donation amount**, despite only ranking **8th** in **total donation amount**.
- **Malibu**, the dominant leader in donor count among cities, also dominates the **total donation amount in the last 6 fiscal years.**
- However, when it comes to **average donation amount** in the last 6 fiscal years, **Henrieville (UT), War (WV), and Stuttgart (AR)** are the leaders, with no overlap in top cities by total vs.average.

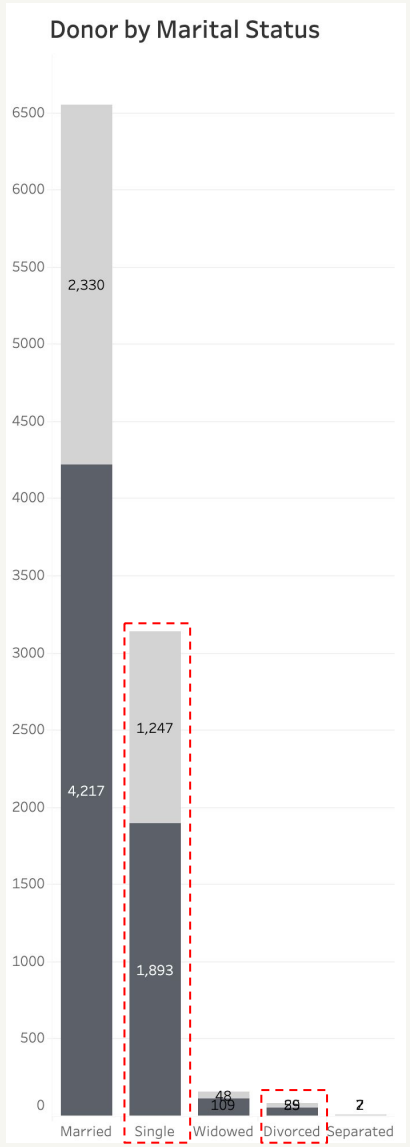
Data Visualization

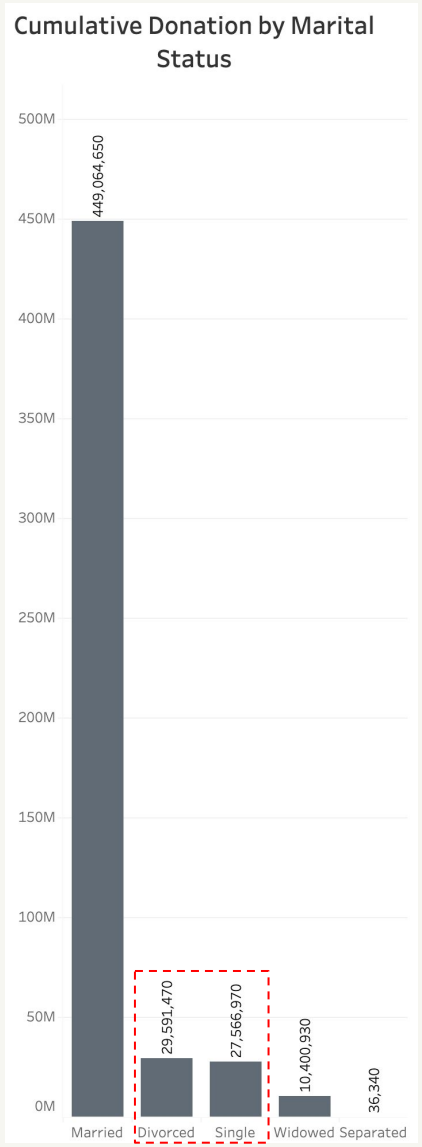


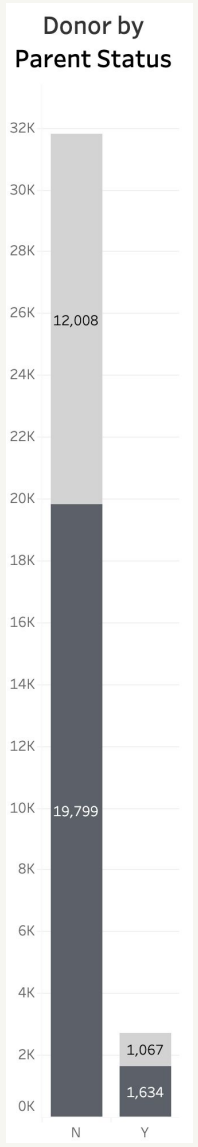
Key Highlights

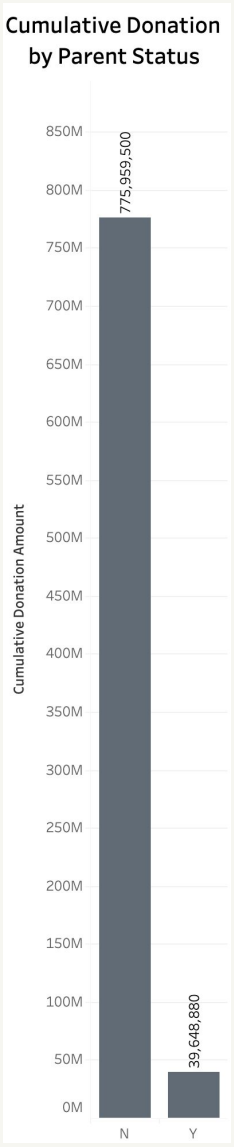
- The total donation amount shows a **rising trend from 2018 to 2021**(a total of 7.4% growth), peaking in **2021**.
- From **2022** onward, donations **declined sharply**, reaching their lowest point in **2023**(a decline of approximately 13%).

Data Visualization





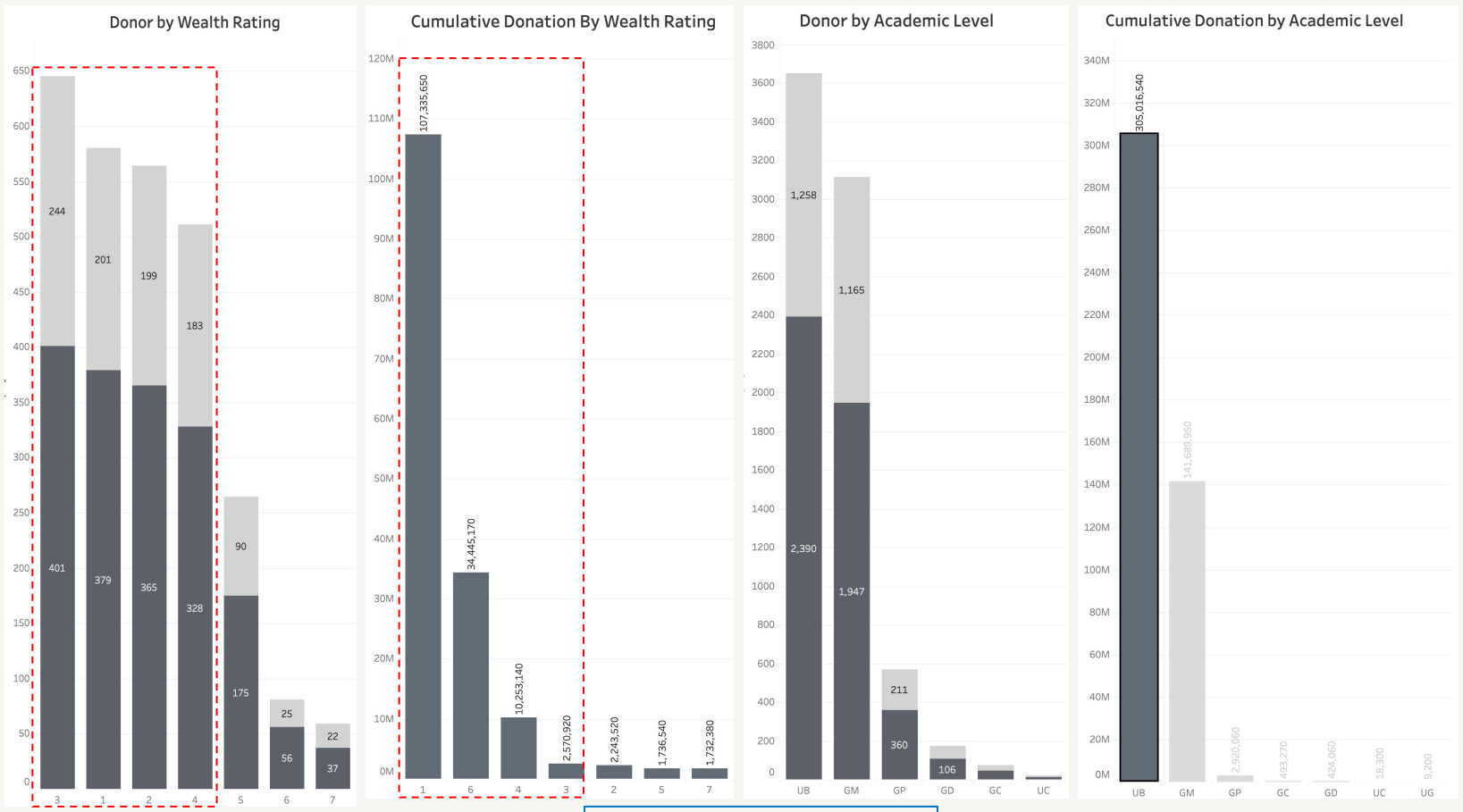




Key Highlights

- Donors with a **Married** status **significantly outnumber others**, with over **TWICE** as many married donors than all the groups combined. Their **cumulative donation amount** is also more than **SIX times** that of all other status combined.
- **Single** donors are the **second** largest group by **count**, and while **Divorced** members have a much smaller representation *(3,140 single vs. 84 divorced)*,the ranking is reversed in **cumulative donations** between divorced and single donors, at **$29M and $27M**, respectively.
- Regarding Parental status, **Non-parents dominate**, with nearly **11 times** more donors and contributing approximately **$740M more** in total donations.

Data Visualization

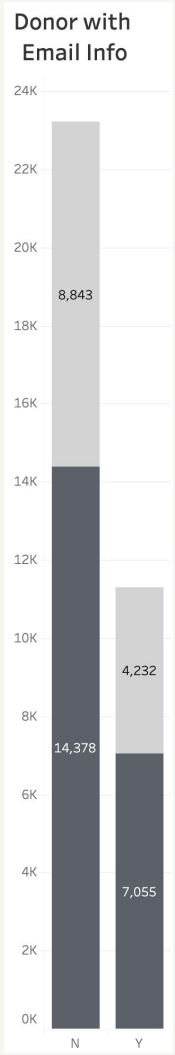


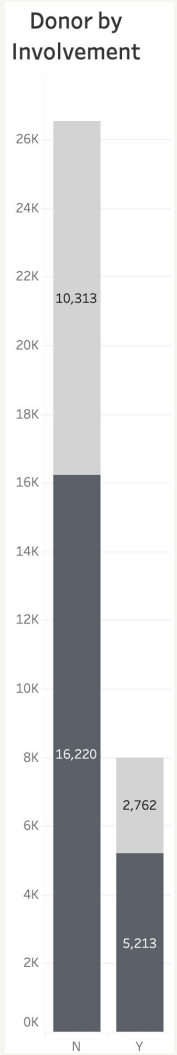
Disclaimer:
82% of the Wealth Rating data and 78% of the Academic Level data were missing. This chart excludes the missing values and is based on the analysis of the available data.

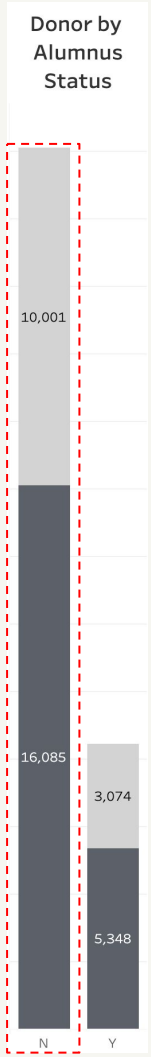
Key Highlights

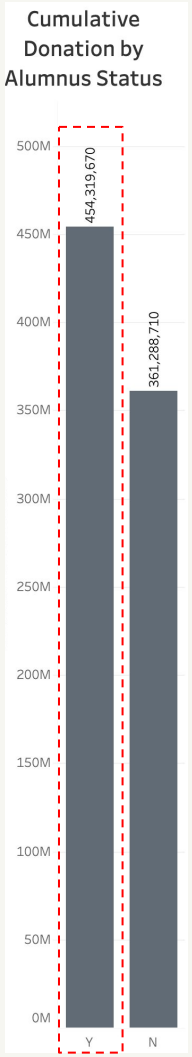
- **Tier '3'**, representing donations between $50,000 and $99,999, leads in **donor count**, with the rank order of tiers being **3 > 1 > 2 > 4**.
- For **cumulative donations**, the ranking changes, with **tier 1 ('$1 - $24,999')** emerging as the dominant leader, with the rank order **1>6>4> 3**.
- **UB** and **GM** lead both **donor count** and **cumulative donation amounts**, with **UB** especially dominating in **cumulative donations**.

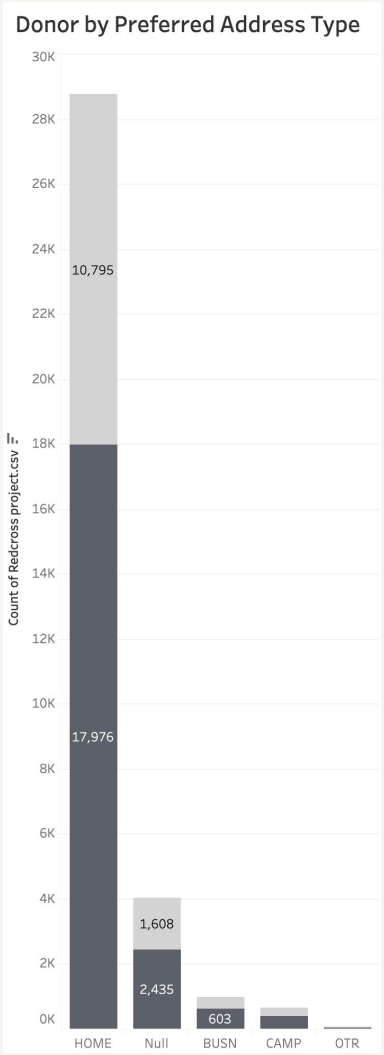
Data Visualization

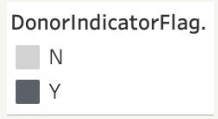








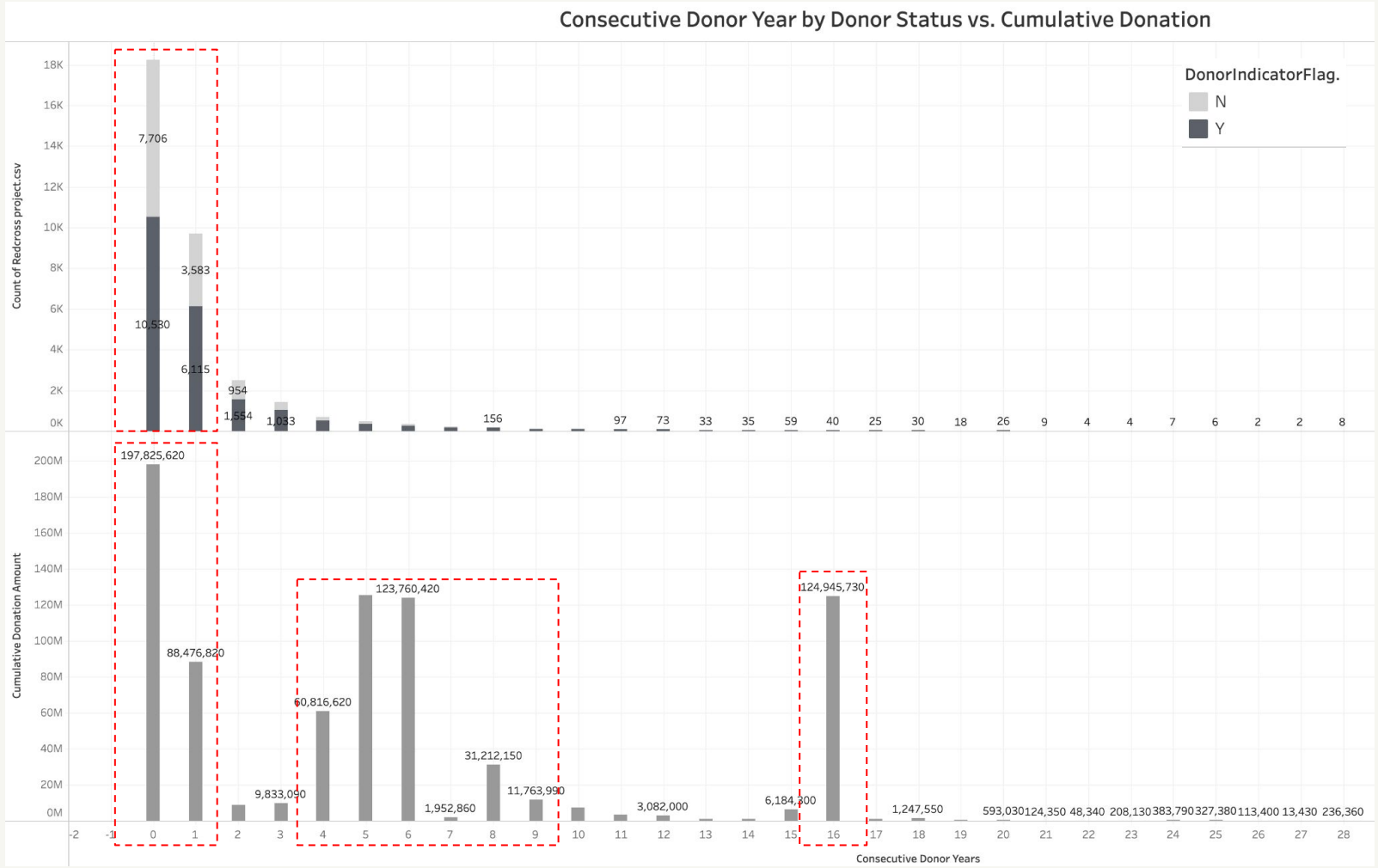




Key Highlights

- **Email** information is **missing** for over **two-thirds** of donors.
- Only **24%** of donors actively engage in **organizational activities and events**.
- **Non-Alumni** donors **outnumber** Alumni by a **3 times**.
- However**, Alumni** contributions exceed those of Non-Alumni in **cumulative donation amount**.
- **Home address** is the preferred communication channel for receiving information.

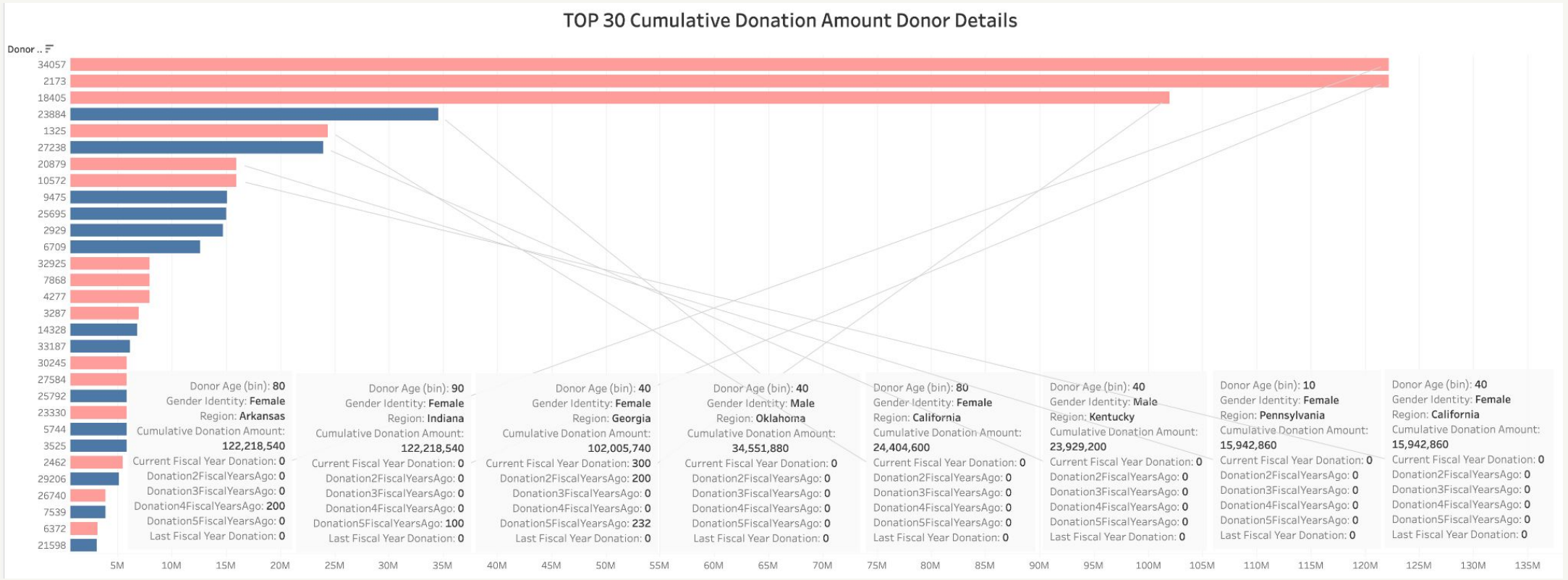
Data Visualization



Key Highlights

- The **majority** of donors fall within the **0-1 year consecutive donation years**.
- However, donors with 4-6, 8-9, and 16 consecutive years of donation contribute significantly to the cumulative donation amount, representing **significant value despite lower count in status**.

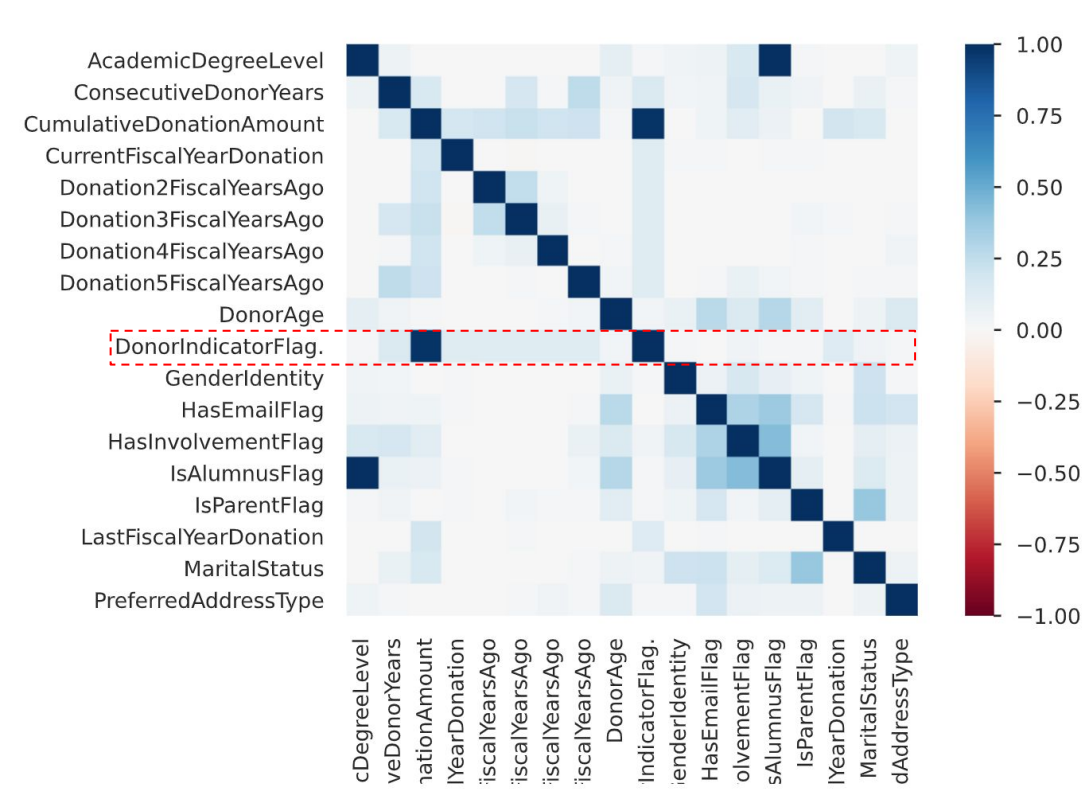
Data Visualization



Key Highlights

- Analysis of top unique donors by cumulative donation amount reveals that the **top three donors significantly outpace the rest**, with the **#1** donor contributing nearly **3.5 times more ($88M)** than the #4 donor.
- The majority of top donors are aged **40's** or **80-90's.**
- Most **have NOT** donated within the **past 5 fiscal years**.

Data Visualization



Key Highlights

- Analysis revealed no significant correlations among variables (all R-values were < 0.3)
- The only exception was Cumulative Donation Amount, which showed a perfect correlation (R = 1.0) as it directly feeds into the target variable.

Insights from Data

✅Target Female Donors:

Since **females** contribute more than males, targeting female donors for engagement and re-engagement could be beneficial. Developing tailored communication or campaigns aimed at females, particularly in their **40s**, may help increase cumulative donations further.

✅Engage with High-Value Older Donors:

The **older age groups (80s and 90s)** contribute disproportionately in donations **despite their small number**. They may require different communication strategies and other long-term engagement opportunities. Consider personalized outreach to these donors, **especially if they have not donated in recent years**.

✅Focus on Married Donors for Increased Contributions:

Married donors are significantly more likely to donate and contribute larger sums. Engaging married donors through **couples-focused campaigns** or incentives could amplify donations. Additionally, create specialized programs for **non-parents**, who form a significant donor base.

✅Leverage State-Based Campaigns:

California should remain a priority, but also explore the specific characteristics of donors in **Indiana** and **Arkansas**, where high-value donations are coming from despite lower donor counts. Tailored campaigns in these areas could yield high returns.

Review donor cultivation strategy in **Malibu**—focus on quality engagement over volume.Tailored stewardship and recognition for smaller high-impact cities like **Taswell** and **Marietta**.Develop a donor appreciation program for **military** and **diplomatic personnel**—recognizing their unique contributions.

Insights from Data

✅Optimize Donation Tier Strategies:

Focus on increasing the number of donors in **Tier 1 ($1 - $24,999)**, since this tier leads in cumulative donations. **Incentivizing smaller but more frequent donations** could sustain long-term funding.
For **Tier 3 ($50,000 - $99,999)**, efforts should focus on converting these mid-level donors into high-value donors, potentially moving them to Tier 1 for higher cumulative donations.

✅Reconnect with Inactive Top Donors:

Many top donors have not contributed in **recent years**. **Re-engage** these high-value donors with personalized outreach, perhaps showcasing the impact of their past contributions and presenting opportunities for renewed involvement or legacy gifts.

✅Improve Donor Engagement:

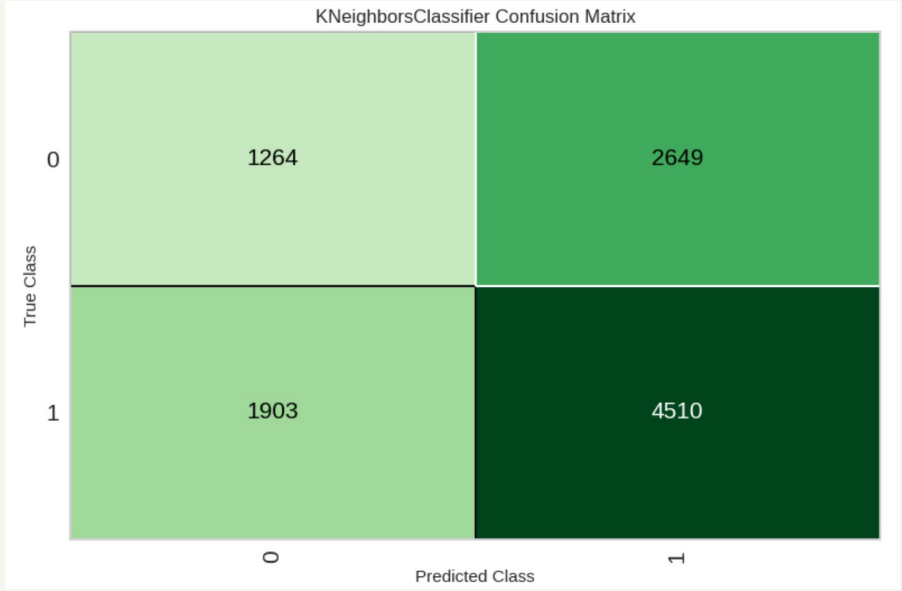
With 76% of donors **not** participating in activities, initiatives to boost engagement (e.g., exclusive events, webinars, or volunteer opportunities) could be a focus. Personal invitations or offers for VIP experiences could incentivize participation.

Since **home** address is the preferred communication channel, consider a **non-digital** (direct mail) strategy for fundraising appeals and event invitations, and complement this with digital outreach when possible. Also, given that **email data** is missing for over two-thirds of donors, **prioritize gathering this information** to enhance **digital engagement** capabilities, allowing for more targeted and cost-effective communication.

Donor Prediction Classification Model Analysis

Pycaret classification model was set up to compare and evaluate all the algorithms to predict the donor indicator

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.5759	0.5411	0.5759	0.5539	0.5592	0.0494	0.0507
1	0.5509	0.5175	0.5509	0.5328	0.5389	0.0073	0.0074
2	0.5529	0.5174	0.5529	0.5374	0.5429	0.0169	0.0170
3	0.5562	0.5303	0.5562	0.5419	0.5471	0.0263	0.0266
4	0.5546	0.5245	0.5546	0.5348	0.5411	0.0113	0.0115
5	0.5712	0.5399	0.5712	0.5543	0.5595	0.0516	0.0523
6	0.5484	0.5097	0.5484	0.5273	0.5340	-0.0040	-0.0041
7	0.5517	0.5143	0.5517	0.5337	0.5397	0.0091	0.0092
8	0.5558	0.5175	0.5558	0.5361	0.5423	0.0139	0.0141
9	0.5799	0.5487	0.5799	0.5601	0.5652	0.0627	0.0641
Mean	0.5598	0.5261	0.5598	0.5412	0.5470	0.0244	0.0249
Std	0.0108	0.0126	0.0108	0.0105	0.0100	0.0212	0.0217



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	K Neighbors Classifier	0.7742	0.8403	0.7742	0.7711	0.7710	0.5087	0.5113

- Among all the evaluated algorithms, '**knn**' was selected considering it having the lowest overall error including Accuracy, AUC, Recall, etc.
- Machine Learning was created using '**knn**' and tested for a 10-fold cross-validation.
- The metrics of the end result was as shown on the right.

Questions?