DATA ANALYSIS PORTFOLIO

PREPARED BY:

Table of Contents

AH

Professional Background

As a certified data analyst, I completed the Udacity Nanodegree program where I acquired skills in data wrangling and data visualization designs using SQL, Python and its libraries such as Pandas, Matplotlib, and Seaborn. Additionally, I also successfully completed the EntryLevel Data Analyst Level 1 course where I learned to analyze data and create visualizations using Excel/Google Sheets and Tableau.

I have worked on a few projects where I demonstrated my ability to analyze data and design visualizations to effectively communicate insights. I am skilled in SQL, Python, Excel/Google Sheets, and Tableau.

Recently, I completed a data visualization internship at Excelerate where I collaborated with my team to analyze data from the marketing team at GlobalShala. Our analysis helped the finance team make cost-saving recommendations. My most recent project is contained in this portfolio, and was completed using SQL and Tableau as part of the requirements for earning my certification of completion for EntryLevel's Data Analyst 2: SQL course.

AH

Introduction

Hypothetically, I am a data analyst working for the charity organization, **Education for All**. I have been asked by the head of fundraising to present the data on donor insights and donation rates.

Within the fundraising team, my objectives are to:
- Increase the number of donors in our database
- Increase the donation frequency of our donors
- Increase the value of donations in our database.

The business problem is that our organization, Education for all, has faced a challenge in meeting our fundraising target for the year.

In two weeks time, I will be presenting insights from our donation data at the next strategy meeting for the following year. The goal is to inform my fundraising strategy and increase donations. The insights will be used to identify areas of improvement in our fundraising efforts and to develop a more effective fundraising strategy.

I used the datasets EFO_Donation_Data and EFO_Donor_Data to answer the business problem. The EFO_Donation_Data contains the following donor data: id,first and last name, email address, gender, job field, donation amount, state of residence (US), and t-shirt size. While the EFO_Donor_Data contains the following donor data: id; frequency of donation; university attended, car make; second language; favorite color; favorite movie genre.

The data cleaning and analysis was done using SQL, and the insights were presented through data visualization designs with Tableau.
I also applied Root Cause Analysis to find answers to the business problem, and sought answers to the following questions as well.
1. For each category of donation frequency, what is the total amount of donations and the total number of donors?
2. What is the sum, the maximum and minimum amount of donations?
3. Which gender donated more and which one has the highest count of donors?
4. What are the top 5 job fields in terms of total donations, and how much has each of these fields donated in total?
5. A list of our top 10 donors full names.
6. For the 10 states with the lowest total donations, what is the total amount of donations and the total number of donors for each donation frequency category?
7. Counts of donors sharing the same favorite movie genre and its first subgenre?

Data Design and Cleaning

The two datasets that were collected for analysis to better understand the business problem, EFO_Donation_Data and EFO_Donor_Data, are available for download as SQL schema files [here](#) and [here](#) respectively.

The EFO_Donation_Data contains the following column headers: id,first name, last name, email address, gender, job field, donation amount, state of residence (US), and t-shirt size. While the EFO_Donor_Data contains the following column headers: id; frequency of donation; university attended, car make; second language; favorite color; favorite movie genre.

These two datasets both contain 1000 rows of data and the id columns being both the primary and foreign keys of each dataset. Hence, they both have no duplicate rows. My cleaning efforts using SQL include the following:

- Manipulating nulls instead of deletion: I used an imputation method to fill in Null values as "Unknown" for all the three columns that have quite some numbers of missing entries. This helped to retain the rows while processing the integrity of the dataset.

```
UPDATE Donor_Data2
SET car = 'Unknown'
WHERE car ISNULL;

UPDATE Donor_Data2
SET university = 'Unknown'
WHERE university ISNULL;

UPDATE Donor_Data2
SET second_language = 'Unknown'
WHERE second_language ISNULL;
```

- I also replaced this value "(no genres listed)" in the `**movie_genre**` column with the value "Unknown" inorder to have a consistent representation for all missing entries.

```
UPDATE Donor_Data2
SET movie_genre = 'Unknown'
WHERE movie_genre = '(no genres listed)';
```

- Lastly, I tidied the EFO_Donor_Data by splitting the different variables, that were combined in a single column `movie_genre`, into 6 different columns namely: `genre`, `subgenre1`, `subgenre2`, `subgenre3`, `subgenre4`, `subgenre5`.
I did the first split with SQL using the following statement:

```
ALTER TABLE Donor_Data2
ADD COLUMN genre TEXT GENERATED ALWAYS AS (CASE when instr(movie_genre, '|') > 0
                                           THEN substr(movie_genre, 1, instr(movie_genre, '|') -1)
                                           WHEN movie_genre = 'Unknown' THEN 'Unknown'
                                           ELSE movie_genre END);
```

The above syntax created the first generated column `genre`. However, due to the high complexity of SQL syntax to achieve the same results for the rest of the 4 columns to be created, I downloaded the table as CSV data, and used Excel's `Text to Column` tool.

H	I	J	K	L	M
genre	subgenre1	subgenre2	subgenre3	subgenre4	subgenre5
Comedy	Drama	Musical	Unknown	Unknown	Unknown
Drama	Unknown	Unknown	Unknown	Unknown	Unknown
Comedy	Romance	Unknown	Unknown	Unknown	Unknown
Action	Drama	Mystery	Sci-Fi	Thriller	IMAX
Children	Comedy	Fantasy	Romance	Unknown	Unknown
Action	Adventure	Romance	Thriller	Unknown	Unknown

Finally, after the data was successfully cleaned, I continued Exploratory Data Analysis (EDA) in **SQLite Online**.

AH

Root Cause Analysis

The business problem as defined earlier is that our organization, **Education for al**l, has faced a challenge in meeting our fundraising target for the year. Within the fundraising team, my objectives are to:
- Increase the number of donors in our database,
- Increase the donation frequency of our donors, and
- Increase the value of donations in our database.

In order to find the root cause of this business problem, I used the **5 Whys**' approach also known as the Root Cause Analysis.

1. **Q: "Why didn't we meet our fundraising target for the year?"**
   A: We received a very low sum of donations from some states. Specifically, 32 states had donation sums that were below the average - 5083.37 USD.

2. **Q: "Why did we have lesser donations from some states?"**
   A: This was due to a low number of donors in those states.

3. **Q: "Why are there few donors in those states?"**
   A: Our fundraising team's outreach and advertising efforts have not been effective enough in these states.

In summary, the underlying cause of this issue was found to be that our organization's fundraising campaigns and solicitation efforts were not effectively reaching potential donors in some states.

A chart showing all the states and their respective sum of donations, along with the other insights found, is presented in the [Visualization Chapter](#).

AH

Queries

Question 1

For each category of donation frequency, what is the total number of donors, the total amount donated, and the average amount donated?

Query 1:

```sql
SELECT d2.donation_frequency,
       COUNT(d2.id) donor_counts,
       SUM(d1.donation) total_donation,
       ROUND(AVG(d1.donation),2) average_donation
FROM Donation_Data d1
JOIN Donor_Data2 d2
ON d1.id = d2.id
GROUP BY 1
ORDER BY 3 DESC;
```

Result:

donation_frequency	donor_counts	total_donation	average_donation
Yearly	259	65667	253.54
Once	264	64586	244.64
Monthly	232	59680	257.24
Weekly	245	59152	241.44

Question 2

What is the sum, the maximum and minimum amount of donations?

Query 2:

```
SELECT SUM(donation) sum_donation,
       MAX(donation) max_donation,
       MIN(donation) min_donation
FROM Donation_Data;
```

Result:

sum_donation	max_donation	min_donation
249085	500	5

Question 3

Which gender donated more and which one has the highest count of donors?

Query 3:

```
SELECT gender,
       SUM(donation) total_donation,
       COUNT(gender) gender_count
FROM Donation_Data
GROUP BY 1;
```

Result:

gender	total_donation	gender_count
Female	121457	508
Male	127628	492

Question 4

What are the top 5 job fields in terms of total donations, and how much has each of these fields contributed?

Query 4:

```sql
SELECT job_field,
       SUM(donation) total_donation
FROM Donation_Data
GROUP BY 1
ORDER BY 2 DESC
LIMIT 5;
```

Result:

job_field	total_donation
Human Resources	23060
Research and Development	22862
Product Management	22798
Business Development	22266
Engineering	21968

AH

Question 5

A list of our top 10 donors' full names?

Query 5:

```
SELECT first_name || ' ' || last_name AS full_name,
       gender,
       job_field,
       state,
       donation
FROM Donation_Data
ORDER BY 5 DESC
LIMIT 10;
```

Result:

full_name	gender	job_field	state	donation
Beverlie Andriesse	Male	Support	Michigan	500
Wallie Leather	Male	Product Management	New York	500
Clevie Camilletti	Female	Legal	Virginia	499
Peder Rilton	Female	Sales	Delaware	499
Worthy Le feaver	Male	Sales	Wisconsin	498
Amalea Knill	Male	Research and Development	New York	497
Tonnie Stockney	Male	Support	California	494
Corbett Lansdale	Female	Product Management	California	494
Nathaniel McGenn	Male	Human Resources	California	494
Beverlee Camacke	Male	Product Management	Maryland	493

AH

Question 6

For the 10 states with the lowest total donations, what is the total amount of donations and the total number of donors for each donation frequency category?

Query 6:

```sql
WITH lowest_donor_states AS (
    SELECT state, SUM(donation) total_donation
    FROM Donation_Data
    GROUP BY 1
    ORDER BY 2
    LIMIT 10
)

SELECT d2.donation_frequency,
       COUNT(d2.id) donor_counts,
       SUM(d1.donation) total_donation
FROM Donation_Data d1
JOIN Donor_Data2 d2
ON d1.id = d2.id
JOIN lowest_donor_states lds
ON d1.state = lds.state
GROUP BY 1
ORDER BY 3 DESC;
```

Result:

donation_frequency	donor_counts	total_donation
Once	13	2817
Weekly	5	1443
Yearly	7	1249
Monthly	6	1104

Question 7

Counts of donors sharing each unique pair of favorite movie genre and immediate subgenre?

Query 7:

```sql
SELECT DISTINCT genre,
                subgenre1,
                counts donor_counts
FROM (SELECT genre,
             subgenre1,
             COUNT(subgenre1) OVER (PARTITION BY genre, subgenre1) counts
      FROM Donor_Data2
     ) like_donors
-- filtering out the donors with unknown genres and subgenres
WHERE genre != 'Unknown' AND subgenre1 != 'Unknown'
ORDER BY 3 DESC;
```

Result:

| genre	subgenre1	donor_counts
Comedy	Drama	77
Crime	Drama	43
Comedy	Romance	35
Drama	Romance	35
Action	Adventure	31
Comedy	Crime	24
Drama	War	23
Action	Comedy	20

The above is just an excerpt of the query result, the list is quite long having 73 unique rows.

Visualizations

Root Cause Analysis Charts

The charts here help to visualize and understand the root cause analysis process I did previously.

Total Donations ($) Per State
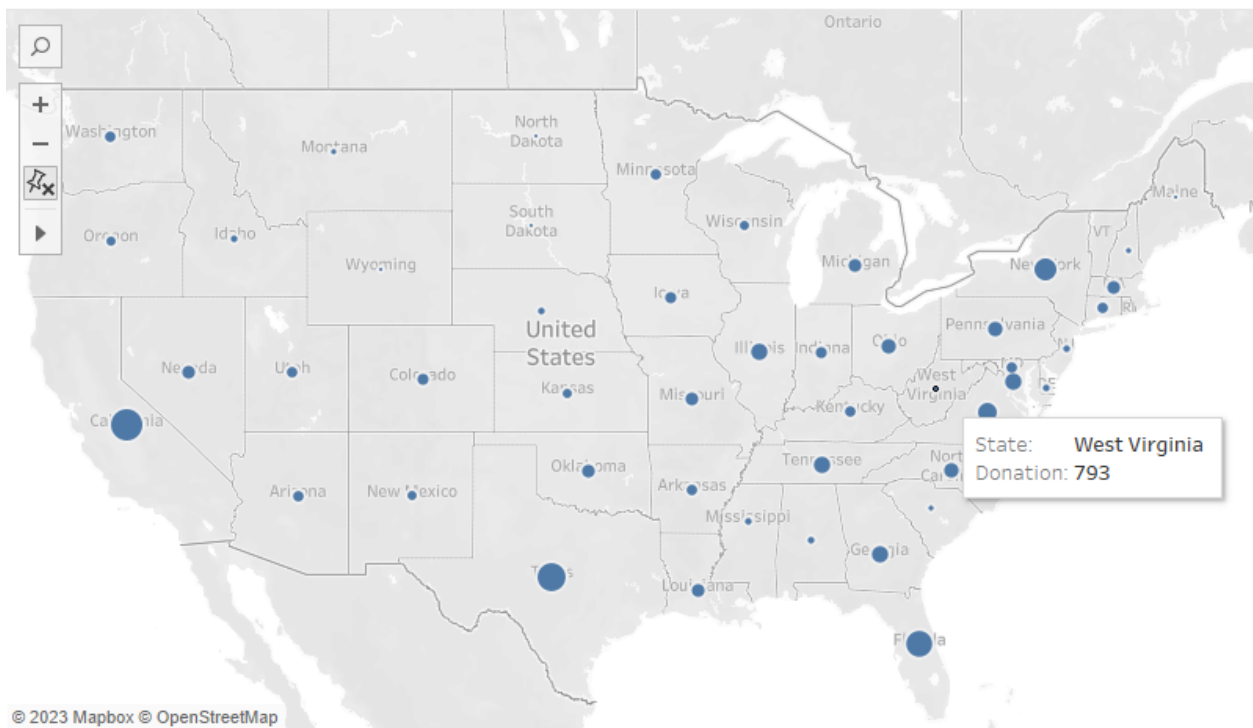
This map chart shows the distribution of the total amount of donations received from donors in each state in the USA. There are about 32 states in the chart whose sum of donations is below the average - 5083.37 USD.

Total Count of Donors Per State
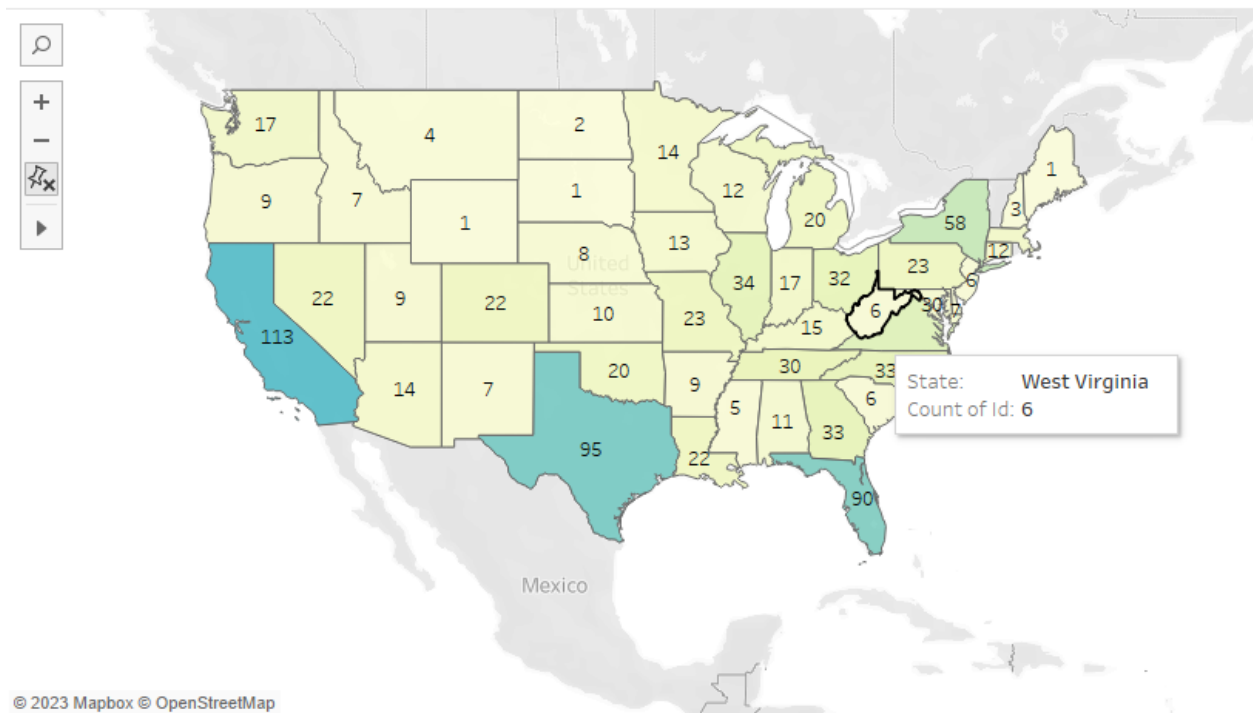
The chart above shows the total count of donors we currently have in each state. Taking West Virginia as a case study: we have just 6 donors there, and a sum donation of $793 from the previous chart. The count of donors has a direct relationship with the total amount received, that is, the higher the donor count; the higher the total amount received, and vice versa. This relationship is displayed in the following scatterplot.
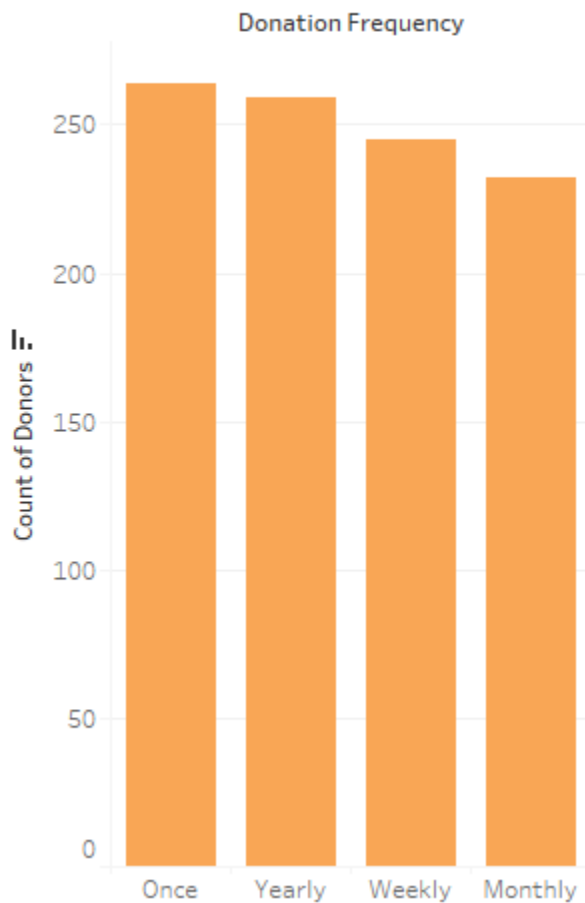
Count of Donors vs Sum of Donations Per State
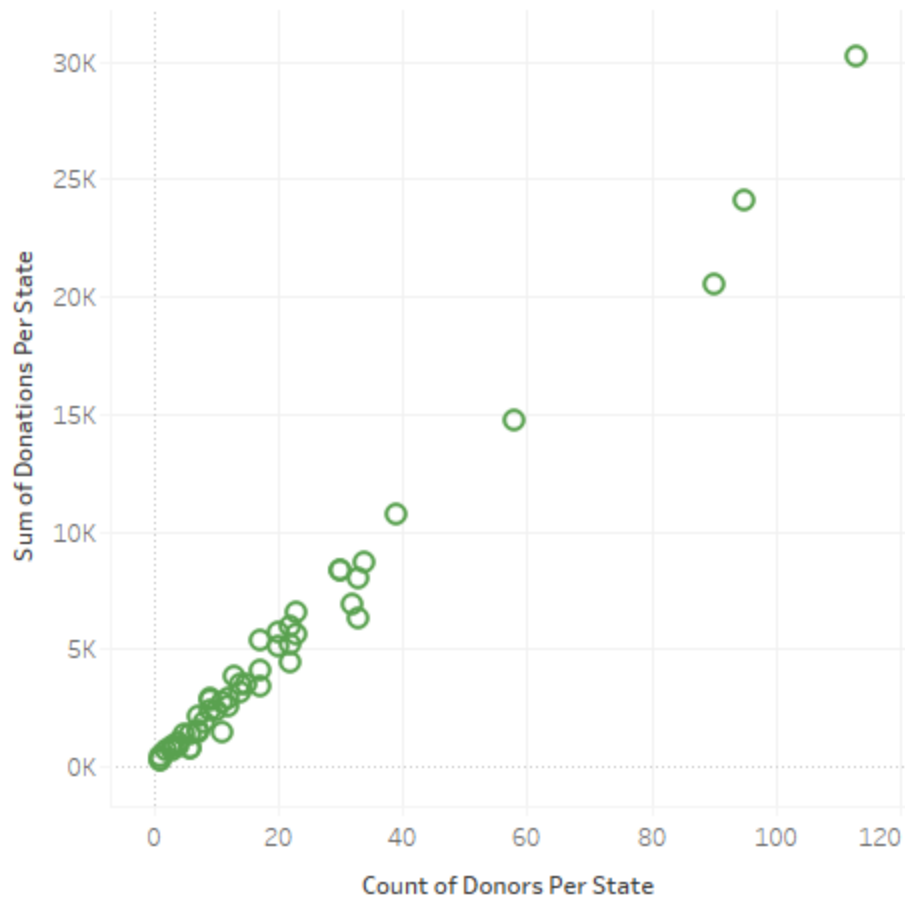
Data Insights

Finding 1

The total number of donors, the total amount donated, and the average amount donated for each category of donation frequency.

### Count of Donors By Their Donation Frequency.

Donors who made a one time donation are the highest in count, followed by the yearly donors, the weekly donors and the monthly donors in that order.

Sum of Donations For Each Donation Frequency

Donation Frequency
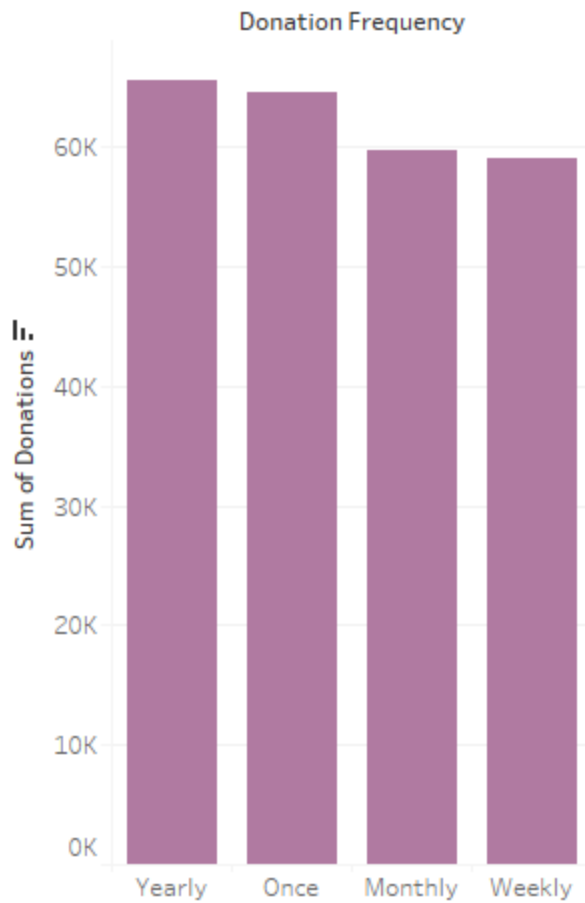
Interestingly, the yearly donors have donated slightly more than the one time donors, and similarly, the monthly donors have a higher total amount of donations than the weekly donors. Overall, the yearly and one time donors together are more both in number and total donations than the monthly and weekly donors combined.

AH

Average Donations By Frequency.

Donation Frequency
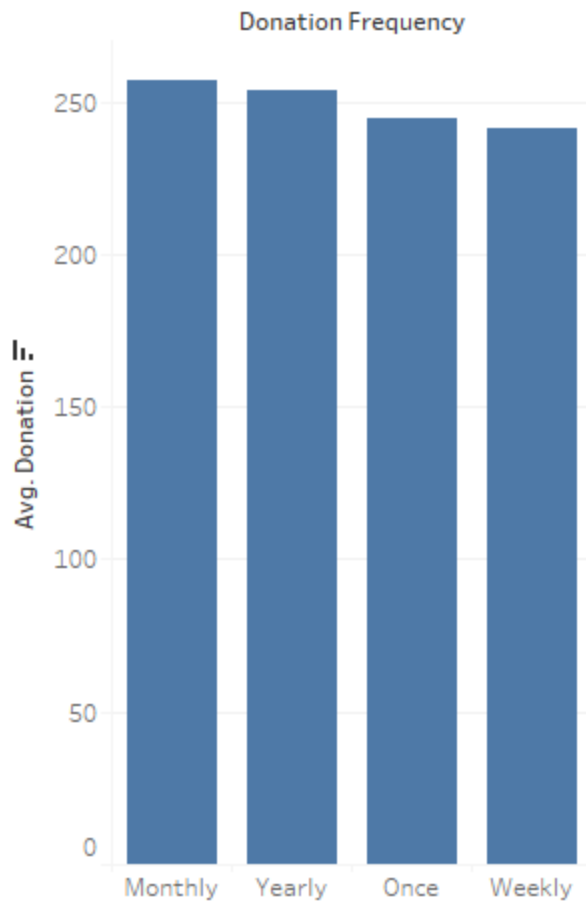
There is a slight difference in averages with the highest value for the monthly donors, then the yearly, the one time and the weekly donors in descending order.
The insights observed so far shows that the donation frequency adopted by any of our donors has not been a cause of the business problem.

Finding 2

The sum, the maximum and minimum amount of donations.

⠿ sum_donation	max_donation	min_donation
249085	500	5

The total amount received from all donations by the end of this year is $249,085. The maximum amounts and minimum amounts received from any single donor are $500 and $5 dollars respectively. This maximum amount was donated each by two donors, one from Michigan and the other from New York. And the minimum amount was also donated by two donors, one from Alabama and the other from Oklahoma.
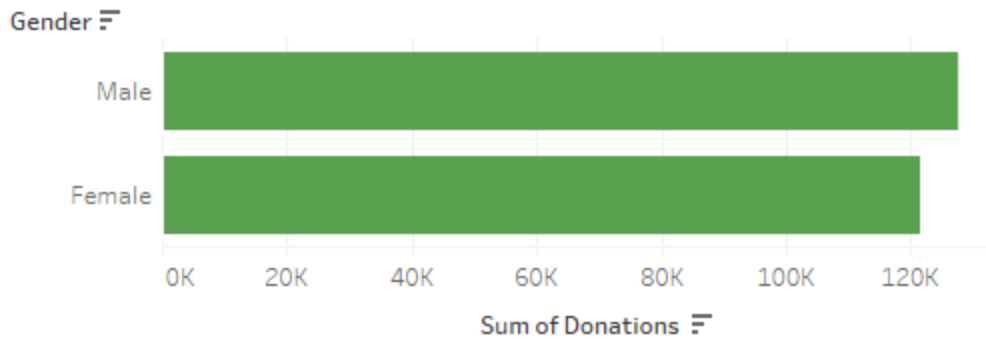
Finding 3

Analysis of donation and donor counts by gender.

Count of Donors by Gender
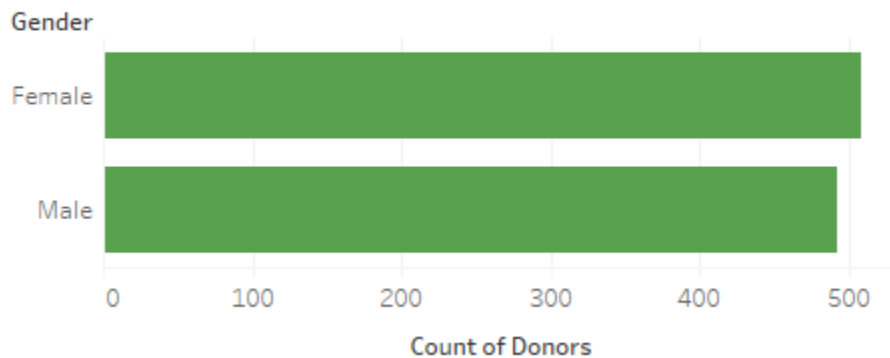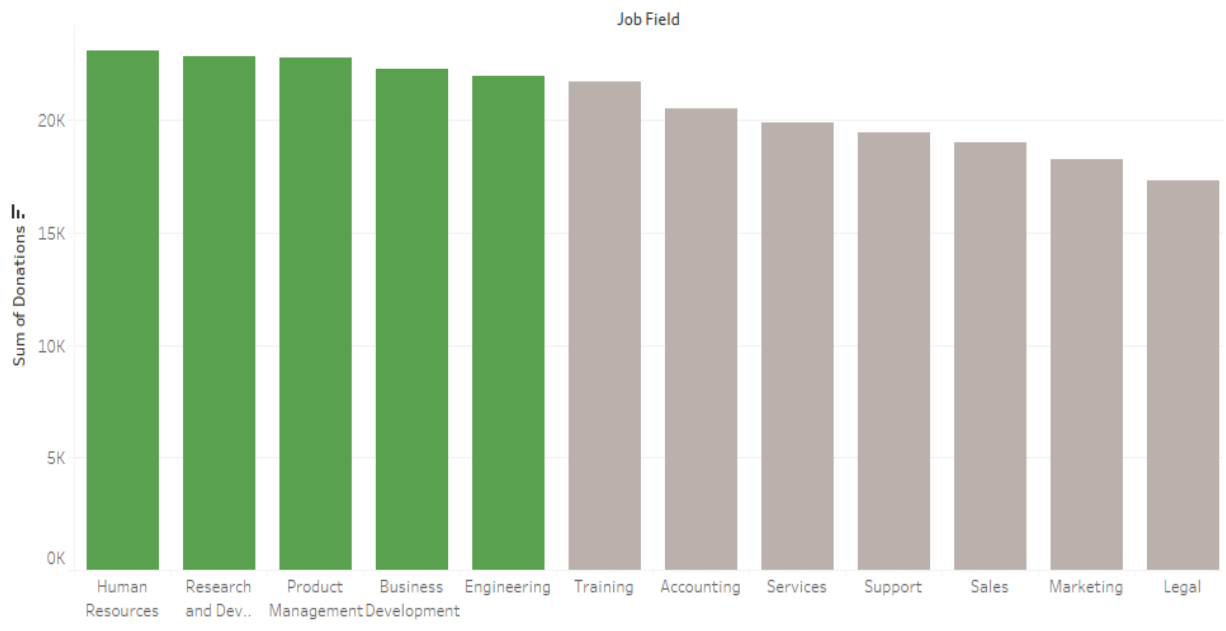
Sum of Donations by Gender

Although the females were more than the males with a count difference of 16, the male donors took the lead by a margin amount of $6,171.

Finding 4

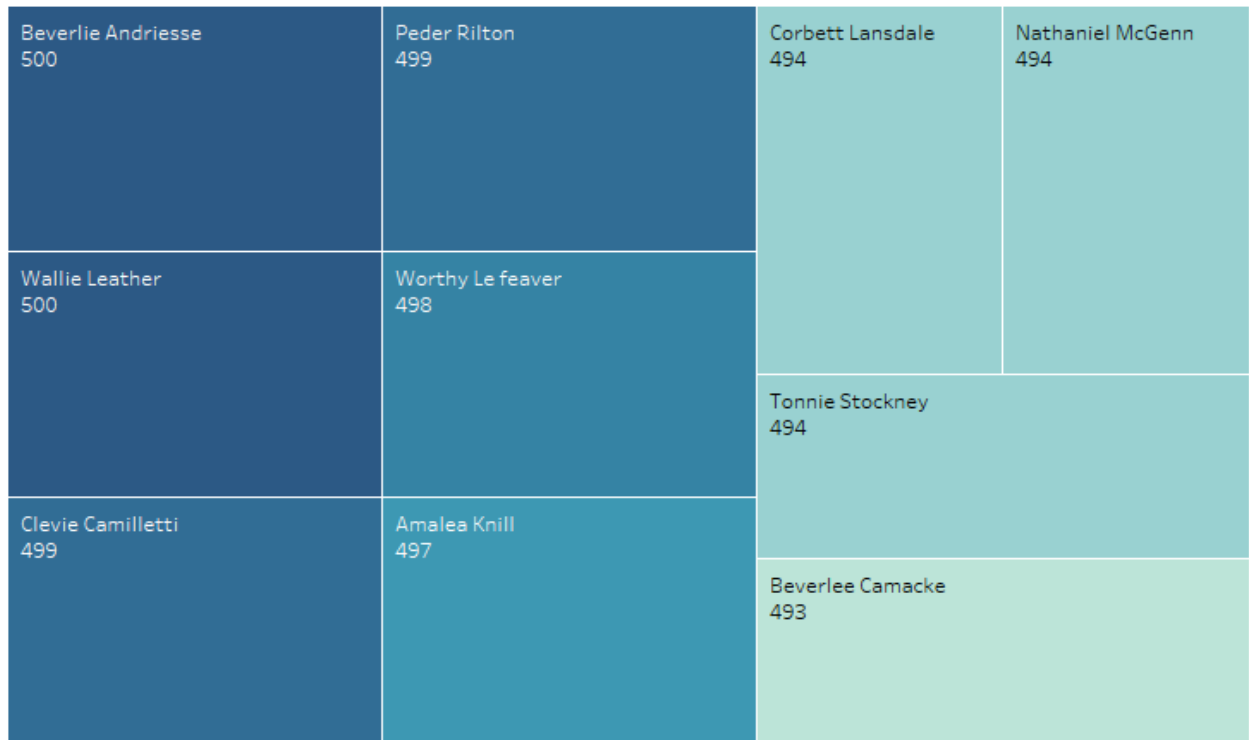The top 5 job fields in terms of total donations.

By grouping the sum of donations by job fields, the top 5 are Human Resources, Research and Development, Product Management, Business Development, and Engineering. Though the donors in these listed job fields have made the highest financial contributions, other job fields have performed well too.

Finding 5

The top 10 donors and the amounts they donated each.

Top 10 Donors

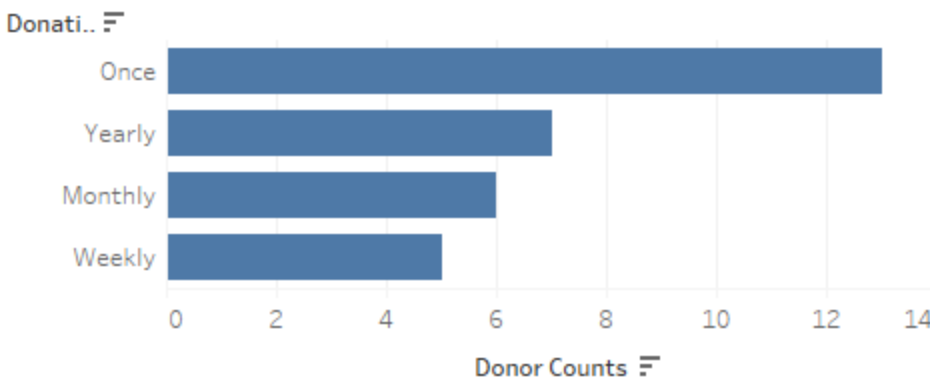| Beverlie Andriesse 500	Peder Rilton 499	Corbett Lansdale 494	Nathaniel McGenn 494
Wallie Leather 500	Worthy Le feaver 498	Tonnie Stockney 494	
Clevie Camilletti 499	Amalea Knill 497	Beverlee Camacke 493	

Finding 6

The total amount of donations and the total number of donors for each donation frequency category, with regard to the ten states having the least sum of donations.

Sum of Donations For Each Donation Frequency

Donati.. ≡

Once	
Weekly	
Yearly	
Monthly	

0   500   1000   1500   2000   2500

Total Donation ≡

AH
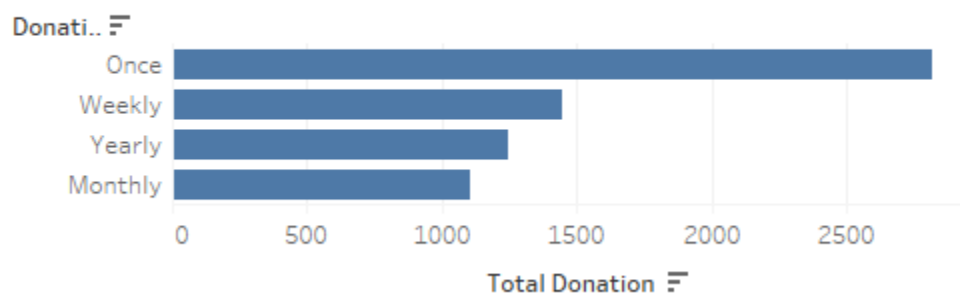
Count of Donors by Donation Frequency

Donati.. ⫶

Donor Counts ⫶

Just as we have seen before in Finding 1, and also in this analysis of donors within the ten states that have the least sum of contributions, the majority of donors are either one-time or yearly contributors.

Summary of Findings

Here are the summary of insights that can help to solve the business problem:

- ☑ The underlying cause of the business problem, using the root cause analysis, was found to be that our organization's fundraising campaigns and solicitation efforts were not effectively reaching potential donors in some states.
- ☑ The majority of our donors are either one-time or yearly contributors, most especially within the states where we received very low amounts of donations.
- ☑ The total amount received from all donations by the end of this year is $249,085. The maximum and minimum of donations received from a single donor are $500 and $5 respectively.
- ☑ Although the females were more than the males, the sum of donations received from male donors was more than that received from female donors.
- ☑ All the donors from each job field have collectively contributed almost equally across the fields.
- ☑ Lastly, all the selected top 10 donors donated not less than $493.

AH

Recommendations

1. We should utilize multiple platforms to reach out to potential donors in the states where we have not received much donations.We can engage them via email campaigns, text message campaigns, and social media platforms like Facebook, Instagram, and LinkedIn.
2. We should also optimize the donation process for mobile devices to make it easier for donors to give on the go.
3. We should increase our efforts in keeping our donors engaged and interested in supporting our cause.
4. We can host events to appreciate our donors or to encourage larger donations in person.

These recommendations can be adopted as a strategy to help solve the business problem, and meet our objectives.