Bryan D'Amico
IST652-Scripting for Data Analysis
Homework 1 Report: Structured Data

Analysis of the Donors Data

Data Set

The data set used for this assignment was the donors data set provided in the 2U learning environment. The donors data contains 3210 rows and 24 columns of demographic and donation information on donors to a charity. The following table provides the original name of the variables found in the data, the new name that was created during the data cleaning process or if that variable was removed, and a description of the information contained in that variable. All variable descriptions were sourced from the data dictionary provided alongside the donors data in the 2U learning environment.

Variable Name	New Name	Description
Row Id	*Removed from Data*	Index
Row Id.	*Removed from Data*	ID number
zipconvert_2	Region_2	1 if living in Region 2, 0 if not
zipconvert_3	Region_3	1 if living in Region 3, 0 if not
zipconvert_4	Region_4	1 if living in Region 4, 0 if not
zipconvert_5	Region_5	1 if living in Region 5, 0 if not
homeowner dummy	Homeowner	1 if a homeowner, 0 if not
NUMCHLD	Num_Children	Number of children
INCOME	Houshold_Income	Seven categories representing household income
gender dummy	Gender	1 if Female, 0 if Male
WEALTH	Wealth_Rating	Based on median family income and pop. statistics from each area. 9 is highest wealth group, 0 is lowest wealth group
HV	Avg_Home_Value	Average home value in potential donor's neighborhood in $ thousands
Icmed	Median_Family_Income	Median family income in potential donor's neighborhood in $ thousands
Icavg	Avg_Family_Income	Average family income in potential donor's neighborhood in $ thousands
IC15	Low_Income_Pct	Percent categorized as "low income" in potential donor's neighborhood
NUMPROM	Num_Promotions	Lifetime number of promotions received to date
RAMNTALL	Lifetime_Gifts	Dollar amount of lifetime gifts to date
MAXRAMNT	Largest_Gift	Dollar amount of largest gift to date
LASTGIFT	Most_Recent_Gift	Dollar amount of most recent gift
totalmonths	Months_Since_Donation	Number of months since the last donation
TIMELAG	Months_Btw_Gift1_Gift2	Number of months between first and second gift
AVGGIFT	Avg_Gift_Amt	Average dollar amount of gifts to date
TARGET_B	*Removed from Data*	Classification response variable
TARGET_D	*Removed from Data*	Prediction response variable

Description of Data Exploration and Cleaning Steps

Cleaning Steps

The donors data set was imported in the form of a pandas DataFrame using pd.read_csv. After creating the DataFrame, the variables "Row Id", "Row Id.", "TARGET_B", and "TARGET_D" were dropped since they would not be used to answer any of the questions generated. Many of the variable names in the data set were

confusing without the data dictionary, so they were renamed to be more approachable for someone else viewing the program. Lastly, the DataFrame was checked for any missing values, of which there were none.

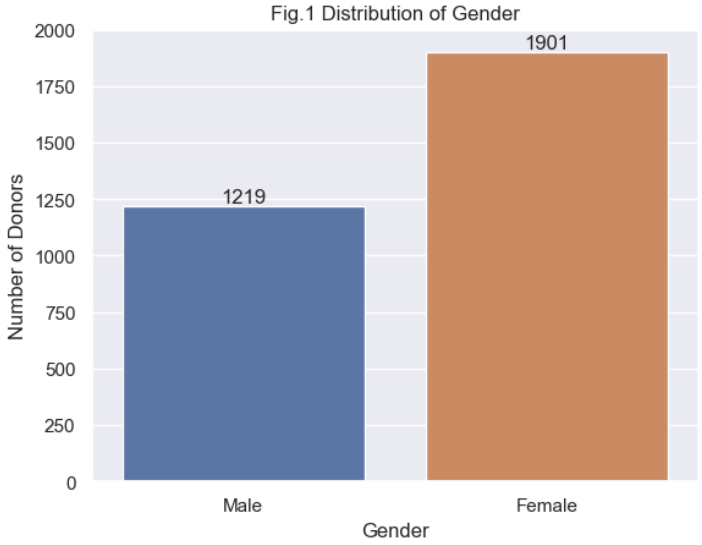
Exploratory Data Analysis

The exploratory data analysis focused on generating and visualizing the distributions of the various demographic metrics contained in the data set. These distributions gave some insight into the data set and helped guide the development of questions based on what variables would be the most useful in terms of creating groupings within the data.

Distribution of Gender

The gender variable represents a female donor with a 1 and a male donor with a 0. The data set contains significantly more female donors than it does male donors.

	Gender	Count
0	1	1901
1	0	1219



Distribution of Wealth Ratings

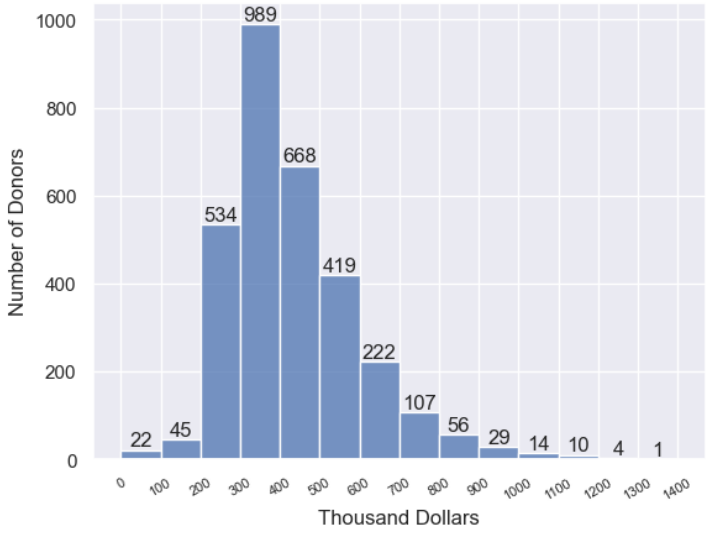
There are 10 different wealth ratings, with 0 representing the lowest wealth group and 9 representing the highest wealth group. These wealth ratings are calculated based on median family income and population statistics calculated from each area where the donors live. The number of donors with each wealth rating is fairly consistent, with the exception of those with a wealth rating of 8. That group has contains about ten times as many donors as any of the other groups.

	Wealth_Rating	Count
9	0	112
8	1	138
7	2	138
5	3	162
6	4	153
2	5	186
4	6	162
3	7	180
0	8	1700
1	9	189

Distribution of Average Family Income

This variable contains the average family income for the neighborhood where the donor lives recorded in thousands of dollars. The great majority of donors reside in very wealthy neighborhoods, with average family incomes between $200,000 and $700,000 per year. There are very few donors living in neighborhoods with average family incomes that are less than $100,000 or more than $1,000,000.

Fig.3 Distribution of Average Family Income

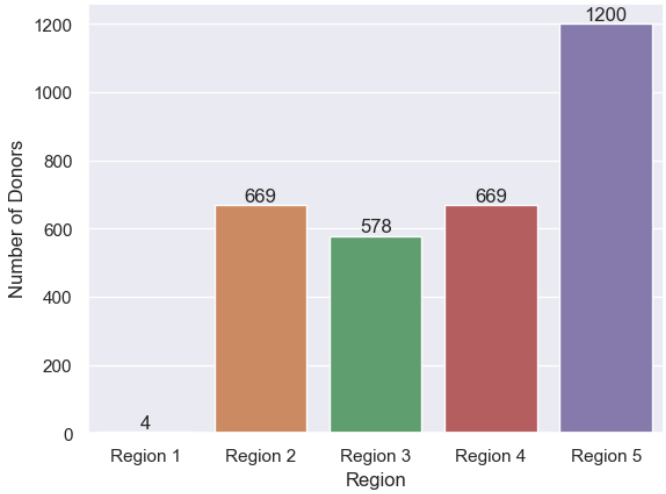


Distribution of Donors by Region

The region variables contain a 1 if the donor lives in that region, and 0 if not. If none of the region variables contain a 1 for a donor, then that donor lives in region 1 which is not listed as a variable. Since this information is spread across four variables, the data was summarized to show the distribution more clearly. Region 1 contains almost no donors, while region 5 contains twice the number of donors as regions 2, 3, or 4.

	Region	Number of Donors
0	Region 1	4
1	Region 2	669
2	Region 3	578
3	Region 4	669
4	Region 5	1200

Fig.4 Distribution of Donors by Region



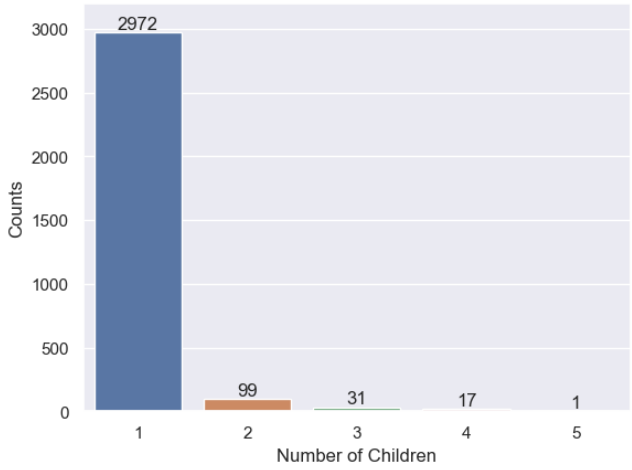
Distribution of Number of Children

The overwhelming majority of the donors' families have only one child. This distribution is too skewed to draw accurate conclusions from groupings based on the number of children. Since there are almost 3,000 families with one child, it would not make sense to compare the "one child" group with the "five children"

group that would only contain 1 donor family. This variable will not be used to create groups in the data analysis.

	Number_of_Children	Counts
0	1	2972
1	2	99
2	3	31
3	4	17
4	5	1

Fig.5 Distribution Number of Children Present in Donor Families

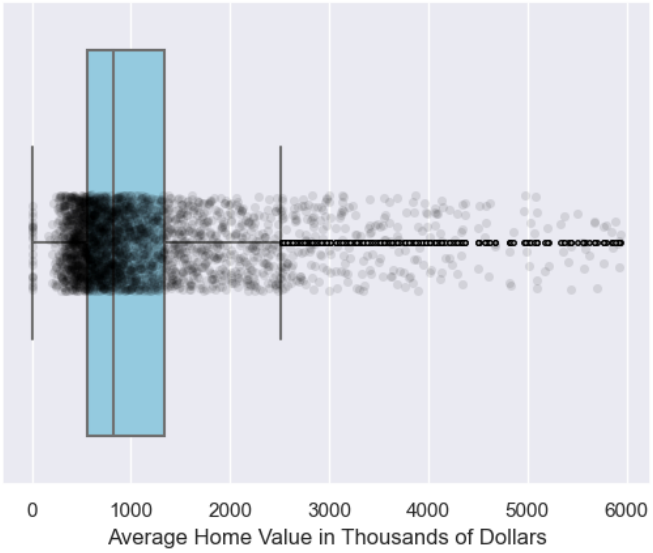


Distribution of Average Home Values

The distribution of average home values shows how different the property values are for the neighborhoods where the donors live. The median average home value is $822,000, while the mean average home value is $1,141,361.86. This implies that the mean of all the average home values is being affected by the presence of outliers. This is confirmed with a maximum average home value of $5,945,000. This means there is a donor living in a neighborhood where the average home value is almost $6 million. Since the mean is being skewed larger by the presence of outliers, the median is a better representative for a measure of central tendency.

```
The mean average home value is $1,141,361.86
The min average home value is $0.00
The 25% average home value is $556,000.00
The 50% average home value is $822,000.00
The 75% average home value is $1,338,750.00
The max average home value is $5,945,000.00
```

Fig.6 Distribution of Average Home Values

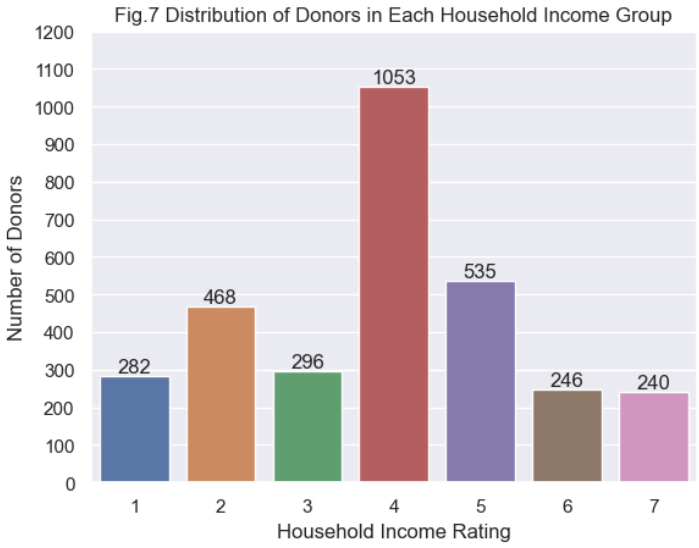


Also, there are a few rows where the minimum average home value is $0. These rows were isolated and examined, and there were a few other variables in these rows that also contained a value of 0. This may indicate that data was not available for these variables, so they were recorded as 0. Without more information, however, it is not possible to properly cleanse these values.

Distribution of Household Income Ratings

All but one of the household income rating groups are fairly consistent in size, similar to the distribution of wealth ratings. There are over 1,000 donors with a household income rating of 4. While this group does contain about one-third of the data, there are enough members of the other groups to still be able to calculate summary statistics and make comparisons.

	Household_Income	Counts
4	1	282
2	2	468
3	3	296
0	4	1053
1	5	535
5	6	246
6	7	240



Questions

Five questions were developed based on this data set.

Question 1: What is the median average home value of the donors grouped by each wealth rating?

Unit of Analysis: Wealth Ratings

Comparison: For each wealth rating, compute the median average home values of the neighborhoods where the donors reside

Question 2: How do the average home values of the donors' neighborhoods compare by gender?

Part 1

Unit of Analysis: Gender

Comparison: For each gender, compute the median average home values of the neighborhoods where the donors reside.

Part 2

Unit of Analysis: Wealth Rating by Gender

Comparison: For each combination of gender and wealth rating, compute the median average home values of the neighborhoods where the donors reside.

Question 3: What regions are the most profitable in terms of donations?

Unit of Analysis: Region

Comparison: For each region, compute the mean lifetime gift amount, the mean largest gift amount, the mean average gift amount, and the mean months since the last donation.

Question 4: How does gender and household income rating correlate with the amount and/or frequency of donations?

Unit of Analysis: Household Income Rating by Gender

Comparison: For each combination of household income rating and gender, compute the mean number of months between the first two gifts, the mean lifetime gift amount, the mean largest gift amount, and the mean average gift amount.

Question 5: How does the average family income correlate with the amount and/or frequency of donations?

Unit of Analysis: Average Family Income

Comparison: The average family income was plotted against the lifetime gift amounts, the largest gift amounts, the months between the first two gifts, and the average gift amount to see if there was a positive or negative correlation between each pairing of variables.

Description of the Program

The program was created in the form of a report style Jupyter notebook. It guides the reader through the data importing, cleaning, and exploration steps previously described. The program then answers each of the five questions by generating data frames that contain the information needed to answer the question and visualizing the results for clarity. The program also writes files based on generated data frames that would be helpful for someone who wanted to record the results that answer the questions. The median average home values are calculated for each wealth rating and then visualized with a regression plot and a supplementary box plot. The median average home values are calculated for each gender, then for each group of gender and wealth rating. These are visualized with a boxplot and a side-by-side box plot to demonstrate the differences between the various groups. Since region 1 only had 4 donors, the means of various donation statistics were calculated for regions 2 through 5 and visualized with a series of bar plots. To supplement this data, the total lifetime gift amount and total number of promotions were also calculated by region. The means of various donation statistics were also calculated for each group of gender and household income. This data was visualized using a series of side-by-side box plots. Lastly, a series of scatter plots were created to identify any possible correlations between average family income and the amount or frequency of donations.

Output and Analysis

Question 1: What is the median average home value of the donors grouped by each wealth rating?

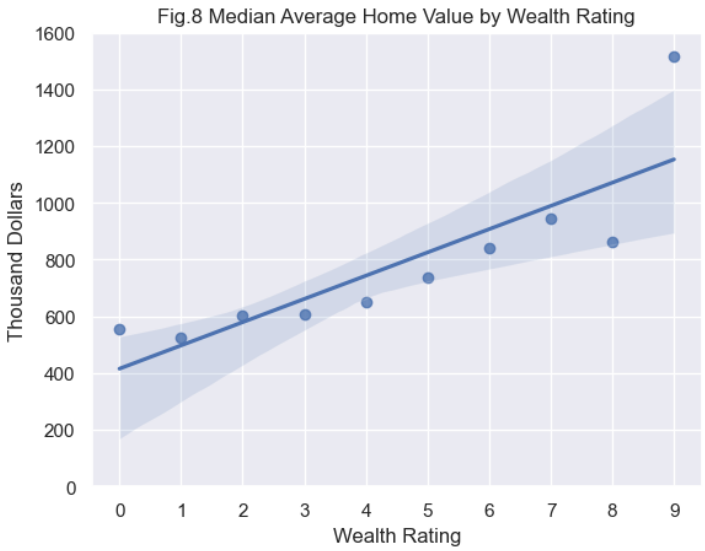
The program groups the donors by each wealth rating, calculates the median average home value for each group, and prints the result as a series of statements. A tabular version is also written to a separate file.

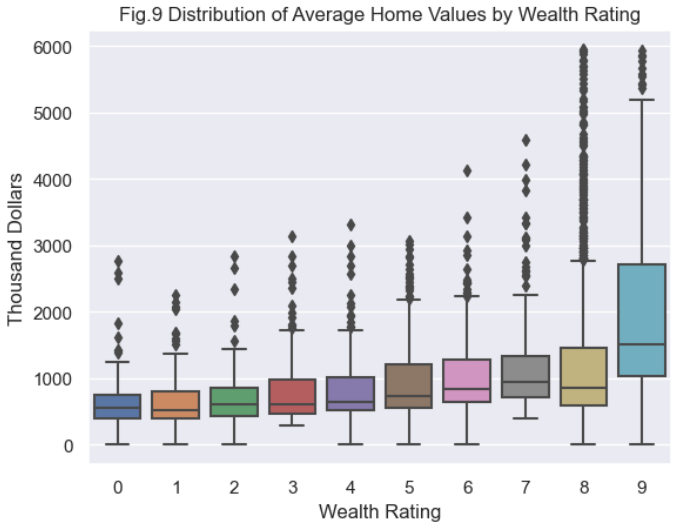
```
The median average home value for donors with a wealth rating of 0 is $555,000.00
The median average home value for donors with a wealth rating of 1 is $525,500.00
The median average home value for donors with a wealth rating of 2 is $600,500.00
The median average home value for donors with a wealth rating of 3 is $606,500.00
The median average home value for donors with a wealth rating of 4 is $651,000.00
The median average home value for donors with a wealth rating of 5 is $736,500.00
The median average home value for donors with a wealth rating of 6 is $839,500.00
The median average home value for donors with a wealth rating of 7 is $946,000.00
The median average home value for donors with a wealth rating of 8 is $863,500.00
The median average home value for donors with a wealth rating of 9 is $1,516,000.00
```

HomeValueByWealthReport

Wealth Rating	Median Average Home Value ($ thousands)
0	555.0
1	525.5
2	600.5
3	606.5
4	651.0
5	736.5
6	839.5
7	946.0
8	863.5
9	1516.0

The median average home values increase as the wealth rating of the donors increases. The median average home value of donors with a wealth rating of 9 was significantly greater than the median average home value of any other wealth rating group. The median average home value for a wealth rating of 9 was greater than the 75th percentile for any of the other wealth rating groups. There are also two instances where the median average home value decreases when increasing to the next wealth rating. These occur at wealth ratings of 1 and 8.



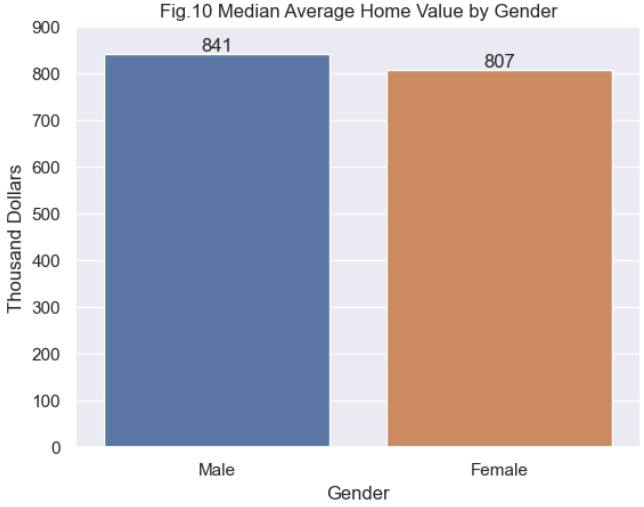


Question 2: How do the average home values of the donors' neighborhoods compare by gender?

The overall median average home value is greater for male donors than for female donors.

	Gender	Median_Avg_Home_Value
0	0	841.0
1	1	807.0

Fig.10 Median Average Home Value by Gender

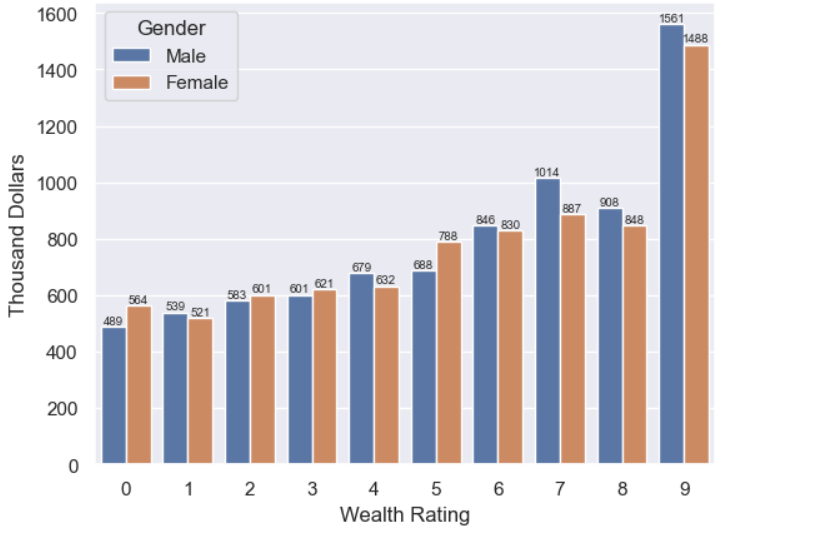


The two gender groups were then broken down into wealth rating groups to both link this question to the first, and to see if any additional information or patterns could be uncovered. The data frame created to answer this question was written to a separate file. For most of the wealth ratings, the median average home value was still greater for male donors than for female donors. However, for donors with wealth ratings of 0, 2, 3, and 5, female donors have a greater median average home value than male donors.

HomeValueByWealthAndGenderReport

Gender	Wealth Rating	Average Home Value ($ thousands)
Male	0	489.5
Male	1	539.0
Male	2	583.0
Male	3	601.0
Male	4	679.0
Male	5	688.5
Male	6	846.5
Male	7	1014.5
Male	8	908.5
Male	9	1561.0
Female	0	564.0
Female	1	521.0
Female	2	601.0
Female	3	621.5
Female	4	632.5
Female	5	788.0
Female	6	830.0
Female	7	887.0
Female	8	848.0
Female	9	1488.5

Fig.11 Difference in Median Average Home Value by Gender in Each Wealth Rating Group



Question 3: What regions are the most profitable in terms of donations?

The answer to this question will allow the charity to reduce costs by better understanding which area to focus on in terms of gathering donations. From the exploratory data analysis, it was already seen that Region 1 only had 4 donors. This region is clearly not the most profitable. There could be something about this region that does not resonate with the cause of the charity. Therefore, the regions that are already actively engaging in the charity were the focus. Regions 2, 3, 4, and 5 were compared in terms of their mean lifetime gifts, mean largest gift, mean average gift amount, and the mean number of months since the last donation. The data frame containing these values was written to a separate file.

RegionStatsReport

Region	Mean Lifetime Gifts ($)	Mean Largest Gift ($)	Mean Avg Gift Amount ($)	Mean Months Since Donation
Region_2	102.77	16.01	10.61	31.201793721973100
Region_3	106.86	16.69	9.73	30.951557093425600
Region_4	118.65	15.98	10.3	31.328849028400600
Region_5	111.88	17.38	11.43	31.088333333333300

Figure 12a displays the mean lifetime gift and mean largest gift amounts for each region. Region 4 has the greatest mean lifetime gift amount. The donors in this region have contributed the most to the charity overall on average. Region 2 has the least mean lifetime gift amount. Region 5, however, has the greatest mean largest gift amount, surpassing region 4 by $1.40. So, although region 4 gives more money in total over time on average, region 5 gives the greatest single donation values on average. Interestingly, region 4 has the lowest mean largest gift amount. This suggests that the donors in region 4 must give smaller donations more frequently.

Fig.12a Gift Statistics for Each Region

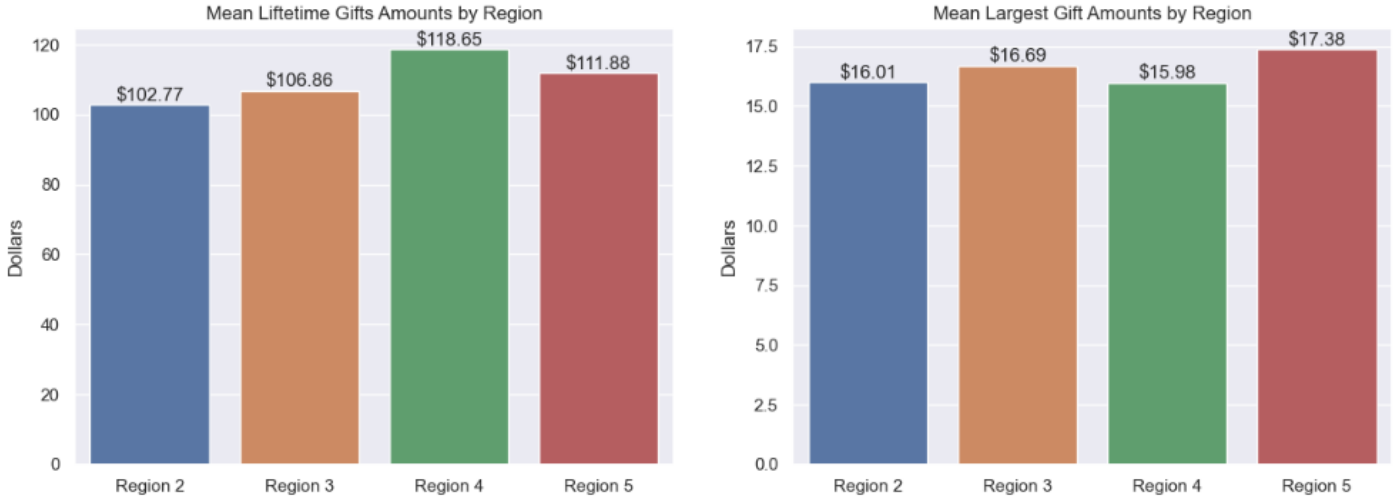
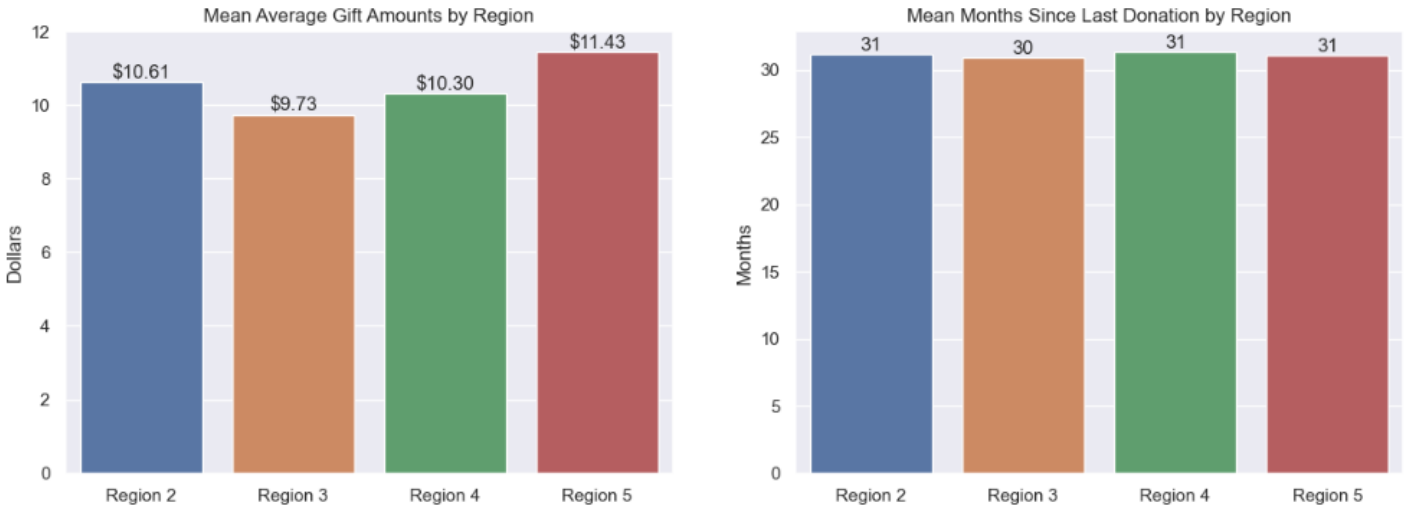


Figure 12b reveals that region 5 contributes the greatest donation amounts on average, while region 3 contributes the smallest donation amounts on average. Also, on average, there does not seem to be much of a difference in the time since a donor gave their last donation in each region. All the regions have an average between 30 and 31 months.

Fig.12b Gift Statistics for Each Region

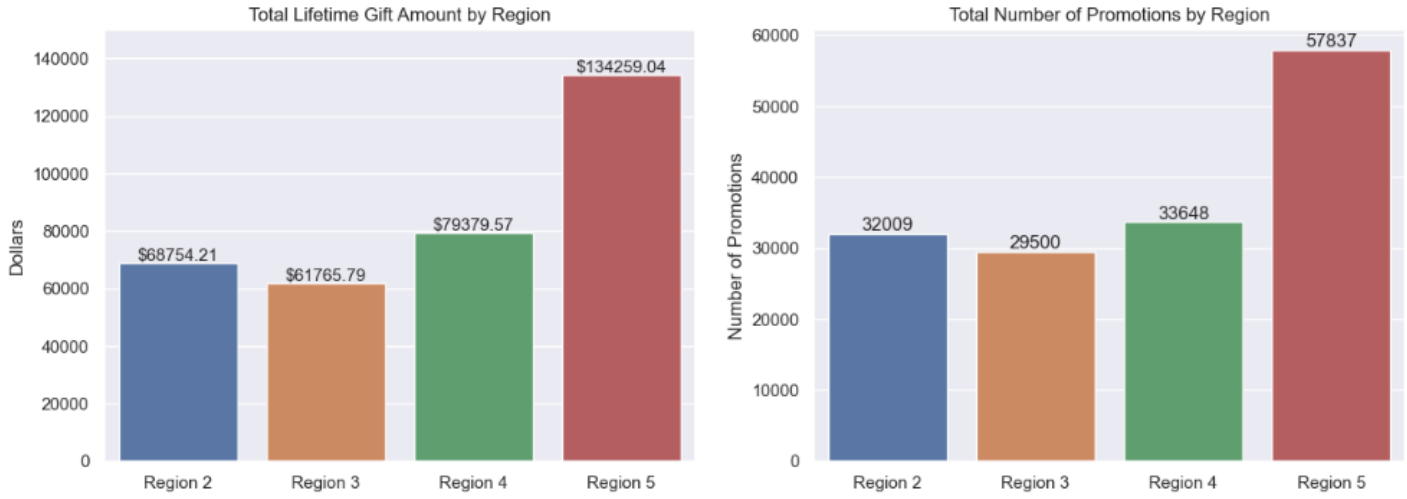


Based on the previous visualizations, it would appear that regions 4 and 5 are currently the most profitable regions for the charity in terms of the averages calculated. The donors in region 5 already contribute the largest donations on average each time they donate, so if they are appealed to for more frequent donations, this would only increase the source of income for the charity.

For the purpose of trying to establish a single region that is currently the most profitable, additional metrics will be considered. This time, the number of donors and the total amount of money donated by each region will be considered.

	Region	Number of Donors	Total_Lifetime_Gifts	Total_Num_Promotions
0	Region 2	669	68754.21	32009.0
1	Region 3	578	61765.79	29500.0
2	Region 4	669	79379.57	33648.0
3	Region 5	1200	134259.04	57837.0

Fig.13 Total Lifetime Gifts and Number of Promotions for Each Region



Region 5 is the most profitable for the charity. They have the most donors, give the largest gifts on average, and have raised almost double the amount of any of the other regions. Region 5 should definitely be the main target of future advertising campaigns and fundraising events to generate even more donations.

Since the total lifetime gift amounts seem to match the number of promotions in that region, increasing the number of promotions in regions 2, 3, or 4 should show a proportional return in donations.

Question 4: How does gender and household income rating correlate with the amount and/or frequency of donations?

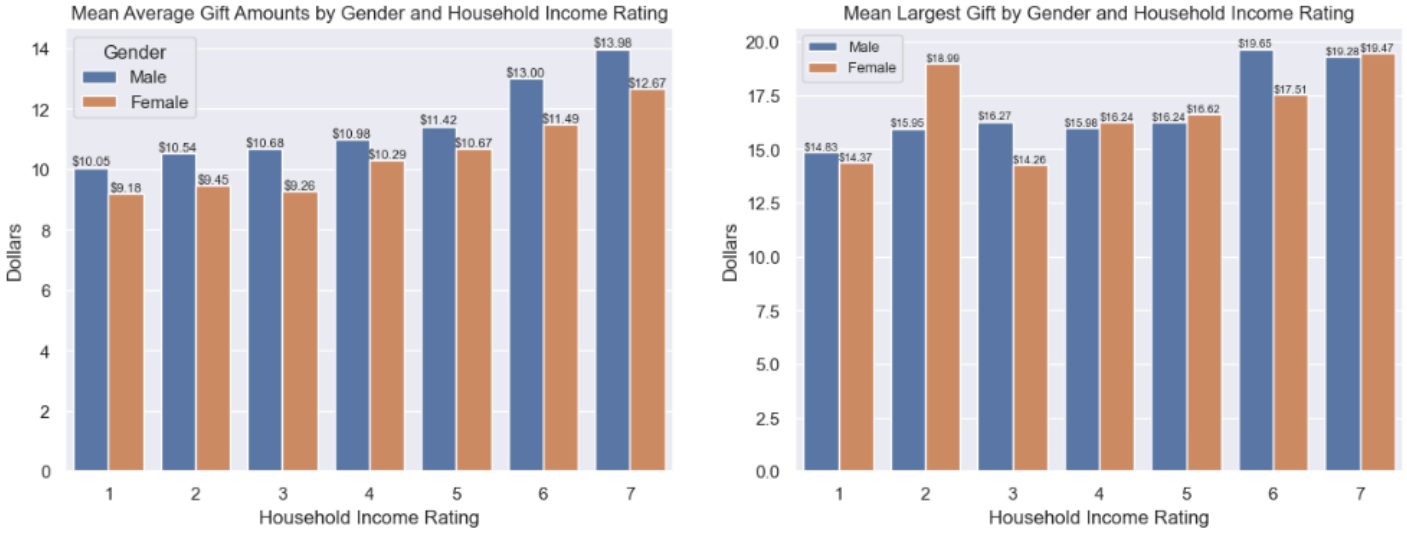
The donors were grouped by both gender and household income rating before calculating the mean number of number of months between the first two gifts, the mean lifetime gift amount, the mean largest gift amount, and the mean average gift amount. The resulting data frame was written to a separate file.

GenderIncomeStatsReport

Gender	Household Income Rating	Mean Months Between First Two Gifts	Mean Lifetime Gift Amount ($)	Mean Largest Gift Amount ($)	Mean Avg Gift Amount ($)
Male	1	7.242990654205610	107.1	14.83	10.05
Male	2	7.064705882352940	109.67	15.95	10.54
Male	3	7.217741935483870	104.01	16.27	10.68
Male	4	6.882198952879580	102.47	15.98	10.98
Male	5	6.990740740740400	95.09	16.24	11.42
Male	6	6.778761061946900	121.02	19.65	13.0
Male	7	6.168224299065420	96.76	19.28	13.98
Female	1	6.645714285714290	113.15	14.37	9.18
Female	2	6.389261744966440	133.54	18.99	9.45
Female	3	6.895348837209300	97.63	14.26	9.26
Female	4	6.977645305514160	115.16	16.24	10.29
Female	5	7.153605015673980	110.84	16.62	10.67
Female	6	5.87218045112782	110.81	17.51	11.49
Female	7	7.330827067669170	105.11	19.47	12.67

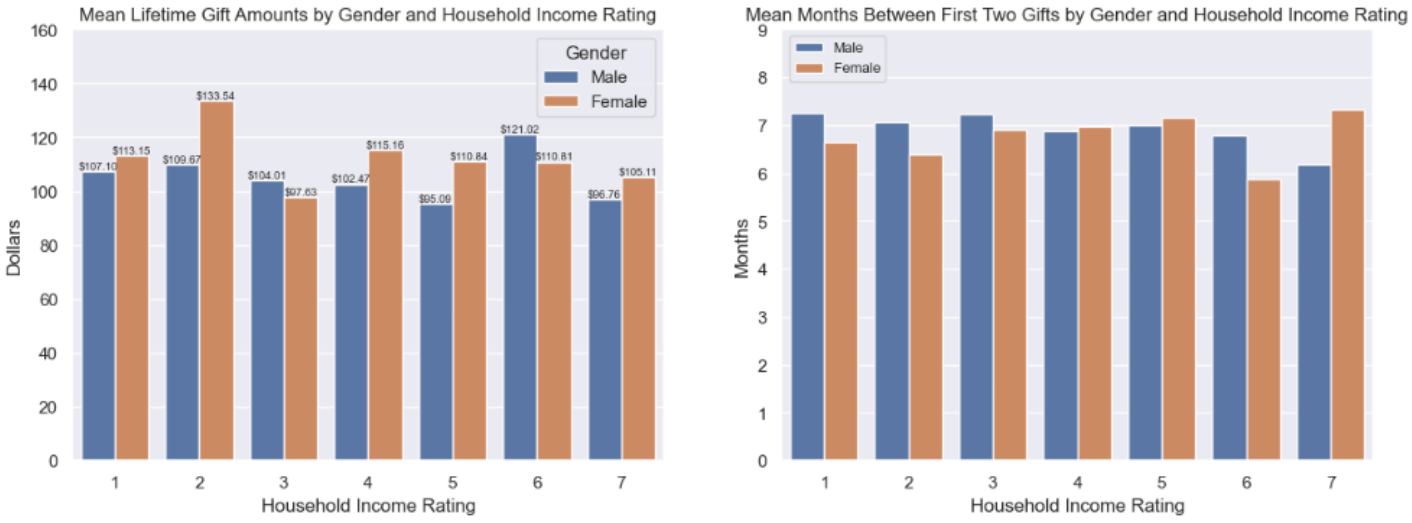
The first of the two plots reveals that the mean average gift amount steadily increases as the household income rating increases. The donors belonging to households that have more income are donating greater amounts on average to the charity. Also, male donors donate a greater gift amount on average than female donors for all the household income ratings. The second plot reveals that there is not as much of a direct positive correlation between the mean largest gift amount and the household income rating. Female donors with a household income rating of 2 have one of the greatest mean gift amounts of all the wealth ratings.

Fig.14a Gift Statistics for Each Household Income Rating by Gender



The mean lifetime gift amounts are also somewhat consistent across the household income ratings. There is not a significant trend that would reveal that donors give more or less overall depending on their household income. Across all household income ratings, female donors usually give more overall when compared to male donors in the same household income rating. The time between the first and second donation is also consistent across all the household income ratings and gender. On average, most donors, regardless of their gender, will make their second donation around 6 to 7 months after their first donation.

Fig.14b Gift Statistics for Each Household Income Rating by Gender



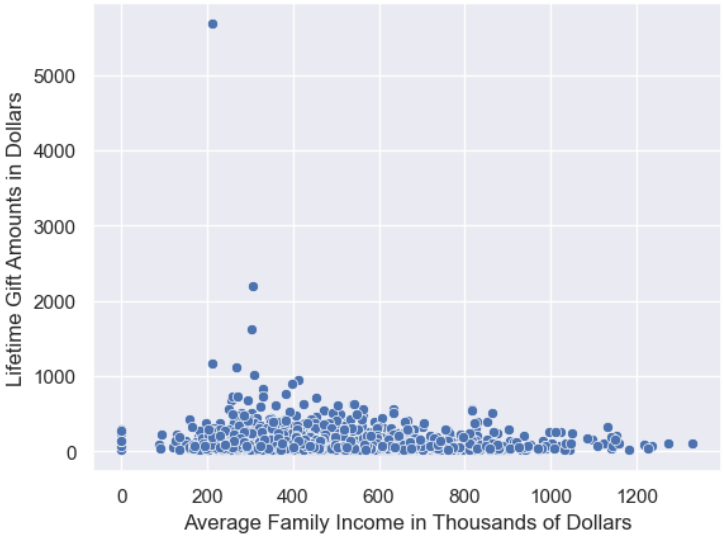
Based on the previous metrics, it would be most profitable for the charity to target donations from female donors with a household income rating of 2 and male donors with a household rating of 6. These groups contribute the most donations when comparing across both demographics in terms of mean lifetime gift amounts and mean largest gift amounts.

Question 5: How does the average family income correlate with the amount and/or frequency of donations?

Several visualizations will be created to form an answer to this question. If there are any strong trends related to the average family incomes in the neighborhoods where the donors live, the charity would be able to make advertisement decisions based on this information to increase the amount and/or frequency of donations they receive.

Figure 15 shows the relationship between average family income and the lifetime gift amounts of the donors. Almost all the points are located in a horizontal band in the range of $0 to $500. The overall amount gifted to the charity by a donor does not seem to change according to the average family incomes of the neighborhoods where they live.

Fig.15 Lifetime Gift Amount by Average Family Income



The rest of the visualizations tell a similar story. There were no positive or negative correlations between average family income and the other donation metrics.

Fig.16 Largest Gift Amount by Average Family Income

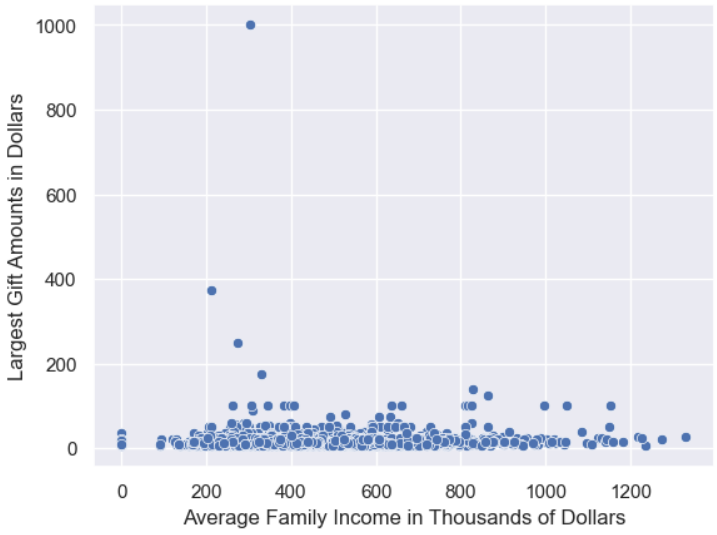


Fig.17 Months Between First Two Gifts by Average Family Income

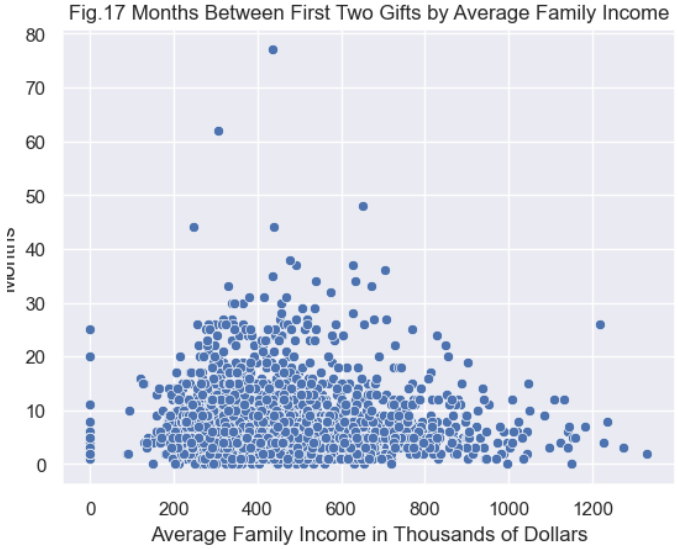


Fig.18 Average Gift Amount by Average Family Income

Final Conclusions

This data has the ability to be leveraged in a way that could guide future promotion campaigns so they could more efficiently increase the number of donations received and get a greater return on advertising costs. The most important takeaways were revealed by the answers to questions 3 and 4. Region 5 is the most profitable region in terms of donations. This region should continue to receive regular promotions to keep a steady stream of donations. Other regions show similar potential, as the number of promotions seem to line up with the total lifetime gift amounts of each region. This theory could be tested by increasing the number of promotions in one of the regions to see if the total lifetime gift amount begins to show a noticeable increase compared to the other regions. Question 4 provides more granularity to decisions about promotions. Since an increase in household income correlates with an increase in the average gift amounts, certain household ratings within each region could be specifically targeted with promotions. Also, leveraging the gender information could help target those donors who have already contributed the most as being willing to contribute more in the future.

This analysis should be the beginning of a greater conversation based on this data set. This report should inspire more data to be collected and give stakeholders the confidence to experiment with new promotion techniques based on its findings. As future data is collected, these and additional analyses could be run to see if the promotions are having the intended effect or if any other changes could be made to help increase the amount and frequency of donations.