Homework 1

Due Date: 5/9/2021

Group Roles:

In our initial group meeting, we performed our data exploration and generated a list of questions around the data set. We answered the first question as a group then broke out and answered the following three questions individually. We then regrouped and combined our findings, performed our debugging, and worked together our final report.

Data Definition:

The data set utilized for our analysis was the Donors Data. Attached is a copy of the original dataset:

This data set is comprised of a list of donors and various elements of their household, incomes and donation history.

Data Exploration:

Once the CSV file was read into our program, we performed the following steps in data exploration:

1. Examined the shape of the data utilizing the .shape function.

2. From there, we examined the formatting of the data by using the .head function to view the top five rows of the data set. We recognized some columns that were not meaningful and removed them using the del function.

3. We utilized the .describe function to analyze the descriptive stats of each column.

4. Additionally, we renamed the columns so that they were more intuitively understood. Below is a list of the variables with their definitions:

Element	Definition
Region_1, Region_2, Region_3, Region_4	Region
homeowner	(1 = homeowner, 0 = not a homeowner)
numchildren	Number of children
income_d	Household income (7 categories)
gender	Gender (0 = Male, 1 = Female)
wealth_d	Wealth Rating
homevalue	Average Home Value in potential donor's neighborhood in $ thousands
income_med	Median Family Income in potential donor's neighborhood in $ thousands
income_avg	Average Family Income in potential donor's neighborhood in $ thousands
lowincome_perc	Percent categorized as “low income” in potential donor's neighborhood
numpromos	Lifetime number of promotions received to date

donations_total	Dollar amount of lifetime gifts to date
donations_max	Dollar amount of largest gift to date
donations_last	Dollar amount of most recent gift
donations_months_since_last	Number of months since last donation
donations_months_between_first_second	Number of months between first and second gift
donations_avg	Average dollar amount of gifts to date
Donor	Classification Response Variable (1 = Donor, 0 = Non-donor)
Donation_Amount	Prediction Response Variable (Donation Amount in $).

Once we the data was cleaned and understood, we came up with 4 unique questions around the data set.

Question 1: Which region has the highest donation total and what gender contributed the most of that donation amount?

To answer this question, we created a subset of the original data frame, pulling the specific columns of just regions, gender, and donation totals. We then utilized the .pivot_table function and grouped the data frame by region and gender to gain insight on the donation amounts by these two variables. This generated the below output:

					donations_total
Region_1	Region_2	Region_3	Region_4	gender	
0	0	0	0	0	226.00
				1	63.00
			1	0	49136.63
				1	85122.41
		1	0	0	24932.32
				1	54447.25
	1	0	0	0	25143.50
				1	36622.29
1	0	0	0	0	27274.51
				1	41479.70

The conclusion that we were able to draw from this is that region 4 has the highest donation total and females donated the most in the region at the amount of $85,122.41.

Question 2: Are there correlations between the average home value, median income, average income, and low income percentage? Would it make sense to combine any of them?

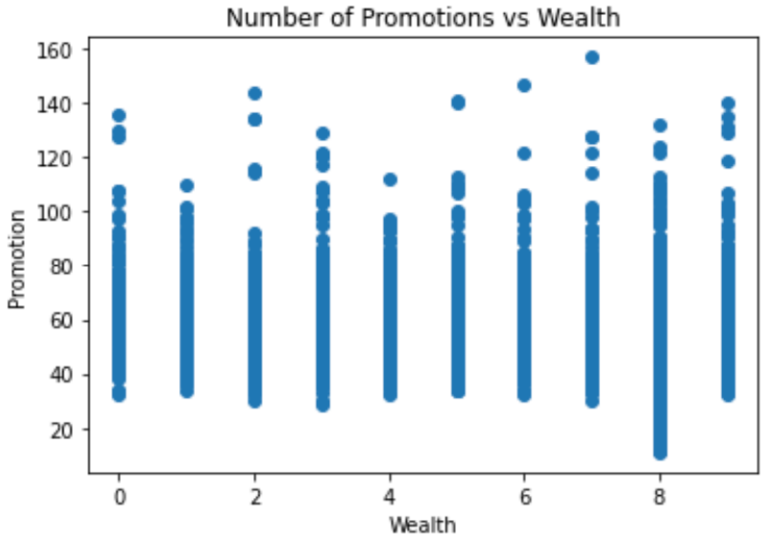
A subset data frame was created, containing only the variables mentioned in the question. This allowed us to generate a correlation matrix of the subset to draw a conclusion. This was done using the .corr function. Below is the output of this matrix:

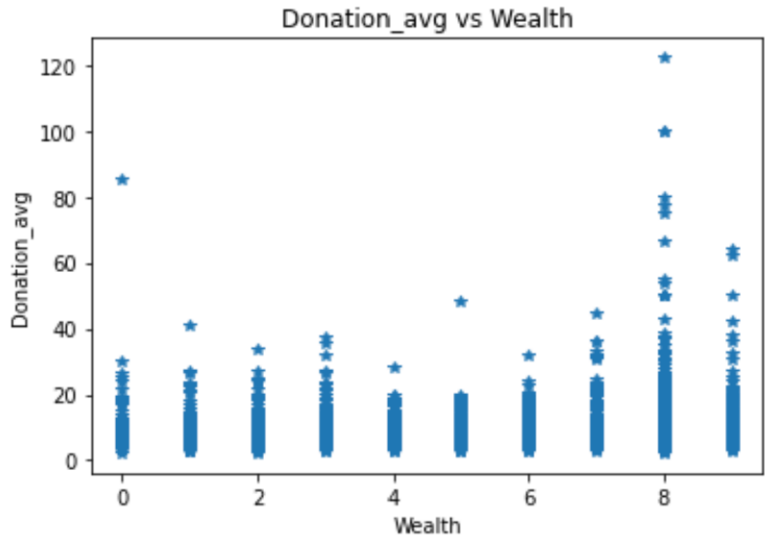
	homevalue	income_med	income_avg	lowincome_perc
homevalue	1.000000	0.733773	0.748237	-0.398843
income_med	0.733773	1.000000	0.972409	-0.666637
income_avg	0.748237	0.972409	1.000000	-0.680754
lowincome_perc	-0.398843	-0.666637	-0.680754	1.000000

Through this, we were able to see that income_med and income_avg are almost a perfect positive correlation and one could be removed to reduce noise in the dataset.

Question 3: How does wealth affect donation amounts on average and promotions received?

To create a subset data frame to answer this question, we broke the original data frame down into three separate data sets, then joined them together using the .concat function. This resulted in a dataset containing the number of promotions, average donations, and wealth levels of the donors. We then plotted the correlation between wealth and promotions (shown below) to answer a portion of the original question. To answer the second part of the question we plotted a similar chart, except this time reflecting the correlation between wealth and donations on average.





These two visuals enabled us to come to the conclusion that there is definitely a relationship between wealth and donation amount. As the wealth rating increases, the donation amount on average also increases. However, in terms of number of promotions received vs. wealth, there is not a distinct relationship. The plot shows that regardless of wealth rating, all the promotions received is relatively the same.

Question 4: Is there any correlation between the number of kids in a household and their donation practices?

For our final question, we created a subset data frame containing the number of children in a household, their donation totals, max, last, and average donations. We first explored this data set using the .describe function and then created a correlation matrix out of it.

	numchildren	donations_total	donations_max	donations_last	donations_avg
numchildren	1.000000	-0.051376	-0.017954	-0.013765	-0.018444
donations_total	-0.051376	1.000000	0.505979	0.202602	0.182302
donations_max	-0.017954	0.505979	1.000000	0.457395	0.481323
donations_last	-0.013765	0.202602	0.457395	1.000000	0.864480
donations_avg	-0.018444	0.182302	0.481323	0.864480	1.000000

The correlation matrix did not present and indications that there was a relationship between the number of children in a household and their donations, so we created another table for some additional insight. This was a separate data frame which grouped the data set by the number of children in the household and took an average of the other variables.

	numchildren	donations_total	donations_max	donations_last	donations_avg
0	1	112.153469	16.753149	13.565781	10.727186
1	2	75.487374	14.434343	12.373737	9.943455
2	3	91.266129	15.290323	13.419355	9.644536
3	4	46.588235	14.647059	13.294118	10.814566
4	5	33.000000	10.000000	7.000000	6.600000

Through this analysis we were able to gain further insight on how the number of children in a household impacts the donation levels of the donor. We can see that households with 1-4 children donate $10.28 per donation, while households with 5 children donate less on average at $6.60 per donation. The total number of donations tell a similar story, where households with 1 child donate the most in total and 3 children totaling in second. Households with 4-5 children donate on a much more infrequent basis than households with fewer children.