Nick Videtti

IST-659 Summer 2022

Homework 1

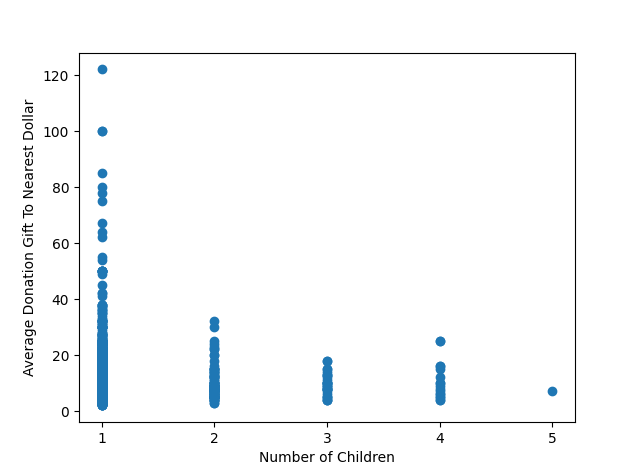
**Data**

The data used for this homework assignment were the data found in the donors\_data.csv file that was supplied in the Assessments section of the course. This was read into Python using the csv module and was saved as a pandas DataFrame. Column headers were assigned from the first row of the data and row indices were set to the Row Id column, which is defined as the index in the data dictionary for this data source. Since it seems to be much easier to append lists than it is to append rows to pandas DataFrames, the data were initially read in as a list of lists, then converted to a pandas DataFrame. For numerical analyses, many of the variables needed to be converted to integers or floats. The approach to this was adding a column to the DataFrame that was a copy of the original, but was in a numerical type. Lastly, for some scatter plots, donation amounts were rounded to the nearest dollar in order to keep from too many axis ticks.

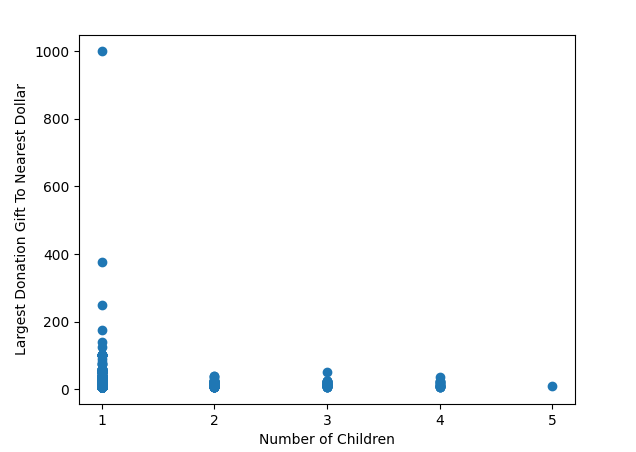
**Question 1 - Does the number of children a donor has impact the size of their donations?**

There were a few approaches taken to answering this question, none of which seemed to lead to an answer to this question.

The first approach was looking at the number of children each donor had and their average donation value, then graphing those on a scatter plot using the matplotlib.pyplot module, which resulted in the plot below.



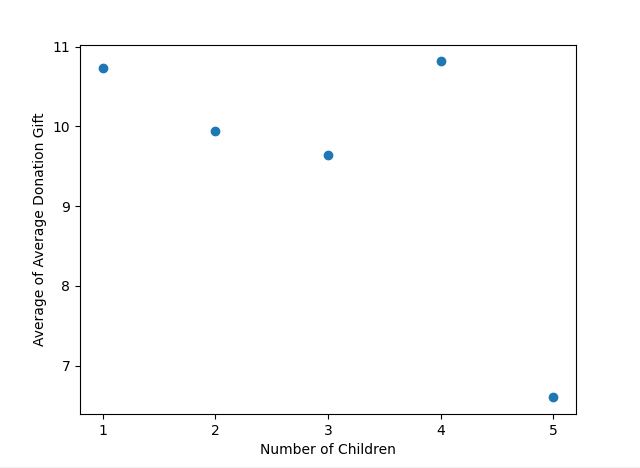
The next approach was looking at the number of children each donor had and their largest donation value, then graphing those on a scatter plot using the matplotlib.pyplot module, which resulted in the plot below.

****

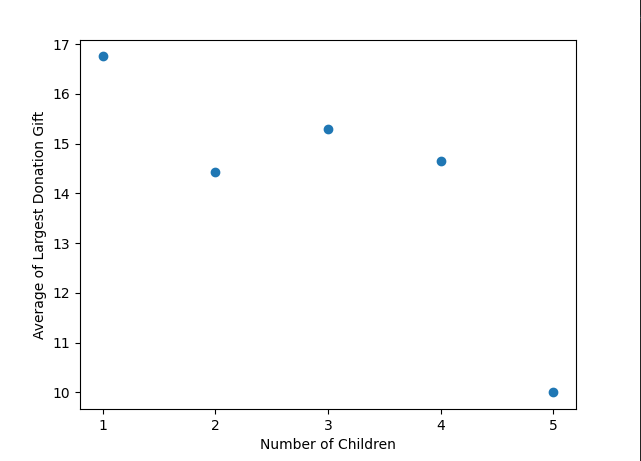
After this, the next approaches were to find the number of children corresponding to the donor with the smallest average donation value, as well as the donor with the largest average donation value. This was done by finding the index of each and using that to find the number of children in the corresponding index. It turned out that both the donor with the smallest average donation value and the donor with the largest average donation value had exactly 1 child.

The scatter plots indicated that there may be an indirect relationship between number of children and donation size, but the results of the other two approaches helped to remind that the scatter plots also tell us that there are many more donors with one child than with multiple children. To account for this, the next step was to repeat the four approaches, this time averaging by the number of children.

Approach 1 with aggregation:



Approach 2 with aggregation:



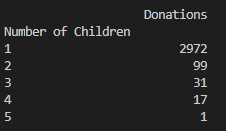
Approach 3 with aggregation:

The smallest average of the average donation amount belonged to the group of donors with 5 children.

Approach 4 with aggregation:

The largest average of the average donation amount belonged to the group of donors with 4 children.

This seems to be much more useful, but it is still important to look at the number of donations by each group to get an idea of how much the differences in group sizes were affecting the results. The data were aggregated to count the number of donations by group, and the results were as follows.



This could explain quite a bit of the results found in each approach to this question. Only one donor had 5 children, so it is difficult to make any conclusions based on that. Also, it appears that a vast majority of donations came from those with 1 child. This may actually be the most indicative metric we have discovered thus far.

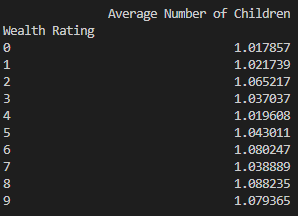
However, this could be due to the fact that not many people have a large amount of children, or the fact that the more children somebody has, the less disposable income they have to donate.

**Question 2 - Does the number of children a donor has affect their wealth status?**

Going back to the hypothesis made at the end of question 1, this somewhat builds off question 1. The wealth rating is given for each donor on a scale of 0-9, with 9 being the most wealthy. Multiple approaches were also taken for this question.

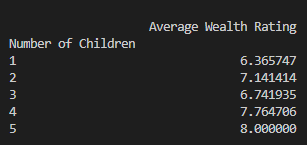
Approach 1: Average Number of Children by Wealth Rating

The DataFrame was aggregated into groups for the wealth rating. This time, the metric was average number of children. The results were as follows.(Number of children needed to be converted from a string to an integer for this analysis to work.)



Approach 2: Average Wealth Rating by Number of Children

Essentially the flipped version of approach 1, the second approach was to aggregate the DataFrame into groups for the number of children. The metric this time was average wealth rating. As stated above, wealth rating ranges from 0-9, where 9 is most wealthy. This approach was to try and limit the effect of the uneven group sizes for number of children. The results were as follows. (Wealth rating needed to be converted from a string to an integer for this analysis to work.)



Much like question 1, multiple approaches led to not much of a conclusion. There are certainly an uncommonly large proportion of donors with exactly one child, but it is still unknown as to why that is, based on the analysis results.

**Output Files:**

For each question, the last approach was saved as a separate DataFrame and was read into a new csv file using a writer from the csv module. The files were saved to a directory specified by user input.

**Other Details:**

See and run Python script and see output for more details. Both are very heavily commented. For easier reading, 5 seconds were put between execution of certain chunks of code, and 15 seconds were put between questions 1 and 2. Note that the program is not done running until “End of Homework 1” is printed to the terminal.