George Smith

HW 1: Structured Data

# Introduction

The purpose of this assignment is to write a program that will read in the data from a file, such as the .csv file saved from excel. This will be in a format that is structured with lines of data representing one type of unit (i.e., one donor in the donors file). The program will represent the data as Python data structures. The overall structure will be either:

* A list of dictionaries, or some combination of lists, dictionaries and numpy arrays
* A pandas data frame

Data exploration and cleaning has been conducted on this data. The program will do some processing to convert the data to a form that will answer at least two questions, as described below, and write files with the data suitable for answering each question. Graphing has also been completed as part of the data exploration process

## Data

The data analyzed is the data donor’s file. The file shows various donation measures with other variables such as income, family size, and homeownership. The source of this data came directly from the LMS and was imported using pandas into a data frame.

## Data Cleaning

In order to have a clean data frame to work with for the analysis, data cleaning was a necessary first step. The following columns were removed as they were deemed not useful for the analysis:

* Row Id
* Row Id
* zipconvert\_2
* zipconvert\_3
* zipconvert\_4
* zipconvert\_5

After dropping the unnecessary columns each column was renamed in order to follow a similar naming convention, as well as to provide names that are more descriptive to what the field represents than the originally provided values.

## Data Questions

Data analysis was conducted to answer the following questions:

1. Does the number of promotions someone has received influence the number of children they have?
2. Does wealth rating influence the average gift amount?

Question 1:

In order to determine if the number of promotions someone has received influences the number of children they have, 5 ranges of promotions were created with values of 0-25, 25-50, 50-75, 75-100, and 100 – 125. A plot was created to view the distribution of number of promotions someone has received and the number of children they have.

Chart

Description automatically generated with medium confidence

Based on the above plot it is evident that the data doesn’t contain many records in the 75-100, and 100-125 promotion range. As a result, it was necessary to calculate the mean value of children by each representative promotion group.

Table

Description automatically generated

Based on the above results it is reasonable to conclude that as the number of promotions increases the number of children decreases.

Question 2:

In order to determine if wealth rating influences the average gift amount, a plot was created to visualize the distribution of wealth rating and average gift amount.

Table

Description automatically generated with medium confidence

The above plot displays that most records are a part of wealth group 8. As a result, the mean average gift amount was calculated by wealth group.

Table

Description automatically generated

By analyzing the above output, it is evident that there is no definitive relationship between distribution of wealth rating and average gift amount.