Homework Assignment #1

Analysis of Donors data using Python Pandas

Syracuse University
Prof. Landowski
Student: Patrick Walsh
Date: 5/6/2023

Data and its source

The donors dataset is a CSV file with demographic information on donors and the gift amounts the donate. Demographic information includes income, wealth, number of children, and other information about each donor. The dataset was provided by Syracuse University.

Description of your data exploration and data cleaning steps

We first load the dataset using Pandas and check for null values. If there are null values, we should drop them from our analysis.

```
1 import pandas as pd
2
3 file = './donors_data.csv'
4
5 df = pd.read_csv(file)
```

	Row Id	Row Id.	zipconvert_2	zipconvert_3	zipconvert_4	zipconvert_5	homeowner dummy	NUMCHLD	INCOME	gender dummy	...	IC15	NUMPROM	RAMNTALL	MAXRAI
0	1	17	0	1	0	0	1	1	5	1	...	1	74	102.0	
1	2	25	1	0	0	0	1	1	1	0	...	4	46	94.0	
2	3	29	0	0	0	1	0	2	5	1	...	13	32	30.0	
3	4	38	0	0	0	1	1	1	3	0	...	4	94	177.0	
4	5	40	0	1	0	0	1	1	4	0	...	7	20	23.0	

5 rows × 24 columns

```
1 df.dropna()
```

After analyzing each column and what it represents, we select only the columns we will be using to answer our comparison questions.

```
1 print(list(df.columns))
```

```
['Row Id', 'Row Id.', 'zipconvert_2', 'zipconvert_3', 'zipconvert_4', 'zipconvert_5', 'homeowner dummy', 'NUMCHLD', 'INCOME',
'gender dummy', 'WEALTH', 'HV', 'Icmed', 'Icavg', 'IC15', 'NUMPROM', 'RAMNTALL', 'MAXRAMNT', 'LASTGIFT', 'totalmonths', 'TIMELA
G', 'AVGGIFT', 'TARGET_B', 'TARGET_D']
```

'Row Id' and 'Row Id.'

These appear to be column headers as row indexes from other sources and don't provide much value by themselves. We will omit these from our analysis.

The zipconvert columns

These appear to be one-hot-encoded values to show which of the 5 zip codes a donor lives.

'gender dummy' and 'homeowner dummy' are binary values.

Assumption for homeowner:

1 = homeowner, 0 = non-homeowner.

Assumption for gender: 1 = man, 0 = woman.

Wealth

The values from this column are 0-9 and appear to be brackets of wealth, i.e., a net worth of $0-$50k, $51k-$80k. However, there is no information to indicate what these wealth brackets are.

Assumption: The higher the value for 'wealth', the higher the net worth of the donor.

Income

Similar to 'wealth', the 'income' column appears to represent brackets of income for the donors. The values range from 1-7 but there is nothing to indicate what values these income brackets equal.

Assumption: The higher the 'income' bracket, the higher the income of the donor.

Home value

The home value column appears to be the monthly mortgage/rent of a home. This is based on the relatively low values in this column. The highest value is 5945, which is too low to be the total value of a home.

There were a few home values of zero, which may indicate that the mortgage was paid off, but for the purposes of our analysis, we treat these zeros as outliers and thus drop these rows.

```
1 df = df[df.HV != 0]
2 df.shape
```

```
(3097, 24)
```

Number of children

This column is self-explanatory. The vast majority of donors had only 1 child, and only 1 had 5 children. No donors had zero children.

Average gift

This column is the average of gifts per donor. Assumption: This number represents the average number of gifts per donor.

Our analysis will only require the following columns, so we will exclude the rest.

	AVGGIFT	INCOME	gender dummy	NUMCHLD	WEALTH	homeowner dummy	HV
0	4.857143	5	1	1	9	1	1399
1	9.400000	1	0	1	7	1	698
2	4.285714	5	1	2	8	0	828
3	7.080000	3	0	1	4	1	1471
4	7.666667	4	0	1	8	1	547
...	...	...	...	...	...	...	...
3115	7.500000	4	0	1	8	1	697
3116	26.500000	2	0	1	8	0	590
3117	11.157895	3	0	1	7	1	3129
3118	20.000000	7	1	1	8	1	1345
3119	9.250000	5	1	1	4	1	882

3097 rows × 7 columns

Here are some statistics on the chosen columns, to include the mean, standard deviation, min and max, and quartiles.

```
1 df.describe()
```

	AVGGIFT	INCOME	gender dummy	NUMCHLD	WEALTH	homeowner dummy	HV
count	3097.000000	3097.000000	3097.000000	3097.000000	3097.000000	3097.000000	3097.000000
mean	10.694372	3.894091	0.610591	1.069422	6.405554	0.769454	1149.838231
std	7.452061	1.638844	0.487695	0.348527	2.538122	0.421250	945.007453
min	2.138889	1.000000	0.000000	1.000000	0.000000	0.000000	163.000000
25%	6.368421	3.000000	0.000000	1.000000	5.000000	1.000000	562.000000
50%	9.062500	4.000000	1.000000	1.000000	8.000000	1.000000	826.000000
75%	12.800000	5.000000	1.000000	1.000000	8.000000	1.000000	1343.000000
max	122.166667	7.000000	1.000000	5.000000	9.000000	1.000000	5945.000000

At least two clearly stated comparison questions with the unit of analysis, the comparison values, and how they are computed

Question 1:

What is the gender and number of children of homeowners with home values in the top (75%) quartile?

Unit(s) of analysis: gender, number of children, homeowner, and home value.

Comparison: compute the mode for gender and average for number of children for homeowners in the 75% quartile.

Output: mode for gender, average for number of children in the stated subset.

```
The mode for gender (man=1, woman=0):
1

Average number of children:
1.0692307692307692

Statistics for the home value column:
count     650.000000
mean     2466.061538
std      1082.179516
min      1343.000000
25%      1663.000000
50%      2121.000000
75%      2903.000000
max      5945.000000
```

Question 2

How do wealth and income affect average number of gifts?

Unit(s) of analysis: wealth, income, and average number of gifts.

Comparison: find the average for wealth and income in the bottom (25%) quantile and top (75%) quantile and their average number of gifts in these quantiles.

Output: average number of gifts for wealth and income in the bottom and top quantiles.

```
Average number of gifts among richest donors (75% quantile):
12.39993268808217

Average number of gifts among poorest donors (25% quantile):
9.721989189392765
```

Brief description of the program

The program is written in Python in a Jupyter Notebook and loads the donors dataset from the donors_data.csv file. The program requires importing the Pandas library and uses a number of different Pandas functions to filter and analyze the dataset. The answers provided in this document are fully spelt out in the accompanying Jupyter Notebook.

Description of the output and analysis of that output.
The output consists of a series of print() statements which are produced to show the answers to the questions presented about the data.

Final conclusion about the data and what your most important takeaway was
Based on the analysis, the top home values belong to men with an average of 1.1 children. This suggests a negative correlation between number of children and home values.

Additionally, the richest donors on average give a higher number of gifts. This suggests a positive correlation between wealth (wealth and income) and number of gifts donated.