Homework 1 Report

Structured Data: Donors

[1]

Marriah Lewis

08.06.2021

IST 652 Summer 2021

[1] *Two Parents, One Salary: How to Make a Single-Income Household Work in 2021. (2021). [Image]. https://www.parents.com/parenting/money/two-parents-one-salary-how-to-make-a-single-income-household-work/*

INTRODUCTION

The purpose of this assignment is to create a program based on a donors dataset. The program will represent the data as Python data structures. After cleaning and preparing the data, the following four questions below will be answered.

Questions

1. Does the amount of promotions determine the amount donated?
2. Does the number of promotions sent determine the frequency of donations?
3. Does the number of children impact the number of donations?
4. Does the number of promotions sent impact the time since the last donation?

Packages used:

import pandas as pd

from matplotlib import pyplot as plt

import matplotlib.gridspec as gridspec

DATA

#Importing the dataset

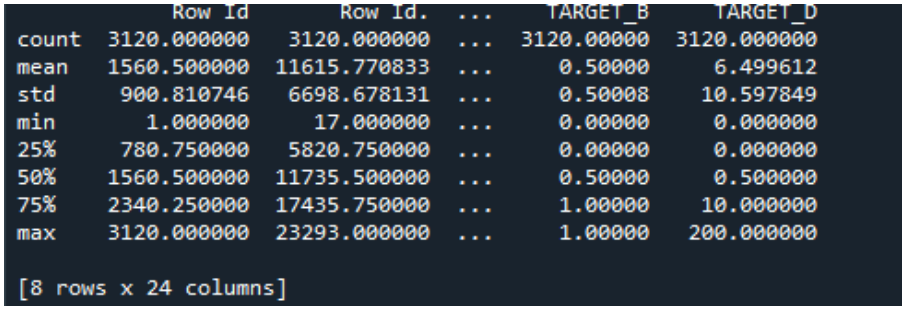
donors= pd.read_csv('C:/Users/lewis/OneDrive/Documents/data/donors_data .csv')

#Top five rows

```
In [1]: runfile('C:/Users/lewis/OneDrive/Documents/Python Scripts/
Marriah_Lewis_Homewrok_1.py', wdir='C:/Users/lewis/OneDrive/Documents/Python
Scripts')
   Row Id  Row Id.  zipconvert_2  ...   AVGGIFT  TARGET_B  TARGET_D
0       1       17             0  ...  4.857143         1       5.0
1       2       25             1  ...  9.400000         1      10.0
2       3       29             0  ...  4.285714         1       5.0
3       4       38             0  ...  7.080000         0       0.0
4       5       40             0  ...  7.666667         0       0.0

[5 rows x 24 columns]
```

#Provide a summary of stats pertaining to the df



	Row Id	Row Id.	...	TARGET_B	TARGET_D
count	3120.000000	3120.000000	...	3120.00000	3120.000000
mean	1560.500000	11615.770833	...	0.50000	6.499612
std	900.810746	6698.678131	...	0.50008	10.597849
min	1.000000	17.000000	...	0.00000	0.000000
25%	780.750000	5820.750000	...	0.00000	0.000000
50%	1560.500000	11735.500000	...	0.50000	0.500000
75%	2340.250000	17435.750000	...	1.00000	10.000000
max	3120.000000	23293.000000	...	1.00000	200.000000

[8 rows x 24 columns]

Cleaning/Preparation

First, the unnecessary rows were removed by creating another dataframe with just the data that is needed. Secondly, pd.concat was used to concatenate pandas objects along axis=1 with passed keys as the outermost level. The headers were renamed as such:

['homeowner',

'numchildren',

'income_d',

'gender',

'wealth_d',

'homevalue',

'income_med',

'income_avg',

'lowincome_perc',

'numpromos',

'donations_total',

'donations_max',

'donations_last',

'donations_months_since_last',

'donations_months_between_first_second',

'donations_avg'])

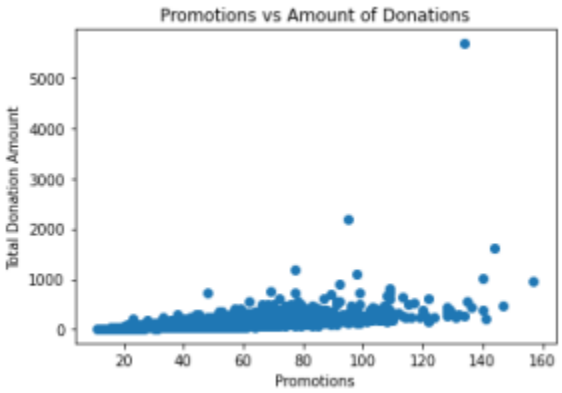
A '_d' was added for the discretized variables to stay organized. Ran another describe() command to check the data.

	homeowner	...	donations_avg
count	3120.000000	...	3120.000000
mean	0.770192	...	10.690713
std	0.420777	...	7.443980
min	0.000000	...	2.138889
25%	1.000000	...	6.356092
50%	1.000000	...	9.000000
75%	1.000000	...	12.811652
max	1.000000	...	122.166667

[8 rows x 16 columns]

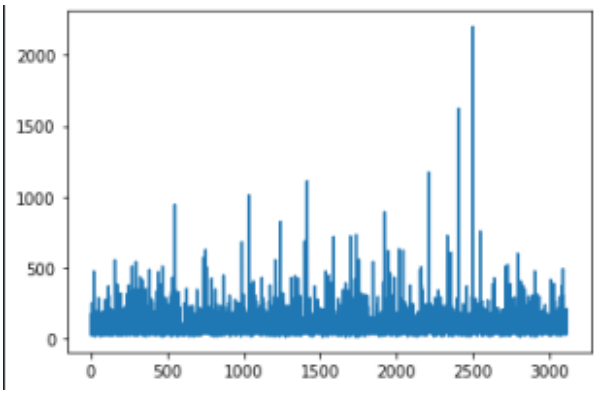
RESULTS/VISUALIZATIONS

1. Does the amount of promotions determine the amount donated?



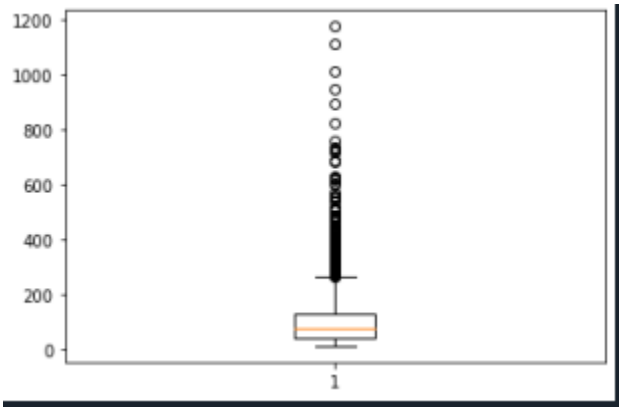
#There is a huge separation pertaining to the outlier.

Further assessment is needed, the outlier was removed so the data is normally distributed. Before completely changing the dataframe, a test dataframe was created with certain parameters set.



#Based on the plot only adding items to the df if the donation total is less than 2500.

There are still some outliers; a different parameter (<=1500) was set for re-examination.

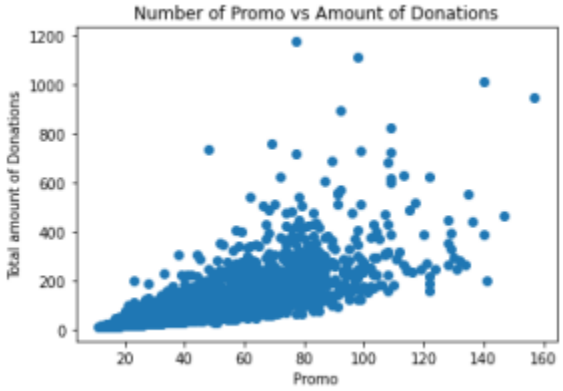


#boxplot for re-examination of the outliers

```
[8 rows x 16 columns]
homeowner                                3117
numchildren                              3117
income_d                                 3117
gender                                   3117
wealth_d                                 3117
homevalue                                3117
income_med                               3117
income_avg                               3117
lowincome_perc                           3117
numpromos                                3117
donations_total                          3117
donations_max                            3117
donations_last                           3117
donations_months_since_last              3117
donations_months_between_first_second    3117
donations_avg                            3117
dtype: int64
```

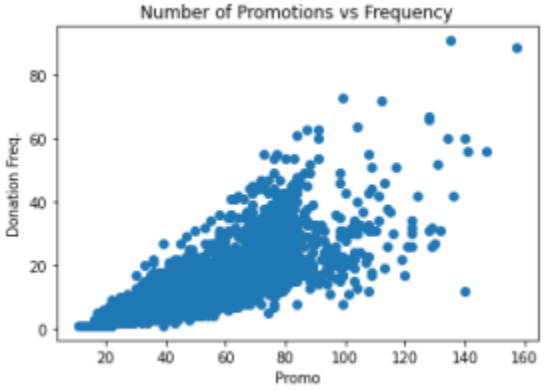
#Used count() to calculate how many times a value appears within the list.

Based on the results from the count() command, the variables were reassigned.

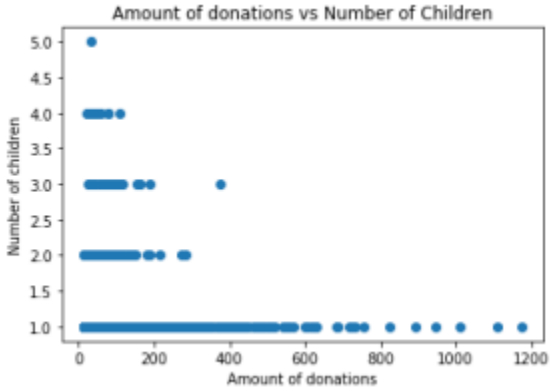


#Better visual and further clarity on question #1

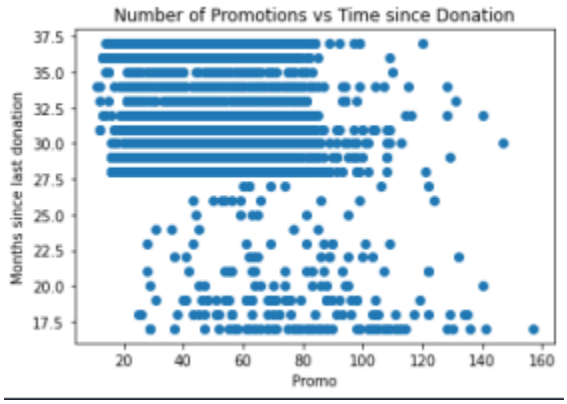
2. Does the number of promotions sent determine the frequency of donations?



3. Does the number of children impact the number of donations?



4. Does the number of promotions sent impact the time since the last donation?



CONCLUSION

Based on the data presented in this report, the number of donations increases with the number of promotions. The same results are shown when comparing the number of promotions vs the frequency of donations. Also, when the number of children increases the amount of donations given decreases. Another comparison that would be interesting to compare would be the number of children vs the income average to determine the amount of donations. Lastly, the number of promotions sent did not have a significant difference on the time since the last donation which could be due to if a person is willing to donate before they are willing to donate again no matter the amount of promotions.