SCRIPTING FOR DATA ANALYSIS

HOMEWORK 1_STRUCTURED DATA

IST652_WINTER_2022

Contents

OVERVIEW 3
ANALYSIS 3
DATA SOURCE & CLEAN UP 4
DATA EXPLORATION 5
KEY NOTES 7
The errors I got are 9
References 9

OVERVIEW

The main purpose of this document to provide a report as part of Homework 1, this one primarily provides the information how the data has been processed for analysis and the kind of analysis done by using the learnings from the class. The computation part has been done using basic python commands with the Jupyter note book .

I've used the Donors_data.csv structured data which has been presented as part of the course homework.

ANALYSIS

The data provided in the CSVfile has information about donations done by certain number of people, it has few details such as gender, wealth type, whether they have own house, number of children and income categories along with the elements like donations given, promotions, ..etc.,

By looking at the file and the data dictionary it has provided an idea that one can extract the information and trend how the donations are contributed by people.

Like whether there were more donations who has children less than 3 and whether female vs male provided, and the most important part is that were donations provided based on promotions activity.

However, one can follow the principles before analyzing any type of data, I've also followed few important steps

Step 1: What format the data is

Step 2: Whether we can consider whole data as is for analysis or do I need to

a. Clean up the junk information
b. Remove or rename some of the columns as it's a CSV file data
c. Do I consider sample data or need to train the data? Or train and test the data
d. Any principles of descriptive statistics can be applicable
e. What could be measure

DATA SOURCE & CLEAN UP

The donors_data.csv is the source of the data and since it has all the integer and float type of data (no strings or varchar) I thought they would be some blanks or NIL kind of columns.

First importantly wished to visualize what kind of information I can extract by analyzing the data

I've given preference to the questions that are mentioned in the instruction of the homework.

Imported the required libraries to load, read the data.

Executed few commands (reutilized the week5's activities base code to start with)

Such as no of columns and rows

There are 3120 rows and 24 columns of information

```
(3120, 24)
```

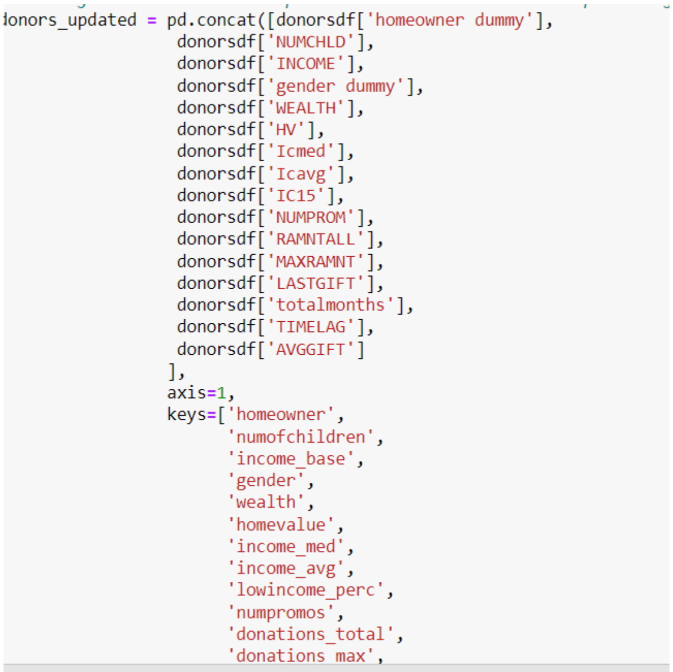
Additionally executed few more basic steps to understand the data

Column names and column types and the descriptive statistics

While doing so could figure out few column names are with suffix with dummy, _d_C and some of the naming conventions are not providing exact understanding of the data for the column has

Hence renamed the few column name and removed the columns that are not going to provide any value addition to the analysis

Like the following (screenshot of snippet for reference)



```
donors_updated = pd.concat([donorsdf['homeowner dummy'],
                 donorsdf['NUMCHLD'],
                 donorsdf['INCOME'],
                 donorsdf['gender dummy'],
                 donorsdf['WEALTH'],
                 donorsdf['HV'],
                 donorsdf['Icmed'],
                 donorsdf['Icavg'],
                 donorsdf['IC15'],
                 donorsdf['NUMPROM'],
                 donorsdf['RAMNTALL'],
                 donorsdf['MAXRAMNT'],
                 donorsdf['LASTGIFT'],
                 donorsdf['totalmonths'],
                 donorsdf['TIMELAG'],
                 donorsdf['AVGGIFT']
                ],
                axis=1,
                keys=['homeowner',
                      'numofchildren',
                      'income_base',
                      'gender',
                      'wealth',
                      'homevalue',
                      'income_med',
                      'income_avg',
                      'lowincome_perc',
                      'numpromos',
                      'donations_total',
                      'donations max',
```

Figure 1:Column concat, rename, remove_reference

By doing reduced the no of columns from 24 to 16 and named the data frame as **“donors_updated”**

Note: I’ve removed Zip converter completely as I could not find out better a way to concatenate the four columns to achieve a single zip code and though which location people donated more.

DATA EXPLORATION

I was considering the following questions for analysis

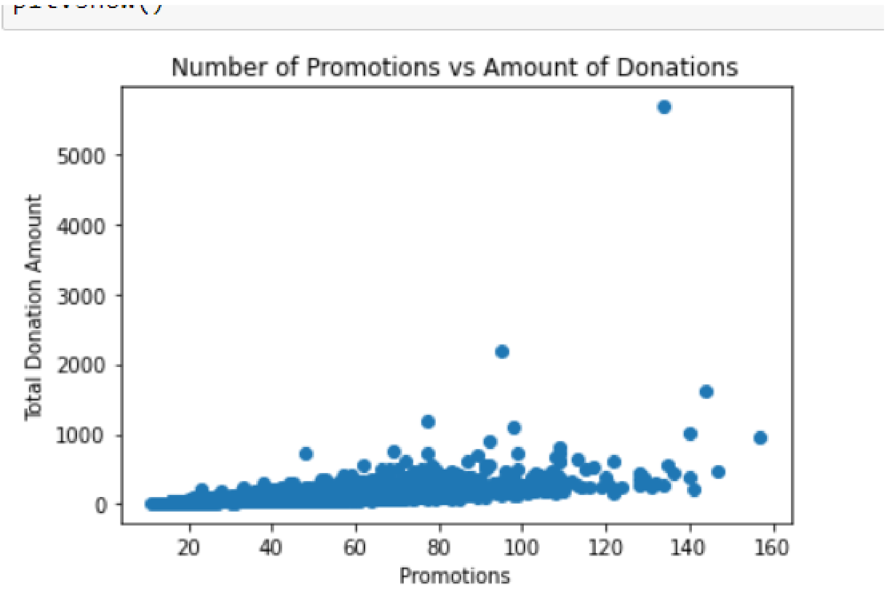
1. Compare donors by the number of promotions with the total amount of donations and the frequency of donations.
2. Compare the number of months since the last donation to the donation amounts

Need to figure what kind of data is needed

# 1 -- the total number of promotions

# 2 -- the total amount of donations

Hence tried to get the graphical representation how the data is picturized



```
plt.show()
```

Figure 2No of promotions vs total donation amount

From the above graph we can understand the range of donations done based on promotions, between 2000 to 5000 there is not much data, as this information can't add any value for analysis.

The data between 1000 to 2000 is something we can consider for analysis, such mechanisms called are skewing data and figuring out the outliers, how far the data can be used as is or how much data (which region) needs to be considered for analysis instead of testing against whole data

Like wise did couple of more tests (all tests are in the code file, to avoid duplication I did not include all the steps and graphs here with this document)

Finally concluded that, to consider the region where the data is denser than the less spread region. i.e., below 1000 and between 1000 and 2000.

Considered few more comparisons like

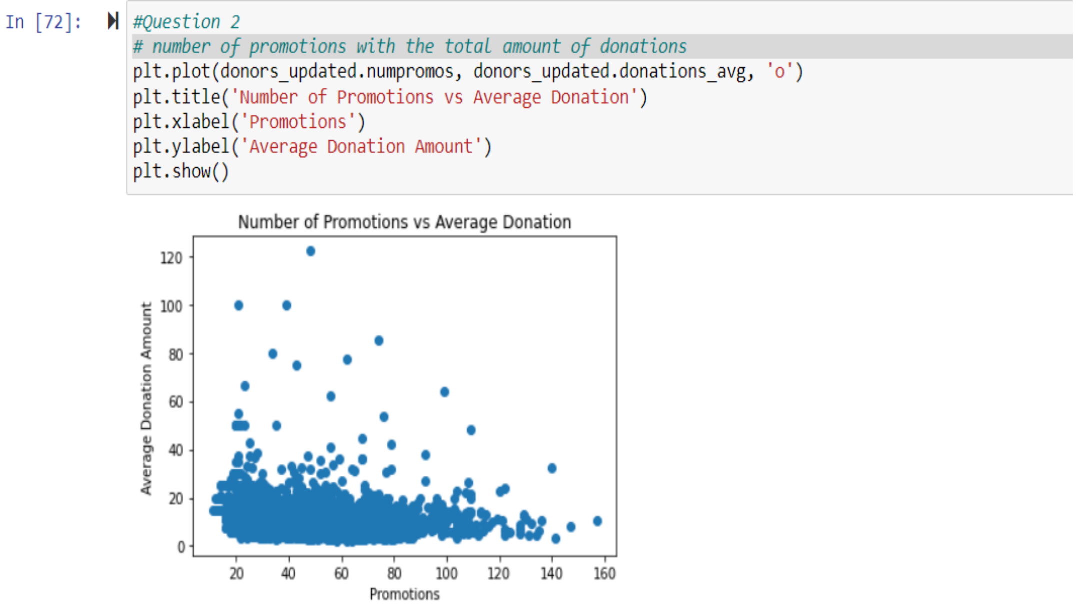
a) Homeowner and non-home owner in terms of donations, that graph has provided a good understanding as well

Yes this has certainly impacted the people who owned home are the people who contributed donations.

Question: Whether there was an impact of promotions to encourage people to contribute

Yes it’s true the promotions attracted the community to donate as the data shows the same

Below is the representation



```
In [72]:  #Question 2
          # number of promotions with the total amount of donations
          plt.plot(donors_updated.numpromos, donors_updated.donations_avg, 'o')
          plt.title('Number of Promotions vs Average Donation')
          plt.xlabel('Promotions')
          plt.ylabel('Average Donation Amount')
          plt.show()
```

Figure 3: Promotion Average Donations

KEY NOTES

The main aspect of this data analysis how better we can gauge the data to get the information, this csv is very less data in terms of real-world data(millions of rows) here we have 3126.

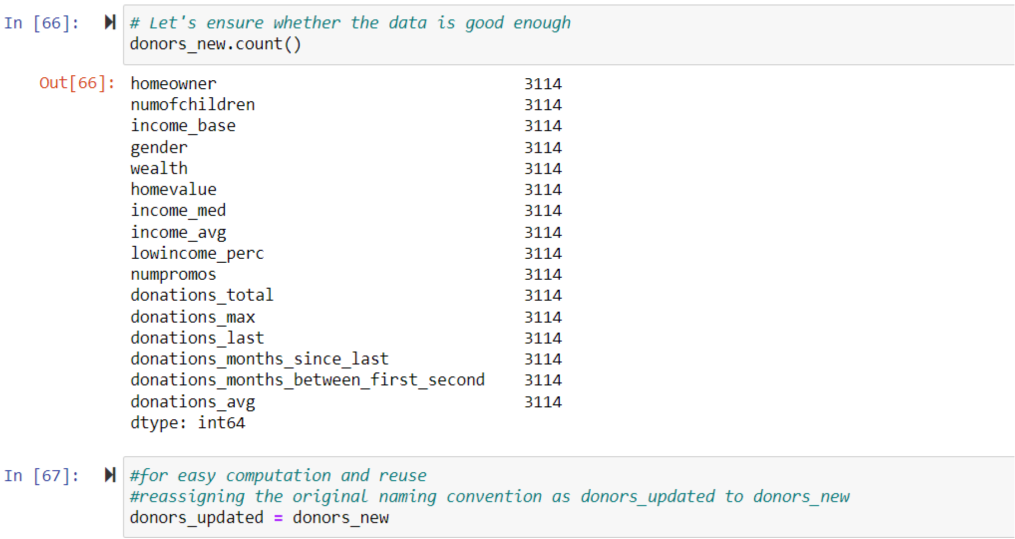
With that 3126 any data analyst needs to figure whether I can use whole, likewise I too tired and per my best of knowledge the box plots provided me a greater insight what and where I need to focus to obtain the better results.(i.e. , Outliers, Spread, Range)

Got the box plots plotted and understood the range of data, renamed the column information to be more precise and tested the same.

Since it’s a small amount of data, and for ease of homework purposes the renamed (donors_new) data frame has been reassigned to (donors_updated) to avoid the coding errors.

Considered the source as Donors_data

Updated to donors_updated (by renaming the columns and removing the columns)

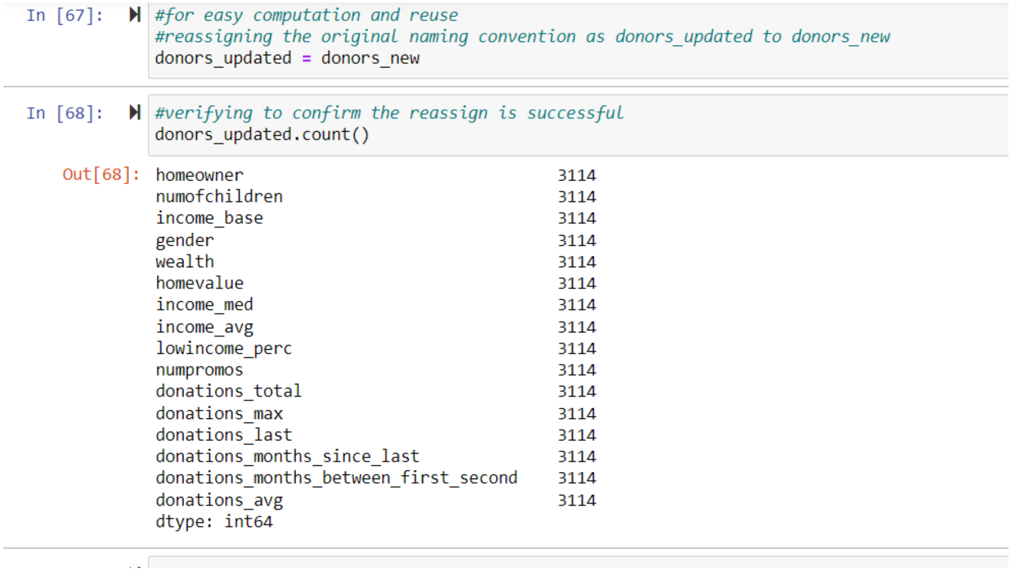


```
In [66]:  # Let's ensure whether the data is good enough
          donors_new.count()

Out[66]:  homeowner                                3114
          numofchildren                            3114
          income_base                              3114
          gender                                   3114
          wealth                                   3114
          homevalue                                3114
          income_med                               3114
          income_avg                               3114
          lowincome_perc                           3114
          numpromos                                3114
          donations_total                          3114
          donations_max                            3114
          donations_last                           3114
          donations_months_since_last              3114
          donations_months_between_first_second    3114
          donations_avg                            3114
          dtype: int64

In [67]:  #for easy computation and reuse
          #reassigning the original naming convention as donors_updated to donors_new
          donors_updated = donors_new
```

Re- analyzed the data to "donors_new" after the box plots to focus on the 3114-row information and assigned the name as "donors_updated".



```
In [67]:  #for easy computation and reuse
          #reassigning the original naming convention as donors_updated to donors_new
          donors_updated = donors_new

In [68]:  #verifying to confirm the reassign is successful
          donors_updated.count()

Out[68]:  homeowner                                3114
          numofchildren                            3114
          income_base                              3114
          gender                                   3114
          wealth                                   3114
          homevalue                                3114
          income_med                               3114
          income_avg                               3114
          lowincome_perc                           3114
          numpromos                                3114
          donations_total                          3114
          donations_max                            3114
          donations_last                           3114
          donations_months_since_last              3114
          donations_months_between_first_second    3114
          donations_avg                            3114
          dtype: int64
```

Likewise, few more examples I've tried few comparisons

1) Whether the wealth ==9 were they more contributions

1.a) People with wealth <5 how the contributions are done

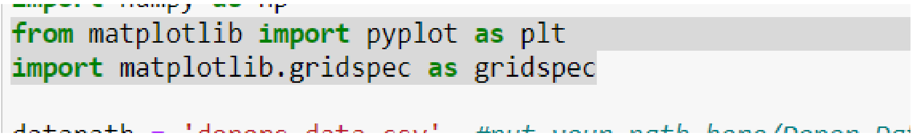
2) Does there any variation based on the gender (female vs male)

3) Is there any trend that people with 2 child have more donations contributed than people with 5 children.

For all these the output shows a decent amount of information to conclude that wealth rank, no of children has influence on the donations whereas gender comparison is doesn't matter much as both male and female contribution doesn’t have a greater variation (graphical representation is with the code file)

The errors I got are

1) Graphs I could not get properly for the first instance even though I’ve imported matplot I was able to plot successfully after executing the following



```
import numpy as np
from matplotlib import pyplot as plt
import matplotlib.gridspec as gridspec

datapath = 'donors_data.csv'  #put your path here(Donor_Da
```

2) Couple of items I mistyped the dataframe name like donors to donor and donor_updated to zigzag occurred

3) Initially thought pie charts or ggplots would provide better details of the data however back to basics of statistics as part of applied data science program the statistics class always emphasis on range of the data, spread and dense and outliers. Hence box plots would be best option for me

I thought to get a heatmap, but I consistently got errors which says that data needs to be corrected and can’t be executable with commands used hence removed all the heatmap concepts to get the plots various wealth ranges vs contributions

References

1) 2su’s week 5 base code

2) Kaggle.com

3) realpython.com