PS3_Han

September 17, 2021

1 Introduction

This mini-project will serve as a post ad hoc analysis prep for my first year paper. After submitting it to the department, my advisors and I found some potential errors or places for improvement. This assignment is a perfect opportunity for me to check whether our initial findings have some support by visuals. I will explain the potential problems one by one and talk about the findings from each graph.

Essentially, my first-year paper uses a 4 year panel from 183 food banks to explore the moderators on the negative relationship between number of fundraising employees and each employees' efficiency. We introduced a new measurement, revenue per fundraising employee, as the DV.

[1]:
```
#first import the data and packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# Read in data from Excel workbook directly
foodbank = pd.read_excel('C:/Users/MSB/Dropbox/Food Bank Research with Ruby/
 ↪Aggregated Food Bank Data - Market and Tech.xlsx',
                         header=0)

#Subset data to 2016-2019
foodbank = foodbank[foodbank['Year']>2015]

#look at data
foodbank.head(n=10)
```

[1]:
```
     Year  FBID                      FB Name State region            division  \
2    2016     2        High Plains Food Bank    TX  South  West South Central
3    2017     2        High Plains Food Bank    TX  South  West South Central
4    2018     2        High Plains Food Bank    TX  South  West South Central
5    2019     2        High Plains Food Bank    TX  South  West South Central
8    2016     4  Atlanta Community Food Bank    GA  South      South Atlantic
9    2017     4  Atlanta Community Food Bank    GA  South      South Atlantic
10   2018     4  Atlanta Community Food Bank    GA  South      South Atlantic
```

```
11  2019     4  Atlanta Community Food Bank    GA  South      South Atlantic
14  2016     5     Golden Harvest Food Bank    GA  South      South Atlantic
15  2017     5     Golden Harvest Food Bank    GA  South      South Atlantic

        City  Offices  Total Physical Plant  size of service area (sq miles)  \
2   Amarillo   2392.0              144159.0                          28490.04
3   Amarillo   2392.0              144159.0                          28490.04
4   Amarillo   2392.0              144159.0                          28490.04
5   Amarillo      NaN                   NaN                               NaN
8    Atlanta  36707.0              190872.0                           9261.56
9    Atlanta  36707.0              197520.0                           9261.56
10   Atlanta  36707.0              197520.0                           9261.56
11   Atlanta      NaN                   NaN                               NaN
14   Augusta  10394.0               96657.0                          14197.47
15   Augusta  10394.0               96657.0                          14197.47

    ...  FB Likes  Twitter Likes  ST_social media  #ofnew_ind_donors  \
2   ...    4934.0         3676.0             0.25                987
3   ...    5301.0         3965.0             0.25               1341
4   ...    5745.0         3960.0             0.25                875
5   ...    5745.0         3960.0             0.25                734
8   ...   13060.0         7922.0             0.40              13489
9   ...   14462.0         8563.0             0.25              14558
10  ...   16035.0         9074.0             0.20              13466
11  ...   16035.0         9074.0             0.20              10268
14  ...    5589.0          480.0             0.33               3067
15  ...    5744.0          584.0             0.25               1140

    #ret_ind_donors  %new_ind_donors  total_pop  left_ind_donors  exist_2002  \
2            8419.0         0.097897   477370.0           1663.0           1
3           10400.0         0.142569   477236.0           -994.0           1
4            3627.0         0.074525   477268.0           8114.0           1
5            3375.0         0.163039   475188.0           1127.0           1
8           24000.0         0.357618  5849704.0          13719.0           1
9           22209.0         0.388327  5939415.0          15280.0           1
10          23657.0         0.366252  6020216.0          13110.0           1
11          23296.0         0.276594  6105046.0          13827.0           1
14           6710.0         0.291401  1303758.0           3815.0           1
15           8639.0         0.116600  1309594.0           1138.0           1

    year_min
2       2012
3       2012
4       2012
5       2012
8       2012
9       2012
```

```
10      2012
11      2012
14      2012
15      2012

[10 rows x 79 columns]
```

[2]:
```python
list(foodbank.columns)
```

[2]:
```
['Year',
 'FBID',
 'FB Name',
 'State',
 'region',
 'division',
 'City',
 'Offices',
 'Total Physical Plant',
 'size of service area (sq miles)',
 'Advocacy: Full Time',
 'Advocacy: Part-Time',
 'Advocacy: FTE',
 'Agency Relations: Full Time',
 'Agency Relations: Part-Time',
 'Agency Relations FTE',
 'Communications/Marketing: Full Time',
 'Communications/Marketing: Part-Time',
 'Communications/Marketing: FTE',
 'Dedicated Nutrition Staff (Registered Dieticians, paid interns) : Full Time',
 'Dedicated Nutrition Staff (Registered Dieticians, paid interns) : Part-Time',
 'Dedicated Nutrition Staff (Registered Dieticians, paid interns) : FTE',
 'Development/Fund Raising: Full Time',
 'Development/Fund Raising: Part-Time',
 'Development/Fund Raising: FTE',
 'Food Sourcing: Full Time',
 'Food Sourcing: Part-Time',
 'Food Sourcing: FTE',
 'SNAP (Food Stamp) Outreach: Full Time',
 'SNAP (Food Stamp) Outreach: Part-Time',
 'SNAP (Food Stamp) Outreach: FTE',
 'Technology: Full Time',
 'Technology: Part-Time',
 'Technology: FTE',
 'Warehouse Operations: Full Time',
 'Warehouse Operations: Part-Time',
 'Warehouse Operations: FTE',
 'Other: Full Time',
```

```
'Other: Part-Time',
'Other: FTE',
'TotalFTE',
'Pounds Distributed',
'Pounds transferred',
'Pounds trash',
'Program Expenses',
'Management Expenses',
'Development/ Fundraising Expenses',
'Total Operating Expenses',
'Earned Revenue',
'Government Support',
'Fundraising Revenue ($) from Individuals',
'Fundraising Revenue ($) from Corporations',
'Fundraising Revenue ($) from Foundations',
'Fundraising Revenue ($) from Social Organizations',
'Total Fundraising Revenue ($)',
'# of Individual Donors',
'# of Corporate Donors',
'# of Foundation Donors',
'# of Social Organization Donors',
'Total # of Donors',
'Cost per Dollar Raised',
'Cost per Meal',
'Annual Meal Gap',
'Food Insecure Rate',
'Wages & Benefits',
'Agencies served',
'Volunteer FTE',
'OE_DF Expenses',
'OE_Tech Expenses',
'FB Likes',
'Twitter Likes',
'ST_social media',
'#ofnew_ind_donors',
'#ret_ind_donors',
'%new_ind_donors',
'total_pop',
'left_ind_donors',
'exist_2002',
'year_min']
```

2 Graph 1 - DV and IV

The current DV is revenue per fundraising employee, and IV is the number of full-time equivalent employees in the development and fundraising department. However, we just realized that this department might to more than raising funds, i.e. they could be responsible for generating rev-

enue from all three identified revenue streams, government support, earned revenue, and private support (which we referred to as fundraising). Therefore, we thought about combining all of the revenues as the output of the department. I would like to check the relationship of between the two new revenue streams and the DV. As a reference, I would also include the original DV on the same plot.

[4]:
```
# Subset a data for this plot

fb_plot1 = foodbank[['Government Support', 'Earned Revenue', 'Total Fundraising␣
↪Revenue ($)',
                        'Development/Fund Raising: FTE']]

#Drop those observations with zero dfFTE
fb_plot1 = fb_plot1[fb_plot1['Development/Fund Raising: FTE']>0]

# further cleaning - take out nan's
fb_plot1.dropna(inplace = True)
```

[5]:
```
#Generate the current DV
fb_plot1['RevperEmp'] = fb_plot1['Total Fundraising Revenue ($)']/
↪fb_plot1['Development/Fund Raising: FTE']

#generate new DVs
fb_plot1['GvtSptperEmp'] = fb_plot1['Government Support']/fb_plot1['Development/
↪Fund Raising: FTE']
fb_plot1['EarnedRevperEmp'] = fb_plot1['Earned Revenue']/fb_plot1['Development/
↪Fund Raising: FTE']
```

[6]:
```
# scatter plot with fitted lines
columns = ['RevperEmp','GvtSptperEmp','EarnedRevperEmp']
color = ['navy','darkred','darkgreen']

plt.scatter(fb_plot1['Development/Fund Raising: FTE'],␣
↪fb_plot1['RevperEmp'],alpha = 0.2, color='blue', label="Fundraising")
plt.scatter(fb_plot1['Development/Fund Raising: FTE'],␣
↪fb_plot1['GvtSptperEmp'],alpha=0.2, color='red', label="Government")
plt.scatter(fb_plot1['Development/Fund Raising: FTE'],␣
↪fb_plot1['EarnedRevperEmp'], alpha=0.2, color='green', label = "Earned")
for y, i in zip(columns,color):
    plt.plot(np.unique(fb_plot1['Development/Fund Raising: FTE']),
         np.poly1d(np.polyfit(fb_plot1['Development/Fund Raising: FTE'],
                              fb_plot1[y], 1))(np.unique(fb_plot1['Development/
↪Fund Raising: FTE'])),
         color=i, linestyle="--", linewidth=2)

plt.ylabel('Revenue')
plt.xlabel('dfFTE')
```

```
plt.title('Revenues and dfFTE')
plt.legend()
```

[6]: <matplotlib.legend.Legend at 0x2bd1eb88fa0>

output_6_1.png

The above graph shows that all three DVs have similar trends with the IV of interest, serving as a baseline validation for the intention to combine the three revenue streams as a final DV. The negative trend is slightly more obvious for earned revenue than the other two DVs.

3 Graph 2 - adding a new control variable

Further exploration of the literatue made us find out that employee turnover is also a factor that can affect the relationship between fundraising revenue per employee and number of employees. Our data base does have employee retention as a type of report. Therefore, in this section, I will first extract this variable from 4 different reports and combine and convert them into a single excel file, then explore this variable's distribution.

[7]:
```
#import the raw data
ret1619 = pd.read_excel('C:/Users/MSB/Dropbox/Food Bank Data/FA Data/Human␣
↪Resources/Food Bank Staffing/Paid Staff/Employee Retention - FY2019.xlsx',
                        sheet_name="FB Details", skiprows=5, header=0)
ret1619.head(10)
```

[7]:
```
   Org Id                              Org Name         City State  \
0       1                    Roadrunner Food Bank  Albuquerque    NM
1       2                   High Plains Food Bank     Amarillo    TX
2       3               Food Bank of Alaska, Inc.    Anchorage    AK
3       4             Atlanta Community Food Bank      Atlanta    GA
4       5                Golden Harvest Food Bank      Augusta    GA
5       6                 Central Texas Food Bank       Austin    TX
6       7                      Maryland Food Bank    Baltimore    MD
7       8  Community Food Bank of Central Alabama   Birmingham    AL
8       9             The Greater Boston Food Bank       Boston    MA
9      10              Food Bank For New York City        Bronx    NY

                             Epg2  Full Time Staff  \
0                    Orange/Papaya             66.0
1                      Banana/Pear             41.0
```

2	Orange/Papaya	32.0
3	Blueberry	156.0
4	Banana/Pear	53.0
5	Blueberry	126.0
6	Blueberry	113.0
7	Pineapple/Strawberry/Watermelon	22.0
8	Apple	112.0
9	Apple	105.0

	# of Employees that Left FB	Employee Retention	Full Time Staff.1	\
0	33.0	0.500000	65.0	
1	12.0	0.707317	42.0	
2	12.0	0.625000	32.0	
3	156.0	0.000000	158.0	
4	14.0	0.735849	43.0	
5	40.0	0.682540	118.0	
6	23.0	0.796460	112.0	
7	7.0	0.681818	23.0	
8	24.0	0.785714	105.0	
9	40.0	0.619048	117.0	

	# of Employees that Left FB.1	Employee Retention.1	Full Time Staff.2	\
0	18.0	0.723077	59	
1	6.0	0.857143	41	
2	12.0	0.625000	27	
3	38.0	0.759494	167	
4	16.0	0.627907	44	
5	32.0	0.728814	118	
6	31.0	0.723214	108	
7	7.0	0.695652	19	
8	35.0	0.666667	98	
9	44.0	0.623932	141	

	# of Employees that Left FB.2	Employee Retention.2	Full Time Staff.3	\
0	36	0.389831	65.0	
1	2	0.951220	43.0	
2	9	0.666667	25.0	
3	47	0.718563	142.0	
4	12	0.727273	46.0	
5	22	0.813559	87.0	
6	36	0.666667	103.0	
7	7	0.631579	23.0	
8	27	0.724490	99.0	
9	30	0.787234	140.0	

	# of Employees that Left FB.3	Employee Retention.3
0	24.0	0.630769

```
1                          14.0              0.674419
2                           6.0              0.760000
3                          38.0              0.732394
4                          11.0              0.760870
5                          12.0              0.862069
6                          64.0              0.378641
7                           4.0              0.826087
8                          27.0              0.727273
9                          35.0              0.750000
```

[8]:
```
list(ret1619.columns)
```

[8]:
```
['Org Id',
 'Org Name',
 'City',
 'State',
 'Epg2',
 'Full Time Staff',
 '# of Employees that Left FB',
 'Employee Retention',
 'Full Time Staff.1',
 '# of Employees that Left FB.1',
 'Employee Retention.1',
 'Full Time Staff.2',
 '# of Employees that Left FB.2',
 'Employee Retention.2',
 'Full Time Staff.3',
 '# of Employees that Left FB.3',
 'Employee Retention.3']
```

[9]:
```
#only keep the data that I need
ret1619=ret1619[['Org Id','Employee Retention','Employee Retention.1','Employee␣
↪Retention.2','Employee Retention.3']]
```

[10]:
```
#rename the column names so that they indicate year
ret1619 = ret1619.rename(columns={'Employee Retention':'2019','Employee␣
↪Retention.1': '2018',
                        'Employee Retention.2':"2017",'Employee Retention.3':
↪'2016'})
```

[11]:
```
list(ret1619.columns)
```

[11]:
```
['Org Id', '2019', '2018', '2017', '2016']
```

[12]:
```
# reshape the data to a panel
ret1619 = pd.melt(ret1619, id_vars="Org Id",␣
↪value_vars=['2016','2017','2018','2019'])
```

```
ret1619
```

[12]:

```
     Org Id variable     value
0         1     2016  0.630769
1         2     2016  0.674419
2         3     2016  0.760000
3         4     2016  0.732394
4         5     2016  0.760870
..      ...      ...       ...
799     563     2019  0.750000
800     603     2019  0.846154
801     611     2019  0.875000
802     627     2019  0.735294
803     726     2019  0.727273

[804 rows x 3 columns]
```

[13]:

```
#rename the columns
ret1619 = ret1619.rename(columns={'variable':'Year','value':'Retention'})
```

[14]:

```
# save it as an excel file so that I can incoporate it into my original excel␣
↪sheet for later analysis
ret1619.to_excel("employee retention 16-19.xlsx")
```

[15]:

```
# now explore the distribution of this variable
plt.style.use('ggplot')
sns.histplot(ret1619, x = 'Retention', kde=True)
plt.title('Distribution of Employee Retention Rate 16-19')
```

[15]:

```
Text(0.5, 1.0, 'Distribution of Employee Retention Rate 16-19')
```

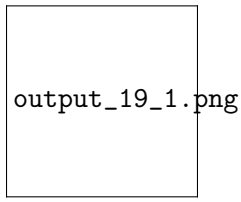
output_17_1.png

The fact that there are negative retention indicates the possibilty of reporting errors. After going back to the original dataset and manually calculating the retention rates for those foodbank-years with negative retention, I did find such errors. Therefore, I decide to take out these values and replot.

[16]:

```
plt.style.use('ggplot')
sns.histplot(ret1619[ret1619['Retention']>0]['Retention'], kde=True)
plt.title('Distribution of Employee Retention Rate 16-19')
```

[16]: Text(0.5, 1.0, 'Distribution of Employee Retention Rate 16-19')



It seems that I would need to transform this variable when including it into my model. Since it is left-skewed, I may need to use square root transformation.

4 Graph 3 - Food Price and Pounds Distributed

In this section, I would like to explore the reasons behind food price's moderating role on the relationship b/w DV and IV. Specifically, my analysis has told me that food banks located in a high food price region enjoy a lower reduction in DV. As a post ad hoc analysis, we think that this might result from the fact that food banks in low food price regions do not have much going on. Therefore, I would like to explore the relationship between food price and pounds of food distributed for each year.

[17]:
```
from bokeh.plotting import figure
from bokeh.io import output_notebook, show
from bokeh.palettes import Spectral4 as palette
from bokeh.transform import factor_cmap

output_notebook()
```

[28]:
```
# Create a subset of data that I would need for this graph
fb_plot2 = foodbank[['Cost per Meal', 'Pounds Distributed', 'Year']]
fb_plot2 = fb_plot2.astype({'Year':str})
fb_plot2.dtypes
```

[28]:
```
Cost per Meal         float64
Pounds Distributed      int64
Year                   object
dtype: object
```

[29]:
```
# First create the lists that I would need in the graph
year = ['2016','2017','2018','2019']
p = figure(title = 'Local Food Price and Pounds Distributed 16-19')
p.xaxis.axis_label = 'Cost per Meal'
p.yaxis.axis_label = 'Pounds Distributed (Annually)'
```

[30]:
```

```

```
p.scatter('Cost per Meal', 'Pounds Distributed', source = fb_plot2, legend_group␣
↪='Year', fill_alpha=0.4, size=12,
        color = factor_cmap('Year', 'Spectral4', year))

p.legend.title = "Year"

show(p)
```

The above graph demonstrates an overall upward relationship between cost per meal and pounds distributed. Our intuition was correct - food banks located in low food price regions are less active in serving the food insecure population. In addition, as time progresses, dots are moving towards the right, and the dispersion between dots becomes larger. These observations indicate the inflation of food prices as well as the enlarging differences in capabilities among food banks.