Shubhangi Singhal
IS 590 DV FA19

Assignment 3

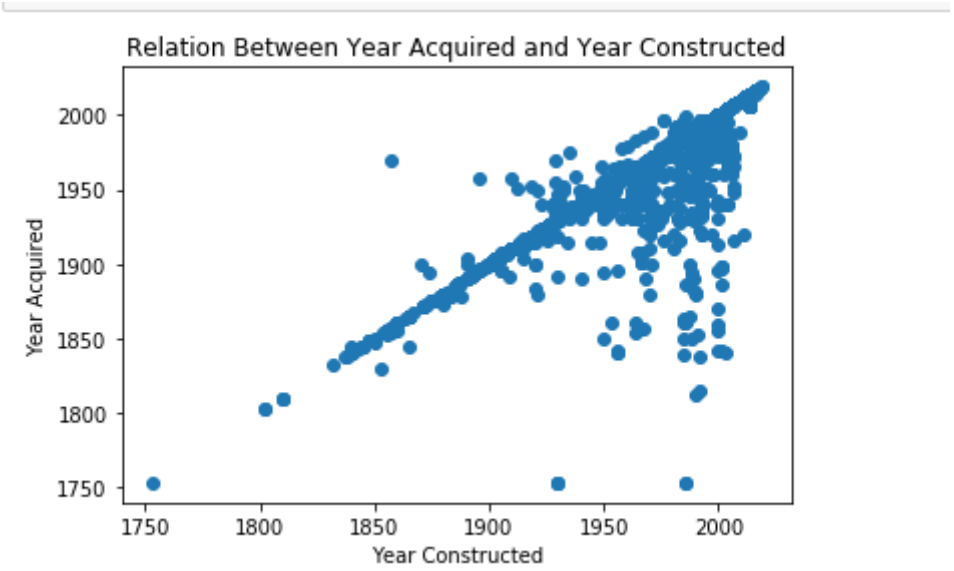
Using matplotlib, construct a visualization of the Illinois Building Inventory that communicates the following information:

- Relationship between the year acquired and the year constructed
- Total square footage as a function of congressional district
- Average square footage per floor as a function of congressional district
- Square footage for the five most common departments as a function of year

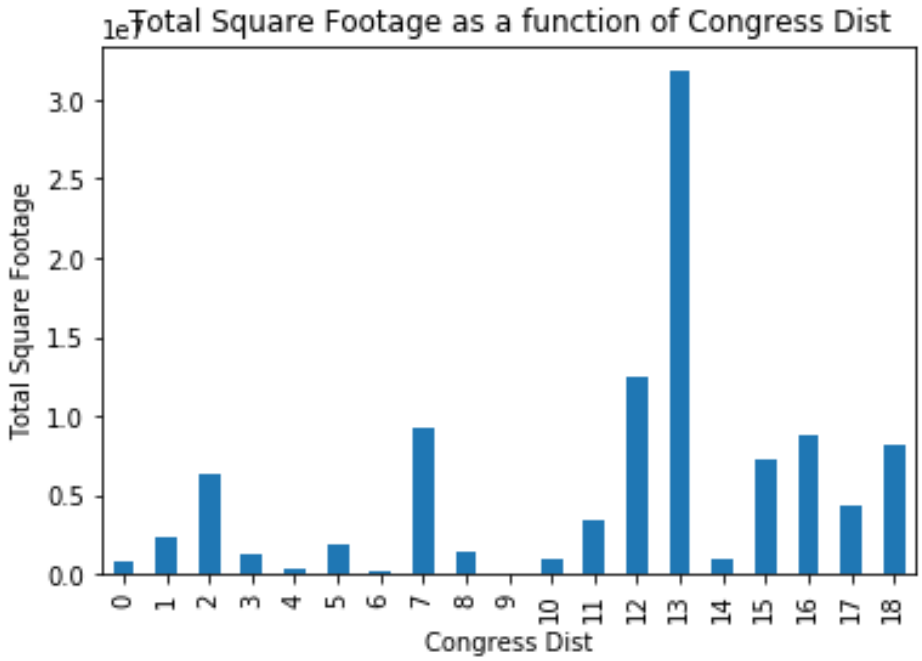
Each component will be worth 5 points and *must* be a completely communicative visualization -- including labels and a one paragraph write up of successes and shortcomings in your approach. Submit a notebook to Moodle. All source code must be in these files.

Shubhangi Singhal
IS 590 DV FA19

Q.1 – This is a scatter plot depicting the relationship between the Year Acquired and Year Constructed columns of the Dataframe "df" which is basically the Building Inventory.csv file. The x axis label is Year Acquired and the Y axis label is Year Constructed. The default colour for the scatter plot is blue. The title is "Relation Between Year Acquired and Year Constructed". I utilised a scatter plot to depict this information because I had to define a relationship between 2 similar variables which have the same unit – "Years". The points on the line at an angle of 45 degrees from the x axis shows that these are the points for which the year constructed and the year acquired is the same. Whereas, the all the points outside of this linear line depict that the year of acquiring is not the same as year of construction.



Q.2 – This is the second visualization which is a bar graph between the total square footage and the congress dist. The x axis is the Congress District column and the y axis represents the Total Square Footage that we calculated using the sum() function. I have chosen to represent these two variables using a bar graph because the district numbers are not continuous enough so that I can plot a histogram using different intervals. Bar graph can easily show 18 different district values and their corresponding Total Square Footage. The x and y labels represent the names of the columns from the Dataframe df. The title is "Total Square Footage as a function of Congress Dist".



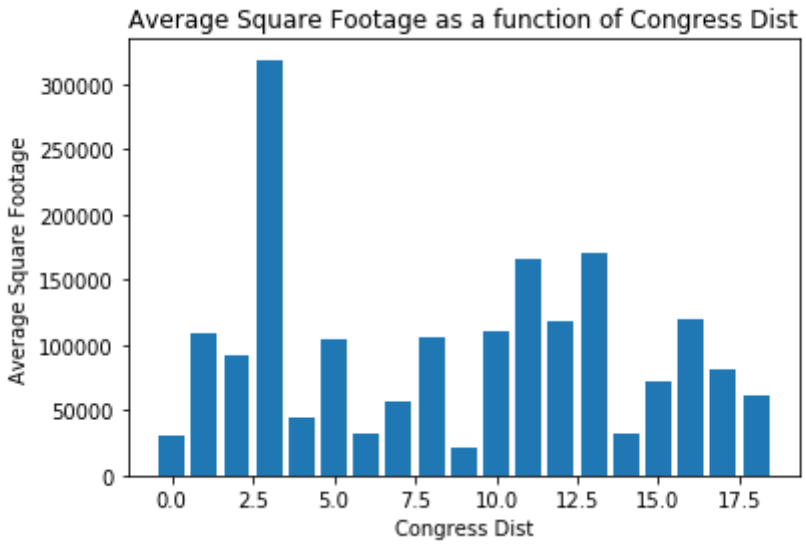
Q.3 Below is the third visualization which is a bar graph between the average square footage and the congress dist. The x axis is the Congress District column and the y axis represents the Average Square Footage that we calculated using the sum(), function. I have chosen to represent these two variables using a bar graph because the Congress District numbers are not continuous enough that I can plot a histogram, using different intervals. Bar graph can easily show 18 different district values and their corresponding Average Square Footage. The x and y labels represent the names of the columns from the Dataframe df. The title is "Average Square Footage as a function of Congress Dist".

Out[43]:

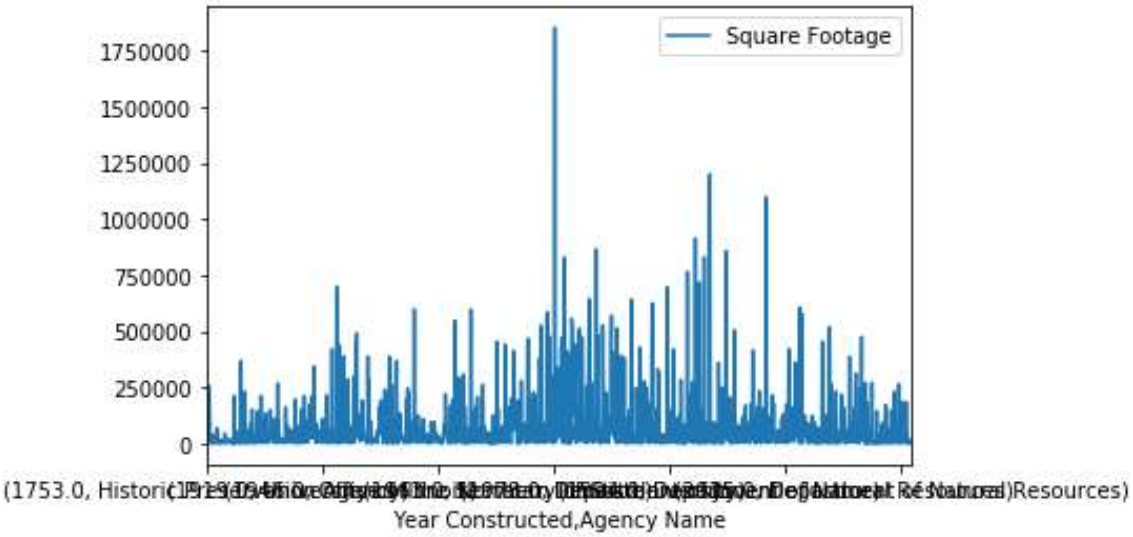
Congress Dist	Square Footage	Total Floors
0	793124.0	216
1	2351762.0	134
2	6285684.0	647
3	1293687.0	72
4	317741.0	21
5	1809022.0	113
6	145659.0	28
7	9211651.0	517
8	1461548.0	312
9	77978.0	11
10	929267.0	213
11	3467894.0	411
12	12444775.0	2339
13	31849251.0	2896
14	857108.0	319
15	7209934.0	1872
16	8725880.0	1654
17	4347743.0	924

Out[46]:

	Zip code	Congress Dist	Rep Dist	Senate Dist	Year Acquired	Year Constructed	Square Footage	Total Floors	Floors Above Grade	Floors Below Grade	Average_totalfloor
count	8862.000000	8862.000000	8862.000000	8862.000000	8597.000000	8573.000000	8.805000e+03	8862.000000	8862.000000	8862.000000	8805.
mean	61821.076845	13.404085	92.303318	46.408599	1972.243573	1970.391928	1.155032e+04	1.636087	1.449334	0.161589	ir
std	1095.203357	4.037936	23.568457	11.781038	27.549380	29.326256	3.828480e+04	1.537603	1.286898	0.392717	Nal
min	1235.000000	0.000000	0.000000	0.000000	1753.000000	1753.000000	9.000000e+00	0.000000	0.000000	0.000000	9.
25%	61105.000000	12.000000	79.000000	40.000000	1959.000000	1957.000000	2.400000e+02	1.000000	1.000000	0.000000	220.
50%	62023.000000	14.000000	97.000000	49.000000	1976.000000	1975.000000	1.664000e+03	1.000000	1.000000	0.000000	1280.
75%	62650.000000	16.000000	110.000000	55.000000	1993.000000	1991.000000	6.528000e+03	2.000000	1.000000	0.000000	4800.
max	68297.000000	18.000000	119.000000	60.000000	2019.000000	2019.000000	1.200000e+06	31.000000	30.000000	4.000000	ir


Average Square Footage as a function of Congress Dist
Average Square Footage
300000
250000
200000
150000
100000
50000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Congress Dist

Q4-



Shortcomings with this visualization – As I understand from the question, I require to plot the Total Square Footage for top 5 departments for each Year in the "Year Constructed" column. My visualization showcases the relationship between three variables "Year Constructed", "Agency Name" and "Total Square Footage". I have used the line graph to plot this relation.

Line graph with three variables, where as it should have been between 2 variables.

Top five departments for each year extracted. But could not understand the way to extract the year and total footage data for each of the five. Tried using isin() function and multiple OR conditions.

References (Sources for help)–

1. https://medium.com/python-pandemonium/data-visualization-in-python-bar-graph-in-matplotlib-f1738602e9c4 (for plotting bar graphs)

2. https://medium.com/python-pandemonium/data-visualization-in-python-scatter-plots-in-matplotlib-da90ac4c99f9 (for scatter plot)

3. Code done during the class

4. geeksforgeeks.org

5. stackoverflow.com

6. matplotlib.org