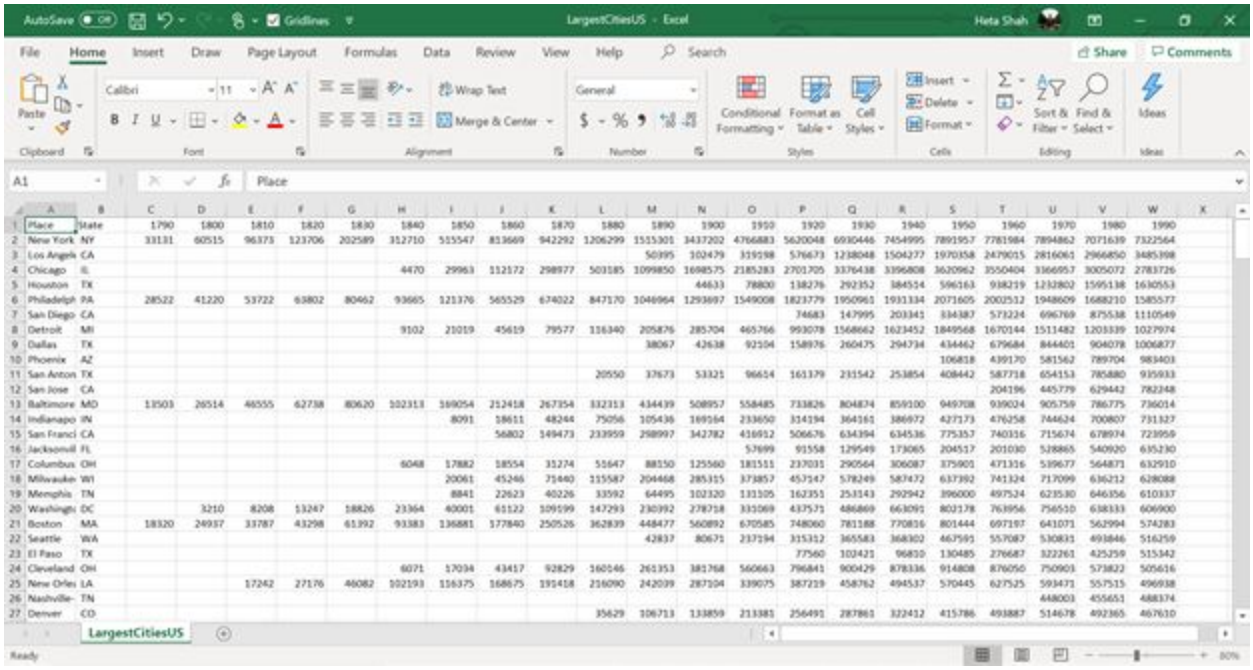
Heta Shah

Report on Population Growth in US Cities

Data from the 1790 Census thru 1990 Census is used for the analysis of Population Growth in the US Cities. This report is a complete documentation of the process included in the analysis. The data is taken from a non-formatted CSV file, formatted using Python scripting, and visualized using MS Power BI to depict the population growth trend.

CSV – LargestCitiesUS.csv snapshot:

The csv contains data about ~250 US cities, states and population over 20 years, ranging from 1790 to 1990. However, the data is not formatted well. To make the data easier to read and understand, I use Python scripting for data formatting.



Place	State	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
New York	NY	33131	60515	96373	123706	202589	312710	515547	813669	942292	1206299	1515301	3437202	4766883	5620048	6930446	7454995	7891957	7781984	7894862	7071639	7322564
Los Angele	CA											50395	102479	319198	576673	1238048	1504277	1970358	2479015	2816061	2966850	3485398
Chicago	IL						4470	29963	112172	298977	503185	1099850	1698575	2185283	2701705	3376438	3396808	3620962	3550404	3366957	3005072	2783726
Houston	TX												44633	78800	138276	292352	384514	596163	938219	1232802	1595138	1630553
Philadelph	PA	28522	41220	53722	63802	80462	93665	121376	565529	674022	847170	1046964	1293697	1549008	1823779	1950961	1931334	2071605	2002512	1948609	1688210	1585577
San Diego	CA														74683	147995	203341	334387	573224	696769	875538	1110549
Detroit	MI						9102	21019	45619	79577	116340	205876	285704	465766	993078	1568662	1623452	1849568	1670144	1511482	1203339	1027974
Dallas	TX											38067	42638	92104	158976	260475	294734	434462	679684	844401	904078	1006877
Phoenix	AZ																	106818	439170	581562	789704	983403
San Anton	TX										20550	37673	53321	96614	161379	231542	253854	408442	587718	654153	785880	935933
San Jose	CA																		204196	445779	629442	782248
Baltimore	MD	13503	26514	46555	62738	80620	102313	169054	212418	267354	332313	434439	508957	558485	733826	804874	859100	949708	939024	905759	786775	736014
Indianapo	IN							8091	18611	48244	75056	105436	169164	233650	314194	364161	386972	427173	476258	744624	700807	731327
San Franci	CA								56802	149473	233959	298997	342782	416912	506676	634394	634536	775357	740316	715674	678974	723959
Jacksonvil	FL													57699	91558	129549	173065	204517	201030	528865	540920	635230
Columbus	OH						6048	17882	18554	31274	51647	88550	125560	181511	237031	290564	306087	375901	471316	539677	564871	632910
Milwauke	WI							20061	45246	71440	115587	204468	285315	373857	457147	578249	587472	637392	741324	717099	636212	628088
Memphis	TN							8841	22623	40226	33592	64495	102320	131105	162351	253143	292942	396000	497524	623530	646356	610337
Washingt	DC		3210	8208	13247	18826	23364	40001	61122	109199	147293	230392	278718	331069	437571	486869	663091	802178	763956	756510	638333	606900
Boston	MA	18320	24937	33787	43298	61392	93383	136881	177840	250526	362839	448477	560892	670585	748060	781188	770816	801444	697197	641071	562994	574283
Seattle	WA											42837	80671	237194	315312	365583	368302	467591	557087	530831	493846	516259
El Paso	TX														77560	102421	96810	130485	276687	322261	425259	515342
Cleveland	OH						6071	17034	43417	92829	160146	261353	381768	560663	796841	900429	878336	914808	876050	750903	573822	505616
New Orles	LA			17242	27176	46082	102193	116375	168675	191418	216090	242039	287104	339075	387219	458762	494537	570445	627525	593471	557515	496938
Nashville-	TN																			448003	455651	488374
Denver	CO										35629	106713	133859	213381	256491	287861	322412	415786	493887	514678	492365	467610

You can download the LargestCitiesUS.csv file here.

Python scripting for CSV formatting:

1. The script will merge the “Place” and “State” columns into 1 column with the title “State-City”.
2. We transpose data about years from the header row to 1 column called “Year”.
3. We also transpose the data about population from the content rows to 1 column titled “Population”.
4. We add a new column called “Rank-Population” to rank the cities according to the population.

Following is the code for the same:

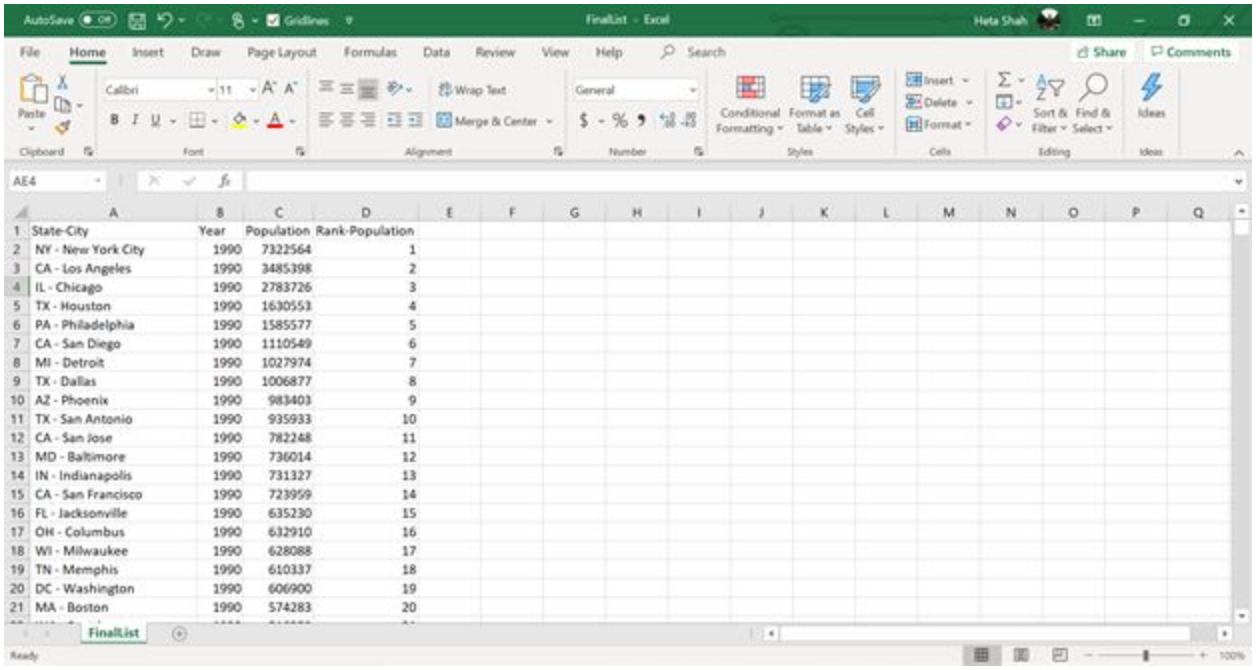
```
# Importing required libraries:
import csv
import operator
import copy
from token import EQUAL
from builtins import sorted

# Reading csv file (with comma as a delimiter) to get the Cities'
data from the absolute path:
reader =
csv.reader(open('C:\\Users\\heta2\\Desktop\\GitHub_Projects\\Power-Up
-with-MicrosoftPowerBI\\US Cities Growing
Population\\LargestCitiesUS.csv'), delimiter = ",")

# Initializing variables for columns in csv:
col1 = 'State-City'
col2 = 'Year'
col3 = 'Population'
col4 = 'Rank-Population'

# Code counter variables:
rows_so_far = 0
c = 0

# making an empty 2D data array for rows and columns:
pool = []
pool.append([])
```

```
# iterating over each row in csv file:
for row in reader:
      # if it's a header row
     if rows_so_far == 0:
           rows_so_far += 1
           header = row
            #populating headers
           for j in range (0,4):
                if j == 0:
                     pool.append([])
                     pool[0].append(col1)
                if j == 1:
                     pool[0].append(col2)
                if j == 2:
                     pool[0].append(col3)
                if j == 3:
                     pool[0].append(col4)

      # if it's a data row (not header)
     else:
           for i in range(len(row)-2):
                n = len(pool)
                if not row == []:
                     if i == 0 or i >= 1:
                          item = copy.deepcopy(row)
                          r = copy.deepcopy(row)
                          for j in range (0,4):
                               if item[i+2] is not '':
                                    if j == 0:
                                         r[0] = item[j+1]+' -
'+item[j]
                                         pool.append([])
                                         pool[n-1].append(r[0])
                                    if j == 1:

pool[n-1].append(int(header[i+2]))
                                    if j == 2:
                                         if item [i+2] == '':
```

```
pool[n-1].append(int(0))
                                                else:

pool[n-1].append(int(item[i+2]))
                                           if j == 3:
                                                pool[n-1].append(int(0))
rows_so_far += 1
# verify number of rows
n = len(pool)

# don't copy header as data
output = pool[1:n-1]

# descending sort data in csv according to Year and Population
output.sort(key = lambda d: (d[1],d[2]), reverse = True)

# make an empty 2D array for rows and columns
outdata = []
outdata.append([])
outdata[0] = pool[0]
outdata[1:n-1] = output[0:n-2]

# Writing formatted data to csv file for use in MS Power BI:
mycsv =
csv.writer(open('C:\\Users\\heta2\\Desktop\\GitHub_Projects\\Power-Up
-with-MicrosoftPowerBI\\US Cities Growing Population\\FinalList.csv',
'w', newline=''))
for row in outdata:
     e = outdata.index(row)
     if row[1] != c and e != 0:
          v = 1
          c = row[1]
          row[3] = v
     else:
          if row[1] == c and e != 0:
               v += 1
               row[3] = v

     mycsv.writerow(row)
```

Snapshot of the formatted “FinalList.csv”:

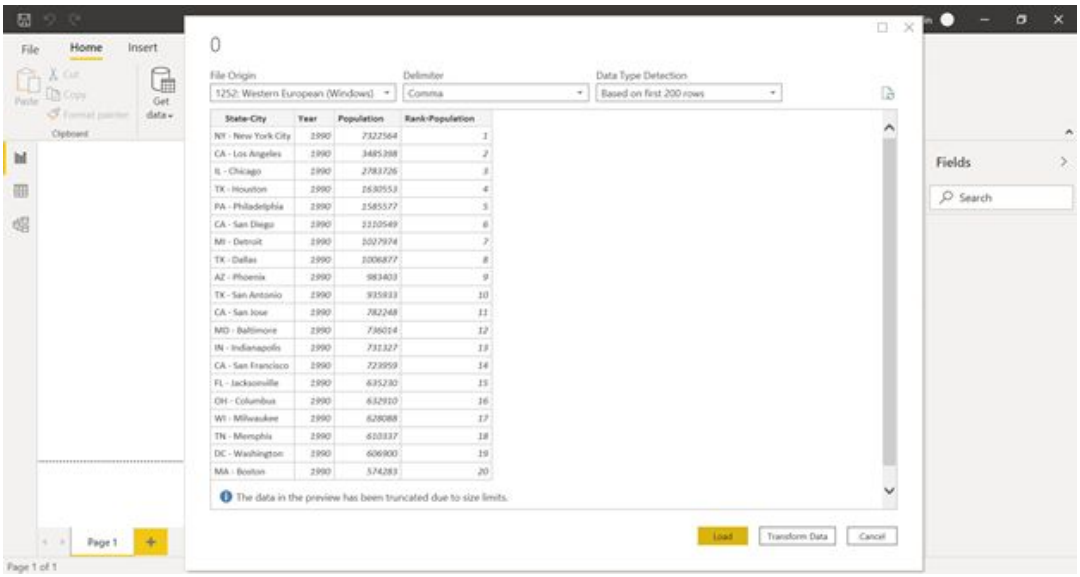
The formatted csv looks as follows:



	A	B	C	D
1	State-City	Year	Population	Rank-Population
2	NY - New York City	1990	7322564	1
3	CA - Los Angeles	1990	3485398	2
4	IL - Chicago	1990	2783726	3
5	TX - Houston	1990	1630553	4
6	PA - Philadelphia	1990	1585577	5
7	CA - San Diego	1990	1110549	6
8	MI - Detroit	1990	1027974	7
9	TX - Dallas	1990	1006877	8
10	AZ - Phoenix	1990	983403	9
11	TX - San Antonio	1990	935933	10
12	CA - San Jose	1990	782248	11
13	MD - Baltimore	1990	736014	12
14	IN - Indianapolis	1990	731327	13
15	CA - San Francisco	1990	723959	14
16	FL - Jacksonville	1990	635230	15
17	OH - Columbus	1990	632910	16
18	WI - Milwaukee	1990	628088	17
19	TN - Memphis	1990	610337	18
20	DC - Washington	1990	606900	19
21	MA - Boston	1990	574283	20

Visualization using MS Power BI:

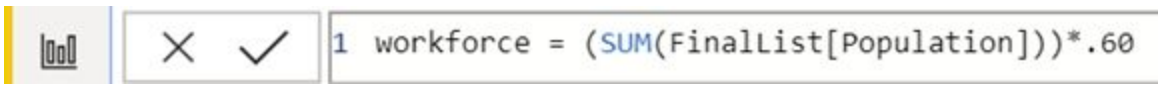
Launch MS Power BI desktop and import “FinalList.csv” using “Get Data -> Text/CSV”. Browse for the location where the new csv file is stored and select it. On the landing screen that opens, click on “Load”:



You can verify the import by switching to Data view from Report view.

In the Report view, select “Slicer” from the Visualizations section and then select the “Rank-Population” column from the Fields section. This slicer allows us to filter based on Rank and shows values from 1-100 by default:

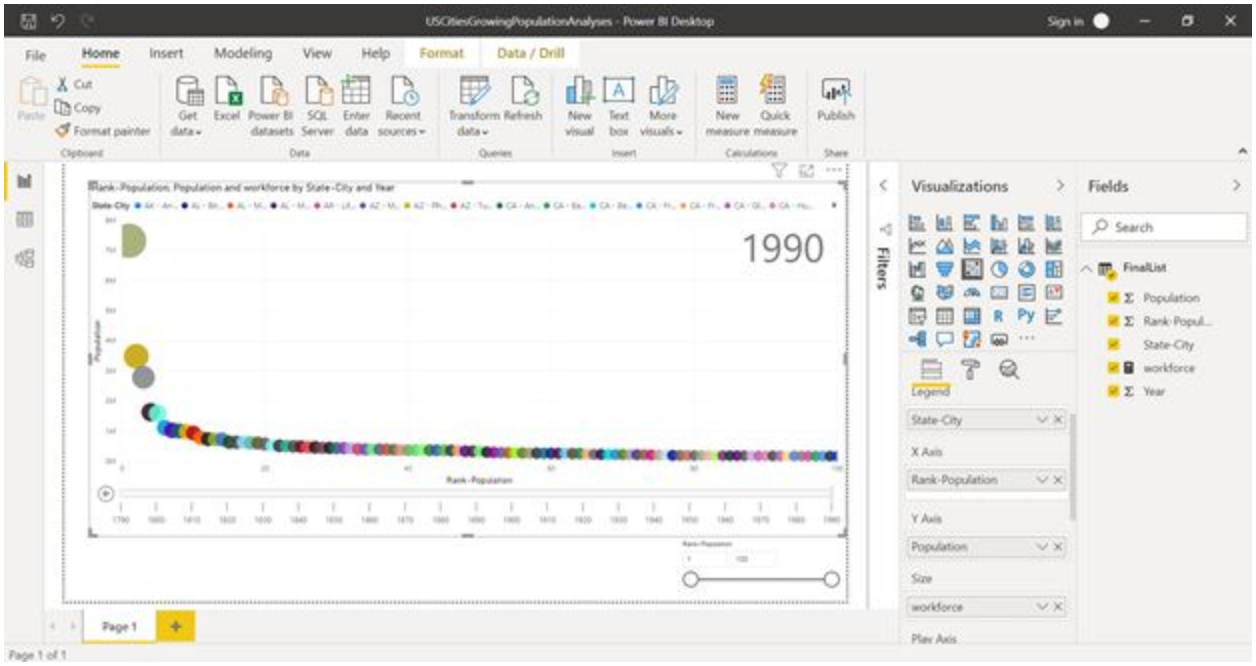
Creating a new measure to calculate “Workforce” – From the fields list, click on the ellipses next to FinalList and click “New Measure”. In the formula bar that opens, type the following formula: `workforce = (SUM(FinalList[Population]))*.60`



Preparing the Scatter chart – Select the “Scatter Chart” from the Visualizations pane.

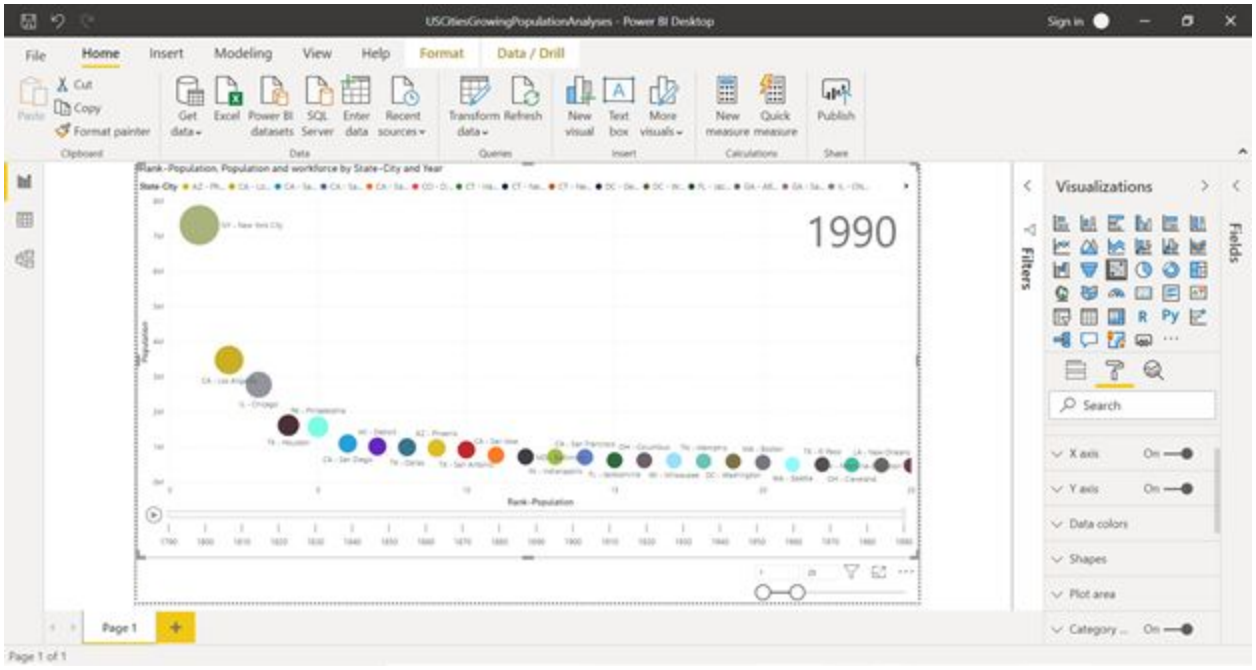
1. Drag the “Rank-Population” from fields to “X-axis” (under visualizations pane)
2. Drag “Population” from fields to “Y-axis”
3. Drag “State-City” from fields to “Legend” section
4. Drag “workforce” from fields to “Size” section
5. Drag “Year” from fields to “Play” section

The resulting visualization should look as follows:



The bubbles on the visualization represent each city.

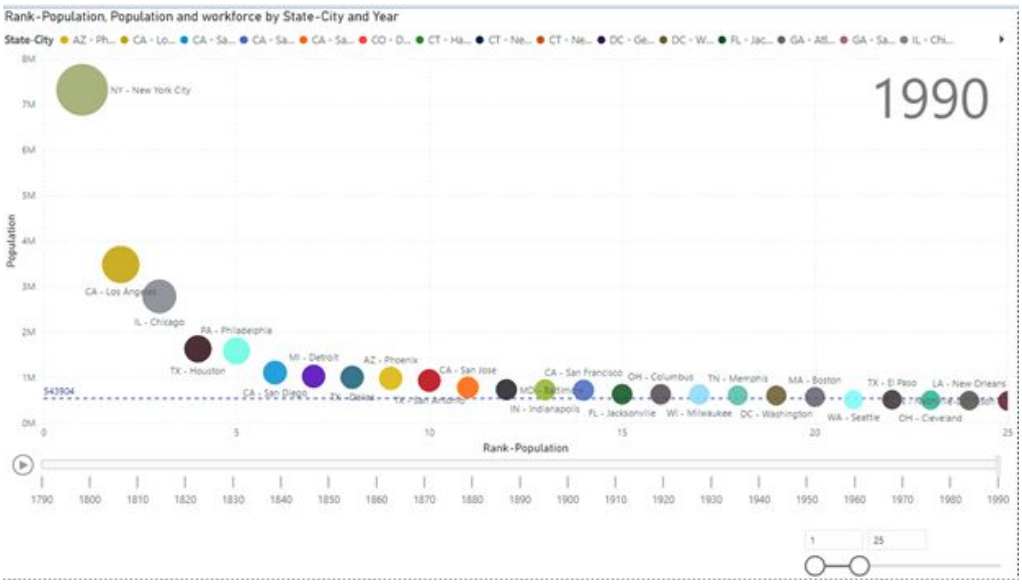
However, to make the visualization more descriptive, reduce the ranks slider to top **25**. Then, under the “Format” section (below visualizations pane), toggle “Category Labels” on. Following is the difference:



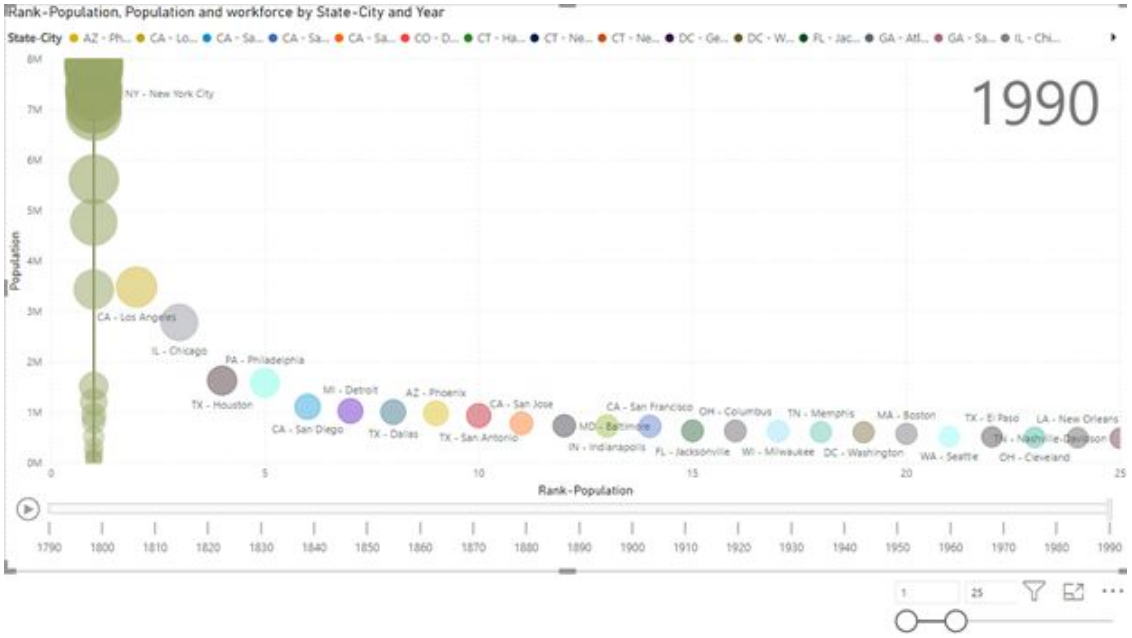
Each bubble now has a name of the city associated with it. Hit the Play button at the bottom left corner of the scatter chart to see the bubbles in motion. You can check out this video to see the workbook.

Analyses:

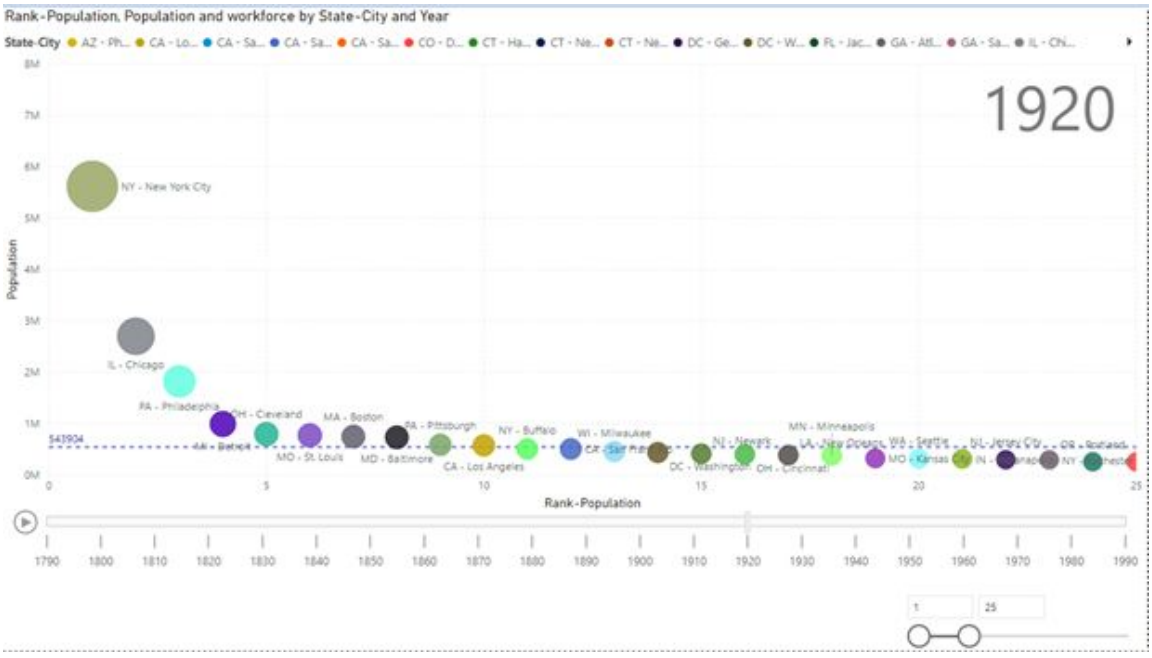
1. The average population (indicated by the dashed blue line) of these US cities is ~0.5 million people:



2. As seen, New York City's population has been only growing between 1790 – 1990. Since then, it is also called "The city that never sleeps", "The Center of the Universe" and "The Big Apple":



3. The beginning of the 20th century marked an increase in the population of most US cities since World War 1 ended. This flourished a lot of commerce and culture, hence, attracting more crowd and attention:



4. Other cities that saw an increase in population by 1990 belonged to the states of California, Texas and Illinois. These states supported a large growth in business.
5. Mortality rates were growing in the cities of Baltimore (MD), Boston (MA), New Orleans (LA) and Philadelphia (PA) due to pneumonia, tuberculosis (TB) and diarrhea in the 19th and the 20th century.