Web scraping & Animated Data Visualization

Aim:

Step 1: Web scraping: Extracting the information on 'US City and Population' from website using Python

Step 2: Data cleaning

Step 3: Animated PowerView based visualization of data

Prerequisites:

a. Notepad++
b. Python IDE/Command Prompt
c. MS Excel(Power view enabled)

Dependencies:

a. Python version 2.7.9
b. BeautifulSoup
   The scraper used Python's BeautifulSoup toolkit to parse the site's HTML and extract the data. Also, the Requests library was used to open the URL, download the HTML and pass it to BeautifulSoup.
c. easy_install
   For automatic download, build, install, and manage Python packages.
d. lxml
   For processing HTML/XML

Process

Step 1: The data was available on the website in data table. Using **Python program** named 'Web data Scraping', data was scraped and stored in comma separated file named 'Largest_Cities_CSV.csv' The python script is available in the repository.

The preview of output:

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 Angel	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
Philadelpl	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 Antor	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

Step 2: As can be seen in the image, there are many blanks and data needs to be formatted to make it appropriate for power view. Using python program named 'Data Preprocessing' Following changes were committed in the scraped CSV data:

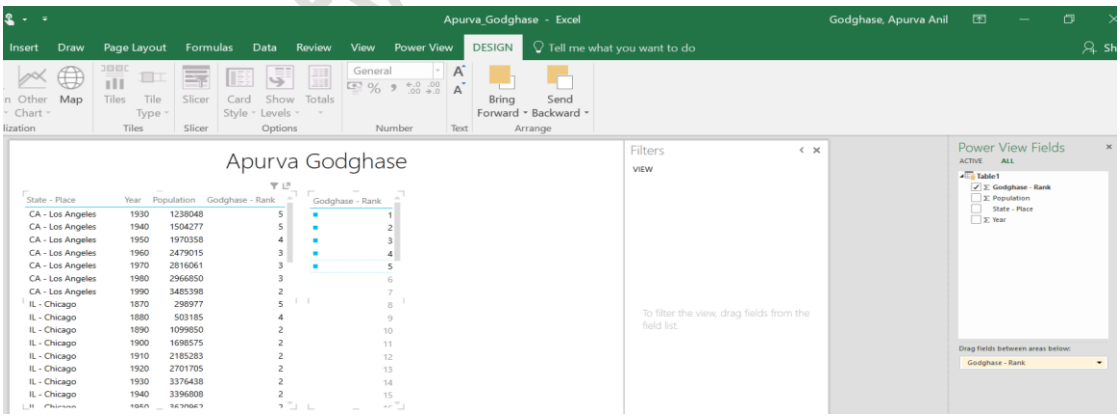
- Merge the columns 'Place' and 'State' and make it one column with format 'State-Place'
- Transpose 'Years and Population' i.e. after running this python code the newly created csv file will have 'Years' and 'Population' as columns instead of rows
- A new column 'Student Last Name - Rank' will be created which will rank the 'State-City' based on population.

The output looks like follows:

State - Place	Year	Populatior	Godghase - Rank
NY - New York City	1990	7322564	1
CA - Los Angeles	1990	3485398	2
IL - Chicago	1990	2783726	3
TX - Houston	1990	1630553	4
PA - Philadelphia	1990	1585577	5
CA - San Diego	1990	1110549	6
MI - Detroit	1990	1027974	7
TX - Dallas	1990	1006877	8
AZ - Phoenix	1990	983403	9
TX - San Antonio	1990	935933	10

Power view can be created in Excel not in CSV so the output file is saved as Excel Workbook(*.xlsx) and the file is formatted in tabular form.

Step 3: By default, the year & population columns are summarized, in order to get the original values for these columns, 'do not summarize' was selected for the year and population columns. Same for the 'Rank column'. Slicer was added to select any specific entry in the rank column, as it filters the data in the table based on that rank. Then, 5 largest population against a year are displayed.



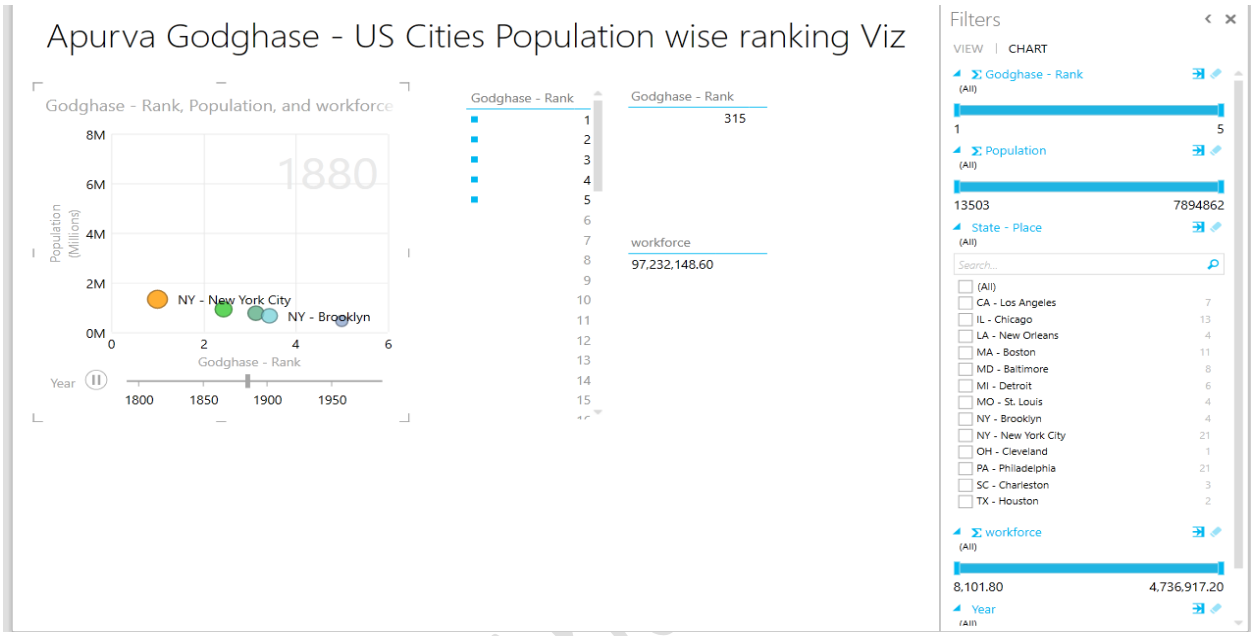
A new calculated field named 'Workforce 'is created and added to the canvas

Formula:

=(SUM(PopulationData[Population]))*.60

Animated Power view bubble charts are displayed as follows:

Between 1850 to 1900



Between 1900 and 1950