PULITZER CLUSTERING USING MACHINE LEARNING

Final Report - 7th Jan 2018
By: Subir Das

Objective, Clients and Goal

Objective:

As we all know that "Democracy Dies in Darkness[1*]" and "Truth is hard to Find[2*]" are the cornerstone of the economy. With this mind, the goal is to identify different types of Newspapers Segments based on Pulitzer prize and then identify ways to increase daily circulations by improving visibility and gaining new insights. It may also act as a catalyst for further boosting readers confidence in the print media.

Clients:

Print Media like New York Times (www.nytimes.com) and Machine Learning Community.

Goal:

Increase daily circulations and gaining new insights by identifying patterns thus further improving readers' confidence in the print media.

Credit and Copyrights:
1 – Washington Post and 2– New York Times hold the respective copyright and credits.
* Source Internet

Data Source

The data for this project is collected from different internet resources as listed below. The base data also called raw data ("Pulitzer Dataset") is made available by FiveThirtyEight. It is a simple table mapping newspaper with daily circulation and Pulitzer. Pulitzer Dataset is combined with many other dataset in order to find socioeconomic correlations. Additional data source and their formats are described below.

<u>Additional Data Sources used</u>:

1) Raw Data is made available by FiveThirtyEight.
   https://github.com/fivethirtyeight/data/blob/master/pulitzer/pulitzer-circulation-data.csv

2) Crime Data by State and US: The idea is to check if there is any correlation between crime rate and Pulitzer. The dataset is made available by US Departments.
   a. http://www.usa.com/rank/us--crime-index--state-rank.htm: The data is a simple excel sheet mapping crime index to US state by population.
   b. https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-1 : This is also an excel sheet mapping year, US population to crime index and types of crime

Data Contd…

Additional Data Sources used:

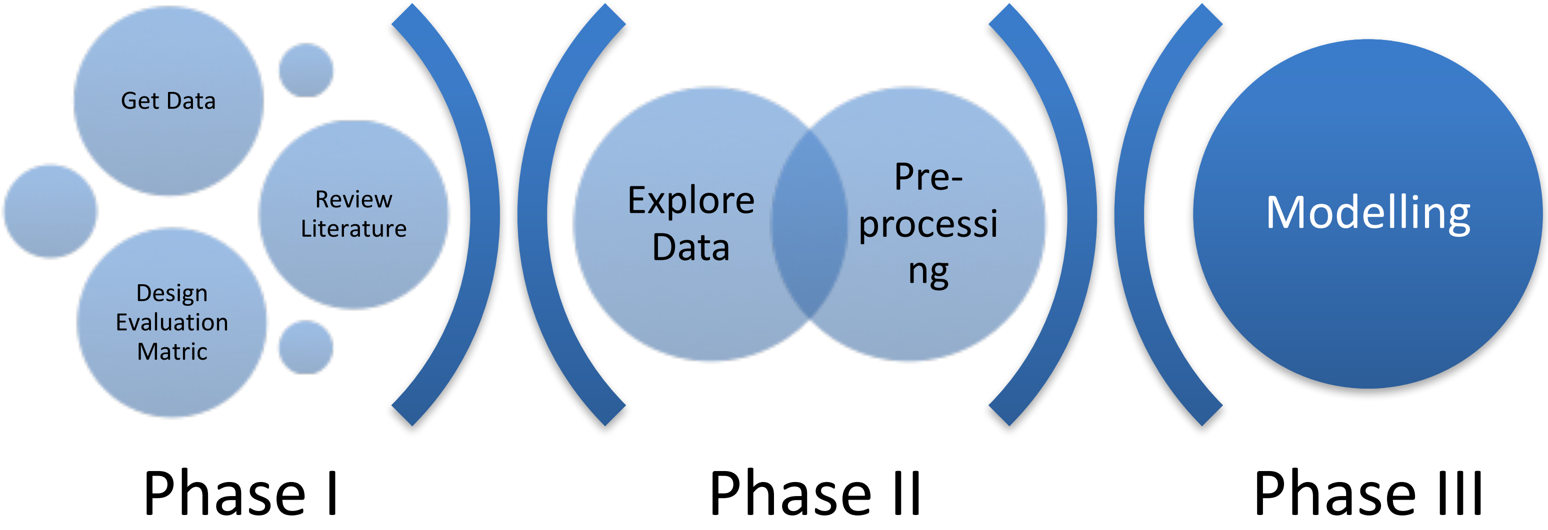
3) GDP for US and States: https://www.usgovernmentrevenue.com/download_multi_year_2000_2014USb_19c1li101mcn_F1cF0t: This is a CSV file mapping GDP by state by year dataset. We have 50+1+1 such files.

4) Population by State and US: https://www.census.gov/data/datasets/2016/demo/popest/state-total.htm : This is also an excel based data file mapping US stats to population to year. We have two files to cover 22 decades i.e. 1 file per census.

Solution Approach

Machine Learning: K-Means and PCA to help with the dimension reduction and newspapers' segmentation.

The solution approach is specifically designed to achieve a controlled and structured approach to minimize data quality issues that may be present or introduced to protect personally identifiable information(PII). The solution is sub-divided into three phases as listed below.



Solution Approach…

1. Data Assembly - Phase I: This phase of the project is designed to gather and do basic cleanup like join, merge, add or update attributes.

2. Explore and Preprocessing – Phase II: This phase of the project is designed to validate and explore the dataset for all the problems listed in the “Problem” section of this proposal.

3. Modelling and Evaluation Phase III: In this phase of the project I will be exploring various machine learning algorithms to find the best model to cluster the newspapers.

Technology Stack: Cassandra, Google Cloud Platform, Python, MS-Excel, Atom
Configuration Management System: GitHub
Testing and Traceability: IPython Notebook

Data Wrangling – Phase I

Final Dataset:

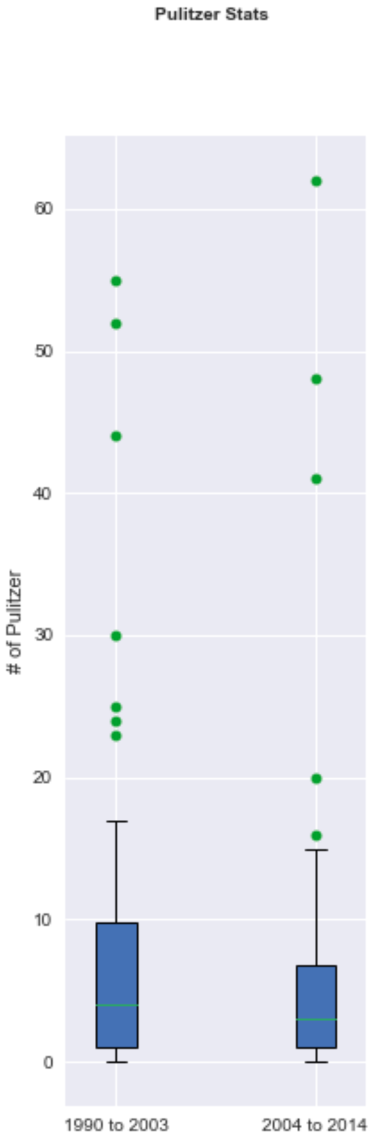
The final Pulitzer dataset went through industry standard data validation and verification(V&V) along with transformation before it was used for exploration and pre-processing. The final dataset was created after merging and transforming 58 datasets. For Audit, timestamp and unique identification(UUID) was added to all the 58 pandas dataframes. And to make it persistent and reproducible, domain level information was stored in Cassandra cluster running in Google Cloud. Please see the below list for different type of operations as performed.

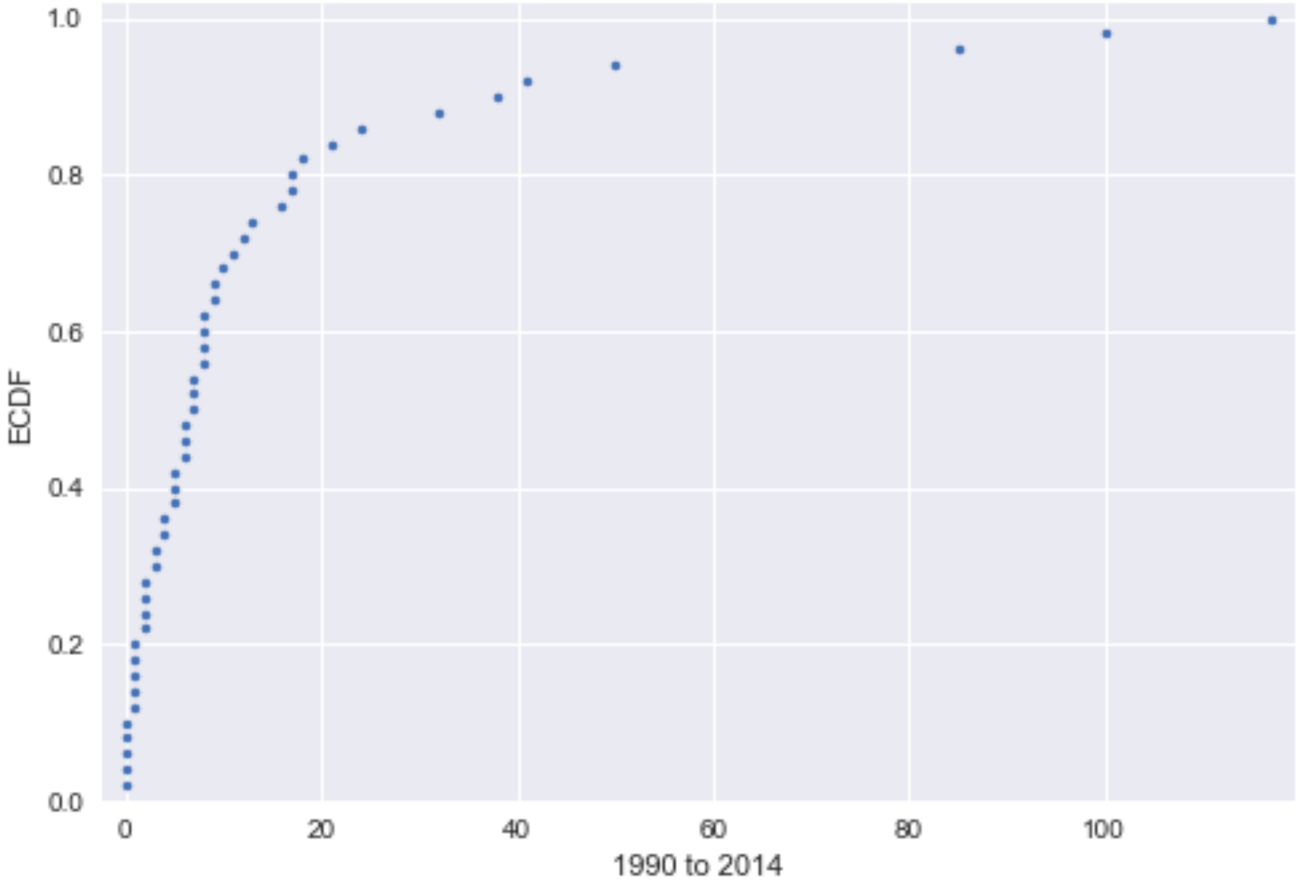
- Pulitzer Dataset (1 data table) : Datatype Conversion, Removing special char {, % / \} , pandas.join|merge|concat|pivot
- Crime Data (2 Data tables): Missing Values, Datatype Conversion, Column Hacks
- GDP for US and States (53 Data tables) : CSV Bad Lines, Missing Data, Column Hacks, pandas.join|merge
- Population by State and US (2 Data tables) : Selective columns reading, Datatype Conversion

Wrangling Highlights:

1) Google Cloud Platform and Multi Threading to handle massive volume of data
2) Cassandra Storage for Persistency and Reproducibility
3) Audit details for Traceability

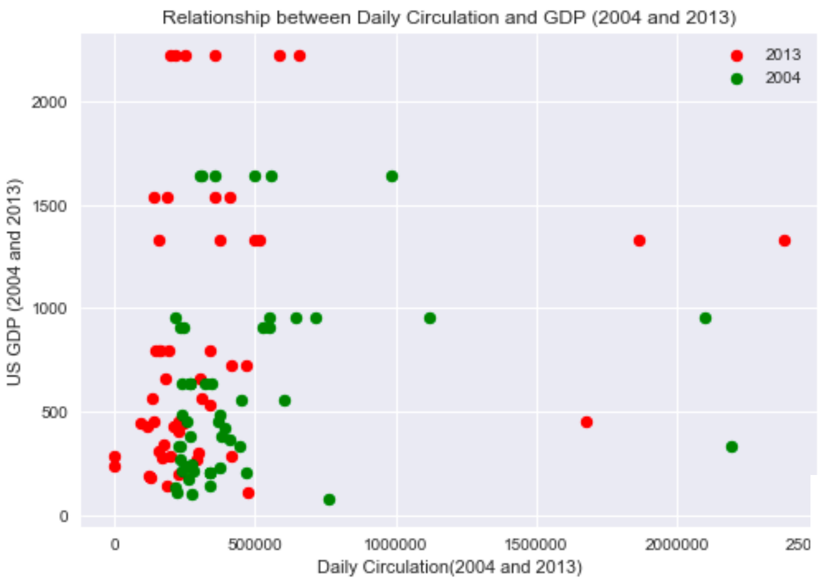
Data Exploration – Phase II

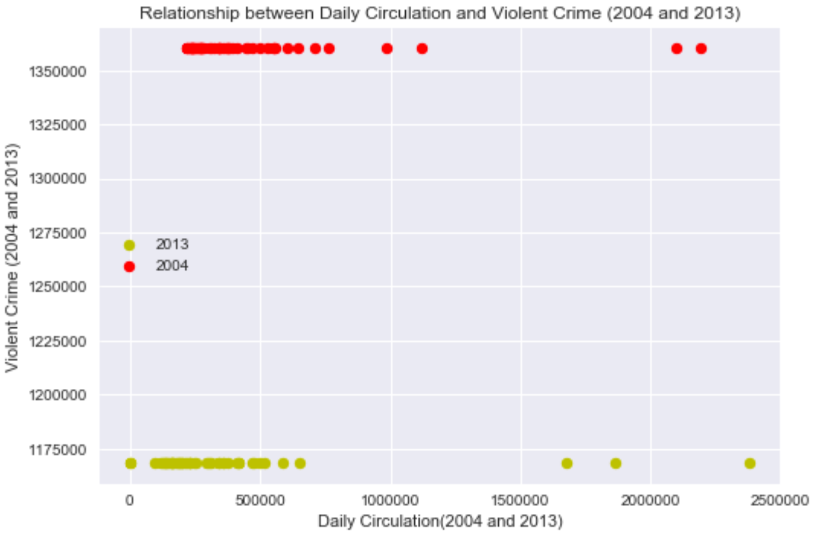


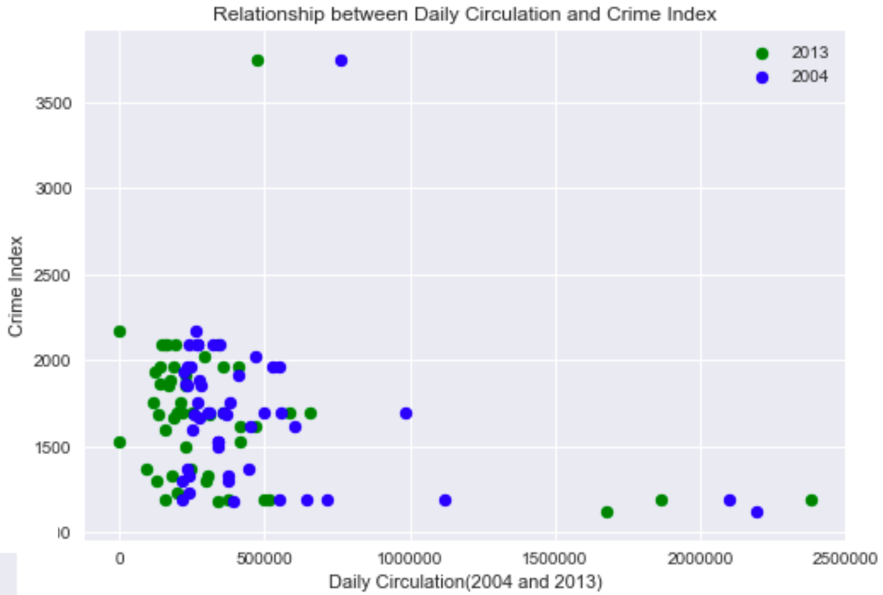


Based on initial data review, it appears that we have outliers and Pulitzer data tends to follow a 80-20 rule i.e. 80% of the newspaper population is under 10 Pulitzer Prizes and 20% of the population is above 10 Pulitzer Prizes.

Data Exploration – Phase II

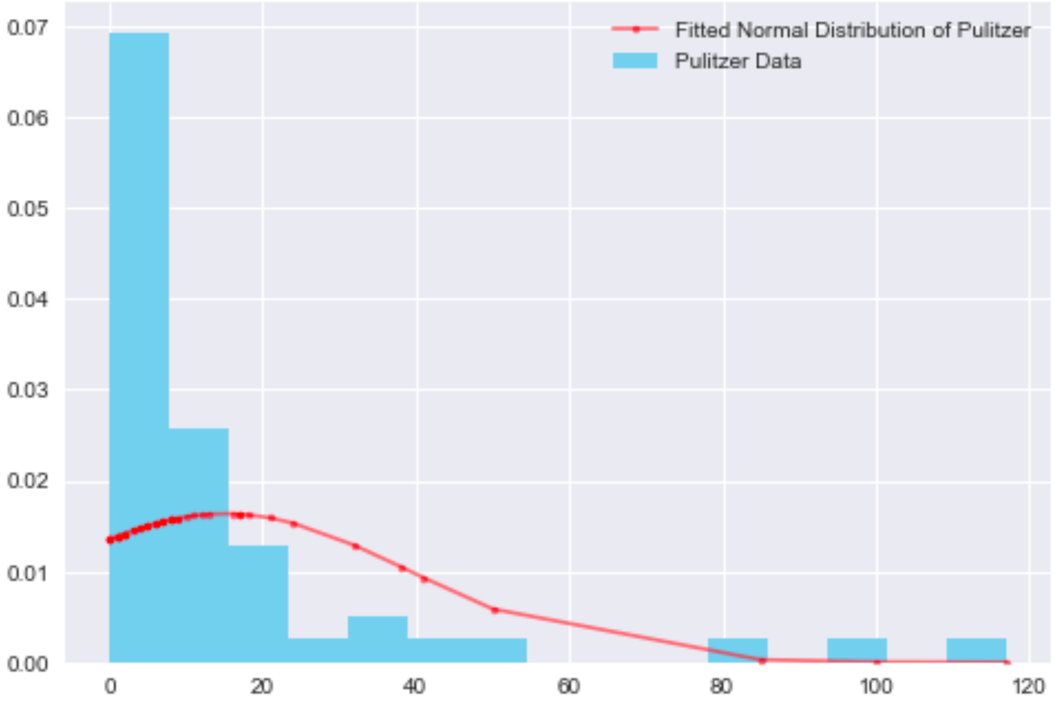


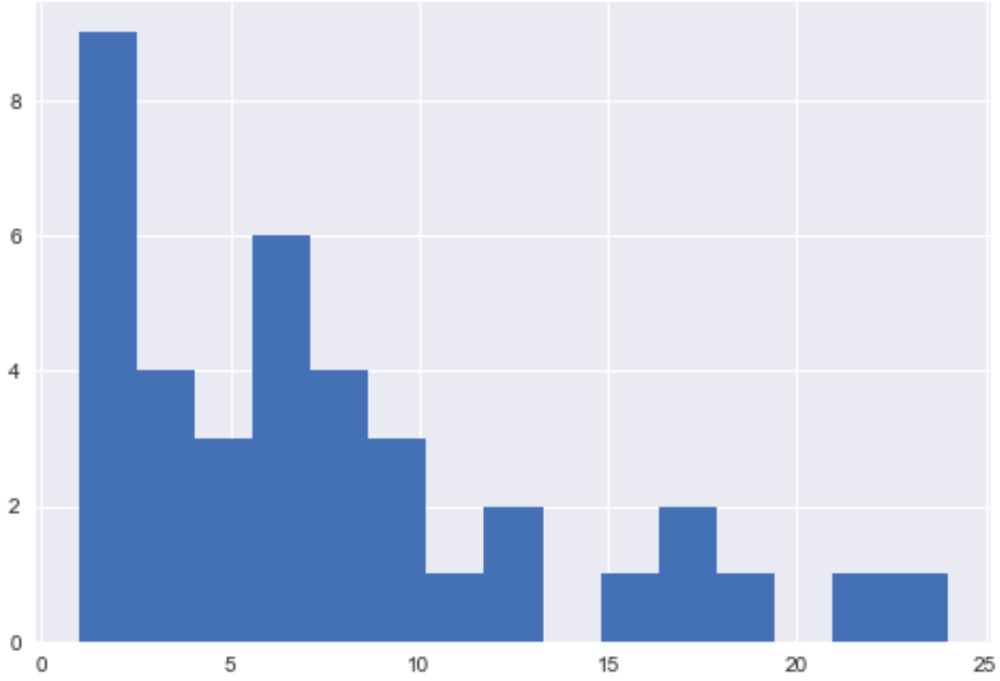




It appears that we have positive relationship between Crime Index, US GDP and Daily Circulation not with Violent Crime.

Data Exploration – Phase II





Summary Statistics:

1. Mean : 15.06
2. Variance : 595.0164
3. Skewness: 2.8072431554058634
4. Kurtosis : 7.563560952747379
5. Median: 7.0
6. pValue: 1.5546437888e-11

Observation: It looks like that Pulitzer Data is not a normal distribution. This makes sense as Pulitzer is a highly respected price.

Modeling – Phase III – K- Means and PCA

The data was prepared by **removing commas, changing datatypes and US state name to two digit code. Please see the below snips for details.**

Before pre-processing:

In [4]: `df_pulitzer.sample(7)`

Out[4]:

	newspaper	state	dailycirculation_2004	dailycirculation_2013	changedailycirculation_2004_2013	winnfinalists_1990_2003	winnfinalists_2004_2014	winnfin
47	Columbus Dispatch	Ohio	259,127	137,148	-47%	1	0	
8	New York Post	New York	642,844	500,521	-22%	0	0	
29	Newark Star Ledger	New Jersey	395,000	340,778	-14%	2	6	
30	Chicago Sun-Times	Illinois	453,757	470,548	4%	1	1	
20	San Jose Mercury News	California	558,874	583,998	4%	4	2	
6	Chicago Tribune	Illinois	603,315	414,930	-31%	23	15	
33	Minneapolis Star Tribune	Minnesota	377,058	301,345	-20%	4	4	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 18 columns):
newspaper                          50 non-null object
state                              50 non-null object
dailycirculation_2004              50 non-null object
dailycirculation_2013              50 non-null object
changedailycirculation_2004_2013   50 non-null object
winnfinalists_1990_2003            50 non-null object
winnfinalists_2004_2014            50 non-null object
winnfinalists_1990_2014            50 non-null object
gdp_2004                           50 non-null object
gdp_2013                           50 non-null object
gdp_2014                           50 non-null object
crimeindex                         50 non-null object
violentcrime_2004                  50 non-null object
violentcrime_2013                  50 non-null object
violentcrime_2014                  50 non-null object
popestimate2004                    50 non-null object
popestimate2013                    50 non-null object
popestimate2014                    50 non-null object
dtypes: object(18)
memory usage: 7.1+ KB
```

After pre-processing:

In [8]: `df_pulitzer.sample(7)`

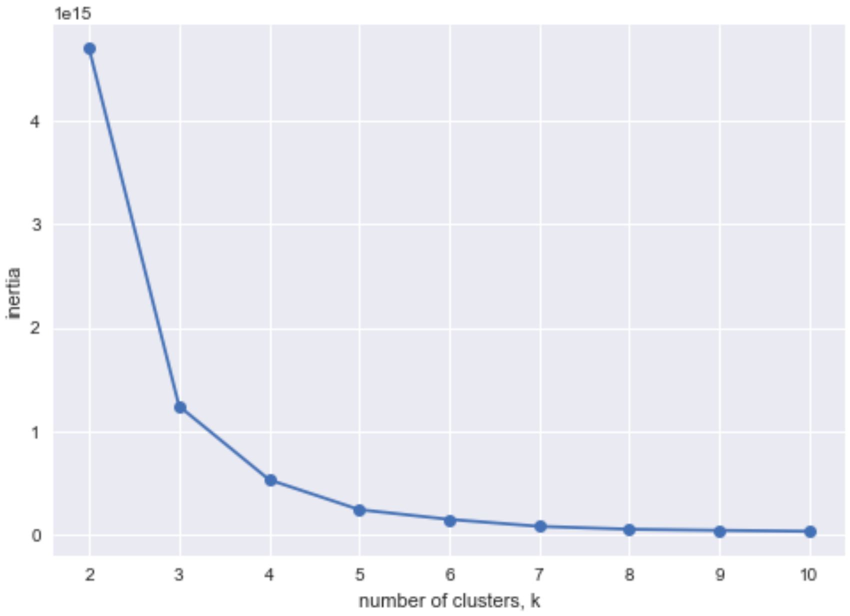
Out[8]:

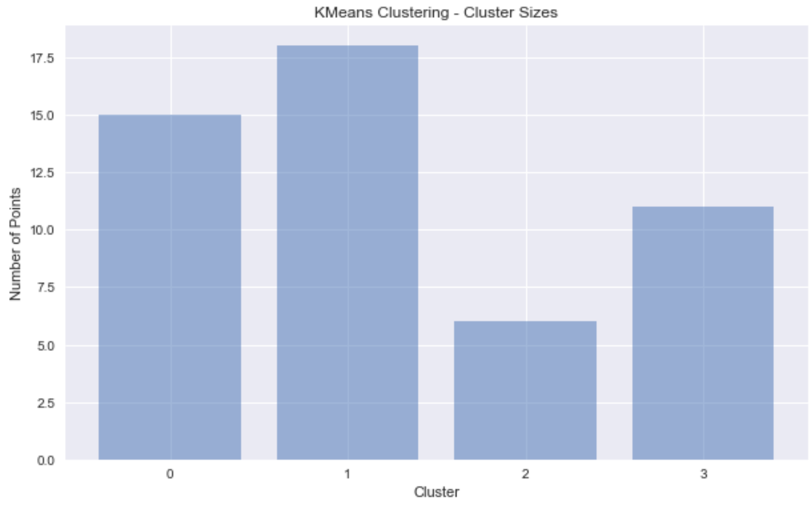
	newspaper	state	dailycirculation_2004	dailycirculation_2013	changedailycirculation_2004_2013	winnfinalists_1990_2003	winnfinalists_2004_2014	winnfinalists
0	Orlando Sentinel	FL	269269.0	161070.0	-40%	5.0	2.0	
5	Fort Woth Star-Telegram	TX	237318.0	188593.0	-21%	1.0	0.0	
37	New Orleans Times-Picayune	LA	262008.0	0.0	-100%	5.0	3.0	
19	Charlotte Observer	NC	231369.0	137829.0	-40%	1.0	3.0	
21	Tampa Tribune	FL	238877.0	191477.0	-20%	0.0	0.0	
40	St. Louis Post-Dispatch	MO	281198.0	167199.0	-41%	4.0	3.0	
44	Newsday	NY	553117.0	377744.0	-32%	12.0	6.0	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 18 columns):
newspaper                          50 non-null object
state                              50 non-null object
dailycirculation_2004              50 non-null float64
dailycirculation_2013              50 non-null float64
changedailycirculation_2004_2013   50 non-null object
winnfinalists_1990_2003            50 non-null float64
winnfinalists_2004_2014            50 non-null float64
winnfinalists_1990_2014            50 non-null float64
gdp_2004                           50 non-null float64
gdp_2013                           50 non-null float64
gdp_2014                           50 non-null float64
crimeindex                         50 non-null float64
violentcrime_2004                  50 non-null float64
violentcrime_2013                  50 non-null float64
violentcrime_2014                  50 non-null float64
popestimate2004                    50 non-null float64
```

K-Means and PCA to help with the newspapers' segmentation and dimension reduction.

Modeling – Phase III – K- Means and PCA





K-Means and PCA to help with the newspapers' segmentation and dimension reduction.

K-Means: We plan to use the K-Means Clustering to maximize the distance between centroids and minimize the distance between data points and the respective centroid for the cluster they are in.

```
from sklearn.cluster import Kmeans
Ks=range(2,11)
inertias = []
for k in Ks:
    model = KMeans(n_clusters=k)
    model.fit(X)
    inertias.append(model.inertia_)
```

Observation:

We can see that between K=4 and K=5 the error did not drop significantly. We can safely select K=4 as our best case value of k for K-Means clustering algorithm.

Modeling – Phase III – K- Means and PCA

PCA: We plan to use the same number of components as K-Means i.e. 4.

```
from sklearn.decomposition import PCA
pca = PCA(n_components =4)
pca.fit(X)
pca_features=pca.transform(X)

df_pulitzer['x']=pca_features[:,0]
df_pulitzer['y']=pca_features[:,1]

df_result=df_pulitzer.groupby(['cluster','newspaper'])['state'].count().unstack('cluster')
df_result = df_result.fillna(0).reset_index()
```

cluster	newspaper	0	1	2	3
0	Arizona Republic	0.0	1.0	0.0	0.0
1	Atlanta Journal Constitution	0.0	0.0	0.0	1.0
2	Baltimore Sun	0.0	1.0	0.0	0.0
3	Boston Globe	0.0	1.0	0.0	0.0
4	Boston Herald	0.0	1.0	0.0	0.0
5	Charlotte Observer	0.0	0.0	0.0	1.0
6	Chicago Sun-Times	0.0	0.0	0.0	1.0
7	Chicago Tribune	0.0	0.0	0.0	1.0
8	Cleveland Plain Dealer	0.0	0.0	0.0	1.0
9	Columbus Dispatch	0.0	0.0	0.0	1.0
10	Daily Oklahoman	0.0	1.0	0.0	0.0
11	Dallas Morning News	1.0	0.0	0.0	0.0
12	Denver Post	0.0	1.0	0.0	0.0
13	Detroit Free Press	0.0	0.0	0.0	1.0
14	Detroit News	0.0	0.0	0.0	1.0
15	Fort Woth Star-Telegram	1.0	0.0	0.0	0.0
16	Houston Chronicle	1.0	0.0	0.0	0.0
17	Indianapolis Star	0.0	1.0	0.0	0.0

<u>Conclusion</u>:
We have mapped the newspapers into 4 clusters based on Crime, GDP and Population.

Summary and Next Steps

Summary:

The clustering model was designed to produce cluster of newspapers based on Crime, GDP, Population, Daily Circulation and Pulitzer. It was tested and statistically proved that Pulitzer price do not follow a Normal Distribution, which makes perfect sense else it will be easy to predict prices. The clustering of the newspaper seems to suggest that print media clusters try to form around highly dense population.

1. A cluster size of 4 for K-Mean was our best size.
2. A component of 4 was our best size for PCA
3. Pulitzer Data is NOT normally distributed

Next Steps:

The identified newspaper cluster model needs analysis using Graph(network analysis) for finding the economic zones. It should be further be extended to cluster newspaper Article SOURCE for pattern recognition followed by a ML classification (chaining model outputs/logistic regression) to boost news quality.