**Assignment 4**

**Question: PREPROCESS THE DATASETS given along with this assignment**

1. **Identify the type of each data attribute. Indicate if the type of attribute is Nominal, Ordinal, Interval or Ratio**

***For Dataset 1 (movies)***

| **Attribute** | **Type** | **Reason** |
| --- | --- | --- |
| **Year** | **Interval** | Since there is a difference between two values is meaningful and hence this attribute lies in the Interval Type and can never be 0 |
| **Length** | **Ordinal** | Since it is a measurable attribute and hence this attribute lies in the Ordinal Type |
| **Title** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Genre** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Actor** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Actress** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Director** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Popularity** | **Ratio** | Since the attribute has all the properties of an interval variable, and also has a clear definition of 0.0 and hence this attribute lies in the Ratio Type |
| **Awards** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |
| **Image** | **Nominal** | Since the attribute is not ordered and it lies in the category of a group of something and hence this attribute lies in the Nominal Type |

***For Dataset 2***

| **Attribute** | **Type** | **Reason** |
| --- | --- | --- |
| **Column1** | **Ratio** | Since the attribute has all the properties of an interval variable, and also has a clear definition of 0.0 and hence this attribute lies in the Ratio Type |
| **Column2** | **Ratio** | Since the attribute has all the properties of an interval variable, and also has a clear definition of 0.0 and hence this attribute lies in the Ratio Type |

**ii. Load the given dataset into Spark Data-Frames and perform the following preprocessing on it.**

1. **Handle missing values**

***For Dataset 1 (movies)***After checking for NULL values in the dataframe we have found out that "Length" has 67 null values, "Genre" has 2 null values, "Actor" has 8 null values, "Actress" has 378 Null values, "Director" has 253 Null values and "Popularity" has 6 null values, the rest has no null values.

I impute those null values using different techniques.

1. ***Length column:***

Since the movies with the same genre has almost the same length on average, so I have decided to take the mean of length of each Genre and impute the values for the Length Attribute for the Genre it has.

1. ***Genre column:***

Genre has 2 missing values here I will check the Director and mode of the movies he/she directs and put it there

1. **Actor**

Actor has 8 missing values here I will check the Genre where the Actor is same and for other values and put it in there

1. **Actress**

#### Actress has 378 missing values here I will check with the same technique as used for Actor and then see what can we do to impute the values first

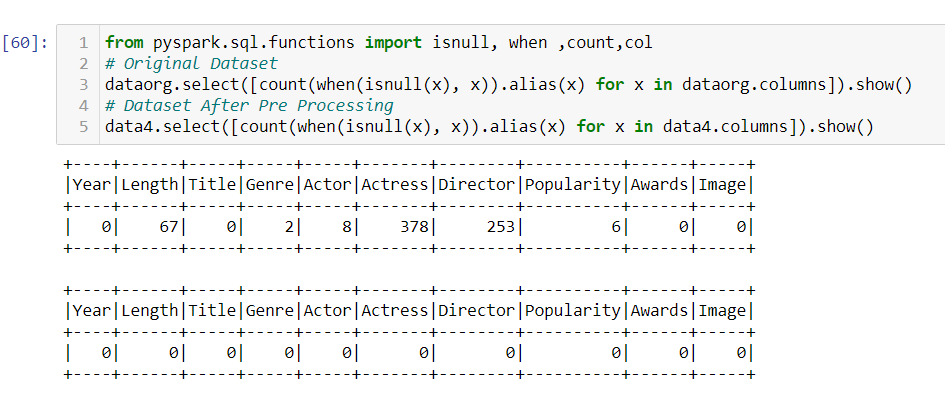
1. **Director**

#### Director has 253 missing values here I will check with the same technique as used for Actor and then see what can we do to impute the values first

1. **Popularity**

Popularity has only 6 missing values so here I will purge the values with the mean of Popularity

# New data at end of this part has no null values as Shown in fig below(taken from code):



***For Dataset 2***

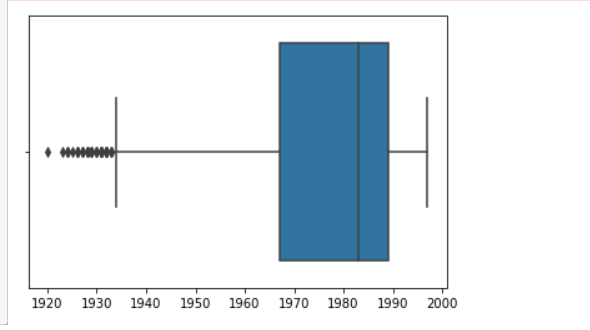
In dataset 2 there was no null value so no imputation was needed for this dataset.

***Note: I used too many partitions. So i did the remaining parts on the Pre Processed dataset on a new Session and perform rest of tasks on that dataset.***

1. **Identify if an attribute has outliers or noise**

To detect and see outliers in datasets I used two techniques: Box Plot and Z-Score

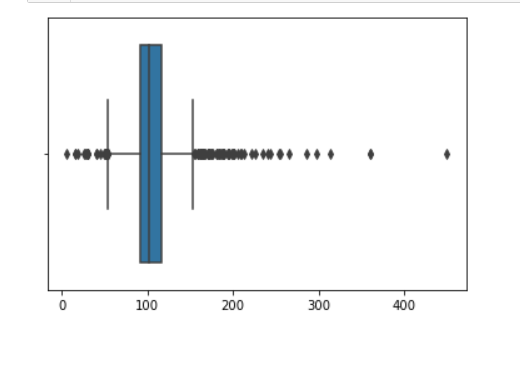
***For Dataset 1 (movies)  
  
1. Year Column***The **Box Plot** shows that the values before 1930 are clearly Outliers.



And By **Z-score Method** I found out these Outlier for Year [1920, 1923, 1924, 1924, 1924]

1. ***Length Column***

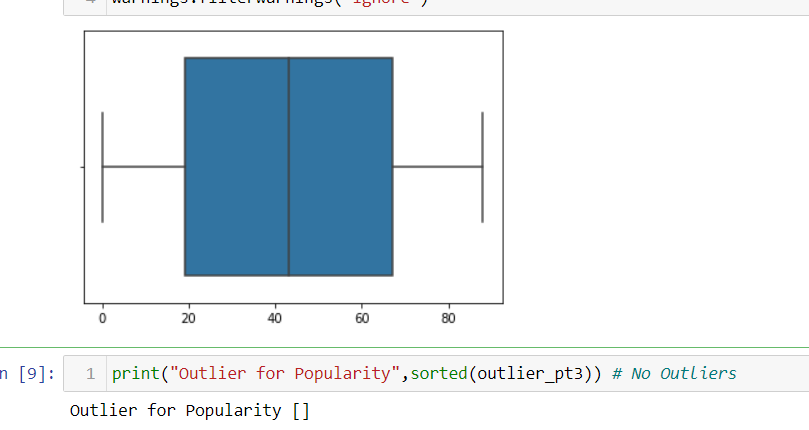
The **Box Plot** shows that the value between 90 to approximately 120 are not outliers the dotted shows that these are outliers for example 5,15, 200 all of them are outliers



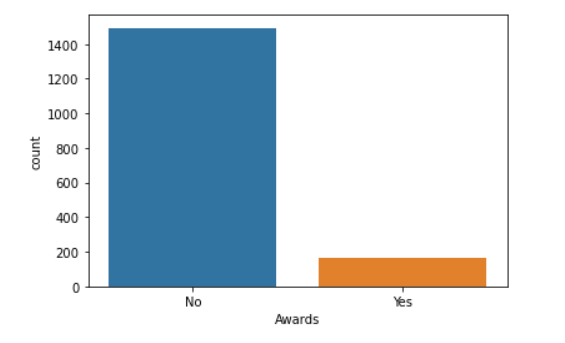
And By **Z-score Method** I found out these Outlier for Length [5, 15, 198, 199, 200, 201, 201, 206, 208, 210, 213, 221, 226, 234, 240, 243, 253, 255, 265, 286, 298, 313, 360, 360, 450]

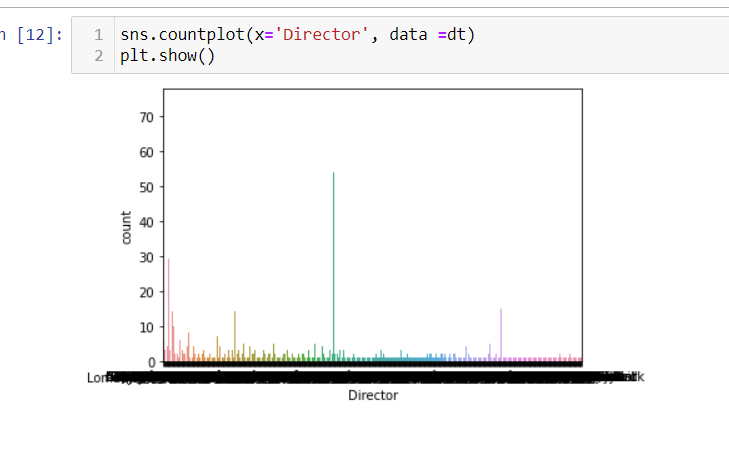
1. **Popularity Column**

Both methods shows that there is no single outlier in popularity. One may see in snippet given below:

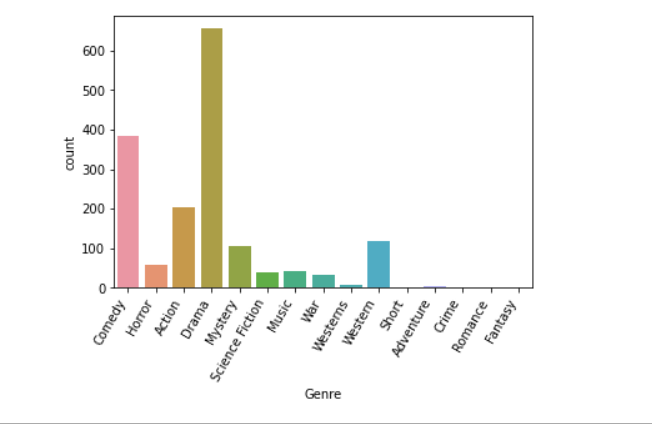


**For rest of columns given below I show outliers using sns plot**

1. ***Awards column  
   ***
2. ***Director column***

******

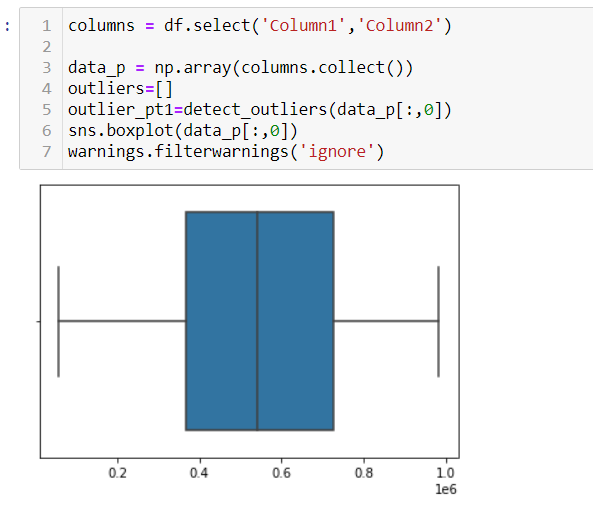
1. ***Genre column***



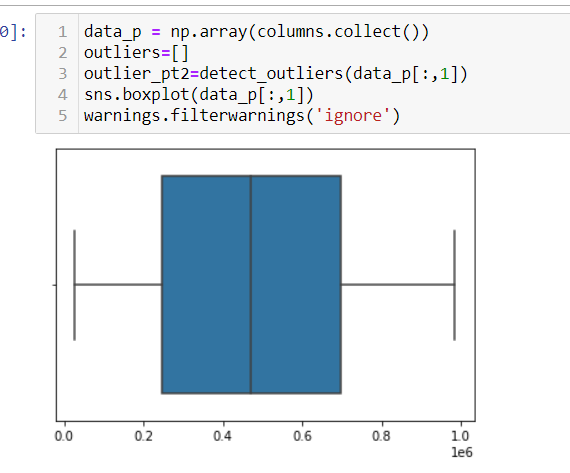
***For Dataset 2***

In both columns of dataset two there were no outliers in dataset. One may see my output for detecting outliers in figures below:

**For Column 1**



**For Column 2**

****

**c) Apply measures of the central tendency and dispersion to analyze numeric attributes. That is, compute**

**the mean, median, mode, range, variance,correlation for the attribute. Don’t just give values explain**

**analyze them.**

***For Dataset 1 (movies)***

***Mean:***

Mean is the average of dataset.

**Mean of Year is :** 1975.814121907061

**Mean of Length is :** 105**.**07724803862402

**Mean of Popularity is :** 43.090525045262524

**Median:**

The median is the middle of the set of numbers

**Median of Year is :** 1983.0

**Median of Length is :** 102.0

**Median of Popularity is :** 43.0

**Mode:**

Mode is the most repeated value in column(s)

**Mode of Year is** : [1991]

**Mode of Length is :** [90]

**Mode of Popularity is :** [8]

**Variance:**

Variance is a measure of how data points differ from the mean.

**Variance(Year)** 290.99682943682205

**Variance(Length)** 899.6896813110241

**Variance(Popularity)** 713.0811722483157

**Range:**

The difference of largest and smallest value in a feature set / column is range

**Range of Year is :** 77

**Range of Length is :** 445

**Range of Popularity is :** 88

**Correlation:**

Correlation Between Year and Length 0.031795051625372654

Correlation Between Year and Popularity -0.09051627627705547

Correlation Between Length and Popularity -0.009556346766471775

All numeric columns have very weak correlation, relation b/w (year, popularity) is weak and negative (which suggests that once one price decrease, different increase and vice versa) similarly between (popularity, length) relation is negative and weak. relation between year and length is positive however very weak because it is extremely on the point of zero

**Analysis of Length Column:**

As mean of length is greater than median of length column, it means data is positive/right skewed distributed.

**Analysis of Year Column:**

As mean of year is less than median of year it means data is negatively skewed distributed.

**Analysis of Popularity Column:**

as mean of popularity and its median are same it means this column is of normal distribution.

***For Dataset 2***

***Mean:***

Mean is the average of dataset.

**Mean of Column 1 is :**  525659.3564

**Mean of Column 2 is :**  477365.5364

**Median:**

The median is the middle of the set of numbers

**Median of Column 1 is :** 539262.0

**Median of Column 2 is :**  471640.5

**Mode:**

Mode is the most repeated value in column(s)

**Mode of Column 1 is** : 160154.

**Mode of Column 2 is :** 144848.

**Variance:**

Variance is a measure of how data points differ from the mean.

**Variance(Column 1):**  4.873847879620425

**Variance(Column 2) :**  5.468063424528494E10

**Range:**

The difference of largest and smallest value in a feature set / column is range

**Range of Column 1 is :**  928001.0

**Range of Column 2 is :**  958924.0

**Correlation:**

Correlation Between Column1 and Column2 0.13278911382081576

The correlation is positive but weak correlation.

**Analysis of Column1:**

As mean of column 1 is less than median of column 1 it means data is negatively skewed distributed.

**Analysis of Column2:**

As mean of column 2 is greater than median of column 2 , it means data is positive/right skewed distributed.

**d) Would you apply preprocessing techniques like discretization or normalization on any attribute? Explain**

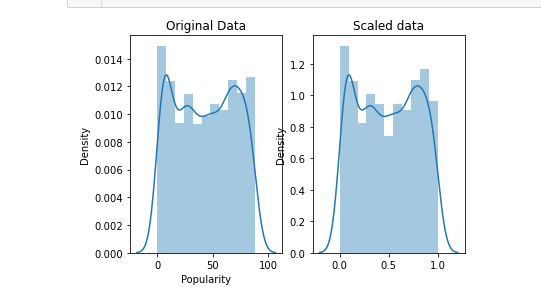
**your answer. If yes, then apply the technique and share the results.**

***For Dataset 1 (movies)***

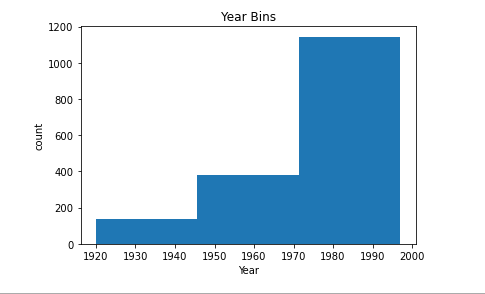
The choice of discretization or normalization depends on the machine learning / data mining model we use to solve the problem.

I chose to do discretization of year, length column and normalize popularity column

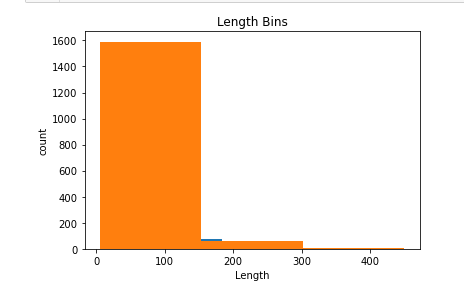
**Popularity Column**

. ******

**Year Column**

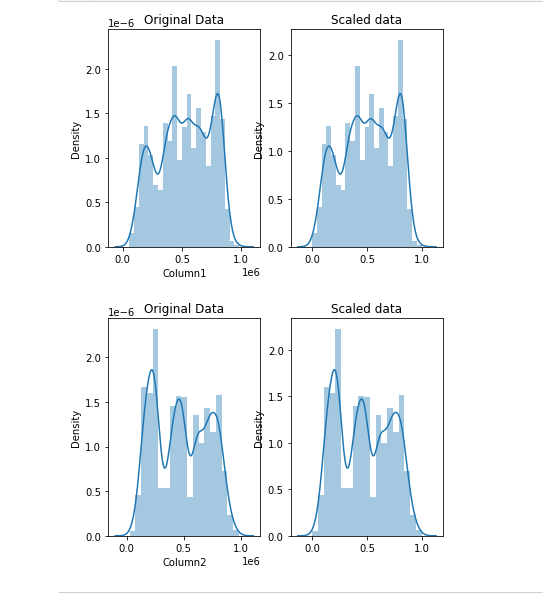


**Length Column**



***For Dataset 2***

The values of both columns of this dataset was of high range so I normalize them.

****

**Assignment 5:**

**CLUSTER THE DATA using the PySpark built-in K-means clustering algorithm (this is provided in the Spark**

**Library). Run your algorithm for various values of K and different values of convergence, show the results in your report.**

**Use measures such as SSE(the sum of square error), silhouette co-efficient, and NMI (normalized mutual Index) to analyze**

**the clustering results.**

**a) Cluster the Movies dataSet using atmost three attributes to avoid curse of dimentionality. You can select the**

**attribute based on the preprocessing.**

**Results of clustering:**

**When K=2**

Silhouette with squared euclidean distance for k = 2 = 0.4579375514358384

Cluster Centers:

[1974.1822542 104.87410072 66.37170264]

[1977.46780073 105.28311057 19.4981774 ]

Within Set Sum of Squared Error = [**17.04051280316536**, 50.9421004191285]

Normalized Mutual Index = [1.0, 1.0]

**When k=3**

Silhouette with squared euclidean distance for k = 3 = 0.5161197274087755

Cluster Centers:

[1974.25725095 100.79445145 66.28121059]

[1977.80783818 100.89759798 19.68900126]

[1971.12328767 196.89041096 44.73972603]

Within Set Sum of Squared Error = [**18.839894002183993**, 51.52139232461058, 99.71685472643122]

Normalized Mutual Index = [1.0, 1.0, 1.0]

**When k=4**

Silhouette with squared euclidean distance for k = 4 = 0.49726787009203516

Cluster Centers:

[1974.45726496 95.87037037 66.36039886]

[1977.85615491 98.80636238 18.46749654]

[1964.78571429 286. 45.71428571]

[1974.11926606 143.90366972 49.65137615]

Within Set Sum of Squared Error = [**21.752484851975108**, 53.176018198645956, 185.91314902284842, 49.08192081369391]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0]

**When k=5**

Silhouette with squared euclidean distance for k = 5 = 0.44423683857046137

Cluster Centers:

[1976.32692308 124.44871795 62.00320513]

[1981.38498403 103.12779553 16.3370607 ]

[1969.5 212.67307692 42.36538462]

[1944.45333333 84.56444444 49.68 ]

[1984.26923077 91.9479638 64.36199095]

Within Set Sum of Squared Error = [**20.09434238144985**, 53.46932064427942, 115.47373161159942, 67.10316980240844, 50.79216710744815]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0]

**When k=6**

Silhouette with squared euclidean distance for k = 6 = 0.3801598342095096

Cluster Centers:

[1977.66544118 124.29044118 65.05882353]

[1977.60117302 119.19354839 14.95014663]

[1983.8714653 93.57583548 67.73521851]

[1943.41176471 87.26244344 51.67420814]

[1983.81933842 88.50636132 22.63104326]

[1970.14634146 224.19512195 47.41463415]

Within Set Sum of Squared Error = [18.369204856701778, 55.065193624338505, 15.372636723533503, 67.52298584066247, **11.770287731836257**, 87.62301287609797]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**When k=7**

Silhouette with squared euclidean distance for k = 7 = 0.37064103114728414

Cluster Centers:

[1983.2997543 89.86240786 20.5970516 ]

[1943.52252252 86.97297297 51.13063063]

[1971.03225806 177.88709677 46.96774194]

[1968.54545455 302.09090909 43.72727273]

[1977.66770186 117.66149068 66.38198758]

[1985.41463415 90.24695122 66.80792683]

[1978.28196721 119.3147541 16.34754098]

Within Set Sum of Squared Error = [52.33288269445365, 56.71060417955538, 81.41618680253417, 184.29698902936332, 57.479119565179445, 49.82587038335262, **19.32969387509863**]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**When k=8**

Silhouette with squared euclidean distance for k = 8 = 0.408660188048697

Cluster Centers:

[1980.52494062 109.03562945 11.40142518]

[1962.1031746 65.20634921 24.69047619]

[1979.7260274 124.21004566 68.0913242 ]

[1963.45454545 301.54545455 51.36363636]

[1972.796875 176.484375 37.921875]

[1984.17957746 92.53873239 72.69014085]

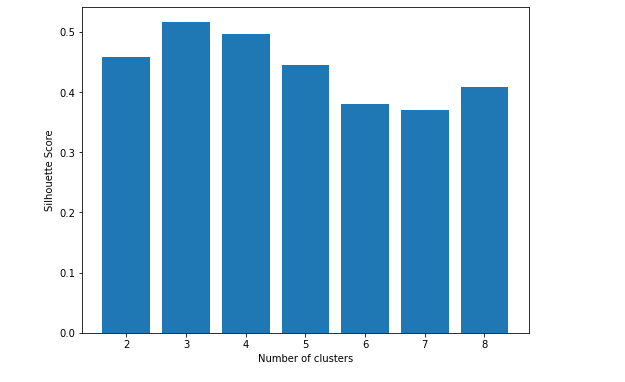
[1945.19786096 95.40106952 61.87165775]

[1983.25217391 98.71884058 38.75942029]

Within Set Sum of Squared Error = [57.41980645916998, 70.67513301584772, 23.198363087883113, 185.8388842763831, 86.0406923055706, 47.707030371588004, 61.87608287966914, **12.201708090233165**]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**Graphical View of selecting k using silhouette score:**

****

From above results I conclude that best value for k is 3 using silhouette score

From above results I conclude that best value for k is 6 using SSE score

1. **Cluster the DataSet 2 for different value of K**

**Results of clustering:**

**When K=2**

Silhouette with squared euclidean distance for k = 2 = 0.5722982025192378

Cluster Centers:

[603467.33476764 692357.33562823]

[458086.20926756 290653.67488789]

Within Set Sum of Squared Error = [**248617.06196427115**, 500586.6346872325]

Normalized Mutual Index = [1.0, 1.0]

**When K=3**

Silhouette with squared euclidean distance for k = 3 = 0.5434385882340155

Cluster Centers:

[500724.52964602 257534.00752212]

[741651.85074627 672435.42146411]

[339951.38109527 644174.18379595]

Within Set Sum of Squared Error = [**512582.65575508965**, 76867.23566034321, 529432.2935122066]

Normalized Mutual Index = [1.0, 1.0, 1.0]

**When K=4**

Silhouette with squared euclidean distance for k = 4 = 0.6241291870044543

Cluster Centers:

[276208.36804049 324984.9978308 ]

[656487.59757024 256869.23766135]

[408317.85826002 728839.13489736]

[754892.14878622 668343.1550509 ]

Within Set Sum of Squared Error = [649222.1191778906, 417295.15622794675, 475244.79559665656, **176273.20701083774**]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0]

**When K=5**

Silhouette with squared euclidean distance for k = 5 = 0.5958639253140531

Cluster Centers:

[185068.55481283 361389.57486631]

[727333.54274611 199761.10492228]

[407572.09819639 734796.93887776]

[754643.66974596 663964.82832948]

[457586.58241758 309784.48858833]

Within Set Sum of Squared Error = [715715.4233419113, 452139.76071453246, 477500.81066558615, **175039.62689312967**, 506087.8025608831]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0]

**When K=6**

Silhouette with squared euclidean distance for k = 6 = 0.5723126844748061

Cluster Centers:

[184934.31325301 361668.28380187]

[727295.92480211 195902.6055409 ]

[396597.25747863 726466.93803419]

[715506.48289474 764865.74078947]

[454590.0637931 306500.08017241]

[777051.80751174 549853.19092332]

Within Set Sum of Squared Error = [715732.2365373013, 455936.09708680335, 486017.3435431596, **258369.64984801252**, 510501.0356888243, 82844.16610247426]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**When K=7**

Silhouette with squared euclidean distance for k = 7 = 0.5941145934729072

Cluster Centers:

[227982.62539022 368551.04890739]

[733836.48551724 195801.04689655]

[350673.1633888 674780.2526475]

[818646.6707132 720584.62670713]

[452656.93274854 218856.11695906]

[647538.57317073 464396.99237805]

[568024.86850153 788161.06727829]

Within Set Sum of Squared Error = [673419.8012099692, 454852.69005164586, 521701.653040664, **152192.58025315902**, 572839.5557966921, 237946.5181587605, 300523.2797566922]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**When K=8**

Silhouette with squared euclidean distance for k = 8 = 0.604037718011563

Cluster Centers:

[226248.8197065 368897.96750524]

[733112.48834019 196186.4005487 ]

[452854.535 588127.77 ]

[818845.6402439 721291.26829268]

[261567.67168675 736654.5753012 ]

[666014.59284497 464017.62691652]

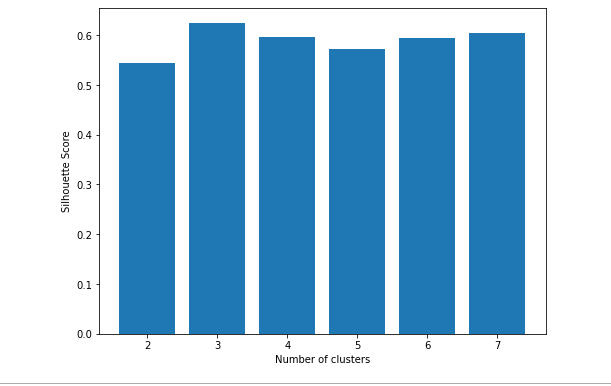
[567542.64263804 788867.46319018]

[452004.36811594 220538.7 ]

Within Set Sum of Squared Error = [674872.9921106655, 454599.8126066075, 415888.8669791596, **152587.12955588306**, 588150.9017209816, 223172.12450765586, 301317.2173120796, 583880.6417879385]

Normalized Mutual Index = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

**Graphical View of selecting k using silhouette score:**

****

From above results I conclude that best value for k is 3 using silhouette score

From above results I conclude that best value for k is 7 using SSE score