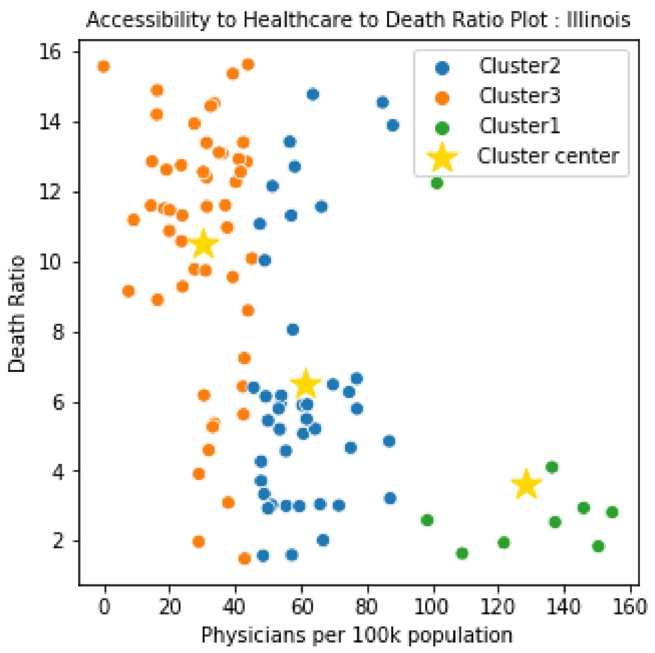
Unsupervised learning - K Means Clustering on Illinois data

In [2]:
```python
import os
from os import listdir
from os.path import isfile, join
codepath=r'C:\Users\NehaS\Desktop\CS418\CHSI EDA'
os.chdir(codepath)
%run clusterplot.py
```



Ratio of Deaths Vs Physician Rate have been clustered by Kmeans, an inversely proportional relationship shows that staffing of medical depts is one of the top concerns.

K-Means Clustering and Multinomial Bayes Classifier for Infant Mortality Rate

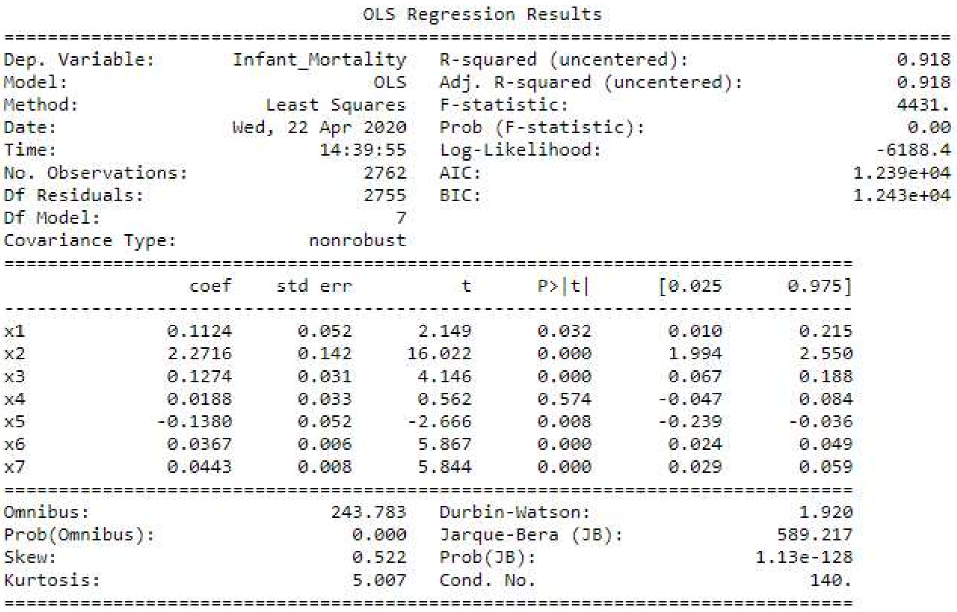
Clustering the counties in Illinois by high and low infant mortality rates and training a Multinomial Naive Bayes classifier to predict the infant mortality rate for new counties, we get a very poor initial accuracy on the model.

In [3]:
```python
%run before_opt.py
```

```
The initial accuracy of the Multinomial Naive Bayes Classifier on the clustered model is :
0.6
```

Backward Elimination to improve model accuracy

After backward elimination, taking the significance level as 0.05. Only Low Birth Weight and Very Low Birth Weight have an effect on the infant mortality for counties in Illinois after 5 iterations. This will change from State to State. Hence it is important to run this on different States to see which features have most impact which is a useful insight to healthcare agencies. Although, such a large increase in accuracy for this model is superficial. Since there is a high bias in this model, the accuracy can vary a great deal. Hence we train and test the same model pipeline on world data.



```
                                 OLS Regression Results
==========================================================================================
Dep. Variable:       Infant_Mortality   R-squared (uncentered):                   0.918
Model:                            OLS   Adj. R-squared (uncentered):              0.918
Method:                 Least Squares   F-statistic:                              4431.
Date:                Wed, 22 Apr 2020   Prob (F-statistic):                        0.00
Time:                        14:39:55   Log-Likelihood:                         -6188.4
No. Observations:                2762   AIC:                                  1.239e+04
Df Residuals:                    2755   BIC:                                  1.243e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1124      0.052      2.149      0.032       0.010       0.215
x2             2.2716      0.142     16.022      0.000       1.994       2.550
x3             0.1274      0.031      4.146      0.000       0.067       0.188
x4             0.0188      0.033      0.562      0.574      -0.047       0.084
x5            -0.1380      0.052     -2.666      0.008      -0.239      -0.036
x6             0.0367      0.006      5.867      0.000       0.024       0.049
x7             0.0443      0.008      5.844      0.000       0.029       0.059
==============================================================================
Omnibus:                      243.783   Durbin-Watson:                   1.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              589.217
Skew:                           0.522   Prob(JB):                    1.13e-128
Kurtosis:                       5.007   Cond. No.                         140.
==============================================================================
```

In [4]:
```python
%run after_opt.py
```

```
Accuracy of the model affter backward elimination :
0.9
```

Results on world data

In the world data, all factors satisfy the significance level except for pregnancies under 18. This makes sense, although it is unethical for an age of under 18, a lower age plays a role in producing a healthy offspring hence it does not affect the classifications. Accuracy is lower but the model is more stable and each run doesn't modify the accuracy by more than 1%.

In [5]:
```python
%run world_data.py
```

```
Accuracy of the model on world data :
0.8592057761732852
```