

The above bar graph shows number of people in different stages of prostrate cancer from the given dataset

Below is the visualisation formed by attempting to cluster different stages of prostrate cancer based on procedure\_code and patient\_age .

The **red cluster indicates early stage** which are plenty in number from the above graph and are spread throughout.

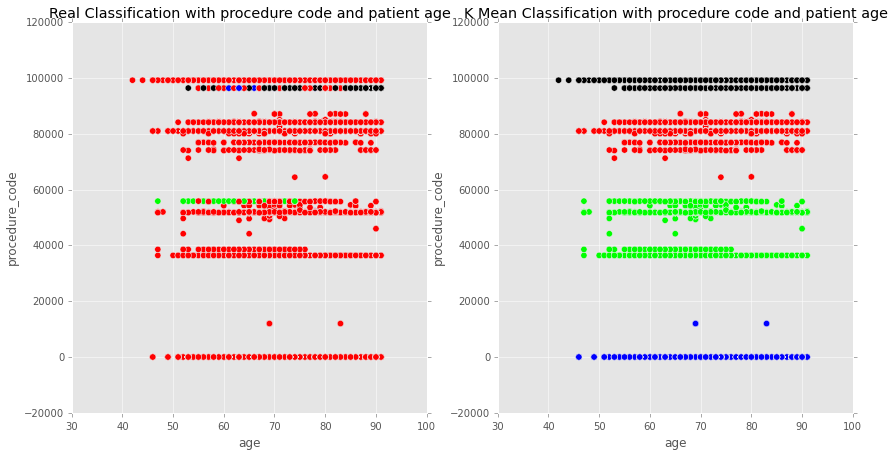
The **green cluster indicates early stage with intervention** which shows that it tries to fall into procedure\_code 55866 and according to real classification is predominant amongst the younger age.

The **black cluster indicates late stage** which shows that procedure\_code is 96402 according to kmeans and real classification and also according to real classification its slightly falls inder higher age bucket

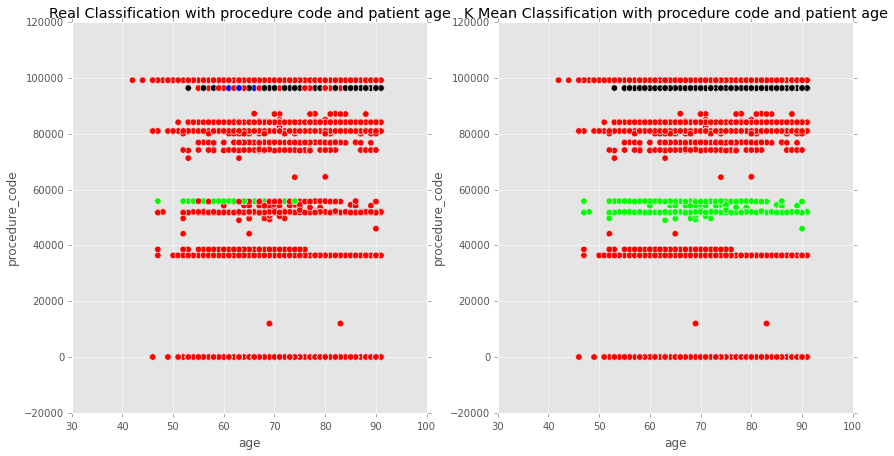
The **blue cluster indicates very late stage** which shows that procedure\_code is 96402 according to kmeans and real classification .

**Kmeans** shows us the clustering made by the kmeans ML algorithm given patient age and procedure code as features

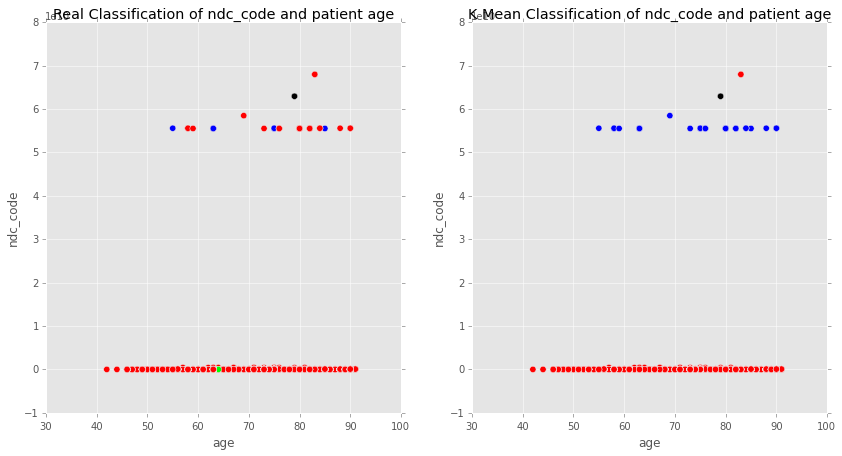
**real classification** is how the data in dataset looks when plotted .



the above plot k=4 and accuracy is around 0.289



the above plot is with k=8 and we see an accuracy of 0.897 mainly because here multiple clusters are put into the same bucket that is early stage (as the cluster is unbalanced).The smaller clusters of late ,very late stage happen to find a single smaller smaller cluster that appropriately fits it.



The above plot is with k=4 and a plot which takes in ndc\_code and age as parameters .

Found most of the data in the dataset having null values in ndc\_code which we have replaced with 0 for the algorithm to run.

ndc\_code for early stage is null,while it has a higher value for late and very late stage.

The below plot tried incorporating all 3 features that is age,procedure\_code and ndc\_code, used Principal Component Analysis to transform it into 2d space.

