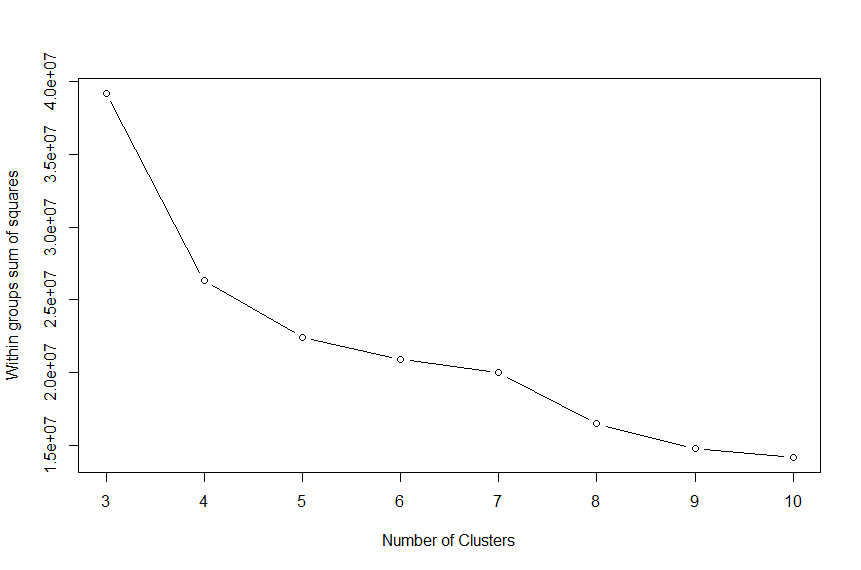
For the medical data set, I decided to not use any categorical variables given a couple of reasons. First, given the size of the data about 1.7 GB, one hot encoding would have taken quite a deal of time and second, some categorical variables had too many categories. It would be interesting to see how the results change if we included such categorical variables with low number of categories like gender, entity code (individual/organization), and medicare participation indicator (Y/N), and place of facility (facility/non facility) would affect the results, but inclusion of categorical data in clustering is generally discouraged given the technique utilizes euclidean distance which relies on points on a non discrete space. The numeric variables that I did pick included line\_service\_cnt, bene\_unique\_cnt (number of distinct Medicare beneficiaries), bene\_day\_srv\_cnt (number of distinct Medicare beneficiary services per day), average\_Medicare\_allowed\_amt, average\_submitted\_chrg\_amt, and average\_Medicare\_payment\_amt. I decided not to use the standard deviation measures as it would be strange to standardize these values. I then took the numeric data, standardized it, and used a scree plot of within sum of squares to determine the appropriate number of clusters, which turned out to be seven. I then ran k means clustering for 50 iterations.



Below is a visualization of clusters by the attributes. The most interesting is cluster 5 and 2. Cluster 5 seems to provide a lot of services and is in line with the average payment, while cluster 2 seems to provide just as much services as the rest but submit more for reimbursement.

