clusters

Cluster exploration!!

time to have a look at if some of these variables with many many categories can be simplified

Remaining categories include:

Full:

inv.mlsto timespan no.users Discipline pc.pro client.totinv code.director Num.disc Business Biz.type

Many NA:

code.contact code.client Project.Value JD.Second Billing.Type Post.Code code.ProjEng Billing.Type majority.pos pc.majpos

Create reduced dataset and have a think about what you want to cluster against each other

What kind of clusters are you looking for? Within the numeric variables:

Numeric variables in reduced dataset are: inv.mlsto, timespan, no.users, pc.pro, client.totinv, Num.disc, Project.Value, pc.majpos

Numeric variables worth considering in full dataset: Num.days, mean.peeps, hours.perday, balance.mlsto, hrs.mlsto, cost.mlsto,

So, lets choose:

* inv.mlsto
* timespan
* no.users
* pc.pro
* client.totinv
* Num.disc
* Num.days
* mean.peeps
* hours.perday
* balance.mlsto
* hrs.mlsto
* cost.mlsto

Many NA - do second \* pc.majpos \* Project.Value

Before clustering, need to standardise variables, and have a look at 'normality' of other variables

taking the log worked for: \* cost.mlsto, hrs.mlsto, hours.perday, Num.days, client.totinv, no.users, inv.mlsto \* sqrt of timespan

so we end up with a numeric data set with the above variables log transformed. Then scale each by subtracting the mean of each variable and then dividing by the sd of each variable.

Now to do hierarchical clustering

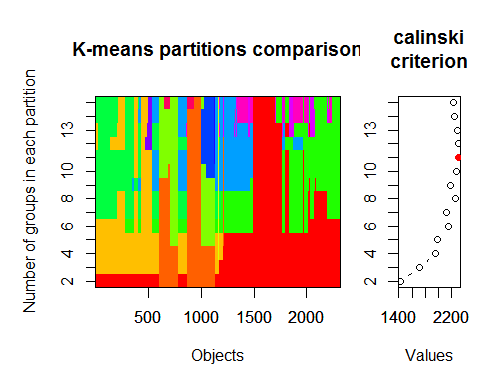
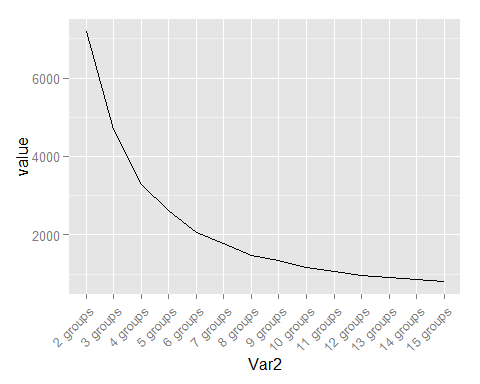
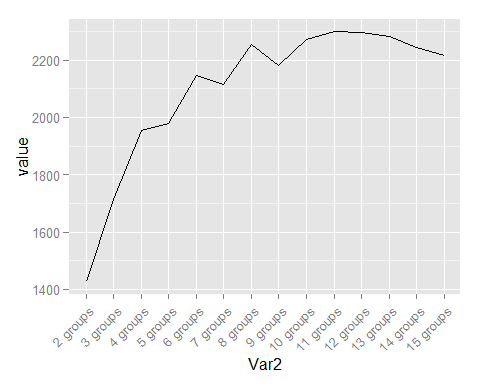
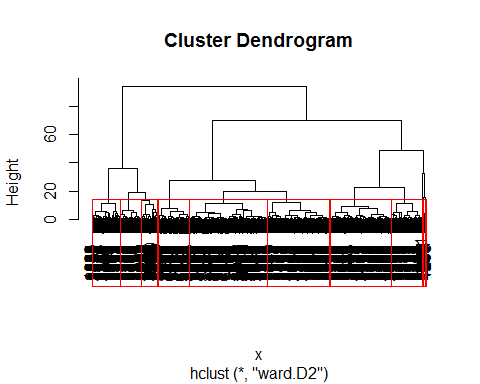
Now look at properties of groups using k means clustering

Very interesting, now want to optimise number of clusters- refer wards tutorial

Analysis shows that inv.mlsto, balance.mlsto, return.pdol, and hrs.mlsto should be used for clustering and that 17 clusters is ideal. I'm not completely happy with the clustering. I think I should now add a new variable: inv.mlsto x return.pdol This will give a 'success' scale, weighing large jobs even with a smaller return.pdol as quite good and tiny jobs with a good return.pdol as average to low.

this means I will need to check normality, scale the variable before adding it to clust

Repeat clustering with new suc and revised bi-modal return.pdol

now want 11 clusters, based on k means optimisation, have a look at average characteristics

## Using groups as id variables