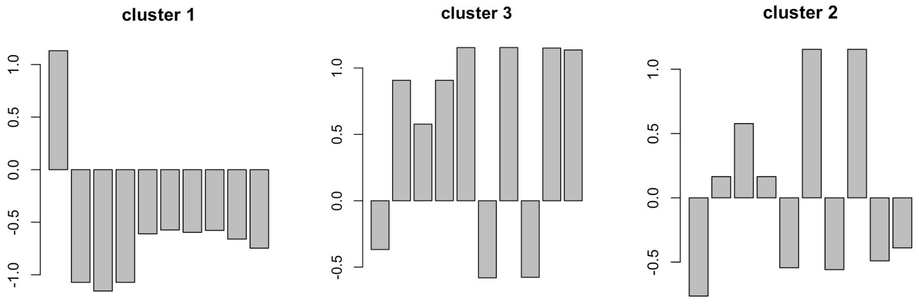
1. Data Column to use: In order to capture the variation in the dataset as much as possible, I used all of the numerical values available in this dataset. Furthermore, because gender is easy to code into number, I also use gender (Male 0, Female 1). Therefore, columns I used are: nppes_provier_gender, line_srvc_cnt, bene_unique_cnt, bene_day_srvc_cnt, average_Medicare_allowed_amt, stdev_Medicare_allowed_amt, average_submitted_chrg_amt, stdev_submitted_chrg_amt, avg_payment_amt, stdev_payment_amt.
2. To standardize the data: I run a mapReduce job to obtain the max and min value for each column, and then use (value - min)/(max - min) formula to standardize data.
3. We can visualize the centroid below:



Interpretation:

As we can see from the comparison of three clusters (values after standardization), the first cluster is predominantly male, and other values are very low. So the first cluster can be **male providers who are low on number of services and amount paid.**

The second cluster is predominantly female and high on standard deviation columns. So the second cluster can be **female providers who are very uncertain in terms of number of services and payment amount.**

The third cluster has more female than male and is low on standard deviation and high on means. This cluster means **providers that are high on payment amount and number of beneficiaries with certainty.**

Even without the gender, we can still differentiate three clusters as those high in payment amount and number of beneficiaries, those low in them and those uncertain about them (indicated by high standard deviations).