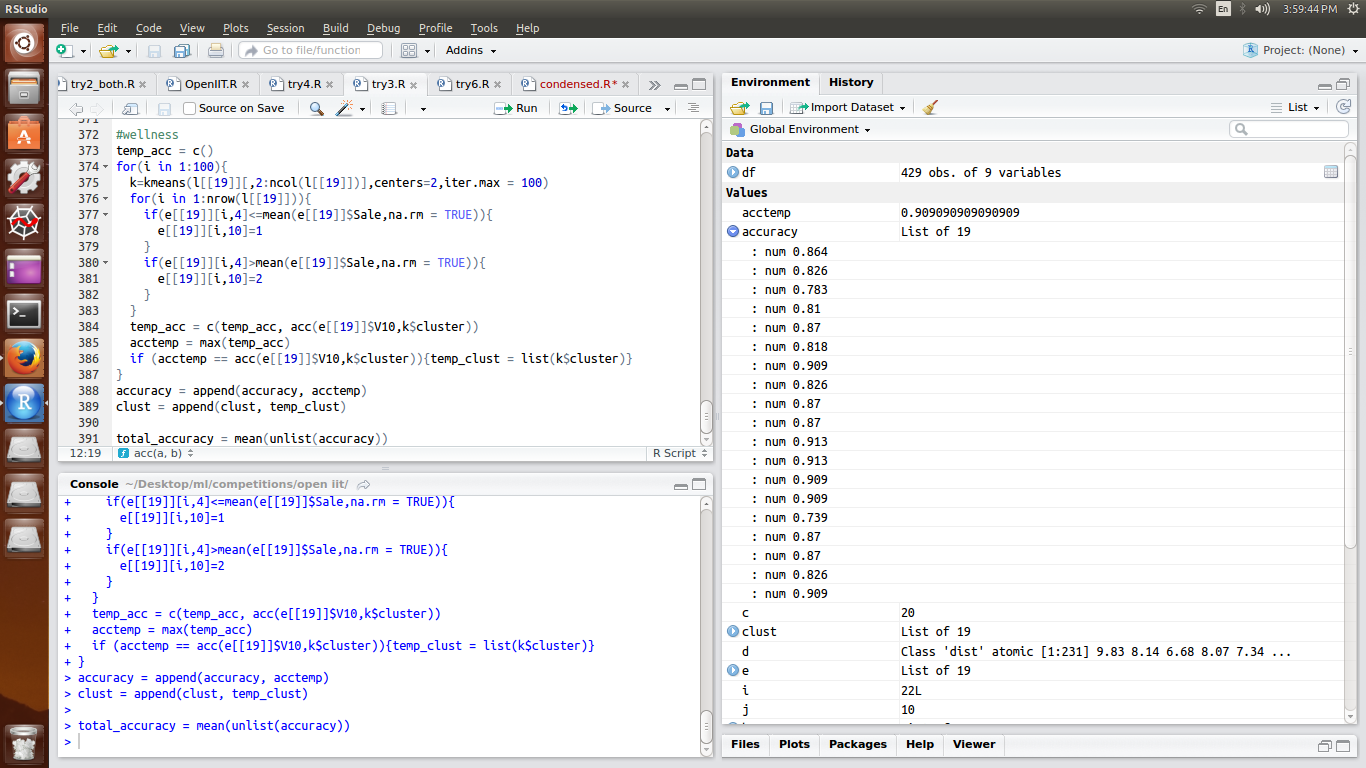
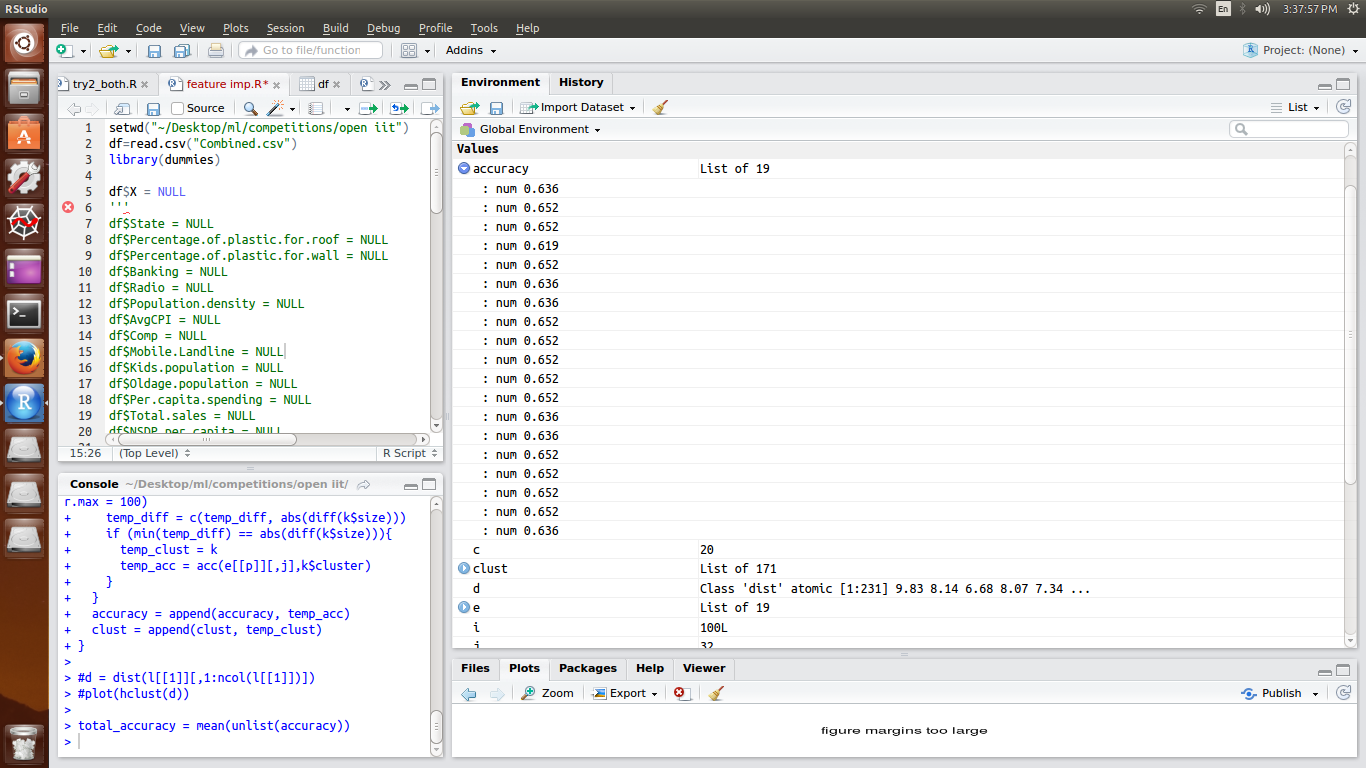
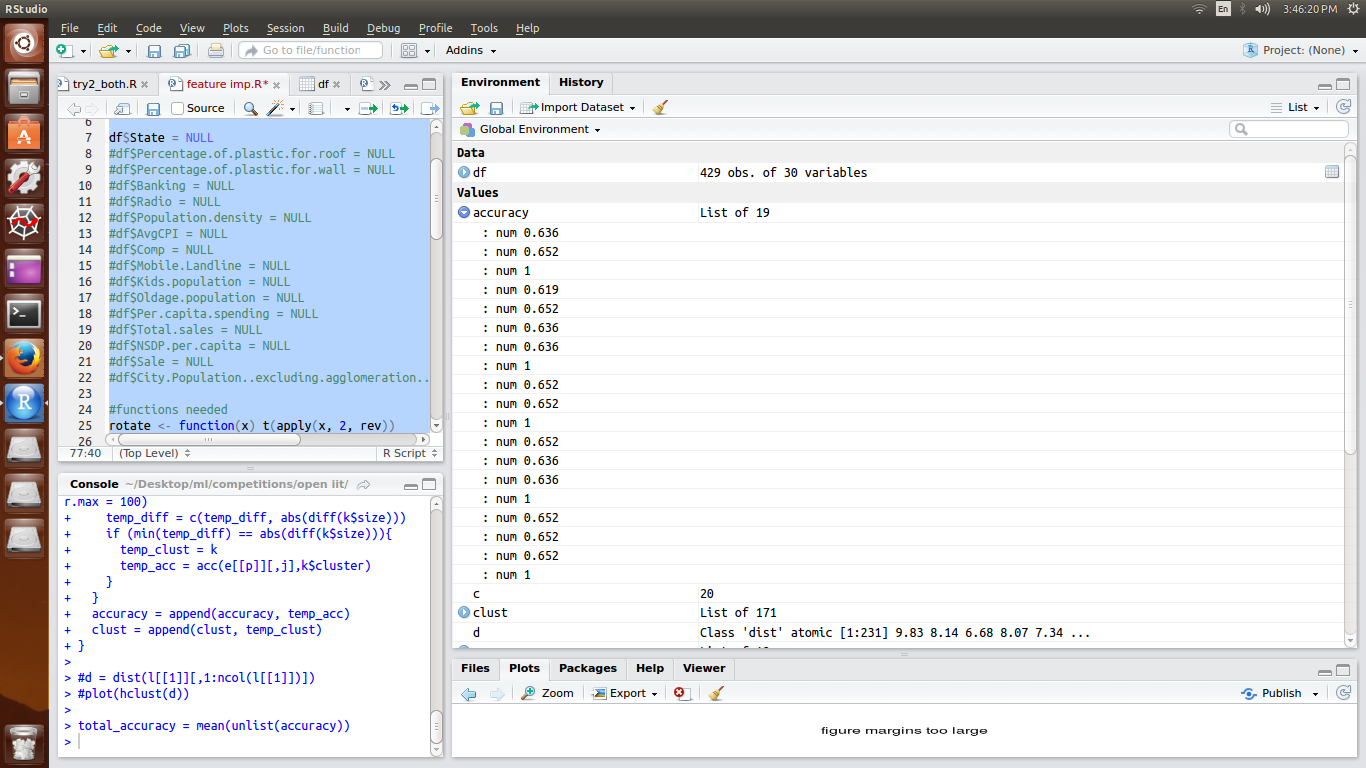
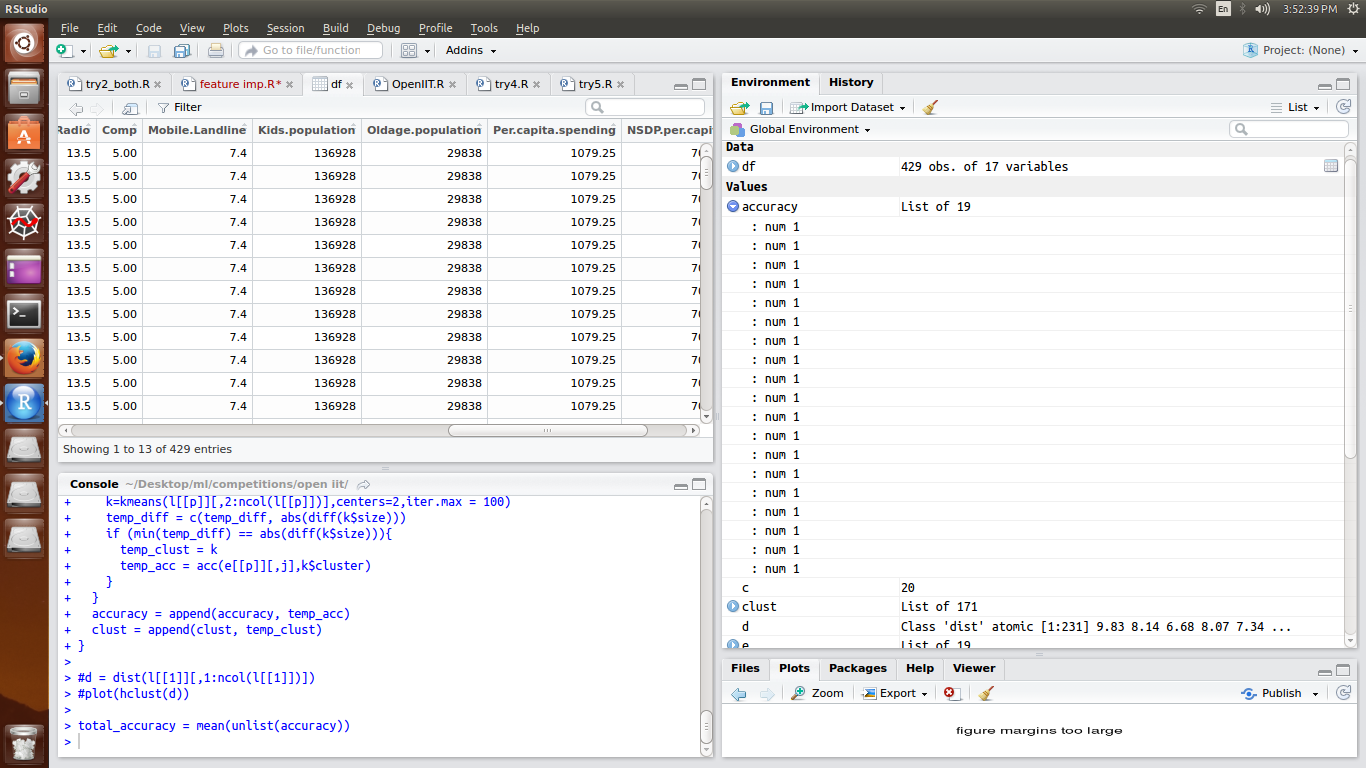
* We first looked at data of each SKU from the given data and clustered them using Kmeans. We found optimal number of clusters by observing number of data points in each cluster and applied that number of clusters.
  + We scaled the data and applied one hot encoding for categorical data.
  + But if we apply Kmeans on the same data again and again the number of data points in each cluster is varying for every time we clustered.
  + So to overcome that problem we wrote the clustering in loop and observed the clusters.
  + In loop we choose the best cluster(fixed number of clusters) based on the difference between number of data points in clusters and stored them in list for all 19 SKUs.
* We wanted to know on what basis Kmeans is clustering. So our idea is to assume that its clustering based on one feature and finding the accuracy of that being the reason.
  + So we first considered that its clustering on the basis of sales SKU in cities/Towns and divided the dato into 2 parts.i.e. low sales and High sales based on mean of sales of SKU.
  + We compared the clusters formed with the group of high and low sales that we made.
  + We found that 76% of the data matched. So from this we can say that, “For each SKU its clustering based on sales with 76% accuracy”
  + Found total accuracy by taking mean of each SKU
* We tried improving the accuracy of sales by taking various types of datasets:
  + We took the additional data with features like population density, percentage of population using plastic roof and wall, banking, mobiles, TV, Computer etc,...
  + For that data we got the accuracy of around 75%
  + Since the accuracy decreased we removed some of the features from the data and kept features like average CPI, population density, percentage of people using plastic roof, plastic wall, banking, radio, landline, kids population, old age population, per capita spending, NSDP per capita and Sales.
  + On that data we got accuracy around 78.9%.
* Till now we tried to observe the accuracy based on sales, now we tried based on groups of locations.i.e. Cities and Towns.
  + We similarly made 2 groups .i.e. Towns and Cities.
  + Now we found the accuracy with clusters and we found that accuracy is 80%
  + Now we removed some features and surprised to find accuracy of 100%
  + From this we can say that, “For each SKU if we make them into 2 clusters, one cluster is full of cities and other is full of towns.”
  + For the new features we still got sales with accuracy of 80%.



**Important Features:**

* If we have clustered with all the features we are getting accuracy around 64%. So I tried removing one features and found the accuracy.
* This shows the accuracy of individual SKU.
* This is what we obtained after removing 4 features…… Population, States, Teen Population, Working Population
* After that i tried to remove one more feature……...Backward Elimination.
* 
* Final Accuracy
* Finally found the important features as:
  + Type of location( city/ town )
  + Average CPI
  + Population density
  + Percentage of plastic for roof
  + Percentage of plastic for wall
  + Banking
  + Radio
  + Comp
  + Mobile landline
  + Kids population
  + Old Age population
  + Percapita spending
  + NSDP percapita