Shubham Saxena
ss4017@rit.edu

Running the program:
In the folder where all the submitted files are present:

javac *.java
java HW06_Saxena_Shubham_program <filename.csv> <number of clusters desired>

Answer 1:
Planned my approach on my own. Jaydeep Untwal helped me.

Answer 2:
ID. This is unique. It would be incorrect to include it in the mean calculation/distance calculations.

Answer 3:
All except the ID. (Milk PetFood Veggies Cereal Nuts Rice Meat Eggs Yogurt Chips Beer Fruit)

Answer 4:
Cluster 1 size:7 Guest ids:[8, 12, 9, 13, 1, 4, 22]
Cluster 2size:10 Guest ids: [7, 28, 20, 16, 3, 33, 21, 29, 10, 23]
Cluster 3 size:16 Guest ids: [6, 27, 2, 32, 24, 15, 30, 14, 17, 26, 18, 31, 11, 25, 5, 19]

Shubham Saxena
ss4017@rit.edu

```
<terminated> HW06_Saxena_Shubham_program [Java Application] C:\Program Files\Java\jre1.8.0_60
[Cluster size:7
[8[1.0, 5.0, 8.0, 8.0, 10.0, 4.0, 0.0, 2.0, 4.0, 5.0, 1.0, 9.0]
, 12[2.0, 7.0, 7.0, 9.0, 10.0, 8.0, 0.0, 0.0, 4.0, 6.0, 0.0, 8.0]
, 9[2.0, 3.0, 9.0, 9.0, 6.0, 7.0, 0.0, 1.0, 2.0, 7.0, 1.0, 9.0]
, 13[1.0, 4.0, 7.0, 8.0, 5.0, 10.0, 0.0, 0.0, 3.0, 6.0, 0.0, 9.0]
, 1[2.0, 5.0, 8.0, 9.0, 9.0, 9.0, 1.0, 2.0, 3.0, 5.0, 0.0, 5.0]
, 4[3.0, 6.0, 10.0, 6.0, 7.0, 9.0, 0.0, 1.0, 3.0, 3.0, 1.0, 8.0]
, 22[0.0, 4.0, 9.0, 7.0, 8.0, 9.0, 1.0, 0.0, 1.0, 3.0, 0.0, 6.0]
]
, Cluster size:10
[7[8.0, 7.0, 10.0, 7.0, 4.0, 0.0, 4.0, 5.0, 4.0, 6.0, 5.0, 4.0]
, 28[5.0, 1.0, 3.0, 7.0, 8.0, 3.0, 5.0, 6.0, 5.0, 8.0, 6.0, 1.0]
, 20[9.0, 2.0, 9.0, 9.0, 8.0, 0.0, 2.0, 2.0, 3.0, 7.0, 8.0, 2.0]
, 16[8.0, 9.0, 3.0, 7.0, 7.0, 3.0, 1.0, 3.0, 4.0, 6.0, 5.0, 3.0]
, 3[6.0, 5.0, 6.0, 10.0, 8.0, 2.0, 3.0, 2.0, 9.0, 9.0, 4.0, 4.0]
, 33[6.0, 4.0, 5.0, 10.0, 6.0, 2.0, 5.0, 4.0, 4.0, 8.0, 4.0, 1.0]
, 21[8.0, 5.0, 7.0, 9.0, 6.0, 0.0, 1.0, 2.0, 5.0, 8.0, 5.0, 1.0]
, 29[7.0, 4.0, 4.0, 7.0, 7.0, 2.0, 2.0, 4.0, 7.0, 7.0, 7.0, 5.0]
, 10[9.0, 6.0, 4.0, 9.0, 7.0, 1.0, 3.0, 4.0, 5.0, 7.0, 8.0, 1.0]
, 23[5.0, 6.0, 6.0, 8.0, 6.0, 1.0, 2.0, 2.0, 6.0, 4.0, 7.0, 3.0]
]
, Cluster size:16
[6[10.0, 3.0, 8.0, 5.0, 4.0, 5.0, 2.0, 5.0, 4.0, 1.0, 4.0, 1.0]
, 27[8.0, 1.0, 7.0, 7.0, 4.0, 8.0, 2.0, 8.0, 6.0, 1.0, 4.0, 8.0]
, 2[9.0, 7.0, 7.0, 9.0, 5.0, 6.0, 4.0, 3.0, 7.0, 5.0, 4.0, 8.0]
, 32[10.0, 5.0, 8.0, 10.0, 3.0, 3.0, 6.0, 3.0, 7.0, 3.0, 5.0, 8.0]
, 24[9.0, 3.0, 7.0, 9.0, 4.0, 6.0, 3.0, 7.0, 9.0, 5.0, 2.0, 4.0]
, 15[6.0, 8.0, 8.0, 8.0, 7.0, 7.0, 3.0, 9.0, 4.0, 6.0, 6.0, 7.0]
, 30[8.0, 8.0, 6.0, 8.0, 5.0, 8.0, 5.0, 8.0, 7.0, 3.0, 6.0, 3.0]
, 14[10.0, 5.0, 10.0, 8.0, 7.0, 7.0, 6.0, 9.0, 4.0, 3.0, 5.0, 6.0]
, 17[9.0, 5.0, 9.0, 7.0, 4.0, 7.0, 2.0, 7.0, 4.0, 3.0, 6.0, 3.0]
, 26[7.0, 6.0, 9.0, 9.0, 5.0, 8.0, 4.0, 8.0, 4.0, 2.0, 5.0, 4.0]
, 18[10.0, 4.0, 6.0, 9.0, 5.0, 6.0, 5.0, 8.0, 5.0, 3.0, 4.0, 6.0]
, 31[9.0, 5.0, 6.0, 9.0, 7.0, 9.0, 3.0, 7.0, 7.0, 3.0, 4.0, 5.0]
, 11[9.0, 6.0, 8.0, 8.0, 8.0, 7.0, 6.0, 6.0, 5.0, 3.0, 2.0, 5.0]
, 25[10.0, 6.0, 7.0, 8.0, 6.0, 7.0, 3.0, 6.0, 4.0, 4.0, 1.0, 7.0]
, 5[7.0, 6.0, 9.0, 8.0, 4.0, 8.0, 4.0, 6.0, 7.0, 4.0, 3.0, 5.0]
, 19[9.0, 5.0, 7.0, 8.0, 4.0, 9.0, 4.0, 5.0, 5.0, 2.0, 2.0, 6.0]
]
]
```

Answer 5:

Looking at the mean of clusters I noticed that the guests in the 3rd cluster are probably vegetarians and do not like beer. cluster center:[4.8, 8.2, 8.0, 7.8, 8.0, 0.2, 0.8, 2.8, 5.0, 0.4, 7.7] Label: “Weight watchers”. (meat eggs and beer are high in calories. And these guests have highest fruit purchases).

Answer 6:

True number of clusters are 3. (Used weka to verify).
Size of the smaller cluster was always less than or equal to 2. (mostly 1) until iteration 30. After this we see the sizes are much larger. Which implies the true clusters are 3.

1. c1 size:1 c2 size:1
2. c1 size:1 c2 size:1
3. c1 size:1 c2 size:1
4. c1 size:2 c2 size:1
5. c1 size:3 c2 size:1

Shubham Saxena
ss4017@rit.edu

6. c1 size:4 c2 size:1
7. c1 size:5 c2 size:1
8. c1 size:1 c2 size:1
9. c1 size:2 c2 size:1
10. c1 size:6 c2 size:2
11. c1 size:8 c2 size:1
12. c1 size:1 c2 size:1
13. c1 size:2 c2 size:1
14. c1 size:3 c2 size:1
15. c1 size:3 c2 size:2
16. c1 size:5 c2 size:1
17. c1 size:1 c2 size:1
18. c1 size:9 c2 size:1
19. c1 size:4 c2 size:1
20. c1 size:10 c2 size:1
21. c1 size:5 c2 size:1
22. c1 size:6 c2 size:1
23. c1 size:6 c2 size:1
24. c1 size:11 c2 size:1
25. c1 size:12 c2 size:2
26. c1 size:14 c2 size:1
27. c1 size:7 c2 size:1
28. c1 size:8 c2 size:1
29. c1 size:9 c2 size:1
30. c1 size:15 c2 size:1
31. c1 size:16 c2 size:10
32. c1 size:26 c2 size:7

Answer 7:

We would have to compare all the members of all the clusters everytime. That is we will have check the nearest pair from different clusters.

Answer 8:

7 hrs. (2 hrs to code 5 to debug one small very fundamental mistake)

Question 9. Write a short answer question for the next midterm exam. As if your question is used, you get
the points on the exam. Part of the reason I ask this is to be sure you think about the questions that might be on the next exam. (½)

Question: When would you use K means and when would you use agglomerative clustering.

Answer: Agglomerative clustering should be used when our clusters may have subclusters like Species taxonomy. Ie we have a hierarchical structure.

Shubham Saxena
ss4017@rit.edu

K means should be used when we need to consider and keep outliers and noise. Works well for circular cluster shape.