Reasoning on Factors that Influence Nutrition Intake in the American Population

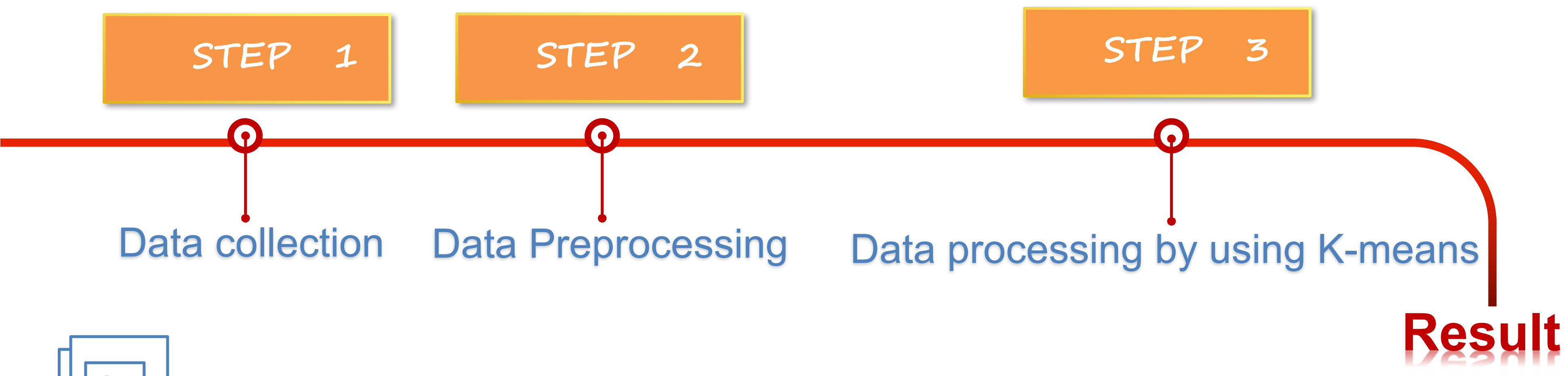
Author : Yuhao Peng
Advisor : Michela Taufer

Overview

Motivation

The purpose of this research is to identify whether age, race and gender influence Nutrition intake in the American Population based on NHANES data. Nowadays fat becomes gradually popular. Therefore I analyze ten type of main nutrient impact by using 2014 the newest version NHANES data. I do Preprocessing data and Clustering with K-means analysis by using MapReduce programming model, Finally prompt a group of people to pay attention to there Nutrition Intake by analyzing k cluster center.

Goal



STEP 1 — Data collection

STEP 2 — Data Preprocessing

STEP 3 — Data processing by using K-means

Result

01 Data collection

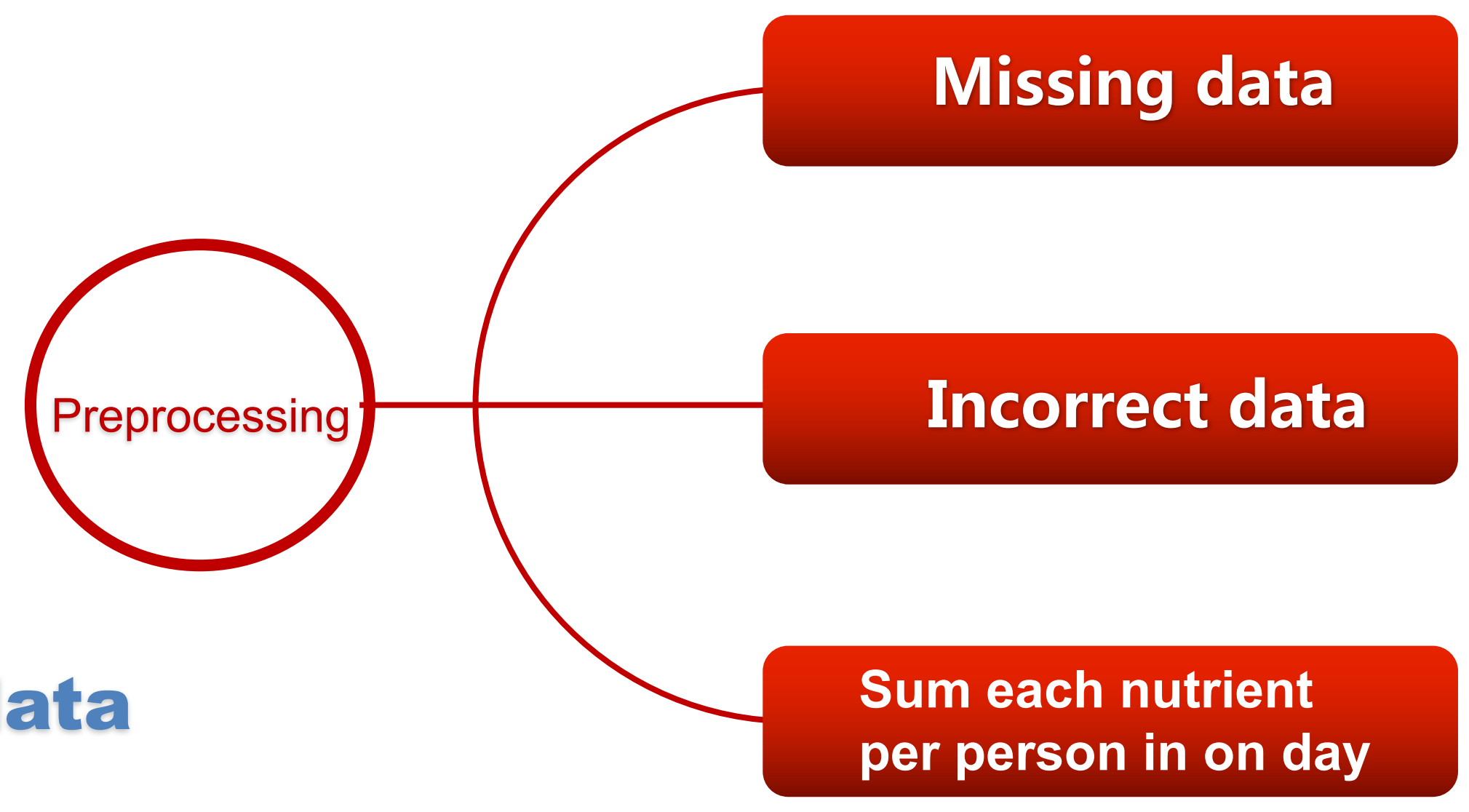
The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.

1.Download NHANES data from http://wwwn.cdc.gov/nchs/nhanes
2. Convert SAS datatype to CSV datatype

Reference from : Michael Wyat
https://github.com/TauferLab/NHANES-Analytics

02 Data Preprocessing

Data Preprocessing has three aspects. Preprocessing data make result more accurate



Preprocessing

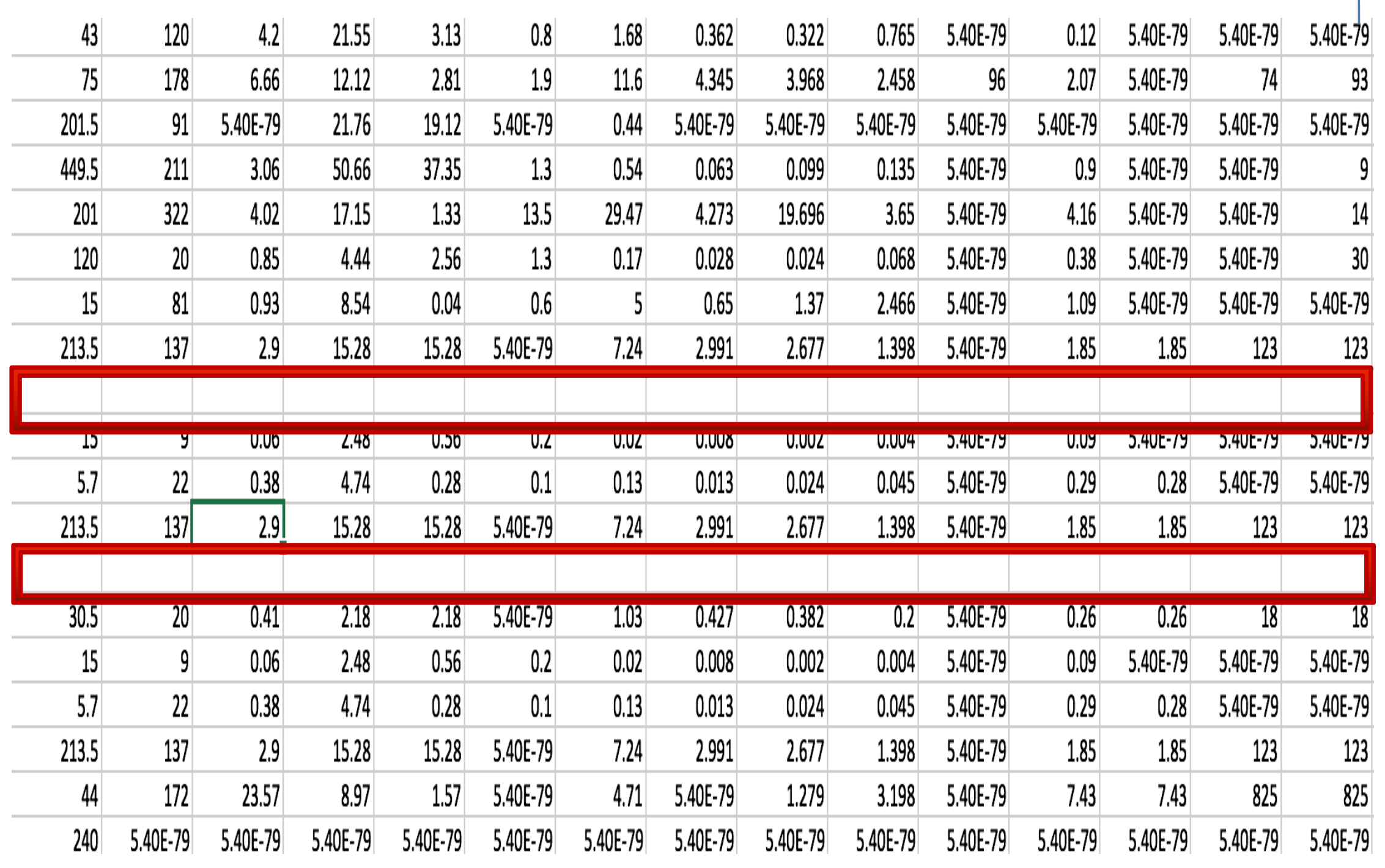
Missing data

Incorrect data

Sum each nutrient per person in on day

2.1 Preprocessing Missing data

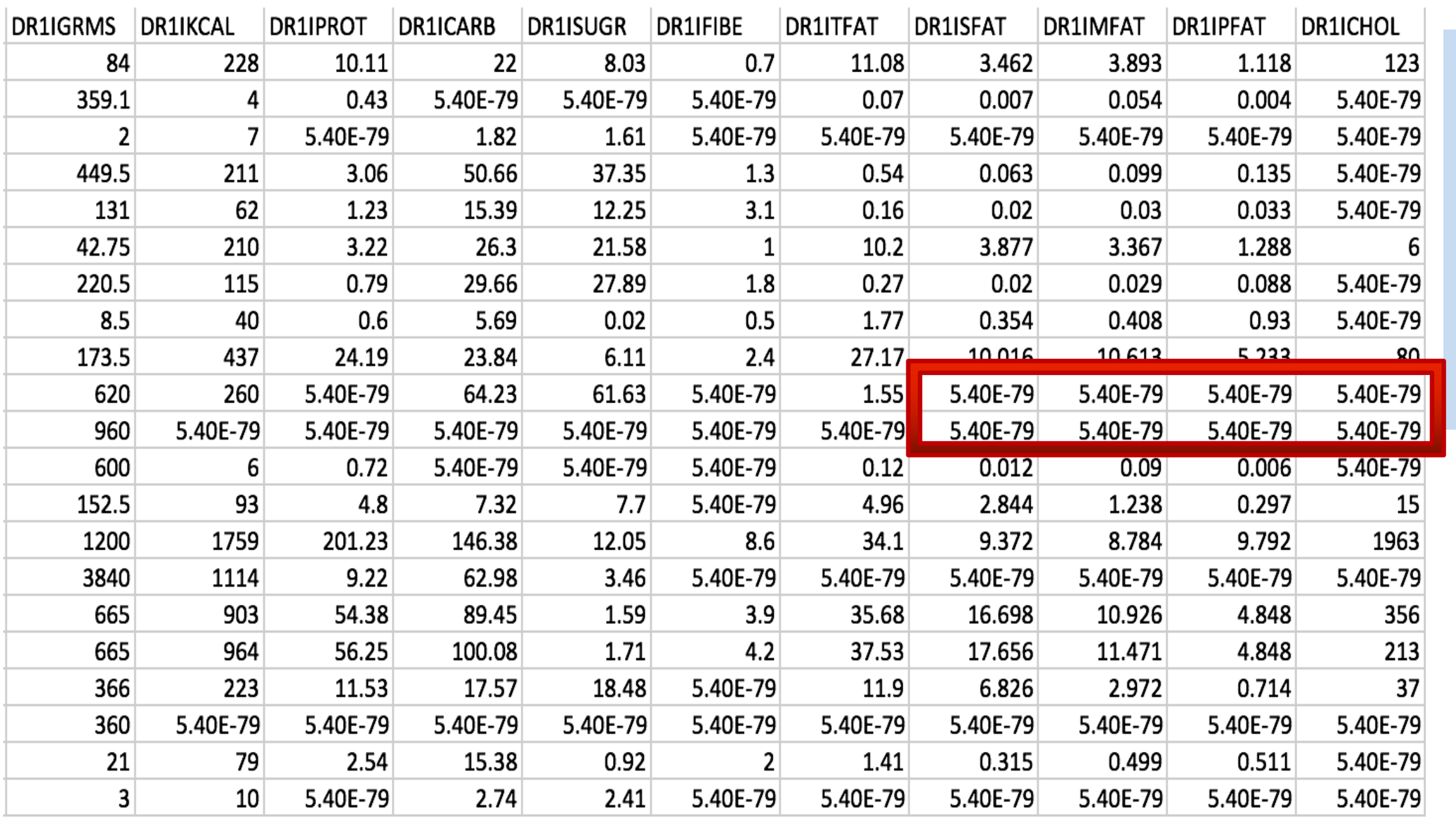
From the first day Individual Foods (DR1IFF_H.csv) in 2014, there are 131394 data in each column and there are 696 missing data in each column. So the rate of missing data is 0.53%. Because the rate is less than 10%, drop all Missing data, they are not influence the result.



43	120	4.2	21.55	3.13	0.8	1.68	0.362	0.322	0.765	5.40E-79	0.12	5.40E-79	5.40E-79	5.40E-79
75	178	6.66	12.12	2.81	1.9	11.6	4.345	3.968	2.458	96	2.07	5.40E-79	74	93
201.5	91	5.40E-79	21.76	19.12	5.40E-79	0.44	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
449.5	211	3.06	50.66	37.35	1.3	0.54	0.063	0.099	0.135	5.40E-79	0.9	5.40E-79	5.40E-79	9
201	322	4.02	17.15	1.33	13.5	29.47	4.273	19.696	3.65	5.40E-79	4.16	5.40E-79	5.40E-79	14
120	20	0.85	4.44	2.56	1.3	0.17	0.028	0.024	0.068	5.40E-79	0.38	5.40E-79	5.40E-79	30
15	81	0.93	8.54	0.04	0.6	5	0.65	1.37	2.466	5.40E-79	1.09	5.40E-79	5.40E-79	5.40E-79
213.5	137	2.9	15.28	15.28	5.40E-79	7.24	2.991	2.677	1.398	5.40E-79	1.85	1.85	123	123
15	9	0.06	2.48	0.56	0.2	0.02	0.008	0.002	0.004	5.40E-79	0.09	5.40E-79	5.40E-79	5.40E-79
5.7	22	0.38	4.74	0.28	0.1	0.13	0.013	0.024	0.045	5.40E-79	0.29	0.28	5.40E-79	5.40E-79
213.5	137	2.9	15.28	15.28	5.40E-79	7.24	2.991	2.677	1.398	5.40E-79	1.85	1.85	123	123
30.5	20	0.41	2.18	2.18	5.40E-79	1.03	0.427	0.382	0.2	5.40E-79	0.26	0.26	18	18
15	9	0.06	2.48	0.56	0.2	0.02	0.008	0.002	0.004	5.40E-79	0.09	5.40E-79	5.40E-79	5.40E-79
5.7	22	0.38	4.74	0.28	0.1	0.13	0.013	0.024	0.045	5.40E-79	0.29	0.28	5.40E-79	5.40E-79
213.5	137	2.9	15.28	15.28	5.40E-79	7.24	2.991	2.677	1.398	5.40E-79	1.85	1.85	123	123
44	172	23.57	8.97	1.57	5.40E-79	4.71	5.40E-79	1.279	3.198	5.40E-79	7.43	7.43	825	825
240	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79

2.2 Preprocessing Incorrect data

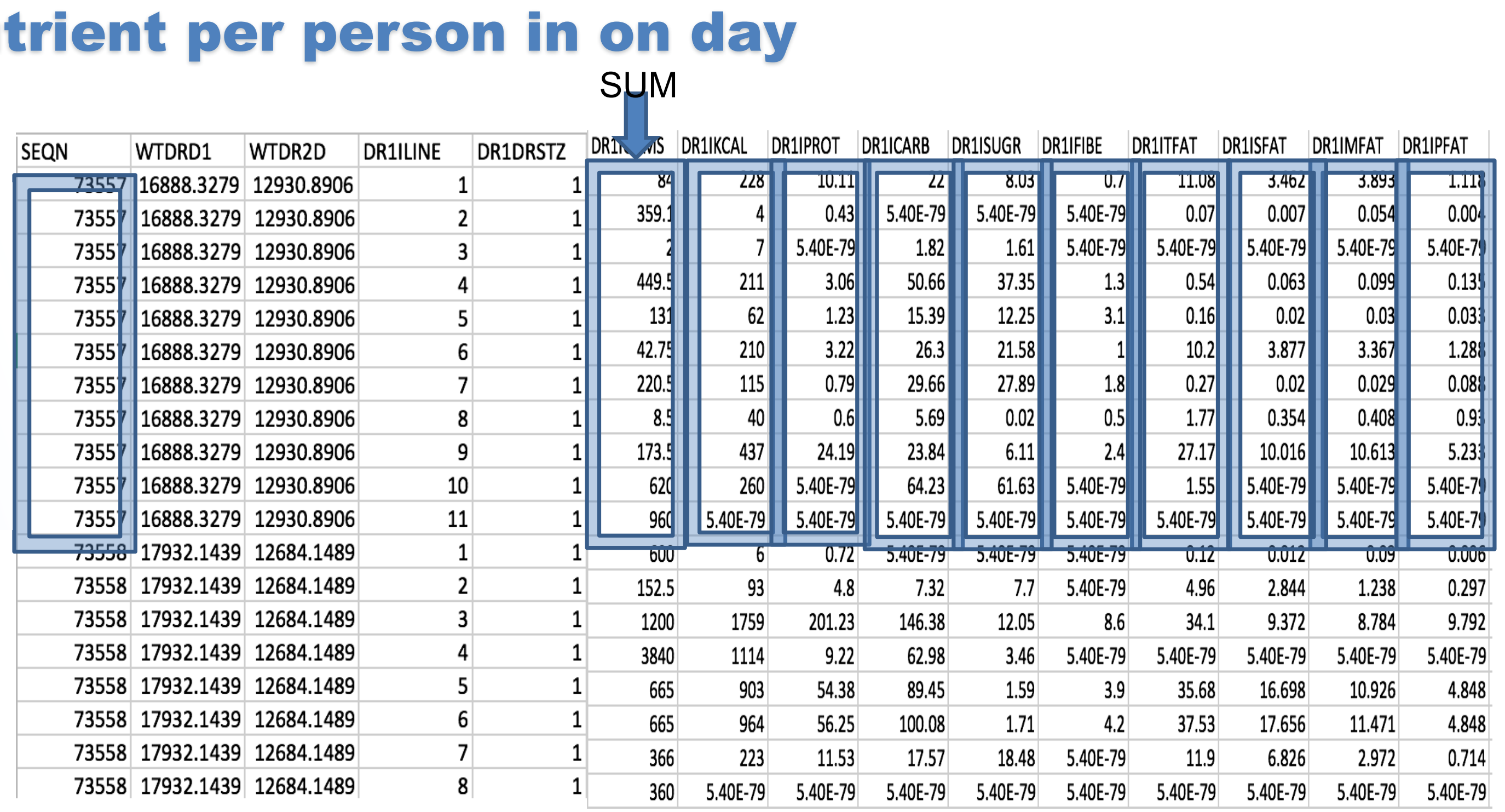
There are lots of value 5.40E-79 data in the first day Individual Foods (DR1IFF_H.csv. It is almost zero. Perhaps one object is kid or vegeterian, who may have value 0 in one type nutrient. The average of occurrence is 26735 in each column of these nutrient. The rate of 5.40E-79 data is 20.3%. So we can replace it with zero with very little influence.



DR1IGRMS	DR1IKCAL	DR1IPROT	DR1ICARB	DR1ISUGR	DR1IFIBE	DR1ITFAT	DR1ISFAT	DR1IMFAT	DR1IPFAT	DR1ICHOL
84	228	10.11	22	8.03	0.7	11.08	3.462	3.893	1.118	123
359.1	4	0.43	5.40E-79	5.40E-79	5.40E-79	0.07	0.007	0.054	0.004	5.40E-79
2	7	5.40E-79	1.82	1.61	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
449.5	211	3.06	50.66	37.35	1.3	0.54	0.063	0.099	0.135	5.40E-79
131	62	1.23	15.39	12.25	3.1	0.16	0.02	0.03	0.033	5.40E-79
42.75	210	3.22	26.3	21.58	1	10.2	3.877	3.367	1.288	6
220.5	115	0.79	29.66	27.89	1.8	0.27	0.02	0.029	0.088	5.40E-79
8.5	40	0.6	5.69	0.02	0.5	1.77	0.354	0.408	0.93	5.40E-79
173.5	437	24.19	23.84	6.11	2.4	27.17	10.016	10.613	5.233	80
620	260	5.40E-79	64.23	61.63	5.40E-79	1.55	5.40E-79	5.40E-79	5.40E-79	5.40E-79
960	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
600	6	0.72	5.40E-79	5.40E-79	5.40E-79	0.12	0.012	0.09	0.006	5.40E-79
152.5	93	4.8	7.32	7.7	5.40E-79	4.96	2.844	1.238	0.297	15
1200	1759	201.23	146.38	12.05	8.6	34.1	9.372	8.784	9.792	1963
3840	1114	9.22	62.98	3.46	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
665	903	54.38	89.45	1.59	3.9	35.68	16.698	10.926	4.848	356
665	964	56.25	100.08	1.71	4.2	37.53	17.656	11.471	4.848	213
366	223	11.53	17.57	18.48	5.40E-79	11.9	6.826	2.972	0.714	37
360	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
21	79	2.54	15.38	0.92	2	1.41	0.315	0.499	0.511	5.40E-79
3	10	5.40E-79	2.74	2.41	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79

2.3 Sum each nutrient per person in on day

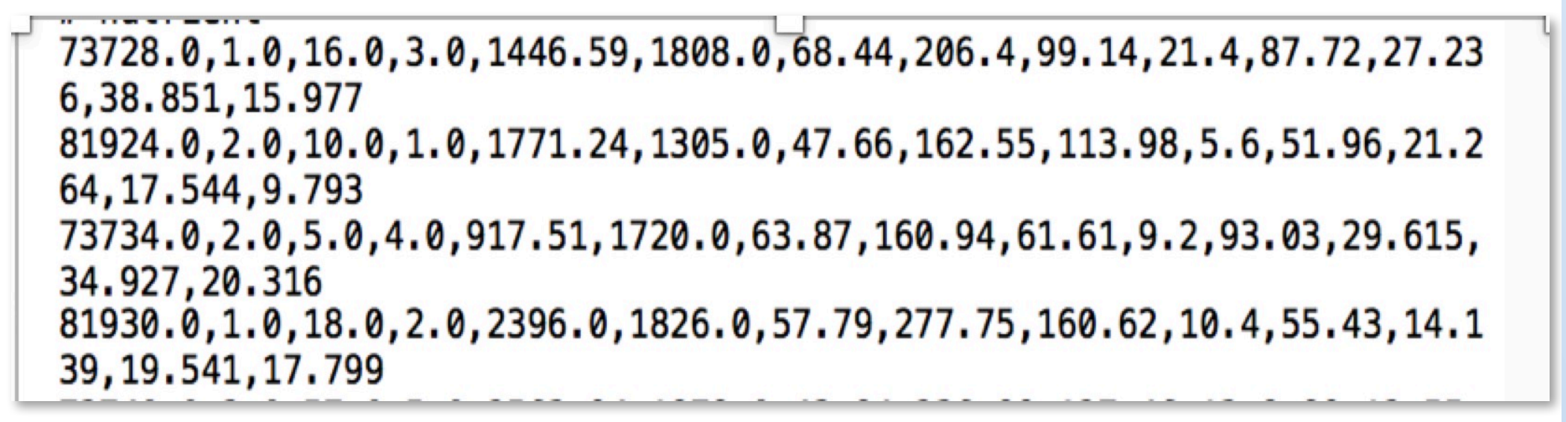
Every SEQN represent a person. Sum each column to get each nutrient total.



SUM

SEQN	WTDRD1	WTDR2D	DR1ILINE	DR1DRSTZ	DR1IGRMS	DR1IKCAL	DR1IPROT	DR1ICARB	DR1ISUGR	DR1IFIBE	DR1ITFAT	DR1ISFAT	DR1IMFAT	DR1IPFAT
73557	16888.3279	12930.8906	1	1	84	228	10.11	22	8.03	0.7	11.08	3.462	3.893	1.118
73557	16888.3279	12930.8906	2	1	359.1	4	0.43	5.40E-79	5.40E-79	5.40E-79	0.07	0.007	0.054	0.004
73557	16888.3279	12930.8906	3	1	2	7	5.40E-79	1.82	1.61	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
73557	16888.3279	12930.8906	4	1	449.5	211	3.06	50.66	37.35	1.3	0.54	0.063	0.099	0.135
73557	16888.3279	12930.8906	5	1	131	62	1.23	15.39	12.25	3.1	0.16	0.02	0.03	0.033
73557	16888.3279	12930.8906	6	1	42.75	210	3.22	26.3	21.58	1	10.2	3.877	3.367	1.288
73557	16888.3279	12930.8906	7	1	220.5	115	0.79	29.66	27.89	1.8	0.27	0.02	0.029	0.088
73557	16888.3279	12930.8906	8	1	8.5	40	0.6	5.69	0.02	0.5	1.77	0.354	0.408	0.93
73557	16888.3279	12930.8906	9	1	173.5	437	24.19	23.84	6.11	2.4	27.17	10.016	10.613	5.233
73557	16888.3279	12930.8906	10	1	620	260	5.40E-79	64.23	61.63	5.40E-79	1.55	5.40E-79	5.40E-79	5.40E-79
73557	16888.3279	12930.8906	11	1	960	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
73558	17932.1439	12684.1489	1	1	600	6	0.72	5.40E-79	5.40E-79	5.40E-79	0.12	0.012	0.09	0.006
73558	17932.1439	12684.1489	2	1	152.5	93	4.8	7.32	7.7	5.40E-79	4.96	2.844	1.238	0.297
73558	17932.1439	12684.1489	3	1	1200	1759	201.23	146.38	12.05	8.6	34.1	9.372	8.784	9.792
73558	17932.1439	12684.1489	4	1	3840	1114	9.22	62.98	3.46	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79
73558	17932.1439	12684.1489	5	1	665	903	54.38	89.45	1.59	3.9	35.68	16.698	10.926	4.848
73558	17932.1439	12684.1489	6	1	665	964	56.25	100.08	1.71	4.2	37.53	17.656	11.471	4.848
73558	17932.1439	12684.1489	7	1	366	223	11.53	17.57	18.48	5.40E-79	11.9	6.826	2.972	0.714
73558	17932.1439	12684.1489	8	1	360	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79	5.40E-79

We collect 8661 sets of data after preprocessing data.
Here is the sample of data



```
73728.0,1.0,16.0,3.0,1446.59,1808.0,68.44,206.4,99.14,21.4,87.72,27.23
6,38.851,15.977
81924.0,2.0,10.0,1.0,1771.24,1305.0,47.66,162.55,113.98,5.6,51.96,21.2
64,17.544,9.793
73734.0,2.0,5.0,4.0,917.51,1720.0,63.87,160.94,61.61,9.2,93.03,29.615,
34.927,20.316
81930.0,1.0,18.0,2.0,2396.0,1826.0,57.79,277.75,160.62,10.4,55.43,14.1
39,19.541,17.799
```

Order:
SEQN,SEX,AGE,RACE Grams, Energy, protein, carbohydrate, Total sugars, Dietary fiber, Total fat , Total saturated fatty acids, Total monounsaturated fatty acids, Total polyunsaturated fatty acids.

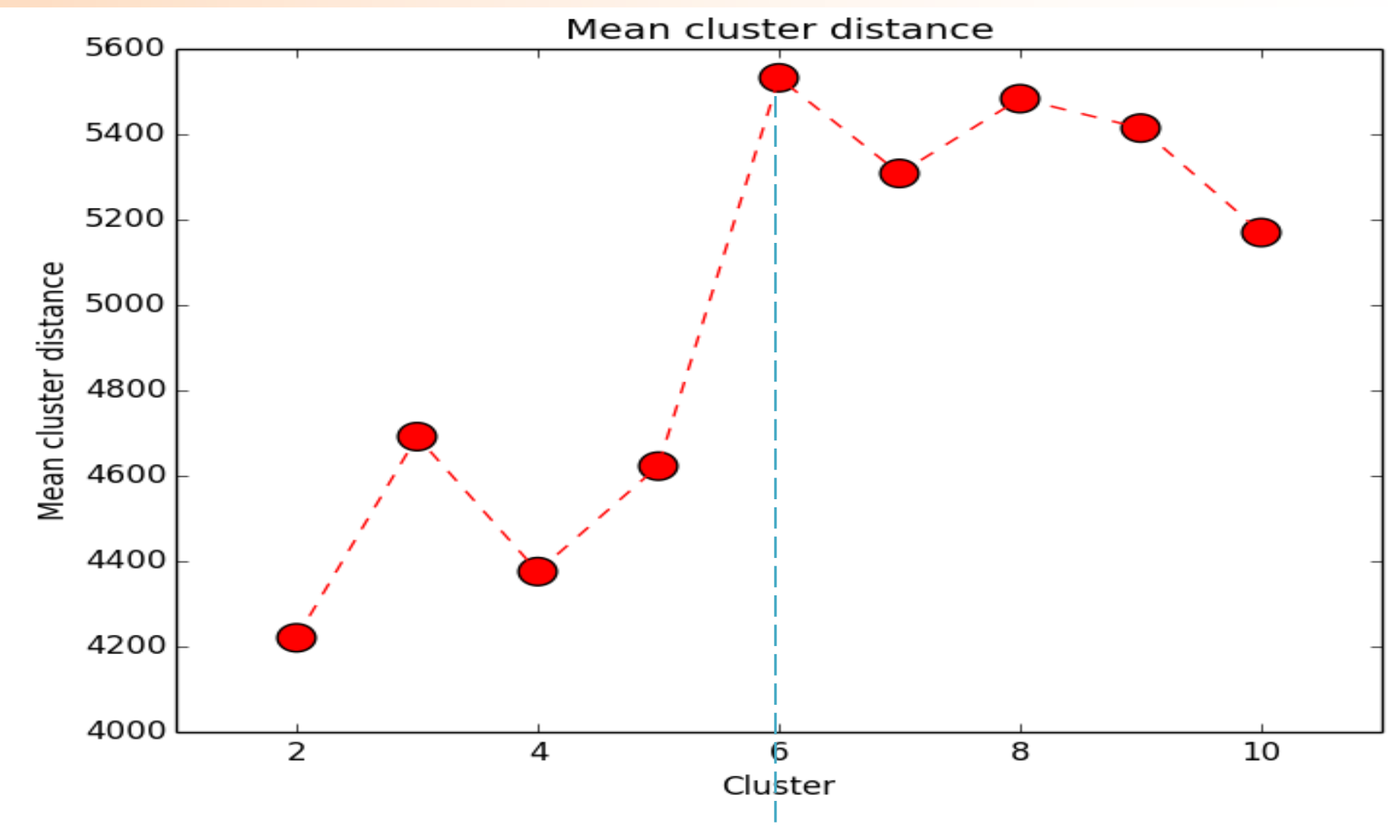
03 Data processing by using K-means

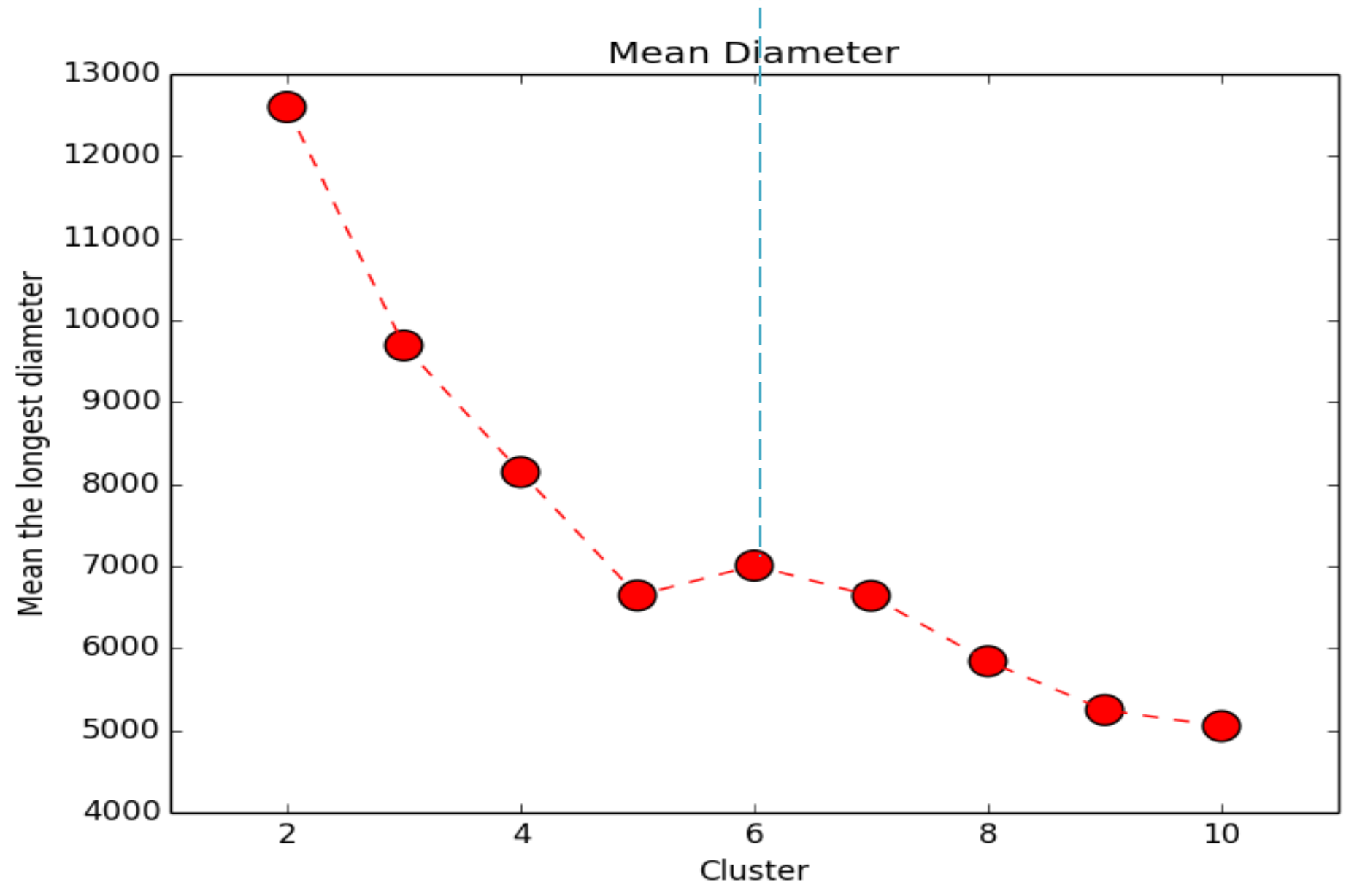
K-Means is one of the most popular "clustering" algorithms.
K-means stores k centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.
K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters.

3.1 deciding K in K-means

We mainly consider two aspects, one is Cluster Distance Mean, the other one is Mean Diameter. Large Mean Cluster Distance increase cluster diversity. Small Mean Diameter increase cluster accuracy.

Compared these two graph, I think six means would be better.





3.2 Conclusion

After K-means, I find there is a cluster center listed below has gotten the most nutrition.

[1.086, 36.236, 3.00, 7773.360, 5959.860, 223.834, 659.770, 302.647, 37.255, 227.238, 71.967, 81.091, 52.808]

In this cluster, the first value symbolize male, since male is one and female is two in NHANES dataset, the second value symbolize 36 years old, and the third one symbolize that race is not the key reason of nutrition intake. Because there are five races (1.0-5.0) and the value is the average races of cluster. It is nearby 2.5 although it has some bias.

Result

Male around 36 years old need to pay attention to their nutrition intake!