PART I:

1. ***Using all of the variables except name, and rating, run the k-means clustering algorithm with k=2. Describe the characteristics of the cereals within the cluster***

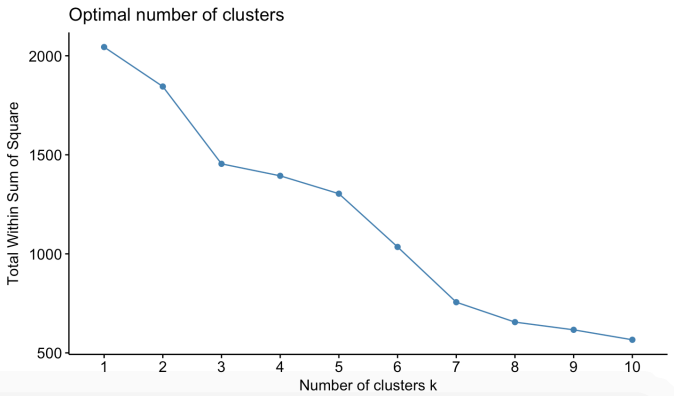
To run the kmeans test, the first step is to remove the na value of the dataset. For simplicity I used na.omit instead of replacing them with the average value of column. After that I dummied the variables manuf and type and normalized the rest. The result were as follows:

```
K-means clustering with 2 clusters of sizes 36, 38

Cluster means:
     Calories    Protein        Fat      Sodium      Fiber      Carbo     Sugars     Potass     Vitamins       Shelf      Weight        Cups       Cold    Nabisco
1 -0.03215791  0.1681824 -0.3310734 -0.09904820  0.2484544  0.2550297 -0.1713645  0.1405036  0.005051172 -0.09293397  0.09248862 -0.04458985 -0.1227058  0.2822196
2  0.03046539 -0.1593307  0.3136485  0.09383513 -0.2353778 -0.2416071  0.1623453 -0.1331087 -0.004785321  0.08804271 -0.08762080  0.04224301  0.1162476 -0.2673659
       Quaker   Kelloggs GeneralMills    Ralston       AHFP     ManufA     ManufG     ManufK     ManufN     ManufP     ManufQ     ManufR       TypeC       TypeH
1 -0.3210386  0.7040532   -0.6460338  0.3388740  0.1227058  0.1227058 -0.6460338  0.7040532  0.2822196 -0.3695814 -0.3210386  0.3388740 -0.1227058  0.1227058
2  0.3041418 -0.6669978    0.6120320 -0.3210386 -0.1162476 -0.1162476  0.6120320 -0.6669978 -0.2673659  0.3501298  0.3041418 -0.3210386  0.1162476 -0.1162476
```

As we can observe above, there are two clusters of size 36 and 38, fairly similar. The characteristics of cluster 1 are that they are cereals that are high in protein, fiber, carbo, potassium, but they are low in fat and sugars. They are also most likely produced by Kellogg's, by manufacturer K and probably not by Quaker or General Mills.

On the other hand, cluster 2 cereals are high in fat, sugar and low in protein, fiber, carbo, potassium, and most likely produced by General Mills or Quaker.

2. ***Vary the value of k. What is the value of k that provides the best clusters based on the elbow method? Why do you think so? Describe the characteristics of the cereals within the cluster when the optimal k is chosen***



From this graph, one can see that the optimal value of cluster for cereals is most likely 7.

3. ***With the value of k chosen in question 2, use cluster membership to predict rating by a simple linear regression. Describe the relationship that you uncovered in this prediction***

To be able to perform this regression, a kmeans test was performed with the optimal 7 clusters and the results of this were added to the data table as a new column. A simple linear regression as then made using the rating of the cereals as the dependent variables and the cluster as the independent.

```
Call:
lm(formula = Rating ~ cereals4$km.cereals7.cluster, data = cereals4)

Residuals:
    Min      1Q  Median      3Q     Max
-23.743  -7.784  -0.493   5.519  45.383

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     40.002      2.634  15.189  < 2e-16 ***
cereals4$km.cereals7.cluster2    1.783      5.076   0.351   0.7264
cereals4$km.cereals7.cluster3   14.849     11.778   1.261   0.2118
cereals4$km.cereals7.cluster4    8.320      4.132   2.014   0.0481 *
cereals4$km.cereals7.cluster5    2.563      5.076   0.505   0.6152
cereals4$km.cereals7.cluster6   28.653      5.770   4.966 4.98e-06 ***
cereals4$km.cereals7.cluster7   -5.517      3.595  -1.534   0.1296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.48 on 67 degrees of freedom
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3309
F-statistic: 7.016 on 6 and 67 DF,  p-value: 8.064e-06
```

Having cluster 1 as the reference group, one can see that that the cluster with the highest rating is cluster 6 which has on average a higher rating of 28.653 followed by cluster 3 with an average higher rating of 14.849. Still compared to cluster 1, only cluster 7 has lower average rating of -5.517.

Furthermore, from this one can see that the most significant clusters are cluster 6 and 4 and that the overall regression is significant as the p-value is very low (8.064*10^-6).

PART II:

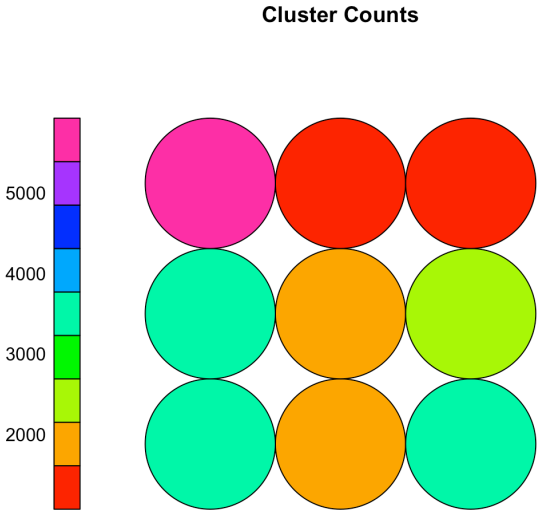
4. Apply the Kohonen clustering algorithm to the dataset (excluding income). Use 3x3 topology.

```
# Kohonen Clustering
som.adult <- som(as.matrix(adult3),
grid=somgrid(3,3),
rlen=170,alpha=c(0.3,0.00),radius=2)
```

This is the code to run the actual kohonen network. To be able to run this code, the first thing I did was to merge the factor columns that had many factors with common characteristics. For example: in the matital.status column there were many types of married that were grouped as married or in the native.country column countries from each continents were merge to avoid having 42 different levels. See the r code to have details on all the mergers.

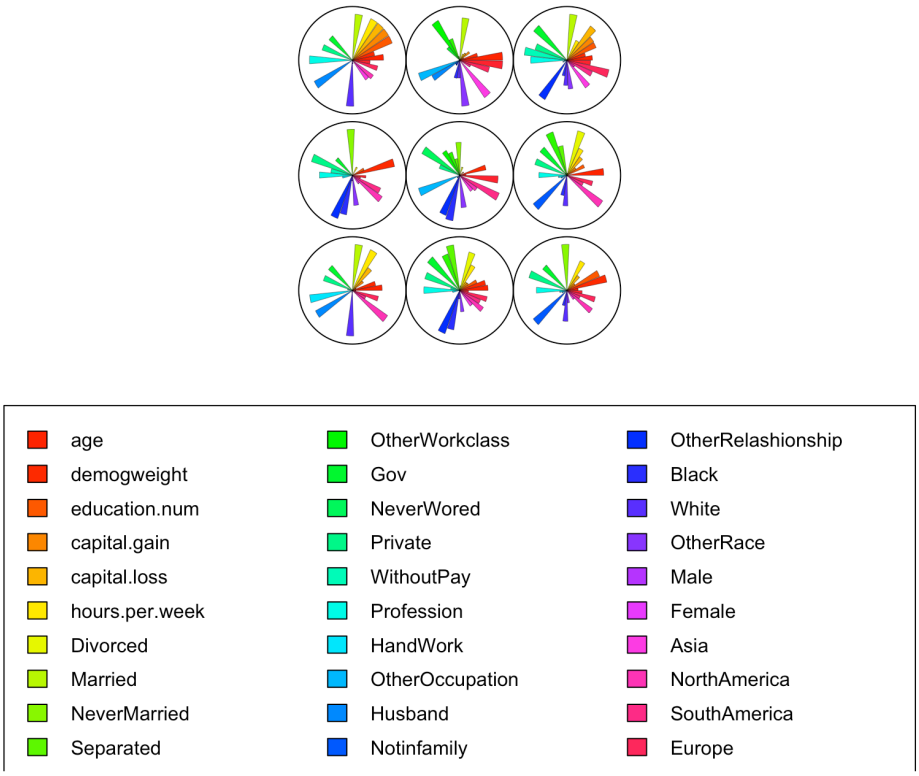
Once that was done, each factor became a column in a new data with only 0 and 1's as the kohonen network only accepts numerical values and the rest of the variables were normalized. Only then could the above line of code be run.

5. Generate numerical summaries for the clusters (e.g., cluster mean summary). Describe the characteristics of each cluster



This first diagram represents the number of individuals in each cluster. As one can see the colour represents the amount of observations in each cluster ranging from under 2000 up to over 5000. The variety of colour shows graphically that there is not a fair amount of observations in each cluster. For example, number 2 and 3 have under 2000 while number 1 has over 5000.

Codes plot



This next diagram represents the characteristics of each cluster. In the below box is the colour assigned to each characteristic and each circle represents a cluster. Inside the cluster the length of the bar assigned to such colour represents the importance of such variable in such cluster. For example: there are a lot of males in cluster 2 as the purple line is relatively intense or there are a lot of white people in cluster 1 as the blue-purple line is very intense.

6. ***Do you see any relationship between cluster membership and income? If you can, describe the relationship***

	1	2	3	4	5	6	7	8	9
<=50K.	69.35	91.82	90.82	98.38	98.93	87.70	44.41	81.67	50.05
>50K.	30.65	8.18	9.18	1.62	1.07	12.30	55.59	18.33	49.95

The above visual represent the percentage of people in both types of income in each cluster. As one can see there is a clear relationship. Above 87% of people in cluster 2,3,4,5,6 have income equal or higher than 50K and in cluster 8 about 80% do. This shows that is you are at an income above 50K the you will most likely be in such cluster.

PART III:

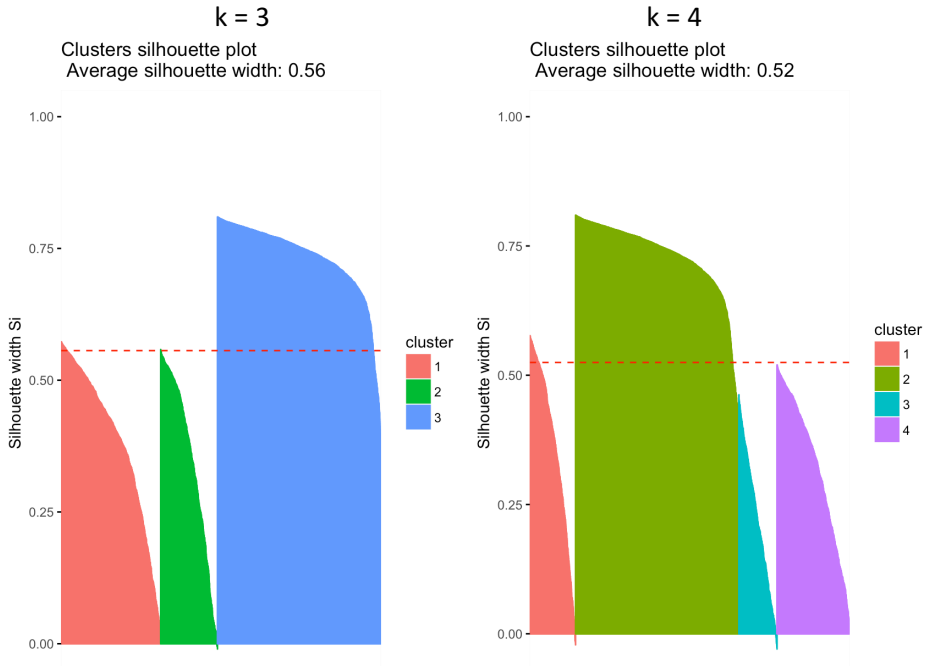
7. Run k-means with k=3 and k=4 to generate two cluster models with the training dataset.

To be able to the kmean clustering, the first step was to clean the data set and omit the na values. The column referring to interest rates was then also removed as it is high correlated with the request amount and keeping it would mean assigning more importance to such variable than to the others. Then as all the values are numerical, they were normalised.

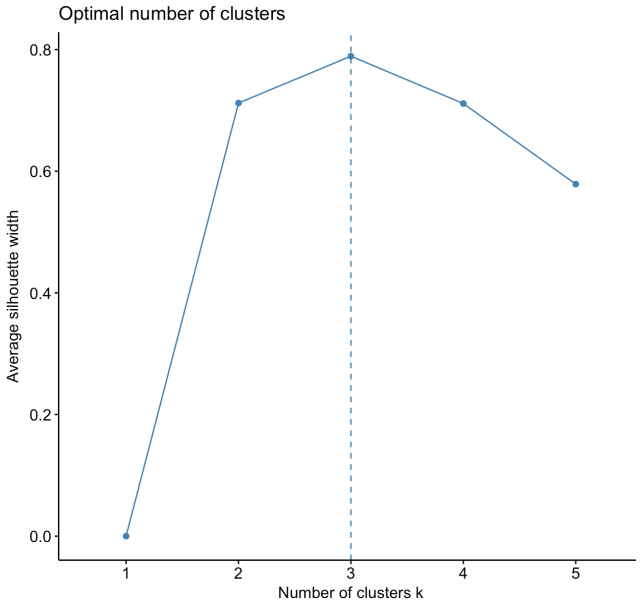
The results of the 2 kmeans are as follow:

```
K-means clustering with 3 clusters of sizes 1089, 1788, 623

Cluster means:
  Debt-to-Income Ratio FICO Score Request Amount approvaldummy
1            0.2105971  0.5062548      0.1694725             0
2            0.1507472  0.7165566      0.3091529             1
3            0.2540418  0.6075315      0.6281632             0

K-means clustering with 4 clusters of sizes 586, 1202, 1089, 623

Cluster means:
  Debt-to-Income Ratio FICO Score Request Amount approvaldummy
1            0.1565920  0.7244047      0.5500780             1
2            0.1478978  0.7127305      0.1916970             1
3            0.2105971  0.5062548      0.1694725             0
4            0.2540418  0.6075315      0.6281632             0
```

8. Generate a silhouette plot of both cluster models and calculate the mean silhouette values for each cluster in both cluster models and the overall mean silhouette for each cluster model. Which cluster model is preferred? Why?



The diagram to the left show the silhouette plot for k = 3 and the right it shows the silhouette plot for k = 4. From his graph, one can see that k = 3 has a higher average silhouette value and passes the 0.5 threshold of acceptance and that it therefore shows that 3 clusters are a better representation of the data.



The above plot is just another way to determine what the ideal number of clusters is with the silhouette method. This simply confirms that 3 is probably the ideal number of clusters.

9. Calculate the pseudo-F statistics for the two cluster models. Which cluster model is preferred? Why?

Using a programmed function, here is the output of the pseudo-F-statistic for model from k = 1 to k = 5.

```
$k
[1] 1 2 3 4 5

$W
[1] 1223.0677  316.5140  228.3232  177.6428  149.4005

$pF
[1]       Inf 10018.907  7617.756  6857.967  6279.206
```

From this one can see that between k = 3 and k = 4 the ideal number of cluster seems to be 3 as the value is highest and the higher the value the higher the evidence for clusters.

10. ***With the test dataset, apply k-means clustering with k=3. Perform cluster validation. Do you observe any differences between clusters created by training and test dataset? If so, how?***

	Cluster 1	Cluster 2	Cluster 3
Test Data Mean	0.2008280	0.1667789	0.18097416
Train Data Mean	0.2042792	0.2464205	0.14622483
Test Data Std Dev	0.1364496	0.1355737	0.12627219
Train Data Atd Dev	0.1690470	0.1733856	0.07538752

This is the result of running cluster validation. As one can see, the means of the clusters for the test and the train are relatively similar. In fact, they are pretty much the same for cluster 1 and vary a little for cluster 2 and 3. However the difference in mean for both of these is included in the standard deviation, making me think that the clusters are probably good as the data is clustered in the same way for both datasets.