Assignment 1 - Clustering with K-means

- Ryan Jones

Introduction -

Investigation: I am going to investigate the link between age and the number of hours people work per week. Then I will investigate the differences in the results based on what sex each person is. There are different methods to cluster data. I am going to discuss a few of them in this report and compare the performance of each.

The clustering problem is described as the problem of locating homogenous groupings of data points in each data collection in its most basic form. Each of these groups is referred to as a cluster, which is defined as a place where the density of items is higher locally than in other areas. (Jain, Murty and Flynn, 1999)



Table

Description automatically generatedFigure 1 (Bação, Lobo and Painho)

The outcomes of the two methods in dataset DS1 are shown in Figure 1. The structural coherence of the clustering varies relatively little, indicating that SOM is more stable than k-means. When standard deviation is low (all observations are close to the clusters centroids), k-means performs poorly, failing structural coherence in more than half of the tests. The SOM, on the other hand, only gets the appropriate structure 10% of the time. K-means enhances the percentage of runs in which the structural coherence is correct as the standard deviation increases. Nonetheless, it never reaches the 100 percent threshold, where SOM scores in every run with a standard deviation of 0.2 to 0.9. (Bação, Lobo and Painho)

Although SOM is the better performing method (SOM Clustering Method, RPubs), I will still be using k-means for the purpose of this report. Before I started preparing the data, I wanted to make sure that I knew what I needed from the data. To find this out I also needed to know how the k-means algorithm worked. So, I sourced some code from GitHub (kshitijved, Github).

Therefore, I am going to walk through my research into the k-means algorithm and how I came to develop the solution for the adult census dataset.

K-Means Clustering with Iris Dataset (Without Elbow Method)

Graphical user interface, text

Description automatically generatedImporting libraries

Figure 2 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm\_Iris)

Graphical user interface, text

Description automatically generatedCapturing the current working directory to ensure no matter where the code is run it should work. Also, as you can see the first 5 rows of the dataset are outputted here, this can be helpful to make a better decision on the type of approach to use with the dataset.

Figure 4 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm\_Iris)

The next job is plotting the data. Which includes selecting points of interest (two columns in this case, which I have selected **above** x1 and x2), then plotting the dataset onto the scatter graph shown **below**.

Chart, scatter chart

Description automatically generatedFigure 5 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm\_Iris)

Now that the data is clearly being plotted on the graph clearly it is time to implement k-means. The two separate arrays are not zipped together into **X** which is then plotted on the graph though iteration. Now showing k-means, the for loop iterates through each pair between the two (now zipped) arrays.

The iteration can also be explained like this:

First iteration: array([[0, 0]])

Second iteration: array([[1, 1]])

Third iteration: array([[2, 2]])

Chart, scatter chart

Description automatically generated

Figure 6 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm\_Iris)

K-Means Clustering with Iris Dataset (With Elbow Method)

Although the above implementation is correct, I need to implement the elbow method to show how I came up with the value for n\_clusters as 3. Using the same Iris dataset (R.A. Fisher, Iris) I will demonstrate the elbow method below.

K-Means is an unsupervised machine learning algorithm that groups data into k number of clusters. The number of clusters is user defined and the algorithm will attempt to group the data even if this number is not optimal for the specific case.

Therefore, the Elbow Method was developed. It is a very popular technique which helps its users decide on how many clusters to use for the K-Means model. The idea is to run K-Means clustering for a range of clusters k (let’s say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned centre(distortions).

Graphical user interface, text

Description automatically generatedImporting libraries and capturing the current working directory.

Figure 7 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm-ElbowMethod\_Iris)

Text

Description automatically generated with medium confidenceReading csv and changing species values to numeric values.

Figure 8 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm-ElbowMethod\_Iris)

Graphical user interface, text, application

Description automatically generatedRunning K-Means with a rand of k and collecting distortions into a list.

Figure 9 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm-ElbowMethod\_Iris)

Chart, line chart

Description automatically generatedNow to plot the graph to show the optimal k value, the k values are used as the x label and the distortions as the y label. The elbow is the number 3 because that is when the slope reaches the ‘elbow’ point. Which means that n\_clusters = 3 is the optimal cluster value for this dataset. Now I can run a K-Means using the new n\_clusters value.

Figure 10 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm-ElbowMethod\_Iris)

Now that the Elbow method has been implemented and the K-Means algorithm run again using the value given for n\_clusters. This is a more accurate method of completing a K-Means cluster on a data set, because the value for n\_clusters is justified.

Chart, scatter chart

Description automatically generatedFigure 11 (R.A. Fisher, Iris; RyanJones999, K-Means-Algorithm-ElbowMethod\_Iris)

Data Manipulation -

Now to prepare the adult dataset.

Graphical user interface, text

Description automatically generatedImporting libraries and capturing the current working directory.

Figure 12 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Loading the dataset in and creating the main dataframe.

Figure 13 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)Graphical user interface

Description automatically generated

Table

Description automatically generatedChecking the columns and their datatypes.

Figure 14 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Graphical user interface, application

Description automatically generatedCreating a copy of the main dataframe so that I can edit it without disturbing the master copy.

Figure 15 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Creating list of columns, I want to drop. These columns are not going to be included in my investigation, so I do not need them.

Graphical user interface, text, application, Word

Description automatically generated

Figure 16 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

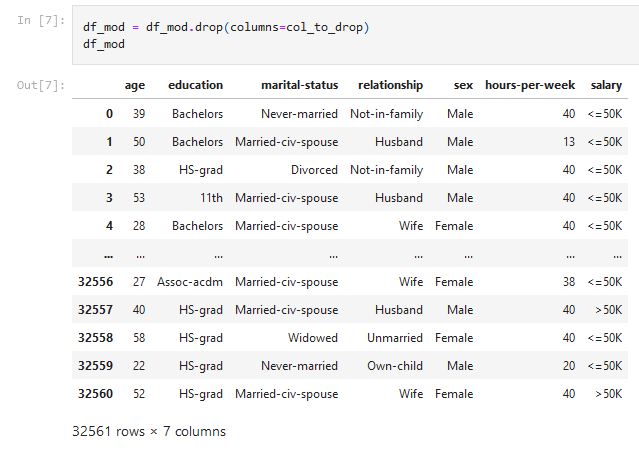
Dropping unneeded columns

Figure 17 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Graphical user interface, text, application, email

Description automatically generatedWith this result I can see that I still need to sort the data into more of a binary approach if the clustering is going to work. Now that I have made a copy of the edited dataset, I can start to pick points of interest.

Figure 18 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

As you can see, I am now left with three columns or points of interest. age, sex and hours-per-week. I will be able to build the scatter graph showing how age relates to how many hours people work. But I also have the possibility of separating the data into two scatter graphs which would also show not just how age affects, but also sex.

Now I am going to split this data into Male and Female and change the value of the cells to 1 or 0.

Text

Description automatically generated with low confidence

Figure 20 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Separated Male Data into its own data frame.

Figure 21 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)A picture containing text

Description automatically generated

Table

Description automatically generated with low confidenceSeparated Female Data into its own data frame.

Figure 22 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Text

Description automatically generatedWriting changes to an excel file ready for K-Means Clustering

Figure 23 (Kohavi and Becky, ; RyanJones999, Data\_Manipulation)

Final Implementation –

Now the data has been prepared, I am going to apply what I have learnt about K-Means and the Elbow Method to the adult dataset. With the aim of demonstrating the relationship between how many hours they work per week depending on age. Then illustrate the difference between the results depending on a person’s sex.

Text

Description automatically generatedImporting Libraries

Figure 24 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Graphical user interface, text, application, email

Description automatically generatedCreating Functions. I decided to use functions for this as it seemed to speed up the processing time, and it helped reference variables when programming the solution.

Figure 25 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Table

Description automatically generatedReading Files and displaying the head rows.

Figure 26 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

The xlsx writer library seems to automatically put an autoincrement column in which in this case

Graphical user interface, text, application, email

Description automatically generatedis not necessary so I can remove it here. Also, for the purposes of what I am investigating, I do not actually need the 'sex' column. For recognising what sex each data set is representing I have named them appropriately. So I can use the DelUnCol method to delete these unneeded columns.

Figure 27 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Final Implementation – Male & Female

Graphical user interface, text, application

Description automatically generatedProducing KMeans model for the Final\_Result Worksheet DataSet which includes the data for both sex's combined.

Figure 28 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Chart, line chart

Description automatically generatedThe optimal number of clusters is 2, as the elbow method above demonstrates. The 'elbow' point on the graph is at 2.

Figure 29 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Calling functions to draw the graphGraphical user interface, text, application

Description automatically generated

Figure 30 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

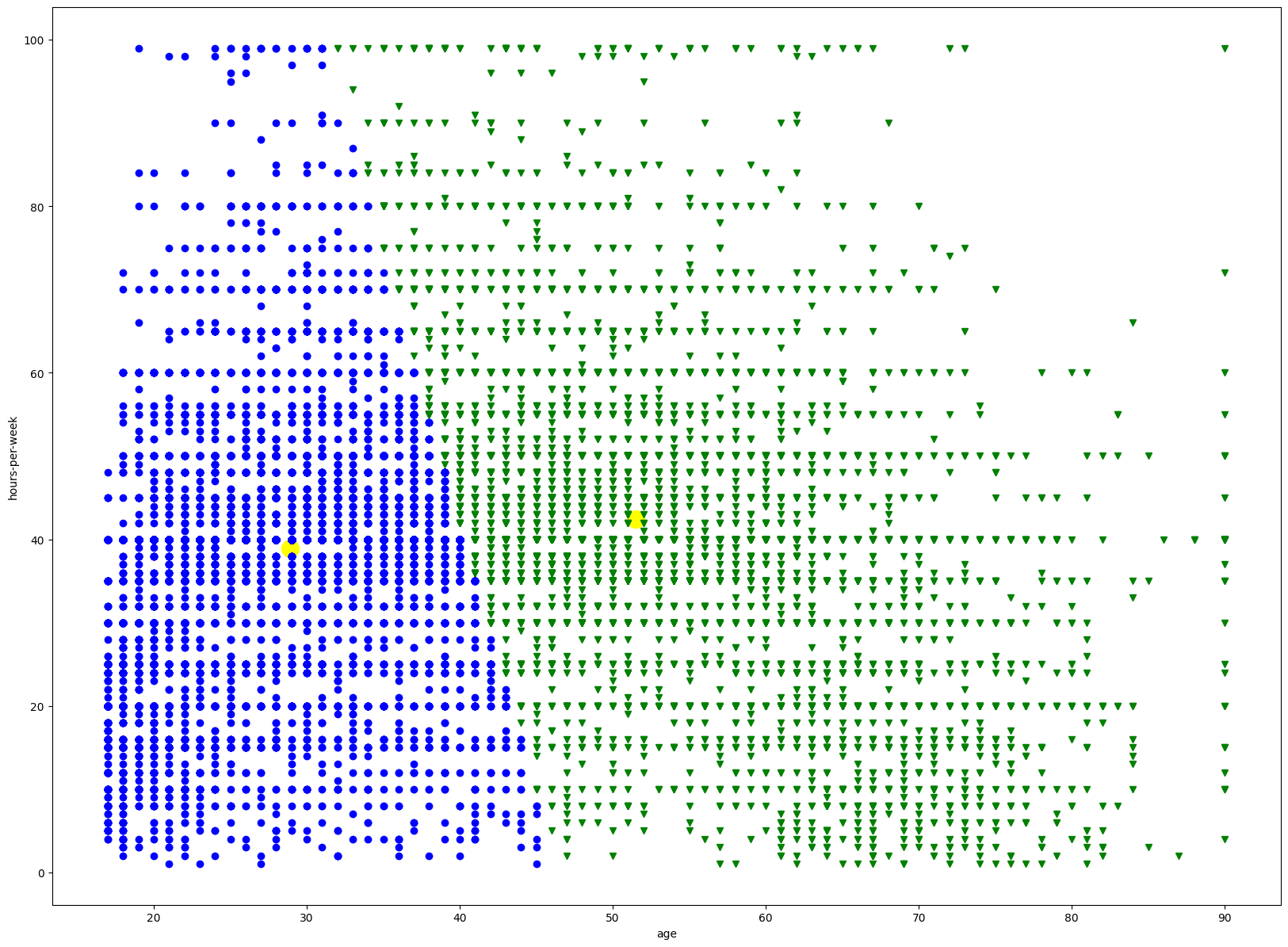
Graph for the Final\_Result worksheet dataset which includes the Male and Female datasets combined.

Figure 31 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Final Implementation – Male

Graphical user interface, text, application

Description automatically generatedProducing KMeans model for the Final\_MaleData Worksheet DataSet.

Figure 32 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Chart, line chart

Description automatically generatedThe optimal number of clusters is 2, as the elbow method above demonstrates. The 'elbow' point on the graph is at 2

Figure 33 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Building the KMeans Model for the Final\_MaleData worksheet dataset.

Figure 34 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

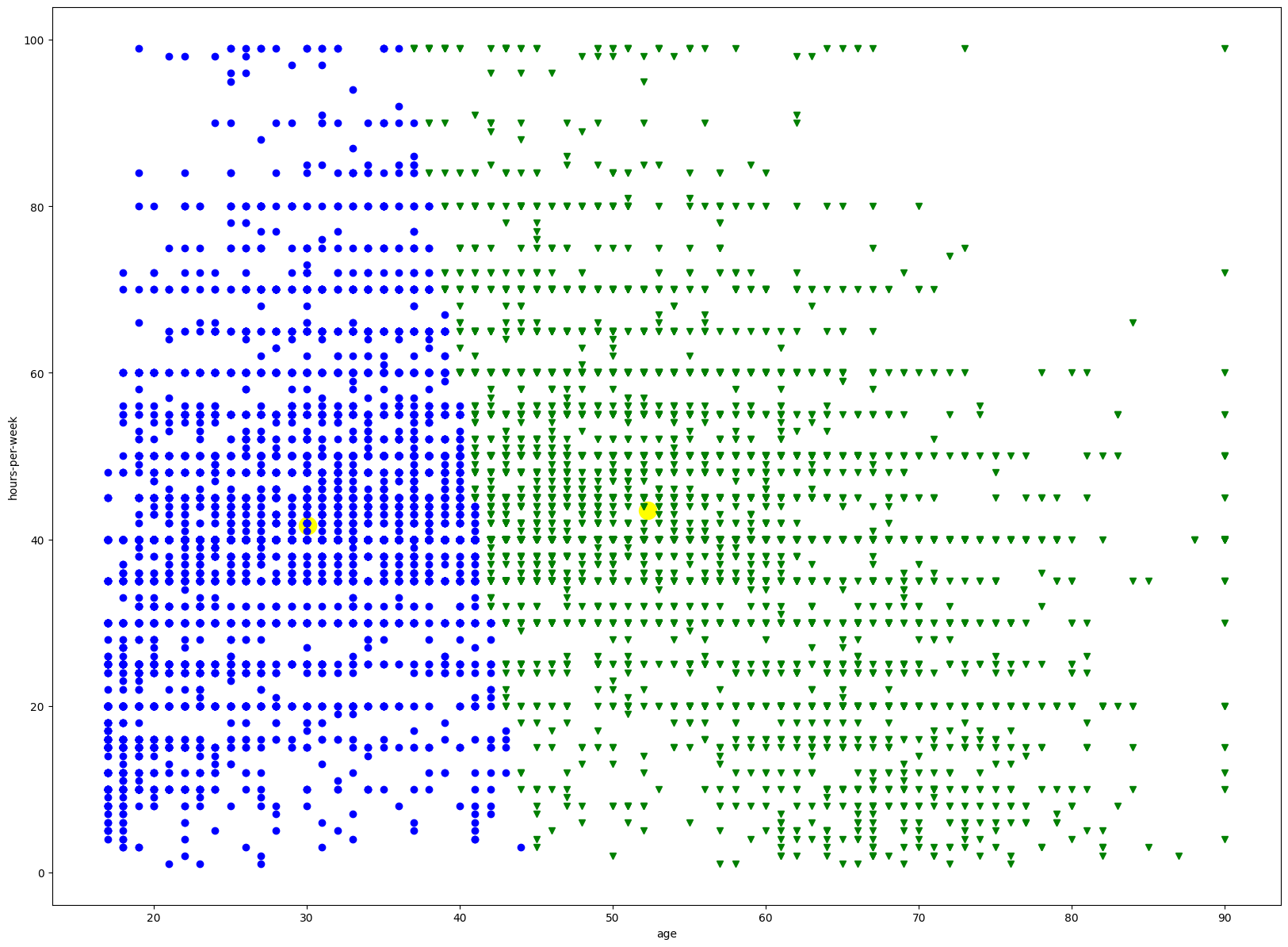
Graph for the Final\_MaleData worksheet dataset.

Figure 35 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Final Implementation – Female

Producing KMeans model for the Final\_FemaleData Worksheet dataset.Graphical user interface, text, application

Description automatically generated

Figure 36 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Chart, line chart

Description automatically generatedThe optimal number of clusters is 2, as the elbow method above demonstrates. The 'elbow' point on the graph is at 2.

Figure 37 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Building the KMeans Model for the Final\_FemaleData worksheet Graphical user interface, text, application

Description automatically generateddataset.

Figure 38 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

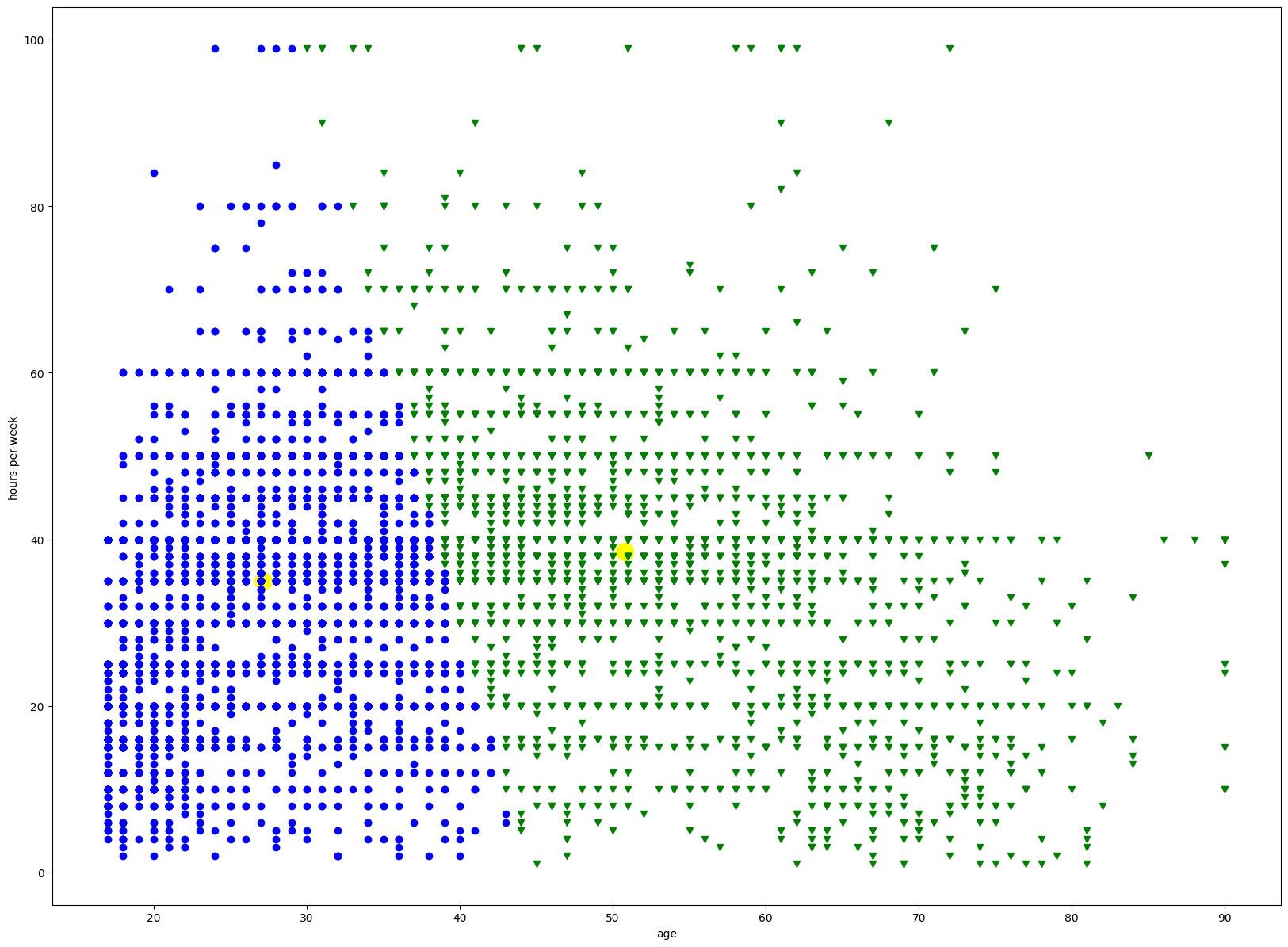
Graph for the Final\_FemaleData worksheet dataset.

Figure 39 (Kohavi and Becky, ; RyanJones999, K-Means-Cluster\_Census)

Observations –

As you can see the clustering has worked as planned, it has taken n = 9 for the number of clusters and successfully clustered the dataset into groups. The clustering has been accurate throughout the combined male & female data but also the separate datasets as well, and you can see clear differences between the two sexes because of this.

The relationship between age and hours-per-week worked is clearly shown and the results on all three datasets show that as people get older there are less hours worked. However, when looking at the comparison of hours-per-week worked and age depending on sex. The scatter graph clearly illustrates that on average Males work more hours-per-week across the board. Whereas the Female dataset scatter graph, clearly shows that on average Females work less hours-per-week; the difference between the two in terms of hours-per-week worked it extraordinary; especially as the age increases.

Conclusion –

After researching K-Means and now implementing it, I have found it a very useful technique of recognising groups in datasets. Not only does it help to make data more understandable, but it provides a method in which to be able to recognise patterns in data which can be fundamental for business, governing bodies and agencies alike as well as health services and scientists to be able to improve their services; research new methods to tackle issues which become obvious after reviewing the pattern. Research into conditions/ diseases like Diabetes and Cancer may rely on understanding patterns from surveys. K-Means is perfect for this. (RyanJones999, K-Means-Cluster\_Census)

To improve my method, I would try and create a function to be able to handle strings in the same way integers and floats are handled. This may help the program be more adaptable to different scenarios and would also prevent having to edit data where unnecessary (changing values from string to int/ float). For the purposes of this it would have been unneeded, but for future projects I would consider this.

To improve K-Means I would suggest that K-Means takes a different approach to initialization and repeating the K-Means algorithm. According to ((Fränti and Sieranoja, 2019)), he processing time is also multiplied by a factor of R when k-means is repeated. By separating the data into random subgroups, this can be compensated for. For example, dividing the data into R subsets of size N/R would result in a total processing time that is nearly the same as a single run.

References:

matplotlib Documentation  
. Available at: <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html?highlight=scatter#matplotlib.pyplot.scatter>

sklearn.cluster.KMeans Documentation  
. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

SOM Clustering Method  
. Available at: <https://rpubs.com/loveb/som>

Bação, F., Lobo, V. and Painho, M. 'Clustering census data: comparing the performance of  
self-organising maps and k-means algorithms'

Jain, A., Murty, M. and Flynn, P. (1999) Data Clustering: A Review ACM Press.

Kohavi, R. and Becky, B. Adult Census Data Set  
. Available at: <https://archive.ics.uci.edu/ml/datasets/adult>

kshitijved kmeans\_clustering\_algorithm. Available at: <https://github.com/kshitijved/kmeans_clustering_algorithm.git>

R.A. Fisher Iris Data Set. Available at: <https://archive.ics.uci.edu/ml/datasets/iris>

RyanJones999 (a) Data\_Manipulation. Available at: <https://github.com/RyanJones999/AI-Assignment-KMeans.git>

RyanJones999 (b) K-Means-Algorithm\_Iris. Available at: <https://github.com/RyanJones999/AI-Assignment-KMeans.git>

RyanJones999 (c) K-Means-Algorithm-ElbowMethod\_Iris. Available at: <https://github.com/RyanJones999/AI-Assignment-KMeans.git>

RyanJones999 (d) K-Means-Cluster\_Census. Available at: <https://github.com/RyanJones999/AI-Assignment-KMeans.git>

Fränti, P. and Sieranoja, S. (2019) 'How much can k-means be improved by using better initialization and repeats?', *Pattern recognition,* 93, pp. 95-112. doi: 10.1016/j.patcog.2019.04.014.