Exploratory Analysis and Similarity Estimation of Income and Iris dataset

Abhinav

Introduction

The objective is to calculate a similarity measure between objects for given datasets (Iris and Income) and output the k nearest objects and their similarity scores. The report aims to cover the 3 phases of the project

1. Exploratory analysis (of the income dataset)
2. Program description and design choices
3. Analysis of the result

Exploratory analysis of the income dataset

The first phase is basically aimed at getting a better understanding of the data that is being dealt with. The detected patterns and the observed inferences will later aid in deciding design features and will also play a part in understanding the final results of the project.

This project employs the visualization technique to explore the given data for patterns and trends. The histogram plot, plot of values and the frequency of their occurrence, is one of the common methods used which help assess key characteristics of the data distribution such as the peak, spread, symmetry and outliers for the numerical data. The barplot method has been applied to visualize the nominal and ordinal features of the given dataset, which provides us details on the count and the ratio of distribution among the classes present.

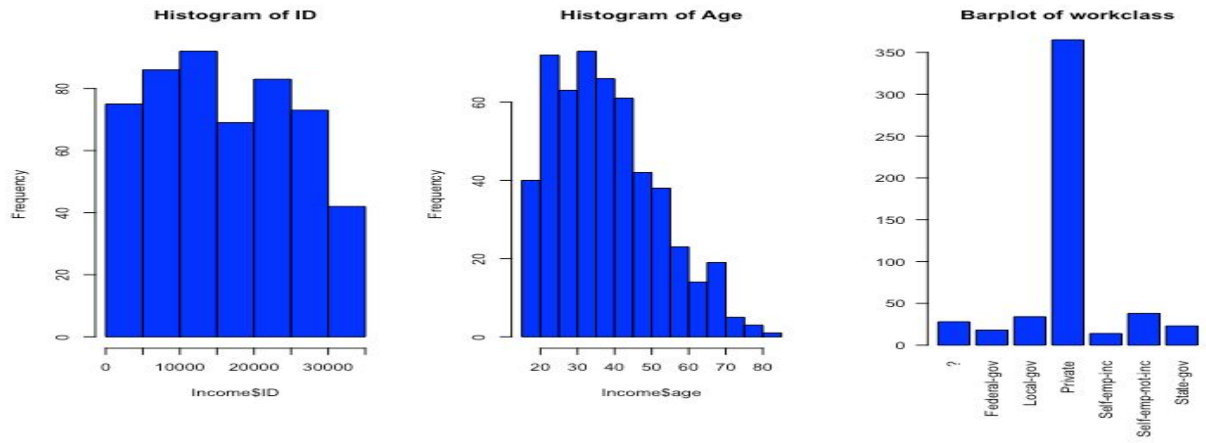


Fig 1.1

From the above figure (*fig 1.1*), it can be observed that the plot of "age" is skewed a bit right-skewed, that is most of the sample values seem to be clustered around the relatively younger population (20 to 40 years) with a peak at the category of 30-35 bin. It can seen from the barplot of "work class" that the current sample consists majorly of people employed in the private sector, which suggests that the final hypotheses made might have a relation with the distribution of data observed here. It can also be seen that the work class feature contains missing data which will have to handled during analysis.

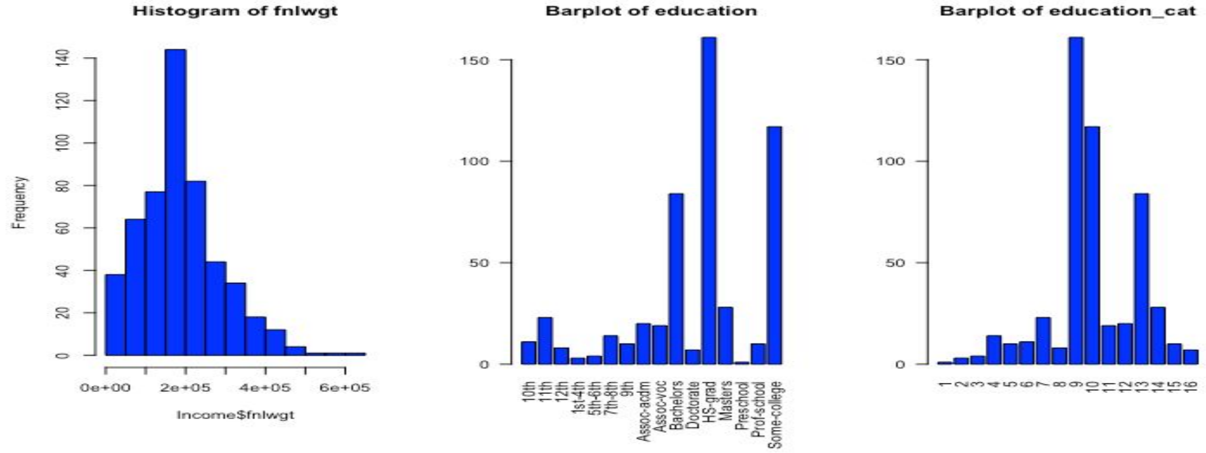


Fig 1.2

The feature “final weight” appears to have a gaussian distribution with a mean close to 20,000 as can be seen in the above figure (*fig 1.2*). On first glance the other two plots in the figure, “education” and “education_cat”, appear different, but looking closer it can be observed that both graphs have a similar peak value, equal number of categories, and the classes map on a one to one basis (*fig 1.3*). On further observation, it can be noticed that the numbers in education category is assigned to increasing level of classes in education, making it an ordinal variable.

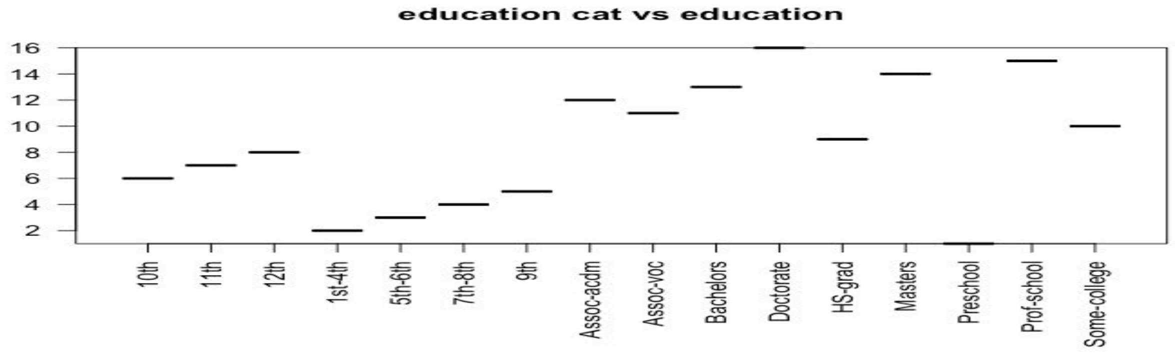


Fig 1.3

The barplots in the figures below (*fig 1.4 and fig 1.5*) provides a view on the distribution of data in features such as “marital status”, “occupation” , “relationship”, “gender”, “class” and “race”. Also Occupation contains missing data which will have to handled later.

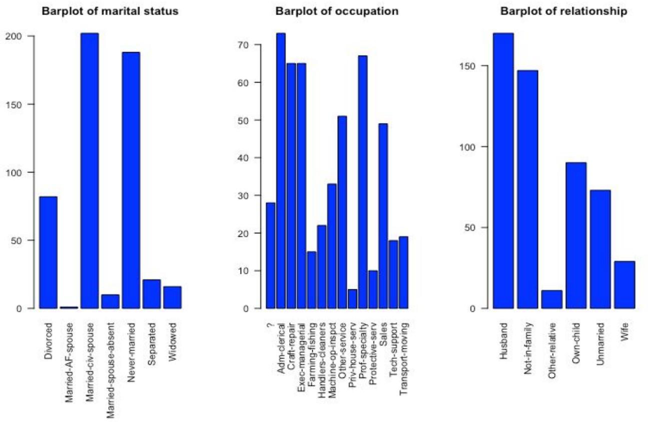


Fig 1.4

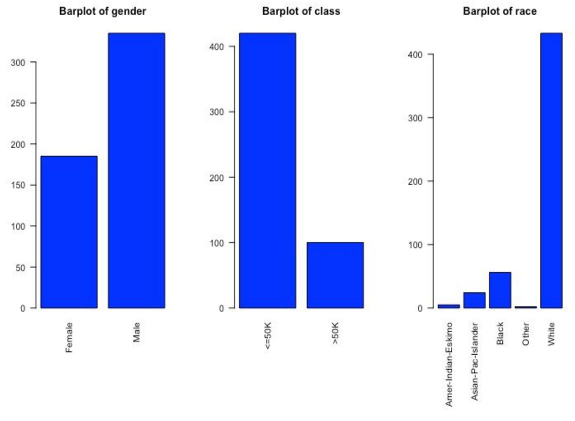


Fig 1.5

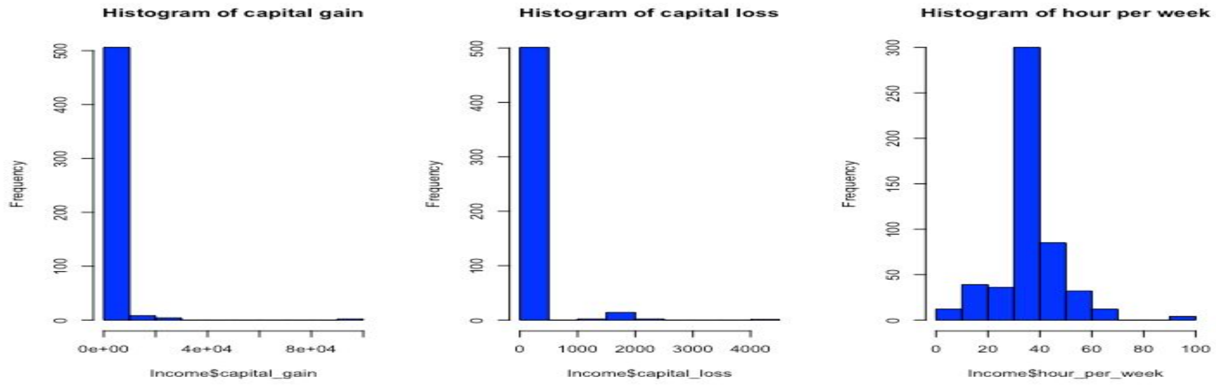


Fig 1.6

The histogram plot of both “capital gain” and “capital loss” (*fig 1.6*) both show a highly right-skewed data with majority of sample holding the value zero, which indicates that applying a log transformation to the data might give better insight into it. The plot also indicates the possible presence of outliers in both the data, with instances of data values that are far away from the rest of them. The “hour per week” plot shows a normal distribution centered near 40 and a sharp and symmetric drop on either side. The bar plot of native countries shown below (*fig 1.7*) indicates that the data majorly consists of US citizens and the results might be representative of that fact.

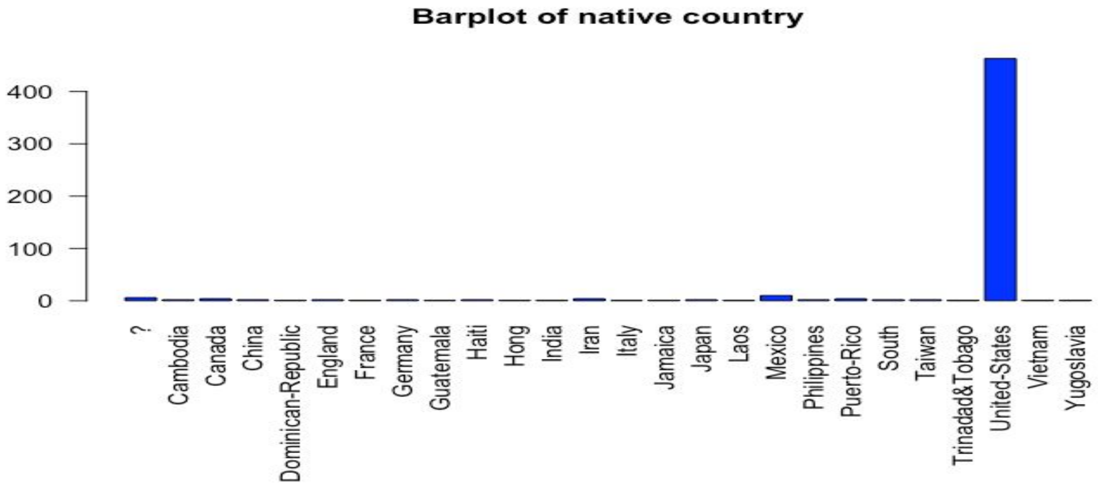


Fig 1.7

Some of the other visualization techniques that have been utilized are scatter plots (continuous vs continuous data) and boxplots (continuous vs noncontinuous data) to discover trends and relationship between the given features.

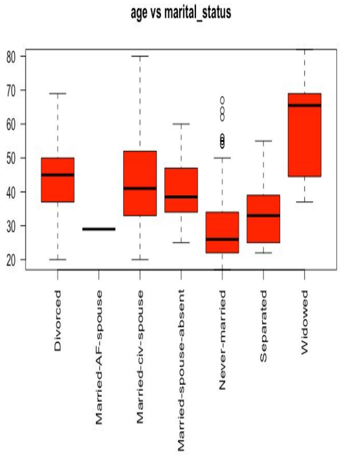


Fig 1.8

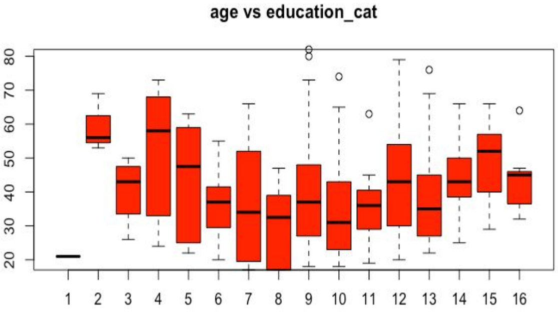


Fig 1.9

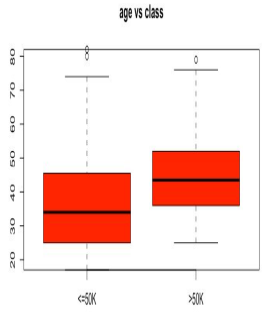
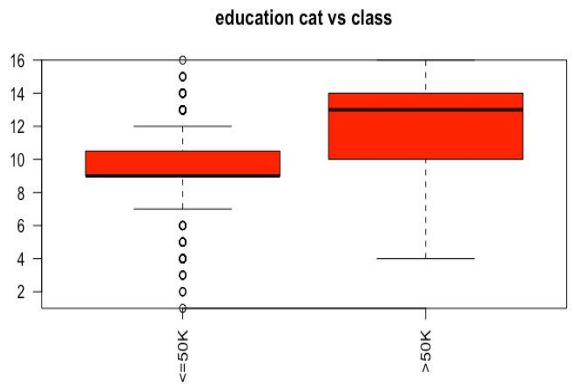


Fig 1.10

The plot of age against marital status (*fig 1.8*) shows trends that conform with the norm, the median of people who are widowed tend to be quite old (around 65) and the median of people who are unmarried tend to be quite young (around 25). It also interesting details such as the median age of people who are separated is close to 30 and divorcees near 45.

Similarly in the next plot (*fig 1.9*), it can seen that a lot of people with high school education or below tend to be very old, indicative of how education was probably not the focus during their days. As we move from 12th grade graduate to above, the pattern shows an increase in median age which probably indicates the increase in importance of education, the number of years it takes for each level. The third plot (*fig 1.10*) suggests that a higher ages indicates a greater chance of salary above $50,000.



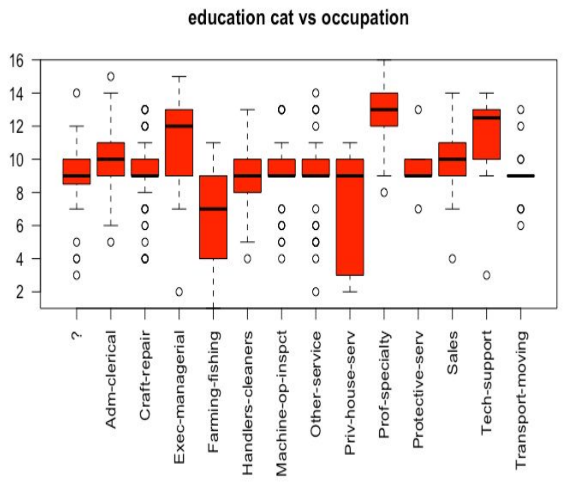
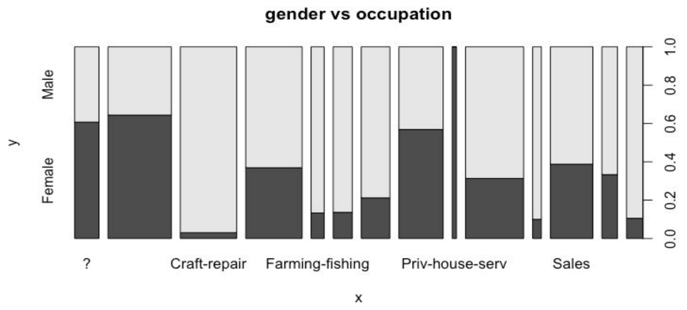


Fig 1.11

Fig 1.12

Few of the other interesting plots (*fig 1.11, fig 1.12, fig 1.13 and fig 1.14*) are shown above and below. While higher education category appears to have a direct correlation to pay greater than $50k, it is to be noted that there exists a few outliers in either extremes of education in the “less than $50k” category. Also, while the general median of education category for work lies around 9 (High school graduate), certain occupation such as sales, managerial executive and speciality requires a much higher level of education. The plots below describe the gender proportion in various occupations and work classes. It is to be notes that the observation noted are influenced by the nature of the data (eg: majority of the people concerned are white american population.)



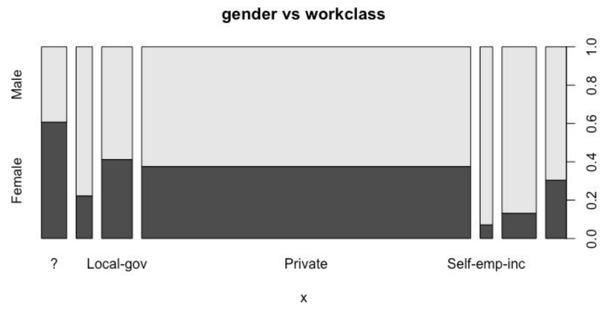


Fig 1.13

Fig 1.14

Program description and design choices

This section intends to describe and explain the decision behind the design of the project. After analysing the Income dataset and Iris dataset, it was observed that there was considerable difference between them and the initial processing and handling of the data required separate steps, which will be clarified below. Therefore two programs have been written and each handles a particular database.

Handling of Iris Dataset

The Iris dataset consists of four features with numeric data and one feature with nominal data. Since one of the project requirement is to not use the class for similarity calculations, only the four numeric features were chosen. From the histogram plots (*fig 2.1*) it can be observed that all features display normal distributions although some are multimodal. This shows that no transformations are required for the data although normalization is necessary to provide equal weightage, since the value of length tends to be twice as much as the width. The min-max normalization technique was used to scale down the information held by all four features to a range of (0, 1).

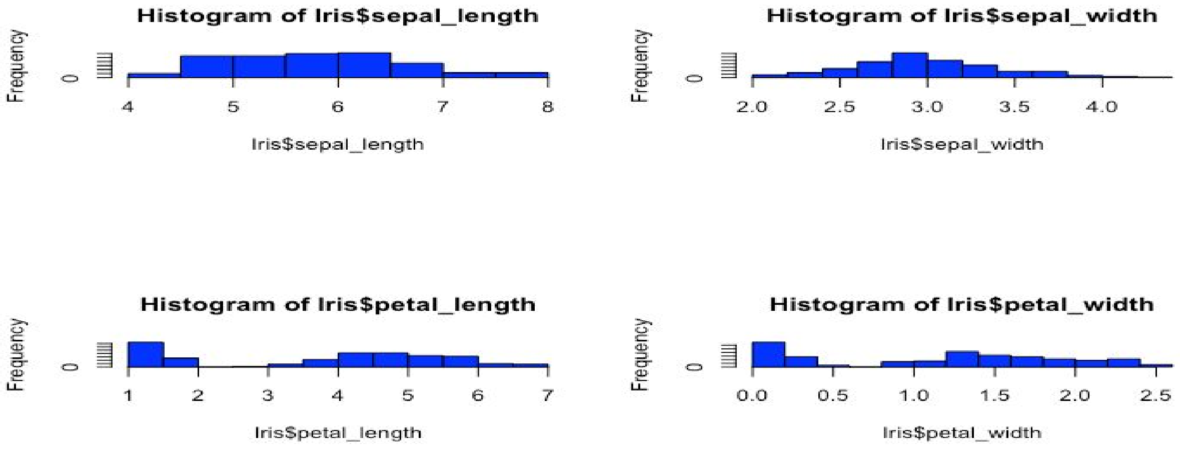


Fig 2.1

Furthermore the data does not contain any missing values or outliers that needed to be handled. Finally, the measure of Euclidean distance and Cosine similarity were chosen to be the two methods used to estimate the proximity measure of each object to all the other objects. In the case of Euclidean distance, similarity was calculated using the formula s = 1/(1+d), where ‘d’ is distance and ‘s’ is similarity.

Handling of Income Dataset

The Income dataset contains 16 features in total split into 5 ratio features, 2 ordinal features and 9 nominal features. It can be observed below (*fig 2.2*) that the column “ID” consists of unique values used to number the objects. As the feature contains no particular information about each object, the column was ignored in the similarity estimation process. The nominal feature of “class” was also removed from the calculation as per the project requirement. The third feature that was removed from the dataset was the ordinal feature “education”. It was earlier observed (*fig 1.3*) that education and education_cat were redundant categories and since education_cat provided a numerical ordering of the education levels, it was retained for ease of calculation.

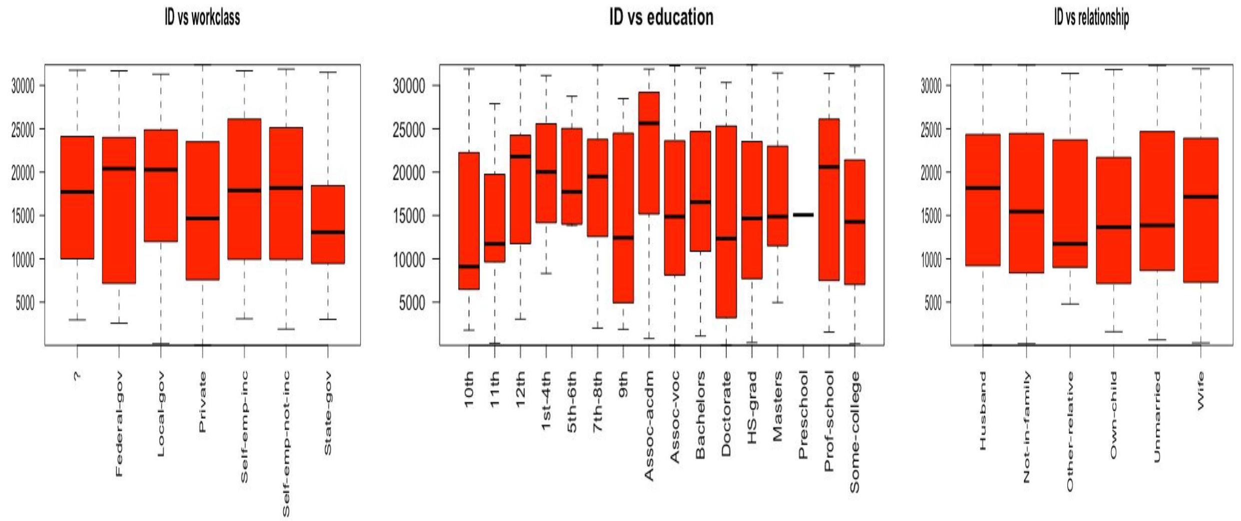


Fig 2.2

During the exploratory analysis, it was also detected that the histogram plot of capital gain and capital loss were highly right-skewed (*fig 1.6*) and possibly contained outliers. The data in capital gain majorly consisted of the value zero and few other values ranging from 10^3 to 10^5 which was not representative of the similarity amongst them. Therefore a log transformation was performed on the data to bring it to a scale that depicted the comparative measure better (*fig 2.3*). It can be seen that the few possible outliers (values of 99999) which looked like placeholders for much bigger values were also handled by the log transformation. Similar results can be viewed for the log transformation of the capital loss feature.

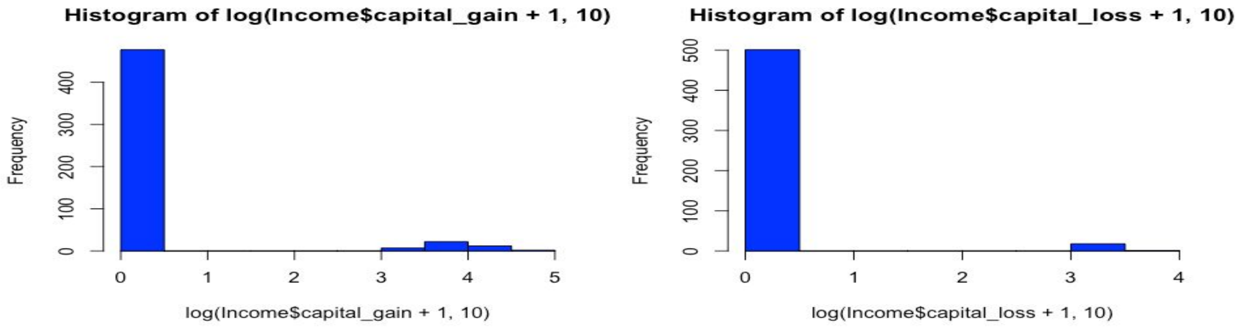


Fig 2.3

All the ratio attributes were then normalized using the min-max normalization technique to bring the values down to the same range (0-1), concluding the phase of data preprocessing. To calculate the similarity for the ordinal features, cosine similarity and Euclidean distance were chosen, with the formula s=1/(1+d) was used to calculate the similarity value from distance in the second scenario. The ordinal similarity was estimated using the formula s = 1 - (|p-q|/(n-1)), where p and q are values of the two objects and n is the total number of values. Finally the similarity measure used for nominal data was s=1 if values are the same else s=0. The final similarity value was obtained by getting the weighted sum of the three individual measures, where the weights were distributed according to the number features each category represented. As seen earlier (*fig 1.1, fig 1.4*) missing data was present in two nominal features. It's occurrence was handled by ignoring the attribute and appropriately changing the weights for those particular scenarios.

Analysis of the Result

This phase is aimed at analysing and making inferences from the result of the program. The four questions that it addresses are

A. Describe the distribution of proximities between each example and its (closest) nearest neighbor and how does this distribution change as *k* increases ?

B. Do any of these results differ when the proximity measures are changed?

C. Is there one example which is the closest to the largest number of other examples?

D. You did not use the ‘class’ attribute in the proximity function - but for each class, do you observe any differences for item A. above?

Proximity distribution and its variations (with k and measures)

This section answer question A and B specified above. The distribution of the proximities can be observed by plotting the proximity against their frequency (histogram plot). The distribution of the closest proximity (proximity 1) in the Iris dataset for both the similarity measures can be seen in the figure shown below (*fig 3.1*).

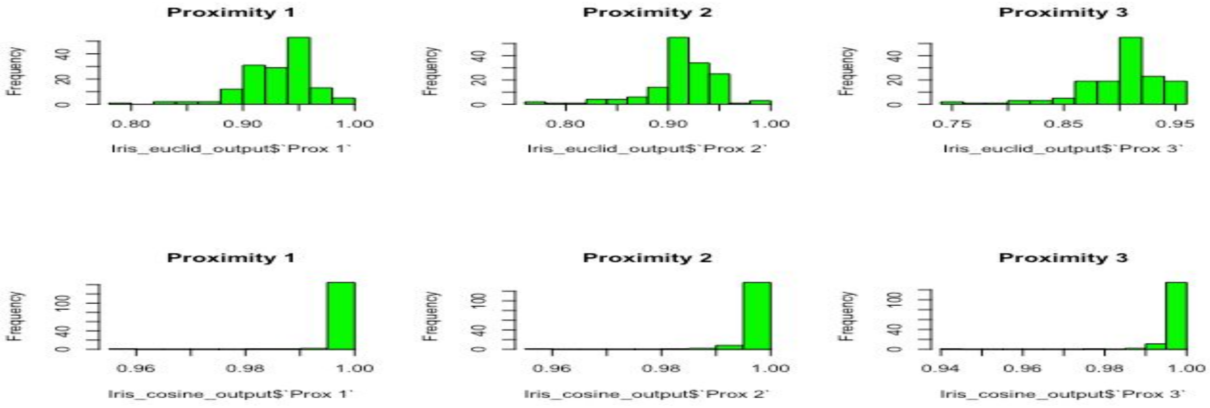


Fig 3.1

The closest proximity plot for euclidean distance measure is a normal distribution with peak at 0.95, gradual descent towards the left and a sharper decline at the right. As we look at the next closest data point (as k increases), the distribution is still normal but it can be observed that the peak decreases (0.90 for k=2 and 0.875 for k =3) and the distribution seems to even out a little.

Meanwhile, as the similarity measure changes to cosine similarity, the closest proximity plot is a left-skewed plot with most of the values greater than > 0.995 and rare occurrences of values between .985 and .995. As k increases, the plot is almost unchanged, with the peak remaining constant but the size of the tail slowly increases.

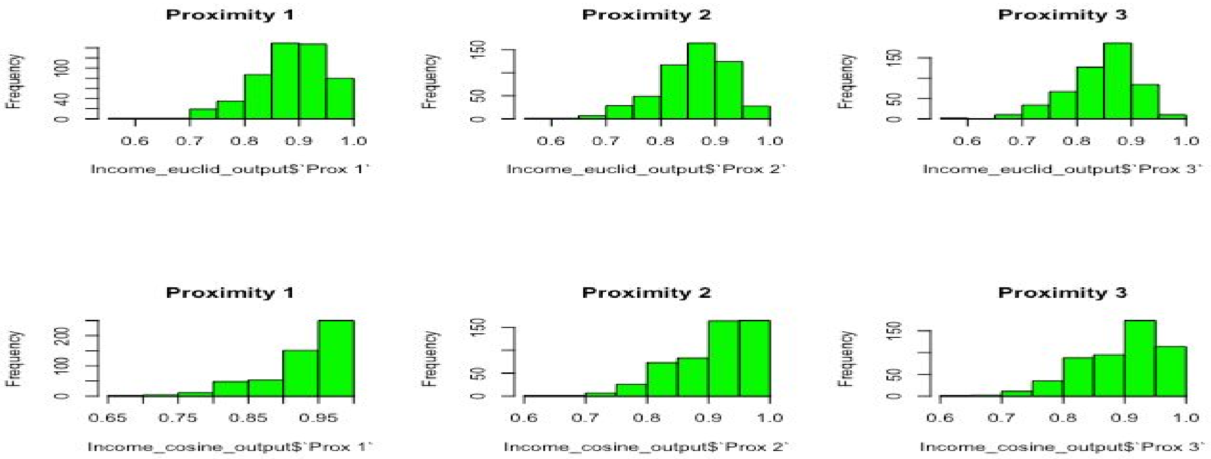


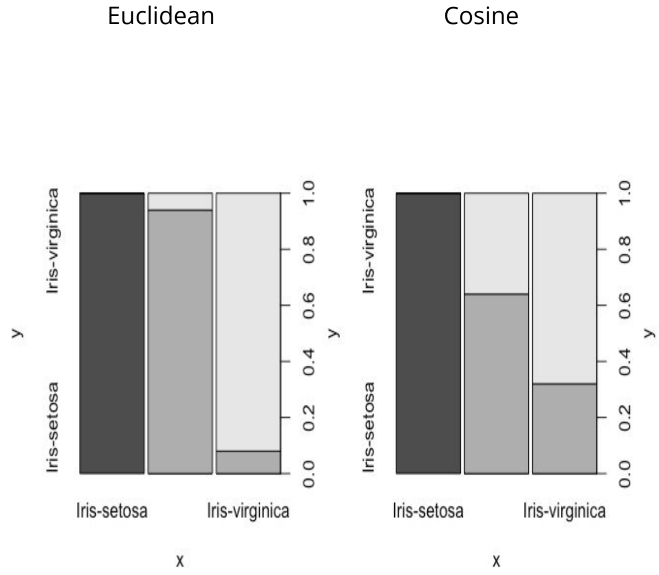
Fig 3.2

Similar plot for the income dataset can be seen in the figure (*fig 3.2*) above. The closest proximity for the euclidian measure has a normal distribution centered at 0.9 and symmetric descent on either side with a spread spanning 0.7 to 1. As k increases the peak remains almost unchanged while the plot starts becoming more left-skewed.

On changing to cosine similarity, it can be noticed that the distribution is again left-skewed (similar to iris dataset) peak between 0.95 and 1. Increase in the value of k causes the peak to shift leftwards and an increase in the size of the tail.

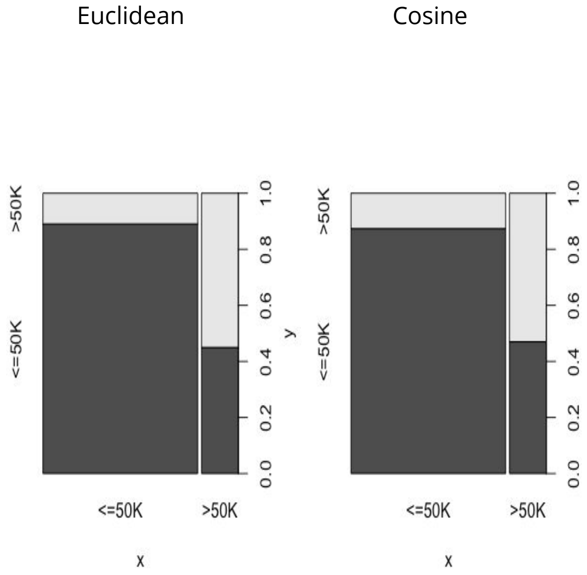
Distribution of current class vs closest example’s class

The distribution of the class of an data with the class of its closest example for Iris and Income data set can be seen below in *fig 3.3 and 3.4* respectively. Looking at the Iris dataset, it can be observed that in both the euclidean and the cosine estimation, Setosa’s closest match was always a Setosa. In the euclidean estimation, both Versicolor and Virginica mostly mapped to their own class with a little overlap, whereas the cosine estimation saw a lot of overlap. For the Income dataset, both the euclidean and the cosine distribution look the same, with about 50% of the people greater than 50K find their closest example in the other class.


Euclidean

Cosine

Fig 3.3


Euclidean

Cosine

Fig 3.4

The example which is closest to the largest number of example is

Iris and Euclidean = 47   Iris and Cosine = 35,42,45,81,143

Income and Euclidean = 146   Income and Cosine = 146