Implementing Cosine and Jaccard Similarities

by:

Darius Kharazi

September 13, 2017

The Ohio State University

1. Exploratory Analysis

The "class" variable seemed to be a good indicator of income under some analysis and assumption. There needed to be a large enough sample size in each class, ">50K" and "<=50K," in order to ensure any analysis could be conducted. Although only 100 entries existed with class equal to ">50K," and 420 entries existed with class equal to "<=50K," we should assume that the sample sizes for each group are large enough in order to perform further analysis. Most likely, the skewness is a result of the actual data and not the collection of data, since fewer people are wealthier than the given class.

After analysis of the distribution of classes given an individual's marital status, it seems that a smaller percentage of people from the ">50K" class have never been married; whereas the largest percentage of people from the "<=50K" class have never been married (Figure 1). This could indicate some potential interaction between the "marital status" variable and "class" variable. Additionally, the largest percentage of individuals from each class are part of the "private class," which seems to not indicate any interaction occurring between the two classes in the scope of an individual's work class (Figure 2). However, it seems that nearly 70% of individuals indicated that they are "private class." Therefore, it seems reasonable to divide individuals by "private class" and "other class," rather than including the numerous amounts of work classes in the analysis. Also, it seems like a higher percentage of individuals from the ">50K" class are "Prof-specialty" and "Exec-managerial", compared to the "<=50K" class. The higher percentage of individuals from the "<=50K" class are "Adm-clerical" and "Other-service" (Figure 3). This finding may be a consequence of the differing sample sizes between the classes, but could indicate an interaction between the variables. Furthermore, the highest percentage of people from class ">50K" consider their relationship status to be either "husband" or "wife;" whereas the highest percentage of people considered themselves "not-in-family" from the class "<=50K." Similar to other variables, this may indicate an interaction between relationship and class (Figure 4). Interestingly, the highest percentage of race from both classes is "white." However, nearly 83% of individuals indicated that they are white; and, therefore, it seems reasonable to divide individuals into either "white race" or "other race" for our analysis (Figure 5).

After analysis of the makeup of the dataset's native country variable, it seems that 90% of individuals are from the United States, which is extremely high and may indicate reason for variable manipulation. Also, nearly 90% of individuals indicated that they had a 0% capital gain, and nearly 95% of individuals indicated that they had a 0% capital loss. These large quantities may also indicate reason for variable manipulation. Lastly, it seems that there is a large division amongst individuals who work forty hours per week (or less) and more than 40 hours per

week. There also seems to be a few outliers for people who work close to 100 hours per week. For both of these reasons, it may be reasonable to perform analysis differentiating the two groups with some sort of variable transformation.

The dataset seems to include interesting conditional information, as well, which may be important to consider during further analysis. For example, the dataset contains nearly double the amount of men compared to women. Also, there is a smaller percentage of people married that have a college education, but less than a master's degree. However, there is a higher percentage of people married who have never been to college. The highest percentage of people married are those who have a master's degree or greater, but that could be a result of a small sample size. Also, nearly 43% of people who are white are married in the dataset; whereas nearly 30% of people who aren't white are married in the dataset. This could also be a result of the smaller sample size of people who are not white. Additionally, nearly 37% of people who are white have a business-related occupation, and 31% of people who are not white have a business-related occupation. Again, the slight difference could be a result of the smaller sample size of people who are not white. Lastly, 45% of both people whose native country is the United States and people whose native country is not the United States have less than a college-level education. Furthermore, nearly 7% of people whose native country is the United States have a master's degree or further education; whereas nearly 16% of people whose native country is not the United States have a master's degree or further education. Therefore, nearly 47% of people from the United States have some sort of college education, but less education than a master's degree; and only around 38% of people who are not from the United States have some sort of college education, but less education than a master's degree.

2. Description of the Program

After exploratory analysis, data preprocessing needed to be performed. In the early stages of program implementation, it made sense to perform variable manipulation, primarily due to the reasons mentioned in the exploratory analysis. Since certain categorical variables contained such high percentages of the total makeup at certain levels, it seemed reasonable to transform these variables. Also, variable transformation seemed reasonable for variables that contained quite a few outliers, such as the "fnlwgt" and "hours per week" variables, perhaps to understand the outliers more and capture this information. Lastly, since so many data entries had missing values, data deletion was not much of an option. Manipulating and creating a new "other" level for certain variables, such as "occupation," seemed like a reasonable approach for handling missing values. Because of the huge quantity of variables that needed their own variables for distinct levels, the variable manipulation process engendered binary data.

After data preprocessing, cosine and jaccard similarity matrices and cosine and jaccard output matrices are created for two separate purposes. The cosine and jaccard similarity matrices are two symmetrical matrices with the diagonal values equal to 1. Their purpose is to demonstrate the cosine and jaccard similarity values for each data entry in comparison with other data entries. After these similarity matrices are calculated, the output matrices poll the corresponding data and identify the k-most similar data entries in relation to every other data entry. The parameter "k" is adjustable, and determines the limit of indices and proximities that should be included in the output matrix. Each proximity that is included in the output matrix provides its corresponding index from the income dataset, as well.

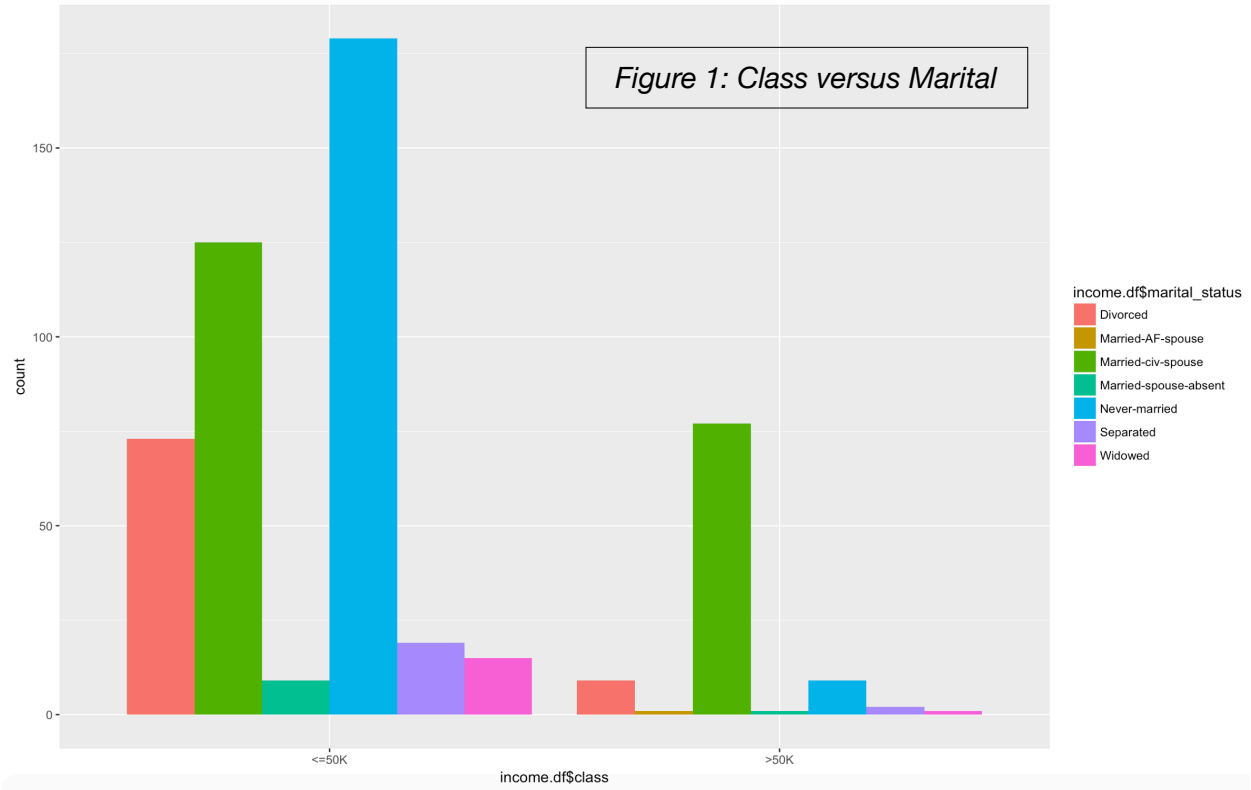
Since each variable contained binary data after the variable transformation process, it seemed reasonable to calculate cosine and jaccard proximities for each data entry. Cosine similarity seemed reasonable for analysis of each data entry's vector in the cosine similarity matrix. Additionally, calculating jaccard similarities seemed reasonable given our transformed binary data, since jaccard similarities should only be calculated on binary data.

3. Analysis of the Results

The distribution of proximities between each example and its first nearest neighbor follows a similar shape. However, the data will be distributed among different proximity values, since two different proximity functions were used in the program. When k=5, the application will be gathering proximity values from the tail of the distribution containing the total proximity values for each entry. As "k" increases, the application will be gathering proximity values from the area closest to the mean. Essentially, the total proximity values for each entry follows a normal distribution, and the distribution of the gathered proximity values will begin to follow this normal distribution as "k" increases. Additionally, the class attribute seems to follow a similar pattern, although it was not used in the proximity function.

After running the program, we can see that a few data entries are very similar to each other, according to both the cosine and jaccard proximity functions. Some data entries are perfectly similar to other data entries, since continuous variables were manipulated to represent variables of categorical type after data preprocessing. For example, the entry with ID=9364 is perfectly similar to the entry with ID=12510, according to the Jaccard proximity function (Figure 7). Although a few of the variables from the original dataset are not exactly equal, the overall proximity of these values prove the effectiveness of the approach taken to transform variables and group certain values together during data preprocessing.

4. References



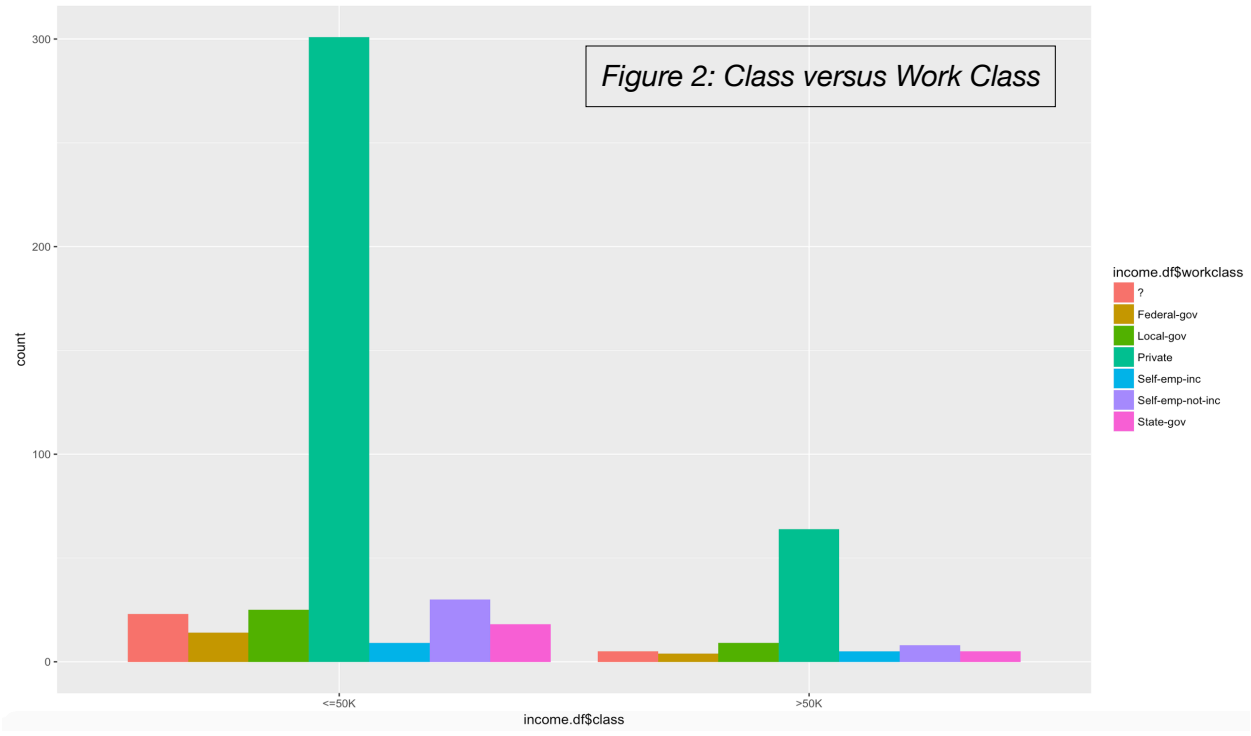


Figure 3: Class versus Occupation

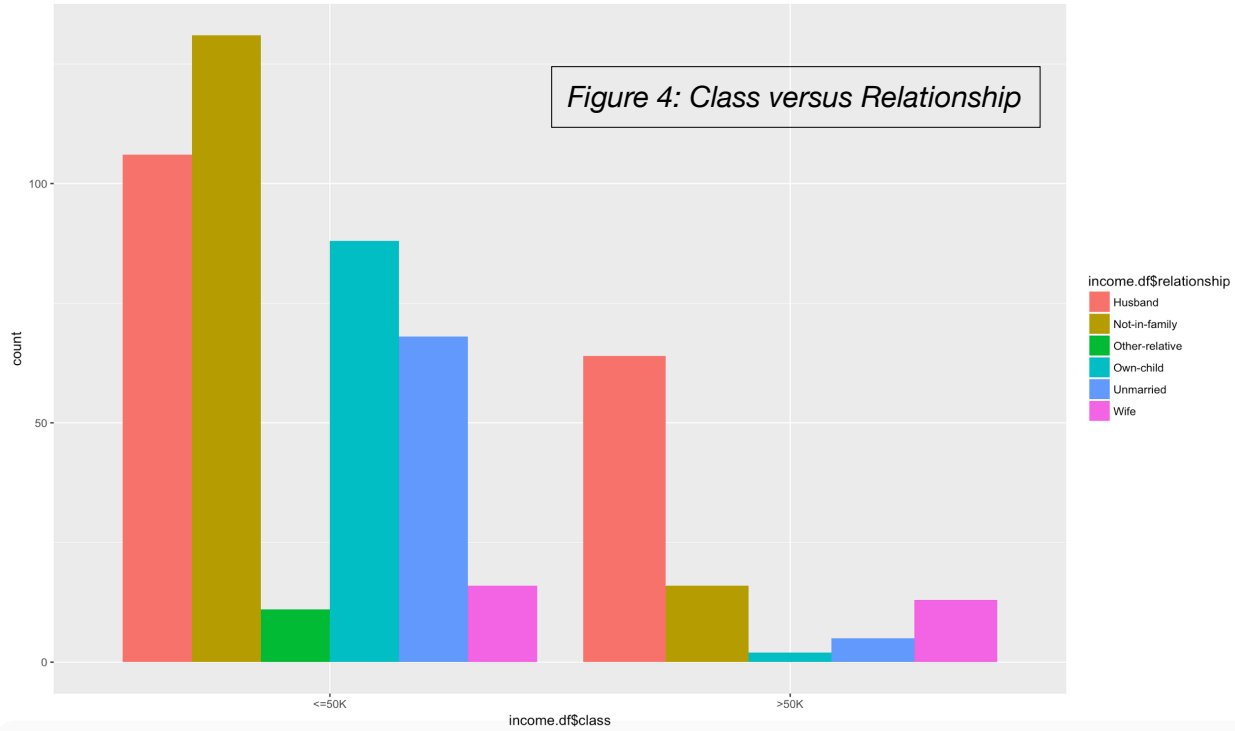
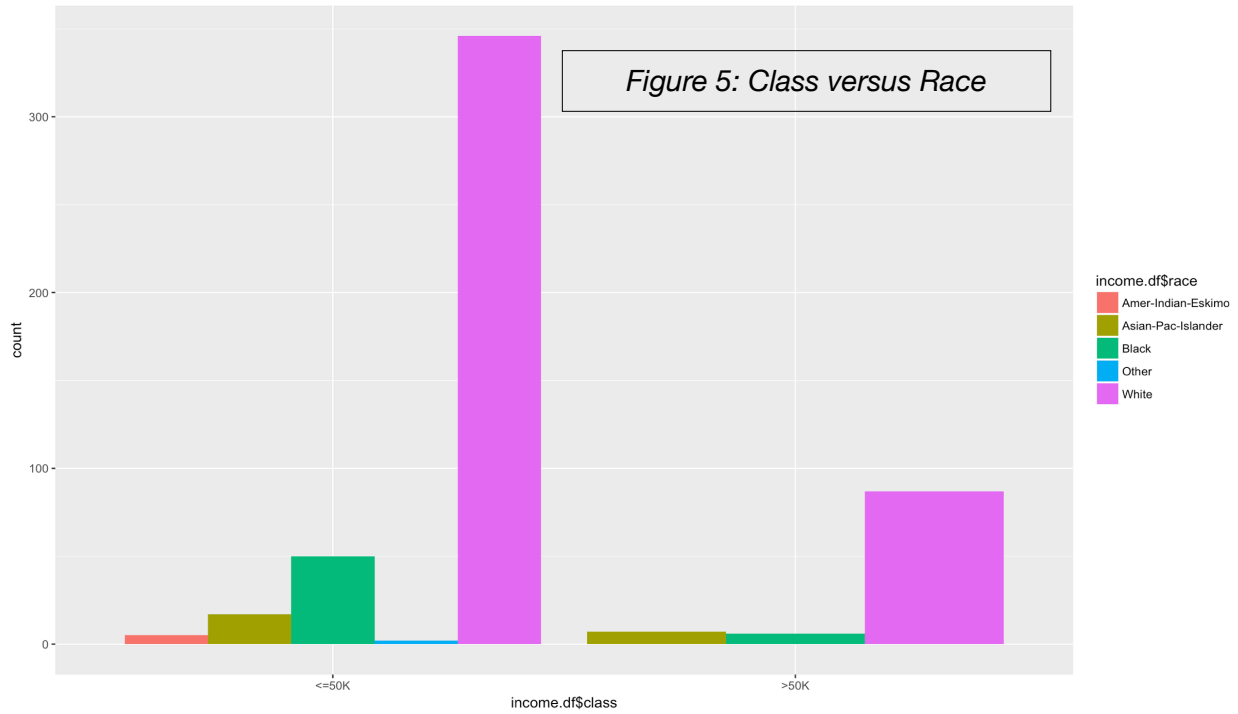


Figure 4: Class versus Relationship



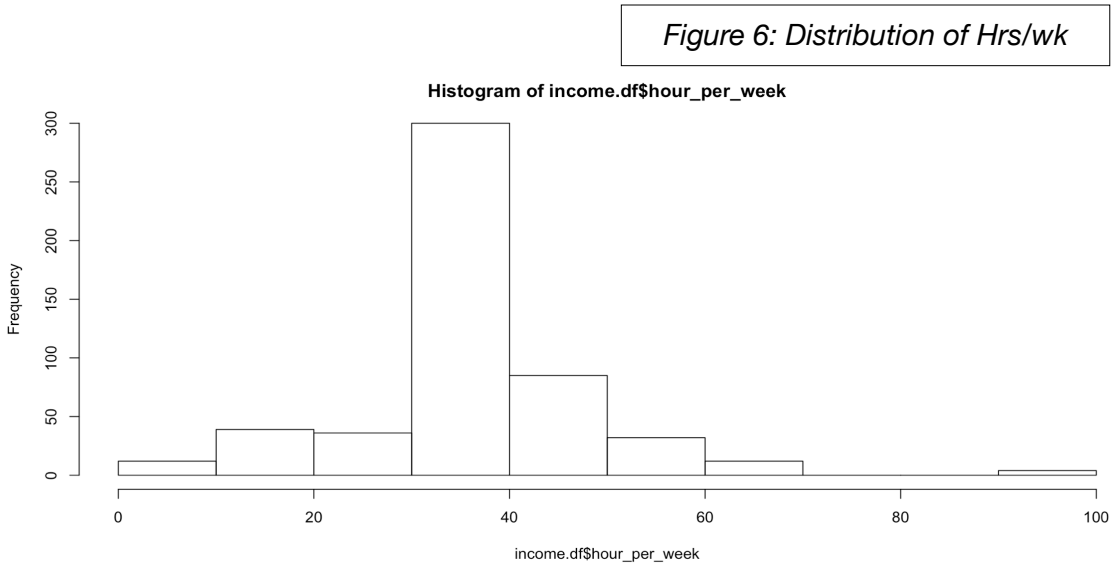


Figure 7: Proximity Comparison

	ID	age	workclass	fnlwgt	education	education_cat	marital_status	occupation	relationship	race	gender	capital_gain	capital_loss	hour_per_week	native_country
1	9364	38	Private	197077	HS-grad	9	Married-civ-spouse	Other-service	Husband	White	Male	0	0	40	United-States
348	12510	36	Private	193855	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States