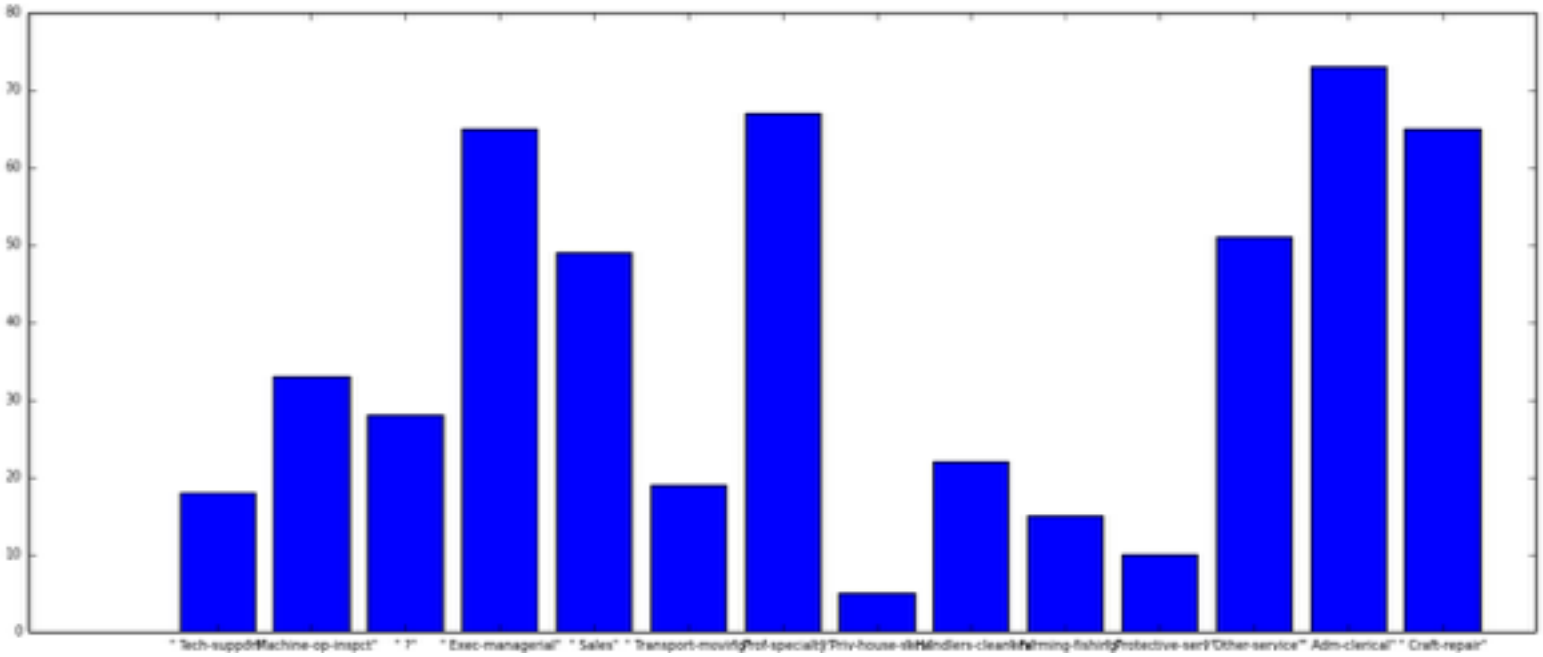
Sam Farren: Lab 1 Report - Income Dataset

Section 1:

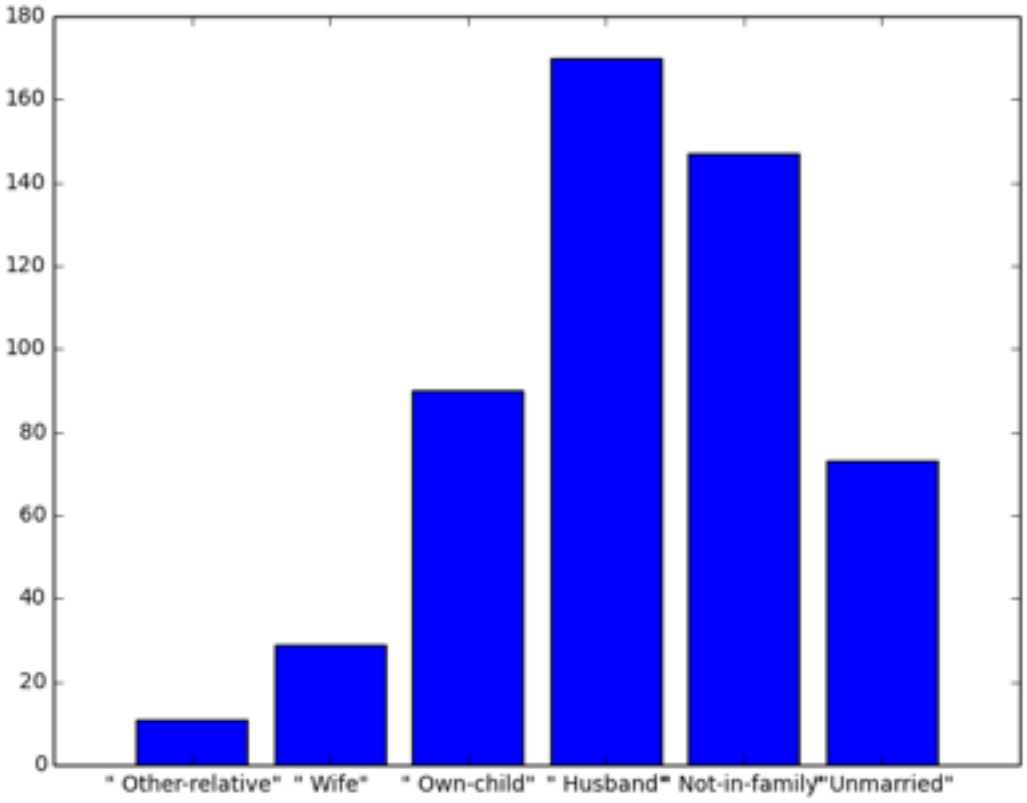
After first glance at the income dataset, there were so many attributes to take into account that I tried to see if any didn't provide much value to the distinguishing characteristics of each entry. Finding such attributes could reduce the dimensionality of the data and make it easier to analyze. The attributes that were reviewed were as follows:

> "ID","age","workclass","fnlwgt","education","education_cat","marital_status","occupation", "relationship","race","gender","capital_gain","capital_loss","hour_per_week", "native_country","class"

I ended up going through each attribute and deciding if it was worth keeping for the proximity calculations and data exploration/visualization step. The first attribute was ID and was solely used for distinguishing each entry from the others, so it was kept but not used in calculations such as the proximity measure as it didn't provide any meaningful data. I later found a dependency among the data for two of the attributes: education and education_cat. Education_cat was dependent on education for its value, and was already a quantitative value that could be used in calculations. Therefore, I decided to discard education and just use education_cat as this would reduce the dimensionality of the data by 1. Another attribute that was dropped from the set was "fnlwgt", because the description stated that they were arbitrary values. Random values don't give any significant meaning to data, so dropping it was justified to again reduce the dimensionality by 1. One of the attributes that posed a hard decision was the work class attribute. This was because there was missing data for it and most had a value of "Private". Specifically 71% of the data had "Private" and 6% of the data for this attribute was missing entirely. With such a non-normal distribution and a good amount of missing data, I made the decision to not include work class in the proximity function and also found it not valuable to compare amongst other attributes to check for correlation and other relationships. Another tough decision needed to be made with occupation. It seemed to be pretty sparse, so I decided to analyze the counts of each job. The following graph was produced from a python dictionary of the counts of each occupation:



Notable among this distribution is that “Exec-managerial”, “Prof-specialty”, “Adm-clerical”, “and Craft-repair” accounted for half of the total number of occupations. The problem with this is that there are 14 different jobs listed, and in order to do any numerical computations would require a transformation of these jobs into bins which would produce a nominal representation of this set of data, with each occupation being represented by a discrete number 1-14. With so many subcategories and ambiguous, non-continuous data, it was a difficult problem to account for without skewing the data and losing information. One other major decision that was made was with the relationship attribute. I didn’t know how many subcategories there were of this, because there are many types of relationships between people. Since these questions were raised I went and obtained more specific data than that of what just my eyes could see. The following graph is the counts of each type of relationship among the data:



(Figure 2: Relationship occurrence Histogram)

While this was a normal distribution based off the graph depicted above, these are categorical subcategories and can be arranged into any order to make the graph look any way one wants to. I decided just to include it and go with a binary check, testing if two objects had different relationship values then the dissimilarity was increased by 1.

My first analysis I did came from an initial thought about if higher education was related to the pay of a person. When I had this thought I went into the records and noticed about 1 out of

every 2 people with a bachelors degree made above 50,000 while most with less than that didn't make above 50K. To get a more concrete answer I took a simple ratio of (# with bachelors above 50K / Total number with Bachelors). This calculation gave a percentage of 44.19%. The other calculation was a ratio of (# of people with less than a bachelors that make above 50K / Total number of people with less than a bachelors) and produced a value of 11%. This value proved to be a very significant difference and confirmed my original hypothesis that those who have a bachelors degree (or higher) are more likely to make above 50,000 dollars a year.

Section 2: A description of your program, including discussions on design choices made. For example, how did you choose to handle missing values or outliers for these datasets? Did you transform any of the attributes? For the income dataset, you should justify your choices based on the results of the exploratory analysis above.

Design choices in my analysis were briefly mentioned above. Outliers were kept in the data but dealt with by transforming continuous data into discrete data represented by bins. This is discussed more in depth below for each attribute. In the case of missing data, I opted to not try and extrapolate values for it. If data was missing, they were simply ignored in the data set.

One of the decisions I had to make early on was what data was going to be used for the proximity measures and what and how I should transform the data for the proximity measure. One of the big things when transforming the data was that a lot of it was sporadic and not normal, so for many attributes I felt the need to transform into discrete ranges. For example, capital_gain ranged from 0 to 99999, so the range 1-10000 got a value of 2, and range 10000-50000 got a value of 3 etc. However, attributes such as gender, race, marital_status, relationship, occupation, native_country, and workclass, were all categorical data. It was data that the only thing you could really do was compare and see if the value was the same between two objects. Two different values for and object didn't hold much value. Therefore a simple binary comparison was justified for proximity calculations and other relationships. All in all, the following transformations were made with the justifications for why below:

Education_cat: Education_cat was reduced from around 10 discrete values to 5, the bins being discrete values numbered 1-5. This scale was also continuous and the order of the numbers held meaning, with 1 having a lower education and 5 having the highest education. A 1 was mapped to all individuals that had an education at or less than a high school level. A 2 was for those who had an associates degree or some college experience. A 3 was a bachelors degree, a 4 was a masters degree, and anything above a masters(doctorate or specific grad school study) was mapped to a 5. I felt justified in doing this because the original scale was a distorted and didn't represent the value of the educations correctly.

Marital_status: Marital_status was also simplified. There were multiple groups however I chose to reduce it and generalize it to only three categories. These were 3 for married-civ-spouse, 2 for Divorced, and 1 for anything else. The proximity calculation was then based upon a simple nominal comparison between each object.

Race: There were four specific categories for race with a fifth being "other". Therefore the transformation could just map the five categories to discrete integers numbered 1-5. These were White -> 1, Black-> 2, Asian-Pac-Islander -> 3, Amer-Ind-Eskimo -> 4, and Other -> 5. Again when computing similarity for proximity a nominal comparison between objects was used.

Gender: With only two choices for Gender, It made sense to transform it into a binary attribute with Male being a 0, and Female being a 1. This way when comparing for similarity one can simply use a binary comparison.

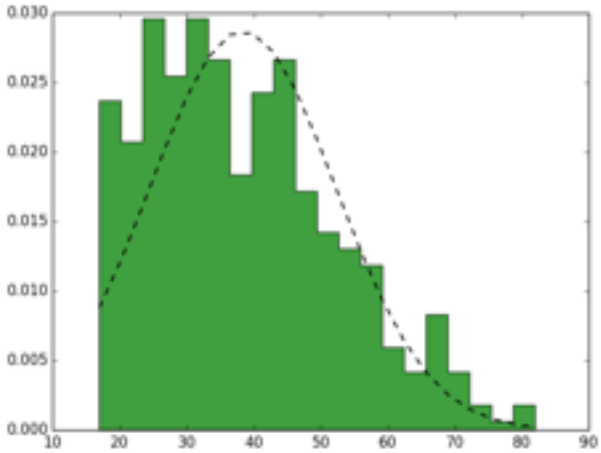
Native_Country: Native Country was a tougher decision since there were so many different international countries. I simplified it by comparing only those that were in the united states and those that were International. This was a simple binary mapping of those from the US to a 0, and those that were international to a 1. The comparison for Native Country then used a binary comparison.

Capital_Gain: Capital gain ended up having a couple of outliers. A lot of the values were 0, but there was one that was 99999 which skewed the mean. Since there was a high concentration of 0's I split up and placed different ranges into bins to try and normalize it a little bit. These transformations were: 0 ->1, (<=10000) -> 2, (<=50000) -> 3, and (>50000) -> 4. Nominal comparisons for similarity were then used.

Capital_Loss: Capital_loss was similar to Capital_gain in that there were a lot of zeroes with a few outliers. The same technique was used where ranges were placed into bins. These were: 0 -> 1, (<=2000) -> 2, (<=4000) -> 3, (>=4000 -> 4). The same comparison was then used with a nominal similarity.

Hours_Per_Week: Hours per week was also transformed into similar bins. There were a couple outliers where people worked upwards of 70 and 80 hours, but most of the sample fell around 40 hours. The transformation used was: (<=30) -> 1, (<=40) -> 2, (<=50) -> 3, (>=50) -> 4. Again, a nominal comparison was used for the similarity of how many hours each person worked per week.

One that was not in the above list was Age. I was considering mapping age to bins to simplify the data since it was a continuous ratio attribute. However, I felt there would have been a loss of data if that was done. The following histogram was created from the age data:



(Figure 3: Age Occurence with Expected Distribution)

The graph showed me the distribution, and while it was skewed to the left, I felt that moving it into bins would result in a high loss of information. Therefore, I decided to keep the attribute as is.

Section 3: Analysis of the results. Some examples of analysis you might conduct (feel free to add other ideas):

A. How do you describe the distribution of proximities between each example and its first nearest neighbor? How does this distribution change as k increases?
B. You did not use the class attribute in the proximity function - but for each class, do you observe any differences for part A above?
C. Is there one example which is the closest to the largest number of other examples?
D. Do any of these results differ when you change the proximity measure?

After preprocessing, and many transformations of the data, the proximities for the income dataset finally were output. The first similarity function implemented was a form of jacquard coefficient similarity. In each iteration of the computation, the top five similarities for each id were found and sorted in descending order. Higher numbers meant closer similarity in this computation. The following equation was used for the first similarity calculation:

similarity = (gender) + (relationship) + (country) + (race) + (education_cat) + (marital) + (capital_gain) + (capital_loss) + (hours) + (float(age)/100)

All except for age, were simple nominal comparisons between the data. If a rows attribute was equal to another rows attribute, then the similarity was incremented by 1, if it wasn't then the similarity didn't change. For example, if two rows being compared were both female, then the value of gender would be set to 1. If one was male and one was female, the value of gender would be set to 0. Age was the only one not checked nominally. A difference was taken then divided by 10 to keep the value below 1 and above 0. The top similarities had values around 5, meaning that around 5 of the attributes had the same value. Nothing was weighted any higher than the other attributes, however, age differences were divided by 100 so it was weighted a little less than other attributes.

When changing the proximity measure my results did change. The id's were different when between the first similarity test and the cosine similarity. For example, id=25668 with k=4, had different similar id's. This can clearly be seen below.

```
Record 1
Compared ID: "25668"
id:"23382"  1-prox: 5.94
id:"29286"  2-prox: 5.78
id:"17822"  3-prox: 5.69
id:"22934"  4-prox: 5.69
```

```
Compared ID: "25668"
id:"31146"  1-prox: 1.0
id:"26324"  2-prox: 1.0
id:"3039"  3-prox: 1.0
id:"32378"  4-prox: 1.0
```

```
Record 2
Compared ID: "13316"
id:"26111"  1-prox: 5.13
id:"14594"  2-prox: 5.11
id:"9983"  3-prox: 5.05
id:"1542"  4-prox: 4.99
```

```
Compared ID: "13316"
id:"30619"  1-prox: 1.0
id:"8009"  2-prox: 1.0
id:"10672"  3-prox: 1.0
id:"32352"  4-prox: 1.0
```

In conclusion, this lab introduced the class in a very practical manner. It can be a very daunting task to reduce the dimensionality of data in order to analyze it without losing valuable information. Choices need to be justified and greatly thought out or analysis can be quickly discredited by someone who sees faults.