Project Stage 1. Enter your group information and do data understanding and cleaning.

Started on Fri Feb 12. Pls submit by midnight Wed Feb 24 by posting your report on your group's homepage. Please do not post earlier than 10 pm of Wed Feb 24, to prevent "cross pollination" of ideas.

Enter your group information

Please go to this Google doc to enter your group's information.

Data understanding and cleaning

Step 1. Download data from the link http://pages.cs.wisc.edu/~sanjibkd/784-stage1/
The data contains two files:

(a) elec_pairs_stage1.txt that has 20K labeled product pairs, and
(b) readme.txt that describes how the data looks like.

Step 2. You will submit a report (it can be in pdf or google doc format). The report should list the following items:

+ The names and emails of members in your group.

+ The list of all attributes that you find in the data. For example, suppose your data has only two tuple pairs (a,b) and (c,d), where a, b, c, d are products. Suppose

a has attributes x and y,
b has attributes x and z,
c has attributes x, y, u, and
d has attributes z, v.

Then you should list x, y, u, z, v. For each attribute, list the number of products it appears in. You should list the attributes in decreasing value of this number. This will help you understand the attributes in the data, and how common an attribute is.

+ For each attribute A of the top 10 most frequent attributes, discuss the following:
- missing values: report the percentage of missing values for A. For example, if there are 20 products but A has value in only 5 of them, then the percentage of missing values is 75%. (Report both the fraction and the percentage.)

- you often have to fill in these missing values somehow for machine learning in subsequent steps. Discuss solutions you may use to fill in these missing values (you don't have to fill in these values; I'm only asking for a discussion of possible solutions).

- classify the attribute as numeric, textual, categorical, or boolean. If you can't classify, discuss why (e.g., an attribute has values 1, 3, and medium, so it's neither numeric nor categorical).

- if the attribute is textual, report the average length of its values, report the minimal and the maximal length of its values (length is measured in the number of characters in the value).

- find and report possible outliers and anomalies among the attribute values. For example, if attribute price typically has values in the range $1-20, and then there is a value of $200, then this value is an outlier, and it can also be an anomaly (that is, an incorrect value in this case). You can detect outliers by creating a histogram on some property of the attribute values. For example, a histogram on price values will help detect price outliers. As another example, if an attribute is textual, then a histogram on the length of the values (as measured by the number of characters in the value) can help detect values that are very long or very short. Show at least two histograms that your team has created.

- if the attribute value is supposed to follow a certain format (e.g., dates), then discuss if all values follow the same format, or if there is some problem with the format and we will have to standardize the formats later.

- are there synonyms among attribute values? For example, an attribute "book-type" may have values "soft cover" and "paperback", which are synonyms.

- sometimes attribute values are "sprinkled" all over the item. For example, a book may have an attribute "publisher", but its value is missing. Instead, the book title contains the publisher (e.g., "Principles of Data Integration by Springer"). Do you have this problem with this attribute?

- do you see any other data quality problems with this attribute? You may want to read this paper, which is a bit outdated but can give you ideas about data quality problems. Just ignore stuff you do not yet know when you read the paper.

+ List any software tools that you have used to do the above data understanding and cleaning. For example, if you have used a particular Python package, pls list the name of the package.

+ Bonus points (you don't have to do this to receive the full credit): Do you find any label (MATCH, MISMATCH) that is incorrect? Is there any quick and dirty way for you to find such incorrect labels? (Note that these labels say whether the two products match or not.)
