Assignment #1

Description: Along with this description is a file named – student-mat.csv. It is a comma delimited file of Mathematics test scores for students. It is a well known data set from the UCI Data Repository. You can open it in a spreadsheet and see all of the values about the students in the experiment.

I have this theory that this data set does not come from a good sample. There are two attributes that I believe are not good samples representing the general population.

- Family size (column 5) has two values GT3 (greater than 3) and LE3 (Less than or equal to 3). I know from census data that 25% of all family have three or less family members.
- Romantic Interest (ie is the student in a relationship) (column 23) has values yes and no. Other data that I have tells me that 30% of high school students are in relationships.

Your assignment is as follows:

1. Create histograms graph using ThinkStats2 package for the family size and romantic attributes.
2. Create a short 1-page (and 1 page only) report on your analysis to my theory about the data set. The two plots should be included in the 1-page.

Hint: In order to get this to work, you'll need to get a columns worth of data into a sequence. Here is code to show you how to do that. Put your Python code and the data file in the same folder.

This will read the file and create the sequence (call dataList). In the last line there is a [4]. That is the column number being put into the sequence. But Python starts counting at 0, so 4 is the 5th column (ie Family Size).

```
dataList = list()
F = open("student-mat.csv", "r")
for line in F:
    spLine = line.split(",")
    dataList.append(spLine[4])
```

Rubric:

1. Grammar and Spelling 20%
2. Correct Histograms 40%
3. Analysis 40%

Due: The end of week 4.