Create a text document listing possible questions your data set can answer. Include the following components in your write-up:

1. Brief description of your data set (age of the data, number of observations and attributes)
2. At least 1 yes-no or binary-choice question (Does it...? Is it...?)
3. At least non-binary question (What is...? How many...? When does....?)
4. Future improvements to the data set: What other related data could be collected?
5. Source citation for your data set

The data set I am using is the Heart Disease dataset from the UCI Machine Learning dataset. The file I processed was part of Hungarian data. There were originally 294 rows (data from 294 individuals) with 76 attributes gathered. These were processed (by the publishers of the data) down to 14 attributes for 294 individuals. There are also similar files from the Long Beach CA VA facility (200 individuals, 14 attributes), Cleveland (303 individuals with 14 attributes) and Switzerland (123 individuals with 14 attributes). One of the recorded attributes is the presence of heart disease. That can be used as part of the machine learning for training.

After my processing, two of the attributes were dropped (ca – number of vessels colored by fluoroscopy, and “thal”) since these were mostly missing data (more than 80%). My processed data set had

The binary question I would like to ask is, “does the data predict heart disease in a useable way?” A non-binary question would be” what are the important attributes needed to predict heart disease?” Also, “when does gender make a difference in the presence of heart disease?”

1. Hi -
2. I used the Heart Disease data set. I kept the "num" column unchanged. I changed the "sex" column to categories of "Male" and "Female". Then I dropped the "Male" category since it could be reconstructed from the opposite of the "Female" category.
3. I took the "fbs" column, imputed the missing data, then changed the "abs" column to categories of "above 120 mg/dl" and "below 120 mg/dl". I dropped the "below 120 mg/dl" category column since it can be reconstructed from the opposite of the "above 120 mg/dl" column.
4. I took the “resting” column, imputed missing data, then created columns of “normal”, “ST-T abnormal”, and “left v htrophy”. I dropped the “normal” category column since that can be recreated from the other two.
5. I took the “exang” column, imputed missing data, then created columns of “xrcise induced” and “not exercise induced”. I dropped the “not exercise induced” category column since that could be recreated from the other.
6. I took the “slope” column, imputed missing values, and created categories of “upslope”, “flat”, and “downslope”. I dropped the “downslope” category column since it can be recreated for the other two.
7. I took the “cp” column, imputed missing data, then combined “typical angina” and “atypical angina” into a category of “angina”. I also made categories of “non-angina” and “asymptomatic”. I dropped the “asymptomatic” category column since it can be recreated from the other two.
8. I took the “age” data, imputed missing values, and then binned it into categories of “28 to 34”, “35 to 42”, “43 to 50”, “51 to 58” and “59 to 66”. It seemed easier to keep all categories rather than deleting one and building it up from a combination of the other 4.  I took the “oldpeak” data, imputed missing values, replaced outliers with the mean, and min-max normed it. The data was not normally distributed, so a min-max normalization seemed better.
9. I did the same things with the “restbps”, “chol” and “thatch” columns: imputed missing values, replaced outliers with the mean, and min-max normalized the data.
10. I dropped the “ca” and “thal” columns since they were predominantly missing data values.