individual records in the file, after missing data has been excluded.

Read the data and remove all records with missing data.

```
DF=pd.read_csv(
"https://archive.ics.uci.edu/ml/machine-learning-databases/
    heart-disease/processed.cleveland.data",
    header=None,na_values="?")
DF=DF.dropna(axis=0)
```

For easy reference, we'll label the columns based on the names given by the author of the data set and then make a list of the ones we want to mark as categorical.

```
DF.columns=["age","sex","cp","trestbps","chol","fbs","restecg",
    "thalach","exang","oldpeak","slope","ca","thal","num"]
categorical=["sex","cp","fbs", "restecg","exang","slope",
    "thal"]
```

Here's what the first few lines of the data look like. The data wraps around because of the page width.

```
print(DF[:3])
```

```
    age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  slope
ca  thal  num
0  63.0  1.0  1.0     145.0  233.0  1.0      2.0    150.0    0.0      2.3    3.0
0.0   6.0    0
1  67.0  1.0  4.0     160.0  286.0  0.0      2.0    108.0    1.0      1.5    2.0
3.0   3.0    2
2  67.0  1.0  4.0     120.0  229.0  0.0      2.0    129.0    1.0      2.6    2.0
2.0   7.0    1
```

The **Y** variable is **"num"**, set to 1 if its value is nonzero, and 0 otherwise.

```
M=np.array(DF["num"])
Y=np.array([1 if x>0.5 else 0 for x in num])
```

The feature matrix **X** is built by dropping the column labeled **"num"** and keeping the rest of the data. Categorical features are encoded using a one-hot encoding. The easiest way to do this is with the function **get_dummies**.

```
X=np.array(pd.get_dummies(DF.drop("num",1),
                          columns=categorical))
print(X.shape)
X
```

```
(297,25)
array([[ 63., 145., 233., ...,   0.,   1.,   0.],
       [ 67., 160., 286., ...,   1.,   0.,   0.],
       [ 67., 120., 229., ...,   0.,   0.,   1.],
```