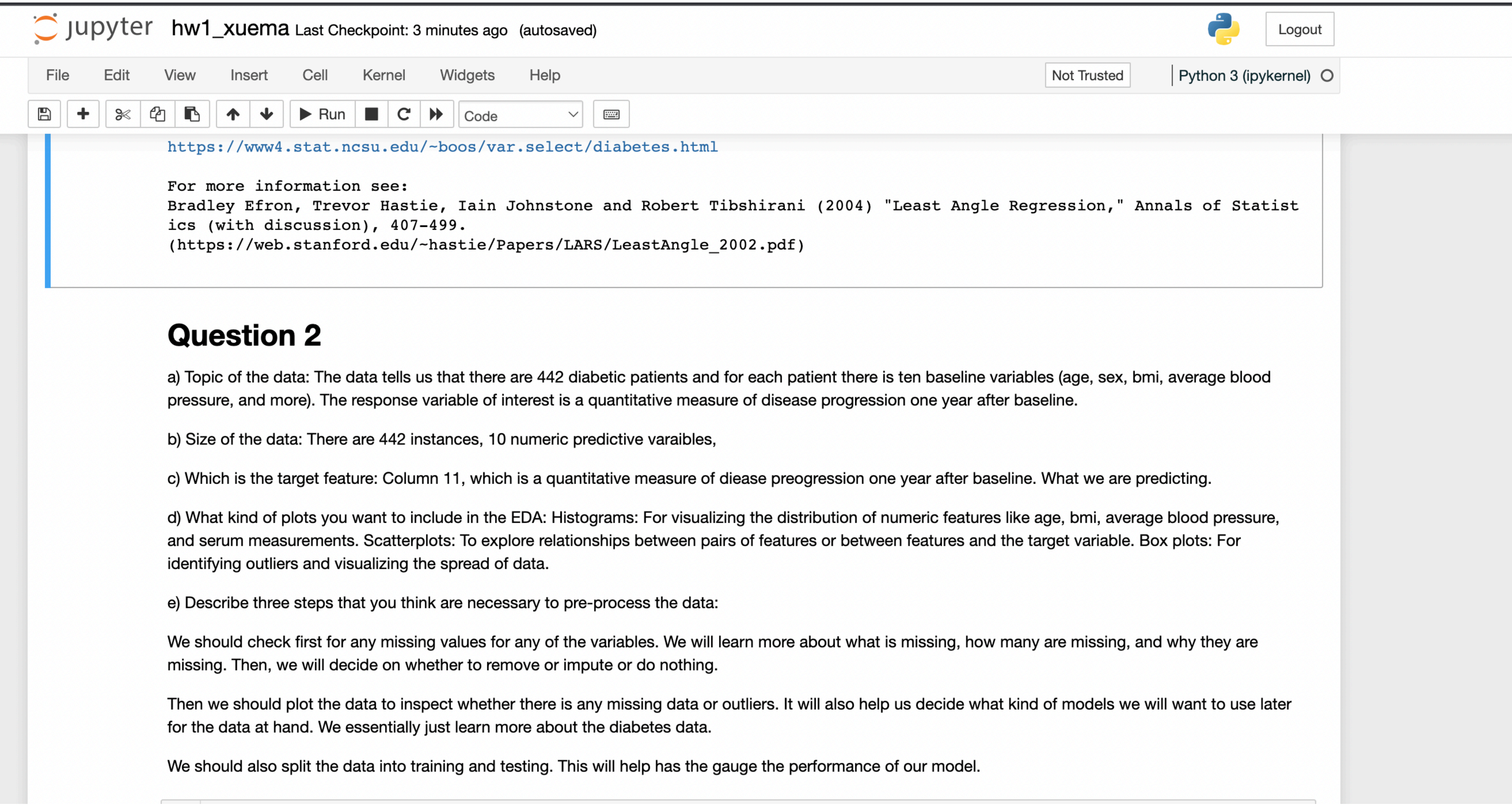

```
https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statist
ics (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)
```

Question 2

a) Topic of the data: The data tells us that there are 442 diabetic patients and for each patient there is ten baseline variables (age, sex, bmi, average blood pressure, and more). The response variable of interest is a quantitative measure of disease progression one year after baseline.

b) Size of the data: There are 442 instances, 10 numeric predictive varaibles,

c) Which is the target feature: Column 11, which is a quantitative measure of diease preogression one year after baseline. What we are predicting.

d) What kind of plots you want to include in the EDA: Histograms: For visualizing the distribution of numeric features like age, bmi, average blood pressure, and serum measurements. Scatterplots: To explore relationships between pairs of features or between features and the target variable. Box plots: For identifying outliers and visualizing the spread of data.

e) Describe three steps that you think are necessary to pre-process the data:

We should check first for any missing values for any of the variables. We will learn more about what is missing, how many are missing, and why they are missing. Then, we will decide on whether to remove or impute or do nothing.

Then we should plot the data to inspect whether there is any missing data or outliers. It will also help us decide what kind of models we will want to use later for the data at hand. We essentially just learn more about the diabetes data.

We should also split the data into training and testing. This will help has the gauge the performance of our model.