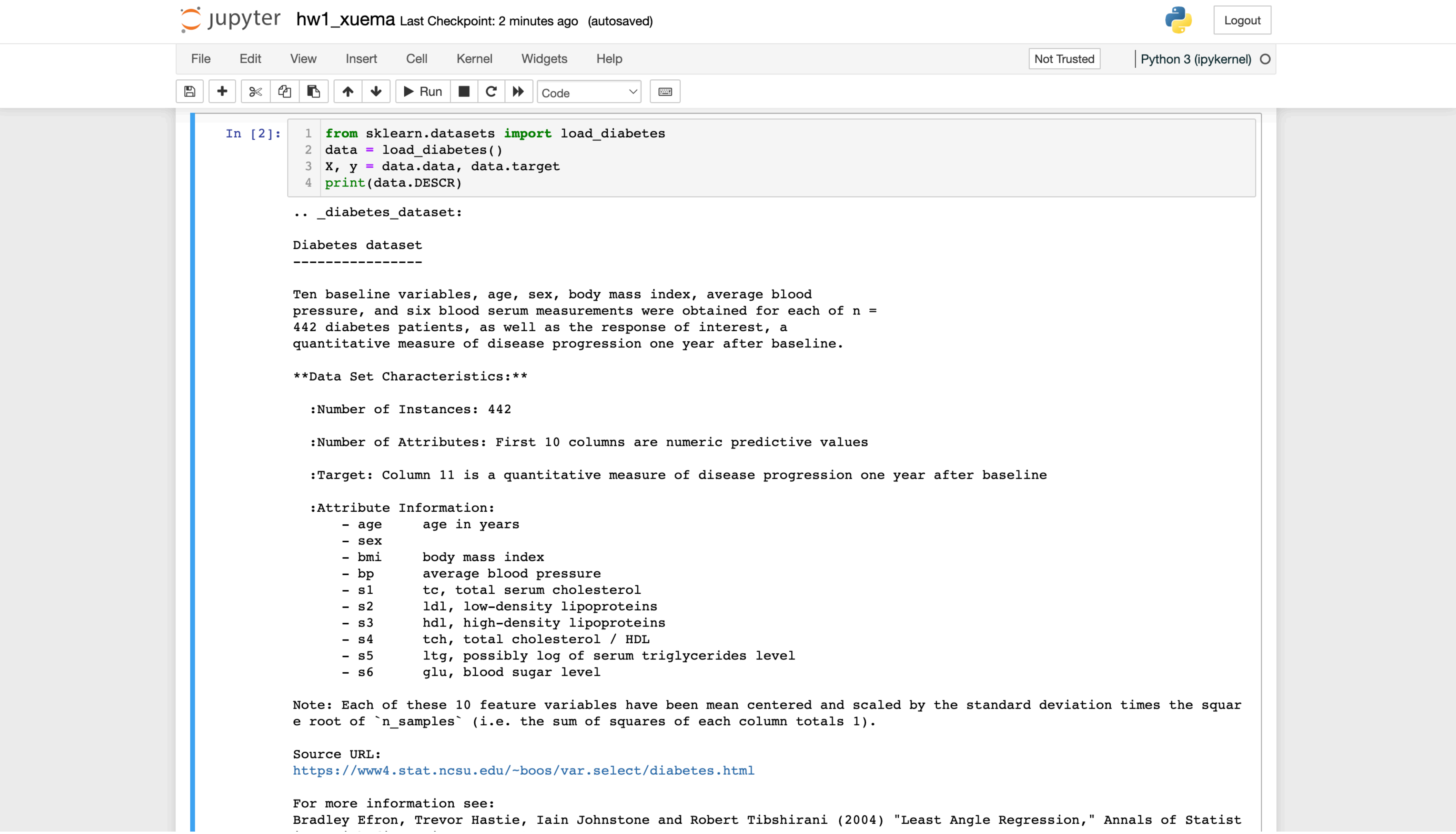
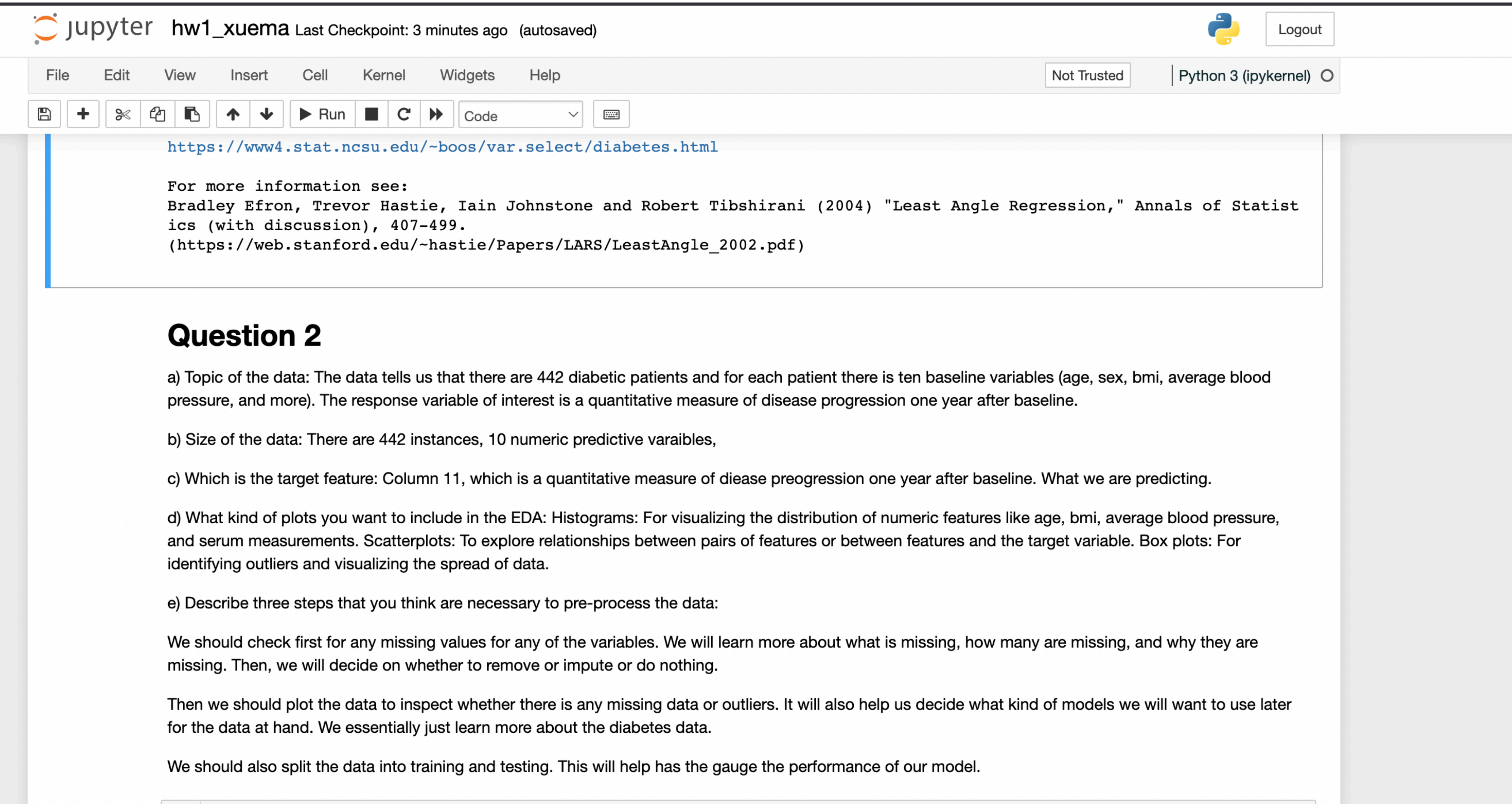


In [2]:

```python
from sklearn.datasets import load_diabetes
data = load_diabetes()
X, y = data.data, data.target
print(data.DESCR)
```

```
.. _diabetes_dataset:

Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one year after baseline

  :Attribute Information:
      - age     age in years
      - sex
      - bmi     body mass index
      - bp      average blood pressure
      - s1      tc, total serum cholesterol
      - s2      ldl, low-density lipoproteins
      - s3      hdl, high-density lipoproteins
      - s4      tch, total cholesterol / HDL
      - s5      ltg, possibly log of serum triglycerides level
      - s6      glu, blood sugar level

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times the squar
e root of `n_samples` (i.e. the sum of squares of each column totals 1).

Source URL:
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
```



```
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.

Question 3

matrix A matrix B

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

4 by 2 2 by 4

A) No, we cannot add matrix A and B together because the dimensions do not match. The matrices being added together must have the same dimensions.

B) Yes, we can add matrix A^T and B together because the dimensions match after the transpose of A. The transpose of a matrix is a flipped version of the original matrix. So, the dimensions of A becomes 2 by 4 after it is transposed

C) A • B

matrix A matrix B

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} = \begin{bmatrix} 11 & 14 & 17 & 20 \\ 23 & 30 & 37 & 44 \\ 35 & 46 & 57 & 68 \\ 47 & 62 & 77 & 92 \end{bmatrix}$$

a) $1 \cdot 1 + 2 \cdot 5 = 11$ b) $1 \cdot 2 + 2 \cdot 6 = 14$ c) $1 \cdot 3 + 2 \cdot 7 = 17$ d) $1 \cdot 4 + 2 \cdot 8 = 20$

e) $3 \cdot 1 + 4 \cdot 5 = 23$ f) $3 \cdot 2 + 4 \cdot 6 = 30$ g) $3 \cdot 3 + 4 \cdot 7 = 37$ i) $3 \cdot 4 + 4 \cdot 8 = 44$

j) $5 \cdot 1 + 6 \cdot 5 = 35$ k) $5 \cdot 2 + 6 \cdot 6 = 46$ l) $5 \cdot 3 + 6 \cdot 7 = 57$ m) $5 \cdot 4 + 6 \cdot 8 = 68$

n) $7 \cdot 1 + 8 \cdot 5 = 47$ o) $7 \cdot 2 + 8 \cdot 6 = 62$ p) $7 \cdot 3 + 8 \cdot 7 = 77$ q) $7 \cdot 4 + 8 \cdot 8 = 92$

$B \cdot A$

matrix B   matrix A

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 50 & 60 \\ 114 & 140 \end{bmatrix}$$

a) $1 \cdot 1 + 2 \cdot 3 + 3 \cdot 5 + 4 \cdot 7 = 50$

b) $1 \cdot 2 + 2 \cdot 4 + 3 \cdot 6 + 4 \cdot 8 = 60$

c) $5 \cdot 1 + 6 \cdot 3 + 7 \cdot 5 + 8 \cdot 7 = 114$

d) $5 \cdot 2 + 6 \cdot 4 + 7 \cdot 6 + 8 \cdot 8 = 140$

D) We cannot calculate A^{-1} inverse because the matrix must be a square, which means it must have the same number of rows and columns. Matrix A is 4 by 2 so there is no inverse.

We also cannot calculate $(AB)^{-1}$ inverse. $A \cdot B$ is actually a 4 by 4 square, but the issue is that it's determinant is 0 so we cannot find the inverse of $(A \cdot B)^{-1}$.
I used a determinant calculator on google.

Question 4

G-C = 60%

A-T = 40%

a) The probability that two successive nucleotides in this region are G-C is 36%.

$60\% \cdot 60\% = 36\%$

b) The probability that two successive nucleotides in this region are A-T is 16%.

$40\% \cdot 40\% = 16\%$

c) The probability that two successive nucleotides in this region are not the same is 48%

$100\% - (36\% + 16\%) = 48\%$

Question 5

Calculate the probability of deafness among white cats

$P(D) = ?$

a) $P(B) = 30\%$

$P(BB) = 10\%$

$P(N) = 60\%$

$P(D|BB) = 73\%$

$P(D|B) = 40\%$

$P(D|N) = 19\%$

$P(D) = P(B) \cdot P(D|B) + P(BB) \cdot P(D|BB) + P(N) + P(D|N)$

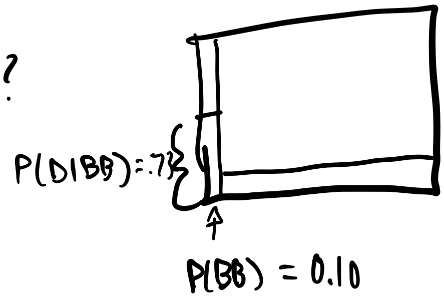
$P(D) = 0.40 \cdot 0.30 + 0.73 \cdot 0.10 + 0.60 \cdot 0.19$

$P(D) \quad 0.12 + 0.073 + 0.114$

$P(D) = 0.307$

$P(D) = 30.7\%$

b) $P(BB|D) = ?$



$$P(BB|D) = \frac{P(BB) \cdot P(D|BB)}{P(BB) \cdot P(D|BB) + P(B) \cdot P(D|B) + P(N) \cdot P(D|N)}$$

$$P(BB|D) = \frac{0.073}{0.073 + 0.12 + 0.114} = 0.2378$$

$P(BB|D) = (23.78\%)$

$P(N|D) = ?$

$$P(N|D) = \frac{P(N|D) \cdot P(D|N)}{P(N|D) \cdot P(D|N) + P(BB) \cdot P(D|BB) + P(B) \cdot P(D|B)}$$

$$P(N|D) = \frac{0.60 \cdot 0.19}{0.60 \cdot 0.19 + 0.73 \cdot 0.10 + 0.40 \cdot 0.30}$$

$$P(N|D) = \frac{0.114}{0.114 + 0.073 + 0.12}$$

$$P(N|D) = 0.3713 = 37.13\%$$