Data set contaminated by the robot

We work with the data set `diabetes` accessible in python. The initial data consists of $n = 442$ patients and $p = 10$ covariates. The output variable Y is a score reflecting the disease progressing. For fun, a bad robot has contaminated the data set by adding 200 inappropriate exploratory variables. Since simple noising the data was not sufficient for the robot, he arbitrarily permuted the variable. To complete the picture, the robot has erased any trace of his villainous act and thus we do not know which variables are relevant. The new data set contains $n = 442$ patients and $p = 210$ covariates denoted by X. Are you capable to resolve the enigma created by the playful machine and retrieve the relevalnt variables?

(Q1) Import the data set `data_dm3.csv` accessible by the link `https://bitbucket.org/portierf/shared_files/downloads/data_dm3.csv`. The last column is the output variable Y. The other columns are the exploratory variables. Provide the number of the exploratory variables and the number of the observations.

(Q2) Are the exploratory variables centered? Normalized? And the output variable? Provide a scatter plot of four randomly chosen exploratory variables and the output variable (a scatter plot or a bi-plot/pairwise-plot plots all existing). Comment the obtained graphs.

(Q3) Train and test sample. Create two samples: one to learn the model X_{train} and one to test it X_{test}. Put 20% of the data set in the test sample. Provide the size of each of the 2 samples. Note that the new sample of the covariates X_{train} is not normalized. In what follows, please pay attention to include the intercept in the regression models.

(Q4) Provide the covariance matrix for X_{train}. Plot the eigenvalues of the covariance (or correlation) matrix in descending order. Explain why does it make sense to keep only first PCA variables. In what follows, we will keep 52 variables.

(Q5) Following the observations of the question (Q4), apply the method "PCA before OLS" that consists in applying OLS with Y and the 52 first principal components. Run linear regression (with intercept), then plot the values of the coefficients (but not for the intercept). On another graph, do the same using the classical OLS.

(Q6) To prepare for the cross-validation, split randomly the train sample in 4 equal parts (called “folds”). Provide the numbers of the observations falling into each fold.

(Q7) For the two methods (OLS and PCA before OLS): Plot the residuals of the prediction for the test sample. Plot their density (one can use a histogram for example). Calculate the determination coefficient for the test sample. Calculate the prediction risk for the test sample.

(Q8) Using the function `lassoCV` of the library `sklearn`, choose the regularization parameter for the LASSO. Provide the corresponding prediction risk value.

(Q9) Provide the variables selected by the LASSO. How many are they? Apply the OLS method to the selected variables. This method is called Least-square LASSO. Provide the prediction risk for the test sample.

(Q10) Implement RIDGE regression and compare it to the others algorithm. Consider a sufficiently large grid to select the best regularization parameter using cross validation (make sure that the selected regularization parameter should not lie at the boundary of the grid).