The dataset provide the results of four diagnostic methods, “Hinselmann”, “Schiller”, “Cytology” and “Biopsy”. Since the gold standard for cervical cancer diagnosis is usually biopsy, I choose “Biopsy” as the target, and exclude the results of the other three methods (columns “Hinselmann”, “Schiller” and “Cytology”).

Columns “STDs: Time since first diagnosis” and “STDs: Time since last diagnosis” are also excluded, since there are too many missing data in these two columns.

To test which features are correlated to the target variable “Biopsy”, I employ t-test to compare the mean between “cancer” (“Biopsy” = 1) and “no cancer” (“Biopsy” = 0) people for the features with numeric data. Take feature "STDs (number)" for example. The null hypothesis is, there is no difference in number of STDs between people with or without cancer. The alternate hypothesis is there is a difference in number of STDs between people with or without cancer. The statistical results show significant difference ($p < 0.05$, reject the null hypothesis) between “no cancer” and “cancer” people for features "Hormonal Contraceptives (years)", "STDs (number)" and "STDs: Number of diagnosis". These suggest that these features are correlated with the target variable. For some features, such as "Smokes (years)", the result does not show statistical significant difference, since the variance is too big.

I use chi-squared test to test to whether there is a significant relationship between features with categorical data and the target variable. The null hypothesis is, there is no relationship between the feature and the target variable, and they are independent. The alternative hypothesis is, there is a relationship between the feature and the target variable. A contingency table is made for each feature and the target variable. And chi-squared test is performed based on the contingency table. The results show that features "STDs", "STDs: condylomatosis"，"STDs: vulvo-perineal condylomatosis"，"STDs: HIV", "Dx: Cancer", "Dx: CIN", "Dx: HPV" and "Dx" are correlated with the target variable ($P<0.05$, reject the null hypothesis). For feature “STDs: genital herpes”, there is only one case with genital herpes, and this person also get cervical cancer.