Problem II: Playing with data

• Decide on your programming language (why wait?)
• Select a small dataset (e.g., Iris from the UCI repository)
• Compute some statistics or plot the data in some way, with your own code • Interpret the statistics or plots (what do they tell you about the data?)

Solution:
Programming language: **Python**. Because I am familiar with the language and also due to its rich ecosystem of data-centric libraries like (NumPy, Pandas, matplotlib), ease of learning, and strong community support.

Selected Dataset is of **Cervical Cancer** it has 835 data points and 36 features.

Part 1 : Age
Figure 1 shows how age is not a key factor as cervical cancer is found in all ages with younger women presenting more frequently with cervical cancer. It also shows that there are more healthy cases than cancerous cases.

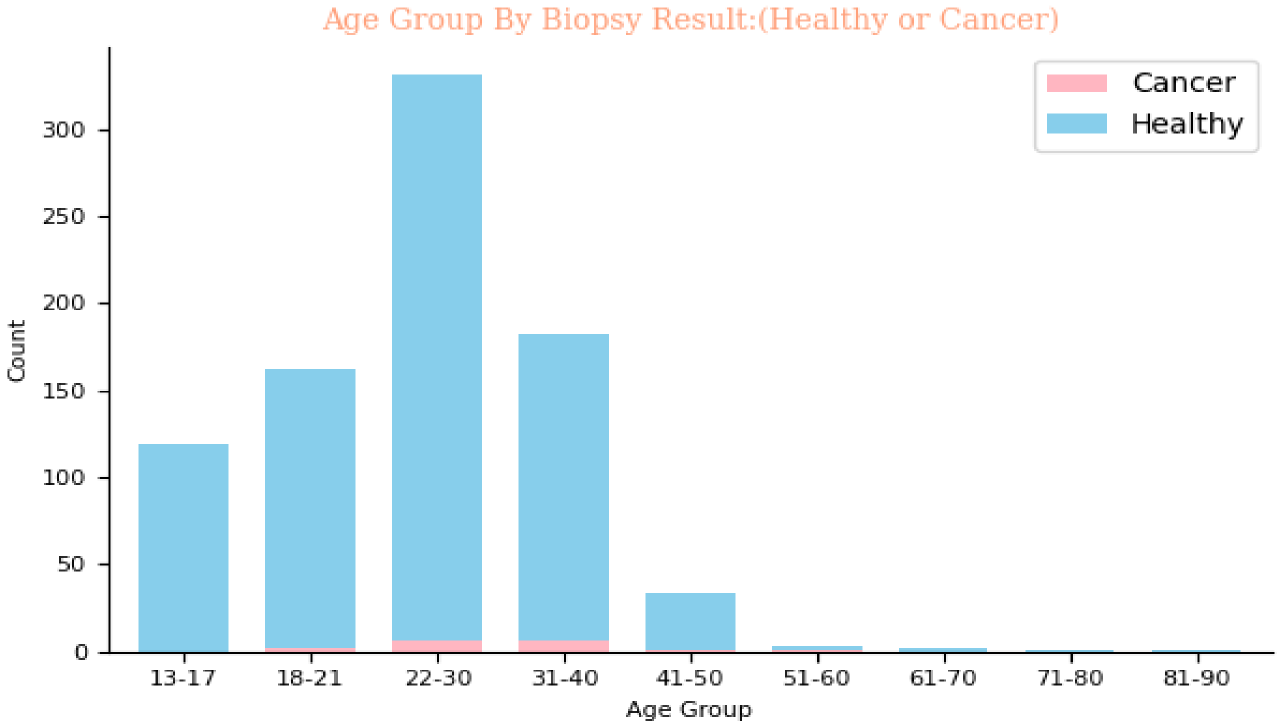

Figure 1: Age Group By Biopsy Result

Problem II: Playing with data

Part 2: STDs

Cervical cancer is extremely rare in women younger than age 20. However, many young women become infected with multiple types of human papillomavirus, which then can increase their risk of getting cervical cancer in the future.

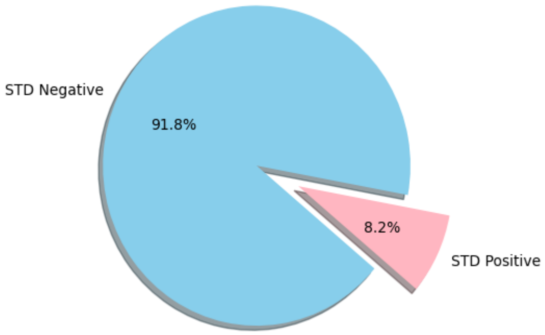

Figure 2 (a): Diagnosed with STD

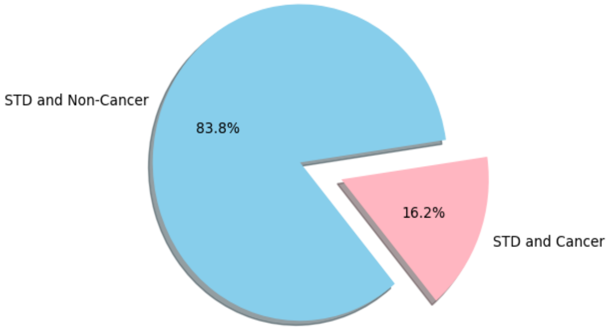

Figure 2 (b): Diagnosed with STD and Cervical Cancer

Conclusions: Figure 2 (a) and Figure 2 (b) shows how STDs are a known risk factor for cervical cancer.

Part 3 : Smoking

Smoking increases the risk of cervical cancer. Women who smoke are about twice as likely to get cervical cancer as those who don't. The risk increases with the duration and intensity of smoking.

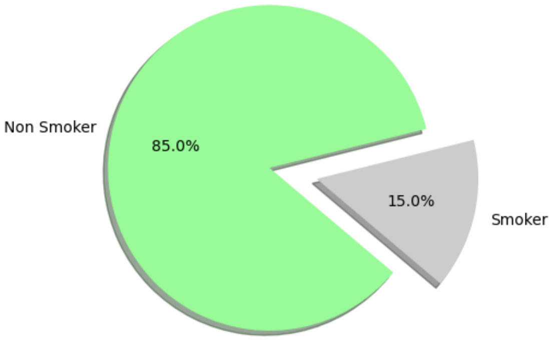

Figure 3 (a): Smoker vs Non Smoker

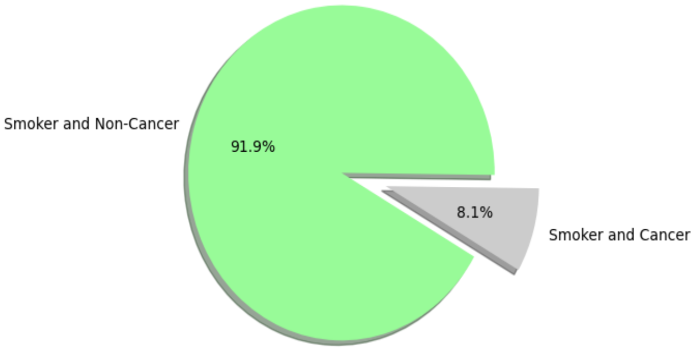

Figure 3 (b): Smokes and Diagnosed with Cancer

Conclusions: Figure 3 (a) and Figure 3 (b) shows that women who smoke have a risk of getting cervical cancer.