ASSIGNMENT 1: CIS335

—

1 High level task

Do some basic preliminary analysis on on a cancer dataset.

2 Deadline

See BlackBoard for deadlines

3 Dataset

Download the data from here. It should be under the assignment 1 folder

4 What to submit

- The report in pdf, named firstname_lastname.pdf
- Download the colab notebook as .ipynb file and submit that as well. Name it firstname_lastname.ipynb

5 What you should have done already

- Have gone through the lecture slides for "Getting to Know Your Data"
- Have finished working with the Google Colab Demo
- Downloaded all the assignment 1 materials from from here

6 What is expected

In a single pdf include sections that contain answers and discussions to the following:

- Have a Section titled *Statistical Distribution of Attributes*. There, have a table that lists the mean, median, Q1 and Q3 for at least three of the numeric attributes (you choose randomly any five) in the dataset. You have to present this information in a table. Absence of a table presentation will result in 50% penalty. (25)
- Have a Section titled *Outlier identification*. Choose one of the attributes that you selected in the previous section. Calculate the IQR distance and use that to identify the outlier rows in the data. Provide a screenshot of the result you get when identifying outliers using this particular attribute's IQR distance. (25)

- Have a Section titled *Correlations*. Use scatter plots to determine if there's correlation between the three numeric attributes you chose. Include the plots in the section and describe if you find any positive, negative or no correlation. (25)
- Have a Section titled *Boxplot*. Use boxplots for the three numeric attributes. Describe the distribution and see if there any any differences among them. (25)