Lab 2 : Analyzing Medical Data

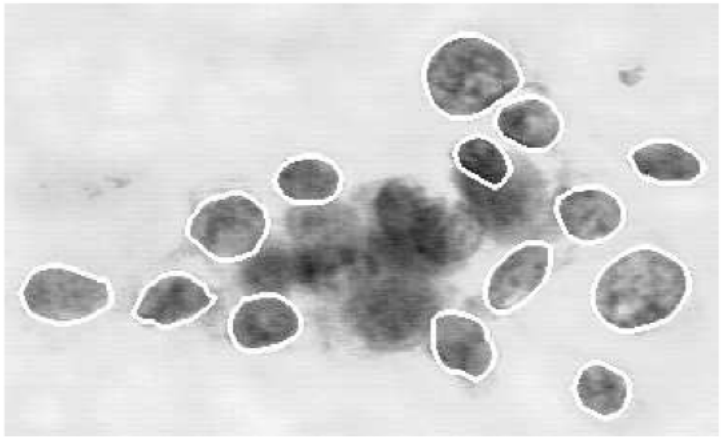
Lab Aims

In this lab, you will use your new knowledge to propose solutions that will be applied to real-world scenarios. To succeed, you will need to import data into Python, answer questions using the data, and the visualization techniques such as histograms and density plots to understand the data patterns. Now, some quick commands that you can use to do these charts.

- *sns.histplot* - Histograms show the distribution of a single numerical variable.
- *sns.countplot*- Shows the counts of observations in each categorical bin using bars.
- *sns.barplot* - Bar charts are useful for comparing quantities corresponding to different groups.
- *sns.kdeplot* - KDE plots (or 2D KDE plots) show an estimated, smooth distribution of a single numerical variable (or two numerical variables).
- *sns.jointplot* - This command is useful for simultaneously displaying a 2D KDE plot with the corresponding KDE plots for each individual variable.

1 Medical Scenario

You'll work with a real-world dataset containing information collected from microscopic images of breast cancer tumors, similar to the image below.



Each tumor has been labeled as either benign (noncancerous) or malignant (cancerous).

In these datasets, each row corresponds to a different image. Each dataset has 31 different columns, corresponding to:

- The column (**'Diagnosis'**) that classifies tumors as either benign (which appears in the dataset as **B**) or malignant (*M*),
- and 30 columns (extracted features) containing different measurements collected from the images.

1.1 Load the data

Your first assignment is to import and configure the Python libraries that you will need to complete this exercise. Then, you will load two data files.

- Load the data file corresponding to benign tumors into a DataFrame called **cancer_b_data.** Use the *"Id"* column to label the rows.
- Load the data file corresponding to malignant tumors into a DataFrame called **cancer_m_data**. Use the *"Id"* column to label the rows.

1.2 Review the data

(i) Print the **first and last 5 rows** of each data frame (benign and malignant tumors).

(ii) Print the **list of columns/features**.

(iii) Count the observations in each data frame.

(iv) Concatenate both data frames into one data frame called **cancer_data**.

(v) Count the observations in the **cancer_data** data frame.

(vi) What is the largest value for 'Perimeter (mean)' for benign tumors and its corresponding *'Id'*?

(vii) What is the average value and standard deviation of **'Radius (mean)'**?

(viii) What is the value for **'Radius (mean)'** for the tumor with Id 842517?

1.3 Investigating differences

(ix) create a single figure containing the distribution in values for 'Area (mean)' for both benign and malignant tumors.

(x) In fact, the 'Area (mean)' column can be used to understand the difference between benign and malignant tumors. Based on the histograms above:

- Do malignant tumors have higher or lower values for 'Area (mean)' (relative to benign tumors), on average?
- Which tumor type seems to have a larger range of potential values?

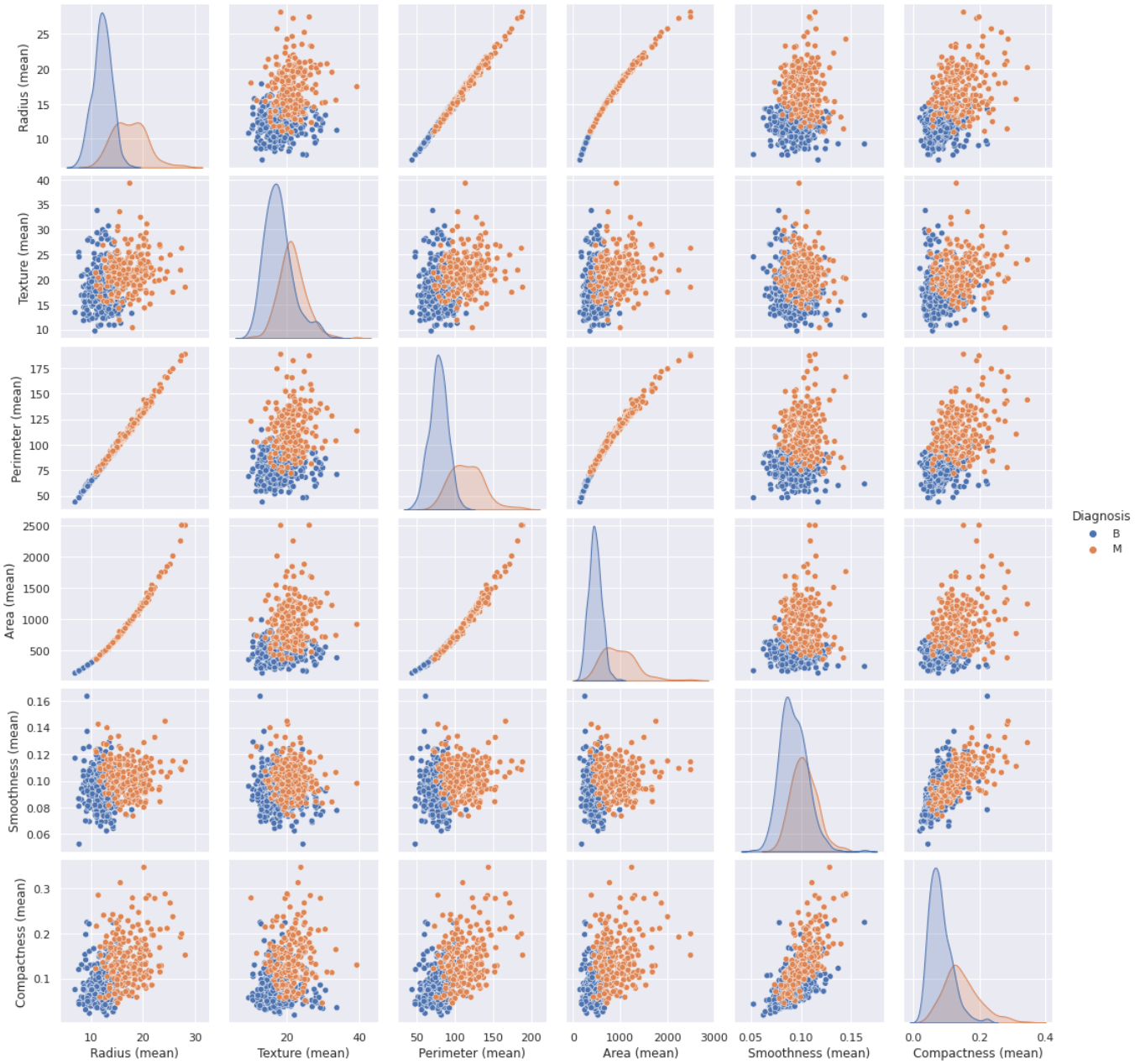
(xi) Create a single figure containing the KDE plots (that show the density) of 'Radius (worst)' for both benign and malignant tumors.

(xii) A hospital has recently started using an algorithm that can diagnose tumors with high accuracy. Given a tumor with a value for 'Radius (worst)' of 25, do you think the algorithm is more likely to classify the tumor as benign or malignant?

(xiii) Create a two-dimensional (2D) KDE plot that show the distribution in values for 'Radius (worst)' and 'Area (mean)' for both benign and malignant tumors.

(xiv) Provide a pair-plot that plots all pairwise relationships in the data-set by using the corresponding method of sea-born library. However, it will take long time. Therefore, you can test it with only the following columns: *'Diagnosis','Radius (mean)', 'Texture (mean)', 'Perimeter (mean)', 'Area (mean)', 'Smoothness (mean)', 'Compactness (mean)'*.

(xv) Provide the same pair-plot, but now use the "Diagnosis" feature ('B' or 'M') in order to map plot aspects to different colors. What can you conclude? You have to obtain the same results of the following figure.



1.4 More Investigation

- Create a boxplot of the 'Diagnosis' feature in the function of the 'Area (mean)' feature. Is the median of 'Area (mean)' greater for benign or malignant?
- Create a boxplot of the 'Diagnosis' feature in the function of the 'Perimeter (mean)' feature. Is the median of 'Perimeter (mean)' greater for benign or malignant?
- Create a 3D scatter plot variables (x,y,z): where 'Perimeter (mean)', 'Radius (mean)', and 'Area (mean)' represent the axes x, y and z, respectively. You will need the following code:

```
ax = plt.axes(projection='3d')
ax.scatter3D (x,y,z)
```

- What can you conclude from this plot?
- Which features might include outliers? You have to justify this by a visual graph.
- Create a heatmap to visualize the correlation coefficients between the extracted features.

Good Luck