Assignment 1

Goal:

Exploratory data analysis (EDA) is an essential part of analyzing data sets and helps summarize their main characteristics, often accompanied by visualization methods. EDA is amongst the first steps in the pipeline of any dataset analysis specially for medical image/signal analysis. EDA will help provide insights regarding the distribution of the data, detection of outliers, batch effects in data, etc. In this assignment, you will perform EDA and on a medical imaging dataset.

Data:

The data for this assignment is from the Prostate Imaging: Cancer AI (PI-CAI) challenge. The aim of the challenge is to detect and segment clinically significant prostate cancer (csPCa) in multi-parametric MR images including axial T2-weighted (T2W), high b-value diffusion-weighted (HBV), and apparent diffusion coefficient (ADC). To allow for robustness and generalizability of the algorithms, the data includes both patients with benign disease (e.g. benign prostatic hyperplasia) and prostate cancer, and the data is from multiple centers and multiple MRI vendors. The labels include segmentations of the prostate gland as well as segmentations of cancer lesions.

Steps:

1. **Knowledge Gathering:** Make yourself familiar with the problem itself. A preliminary knowledge of prostate anatomy, prostate cancer, and MRI imaging will be helpful for the follow-up parts below. Visit the PI-CAI challenge at https://pi-cai.grand-challenge.org/. Join the challenge to have access to "Datasets: Imaging, labels", and read the information and instructions on this section as well as "Overview and Goals" section to have a better understanding of data/labels. Write a short paragraph in your own language about the challenge and the data.

2. **Download Dataset:** For this assignment, you are going to use the "Public Training and Development" portion of the Dataset, which is provided in 5 folds (~5.5GB each). If you have enough storage, download the whole dataset (1500 cases), otherwise, just use fold0 (300 cases). The labels including prostate and lesion segmentations are provided separately on a GitHub page (link available on the Dataset section under "Annotations"). Clone/download the GitHub to have access to annotations for all 1500 cases. Both MRI volumes (in MHA format) and annotations (in compressed Nifti format) are directly readable via 3D Slicer and SimpleITK library in Python.

3. **Visualize Sample Data:** The first step in exploration is to visualize the imaging data and the annotations for a sample case. For this part, locate the data/annotations for patient ID of 10522 and study ID of 1000532 (`10522_1000532`)

a) Open all 3 MRI volumes (T2W, ADC, HBV) for this case in 3D Slicer. Set the view layout to *"four-up"*, view all volumes in Axial orientation, and save the screenshot of the Slicer view. You can control the visibility in the Data module.

b) Open the whole prostate gland delineation and the human expert delineation of the lesions for this case. Set the view layout to *"side by side"*, show the axial view of both annotations, and save a screenshot.

c) Using Volume Information in the Volumes module, convert both annotations to *"label maps"*. Then in Data module, right click on each label map and convert it to *"segmentation node"*. This way, you would be able to overlay the annotations on the volumes. Set the view layout to *"four-up"*, view all 3 MRI sequences on Slices, and overlay the prostate segment and lesion segment on top. Also, show the label maps in 3D from the Data module and save the screenshot. Your screenshot will be similar to Figure 1. Save the Slicer scene for this step as well.
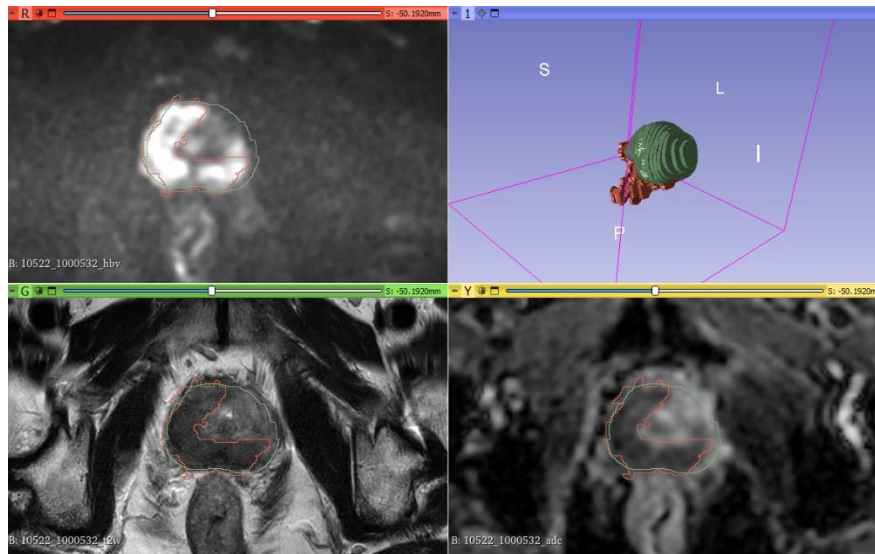

Figure 1. Sample screenshots of the 3 MRI sequences with overlaying prostate and lesions annotations.

4. **EDA on Clinical Information**: The clinical information about the cases in the dataset is stored in `picai_labels\clinical_information\marksheet.csv`. Write a Python script to read the file and plot the distribution of (a) prostate volumes, (b) case ISUPs, and (c) case csPCas for the whole dataset.

5. **EDA on Volumes:** ITK is an open-source, cross-platform library that provides developers with an extensive suite of software tools for image analysis. Among them, SimpleITK is a simplified layer built on top of ITK, intended to facilitate its use in rapid prototyping. To make yourself familiar with SimpleITK, check SimpleITK python notebooks: http://insightsoftwareconsortium.github.io/SimpleITK-Notebooks/
Use SimpleITK to extract the following information from the dataset (either fold0 or the whole dataset) to get insight on the variability among cases:

a) *Image dimension*: the number of voxels in the volume, represented by 3 values in x, y, and z axes (D_x, D_y, D_z). Plot the distribution of each value among cases for each MRI sequence.

b) *Voxel spacing*: the size of each voxel in the image in millimeters, represented by 3 parameters in x, y, and z axes (S_x, S_y, S_z). Plot the distribution of each parameter among cases for each MRI sequence. For each plot, indicate the spacing which the majority of cases share and report them as the "reference spacing" ($S_{x\text{-ref}}$, $S_{y\text{-ref}}$, $S_{z\text{-ref}}$).

c) *Prostate and lesions volume*: Calculate an estimate of the volume of the prostate and lesions. To do this, use the number of active voxels within the segmentations, and the physical size of each voxel (the spacing information you extracted in (b) for each image). Plot the distribution of estimated prostate volume and lesion volume separately.

6. **Volume Re-slicing**: resample the volume to change the voxel spacing. As you saw in part 5.b, the spacing of MRI sequences are different for each case and among cases. By resampling, you will end up with homogenous spacing over all dataset (the same spacing among all data volumes). For this part, resample all the volumes (MRI sequences and annotations) for the case you already visualized in 3d Slicer (`10522_1000532`) to have the reference spacing you've found in 5.b. You can use `ResampleImageFilter` method from SimpleITK. Please note that re-slicing should result in approximately the same anatomy (physical dimensions of tissue DO NOT change by resampling).

What to Upload:
- Put your screenshots, plots, and your responses in a pdf report file and name it as `firstName_lasNmae.pdf`.
- Save your Python codes and Slicer scene.
- Compress all the above files, name it as `firstName_lasNmae.zip` and upload it to OnQ

Good luck