PNEUMONIA DETECTION

CAPSTONE PROJECT REPORT
by AIML Sep 19 B Group3

PROJECT INTERIM REPORT #1

Sep 7th, 2020

greatlearning
Power Ahead

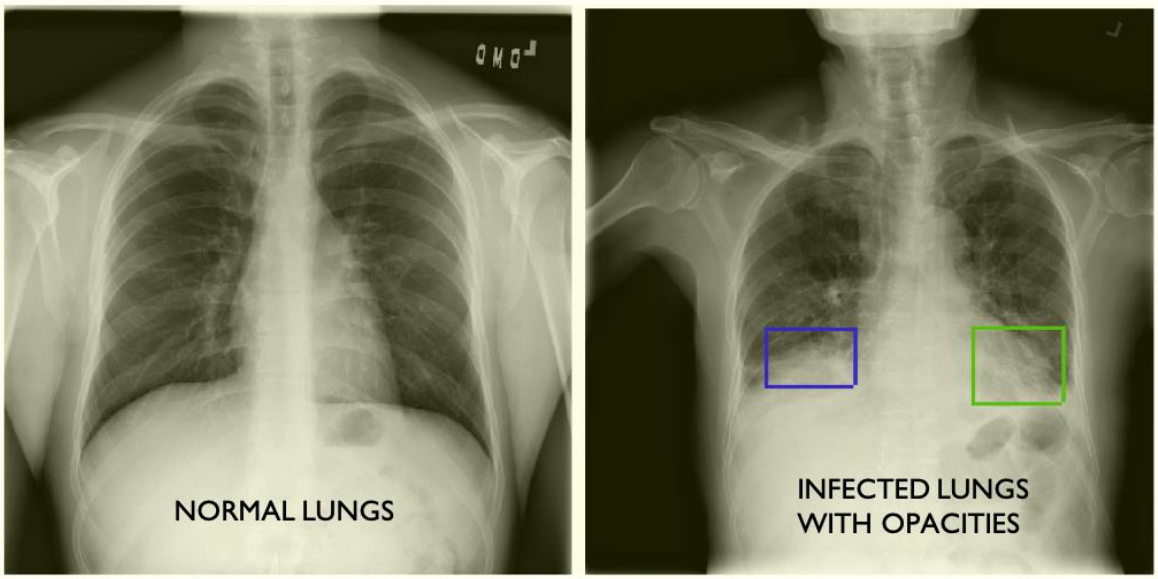
CONTENTS

Project Background 1
Background 2
Project Statement 2
Notes on Chest X-Ray Images 3
Exploratory Data Analysis 7
Data Report 7
Basic Data Report 7
DICOM Data Report 9

BACKGROUND

PROJECT STATEMENT

In this capstone project, the goal is to build a pneumonia detection system by locating the position(s) or zone(s) of inflammation in sampled patient Chest X-Ray(CXR) images like the ones shown below. The project explores a variety of deep neural network architectures to identify optimum detection via object detection as well as sematic segmentation methods.

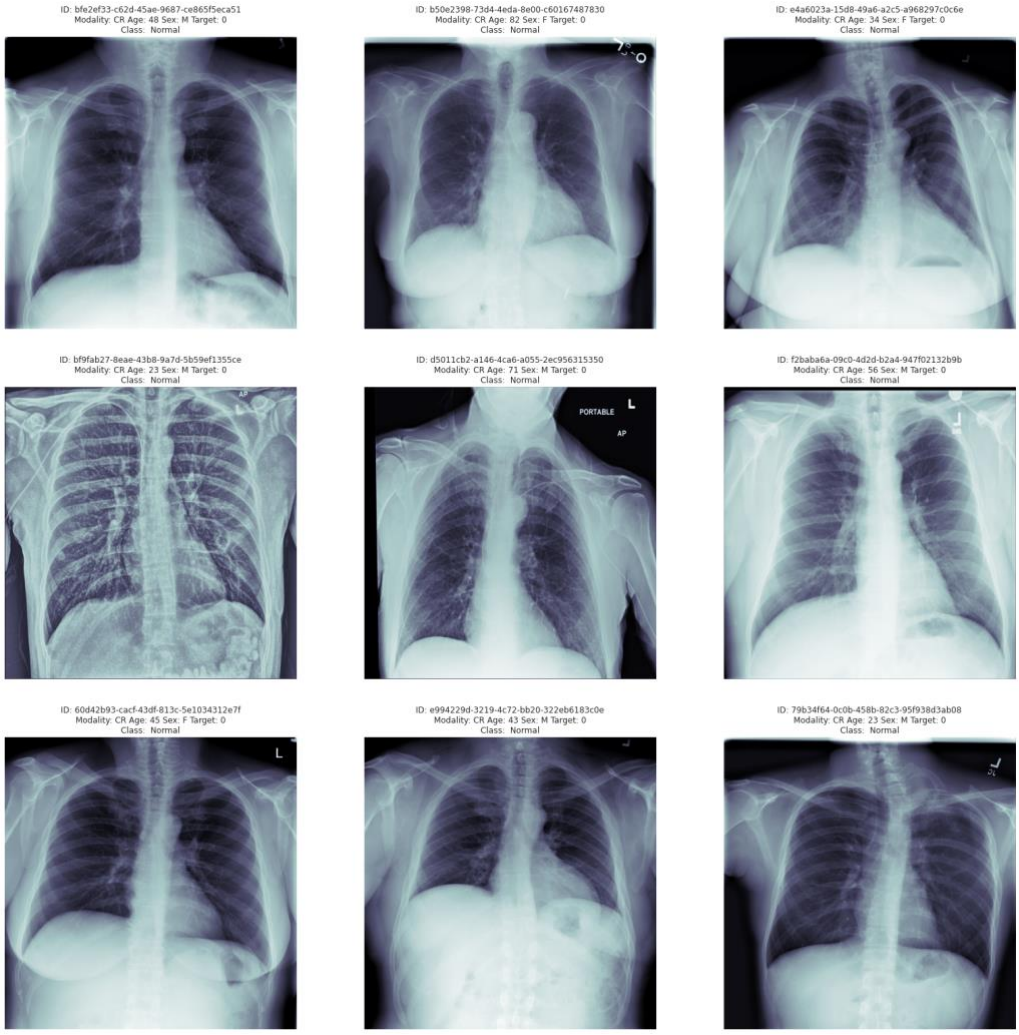


In X-Ray images, tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear **black** in the image, indicating transparency. Dense tissues such as bones absorb X-rays, indicate opacity and appear **white** in the image. We are detecting lung opacities in the images which are then indicative of pneumonia. X-ray images in the dataset given are identified either as "*Normal*" lungs or as pneumonia infected *"Lung opacity"* via annotation boxes via inputs from practicing radiologists. The "lung opacities" typically appear **grey** and "hazy" in X-ray images with partial transparency and with the lack of a clear boundary. The additional complexity in the dataset given is that there are also X-ray images where lung opacities are not pneumonia related such as in the case of lung tumors that maybe cancer related or in some cases, chest cavity filled with fluid or even in the case of patients with heart enlargement.

PROJECT BACKGROUND

CHEST X-RAY IMAGES
(Notes from Dataset Visualization and Research on CXR Images)

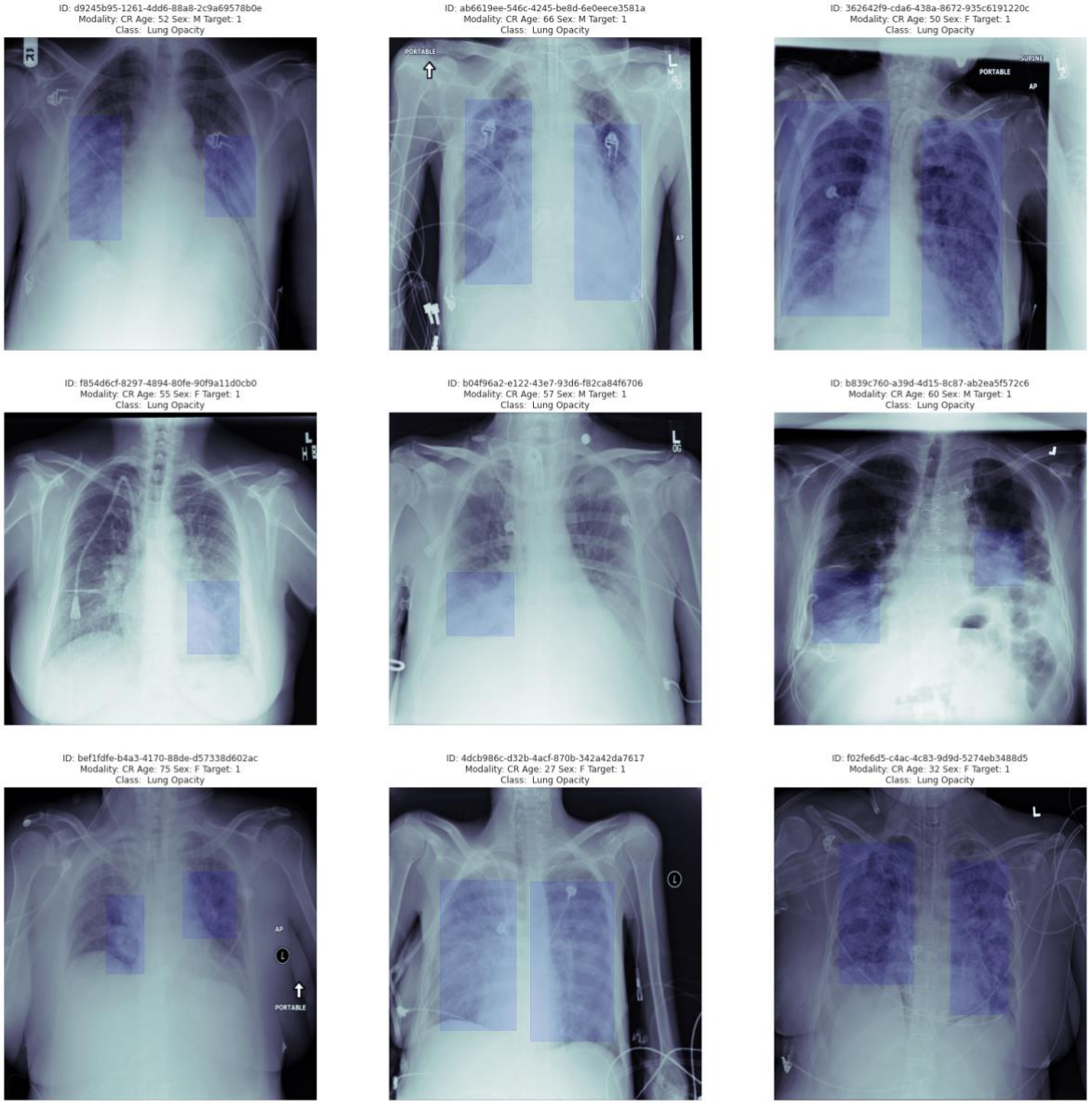
The lung infection or patient afflictions observed in the CXRs images are of many kinds. Many of them are not indicative of pneumonia and are in fact classified in our dataset as "Not normal / No lung opacities". This extra third class of images had abnormalities on the image and oftentimes may mimic the appearance of true pneumonia. Let's look at the various possibilities to understand our dataset fully and that will help us clean up and augment data going forward and thus enable eventually a more robust set of ML pneumonia detection models. Let's 1st look at the CXRs belonging to *"Normal"* class which is associated with *"Target"*=0 in the dataset showing mostly black and clear lung air cavity with no grey zones to identify infections or abnormalities *(figure is code output)*



PROJECT BACKGROUND

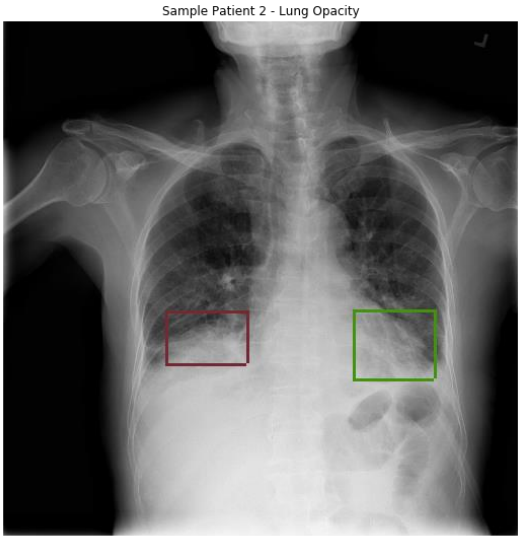
CHEST X-RAY IMAGES (contd.)

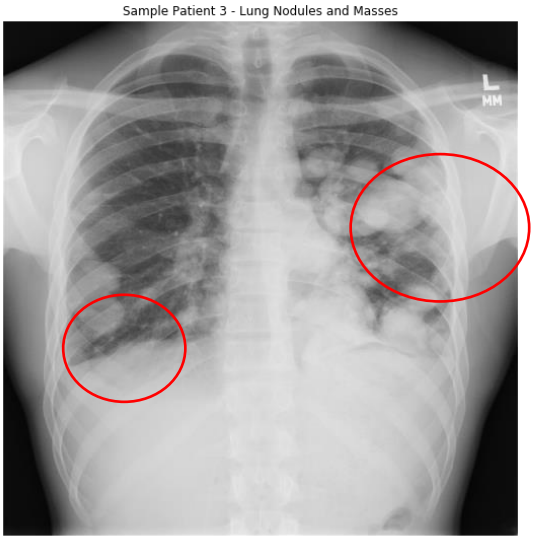
Let's now look at the CXR images identified as *"Lung opacity"* class and *"Target"* =1 in given dataset and with the box annotations to show the zones where lung opacities or grey areas are observed. Note that the grey areas are mostly hazy and with no shape or clearly defined boundaries (figure below is from code output)



CHEST X-RAY IMAGES (contd.)

Let's now take a closer look at image comparison below between a typical *"Lung opacity"* CXR sample image indicating pneumonia and an image type of CXR with abnormalities observed in the form of lung nodules and masses which could be indicative of tumour / cancer. Referring to the CXR image on the right below, the observed grey segments within the lung cavity space are identified as lung nodules/masses and have more well defined shapes (oval) and boundaries (figure is from research). This (CXR on right) would be not be identified as normal nor as pneumonia (*"Not Normal / No Lung Opacity"*). Contrast this with the image on left where the *"Lung opacity"* grey areas are hazy and with no shape or boundaries. These could be potentially confusing for the image segmentation or object detection algorithm we define and need perhaps adequate feature extraction (pointing to deep layered network architecture?)

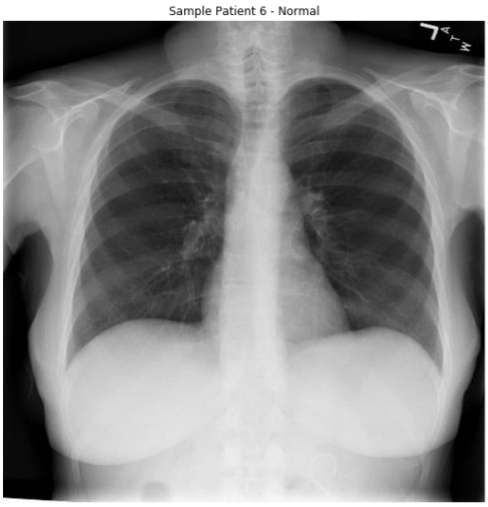


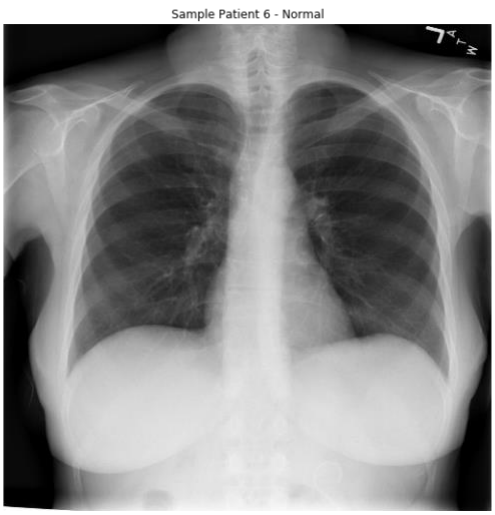


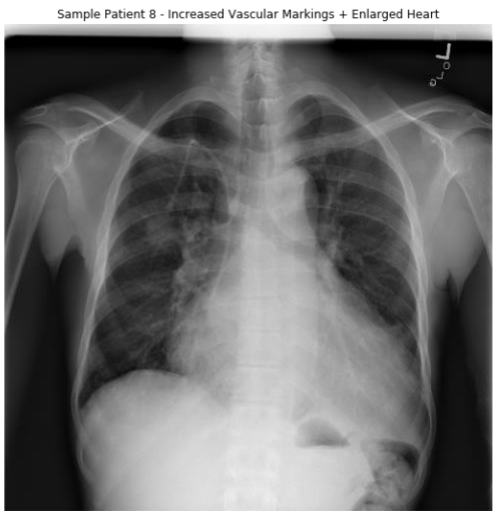
Note: Figure above and below are from Research

PROJECT BACKGROUND

More examples of "non-Pneumonia" abnormalities are shown in the 2 figures below where the images on the right in each figure show abnormal CXRs and are contrasted vs normal CXRs which are shown on the left side of the figure. In the 1st figure below, the image on right shows actually a CXR with chest cavity filled with fluid. In the 2nd image, the grey areas appearing to infringe into lung cavity zone due to an enlarged heart and vascular markings. Obviously in all above cases, no Pneumonia is diagnosed.







We thus need to be careful in data preparation and also augmentation for training our models while segregating the pneumonia patient CXR images vs the normal CXRs.

DATA REPORT

The X-ray medical images are stored in DICOM format files (*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data.

Details about the data and the dataset files are extracted from the link,

https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

BASIC DATA SUMMARY

- Images for testing and training are in the folders: stage_2_train_images and stage_2_test_images.
- Training data 1: "stage_2_train_labels.csv" contains the box annotation coordinates and the targets (1-for pneumonia patient and 0-not pneumonia)
- Training data 2: "stage_2_detailed_class_info.csv" contains detailed information about the classes in the training set ("Normal", "Lung opacities" and "No lung opacity / Not Normal" classes)
- Both metadata from CSVs are loaded into dataframes and observed to have 30227 patient samples info obtained from CXRs of 26684 unique patients. The dataframes are later merged into one with 'patient id' as the common identifier.
- The "train_labels" dataframe is seen to have 6 columns with:
    - 4 numeric real data corresponding to rectangle location and dimensions of the annotated box identified for each pneumonia patient,
    - 1 target class column which is a classifier for patient as pneumonia patient or not (1-yes / 0-no)
    - 1 column for the 'patient id' object data
- The "class info" dataframe has the 'class' identified for all the CXR images:
    - 3 classes observed: "Normal", "Lung Opacity" and "No Lung Opacity / Not Normal
- The 'training labels' data has 68.4% null data for the box identifier of patient CXR which should ideally correspond to number of patients identified with no pneumonia. This is confirmed thru the class distribution count plot below

EXPLORATORY DATA ANALYSIS

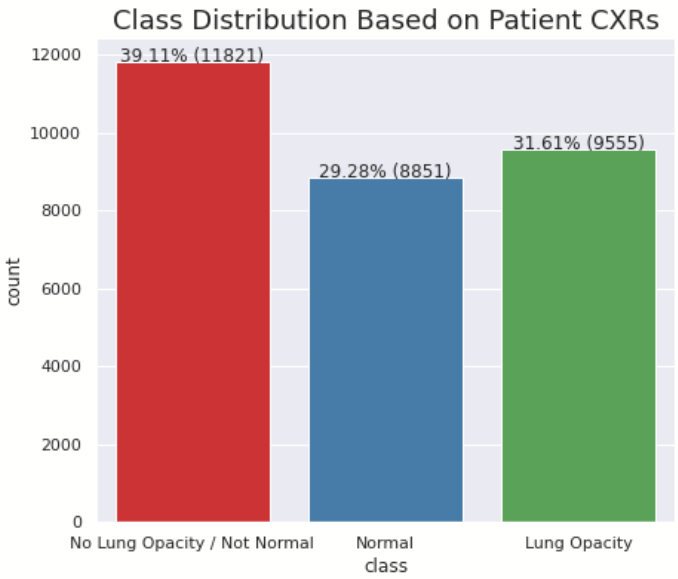


Figure is Code Output

- The non-Pneumonia patients (blue + red bars in above figure) add up to 68.4% which exactly corresponds to the above mentioned Null data in the dataframe.
- Let's look at the data distribution across the "Normal" and the "Lung Opacity" classes. We may consider just these classes for initial training of our models before adding samples from the 3rd class ("Not Normal / No Lung Opacity")

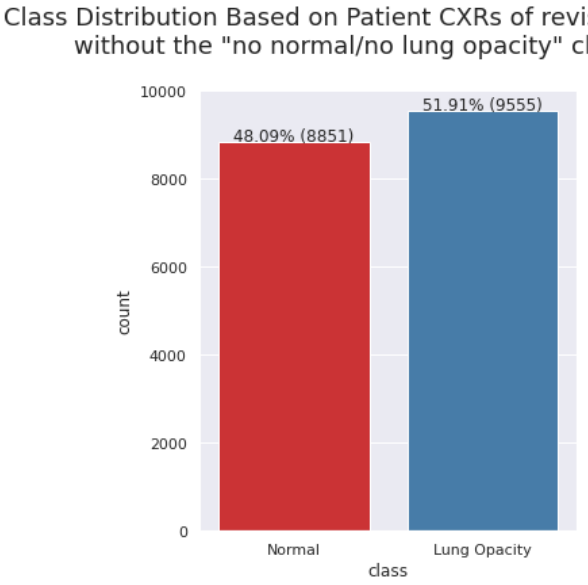


Table: Class Distribution of Unique Patient Records

Target	Class	CXRs_count per patient	Total count	%
0	No Lung Opacity / Not Normal	1	11821	44.3%
0	Normal	1	8851	33.2%
1	Lung Opacity	1	2614	9.8%
1	Lung Opacity	2	3266	12.2%
1	Lung Opacity	3	119	0.4%
1	Lung Opacity	4	13	0.1%
		TOTAL	26684	100%

- Some of the Pneumonia patients with lung opacities identified have multiple CXR records as shown by the table above. ~56% of the Pneumonia patients have atleast 2 CXRs taken.
- ~78% of the patients sampled are either normal or have abnormalities in their CXRs that are not Pneumonia related. These patients have just one CXR taken based on the dataset sample

EXPLORATORY DATA ANALYSIS

Based on the annotated box centres to denote lung opacities indicated by figure below, most of the infected zones seem to be centred around the range (Xc=200 to 400, Yc= 400 to 700) on the left and the range (Xc=600 to 800, Yc=400 to 700) on the right lung. The symmetry is intuitive since both lungs are equally vulnerable to Pneumonia and will have lung opacities accordingly.

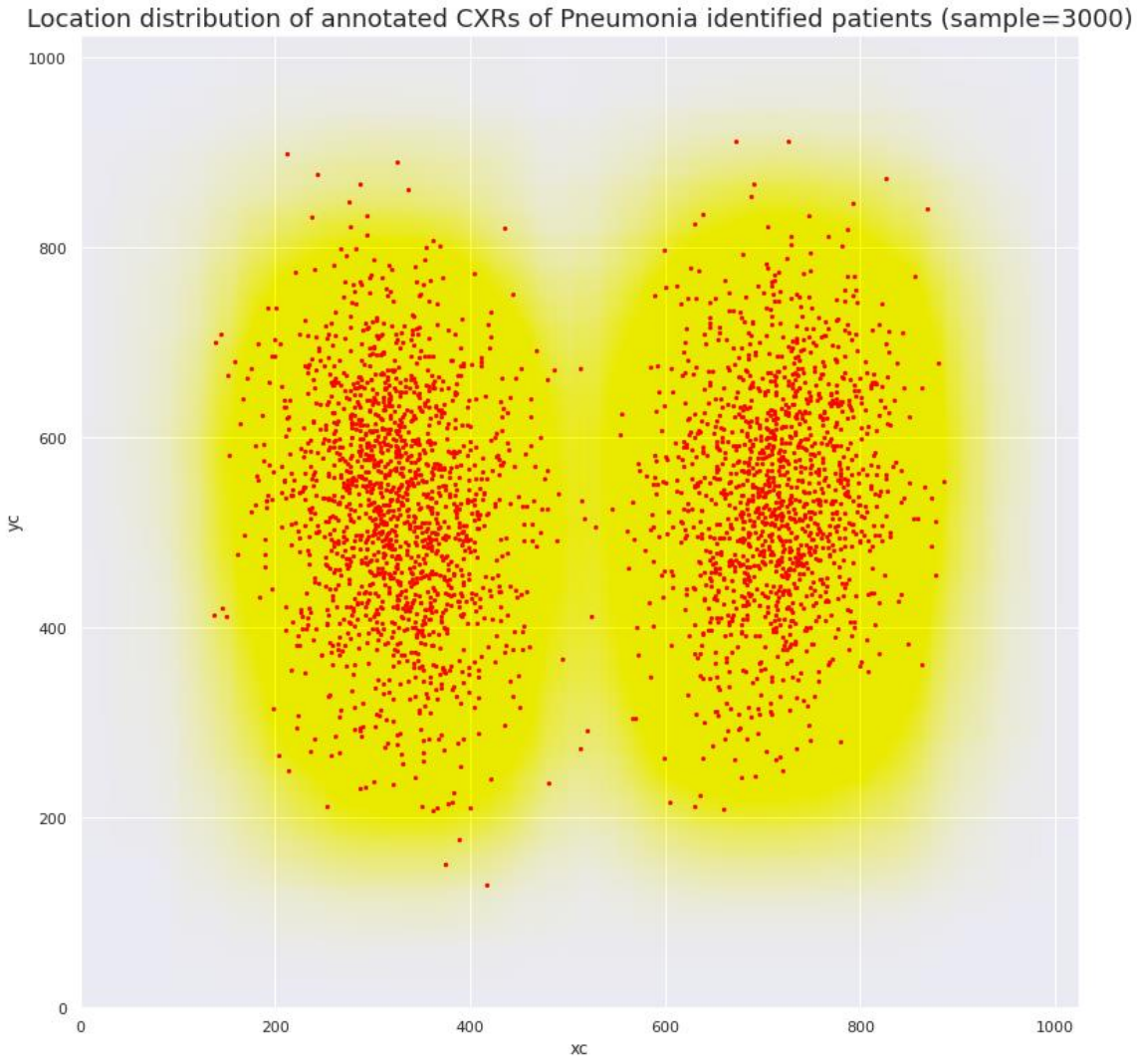
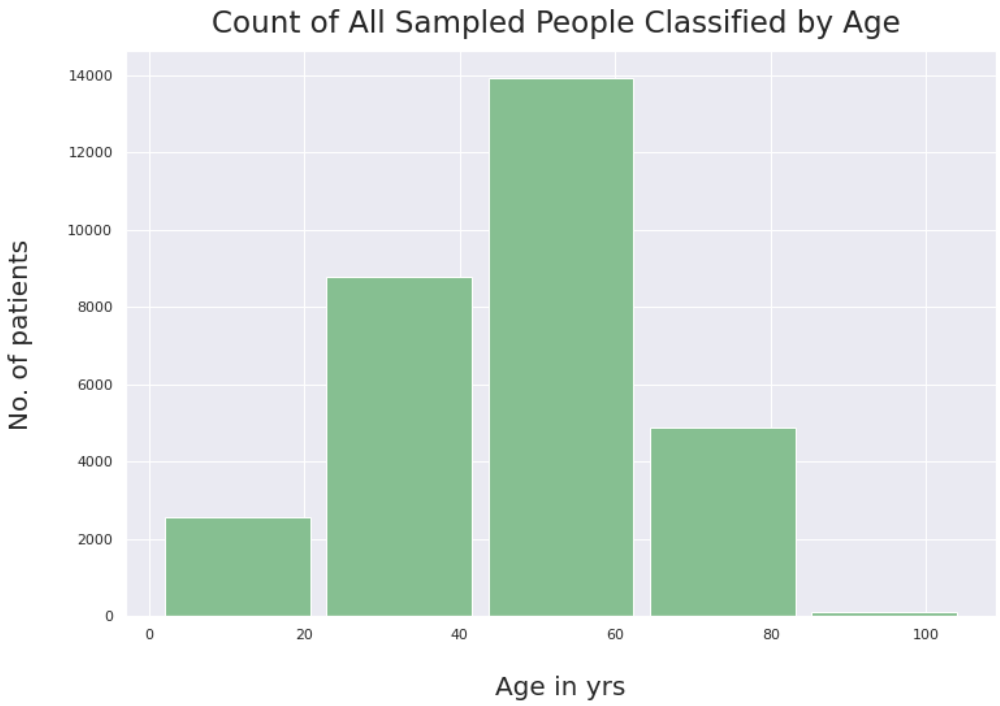
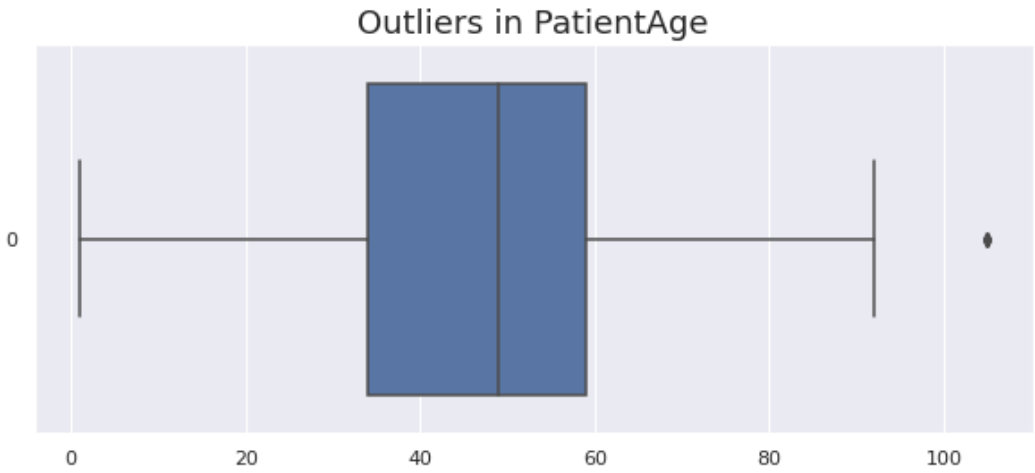


Figure Extracted from Code Output

DICOM Data: Age Based Statistics

- A larger proportion of patients (~46%) are from the age group 40-60
- A few outliers are seen (5) well above the age of 100 (max=155) and the age is capped off at 105 for these outliers during clean up



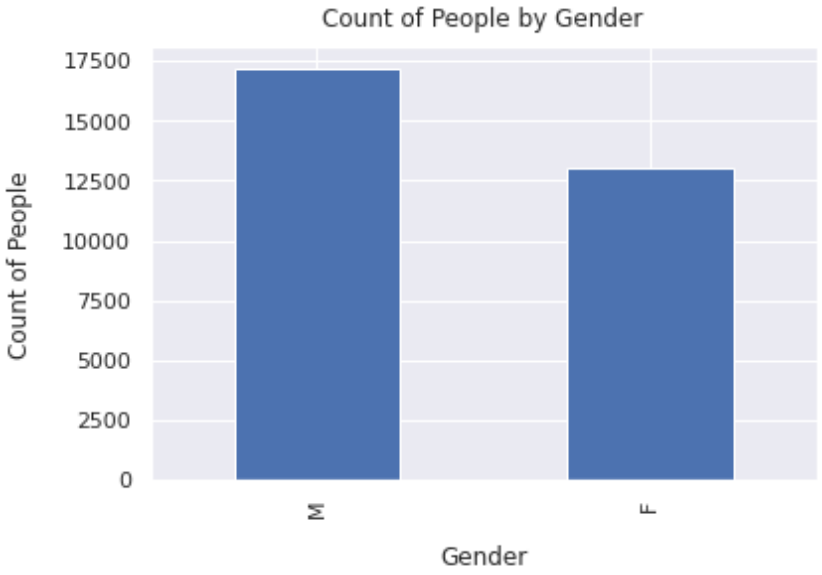


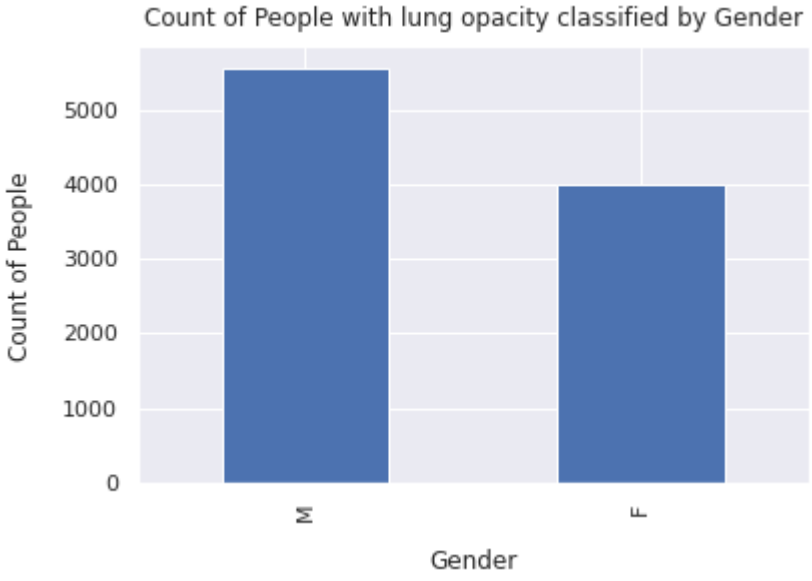
DICOM Data: Age Based Statistics (contd.)

- By comparing the 2 bar plots (below here and above in previous page), A distinctly higher percentage (~56%) of patients sampled in the age group 60-80 are showing lung opacities in their CXRs indicating a relatively higher propensity for Pneumonia infection in this age group.
- In other age groups, the %age of patients sampled showing Pneumonia infection are relatively much smaller (<25%)

DICOM Data: Gender Based Statistics

- A higher %age of males (~56%) are sampled in the dataset indicative perhaps of the higher incidence of such illnesses in males
- The proportion of patients infected by Pneumonia is the same across males or females (~32% or ~1/3rd of patients sampled)





DATA PREPARATION

MODEL BUILD