**Great Learning**

**Capstone Project – Interim Report (Milestone 1)**

**Computer Vision - Pneumonia Detection Challenge**

**Computer Vision Group 4**

**Group members –** Ajay Devnani, Manojkumar Sivaraman, Naveen Kumar Kaginelli, Sumeet Kumar, Umesh Singh Kushwaha

**Project Abstract –**

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it and the infection causes inflammation in the air sacs in your lungs, which are called alveoli. The alveoli fill with fluid or pus, making it difficult to breathe. Typically, X-ray helps your doctor look for signs of inflammation or opacities in your chest which when present can indicate the Pneumonia infections. Since Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally, it is crucial to identify and react swiftly if there are any infections identified.

As the detection and reaction time is vital and the infection is detected using the X-ray Images, Image processing techniques can be leveraged from the emerging AI technology on these images to predict the presence of opacities. Powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making. This will also assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (eg, radiology).

For this purpose, this project uses the **XXXXXXXXX** image processing techniques to create the Pneumonia prediction model which can predict Pneumonia on the patients with a Accuracy of **XX.XX%** there by helping the doctors to react quickly to save lives.

**Highlights in 3 bullet points –** xxxxxxxxxxxxxx

**Problem Statement:**

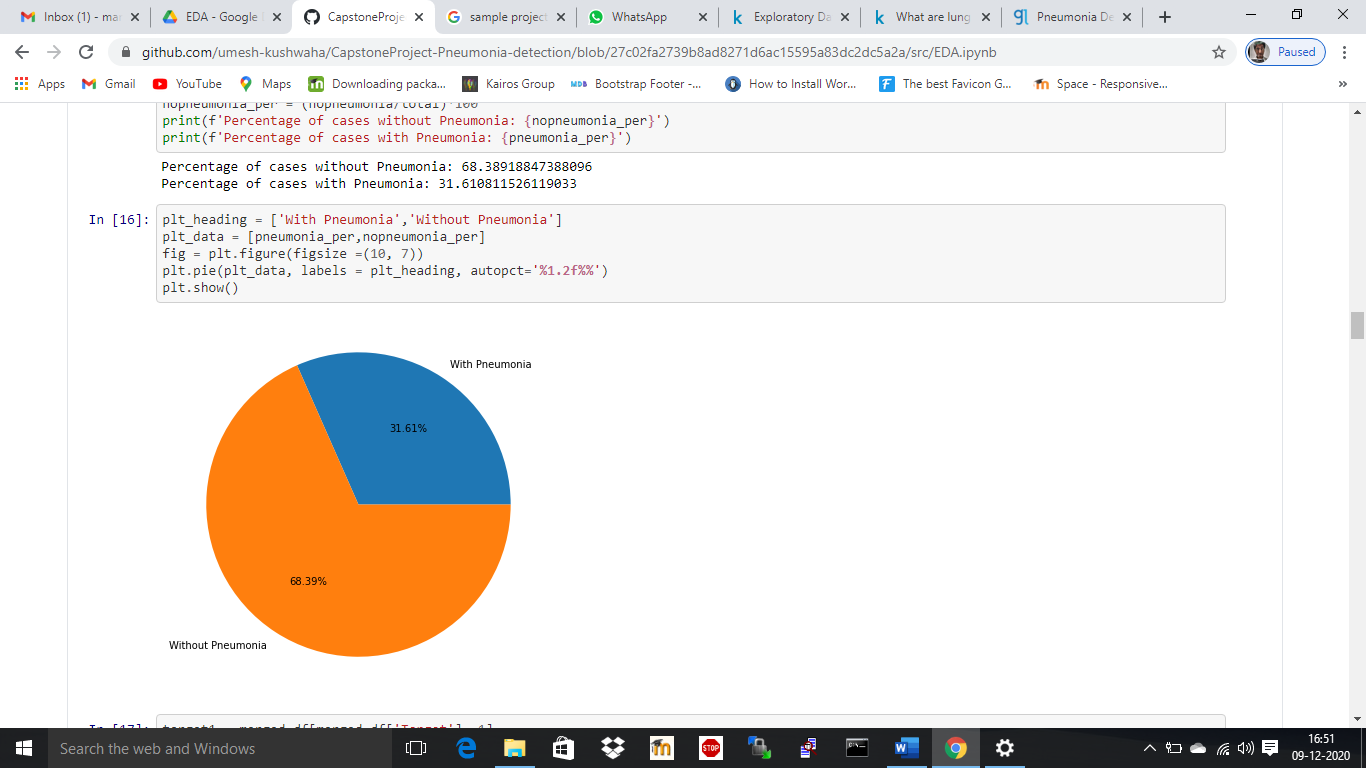
1. **Problem statement:** The problem is about detecting bounding boxes for lung opacity corresponding diagnosis of Pneumonia on chest radiographs (images). Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image. While we are theoretically detecting “lung opacities”, there are lung opacities that are not pneumonia related. In the data, some of these are labeled “Not Normal No Lung Opacity”. This extra third class indicates that while pneumonia was determined not to be present, there was nonetheless some type of abnormality on the image and oftentimes this finding may mimic the appearance of true pneumonia. The original medical images are stored in a special format called DICOM files (\*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data.

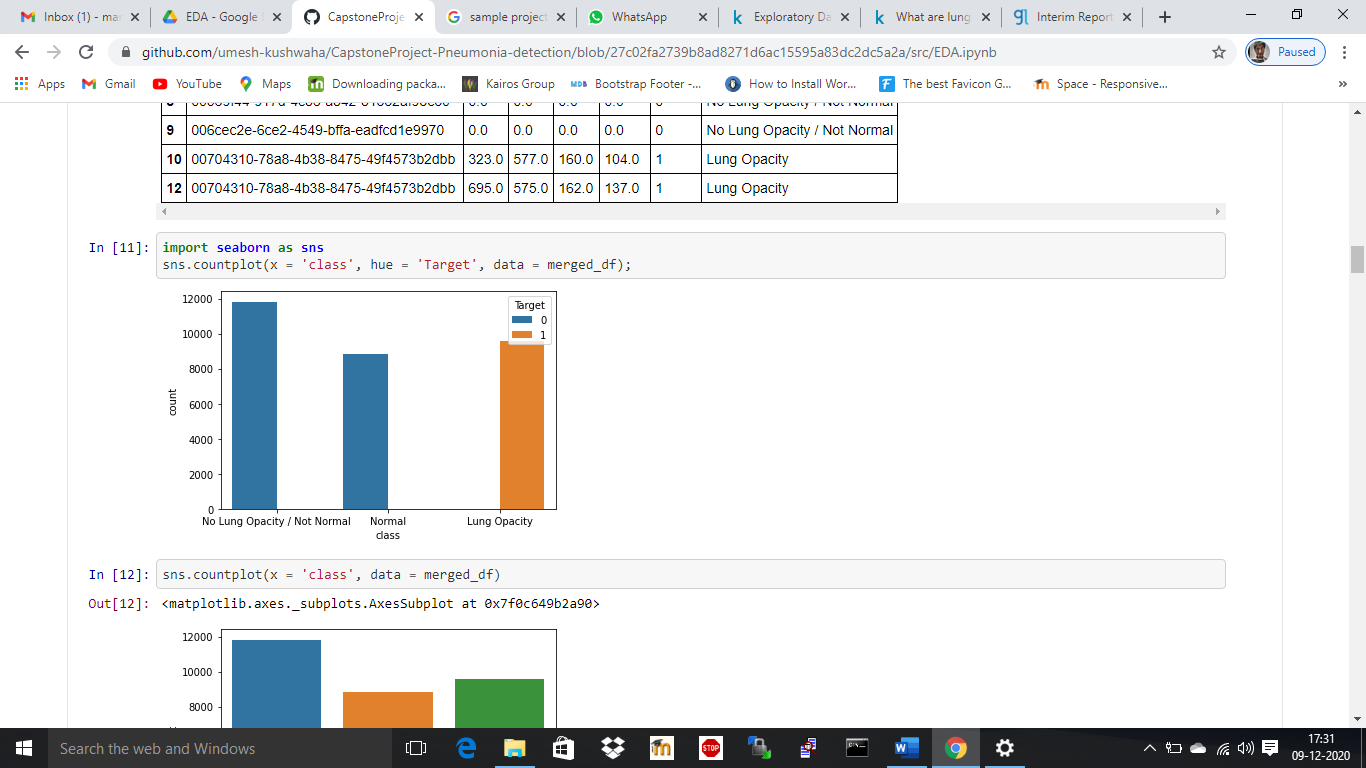
All lung opacities may not attribute to Pneumonia as the Pneumonia is one of the several diseases that can occur on a chest radiograph. A radiograph may contain one or more than1 bounding boxes for any possible Pneumonia case.

1. **Summary of problem statement, data and findings**
2. **Dataset:** In class dataset, information is given about the positive or negative class associated with a particular patient ID. In train dataset, information is given about the bounding box (x, y, w, h) comprising evidence of pneumonia.
3. **stage\_2\_train\_images:** It contains set of raw medical images (DICOM files) for training model. The DICOM files contain a combination of header metadata as well as underlying raw image arrays for pixel data.
4. **stage\_2\_test\_images:** It contains set of raw medical images (DICOM files) for testing the model. The file contains a combination of header metadata as well as underlying raw image arrays for pixel data**.**
5. **stage\_2\_train\_labels.csv:** This CSV file contains detailed information about the labels (Patient ID, bounding boxes for lung opacity and target 1 or 0 indicate the presence of abnormality i.e. Pneumonia)
6. **stage\_2\_detailed\_class\_info.csv:** This CSV files contains information regarding three possible classes in the data - namely normal, lung opacity and no lung opacity (not normal).
7. **DICOM files:** It contains a combination of header metadata as well as underlying raw image arrays for pixel data.
8. **Findings:** The most important thing is that a given patient ID may have multiple boxes if more than one area of pneumonia is detected.
9. **xx:** xxxxx
10. **Summary of the Approach to EDA and Pre-processing**

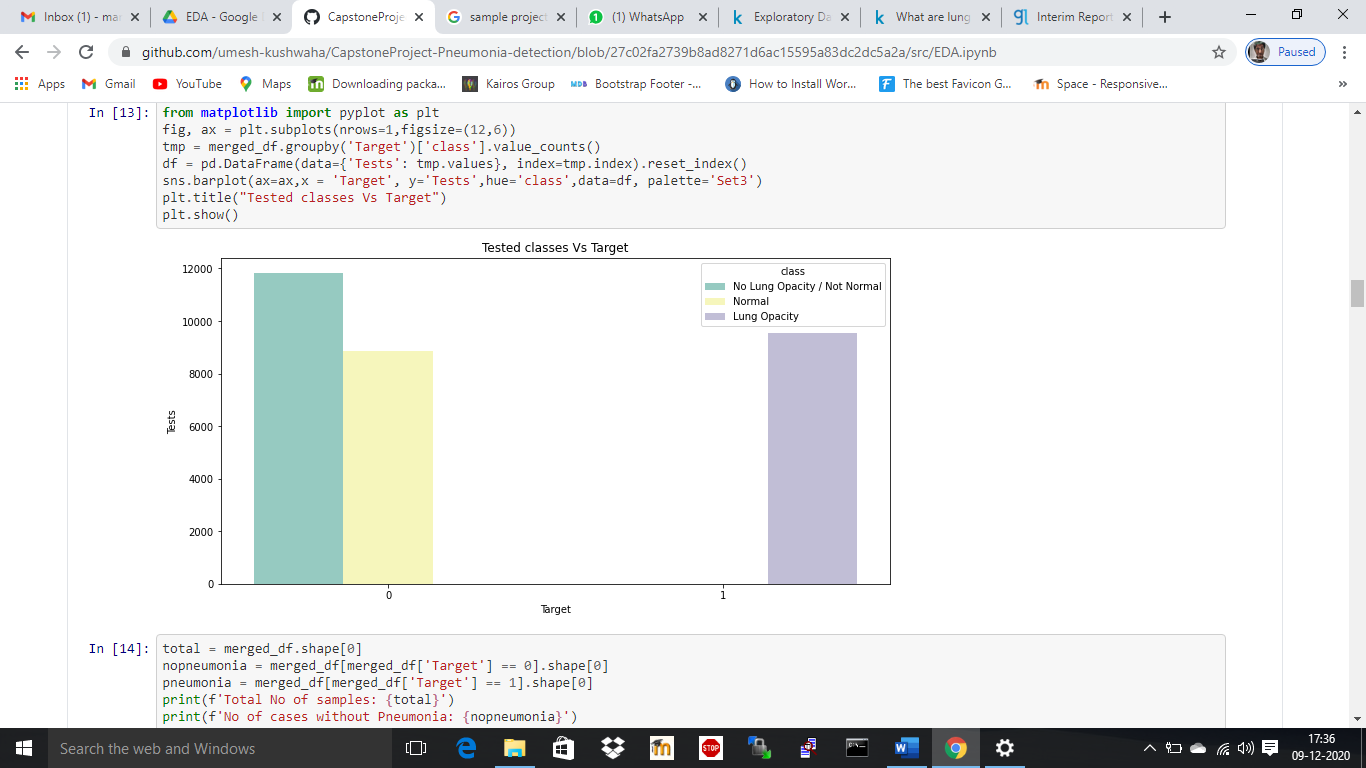
**Exploratory data analysis (EDA) –**

1. **Data information:** Out of the total class information of 30,227 patient IDs, 29.3% (8,851) belongs to ‘Normal’ class, 31.6% (9,555) comes constitutes ‘Lung Opacity’ class and 39.1% (11,821) is ‘No Lung Opacity / Not Normal’ class. Hence, 31.6% are with pneumonia and the remaining 68.4% are without pneumonia.



On plotting the class values and its corresponding counts segregated with the target values, we observe that the count of patients with No Lung opacities/ Not normal is high than the pneumonic or normal patients. Also, the count of normal class is less than other 2 classes indicating that the data has a greater number of Ill health patients.

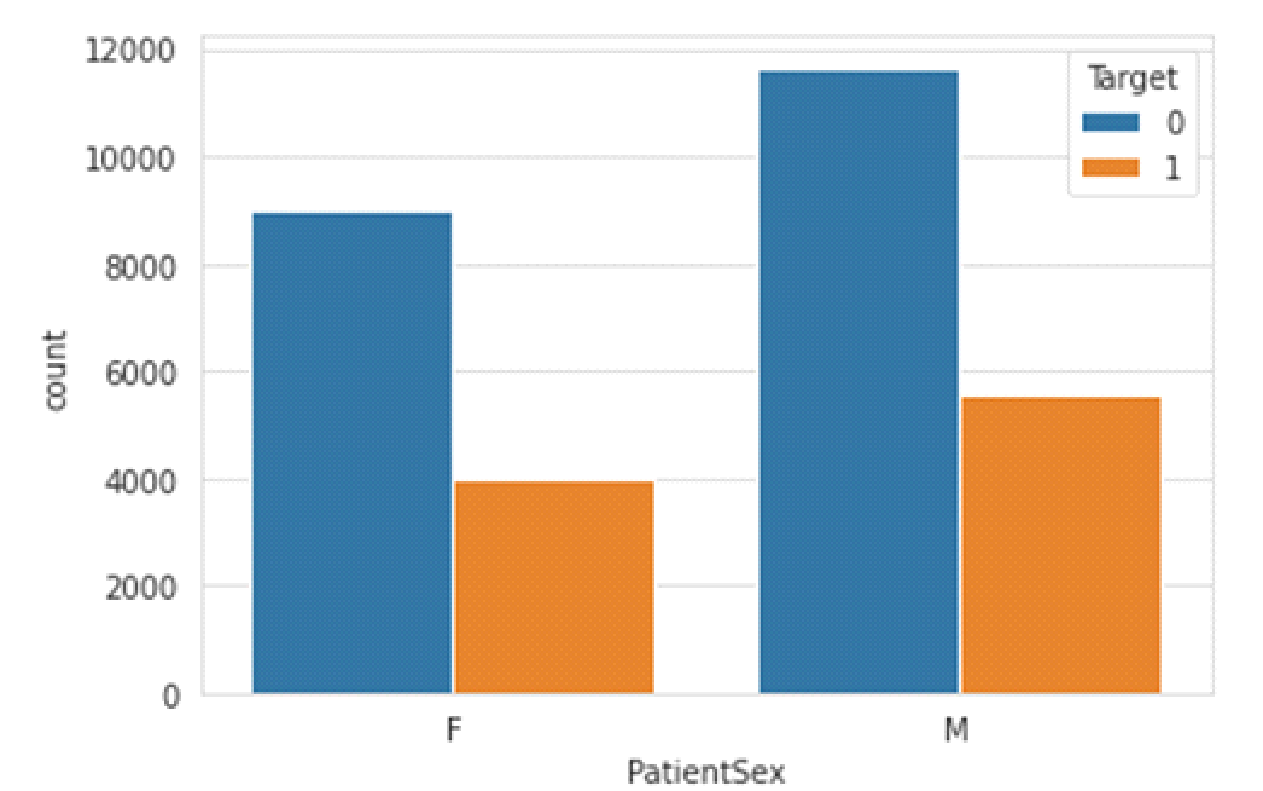
The same can be observed when plotting the count of Target values segregating the classes.



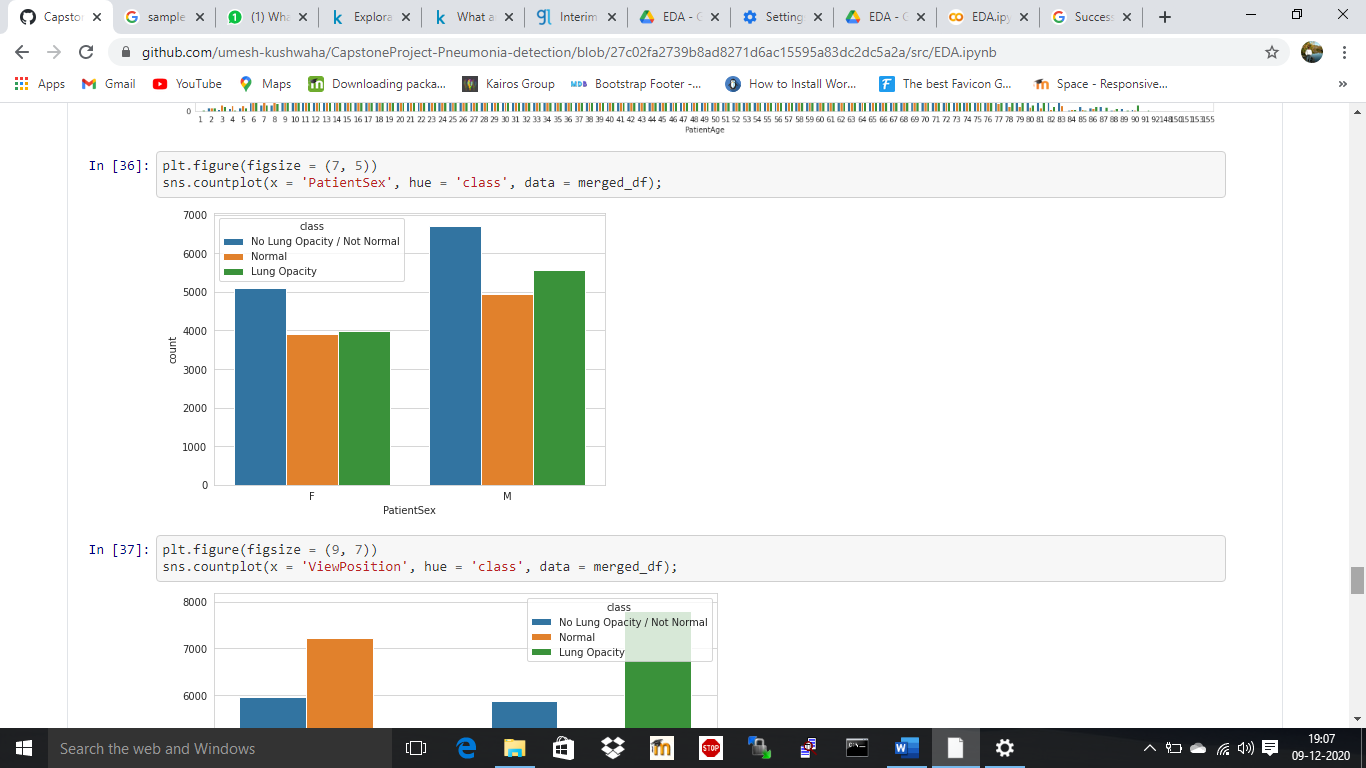
1. **Train set:** Out of 30,227 total images, 26,684 images are available in the training set and all 26,684 images are unique (equal to unique patient IDs).
2. **Bounding box:** Out of 26,684 images available in the training set, 23,286 images have only 1 bounding box, 3,266 images have 2 bounding boxes, 119 images have 3 boxes and 13 images have 4 boxes. Hence, 3,398 patients have more than 1 bounding box. If any given patient may potentially have many boxes, only if there are several different suspicious areas of pneumonia.

|  |  |
| --- | --- |
| **No of Occurrences** | **Count of the Patient ID** |
| 1 | 23286 |
| 2 | 3266 |
| 3 | 119 |
| 4 | 13 |

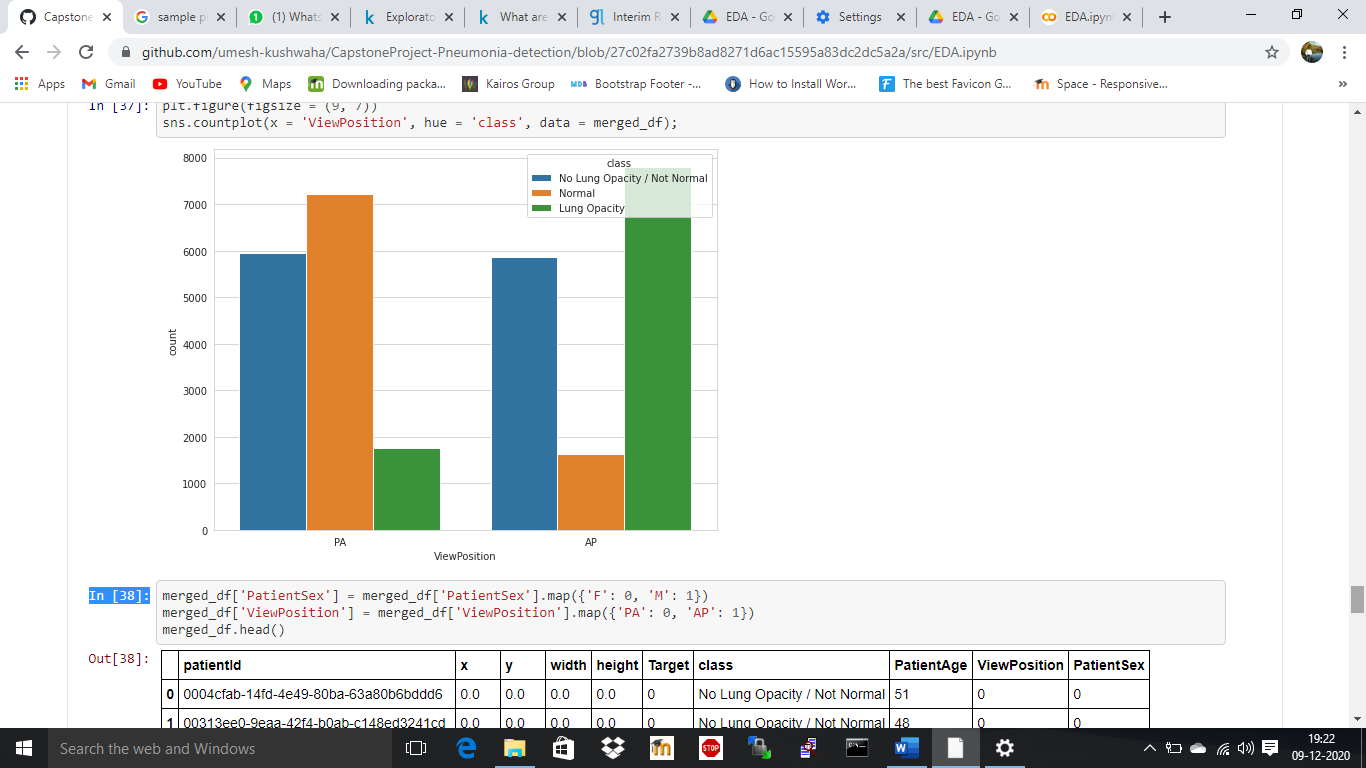
1. **Data characteristics:** We have different parameters or characteristics of available information – patient age, sex, body part examined, view position, rows and columns, pixel spacing, etc.
2. **Correlation:** We have observed that ‘Target’ and ‘View Position’ have a higher correlation and stands at 0.42.
3. **Gender and view position mix:** Out of total 9,555 cases of Lung opacity, ~60% is male and rest 40% is female. Out of total, ~81% cases have been identified from AP view.



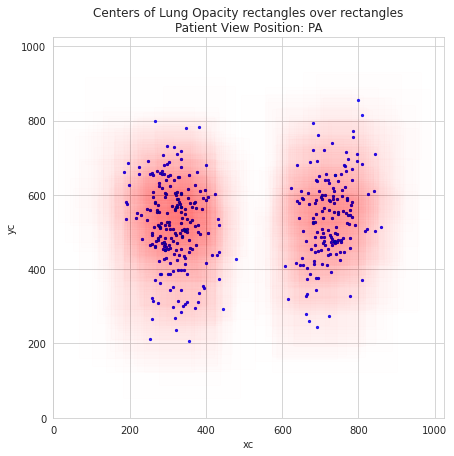
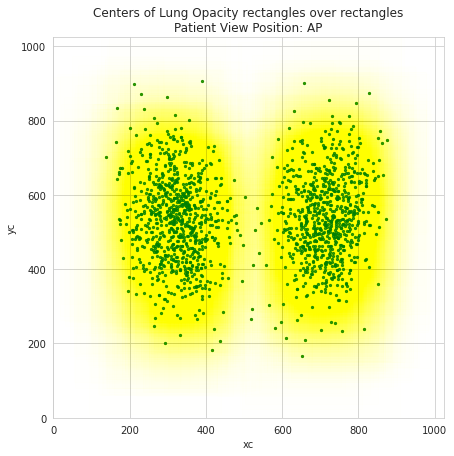
On doing further analysis with the classes, we can see that for both male and female have almost equal numbers of Normal and Lung opacity cases, whereas the Not Normal cases are more indicating that there are patients with other lung related illness other than pneumonia.



The PA position has very minimal pneumonia infection when compared to the AP position. The normal cases are more in PA view position and at the same time the Not normal cases are distributed equally between AP and PA position.

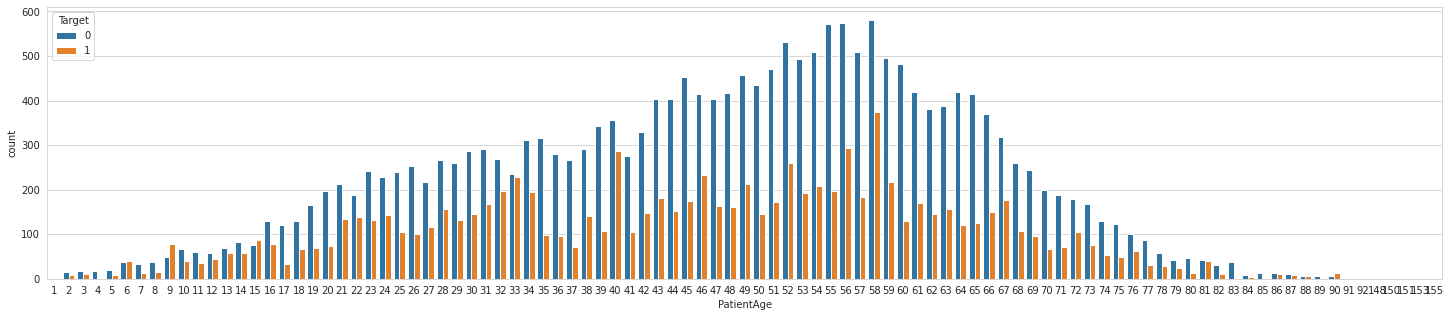


Further for analysis purpose, a scatter plot is for the Pneumonia persons for different views as below. The plot indicates that the bounding boxes for AP are concentrated in the middle, whereas for the PA cases, its scattered all over the lungs.

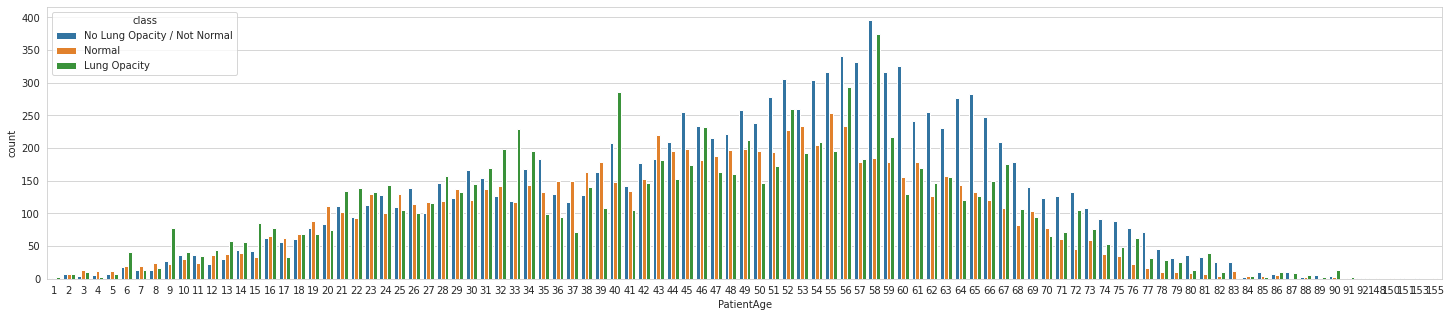


1. **Other observations:** Pneumonia cases are higher between the age group of 40-60 years.

Plotting the count of patient age with Hue as the target values, we observe that the infected patients spread across the ages. Most of the patients age between 40 to 60 years.

****

The same is done with the class values and the patient age mentioned below, which indicates that the lung opacities are spread across the patient age and have peaks within age 58 to 59 years. Similarly, the Not normal cases are found mostly in age between 40 to 60 years.



**Pre-processing –**

1. **Image scale:** The images have been windowed and leveled already, as have been rescaled to 8-bit encoding and the resolution has been rescaled to (1024, 1024).
2. **Data imbalance:** There is no discrepancy in the data, as the data in the class csv and label csv is same.
3. **xxxx:** Cxxxxxxxa
4. **Deciding Models and Model Building**

**VGG**

Xxxxx

**Mask RCNN**

xXxx

1. **How to improve your model performance?**

Xxxxxxxxxxxxxx

Xxxxxxxxx

Xxxxxxxx

**Future course of action – in the final report**

Xxxxx

Xxxxxxxxxx

Screenshot

xXxx

Xxxxxx

Xxxxx

1. **xxxxxxxxxxxx**

**Appendix**

Code

Xxxxx

Xxxxxxxxxx

Xxxxxxxxx

Xxxxxxx