**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

**Deliverables:**

1. **Summary of problem statement, data and findings**
   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. **Dataset:** In class dataset, information is given about the positive or negative class associated with a particular patient ID. In the train dataset, information is given about the bounding box (x, y, w, h) comprising evidence of pneumonia.

The data is organized in several folders,

1. **stage\_2\_train\_images -** 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.
2. **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)
3. **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*
4. **stage\_2\_test\_images -** Contains set of raw medical images (DICOM files) for testing the model. The files contain a combination of header metadata as well as underlying raw image arrays for pixel data
   1. **Findings:** xxxxxxx
   2. **Xxxxxxxxx**
   3. xxxxxxxxxxxxxx

Code

Xxxxx

Xxxxxxxxxx

Xxxxxxxxx

Xxxxxxx

Screenshot

xXxx

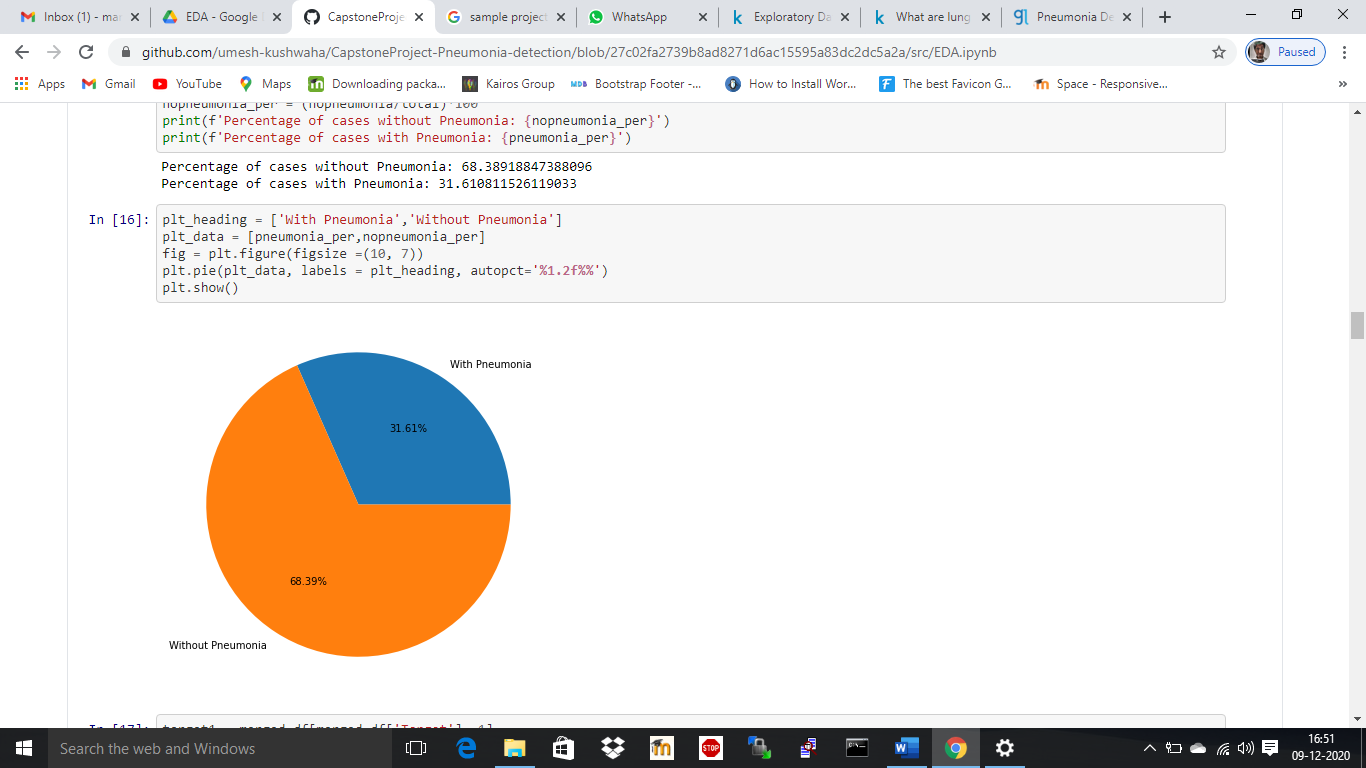
1. **Summary of the Approach to EDA and Pre-processing**
   1. **Exploratory data analysis (EDA) –**
      1. **Data analysis on the given csv files**

Out of the total class information of 30,227 patients, 9555 are identified as having pneumonia which contributes around 31.61% of the total data and 20672 has been identified as not infected which is 68.39%. This is as indicated below

Total No of samples: 30227

No of cases without Pneumonia: 20672

No of cases with Pneumonia: 9555

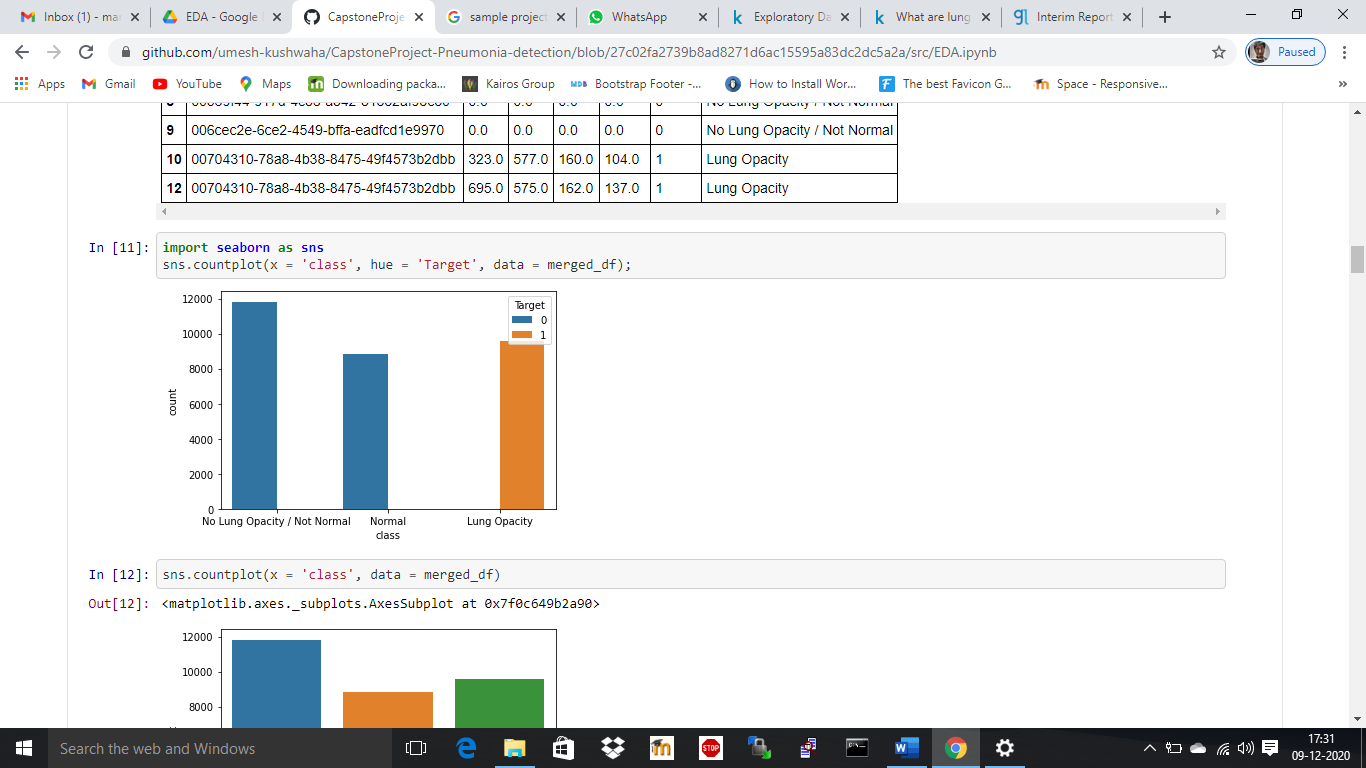


There are no Missing data in both the Class and Target dataset as shown below

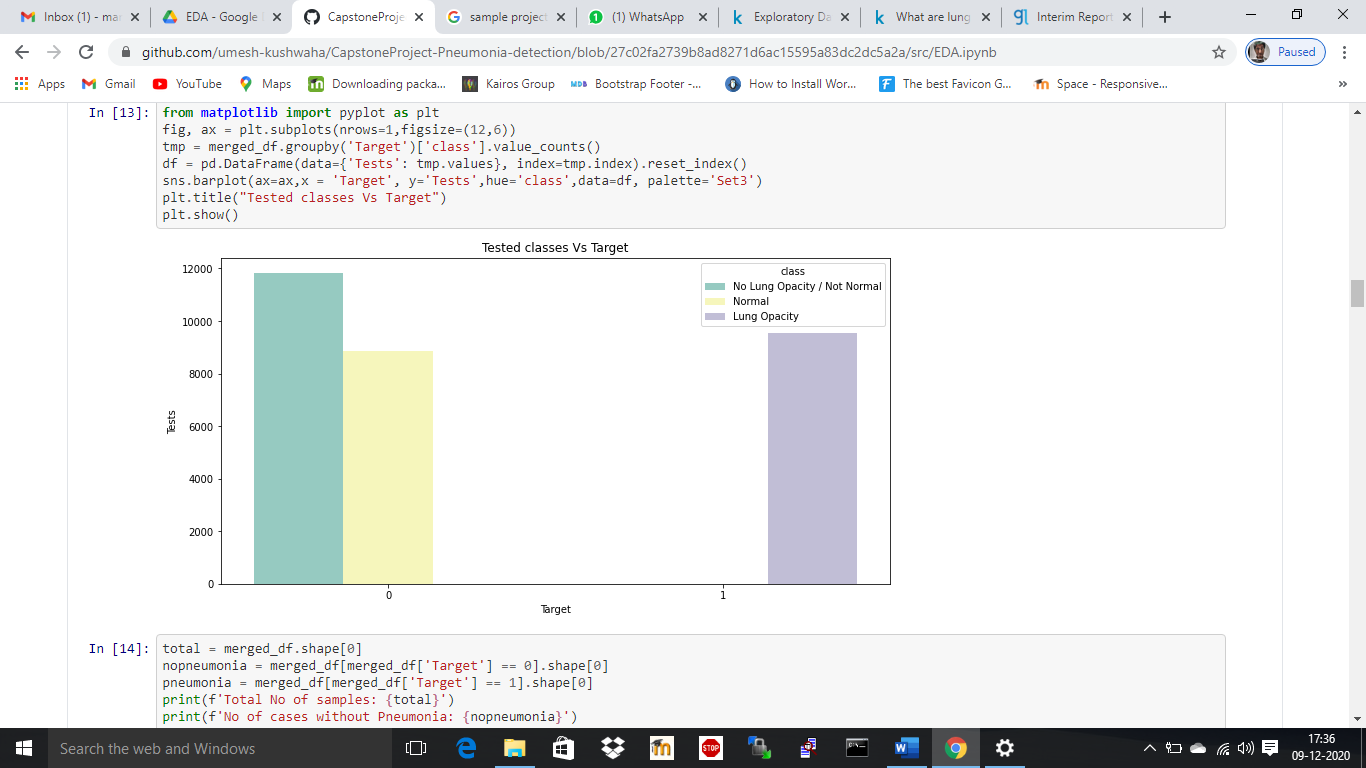
|  | **height** | **width** | **y** | **X** | **Target** | **patientId** |
| --- | --- | --- | --- | --- | --- | --- |
| **Total** | 20672.00 | 20672.00 | 20672.00 | 20672.00 | 0.0 | 0.0 |

|  | **class** | **patientId** |
| --- | --- | --- |
| **Total** | 0.0 | 0.0 |

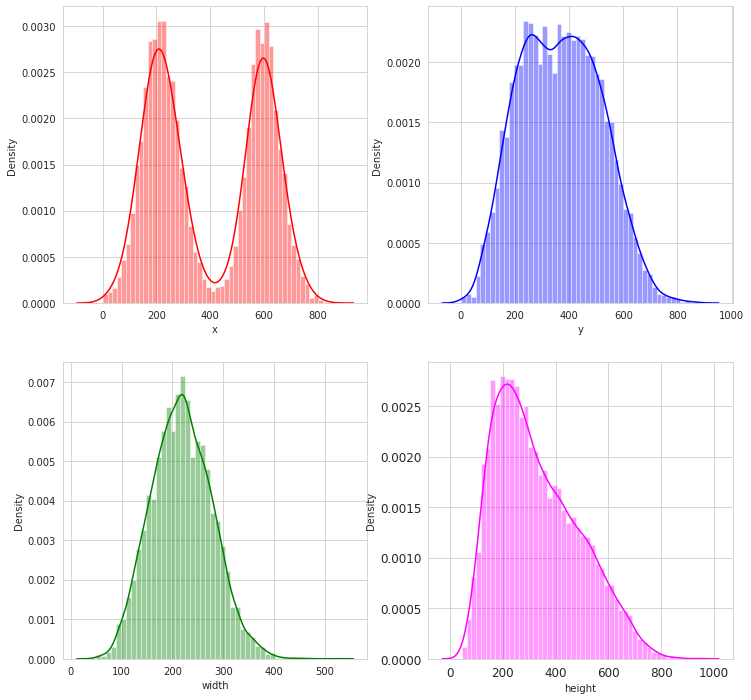
On plotting the class values and it corresponding counts segregated with the Target values, we observe that there are count of patients with No Lung opacities but are not normal is high in count than the Pneumonic or normal patients. Also, the count of normal class is less than other 2 classes indicating that the data has more number of Ill health patients



The same can be observed when plotting the count of Target values segregating the classes as below/

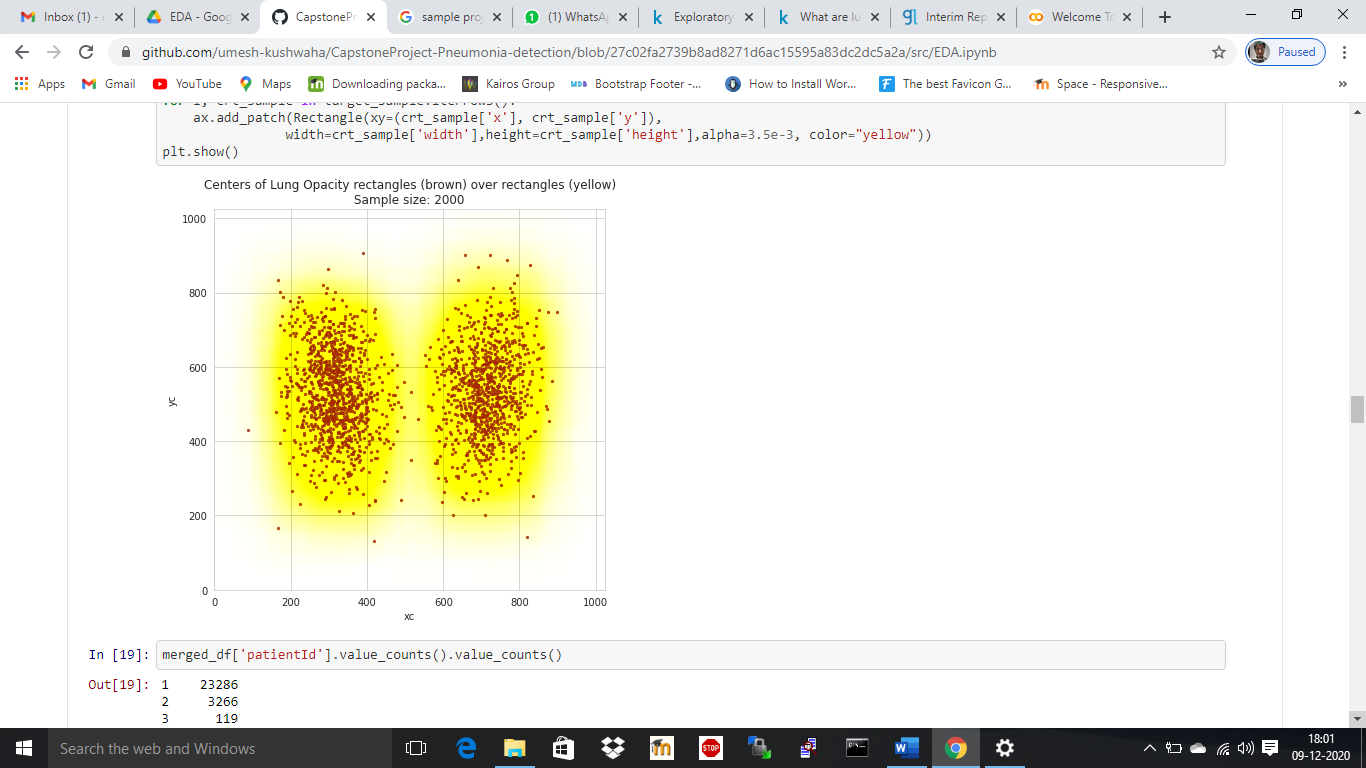


To analyze the distribution of the X, Y, Width and Height data, a plot of distribution is done for these variables and the below is observed



* The X value has a typical Binomial distribution indicating that there are typically two bounding boxes in the left and right lungs for the patients X-ray images.
* The Y value has a slight binomial distribution indicating that the boundary boxes spread across the images concentrating at the middle part of left and right Lungs
* The width of the boundary boxes has a Normal distribution indicating that most of the images has width between the range 150 to 300
* The height has a skewed distribution again falling in the same range as width. Most of the boxes falls between the range of 150 to 300

To understand more on the distribution of the X and Y, centers of the bounding boxes can be found by averaging the X and width values and Y and Height values and plotting them as a scatter plot



The plot indicates that the opacities are distributed mostly in the middle portion of both Right side and left side of the lungs.

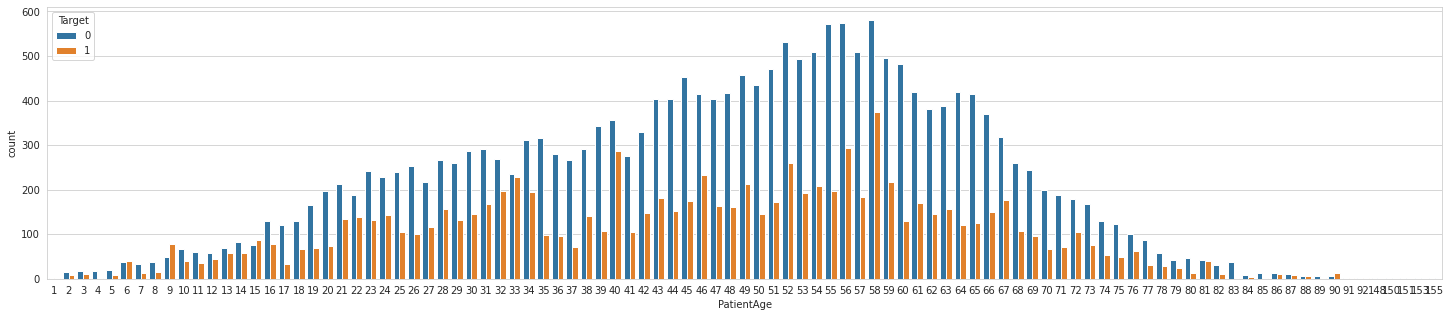
On checking the number of occurrences for each patient id, we observe the below

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

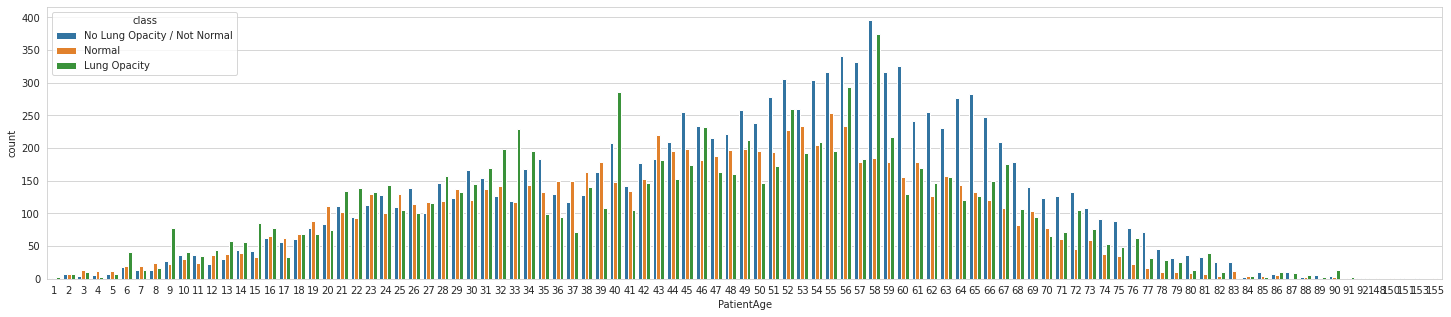
* There are 13 Patient id which has 4 occurrences. This indicates that there are 4 bounding boxes for the patient
* 119 patients have 3 bounding boxes and 3266 patients have 2 bounding boxes.
* 23286 patients have either 1 or no bounding boxes.
  + 1. **Data analysis on the Meta data of the DICOM Image files**

From the Image files Meta data, Patient Age, the Image View Position and Patient sex are the useful information which are extracted and merged with the CSV data. Performing further analysis with these data

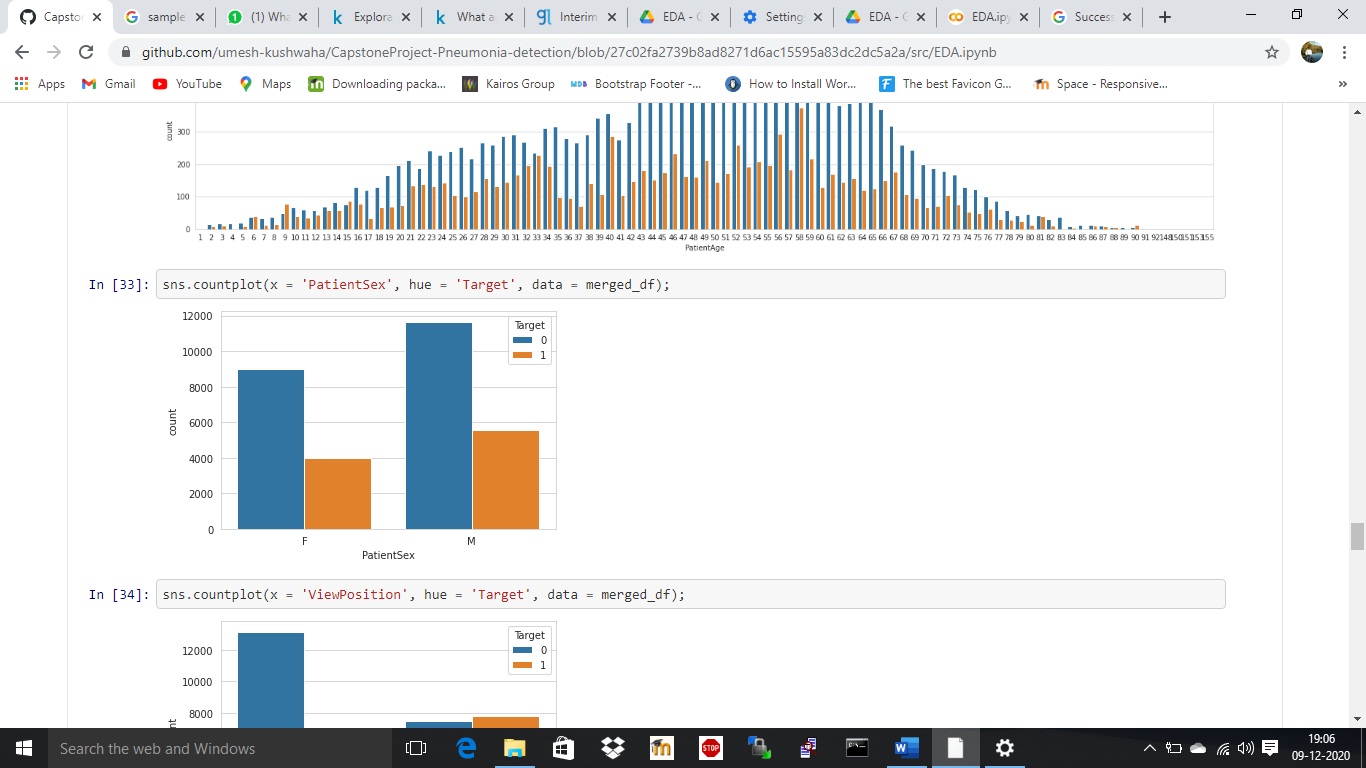
Plotting the count of patient age with Hue as the target values, we observe that the infected patients spread across all the ages. Most of the patient has ages around 40 to 60.

****

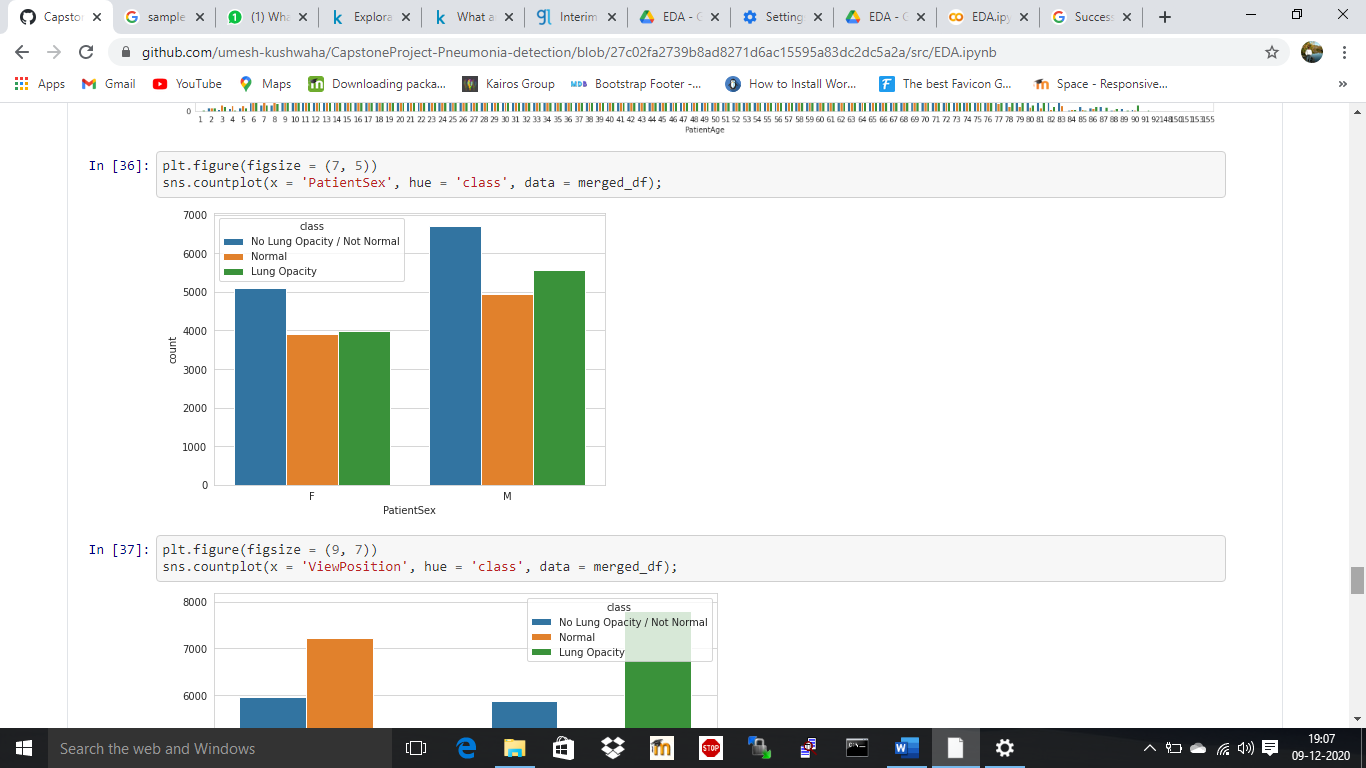
The same is done with the class values and the Patient age as below which indicates the Lung opacities are spread across the patient age and has peeks within the ages 58 to 59. Similarly, the Not normal cases are found mostly in ages between 40 to 60



The same analysis is done for the Patient Sex and found that data provided has more number of Male patient than the female patient with more infected cases in Male category



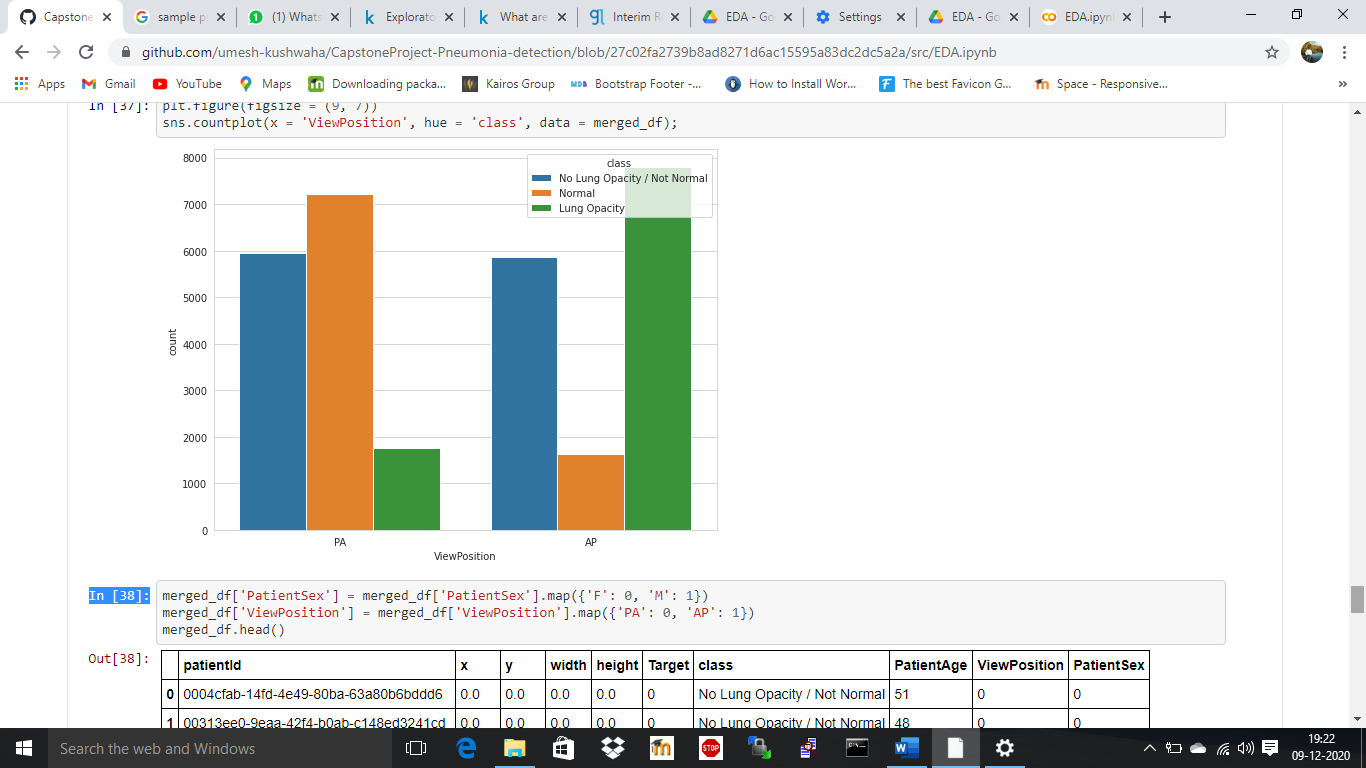
On doing further analysis with the classes, we can see that for both Male and Female has almost equal number of Normal and Lung opacity cases where as the Not Normal cases are more indicating that there are patients with other lung related illness other than Pneumonia.



Another important feature from the meta data is the view position of the patient. There are 2 different view positions Anterior-Posterior (AP) and Posterior-Anterior (PA). PA indicates that the X-Ray is taken when the patient is in standing position. Sometimes it is not possible for radiographers to acquire a PA chest X-ray. So, the X-Rays are taken in AP view position. This is usually because the patient is too unwell to stand. Plotting the given data with the view positions to get the details further

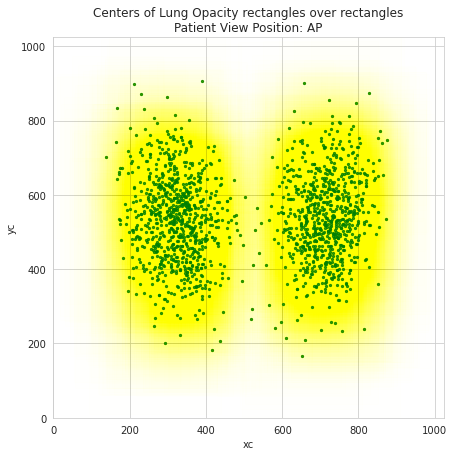


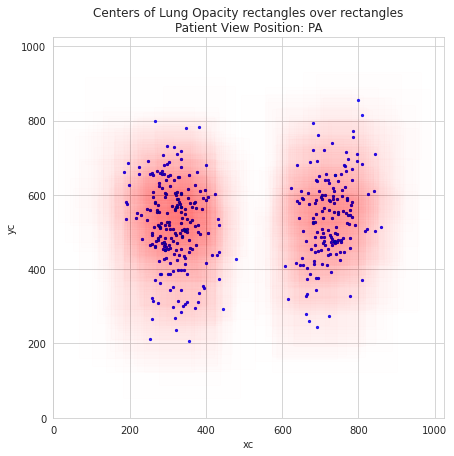
As expected, the patient with AP view position has been observed with Pneumonia. Plotting further with classes,



The PA position has very minimal pneumonia infection when compared to the AP position. The normal cases are more in PA view position 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 view as below. The plot indicates that the bounding boxes for AP are concentrated in the middle where are for the PA cases, its scattered all over the lungs.





1. **Deciding Models and Model Building**

Xxxxxxxxxxxxxx

Xxxxxxxxxxx

Xxxxxxxxxx

Code

Xxxxx

Xxxxxxxxxx

Xxxxxxxxx

Xxxxxxx

Screenshot

xXxx

Xxxxxxxxxx

Xxxxxxxxx

Xxxxxx

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

Xxxxxxxxxxxxxx

Xxxxxxxxx

Xxxxxxxx

Code

Xxxxx

Xxxxxxxxxx

Xxxxxxxxx

Xxxxxxx

Screenshot

xXxx

Xxxxxx

Xxxxx

1. **xxxxxxxxxxxx**