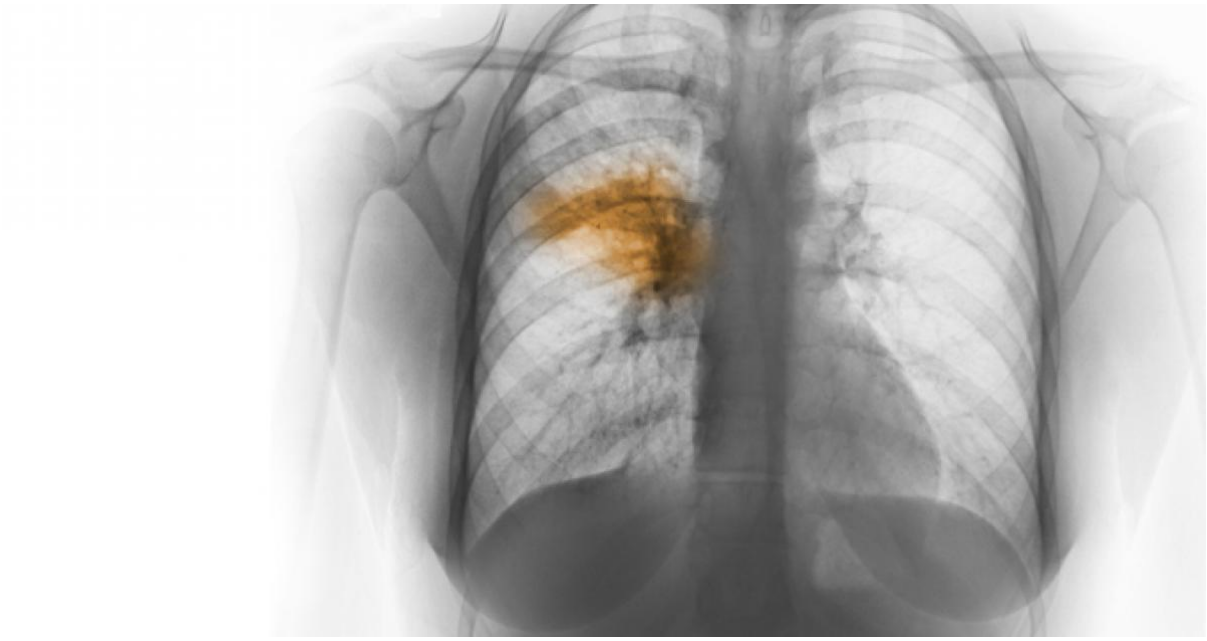
PROBLEM STATEMENT

In medicine, the next frontier for AI is anomaly localization in medical imaging. Localization of anomalies refers to both predicting anomalies and their boundaries. Automatic detection algorithms to locate the position of inflammation in an image can help the physicians in making the better clinical decisions. In this project we analyze data with the knowledge of EDA, we build a detection model and we present our findings based on the evaluations with the RSNA Pneumonia Detection Challenge dataset.

OVERVIEW OF PNEUMONIA

Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are comprised of small sacs called alveoli that, as a healthy person breathes, fill up with oxygen. The alveoli are filled with pus and fluid when a person has pneumonia, which makes breathing difficult and reduces oxygen intake.

The single greatest bacterial cause of death in children worldwide is pneumonia. In 2017, pneumonia killed 808,694 children under the age of 5, accounting for 15 percent of all deaths by children under the age of five. Children and families worldwide are afflicted by pneumonia, but it is most common in South Asia and sub-Saharan Africa. It can be avoided with easy procedures, and managed with low-cost, low-tech treatment and care. Children can be shielded from pneumonia.



In 2015 spending for maternal, infant, and child survival, the cost of antibiotic care for all children with pneumonia is estimated at about US$ 109 million per year among 66 countries. The expense requires antibiotics and diagnostics for the treatment of pneumonia.

CAUSES AND TRANSMISSION

According to WHO, pneumonia is caused by a number of infectious agents, including viruses, bacteria and fungi. The most common are:

- Streptococcus pneumoniae – the most common cause of bacterial pneumonia in children;
- Haemophilus influenzae type b (Hib) – the second most common cause of bacterial pneumonia;
- respiratory syncytial virus is the most common viral cause of pneumonia;
- in infants infected with HIV, Pneumocystis jiroveci is one of the most common causes of pneumonia, responsible for at least one quarter of all pneumonia deaths in HIV-infected infants.

Pneumonia can be spread in a number of ways. The viruses and bacteria that are commonly found in a child's nose or throat, can infect the lungs if they are inhaled. They may also spread via air-borne droplets from a cough or sneeze. In addition, pneumonia may spread through blood, especially during and shortly after birth. More research needs to be done on the different pathogens causing pneumonia and the ways they are transmitted, as this is of critical importance for treatment and prevention.

TREATMENT AND PREVENTION

Pneumonia should be treated with antibiotics. Amoxicillin-dispersible tablets are the antibiotic of choice. In most cases of pneumonia, oral antibiotics are needed, which are mostly administered at a health clinic. These cases may also be diagnosed and treated at the neighborhood level by qualified community health professionals with affordable oral antibiotics. Only for serious cases of pneumonia is hospitalization recommended.

DIAGNOSTIC PROCEDURE

Doctor will diagnose pneumonia based on your medical history, a physical exam, and test results. Sometimes pneumonia is hard to diagnose because symptoms may be the same as a cold or flu. Patient may not realize that his/her condition is more serious until it lasts longer than these other conditions.

If doctor thinks patient may have pneumonia, he or she may do one or more of the following tests.

- Chest X-ray to look for inflammation in patient’s lungs. A chest X-ray is often used to diagnose pneumonia.
- Blood tests, such as a complete blood count (CBC) to see whether patient’s immune system is fighting an infection.
- Pulse oximetry to measure how much oxygen is in his/her blood. Pneumonia can keep patient’s lungs from moving enough oxygen into his/her blood. To measure the levels, a small sensor called a pulse oximeter is attached to patient’s finger or ear.

DATA AND FINDINGS

In 2018, RSNA organized an AI challenge to detect pneumonia, one of the leading causes of mortality worldwide, as part of its efforts to help improve artificial intelligence (AI) instruments for radiology. RSNA Pneumonia dataset consists of 29684 thousand images. All the images are in dicom format. There are 26684 images for training and 3000 images for testing.

Dicom images: In a special format called DICOM files (*. dcm), medical images are stored. They contain a mix of header metadata as well as pixel data underlying raw image arrays.

There are three classes featured in the dataset Normal, Not normal/No opacity and Lung opacity. Normal class indicates there is no anomaly in the lungs. Not normal/No opacity indicates while it was decided that pneumonia was not present, there was still some sort of picture abnormality and sometimes this finding could mimic the appearance of true pneumonia. Lung opacity class indicates there is definite pneumonia in the lungs. These three classes are divided as two target variables 0 and 1 the images with lung opacity comes under target 1 and 0 is assigned to other two classes.

Along with the images two csv files are provided. Detailed class info file consists of image name and the class it belonged to. Train labels file consists of the bounding box coordinates belonging to each image. Bounding box coordinates are given in the following format as follows,

- x -- the upper-left x coordinate of the bounding box.
- y -- the upper-left y coordinate of the bounding box.
- width -- the width of the bounding box.
- height -- the height of the bounding box.

With these bounding box coordinates target column is provided which discriminates classes into categories of 0 and 1. All the images in the dataset belongs to stage 2 pneumonia.

EDA AND PREPROCESSING

There are two datasets each dataset has total 30227 number of rows and the trian_labels.csv has 6 columns but in class_info.csv has 2 columns.

```
print("Bounding Box Shape:", bbox_df.shape)
print("Class info Shape:", class_df.shape)
```

```
Bounding Box Shape: (30227, 6)
Class info Shape: (30227, 2)
```

As we can see below, in train_labels.csv file contains patientId which is unique value per patient, for each patientId, it has one target column and 4 values which are corresponding abnormality bounding box defined by the upper-left hand corner ‘x’ and ‘y’ coordinate and its

corresponding width and height. Target column has two values 0 and 1. 0 is for No Lung Opacity / Not Normal, Normal and 1 is for Lung opacity.

	patientId	x	y	width	height	Target
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	NaN	NaN	NaN	NaN	0
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	NaN	NaN	NaN	NaN	0
2	00322d4d-1c29-4943-afc9-b6754be640eb	NaN	NaN	NaN	NaN	0
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	NaN	NaN	NaN	NaN	0
4	00436515-870c-4b36-a041-de91049b9ab4	264.0	152.0	213.0	379.0	1

In the class_info.csv file there are two columns patientId and class column that describes the three conditions of lungs which are Normal, Not Normal/No Lung Opacity and lung opacity. As we can see below

	patientId	class
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	No Lung Opacity / Not Normal
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	No Lung Opacity / Not Normal
2	00322d4d-1c29-4943-afc9-b6754be640eb	No Lung Opacity / Not Normal
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	Normal
4	00436515-870c-4b36-a041-de91049b9ab4	Lung Opacity

We have calculated each value for 3 different class and we got only 23.5 percent patient are Normal and 76 percent Not Normal and Lung opacity.

```
[ ]  data['class'].value_counts()

Lung Opacity                  16957
No Lung Opacity / Not Normal  11821
Normal                         8851
Name: class, dtype: int64

data['class'].value_counts()*(100.0)/len(data.index)

Lung Opacity                  45.063648
No Lung Opacity / Not Normal  31.414600
Normal                        23.521752
Name: class, dtype: float64
```

Now when it comes to the bounding boxes dataset which has missing value in x, y, height and width column

Total number of missing values in each column are same that is 20672. As shown below

```
def missing_data(data):
  null_data = data.isnull().sum()
  num_rows = len(data.index)
  percent_null = 100.*null_data/num_rows
  return pd.concat([null_data, percent_null.round(1)], axis=1, keys=['Missing', 'PercentMissi

missing_data(data)
```

	Missing	PercentMissing
patientId	0	0.0
class	0	0.0
x	20672	54.9
y	20672	54.9
width	20672	54.9
height	20672	54.9
Target	0	0.0

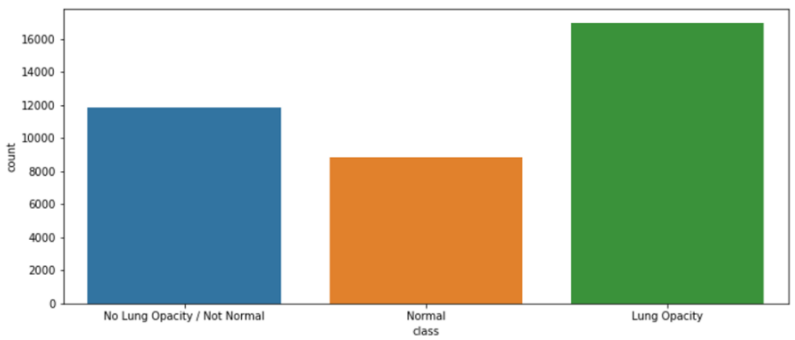
The total percentage of missing is 54.9 percent and also, we can see from below in the target column 0 means No Lung Opacity value is same with the missing values.

```
print("Target 0")
print("="*20)
print(data[data.Target == 0].count())
print("")
print("Target 1")
print("="*20)
print(data[data.Target == 1].count())
```

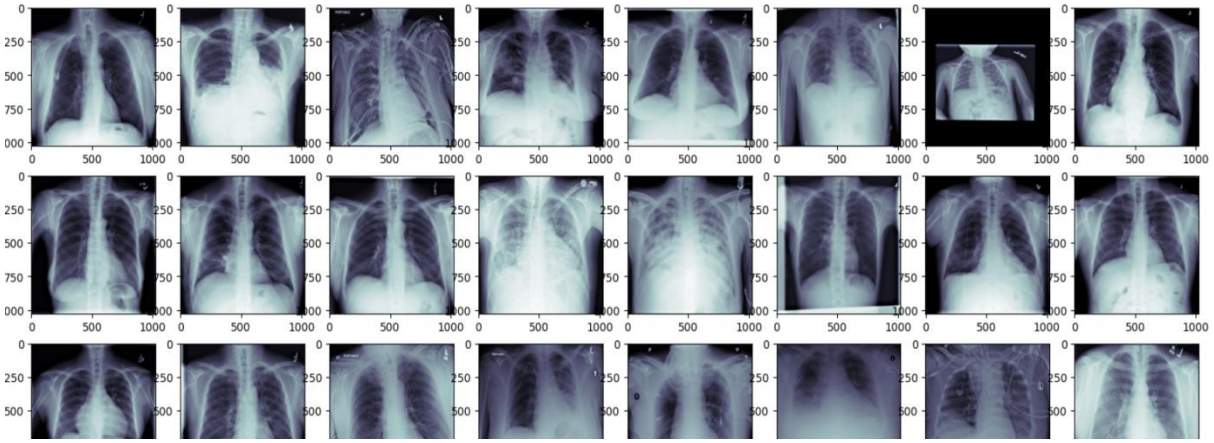
```
Target 0
====================
patientId    20672
class        20672
x            20672
y            20672
width        20672
height       20672
Target       20672
dtype: int64
```

So, the remaining 16957 are the positive means Lung Opacity case. Above bar graph of classes using count shows that.

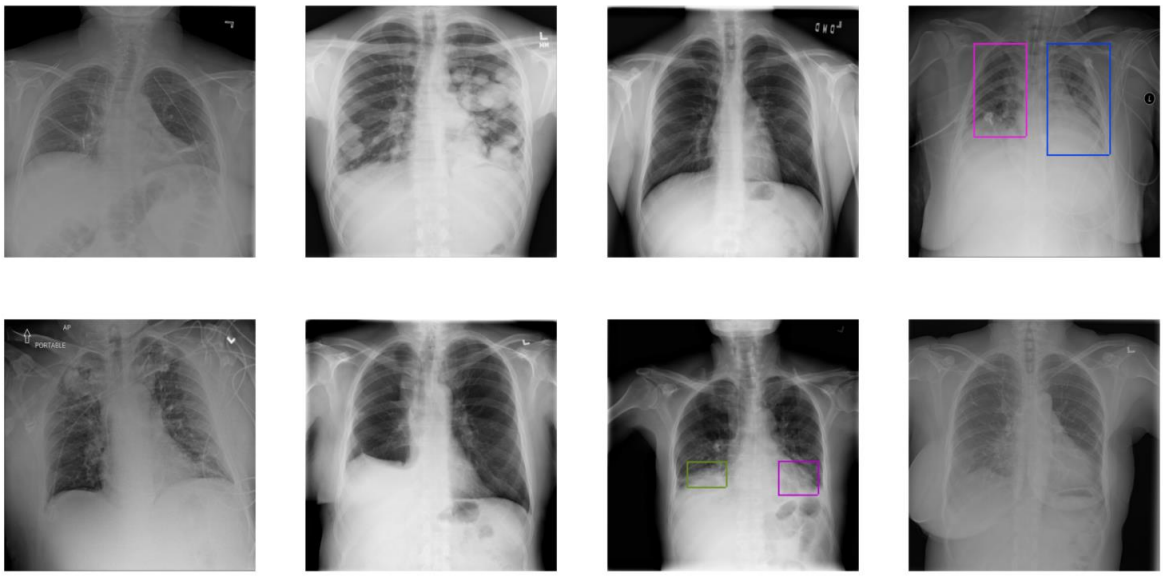
```
fig=plt.figure(figsize=(12, 5))
countplot = sns.countplot(data['class'])
```



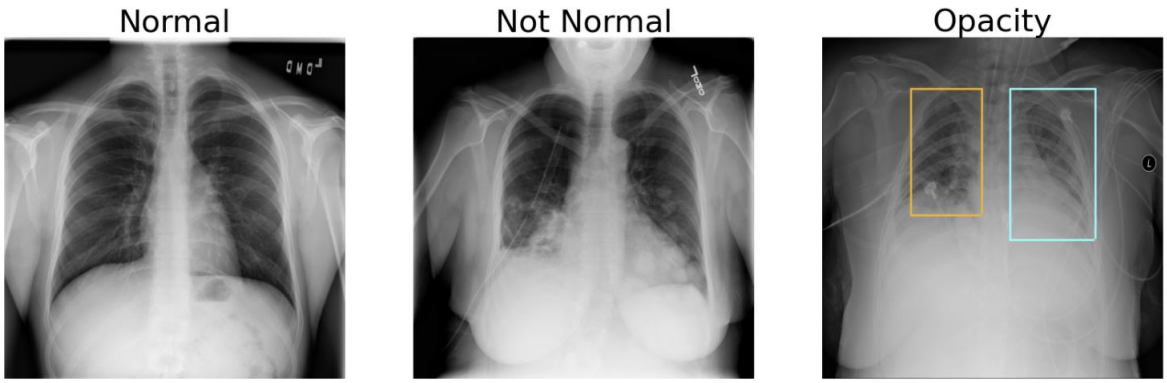
There are 26684 training images are there and 3000 test images. Visualizations of the few samples are shown below.



Now, the visualization of the images with the bounding box is presented below. From that we can see that there are images with different numbers of bounding boxes. And also, maximum images don't have bounding boxes.

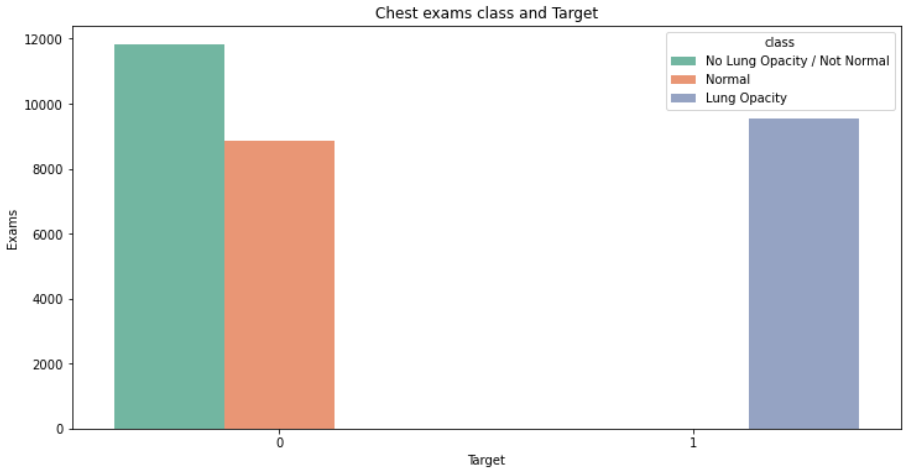


As we can determine from above visualizations task in hand is a regression problem. There is a need for building a feasible model that can regress the bounding box in the images. And there is an extra class which is Not normal/ No lung opacity. This class shows there is a anomaly in the lungs which can be easily misread as pneumonia. So, there is a need to examine that class little more briefly. The visualization which shows the all three classes together is shown below. And the bar plot of the target class which shows all the three class is shown below.



In [12]:

```
fig, ax = plt.subplots(nrows=1,figsize=(12,6))
tmp = comb_bounding_box_df.groupby('Target')['class'].value_counts()
df = pd.DataFrame(data={'Exams': tmp.values}, index=tmp.index).reset_index()
sns.barplot(ax=ax,x = 'Target', y='Exams',hue='class',data=df, palette='Set2')
plt.title("Chest exams class and Target")
plt.show()
```



There is patients and other information available in the meta data of the Dicom images. Visualizations of that data may give better understanding of the pneumonia disease itself.

[53]:

```
image_meta_df['class'] = detail_class_df['class']
image_meta_df['patientId'] = detail_class_df['patientId']
image_meta_df.head()
```

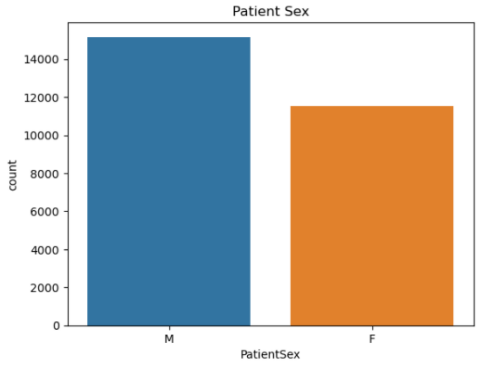
Out[53]:

	PatientAge	BodyPartExamined	ViewPosition	PatientSex	path	class	patientId
0	8	CHEST	PA	M	../input/rsna-pneumonia-detection-challenge/st...	No Lung Opacity / Not Normal	0004cfab-14fd-4e49-80ba-63a80b6bddd6
1	16	CHEST	AP	M	../input/rsna-pneumonia-detection-challenge/st...	No Lung Opacity / Not Normal	00313ee0-9eaa-42f4-b0ab-c148ed3241cd
2	62	CHEST	PA	M	../input/rsna-pneumonia-detection-challenge/st...	No Lung Opacity / Not Normal	00322d4d-1c29-4943-afc9-b6754be640eb
3	65	CHEST	AP	M	../input/rsna-pneumonia-detection-challenge/st...	Normal	003d8fa0-6bf1-40ed-b54c-ac657f8495c5
4	26	CHEST	AP	F	../input/rsna-pneumonia-detection-challenge/st...	Lung Opacity	00436515-870c-4b36-a041-de91049b9ab4

Above picture shows the data frame of the extracted dicom data.

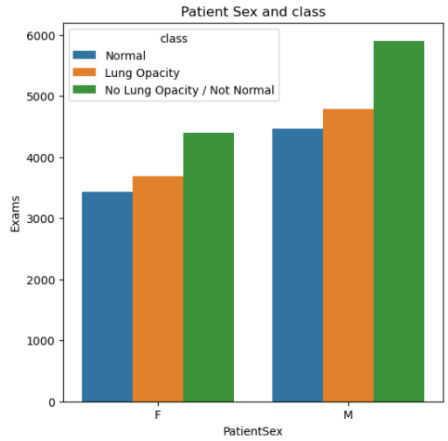
As we can see that there is gender of the patients, we can explore the diversity in the given data.

[56]:
```
sns.countplot(image_meta_df['PatientSex'])
plt.title("Patient Sex")
plt.show()
```



Now with this information we can visualize class diversity between male and female.

[55]:
```
tmp = image_meta_df.groupby(['class', 'PatientSex'])['patientId'].count()
df1 = pd.DataFrame(data={'Exams': tmp.values}, index=tmp.index).reset_index()
tmp = df1.groupby(['Exams','class', 'PatientSex']).count()
df3 = pd.DataFrame(data=tmp.values, index=tmp.index).reset_index()
fig, (ax) = plt.subplots(nrows=1,figsize=(6,6))
sns.barplot(ax=ax, x = 'PatientSex', y='Exams', hue='class',data=df3)
plt.title("Patient Sex and class")
plt.show()
```



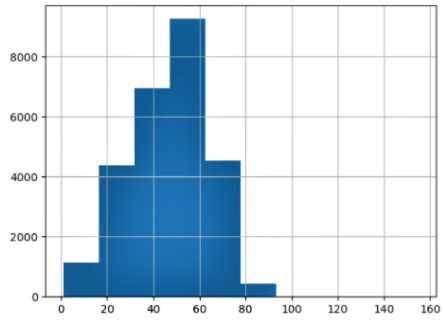
From the above images we can clearly infer that dataset has more male examples than the female examples. In this case there are more men with pneumonia around 4800 when compared around 3300 women has pneumonia. The abnormality in the lungs is very high in the men compared to women. There is a huge difference in Not normal/ No lung opacity class in the between male and female. This shows men tend to have abnormality in lungs more often than the women.

There are other columns which can be explored more in the meta data. The Visualization that summarizes useful meta data such as age, gender, body part examined and view position of the machine can be seen in below picture.

[52]:
```
DCM_TAG_LIST = ['PatientAge', 'BodyPartExamined', 'ViewPosition', 'PatientSex']
def get_tags(in_path):
    c_dicom = pydicom.read_file(in_path, stop_before_pixels=False)
    tag_dict = {c_tag: getattr(c_dicom, c_tag, '')
          for c_tag in DCM_TAG_LIST}
    tag_dict['path'] = in_path
    return pd.Series(tag_dict)
image_meta_df = image_df.apply(lambda x: get_tags(x['path']), 1)
# show the summary
image_meta_df['PatientAge'] = image_meta_df['PatientAge'].map(int)
image_meta_df['PatientAge'].hist()
image_meta_df.drop('path',1).describe(exclude=np.number)
```

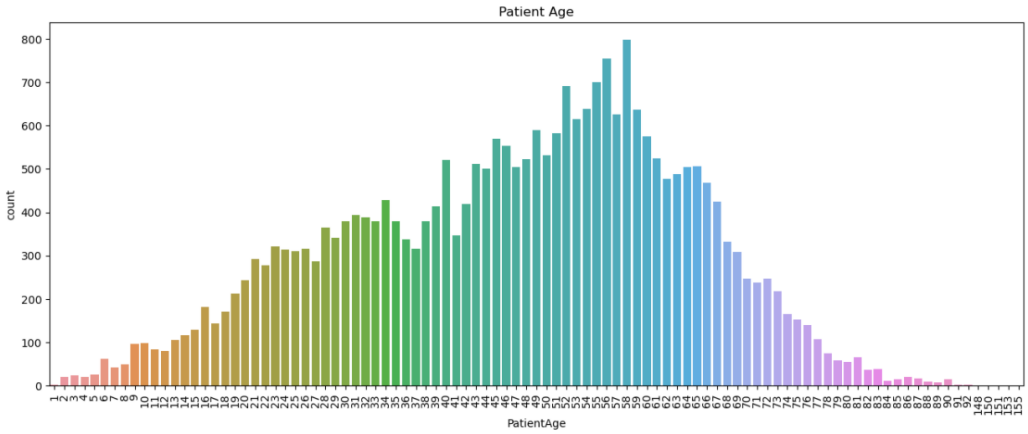
Out[52]:

	BodyPartExamined	ViewPosition	PatientSex
count	26684	26684	26684
unique	1	2	2
top	CHEST	PA	M
freq	26684	14511	15166



Above picture gives the histogram of the age distribution, we can see most of the patients are above 40years and below 65 years. Detailed plot is examined below.

[54]:
```
fig, (ax) = plt.subplots(nrows=1,figsize=(16,6))
sns.countplot(image_meta_df['PatientAge'], ax=ax)
plt.title('Patient Age')
plt.xticks(rotation=90)
plt.show()
```

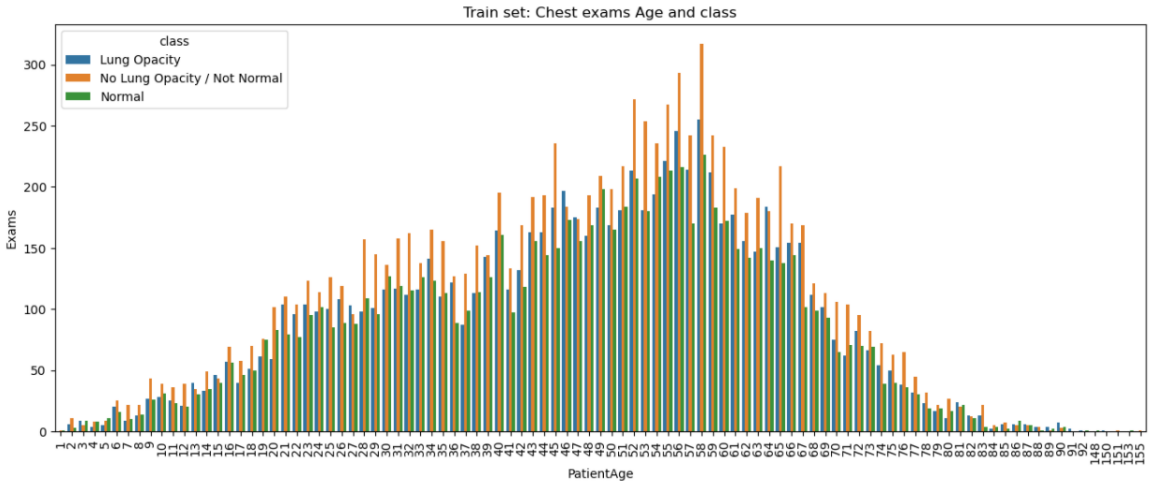


From the above visualization we can clearly see that patients of age between 50 and 60 are more effect with the dominant being age 58. This clearly shows that pneumonia becomes more prominent at the age above 50. And we can also see that the dataset that is provided consists of people aged between 40 and 65.

We can explore a bit further by comparing Age with the class. The visualization that summarizes that relationship is plotted in below picture.

```
tmp = image_meta_df.groupby(['class', 'PatientAge'])['patientId'].count()
df1 = pd.DataFrame(data={'Exams': tmp.values}, index=tmp.index).reset_index()
tmp = df1.groupby(['Exams','class', 'PatientAge']).count()
df3 = pd.DataFrame(data=tmp.values, index=tmp.index).reset_index()


fig, (ax) = plt.subplots(nrows=1,figsize=(16,6))
sns.barplot(ax=ax, x = 'PatientAge', y='Exams', hue='class',data=df3)
plt.title("Train set: Chest exams Age and class")
plt.xticks(rotation=90)
plt.show()
```



From the above plot we can infer that Abnormalities in the lungs are more prominent in the people of age between 45 and 65. So of no surprise the lung opacities are maximum in that age group.

To summarize the Exploratory data analysis. We can say we are dealing with class imbalance as the maximum images don't have the bounding boxes. And we can say that there is a need for the bounding box regressor model. The metadata provided can give insight to the disease as whole. The visualizations of the both original data and dicom meta data is extremely useful for the stakeholders and also model evaluation.