Chest X-ray AI Project Update

Progress update on our AI initiative for lung disease detection

What We Are Doing

- Our goal is to leverage AI for **faster diagnosis** of lung diseases from X-ray images, enhancing support for physicians in their decision-making processes.
- By implementing advanced algorithms, we aim to significantly reduce diagnosis time, thereby improving patient outcomes and streamlining healthcare workflows.

Dataset Overview and Updates

Our dataset now includes disease labels and bounding boxes, with a new area column to enhance analysis. This will improve accuracy in detecting lung diseases from X-ray images.

```
Shape: (984, 9)
       Image Index Finding Label     Bbox [x              y              w  \
0  00013118_008.png    Atelectasis  225.084746     547.019217      86.779661
1  00014716_007.png    Atelectasis  686.101695     131.543498     185.491525
2  00029817_009.png    Atelectasis  221.830508     317.053115     155.118644
3  00014687_001.png    Atelectasis  726.237288     494.951420     141.016949
4  00017877_001.png    Atelectasis  660.067797     569.780787     200.677966


          h]   Unnamed: 6   Unnamed: 7   Unnamed: 8
0   79.186441         NaN          NaN          NaN
1  313.491525         NaN          NaN          NaN
2  216.949153         NaN          NaN          NaN
3   55.322034         NaN          NaN          NaN
4   78.101695         NaN          NaN          NaN
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 984 entries, 0 to 983
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Image Index    984 non-null    object
 1   Finding Label  984 non-null    object
 2   Bbox [x        984 non-null    float64
 3   y              984 non-null    float64
 4   w              984 non-null    float64
 5   h]             984 non-null    float64
 6   Unnamed: 6     0 non-null      float64
 7   Unnamed: 7     0 non-null      float64
 8   Unnamed: 8     0 non-null      float64
dtypes: float64(7), object(2)
memory usage: 69.3+ KB
None
            Bbox [x           y           w          h]  Unnamed: 6  Unnamed: 7
count  984.000000  984.000000  984.000000  984.000000         0.0         0.0
mean   398.806111  405.425364  256.334708  252.302547         NaN         NaN
std    222.700868  166.309995  167.629620  159.443635         NaN         NaN
min      5.417989   12.837934   27.306667   21.617778         NaN         NaN
25%    203.093333  293.869045  136.533333  115.674074         NaN         NaN
50%    340.249735  412.850794  214.340942  216.949153         NaN         NaN
75%    607.959365  521.641995  311.832381  367.902430         NaN         NaN
max    905.887831  876.980783  901.120000  873.379894         NaN         NaN

       Unnamed: 8
count         0.0
mean          NaN
std           NaN
min           NaN
25%           NaN
50%           NaN
75%           NaN
max           NaN
Unique images: 880
Unique diseases: 8
Finding Label
Atelectasis     180
Effusion        153
Cardiomegaly    146
Infiltrate      123
Pneumonia       120
Pneumothorax     98
Mass             85
Nodule           79
Name: count, dtype: int64
```
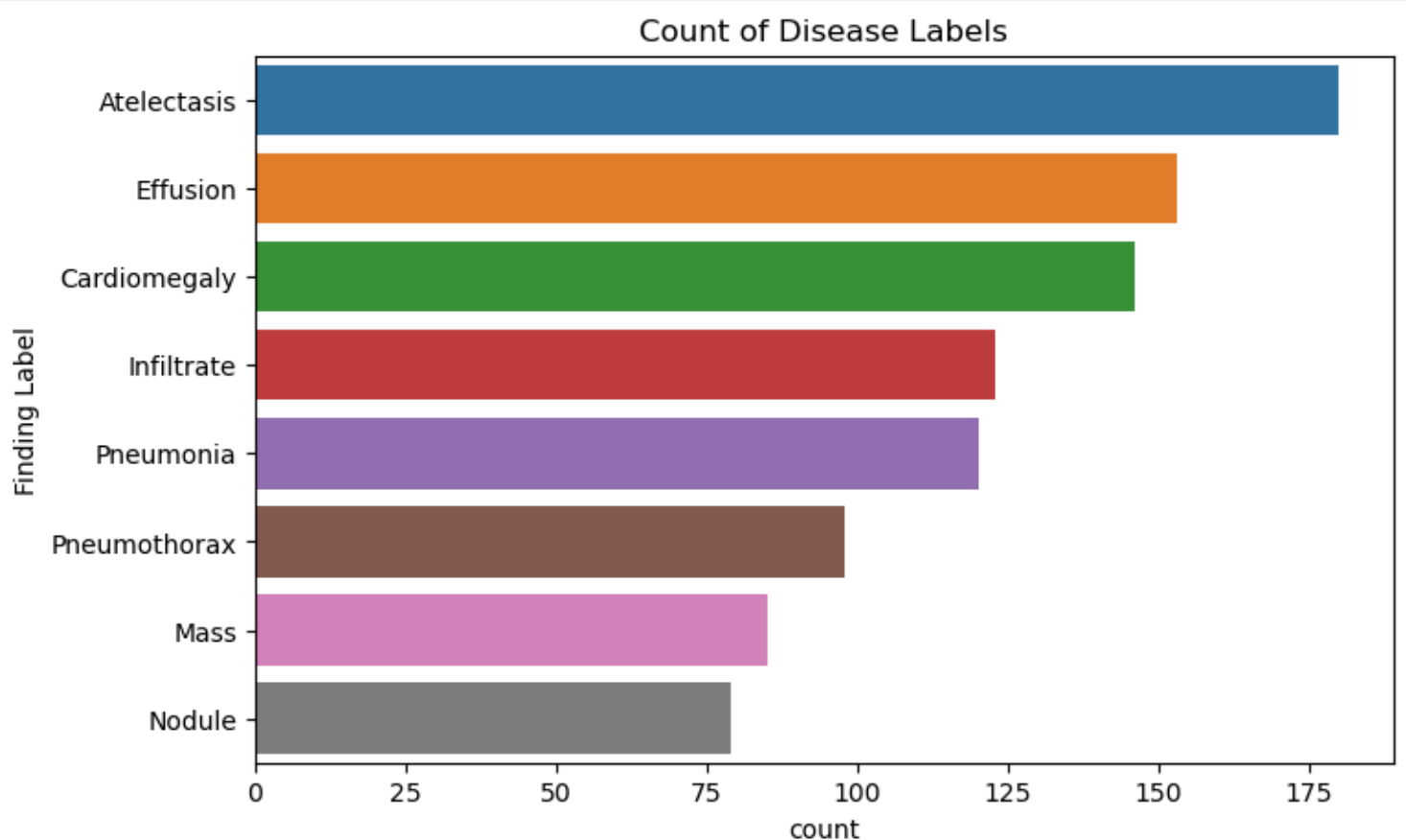
Exploratory Data Analysis Overview

Count of Disease Labels

Distribution of Bounding Box Areas


```python
# === Part 5: Filtering examples ===
big_boxes = df[df["area"] > 50000]
print("Large boxes:", big_boxes.shape)


small_boxes = df[df["area"] < 5000]
print("Small boxes:", small_boxes.shape)


multi_labels = df.groupby("Image Index").size()
print("Images with multiple findings:", (multi_labels > 1).sum())
```
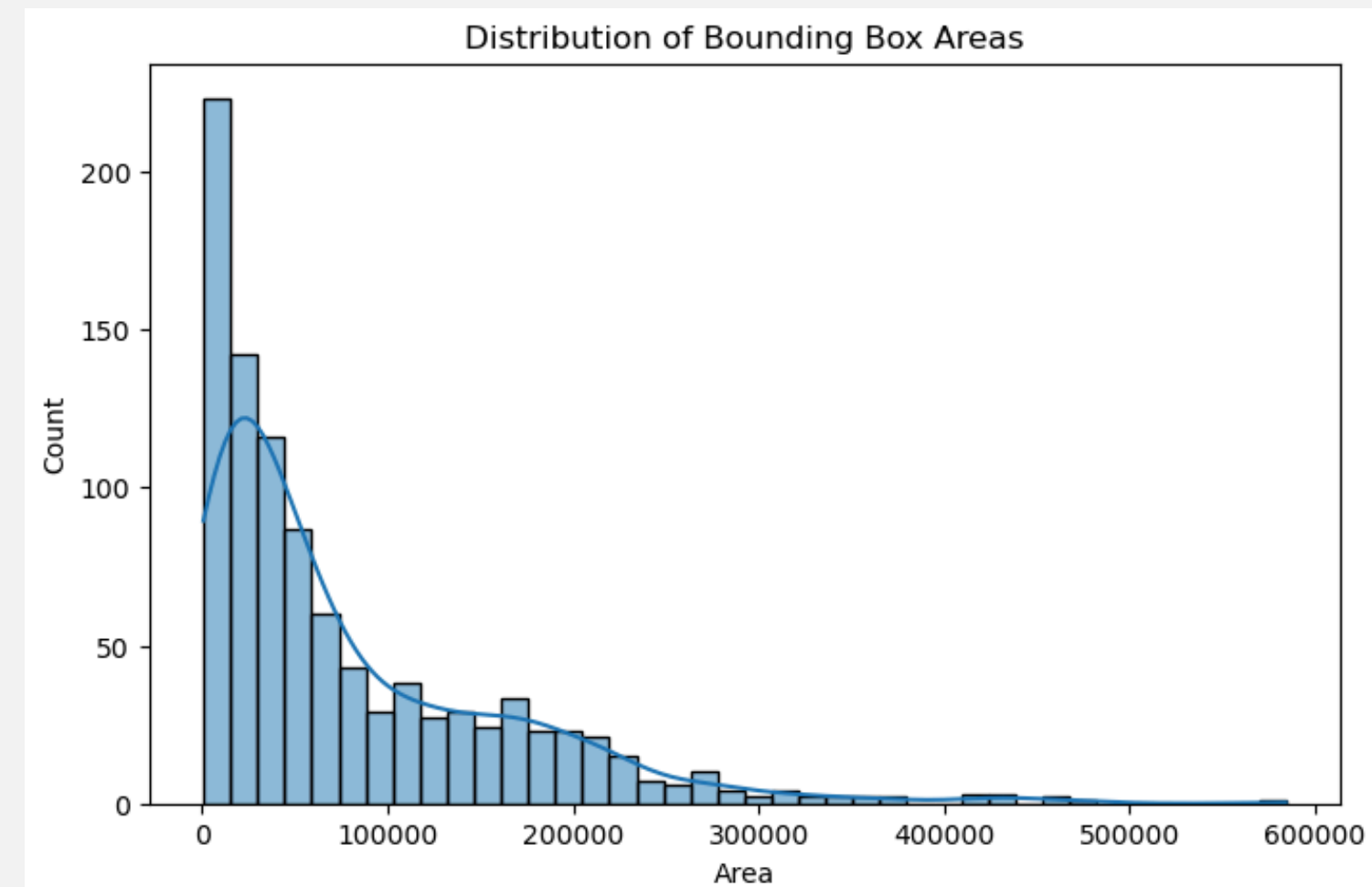
```
Large boxes: (473, 10)
Small boxes: (60, 10)
Images with multiple findings: 93
```
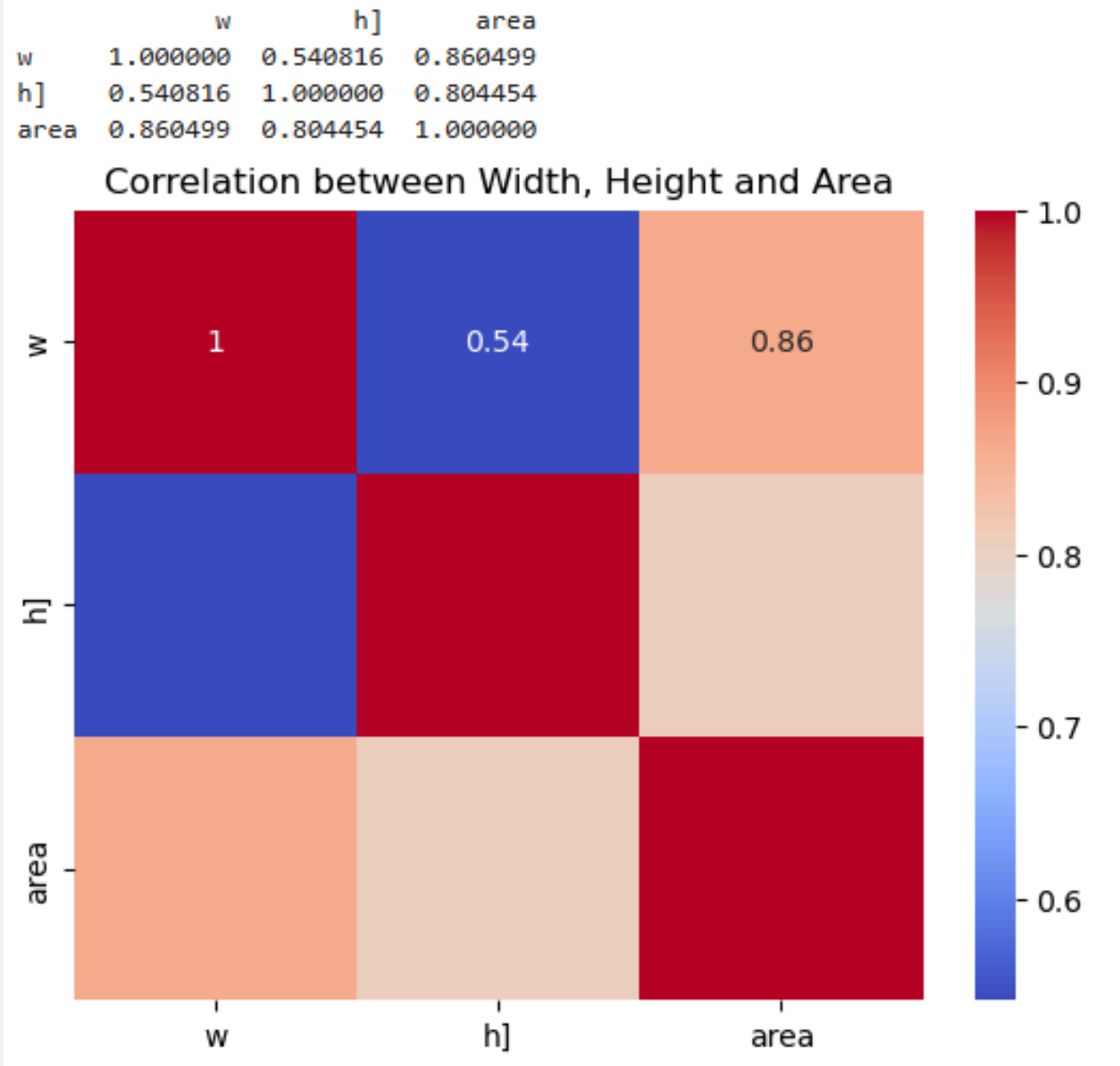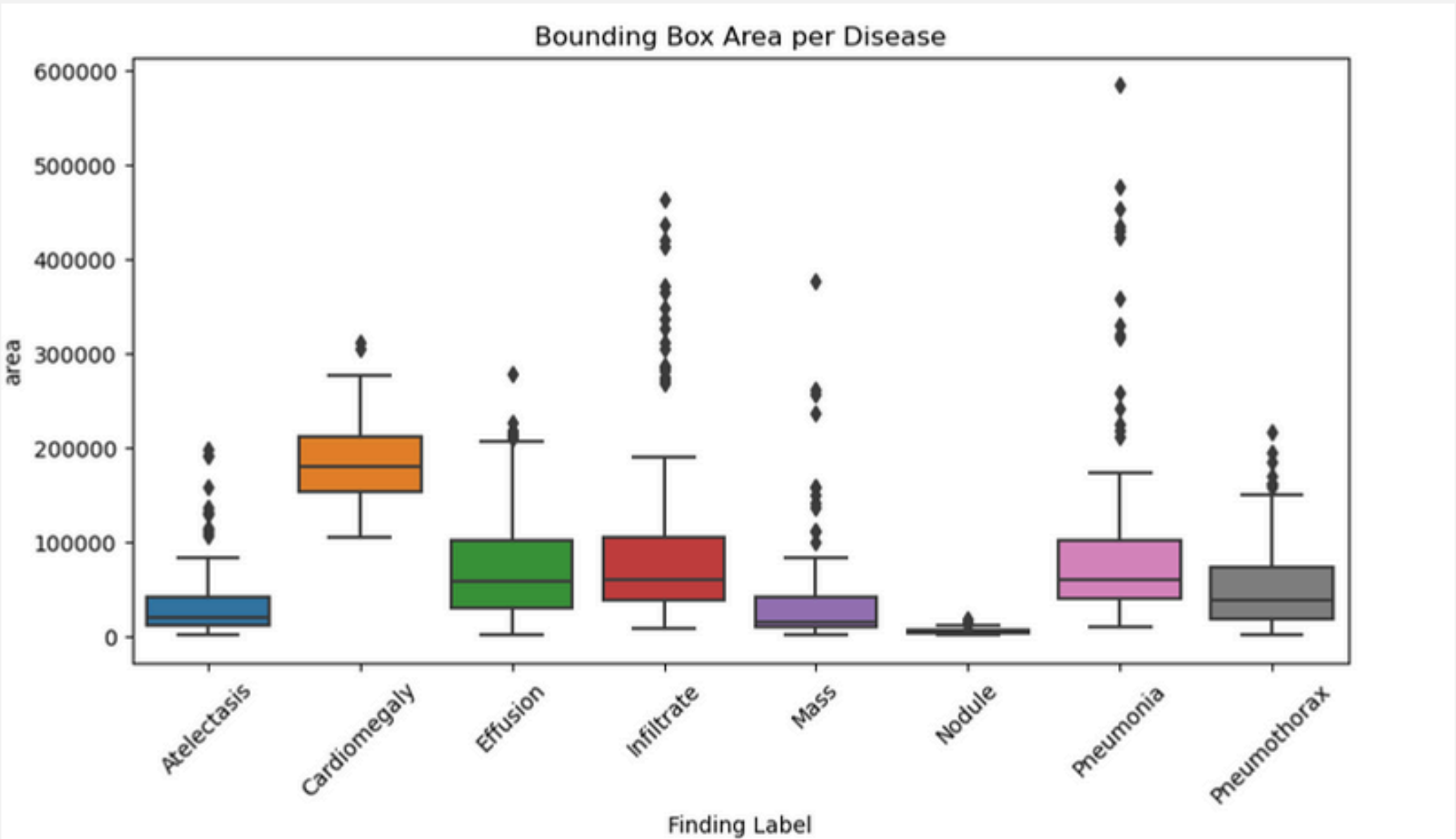
w	h]	area
86.779661	79.186441	6871.772479
185.491525	313.491525	58150.021258
155.118644	216.949153	33652.858374
141.016949	55.322034	7801.344441
200.677966	78.101695	15673.289285

Bounding Box Area per Disease

Correlation between Width, Height and Area

	w	h]	area
w	1.000000	0.540816	0.860499
h]	0.540816	1.000000	0.804454
area	0.860499	0.804454	1.000000

|	Bbox [x	y	w	h]	x_center	y_center
0	225.084746	547.019217	86.779661	79.186441	268.474576	586.612437
1	686.101695	131.543498	185.491525	313.491525	778.847458	288.289261
2	221.830508	317.053115	155.118644	216.949153	299.389831	425.527691
3	726.237288	494.951420	141.016949	55.322034	796.745763	522.612437
4	660.067797	569.780787	200.677966	78.101695	760.406780	608.831634

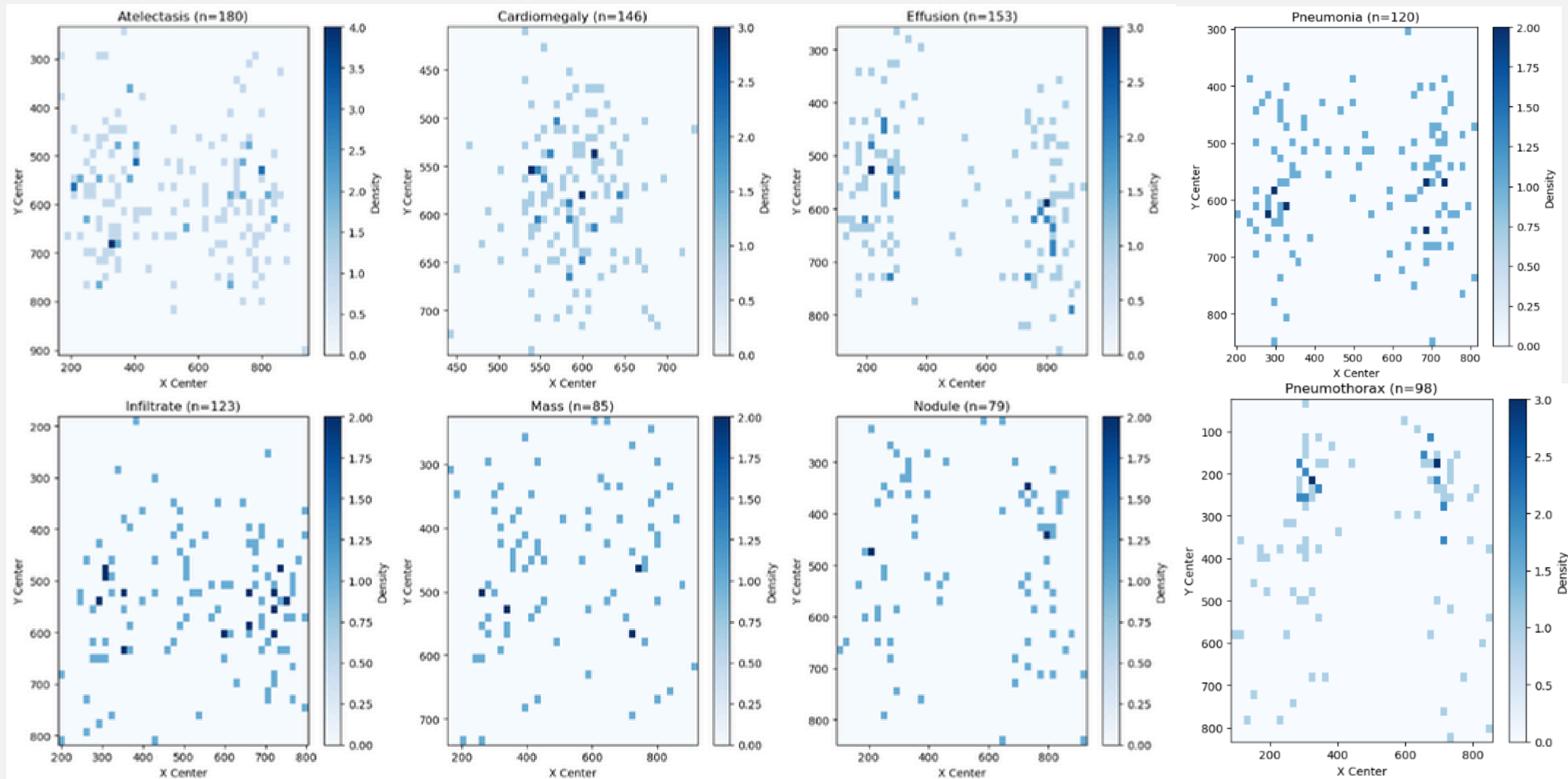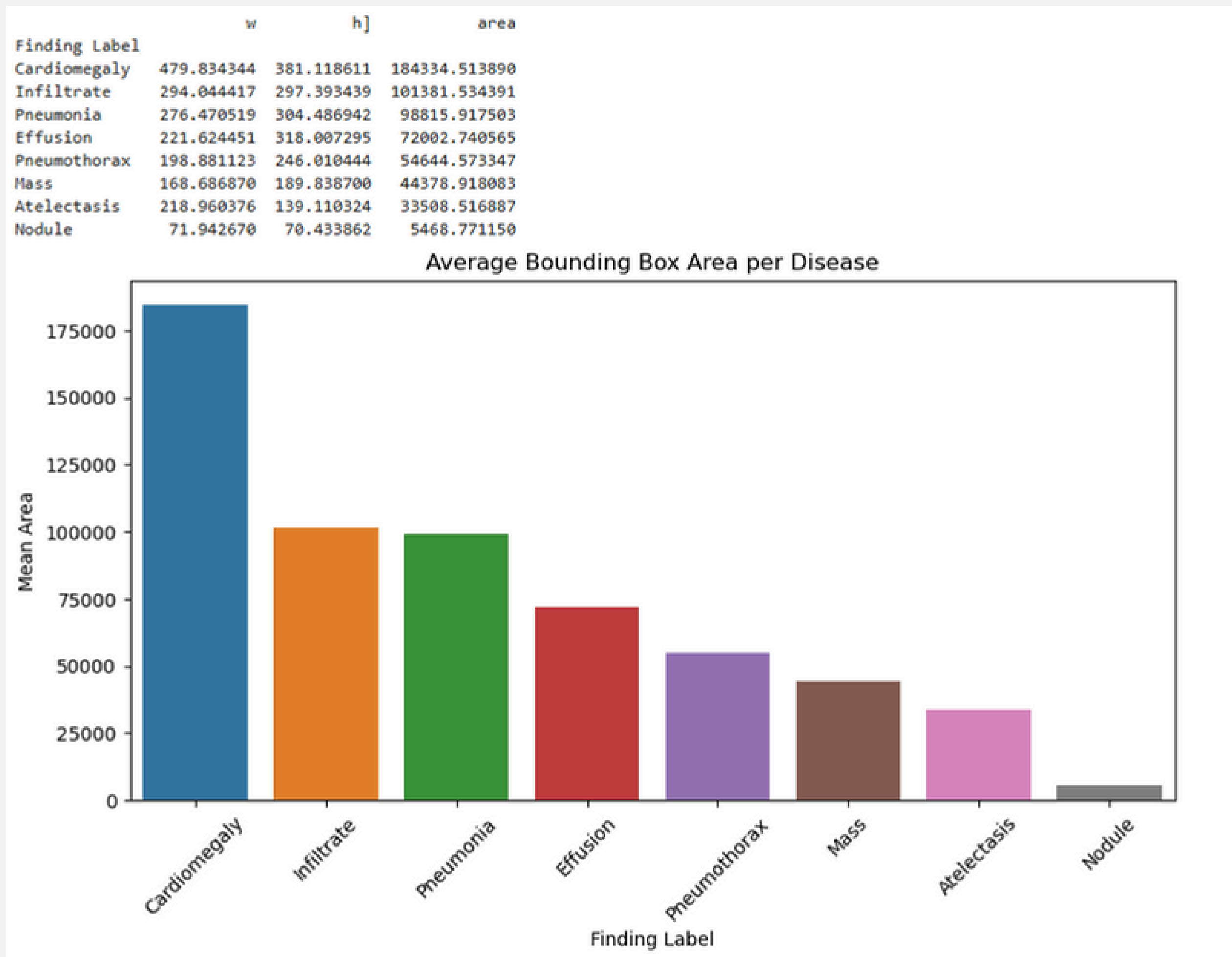Bounding Box Centers by Disease Label
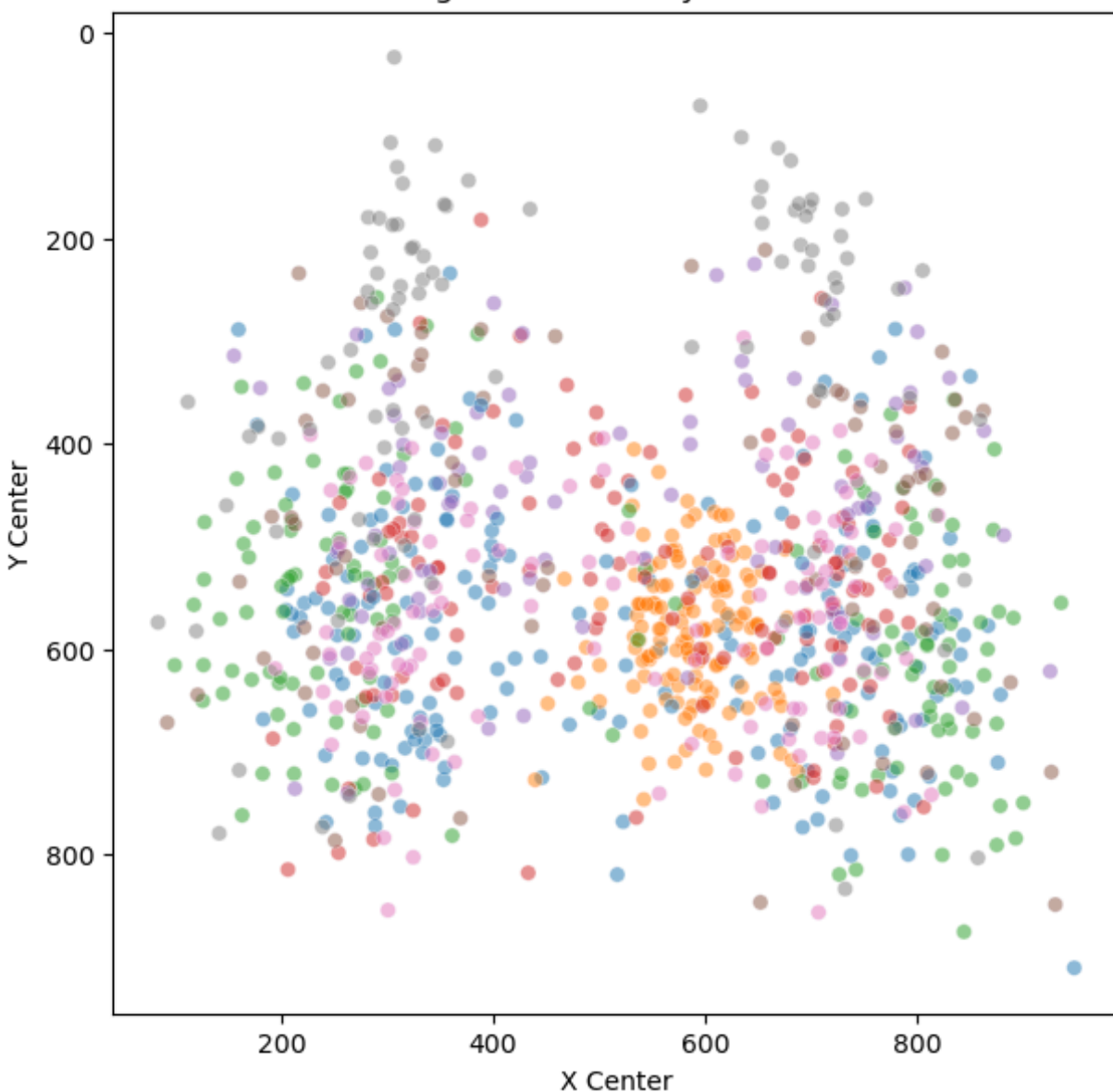
Bounding Box Density Across Image
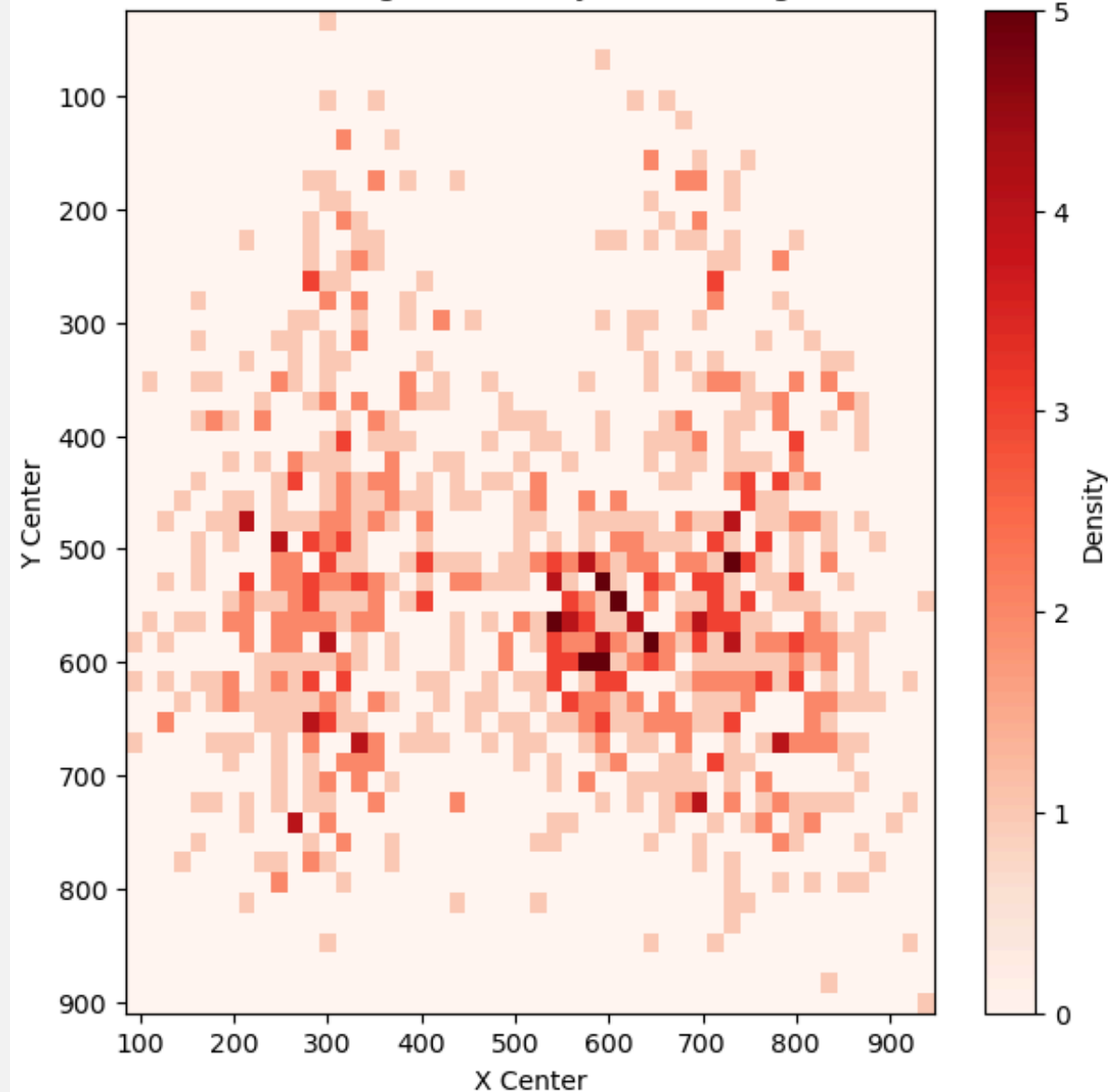
Bounding Box Centers with Size (scaled by area)
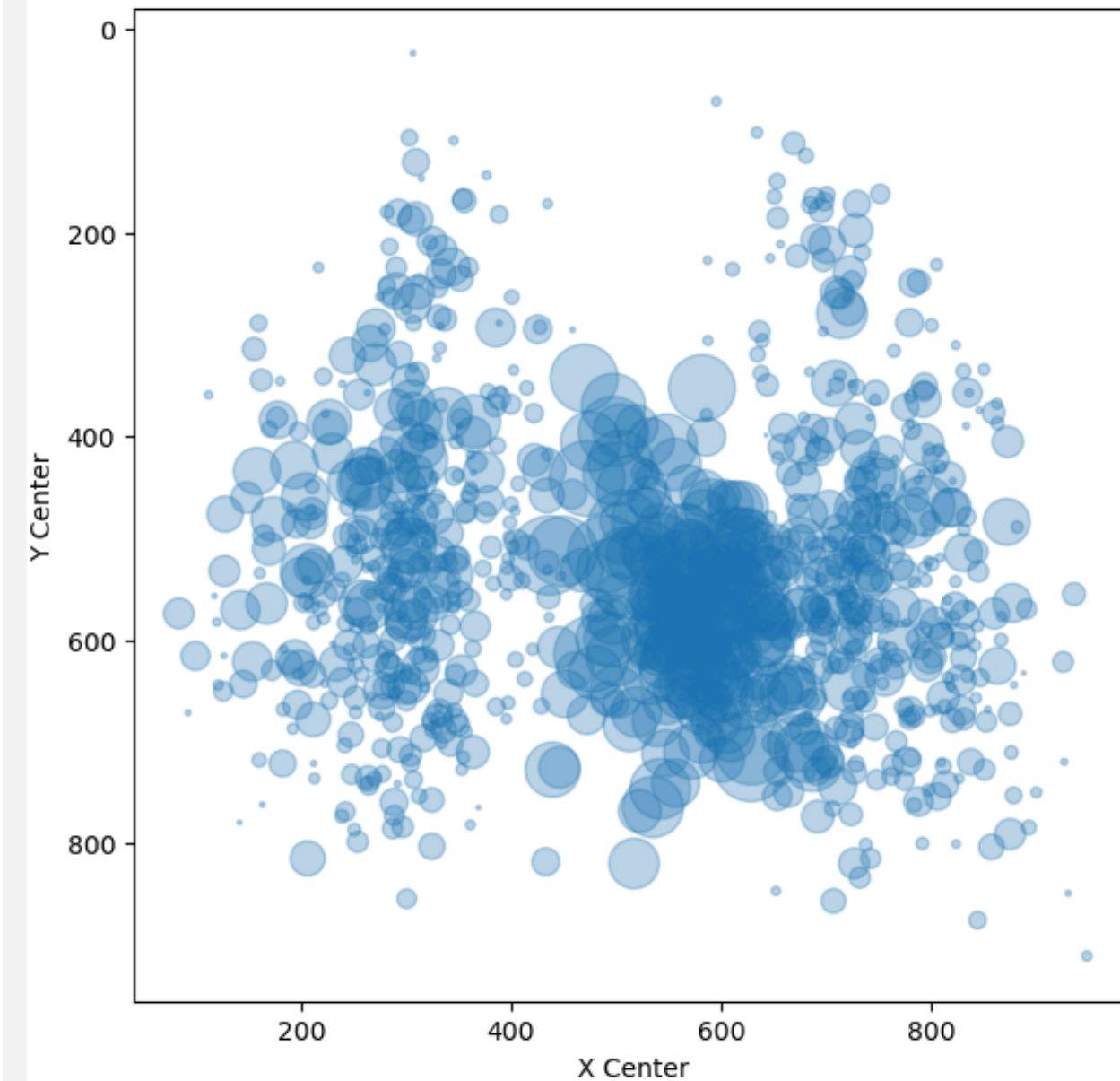
- Atelectasis
- Cardiomegaly
- Effusion
- Infiltrate
- Mass
- Nodule
- Pneumonia
- Pneumothorax

```
                    w              h]          area
Finding Label
Cardiomegaly    479.834344    381.118611    184334.513890
Infiltrate      294.044417    297.393439    101381.534391
Pneumonia       276.470519    304.486942     98815.917503
Effusion        221.624451    318.007295     72002.740565
Pneumothorax    198.881123    246.010444     54644.573347
Mass            168.686870    189.838700     44378.918083
Atelectasis     218.960376    139.110324     33508.516887
Nodule           71.942670     70.433862      5468.771150
```


Average Bounding Box Area per Disease

This collection illustrates our visual insights: heatmaps, scatter plots, and annotated X-ray examples.