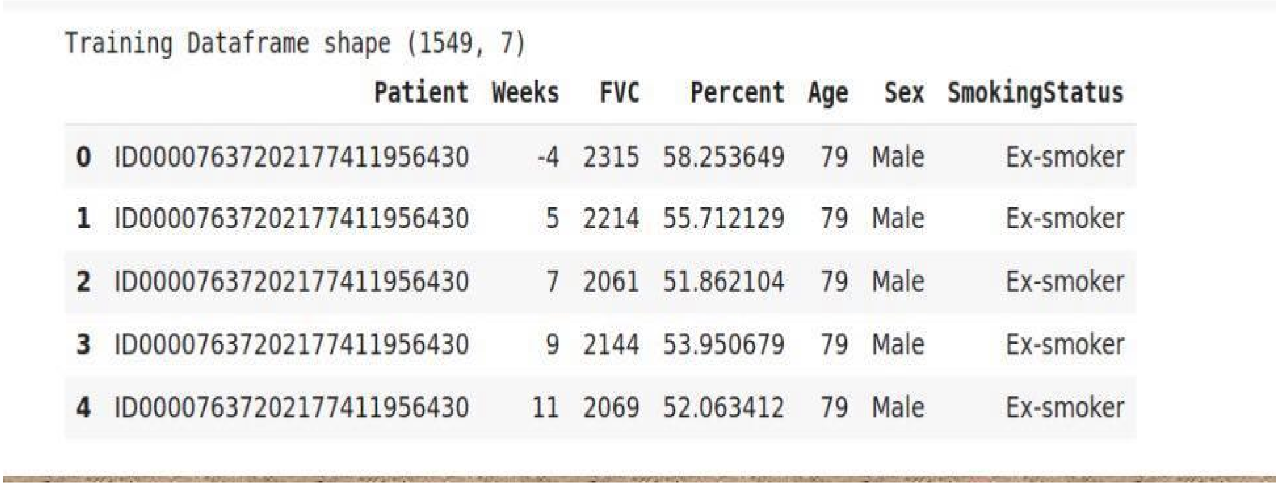
1.1. Data Sets

Data set contains two main categories tabular data and CT-Scan

This is a snapshot from the dataset which is consist of 176 patients and each patient has a specific number of weeks or observations
The weeks column contains an equation of time which shows the difference between the date of FVC scan and CT Scan.

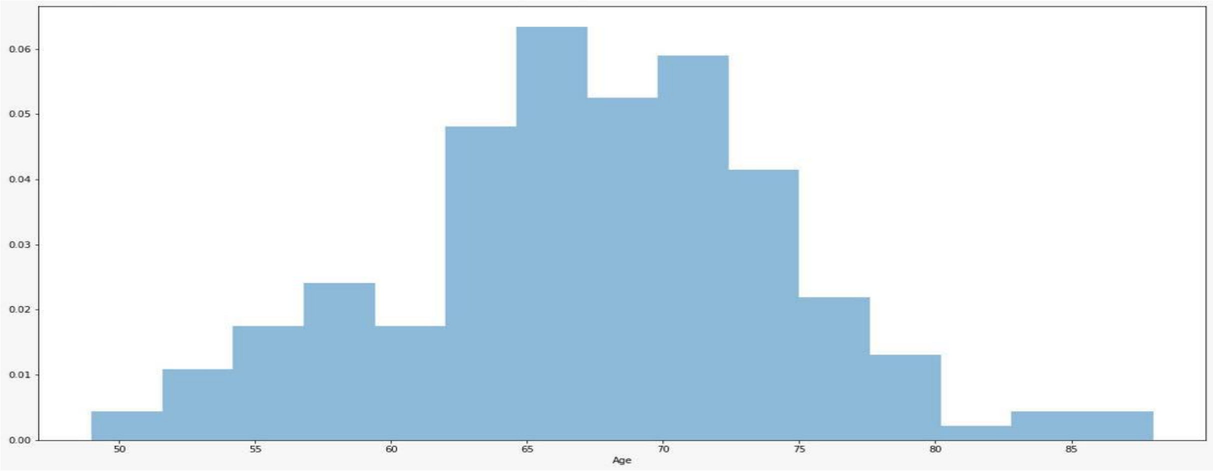
For example

In the case no.(0)the patient did the FVC scan before 4 weeks of doing the CT Scan

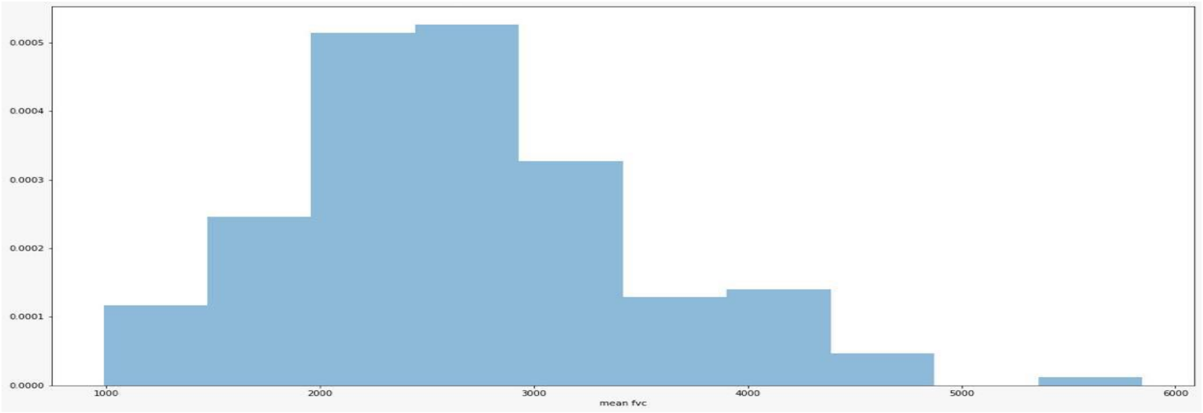


```
Training Dataframe shape (1549, 7)
                         Patient  Weeks   FVC    Percent  Age   Sex  SmokingStatus
0  ID00007637202177411956430     -4  2315  58.253649   79  Male      Ex-smoker
1  ID00007637202177411956430      5  2214  55.712129   79  Male      Ex-smoker
2  ID00007637202177411956430      7  2061  51.862104   79  Male      Ex-smoker
3  ID00007637202177411956430      9  2144  53.950679   79  Male      Ex-smoker
4  ID00007637202177411956430     11  2069  52.063412   79  Male      Ex-smoker
```

√ EDA: Age distribution

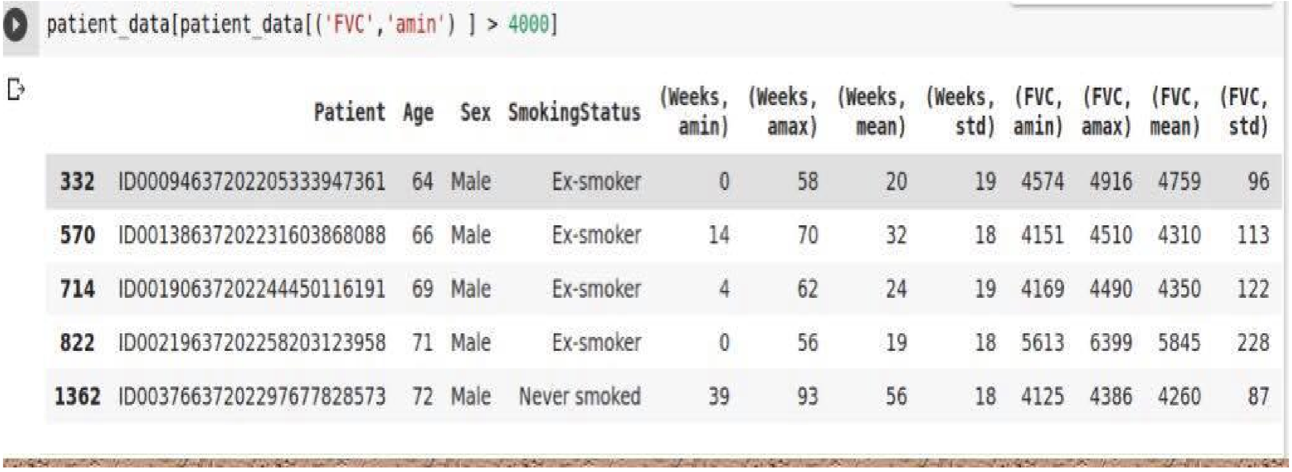


√ EDA: mean FVC distribution



√ **EDA: FVC ¡ 4000**

It is not scientifically valid that there are patients who have FVC greater than 4000.

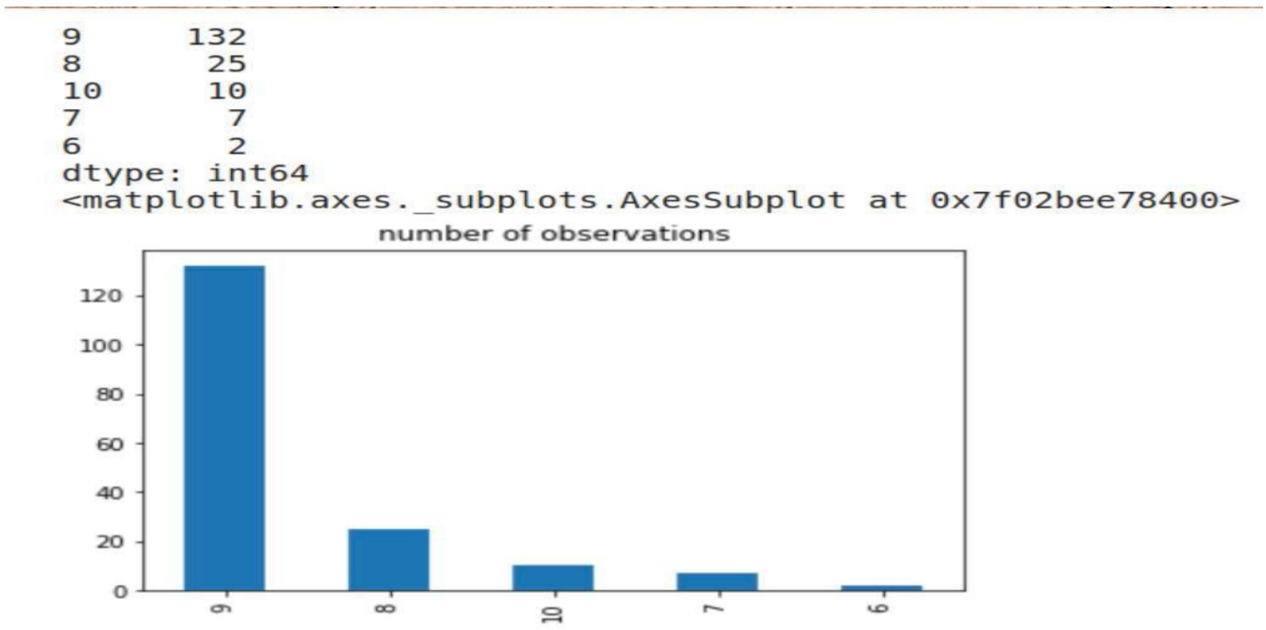


```
patient_data[patient_data[('FVC','amin') ] > 4000]
```

	Patient	Age	Sex	SmokingStatus	(Weeks, amin)	(Weeks, amax)	(Weeks, mean)	(Weeks, std)	(FVC, amin)	(FVC, amax)	(FVC, mean)	(FVC, std)
332	ID00094637202205333947361	64	Male	Ex-smoker	0	58	20	19	4574	4916	4759	96
570	ID00138637202231603868088	66	Male	Ex-smoker	14	70	32	18	4151	4510	4310	113
714	ID00190637202244450116191	69	Male	Ex-smoker	4	62	24	19	4169	4490	4350	122
822	ID00219637202258203123958	71	Male	Ex-smoker	0	56	19	18	5613	6399	5845	228
1362	ID00376637202297677828573	72	Male	Never smoked	39	93	56	18	4125	4386	4260	87

√ **EDA: number of observations**

All patients have in common that they have number of observations around 10 weeks per patient.



```
9      132
8       25
10      10
7        7
6        2
dtype: int64
<matplotlib.axes._subplots.AxesSubplot at 0x7f02bee78400>
```

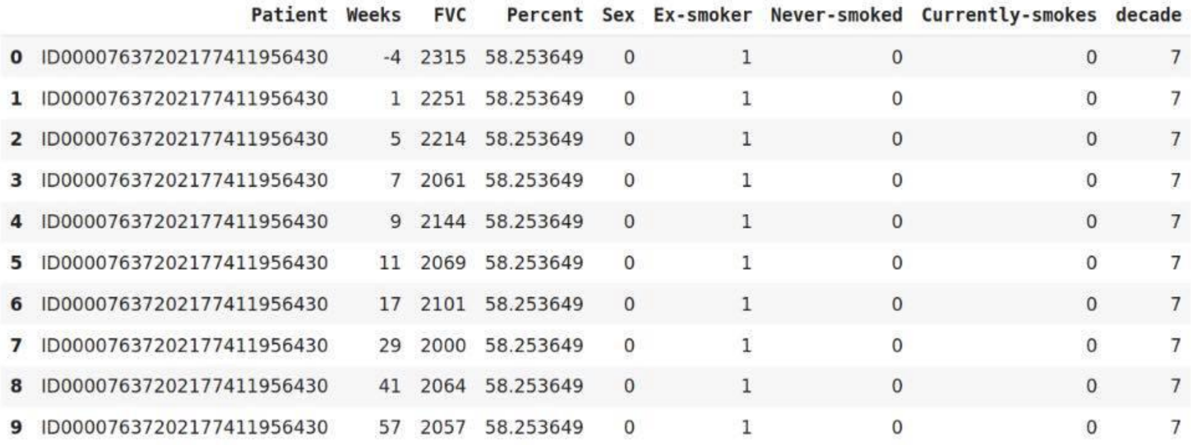
√ **Data transformation**

•remove patient that has FVC value greater than 4000.

•interpolate the missing weeks thus becomes the number of observations exactly 10 per patient.

•convert categorical feature into a one-hot vector.

√ **Tabular Data after transformation**



	Patient	Weeks	FVC	Percent	Sex	Ex-smoker	Never-smoked	Currently-smokes	decade
0	ID00007637202177411956430	-4	2315	58.253649	0	1	0	0	7
1	ID00007637202177411956430	1	2251	58.253649	0	1	0	0	7
2	ID00007637202177411956430	5	2214	58.253649	0	1	0	0	7
3	ID00007637202177411956430	7	2061	58.253649	0	1	0	0	7
4	ID00007637202177411956430	9	2144	58.253649	0	1	0	0	7
5	ID00007637202177411956430	11	2069	58.253649	0	1	0	0	7
6	ID00007637202177411956430	17	2101	58.253649	0	1	0	0	7
7	ID00007637202177411956430	29	2000	58.253649	0	1	0	0	7
8	ID00007637202177411956430	41	2064	58.253649	0	1	0	0	7
9	ID00007637202177411956430	57	2057	58.253649	0	1	0	0	7

CT-Scan

every single patient has several slices and that number varies from one patient to another.

some patients have around 20 slices while other have around 1,000 slices.

why this gap？how to handle it？

> We use it because it is a computed field which approximates the patient’s FVC as a percent of the typical FVC for a person of similar characteristics

√ CT scan: how to handle

- **first approach:**

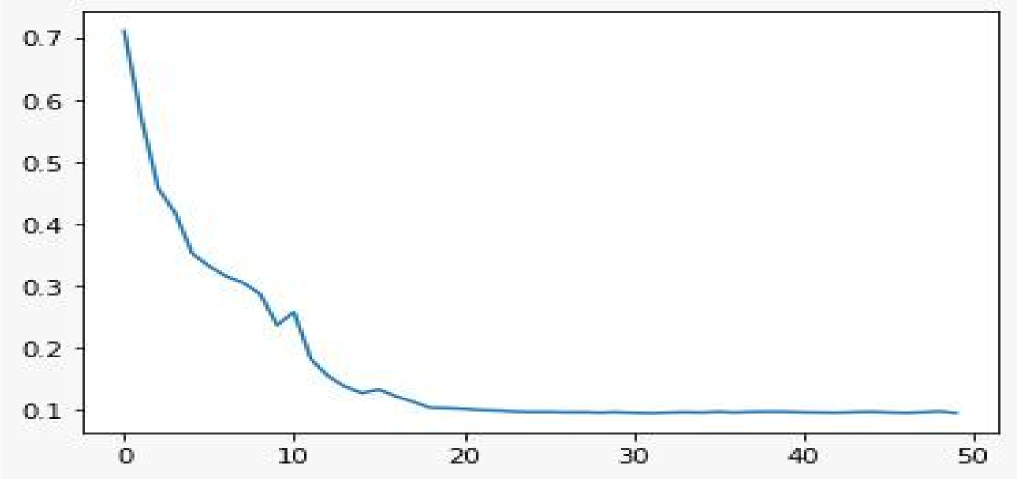
for each CT scan image we can segment it to find the lung parts and the fibrosis disease regions and based on that we can calculate the percentage of the fibrosis disease for each lung which will help us predict the FVC.

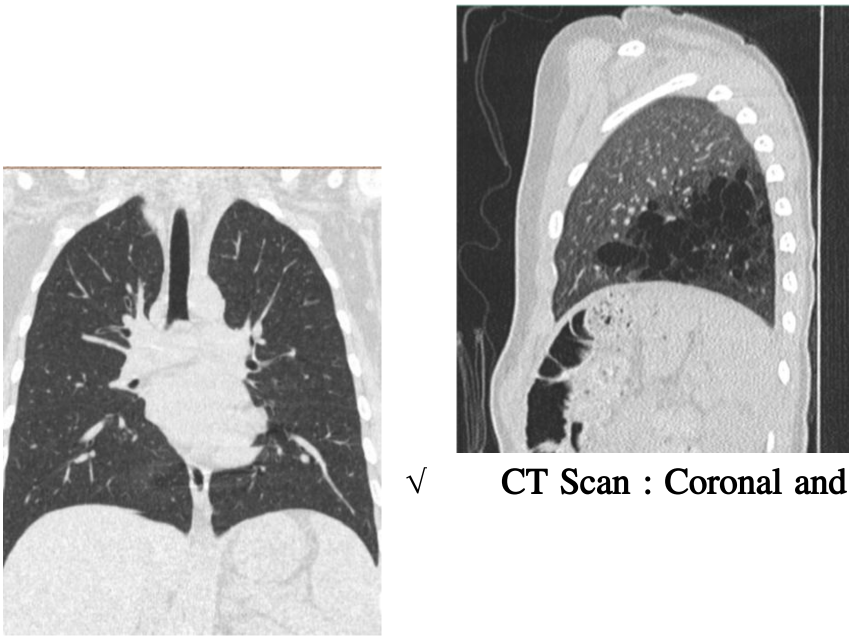
- **second approach:**

We need a constant number of slices for each patient so we need to convert the slices into two main views Sagittal and Coronal, then we can use a neural network to extract lung features that will help predict FVC.

As for us we chose to use the second approach Because in the first one we couldn't make a segmentation for a lung fibrosis, so we convert from axial view into coronal and sagittal view.

After converting into coronal and sagittal all patient will have the same number of slices, But we couldn't just convert into another view and pass these slices into CNN, but after converting and before the feature extraction phase we need to segment the lung But the U-net model that we have can't segment the lung from these views it's only provide a segmentation from axial view So we need to build our own U-net model But there are some problem, we don't have a dataset to train our model So we manually segment the lung and make our own dataset then we train the U-net model and we got a validation loss around 0.06.





√ CT Scan : Coronal and Sagittal View

√ **CT Scan : Segmentation**

Binary Segmentation

First, we analyze the color space of the ct scans then we use Kmeans with k = 2 to find two mean color values and if a given pixel is below the mean of the two-color values then the pixel belongs to the lungs otherwise it's a background then we can do erosion to remove unwanted pixels outside the lungs, then dilation to add pixels around the lungs.

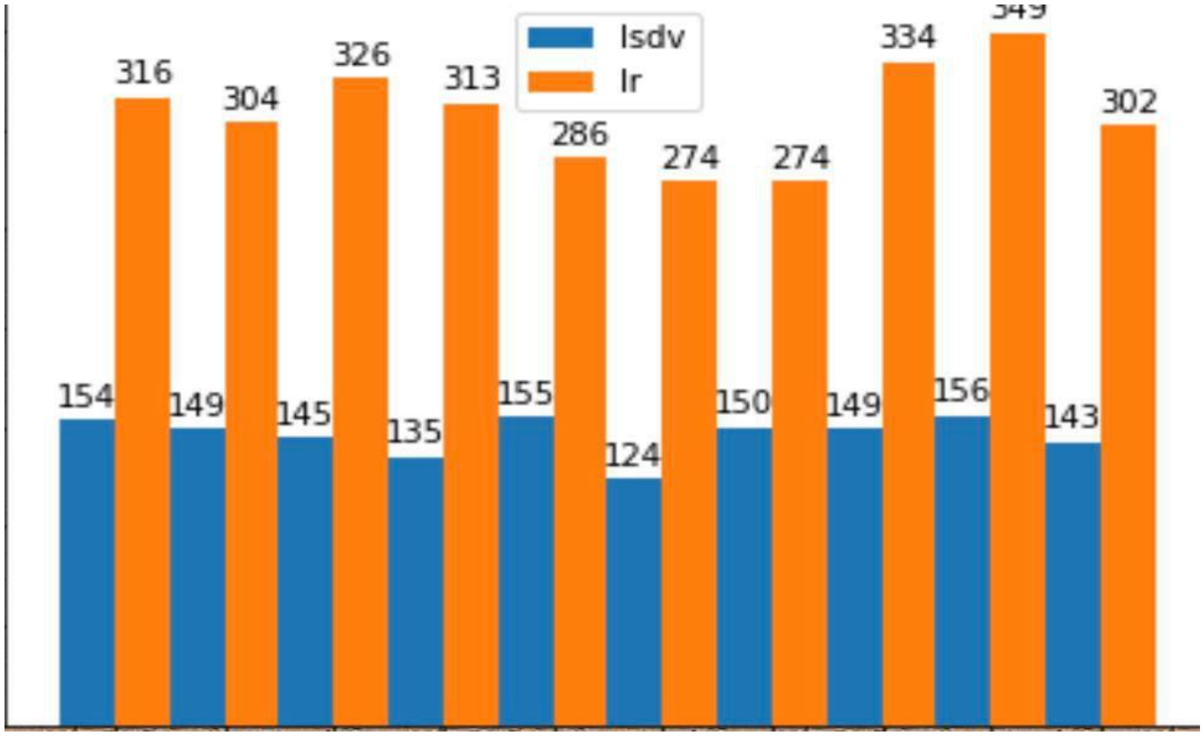
U-Net

We use a deep neural network called u-net which is a convolutional network architecture for fast and precise segmentation of images. The model is trained to segment right and left lungs separately including air-pockets, tumors, and effusions, and excluding the trachea. We can use the same model to train it to segment fibrosis disease regions.

1.2. Results

After used the new feature the loss became 136

Panel data model vs liner regression



2. conclusion

predict the FVC value using a panel data model.

we analyze the tabular data and understand it very Well

we convert the axial image into coronal and Sagittal view

Then segment it to extract the lung features from it

Finally, we add the extracted features into the tabular data and then train the model with that newly data and we got a MAE = 136