Quantitative_assessment_data

March 21, 2024

1 Quantitative Assesment of NLST DATA

1.1 Introduction

The National Lung screening trial was a large study involving more than 50,000 patients over several years. Patients were screened three times annually and were assigned to either low dose-helical CT or standard chest X-ray. This was done to assess whether those exposed to low dose CT had better outcomes than those exposed to X-rays. cdas.cancer.govThe study indicated that patients who received low dose helical CT scans had lower risk of dying from lung cancer than those who received X-rays only.NLST

1.2 Purpose:

The purpose of this report was to perform a quantitative assessment of the data avialable and used for this project.

The images are stored and acessible publicly at the following websites:

https://www.cancerimagingarchive.net/collection/nlst/
https://portal.imaging.datacommons.cancer.gov/explore/

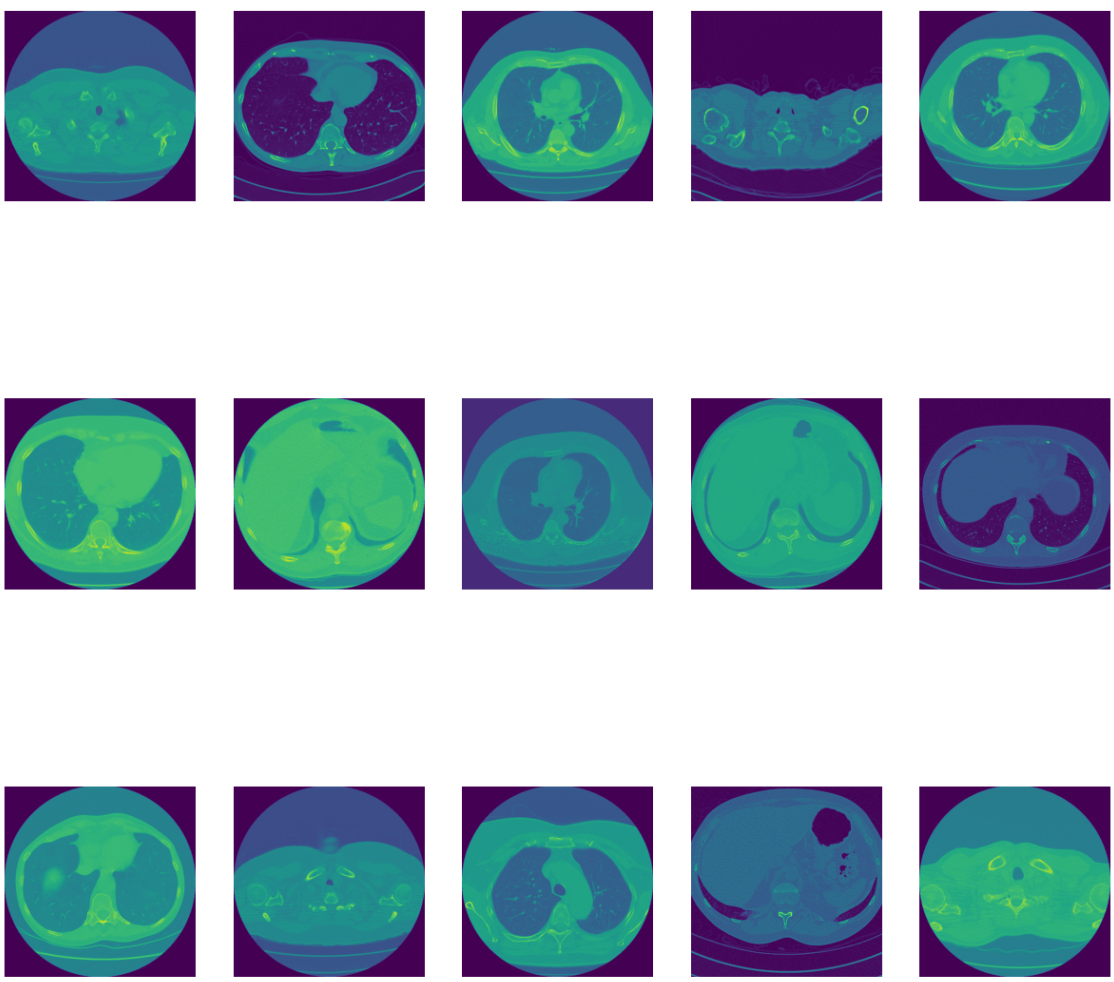
And there are a few pieces of data needed to navigate the image data.

- patient data sets
- patient data dictionaries to explain the data sets

The patient data dictionaries can be found at : https://cdas.cancer.gov/datasets/nlst/
And the data sets associated with those dictionaries : https://www.cancerimagingarchive.net/collection/nlst/

The data covered in this project:

The `People` dataset which contains demographic information about participants.
The `Abnormalities` dataset, which contains information about locations of abnormalities for patients.
The `Lung Cancer data set`, which contains information about the states of the cancers if available, for example staging information.
And the images themselves. Which consisted of the DICOM files as well as the transformed to j-peg versions of them.



The above images are the loaded CT scans from the NLST study.

1.3 Metrics to be assessed:

1.3.1 Accuracy/Completeness:

`metrics` function: Using the dtype functions to produce the various data types of the columns.
`patientID` function: Verifying the patient ID numbers are present.

1.3.2 Missingness:

`missing_plots` function: Using the missing no package to visualize null and empty columns. https://pypi.org/project/missingno/

1.3.3 Correlations:

`heats` function: From the missing no package. A heatmap is produced using the missing no nullity function. Which outputs how likely two events are to occur together.
* -1 meaning they are mutually exclusive.
* 0 meaning there is no relationship
* 1 if there is a strong relationship.
`corrs1` function: Plots The Pearson product-moment correlation

1.3.4 Interesting Plots

Any additional visualizations gathered from each data set.

1.4 The data:

1.4.1 Patients

"The Participant dataset is a comprehensive dataset that contains all the NLST study data needed for most analyses of lung cancer screening, incidence, and mortality. The dataset contains one record for each of the ~53,500 participants in NLST." It is associated with a pdf file called the participant data dictionary. Link to dictionary can be found here:

https://cdas.cancer.gov/datasets/nlst/

The data set can be obtained here: https://www.cancerimagingarchive.net/collection/nlst/

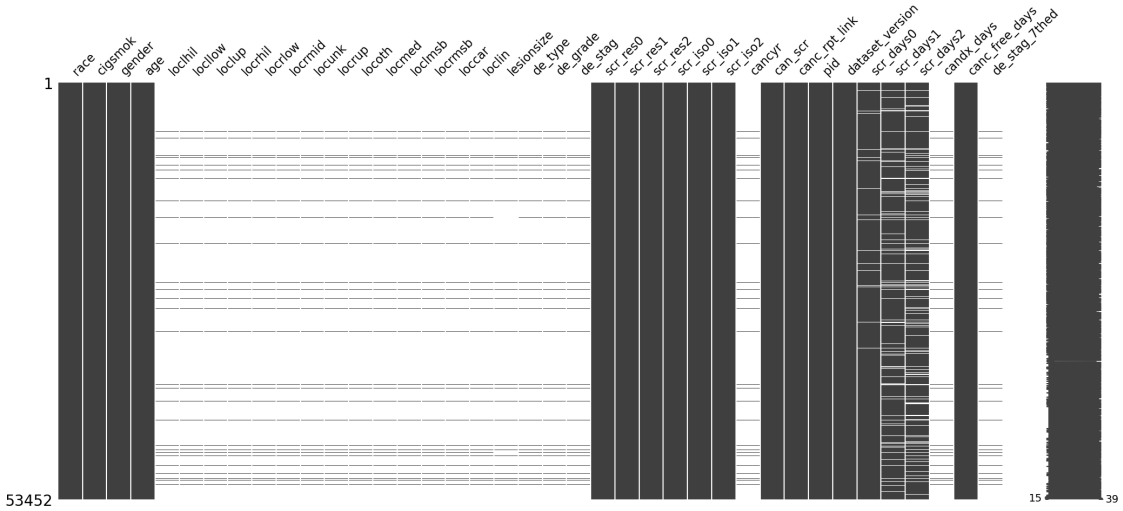
1.4.2 Accuracy/Completeness

How many null and non null columns are there? And what is the shape of the data (breadth and depth)?

```
The shape of the data set: 53452 rows, and 39 columns
The Counts of various columns
 Counter({dtype('float64'): 24, dtype('int64'): 14, dtype('O'): 1})
There are: 0 Duplicate columns

 (None, None)

There are 0 null Patient IDS
and 53452 total patient records
As well as 53452 unique IDs present
 The % of null columns is: 52.35135477408919 %
```

1.4.3 Missingness

```
This plot uses the missingno package, and plots the missing data vs the non
missing in any dataframe.
    Empty (NAN) values are white and non empty values are black
```

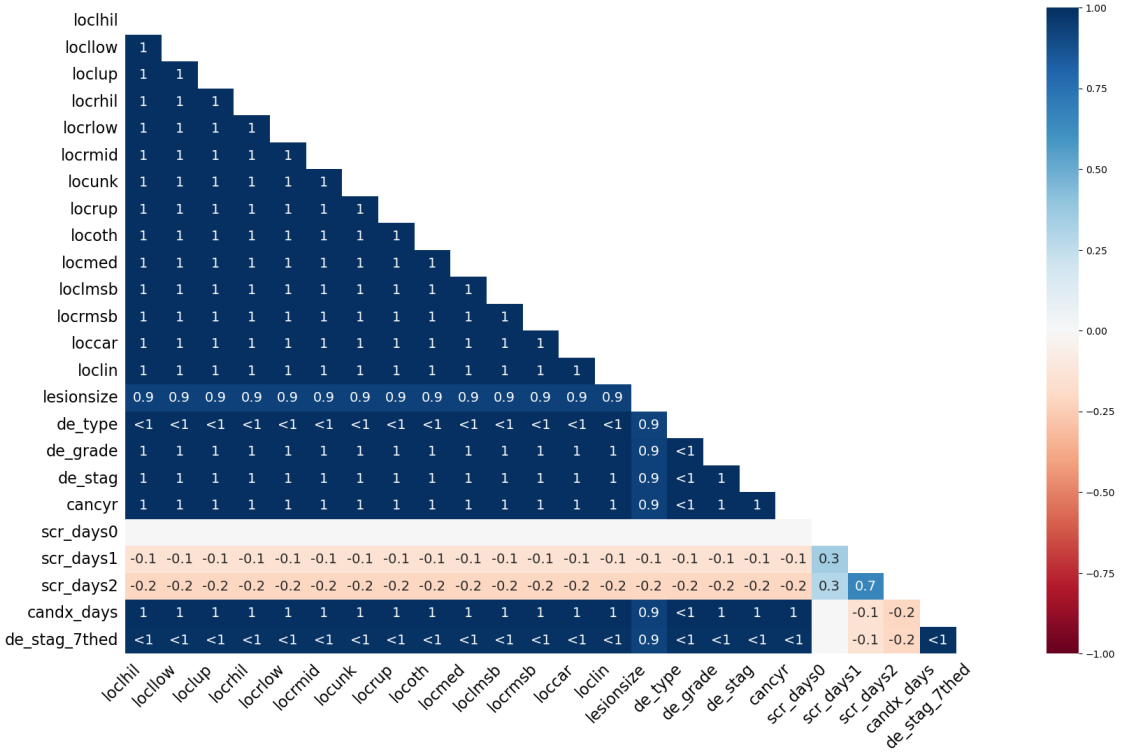


In the above plot most of the null columns are the ones associated with a positive cancer diagnosis. For example lesion size or location. Most patients involved in the study did not have cancer diagnosed during the study period. Sometimes empty columns are not a bad thing.

1.4.4 Nullity

The matrix below is known as a nullity matrix

```
Nullity is -1 if two variables are mutually exclusive.
           0 if there is no effect.
           1 if there is a strong relationship.
```

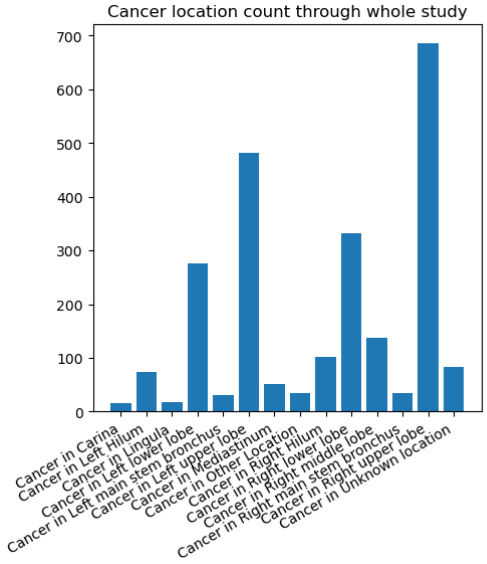


It would make sense that the locations of cancers which are the columns starting with "lo" would have a value close to 1. The way that data was recorded involved putting a 0 where the cancer was not for every one of those columns. And a 1 where it was present location wise in the lung area. A negative Cancer patient had NA recorded in those columns instead. src_days1 and src_days2 are days since randomizaton at the time of screening. Since they should be extremely random, there should be no correlation.

1.4.5 Correlations

```
The % of values that are highly positively correlated (above 0.8)
36.36363636363637 %
 The % of values that are highly negatively correlated (below -0.8) 0.0 %
```

1.4.6 Interseting plots

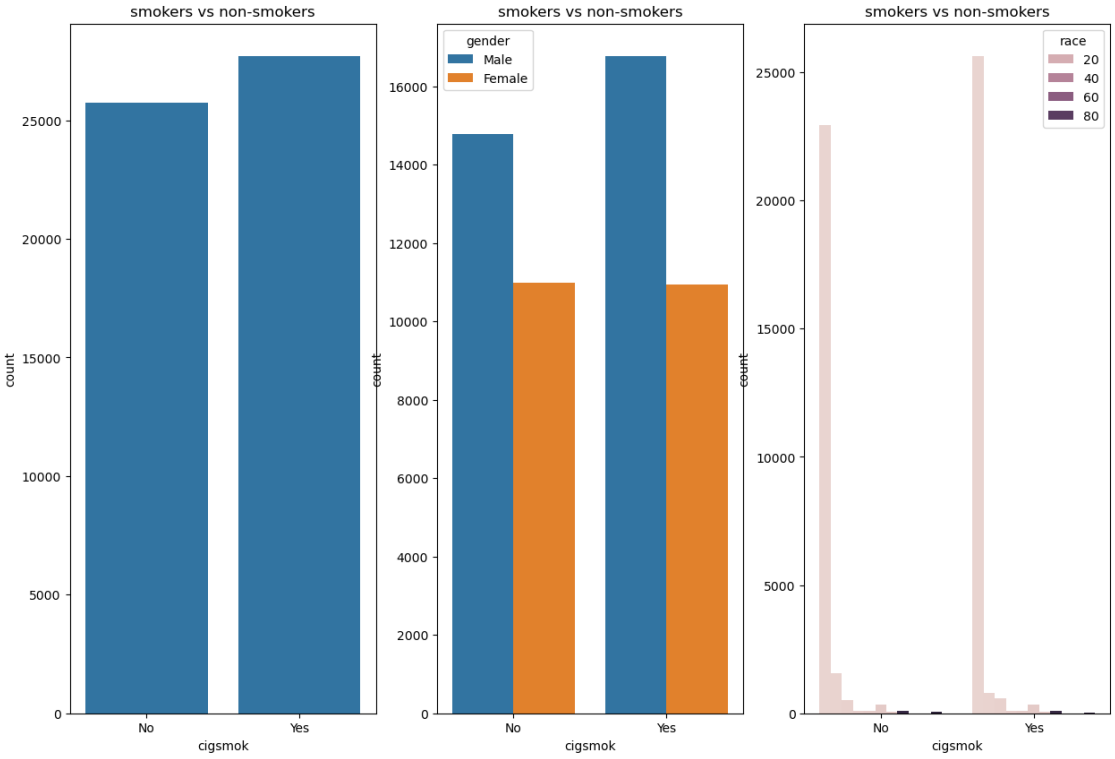


The cancer locations category marked
a correct 'hit' with a 1.
So this only applied to patients who
had cancerous abormalities.
In this case. Any missing values
should mean there was no cancer found at all.

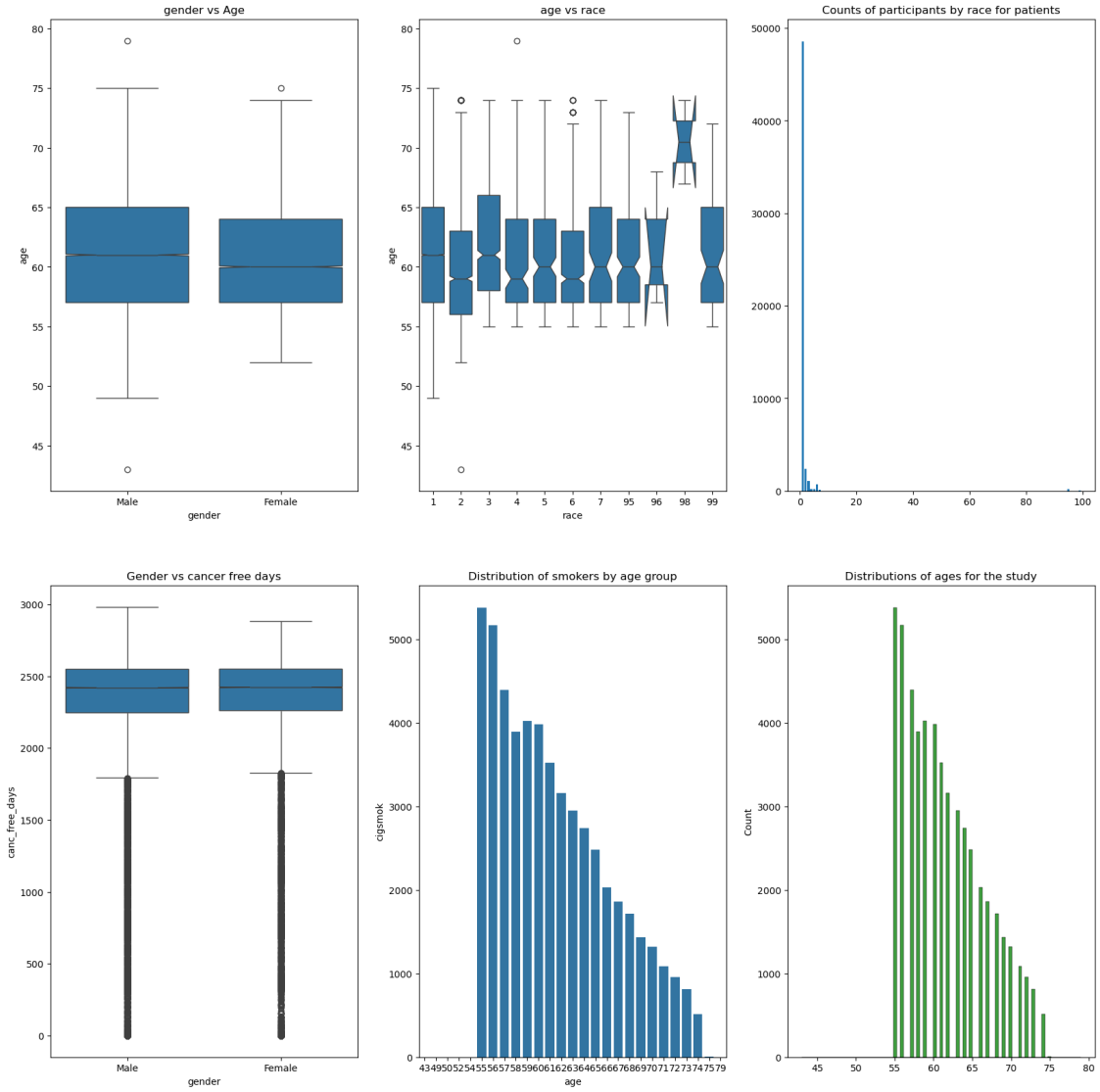
The total number of positive 'hits':2354
Note.
A patient may have a
positive result in more than 1 location.

Where patients most often had lung cancer found during the study:

Cancer in Carina ,16,
Cancer in Left Hilum ,73,
Cancer in Lingula,17,
Cancer in Left lower lobe,275,
Cancer in Left main stem bronchus,31,
Cancer in Left upper lobe,482,
Cancer in Mediastinum,52,
Cancer in Other Location,34,
Cancer in Right Hilum,102,
Cancer in Right lower lobe,331,
Cancer in Right middle lobe,138,
Cancer in Right main stem bronchus,34,
Cancer in Right upper lobe,686,
Cancer in Unknown location,83,



The general ratio of smokers to non smokers was about even. But men were smokers more often. And the data for smokers by race shows that most of the study was comprised of one group.



`0.9127295971122934 % of patients were this race`

`0.0446692109567408 % of patients belonged to the next most frequent category`

Looking over all patients as a whole. The distributions of patients by age appears to be fairly evenly distributed in terms of race and gender. But most of the study participants did fall into one category. 91% of the study participants were in the same category. The next highest category was 5%. And most people who participated in the study were in their 50s.

1.4.7 The Lung Cancer Data Dictionary/ Data set

"The Lung Cancer dataset (~2,100, one record per lung cancer) contains information about each lung cancer instance diagnosed during the trial, including multiple primary tumors in the same individual. It focuses on characteristics of the cancer, including information not available in the Participant dataset."

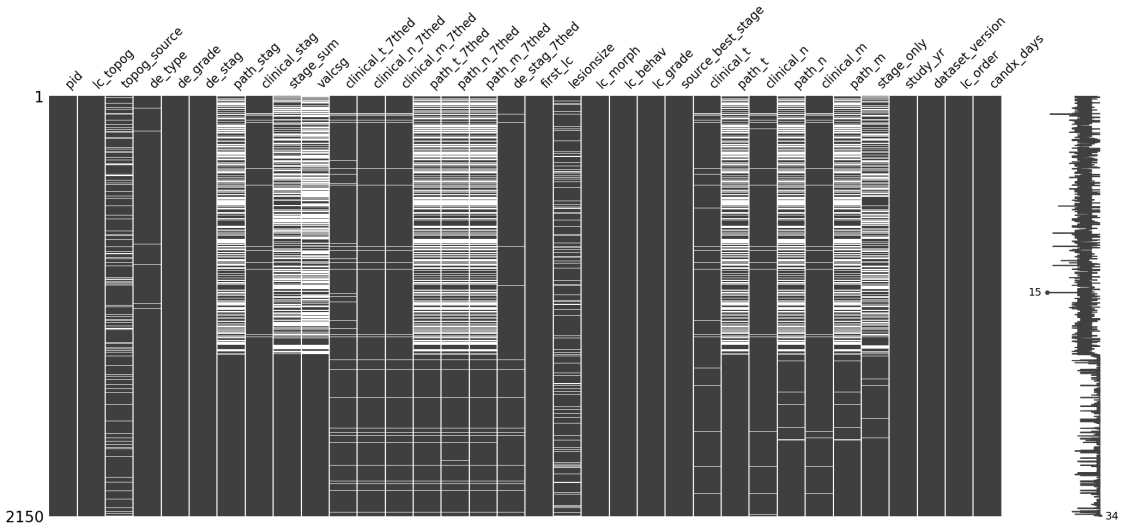
https://cdas.cancer.gov/datasets/nlst/

1.4.8 Accuracy/Completeness

```
The shape of the data set: 2150 rows, and 34 columns
The Counts of various columns
 Counter({dtype('float64'): 22, dtype('int64'): 10, dtype('O'): 2})
There are: 0 Duplicate columns

 (None, None)

There are 0 null Patient IDS
and 2150 total patient records
As well as 2058 unique IDs present
 The % of null columns is: 10.510259917920656 %
```

1.4.9 Missingness

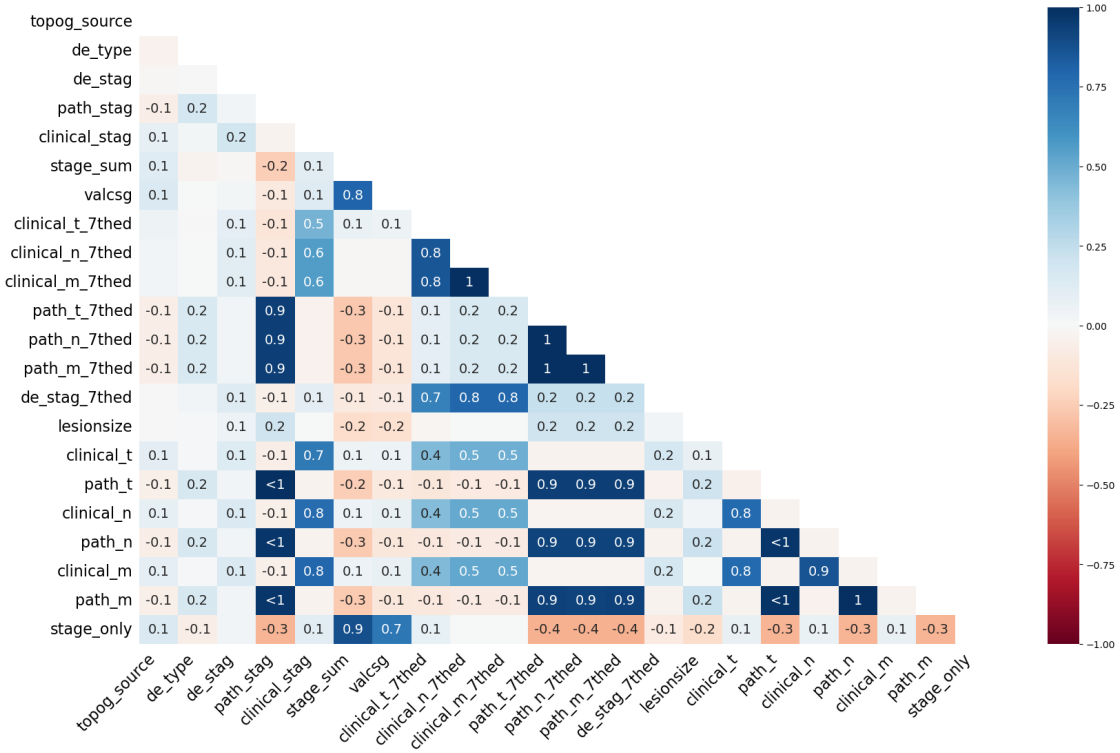
```
This plot uses the missingno package, and plots the missing data vs the non
missing in any dataframe.
    Empty (NAN) values are white and non empty values are black
```



The reasons for the null values in this plot are more nuanced. For example it could be that a patient did not provide the information or the information was lost.

1.4.10 Nullity

```
Nullity is -1 if two variables are mutually exclusive.
           0 if there is no effect.
           1 if there is a strong relationship.
```

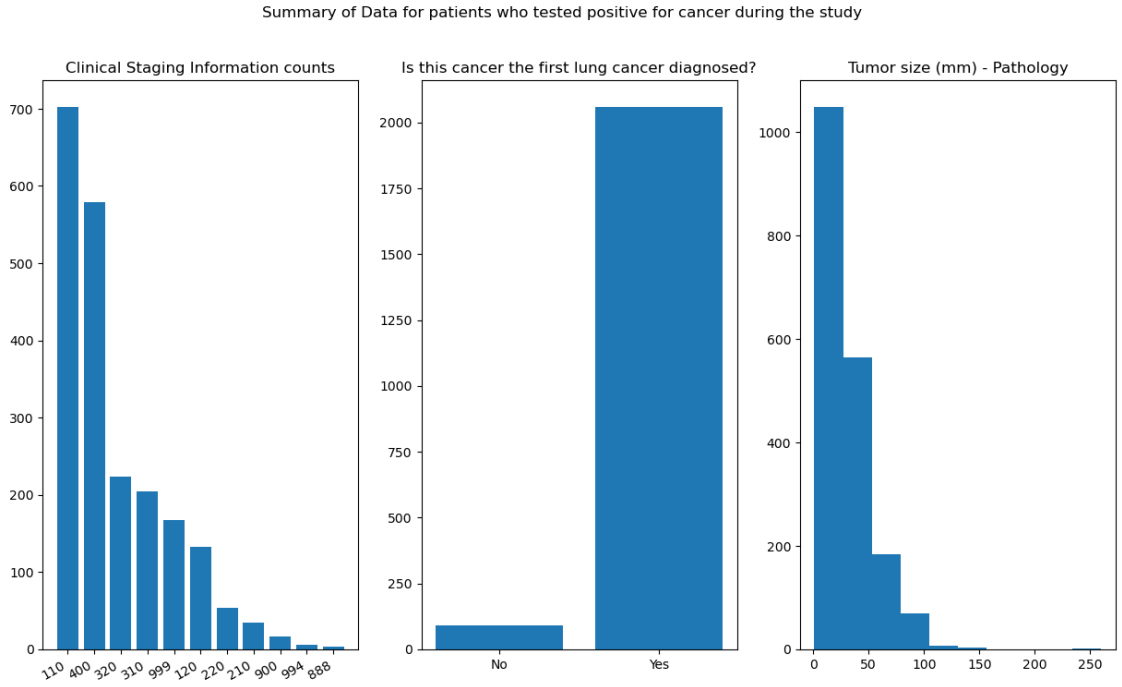


Most of the columns in this data set have very small nullity values.

1.4.11 Correlations

```
The % of values that are highly positively correlated (above 0.8)
5.671077504725898 %
 The % of values that are highly negatively correlated (below -0.8) 0.0 %
```

1.4.12 Interesting plots



Most of the Cancerous abnormalities fell into stage 1, characterized by not much spread. Followed by stage 2 which means they found tumors but just in one lung. And for all but a few participants, this was their first lung cancer diagnosis. Additionally most tumors were very small in size.

1.4.13 The Abnormalities data dictionary and dataset

"The Spiral CT Abnormalities dataset (~177,500, one record per abnormality on CT) contains information about each abnormality observed on the Spiral CT screening exams."

https://cdas.cancer.gov/datasets/nlst/

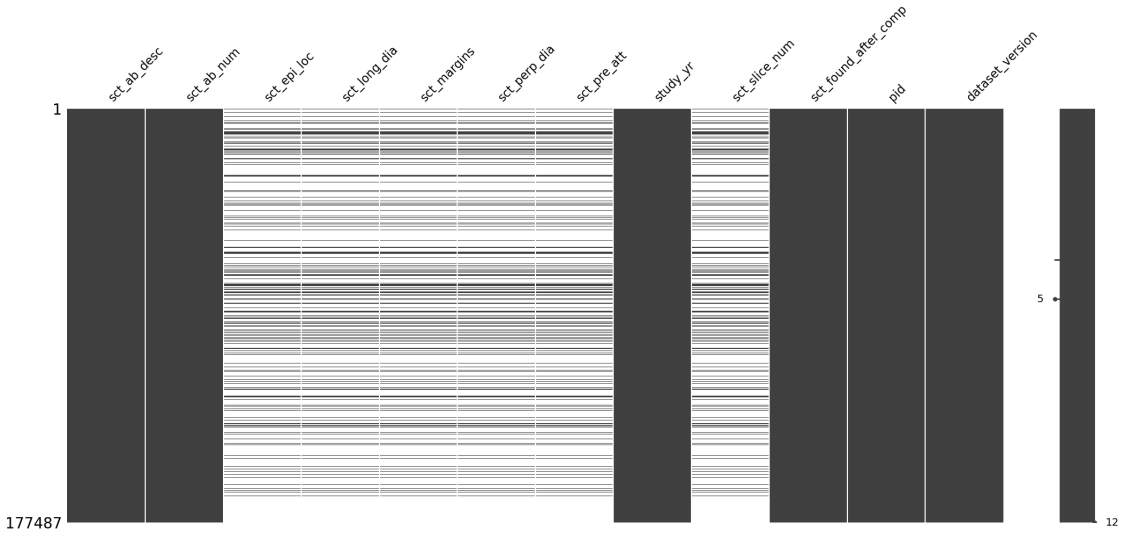
1.4.14 Accuracy/Completeness

```
The shape of the data set: 177487 rows, and 12 columns
The Counts of various columns
 Counter({dtype('float64'): 7, dtype('int64'): 4, dtype('O'): 1})
There are: 0 Duplicate columns

 (None, None)

There are 0 null Patient IDS
and 177487 total patient records
As well as 24517 unique IDs present
 The % of null columns is: 40.04434127569906 %
```

1.4.15 Missingness

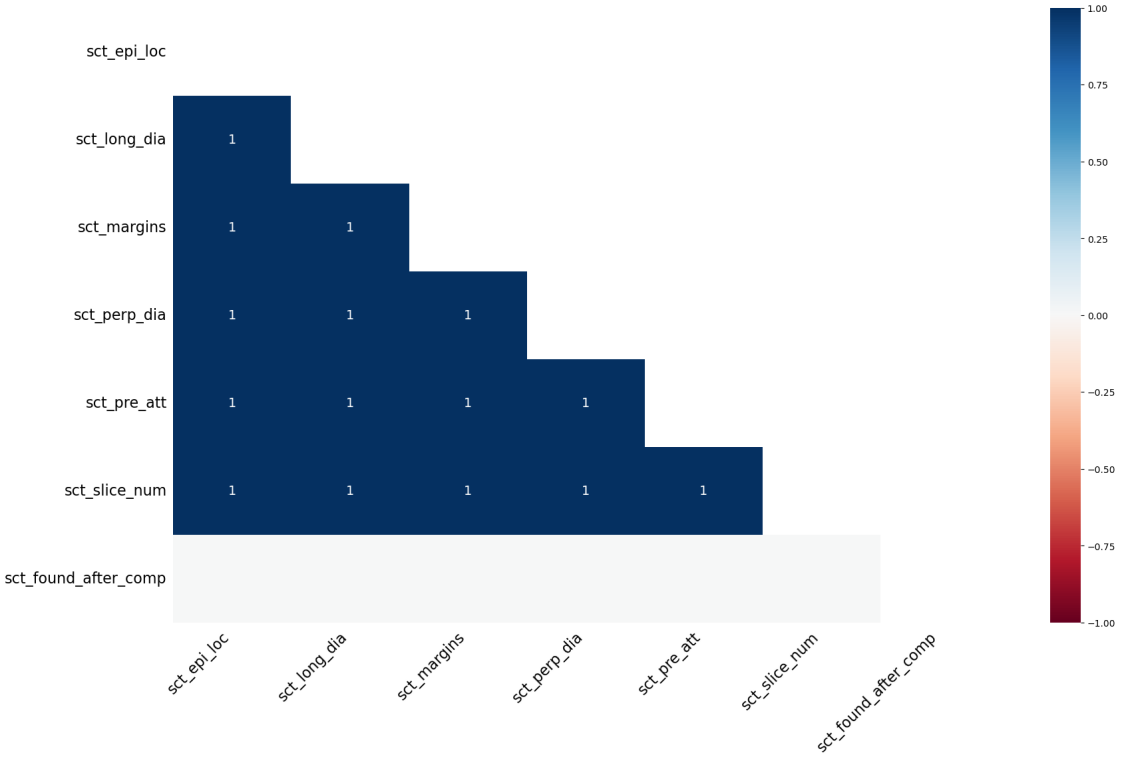
This plot uses the missingno package, and plots the missing data vs the non missing in any dataframe.

Empty (NAN) values are white and non empty values are black



1.4.16 Nullity

```
Nullity is -1 if two variables are mutually exclusive.
          0 if there is no effect.
          1 if there is a strong relationship.
```

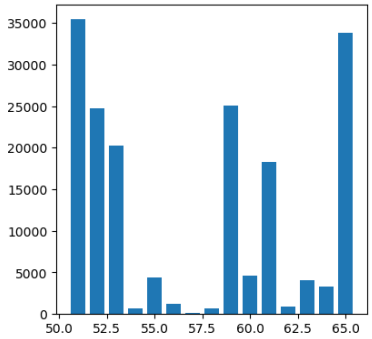


1.4.17 Correlations

```
The % of values that are highly positively correlated (above 0.8)
28.57142857142857 %
 The % of values that are highly negatively correlated (below -0.8) 0.0 %
```

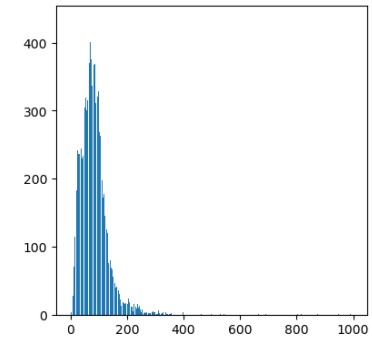
1.4.18 Interesting Plots

SCT abnormality data dictionary



The data in this figure is from the study years T0,T1,T2.
And comes from the combination of all values for said
study years.

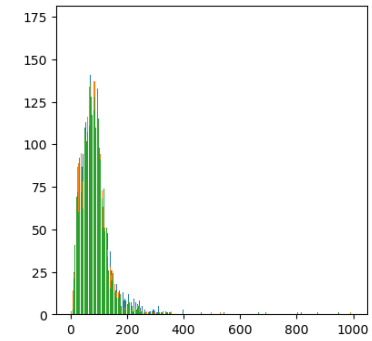
Abnormality description
The type of the abnormality.
Note that the LSS screening forms use
a different numbering system than what is used in this variable.
51="Non-calcified nodule or mass (opacity >= 4 mm diameter)"
52="Non-calcified micronodule(s) (opacity < 4 mm diameter)"
53="Benign lung nodule(s) (benign calcification)"
54="Atelectasis, segmental or greater"
55="Pleural thickening or effusion"
56="Non-calcified hilar/mediastinal adenopathy or mass (>= 10 mm on short axis)"
57="Chest wall abnormality (bone destruction, metastasis, etc.)"
58="Consolidation"
59="Emphysema"
60="Significant cardiovascular abnormality"
61="Reticular/reticulonodular opacities, honeycombing, fibrosis, scar"
62="6 or more nodules, not suspicious for cancer (opacity >= 4 mm)"
63="Other potentially significant abnormality above the diaphragm"
64="Other potentially significant abnormality below the diaphragm"
65="Other minor abnormality noted



When a patient has a CT scan performed,
each slide in that set is annotated with a number.

This plot shows the counts of the CT slice numbers containing an abnormality's
greatest diameter, for non-calcified nodules
or masses with >= 4 mm diameter

It seems that most abnormalities were found
within the first hundred or so slides.
But most data sets contained about 200 images.
And there are few images that contain mising values,
the category for missing is 999.



An overlay of the three study years shows
that year-to-year there
is not much change for where
abnormalities are found.

```
Counter({2: 60438, 1: 59283, 0: 57766})
```

1.4.19 Images

Out of 11 TB of images. There were about 600 patients who fit into the positive category. Of those, I was able to obtain data for roughly 530. Which turned into 1730 or so images. For the negative group. Despite the larger sample pool from which to choose. For the amount of patients I queried and processed (~1500) Many were filtered out for one reason or another (namely too few images in their files, it was how I excluded X ray images) And I ended up with about 130 patient

IDs that fit the proper metrics.

```
2115
```

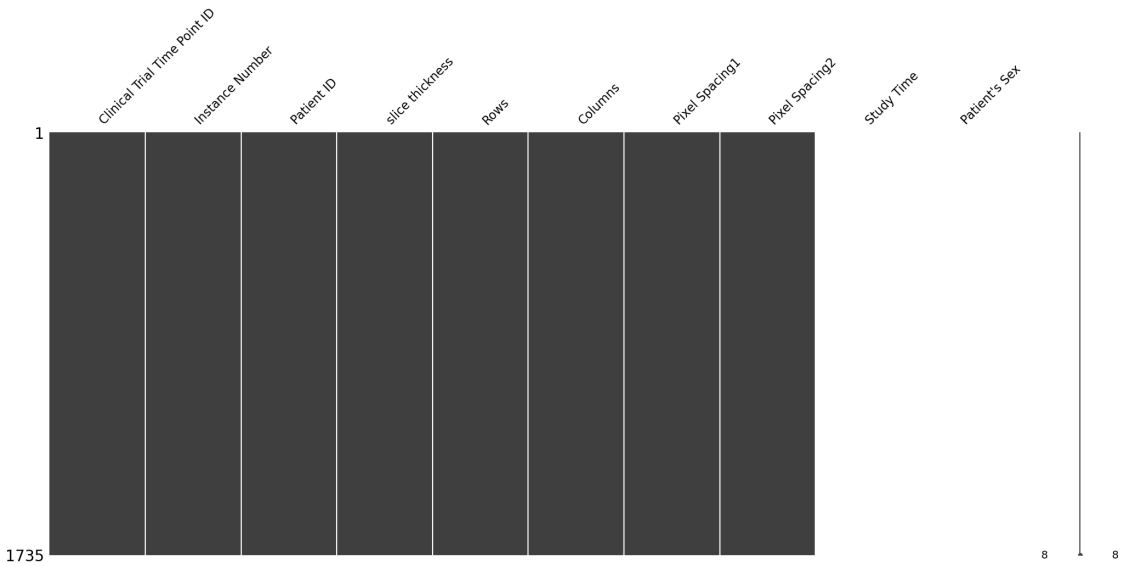
1.4.20 Accuracy/Completeness

```
The shape of the data set: 1735 rows, and 10 columns
The Counts of various columns
 Counter({dtype('int64'): 7, dtype('float64'): 2, dtype('O'): 1})
There are: 18 Duplicate columns

 (None, None)

The shape of the data set: 1735 rows, and 10 columns
The Counts of various columns
 Counter({dtype('int64'): 7, dtype('float64'): 2, dtype('O'): 1})
There are: 0 Duplicate columns

 (None, None)

There are 0 null Patient IDS
and 1735 total patient records
As well as 129 unique IDs present
 The % of null columns is: 0.0 %
There are 0 null Patient IDS
and 1735 total patient records
As well as 566 unique IDs present
 The % of null columns is: 0.0 %
```

1.4.21 Missingness

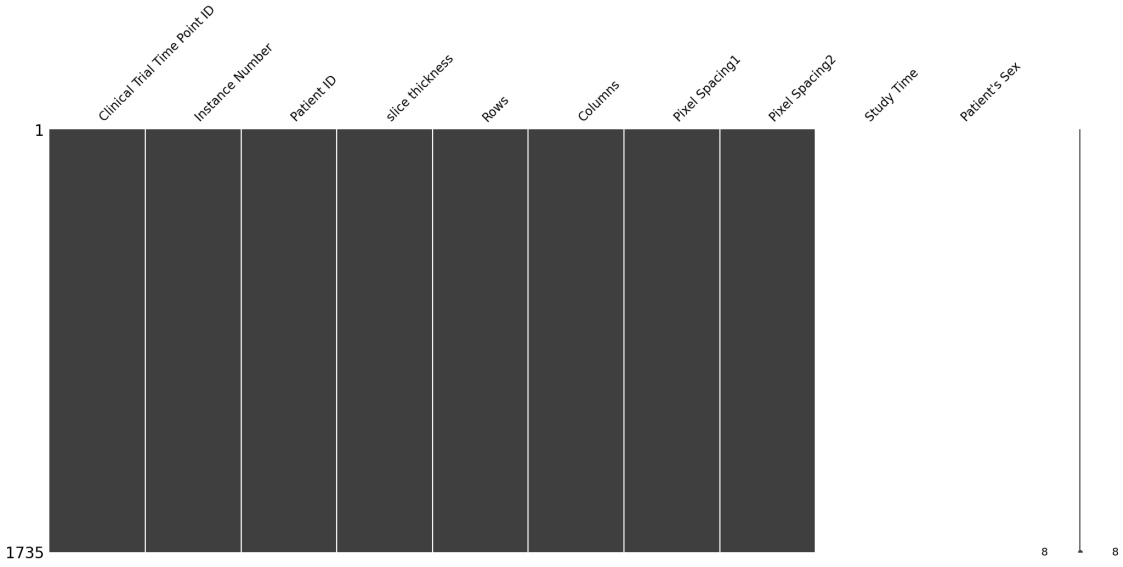
This plot uses the missingno package, and plots the missing data vs the non missing in any dataframe.

Empty (NAN) values are white and non empty values are black



This plot uses the missingno package, and plots the missing data vs the non missing in any dataframe.

Empty (NAN) values are white and non empty values are black

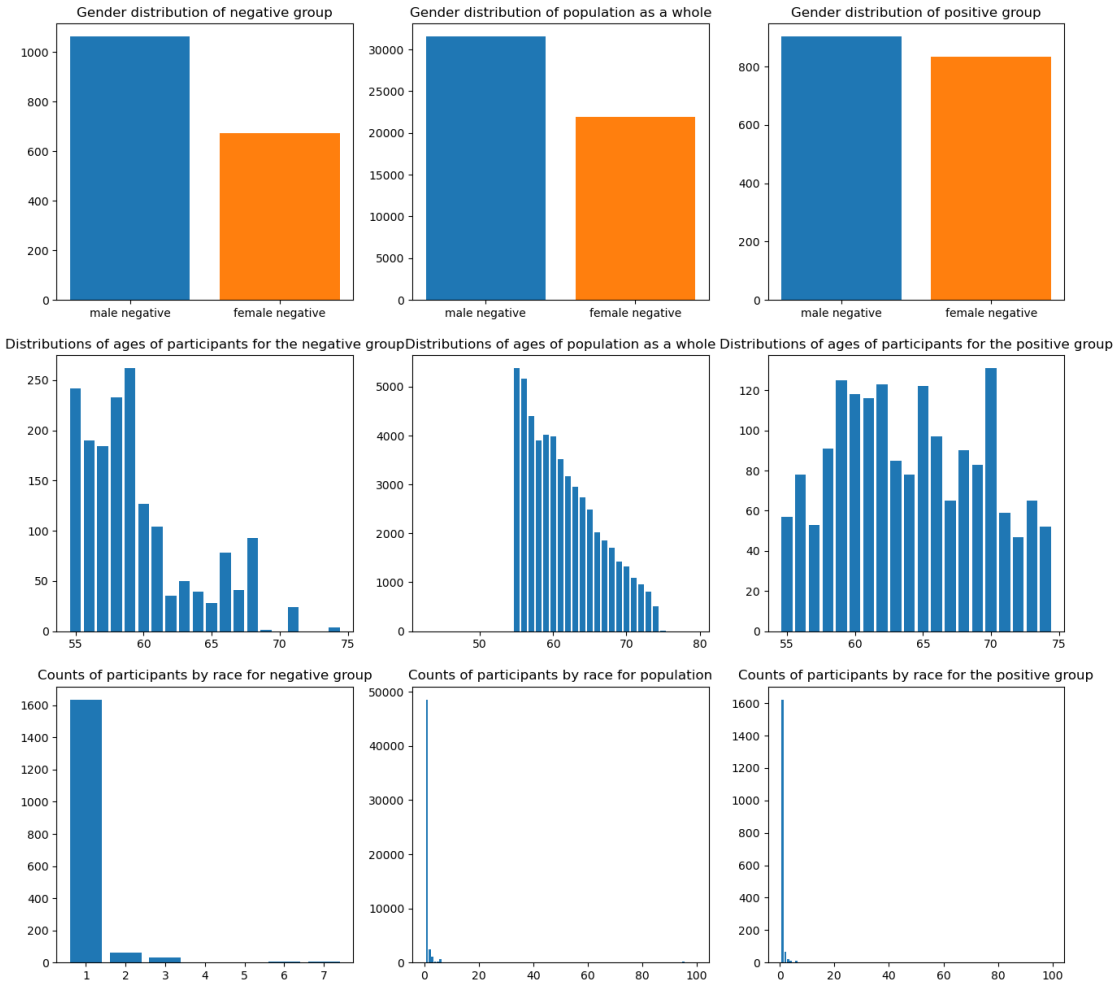


1.4.22 Correlations

```
The % of values that are highly positively correlated (above 0.8)
14.285714285714285 %
 The % of values that are highly negatively correlated (below -0.8) 0.0 %
```

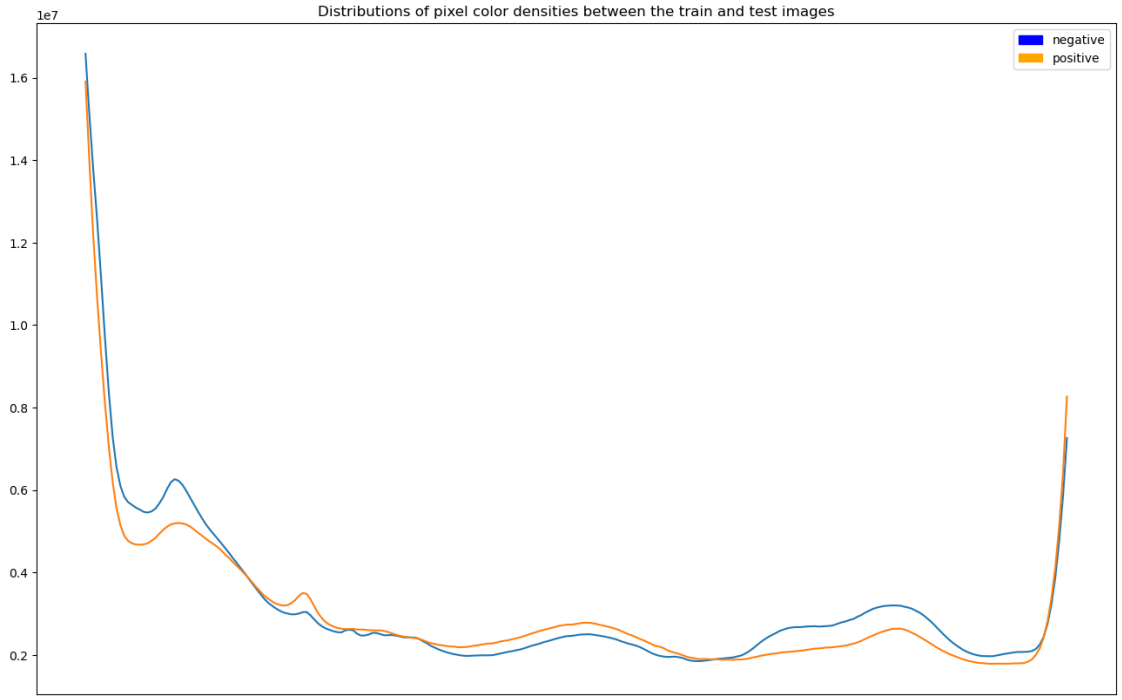

```
The % of values that are highly positively correlated (above 0.8)
14.285714285714285 %
 The % of values that are highly negatively correlated (below -0.8) 0.0 %
```


1.4.23 Interesting Plots

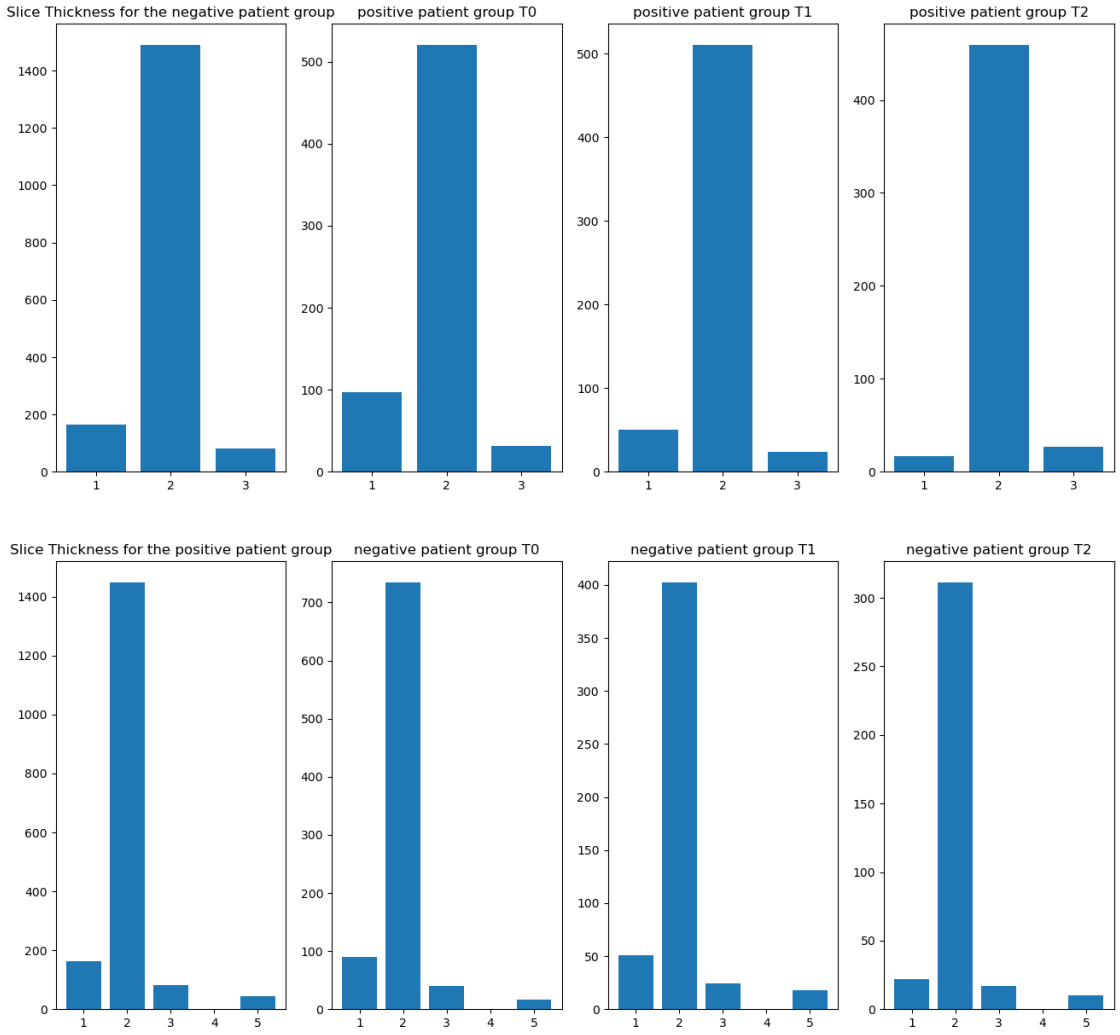


The distribution of males to females and distribution of ages of participants in the negative Cancer group was actually less well proportioned than in the positive cancer group.

```
Counter({2: 1489, 1: 164, 3: 82})
```



An interesting thing to look when working with images is the pixel density. The above plot shows the concentration of pixels in the range of 0-255.For all of the images in the train and test group. The values do not differ very much.




```
[NbConvertApp] Converting notebook Quantitative_assessment_data.ipynb to PDF
[NbConvertApp] Support files will be in Quantitative_assessment_data_files/
[NbConvertApp] Making directory ./Quantitative_assessment_data_files
[NbConvertApp] Writing 40317 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 2170670 bytes to Quantitative_assessment_data.pdf
```