



Data Preprocessing with Python

Machine Learning course, Chapter 3: Data Science, 3rd Session
Exercise number 12

SAMIRA SHEMIRANI #148

08/21/2023

Table of contents

O1

Dataset

In this section, we will have a short description about the dataset.

O2

Methods

Analysis of dataset features visually and description of coding parts are in this section.

O3

Plots

Scatter / Overlap Scatter / Bar plots, will be here.

O4

Conclusion

A brief conclusion about what was done and what was achieved.





01 Dataset

Breast Cancer



What's in the Dataset?

The dataset on which we intend to perform pre-processing is called the female “Breast Cancer” dataset.

- 4024 female patients have their information recorded in this dataset.
- 16 features have been examined for each female patient.
- So, it can be said that this dataset is a table with 4024 rows and 16 columns.
- This dataset is 4024×16 .

The format of this file is “.csv”, and here, in the next column, we name the features of each column:

1. Age
2. Race
3. Marital Status
4. T Stage
5. N Stage
6. 6th Stage
7. Differentiate
8. Grade
9. A Stage
10. Tumor Size
11. Estrogen Status
12. Progesterone Status
13. Regional Node Examined
14. Regional Node Positive
15. Survival Months
16. Status





O2

Methods

Features and histograms.

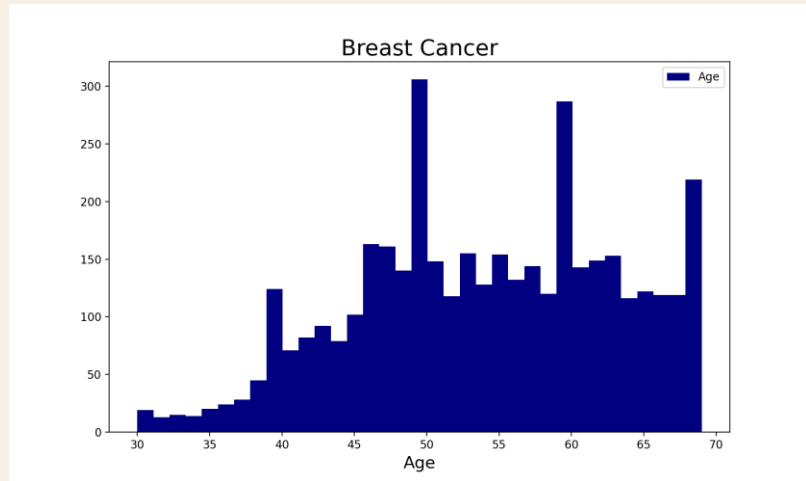
How to code for preprocessing?



Features Description (x axis)

Age

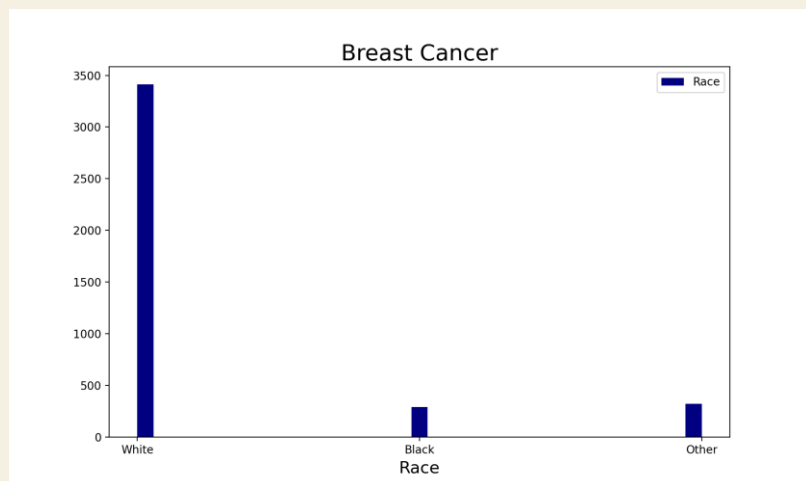
It shows the age of each patient and varies from 30 to 69 years old.



Race

It shows the race of each patient and this column has three different values of White, Black, and Other.

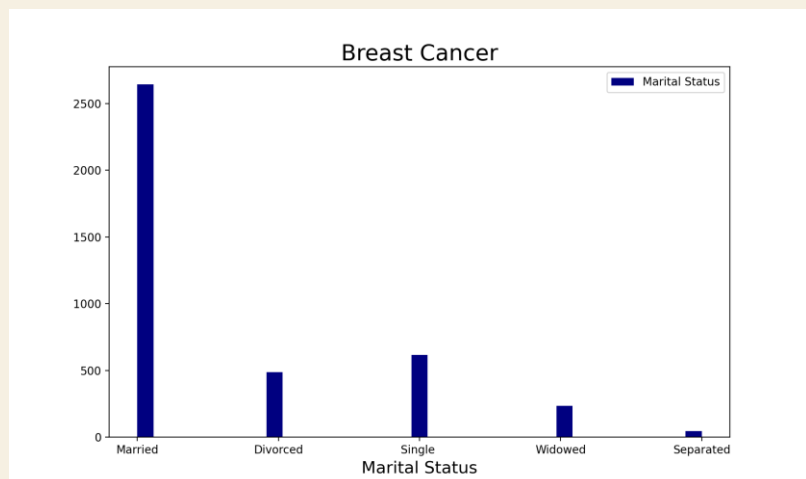
White	3413
Black	291
Other	320



Marital Status

It shows the marital status of each patient and this column has five different values of Married, Single, Divorced, Widowed, and Separated.

Married	2643
Single	615
Divorced	486
Widowed	235
Separated	45



Features Description (x axis)

T Stage

No Histogram Plot obtained
(probably Jupyter Notebook problem)

T followed by a number from 0 to 4 describes the main (primary) tumor's size. Higher T numbers mean a larger tumor and/or wider spread to tissues near the breast.

TX: Primary tumor cannot be assessed.

T0: No evidence of primary tumor.

T1	1603
T2	1786
T3	533
T4	102

N Stage

N Values for Breast Cancer Staging.

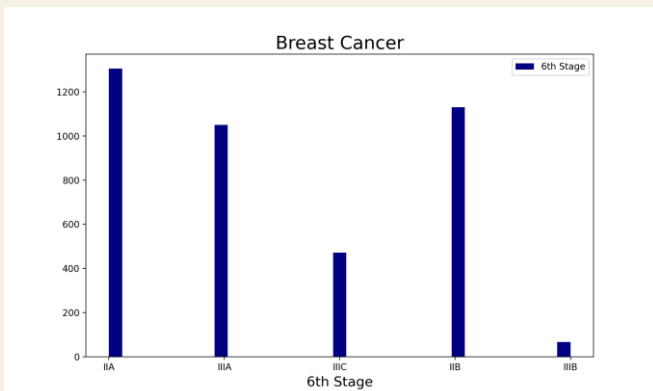
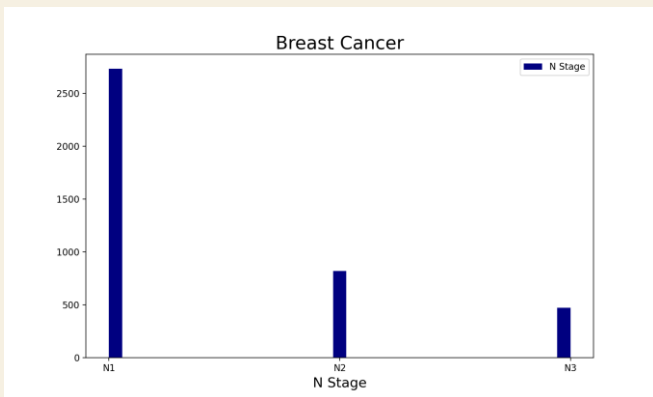
NX: Lymph nodes are unable to be evaluated.
N0: There is no spread to nearby lymph nodes.
N1: Cancer has spread to fewer than 3 lymph nodes located on the underarm or has spread to any number of lymph nodes located near the breastbone.

N1	2732
N2	820
N3	472

6th Stage

Stage I, Stage IIA and Stage IIB (early) refer to early breast cancer. Stage IIB (advanced), Stage IIIA, Stage IIIB, Stage IIIC and Stage IV refer to advanced breast cancer (locally advanced breast cancer or metastatic breast cancer).

IIA	1305
IIB	1130
IIIA	1050
IIIB	67
IIIC	472



Features Description (x axis)

Differentiate & Grade

Grade 1 means that the cells are **well-differentiated**, meaning they look like normal cells (In this case, their appearance is similar to normal breast tissue, and they are growing and spreading slowly).

Grade 2 or **moderately differentiated**; The cells are growing at a speed of and look like cells somewhere between grades 1 and 3.

Grade 3 or **poorly differentiated**; The cancer cells look very different from normal cells and will probably grow and spread faster.

Lower grade cancer cells tend to be slow growing and are less likely to spread. **High grade** means that the cancer cells are poorly differentiated or undifferentiated.

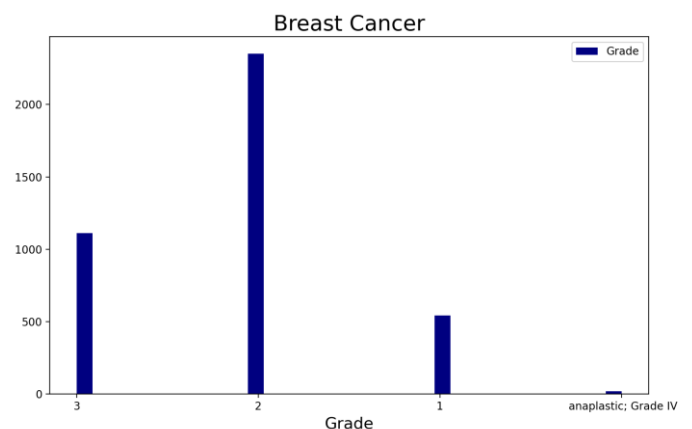
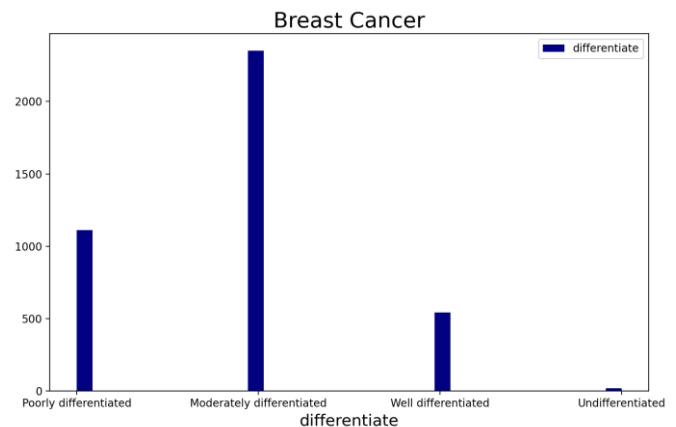
Secondary breast cancer is also called **metastatic breast cancer**, **advanced breast cancer**, or **stage 4 breast cancer**; It means that the cancer has spread to other distant parts of the body, such as the liver or bones. Breast implant-associated **anaplastic** large cell lymphoma is also called BIA-ALCL. It's a rare form of lymphoma that occurs in some people who've had breast implants. It's a type of immune system cancer and isn't breast cancer.

A Stage

There are two types of A Stage;

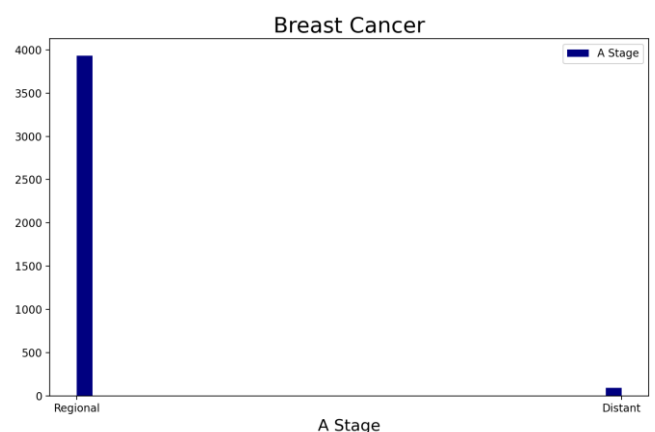
- 1) **Regional**: The lymph nodes, primarily those in the armpit, are involved.
- 2) **Distant**: The cancer is found in other parts of the body as well.

Regional	3932
Distant	92



1	543
2	2351
3	1111
Grade IV	19

Well differentiated	543
Moderately differentiated	2351
Poorly differentiated	1111
Undifferentiated	19



Features Description (x axis)

Tumor Size

It shows the size of the tumor which is inside the body and it varies from 1 to 140 (cm).

Estrogen Status

The cells of this type of breast cancer have receptors that allow them to use the hormone estrogen to grow. Treatment with anti-estrogen hormone (endocrine) therapy can block the growth of the cancer cells.

Positive	3755
Negative	269

Progesterone Status

This type of breast cancer is sensitive to progesterone, and the cells have receptors that allow them to use this hormone to grow. Treatment with endocrine therapy blocks the growth of the cancer cells.

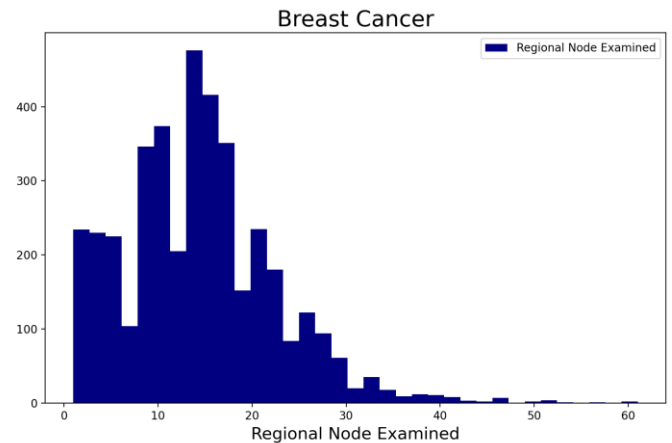
Positive	3326
Negative	698



Features Description (x axis)

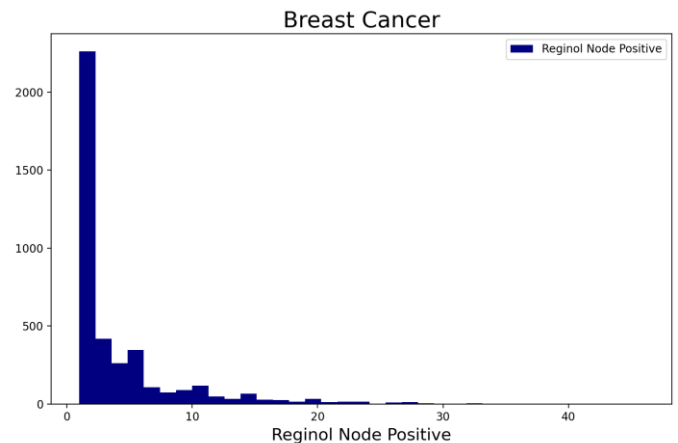
Regional Node Examined

Regional lymph nodes relevant to breast cancer; The axillary lymph nodes receive the majority of the lymphatic drainage from all quadrants of the breast; the remainder drains to the internal mammary, infraclavicular, and/or supraclavicular lymph nodes. It varies from 1 to 61 in this dataset.



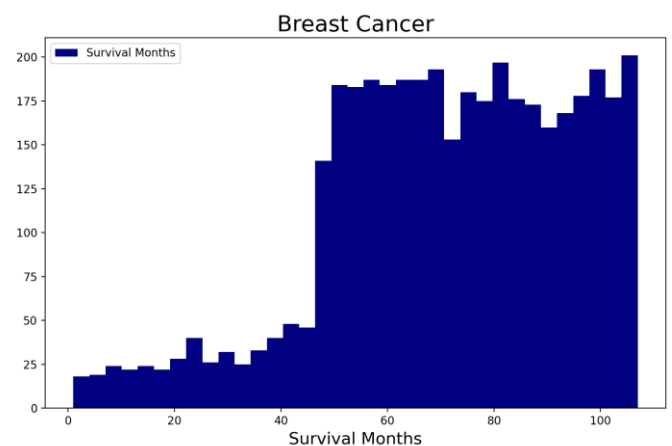
Reginol Node Positive

Lymph node-positive breast cancer is a type of cancer that has spread from the original tumor to the nearest lymph nodes, near or in patient armpit. This is known as regional spread or locally advanced breast cancer. This value varies from 1 to 46 in this dataset.



Survival Months

Survival Months of the patients varies between 1 to 107 months in this dataset.

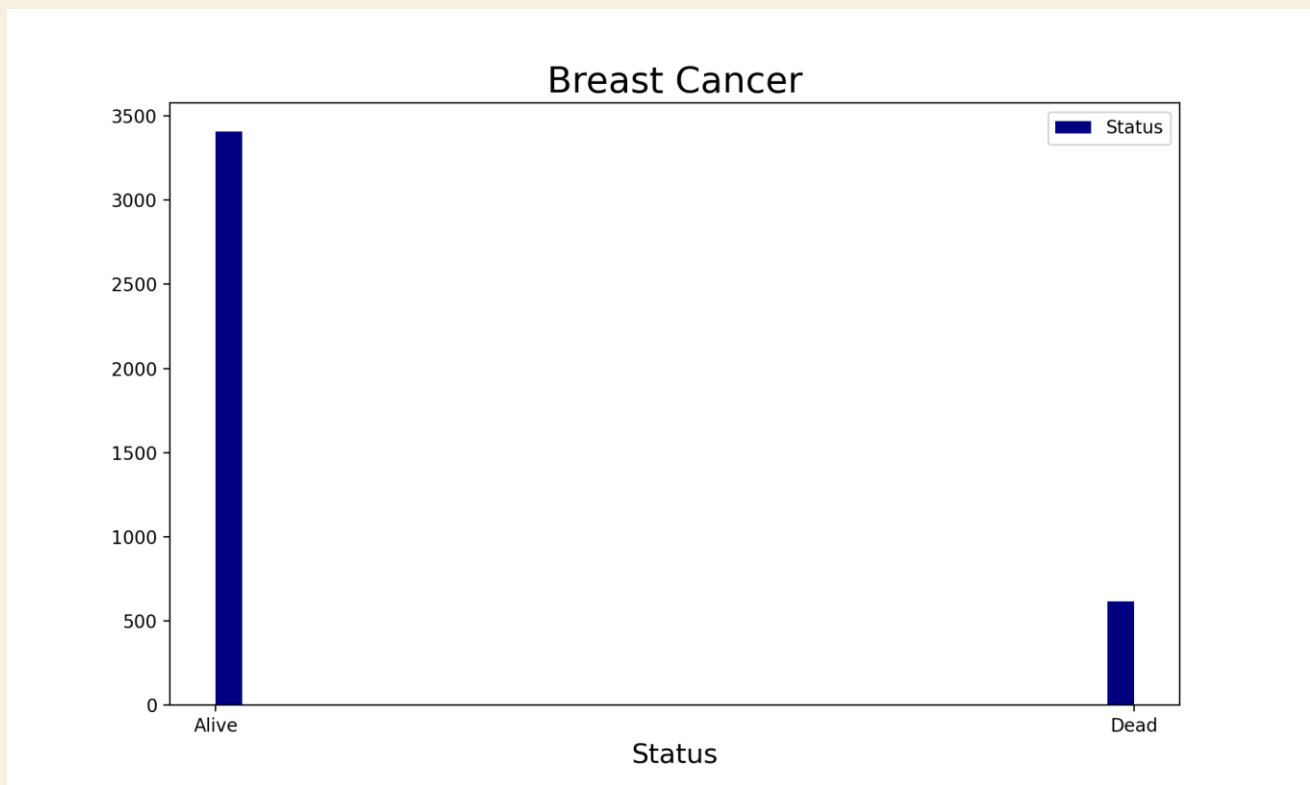


Target Description (y axis)

Status

This value indicates whether the patient is dead or alive.

Alive	3408
Dead	616



What we've done in the coding part?



Dataset

Reading the data and creating the DataFrame



.describe()

Finding out some info's & if there is any MV's



Missing Values

Finding MV's and removing them (.dropna())



Plotting plots

Implementing plots codings

For the coding part, first we checked the dataset in Excel and Notepad++, and then we went to the Jupyter coding environment and did the following:



What's going on in our Jupyter Notebook?

Libraries

1 – We imported important libraries.

```
1 # importing libraries
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
```

Dataset Reading

2 – We called the dataset into the coding environment, Jupyter Notebook.

```
1 # uploading and showing our data
2 data = pd.read_csv ("D:/IMT/3- Data Science/3- Data pre-processing (2)/Breast_Cancer.csv")
3 data
```



What's going on in our Jupyter Notebook?

Features & Target Recognition

3 – We decided which columns of Features are supposed to be our Target and which ones will remain as they are to be Features.

Features																Target
	Age	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Reginol Node Positive	Survival Months	Status
0	68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive	24	1	60	Alive
1	50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive	14	5	62	Alive
2	58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive	14	7	75	Alive
3	58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	18	Positive	Positive	2	1	84	Alive
4	47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	41	Positive	Positive	3	1	50	Alive
...

Creating DataFrame

4 – We made a DataFrame of the desired columns and rows.

```
1 df = pd.DataFrame (data) # DataFrame with all the columns
2 df
```



What's going on in our Jupyter Notebook?

.describe()

5 – We used to describe command to understand the statistical information of the dataset. Here we will find out if there is a Missing Value (MV).

```
1 df.describe() # describe in the numeric way of the numeric columns
```

```
1 df.describe(include=object) # describe in the alphabetic way of the verbal columns
```

	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Estrogen Status	Progesterone Status	Status
count	4024	4024	4024	4024	4024	4024	4024	4024	4024	4024	4024
unique	3	5	4	3	5	4	4	2	2	2	2
top	White	Married	T2	N1	IIA	Moderately differentiated	2	Regional	Positive	Positive	Alive
freq	3413	2643	1786	2732	1305	2351	2351	3932	3755	3326	3408

No Missing Values

Count shows; the number of rows/values in each column, if it does not have the same value as the rest, it means there is/are Missing Values. Here there is no MVs.

Unique shows; that there are several different values in each column, for example, we had three races (white, black, and other), now it shows us the value of 3. Here all the values are correct, as we had previously estimated by Notepad++ and Excel.

Top shows: which of the samples has been repeated the most? For instance, in the first column, as we have already noticed, the white race is repeated the most. The rest of the columns are the same.

Frequency shows: that how many times **Top** has been repeated?



What's going on in our Jupyter Notebook?

.dropna()

6- If we noticed in the previous section that we have MVs in the dataset, in this section, we will delete it with the `.dropna()` command.

We didn't have any missing values here, so the `.dropna()` has no effect on this DataFrame.

1	<code>df.describe() # describe in the numeric way of the numeric columns</code>				
No Missing Values	Age	Tumor Size	Regional Node Examined	Reginol Node Positive	Survival Months
count	4024.000000	4024.000000	4024.000000	4024.000000	4024.000000
mean	53.972167	30.473658	14.357107	4.158052	71.297962
std	8.963134	21.119696	8.099675	5.109331	22.921430
min	30.000000	1.000000	1.000000	1.000000	1.000000
25%	47.000000	16.000000	9.000000	1.000000	56.000000
50%	54.000000	25.000000	14.000000	2.000000	73.000000
75%	61.000000	38.000000	19.000000	5.000000	90.000000
max	69.000000	140.000000	61.000000	46.000000	107.000000



What's going on in our Jupyter Notebook?

Plots

7- Then, we plot different plots, which we will explain in the next section (here is a sample code of scatter plotting).

```
1 plt.figure(figsize=(10, 6), dpi=80) # Maximize the plot
2
3 plt.scatter(df_new['Age'], df_new['Tumor Size'], s=50, c='rebeccapurple', alpha=0.5, marker=r'$\clubsuit$', label='luck')
4
5 plt.legend(['Age'], loc = 'best')
6 plt.title ("Breast Cancer", fontsize=20)
7
8 plt.xlabel('Age', fontsize=15)
9 #plt.xticks(rotation=90) # Rotating the x labels
10 plt.ylabel('Tumor Size', fontsize=15)
11
12 plt.grid()
13
14 #plt.savefig ('D:/IMT/3- Data Science/3- Data pre-processing (2)/1-Age&Tumor_Size.png')
15
16 plt.show()
```

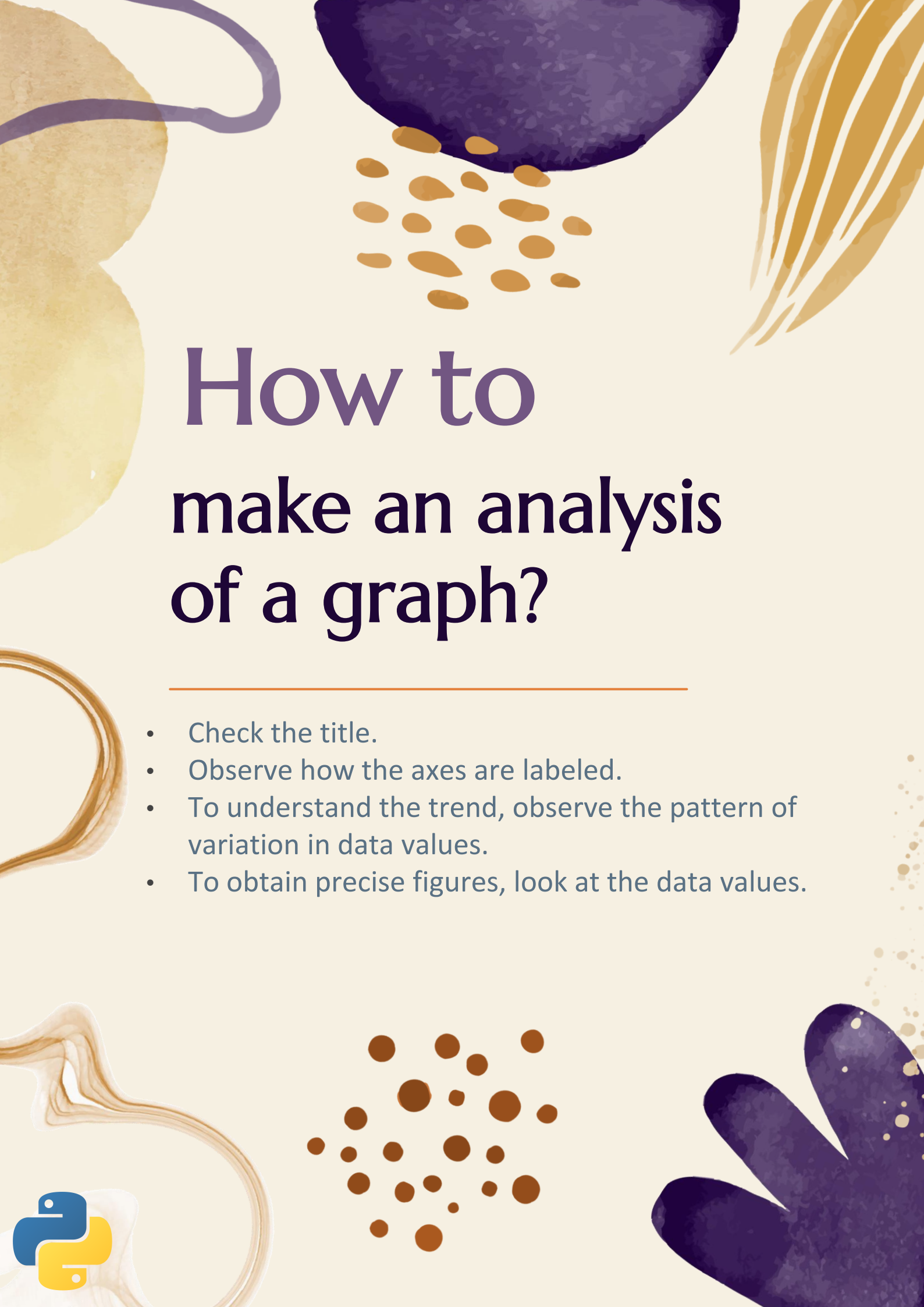


O3

Plots

Scatter / Overlap Scatter / Bar





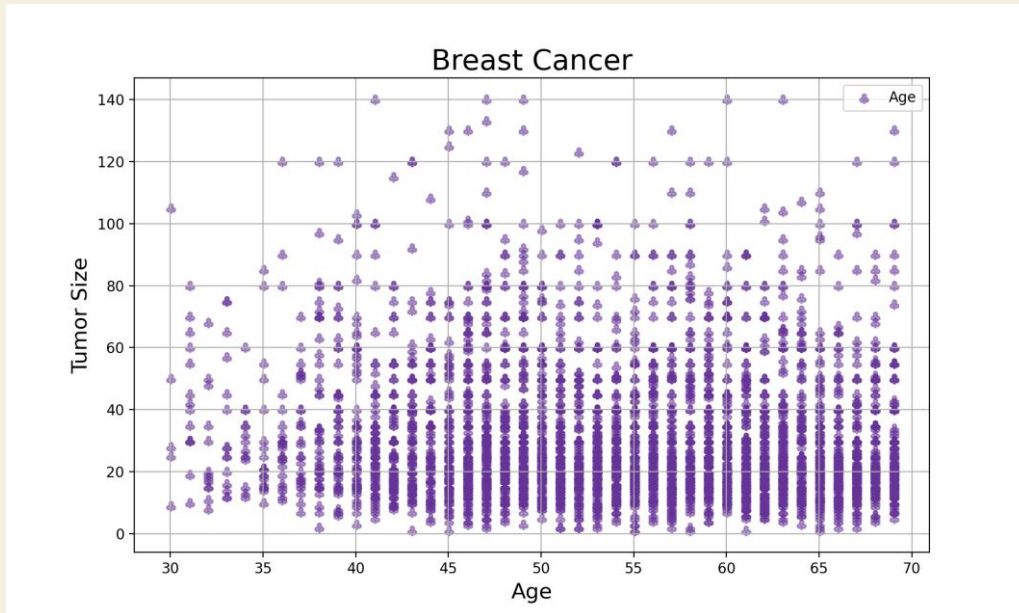
How to make an analysis of a graph?

- Check the title.
- Observe how the axes are labeled.
- To understand the trend, observe the pattern of variation in data values.
- To obtain precise figures, look at the data values.



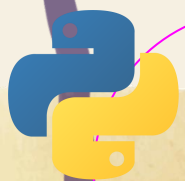
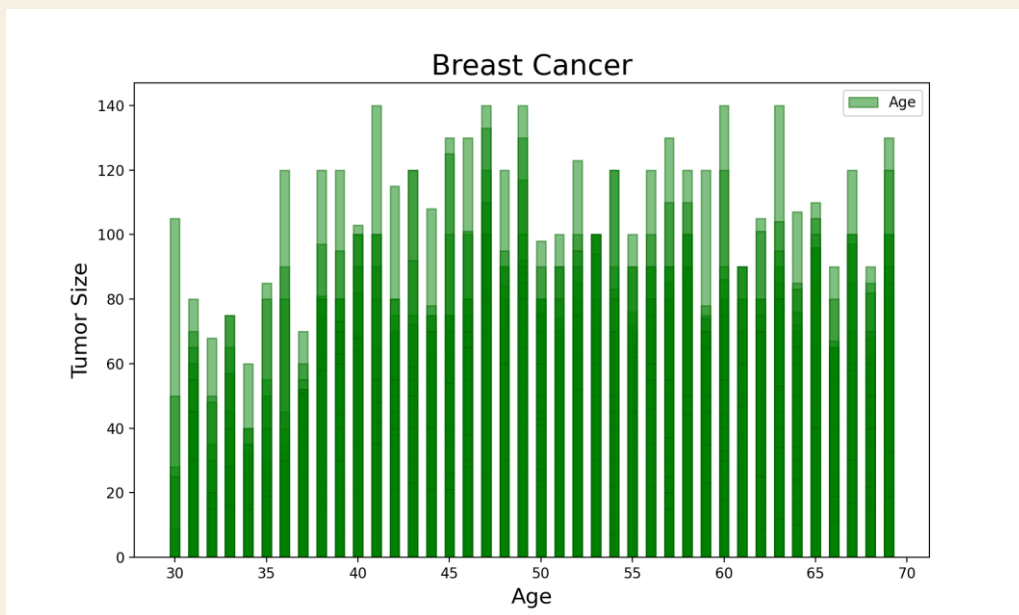
Scatter Plot #1

Age vs. Tumor Size



Bar Plot #1

Age vs. Tumor Size



Plot #1 description



What is the title of the plot showing?

These two graphs (Scatter & Bar) both talk about (increasing) changes in Age compared to (increasing) changes in *Tumor Size* in breast cancer.

What does the x-axis indicate?

The x-axis, Age, is changing from about 30 years old to about 70 years old.

What does the y-axis indicate?

The y-axis, *Tumor Size*, is changing from 0 to about 140 (cm, probably).

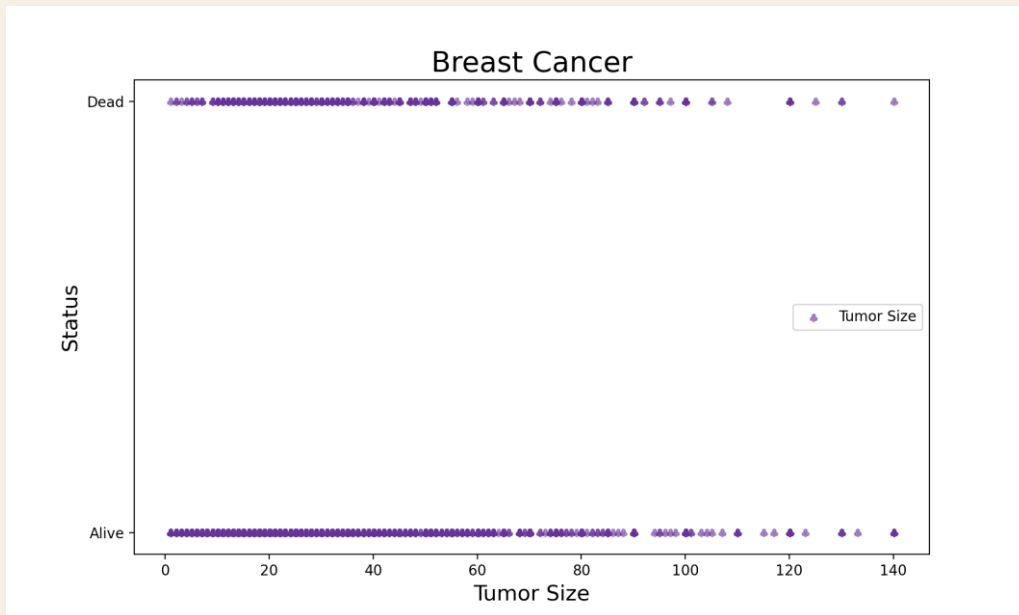
What is the trend of the values?

- This graph is the result of comparing the two features of Age and Tumor Size in relation to each other.
- Since the investigation of these two features is interesting to us, this plot was drawn.
- This chart shows Age with 5-year intervals and Tumor Size with 20-year intervals.
- **The density of the graph** on the x-axis, that is, Age, is in the range of about 35 to about 68, and on the y-axis, that is, the Size of the Tumor, it is in the range of about 3 to about 75.
- **The peak of** Tumor Size generally occurred at 130 and sporadically this peak for Age occurred at 45 to 55.
- It can also be understood from the graph that the most common Tumor Size in patients is size 15 to 40.
- **The trend of the values is**, as the Age goes up from 35 years (that is, up to 60 years), the Tumor Size changes from more than 0 to 50, so that the density of the graph is higher in this interval.



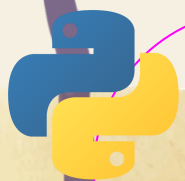
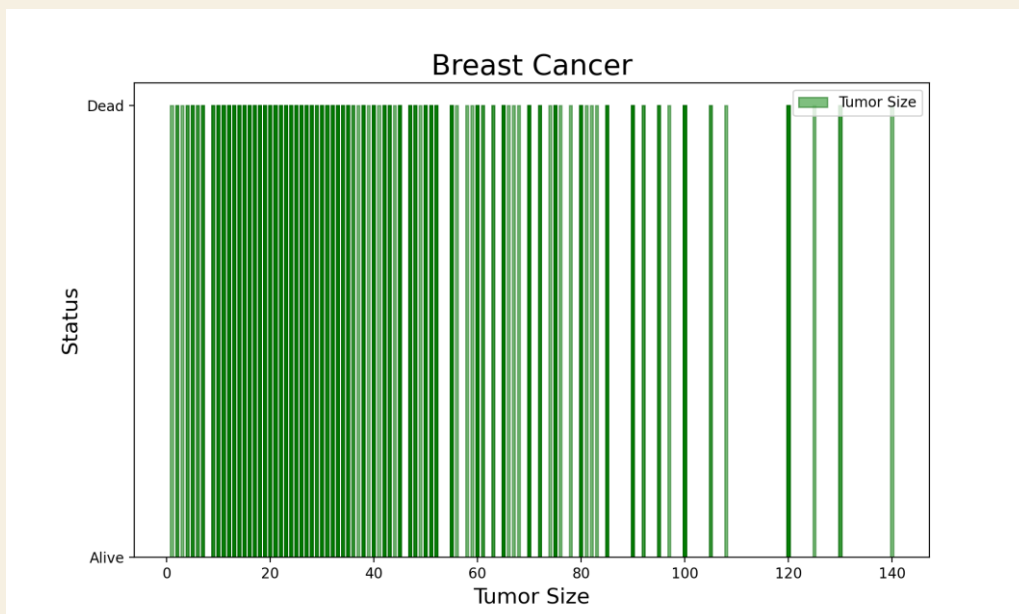
Scatter Plot #2

Tumor Size vs. Status



Bar Plot #2

Tumor Size vs. Status



Plot #2 description



What is the title of the plot showing?

These two graphs (Scatter & Bar) both talk about (increasing) changes in *Tumor Size* compared to being *Alive* or *Dead* (in *Target, Status*).

What does the x-axis indicate?

The x-axis, *Tumor Size*, is changing from 0 to about 140 (cm, probably).

What does the y-axis indicate?

The y-axis, *Status*, is *Alive* or *Dead*, just these two values.

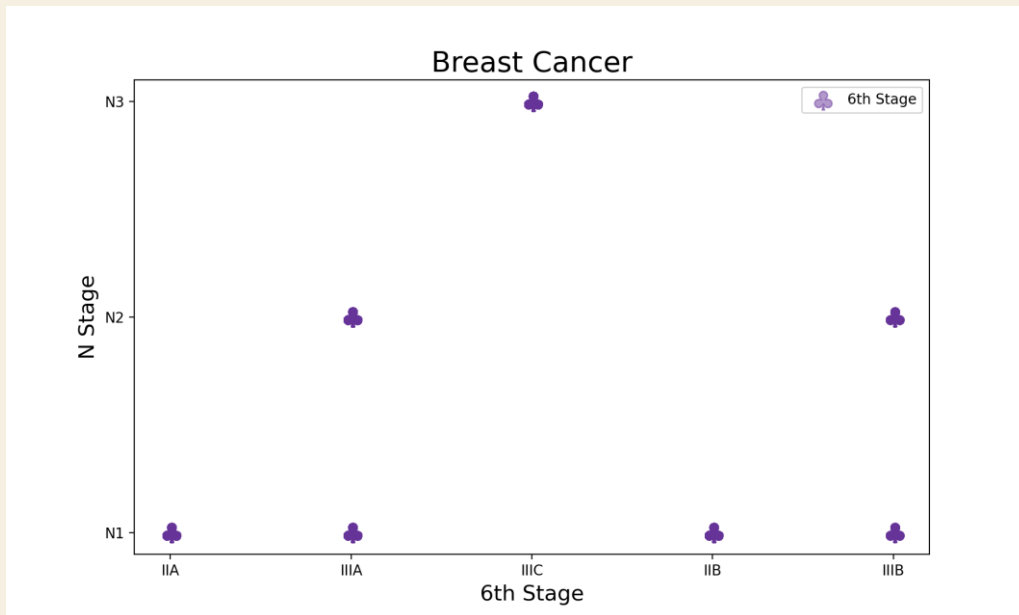
What is the trend of the values?

- This graph shows the *Tumor Size* compared to the patient's *Status*.
- This reviews one of the *features* to the *Target*.
- This chart shows the *Tumor Size* with 20-year intervals, and the *Status*, whether being *Alive* or *Dead*.
- **The density of the graph** on the x-axis, that is, *Tumor Size*, is in the range of about 15 to about 50, and on the y-axis, that is, the *Status*, there is more density in the *Alive* rather than the *Dead*.
- **The peak of** *Tumor Size* generally occurred at 140, and sporadically this peak for *Status* occurred the most in the *Alive* values.
- It can also be understood from the graph that the most common *Tumor Size* in patients, who are *Alive* now, is in the range of 0 to 80, and those who are *Dead*, were in a range of 0 to 70.
- **The trend of the values is**, as the *Tumor Size* increases, the number of patients decreases, which means that most patients have smaller *Tumor Size*. In other words, many patients have a *Tumor Size* of about 15 to about 80, and a few have a larger *Tumor Size*.



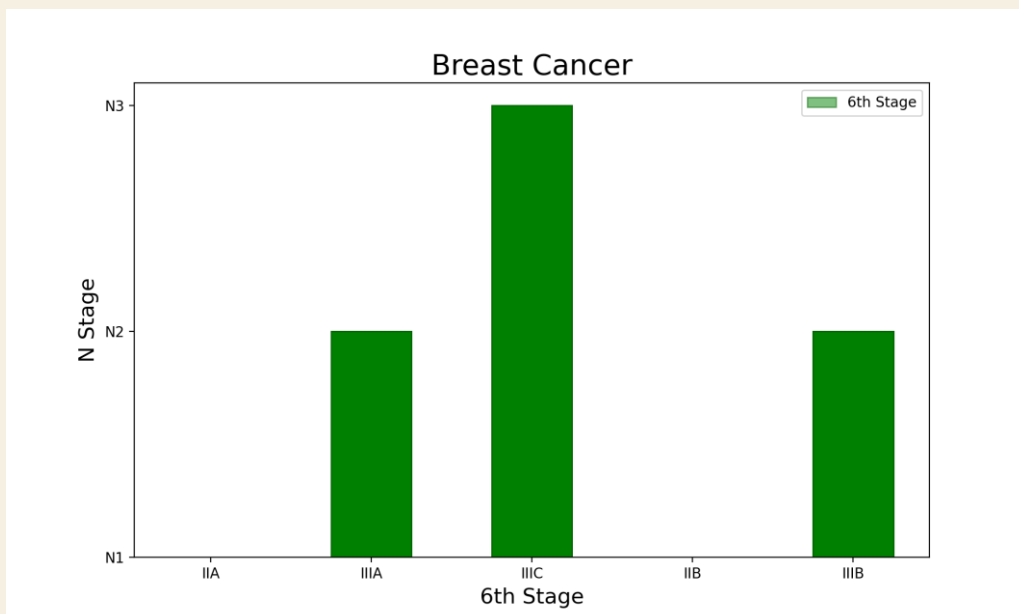
Scatter Plot #3

6th Stage vs. N Stage



Bar Plot #3

6th Stage vs. N Stage



Plot #3 description



What is the title of the plot showing?

These two graphs (Scatter & Bar) both talk about changes in *6th Stage* compared to changes in *N Stage*.

What does the x-axis indicate?

The x-axis, *6th Stage*, indicates the value of five different measurements (IIA, IIB, IIIA, IIIB, and IIIC).

What does the y-axis indicate?

The y-axis, *N Stage*, indicates the value of three different measurements (N1, N2, and N3).

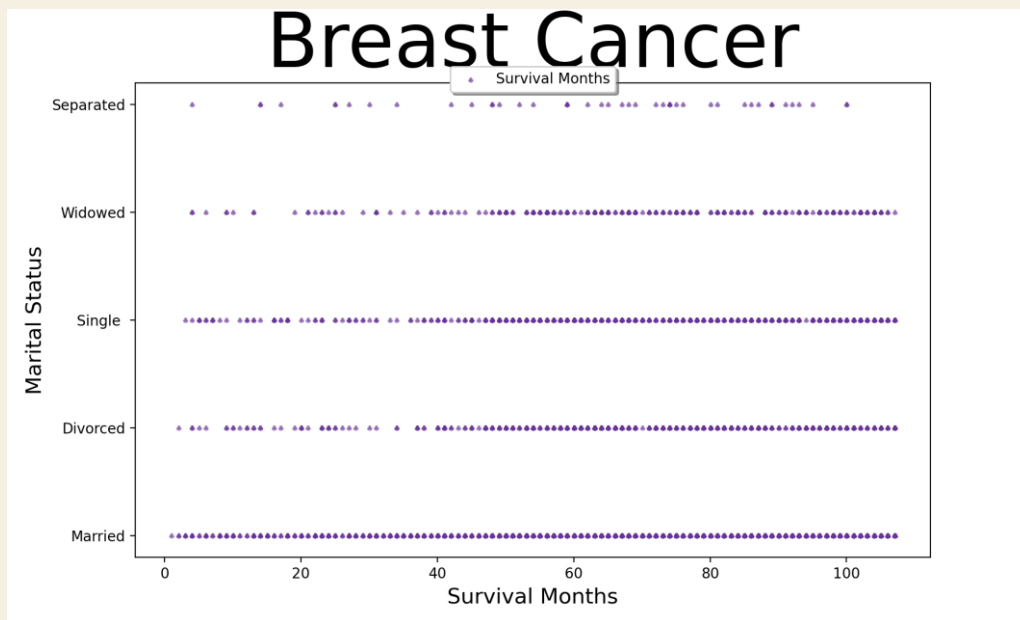
What is the trend of the values?

- This graph shows the *6th Stage* compared to the *N Stage*.
- These are both features of this dataset, which we plotted based on a personal desire to see their changes relative to each other.
- If you pay attention to the Scatter plot, on the x-axis, in IIA there is only value N1 on the y-axis, in IIIA there are values N1 and N2. In IIIC there is a value of N3, in IIB there is only a value of N1, and in the last value, which is a value of IIIB, we have the values of N1 and N2. In other words, for example, in the value of IIIC, we will not have a value of N1 and N2. But if you pay attention to the Bar plot, with the occurrence of IIIC we will have the highest value of *N Stage* axis where N3 occurs. And also, the occurrence of IIA and IIB have the lowest value of *N Stage* axis where N1 occurs. It should be noted that the correct order of *6th Stage* for its values is: IIA < IIB < IIIA < IIIB < IIIC, and the correct order of *N Stage* for its values is as follows: N1 < N2 < N3.
- Some values are not visible in the Bar plot that we can see in the Scatter plot.



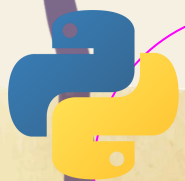
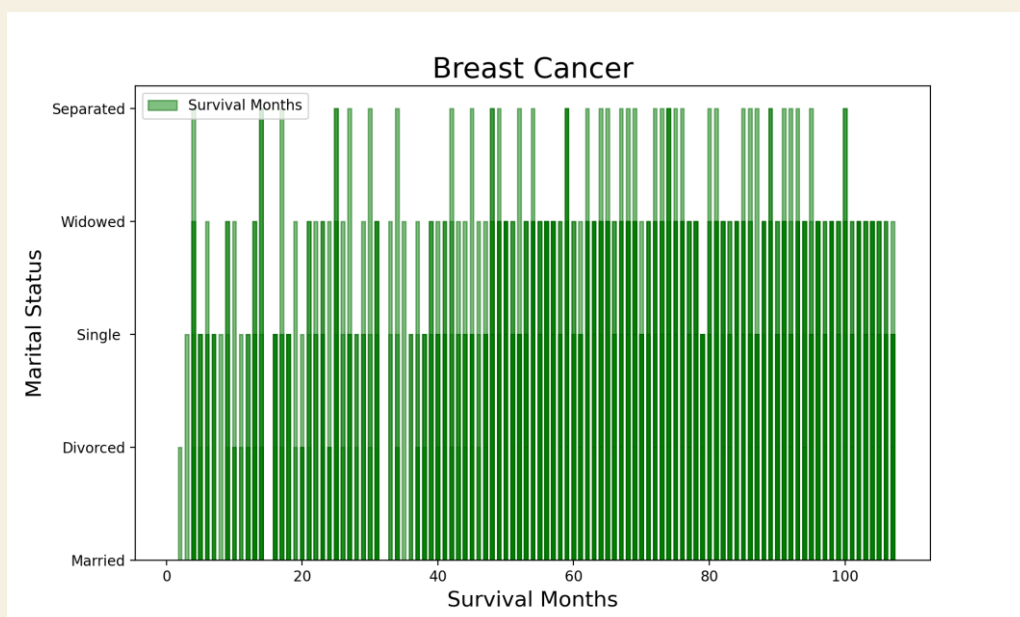
Scatter Plot #4

Survival Months vs. Marital Status



Bar Plot #4

Survival Months vs. Marital Status



Plot #4 description



What is the title of the plot showing?

These two graphs (Scatter & Bar) both talk about (increasing) changes in *Survival Months* compared to changes in *Marital Status*.

What does the x-axis indicate?

The x-axis, *Survival Months*, is changing from about 0 to more than about 100 months.

What does the y-axis indicate?

The y-axis, *Marital Status*, has five different values (Married, Single, Divorced, Widowed, and Separated).

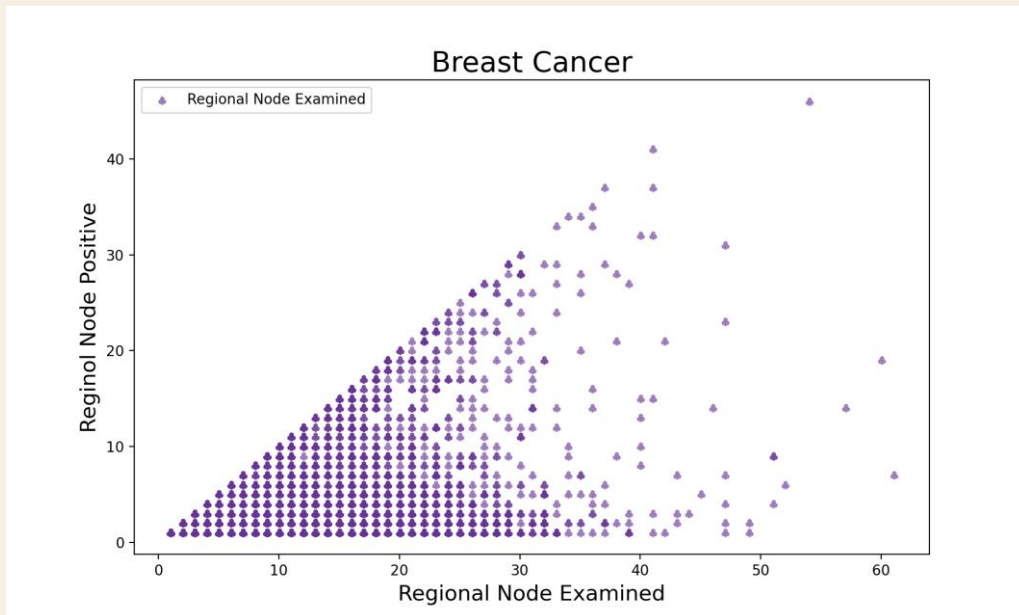
What is the trend of the values?

- This graph is the result of comparing the two features of *Survival Months* and *Marital Status* in relation to each other.
- Since the investigation of these two features is interesting to us, this plot was drawn.
- This chart shows *Survival Month* with 20-year intervals.
- As can be seen, the density of the graph on the x-axis is *higher* in the right part of the graph, i.e., from 50 to 100 *Survival Months* of marriage, and on the y-axis, in the *Married* section.
- The peak of *Survival Months* generally occurred at 40 to 100, and this peak for *Marital Status* occurred at *Married*.
- It can also be seen from the graph that, for *Marrieds*, their *Survival Months* are from 0 to 100, for *Divorcees*, their *Survival Months* are more than 40 to 100, for *Singles*, this amount is similar to *Divorced* and *Married* (maybe something in between), for *Widowed*, most of the *Survival Months* are around 50 to 100, and for *Separated*, these values are scattered, but this dispersion is more concentrated in 60 to 90 *Survival Months*.
- The trend of the values is, as the *Survival Month* *goes up* from 0 month (that is, up to more than 100 Months), the *Marital Status* is *denser* in the *higher* months, so that we assume that it takes time to get survived.



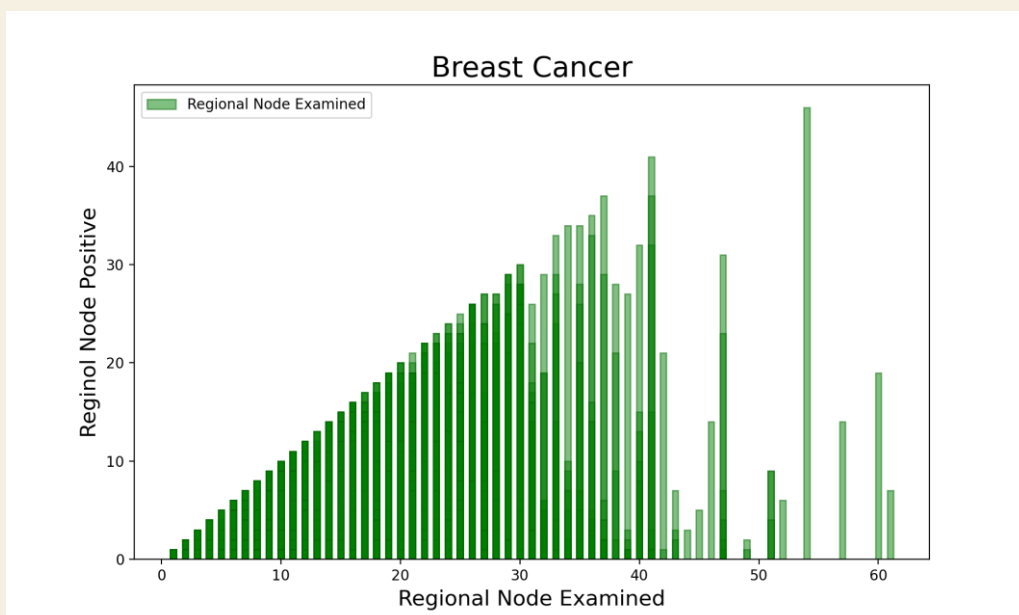
Scatter Plot #5

Regional Node Examined vs. Regional Node Positive



Bar Plot #5

Regional Node Examined vs. Regional Node Positive



Plot #5 description



What is the title of the plot showing?

These two graphs (Scatter & Bar) both talk about (increasing) changes in *Regional Node Examined (RNE)* compared to (increasing) changes in *Reginol Node Positive (RNP)*.

What does the x-axis indicate?

The x-axis, *RNE*, is changing from about 0 to more than about 60.

What does the y-axis indicate?

The y-axis, *RNP*, is changing from about 0 to more than about 40.

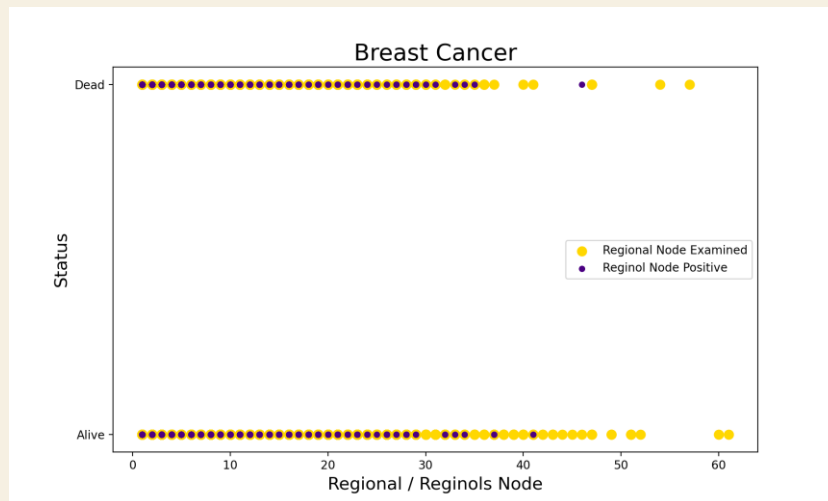
What is the trend of the values?

- This graph is the result of comparing the two features of *RNE* and *RNP* in relation to each other.
- Since the investigation of these two features is interesting to us, this plot was drawn.
- This chart shows both *RNE* & *RNP* with 10-10 intervals.
- As can be seen, for checking the density of the graph, the closer the values are to smaller values, the denser the graph is. For this reason, in fewer *RNEs*, there is more density and more number of *RNPs* than if both of these axes have more values.
- On the other hand, for checking the peak of the values in the plot, it can be said that the highest value for *RNE* has happened around more than 60, and also the highest value has happened for *RNP* around more than 40 or even close to 50.
- The trend of the values is, that we see an upward trend with the increase in the value of the axes. It can also be seen that with this upward increase of *RNE* and *RNP* values relative to each other, the density of the data gradually decreases, and the denser state of the graph is seen more in the initial values.



Scatter Overlap Plot

Regional / Regional Node vs. Status



What is the title of the plot showing?

This Overlap (in Scatter) plot talks about (increasing) changes in *Regional Node Examined (RNE)* & *Regional Node Positive (RNP)* (at the same time) compared to changes in current Status of the patients.

What does the x-axis indicate?

The x-axis, *RNE* & *RNP*, is changing from about 0 to more than about 60.

What does the y-axis indicate?

The y-axis, *Status*, indicates two values (being Alive or Dead).

What is the trend of the values?

- This overlap plot is the result of comparing the two features of *RNE* and *RNP* in relation to the target value, which is *Status*.
- Since the investigation of these two features with the target is interesting to us, this plot was drawn.
- This chart shows both *RNE* & *RNP* with 10-10 intervals.
- In terms of checking the data density, as can be seen from the plot, in *Alive*, *RNP* values are lower than *RNE* values, and although the density of both is relatively the same, *RNP* continues to be more dense up to 35 is, because *RNE* has gone up to about 60. In *Dead*, *RNP* values are still lower than *RNE* values, and although the density of both is relatively the same, *RNP* continues to be more dense up to 30, compared to *RNE*, almost reaching 55.
- On the other hand, for checking the peak values of this plot, it can be said that the highest value for *RNE* has happened around more than 60, and also the highest value has happened for *RNP* around more than about 47.
- The trend of the values is, two *Alive* and *Dead* values can be seen, for each of which there are specific *RNE* and *RNP* values. The *RNE* values shown in yellow appear quantitatively higher than the *RNP* values shown in purple. This graph shows that both *RNE* and *RNP* have been tested for both currently *Alive* and *Dead* patients. This chart does not have any ups and downs to discuss.



O4

Conclusion

What was done.

What was achieved.



Summary & Conclusion

In this report, the Breast Cancer dataset was examined, necessary pre-processing was done and plots were drawn.

Some plots were based on the comparison of Feature and Target, some were based on the comparison of Feature with Feature (which was drawn based on curiosity and analysis of the relationship between them), and some were the result of superimposing two Feature-Target graphs.

Each of these plots showed us relationships between samples, data points, and data within the dataset. Relationships that were not possible to discover in normal mode and only by looking at the table called dataset. These relationships give the reader of the report, even if he does not have expertise and knowledge about this dataset, valuable and categorized information.

The purpose of plotting these data is to discover the relationships between the data.



The end

Do you have any questions?

Samira.Shemirani92@gmail.com



www.linkedin.com/in/samira-shemirani-664302132

