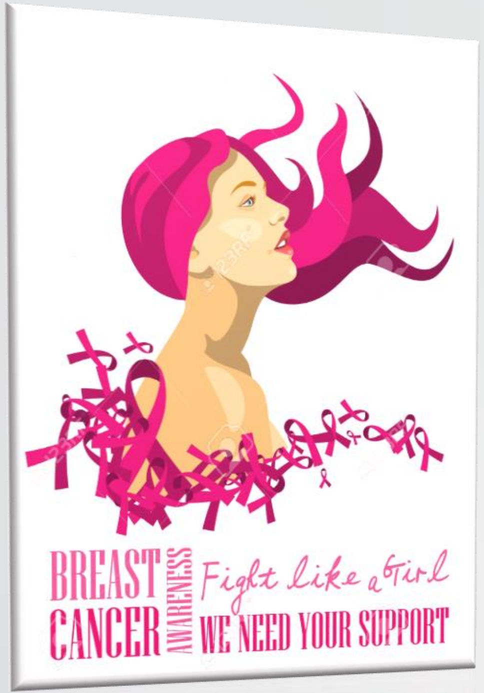


# Project : Breast Cancer Survival

July-2019

Jigna Thacker

GCD- April 2019 Batch

# Project Brief

- **Data set information :**

  The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

- **Attribute information :**

  - **Age of patient** at time of operation (numerical)
  - Patient's **year of operation** (year - 1900, numerical)
  - Number of **positive axillary nodes detected** (numerical)
  - **Survival** status (class attribute)

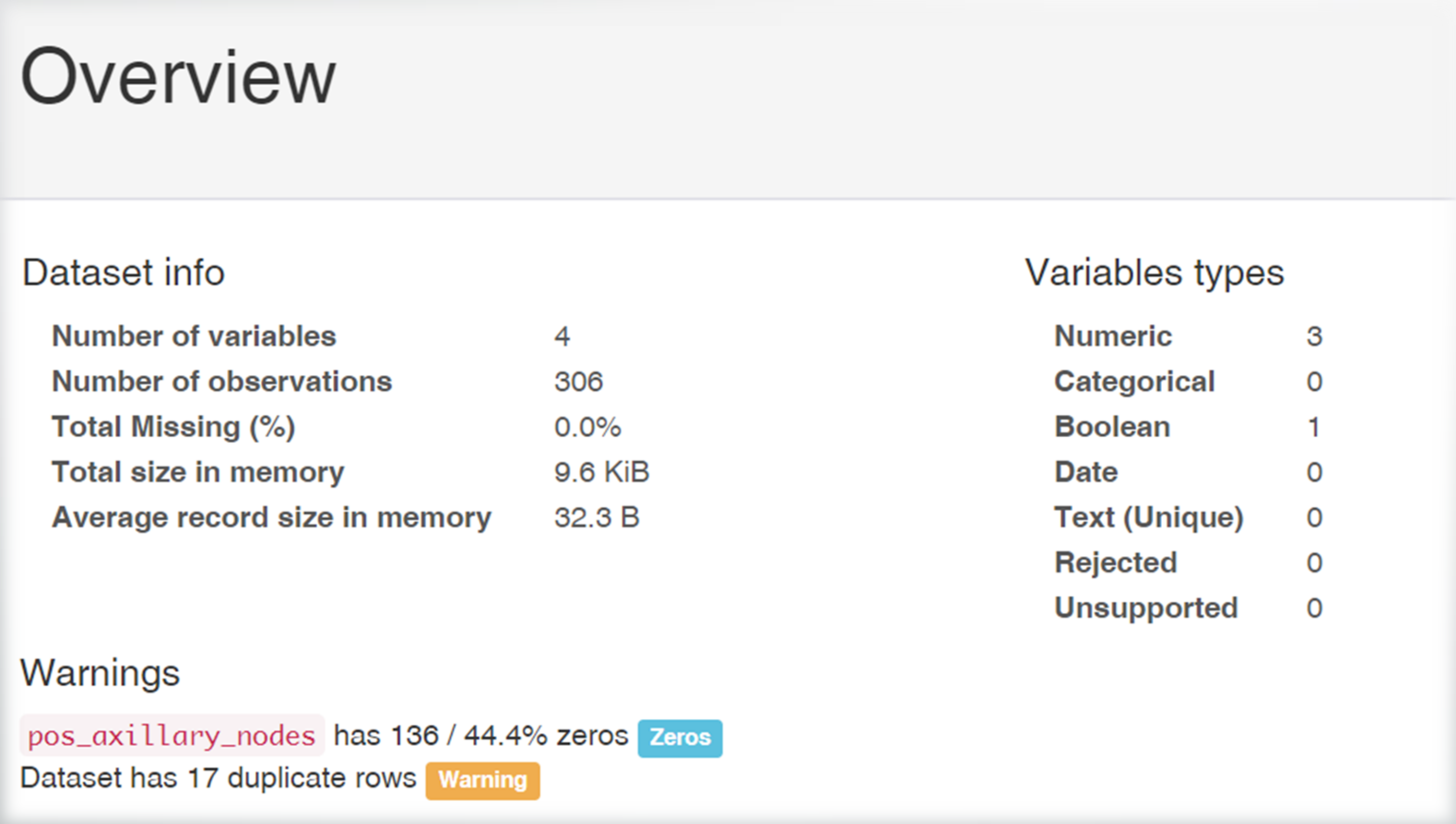
- **Objective :**

  - To predict patient survival after 5 years of operation basis various **attributes** : Age of patient, year of operation and positive axillary node

# Preliminary observations

- Total sample : 306
- Total variables for analysis : 4
- No missing value found in the given 4 variables
- Dataset has 17 duplicate rows
  - Considered as "valid dataset‟  as age is one of the variable and is in completed year

# Profiling using Pandas



## Overview

### Dataset info

| | |
|---|---|
| **Number of variables** | 4 |
| **Number of observations** | 306 |
| **Total Missing (%)** | 0.0% |
| **Total size in memory** | 9.6 KiB |
| **Average record size in memory** | 32.3 B |

### Variables types

| | |
|---|---|
| **Numeric** | 3 |
| **Categorical** | 0 |
| **Boolean** | 1 |
| **Date** | 0 |
| **Text (Unique)** | 0 |
| **Rejected** | 0 |
| **Unsupported** | 0 |

### Warnings

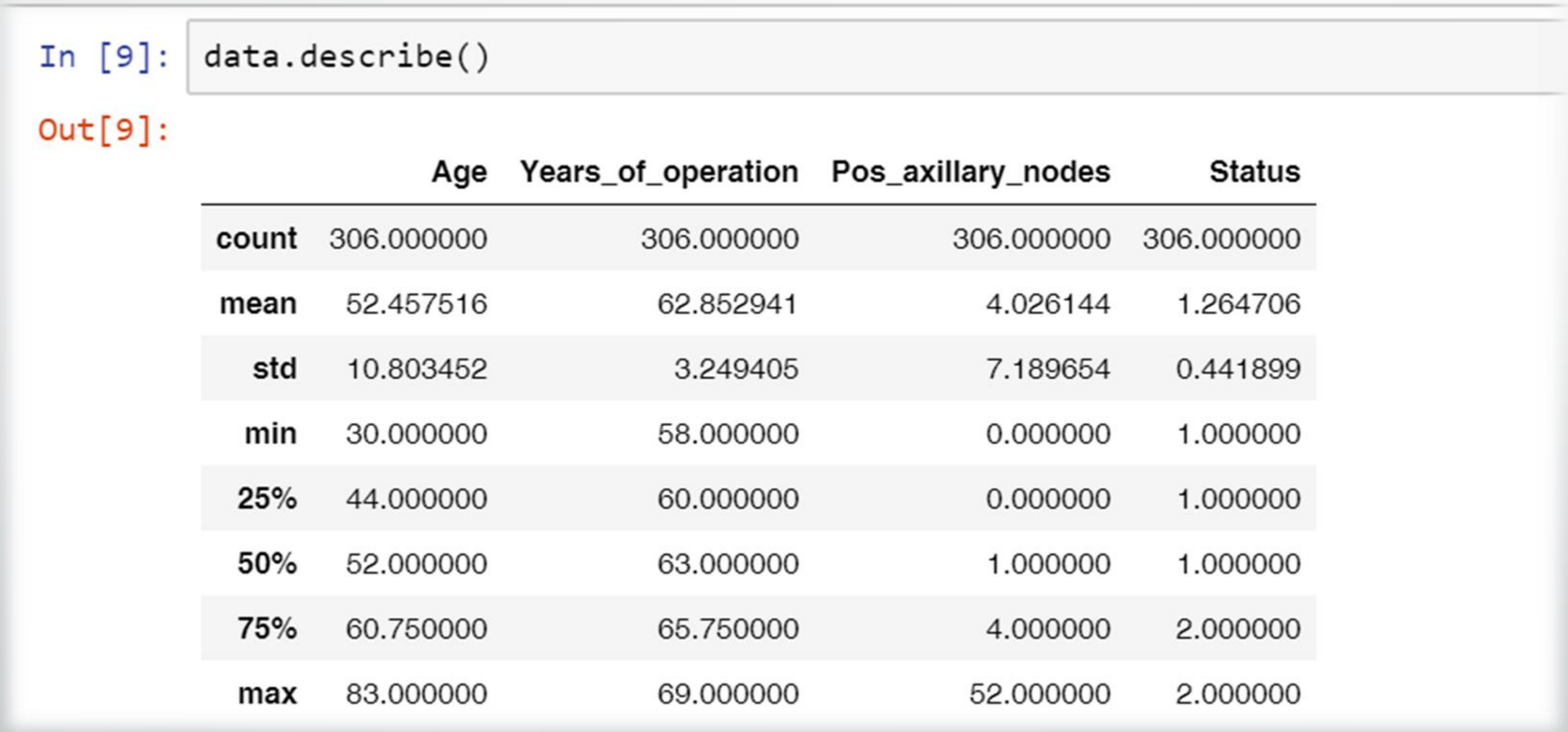
`pos_axillary_nodes` has 136 / 44.4% zeros Zeros

Dataset has 17 duplicate rows Warning

- No missing value

- Data has 17 duplicate values which I have considered as Valid response

# Basic statistics

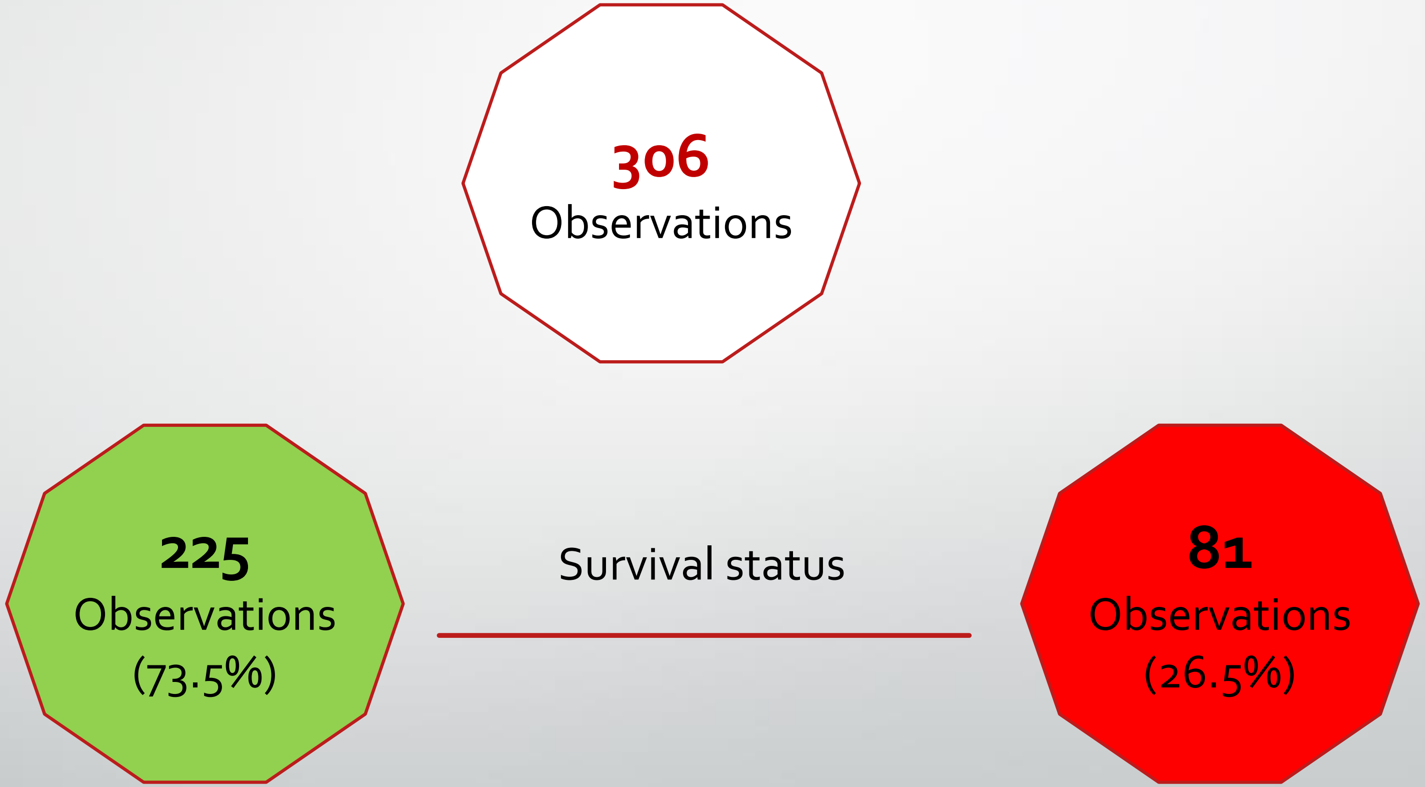


```
In [9]: data.describe()
```

Out[9]:

| | Age | Years_of_operation | Pos_axillary_nodes | Status |
|---|---|---|---|---|
| **count** | 306.000000 | 306.000000 | 306.000000 | 306.000000 |
| **mean** | 52.457516 | 62.852941 | 4.026144 | 1.264706 |
| **std** | 10.803452 | 3.249405 | 7.189654 | 0.441899 |
| **min** | 30.000000 | 58.000000 | 0.000000 | 1.000000 |
| **25%** | 44.000000 | 60.000000 | 0.000000 | 1.000000 |
| **50%** | 52.000000 | 63.000000 | 1.000000 | 1.000000 |
| **75%** | 60.750000 | 65.750000 | 4.000000 | 2.000000 |
| **max** | 83.000000 | 69.000000 | 52.000000 | 2.000000 |

# Basic statistics – Survival Status



**306**
Observations

**225**
Observations
(73.5%)

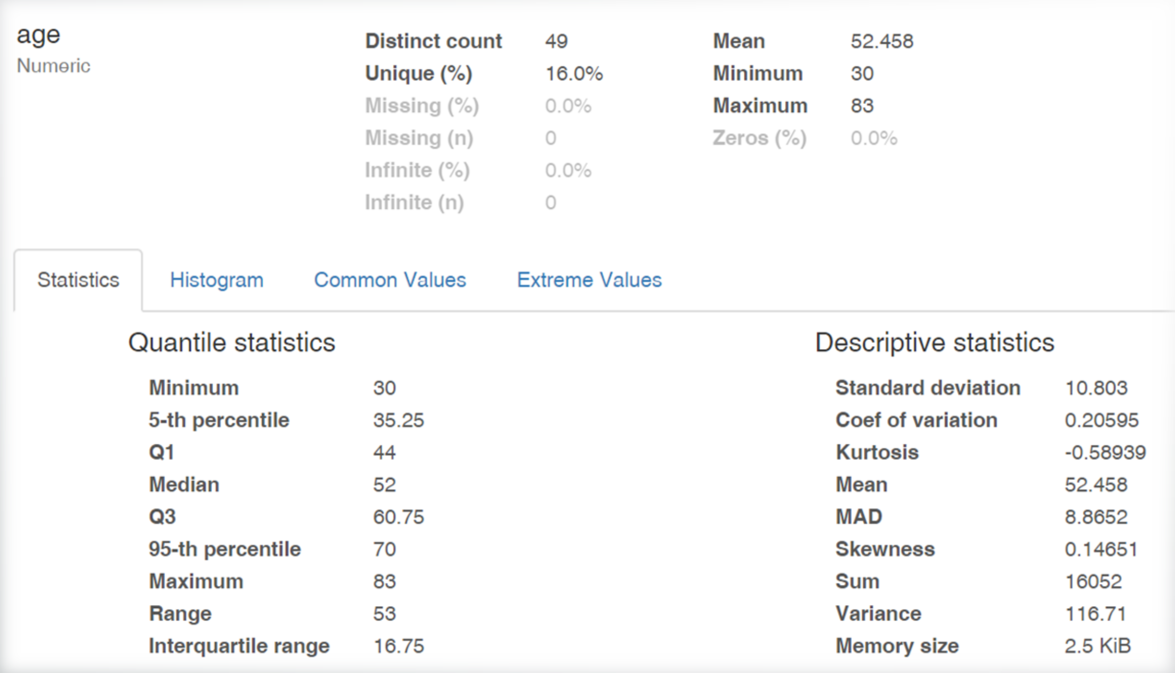
Survival status

**81**
Observations
(26.5%)

**patient survived**
**5 years or longer**

**patient died**
**Within 5 years**

# Survival by Age



age
Numeric

| | | | |
|---|---|---|---|
| Distinct count | 49 | Mean | 52.458 |
| Unique (%) | 16.0% | Minimum | 30 |
| Missing (%) | 0.0% | Maximum | 83 |
| Missing (n) | 0 | Zeros (%) | 0.0% |
| Infinite (%) | 0.0% | | |
| Infinite (n) | 0 | | |

Quantile statistics

| | |
|---|---|
| Minimum | 30 |
| 5-th percentile | 35.25 |
| Q1 | 44 |
| Median | 52 |
| Q3 | 60.75 |
| 95-th percentile | 70 |
| Maximum | 83 |
| Range | 53 |
| Interquartile range | 16.75 |

Descriptive statistics

| | |
|---|---|
| Standard deviation | 10.803 |
| Coef of variation | 0.20595 |
| Kurtosis | -0.58939 |
| Mean | 52.458 |
| MAD | 8.8652 |
| Skewness | 0.14651 |
| Sum | 16052 |
| Variance | 116.71 |
| Memory size | 2.5 KiB |

Minimum age : 30
Maximum age : 85

| Age Class | Count |
|---|---|
| upto 40 Years | 43 |
| 41-50 Years | 93 |
| 51-60 Years | 93 |
| Above 60 Years | 77 |

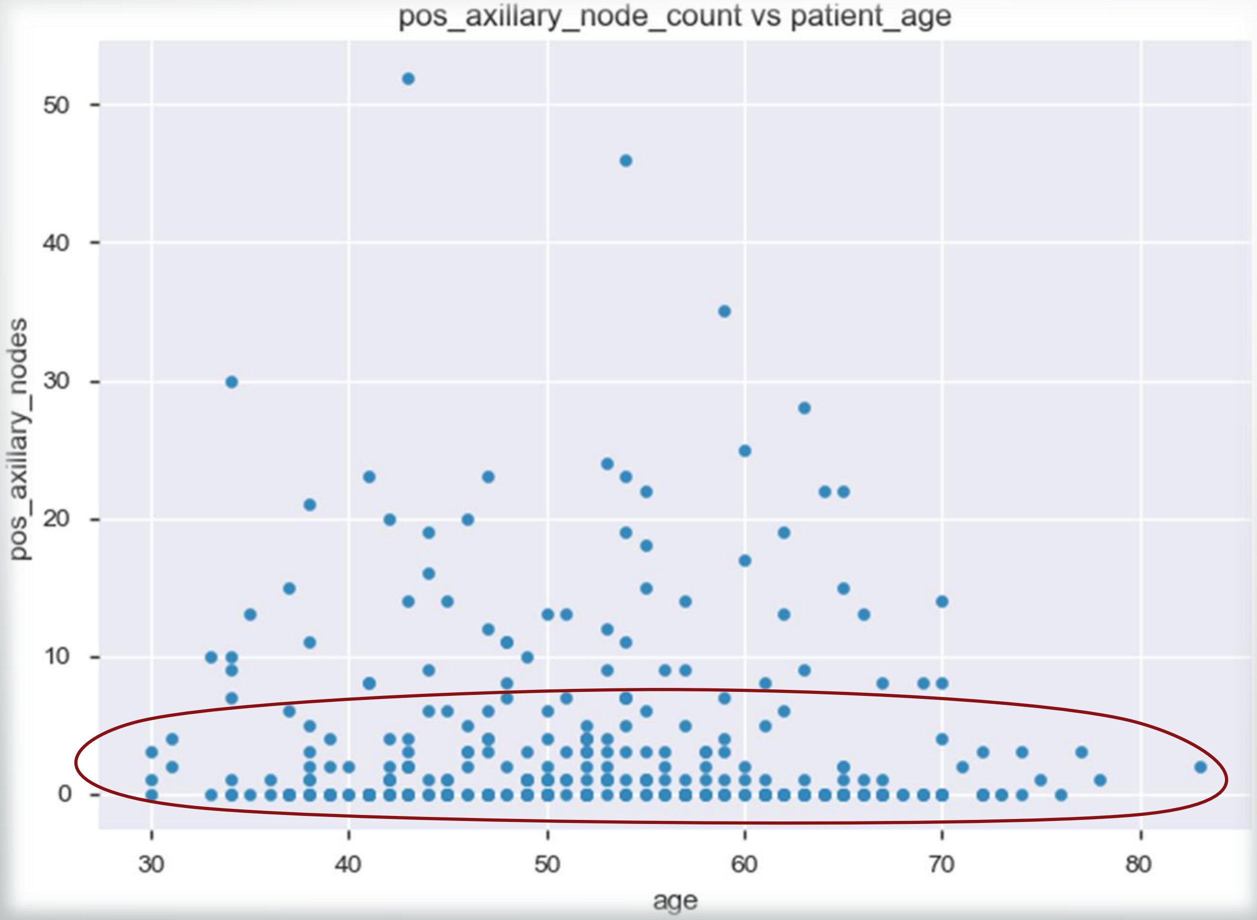
Filtering Survival = 1 i.e. survived more than 5 years = 225

Minimum age : 30
Maximum age : 77

| Age Class | Count |
|---|---|
| upto 40 Years | 39 |
| 41-50 Years | 64 |
| 51-60 Years | 67 |
| Above 60 Years | 55 |

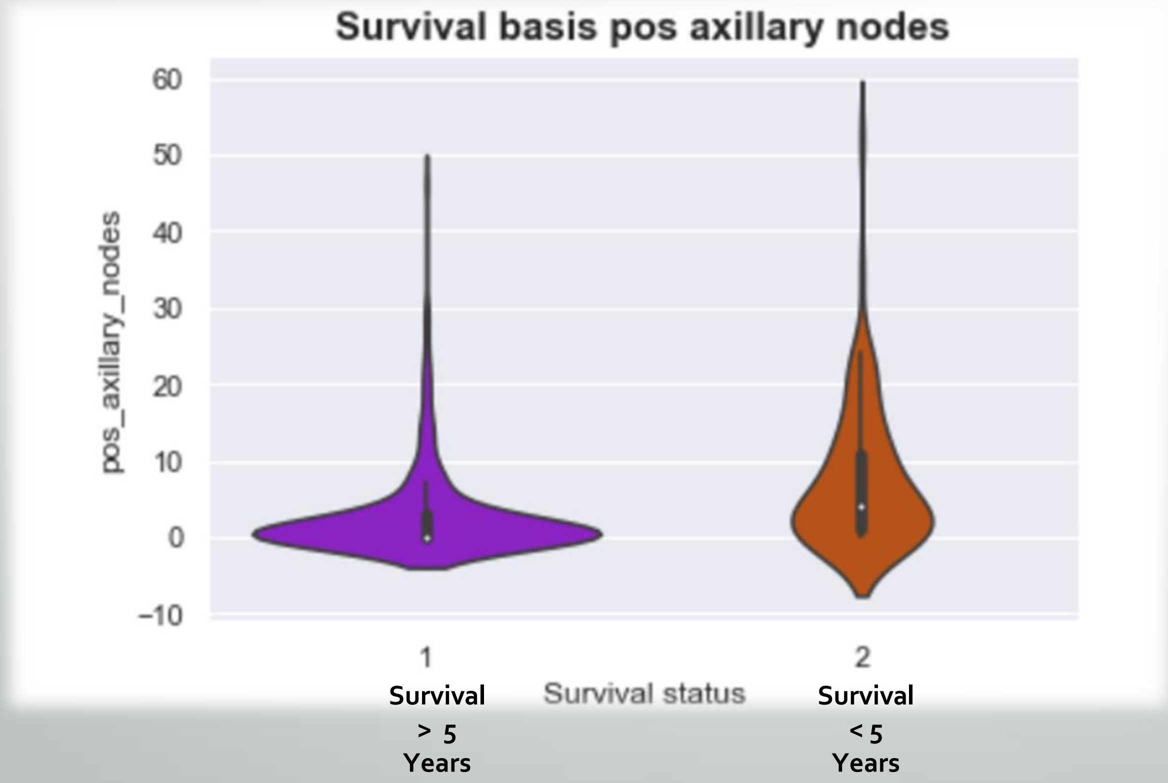
Survival for more than 5 years is higher for upto age 40
Total : 43 cases
Survived more than 5 years : 39 cases

# Positive axillary node detection by age of Patient



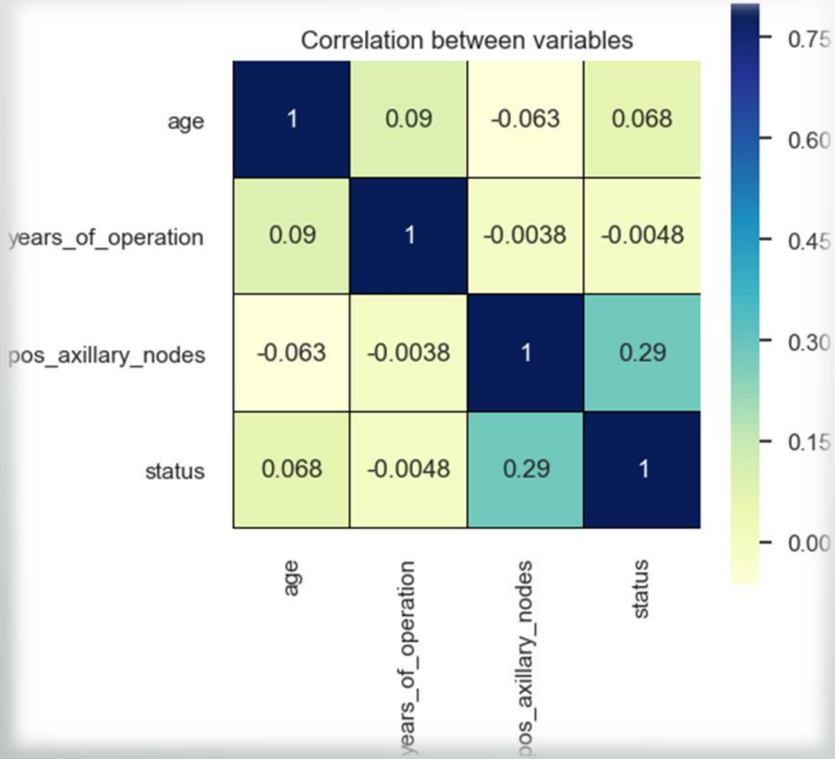
Most of the positive axillary nodes detected at "Zero"

# Survival by Positive axillary node



- Survival for more than 5 years is higher for positive node at “Zero”
- Survival for less than 5 years is where positive axillary nodes are higher than 1
- There are few cases where chances of survival more than 5 years is there even if positive axillary node > 30

# Correlation between variables



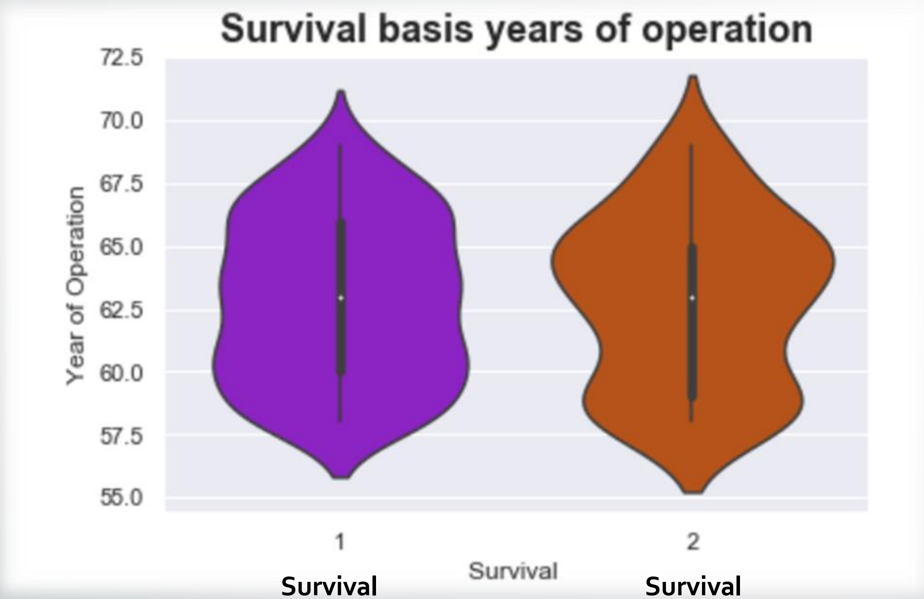
| | age | years_of_operation | pos_axillary_nodes | status |
|---|---|---|---|---|
| age | 1 | 0.09 | -0.063 | 0.068 |
| years_of_operation | 0.09 | 1 | -0.0038 | -0.0048 |
| pos_axillary_nodes | -0.063 | -0.0038 | 1 | 0.29 |
| status | 0.068 | -0.0048 | 0.29 | 1 |

| | | Survival status | | |
|---|---|---|---|---|
| | | Total | 5+ years | < 5 Years |
| Positive Axillary nodes | Total | 306 | 225 | 81 |
| | 0 | 136 | **117** | 19 |
| | 1-5 | 100 | 73 | 27 |
| | >5 | 70 | 35 | 35 |

Positive correlation between axillary nodes and survival.

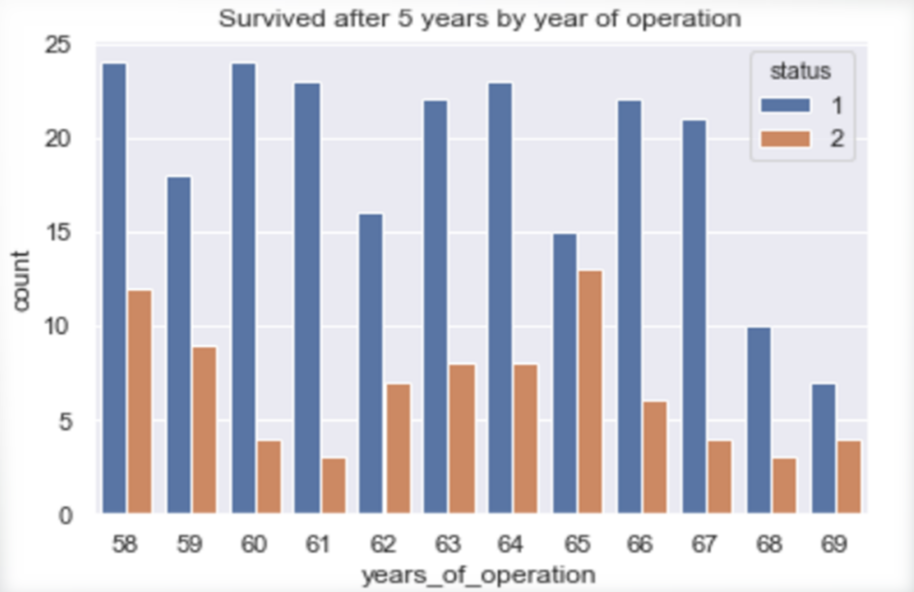
Survival is higher with node=0

# Survival by Year of operation



Survival > 5 Years

Survival < 5 Years



Year of operation is not giving any conclusion on survival more than 5 years.

# Summary

- Looking at basic profiling – there are 17 duplicate cases.

  Considered as "valid dataset“ as age is one of the variable and is in completed year
- In all 73.5 % patients survived more than 5 years
- In all 91% patients who are upto age 40 – survived more than 5 years
- Positive correlation between survival more than 5 years and positive axillary node: With node “zero” survived more than 5 yeas are higher
- Out of all 4 variables - year of operation is not giving any conclusion for survival more than 5 years.

# Thank you