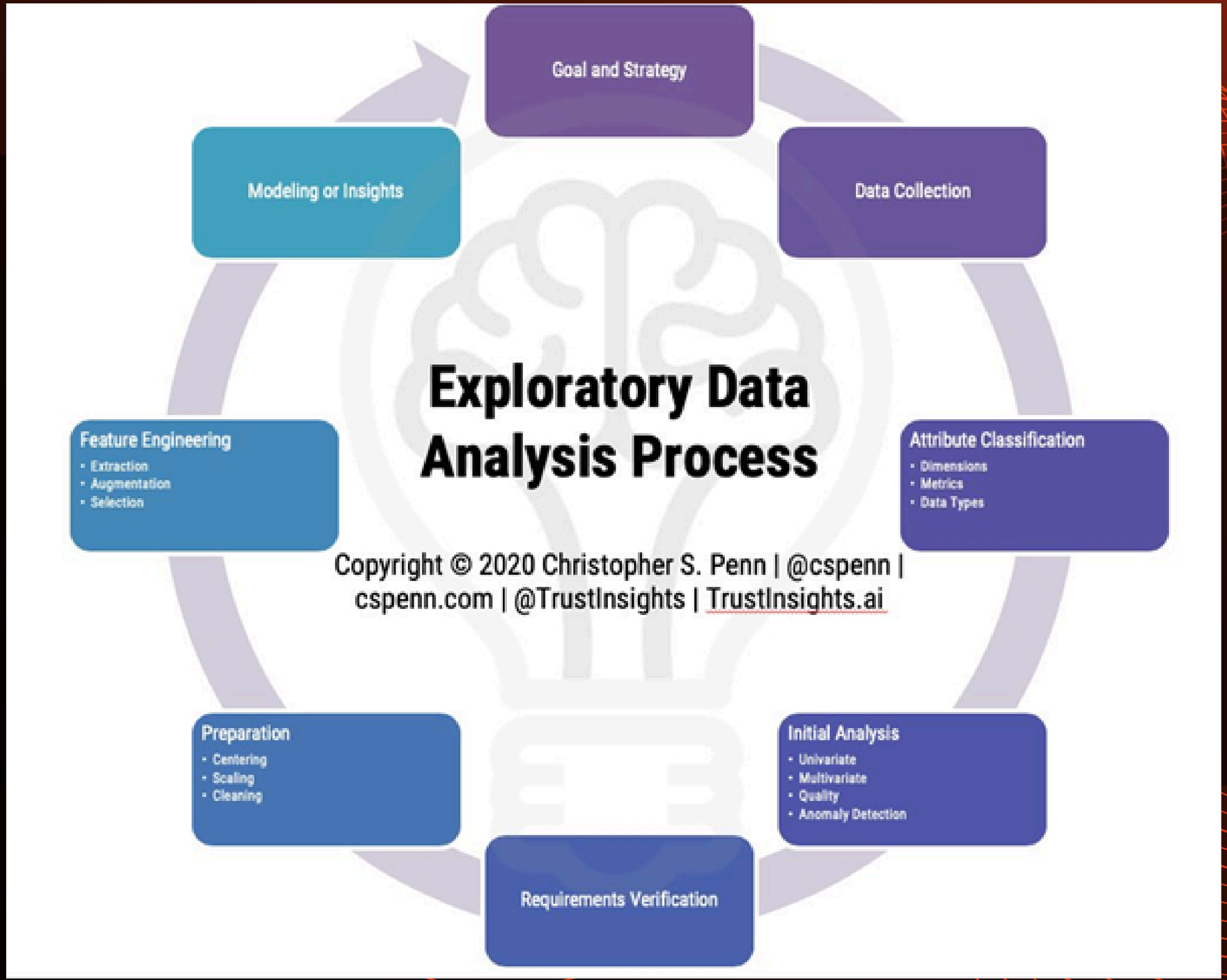
EXPLORATORY DATA ANALYSIS OF HABERMAN CANCER SURVIVAL DATASET

Harshvardhan Mishra – 18BCE2247

Palak Kishore - 18BCE2312

ABSTRACT

Perform Exploratory Data Analysis (EDA) on Cancer Survival Dataset. Exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. This Dataset is the dataset containing the information data about the survival of the breast cancer patients in the Billings Hospitals of University of Chicago. We are trying to analyse the dataset and make observations and predictions about the survival rate.


Goal and Strategy
Data Collection
Attribute Classification
• Dimensions
• Metrics
• Data Types
Initial Analysis
• Univariate
• Multivariate
• Quality
• Anomaly Detection
Requirements Verification
Preparation
• Centering
• Scaling
• Cleaning
Feature Engineering
• Extraction
• Augmentation
• Selection
Modeling or Insights
Exploratory Data Analysis Process
Copyright © 2020 Christopher S. Penn | @cspenn | cspenn.com | @TrustInsights | TrustInsights.ai

INTRODUCTION

- Objectives: to predict Patient survived more than 5 years or die within 5 years based upon their Age, Axil Node,operation year and survival status.
- Using this, we will be able to predict whether the patient will survive more than 5 years or die within 5 years due to breast cancer based upon the Haberman Dataset.
- Our data consists of:
  - number of auxiliary nodes detected (0-52)
  - age (30-83)
  - year of operation
  - survival status (if patients survived 5 years or more after undergoing surgery then they are represented as 1 and patients who survived less than 5 years are represented as 2)

SOFTWARE REQUIREMENTS

- Python (package - Pandas, Numpy, Matplotlib and seaborn)

OPERATIONS

```
In [7]: haberman.info()
```

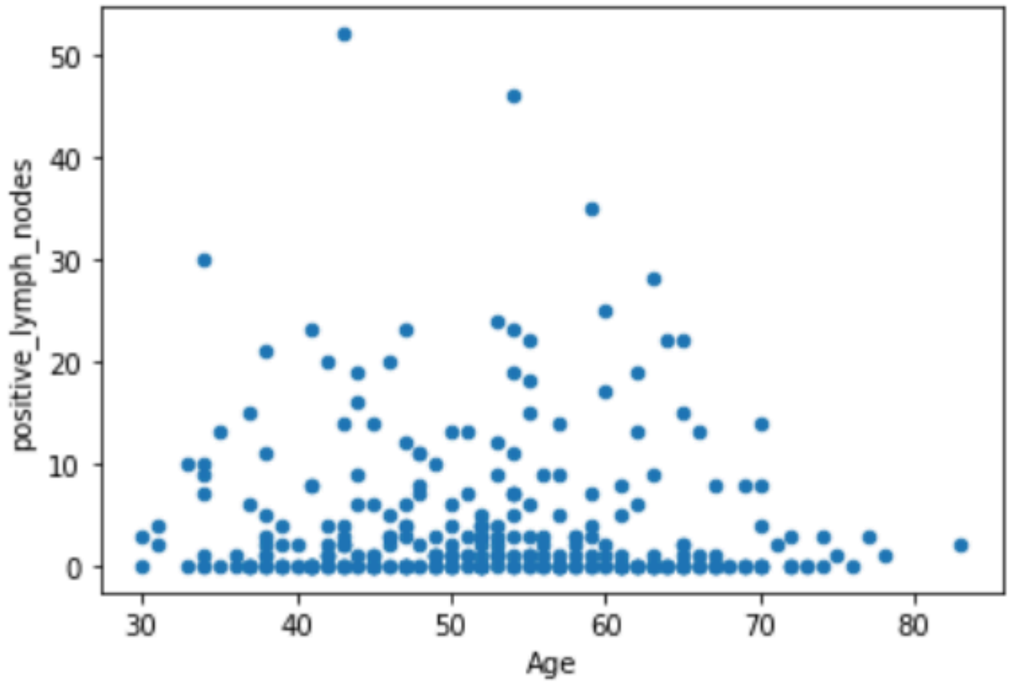
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   305 non-null    int64
 1   Operation_Year        305 non-null    int64
 2   positive_lymph_nodes  305 non-null    int64
 3   Survival status       305 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

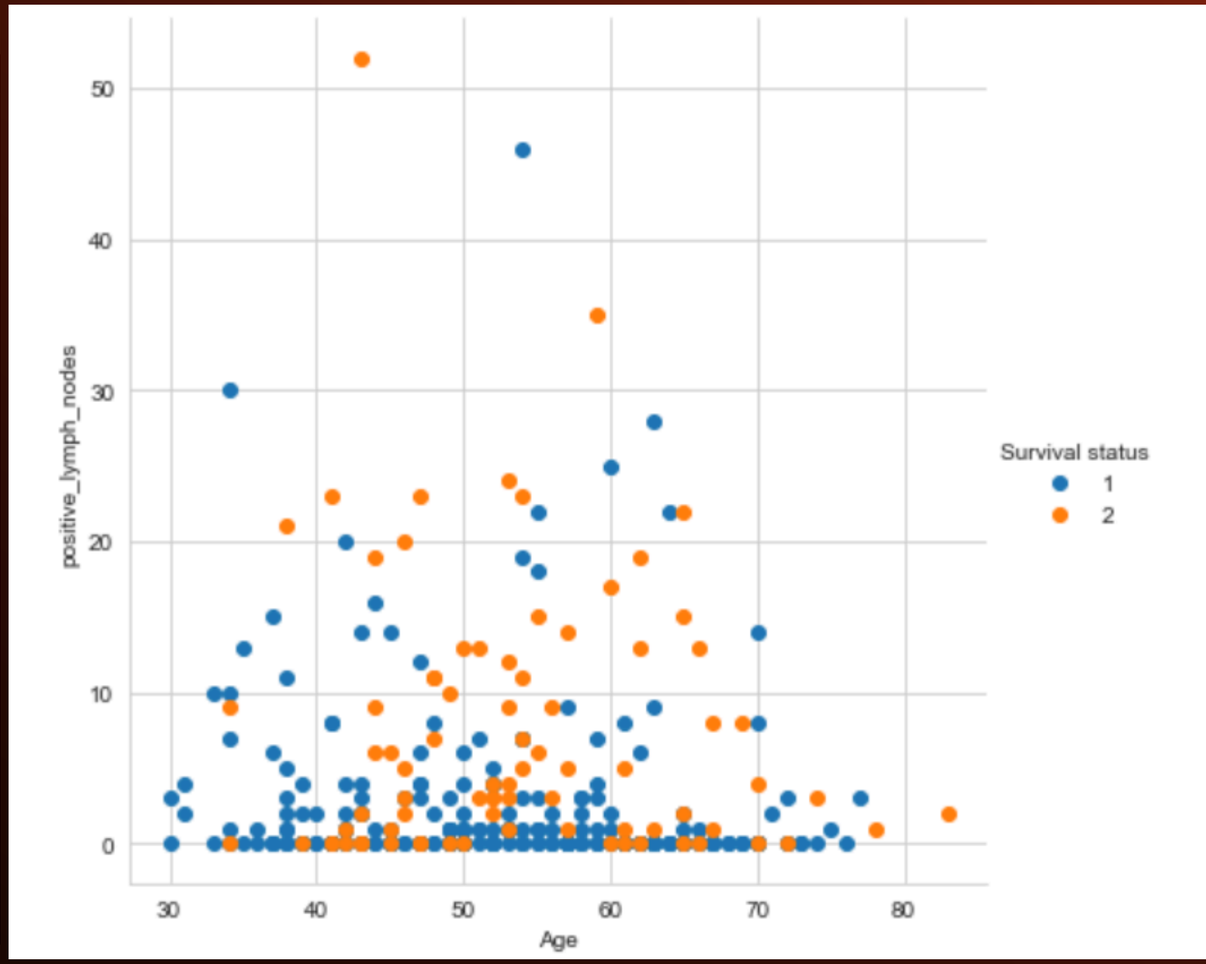
```
In [8]: haberman["Survival status"].value_counts()
```

```
Out[8]: 1    224
        2     81
        Name: Survival status, dtype: int64
```

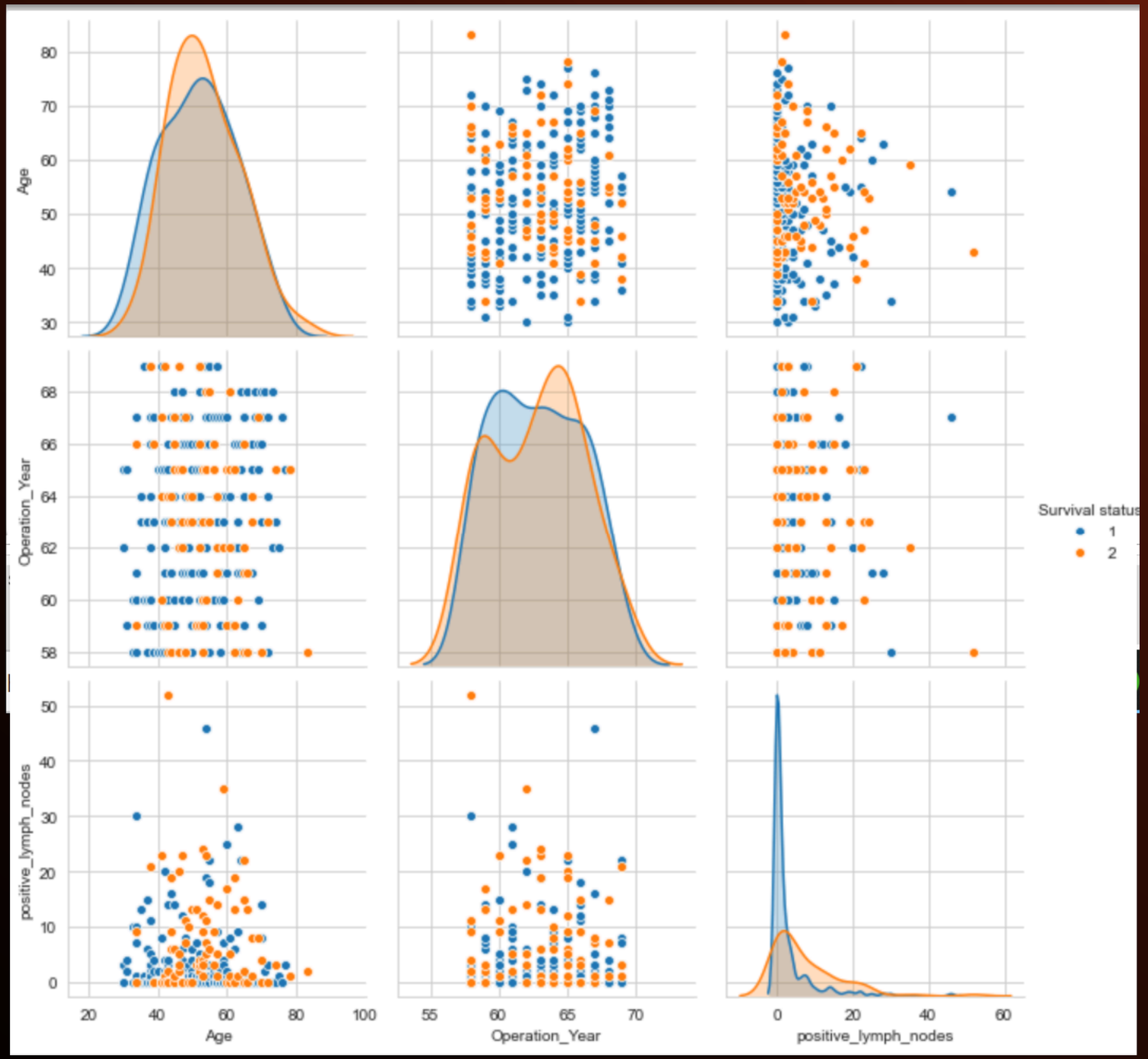
2D-SCATTER PLOT

```
haberman.plot(kind='scatter',x='Age',y='positive_lymph_nodes');
plt.show()
```

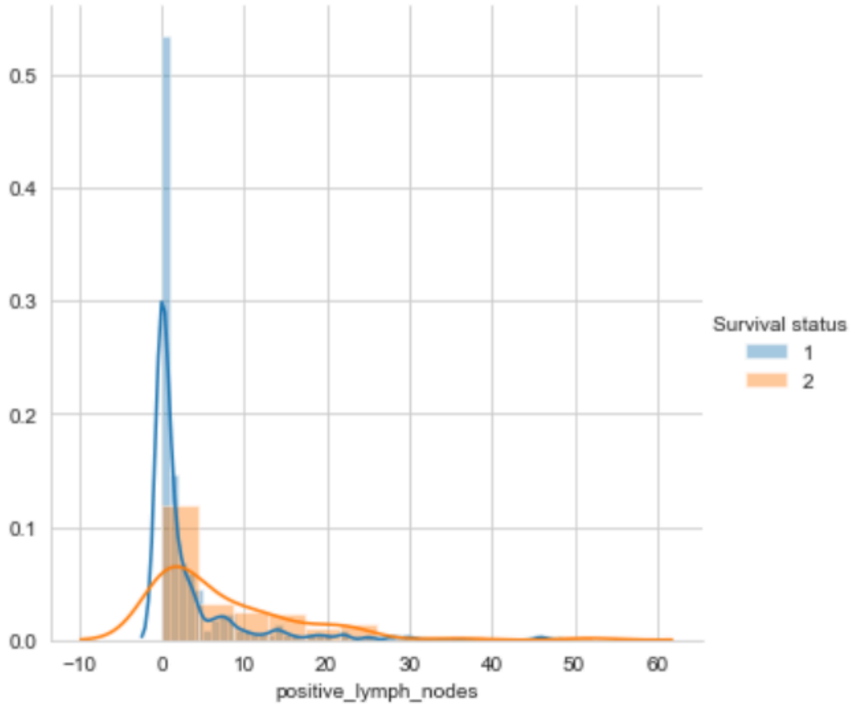




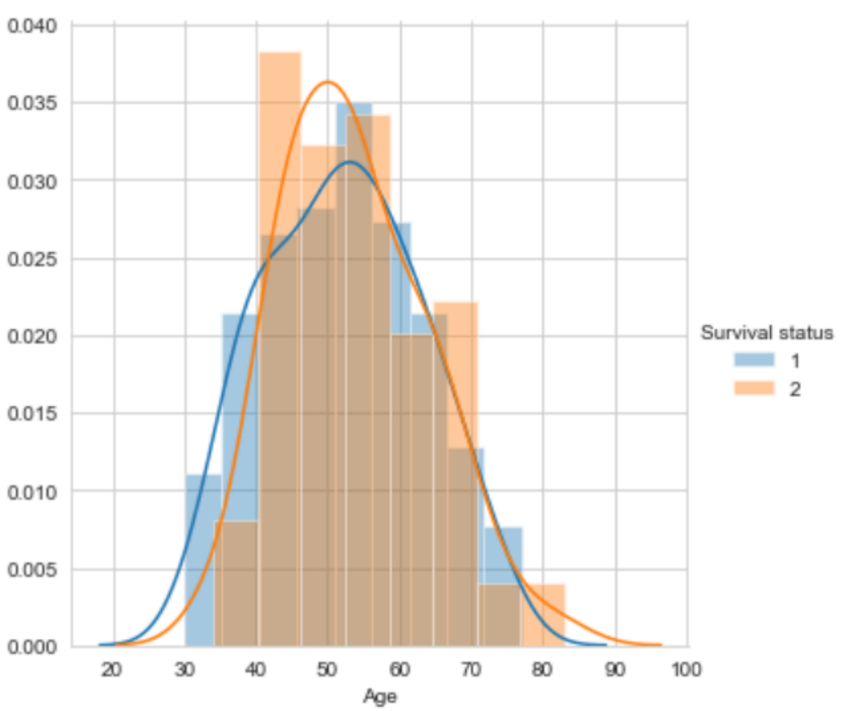
PAIR PLOT


Age
Operation_Year
positive_lymph_nodes
Survival status
1
2

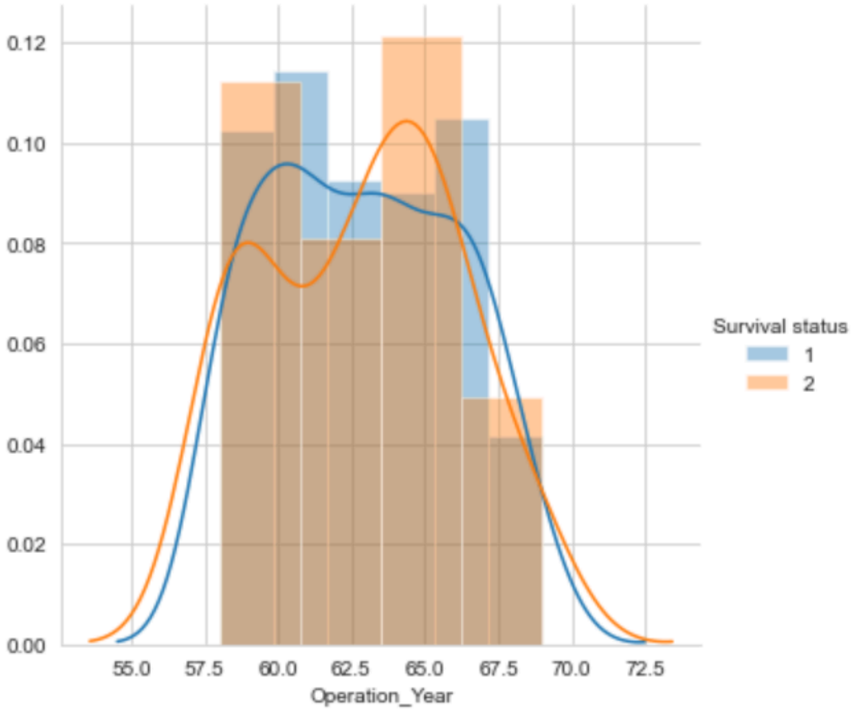
PDF (PROBABILITY DENSITY FUNCTION)



(1)

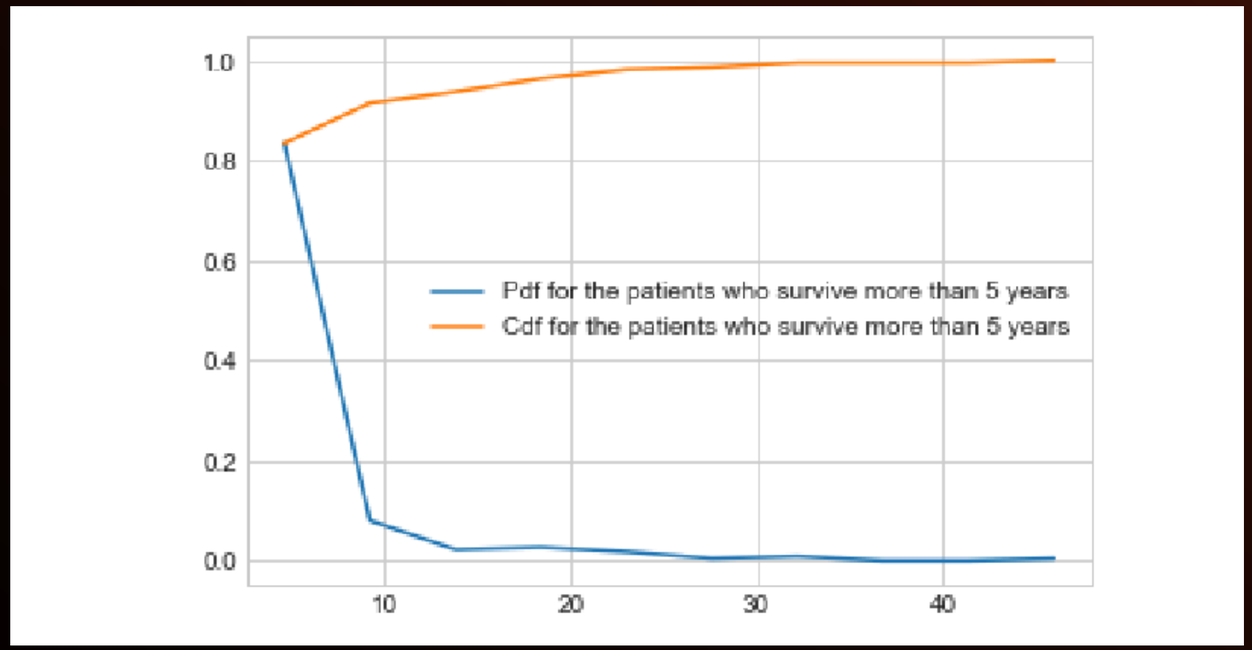


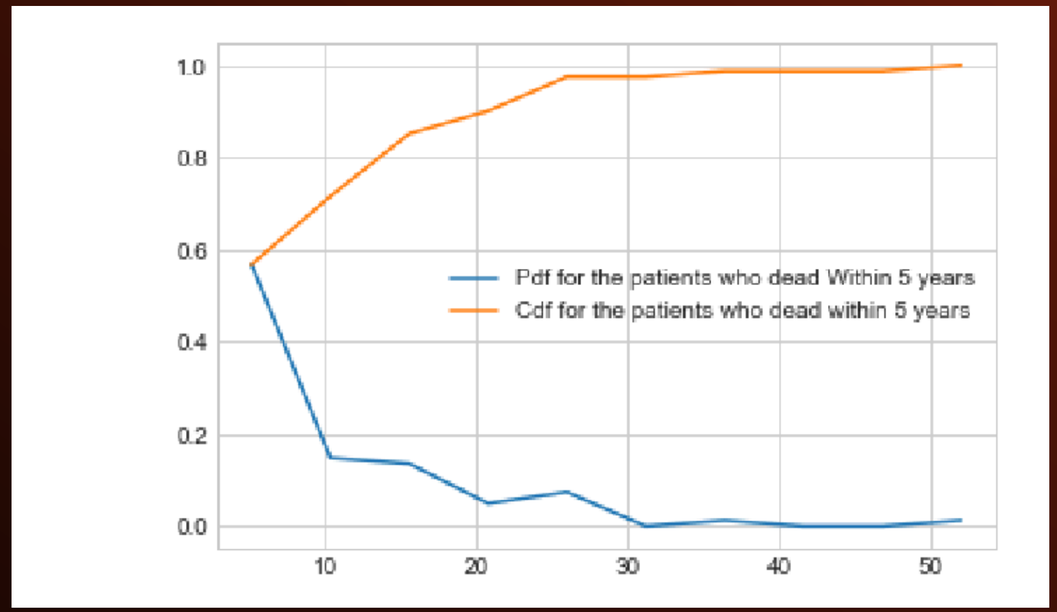
(2)



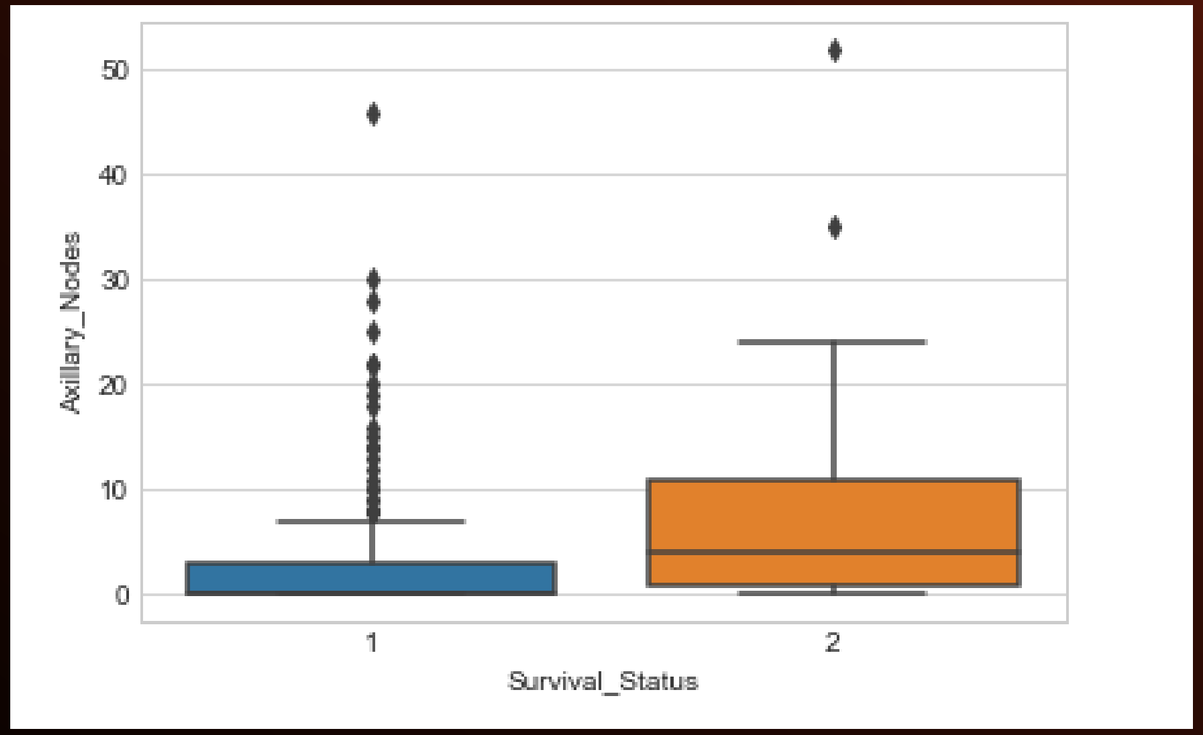
(3)

CDF (CUMULATIVE DISTRIBUTION FUNCTION)

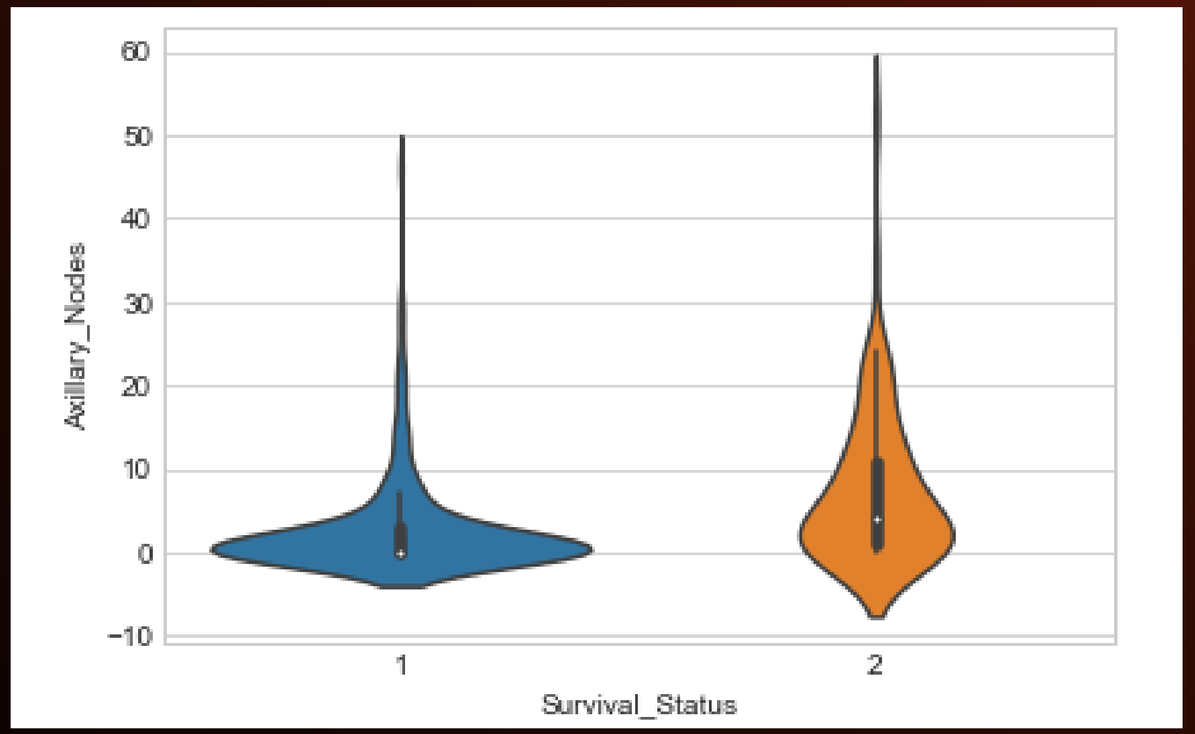




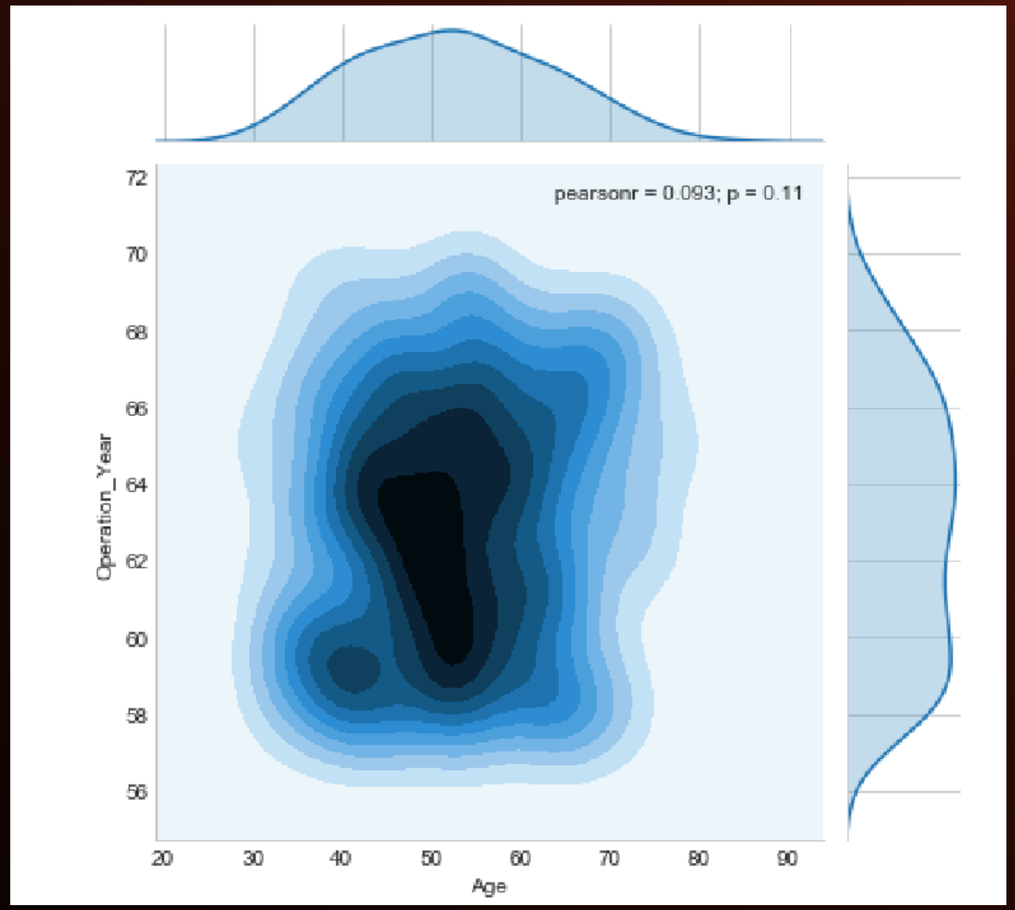
BOX PLOT



VIOLIN PLOT



CONTOUR PLOT



CONCLUSION

1. There are 306 observations with 4 features in the data set.
2. It is an imbalanced dataset with

a. 225 patients belonging to status 1, those who survived for 5 years and longer and

b. 81 patients belonging to status 2, those who survived for less than 5 years.

3. Using scatter plot(Bi-variate analysis)

a. Most of the people have zero positive_lymph_nodes.

b. We cannot distinguish between the people who survived and who didn't survive.

4. Using Pair-plot concept (Bi-variate analysis)

a. positive_lymph_nodes VERSUS Age is the useful plot to atleast get the insight that most people who survived have 0 postive lymph nodes detected.

b. Age and Operation Year have overlapping curves which makes difficult for classifying the survival status.

c. We cannot distinguish the data easily with the help of these plots as most of them are overlapping.

5. Using PDFs(Uni-variate Analysis)

a. both Age and Operation Year are not good features for useful insights as the distibution is more similar for both people who survived and also dead.

b. positive_lymph_nodes is the only feature that is useful to know about the survival status of patients as there is difference between the distributions for both classes(labels). From that distibution we can infer that most survival patients have fallen in to zero positive_lymph_nodes.

c. More number of people are not survived in year of operation of 1965.

6. Using CDFs(Uni-variate analysis)

a. We can observe that almost for all the features the statistics are similar except for positive_lymph_nodes.

b. We can infer that patients above 46 axillary nodes detected can be considered as dead within 5 years. So,People having less number of positive_lymph_node have survived over 5 years.

7. The mean(average) of positive lymph nodes is more for people who died within 5 years than people who have survived for more than 5 years.

8. Mean age of patients who survived is 52 years and who didn't survive is 54 years.

9. Using Box plot and Violin plots-

a. The number of positive lymph nodes of the survivors is highly dense from 0 to 5.

b. Almost 80% of the patients have less than or equal to 5 positive lymph survived more than 5 years.

c. From box plots and violin plots, we can say that more no of patients who are dead have age between 46-62,year between 59-65 and the patients who survived have age between 42-60, year between 60-66.

10. Using Contour plots. There are more number of people who have undergone operation during the year 1959 1964 period and between the ages 42 - 60.