9.16_hebermans_cancer_survival

June 8, 2018

1 Cancer survival dataset Exploratory analysis

dataset from https://www.kaggle.com/gilsousa/habermans-survival-data-set/data

2 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)
1 = the patient survived 5 years or longer
2 = the patient died within 5 year

3 Import data and libraries

```
In [1]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt

        dataset=pd.read_csv("haberman.csv",header=None)
        names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        dataset.columns=names
        #dataset.head(5)
        print('Total number of row and columns'+ str(dataset.shape))
        print('Basic statistics of features \n\n',dataset.describe())

        a=dataset['Survival status'].value_counts()
        print('No of patient dieded within 5 years : '+str(a[2]))
        print('No patient survived 5 years or longer : '+str(a[1]))
```

```
Total number of row and columns(306, 4)
Basic statistics of features

               Age  Year operation  Axillary nodes detected  Survival status
count  306.000000      306.000000               306.000000       306.000000
mean    52.457516       62.852941                 4.026144         1.264706
```

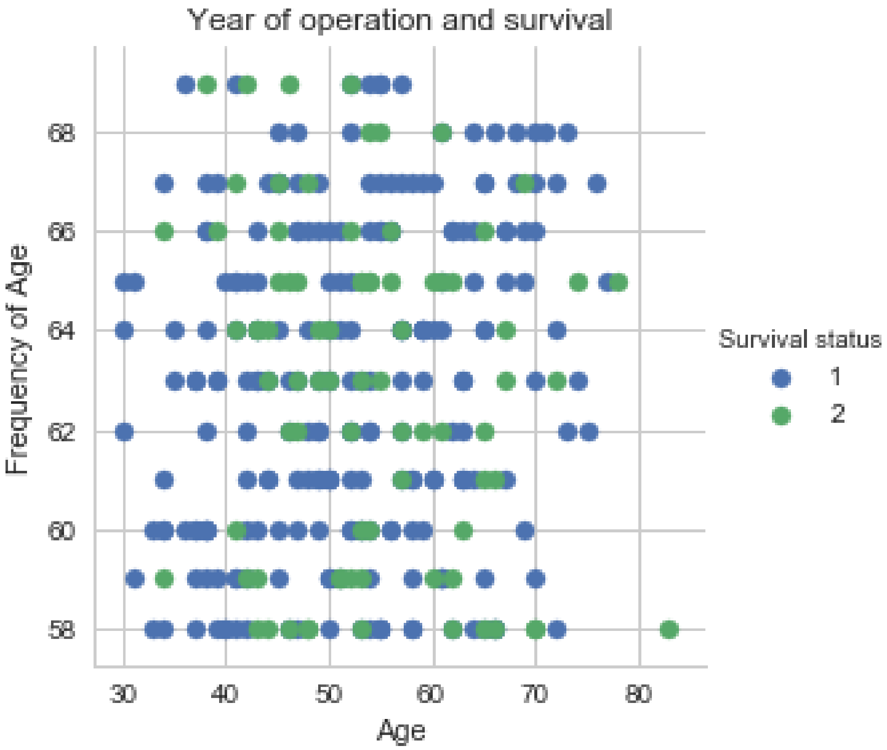
```
std     10.803452        3.249405                 7.189654         0.441899
min     30.000000       58.000000                 0.000000         1.000000
25%     44.000000       60.000000                 0.000000         1.000000
50%     52.000000       63.000000                 1.000000         1.000000
75%     60.750000       65.750000                 4.000000         2.000000
max     83.000000       69.000000                52.000000         2.000000
No of patient dieded within 5 years : 81
No patient survived 5 years or longer : 225
```

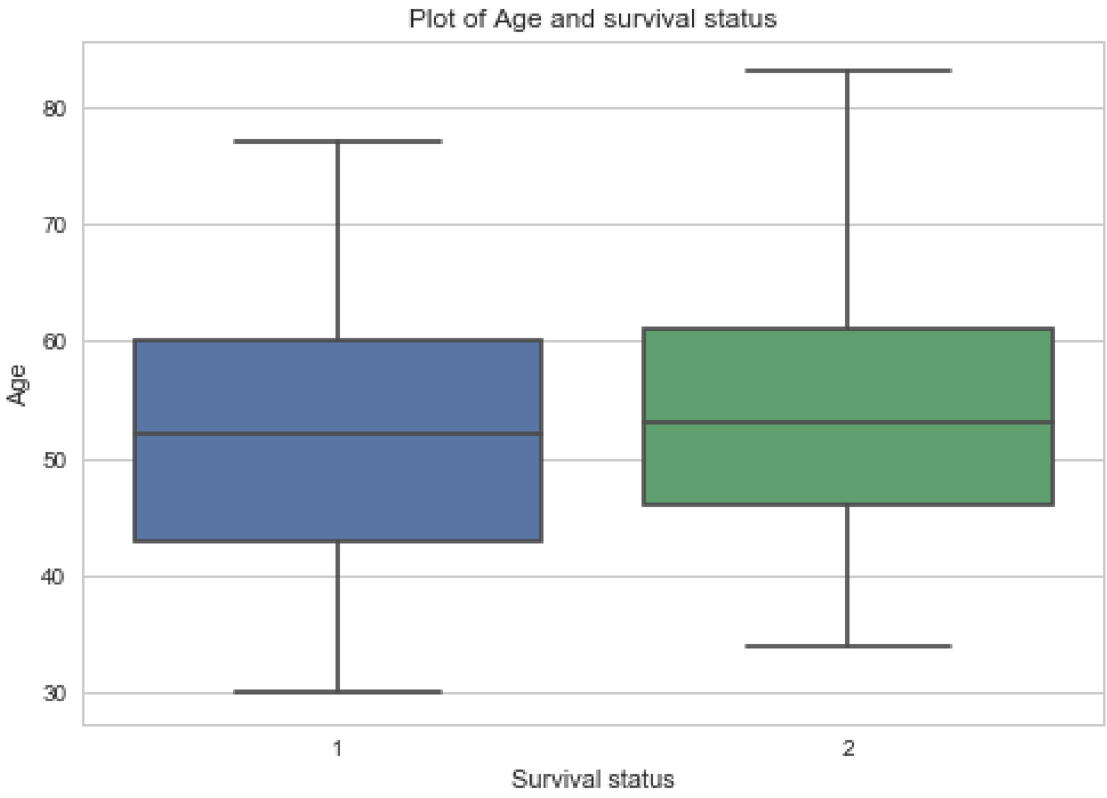
4 Objective

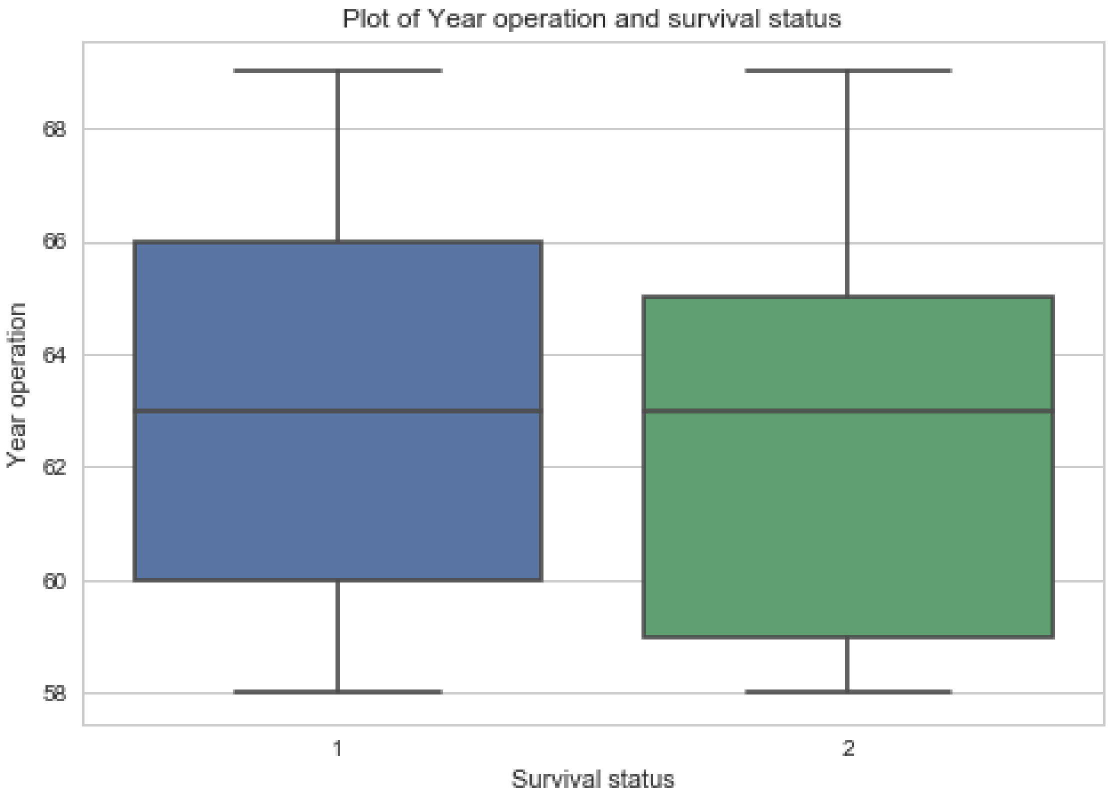
1. To find out pattern of cancer survival with Age of patience, Patient's year of operation, Number of +ve auxilary node detected

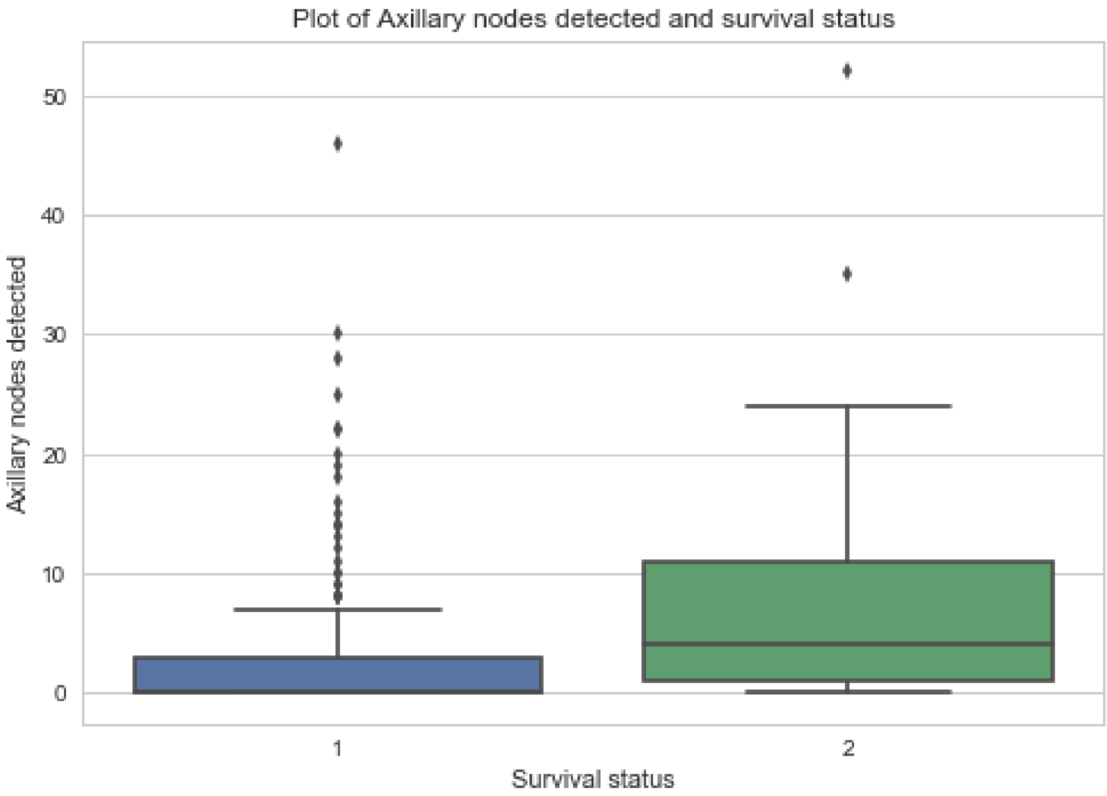
```
In [9]: # Removed the below graphs and could not see any pattern

        names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.set_style("whitegrid")
        sns.FacetGrid(dataset,hue='Survival status',size=4) \
         .map(plt.scatter,"Age", "Year operation").add_legend()
        plt.title("Year of operation and survival")
        plt.ylabel("Frequency of Age")
        plt.show()

        #sns.set_style("whitegrid")
        #sns.FacetGrid(dataset,hue='Survival status',size=4) \
        # .map(plt.scatter,"Year operation", "Axillary nodes detected").add_legend()
        #plt.show()

        #pairplot difficult for many feature(use pca tsne for many catagory)
        #sns.set_style("whitegrid")
        #sns.pairplot(dataset,hue="Survival status",size=4)
        #plt.show()
        #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']

        #25th quantile=X 25% of the data value are less than X
        #median abs deviation  IQR=75%value -25% value
        #boxplot and whisker(min max)
        names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.boxplot(x='Survival status',y='Age',data=dataset)
        plt.title("Plot of Age and survival status")
        plt.show()

        names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.boxplot(x='Survival status',y='Year operation',data=dataset)
        plt.title("Plot of Year operation and survival status")
plt.show()
```

```
names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
sns.boxplot(x='Survival status',y='Axillary nodes detected',data=dataset)
plt.title("Plot of Axillary nodes detected and survival status")
plt.show()
```



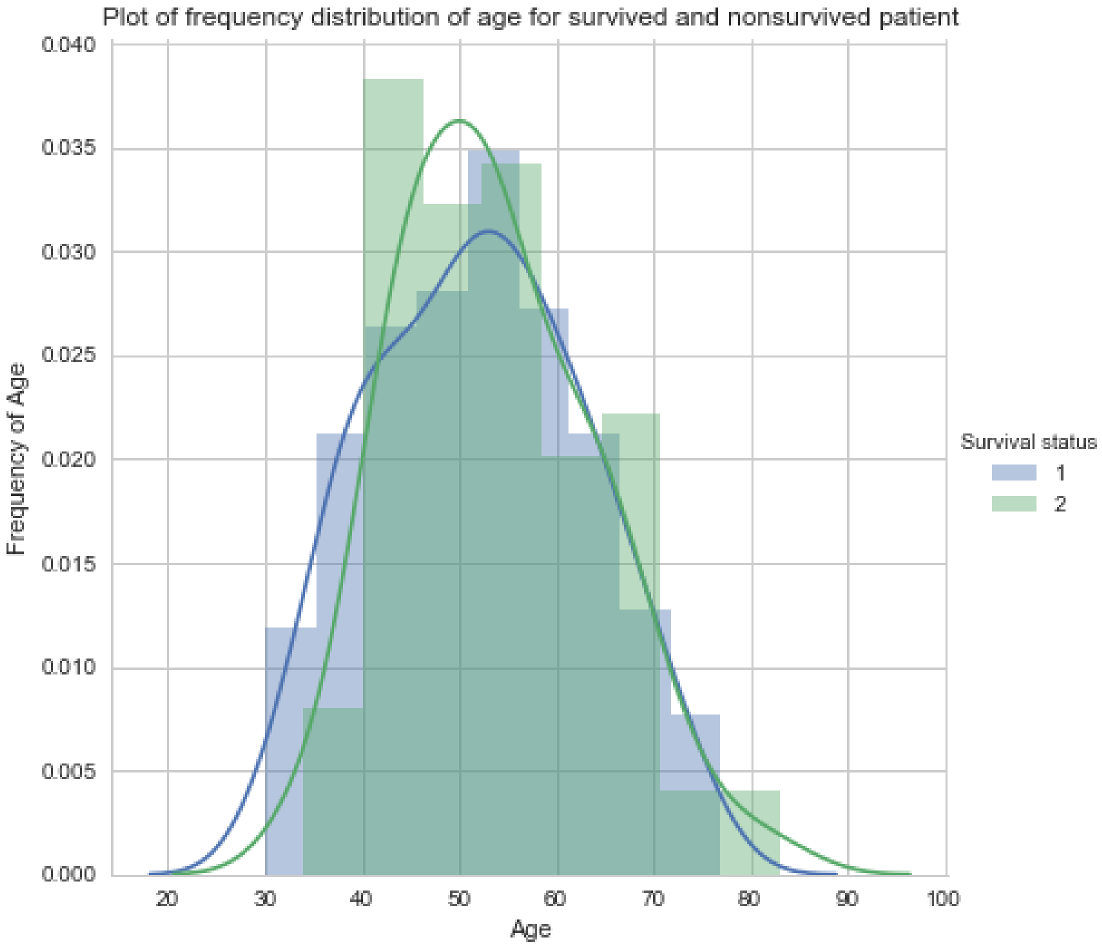

Plot of Age and survival status
Age
80
70
60
50
40
30
1
2
Survival status


Plot of Year operation and survival status
Year operation
68
66
64
62
60
58
1
2
Survival status



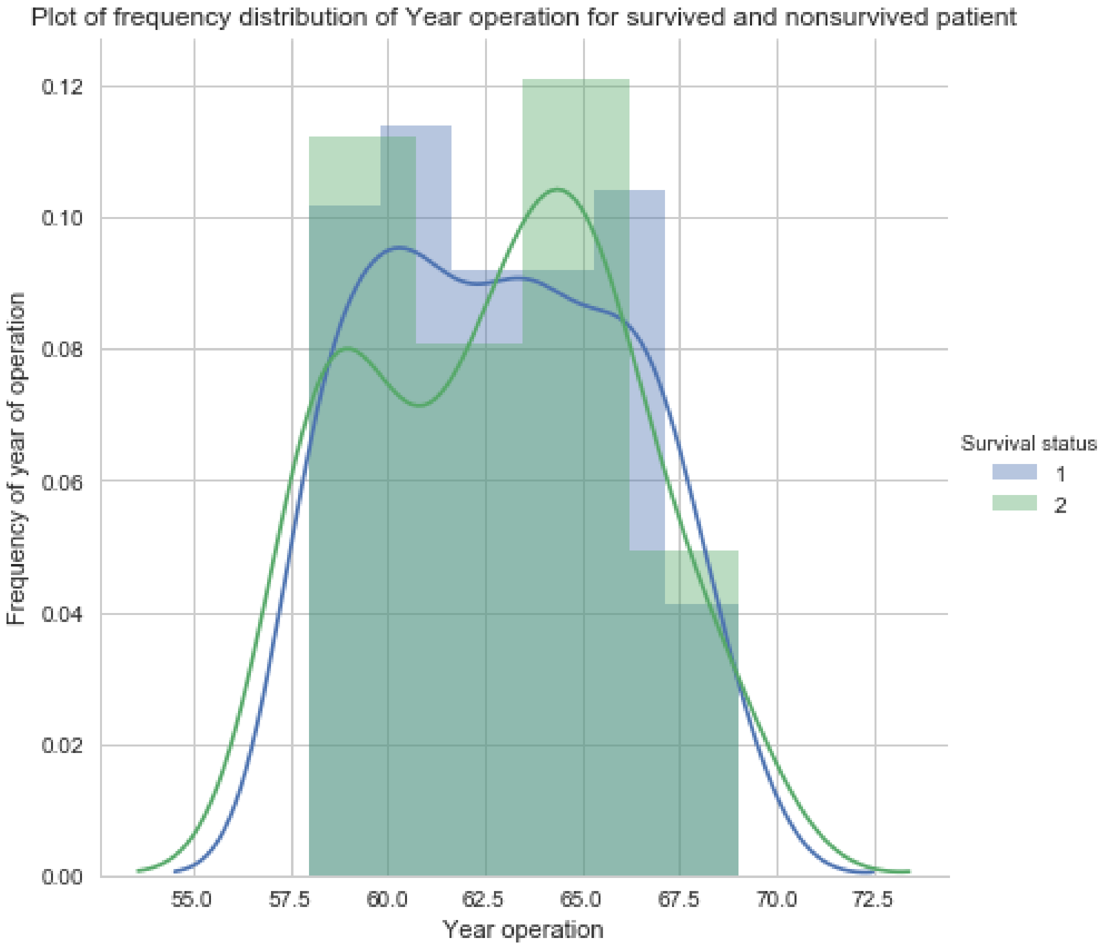
5 Different visualization of data

```
In [3]: #histogram
        #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.set_style("whitegrid")
        sns
        sns.FacetGrid(dataset,hue="Survival status",size=6) \
         .map(sns.distplot,"Age").add_legend()
        plt.title("Plot of frequency distribution of age for survived and nonsurvived patient")
        plt.ylabel("Frequency of Age")
        plt.show()
```



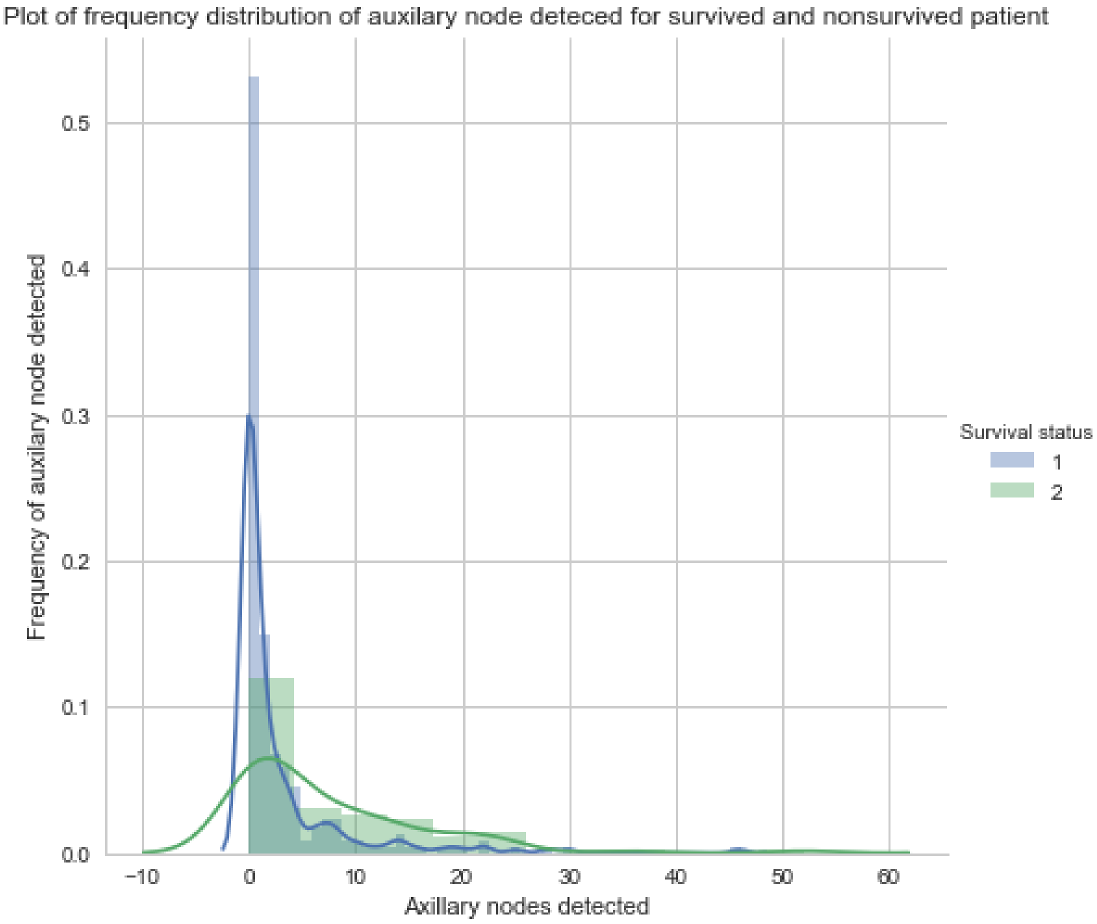
Most of the people who died are between age 35-75 and lived between age 30-75

```
In [4]: #histogram
        #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.set_style("whitegrid")
        sns.FacetGrid(dataset,hue="Survival status",size=6) \
         .map(sns.distplot,"Year operation").add_legend()
        plt.title("Plot of frequency distribution of Year operation for survived and nonsurvive
        plt.ylabel("Frequency of year of operation")
        plt.show()
```



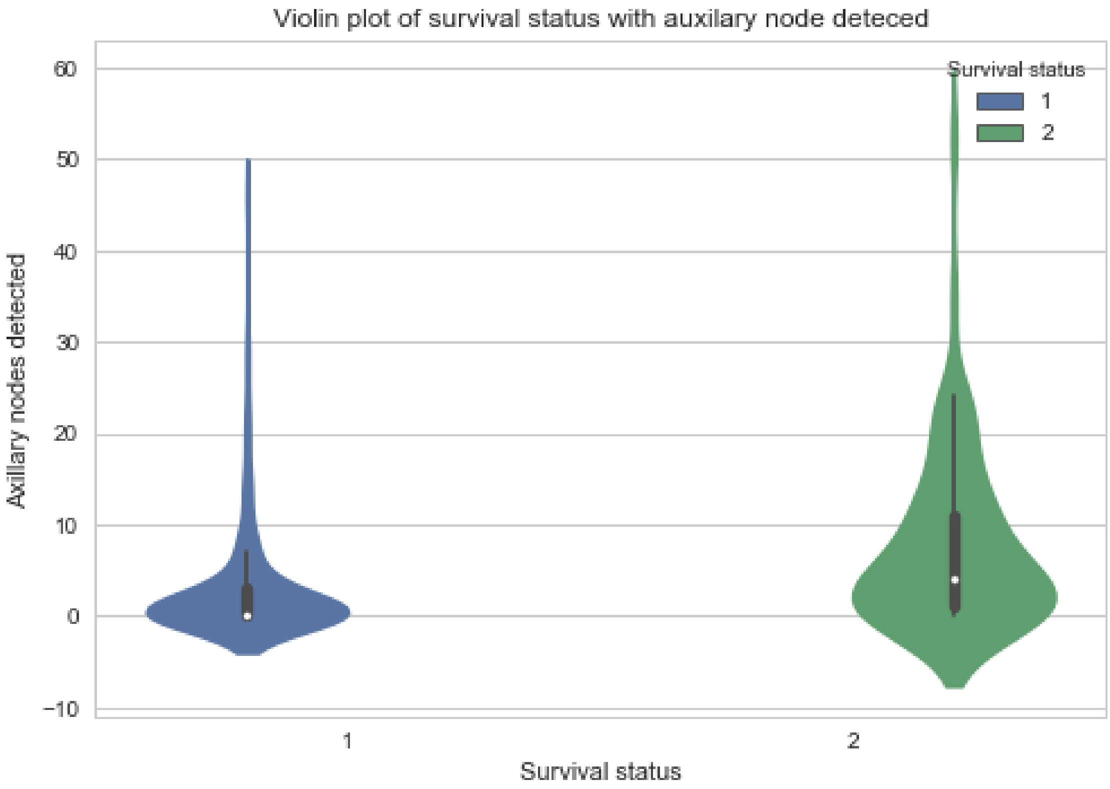
Most of the people have year of opertation between 60-66 survived and 59-65 died

```
In [5]: #histogram
        #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
        sns.set_style("whitegrid")
        sns.FacetGrid(dataset,hue="Survival status",size=6) \
         .map(sns.distplot,"Axillary nodes detected").add_legend()
        plt.title("Plot of frequency distribution of auxilary node deteced for survived and nor
        plt.ylabel("Frequency of auxilary node detected")
        plt.show()
```



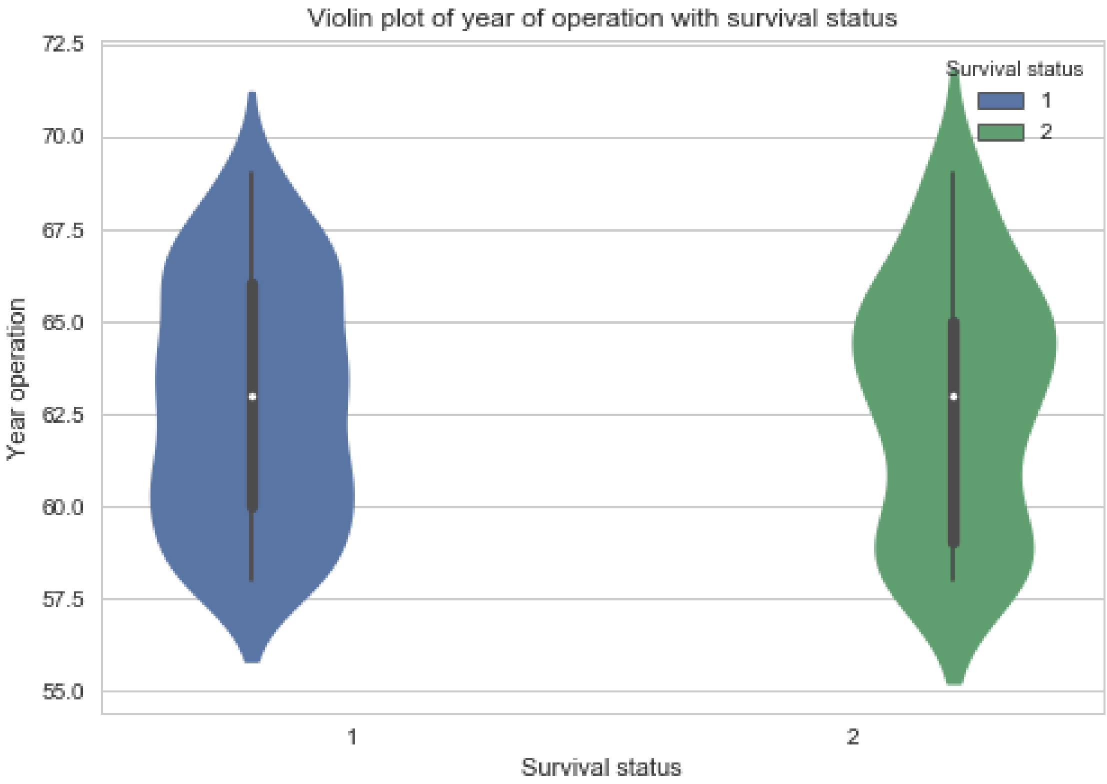
Most of the people survived when no auxilury node detected. This is one of the major indicator

```
In [37]: #violin plot: it plot pdf along with boxplot
         sns.violinplot(x='Survival status',y='Axillary nodes detected',hue="Survival status",d
         plt.title("Violin plot of survival status with auxilary node deteced")
         plt.show()
```



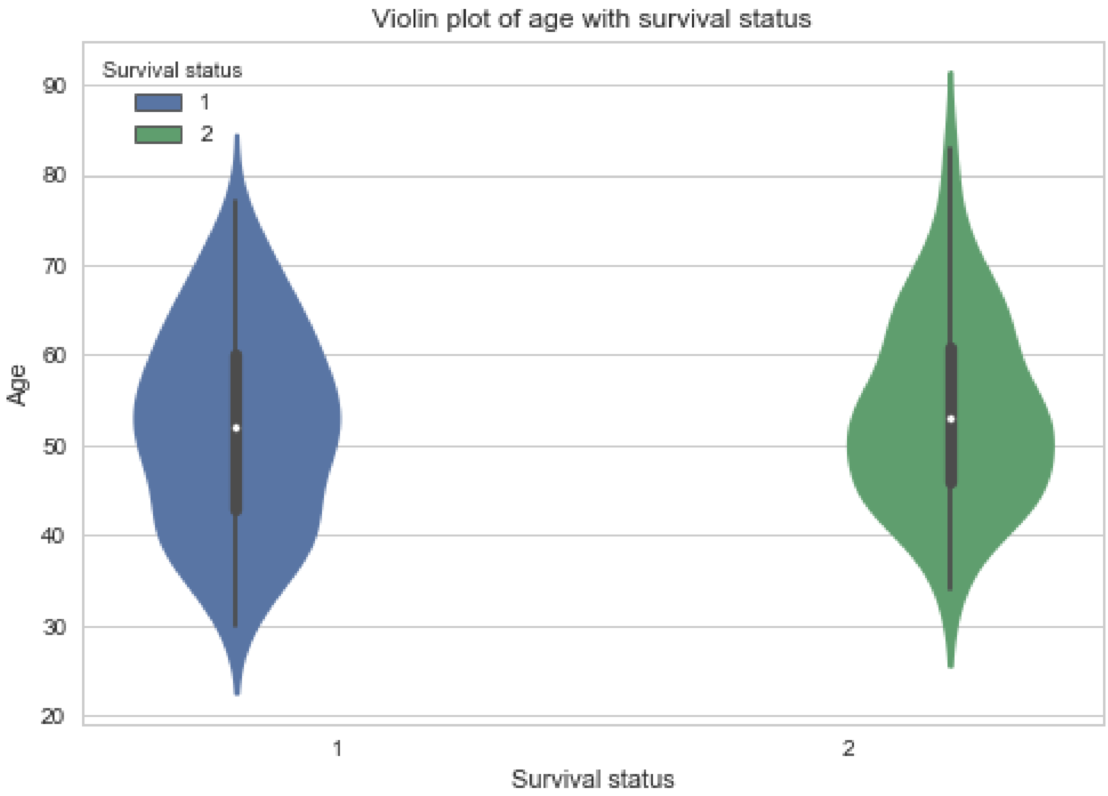
Most of the people survived when no auxilury node detected

```
In [38]: #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
         sns.violinplot(x='Survival status',y='Year operation',hue="Survival status",data=datas
         plt.title("Violin plot of year of operation with survival status")
         plt.show()
```



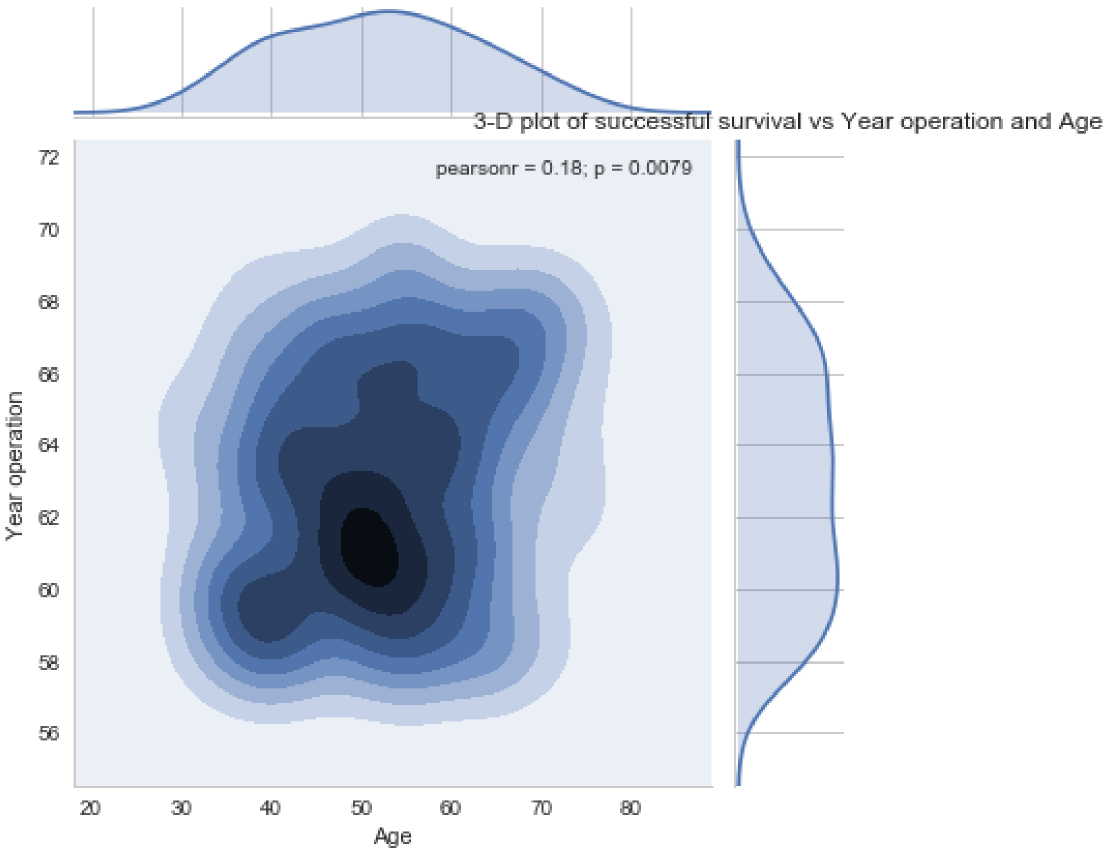
Most of the people have year of oertation between 60-66 survived and 59-65 died

```
In [39]: #violin plot is boxplot along with pdf plot
         #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
         sns.violinplot(x='Survival status',y='Age',hue="Survival status",data=dataset)
         plt.title("Violin plot of age with survival status")
         plt.show()
```



Most of the people who died are between age 35-75 and lived between age 30-75

```
In [41]: #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
         #sns.jointplot(x='Year operation',y='Axillary nodes detected',data=dataset_1,kind="kd
         #plt.show()
         #names = ['Age', 'Year operation', 'Axillary nodes detected', 'Survival status']
         #pdf for 2 dimension, contour plt
         dataset_1=dataset[dataset['Survival status']==1]
         sns.jointplot(x='Age',y='Year operation',data=dataset_1,kind="kde")
         plt.title("3-D plot of successful survival vs Year operation and Age")
         plt.show()
```



People with age between 50-55 and year of operation between 60-62 has a high rate of survival rate

6 Findings/Conclusion

Most of the patients are between age 30-75

Data is quite imbalance as 1/3 of the total died within 5 years

Most of the people who died are between age 35-75 and lived between age 30-75. So maybe between 30-35 years chances of survival high

Most of the people have year of opertation between 60-66 survived and 59-65 died. There is almost no clear distinction

Most of the people survived when no auxilury node detected. This is one of the clear indication that patient with no auxilary node detected has a high chance of survival

People with more age survived less with cancer but they are very close and difficult to distinguish

People with age between 50-55 and year of operation between 60-62 has a high rate of survival rate. These 2 combination is good for survival