EDA on Haberman cancer servival dataset

Dataset description:

- 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:

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year

In [41]:
```
import warnings
warnings.filterwarnings("ignore")
```

In [42]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [43]:
```
df=pd.read_csv("dataset/haberman.csv")
```

In [44]:
```
df.columns=["age","op_year","axil_nodes_det","surv_status"]
```

High level statistics of the dataset:

-lets perform high level statistics -here we are going through following things 1)number of data points in the dataset 2)number of features and understanding those feature 3)number of class labels in the dataset 4)number of data points for each classes

In [45]:
```
df.shape
```

Out[45]:
```
(305, 4)
```

In [46]:
```
df.columns
```

Out[46]:
```
Index(['age', 'op_year', 'axil_nodes_det', 'surv_status'], dtype='object')
```

In [47]:
```
df["surv_status"].value_counts()
```

Out[47]:
```
1    224
2     81
Name: surv_status, dtype: int64
```

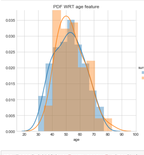
Objective:

-Given a cancer survival dataset ,our task is to classify wheather the patient will going to survive for next five years or not based on given three attributes that is operation year ,year of operation and axilary lymph node

Perform Univariate analysis

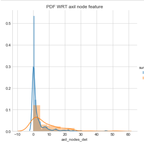
1)Probability density function:

In [48]:
```
sns.FacetGrid(df,hue="surv_status",size=5) \
.map(sns.distplot,"age") \
.add_legend()
plt.title("PDF WRT age feature")
plt.show()
```



In [49]:
```
sns.FacetGrid(df,hue="surv_status",size=5) \
.map(sns.distplot,"op_year") \
.add_legend()
plt.title("PDF WRT opration year feature")
plt.show()
```

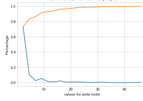
In [50]:
```
sns.FacetGrid(df,hue="surv_status",size=5) \
.map(sns.distplot,"axil_nodes_det") \
.add_legend()
plt.title("PDF WRT axil node feature")
plt.show()
```



2)comulative density function:

In [51]:
```
df_1=df[df["surv_status"]==1]
df_2=df[df["surv_status"]==2]
```

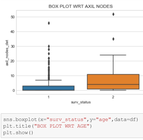
In [52]:
```
counts,bin_edges=np.histogram(df_1["axil_nodes_det"],bins=20,density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)
plt.title("PDF AND CDF for survival people")
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel("values for axile node")
plt.ylabel("Percentage")
plt.show()
```



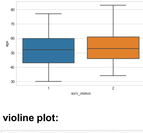
In [53]:
```
counts,bin_edges=np.histogram(df_2["axil_nodes_det"],bins=20,density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)
plt.title("PDF AND CDF for people who will going to die")
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel("values for axile node")
plt.ylabel("Percentage")
plt.show()
```

3) Box plots:

In [54]:
```
sns.boxplot(x="surv_status",y="axil_nodes_det",data=df)
plt.title("BOX PLOT WRT AXIL NODES")
plt.show()
```

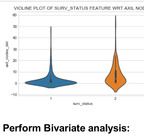


In [55]:
```
sns.boxplot(x="surv_status",y="age",data=df)
plt.title("BOX PLOT WRT AGE")
plt.show()
```



violine plot:

In [56]:
```
sns.violinplot(x="surv_status",y="axil_nodes_det",data=df)
plt.title("VIOLINE PLOT OF SURV_STATUS FEATURE WRT AXIL NODES")
plt.show()
```



Perform Bivariate analysis:

1)2-D scatter plot:

In [57]:
```
sns.set_style("whitegrid")
sns.FacetGrid(df,hue="surv_status",size=4) \
    .map(plt.scatter,"axil_nodes_det","age") \
    .add_legend();
plt.title("2D SCATTER PLOT :AXIL NODES VS AGE")
plt.show()
```

2)pair plots:

In [58]:
```
sns.set_style("whitegrid")
sns.pairplot(df,hue="surv_status",vars = ["age","axil_nodes_det","op_year"],size=4)
plt.title("PAIR PLOT")
plt.show()
```

Observation/conclusion: -haberman cancer servival dataset contains 305 data points and 4 different features •surv_status is class label feature •the dataset is imbalanced •As you can see pdf of all three features ,what we found is none of the feature/attribute is helpful in classifying survival chances •If still you go deep ,with the error you can say axil_nodes_det will be better compared to other two feature •From cdf you can say 90% of survival people have axile node value less than 10 and 60% non-survival people have value greater than 5 •This above statement giving us little idea to differentiate survival and non-survival people •from boxplots we can say that,survival patients have axile value very less, only few have gone above 3 or 4 •axile node values for non survival patients are overlapped with values of survival that mean there is some other reason that they dead ,which can be due to age factore or may be op_age •for more understanding i also plotted box plot of age ,which is giving very little idea that ages of some patients who will going to die are slightly higher compare to who will survive •box plot in violine plot tells us ,values of axile nodes for survival patients are not that much spreaded they are in very little range •where as values of patients who will die have far range as compare to survival patients range •as we see above 2d-scatter plot ,data points are overlapped ,so its difficult to conclude or build simple models ,if you build also that will come with many error . •so our conclusion is we will try pair plots first and find which combination is better if any combination is better then will draw scatter plot of that. •Above pair plot tells, no combination of any feature is giving good idea to classify •so no need to draw scatter plot also