Habermann EDA

June 6, 2019

1 Habermann DataSet EDA

2 1.High Level Statistics

1.1. Loading the data

```
In [11]: import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         import numpy as np


         #Load habermann.csv into a pandas dataFrame.
         column_name = ['Age' , 'Operation_Year' ,'Axil_Nodes' , 'Surv_Status' ]
         haberman = pd.read_csv('haberman.csv' , header = None , names = column_name)
```

```
In [12]: haberman
```

Out[12]:

	Age	Operation_Year	Axil_Nodes	Surv_Status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1
5	33	58	10	1
6	33	60	0	1
7	34	59	0	2
8	34	66	9	2
9	34	58	30	1
10	34	60	1	1
11	34	61	10	1
12	34	67	7	1
13	34	60	0	1
14	35	64	13	1
15	35	63	0	1
16	36	60	1	1
17	36	69	0	1
18	37	60	0	1

```
19    37              63            0             1
20    37              58            0             1
21    37              59            6             1
22    37              60           15             1
23    37              63            0             1
24    38              69           21             2
25    38              59            2             1
26    38              60            0             1
27    38              60            0             1
28    38              62            3             1
29    38              64            1             1
..   ...             ...          ...           ...
276   67              66            0             1
277   67              61            0             1
278   67              65            0             1
279   68              67            0             1
280   68              68            0             1
281   69              67            8             2
282   69              60            0             1
283   69              65            0             1
284   69              66            0             1
285   70              58            0             2
286   70              58            4             2
287   70              66           14             1
288   70              67            0             1
289   70              68            0             1
290   70              59            8             1
291   70              63            0             1
292   71              68            2             1
293   72              63            0             2
294   72              58            0             1
295   72              64            0             1
296   72              67            3             1
297   73              62            0             1
298   73              68            0             1
299   74              65            3             2
300   74              63            0             1
301   75              62            1             1
302   76              67            0             1
303   77              65            3             1
304   78              65            1             2
305   83              58            2             2

[306 rows x 4 columns]
```

1.2. Some high level statistics

```
In [13]: # (Q) how many data-points and features?
         print (haberman.shape)
```

```
(306, 4)
```

```
In [14]: #(Q) What are the column names in our dataset?
         print (haberman.columns)
```

```
Index(['Age', 'Operation_Year', 'Axil_Nodes', 'Surv_Status'], dtype='object')
```

```
In [17]: #(Q) How many data points for each class are present?

         haberman["Surv_Status"].value_counts()
```

```
Out[17]: 1    225
         2     81
         Name: Surv_Status, dtype: int64
```

3 2. Objective

To find and predict whether the patient will survive the given treatment or not

4 3. Univariate Analysis

5 3.1. PDF

```
In [24]: # Age
         sns.FacetGrid(haberman , hue = 'Surv_Status' , size = 4).map(sns.distplot , 'Age').ad
         plt.show();
```

```
C:\Users\MANISH\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'n
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\MANISH\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'n
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```