Exploratory Data Analysis: Haberman's Survival

The Haberman's survival 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.

Objective: To classify/predict a patient survival who had undergone operation for breast cancer in year 1958-1970.

Column Description

Age-Age of patient at time of operation

Operation year-Patient's year of operation

Auxiliary lymph node-Number of positive auxiliary nodes

Survival status - Survived or not

Libraries Used:

1. Pandas
2. Numpy
3. Matplot
4. Seaborn
5. Sklearn
6. Scipy

About Data set:

1. Here we can see that there are 305 rows and 4 columns.
2. There are no missing values in data set.
3. Data Type of all columns is integer.
4. There are total 305 patients. Mean age of patients is 52.53. Minimum age of the patient is 30 and max age of the patient is 83. 25% of people have age less than 44 years and 75% of the people have age more than 61 years.
5. Minimum Auxiliary lymph node in patients is 0 and maximum auxiliary lymph node in patients in 52. 75% of the people out of 305 have auxiliary lymph node more than 4.
6. Here the Target column is survival as we have to predict a patient survival who had undergone surgery for breast cancer. It has numerical value 1 and 2. So, we will convert it to categorical column with values yes and no.

   1 = the patient who survived
   2 = the patient who died
7. 224 patients survived and 81 patients died out of 305 patients.

Visualization

1. Age of patients is normally distributed with peak at 50 years.
2. patients who died have median value of age slightly more than that of people who survived.
3. People who survived have minimum age around 50 and max age around 75.
4. People who died have minimum age around 32 and maximum age above 80 years.
5. People who survived have minimum age around 50 and max age around 75.
6. People who died have minimum age around 32 and maximum age above 80 years.
7. Auxiliary lymph nodes have outliers value above maximum value in people who survived.
8. Auxiliary lymph nodes have outliers value above maximum value in people who died.
9. People who survived have maximum Axillary_lymph_node around 45.
10. People who died have maximum Axillary_lymph_node more than 50.
11. The people who survived have data points more concentrated along zero for axil_nodes
12. .The people who died have more spread along axil_node axis

Statistical Tests

H0: Mean of both samples is same

H1: Mean of both Samples is different

1. Survival and Age - T Test Independent

```
Ttest_indResult (statistic=1.1224778584494715,
pvalue=0.26254798164754417)
```

H0: mean Age of Survived people = Mean Age of died people.
HA: sample means are different.
α = 0.05
Here P value (0.262) is more than alpha i.e. 0.05 so, we will **reject the null hypothesis. Survival is independent of age**

2. Survival and operation year - T Test Independent

```
Ttest_indResult (statistic=-0.07094841414605883,
pvalue=0.9434856159625905)
```

H0: mean of Operation year of Survived people = Mean of operation year of died people.
HA: sample means are different.
α = 0.05 Here P value(0.94) is more than alpha i.e. 0.05 **so we will reject the null hypothesis survival is independent of year of operation of patients.**

3. Survival and Nodes - T Test Independent

```
Ttest_indResult (statistic=5.199154566746234,
pvalue=3.689473427782154e-07)
```

H0: mean of Node of Survived people = Mean of Node of died people
HA: sample means are different.
α = 0.05 Here P value is very much smaller than alpha, so **we will accept the null hypothesis. Survival is dependent on Axillary_lymph_node**.

After performing Statistical Tests we can see that

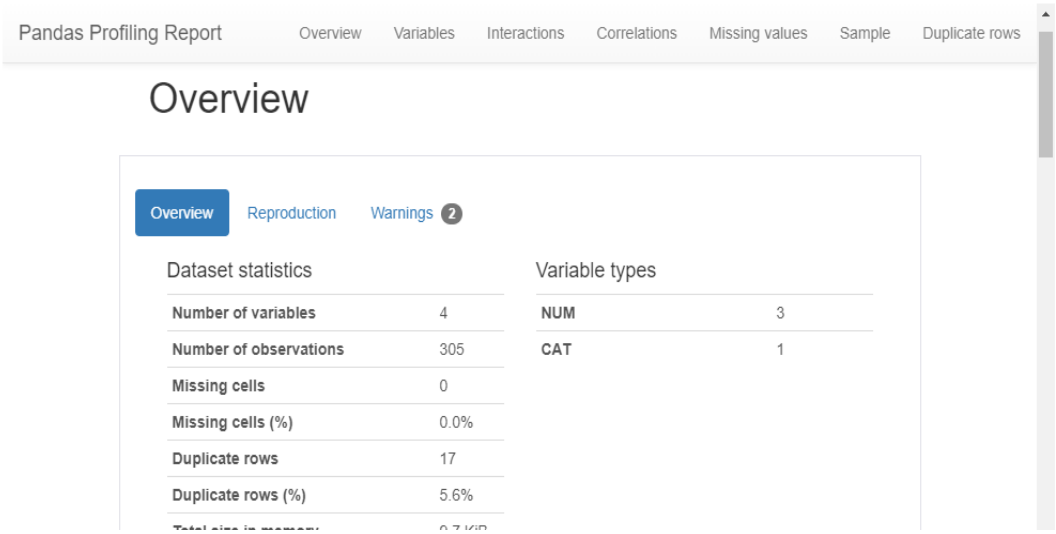
1. **Survival is Dependent on Axillary_lymph_node**
2. **Survival is independent of year of operation of patients**
3. **Survival is independent of age.**

Automated EDA Libraries: It can be used to reduce the time and efforts while Performing EDA on data set.

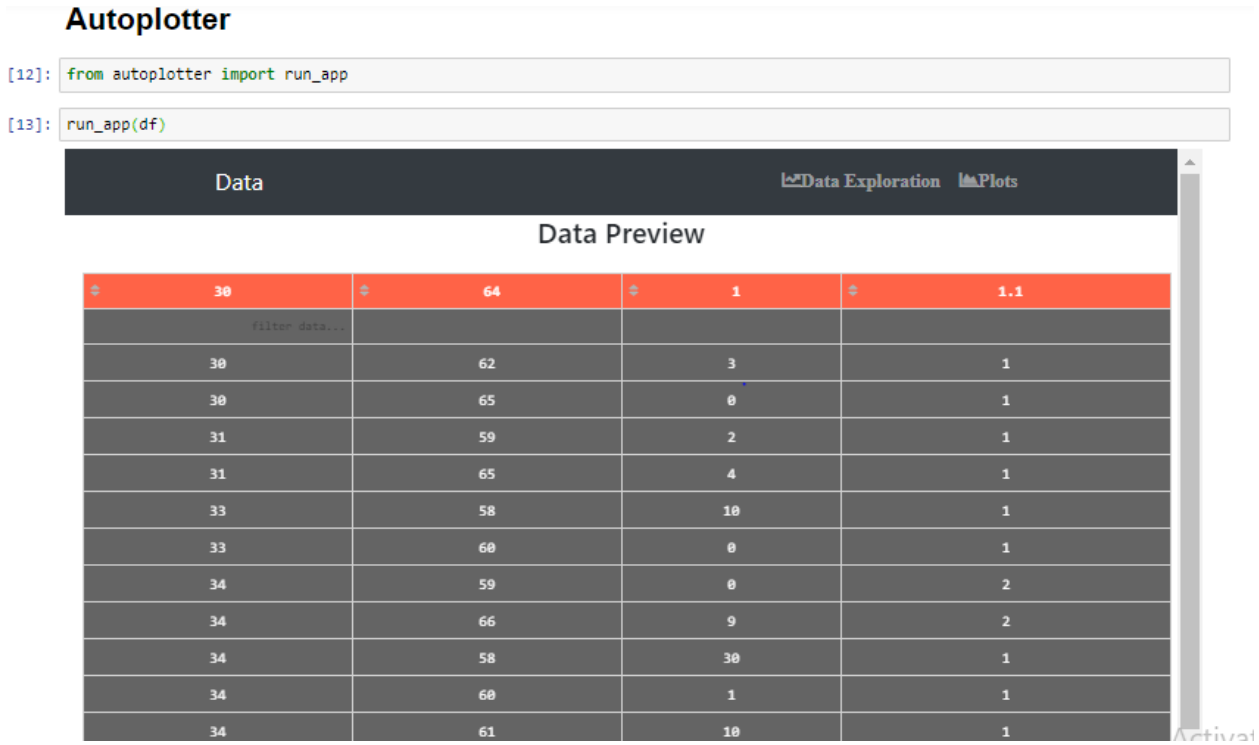
1. **Pandas Profiling:** It is a great tool to create reports in the interactive HTML format which is quite easy to understand and analyze the data.

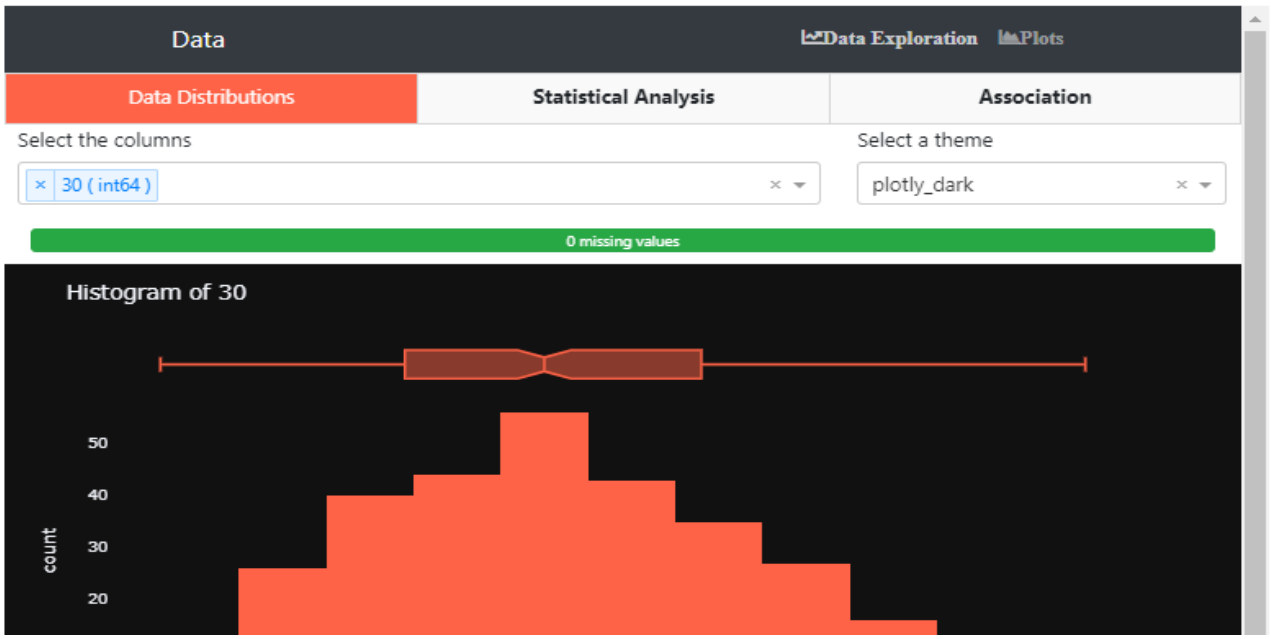
Pandas Profiling Report Consists of the Following Sections:

1. Overview
2. Variables
3. Interactions
4. Co-relations
5. Missing Values
6. Sample



2. **Auto plotter:** auto plotter is a python package for GUI based exploratory data analysis. It is built on the top of dash. We can choose X and Y axis while choosing the plots in auto plotter.




Data
Data Exploration
Plots
Data Distributions
Statistical Analysis
Association
Select the columns
30 (int64)
Select a theme
plotly_dark
0 missing values
Histogram of 30
50
40
30
20
count