International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-5, Issue-2 ,2018 Regular Edition E-ISSN: 2349-7084

AN EDA &PLOTTING TOOLS INTRODUCTION ON IRIS DATA SET

[1]R. Anil Kumar, [2]N. Ravi Kiran, [3]D. Nehemiah

[1]*Assistant Professor, Department of Computer Science and Engineering*
G.Pullaiah College of Engineering and Technology, Kurnool.
[2,3] *B.Tech Final Year Student, Dept. of Computer Science and Engineering,*
G.Pullaiah College of Engineering and Technology, Kurnool.

Received: 26/February/2017, Revised: 26/February/2017, Accepted: 01/February/2017, Published: 06/March /2017

Abstract:-Exploratory data analysis is a task of analyzing data from tools such as statistics, linear algebra, and some plotting techniques it is a very important task for a data set for analyzing data before building an actual machine learning models. It is called exploratory because we understand the data by being Sherlock Holmes. In this paper, we understand some basic plotting tools by using a real-world toy dataset (iris dataset)

Keywords:-IRIS, 2-D Scatterplot, Pair plots, Histogram, PDF, CDF.

1. Introduction

The Iris flower data set is a multivariate data set introduced by the British statistician. Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus" iris data set of 150 points: *https://en.wikipedia.org/wiki/Iris_flower_data_set*. It includes three iris species with 50 samples each as well as some properties of each flower. One flower species is linearly separable from the other two, but the other Two are not linearly separable from each other.

The columns in this dataset are:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species

2. Results and Discussion

2.1 2-D Scatterplot:

Scatter plot is a simple technique that we are scattering over the points of the data set and plotting it on the graph. From the below figure the plot was drawn By taking the values of x-axis of sepal_length and the values of sepeal_width for each flower in the data set of three different flowers

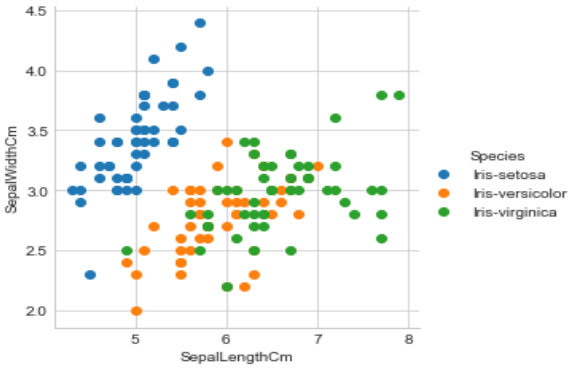

Figure 1. 2-D Scatterplot:

 https://doi.org/10.22362/ijcert/2018/v5/i2/v5i209

The immediate first take away from the above figure is

1. Using sepal_length and sepal_width features, we can distinguish Setosa flowers from others.
2. Separating Versicolor from Virginia is much harder as they have considerable overlap

2.2 Pair plots:

The small hack that is used for visualizing the data that has higher dimension at once is by using pair plots.it says pair the plots since the dataset has 4 features the total number of possible pair plots are 4c2 The below figure will show the pair plot of the iris dataset with four features. Figure can be understood by a 4*4 matrix. All the column X-axis are given at bottom and all the row y-axis are on left side. Matrix was divided into two halves above the diagonal elements, below the diagonal elements both halves are same both were mirror images of one another .institution that has gained from above diagnol is same as below diagonal elements

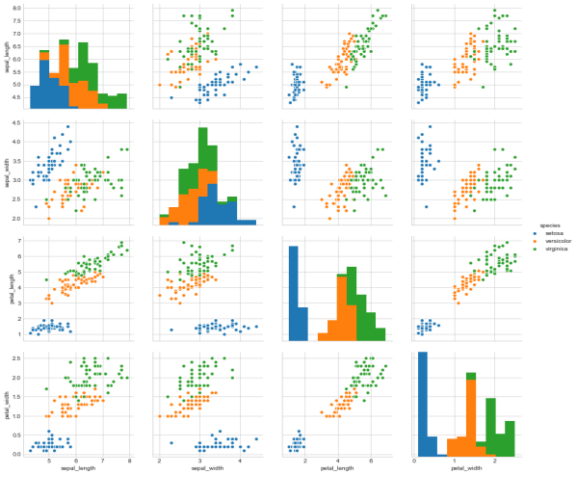

Figure 2. Pair plots:

Observations

1. petal_length and petal_width are the most useful features to identify various flower types.
2. While Setosa can be easily identified (linearly separable), Virnica and Versicolor have some overlap (almost linearly separable).
3. We can find "lines" and "if-else" conditions to build a simple model to classify the flower types

Disadvantages

If Dimensionality increase pair plot fails

2.3 Histogram, PDF, CDF:

1d scatterplot was drawn in below by taking the petal length feature Disadvantages of 1-D scatter plot:

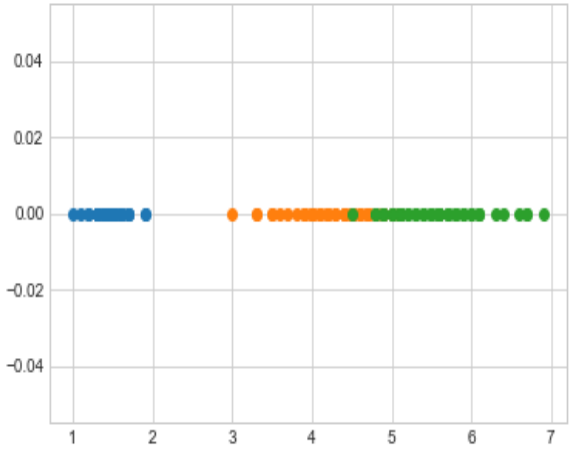

Figure 3. 1d scatterplot

Very hard to make sense as points They are overlapping a lot. Hence histogram is used to make a 1d scatterplot more clear.histrograms will show how many points exist on x-axis.it is actually an x,y plot where x-axis is the variable that you are interested in petal length in this problem context y-axis is how often does a data point apper corresponding X-axis

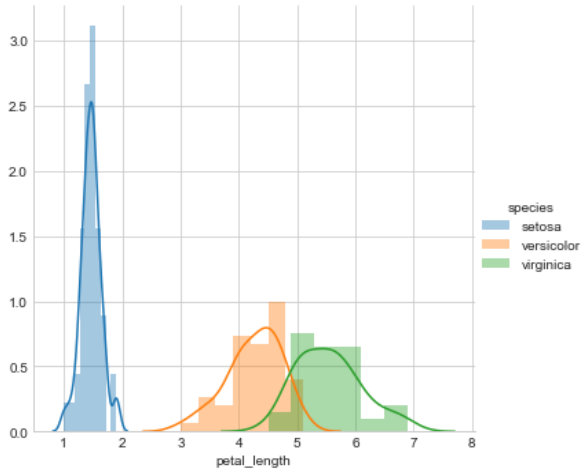

Figure 4. histogram with axis

Blue color is the histogram of petal length,orange color is histogram of Versicolor and green is of Virginia.Pdfs are

 https://doi.org/10.22362/ijcert/2018/v5/i2/v5i209

the smoothed surface of the jagged lines that are shown in above fig Smoothening of the jagged lines can be done by using kernel density function and distribution

2.4 Box-plots:

In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. It another method of visualizing the 1-D scatter plot more intuitively from the below figure the box of Versicolor species is drawn by drawing the middle line as 50^{th} percentile and below line as 25^{th} percentile above line as 75^{th} percentile this box literally tells what is the corresponding value of 25^{th} 50^{th} 75^{th} percentile of petal_length.this will also us what types errors.consider Y-axis of 5 which is roughly around 25^{th} value of Virginia is which means 25 percentile of petal_lengths of Virginia is labeled as Versicolor

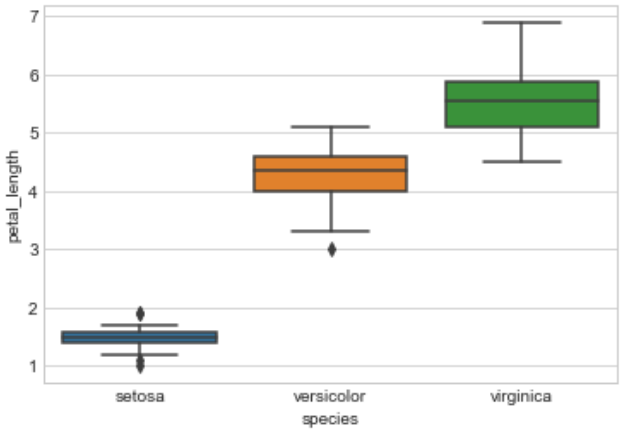

Figure 5. Box-plots:

2.5 Violin plots:

A violin plot combines the benefits of the previous two plots(histogram and box-plot) and simplifies them Denser regions of the data are fatter, and sparser ones thinner in a violin plot. A violin plot is more informative than a plain box plot. In fact while a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data. The difference is particularly useful when the data distribution is multimodal

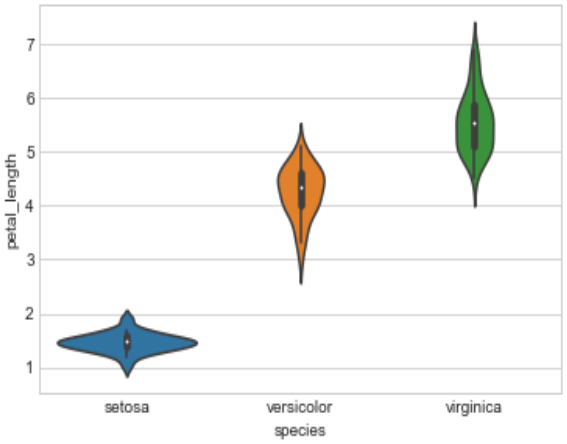

Figure 6. Violin plots:

3. Conclusion and Future Scope

In contrast, EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics is passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data. There is an archival role in the research and manufacturing world for summary statistics, but there is an enormously larger role for the EDA approach.

References:-

[1] http://www.lac.inpe.br/~rafael.santos/Doc /sR/CAP386/IntroEDA-Iris.html
[2] https://www.kaggle.com/lalitharajesh/irisdataset-exploratory-data-analysis
[3] https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python
[4] https://en.wikipedia.org/wiki/Exploratory_data_analysis
[5] https://en.wikipedia.org/wiki/Box_plot
[6] https://en.wikipedia.org/wiki/Violin_plot
[7] https://en.wikipedia.org/wiki/Pair
[8] https://en.wikipedia.org/wiki/Histogram

