Visualizing data using Lattice in R and Seaborn in Python for data science

Om Sehgal

Department of Computer Science, The NorthCap University, Gurugram, Haryana

DOI: 10.29322/IJSRP.9.12.2019.p9609
http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9609

Abstract- Visualization is the graphical representation of information and data. Using visual elements visualization tools helps understand trends, patterns and outliers in data with minimum complexity. There are two types of data visualizations, Univariate visualizations helps understand the distribution of a single variable and Multivariate visualization expresses the relationship between multiple variables. In a world of big data visualizations are crucial to make data-driven decisions by analyzing massive amounts of data. Many visualization methods such as scatter plots, bar charts, histograms, line charts, and pie charts, are widely used to tell stories removing the noise from data and zero in on the useful information. The better you convey your points visually, the better you can leverage that information.

Index Terms- Contrast between Lattice and Seaborn with the support multivariate plots, contrast between Lattice and Seaborn with the support of univariate plots for data science.

I. Introduction

Lattice is a data visualization library of R. Lattice is an application of Trellis graphics, framework for data visualization which is an elegant visualization system with prominence on multivariate data visualization. Lattice is a package for data visualization in R which is recognized as a key component in the R data science stack.
Seaborn is a library built on prime of Matplotlib. It allows one to make their visualizations prettier, and provides us with some of the common data visualization needs (like mapping a color to a variable or using faceting). Seaborn is more integrated for working with Pandas Data Frames.
Both the libraries are easy to understand and implement in their own field of usage. Seaborn has a straightforward syntax whereas for matplotlib there is more complexity with more variables to be defined depending on the user's requirements.

In addition, the published research work also provides a big weight-age to get admissions in reputed varsity. Now, here we enlist the proven steps to publish the research paper in a journal.

The remainder of the paper is as follows:
1) Section 2: Data Overview
2) Section 3: Multivariate Plots
3) Section 4: Univariate Plots
4) Section 5: Conclusions

II. Data Overview

The datasets used for this research purpose are self-made or self-generated. The same dataset has been used for fair comparison between the two libraries of python. This dataset is related to red variants of the Portuguese "Vinho Verde" wine.
Contents:

Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):
12 - quality (score between 0 and 10)

.

III. UNIVARIATE PLOTS

This type of plot is very easy to comprehend and used mostly worldwide containing plots like bar plot, histograms, box plots and many more. They help show the data and summarize its distribution. It describes observations for an individual variable. Also called as one variable at a time plot.

A. Histogram/Density Plot

A Histogram lets you come upon the spread of frequency of a set of continuous data. Allows inspection of underlying distribution.

Figure 1.1 – Histogram with Seaborn

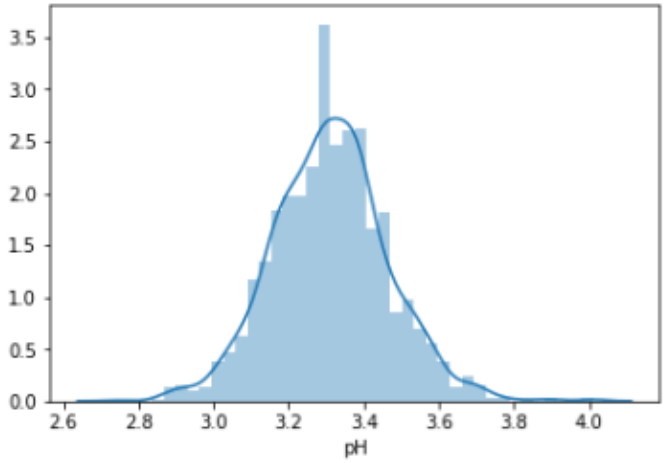
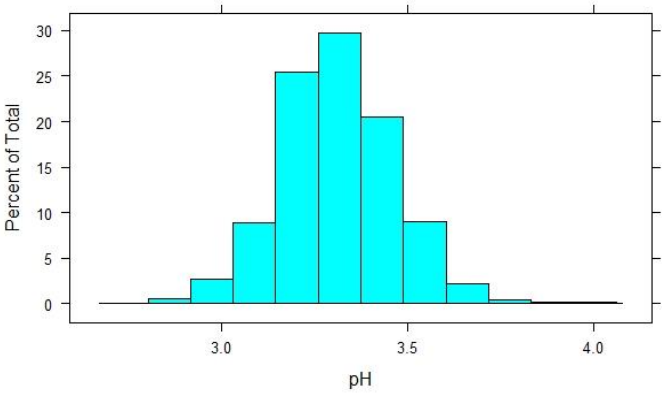


Figure 2.2 – Histogram with Lattice



Given similar datasets to the 2 libraries, we see that lattice's visualization concentrates more on how the data is spread whereas in the visualization by Seaborn, primary focus is on where the spread of data is thick, with the line known as the KDE or Kernel Density Estimate along it, the visualization is able to show how the course of the distribution is.

B. Box Plot/Whisker Plot

Box and whisker plots display the 5 – number summary of a set of data. 5 parts:

- Minimum
- First Quartile
- Median (Second Quartile)
- Third Quartile
- Maximum

Box plots graphically summarize groups of data through their quartiles.

Figure 2.1 – *Box Plot with Seaborn*

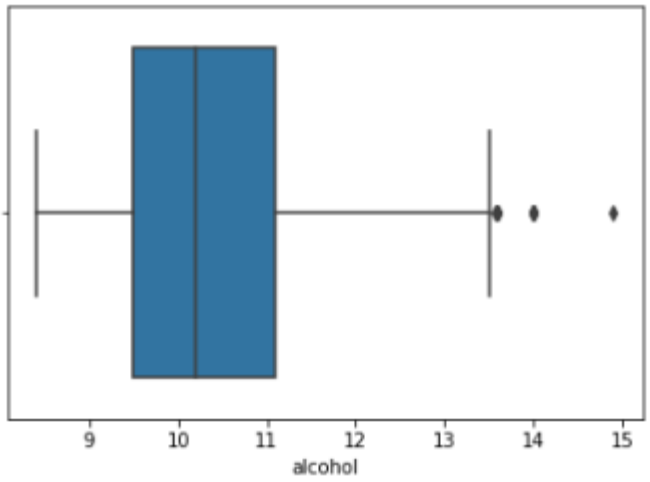
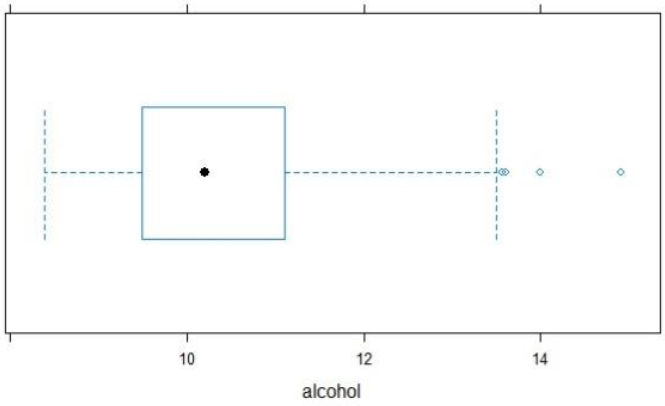


Figure 3.2 – *Box Plot with Seaborn*



When both libraries are provided with the same dataset the lattice's visualization the median is represented by a dot with big intervals on x-axis whereas the visualization with Seaborn color's itself and represents the median by a line and with small intervals on x-axis making the plot easily graspable and captivating.

IV. Multivariate Plots

Multivariate visualizations include the much commonly used scatter plot, heat maps, cluster maps and much more.

A. Scatter Plot

This represents values using dots for 2 different variables. This method of plotting using dots gives the concentration of data.

Figure 3.1 – *Scatter Plot with Seaborn*

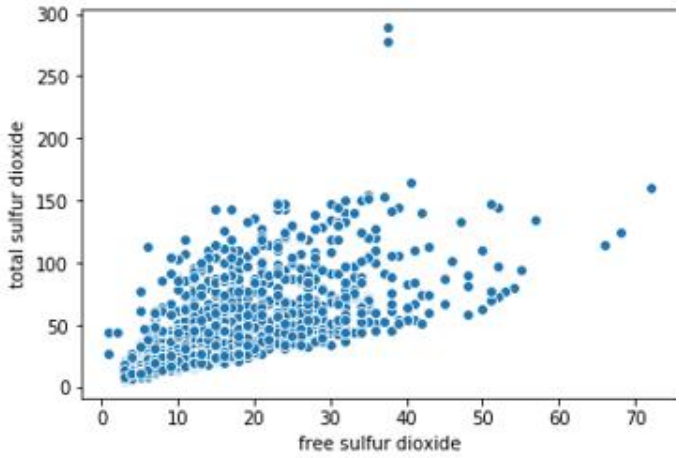
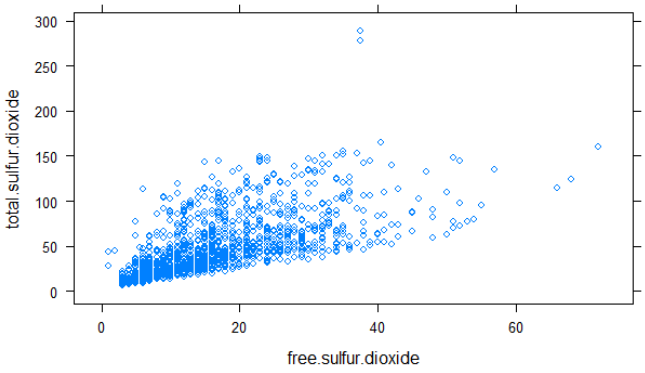


Figure 3.2 – Scatter Plot with Lattice



With the same dataset provided the lattice's visualization has high complexity with less visibility of the dots due to extensive overlapping whereas Seaborn provides us with presentable visualization providing colored dots with better visibility.

B. Heat map

It is a very effective plot where individual values in a matrix are displayed as colors. one can understand the occurrence density of data around an observation. Used often to understand the correlation of the data values.

Figure 4.1 – *Heat map with seaborn*

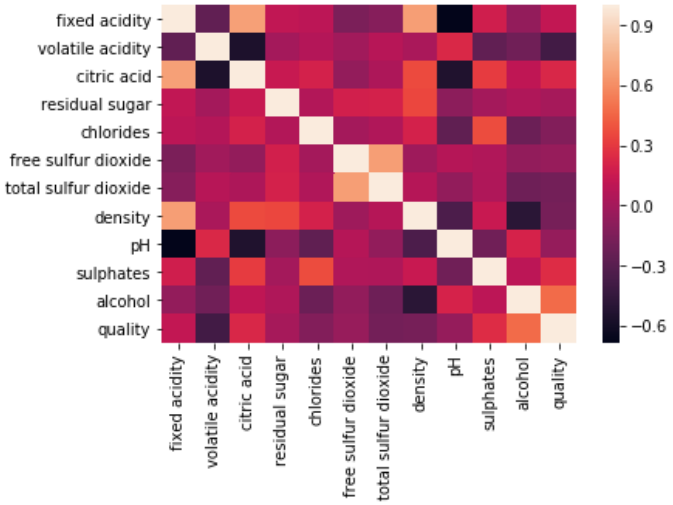
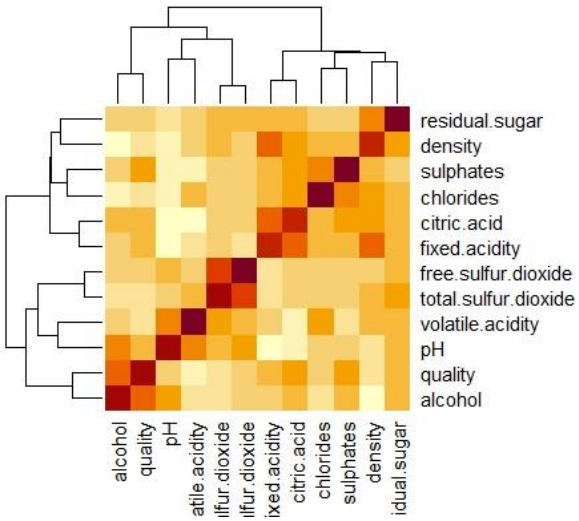


Figure 4.2 – *Heat map with lattice*



Whereas the heat map generated by lattice is poorly detailed and not presentable. Provides us with an hierarchical relationship between values.

V. CONCLUSION

Both the libraries seaborn and lattice are beneficial for visualization of data. Both the libraries produce a simple visualization quickly but when a clear-cut, presentable and explicit visualization is required, Seaborn is a cut above lattice providing each and every trivial characteristic for finer cognizance of the visualization.

Lattice works more with light visualization tools.

To create trailblazing visualizations seaborn could be used. Seaborn provides much more detailing upon the visualizations considering the dataset whereas lattice does not provide that much details under default state. Seaborn requires less effort for producing an explicit visualization compared to lattice which requires much more effort to produce an intricate visualization covering each and every minor detail.

REFERENCES

[1] Deepayan Sarkar "Lattice: Multivariate data visualization with R"
[2] Jessica Hamrick, *"Creating Reproducible, Publication – Quality Plots with Matplotlib and Seaborn"*
[3] Chris Mofitt, *"Choosing a Python Visualization Tool"*
[4] *Kieran Healy "Data Visualization: A practical introduction"*
[5] *Hadley Wickham "R for Data Science"*
[6] *Garrett Grolemund "Hands-On Programming with R"*

AUTHORS

First Author – Om Sehgal, B-Tech (CSE)
The NorthCap University
Omsehgal9211@gmail.com