Analysis of Data in R and Python

Jaswanth Krishna Kumar Patnala[1], Sitha Ramanjaneyulu Thota[2], Venkata Saranya Segu[3]

[1][2][3] Dept. of Computer Science and Engineering, Koneru Lashmaiah Educational Foundation, Guntur

---***---

Abstract - *We all know that R and Python, both are open-source programming languages, developed in early 1990's and the two most popular programming tools for Data Science work. While both languages are competing to be the Data Scientist's language of choice, it is hard to pick the best one out of these two languages i.e. R and Python. Yes, it is true if you just stepped into Data Science and looking for the best language to start with. We cannot pick one but can figure out some strengths and weaknesses of both languages. Even you know their pros and cons, it is your choice to pick one that suits best for your use case. This paper deals with the pros and cons of both the languages in deep by taking some examples. The machine learning algorithm which is being used in this example is already known by all of us. Datasets which we used during this project i.e. in the examples are inbuilt and some are taken in natural.*

Key Words: **Data Science, Programming Tools, Data Scientist, Pros and Cons, Machine Learning, Algorithm, Datasets etc**

1.INTRODUCTION

What is Data Science? Well, in simple words, it is just the study of data which involves storing, analyzing, visualizing, and extracting some useful information from it. The main goal is to gain insights i.e. to gain an accurate and deep understanding of data.

Data…. Data…. Data….! What is Data? Its not something new which just came from out of syllabus. It is the same thing what you know. Its just an information (usually numerical but can be text values also) which is collected from observation. And the Data Scientist, should know how to store, analyze, and extract data, which requires both tools and methods from statistics and machine learning. And, you should spend a lot of time in the process of collecting, cleaning and munging data, because the data is never clean.

Now, how we do all these things i.e. storing, analyzing, and extracting the information from the data which is useful to us. Obviously, instructing a computing device by using a programming language. Here comes the very big question, which programming language should we use/suits for this thing i.e. analysis of data.

In general, R and Python are the most used programming languages (also by Data Scientists) for Data Analysis. Even there are only two languages which are widely used, we will get into a self-doubt at a point of time on why we are using two different languages. Why can't we fix this to one language? R and Python both were developed in early 1990's and open-source languages. If you just stepped into this field i.e. as Data Scientist, you will be in a confusion state on selecting a programming language. It's better to pick one that suits for your use case.

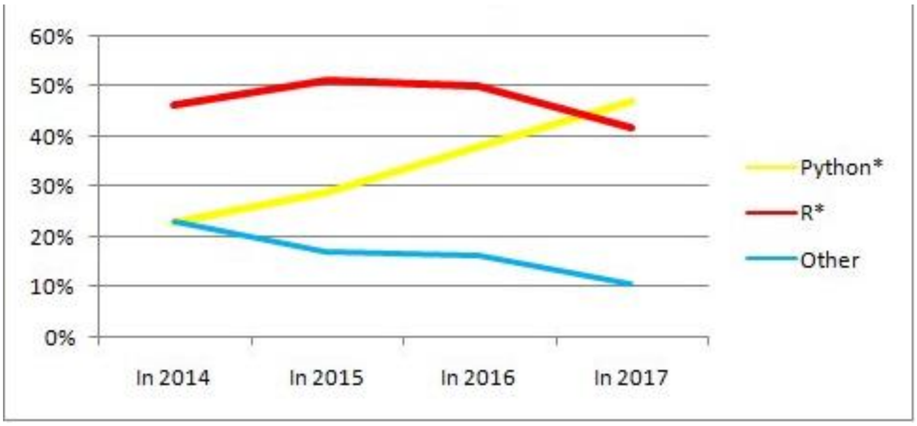
A code to a problem can be written in any language. It is up to your choice. Generally, we look up to a language which is simple, takes less time to implement and which completes in minimum number of lines. Just imagine! You are better in C, Python and JAVA and If you need to write a code to find sum of two numbers. What language will you choose? Will you choose JAVA? You will choose Python/C since it just takes 2 or 3 lines to write/implement the code. Now imagine like you need to solve a series of different problems where adding two numbers is a part in every problem and it should be done multiple times in each problem. Then you will prefer JAVA, since it is an Object-Oriented Programming Language and mainly it has a concept named packages which is best for present problem.

Both R and Python were good at their respective contents, but it is up to you to pick one that suits best for your use case and results best output. We just went through some of the topics of Data Science and found which language (R/Python) suits best for which context.

2. LITERATURE REVIEW

This is the division which deals with R and Python on some of the topics in Data Science. As already said, it is up to you to pick one that suits best for your use case. But we can differentiate two languages i.e. both R and Python, in various topics by mentioning some pros and cons.

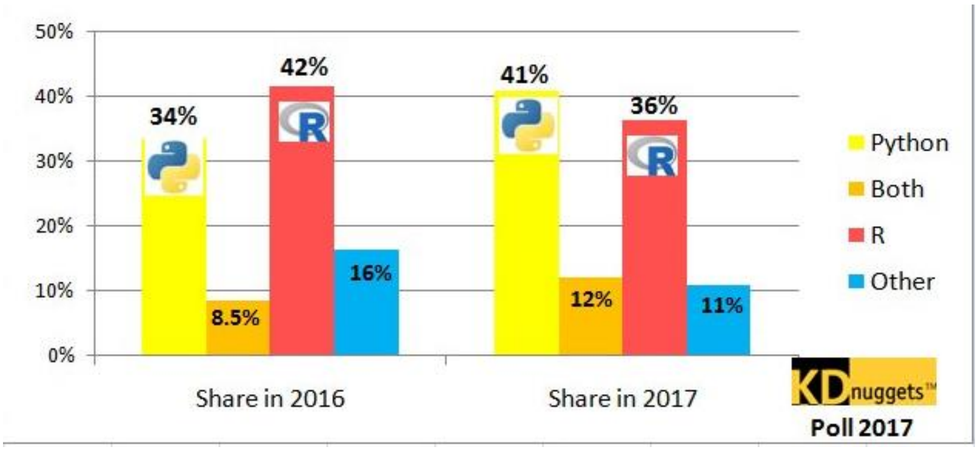
Below is a graph which shows the evolution of R, Python, and other languages for analytics (2014-17).



A picture says more than a thousand words. Visualized data can be understood more effectively and efficiently than raw data. R and Visualization are perfect match. R has different packages for making graphs like ggplot2, ggvis, googleVis, rCharts etc. But ggplot2 is one of the most qualified, refined, multi-skilled and flexible. ggplot2 executes Grammar of Graphics (data frame, aesthetic mappings, geom, Facets, Stats, Scales, Co-ordinate System). In contrast to base R graphics, it allows to add, remove, or alter components in a plot at a high-level abstraction. In the below example diamonds package(built-in) is taken as input and turned it into visualized data using ggplot2 by executing Grammar of Graphics.

The full-fledged programming nature in any subject can be seen in Algorithms or Classifications. Python code is easily understandable, also it helps in building machine learning algorithms easier. Since it is simple in syntax wise, it is faster in developing code than many other programming languages. It also allows the data scientist to test the algorithms quickly without having to implement them. It is better in machine learning algorithms since it consists of libraries such as NumPy, SciPy, Scikit-learn etc. It is also good at data handling since introduction of pandas. In the below example student's dataset is used as input which was taken in natural and it was used to build a k-Nearest Neighbor algorithm, also done predicting the output for various inputs with accuracy.

Below is a graph which shows the evolution of R, Python, both, and other languages for Machine Learning. (2016 - 2017).



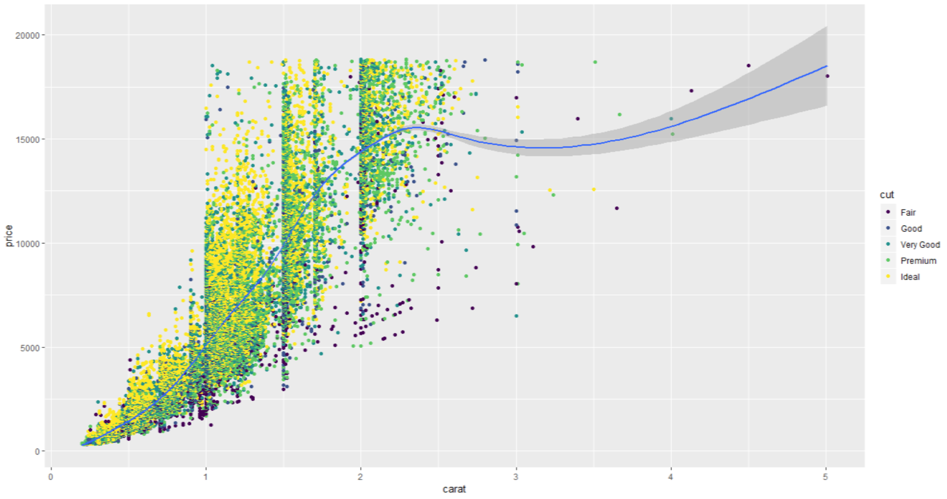
3. MATERIALS AND METHODLOGIES

3.1 Visualization of data in R using 'ggplot2':

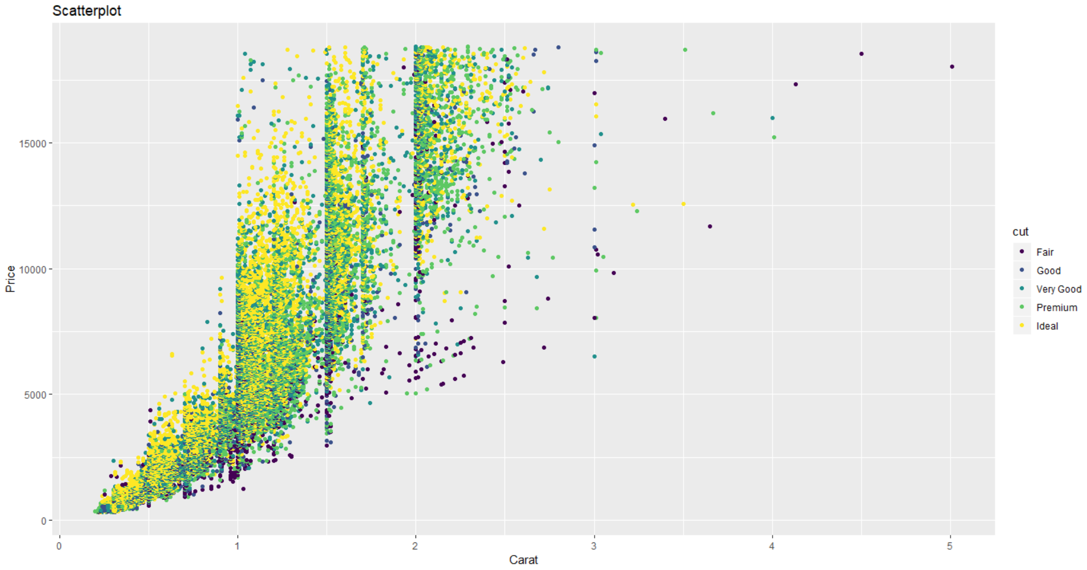
To visualize data, a large set of data set is needed. The data set which we used here is diamonds dataset which is inbuilt in RStudio.

3.2 Results:

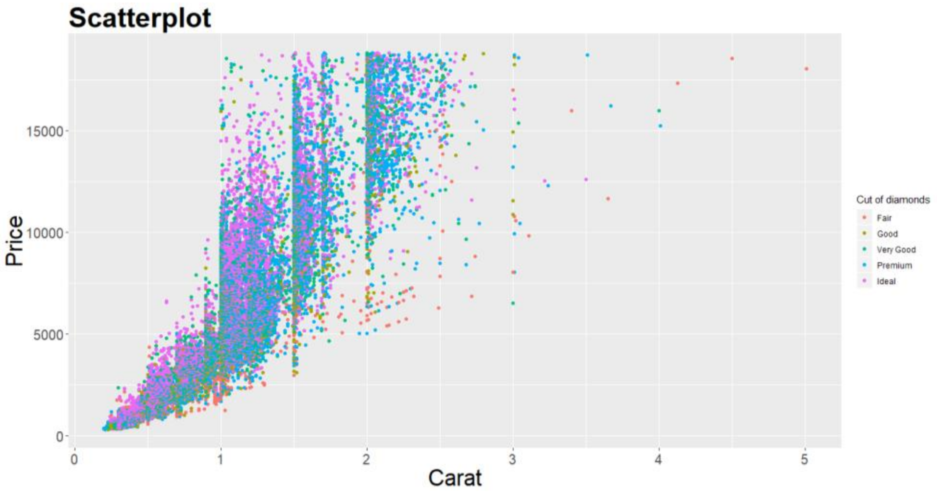
- Layers



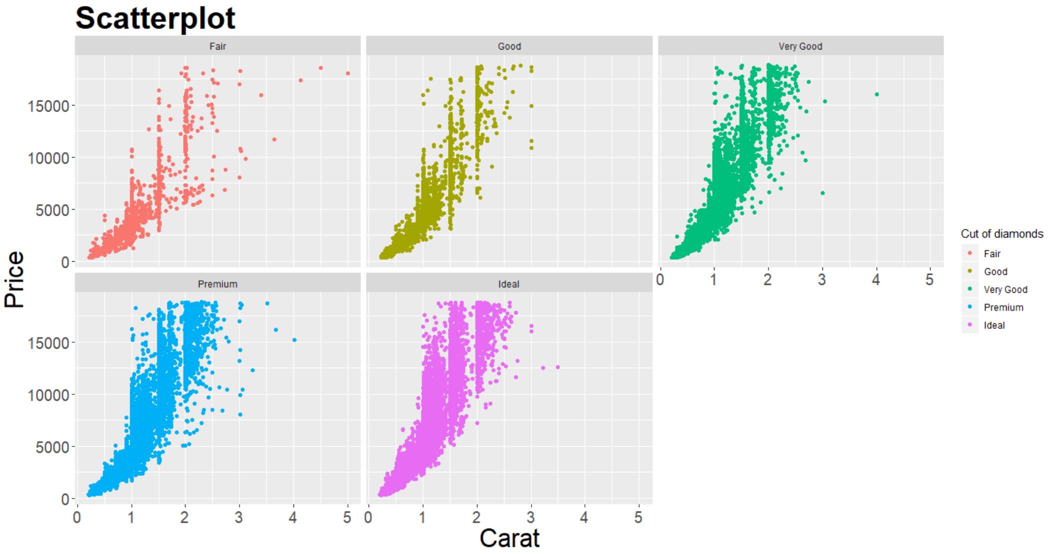
- Labels



- Theme



- Facets



- Look at all the above charts and just imagine telling all the above information through a table or a paragraph to someone. How long will it take to explain all the visualized information? It is impossible to tell the information without these visualizations that too in the today's world where the volume of data is increasing rapidly. This is where R offers an incredible help in doing any exploratory data analysis. Basic graphs such as Histogram, Bar/Line chart, Box plot, Scatter plot can be done easily in R like in any other platforms. But advanced graphs such as Heat Map, Mosaic Map, Correlogram and often 3D Graphs in R can be created quite easily. If you want a Heat Map, you can use the word heatmap, and for Mosaic Map you can use the 'mosaicplot' function.

3.3 Machine Learning Algorithm in Python:

To build up an algorithm, a data set is needed. The data set which we used here is student's dataset which was taken in natural. The detailed dataset is summarized below in table1.

Std.No	1st year CGPA	2nd year CGPA	Category
1	8.5	8.5	C
2	8.2	9	C
3	7.5	7.6	C
4	5.5	4.5	NC
5	9.2	9	C
6	7.8	7.3	C
7	7.3	7.4	NC
8	7.9	7	NC
9	10	6	C
10	6.8	7.1	NC
11	6.5	7.1	NC
12	7.2	7.3	NC

Table 1.

3.4 Results:

The Dataset is trained with 70% of training data while 30% of validation data. The average of the arguments 1st year CGPA and 2nd year CGPA results to C/NC. The algorithm which we build is k-Nearest Neighbor. The test was performed on each row from the validation set which predicts the output with an accuracy of 66.666%. Since the dataset which we took is small, the accuracy was less. If the dataset is large, then we can predict the output with best accuracy.

- As said earlier, Python code is easily understandable. Since it is simple in syntax wise, and easily understandable, any English-speaking person can easily understand the meaning of the code. It also allows the data scientist to test the machine learning algorithms quickly without having to implement them. Developers can implement any changes and can see the results quickly since there is no need of recompiling the source code. And the most important reason why python is good for machine learning is because of its portable and extensible nature. For training the ML models in their respective own machines, most of the data scientists use GPU's (Graphics Processing Units). The portable nature of this programming language i.e. Python is well suited for this.

4. CONCLUSION

As said, it is up to you, as a data scientist it is your job to select a language which fits best to the needs of yours. Before selecting any language just think of the problem which you need to solve. It gives the solution for picking up the right language.

5. ACKNOWLEDGEMENT

This research is supported by the Head of the Department (CSE), through a program called peer mentoring where students can work on their academic interests along with academic curriculum.

6. REFERENCES

[1] Data Science from Scratch (First Principles with Python) by Joel Grus.
[2] Machine Learning by Mitchell, Tata McGraw Hill Education Private Limited
[3] Data Science using R by John Hopkins University in Coursera.
[4] Machine Learning with Python by IBM in Coursera.