Exploratory Data Analysis in Pharo

Oleksandr Zaytsev
olk.zaytsev@gmail.com

Abstract

In this context... We consider this problem P... P is a problem because... We propose this solution... Our solution solves P in such and such way.

Pharo does not have the proper tools for data analysis. One of the most important tools is data frames - tabular data structures for data analysis. In this paper I demonstrate how the existing DataFrame can already be used for exploratory data analysis.

1 Introduction

Context

Problem

Explanatory data analysis is the ...

Known tracks for Stef ►*solutions*◄ here you want to show that you are not an idiot not knowing what have been around

What our solution is Set and OrderedCollection (so that the reader knows where the paper is going)

Contribution of the paper

Paper structure In the first section I provide a brief introduction into the explanatory data analysis. What is it good for? How to do it right? I will also provide you with all the basic knowledge about statistics and data analysis, such as statistical variable, types of variables etc. In the second section I briefly describe the DataFrame - new data structure for data analysis. In the following section I will give a step-by-step example of how to perform EDA on the well-known Iris data set.

2 Problem Description

Context, exposed with the **most precise terms possible** (don't open unwanted doors for the reader)

Probably set the vocabulary before to cut any misinterpretation

Constraints that influenced the solution (because the solution is not universal) *e.g.* our requirements for a solution, possibly not all satisfied. They should be sound and

believable. Analysis of the criteria. Imagine that you are another guy having this problem do the constraint matches yours so that you could apply the solution

Problem

Factual solution tracks, to position... Our solution in a nutshell.

3 Exploratory Data Analysis (EDA)

Here are the main reasons we use EDA:

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables
- assessing the direction and rough size of relationships between explanatory and outcome variables

4 DataFrame class

...

5 Exploring Iris Data Set

Free form, variable number of sections, technical details.

But in general do not mix solution and discussions/possible variation let that for discussion

We can start by loading iris data set from a CSV file.

```
data := DataFrame fromCsv: '/path/to/iris.csv'.
```

Now let's take a look at the first and the last 5 entries in our table. These slices are called *head* and *tail* of a data frame.

```
data head.
data tail.
```

5.1 Univariate non-graphical EDA

To access a single variable we ask a data frame for a specific column, using its name or integer index

```
var := data column: #sepal_width. "Accessing column by its name"
var := data columnAt: 1. "Getting column by index"
```

What we can do with a column depends on a type of statistical variable it represents. If the variable is categorical, we can ..., if its quantitative, we can look at the ... statistics.

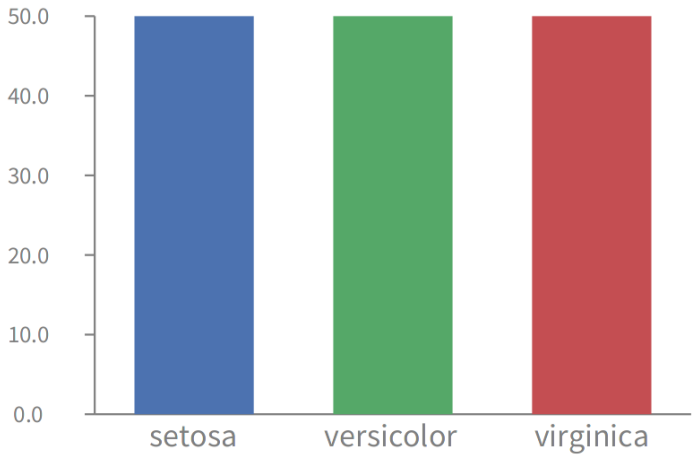
```
var min.
var max.
var range.
var average. "a.k.a. mean"
var median.
var mode.
var stdev.
```

5.2 Univariate graphical EDA

5.2.1 Categorical

Histogram is the only graphical technique that can be used for a categorical variable.

```
var := data column: #species.
var barplot.
```



5.3 Multivariate non-graphical EDA

...

5.4 Multivariate graphical EDA

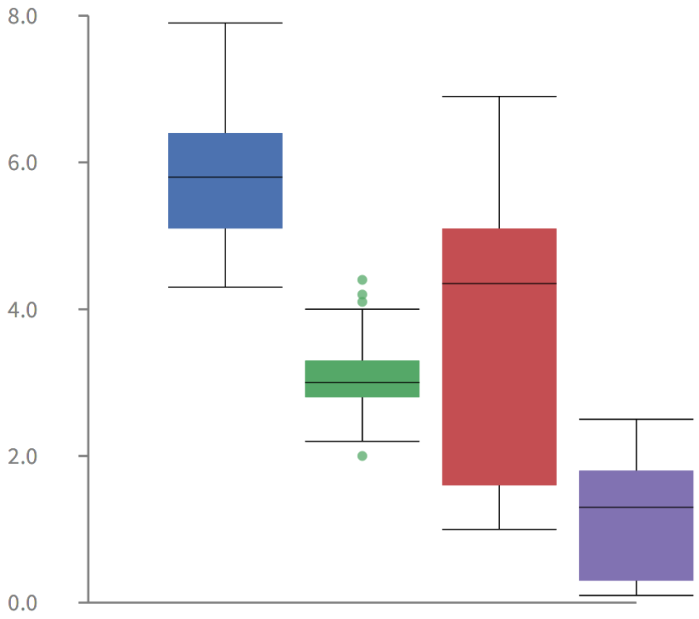


Figure 1: no need to put the file extension!

6 Discussion

Discussion of actual solution *vs.* initial constraints from 2. Explain the space of the solution, why we made it this way.

Evaluation of the solution. How does the solution meet the criteria? Where does it succeed or fails...

7 Related Works

Other solutions in the domain, and a real comparison of our contribution with solutions from other people.

8 Conclusion

In this paper, we looked at problem P with this context and these constraints. We proposed solution S. It has such good points and such not so good ones. Now we could do this or that.

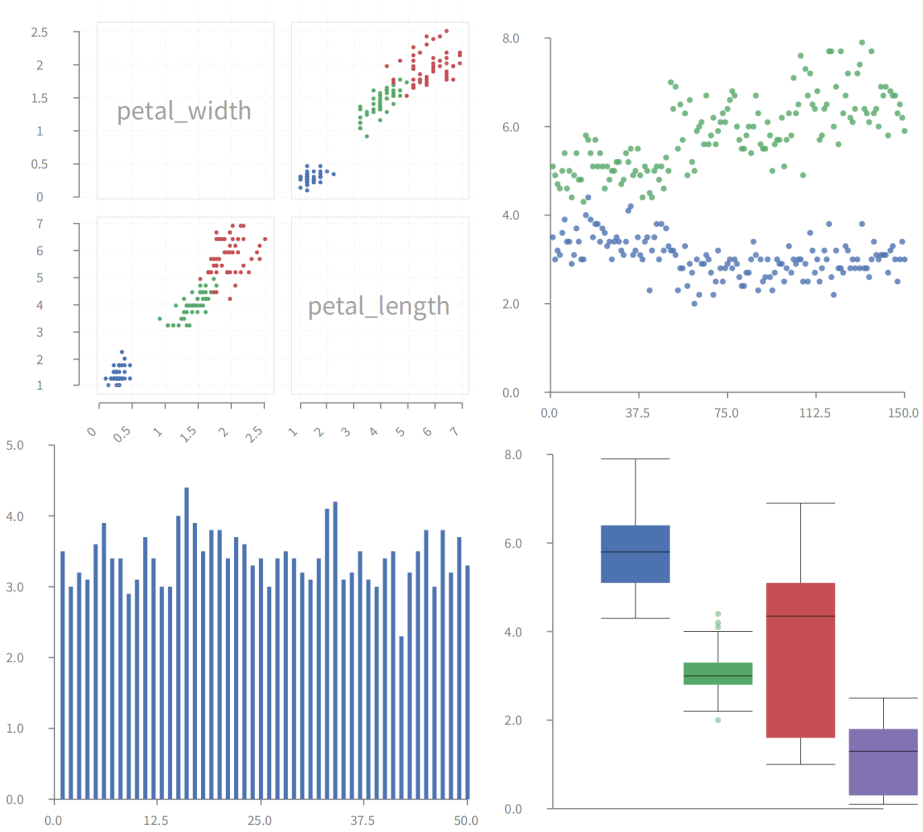


Figure 2: no need to put the file extension!

Acknowledgements

This work was supported by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council, FEDER through the 'Contrat de Projets Etat Region (CPER) 2007-2013', the Cutter ANR project, ANR-10-BLAN-0219 and the MEALS Marie Curie Actions program FP7-PEOPLE-2011- IRSES MEALS (no. 295261).