Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

DATA SCIENCE AND STATISTICS WITH PYTHON

KASPŘÍKOVÁ Nikola (CZ)

Abstract. With Data Science as the recent trend, the tools for efficient manipulation and analysis of data have become important for (training of) professionals in data analysis and statistics. The Python programming language starts to be one of the most popular tools for data analysis, in many cases replacing well-established statistical software packages. We report on a short case study on elementary statistical evaluation of a real-world data on air quality, using Python.

Keywords: Data science, Python, statistics, software

Mathematics subject classification: Primary 97R30; Secondary 68N15

1 Introduction

There is a paper by John M. Chambers [1], the author of the S programming language, the language which can be considered the predeccessor of the widely used R software for the statistical computig [2]. The paper [1] is advocating the so-called greater statistics, which means everything related to learning from data – starting with the planning of the research, including collection of the data and organization of the data, to reporting the results. The paper [1] was published as early as in 1993, and it was probably not the first and definitely not the last such paper. The Data Science has recently become popular as a new discipline related to data analysis. This discipline, actually developing the ideas in [1], aims at the ability to efficiently operate with the data technologies and software tools, ask the right questions about the problem to be analysed and present the results appropriately. The Python programming language [3] is considered as one of the primary tools for Data Science. Many profiles of jobs related to quantitative methods and data analysis include Python as one of the required skills. This fact can be easily checked for example at the MathJobs website [4], run by the American Mathematical Society, and similarly at the well-known ResearchGate website [5].

These facts should be reflected in the training of professionals in mathematics, data analysis and statistics. And indeed many top-class universities have already started including the Data Science courses in the curriculum. There are textbooks like [6] (UC Berkeley), [7] (MIT) and many others and there are initiatives focused at including computational thinking in the curriculum, see e. g. [8].

The goal of this paper is to assess the suitability of using Python for data analysis and statistics courses. We report on a short case study on elementary statistical evaluation of a real-world data on

air quality (the levels of PMO10) in Praha and in Moravskoslezský region in the Czech Republic, using Python. After having discussed the motivation in the introduction, the remaining parts of this paper briefly introduce the data science and the Python programming language for statistics, then the case study is reported, followed by some conclusions.

2 Data Science and Python

2.1 Data Science as the new trend

It is rather difficult to find an objective way of evaluating the popularity of some concept. We try using the Google Trends tool [9] to analyze the "Data Science" search query popularity between December 2012 and November 2017. The figure 1 is the visualization of the data obtained from [9] and it represents the relative interest in the search query at the Google search engine in comparison to the highest value in the time series (the value 100 represents the highest popularity). Obviously "Data Science" has been gaining popularity quite rapidly in the recent years.

The Data Science concept, which has recently started to be used much more often then earlier, may be still finding its precise meaning, but often it refers to (see [6]) parallel computing, dealing with non-standard data formats (such as access log files, e-mail messages), data technologies like XML, JSON, web scraping, NoSQL database technologies (including Map Reduce, Key - value database systems or graph database systems) and others, all aimed at dealing the data efficiently and ability to solve complex real world problems using data analytics.

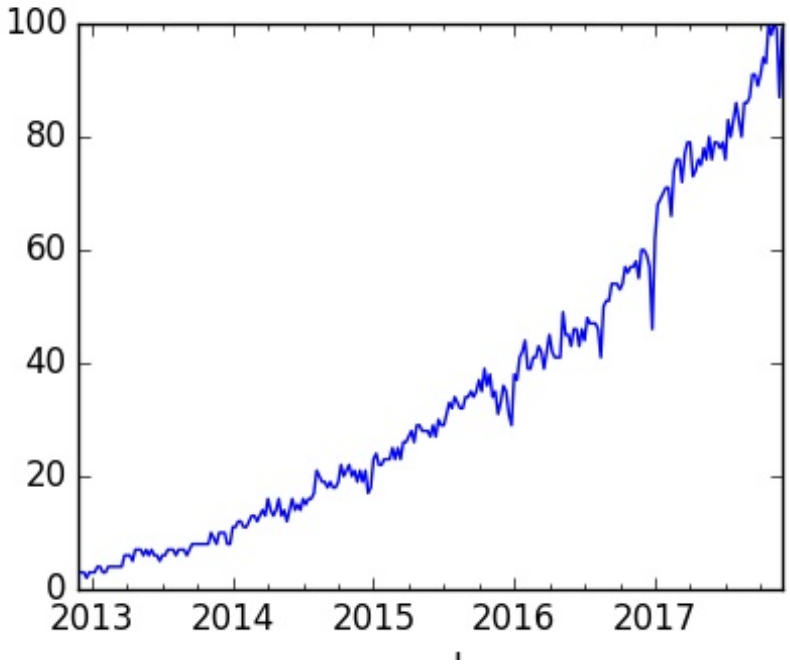

Fig. 1. The "Data Science" query popularity over time in Google Trends.

2.2 Python programming language and statistics

The Python [3] programming language is one of the most often used languages today and it has the following basic features:

- open source software
- interpreted language
- general purpose language
- there exist many extension packages designed for specific needs
- quite easy to learn

The Google Trends tool states "python for data science" as the top related query for the "data science" query. The Python extension packages often used for data analysis, plotting and statistical computing include SciPy [10], pandas [11] and seaborn [12]. There are many papers advocating using Python in data analysis and even in high performance computing, see e. g. [13].

3 Case study: the statistical analysis of air quality data

3.1 Material and methods

We analyze the yearly average levels of PMO10 (measured in $\mu g/m^3$) at the measurement stations in year 2016 in Praha and in the Moravskoslezský region in the Czech Republic. Let's assume that the locations of the stations (13 and 18 respectively) are representative. These two regions are known for troubles with the air quality, which can be generally attributed to heavy transportation, high industry-driven pollution and some geographical characteristics. The dataset has been obtained from the data published by the Czech Hydrometeorological Institute, available at [14].

We will do elementary exploratory statistical analysis - calculate basic descriptive statistical characteristics and obtain the probability density estimate plot. Then we evaluate the difference in the distributions of PM10 levels between the Praha and the Moravskoslezský region using the Wilcoxon rank-sum test.

3.2 Calculation in Python and results

First we read the dataset into the Python computing environment. We use the pandas package [11] for reading the csv file using the following statements.

```
>>> import pandas as pd
>>> dta2 = 'dataCHMU.csv'
>>> df2 = pd.read_csv(dta2)
```

Then, using the describe function, we calculate the basic descriptive statistics for each of the columns of the dataset, which stores the data of Praha and the Moravskoslezský (MS) region.

```
>>> df2.describe()
```

Which results in the following output:

```
            Praha          MS
count   13.000000   18.000000
```

```
mean    22.753846  31.694444
std      2.840368   4.332839
min     19.300000  24.700000
25%     20.400000  28.200000
50%     22.700000  32.900000
75%     25.200000  33.725000
max     26.900000  41.000000
```

The results of the calculations suggest that the levels of PMO10 are rather lower in Praha, with median at around 23 $\mu g/m^3$, than in Moravskoslezský region, with median at around 33 $\mu g/m^3$. The probability density estimate plot (shown in Fig. 2) supports this too. The probability density estimate plot can be conveniently produced using the functions available in the seaborn package. The statements for obtaining the plot are as follows:

```
>>> import seaborn as sns
>>> p1=sns.kdeplot(df2['Praha'], shade=True, color="r")
>>> p1=sns.kdeplot(df2['MS'], shade=True, color="b")
>>> sns.plt.show()
```

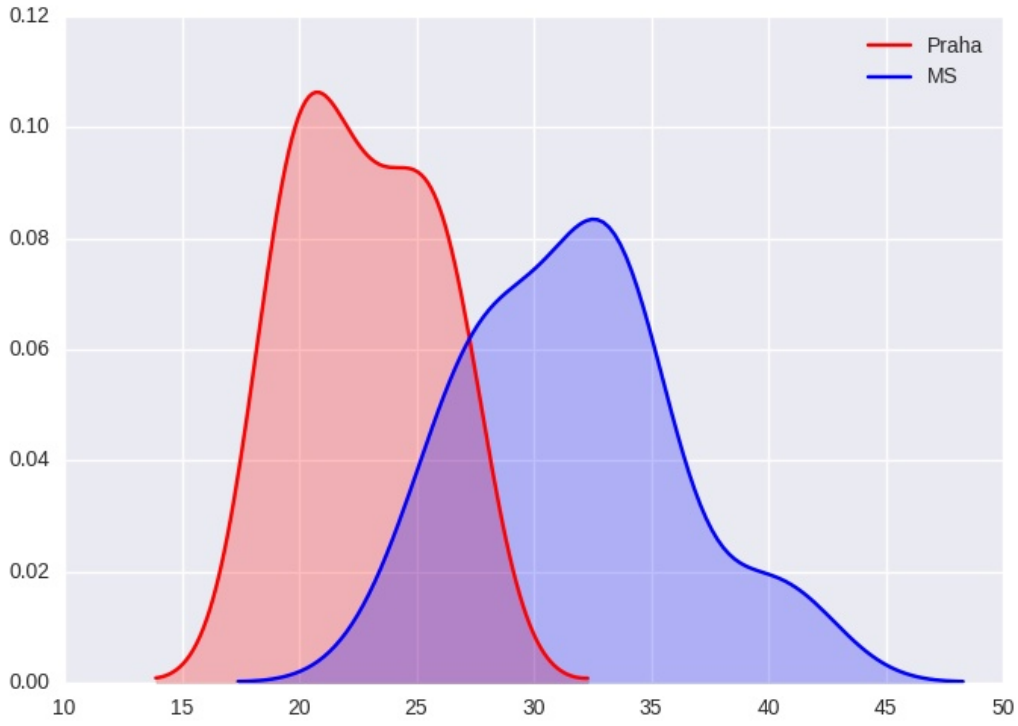

Fig. 2. Probability density estimate of PM10 levels [$\mu g/m^3$] by region in 2016.

Finally we run the Wilcoxon rank-sum test to detect formally if there is the difference in the distributions of PM10 levels between the Praha and the Moravskoslezský region. After importing the stats module from the scipy package, we retrieve directly the p-value of the test.

```
>>> from scipy import stats
>>> round(stats.ranksums(df2['Praha'],df2['MS'])[1],3)
```

Which gives 0.035. Based on this output, it may be concluded at 0.05 confidence level, that the PMO10 levels distribution is different in Praha and in Moravskoslezský region.

4 Conclusions

While perhaps still not offering the full comfort of some professional statistical packages, the Python programming language, with many data analysis related features under development, does already provide convenient tools for data analysis and common statistical calculations and may be easily used in courses on statistics. It has been shown that the syntax needed for elementary statistical analysis does not seem too complex and could easily be mastered, even by non-technical students.

References

[1] CHAMBERS, J. M.: Greater and Lesser Statistics: A Choice for Future Research, *Statistics and Computing*, 3(4), 1993, pp. 182-184.

[2] R CORE TEAM: *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL http://www.R-project.org.

[3] URL: https://www.python.org. Accessed 1/12/2017.

[4] URL: http://www.mathjobs.org. Accessed 1/12/2017.

[5] URL: http://www.researchgate.net. Accessed 1/12/2017.

[6] NOLAN D., LANG, D. T.: *Data science in R. A case studies Approach to computational reasoning and problem solving*, CRC Press, 2015, ISBN 978-1-4822-3481-7.

[7] GUTTAG, J. V.: *Introduction to computation and programming using Python*, The MIT Press, 2013, ISBN 978-0-262-52500-8.

[8] URL: http://www.computerbasedmath.org. Accessed 1/12/2017. x

[9] URL: https://trends.google.com/trends. Accessed 1/12/2017.

[10] URL: https://scipy.org. Accessed 1/12/2017.

[11] URL: http://pandas.pydata.org. Accessed 1/12/2017.

[12] URL: https://seaborn.pydata.org/. Accessed 1/12/2017.

[13] BILINA, R., LAWFORD, S.: Python for unified research in econometrics and statistics. *Econometric Reviews*, 31(5):558–591, ISSN: 0747-4938.

[14] URL: http://portal.chmi.cz/files/portal/docs/uoco/web_generator/tab_reports/automated/tab_2016_01_1Y_CZ.html. Retrieved 1/12/2017.

Current address

Kaspříková Nikola, Ph. D.
University of Economics in Prague, Department of Mathematics
Nám. W. Churchilla 4, 130 67 Praha, Czech Republic
E-mail: nb33@tulipany.cz