DATA ANALYSIS OF IT JOB POSTINGS IN CZECH LABOUR MARKET

Nikola Kaspříková

Abstract

A basic exploratory data analysis of vacancies published at the Czech Labour Offices website is addressed. We analyse the characteristics of the vacancies like the salary, required educational level and others. We address the jobs located in the capital and focus in more detail on jobs in the Information Technology industry. The main contribution of the paper is in the data extraction and the data preparation part and the code which can be used to retrieve the data is shown in the paper, using the XPath formal language expressions to address the relevant parts of the document in the eXtensible Markup Language format. In the outcomes of the elementary statistical analysis, it has been observed, that the vacancies in the Information Technology industry show higher requirements on the level of education and that the median value of the minimum salary offered is at around 40000 CZK, which is considerably higher value compared with the vacancies in all the industries.

Key words: job postings, XPath, exploratory analysis, labour market

JEL Code: J01, C81

Introduction

The labour market is an important part of the economy and it is of major personal interest for many people as well. The labour market is also a popular topic in the research papers. The topic of labour market and human resource management is quite broad. The Czech labour market is addressed recently in papers (Koutna and Janicko, 2018) and (Šafránková and Šikýř, 2017). The paper by Pichault, Lorquet and Orianne (2018) discusses the role of the market intermediaries. Many reports are based on the applications of the sample survey methods in the research, some authors propose the application of advanced quantitative analysis tools, see (Cichoń & Piotrowska, 2018) and (Brandas, Panzaru and Filip, 2016).

In the modern digital economy, there is a vast amount of (sometimes very interesting) information available to support the analysis. The data analysts and statisticians have mostly already mastered the ways how to retrieve the data from the traditional database systems,

usually using the SQL query language. But there have emerged other quite interesting sources of information recently and the tools to be used for these new data sources are rather different form the traditional database query languages. The book by Temple Lang and Nolan (2015) presents the analysis of online job postings from the job portals like monster.com, using the tools for web scraping, i. e. retrieving the contents of interest from the web pages (usually in HTML), designed to be read by humans, not machines. There are many software tools for web scraping, see also the paper (Khalil and Fakir, 2017). In some cases, the data is available in some of the structured text file formats suitable for automated data processing, such as XML or JSON.

This paper reports on the analysis of vacancies published at Czech Labour Offices website. We address the jobs located in the capital and focus on IT related jobs and analyse the characteristics of the vacancies like the salary and the level of education required. The main contribution of the paper is in the data extraction and the data preparation part and the code is shown in the paper, using the XPath formal language expressions.

1 Material and methods

We use the data available from the web portal of the Czech Labour Office. The Czech Labour Office (see Úřad práce České republiky (2018)) is a government institution, which provides the services described in (Czech Labour Office Services, 2018). Among other services, the Office publishes the job offers on the web (Vacancy search, 2018), where the data on the vacancies can be retrieved using the web form. The data on vacancies generally provide an interesting information about the situation of the labour market, it may provide a more accurate description of the reality in comparison with surveys and it reflects the current needs of the employers with respect to the profession and the required skills of prospective employees and it also reflects the expectations with respect to the salary levels. In comparison with other similar data sources, the Czech Labour Office vacancies dataset has the advantage, that it is rather well structured and maintained and quite large, since posting the job offer there is free for the employers.

The dataset on all the vacancies registered by the Czech Labour Office can be obtained from the Czech Labour Office webpage https://portal.mpsv.cz/sz/download. It is possible to get the data on all the vacancies registered by the Czech Labour Office in either the HTML format or the XML format using this webpage. Similarly, one can obtain a dataset of

vacancies registered in a particular region. Files with data from history are available too. For human users of the Labour Office webpages, there is also a page with a form which can be used to conveniently search job offers. Nevertheless, the data obtained by the interactive form is in the form to be read by humans rather than for some automated processing and analysis. In this paper, we address the dataset on vacancies in Prague as of April 20th2018. This dataset contains 14107 jobs.

1.1 Xpath queries

If the dataset in XML format is to be used, then one could use the query in the XPath (the XML Path) language. For example if we are to retrieve the values of the education level required (which is the attribute called "nazev" in the element called "MIN_VZDELANI" in the file), then we could use the query ".//MIN_VZDELANI/@nazev", which could then be used for example in the call of functions in XML or xml2 packages in the R software for data analysis.

If we are to get the information just for the jobs in the Information technology industry (which has attribute "kod" set to value 4 in the "OBOR" element), we could use the following query ".//OBOR[@kod='4']/preceding-sibling::MIN_VZDELANI/@nazev".

From the examples of the Xpath expressions shown above, one can observe, that it is not so difficult to retrieve the data for the analysis, even though the queries may seem rather less intuitive in comparison for example with the standard SQL queries.

2 Job offers characteristics

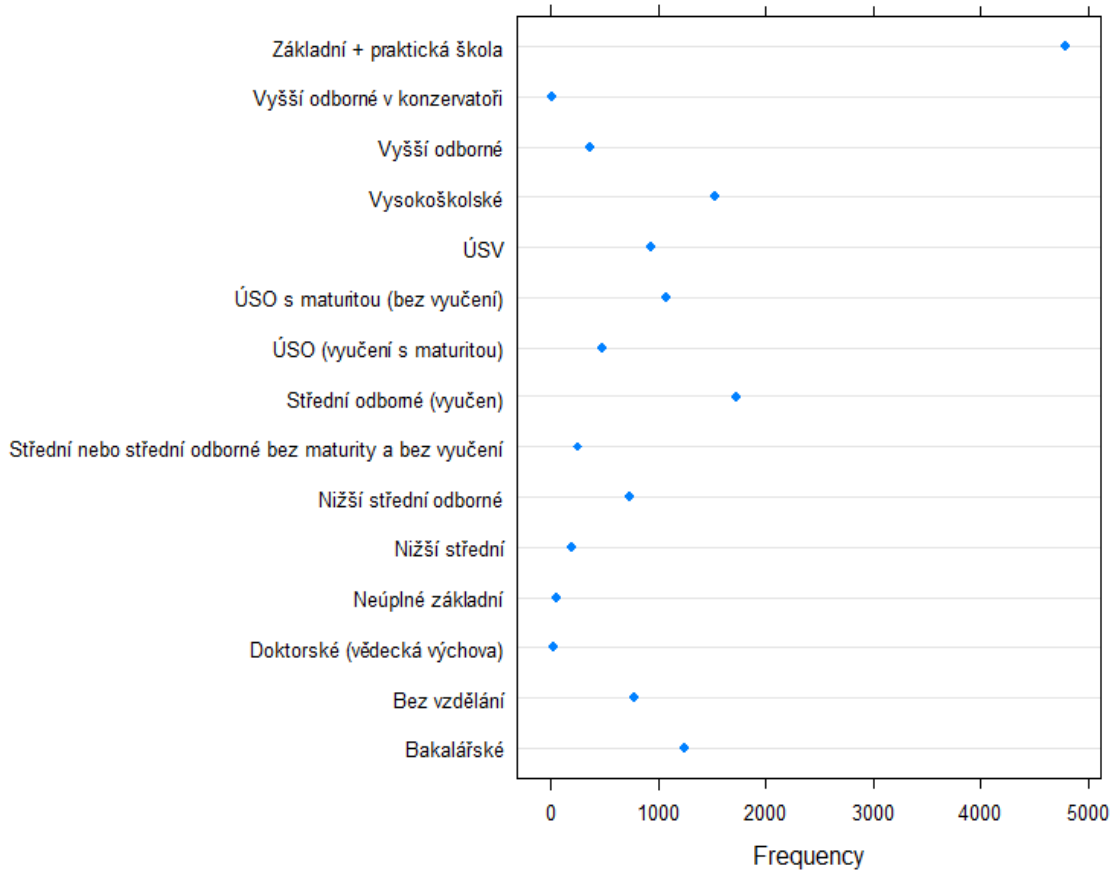
2. 1 Industry of the job

The analysis of the frequencies of the industry of the jobs has shown, that the highest frequency of the jobs offered is in the Trade and tourism (over 3500) and Production and operation (2741). There were 1600 vacancies in the IT industry.

2.2 Education level required

The most often required minimum level of education in all the vacancies was "Základní+praktická škola", which in the Czech system of education is the basic level of education. For more details see Figure 1 (values are relevant for the Czech system of education and are shown in Czech).

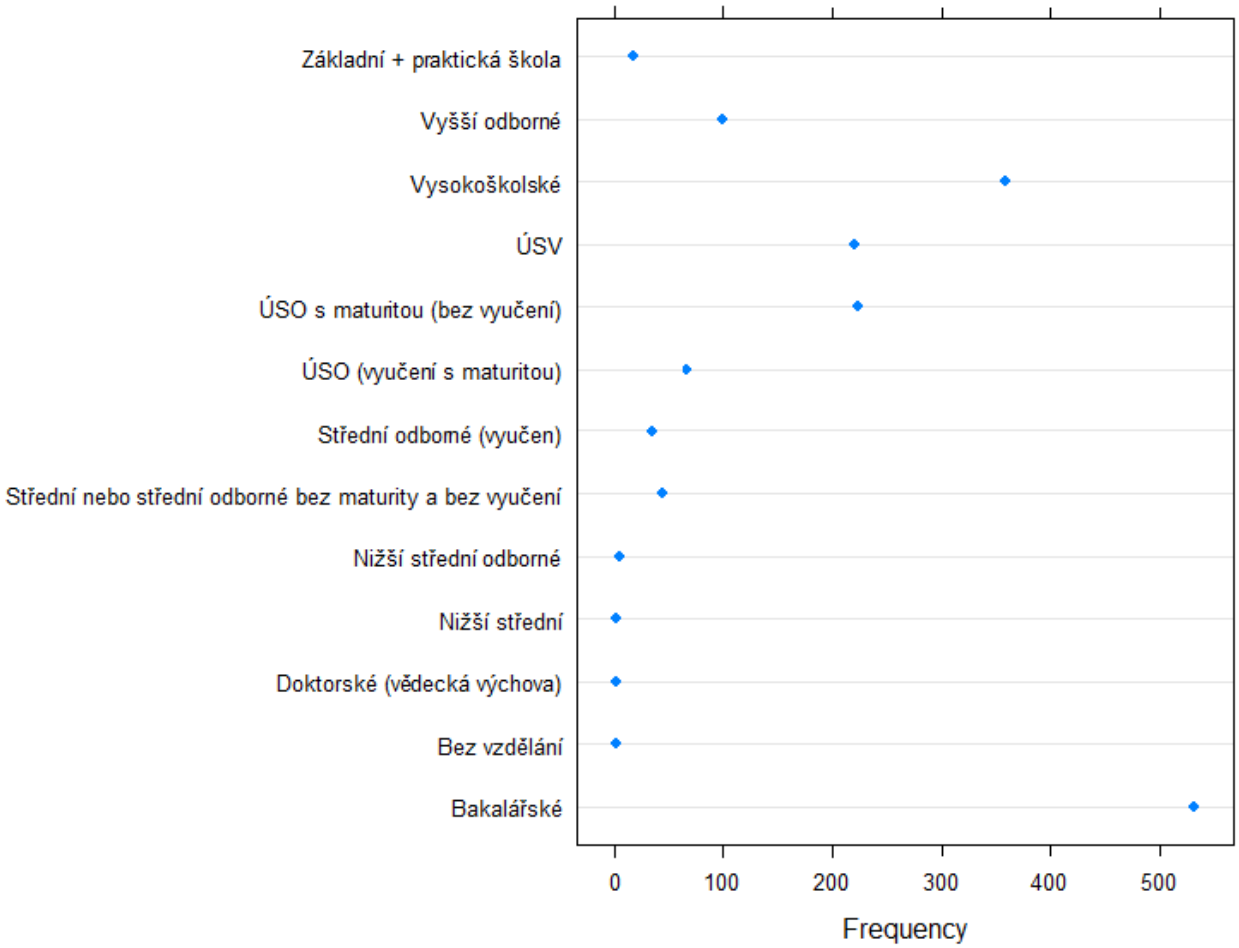
Fig. 1: Frequencies of all vacancies with respect to the minimum education level required



Source: data from mpsv.cz

When just the IT jobs are analysed, the situation is quite different, see Figure 2, which shows that the two most frequent values are "Vysokoškolské" and "Bakalářské", which are the two levels of the university education in the Czech Republic, just before the Ph.D. level.

Fig. 2: Frequencies of IT vacancies with respect to the minimum education level required



Source: data from mpsv.cz

Such outcome of the analysis could have been expected, since the qualification required for the IT jobs is usually rather high.

2.3 Salary

The job postings registered by the Czech Labour Office include minimum and maximum value of the salary for a particular job offer. For the comparison of the values of the basic descriptive statistics for IT jobs and all jobs, see Table 1 and for the plots of the probability density estimates of the salary limits in IT see Figure 3 and Figure 4.

It could be observed, that IT jobs show quite high values of the salary. The median value of the minimum salary offered, is for IT jobs 40000 and 16000 for all the jobs. This may be related to the fact that the IT jobs mostly require higher qualification.

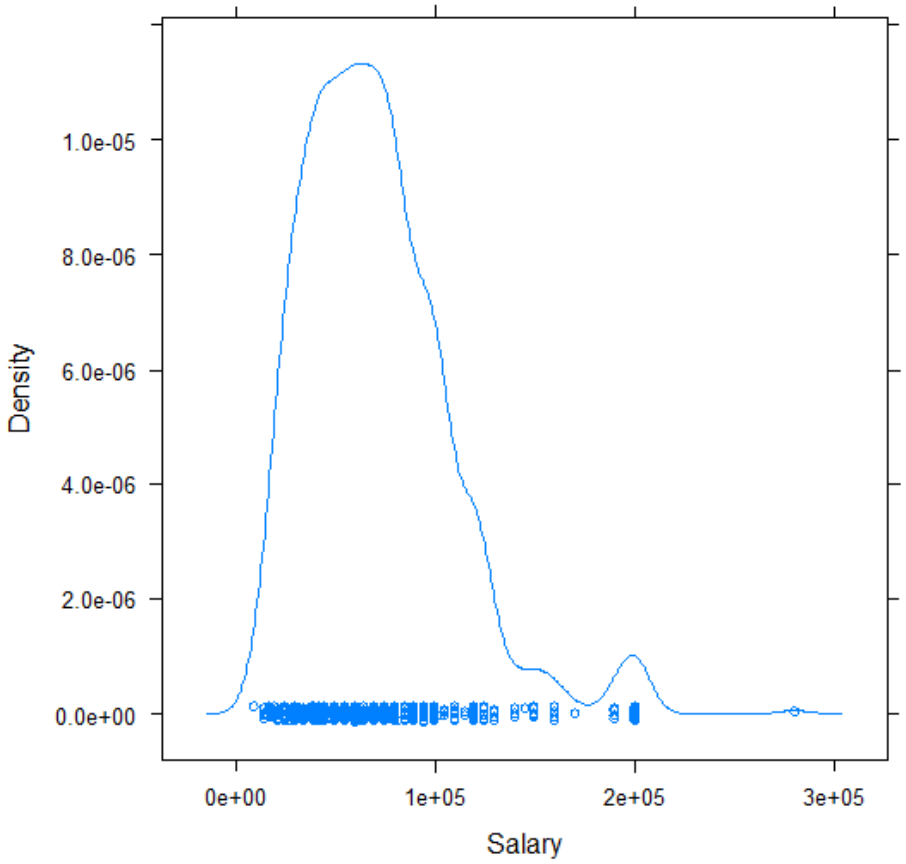
Tab. 1: Basic descriptive summary statistics of minimum salary (in CZK)

	All jobs	IT jobs
1st Quartile	13000	25000
Median	16000	40000
Mean	21749	43099
3rd Quartile	23000	50000

Source: data from mpsv.cz

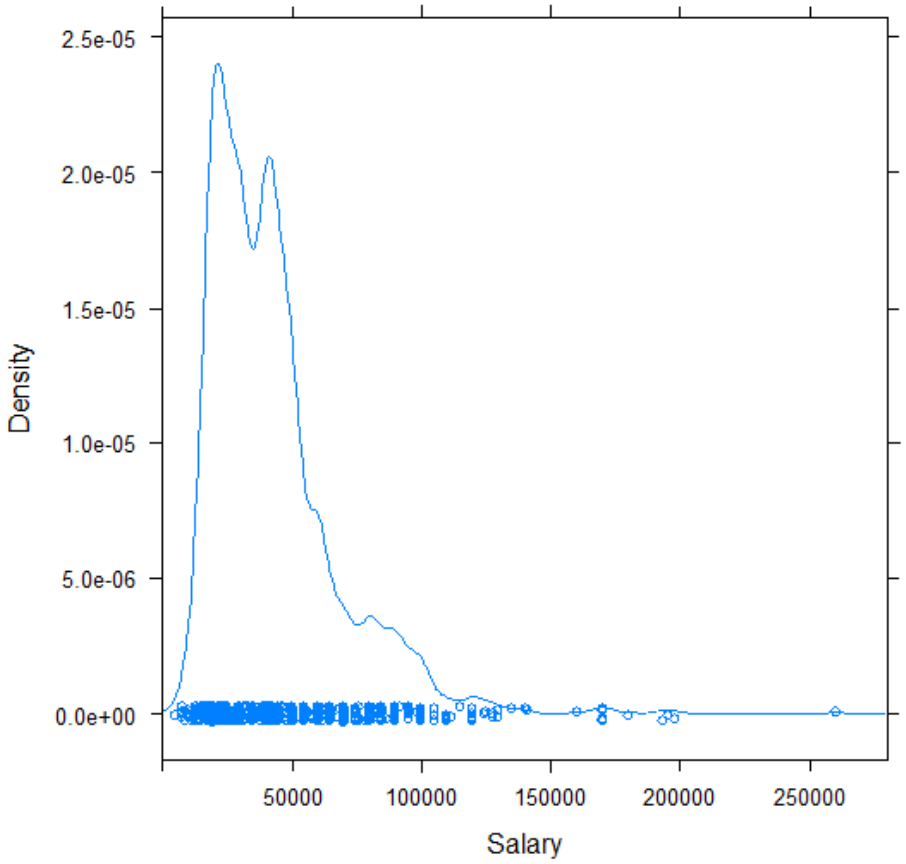
The higher limit of the salary for IT jobs is often between 50000 and 100000 CZK, the lower limit is usually at around 50000 CZK or lower.

Fig. 3: Probability density estimate of the higher limits of the salary in IT jobs



Source: data from mpsv.cz

Fig. 4: Probability density estimate of the lower limits of the salary in IT jobs



Source: data from mpsv.cz

Conclusion

The data on the job offers available from the website of the Czech Labour Office provide valuable information on the situation on the Czech labour market. The data could be conveniently retrieved using the tools for getting the data in modern formats commonly used in the web technologies, namely the XPath language for locating the parts of the data in the XML format.

The elementary analysis of the data on the vacancies located in the capital city has shown that IT jobs much more often require at least bachelor level degree. Regarding the

lower limit for the salary offered, the median value for IT jobs is at around 40000 and 16000 for all the jobs.

As the next step, a multivariate statistical analysis could be performed, which could address other attributes available as well and could extend the analysis to other regions.

Acknowledgment

This paper has been produced with contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

Brandas, C., Panzaru, C., & Filip, F. (2016). Data Driven Decision Support Systems: An Application Case in Labour Market Analysis. *Romanian Journal of Information Science and Technology, 19*(1-2), 65-77.

Cichoń, M., & Piotrowska, I. (2018). Level of academic and didactic competencies among students as a measure to evaluate geographical education and preparation of students for the demands of the modern labour market. *Quaestiones Geographicae, 37*(1), 73-86. doi:10.2478/quageo-2018-0006

Khalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX, 6*, 98-106. doi:10.1016/j.softx.2017.04.004

Koutna, M., & Janicko, M. (2018). Trajectories in the Czech Labour Market: The Role of Information-processing Skills and Education. *Ekonomický časopis, 66*(1), 3-27.

Pichault, F., Lorquet, N., & Orianne, J. (2018). Towards the End of Career Management? HRM and the Growing Role of Labour Market Intermediaries. *Relations industrielles – Industrial Relations, 73*(1), 11-38.

Šafránková, J. M., & Šikýř, M. (2017). Work expectations and potential employability of millennials and post-millennials on the Czech labor market. *Oeconomia Copernicana, 8*(4), 585-599. doi:10.24136/oc.v8i4.36

Temple Lang, D. T., & Nolan, D. A. (2015). *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving.* CRC Press.

Úřad práce České republiky. (2018). In mpsv.cz. Retrieved April 22, 2018, from https://portal.mpsv.cz/upcr

Vacancy search. (2018). In mpsv.cz. Retrieved April 22, 2018, from https://portal.mpsv.cz/sz/obcane/vmjedno

Czech Labour Office Services. (2018). In mpsv.cz. Retrieved April 22, 2018, from https://portal.mpsv.cz/sz/obecne/cinnosti_up

Contact

Nikola Kaspříková

University of Economics in Prague, Department of Mathematics

Nám. W. Churchilla 4, 130 67 Praha

nb33@tulipany.cz