# Using R for Data Analysis of Master Graduates Survey

Marin FOTACHE, Al. I. Cuza University, Iaşi, Romania, fotache@uaic.ro

## Abstract

Graduates survey is a vital tool for testing how well an undergraduate or graduate academic programme performs in relation to the market needs and expectations. Graduates possess the real-world working experience and also do not feel any pressure in assessing their teachers, programmes and schools. In this paper we present our experience in analyzing data gathered from graduates of Business Information Systems master programme. We argue for R as an excellent open-source platform for doing data analysis, presenting its dynamic on the statistical packages market. Some of the most powerful R features for data visualization and exploratory data analysis are presented, as well as some tools for plotting and summarizing Likert scale data.

**Keywords**: Data analysis, R, graduates survey, data visualization

## Introduction

As the market demand changes quickly, adapting curricula is a must for a successful undergraduate and graduate programme. A key point in assessing undergraduate and master programmes performance is graduates views, opinions and proposals. After a few years as professionals, they usually know what current jobs demand, what courses proved to be more useful, what lacks in their formation and how universities can better prepare the students for their careers.

National agencies for accreditation/re-assessment of higher education programmes require for the applicants to survey the graduates and analyze the state of graduates employability and their opinions about the programme. Consequently, graduates survey is a must for all types of academic programmes.

There are lots of technical solutions for publishing, disseminating and filling the survey, and also for survey analysis. In this paper we argue for an increasingly popular data analysis platform, R, focusing on some packages and functions dedicated to exploratory data analysis, graphics and inferential statistics.

## R and data analysis

Data analysis has become extremely popular today. Per se, or related/combined with other buzzword such as big data (Fotache and Strîmbei, 2013), analytics, data science (Kumar, 2012; Dhar, 2013), data analysis experiences an increasing interest not only from academics and researchers, but also from professionals (Musson and Smith, 2013; Amatriain, 2013).

Analyzing the popularity of data analysis software, B. Muenchen (2014) takes into account the following metrics:

- Job advertisements
- Published scholarly articles
- Books
- Blogs
- Web site popularity (reported with tools such as Google Analytics)
- Surveys of use
- Programming activity (using data gathered from public repositories such as GitHub)

- Discussion forums
- Popularity measures (e.g. overall composite score or rank)
- IT research firms (e.g. Gartner, IDC)
- Sales or download measures
- Growth in capability.

Indeed.com is the biggest job site in the U.S. making its sample the most representative of the current job market. It aggregates all the jobs from over 1000 sources - major job boards (Monster, Careerbuilder, Hotjobs, Craigslist), and also newspapers, associations, and company websites.

Figure 1 shows the composite demand for jobs related to data analysis. Among "classical" statistical packages, the best ranked is SAS, followed by R, SPSS, Mathlab, and Stata. There are some explanations for high demand of Java, Python and C/C# professionals. Real-world data analysis must be integrated into business applications, and Java, Python and C# are the most important languages for software development.

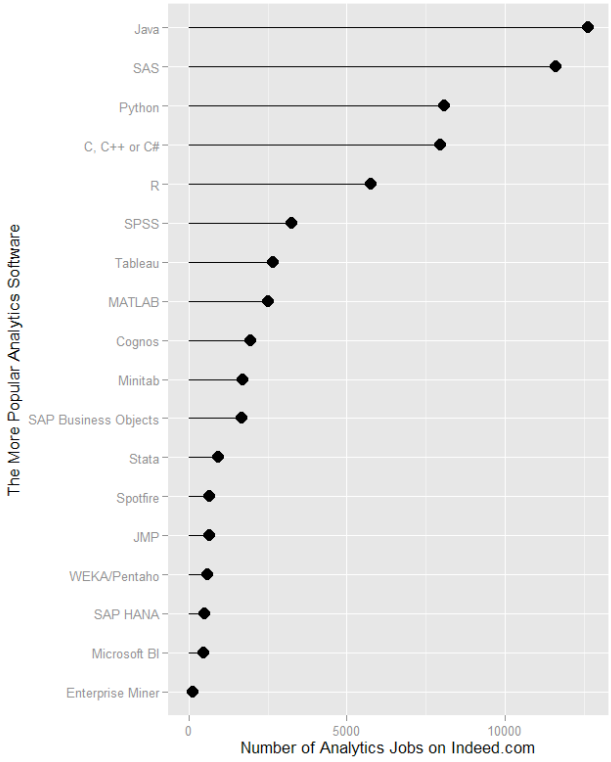


Figure 1. US Job Market for Data Analysis software at beginning of 2014 (Muenchen, 2014)

B. Muenchen (2014) argues that R jobs started exceeding SPSS in mid 2012, estimating that R will catch up with SAS within 1.87-3.35 years. In terms of number of scholarly articles mentioning statistical package, SPSS and SAS still have the lead, despite their steep decline after 2007 (Muenchen, 2014). In recent years, more and researchers urge for using R as main scientific platform for data analysis (Kelley, Lai and Wu, 2008; Dunn, 2011).

We expect R will continue to expand in business world and also in research communities, not only because it is free, but also since there is a huge community of enthusiasts who are eager to contribute with new packages and functions targeting real-world data analysts, data scientists and researchers. Also as new IDE (e.g. R Studio) and GUI (e.g. R commander) tools make it less scary for non-programmers, R will attract new segments of users.

## Data import and cleaning

There are many free survey tools available online with various options for data display, export and processing. As Microsoft SharePoint Portal has been in use for a couple of years (almost ten) at the Faculty of Economics and Business Administration of Al.I.Cuza University in Iasi, we preferred to provide our graduates a similar experience with what they were accustomed to during their years of study. If institutions do not have a SharePoint license, any other survey tool (mostly free of charge) can serve as well, as soon it has a data export option into a general format (.csv, .xls, .xlsx, tab-delimited text, etc.)

R is able to import data from a huge variety of sources/formats. As our survey was implemented in SharePoint, data is available as Excel (.xlsx) file. There are a couple of R packages dedicated to data import from Excel files. We preferred *xlsx* (Dragulescu, 2013).

```
library(xlsx)
df = read.xlsx("SIA2014-01-17_13_00.xlsx", 1, fileEncoding = "UTF-8",
       header=TRUE, stringsAsFactors=FALSE)
```

R is handful at data cleaning, having simple control structures and generous functions for data processing. The basic operations we pursued after import were:

- eliminating the answers with too many empty fields (NA and NULL)
- splitting the main data frame into analysis sections (studies, employment, course assessment, program assessment, etc.)
- "combing"/munging data for further analysis and visualization.

Some of the most useful packages in data cleaning and transformations are *stringr* (Wickham, 2010; Sanchez, 2013), *reshape* (Wickham, 2007), *plyr* (Wickham, 2011), and *dplyr* (Wickham and Francois, 2014).

As there is a single source of data, in our case there is no need of additional operations such as data integration, data curation, etc.

## Basic exploratory data analysis and visualization

Basically exploratory data analysis includes data processing (see previous section), descriptive statistics, identification of outliers, exploration of data distributions, data visualization, identifying associations among variables, etc.

In this section we briefly focus on some elegant techniques for data visualization in R. Starting with histograms and density plots, in *studii.en* data frame there is variable called *age* which points to the current graduates age. The histogram can be plotted with *hist* function:

```
hist(studii.en$age , main = "Current age distribution for master graduates",
     col="lightskyblue", xlab="Age (years)", freq = FALSE)
lines(density(studii.en$age , na.rm=TRUE), col="red", lty=2, lwd=2)
```

Adding (superimposing) a density curve to the histogram requires *lines* function:

```
lines(density(studii.en$age , na.rm=TRUE), col="red", lty=2, lwd=2)
```

The result is on the left side of figure 2. In the reminder of this paper we'll use mainly *ggplot* function from *ggplot2* package (Wickham, 2009), which is extremely versatile and elegant (though intimidating at first sight).

The same task of plotting histogram and density curve can be achieved with the following code:

```
ggplot(studii.en, aes(x=age)) +
  geom_histogram(aes(y=..density..), binwidth=1, colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") + # Overlay with transparent density plot
  ggtitle("Current age distribution for master graduates") +
  theme(text=element_text(size=16)) + xlab("Age (years)")
```

The right side of figure 2 presents the graphic which can be compared with the result of functions *hist* and *lines*.


Figure 2. Histogram and density plot in R with functions *hist/lines* (left) and *ggplot* (right)

An important part of exploratory data analysis consists in comparing data distributions for two or more sub-populations. This could prove the starting point for advancing hypotheses and using inferential statistics such as t-test, ANOVA, regression etc. and their non-parametric counterparts. Graphics can be split, like in the left side of figure 3, or superimposed, like in the right side of the same figure.

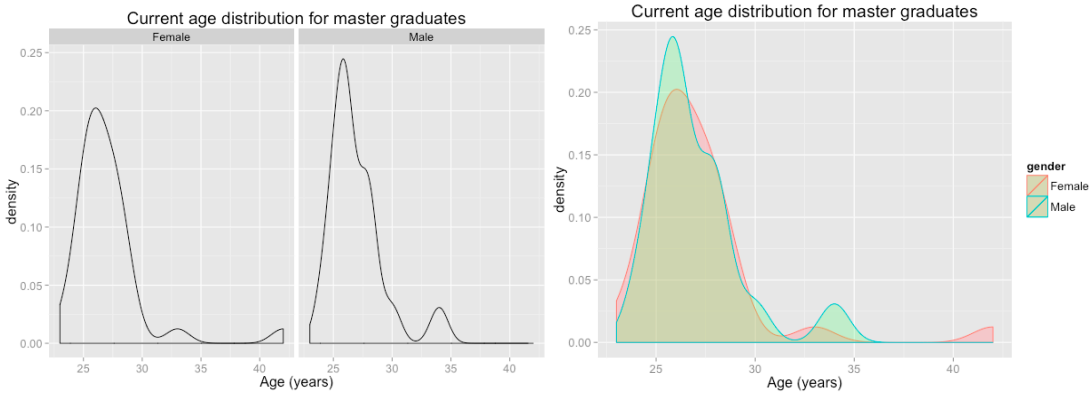


Figure 3. Two ways to graphically compare two density plots (age for males vs. females)

Taking the same distribution of graduates age (some statisticians would argue that using density plot for a discrete variable is somehow dubious), opposing age density curves for females and males is done with *facet_wrap* option:

```
ggplot(studii.en, aes(x = age)) + geom_density() + facet_wrap(~gender) +
```

```
       ggtitle("Current age distribution for master graduates") +
       theme(text=element_text(size=16)) + xlab("Age (years)")
```

Superimposing the same density lines can be achieved as follows:

```
ggplot(studii.en,aes(x=age, color=gender)) +
 geom_density(data=subset(studii.en, gender=="Female"),fill="red", alpha=0.2) +
 geom_density(data=subset(studii.en, gender=="Male"), fill="green", alpha=0.2) +
 ggtitle("Current age distribution for master graduates") +
 theme(text=element_text(size=16)) + xlab("Age (years)")
```

In comparing two or more (sub)populations in terms of distribution shape, mean/median and variance, many times boxplots are preferred to density curves. Again we'll compare the result of a special function – *boxplot* (left side of figure 4):

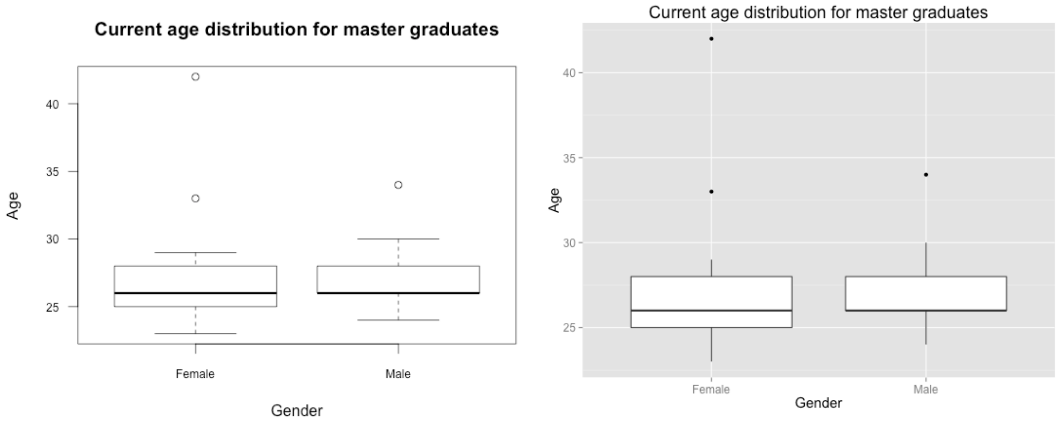


Figure 4. Boxplots with functions boxplot (left) and ggplot (right)

```
boxplot(age ~ gender, data=studii.en,
    main="Current age distribution for master graduates",
    xlab="Gender", ylab="Age", las=1, cex.axis=.75, cex.names = 0.75)
```

with the boxplot obtained with *ggplot* (right side of figure 4)

```
ggplot(data = studii.en, aes(x = gender, y = age)) + geom_boxplot() +
       ggtitle("Current age distribution for master graduates") +
       theme( text=element_text(size=15)) + ylab("Age") + xlab("Gender")
```

Of course, for boxplots the difference between the results of *boxplot* and *ggplot* is not spectacular. The last two types of graphs we deal with in this section are variants of pie chart and bar plot. Pie chart is not a favorite among heavyweight statisticians, but remains still popular for many users. The following *ggplot* syntax is, up to a point, identical for bar plot and circular plot. Option *coord_polar* transforms the bar chart into a circular one (left side of figure 5):

```
ggplot(angajare.en, aes(x = Hired.from, fill = Hired.from)) +
       geom_bar(width = 1) + ggtitle("I got the job in...") +
       guides(fill=FALSE) + theme(text=element_text(size=15,face="bold")) +
       coord_polar(theta = "x")
```

To make the bar chart more attractive, we dive into some options of *ggplot*, so that the text is included inside the bars and the vertical scale is changed so that longer text is displayed properly. Also, in both graphs legend was suppressed by setting *guides(fill=FALSE)*.

```
dat.2 = data.frame(table(angajare.en$Hired.from ))
ggplot(dat.2, aes(x = Var1, fill = Var1)) +
  geom_bar(stat="identity", ymin=0, aes(y=Freq, ymax=Freq), position="dodge") +
  geom_text(aes(x=Var1, y=Freq, ymin=-0.5, ymax=Freq, label=Var1,
       hjust=ifelse(sign(Freq)>0, 1, 0)), position = position_dodge(width=1)) +
  scale_y_continuous(labels = waiver()) + coord_flip() +
  ggtitle("Moment of employment") +
  theme(axis.text.y = element_blank(), text=element_text(size=15)) +
  xlab("hired from...")  + ylab(" ") + scale_fill_discrete(guide=FALSE)
```

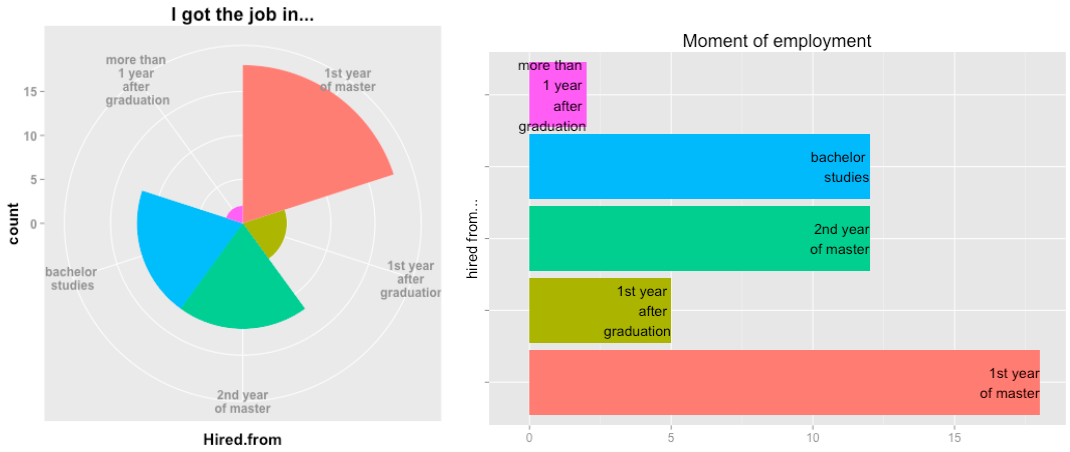


Figure 5. Circular plot and barplot using ggplot

Of course there is huge set of options for drawing necessary graphs in data presentation and analysis. We recommend diving into the references like Wickham (2009), but also on many tutorials and blogs available on the web.

## Likert scale data analysis

On most surveys, preferences (or dislikes) are expressed using Likert scale, with five or more levels. To evaluate the master programme, graduates were asked to assess, using a five-level scale (very bad, bad, average/neutral, good, excellent) the following items:

- the programme (on a general level)
- courses: utility for their professional activity, teaching, link to practice, research content
- professors: teaching, availability, attitude to students
- infrastructure (labs, classrooms, public spaces)
- dean's office attitude towards students/graduates
- internship.

There are many ways of displaying the results from studies using rating scales. Robbins and Heiberger (2011) point out the following categories: tables, bar charts of means, grouped bar charts, divided bar charts, ribbon charts, multiple pie charts, waffle plots, radar plots, and diverging stacked bar charts. Their preference is for diverging stacked bar charts. Fortunately, in R drawing such a kind of plot is easy through package *likert* (Bryer and Speerschneider, 2013).

As in our case the answers were imported from Excel as integers, first they must be converted into factors. After defining the levels, ranking variables are converted and then recoded so that in the subsequent charts levels instead of numbers will label the data.

```
levels = c("very bad", "bad", "average", "good", "excellent")
evGen.en = evaluari [, 92:94]
```

```
for (i in 1:ncol(evGen.en))
{
       evGen.en[,i] = factor(evGen.en[,i], levels=1:5)
       evGen.en[,i] = recode(evGen.en[,i], from=1:5, to=levels)
       evGen.en[,i] = factor(evGen.en[,i], levels=levels, ordered=TRUE)
}
evGen.en <- rename(evGen.en, c( evMaster = "Programme",
       evProfi = "Professors", evDiscipline = "Courses"))
```

The key function is *likert* which builds the ranking model of the data frame.

```
l.evGen.en = likert(evGen.en)
```

Rows of the object *l.evGen.en* refer to ranked variables and each column is dedicated to one ranking level. The value within each cell corresponds to the percentage of responses for that level:

```
> l.evGen.en
       Item very bad      bad   average     good excellent
1   Courses        0 6.122449 34.693878 38.77551  20.40816
2 Professor        0 2.040816  4.081633 40.81633  53.06122
3 Programme        0 2.040816 28.571429 46.93878  22.44898
```

Additional information is provided by *summary* function. Column *low* corresponds to the sum of levels below neutral, column *high* shows to the sum of levels above neutral, and columns *mean* and *sd* show the mean and standard deviation of the results:

```
> summary(l.evGen.en)
        Item      low   neutral     high     mean        sd
2 Professors 2.040816  4.081633 93.87755 4.448980 0.6788846
3  Programme 2.040816 28.571429 69.38776 3.897959 0.7704138
1    Courses 6.122449 34.693878 59.18367 3.734694 0.8606081
```

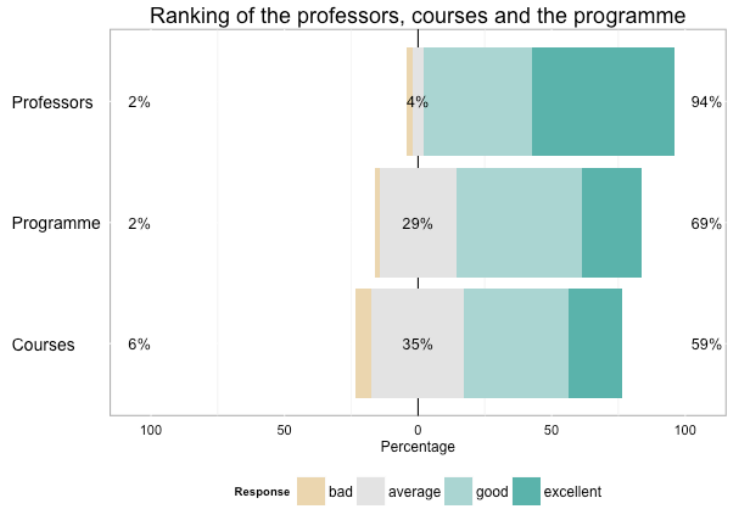


Figure 6. Diverging stacked bar chart

Graphic options are perhaps the main attraction of *likert* package. The diverging stacked bar chart for the overall ranking of the master programme (programme, courses and academics) – figure 6 - is very easy to plot (we removed some detailed options controlling the size of the font on axes and titles):

```
plot(l.evGen.en, text.size=4.5) +
       ggtitle("Ranking of the professors, courses and the programme")
```

There are also some options for visualizing ranked data as density curves (left side of figure 7) or heat maps (right side of the figure):

```
plot(l.evGen.en, type='density')
plot(l.evGen.en, type='heat', wrap=30, text.size=4)
```

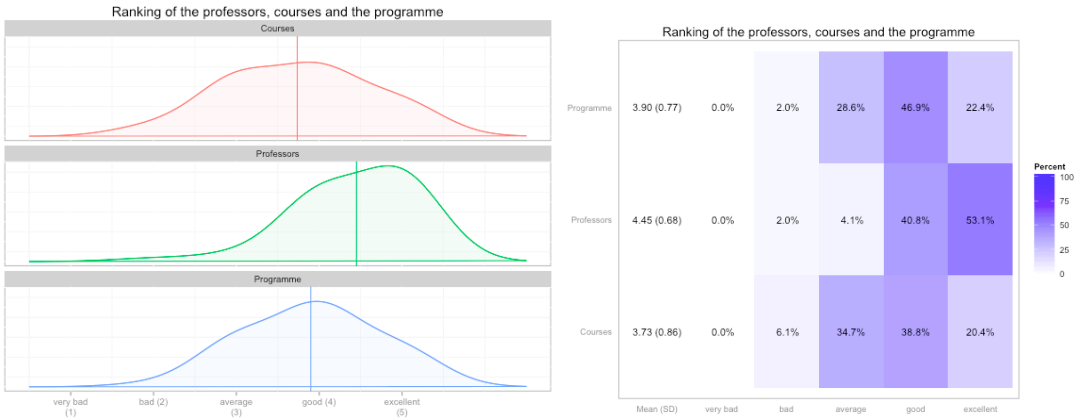


Figure 7. Visualizing ranked data as density plots (left) and heat map (right)

In exploratory data analysis it would be interesting to compare the distribution of data for various subpopulations. For example, to compare the ranking distributions of females versus males, in *likert* function a *grouping* option is needed (see figure 8):

```
l.evGen.en.g1 <- likert(evGen.en, grouping=evaluari$gender)
plot(l.evGen.en.g1)
```

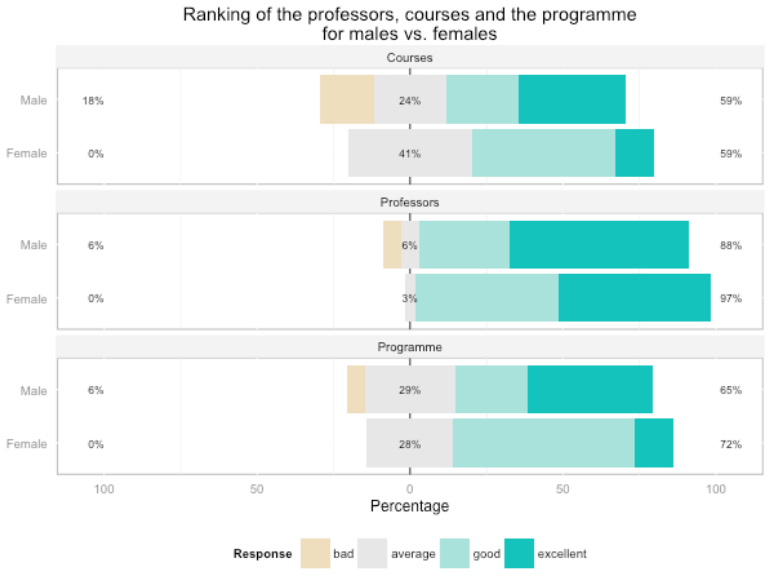


Figure 8. Diverging stacked bar chart with groups

## Inferential statistics

Researchers increasingly use R for its richness of data analysis features (Kelley, Lai and Wu, 2008; Larson-Hall, 2012). As pointed out in section 2, R is an extremely dynamic tool, with new packages issued almost every week. All of the most important statistical tests and data mining options are available in R, finding the appropriate package and function being the main challenge. For example, in order to check the normality for the distribution of a variable, one can choose from an impressive list of normality tests available in different packages:

- Shapiro-Wilk (*shapiro.test*)
- Kolmogorov-Smirnov (*ks.test*)
- Anderson-Darling (*ad.test*)
- Cramer-von Mises (*cvm.test*)
- Lilliefors (*lillie.test*)
- Shapiro-Francia (*sf.test*), etc.

Due to the fact that many of the variables are discrete with values between 1 and 5, basic statistical tests such as t-test and ANOVA must be replaced by their non-parametric counterparts - Wilcoxon rank sum test (Mann–Whitney U test), Kruskal–Wallis test or Friedman test.

Here are some exaamples of statistical tests we applied in analysis of data from IS graduates survey:

- Chi-squared test (*chisq.test*) and Fisher exact test (*fisher.test*) for testing if:
  - Gender is associated with the profile of the graduates employer
  - Males are more prone to work in IT/technical positions than females
- Wilcoxon rank sum test (*wilcox. test*) for testing if:
  - Female respondents rank the programme, courses, academics, etc. significantly different than their male counterparts
  - There is a difference in assessing the programme, courses, academics, etc. between respondents holding a managerial position versus graduates filling non-managerial jobs.

Additionally we used some other functions for computing confidence intervals and effect sizes.

## Open questions and conclusions

Surveying graduates of academic programmes is extremely important for programmes tuning and adaption of market and community requirements. In this paper we outlined the basic R features for analysis of the data gathered from the IS graduates.

There are some concerns about how well the respondents sample describes the populations of all IS graduates. First the number of respondents was not particularly high (the survey is still activated). Also, it is not clear if the survey respondents are a particular/extreme group of all IS graduates (the most contented or the most discontented), or they share the common traits of the entire population.

R also has its limits. Regular users find it intimidating at first. But it has all the features needed to data analysis and visualization. And it is free of charge, which is good while economic crisis is (still) not over and universities stretch their finances to the limit.

## Acknowledgment

The presented solution was developed within ASIGMA (Asigurarea Calității în Învățământul Masteral Internaționalizat: Dezvoltarea cadrului național în vederea compatibilizării cu Spațiul European al Învățământului Superior) project, POSDRU/86/1.2/S/59367

## References


Amatriain, X. (2013), 'Beyond Data: From User Information to Business Value through Personalized Recommendations and Consumer Science', Proc. of the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13). ACM, New York, NY, USA, 2201-2208

Bryer, J. and Speerschneider, K. (2013), 'likert: Functions to analyze and visualize likert type items. R package version 1.1', [Online], [Retrieved February 25, 2014], http://CRAN.R-project.org/package=likert

Dhar, V. (2013), 'Data Science and Prediction,' *Communications of the ACM*, 56(12), 64-73

Dragulescu, A.A. (2013), 'xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.5', [Online], [Retrieved February 1, 2014], http://CRAN.R-project.org/package=xlsx

Dunn, T. (2011), 'Using 'R' in psychology research', *PsyPag Quarterly*, 81, 10-13

Fotache, M. and Strîmbei, C. (2013), 'SQL and Data Analysis. Some Implications for Data Analysts and Higher Education', Proc. of the Globalization and Higher Education in Economics and Business Administration (GEBA 2013), Al.I.Cuza University, Iasi, Romania, 2013

Kelley, K., Lai, K. and Wu, P.J. (2008), 'Using R for Data Analysis. A Best Practice for Research', Best practices in quantitative methods, J. Osborne (ed.), SAGE Publications, Thousand Oaks, CA, USA

Kumar, D. (2012), 'Data Science Overtakes Computer Science?,' *ACM Inroads*, 3(3), 18-19

Larson-Hall, J. (2012), 'A guide to doing statistics in second language research using R', Routledge, New York, [Online], [Retrieved February 15, 2014], http://cw.routledge.com/textbooks/9780805861853/ R/full-version.pdf

Muenchen, B. (2014), 'Job Trends in the Analytics Market: New, Improved, now Fortified with C, Java, MATLAB, Python, Julia and Many More!', [Online], [Retrieved March 5, 2014], http://r4stats.com/2014/02/25/job-trends-improved/

Musson, R. and Smith, R. (2013), 'Data Science in the Cloud: Analysis of Data from Testing in Production', Proc. of the 2013 International Workshop on Testing the Cloud (TTC 2013). ACM, New York, NY, 18-20.

Robbins, N.B. and Heiberger, R.M. (2011), 'Plotting Likert and Other Rating Scales, Proceedings of the Survey Research Methods Section,' Joint Statistical Meetings, American Statistical Association, Arlington, VA, 2011, 1058-1066, [Online], [Retrieved February 18, 2014], http://www.amstat.org/sections/ SRMS/proceedings/

Sanchez, G. (2013), 'Handling and Processing Strings in R', Trowchez Editions, Berkeley, [Online], [Retrieved February 1, 2014], http://gastonsanchez.com/Handling_and_Processing_Strings_in_R.pdf

Wickham, H. (2007), 'Reshaping data with the reshape package,' *Journal of Statistical Software*, 21(12), 1-20

Wickham, H. (2009), 'ggplot2: elegant graphics for data analysis', Springer, New York

Wickham, H. (2010), 'stringr: modern, consistent string processing,' *The R Journal*, 2(2), 38-40

Wickham, H. (2011), 'The Split-Apply-Combine Strategy for Data Analysis,' *Journal of Statistical Software*, 40(1), 1-29

Wickham, H. and Francois, R. (2014), 'dplyr: dplyr: a grammar of data manipulation. R package version 0.1.3', [Online], [Retrieved March 21, 2014], http://CRAN.R-project.org/package=dplyr