NEW DATA VISUALIZATION TOOLS FOR BETTER UNDERSTANDING STATISTICS. EXAMPLES TO USE IN THE CLASSROOM

Cimpoeru Smaranda
University of Economic Studies, Bucharest, Romania
smaranda.cimpoeru@csie.ase.ro

This paper describes some of the most versatile data visualization techniques that are suitable for teaching statistics, but could also be used by each citizen. In the present paper, I focus on Gapminder and Tableau, and for each of the two I present their main characteristics, advantages and learning objectives in teaching statistics that could be attained with the use of these tools. Moreover, some examples of using them in class are detailed, including personal experiences. Each tool's limits and ways of overcoming them are also covered.

INTRODUCTION

Data visualization techniques have gained a lot of popularity in the last years. Being able to read statistical graphs is one of the main competencies of a statistically literate person (as defined by Gal, 2000). This is due to the fact that many times a simple graph is sufficient to tell a story, to say more about a certain set of data than a complex model. In the GAISE Report (2005), it is emphasized that some of the first steps to be followed by students in order for them to understand the process by which statistics works in answering questions are: learning how to graph the data and know when that suffices to answer the questions; learning how to interpret graphical displays of data.

Nowadays, the graphical tools have developed into more complex data visualization techniques that help us understand and perceive information better. These tools emerged as a solution of the decreasing attention span due to exposure to higher and higher amounts of information and as a means of conveying the essentials of a message in a quick and visually pleasant way. The importance of data visualization consists in making patterns in data visible, drawing attention to exceptions and outliers, improving analysis of data over time.

Data visualization tools are used in a multitude of domains and contexts. Their main functionalities include: data analysis; conversion of complex data and abstract information in a friendly format; enforce a message; tell a story effectively in order to reach heterogeneous audiences; illustrate information, systems, correlations, timelines, etc.

In this paper I will concentrate on two graphical tools that can be used in statistics courses, but also by regular citizens to enhance statistical literacy. We will focus our analysis on Gapminder and on Tableau. For each of the two I will explain what type of problems and data-sets these tools are suited for, examples of applying them in class, contribution to students' development of statistical knowledge and learning objectives that could be attained. The end of each section will have a brief overview of main benefits, but also disadvantages and possibilities of overcoming them.

VISUALIZING DEVELOPMENT DATA USING GAPMINDER

One of the most used and popular tools for visualizing data is Gapminder, a free data analysis tool that incorporates most current data on World Development into one space allowing everyone to understand socio-economic and public health related trends in our world.

The idea of creating this tool came to Hans Rosling and his team while conducting a study to assess the knowledge of Swedish university students regarding common developmental indicators such as life expectancy or literacy rate. Surprisingly, although data was public and available, students were not aware of the developments made by countries in Asia or South America after the 1990s.

As nowadays visual information is the norm (Murphy, 2009) for the young generation, teachers have to diversify the range of pedagogical methods used: verbal, numeric, audio, visual, etc. In this context, Gapminder is an ideal tool to "empower instructors in designing dynamic presentation of real life data, energizing students' natural curiosity and creating lasting impressions" (Trieg, 2013).

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE),
July 2016, Berlin, Germany. iase-web.org/Conference_Proceedings.php

There are several examples of using Gapminder in teaching activities. For instance, at NYC School, "Gapminder at the ISchool" is an experimental course for 10th and 11th grade students that use the tool for analyzing 200 years of global history. Students investigate trends and correlations with the aid of Gapminder. They create research questions, deal with the research process and at the end of the course present their final reports. For example, one of the students investigated the following research question: "How do social and political changes affect literacy rate in West Africa?" In this way students develop skills like: Data Analysis, Quantitative Reasoning or Research Process.

Reagarding teaching statistics, I will outline an example of a class activity with Gapminder for an introductory course in statistics (a similar activity is given in http://betterlesson.com). In the beginning of the class the teacher revises concepts regarding discrete data, relationships between variables and trends, giving the floor to the students for real life examples. Moving forward he plays the inspirational video of Hans Rosling: "200 years, 200 countries, 4 minutes". Students are asked to take notes while watching, paying attention to the following: which variables are depicted on the horizontal and on the vertical axis, what do the different colors of the dots represent, what type of relationship is depicted. The novelty of the graphical representation consists in animating the graph with two main elements: a time component – showing the development over time of the two variables and overlapping a bubble chart – each bubble is a country – the bubble's size corresponds to the country's population size, while the color defines the region. Then students are given tasks to perform using Gapminder. For instance, they are asked to think of two variables that could be positively correlated – they can choose from several categories of variables (Economy, Society, Education, Health – detailed below). Then they test their assumption by constructing the graphs with the Gapminder tool.

In Figure 1 I depicted an example of this type of graph showing the correlation between income per person (horizontal axis) and life expectancy (vertical axis). On the right side, from the menu, one can select only some countries of interest that can be depicted. Moreover, changing the variables in the graph can be done easily, as well as the scale. By pressing the button "Play" below the horizontal axis, the graph is animated through time giving a suggestive image for the evolution in time of the relationship between the chosen variables. Usually, this graph generates interesting questions like: "Are more developed countries prone to a higher life expectancy? Does it make sense? Why?", triggering further debate.

The tool is very easy to learn by students, attractive for its friendly interface and can be used with success in introductory classes of correlation.

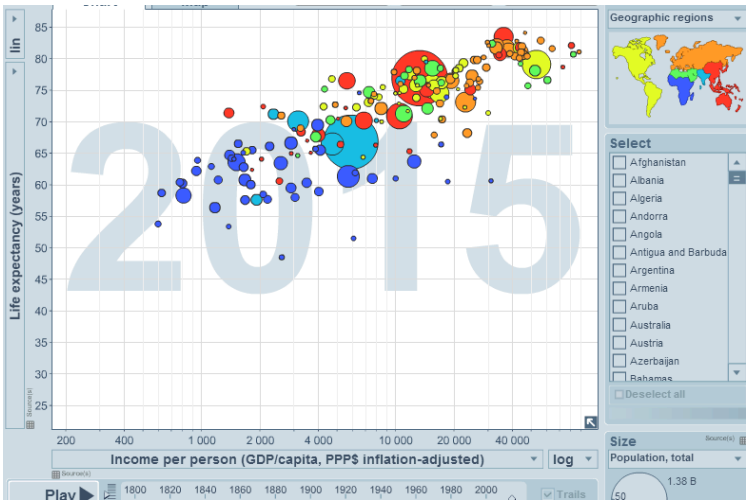

Figure 1 – 200 years that changed the world. Source: http://www.gapminder.org/

Gapminder could also be used in an introductory lesson for Time series. For instance, in Figure 2, we have a Gapminder graph showing the evolution of Income per person over the last 200 years for the key Asian Economies (Japan, China, India), United Kingdom and United Arab Emirates. Students are asked to play the animation and comment on the different patterns. The

teacher could lead the discussion with some additional questions: Is it a linear evolution of the income per person or not? What effects did major economic events have on the income's evolution (like World War, India's independence, 2008 global crisis)? Which country had the most "spectacular" evolution? etc. This type of activity familiarizes students with time series and the concept of trend, thus it could be used successfully as an introduction to the time series chapter.

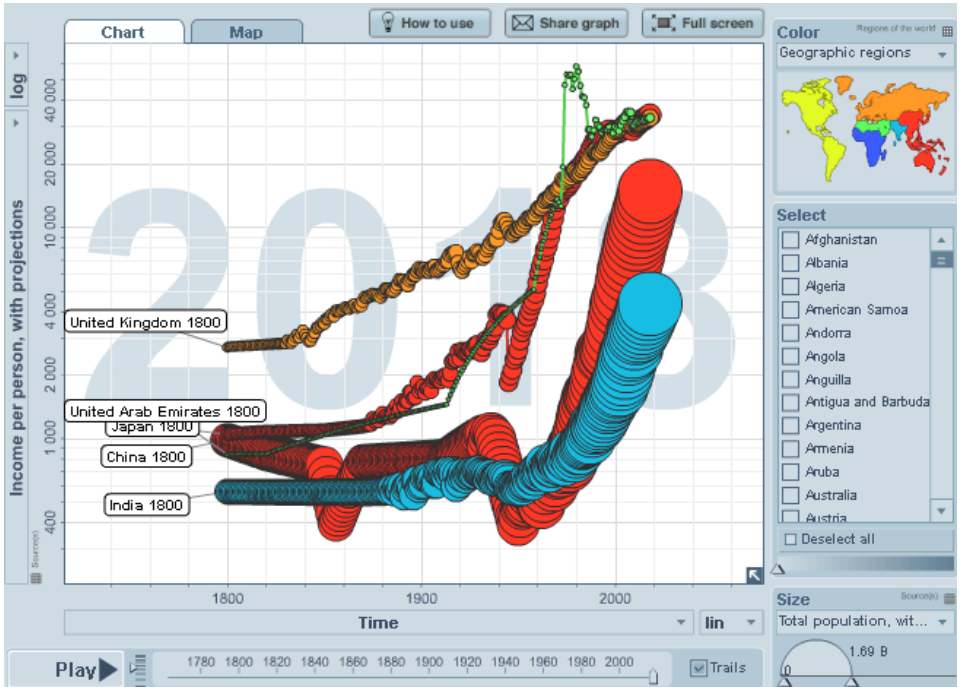

Figure 2 – Evolution of Income per person in time for Asian countries vs. UAE and UK

The benefits of using Gapminder for teaching statistics could be summarized as follows:

- Is a free and friendly, easy to work with software; it can be used either online or downloaded, with data provided in Gapminder World – over 400 indicators on global development.
- Arouses students' curiosity for real world events and thus increases motivation to learn and understand statistics; students are able to explore numbers in the context of the large world around them so it connects them with worldwide events.
- Seizes the power of statistics by expressing 5-dimensional data into one place. The 5 dimensions of the Gapminder graphs are: variable on the horizontal axis, variable on the vertical axis, time, geography (color of the dot), population (size of the dot).
- There is a wide range of quantified data available on Gapminder, so students can chose and switch between variables, depending on their interests. Available sets of data include:
    - Economy: Poverty rate, exports, imports, industry or agriculture as % of GDP
    - Society: Murder rate, Military expenditure
    - Education: Literacy rate, spending / student
    - Environment: CO2 emissions, energy use / person
    - Health: life expectancy, child death rate, fertility rate
    - Population: population growth, median age, population density.

  Students can choose researching various questions (for instance: "Are more developed countries prone to a higher life expectancy?"; "Is there a higher literacy in countries with a greater spending per student?")
- Learning objectives in statistics which can be covered by using Gapminder are:
    - Understanding graphical representation and the meaning of individual points on a scatterplot
    - Identify independent and dependent variables
    - Distinguish between correlation and causation
    - Explore positive and negative correlation

  - Identify outliers
  - Understand correlation trends
- Students can enhance their abstract and quantitative reasoning

However, there are some limitations of Gapminder that we outline below:

- It can be used only to explore trends and relationships among variables; Gapminder is not a software for data analysis like Excel or SPSS; for this type of analysis a dedicated software should be used.
- The graphs are constructed using by default the logarithmic scale, and this could be misleading in what concerns the dispersion between countries (small distances are transformed in larger ones if the log scale is reversed)
- The indicators presented are averages and do not account for the variability existing within a country (a well known example is China which has very different development levels among provinces). However, this could be explained to students as an argument for the importance and effects of variability in data.
- The tool cannot be used with own data; however there are solutions to this limitation, the Google Visualization Api application could be used instead.

In his paper, Le (2013), shares the feedback received from students after their project assignment with Gapminder. The feedback was very positive – it appears that students enjoyed the activity, they took pleasure in working with data of their interest and they were surprised to see how much the world has changed over time. They consider Gapminder as a "valuable tool to learn about issues that face our world" (Le, 2013).

My personal experience with using Gapminder in class is in line with what has been reported by Le (2013). In the introduction of the correlation chapter I play the video by Hans Rosling which raises important issues and debates in class. Students perceive better the abstract notions of correlation, regression, they manage to discriminate between cause and effect and they make the distinction between correlation and causality. I find it very useful as this fits their daily involvement in technology. In the future I plan to further develop the teaching activities that use Gapminder, starting from the examples outlined above.

DATA VISUALIZATION TOOLS – FOCUS ON TABLEAU

In the business context, tools for convincing audiences do not resume anymore to classic presentations including descriptive graphs. Persuasive presentations have to include interactive dashboards and make use of advanced visualization tools. Not only that the tools for constructing this type of graphical representation should be included in the scholar curricula of statistics courses, but learning to work with such an instrument could be in the benefit of the regular citizen, in the ways we will describe hereinafter.

We will focus our analysis on the Tableau software, mainly due to benefits that we will outline briefly.

- Tableau is considered one of the leaders of the graphical technologies (for instance considering the Gartner's Magic Quadrant ranking), its visualization capabilities are diverse and highly insightful.
- The application has a free version for students and teachers, increasing its accessibility to the wide public; it can be used with sample data provided upon first downloading the program, but it allows also use of own dataset.
- It is able to explore data in a very friendly manner, by dragging and dropping data dimensions, variables or parameters through a visual, easy-to-use interface.
- Features such as "bubble maps", "heat maps" or "word clouds" are embedded into a visually driven interactive data-exploration platform, allowing students or regular citizens to change the way they think about data and the information from the world they live in.

Tableau can be used in class in order to attain objectives as: becoming familiar with a visualization tool for data analytics, learning to make inferences, define hypotheses and formulate questions based on visualizations, become familiar and gain proficiency in using the computer for data analysis; enhance students' data analysis, management, and interpretation skills.

For example, Deanna M. Kennedy applies Tableau in teaching the Quantitative Methods and Business Statistics course at the University of Washington. In her interview with Tableau, she mentions that Tableau was introduced as a data analysis tool and the main reason for choosing this software was to increase students' motivation for statistics and make the statistics course enjoyable. Moreover, the program is easy to learn in a 10 week time frame.

For this course, the teacher used a computer and the projector. Students where thought how to create a dashboard from sample data provided by Tableau. They could easily follow and replicate the actions on their computers. Dr. Kennedy reports that students learned quickly the simple techniques showed and started creating more complex graphics.

The evaluation was project based. Students worked in teams of 3 to 5 for a group project which had as main goal using statistical methods to create new knowledge about technology, business or industry. Their topics were very diverse, ranging from the mobile telephone market to the e-book sales or from unemployment to the electric car industry.

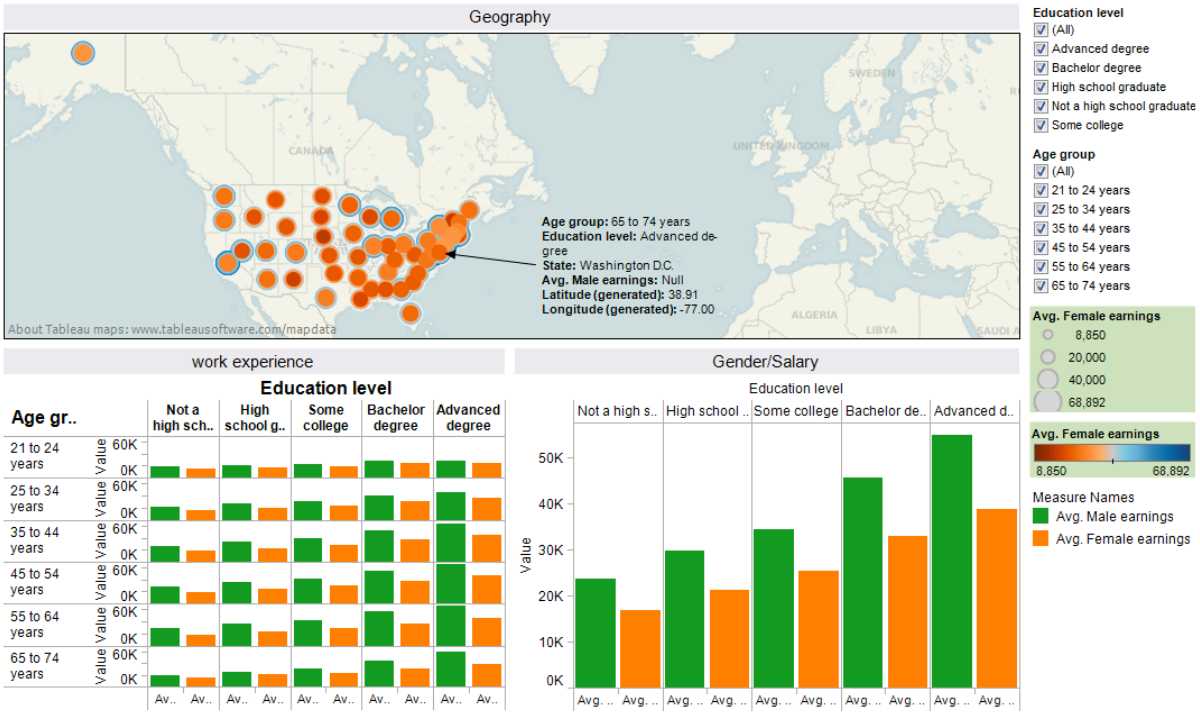

Figure 2 – Example of student project using Tableau
(source: http://www.tableau.com/academic/teaching/university-of-washington)

The final step of the project was to create a dashboard and explain an audience how a manager might use the information shown in the dashboard to make a decision. An example from one of these projects is presented in Figure 2. Basically, the dashboard is centered on the variable Salary which is analyzed through a three-dimensional perspective – educational level, gender, age. The feedback from students was very good, students enjoyed using the program and many of them declare to use beyond the class.

Disadvantages of Tableau are mainly related to its limitations for business purposes (enterprise reporting, scheduling alerting). This however does not have a strong impact on its potential as a data visualization tool to use in class. Dr. Kennedy reports that the most challenging issue in working with Tableau in class was how to organize data in the right format for Tableau. For this task of organizing data, the teacher has to allocate time at the beginning of the course.

Personally, I have not used Tableau in class until now, but I plan to introduce it for a course of "Quantitative Methods of Research", mainly due to its accessibility, easy-to-use interface, and powerful results obtained. The experience gained with Gapminder shows that students respond better to this type of data visualization tool, especially if the audience has not a sound mathematical background, which is the case with some of the students from Social Sciences. In this context, I have designed a teaching activity with the aim of attaining the following learning objectives: data exploration, summarizing data, frequency distribution, graphs. The data set used in

this example comes from a survey conducted by the Foundation Friedrich Ebert together with a Romanian research institute. Their extensive study deals with the attitudes, lifestyle, interests, beliefs of the young generation in Romania (people aged between 15 and 29 years). As the students belong to this target group, they will feel close to the activity and gain much interest in exploring the data.

As a starting activity, the task will be the demographic description of the sample. For this, students will select variables like: gender, residential milieu (urban/rural), education level, region (three historical regions in Romania plus Bucharest), internet connection (yes/no). The frequency distribution tables plus some basic graphs that could be easily obtained using Tableau are depicted in Figure 3.

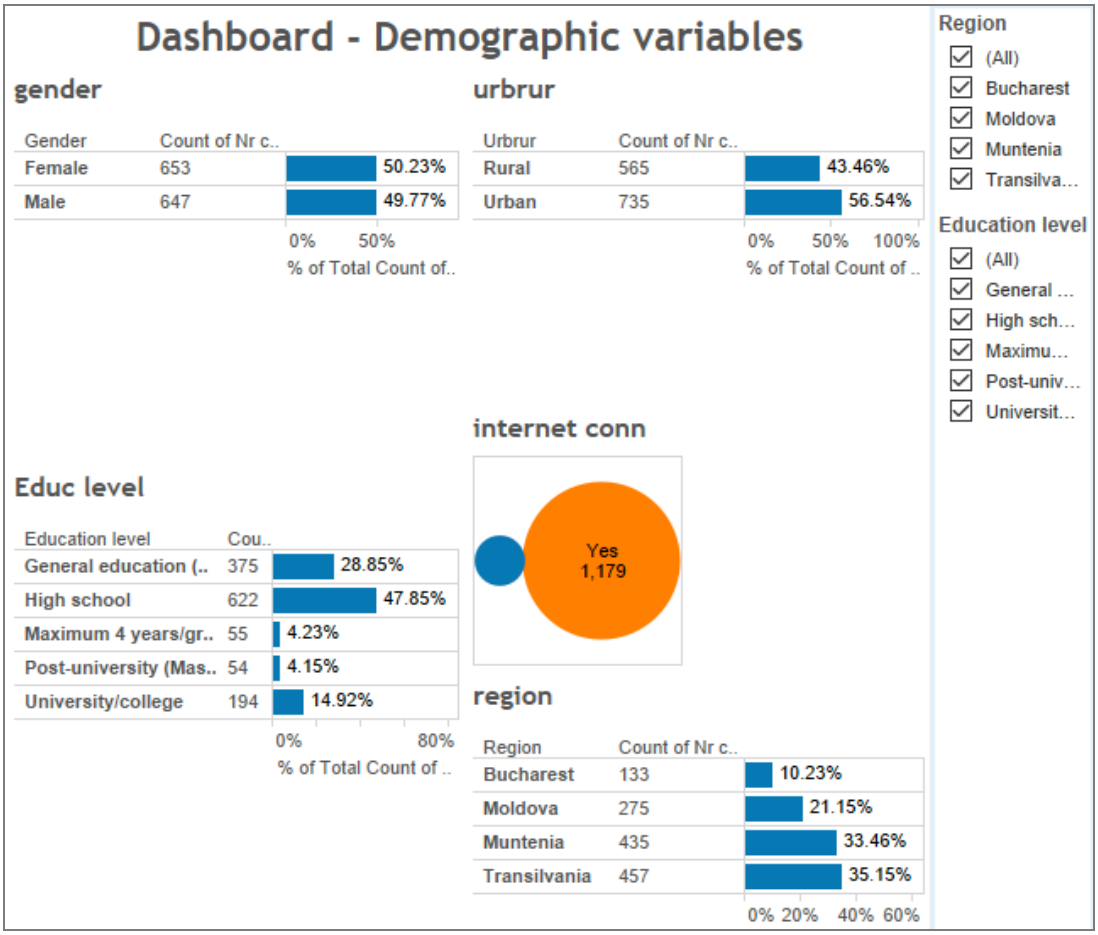

Figure 3 – Dashboard with the demographic variables defining the sample

Through the dashboard, students understand better the multivariate approach of a data analysis task. The teacher should encourage questions from students. For instance: "Do you notice a difference between respondents with different education levels in what concern their access to the Internet?", or "Are young people in certain historical regions more educated than in other regions?", etc. Different assumptions could be easily verified using the dashboard by adding variables as filters and making several selections. In Figure 4 below two selections were made – respondents with a higher level of education (left side) versus the ones with only primary school (right side). Students could easily see that women and respondents with internet access are inclined to a higher education level. This is a basic but good example for teaching students what data exploration means, the richness of data and which are the first steps for transforming data into significant information.

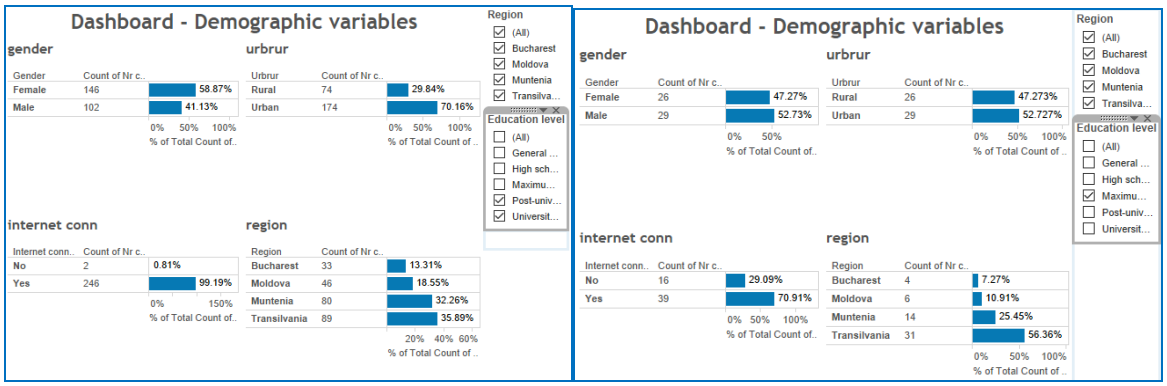

Figure 4 – Two comparative dashboards / Variable used for selection: "Education level"

CONCLUSION

In a world where the volume of information increases exponentially, analyzing and synthetizing data has become a real challenge. The new tools used for data visualization that we presented in this paper aim at uncovering the possible patterns and reveal the true power of data, all done in an easy to understand way, with a significant visual impact.

We have focused on two powerful visualization tools: Gapminder and Tableau. The graphs created with Gapminder are very suggestive and the motion of bubbles through time creates a visual image of economic evolution. Examples of available graphs cover a diversity of domains: wealth and health (income per capita and life expectancy); poverty (income per person and % people below 2$ a day); environment (income per person ad CO2 emissions in tones per person). The tool is a very good way to gain interest of students in class, although we stress the fact that this can be used just as a complementary method for introducing the theme and cannot replace the statistical software to evaluate correlation or regression analysis. We propose an activity of constructing dashboards using Tableau, in order to introduce data exploration to students that do not have a sound mathematical background.

Both tools presented in this paper can be easily used in class, as shown by the pedagogical examples we have briefly outlined. Tableau has the advantage that it can be used with own data, whereas for Gapminder only the built-in data can be applied. On the other hand, Gapminder is very straightforward, while Tableau requires more time to be spent with students in order to learn how to organize data. Neither of the two programs can replace a statistical package like SPSS, but the main advantage of both is that they increase students' interest and motivation in learning statistics, it helps them become familiar with graphs, correlations, develop and test research questions. Own experience has confirmed the increased interest developed by students when using this type of tools.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report*. American Statistical Association. Alexandria VA. www.amstat.org/education/gaise

http://betterlesson.com/

Boyle, E. A., MacArthur, E. W., Connolly, T. M., Hainey, T., Manea, M., Kärki, A., & Van Rosmalen, P. (2014). A narrative literature review of games, animations and simulations to teach research methods and statistics. *Computers & Education*, *74,* 1-14.

Dur, B., & Inanc U. (2012), Analysis of data visualizations in daily newspapers in terms of graphic design, *Procedia-Social and Behavioral Sciences 5,* 278-283.

Gapminder web site: https://www.gapminder.org

Le, Dai-Trang (2013). Bringing data to life into an introductory statistics course with Gapminder, *Teaching Statistics 35.3*, 114-122.

Murphy, S. (2009). *The power of visual learning in secondary mathematics education. Research into Practice MATHeMATics*. Pearson Education.

Rosling, H., Rosling Rönnlud, A., & Rosling O. (2005). New software brings statistics beyond the eye, Statistics, *Knowledge and Policy: Key Indicators to Inform Decision Making. Paris, France: OECD Publishing*, 522-530.

Tableau web site: *http://www.tableau.com/academic/teaching/university-of-washington*