Challenges and Innovations in Statistics Education
Edited by Péter Kovács. University of Szeged, January 2018
ISBN 978-963-306-575-4
Proceedings of Challenges and Innovations in Statistics Education Multiplier Conference of ProCivicStat

EXPLORING CIVIC STATISTICS WITH CODAP

Joachim Engel
Ludwigsburg University of Education, Ludwigsburg, Germany
engel@ph-ludwigsburg.de

Data are abundant, quantitative information about the state of society and the wider world is around us more than ever. In order to root the public debate based on facts instead of emotions and to promote evidence-based policy decisions we as statistics educators are challenged to promote understanding of statistics about society. This report summarizes a hand-on workshop in which participants explored the potential of the freely available web-based data platform CODAP (http://codap.concord.org) and its usefulness for investigating civic statistics data. Materials of the workshop were developed as part of the EU-funded ProCivicStat Project and are freely available (www.procivicstat.org).

INTRODUCTION

Data on important societal topics are becoming increasingly accessible to the general public and to individual citizens or social action groups, on a huge range of topics such as migration, employment, social (in-)equality, demographic changes, crime, poverty, access to services, energy usage, living conditions, health and nutrition, education, human rights, and many others. In order to ground public debate to be based on facts instead of emotions and to promote evidence-based policy decisions, statistics education needs to embrace two areas widely neglected in secondary and tertiary education: understanding of multivariate phenomena and the thinking with and learning from complex data (Engel 2016, Ridgway 2015). The project ProCivicStat, a strategic partnership of six universities funded through the Erasmus+ program of the European Union, explores a subfield we call Civic Statistics which focuses on understanding quantitative and statistical information about society as provided by the media, statistics offices and other statistics providers. Understanding Civic Statistics is required for participation in democratic societies, but involves data that often are open, official, multivariate in nature, and/or dynamic, that is not normally taught in regular mathematics and statistics education, let alone in politics or social studies.
The challenge is multi-facetted. Data literacy for civic engagement involves, among many other aspects, specific statistical knowledge, ICT skills, "data habits of mind" (Finzer 2013), critical thinking, and much more (Engel 2017). In addition to conceptual blueprints for understanding multivariate phenomena in a data-rich world, the EU project ProCivicStat provides authentic and relevant data sets and develops and tests teaching and learning materials for innovative teaching to a wide range of target groups. The ultimate goal of the project is to strengthen civil society, empowering informed citizens to evidence-based decision-making and civil society engagement. Teaching materials, extensive datasets, conceptual representations of civic statistics are available through the website www.procivicstat.org.

COMMON ONLINE DATA ANALYSIS PLATFORM

ProCivicStats addresses the needs of the civil society and aims to empower students to become informed and active citizens. Its target group are not professional statisticians, computer scientists or empirical researchers. To make data about society openly accessible and data visualisation tools manageable for the general public, appropriate easy-to-use digital tools are essential. Over the last decade many innovative data visualization tools have been developed that permit even novice users to engage with data in ways that are impossible using static displays as in textbooks or print media. The Common Online Data Analysis Platform CODAP (http://codap.concord.org) is a recently developed tool for data exploration and visualization that allows the user to do a whole range of own data analysis steps. CODAP is a freely available, web-based environment for data management and visualization of complex data that supports the many desired transformations and restructurings of data. CODAP is designed as an educational tool for novices. There is no need to

install software. Provided with a link to a CODAP document, an internet connection and a browser suffices to get started.

Most data in traditional textbooks are “flat”, i.e. arranged rectangularly in tables with a manageable number of lines (“the cases”) and some columns (“the variables”). The textbooks usually provide only the columns necessary to solve a given problem. In contrast, authentic data from the net, e.g. about the state of society, are often hierarchical (ordered according to countries, regions, continents, by years, etc.) and usually have a complex multivariate structure of correlated variables that are often non-linearly related - unlike the data of current mathematics education in which students learn statistics. Sometimes, data from the internet must be aggregated or disaggregated before analysis, variables must be re-encoded or transformed to allow appropriate visualizations.

Figure 1 shows a screenshot of the top rows of a data table with 96061 records obtained from the United Nation data base on worldwide refugee movements (http://data.un.org/Data.aspx?d=UNHCR&f=indID%3AType-Ref). The first row refers to the number of refugees (n=1) who left Iraq to go to Afghanistan in 2013, the second row lists Iranians coming to Afghanistan etc. Depending on the question of interest, the data have to be rearranged by country of origin or by country of residence before being graphed. When trends over time are of interest, the same data have to be restructured again in a timely order. In other situations data may have to be transformed or aggregated in order to be useful for an illustrative representation or the desired analysis. Cleaning, transforming, and structuring data are necessary skills, but these skills are not taught in the traditional classroom with its focus on inference based statistics with tidy data.

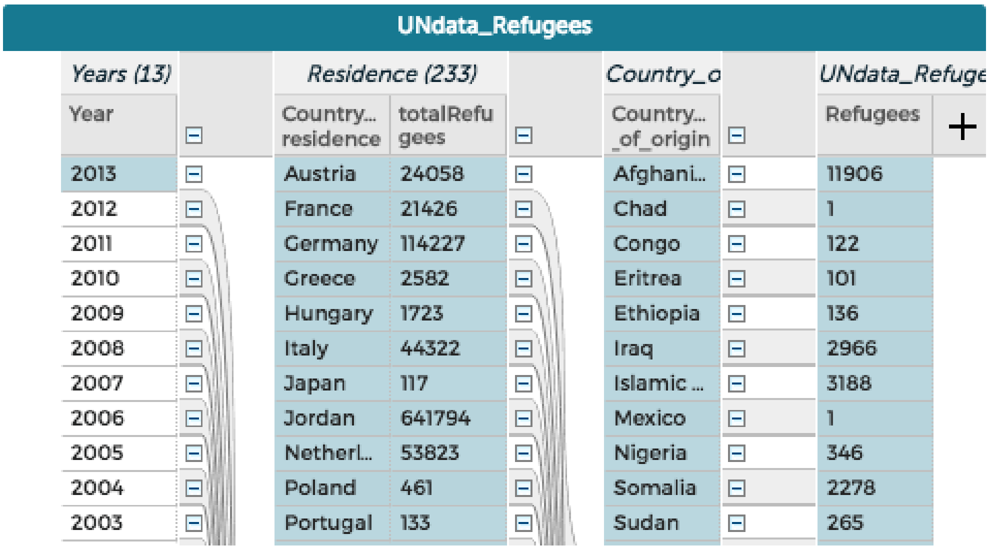
Download Explore Select columns Select sort order Link to this page

80428 records | Page 1 of 1609 |

Country or territory of asylum or residence	Country or territory of origin	Year	Refugees*	Refugees assisted by UNHCR	Total refugees and people in refugee-like situations**	Total refugees and people in refugee-like situations assisted by UNHCR
Afghanistan	Iraq	2013	1	1	1	1
Afghanistan	Islamic Rep. of Iran	2013	36	36	36	36
Afghanistan	Pakistan	2013	34	34	16,825	16,825
Afghanistan	State of Palestine	2013	1	1	1	1
Albania	Algeria	2013	0	0	0	0
Albania	China	2013	12	12	12	12
Albania	Dem. Rep. of the Congo	2013	5	5	5	5
Albania	Egypt	2013	3	3	3	3
Albania	Iraq	2013	5	5	5	5
Albania	Montenegro	2013	2	2	2	2
Albania	Peru	2013	1	1	1	1
	Serbia (and Kosovo:					

Fig. 1: First rows of a data table retrieved from the UN Data Base on worldwide refugee movement

CODAP supports the required transformation and restructuring of the data. Figure 2 shows a display of the first rows of the refugee data, ordered by year (upper level) aggregated at the level of country of residence. Aggregation and restructuring of the data table were made possible in CODAP by simple drag-and-drop data moves. Now, the highlighted data of 2013 can be graphed displaying the number of refugees for each residence country in 2013.



UNdata_Refugees

Years (13)	Residence (233)		Country_o	UNdata_Refuge
Year	Country... residence	totalRefugees	Country... _of_origin	Refugees
2013	Austria	24058	Afghani...	11906
2012	France	21426	Chad	1
2011	Germany	114227	Congo	122
2010	Greece	2582	Eritrea	101
2009	Hungary	1723	Ethiopia	136
2008	Italy	44322	Iraq	2966
2007	Japan	117	Islamic ...	3188
2006	Jordan	641794	Mexico	1
2005	Netherl...	53823	Nigeria	346
2004	Poland	461	Somalia	2278
2003	Portugal	133	Sudan	265

Fig. 2: First rows refugee data restructured with CODAP to explore the yearly distribution of refugees across various host countries

HANDS ON ACTIVITIES

In the following we present and discuss some of the activities participants of the workshop performed. Participants received worksheets (available through the ProCivicStat website [www.procivicstat.org)](http://www.procivicstat.org) with background information about the context, a short technical description of the data including its source and several closed and open questions to guide the exploration to be done individually or in pairs. A link provided on the worksheet leads to a CODAP document containing the required data.

1. Some so rich others so poor – Income distribution in Europe

Promoting Civic Engagement via Exploration of Evidence: Challenges for Statistics Education

ProCivicStat © - Students' Worksheet, 5.101

A few so rich, others so poor - income inequalities in Europe

Joachim Engel   Achim Schiller

Ludwigsburg University of Education
Ludwigsburg, Germany

What is inequality?

Inequality is typically viewed as different people or households having different degrees of living standards. Thus, inequality is concerned with the relative position of different individuals (or households) within a distribution.

Why does it matter?

On July 14, 1789, the French stormed the Bastille, a medieval fortress-prison in Paris in one of the key moments of the French Revolution. The average people were fed up with the — aka the monarchy — and they were protesting the vast inequality between themselves and the upper echelon.

268 years: inequality is still a major issue.

Fig 3: Head of the worksheet on income inequalities in Europe

Why are there in some countries large discrepancies between the rich and the poor while in other countries the income distribution is more equal? What have countries in common that have a large discrepancy between the rich and the poor? Figure 3 shows the head of the worksheet.

The data from EuroStat include the following variables: Country, Year, Population size, Mean Income, Median Income, the Gini-Cofficient for the income data as well decils, quantils and the income share of the 5% lowest and 5% highest earners per country and year.

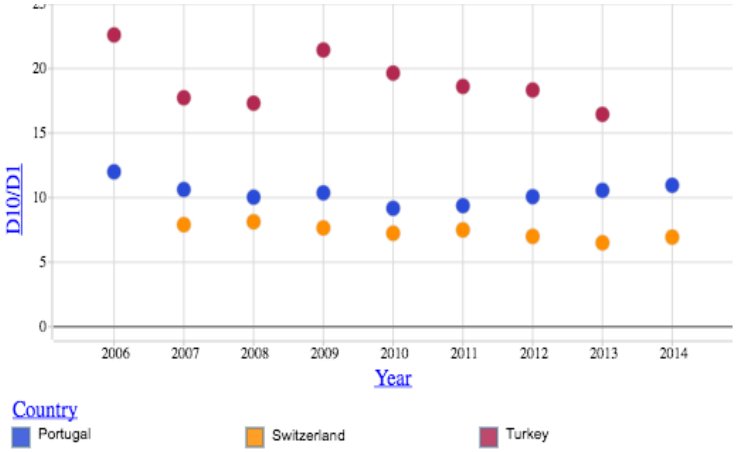


Fig 4: Dotplot of the ratio 10th decile and 1st decile between 2006 and 2014 for Portugal, Switzerland and

One question on the worksheet asked:

To investigate the trend of a newly defined variable D10/D1, defined as ratio of the 10th to the 1st decile, over the years for Portugal, Switzerland and Turkey, what do these countries have in common? How do they differ with respect to D10/D1?

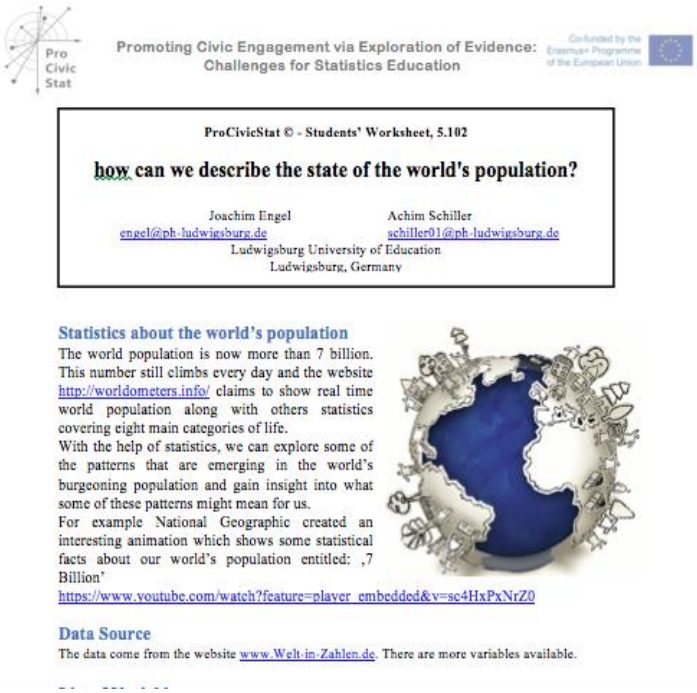
To address this question required the following steps: (1) define a new variable D10/D1 for the quotient of the 10th and 1st decile; (2) plot D10/D1 versus year; (3) select the three countries in question and hide all the unselected cases, resulting in the visualisation of Figure 4.

A possible answer to above questions could be:

The D10/D1 ratio in Switzerland has a minimum of 6.5; the maximum is 8.13. In Portugal the minimum is 9.17 and the maximum is 10.96. Turkey has as minimum 16.45 and as Maximum 21.44. In Switzerland, no real trend can be seen over the years 2007 to 2014, and the rise and fall in value is changing irregularly. In Portugal, the ratio has risen again since 2010, which means that the uneven wealth distribution is increasing. In Turkey, a downward trend can be seen after the financial crisis in 2009, which means that inequality is decreasing. However, there are still clear differences between these countries.

2. How can we describe the state of the world's population

The world population is now about 7.6 billion. With the help of statistics, we can explore some of the patterns that are emerging in the world's burgeoning population and gain insight into what some of these patterns might mean for us (see Figure 5).


Promoting Civic Engagement via Exploration of Evidence: Challenges for Statistics Education

ProCivicStat © - Students' Worksheet, 5.102

how can we describe the state of the world's population?

Joachim Engel engel@ph-ludwigsburg.de
Achim Schiller schiller01@ph-ludwigsburg.de
Ludwigsburg University of Education
Ludwigsburg, Germany

Statistics about the world's population

The world population is now more than 7 billion. This number still climbs every day and the website http://worldometers.info/ claims to show real time world population along with others statistics covering eight main categories of life.
With the help of statistics, we can explore some of the patterns that are emerging in the world's burgeoning population and gain insight into what some of these patterns might mean for us.
For example National Geographic created an interesting animation which shows some statistical facts about our world's population entitled: ‚7 Billion'
https://www.youtube.com/watch?feature=player_embedded&v=sc4HxPxNrZ0

Data Source

The data come from the website www.Welt-in-Zahlen.de. There are more variables available.

Fig. 5: Head of worksheet on world's population

The data are from the website www.Welt-in-Zahlen.de.
17 different variables, representing average values for 222 countries of this planet such as migration, nutrition (in kilocalories per person), fertility, GNP per person, internet access, number of physicians per 1000 and many more.

To investigate the distribution of nutritional intake per person across continents, CODAP allows to draw boxplots for each continent, thus revealing that people in some countries of Africa and Asia live on very low calories. But observe the large spread (Figure 6). Figure 7 shows scatterplots plus least square line, separate for each continent, for GNP per Person versus number of physicians per 1000 people

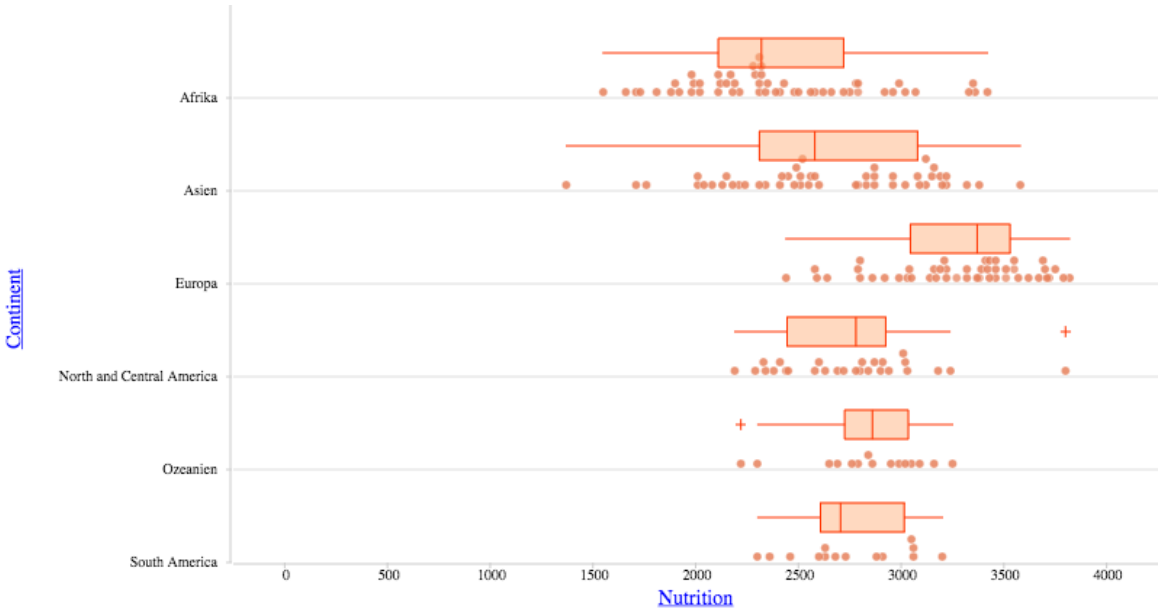
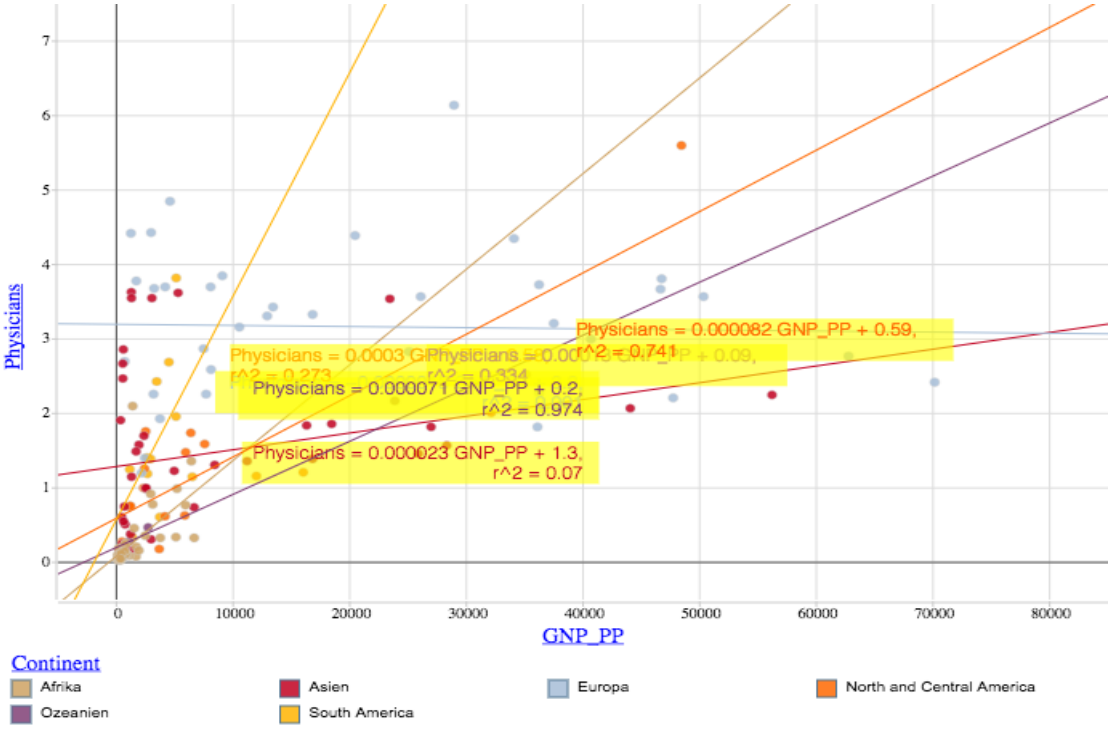


Fig 6: Boxplots representing mean human calorie intake, separated by continent. Each dot represents a different country.



-

Fig. 7: Scatterplot of “number of physicians per 1000 inhabitants” and “GNP per person” in 222 countries with least squares lines. Every dot represents a country, continents are coded by dot color.

3. Are referees in European football racially biased?

Fig. 8: Head of worksheet on investigating possible referee bias

Are players with dark skin tone more likely than light skinned players to receive red cards from referees in European football (see Figure 8)? The data set comprises information on 1419 football players in four professional European football leagues with 19 variables such as number of red, yellow-red and yellow cards received during a player's career, position played, height, weight and a rating of skin color (1=very light, 5 =very dark). While it is easy to create boxplots of the variable RedCard or the newly defined variable RedCardsRate (= RedCards per Game), the tricks question is the search for possible confounding variables, i.e. third explanatory variables that may account for an observed relationship between two variables. Therefore, before looking at boxplots separate for of, say, RedCardsRate at each level of the variable Skintone, one may ask: Are red cards and skin color equally distributed across the four countries (England, France, Germany and Spain)? What about the distribution of the position across players of different skin color? Are players of color more often represented in some positions than in others?

Figure 9 displays the distribution of Position versus Skintone and Country League versus Skintone. The figures reveal that dark skinned players play more often as attackers or midfielders than as defenders or goalkeepers. Also, the French and British League has a higher percentage of coloured players than Germany or France.

These observations matter to our guiding question for potential racial bias of referees. For red cards tend to be given more often to defenders than to midfielders or attackers. Also, referees in England, Spain and France tend to brandish red cards more often than in Germany (Figure 10). All of this calls for caution in jumping to quick conclusions because Country League and Position may well be confounding variables that may have a strong impact on referees red card giving, thus masking the skin tone variable.

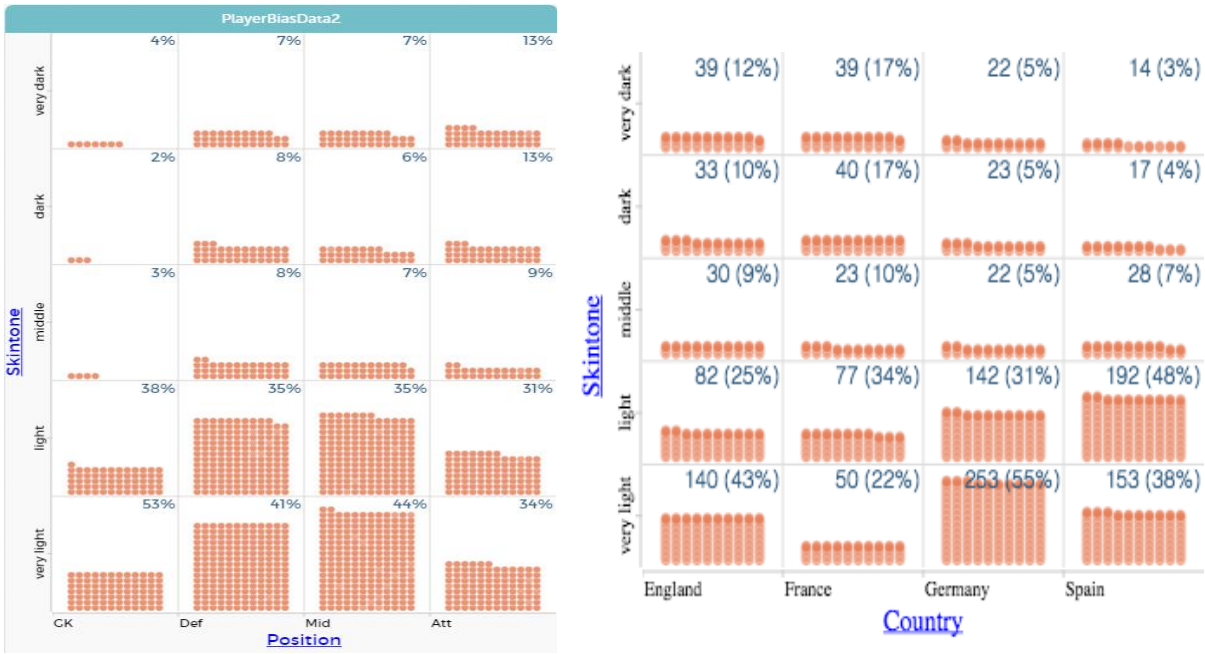


Fig. 9: Bivariate distribution of the nominal variables Position and Skintone (left) and Country League and Skintone (right)

In summary, we conclude by means of the diagrams that colored players are increasingly attackers. The position of midfielder and defender have nearly the same proportion of all skin tones while coloured goalkeeper are underrepresented.
Likewise, the variable RedCardRate is not evenly distributed across the four.

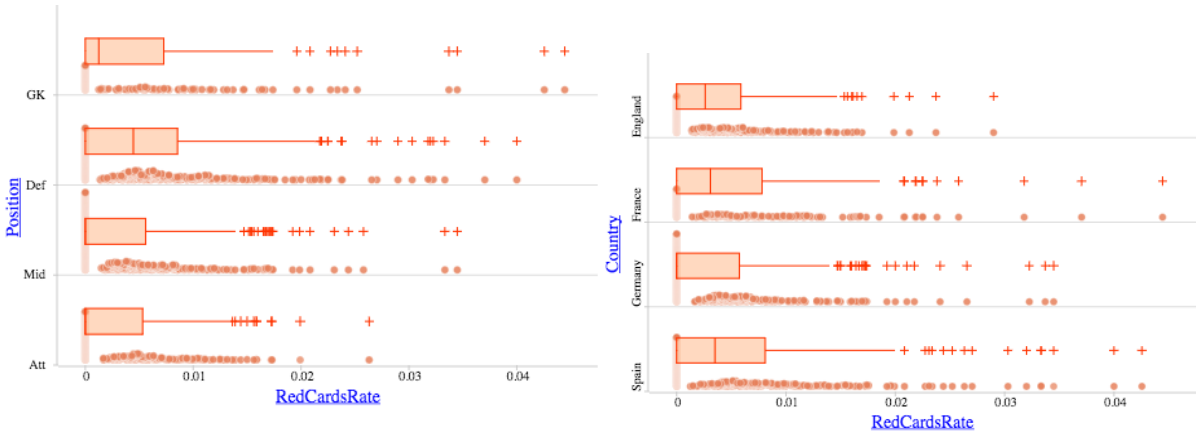


Fig. 10: Distribution of RedCardRate across the position of a player (left(and across country leagues (right)

CONCLUSION

CODAP proved to be a tool, most users manage to work with even without much instructions. Recurring topics for the data analysis based on CODAP were: (1) Comparing distributions, (2) Aggregating data, (3) Restructuring data, (4) Investigating and comparing subgroups, (5) Search for explanatory third variables, and (6) Model functional relationships.

Introducing new digital tools for data analysis to support learning with and from data raises important questions. What is the relative value of curated versus self-selected data sets? In ProCivicStat we provide curated data sets on a whole range of topics related to Civic Statistics. We initiate the analysis with a few closed questions, but also ask for more open exploration which may require from the learner to immerse deeper into the context. On a technical level, new questions may imply the need to define new variables and to restructure the data.
An alternative could be to let learners choose their own topic and have them search for appropriate data sets. It may be more powerful and motivating to learn with the data you have chosen yourself. We tried this path in a seminar with students at our university. However, it is hard to find suitable data you are looking for; if you found them, it is often not at all trivial to import them into the software you are using (even though CODAP allows under some circumstances for web scraping) and the freely chosen data set may not teach what the instructor wants to teach.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by ProCivicStat project, a strategic partnership of the Universities of Durham, Haifa, Ludwigsburg, Paderborn, Porto and Szeged, funded by the ERASMUS+ program of the European Commission. However the views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funding agency.

REFERENCES

Engel, J. (2017). Statistical Literacy for active citizenship: a call for data science education. *Statistics Education Research Journal* 16(1), 44-49.

Engel, J. (Ed.) (2016). Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), July 2016, Berlin, Germany. The Haag, the Netherlands: ISI/IASE. http://iase-web.org/Conference_proceedings.php.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, 7(2). www.escholarship.org/uc/uclastat_cts_tise.

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3) 528–549.