GIRLs in data science: FIRST STEPS IN computational LEARNING with R

Karla Patricia Santos Oliveira Rodriguez Esquerre [1], Herica de Oliveira Santos [2], Júlia Carolina Braz de Freitas Bijos [3] e Caroline Vasconcelos Fernandes [4]

Introduction

The students gender gap in science, engineering and computing courses has decreased, but evasion, especially in the first years of undergraduate courses, is still a problem that affects the female public in general. In addition, women are minorities in research and extension projects in university that require programming.

When speaking about basic education, in Brazil, private schools still have incipient teaching in programming subject while in public schools these practices are practically nonexistent. The technological characteristic of the 21st Century, has favored the formation of digitally active children, although, disconnected with school subjects. They not have the computational and analytical thinking that allows them to analyze and criticize daily life data nor to produce knowledge and technology from images, texts, data and sounds.

On August 2018th, the National Council for Scientific and Technological Development (CNPq) and the Ministry of Science, Technology, Innovations and Communications (MCTIC) opened a call CNPQ/MCTIC 31/2018, aiming to support projects that contribute to the scientific and technological development and innovation in Brazil by encouraging the participation and training of girls and women for the careers of Science, Technology, Engineering, and Mathematics – STEM. A project proposal entitled Gender Diversity in Data Science: Learning through Evidence-Based Practice, coordinated by Professor Karla Esquerre, was awarded. The project main goal is create ways to support girl students of lower secondary school to develop their computational and statistical thinking, approaching them to the area of exact sciences by the surveying, exploration, modeling and data analysis from a variety of traditional and non-traditional data sources related to everyday issues.

_____
[1]GAMMA, PEI, Federal University of Bahia (UFBA), karlaesquerre@ufba.br

[2] PEI, Federal University of Bahia (UFBA), hericaufba@gmail.com

[3]PEI, Federal University of Bahia (UFBA), juliabijos@outlook.com

[4]PEI, Federal University of Bahia (UFBA), carolinefernandes.eq@gmail.com

Objectives

This short paper aims to share with the community the first steps and results of the Girls on Data Science Project, and to show how new concepts based on analytical and computational thinking have been introduced to the students by using R programming.

Material and Methods

The project focus on working with about 500 girl students currently enrolled in 6th to 9th grades and their respective teachers in areas. There are four state schools and one municipal school located in low-income neighborhoods, nearby the Federal University of Bahia. The working team is multidisciplinary and is composed of undergraduate (8) and graduate (4) students, professors (14) and professionals (3) of Exact Sciences and Human Sciences, mostly working as volunteers.

Three main stages are proposed to implement the project: 1- Approaching the School; 2 - Formative and Informative Process; 3 - Connection between Theory and Practice. On the first stage, the importance of data science and women's role are introduced and discussed by using games. The last two stages use PBL method (Problem Based Learning), where the students are led to analyze data-driven problems to gain knowledge and skills on data science. So, STEM are connected to statistics, algorithms, programming, and cross-cutting themes. Instructors guide the students in asking questions to define a problem and help them to analyze and synthesize data.

Each school has disponibility of 25 computers (Google Chromebooks), which have been provided by Google Corporate as part of a Government Education Program. So, programming logic and concepts about computational thinking are worked in computational labs. During these classes, R programming will be used because it is free and open-source; it is available for a wide variety of operating systems; and it provides an integrated development environment for statistical and graphing calculations. Other programming languages, such as Python and Scratch, will be used also to facilitate the understanding of the contents and the development of the activities by the students.

Results and Discussion

Stage 1 revealed that the students have barely heard about Data Science or women scientists and have never thought about exact science careers. The project is in Stage 2,

initiating the formative and informative process. The first and second exercises have already been done in two schools.

The first exercise named "What is Data Science?" was made in two schools and led the kids to solve a problem of finding new friends in a new school. The practice introduced some data science concepts by giving the students autonomy to manipulate data and extract useful information. After the problem being introduced, they were challenged to collect all the data needed: they had to interview five girls, asking about their names, favorite subject on school, favorite sport, if they like to dance and any question that considered a particular characteristic they judge to be a decision factor for choosing a new friend. So, they had to construct a table and then a bar chart (see Figure 1). A simple code and graph will be introduced as shown is this figure.

The second exercise was about big data. The students were proposed to separate confettis by color, to represent it by frequence (see Figure 2) and think about some concepts about sample and inference. They were encouraged to think about how could them deal with millions of data, comparing it with millions confettis in a production and which color were most common in the package. This answers were connected with social media, that produce millions of data by seconds and how their preferences in this media are manipulated by someone. On the next steps we expect them to be able to work the computational and statistical thinking using R, for that, some types of graphs will be used.
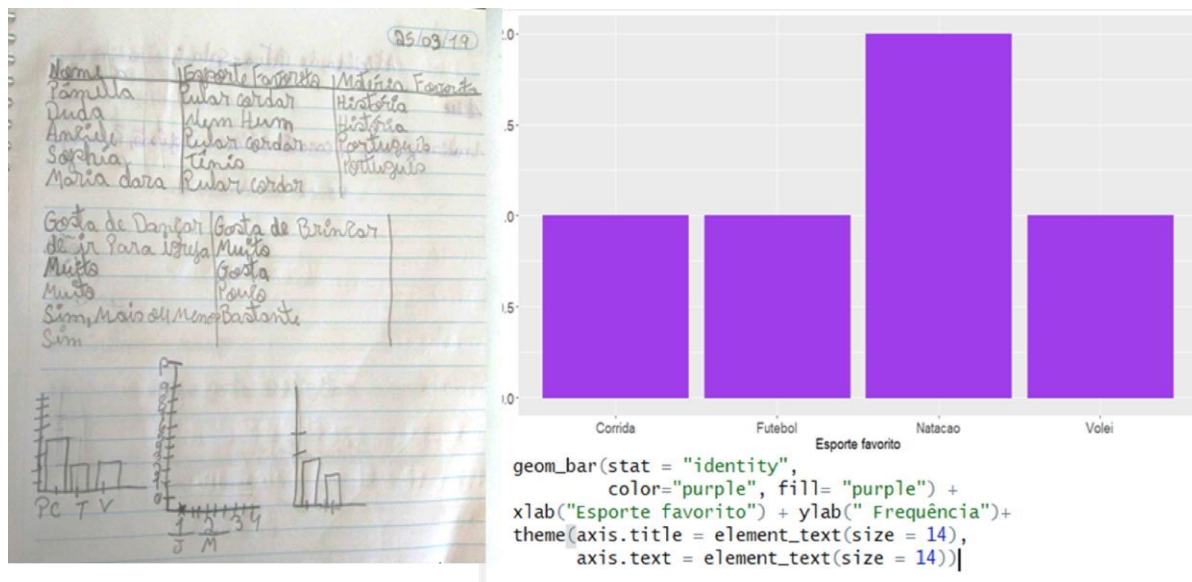

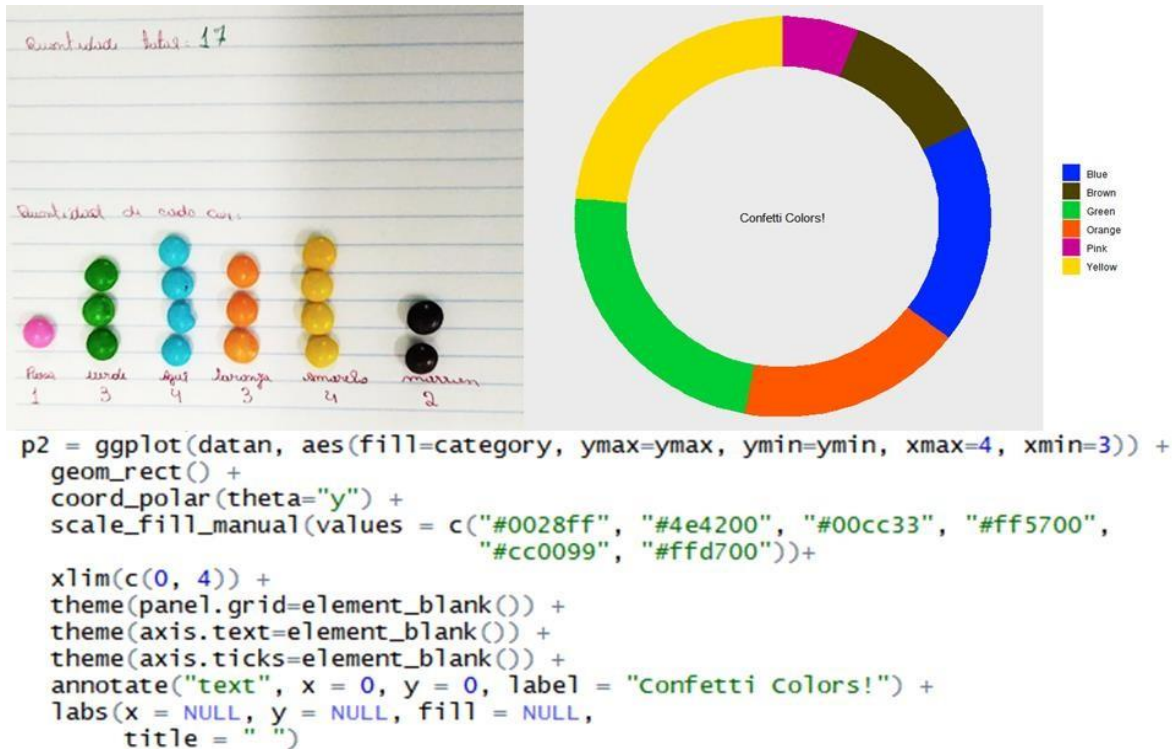**Figure 1-** First Class - Data Science

```
p2 = ggplot(datan, aes(fill=category, ymax=ymax, ymin=ymin, xmax=4, xmin=3)) +
    geom_rect() +
    coord_polar(theta="y") +
    scale_fill_manual(values = c("#0028ff", "#4e4200", "#00cc33", "#ff5700",
                                 "#cc0099", "#ffd700"))+
    xlim(c(0, 4)) +
    theme(panel.grid=element_blank()) +
    theme(axis.text=element_blank()) +
    theme(axis.ticks=element_blank()) +
    annotate("text", x = 0, y = 0, label = "Confetti Colors!") +
    labs(x = NULL, y = NULL, fill = NULL,
         title = " ")
```

Figure 2 - Second class - Big Data

Conclusion

In the first stage, it was possible to notice that girls do not have identification with STEM, as stated by the literature. Through the PBL technique applied in second stage, the girls interest on data analysis could be seeded. Although building graphs intuitively, at the end of the activities they worked using the main steps of working with data: Defining a problem, collecting data, processing data, exploring and analysing data, showing the results and, of course, taking a decision. In the the next step they will learn how to implement this solution with a simple code in R and to build other kids of graphs as well. All project activities, from planning to action, are also contributing to the approximation between university public secondary schools and since undergraduate, graduate and professors.

Acknowledge

References

AGGARWAL, V., SRIKANT, S. Introducing Data Science to School Kids. In: Avaliable in: < https://bit.ly/2yd8V8z> Accessed in 27 de Agosto de 2018.

DAR, P. 19 Inspiring Women Blazing a Trail in the Data Science World. Avaliable In: <https://bit.ly/2G7vYVJ> Accessed in 28 de Agosto de 2018.

MADE WITH CODE. Why coding is a big deal. Avaliable In: <https://bit.ly/1QM3iQr> Accessed in 29 de Agosto de 2018.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.