Intermediate Practice

1st Semester 2020

1. Instructions

- Practice is individual.
- It is due by Friday May 8, 2020.
- You must write a report (in Spanish) and some Python code.
- The report (a Word file) must describe the analysis of the data and its statistics. Do not describe how you implemented your code but rather describe with your own words what perspectives and/or insights can you extract from the analysis of the statistics. For example: what programming language has a better salary? What developer job is more popular? How many women are registered as back-end developers? Etc.
- The code must be in Python. Create a practice.py file and inside it, write a function for every one of the tasks mentioned in the description (see below).
- Create a public GitHub repository and upload your code there.
- Upload the link of the repository to Google Classroom.

2. Description

There is a CSV file named *survey_results_public.csv* (compressed as *survey_results_public.rar*) that contains the data of a survey collected from users of Stack OverFlow in 2019 concerning the following ten variables (the name of the fields in the file are between parenthesis): country (Country), educational level (EdLevel), developer type (DevType), years of experience with coding (YearsCode), annual salary in US dollars (ConvertedComp), average number of working hours per week (WorkWeekHrs), programming language he/she has experience with (LanguageWorkedWith), age (Age), gender (Gender) and ethnicity (Ethnicity). There are data for 88,883 users.

For some variables, the users could respond with more than one answer, with the answers separated with a ; in the file. For example, in programming language he/she has experience with, one user could select at the same time C; C++; JavaScript; Python. In that case, for the statistics, the same user will count for every language he/she chooses. The same applies for any other variable that allows multiple answers.

In other cases, the users could omit one or several answers, and in the file, we can find NaN values or empty string values. In that case, for the statistics, those values must be ignored.

The practice consists of the following small tasks of processing and analyzing of the data contained in the file. For each task, you must write a python function as part of the pratice.py code file.

1. Compute the five-number summary, the boxplot, the mean, and the standard deviation for the annual salary per gender.
2. Compute the five-number summary, the boxplot, the mean, and the standard deviation for the annual salary per ethnicity.
3. Compute the five-number summary, the boxplot, the mean, and the standard deviation for the annual salary per developer type.
4. Compute the median, mean and standard deviation of the annual salary per country.
5. Obtain a bar plot with the frequencies of responses for each developer type.
6. Plot histograms with 10 bins for the years of experience with coding per gender.
7. Plot histograms with 10 bins for the average number of working hours per week, per developer type.
8. Plot histograms with 10 bins for the age per gender.
9. Compute the median, mean and standard deviation of the age per programming language.
10. Compute the correlation between years of experience and annual salary.
11. Compute the correlation between the age and the annual salary.
12. Compute the correlation between educational level and annual salary. In this case, replace the string of the educational level by an ordinal index (e.g. Primary/elementary school = 1, Secondary school = 2, and so on).
13. Obtain a bar plot with the frequencies of the different programming languages.