Individual Assignment

Semester August-December 2021

1. Instructions

- Assignment is individual.
- It is due by Friday November 19, 2021, at midnight.
- You must write a report (in Spanish) and some Python code.
- The report (a PDF file) must describe the analysis of the data and its statistics. DO NOT DESCRIBE HOY YOU IMPLEMENTED YOUR CODE OR ANY ROUTINE, FUNCTION OR LIBRARY YOU USED. **Describe with your own words what perspectives and/or insights can you extract from the analysis of the statistics and for answering the required questions.**
- Name the PDF file as **assignment.pdf** and upload it to Teams.
- Create a Python script for each point, and save it with the corresponding name **assignment_01.py**, **assignment_02.py**, **assignment_03.py**, **assignment_04.py**, **assignment_05.py**

1. Upload all the files directly to Teams. You must upload 6 files.

2. Description

Attached is a CSV file named *survey_results.csv* (compressed as *survey_results.rar*) that contains the data of a survey collected from users of Stack OverFlow in 2019. The data concerns 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 36,736 users.

For the variables **DevType**, **LanguageWorkedWith**, **Gender** and **Ethnicity**, the users could respond with more than one answer, with the answers separated with a ; in the file. For example, in the variable **LanguageWorkedWith**, 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**. Then, you must split the answer of each user. In the example, the answer from the user will count for C, C++, JavaScript, and Python. The same applies for the other variables that allows multiple answers.

For the variable **EdLevel**, if necessary, consider the academic degrees in the following increasing order: 0) I never completed any formal education; 1) Primary/elementary school; 2) Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.); 3) Some college/university study without earning a degree; 4) Associate degree; 5) Bachelor's

degree (BA, BS, B.Eng., etc.); 6) Master's degree (MA, MS, M.Eng., MBA, etc.); 7) Professional degree (JD, MD, etc.); 8) Other doctoral degree (Ph.D, Ed.D., etc.).

The practice consists of the 5 points described below about processing and analyzing the data contained in the file. For each point, you must write a python script and save it as **assignment_01.py**, **assignment_02.py**, … and so on. In each file, copy the sentence of the general problem as a comment in the beginning, create a function for each subsection (a, b, c, and so on) and copy the sentence of each one as a comment for the function.

In the report write the sentence for the point and subsection you are analyzing. **DO NOT ANSWER DIRECTLY THE QUESTION, REPORT THE STATISTICS, METRICS, PLOTS AND DISTRIBUTIONS NECESSARY TO EXPLAIN THE QUESTION, AND DESCRIBE WITH YOUR WORDS WHAT YOUR FINDINGS, INSIGHTS AND CONCLUSIONS ARE ABOUT THE QUESTION.**

Remember, the statistics you can use to analyze the data include frequencies/percentages, five-number summary, mean, variance, standard deviation, Pearson's correlation coefficient, odds ratio, Pearson's phi coefficient, Pearson's chi square test, point biserial correlation and conditional probability. You can also plot the data and statistics using bar/pie plots, scatter plots, boxplots, and histograms. And you can find outliers in the data and/or trim the data to eliminate extreme values.

1. Compute the five-number summary, the mean, and the standard deviation for the annual salary per gender (the data considers three genders, you must compute the statistics for each one) and draw the boxplot.

   Perform the computations with the original data and with the trimmed data at 10% for the salary (you must cut the 10% lowest salaries and 10% highest salaries). Make comparisons between the results with the original data and the ones with the trimmed data.

   Besides, try to give and explanation for the following questions by using the trimmed data at 10% for the salary and by computing all the additional and necessary statistics and drawing the necessary graphs/plots.

   a. Which gender has more answers?
   b. Which gender tends to have higher salaries, and which one tends to have lower salaries?
   c. What are the most popular and less popular programming language per gender?
   d. What are the most popular and less popular developer type per gender?
   e. Is there a relation between gender and salary? (Consider only the genders man/woman)
   f. Is there a relation between gender and age? (Consider only the genders man/woman)

g. Is there a relation between gender and years of experience? (Consider only the genders man/woman)
h. Is there a relation between gender and developer type?
i. Is there relation between gender and the educational level (**EdLevel**)? First consider all the possible values for **EdLevel** and compute the relation. Second, transform **EdLevel** to a binary variable with values: 'higher', and 'nonhigher', and compute the relation. The 'higher' value corresponds to degrees between bachelor and doctorate, the 'nonhigher' to any other value.

2. Compute the five-number summary, the mean, and the standard deviation for the annual salary and draw the boxplots for the 4 most common ethnicities. You first need to find the 4 ethnicities with more answers and compute the statistics for each one.

   Perform the computations with the original data and with the trimmed data at 10% for the salary (you must cut the 10% lowest salaries and 10% highest salaries). Make comparisons between the results with the original data and the ones with the trimmed data.

   Besides, try to give and explanation for the following questions, using the trimmed data at 10% for the salary, only considering the 4 most common ethnicities, and by computing all the additional and necessary statistics and drawing the necessary graphs/plots.

   a. Which ethnicity has more answers?
   b. Which ethnicity tends to have higher salaries, and which one tends to have lower salaries?
   c. What are the most popular and less popular programming language per ethnicity?
   d. What are the most popular and less popular developer type per ethnicity?
   e. Is there a relation between ethnicity and salary? For that, transform the variable **Ethnicity** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular ethnicities, the first pair would be: 'White or of European descent' and 'other'. The second pair would be: 'Hispanic or Latino/Latina' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
   f. Is there a relation between ethnicity and years of experience? For that, transform the variable **Ethnicity** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular ethnicities, the first pair would be: 'White or of European descent' and 'other'. The second pair would be: 'Hispanic or Latino/Latina' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
   g. Is there a relation between ethnicity and developer type?
   h. Is there relation between ethnicity and the educational level (**EdLevel**)? First consider all the possible values for **EdLevel** and compute the relation. Second, transform **EdLevel** to a binary variable with values: 'higher', and 'nonhigher', and compute the relation. The 'higher' value corresponds to degrees between bachelor and doctorate, the 'nonhigher' to any other value.

3. Compute the five-number summary, the mean, and the standard deviation for the annual salary and draw the boxplots for the 4 most common countries. You first need to find the 4 countries with more answers and compute the statistics for each one.

   Perform the computations with the original data and with the trimmed data at 10% for the salary (you must cut the 10% lowest salaries and 10% highest salaries). Make comparisons between the results with the original data and the ones with the trimmed data.

   Besides, try to give and explanation for the following questions, using the trimmed data at 10% for the salary, only considering the 4 most common countries, and by computing all the additional and necessary statistics and drawing the necessary graphs/plots.

   a. Which country has more answers?
   b. Which country tends to have higher salaries, and which one tends to have lower salaries?
   c. What are the most popular and less popular programming language per country?
   d. What are the most popular and less popular developer type per country?
   e. Is there a relation between country and salary? For that, transform the variable **Country** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular countries, the first pair would be: 'United States' and 'other'. The second pair would be: 'Germany' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
   f. Is there a relation between country and years of experience? For that, transform the variable **Country** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular countries, the first pair would be: 'United States' and 'other'. The second pair would be: 'Germany' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
   g. Is there a relation between country and developer type?
   h. Is there relation between country and the educational level (**EdLevel**)? First consider all the possible values for **EdLevel** and compute the relation. Second, transform **EdLevel** to a binary variable with values: 'higher', and 'nonhigher', and compute the relation. The 'higher' value corresponds to degrees between bachelor and doctorate, the 'nonhigher' to any other value.

4. Compute the five-number summary, the mean, and the standard deviation for the annual salary and draw the boxplots for the 4 most common programming languages. You first need to find the 4 programming languages with more answers and compute the statistics for each one.

   Perform the computations with the original data and with the trimmed data at 10% for the salary (you must cut the 10% lowest salaries and 10% highest salaries). Make comparisons between the results with the original data and the ones with the trimmed data.

   Besides, try to give and explanation for the following questions, using the trimmed data at 10% for the salary, only considering the 4 most common programming languages,

and by computing all the additional and necessary statistics and drawing the necessary graphs/plots.

a. Which programming language has more answers?
b. Which programming language tends to have higher salaries, and which one tends to have lower salaries?
c. What are the most popular and less popular developer type per programming language?
d. Is there a relation between programming language and salary? For that, transform the variable **LanguageWorkedWith** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular programming languages, the first pair would be: 'C' and 'other'. The second pair would be: 'Java' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
e. Is there a relation between programming language and years of experience? For that, transform the variable **LanguageWorkedWith** to binary, using four pairs of values one-vs-allothers. For example, for the four most popular programming languages, the first pair would be: 'C' and 'other'. The second pair would be: 'Java' and 'other, and so on with the other pairs. For each pair you must compute the relation between variables.
f. Is there a relation between programming language and developer type?
g. Is there relation between programming language and the educational level (**EdLevel**)? First consider all the possible values for **EdLevel** and compute the relation. Second, transform **EdLevel** to a binary variable with values: 'higher', and 'nonhigher', and compute the relation. The 'higher' value corresponds to degrees between bachelor and doctorate, the 'nonhigher' to any other value.

5. Try to answer the following questions using the original data and the trimmed data at 10% for the salary, and make comparisons:

a. Is there a relation between the educational level and the salary (the higher the education, the higher the salary)?
b. Is there a relation between the age and the salary (the higher the age, the higher the salary)?
c. Is there a relation between the years of experience and the salary (the higher the experience, the higher the salary)?