A picture containing shape

Description automatically generated

Universidad Politécnica de Valencia

Escuela Técnica Superior de Ingeniería Informática

Statistical Study of the

StackOverflow’s Survey 2022

for the Iberian Peninsula.

Logo, company name

Description automatically generated

Author: Rodríguez Díaz, Gabriel  
Tutor: Zarzo Castelló, Manuel

**Introduction**

**Previous Concepts:**

* StackOverflow: This website is one of the most iconic forums for programming related questions and problem solving.

Every year, a survey is carried out in order to study the status of the IT environment.

* Blockchain: Is a technology that allows the possibility of creating shared, immutable ledgers that ensures the veracity and reliability of the transactions.

Its usage is controversial among the developers’ community because it is the core technology for other polemic concepts such as cryptocurrencies or NFTs.

* Operative System: The operative system is the core of a computer. It is the tool that allows people to work easily with computer.

That is why most of the developers have a preference between the Big Three, Linux, Apple and Windows.

**About the dataset:**

For this assignment the dataset chosen was the one providen by the

[2022 Developer Survey](https://survey.stackoverflow.co/2022/). You can get the file from [Kaggle](https://www.kaggle.com/datasets/dheemanthbhat/stack-overflow-annual-developer-survey-2022).

In the original dataset, there are several aspects collected. Some are work related and others are more personal. Nevertheless, in this project we are focusing our attention on an array of variables that may be interesting to analyze and compare.

Summarizing, in the original dataset, there are 78 variables, of which we are using 4 continuous and 4 cualitatives.

The following table shows the variables, its type and a brief description.

|  |  |
| --- | --- |
| **Cuantitative Variables** | |
| **Variable** | **Description** |
| Years of Code  (**X1**) | The years that a person has been coding. It can be useful to know how much years of studying code it takes for someone before they start coding. |
| Years of Code Pro (**X2**) | The years that a person has been coding with some kind of remuneration. |
| Yearly Salary (**X3**) | The amount in **euros** that the surveyed people perceive a **year**. |
| Age (**X4**) | The range of age of each surveyed person. |

TABLE 1. Numerical variables table.

|  |  |
| --- | --- |
| **Cualitative Variables** | |
| **Variable** | **Description** |
| Country (**F1**) | One of the two countries picked for the assignment: **Spain** or **Portugal** |
| Education (**F2**) | The alevel of studies that the person has. |
| OS used for work (**F3**) | Which operative system does the person use to carry out its work. |
| Blockchain opinion (**F4**) | The opinion of each surveyed person about the blockchain technology. |

TABLE 2. Cualitative variables table.

Once we have the cualitative variables, it is time to qualify some things about the variables.

* **Age:** The range of eligible ages is the following. For the sake of simplicity, the option “prefer not to say” has been excluded since there was only one record matching that option.

|  |
| --- |
| **Ranges (Years)** |
| [18,24] |
| [25,34] |
| [35,44] |
| [45,54] |
| [55,64] |

TABLE 3. Ranges of variable Age

* **Years of Code Pro:** In the dataset, there was a variable called “Years Of Experience”. Nevertheless, the survey takes only people who are developers or write code as part of their work, then, we may considere this variable as a representative of the professional experience.
* **Education:**  This variables referes to the level of study of each person. The following table indicates the possible options.

|  |  |
| --- | --- |
| **Education Level (Title)** | **Equivalent in Spanish** |
| Primary/Secondary/None/Something Else | Primaria/Secundaria/Ninguno /Otros |
| Professional degree (JD, MD, etc.) | FP Grado Medio |
| Associate degree (A.A., A.S., etc.) | FP Grado Superior |
| Some college/university study without earning a degree | Estudios de grado sin terminar. |
| Bachelor’s degree (B.A., B.S., B.Eng., etc.) | Grado Universitario |
| Master’s degree (M.A., M.S., M.Eng., MBA, etc.) | Estudios Posgrado |
| Other doctoral degree (Ph.D., Ed.D., etc.) | Doctorado |

TABLE 4. Education Levels and its equivalent in the Spanish System. Where Bachelor´s Degree, Master´s Degree and Doctoral degree are considered high level degrees.

In the dataset there are individual options for Primary, Secondary, None and Something else, but we are grouping them since they are not so relevant and do not teach code.

* **OS used for work:** The options were merged into three options:

|  |
| --- |
| **Operative System** |
| Linux Based |
| Windows |
| Apple |

TABLE 5. OS groups for the variable Operative System used for work.

Without entering in much detail, we are considering the subsystem terminals for Windows, Windows Server, other Linux based systems as Android, BSD; or IOS and MacOS for Apple.

* **Blockchain Opinion:** They are classified in a simple way:

|  |
| --- |
| **Blockchain Opinion** |
| Very Favorable |
| Favorable |
| Neutral |
| Unfavorable |
| Very Unfavorable |
| Unsure |

TABLE 6. Level of support for the blockchain technology.

**Objectives:**

* To study if there exists a relationship between the level of education and the income among the developers.
* To study the increase of the income based on the growth of the working experience.
* To Analyze if there exists a relationship between the used level of education and the opinion about blockchain.
* To review which of the countries has the most people with a Bachelor´s Degree or superior level of education.
* To get an idea of the current aging state of the laboral market in IT.
* To check if there exists a relationship between the salary and the OS used.

**Discussion about the sample and populations:**

The population, as it was said before, was all the people that filled the StackOverflow Developers´ Survey 2022 that is 73268 people.

However, for this assignment, I took a sample that implements the following filters:

* Only the people from Iberian Peninsula (Excluding Andorra, that had only 15 records). This left the sample with 2084 of 73268 records. The dataset was reduced in a 97.155%
* The ones that receive their paycheck in euros. This filtered 501 more people.
* The options “NA” and “Prefer not to say” in Age were removed, then 13 records were removed.
* Excluded the people who did not mark their paycheck as yearly. Based on preliminary observations, this filter will reduce the outliars substantially since this option lead many people to error. After this filter 609 records were discarded.

This filters were applied in the order as it is mentioned above.

That left us with 961 records. Shrinking the size of the dataset in a 98.688%

Even with the filters applied, the sample that was taken contains plenty of useful information, nonetheless, for this assignment, only the previous mentioned variables were choosen.

**Descriptive Statistics**