Universidad Politécnica de Valencia

Escuela Técnica Superior de Ingeniería Informática

STATISTICAL STUDY OF THE STACKOVERFLOW'S SURVEY 2022 FOR THE IBERIAN PENINSULA.

Author: Rodríguez Díaz, Gabriel
Tutor: Zarzo Castelló, Manuel
Group: 1E2

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

2. About the dataset:

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

2022 Developer Survey. You can get the file from Kaggle.

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 cuantitative 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 many years of studying code it takes before someone starts working.
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 level of studies that the person has.
OS used for work (**F3**)	Which operative system does the person use to carry out their work.
Blockchain opinion (**F4**)	The opinion of each surveyed person about the blockchain technology.

TABLE 2. Cualitative variables table.

Now 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 four options:

Operative System
Linux Based
Windows
Apple
Combination

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 as Windows, other Linux based systems as Android, BSD would be Linux; and IOS and MacOS are 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.

3. 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 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.

4. Discussion about the sample and populations:

The population, as it was said before, is the entire Software Development Industry. Since StackOverflow is a huge place where developers around the world discuss about IT, and this survey was filled by 73268 people from any country and older than 18 years old. This dataset is quite representative of the IT industry.

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. Also the option NA in YearOfCode since everyone who answered were supposed to be programmers. These removed 1 record
- 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.
- Finally, we are only taking into account those people who use only one Operative System at their work. Therefore, this will leave us with 495 records.

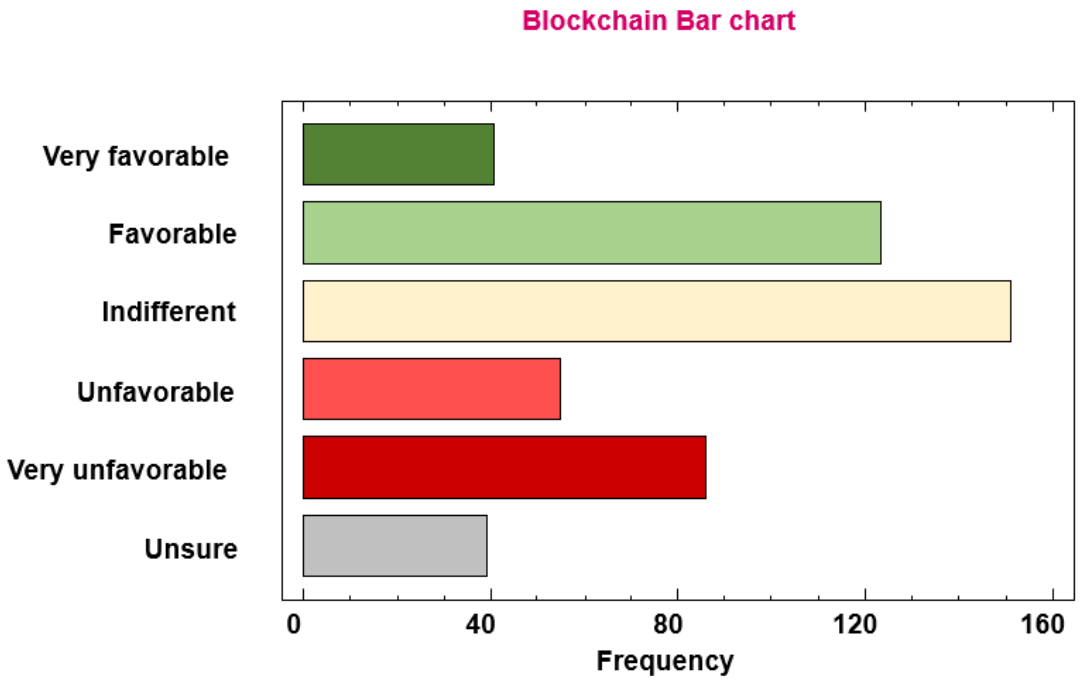
These filters were applied in the order as it is mentioned above. So, we are shrinking the size of the dataset in a 99.324%

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.

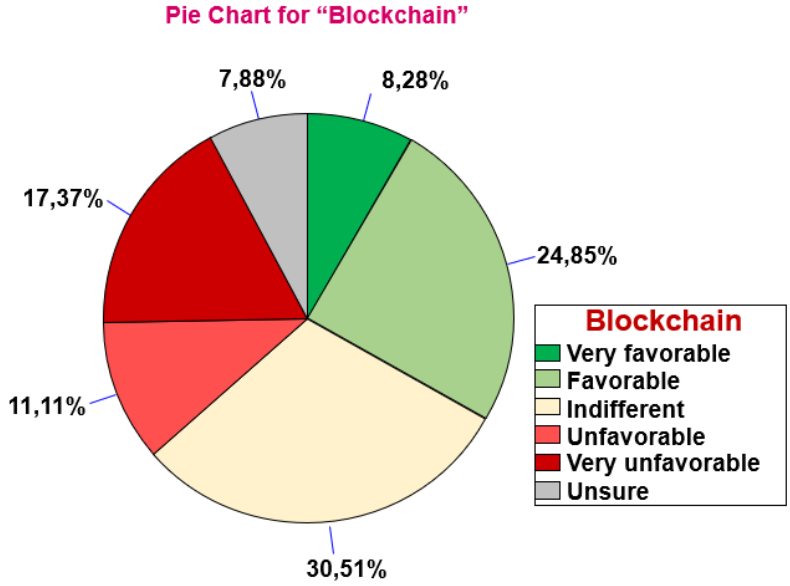
In summary: The current population consists in developers older than 18 years old who live in Spain or Portugal among other characteristics mentioned in the above filters.

Descriptive Statistics

5.1 Represent the bar chart and the pie graph of the variable Blockchain:



PICTURE 1: Bar chart for "Blockchain"



PICTURE 3: Pie chart for "Blockchain"

5.2 Do the categories have a similar frequency?

As we may see in both charts, there are two categories that clearly excel. Those are "favorable" and "indifferent". After those two, we can notice that people are also very unfavorable about blockchain.

Since it is a new technology and kinda hard to understand, there are many people indifferent or unsure about it.

Also, as we mentioned in the definition of the concept, it is very controvertial, which also explains that frequency in an opinion as extreme as very unfavourable.

6.1 Compute the frequency table for the variable Blockchain:

Clase	Valor	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	Favorable	123	0,2485	123	0,2485
2	**Indifferent**	**151**	**0,3051**	**274**	**0,5535**
3	Unfavorable	55	0,1111	329	0,6646
4	Unsure	39	0,0788	368	0,7434
5	Very Favorable	41	0,0828	409	0,8263
6	Very Unfavorable	86	0,1737	495	1,0000

TABLE 7. Frequency table for Blockchain

6.2 What is computed in each of the columns?

Frequency: Or absolute frequency. Shows the number of times each category of Blockchain occurred.

Relative Frequency: Indicates over 1 the frequency for each value in relation with the total of occurrences for every category.

Cumulative Frequency: The current measured frequency taking into account the previous mentioned frequencies.

Cumulative Relative Frequency: The relative frequency with a cumulative criterion.

6.3 Discuss the most relevant results:

It is important to mention that the variable chosen was Blockchain because it is more interesting to analyze than Country.

As we can observe, Indifferent is clearly the dominant category. Its frequency comprehends almost ⅓ of the total. We could combine it with Unsure leaving us with a total of 0,3839 points

We can also discuss the other results in two subgroups which are favorable opinions (Favorable and Very Favorable) and unfavorable opinions (Unfavorable and Very Unfavorable). As it is shown, the favorable opinions have 0,3313 points while the unfavorable group has 0,2848.

Therefore, we can conclude that the vast majority of the people do not have a well-formed opinion about blockchain, and, between those who have, the tendency shows that it has more supporters than detractors.

7.1 Compute a table of crossed frequencies between Blockchain and Country.

	Very Favorable	Favorable	Indifferent	Unfavorable	Very Unfavorable	Unsure	Row Total
Portugal	11	23	29	8	14	7	92
	2,22%	4,65%	5,86%	1,62%	2,83%	2,22%	18,59%
	11,96%	25,00%	31,52%	8,70%	15,22%	7,61%	
Spain	30	100	122	47	72	32	403
	6,06%	20,20%	24,65%	9,49%	14,55%	6,06%	81,41%
	7,44%	24,81%	30,27%	11,66%	17,87%	7,94%	
Column Total	41	123	151	55	86	39	495
	8,28%	24,85%	30,51%	11,11%	17,37%	7,88%	100,00%

TABLE 8. Cross Frequency table for Blockchain and Country

7.2 “Row Percentages” or “Column Percentages”?

In this case, the row percentages option was choosen, since the variable Country has much less categories than Blockchain. Also, it is kind of interesting to know the percentage for each country.

7.3 Explain the difference between Absolute and Relative Frequencies:

Absolute frequency: Indicates the number of time that a value appears. The sum of all absolute frequencies is the total number of observations.

Relative frequency: It is the absolute frequency divided by the total number of values in the data set. It returns a value in the range of 0 to 1. Also, the sum of all the relative frequencies equals to 1.

7.4 Explain the difference between Marginal and Conditional Frequencies:

Marginal frequency: Those that represent the total frequencies of each value of the variable.

Conditional frequency: Are the ones computed based on the values of other variable.

In the case of the **TABLE 8**, the values **marked** represent the conditional frequency of Very Favorable with respect to Portugal. That is, 2,22% of the data that is from Portugal, has Very Favorable as option.

On the other hand, in "Column Total" and "Row Total" we represent the marginal frequencies of each variant with respect to the total amount of observations.

7.5 Is there any relation between Blockchain and Country?

As we can see in the **TABLE 8**, the percentage of the favorable opinions is higher in Portugal. While, in Spain, the percentages of the unfavourable opinions tend to be quite higher than in Portugal.

Also, the values of indifferent and unsure are more or less the same in both countries.

The relationship indicates that people located in Portugal tend to have a better view of the blockchain technology than the ones in Spain.

It could be interesting to remark that Portugal´s government supported the cryptocurrencies in the past, while the Spanish government and the Spanish CNMV criticized the industry and even threatened some blockchain enterprises.

This may be an interesting approach of why there is a correlation between both variables.

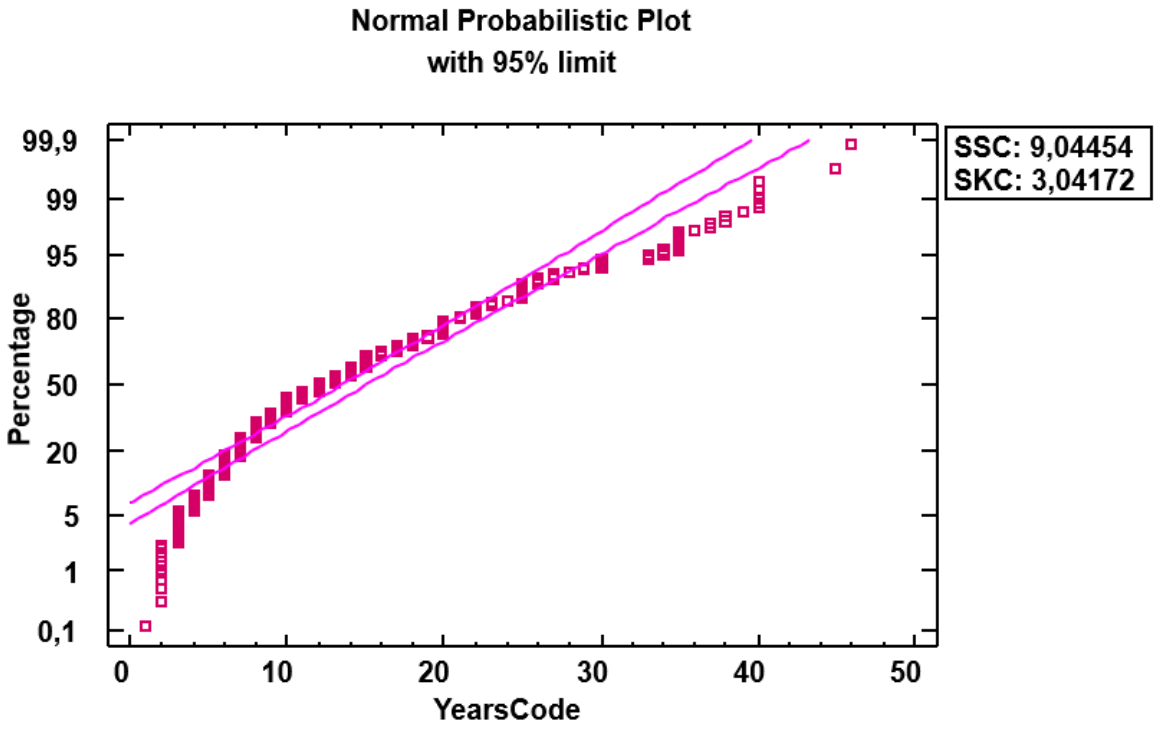
8. Compute a table with the main statistics for each of the 4 cuantitative variables:

Parameter	YearsCode	YearsCodePro	Salary	AvgAge	Type
Range	45	37	200000	38,5	Dispersion
Inter. Range	13	10	29000	10	Dispersion
Average	14,2389	9,81081	49722,8	34,3566	Position
Median	12	8	44500	29,5	Position
Variance %	61,8641	74,5074	58,0181	24,6291	Dispersion
Standard Deviation	8,80874	7,30978	28848	8,46173	Dispersion
S. Skewness Coefficient	9,04454	9,92841	16,6149	4,923	Shape
S. Kurtosis Coefficient	3,04172	4,1722	23,8612	0,69144	Shape

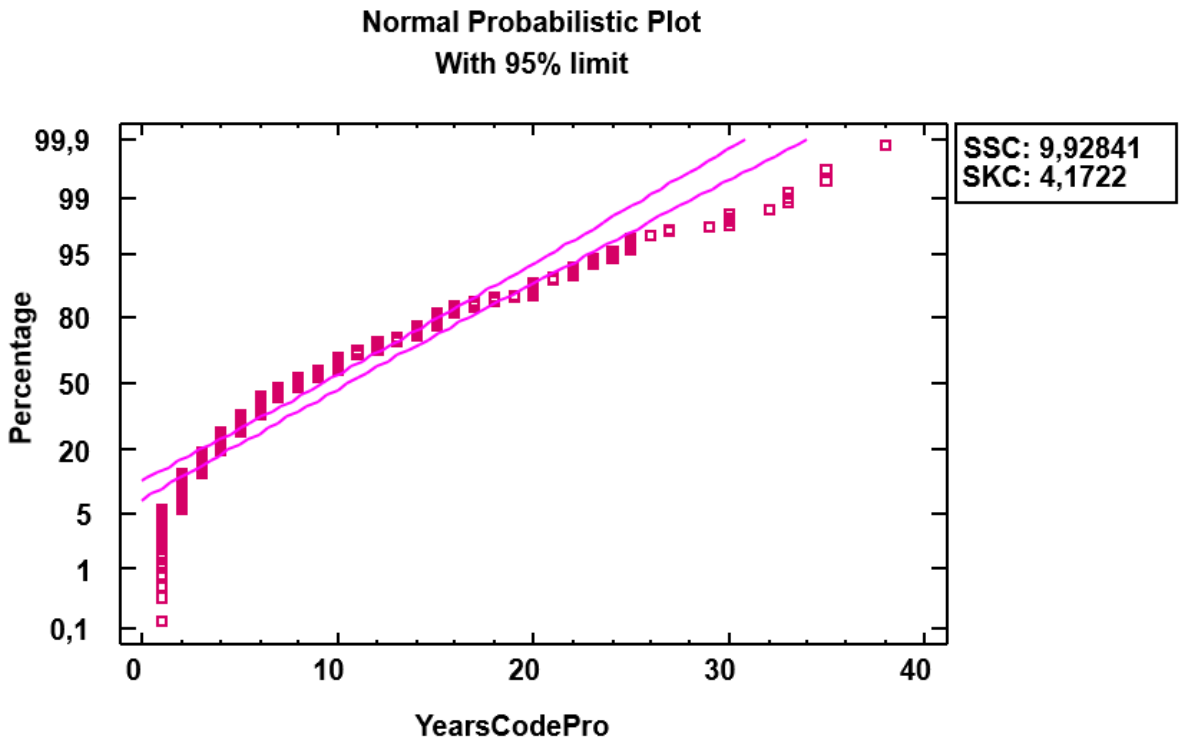
TABLE 9. Analysis of the cuantitative variables

Note that for this analysis I have used a different variable, that is the average of the ages. I replaced the ranges by the average between the minimum and the maximum. For example, for the range [25 ; 34] its value would be 29,5

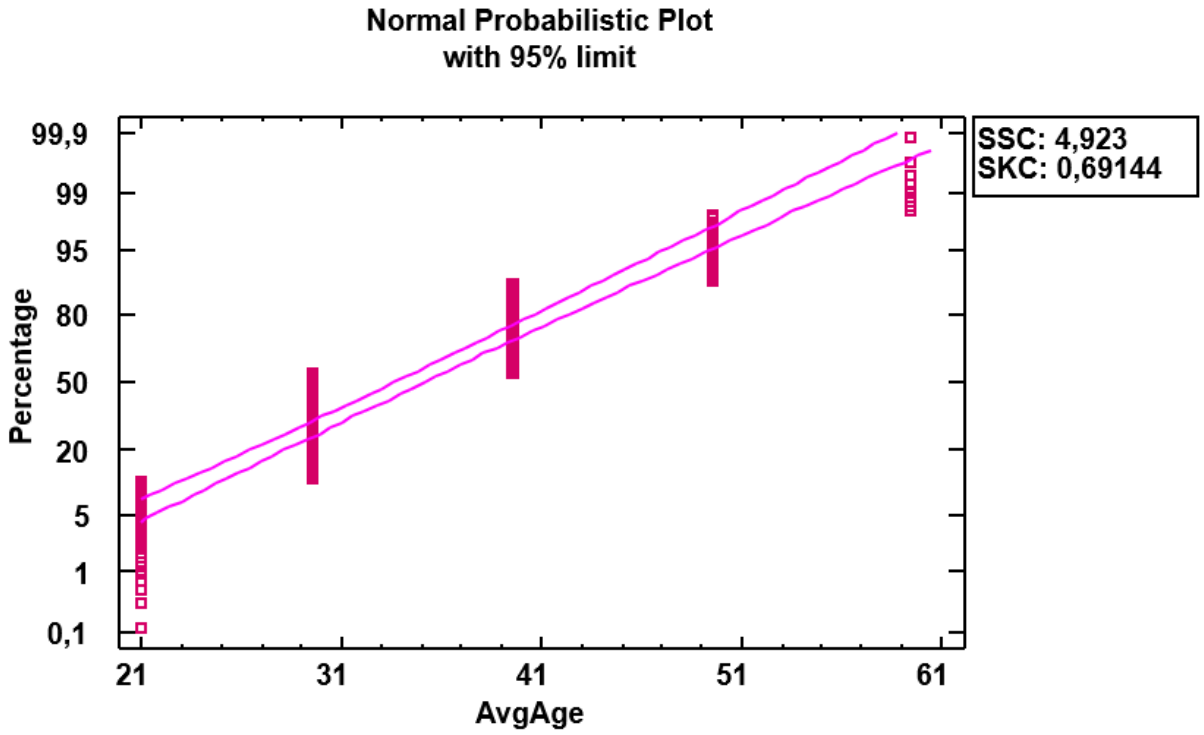
9. Create a normal probabilistic plot and using it along with the Standard Skewness Coefficient and the Standard Kurtosis Coefficient indicate which variable can be taken as X1:



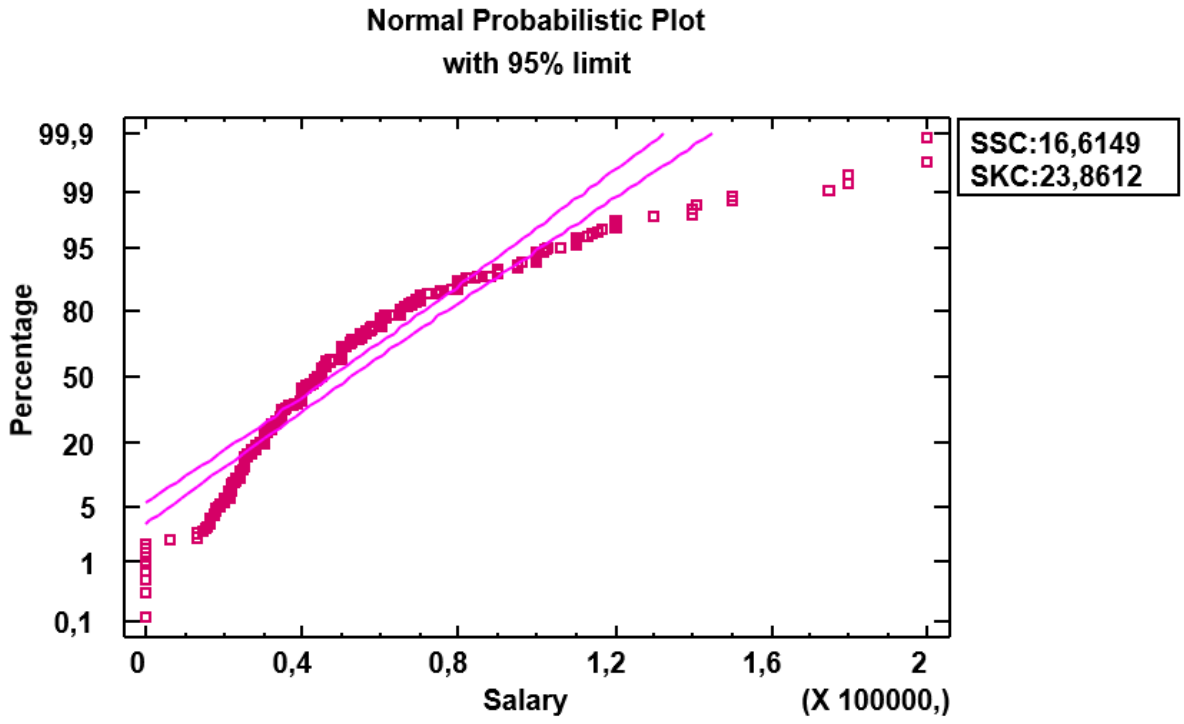
PICTURE 4: Normal Probabilistic Plot for YearsCode



PICTURE 5: Normal Probabilistic Plot for YearsCodePro



PICTURE 6: Normal Probabilistic Plot for AvgAge



PICTURE 7: Normal Probabilistic Plot for Salary

Firstable, we should briefly define the Standard Skewness Cofficient and the Standard Kurtosis Coefficient, since those are the parameters we are considering to choose our variable.

- **Standard Skewness Coefficient:** Is a measure for the skewness of the distribution. Being 0 a perfect symmetry.
- **Standard Kurtosis Coefficient:** Is a measure that indicates the level of -peakedness of a distribution.

To choose a variable based on this criteria, we should look for a variable that has the most coefficients closer to zero. Both Skewness and Kurtosis.

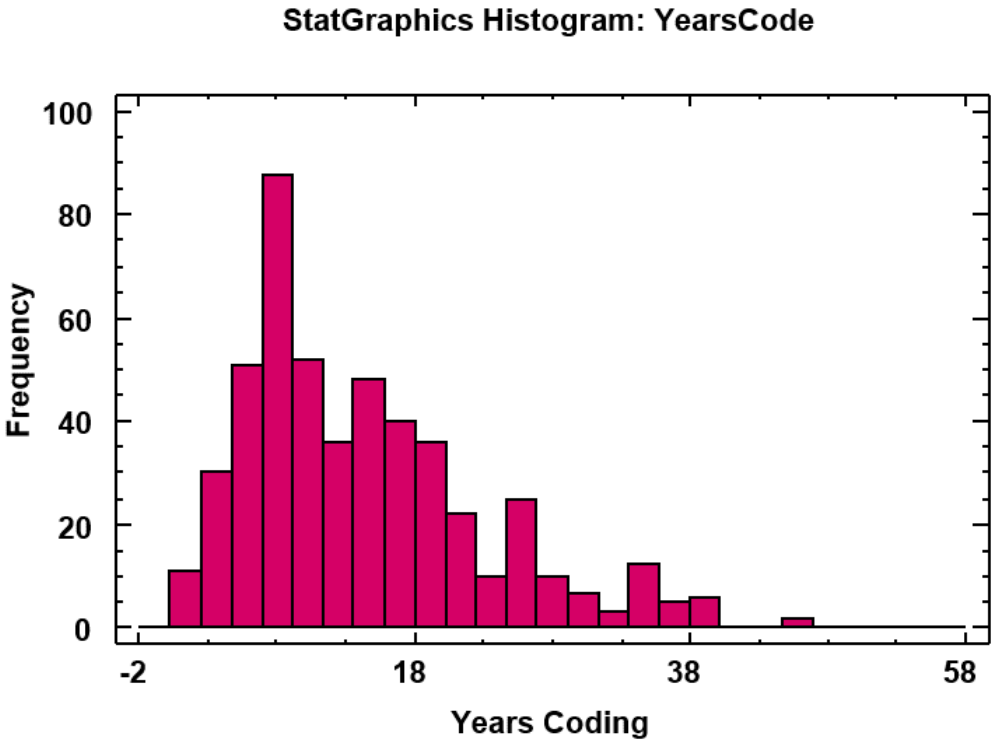
This indicates that the distribution is quite symmetrical and less peaked, which makes it easier to perform statistical computations and interpretate the results.

According to the previous criteria the choosen variable would be AvgAge. Nevertheless, it is important to considerate that it is more like a discrete variable, then we have to take YearsCode.

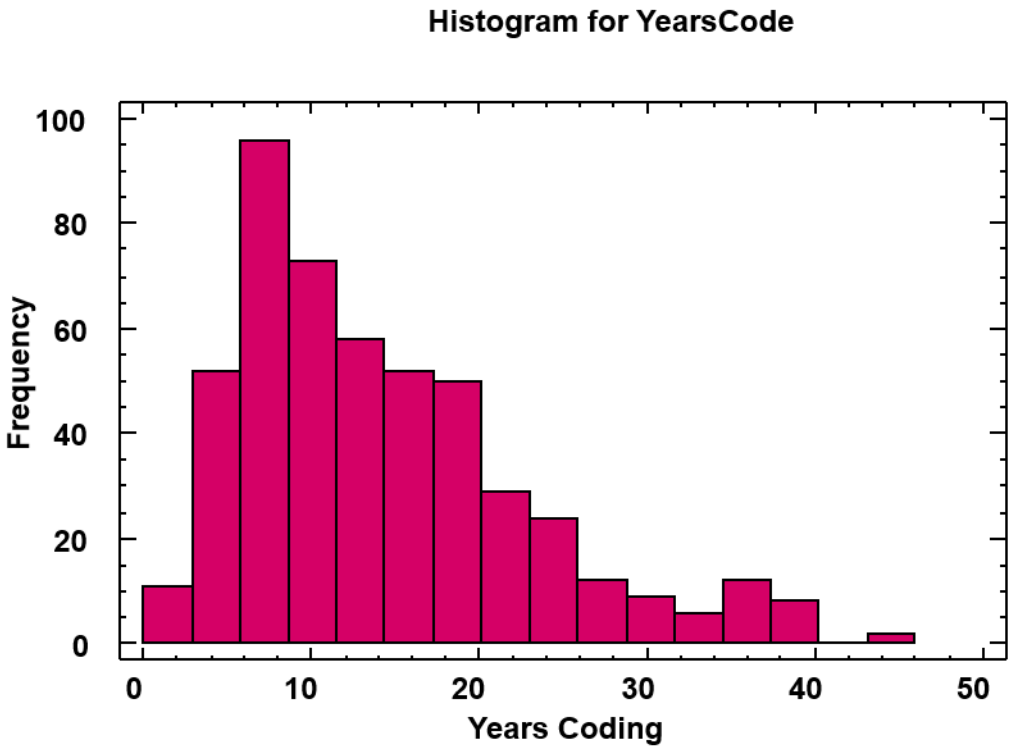
10. Place a histogram for each of the variables. If needed, you may change the intervals arguing why:

YearsCode (Pictures 8-9): In this case the Histogram created by StatGraphics has 27 classes and the limit is set in [-2,58]. I changed it to 16 classes with a limit in 46 starting in zero.

We don´t need a lower limit different from zero and 28 classes is so much. We are looking for less peakness and more symmetry, so 16 classes is fine.

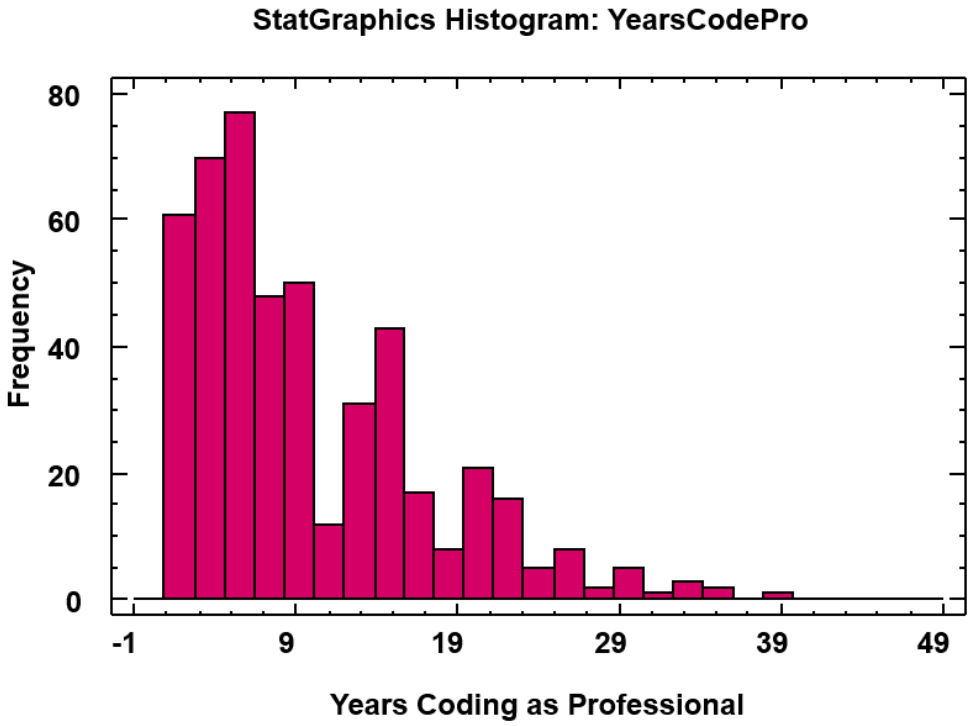


PICTURE 8: Statgraphics Histogram for YearsCode.

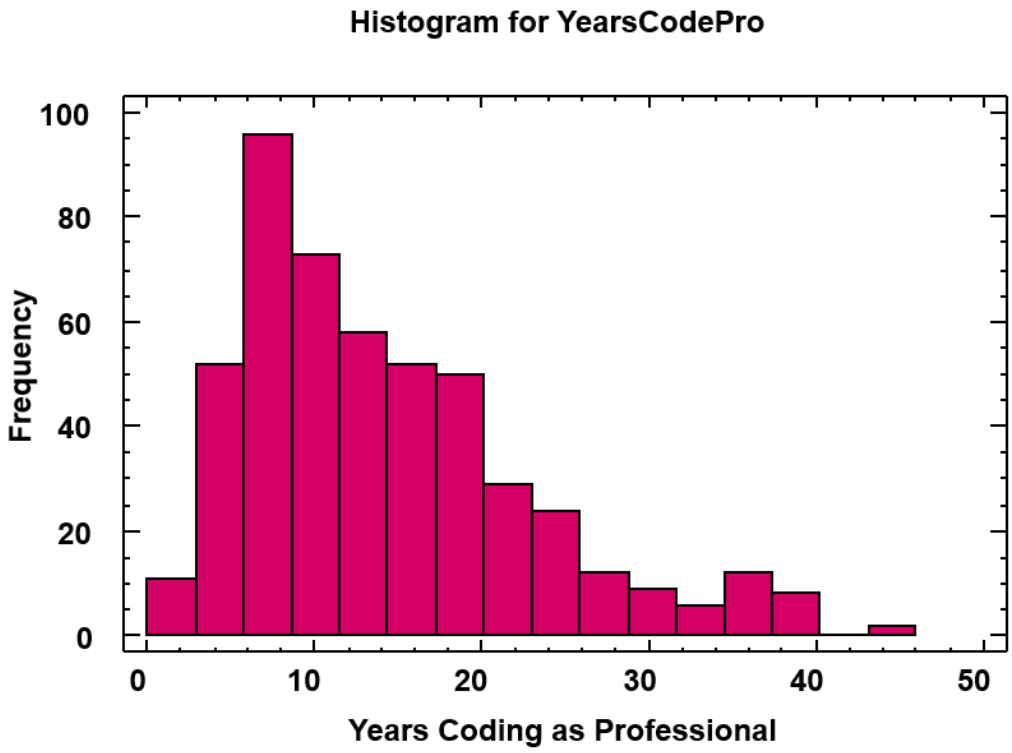


PICTURE 9: Histogram for YearsCode.

YearsCodePro (Pictures 10-11): This case is quite similar to YearsCode. The limits and the classes are adjusted following the same criteria. From 27 classes and limits in [-1,47] to 16 classes with limits in [0,46]

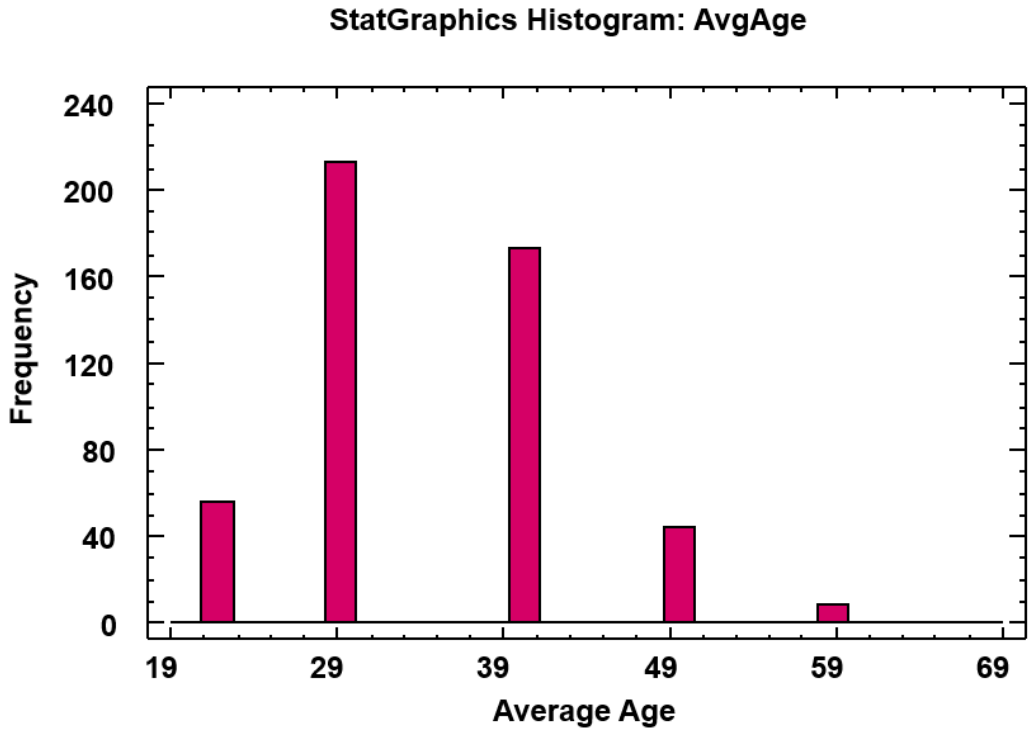


PICTURE 10: StatGraphics Histogram for YearsCodePro

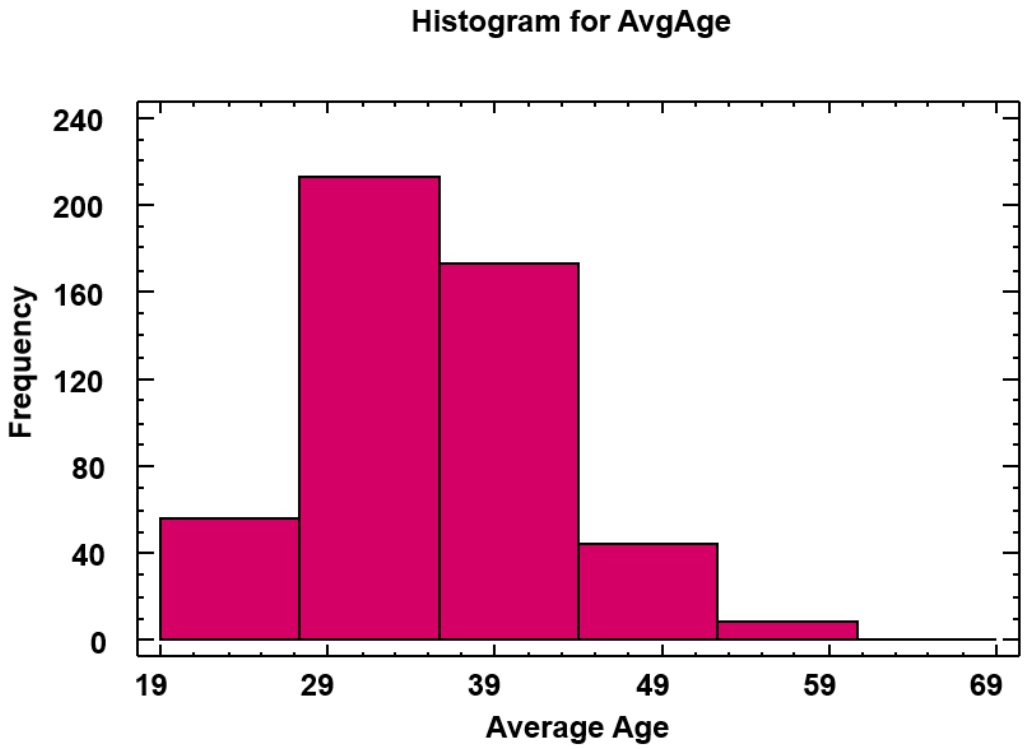


PICTURE 11: Histogram for YearsCodePro.

AvgAge (Pictures 12-13): For this case, the lower limit was set to 19, since it is the minimum age and the classes were drastically reduced from 27 to 16.

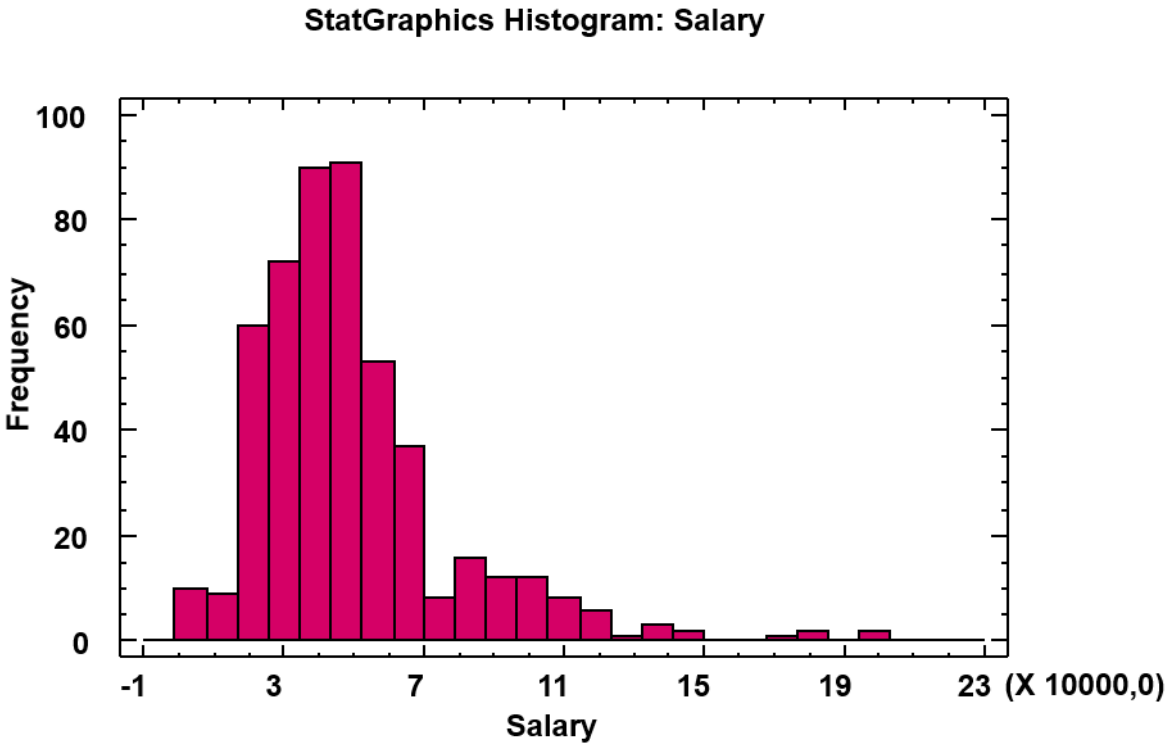


PICTURE 12: StatGraphics Histogram for AvgAge

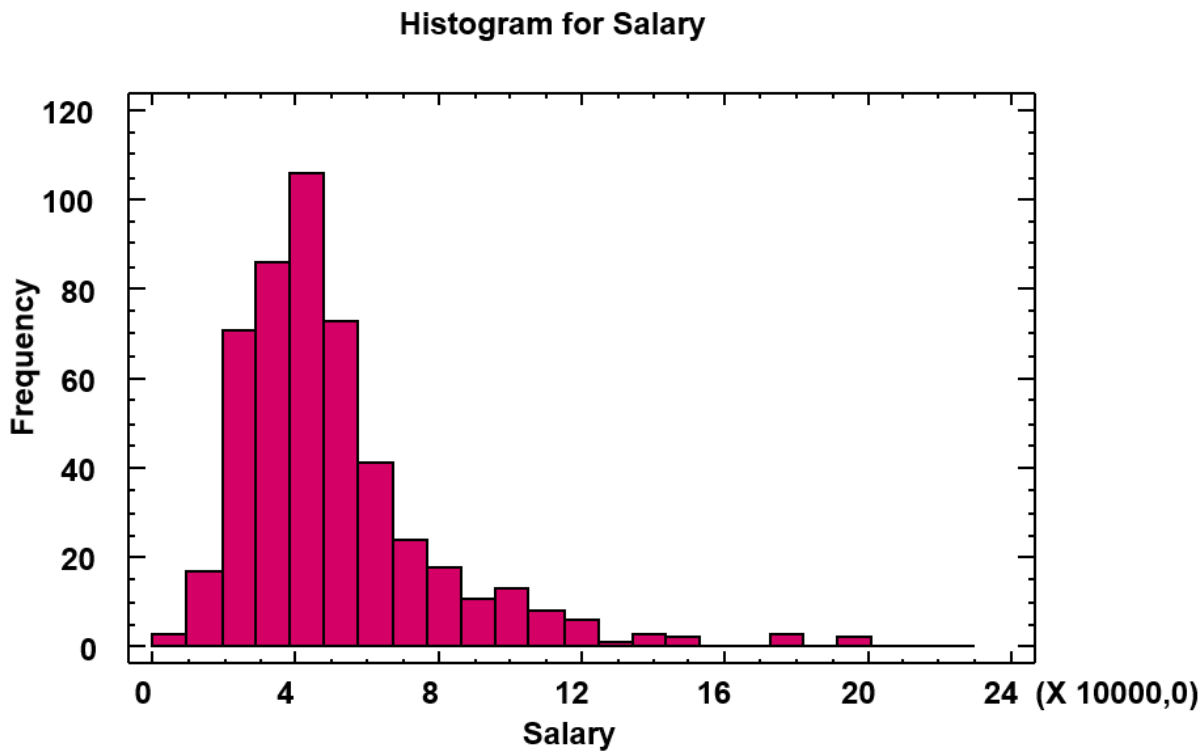


PICTURE 13: Histogram for AvgAge

Salary (Pictures 14-15): Again, the same criteria as in YearsCode. The limit is adjusted from [-10000,230000] to [0,230000] and the classes are reduced from 27 to 24.



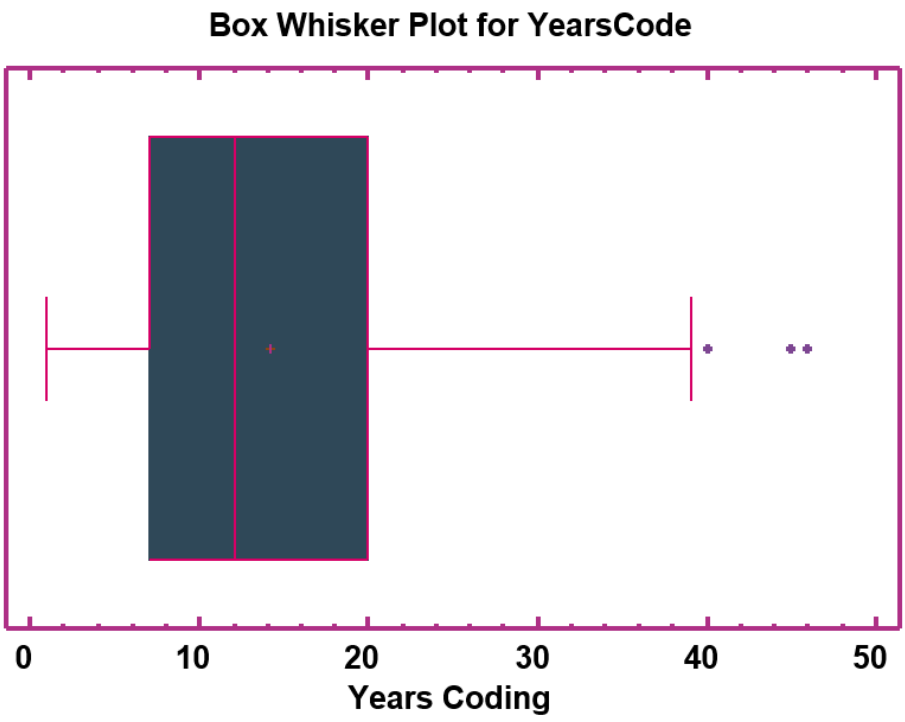
PICTURE 14: StatGraphics Histogram for AvgAge



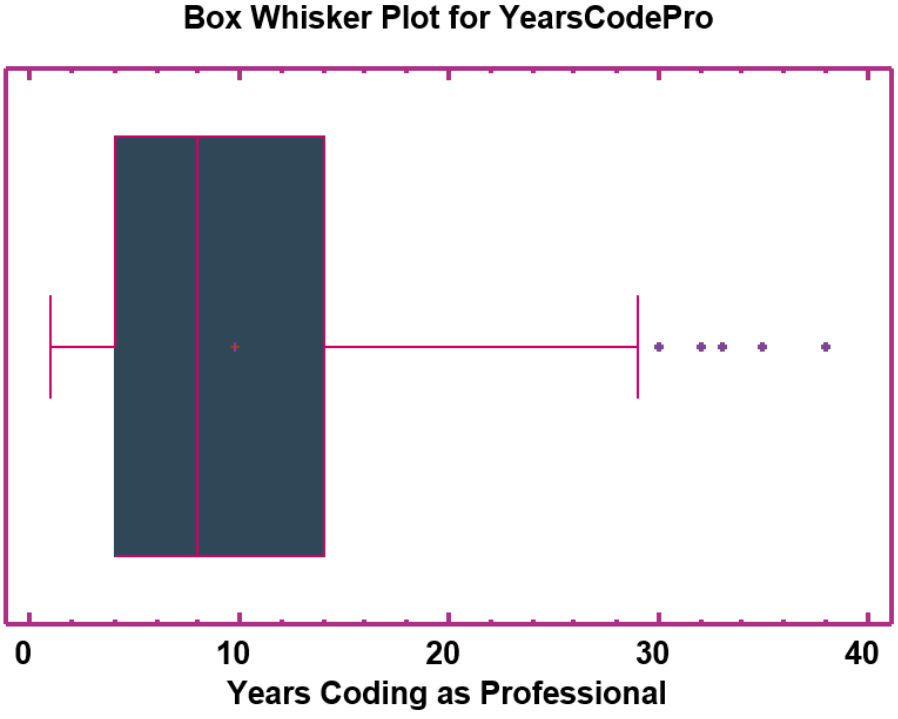
PICTURE 15: Histogram for AvgAge

The general rule for this kind of plot, as it was said before is to reduce the peakedness and the skewness of the distributions. That is why it is important to modify the number of classes. The general rule is to use a number around [5,20], but sometimes it is unavoidable, as it happens in Salary, where 20 classes would be a very abrupt change.

11. Place the box-whisker plots for X_i and indicate similarities and differences between variables:

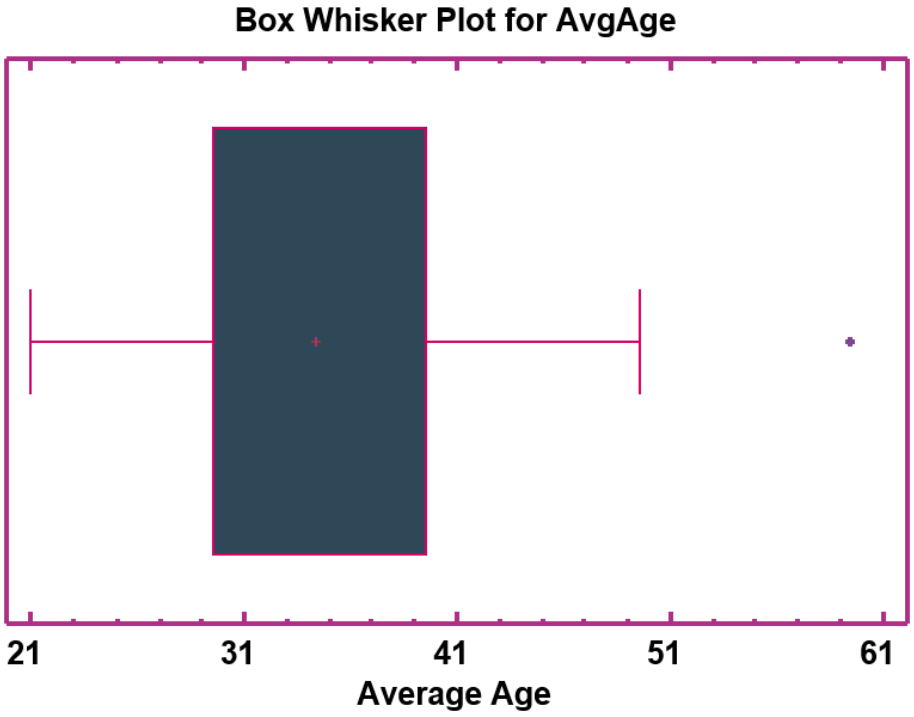


PICTURE 15: Box-Whiskers Plot for YearsCode



PICTURE 16: Box-Whiskers Plot for YearsCodePro

PICTURE 17: Box-Whiskers Plot for Salary



PICTURE 18: Box-Whiskers Plot for Salary

As we can see in the diagrams, we can relate three of the four variables in order to get an idea of the current state of the industry.

First of all, we will analyze YearsCode and YearsCodePro, since both should be related.

Many people start learning code around 2 to 4 years before starting to code professionally.

That is because many people enroll in some type of course that tends to last that time.

Therefore, the results that we see in Pictures 15 and 16 make a lot of sense. The difference between both medians is 2.

Also, the range difference between both makes sense since normally not much people start working as developers without having some formation as it is a specialized job that demands many previous skills.

On the other hand, we can relate Salary, AvgAge and YearsCodePro. Normally in the IT industry and in many others, the salary tends to grow with the working experience, which tends to be bigger depending on the age. That´s why we can see that the median of the salaries is around 40k for a median of working experience of 10 years and a median on average age of 32 years.

Aditionally, we can emphasize that due to the low levels of natality in the occident, along with the population aging, a median of 35 years on AvgAge makes plenty of sense.

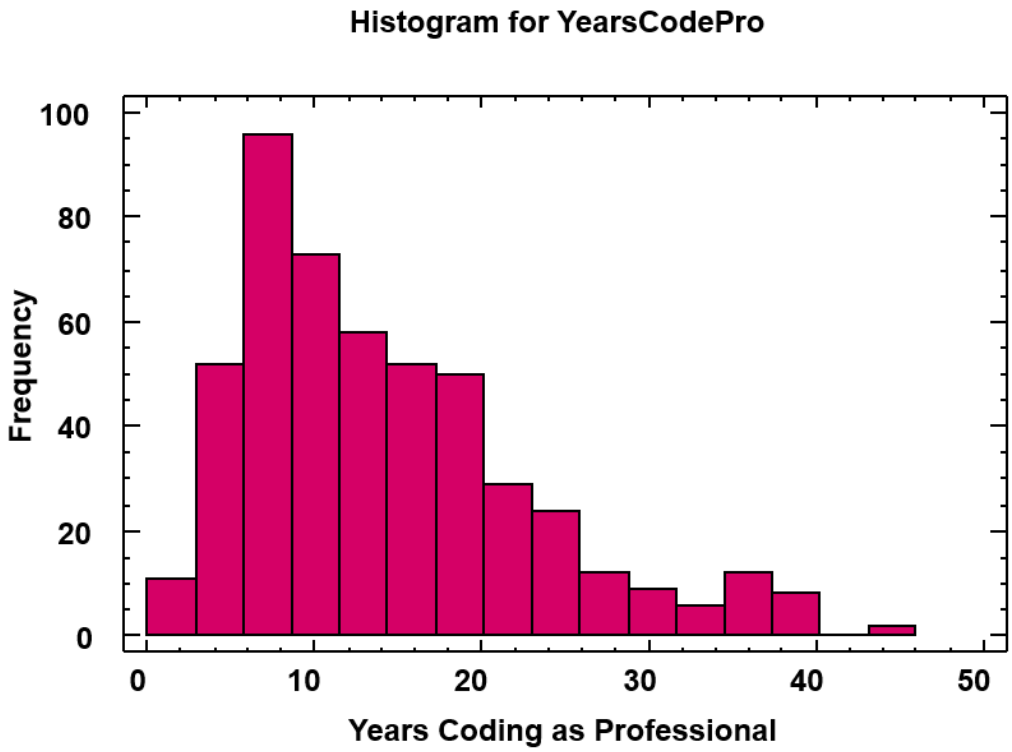
Finally, there are many possible outliers marked that are not considered.

YearsCode and YearsCodePro: Spending around 45 years coding is totally possible.

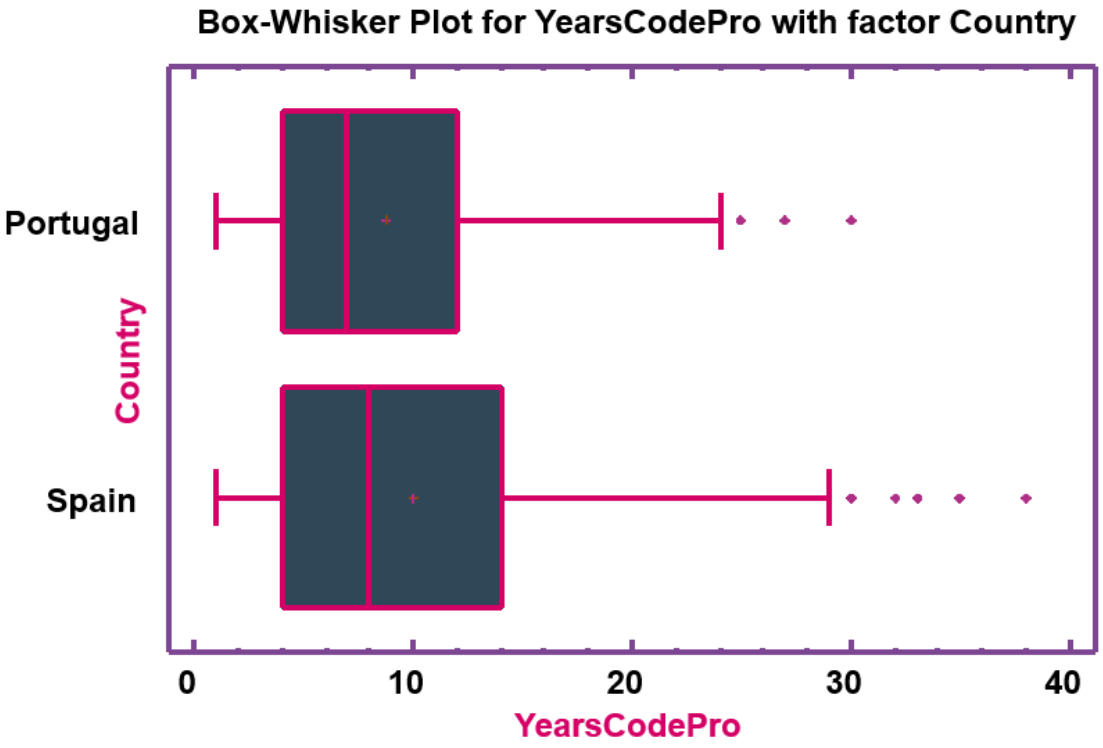
Salary: Earning more than 200000€ is possible for engineering jobs.

AvgAge: None of the values are impossible. Also, by the type of variable, it is protected from casual outliers.

12.1 Place a histogram for YearCodePro with a multiple box-whisker in function of Country:



PICTURE 11: Histogram for YearsCodePro.



PICTURE 19: Multiple Box-Whisker for YearsCodePro with Factor Country.

As the distribution of Picture 11 is more or less symmetric, applying functions would only increase the asymmetry, then, we are taking it as it is.

12.2 Discuss about what the multiple box-whiskers plot consists of:

This plot consists of computing a box whisker plot of one variable, but instead of doing it with all the values, it computes one box whisker plot for each variant of a continuous variable, taking only the data which have that specific variant for each plot. In this case, since Country has two variants, it computed two box-whisker plots.

12.3.1 Compute a table with the SSC and SKC of YearsCodePro in function of Country.

Value	Portugal	Spain
SSC	4,91577	8,61642
SKC	2,36748	3,40381
Count	92	403

TABLE 10. SSC and SKC of YearsCodePro in function of Country variables

12.3.2 Compute a table with the SSC and SKC of YearsCodePro in function of EdLevel.

By taking the values of YearsCodePro and separating them according to which type they belong to, you can use Statgraphics to compute the Standard Skewness and Kurtosis Coefficients of YearsCodePro according to each EdLevel **(Table 4)**. Doing this we obtain the following:

Value	Primary	Professional	Associate	Not Earned	Bachelor	Master	PhD
SSC	1,45663	2,50836	3,4869	3,1369	5,91412	5,52155	0,52498
SKC	0,0572005	1,42359	2,89899	0,480754	2,71079	2,17073	-0,602718
Count	25	33	30	64	174	151	18

TABLE 11. SSC and SKC of YearsCodePro in function of EdLevel variables

This table follows the exact same order as Education level in **Table 4**. Also, primary contains secondary, none and something else

Note: This was done additionally because of a misinterpretation of the exercises.

13. Evaluate the differences between the variants of Country using the information obtained in 12.

- **Position differences: Which has higher median?**
  Portugal has a median of 6,5 and an average of 8,6087 while Spain has a median of 8 and an average of 9,7442.
  Clearly, Spain has higher values.

- **Dispersion differences: Which has higher interquartile range?**
  The interquartile range is 7,5 for Portugal and for Spain has a value of 10. Again, the value for Spain is greater.

- **Distribution differences: Talk about the symmetry.**
  Taking Portugal, a Standard Skewness Coefficient of 4.91 suggests a highly skewed distribution and asymmetric. Also Kurtosis indicates heavy tails and a huge presence of outliars. But in this case, as we stated before **(Activity 11)**, there are no outliars.
  Now, if we compare: In this case, as we have seen previously in **Table 10**, the SSC for Portugal is half of the Spain´s value. Spain has a much more asymmetric distribution.
  In the case of Spain, the skewness and the tails are even higher than for Portugal.

14. Study and discuss the pattern of variation between the variable YearsCodePro and each variant of Country.

Since the distribution for both is quite asymmetric, as said in the previous question, it would be useful to take measures as the 10 and 90 percentile and the Interquartile Range to describe the variation pattern.

14.1 Spain:

The percentile 10 for Spain has a value of 2 points while the percentile 90 has a value of 21.

The variable has a wide range of values, having a difference of 19 between both percentiles. Also, we could have argued about the existence of outliars, but we discarded this option previously.

About the interquartile range, as we know it is 10 we can conclude that the distribution is relatively clustered around the median (8) and few right extreme values.

14.2 Portugal:

The percentile 10 for Portugal has a value of 2 while the percentile 90 has a value of 17.

The variable has also a wide range of values but lower than Spain with a difference of 15 between both percentiles. Here, the outliars are also discarded.

Taking into account the percentiles and knowing that the IQR is 7,5, we can conclude that this distribution is more spreaded than Spain´s. Also, the IQR indicates that the most of the data is around the median (6,5), as it happens in Spain.

Summarizing: We can conclude that Portugal has a distribution that is moderately spread out and in both cases most of the data is clustered around the median.

15. To describe graphically the pattern of variation of the variable Salary, choose the graphic that gives the most information.

PICTURE 17: Box-Whiskers Plot for Salary

15.1 Why would you choose that graphic?

Actually, in my opinion a violin plot would be even better, but it was not an option.

Anyways, the reasons to choose a box-whisker plot are the following:

- High variability, high skewness, high kurtosis.
- Effective display of non-symmetric shapes
- Huge amount of information: Median, quartiles, range and even potential outliers. But in this case, the outliars are taking as realistic extreme values. It is possible to earn 200000€ a year. Even more.

A histogram would also be reasonable, but since there exist those extreme values, they would not be very visible.

A normal probability plot would not be appropriated since the data is far from following a normal distribution.

15.2 Discuss the most relevant information that can be deduced from the graphic:

As we can see in the box-whisker plot, the most important information is the values that can be extracted just by staring at the graphic. Also, we can get useful conclusions based on that information.

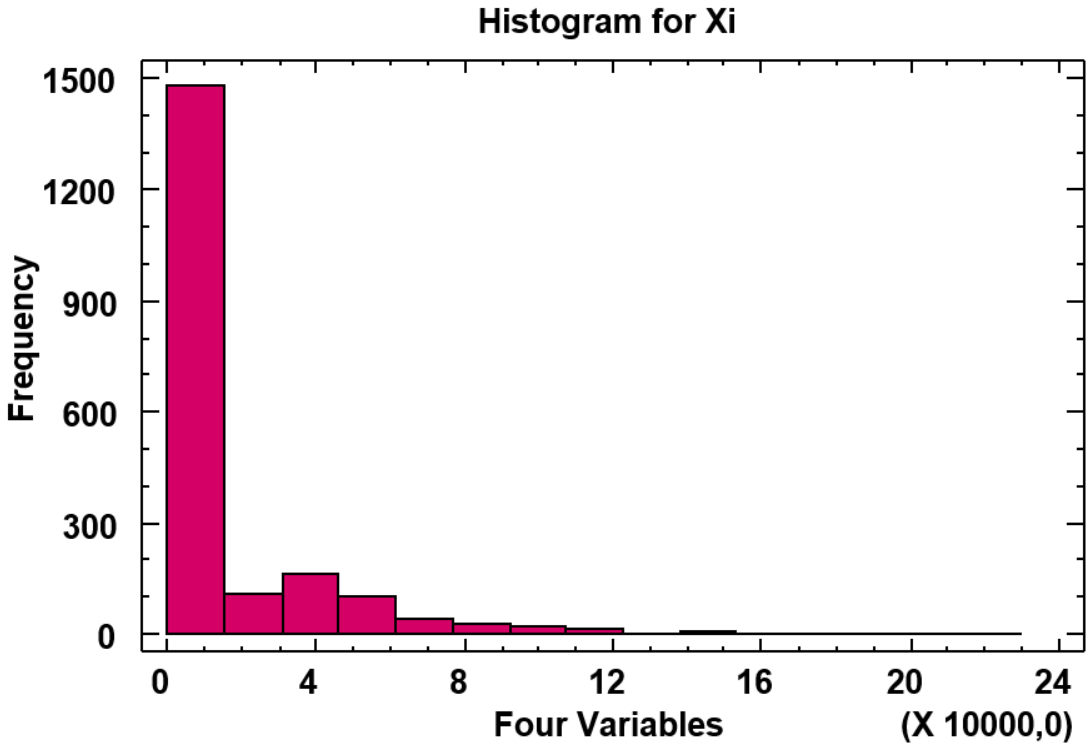
A priori, we can get the following information:

- The median is around 44000€
- As the median is slightly closer to the right bound, we know that the distribution is skewed to the right.
- Q1 is 30000€ while Q3 has a value of 60000€
- The IQR, the has a value of 30000€ aproximatelly.
- Since the whiskers are not so long, we know that the data is mainly clustered in the center of the distribution.
- There are plenty of candidates for outliars, but it does not mean there are outliars since we discarded that possibility previously.

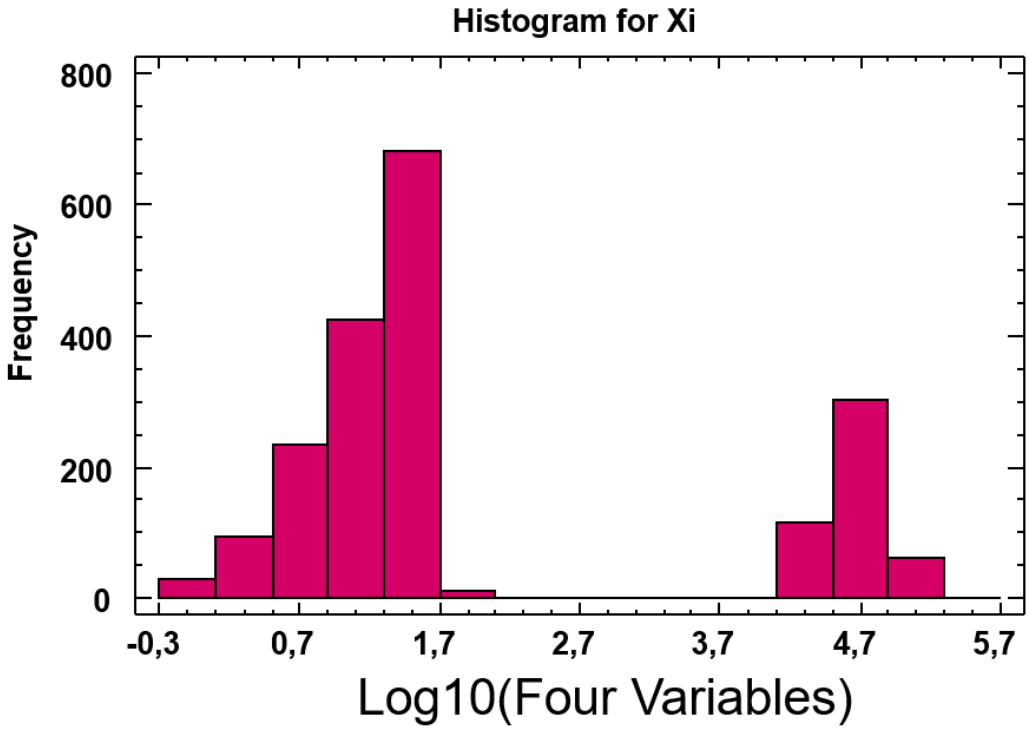
15.3 Are there any anomalous data that should be removed:

As it was settled previously, all the data is valid. It is totally normal for a person to earn yearly a quantity between [0-200000]. A directive or a highly skilled engineer may earn the maximum, while an internship may be working for free to get experience.

16 Join all the X variables and plot a histogram. Convert the data if needed and discuss the results:



PICTURE 20: Histogram for the 4 X variables joined. 15 classes



PICTURE 22: Histogram for the 4 X variables joined with Log10. 15 classes

As we can see in the **Picture 21**, the data is clustered around 0 and 1,5. That is because of the variable salary, which introduces values in a much higher range than YearsCode, YearsCodePro and AvgAge.

While in the YearsCode and YearsCodePro the data is quite similar, and AvgAge has also a small range. On the other hand, in Salary, we are talking about values that are more than 100 times bigger. That is why in **Picture 22** we can see two clear divisions.

For this case, a more interesting histogram may appear if we remove Salary, since it is breaking any coherent relationship we could find between the other 3 variables.