Study and analysis of professional occupations of information technologies, in the state of New York, using Python

Luís Marques[1]

[1] ISCTE, Avenida das Forças Armadas,
1649-026 Lisboa, Portugal
lclms@iscte-iul.pt

Abstract. This document will demonstrate some tests and studies done to the information technology professions of the state of New York. The study will be realized within the scope of data science used some statistical tests, the practical part was done in the IDE Spyder used the language multi paradigm Python.

Keywords - Data science, Data Exploitation, Python

1 Introduction

This work intends to study and demonstrate the employability of some statistical tests on a dataset. For this, a dataset of the professions inherent to information technologies of the state of New York was selected. It was necessary to obtain a dataset to put statistical tests into practice. It is intended that this paper takes a data science approach. For this, in the relative point to this work, the reader is introduced as they relate the data science and some paradigms that have the appearance of law, I speak of the law of Moore that in a great contribution to the society of information and knowledge [1]. This work has two practical parts that intend to show the results of the statistical tests carried out in Python.

1.1 Organization

This work begins by explaining the scope, methodology and explains the problem to be studied. In the demonstrative part of the results is divided into two parts. The first part shows the contribution of the distribution of the data under analysis and the measures of central tendency. The second part addresses some statistical tests of correlation, determination and hypothesis testing.

1.2 Related Work

The objective of this work is to study the professional occupation of the information technology sector. The goal of this job is to investigate how the IT professions are distributed in the state of New York, the motivation for this job is the challenging way to explore the in the Python language is a matter that has gained enormous interest in organizations [2]. This science is not a new one in the era of information and knowledge, has been gaining a great space by Moore's Law [3]. This abundance of information allowed the creation of an area closely linked to the science of data the Big data [3, 4]. A proliferation of data that is everywhere, in the organizations and in the lives of individuals in particular [5]. An enormous demand for online services is a reality, the internet of things, make it even more necessary to deal with all the information in order to take advantage of such organizations [4]. The objective of this work is to study the provision of the professional occupation of the information technology sector.

1.3 Methodology

To explore the data I went to several Python libraries that allowed me to create many things in a short time. The use of these libraries allowed us to save programming time and direct the work to the analysis. Thus, in order not to look at the dataset in macro terms, this work not only looks in a narrow perspective but also demonstrates the results obtained by filtering the original dataset. I used the Anaconda with Spyder is the Scientific Python Development Environment.

1.4 Dataset and graphically

The original dataset was obtained from the New York State Department of Labor website. This dataset is publicly accessible and can be accessed at https://www.labor.ny.gov/stats/lswage2.asp. So, as we already have a first impression of the dataset, I present a summary graph.

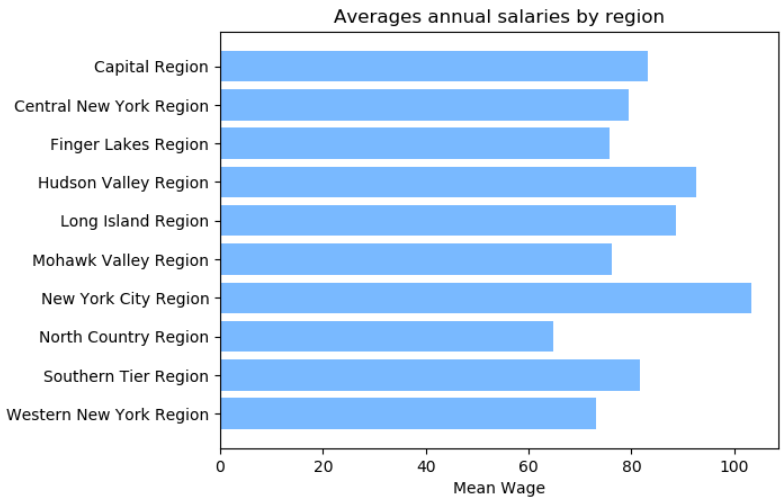

Fig. 1. This figure shows how the variable Mean Wage is distributed by the regions of the state of New York, on the x-axis we can see the scale of the variable on the y-axis of the regions.

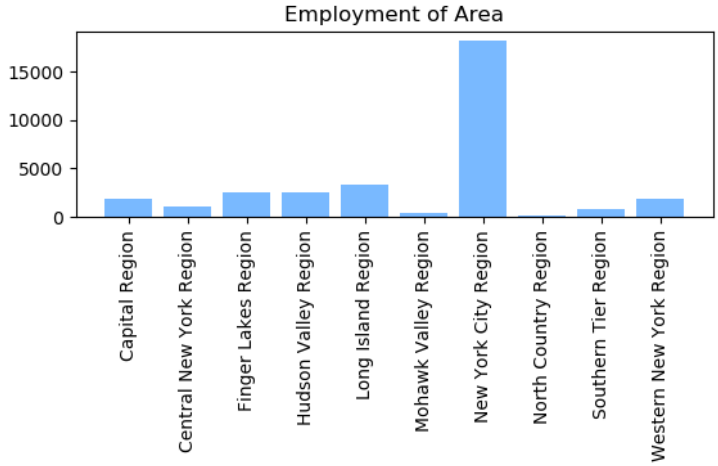

Fig. 2. In this figure, we show the variable Employment. It has its scale on the y-axis and its relation on the x-axis. We have already noticed that the New York region is immensely great compared to other regions.

2 Part I

2.1 Normal

The normal is one of the most important distributions of statistics. This distribution will allow us to see if the values are close to the mean. In addition, allow us to assess how many standard deviations are far from the mean and the asymmetry of the data **[6]**.

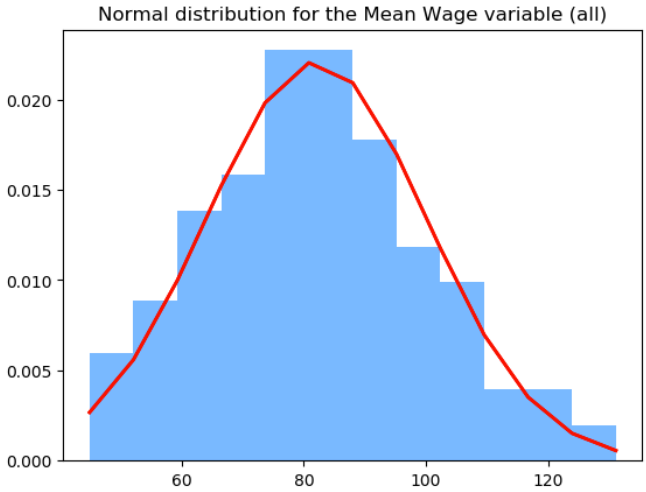

Fig. 3. The graph represented in this figure shows a normal distribution for the variable Mean Wage. So, graphically we can see the curve of normal.

2.2 Cumulative Distribution Function vs Empirical Distribution Function

At this point, we will study how CDF and ECDF relate. In the case of the CDF we will perceive how the data the probability of the distribution of the data. In ECDF we visualize how the distribution of the graph variable is.

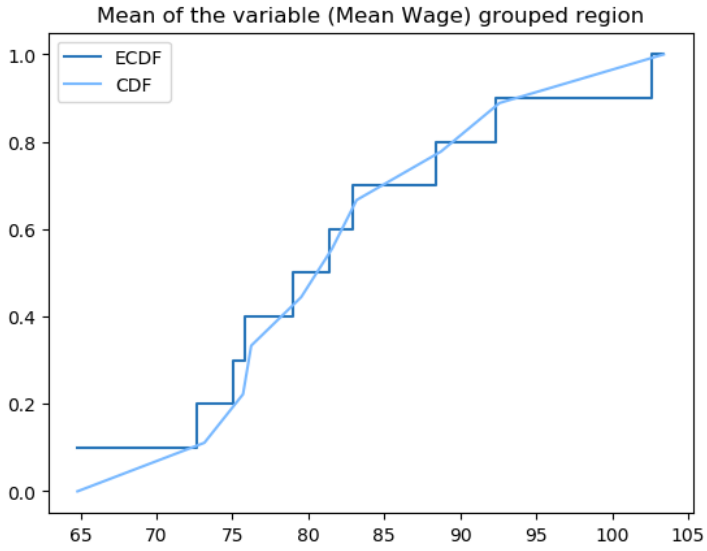

Fig. 4. The graph of this figure shows the probability distribution and cumulative distribution of the Mean Wage variable. This variable was grouped by region. Visually the two distributions are approximate [6].

2.3 Measures of central tendency and graph of identification of possible outliers

A set of plots that allow us to graphically see where the mean is, as well as the visualization of how data can deviate from measures of central tendency.

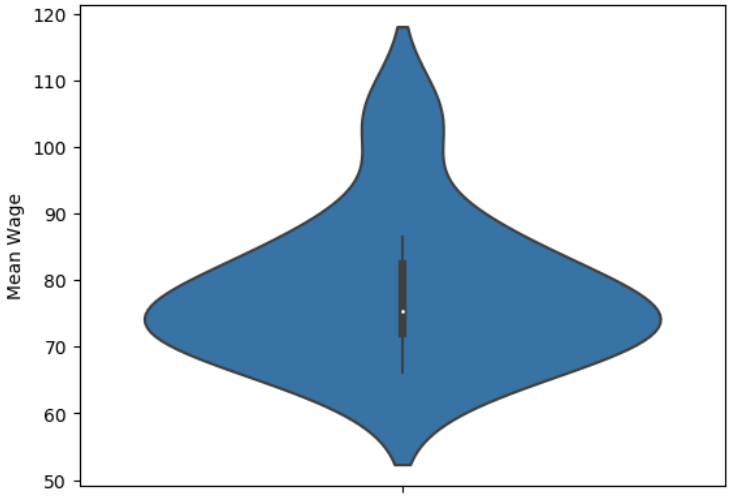

Fig. 5. This chart shows the average earnings per year of Computer Programmers, the average is $ 78,383 per year.

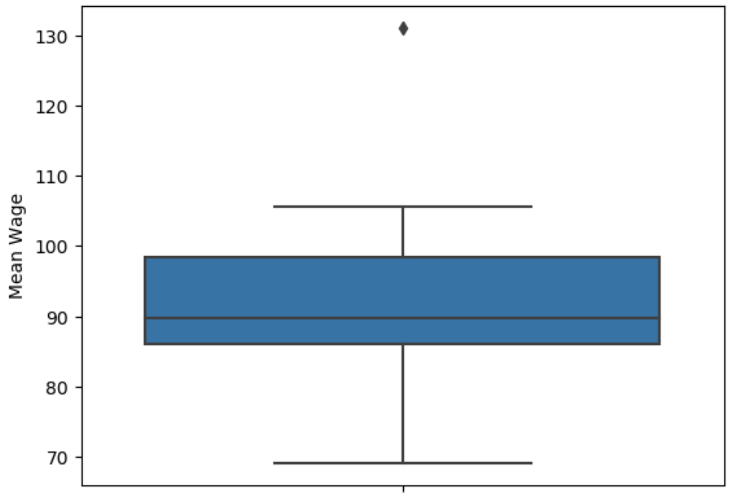

Fig. 6. This plot represents the salaries per year for the Information Security Analysts these professionals earn on average $ 93.90. We can see that in the upper center of the graph this value is demarcated, we will see later in the linear regression the influence of these values. We can not neglect this value because it is a necessary value to the dataset, the value is well identified. The city of New York has higher values than the other regions of the state of New York.

3 Part II

3 Pearson vs Spearman

In this section we will smooth out the coefficients of correlation, for this we use the theorems of Pearson and Spearman. The calculation to be performed for these theorems are of the variables of the number of jobs and the average of annual income, we will use the incomes of the programmers.

Table 1. This table shows the values obtained in the correlation theorems. Spearman is not so influenced by many values beyond the average. We can see from the table that the correction is moderately positive for Spearman and perfect positive for Pearson.

	Correlation
Spearman	0.7212
Pearson	1.0

3.2 Linear regression

In this section, we will calculate and graphically represent the linear regression, so we will understand the relationship between the variables. The challenge in regression is to determine which line best fits the data, i.e. a line so situated in the point cloud that it minimizes the distance from all points to the line.

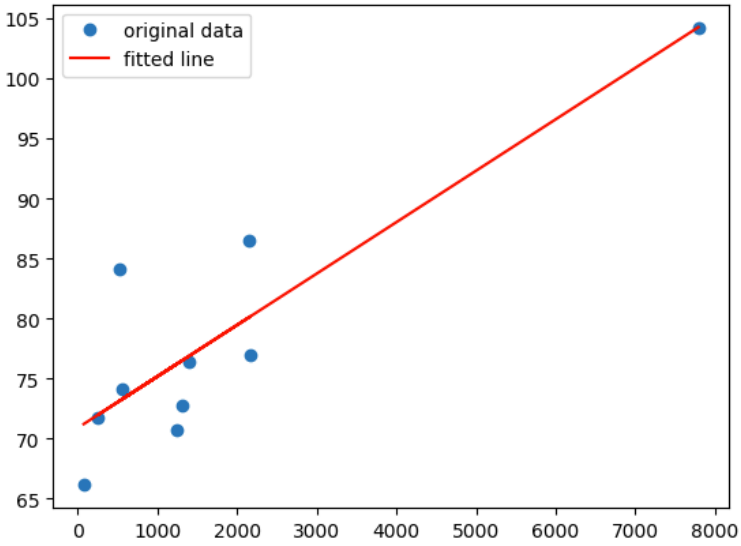

Fig. 7. This graph shows the linear regression for the variables of the number of Employment and Mean Wage. We can see that the real values are concentrated in the lower left. We can observe that there is a value in the upper right corner that greatly influences the equation. This graphic was created with the linregress function of the scipy.stats library. The criterion of least squares was used assumes that the "best adjustment" is achieved when the sum of squares of differences between the real value and the predicted value. The result of the coefficient of determination r^2 was 0.773. Rejection of null hypothesis.

3.3 Kolmogorov–Smirnov test

In this section, we will follow the Kolmogorov-Smirnov test, it is a non-parametric test that will enable us to gauge the determination coefficient between the CDF and ECDF distributions. So, let's use the variables Mean Wage and Employment.

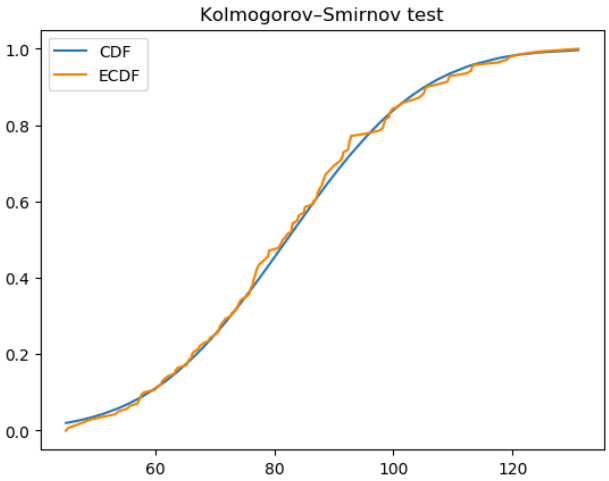

Fig. 8. The graph shown here is the result of the comparison between CDF and ECDF distributions for the variablee Mean Wage. Thus, it is intended to apply the Kolmogorov-Smirnov test to perform a hypothesis test [7]. The result obtained corresponds to KstestResult (statistic = 0.0486, pvalue = 0.892). According to the P-value returned, and because it is a high value, there is strong evidence to accept the null hypothesis.

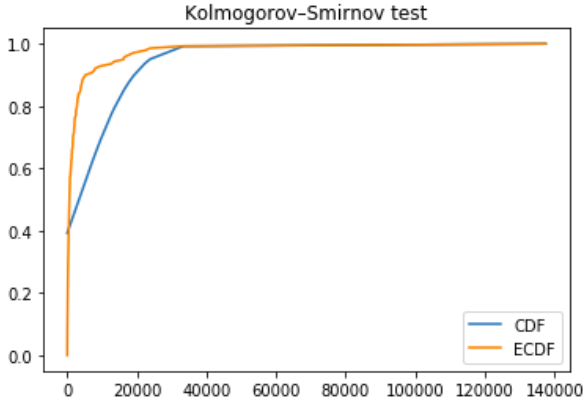

Fig. 9. The graph shown here is the result of the comparison between CDF and ECDF distributions. Thus, it is intended to apply the Kolmogorov-Smirnov test to perform a hypothesis test [7]. The result obtained corresponds to KstestResult (statistic=0.3923, pvalue=0.0). As the P-value is 0, we reject the null hypothesis.

4 Conclusion

This work has shown some statistical tests, however, it had a scope of data exploration. This scope is framed in data science, a science that is not properly new but that leveraged with the technological advance has gained much prominence in the discovery of knowledge [8]. You can see the potentialities of this science if you think, for example, of the internet of things and the immensity of information to be treated [6]. For this, there are several issues to deal with, what drives us this work is that there are intellect, tools and other factors that propel the data science a thriving science for the future.

References

[1] R. R. Schaller, "Moore's law: past, present and future," 1997.

[2] M. A. Waller and S. E. Fawcett, " Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management," *Journal of Business Logistics,* 2013.

[3] M. Barlow, "The Culture of Big Data," *O'Reilly Media, Inc.,* 2013.

[4] K. E. Dierckens, A. B. Harrison and C. K. Leung, "A Data Science and Engineering Solution for Fast K-Means Clustering of Big Data," 2017.

[5] T. Leipziga, M. Manza, D. Schöttlea, K. Ohlhausena, P. Oosthuizenb and G. Leipzigb, "Initialising customer-orientated digital transformation in enterprises," 2016.

[6] G. J. Myatt and W. P. Johnson, Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining,, John Wiley & Sons, Inc. Published , 2014.

[7] M. Krumholz's, "Mark Krumholz's Page," [Online]. Available: https://sites.google.com/a/ucsc.edu/krumholz/teaching-and-courses/ast119_w15/class-10. [Accessed 12 11 2017].

[8] D. Donoho, "50 years of Data Science," 9 2015.

[9] Penn State, [Online]. Available: https://onlinecourses.science.psu.edu/stat501/node/250. [Accessed 12 11 2017].

[10] minitab, "minitab," [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values. [Accessed 10 12 2017].