Nicodemo, Catia; Satorra, Albert

Article

Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data

BRQ Business Research Quarterly

Provided in Cooperation with:
Asociación Científica de Economía y Dirección de Empresas (ACEDE), Madrid

Suggested Citation: Nicodemo, Catia; Satorra, Albert (2022) : Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data, BRQ Business Research Quarterly, ISSN 2340-9436, Sage Publishing, London, Vol. 25, Iss. 3, pp. 283-294, https://doi.org/10.1177/2340944420957335

Methodological Insight

Business Research Quarterly
2022, Vol. 25(3) 283–294

DOI: 10.1177/2340944420957335
journals.sagepub.com/home/brq
SAGE

Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data

Catia Nicodemo[1] and Albert Satorra[2]

Abstract

New challenges arise in data visualization when the research involves a sizable database. With many data points, classical scatterplots are non-informative due to the cluttering of points. On the contrary, simple plots, such as the boxplot that are of limited use in small samples, offer great potential to facilitate group comparison in the case of an extensive sample. This article presents exploratory data analysis methods useful for inspecting variation across groups in crucial variables and detecting heterogeneity. The exploratory data analysis methods (introduced by Tukey in his seminal book of 1977) encompass a set of statistical tools aimed to extract information from data using simple graphical tools. In this article, some of the exploratory data analysis methods like the boxplot and scatterplot are revisited and enhanced using modern graphical computational devices (as, for example, the heat-map) and their use illustrated with Spanish Social Security data. We explore how earnings vary across several factors like age, gender, type of occupation, and contract, and in particular, the gender gap in salaries is visualized in various dimensions relating to the type of occupation. The exploratory data analysis methods are also applied to assessing and refining competing regressions by plotting residuals-versus-fitted values. The methods discussed should be useful to researchers to assess heterogeneity in data, across-group variation, and classical diagnostic plots of residuals from alternative models fits.

JEL CLASSIFICATION: C55; J01; J08; Y10; C80

Keywords

EDA, large data set, ggplot, heat-maps, R

Introduction

The topic of exploratory data analysis (EDA) as a distinctive tool in applied statistics was created by John W. Tukey in his 1977 book "Exploratory Data Analysis." That book renewed the topic of descriptive statistics and enlightened three main strategies that have become crucial in modern data science: (1) graphical presentation, (2) flexibility in viewpoint, and (3) intensive search for simplicity. The methods of EDA do not present *p*-values or standard errors, but instead focus on the sharp visualization of key aspects of the data at hand. Tukey's developments of EDA focused on robust statistics and strategic graphical displays. In this article, we focus on use of boxplots and scatterplots, in combination with new computational tools, for graphical display (heat-maps).

A major feature of modern applied statistics work is the widespread use of graphical displays of the data. This has been grounded on the methods of EDA developed by Tukey in the late 1980s, together with the advancement of graphical capabilities in computer sciences. The discipline of graphical displays has evolved as an entire discipline of statistics (e.g., Chambers et al., 1983; Cleveland, 1993; Downey, 2014; Healey & Enns, 2002; Hoaglin et al., 1983;

[1]CHSEO, Department of Primary Care, University of Oxford, Oxford, UK
[2]Universitat Pompeu Fabra and BGSE, Barcelona, Spain

Corresponding author:
Catia Nicodemo, CHSEO, Department of Primary Care, University of Oxford, Walton Street, Oxford OX2 6GG, UK.
Email: catia.nicodemo@gmail.com

Myatt & Johnson, 2014). All current statistical software (SPSS, Stata, etc.) have ways to display data that were not present just a few years ago.

This article revisits EDA tools, applying the boxplot and the scatterplot to databases with many observations. We focus on methods that assess variation across groups, point to outliers, disclose clustering of cases, and highlight non-linearities in relationships between variables. The role of EDA is to explore graphically data in ways that could reveal structural secrets and to gain new, often unsuspected, insight into the data. The EDA methods discussed align with recent arguments that graphs usually create "pre-attentive" visual processing in the brain, helping to focus on the important message (Camacho et al., 2015; Healey & Enns, 2002; Hussain & Prieto, 2016; Levine, 2018). Cheng et al. (2013) and Schwabish (2014) suggest that researchers who want to disseminate their research to a wider audience of non-specialists should think carefully about how to construct effective graphics. Similar recommendations are made in the papers of Jebb et al. (2017) and Varian (2014). In reflecting on statistics for management, George et al. (2014) conclude that with large data sets it is too easy to get false correlations when using typical statistical tools. The EDA methods can produce a set of visualizations that simplify the understanding of complex data, which will be increasingly useful to researchers in management, business, and social sciences in general, who often face the challenges associated with using large databases in their research.

The classical inferential methods of statistics become less useful in very large samples, since test results are nearly always significant, and the standard errors are very small. EDA could help in this case by emphasizing data visualization and dimension reduction methods. Traditional descriptive statistical analysis faces problems with large data sets. For example, a bivariate relationship that would be clear in a basic scatterplot with relatively few observations may not be readily apparent in a data set containing a million observations (e.g., see Figure 1 where a simple scatterplot is used). A large part of the literature on big data concentrates on computer-intensive methods, such as machine learning or regression trees, that emphasize prediction (for more details about these tools, see Qiu et al., 2016; Al-Jarrah et al., 2015). In our study, rather than big data, we concentrate on the case of large database (large sample size), see De Mauro et al. (2016) for terminology on big data. In contrast, EDA methods focus on methods that serve to explain and describe data and should be among the first steps when analysts want to explore large data sets. We believe that there is gap in the literature of big data analysis on the subject of exploring and presenting statistical features of large data sets. This article focuses on methods that can fill part of this gap. Our main contribution is to show how traditional, simple methods should be used in the context of the new paradigm of big data. It should be

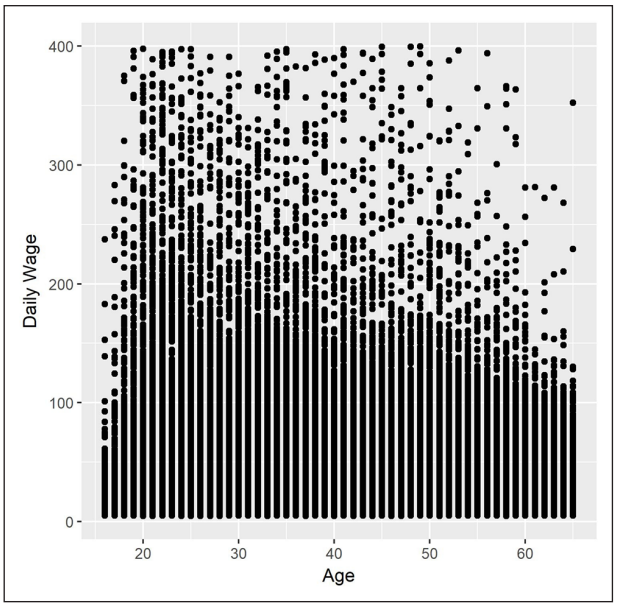

Figure 1. Example of scatter plot of log-wage by age.
Source: MCVL 2010.

recognized that descriptive tools do not give yes/no answer to basic research questions, unlike the classical testing used in econometrics. However, the EDA methods have the ability to show shifts, heterogeneity, outliers, and non-linearities among variables, without the requirement of model assumptions. EDA methods should be seen as complementary, not only prior to the classical econometric methods but also to give support or rejection to a priori posed econometric models. We show how EDA methods are also useful after a regression fit in residual diagnostic plots.

To better convey the practicalities of the methods proposed, we illustrate them by analyzing the variation of salaries in the Spanish Social Security (SS) database. In particular, we address such issues as whether the observed difference in wages can be explained by the following: the rigidity of labor market (temporary vs. permanent contract), discrimination (women vs. men), and age (old vs. young). Previous work involving the Spanish labor market has focused on fitting models to test aspects of labor economic theories (e.g., Dolado et al., 2002; Gehrke & Weber, 2018). In this article, we deviate from this testing approach in favor of descriptive methods that permit direct, visual assessments of the questions explored. The structure of the article is as follows. Section "SS data" describes the database. In section "The comparative boxplots: log-wage by age groups," we discuss group variation in wages using the boxplots. Section "Scatterplot: The heat-map" discusses the use of heat-map scatterplots. In section "The heat-map scatterplot in regression diagnostics," we apply the heat-map approach to a classical residual diagnostic plot, and to the comparison between OLS and Tobit regression. Finally, in the last section, conclusions are presented.

SS data

To illustrate the EDA methods, we use the Spanish SS data on labor, specifically, the "Muestra Continua de Vidas Laborales" (hereafter, MCVL) in 2010.[1] The data comes from the register of the SS System for people active in the labor market. Starting in 2004, SS records have been released for a 4% non-stratified random sample of the population who in that year have had any relationship with Spanish SS (individuals who are working, receiving unemployment benefits, or receiving a pension). Given the structure and magnitude of the MCVL, this is a useful data set with which to illustrate the EDA approach we advocate. The rest of this section is devoted to describing briefly the statistics and economic labor context of this database.

The data set gives information regarding historical relationships of individuals with the SS relating to work, unemployment benefits, and pensions, for around 1 million observations each year. It also contains information regarding the type of contract, sector of activity, qualifications, earnings, date of entering or leaving the job market, part-time or full-time status, and firm size. The MCVL also provides individual characteristics such as age, gender, residence, country of birth, and level of education. Information on educational attainment has improved in recent editions of the MCVL. The main outcome variable of interest in our data set is the earnings of Spanish people. The MCVL only provides information on the "social contribution base" (censored earnings), which captures monthly labor earnings plus 1/12 of bonuses received over the year. The censored earnings variable has minimum and maximum values, which vary over time.[2] We exclude pensioners, the self-employed, individuals receiving unemployment benefits, individuals who report outlying earning values (outside the minimum and maximum values), and individuals with missing information in relevant variables (age, education, type of contract, occupation). This leaves a sample of 541,457 individuals. We calculate the daily wage as our main earnings measure, computed as the ratio between the monthly contribution base and the number of days worked in that particular month. If the individual records more than one job at the same time, we sum the earnings. Other years and variables of course could be considered, but these will be beyond the scope of this article, which is not focused on exploring the Spanish labor market.

We analyze the variation of wage with respect to variables: age, gender, type of contract (fixed [permanent] or temporary), education (primary, secondary, and tertiary), and occupation (high-, medium-, and low-skilled). These variables are traditionally considered in the literature of labor markets (Heckman et al., 2006). Those variables that are found to be significant in a regression analysis of wages are reported below. Note, however, that significance within such a large sample does not provide reliable evidence on the substantive relevance of these variables. As such, visual inspection of the variation of wage with respect to those variables is necessary. We extract information about qualified and non-qualified employees, splitting the sample by qualification (high-, middle-, and low-skill levels) according to the International Standard Classification of Education (ISCED) classification and following the classification proposed by Garcia-Perez (2008) for the MCVL data. We consider the type of contract held (fixed or temporary) and we construct five age groups defined by the following intervals: (15, 25], (25, 35], (35, 45], (45, 55], and (55, 65].[3] Table 1 reports descriptive statistics of the wage by gender, occupation, education, type of contract, and age.

Table 1. Descriptive Statistics of Daily Wage.

	M	Median	*SD*	*N*	%
Total	45.63	51.59	41.27	541,457	100
Gender					
Male	56.07	50.23	30.39	294,132	54.32
Female	46.26	39.89	29.47	247,325	45.68
Age					
(15,25]	35.31	30.52	29.07	64,011	11.82
(25,35]	49.89	44.90	27.99	167,927	31.01
(35,45]	55.20	49.29	29.85	156,692	28.94
(45,55]	57.68	51.47	30.81	106,941	19.75
(55,65]	54.06	48.81	31.81	45,886	8.47
Type of contract					
Temporary	41.41	35.73	31.69	208,954	38.59
Fixed	57.99	52.06	27.65	332,503	61.41
Occupation					
High skill	80.53	87.12	29.78	94,431	17.44
Medium skill	49.25	45.35	26.61	343,481	63.44
Low skill	32.94	30.16	23.10	103,545	19.12
Education					
Primary	39.23	36.61	25.24	117,967	21.79
Secondary	51.26	45.91	29.13	347,429	64.17
Tertiary	72.26	74.46	32.19	76,061	14.05

Table 1 reports descriptive statistics of the wage by gender, occupation, education, type of contract, and age.

Older people generally experience higher wages. Workers with a fixed contract and highly skilled jobs also have increased wages, as do workers who have reached higher education levels. This table shows a wage gap between male and female workers of around 17.50%. This observation of a gender gap in the Spanish labor market is well documented (see, for example, Amuedo-Dorantes & De la Rica, 2006).

The next sections illustrate the use of EDA for assessing the variation of wages across workers' characteristics.

The comparative boxplots: log-wage by age groups

The boxplot was introduced by Tukey (1977) as a way to present graphically the distribution of a continuous variable and also to display the variation of a continuous variable across groups. Boxplots display the first, second, and third

quartile; the interquartile range; and outliers of a database. The information displayed by the boxplot, and most of its variations, is based on the data's median. The 50% central bulk of the data is represented by a box; two whiskers, one at side of the box, represent the two 25% extremes of the data distribution. A line dividing the box represents the median. Symmetry/skewness of the distribution is visualized graphically by the position of the dividing line of the box: symmetry occurs when the line divides the box in two equal halves, and skewness occurs when one half is larger than the other. The boxplot is very useful for visualizing group dispersion in a variable, where the boxplot of each group is set in parallel (vertically, or horizontally), one besides the other. The boxplot is of limited use when the sample size is small, thus displaying variation across groups using the boxplot requires a large sample.

Note that in the case of a very large sample, the *F*-test of an analysis of variance (ANOVA) table will tend to be significant (i.e., *p*-values below the significance level) since any small difference of the population means will be detected as significant, given that the large sample size increases the power of the test. EDA methods and, in particular, the display of parallel boxplots, offer researchers direct visualization of the variation across groups by showing the median and quartile values.

Wages can be presented using raw values or after a log transformation. The advantage of the log transformation for group comparison is the approximate normality of the distribution. Hence, the comparative boxplots will have reduced the skewness of the wage distributions and will make it easier to compare the ends of the wage scale (see Hubert & Vandervieren, 2008). We use the log base 10 transformation; however, it should be noted, that once a researcher finds interesting variation at the log scale, they can easily display the same data in the original scale.

Figure 2 shows parallel boxplots of log-wage for different age groups. For each cohort, the box and the whiskers span the whole variation of log-wage in the group. The edges of the box are the first and third quartile; the median is the dividing line of the box. Points exceeding the solid line of each of the whiskers are cases that could be categorized as outliers.[4] The *y*-axis represents the dependent variable that is common to all boxplots, thus the variation on the level of the dividing line shows graphically the variation of the median across groups. We see that for young people, the distribution of log-wages is not symmetric (the median line does not divide the box in two halves), but near symmetry is observed in the other groups.

A feature we have added to the standard boxplot display is to make the width of the box proportional to the size of the group (number of individuals in each group) in the whole sample. This makes it easy to assess the relevance of a specific group; for example, we see that the age group (55,65) represents a smaller proportion of workers than the age group (35,45).

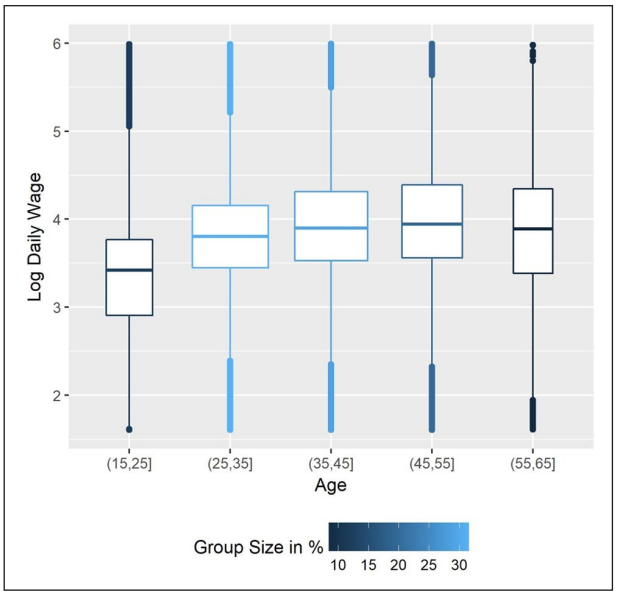

Figure 2. Boxplot of log-wage by age.
Source: MCVL 2010.
The wide of the box is proportional to the size of age group.

The following findings can be drawn from this diagram:

1. There is a slight curvilinear increase in the median level of log-wage with age, with the first age group (15,25] showing the lowest median log-wage that is different from the other groups.
2. There is slight variation across the groups in the dispersion of log-wage, as measured by the interquartile range (the length of the box). The highest dispersion arises in the two most extreme age groups.
3. The width of the boxplots shows the proportion of the population in the group, and we see that the groups (25,35] and (35,45] are the widest.
4. While there is a significant increase in wage when going from group (15,25] to group (25,35] (the youngest groups), there is a slight decrease in the median of log-wage when moving from (45,55] to (55,65].
5. The log transformation normalizes (symmetrizes) the distribution for all groups, except for the youngest group that still shows a long-tail on the right. This asymmetry for the salary of young workers could raise policy discussion.

These findings are congruent with previously established economic empirical research (Cabrales et al., 2017), which states that skills and experience (age would be a surrogate) impact positively on earnings. We now consider group variation when we cross two or more categorical variables.

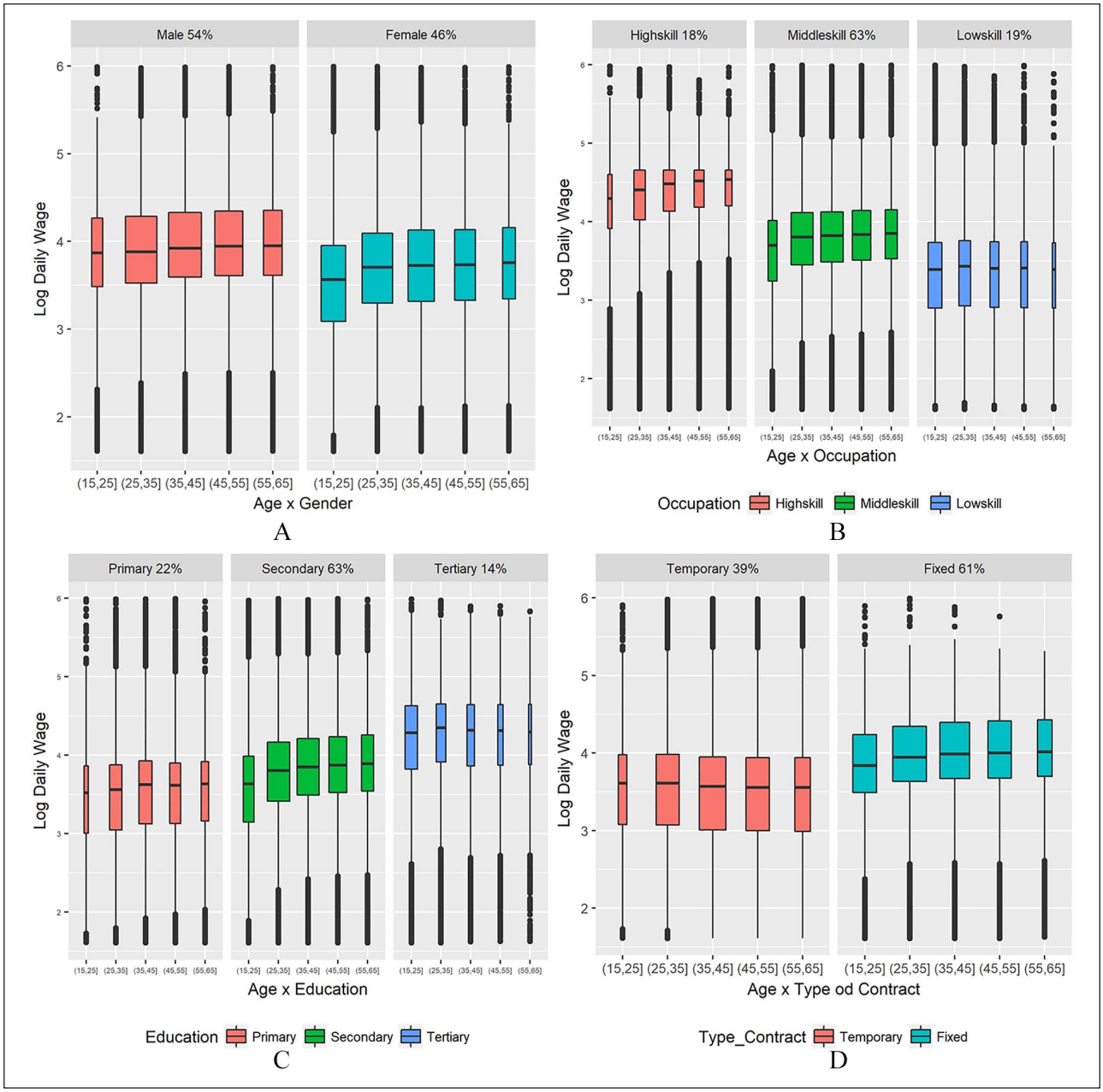

Figure 3. Boxplot of log-wage—Panel A: age and gender; Panel B: age and occupation; Panel C: age and education; and Panel D: age and type of contract.
Source: MCVL 2010.
The wide of the box is proportional to the size of age group.

Variation of log-wage by age groups and a third categorical variable

Log-wage by age can vary when controlling for different additional variables, such as occupation, level of education, type of employment, and gender. Figure 3 displays variation of log-wage on age for each of those aforementioned variables. Panel A shows that there is a gender gap in wages for all age groups except for the two youngest. The gap reaches a maximum of 24 points (of log-wage) for those aged 45 to 55 years. The graph shows that the median (log) salary gap between men and women increases with age, although it is important to note that we do not control for other variables.

Panel B of Figure 3 depicts the variation of log-wage on age for occupations classified as low-, middle-, and high-skilled workers. The proportion of people in the different groups is shown at the top of the graph. We observe the middle skills group is the largest and accounts for 63% of all workers. The width of the boxplots represents the

relative size of the group. The group of young people (ages 15–25 years) is the smallest and has the lowest log-wage, and this trend occurs for the three skill levels. As expected, the group of highly skilled workers is the one with highest earnings.

Panel C shows the variation of log-wage on age for the three educational levels: primary, secondary, and tertiary. As expected, people with a tertiary level of education receive the highest wages (see the median levels of each group). Note that, the size of the group decreases with age for all education levels (as is illustrated by the width of the box). Furthermore, the plot shows that the highest variation of wage by age occurs in the secondary education level.

Panel D shows the variation of log-wage on age for the two types of contract: permanent and temporary. We see that the highest proportion of contracts are permanent, with the highest salaries. In both groups of temporary and fixed contracts, median salaries remain fairly homogeneous across age, though in temporary contracts, we see a slight decrease of salary as age increases.

Figure 3 describes, in a four-display graph, the variation of wage conditional on age and one additional variable (specifically, gender, occupation, education, or type of contract). Note that one could also consider a five-display graph where each display shows the boxplot relative to one or two of the covariates for each age group. The necessity of such additional graphs would arise from the inspection of the ones at hand. As the main purpose of the article is to highlight useful EDA devices, we do not comment further on these additional graphs.

This graph allows researchers to assess directly how the distribution of salaries changes with type of contract in age group, or between men and women. Note that this type of comparison, considering variation of wages in groups defined by crossing many categorical variables, will be possible only in the context of large data. Otherwise, the sub-samples would be small and the boxplot would become non-informative. Variation conditional on a third variable will be introduced in the next section.

Exploring the gender gap in log-wage

In this section, we explore the gender wage gap. Reducing the gender wage gap is an important topic on the European political agenda. The persistence of the gender wage gap is the result of direct discrimination against women and/or a structural inequality, such as segregation in sectors or occupations, access to education and training, or biased evaluation and/or pay systems. We present evidence showing the difference between male and female earnings, considering several factors that could explain not only the wage gender gap but also the selection of women into jobs with certain characteristics.

In Panel A of Figure 4, we see a gender gap in wages for all age groups. The log-wage gaps between males and females persist after controlling for age. The proportion of workers who are young is larger for males than females, while among females there is a smaller proportion of older workers (this is seen from the width of the boxplot). There is a wage gap between younger females' wages than wages of all other female age groups. This pattern does not occur for males. The dispersion of wages is similar across age and gender. Panels A to C of Figure 4 shows the variation of log-wage by gender, controlling for additional characteristics of workers: occupational skills, level of education, and type of contract.

Panel A shows that, while the gender wage gap is not apparent in young groups with temporary contracts, it is clear among workers with permanent contracts. We also see that few women in the older age group hold permanent contracts.

In Panel B, we can see that the boxplots of high-skill groups are thinner. This reflects official statistics which report a lower proportion of workers in the high-skill group. Among high-skill workers, the gender wage gap is minimal. The maximum gender gap arises in the middle skills group and is largest among the young. The gender wage gap is also apparent among low-skill workers but is smaller.

Finally, Panel C shows that when controlling for the level of education, a gender gap in wage is still visible at all levels. The gap is at its highest for the second level of education and greatest for the younger women at this level. There are more women than men in the group defined by lowest age and higher level of education (see the thickness of the boxplot for the tertiary education level for females aged 15–25 years).

Panel A shows that, while the gender wage gap is not apparent in young groups with temporary contracts, it is clear among workers with permanent contract. We also see that few of the women in the higher age group hold permanent contracts.

Panel B shows that the high-skill groups have thinner boxplots. This reflects official statistics which report a lower proportion of workers in the high-skill group. Among high-skill workers, the gender wage gap is minimal. The maximum gender gap arises in the middle skills group and is largest among the young. The gender wage gap persists among low-skill workers but is smaller.

Finally, Panel C shows that, when controlling for the level of education, a gender gap in wage is still visible at all levels. The gap is at its highest for the second level of education and more for the younger women at this level. It is interesting to see that there are more women than men in the group who have a higher level of education and lowest age (see the thickness of the boxplot for the tertiary education level for females aged 15–25 years).

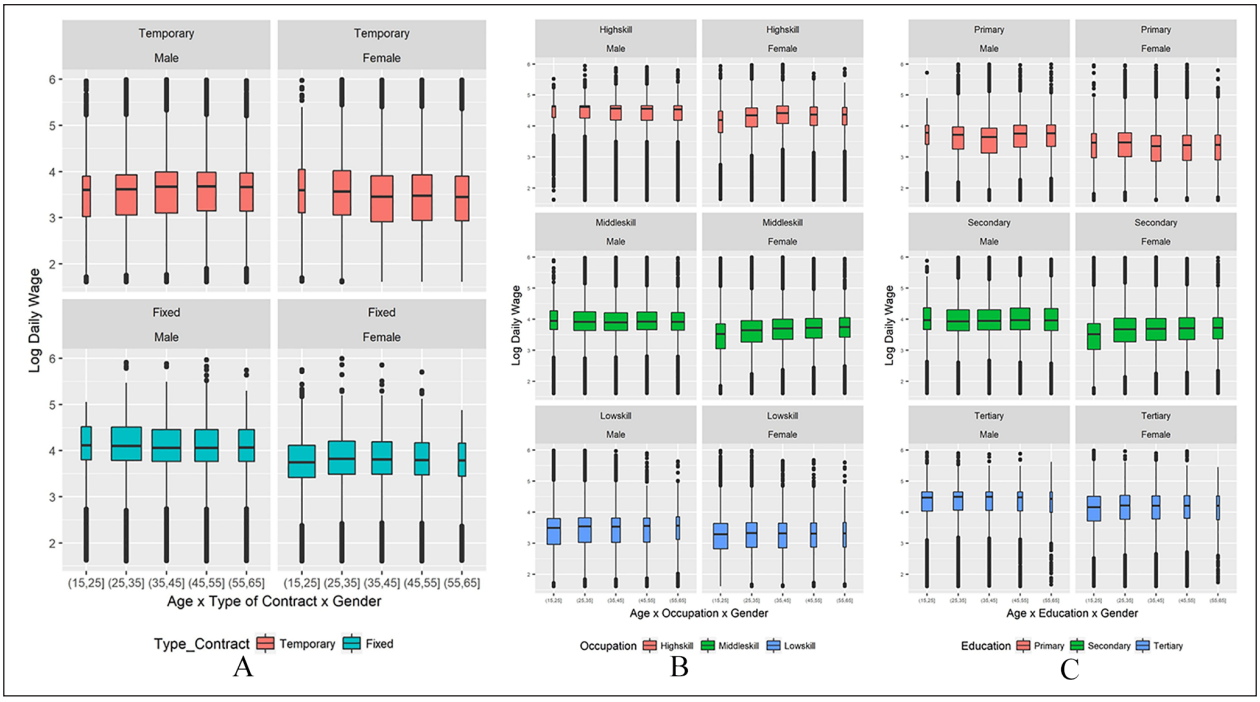

Figure 4. Boxplot of log-wage and gender—Panel A: age, gender and type of contract; Panel B: age, gender and occupation; and Panel C: age, gender and education.
Source: MCVL 2010.
The wide of the box is proportional to the size of age group.

Scatterplot: the heat-map

So far, we have assessed the variation of a continuous variable Y (the log of daily wage) by considering groupings based on several categorical variables Xs. One of the Xs we considered was age, split into several age groups. In our database, however, age is a numerical variable, so a simple scatterplot of daily wage Y[5] on the variable age X could be applied. This is attempted in Figure 1, where, as commented earlier (in section "Introduction"), the large sample sizes produce a cluttering of points and a non-informative graph. Using modern computational tools of density mapping—possible only in the case of a large sample—these data can be presented in a more informative heat-map scatterplot (for more details about heat-map see Barter & Yu, 2018; Wilkinson & Friendly, 2009). The heat-map scatterplot of daily wage on age is shown in Figure 5.[6] The plot shows that the highest density of observations occurs at age around 30 years and salaries around 50 euros, and that there is also a ceiling effect on the daily wage variable that gives rise to a concentration of points. This ceiling effect group will be further discussed below.

Using the heat-map approach, we can assess the variation of daily wage on age when controlling for other variables, as shown in Figure 6. Looking at Panel A of Figure 6, we can see that women have lower wages than men. Yet, we observe that there are few older women working. The modal concentration[7] at the top wage level is of lower intensity for women, showing that there is a smaller proportion of women in this "ceiling effect" group.

Panels B and D show daily wage variation on age when controlling for occupation and education levels. They show that high-skilled workers and those with tertiary level of education have the largest earnings, with many in the ceiling group. The variation of wages across skill and education levels show, as expected, that wages increase with skill and education. Individuals with primary level of education, or in the low-skill group, present larger heterogeneity, not only across wage levels but also across age. Fitting a nonparametric regression line to each of the scatterplots, we see a steady, near-linear increase in wage with age. Panel C shows the variation of salary by age when controlling for the type of contract. We see that in contrast to permanent (fixed) contracts, workers with temporary contracts show significant wage dispersion, concentrated in a small range of age variation. Furthermore, the ceiling effect (of high salary workers) appears only to impact those with permanent contracts. The modal wage for temporary contracts is lower than the modal wage for permanent contract workers (if we discount the ceiling group, then the modal wage for the permanent contracts intersects with the highest modal wage salaries of the nonpermanent contracts). The results in this section are in line with many studies on the Spanish labor market. For more details see, for example, Cabrales et al. (2017).

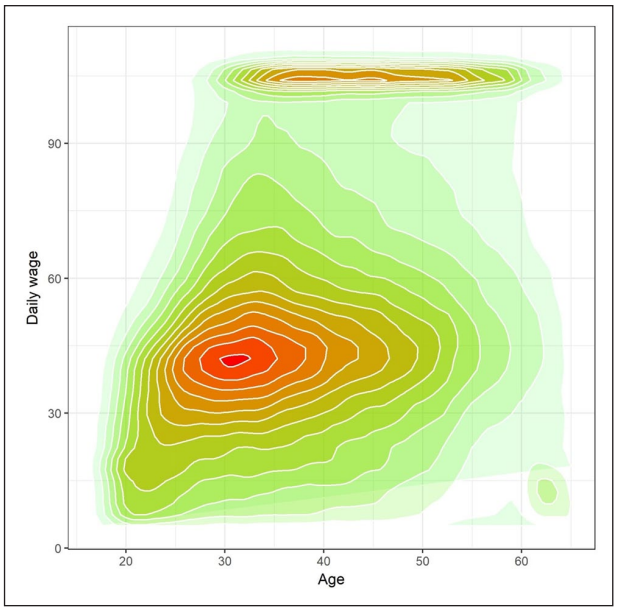

Figure 5. Heat-map of density plot for daily wage by age.
Source: MCVL 2010.

The heat-map scatterplot in regression diagnostics

The relationship between wage and age is now explored using a regression approach. We assume that the expected value of the dependent variables Y is linear on a set of covariates, $X_1, \ldots, X_n$. Violation of the assumption of linearity, however, can invalidate the results of regression analysis. A key diagnostic plot used to test this assumption is the residuals-versus-fitted-y plot. In the context of a very large sample, this diagnostic plot will be cluttered by too many points.

In this section, we illustrate the use of regression residual plots in the context of our labor data, especially when assessing the impact of age, gender, education, and other worker characteristics on our dependent variable, namely, the log of the daily wage.

We fit three regression models using the log daily wage as the dependent variable and the covariates listed in the first column of Table 2. In the covariates, we have categorical variables (represented in the regression by dummies of category) with reference categories detailed in the footnote of the table. The second column of this table shows the estimates for the fitted OLS regression using the entire sample, the third column reports the Tobit regression, and the fourth column reports the OLS regression excluding all the cases where the log daily wage is at the ceiling point 4.506. As expected, given the large sample, all the regression coefficients are statistically significant. We have not reported in the table the standard errors, since the estimates are 0.00 for all the coefficients when rounded at three decimal places (the rounding we use for all the numbers of the table). The significance has been computed using the robust standard errors to adjust for unknown heteroscedasticity.

After a regression fit, it is useful to inspect the residuals-versus-fitted-y-plot. This is the way to asses possible non-linearity and other types of misspecification of the regression equation. Figure 7 reports residuals-versus-fitted y-plot for the three regression models.

In Panel A, we note a bimodality of the bivariate distribution, with a high density of points on the right-hand side of fitted ys and positive residuals. This is illustrated in the graph by a high density mass of points in the upper right section of the graph. The nonparametric curve that fits the mean level of the residuals across the fitted values of Y is shown in the graph by a black line. Strict linearity implies that this line should be a horizontal line at $Y=0$. In the simple scatterplot, it would be very hard to detect problems in the regression, given the large number of points in the graph. In addition to the right censored observations, we see a unimodal distribution with density mass in the center of the bivariate graph.

As mentioned in the section above, earnings in the MCVL data suffer from censoring: any individual with a daily wage above the threshold of 106 euros is recorded to have a wage equal to this threshold. If we eliminate individuals from the regression who have been censored, we observe how the positive residuals disappear in the heat-map of residuals-versus-fitted-y (Panel B of Figure 7). Finally, Panel C reports the heat-map of fitted versus residuals generated using a Tobit model. Note that a researcher will always be faced with the issue of which of the three fitted models to trust. Standard econometrics would recommend the Tobit regression. The problems of OLS analysis when data are censored have been widely reported, see Blundell & Meghir (1987). Note that using EDA methods, we identified modal concentrations among the residuals of the fitted OLS regression, which suggests violation of OLS assumptions and potential problems with the estimation.

The alternative estimates shown in Table 2 should be complemented with the residuals shown in Figure 7. The full picture of the fitted model cannot be assessed without the estimated residuals. Note that OLS estimates do not account for the censoring shown in the residuals of Figure 7 (Panel A) (where we see that a line of censored residuals on the top right). The residuals from a Tobit regression shown in Figure 7 (Panel C) have already accounted for the censoring. The censoring is not visible in Figure 7 (Panel B) because the censored cases have been removed from the estimation.

Conclusion

The analysis of large data sets is increasingly popular in business and social science research, and there are many

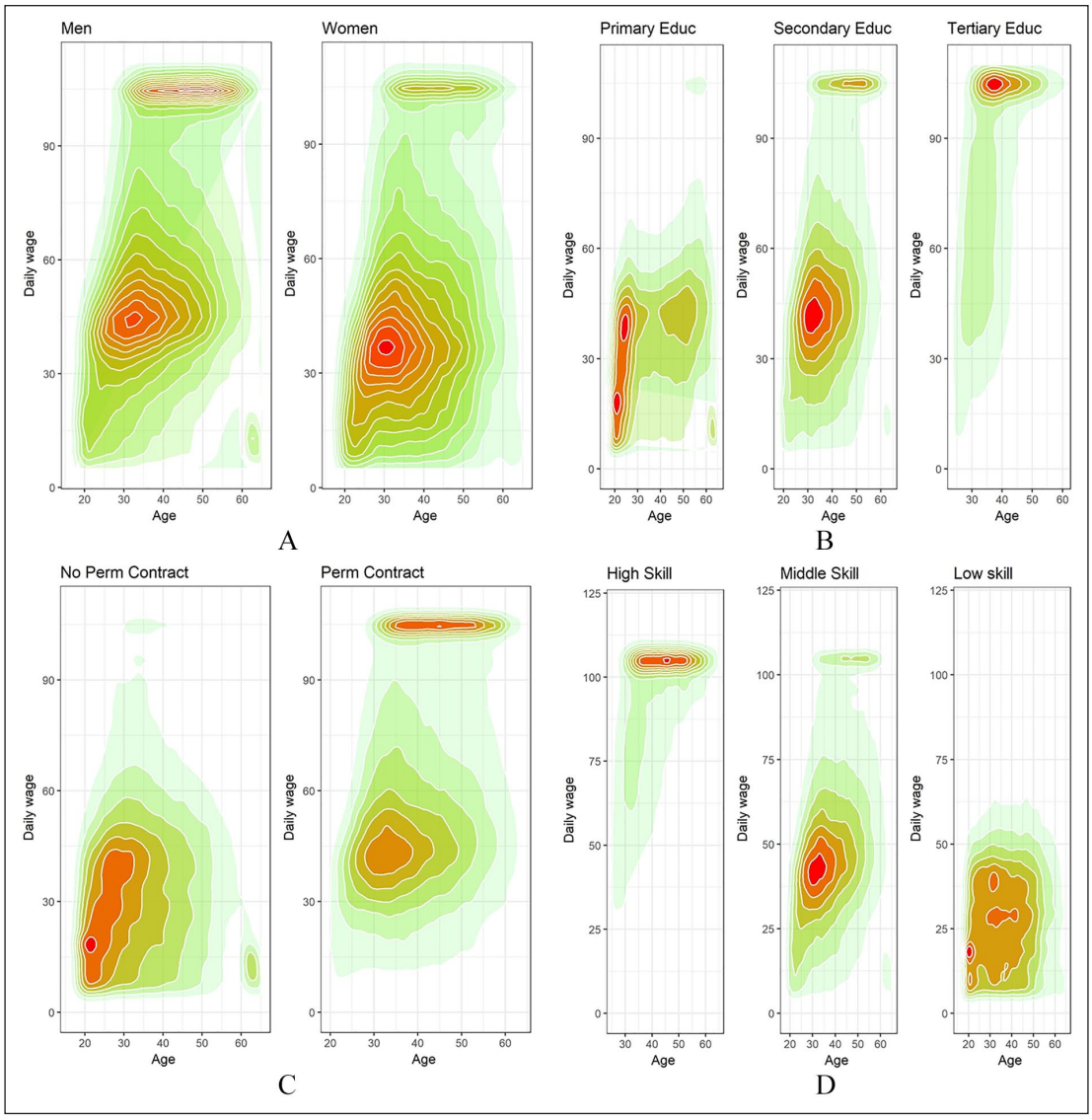

Figure 6. Heat-map of density plot for daily wage and gender—Panel A: gender and age; Panel B: education and age; Panel C: type of contract and age; and Panel D occupation and age.
Source: MCVL 2010.

Table 2. Estimates of the three regression models. The dependent variable in all the models is the log daily wage.[a].

Covariates	OLS (all sample)	Tobit	OLS (sample truncated)[b]
Age	0.052[c]	0.073	0.051
Age square	−0.001	−0.001	−0.001
Education (secondary)	0.059	0.054	0.043
Education (tertiary)	0.165	0.124	0.131
Occupation (middle skill)	−0.405	−0.376	−0.311
Occupation (low skill)	−0.621	−0.632	−0.509
Gender (female)	−0.193	−0.184	−0.170
Contract (fixed)	0.263	0.271	0.252
Industry	0.434	0.512	0.415
Building	0.371	0.423	0.382
Trade	0.175	0.226	0.166
Transport	0.331	0.415	0.319

(Continued)

Table 2. (Continued)

Hotel	0.043	0.053	0.061
Telecommunication	0.313	0.332	0.263
Finance	0.403	0.387	0.296
Service intellectual	0.277	0.207	0.218
Service manual	0.148	0.153	0.164
Public admin	0.361	0.306	0.382
Education	0.094	0.072	0.143
Health	0.357	0.253	0.339
Others	0.058	0.061	0.065
No. of observations	553,103	553,103	486,398
R^2	0.398	0.402	0.304
Adjusted R^2	0.398	0.402	.304

[a]The reference categories for the dummies of education, occupation, gender, type of contract, and sector are, respectively, primary education, high-skill occupation, male and temporary contract, and agriculture. The municipalities fixed effects are also included in all the regressions.
[b]We have excluded in the sample all the cases when the log daily wage it is at the ceiling value at the log of 106 euros (4.663).
[c]All the coefficients of the three regressions are significant at *p*-value level .005 (significance computed using robust standard errors). Standard errors are not in display in the table since they are all equal to 0.000 when rounded at three decimal digits.

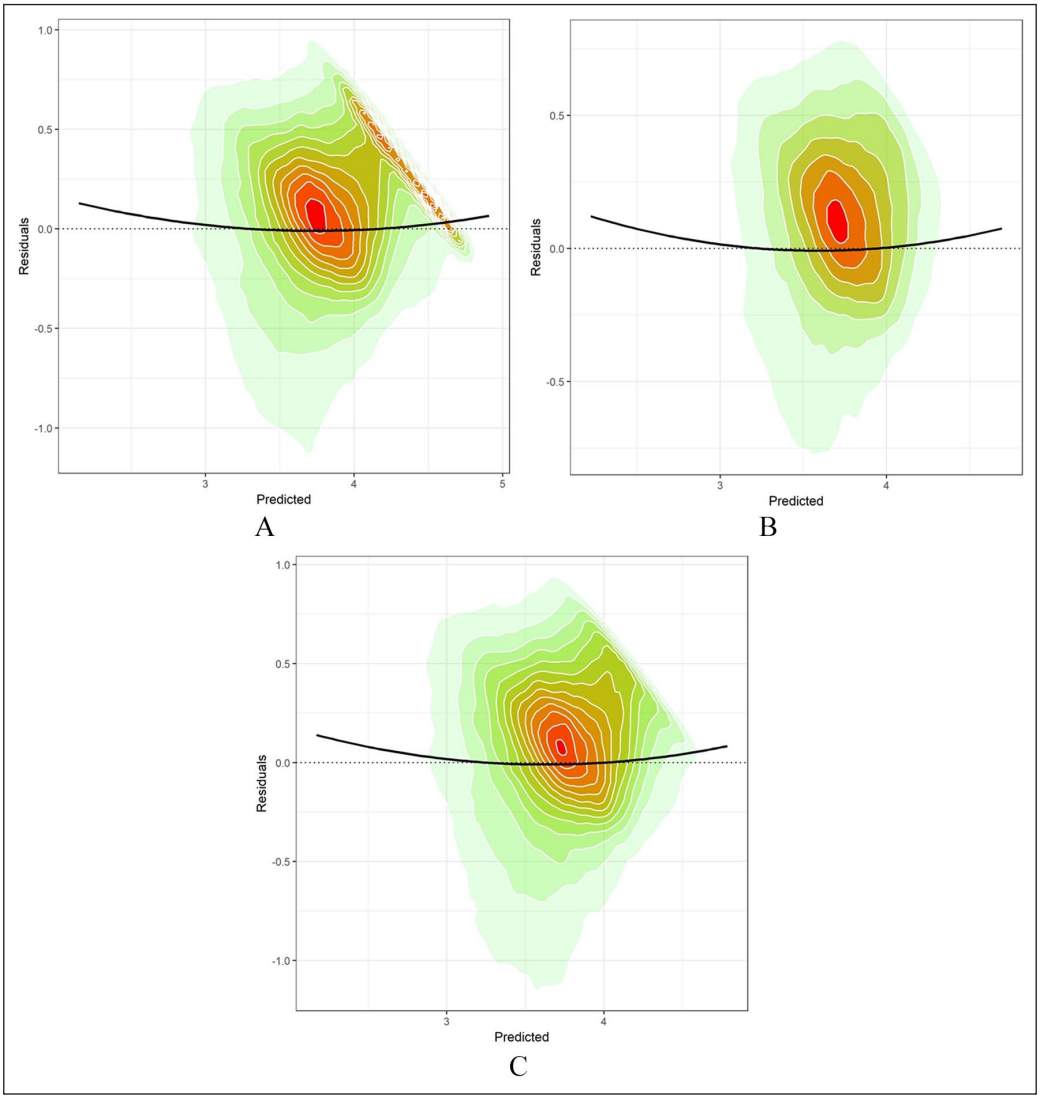

Figure 7. Heat-map of the residuals of regression analysis for daily wage—Panel A: All sample; Panel B; Without censored data; and Panel C: Tobit model (all sample).
Source: MCVL 2010.

new opportunities to extract useful empirical evidence from data. Extensive research in economics has been devoted to the econometrics of regression (or extended regression) models. In this article, we present how EDA methods that focus on graphical displays of data can help researchers to inspect large databases and explore heterogeneity across groups of variables. In particular, we focus on the use of boxplots and heat-maps. EDA methods are useful not only as ways to present descriptive statistics but can also help with robustness check by selecting appropriate models to use in regression analysis.

Using SS data from Spain, this article illustrates how EDA can be used to highlight group variation of critical variables in the labor market and how they interact. We use boxplots and heat-map scatterplots to assess the heterogeneity of earnings on basic covariates, such as gender, contract status, experience, skills, and others. We also show how the heat-map scatterplot can be used for the basic diagnostic plots on regression analysis when the researcher is confronted with a very large data set.

With large databases, there are problems associated with the use of traditional statistical models and EDA offers instruments that can help researchers overcome some of these issues. These methods complement predictive approaches to big data, such as machine learning, random forest. Although it should be noted that EDA methods have limitations, as they do not permit statistical confirmation or refusal of a causal hypothesis. For academic researchers, the EDA methods presented should contribute with intuitive and simple tools to extract useful information from the data, regarding issues such as assessing across-group variation, latent heterogeneity of the data, and post-model-fit analysis. This information should help researchers uncover aspects that contribute to a more solid descriptive analysis and more accurate statistical modeling.

The possible extensions of the EDA methods discussed in this article for large databases are many. We have concentrated just on few features which we feel are immediately available to practitioners using the current free software. One area where we are currently working, and that we feel is also a promising avenue for research, is the visual exploration of longitudinal data in the case of large sample size. The heat-map methods discussed above can serve to detect data heterogeneity on the longitudinal evolution of crucial variables, or residuals of a post model fit.

Authors' Note

The code of the implementation in R of the methods of the paper is provided in the Supplementary Material to this paper.

Acknowledgements

The authors are grateful to the editors, the reviewers, and Stuart Redding for his insightful comments, and the participants at the Big data conferences at UAB 2017 and in Lisbon 2017.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

Notes

1. We use the wave of 2010, where 722,957 individuals were included. To ease the representation of this database, we consider only wages for people aged between 15 and 65 years. More information here http://www.seg-social.es/prdi00/groups/public/documents/binario/190489.pdf
2. For more details see https://docs.google.com/viewer?url=http%3A%2F%2Fwww.mitramiss.gob.es%2Fes%2Fguia%2Fpdfs%2FEvolucixn_bases_cotizacixn_2019.pdf
3. We grouped ages using 10 years interval. Usually this also corresponds to the structure of labor market, as we can see through the graphs.
4. Several approaches for points to be declared the status of outliers have been in the literature (see, for example, Bruffaerts et al., 2014)
5. In the previous section, we used the log transformation of daily wage to symmetrize the distribution. This log transformation is the one mostly used in labor and has the effect of improving the comparativeness of the boxplots displays (Hubert & Vandervieren, 2008).
6. This is a graph that has been produced using the ggplot2 function of the free software R.
7. By modal concentration, we mean a local high concentration of cases in a given point; note that it does not refer to any summary level measure unlike the median or the mean.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87–93.

Amuedo-Dorantes, C., & De la Rica, S. (2006). The role of segregation and pay structure on the gender wage gap: Evidence from matched employer-employee data for Spain. *The BE Journal of Economic Analysis and Policy*, *5*(1), 1498.

Barter, R. L., & Yu, B. (2018). Superheat: An R package for creating beautiful and extendable heat-maps for visualizing complex data. *Journal of Computational and Graphical Statistics*, *27*(4), 910–922.

Blundell, R., & Meghir, C. (1987). Bivariate alternatives to the Tobit model. *Journal of Econometrics*, *34*(1–2), 179–200.

Bruffaerts, C., Verardi, V., & Vermandele, C. (2014). A generalized boxplot for skewed and heavy–tailed distributions. *Statistics and Probability Letters*, *95*, 110–117.

Cabrales, A., Dolado, J. J., & Mora, R. (2017). Dual labour markets and (lack of) on-the-job training: Evidence for

Spain using PIAAC data. *SERIEs, Journal of the Spanish Economic Association*, *8*, 345–371.

Camacho, J., Prez-Villegas, A., Rodrguez-Gmez, R. A., & Jimnez-Maas, E. (2015). Multivariate exploratory data analysis (MEDA) toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, *143*, 49–57.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Wadsworth & Brooks/Cole.

Cheng, S., Shi, Y., Qin, Q., & Bai, R. (2013, October). Swarm intelligence in big data analytics. In H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise . . . X. Yao (Eds.), *International conference on intelligent data engineering and automated learning* (pp. 417–426). Springer.

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review 65*(3), 122–135.

Dolado, J. J., Garcia-Serrano, C., & Jimeno, J. F. (2002). Drawing lessons from the boom of temporary jobs in Spain. *The Economic Journal*, *112*(480), F270–F295.

Downey, A. (2014). *Think Stats: Exploratory data analysis*. O'Reilly Media.

Garcia-Perez, J. I. (2008). The Spanish social security data (MCVL): a user guide for the analysis of transitions. *Revista de Economia Aplicada*, *16*(1), 5–28.

Gehrke, B., & Weber, E. (2018). Identifying asymmetric effects of labor market reforms. *European Economic Review*, *110*, 18–40.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy Management Journal*, *57*, 321–326.

Healey, C. G., & Enns, J. T. (2002). Perception and painting: A search for effective, engaging visualizations. IEEE Computer Graphics and *Applications*, *22*(2), 10–15.

Heckman, J., Lochner, L., & Todd, P. (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. *Handbook of the Economics of Education*, *1*, 307–458.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis* (Vol. 3). Wiley.

Hubert, M., & Vandervieren, E. (2008).An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, *52*(12), 5186–5201.

Hussain, K., & Prieto, E. (2016). Big data in the finance and insurance sectors. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy* (pp. 209–223). Springer.

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265–276.

Levine, S. S. (2018). Show us your data: Connect the dots, improve science. *Management and Organization Review*, *14*(2), 433–437.

Myatt, G. J., & Johnson, W. P. (2014). *Making sense of data I: A practical guide to exploratory data analysis and data mining* (2nd ed.). John Wiley & Sons.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 67.

Schwabish, J. A. (2014). An economist's guide to visualizing data. *Journal of Economic Perspectives*, *28*(1), 209–234.

Tukey, J. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wilkinson, L., & Friendly, M. (2009). *The history of the cluster heat-map*. The American Statistician.