Chapter 15
Exploratory Data Analysis

Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli and Yves Crutain

Learning Objectives

- Why is EDA important during the initial exploration of a dataset?
- What are the most essential tools of graphical and non-graphical EDA?

15.1 Introduction

Exploratory data analysis (EDA) is an essential step in any research analysis. The primary aim with exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualizing and understanding the data usually through graphical representation [1]. EDA aims to assist the natural patterns recognition of the analyst. Finally, feature selection techniques often fall into EDA. Since the seminal work of Tukey in 1977, EDA has gained a large following as the gold standard methodology to analyze a data set [2, 3]. According to Howard Seltman (Carnegie Mellon University), "loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis" [4].

EDA is a fundamental early step after data collection (see Chap. 11) and pre-processing (see Chap. 12), where the data is simply visualized, plotted, manipulated, without any assumptions, in order to help assessing the quality of the data and building models. "Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to explore, and graphics gives the analysts unparalleled power to do so, while being ready to gain insight into the data. There are many ways to categorize the many EDA techniques" [5].

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-43742-2_15) contains supplementary material, which is available to authorized users.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_15

The interested reader will find further information in the textbooks of Hill and Lewicki [6] or the NIST/SEMATECH e-Handbook [1]. Relevant R packages are available on the CRAN website [7].

The objectives of EDA can be summarized as follows:

1. Maximize insight into the database/understand the database structure;
2. Visualize potential relationships (direction and magnitude) between exposure and outcome variables;
3. Detect outliers and anomalies (values that are significantly different from the other observations);
4. Develop parsimonious models (a predictive or explanatory model that performs with as few exposure variables as possible) or preliminary selection of appropriate models;
5. Extract and create clinically relevant variables.

EDA methods can be cross-classified as:

- Graphical or non-graphical methods
- Univariate (only one variable, exposure or outcome) or multivariate (several exposure variables alone or with an outcome variable) methods.

15.2 Part 1—Theoretical Concepts

15.2.1 Suggested EDA Techniques

Tables 15.1 and 15.2 suggest a few EDA techniques depending on the type of data and the objective of the analysis.

Table 15.1 Suggested EDA techniques depending on the type of data

Type of data	Suggested EDA techniques
Categorical	Descriptive statistics
Univariate continuous	Line plot, Histograms
Bivariate continuous	2D scatter plots
2D arrays	Heatmap
Multivariate: trivariate	3D scatter plot or 2D scatter plot with a 3rd variable represented in different color, shape or size
Multiple groups	Side-by-side boxplot

Table 15.2 Most useful EDA techniques depending on the objective

Objective	Suggested EDA techniques
Getting an idea of the distribution of a variable	Histogram
Finding outliers	Histogram, scatterplots, box-and-whisker plots
Quantify the relationship between two variables (one exposure and one outcome)	2D scatter plot +/curve fitting Covariance and correlation
Visualize the relationship between two exposure variables and one outcome variable	Heatmap
Visualization of high-dimensional data	t-SNE or PCA + 2D/3D scatterplot

t-SNE t-distributed stochastic neighbor embedding, *PCA* Principal component analysis

Table 15.3 Example of tabulation table

	Group count	Frequency (%)
Green ball	15	75
Red ball	5	25
Total	20	100

15.2.2 Non-graphical EDA

These non-graphical methods will provide insight into the characteristics and the distribution of the variable(s) of interest.

Univariate Non-graphical EDA

Tabulation of Categorical Data (Tabulation of the Frequency of Each Category)

A simple univariate non-graphical EDA method for categorical variables is to build a table containing the count and the fraction (or frequency) of data of each category. An example of tabulation is shown in the case study (Table 15.3).

Characteristics of Quantitative Data: Central Tendency, Spread, Shape of the Distribution (Skewness, Kurtosis)

Sample statistics express the characteristics of a sample using a limited set of parameters. They are generally seen as estimates of the corresponding population parameters from which the sample comes from. These characteristics can express the central tendency of the data (arithmetic mean, median, mode), its spread (variance, standard deviation, interquartile range, maximum and minimum value) or some features of its distribution (skewness, kurtosis). Many of those characteristics can easily be seen qualitatively on a histogram (see below). Note that these characteristics can only be used for quantitative variables (not categorical).

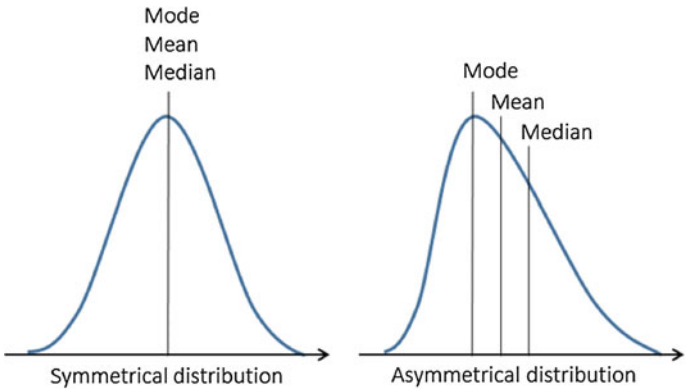

Fig. 15.1 Symmetrical versus asymmetrical (skewed) distribution, showing mode, mean and median

Central tendency parameters

The arithmetic mean, or simply called the mean is the sum of all data divided by the number of values. The median is the middle value in a list containing all the values sorted. Because the median is affected little by extreme values and outliers, it is said to be more "robust" than the mean (Fig. 15.1).

Variance

When calculated on the entirety of the data of a population (which rarely occurs), the variance σ^2 is obtained by dividing the sum of squares by n, the size of the population.

The sample formula for the variance of observed data conventionally has n-1 in the denominator instead of n to achieve the property of "unbiasedness", which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here σ^2). s^2 is an unbiased estimator of the population variance σ^2.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{(n-1)} \tag{15.1}$$

The standard deviation is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable.

The sample standard deviation is usually represented by the symbol s. For a theoretical Gaussian distribution, mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7 % of the probability density, respectively.

Interquartile range (IQR)
The IQR is calculated using the boundaries of data situated between the 1st and the 3rd quartiles. Please refer to the Chap. 13 "Noise versus Outliers" for further detail about the IQR.

$$IQR = Q_3 - Q_1 \tag{15.2}$$

In the same way that the median is more robust than the mean, the IQR is a more robust measure of spread than variance and standard deviation and should therefore be preferred for small or asymmetrical distributions.

Important rule:

- **Symmetrical distribution** (not necessarily normal) **and N > 30**: express results as mean ± standard deviation.
- **Asymmetrical distribution or N < 30 or evidence for outliers:** use median ± IQR, which are more robust.

Skewness/kurtosis
Skewness is a measure of a distribution's asymmetry. Kurtosis is a summary statistic communicating information about the tails (the smallest and largest values) of the distribution. Both quantities can be used as a means to communicate information about the distribution of the data when graphical methods cannot be used. More information about these quantities can be found in [9]).

Summary
We provide as a reference some of the common functions in R language for generating summary statistics relating to measures of central tendency (Table 15.4).

Testing the Distribution

Several non-graphical methods exist to assess the normality of a data set (whether it was sampled from a normal distribution), like the Shapiro-Wilk test for example. Please refer to the function called "Distribution" in the GitHub repository for this book (see code appendix at the end of this Chapter).

Table 15.4 Main R functions for basic measure of central tendencies and variability

Function	Description
summary(x)	General description of a vector
max(x)	Maximum value
mean(x)	Average or mean value
median(x)	Median value
min(x)	Smallest value
sd(x)	Standard deviation
var(x)	Variance, measure the spread or dispersion of the values
IQR(x)	Interquartile range

Finding Outliers

Several statistical methods for outlier detection fall into EDA techniques, like Tukey's method, Z-score, studentized residuals, etc [8]. Please refer to the Chap. 14 "Noise versus Outliers" for more detail about this topic.

Multivariate Non-graphical EDA

Cross-Tabulation

Cross-tabulation represents the basic bivariate non-graphical EDA technique. It is an extension of tabulation that works for categorical data and quantitative data with only a few variables. For two variables, build a two-way table with column headings matching the levels of one variable and row headings matching the levels of the other variable, then fill in the counts of all subjects that share a pair of levels. The two variables may be both exposure, both outcome variables, or one of each.

Covariance and Correlation

Covariance and correlation measure the degree of the relationship between two random variables and express how much they change together (Fig. 15.2).

The covariance is computed as follows:

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{15.3}$$

where x and y are the variables, n the number of data points in the sample, $\bar{x}$ the mean of the variable x and $\bar{y}$ the mean of the variable y.

A positive covariance means the variables are positively related (they move together in the same direction), while a negative covariance means the variables are inversely related. A problem with covariance is that its value depends on the scale of the values of the random variables. The larger the values of x and y, the larger the

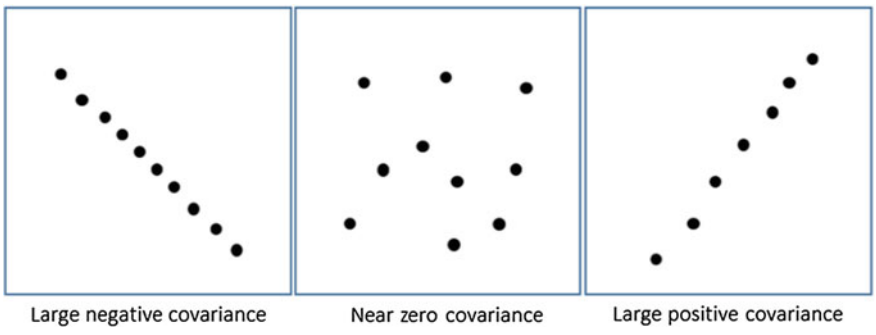

Fig. 15.2 Examples of covariance for three different data sets

covariance. It makes it impossible for example to compare covariances from data sets with different scales (e.g. pounds and inches). This issue can be fixed by dividing the covariance by the product of the standard deviation of each random variable, which gives Pearson's correlation coefficient.

Correlation is therefore a scaled version of covariance, used to assess the linear relationship between two variables and is calculated using the formula below.

$$Cor(x, y) = \frac{Cov(x, y)}{s_x s_y} \tag{15.4}$$

where $Cov(x, y)$ is the covariance between x and y and s_x, s_y are the sample standard deviations of x and y.

The significance of the correlation coefficient between two normally distributed variables can be evaluated using Fisher's z transformation (see the cor.test function in R for more details). Other tests exist for measuring the non-parametric relationship between two variables, such as Spearman's rho or Kendall's tau.

15.2.3 Graphical EDA

Univariate Graphical EDA

Histograms

Histograms are among the most useful EDA techniques, and allow you to gain insight into your data, including distribution, central tendency, spread, modality and outliers.

Histograms are bar plots of counts versus subgroups of an exposure variable. Each bar represents the frequency (count) or proportion (count divided by total count) of cases for a range of values. The range of data for each bar is called a bin. Histograms give an immediate impression of the shape of the distribution (symmetrical, uni/multimodal, skewed, outliers…). The number of bins heavily influences the final aspect of the histogram; a good practice is to try different values, generally from 10 to 50. Some examples of histograms are shown below as well as in the case studies. Please refer to the function called "Density" in the GitHub repository for this book (see code appendix at the end of this Chapter) (Figs. 15.3 and 15.4).

Histograms enable to confirm that an operation on data was successful. For example, if you need to log-transform a data set, it is interesting to plot the histogram of the distribution of the data before and after the operation (Fig. 15.5).

Histograms are interesting for finding outliers. For example, pulse oximetry can be expressed in fractions (range between 0 and 1) or percentage, in medical records. Figure 15.6 is an example of a histogram showing the distribution of pulse oximetry, clearly showing the presence of outliers expressed in a fraction rather than as a percentage.

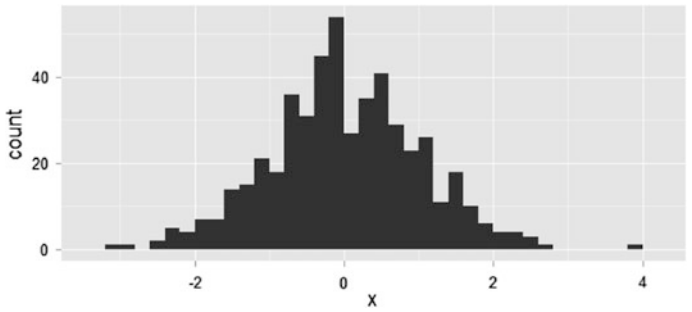

Fig. 15.3 Example of histogram

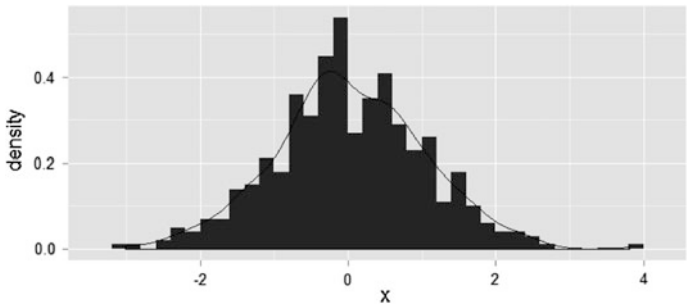

Fig. 15.4 Example of histogram with density estimate

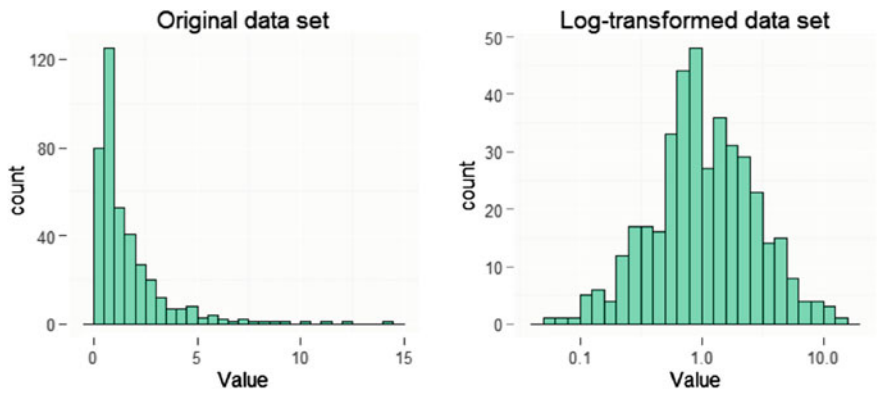

Fig. 15.5 Example of the effect of a log transformation on the distribution of the dataset

Stem Plots

Stem and leaf plots (also called stem plots) are a simple substitution for histograms. They show all data values and the shape of the distribution. For an example, Please refer to the function called "Stem Plot" in the GitHub repository for this book (see code appendix at the end of this Chapter) (Fig. 15.7).

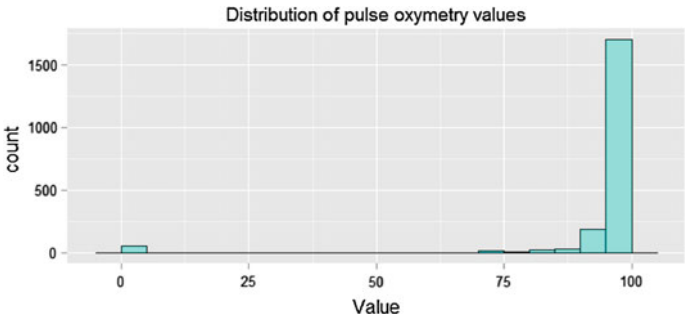

Fig. 15.6 Distribution of pulse oximetry

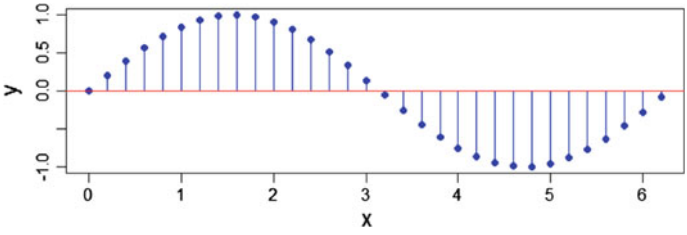

Fig. 15.7 Example of stem plot

Boxplots

Boxplots are interesting for representing information about the central tendency, symmetry, skew and outliers, but they can hide some aspects of the data such as multimodality. Boxplots are an excellent EDA technique because they rely on robust statistics like median and IQR.

Figure 15.8 shows an annotated boxplot which explains how it is constructed. The central rectangle is limited by Q1 and Q3, with the middle line representing the median of the data. The whiskers are drawn, in each direction, to the most extreme point that is less than 1.5 IQR beyond the corresponding hinge. Values beyond 1.5 IQR are considered outliers.

The "outliers" identified by a boxplot, which could be called "boxplot outliers" are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This does not by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. It is also important to realize that the number of boxplot outliers depends strongly on the size of the sample. In fact, for data that is perfectly normally distributed, we expect 0.70 % (about 1 in 140 cases) to be "boxplot outliers", with approximately half in either direction.

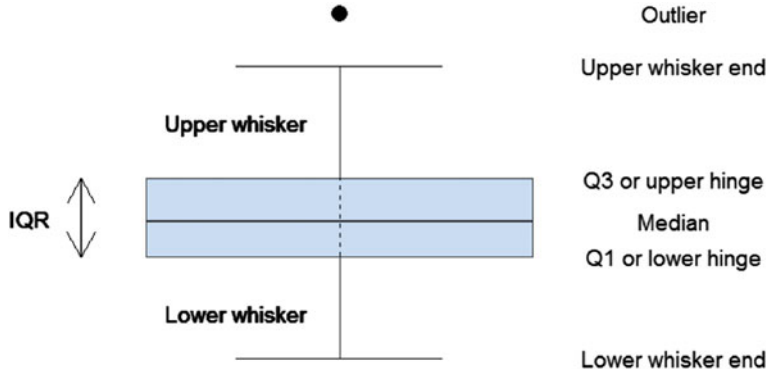

Fig. 15.8 Example of boxplot with annotations

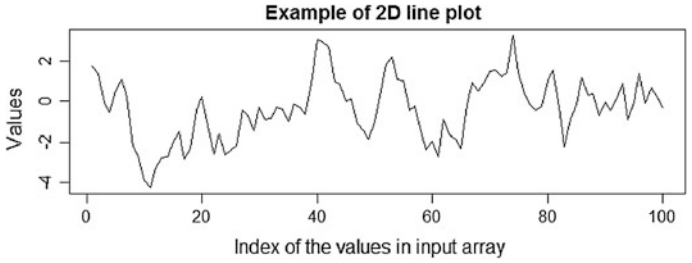

Fig. 15.9 Example of 2D line plot

2D Line Plot

2D line plots represent graphically the values of an array on the y-axis, at regular intervals on the x-axis (Fig. 15.9).

Probability Plots (Quantile-Normal Plot/QN Plot, Quantile-Quantile Plot/QQ Plot)

Probability plots are a graphical test for assessing if some data follows a particular distribution. They are most often used for testing the normality of a data set, as many statistical tests have the assumption that the exposure variables are approximately normally distributed. These plots are also used to examine residuals in models that rely on the assumption of normality of the residuals (ANOVA or regression analysis for example).

The interpretation of a QN plot is visual (Fig. 15.10): either the points fall randomly around the line (data set normally distributed) or they follow a curved pattern instead of following the line (non-normality). QN plots are also useful to identify skewness, kurtosis, fat tails, outliers, bimodality etc.

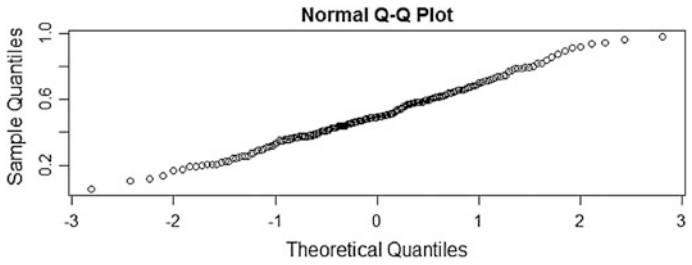

Fig. 15.10 Example of QQ plot

Besides the probability plots, there are many quantitative statistical tests (not graphical) for testing for normality, such as Pearson Chi^2, Shapiro-Wilk, and Kolmogorov-Smirnov.

Deviation of the observed distribution from normal makes many powerful statistical tools useless. Note that some data sets can be transformed to a more normal distribution, in particular with log-transformation and square-root transformations. If a data set is severely skewed, another option is to discretize its values into a finite set.

Multivariate Graphical EDA

Side-by-Side Boxplots

Representing several boxplots side by side allows easy comparison of the characteristics of several groups of data (example Fig. 15.11). An example of such boxplot is shown in the case study.

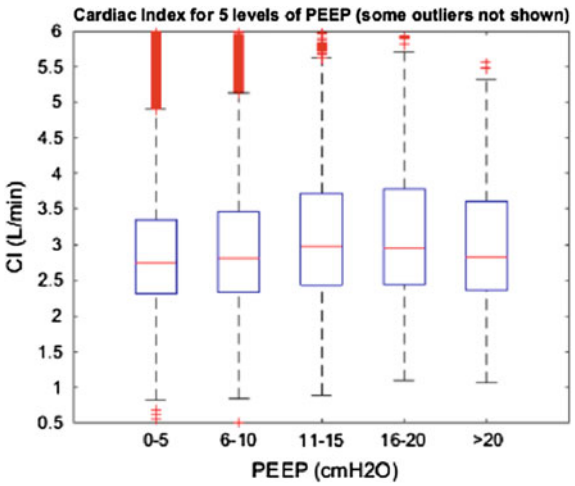

Fig. 15.11 Side-by-side boxplot showing the cardiac index for five levels of Positive end-expiratory pressure (PEEP)

Scatterplots

Scatterplots are built using two continuous, ordinal or discrete quantitative variables (Fig. 15.12). Each data point's coordinate corresponds to a variable. They can be complexified to up to five dimensions using other variables by differentiating the data points' size, shape or color.

Scatterplots can also be used to represent high-dimensional data in 2 or 3D (Fig. 15.13), using T-distributed stochastic neighbor embedding (t-SNE) or principal component analysis (PCA). t-SNE and PCA are dimension reduction features used to reduce complex data set in two (t-SNE) or more (PCA) dimensions.

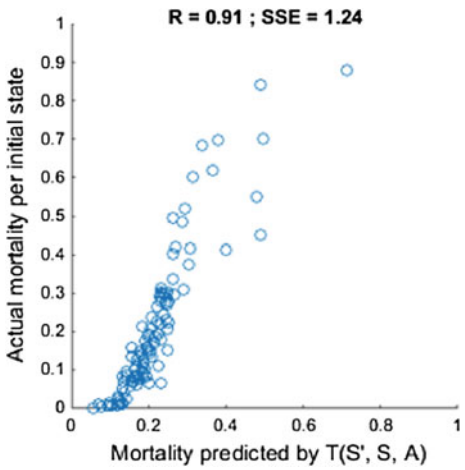

Fig. 15.12 Scatterpolot showing an example of actual mortality per rate of predicted mortality

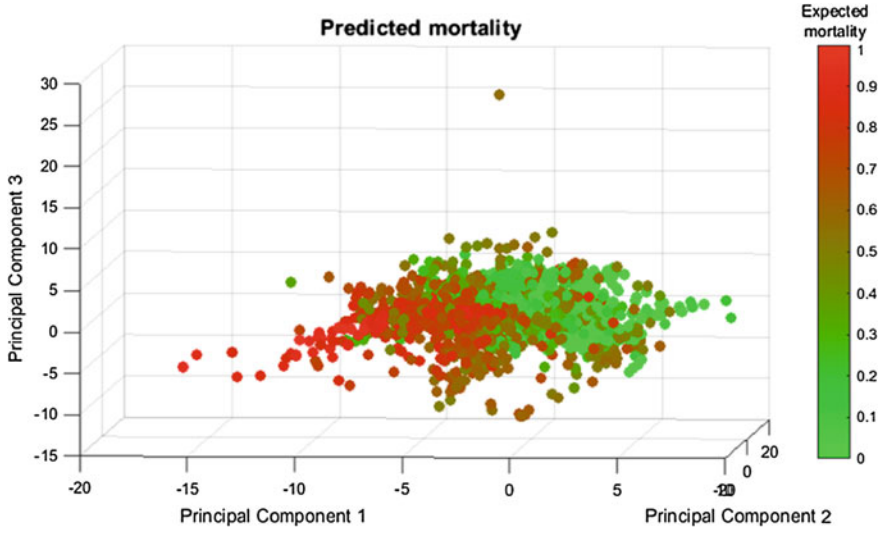

Fig. 15.13 3D representation of the first three dimension of a PCA

For binary variables (e.g. 28-day mortality vs. SOFA score), 2D scatterplots are not very helpful (Fig. 15.14, left). By dividing the data set in groups (in our example: one group per SOFA point), and plotting the average value of the outcome in each group, scatterplots become a very powerful tool, capable for example to identify a relationship between a variable and an outcome (Fig. 15.14, right).

Curve Fitting

Curve fitting is one way to quantify the relationship between two variables or the change in values over time (Fig. 15.15). The most common method for curve fitting relies on minimizing the sum of squared errors (SSE) between the data and the

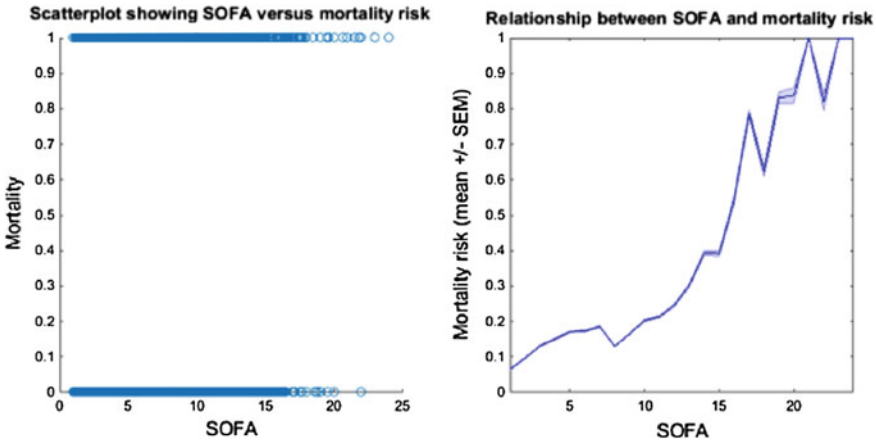

Fig. 15.14 Graphs of SOFA versus mortality risk

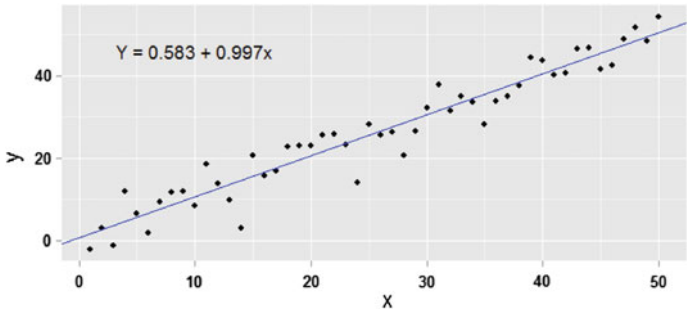

Fig. 15.15 Example of linear regression

fitted function. Please refer to the "Linear Fit" function to create linear regression slopes in R.

More Complicated Relationships

Many real life phenomena are not adequately explained by a straight-line relationship. An always increasing set of methods and algorithms exist to deal with that issue. Among the most common:

- Adding transformed explanatory variables, for example, adding x^2 or x^3 to the model.
- Using other algorithms to handle more complex relationships between variables (e.g., generalized additive models, spline regression, support vector machines, etc.).

Heat Maps and 3D Surface Plots

Heat maps are simply a 2D grid built from a 2D array, whose color depends on the value of each cell. The data set must correspond to a 2D array whose cells contain the values of the outcome variable. This technique is useful when you want to represent the change of an outcome variable (e.g. length of stay) as a function of two other variables (e.g. age and SOFA score).

The color mapping can be customized (e.g. rainbow or grayscale). Interestingly, the Matlab function *imagesc* scales the data to the full colormap range. Their 3D equivalent is mesh plots or surface plots (Fig. 15.16).

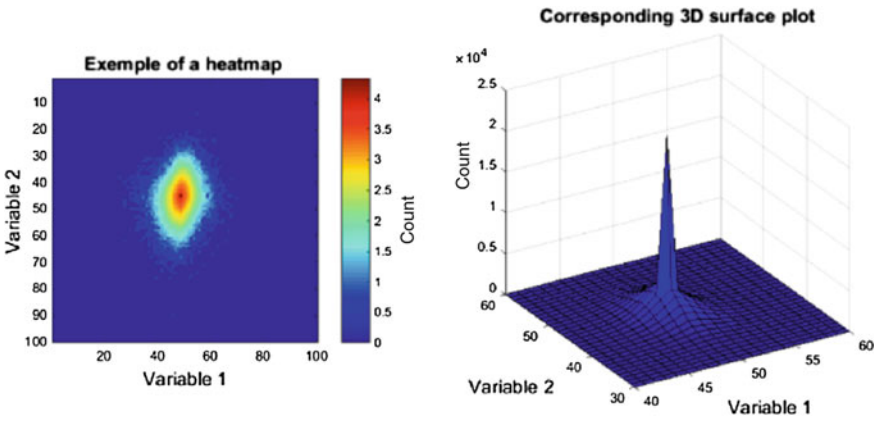

Fig. 15.16 Heat map (*left*) and surface plot (*right*)

15.3 Part 2—Case Study

This case study refers to the research that evaluated the effect of the placement of indwelling arterial catheters (IACs) in hemodynamically stable patients with respiratory failure in intensive care, from the MIMIC-II database.

For this case study, several aspects of EDA were used:

- The categorical data was first tabulated.
- Summary statistics were then generated to describe the variables of interest.
- Graphical EDA was used to generate histograms to visualize the data of interest.

15.3.1 Non-graphical EDA

Tabulation

To analyze, visualize and test for association or independence of categorical variables, they must first be tabulated. When generating tables, any missing data will be counted in a separate "NA" ("Not Available") category. Please refer to the Chap. 13 "Missing Data" for approaches in managing this problem. There are several methods for creating frequency or contingency tables in R, such as for example, tabulating outcome variables for mortality, as demonstrated in the case study. Refer to the "Tabulate" function found in the GitHub repository for this book (see code appendix at the end of this Chapter) for details on how to compute frequencies of outcomes for different variables.

Statistical Tests

Multiple statistical tests are available in R and we refer the reader to the Chap. 16 "Data Analysis" for additional information on use of relevant tests in R. For examples of a simple Chi-square…" as "For examples of a simple Chi-squared test, please refer to the "Chi-squared" function found in the GitHub repository for this book (see code appendix at the end of this Chapter). In our example, the hypothesis of independence between expiration in ICU and IAC is accepted ($p > 0.05$). On the contrary, the dependence link between day-28 mortality and IAC is rejected.

Summary statistics

Summary statistics as described above include, frequency, mean, median, mode, range, interquartile range, maximum and minimum values. An extract of summary statistics of patient demographics, vital signs, laboratory results and comorbidities, is shown in Table 6. Please refer to the function called "EDA Summary" in the

Table 15.5 Comparison between the two study cohorts (subsample of variables only)

Variables	Entire Cohort (N = 1776)		
	Non-IAC	IAC	*p*-value
Size	984 (55.4 %)	792 (44.6 %)	NA
Age (year)	51 (35–72)	56 (40–73)	0.009
Gender (female)	344 (43.5 %)	406 (41.3 %)	0.4
Weight (kg)	76 (65–90)	78 (67–90)	0.08
SOFA score	5 (4–6)	6 (5–8)	<0.0001
Co-morbidities			
CHF	97 (12.5 %)	116 (11.8 %)	0.7
…	…	…	…
Lab tests			
WBC	10.6 (7.8–14.3)	11.8 (8.5–15.9)	<0.0001
Hemoglobin (g/dL)	13 (11.3–14.4)	12.6 (11–14.1)	0.003
…	…	…	…

GitHub repository for this book (see code appendix at the end of this Chapter) (Table 15.5).

When separate cohorts are generated based on a common variable, in this case the presence of an indwelling arterial catheter, summary statistics are presented for each cohort.

It is important to identify any differences in subject baseline characteristics. The benefits of this are two-fold: first it is useful to identify potentially confounding variables that contribute to an outcome in addition to the predictor (exposure) variable. For example, if mortality is the outcome variable then differences in severity of illness between cohorts may wholly or partially account for any variance in mortality. Identifying these variables is important as it is possible to attempt to control for these using adjustment methods such as multivariable logistic regression. Secondly, it may allow the identification of variables that are associated with the predictor variable enriching our understanding of the phenomenon we are observing.

The analytical extension of identifying any differences using medians, means and data visualization is to test for statistically significant differences in any given subject characteristic using for example Wilcoxon-Rank sum test. Refer to Chap. 16 for further details in hypothesis testing.

15.3.2 Graphical EDA

Graphical representation of the dataset of interest is the principle feature of exploratory analysis.

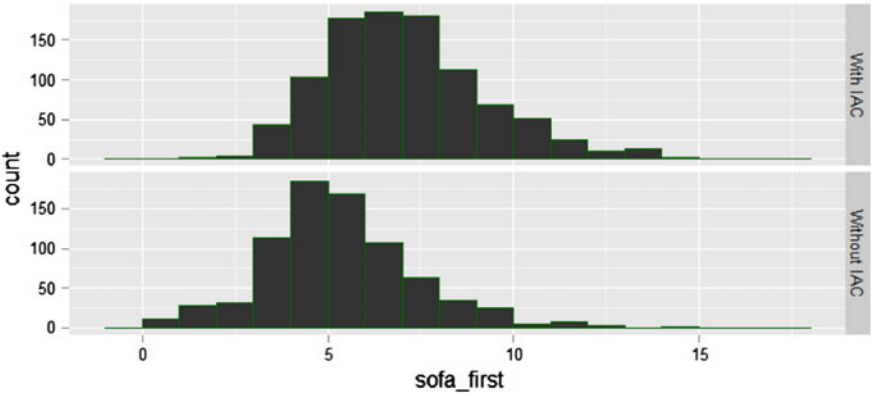

Fig. 15.17 histograms of SOFA scores by intra-arterial catheter status

Histograms

Histograms are considered the backbone of EDA for continuous data. They can be used to help the researcher understand continuous variables and provide key information such as their distribution. Outlined in *noise and outliers,* the histogram allows the researcher to visualize where the bulk of the data points are placed between the maximum and minimum values. Histograms can also allow a visual comparison of a variable between cohorts. For example, to compare severity of illness between patient cohorts, histograms of SOFA score can be plotted side by side (Fig. 15.17). An example of this is given in the code for this chapter using the "side-by-side histogram" function (see code appendix at the end of this Chapter).

Boxplot and ANOVA

Outside of the scope of this case study, the user may be interested in analysis of variance. When performing EDA and effective way to visualize this is through the use of boxplot. For example, to explore differences in blood pressure based on severity of illness subjects could be categorized by severity of illness with blood pressure values at baseline plotted (Fig. 15.18). Please refer to the function called "Box Plot" in the GitHub repository for this book (see code appendix at the end of this Chapter).

The box plot shows a few outliers which may be interesting to explore individually, and that people with a high SOFA score (>10) tend to have a lower blood pressure than people with a lower SOFA score.

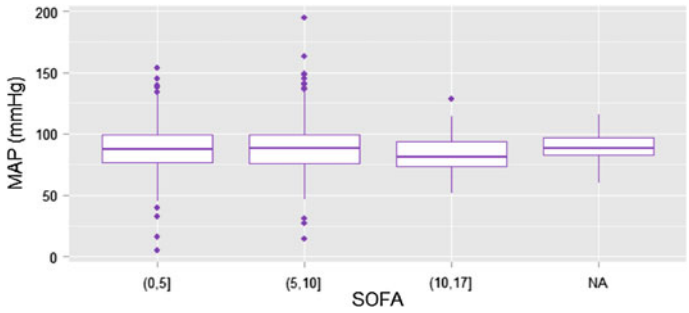

Fig. 15.18 Side-by-side boxplot of MAP for different levels of severity at admission

15.4 Conclusion

In summary, EDA is an essential step in many types of research but is of particular use when analyzing electronic health care records. The tools described in this chapter should allow the researcher to better understand the features of a dataset and also to generate novel hypotheses.

Take Home Messages

1. Always start by exploring a dataset with an open mind for discovery.
2. EDA allows to better apprehend the features and possible issues of a dataset.
3. EDA is a key step in generating research hypothesis.

Code Appendix

The code used in this chapter is available in the GitHub repository for this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website.

References

1. Natrella M (2010) NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH
2. Mosteller F, Tukey JW (1977) Data analysis and regression. Addison-Wesley Pub. Co., Boston
3. Tukey J (1977) Exploratory data analysis. Pearson, London
4. Seltman HJ (2012) Experimental design and analysis. Online http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf
5. Kaski, Samuel (1997) "Data exploration using self-organizing maps." *Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82. 1997.*
6. Hill T, Lewicki P (2006) Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft, Inc., Tulsa
7. CRAN (2016) The Comprehensive R archive network—packages. Contributed Packages, 10 Jan 2016 [Online]. Available: https://cran.r-project.org/web/packages/. Accessed: 10 Jan 2016
8. Grubbs F (1969) Procedures for detecting outlying observations in samples. Technometrics 11(1)
9. Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. The Statistician 47:183–189.