Introduction to Exploratory Data Analysis (EDA)

Using python.

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

Introduction

Take better decisions...

- Determine what variables indicate a higher risk of bankruptcy
- Assest a medical treatment or new drug
- Design better industrial processes
- ...

- Econometrics
- Physics
- Artificial intelligence
- ...

Statistical concepts

Population

An a whole set of **objects** or **persons** in which we are interested

Sample

Any subset of the population.

Type of variables

Qualitative

Measure a characteristic as *gender, religion, profesion, and so on*

Quantitative

Measure an amount or quantity, for instance, the montly waste, average daily minutes in social networks.

Quantitative variables

Discrete (finite)

Can take only integers values

$$X = 0, 1, 2, 3, ... \tag{1}$$

For instance, the number of clicks in a web page, the number of tickets in cinema.

Continuous (infintely)

The variables is fractionable, for instance the weight, the age.

Scales of measurement

Assume that a variable X have two measurements, x_1 and x_2

Properties - Operations

1. $$\frac{x1}{x2}$$
2. $$x1 - x2$$
3. $$x_1 \geq x_2 \text{ or } x_1 \leq x_2$$

Scales of measurement

...

- **Ratio scale variable** fill the properties 1,2 and 3, for instance; consumption, weight, and so on...
- **Interval variable** fill the properties 2 and 3, for instance the year index and IQ test.
- **Ordinal variable** fill only the third property, for instance; the rating of movies in Netflix.
- **Nominal variable** don't fill any property, and only have sense ask if two measurements are different, for instance: political party.

Ratio vs Scale

A natural question will be: **in which differ ratio and interval variable?** An interval variable do not have a real zero, for instance temperature measured in Celsius take as zero the boiled down of water (an ambiguous point), zero not means absence of value. Unlike distance in any unity of measure zero means no-distance.

Scale not support multiplication or division

$$y = \lambda + x \text{ (changes from x)} \tag{2}$$

$$y = \lambda x \text{ (changes from 0)} \tag{3}$$

For what?

This allow us uses statistics adequately, for instance not pretend uses a central tendency measure as mean over a nominal (categorical) variable or neither ratios of interval variables or calculate over they coefficient of variation.

Qualitative description

For the X *nominal* variable, that have k possible unique values, then the i category could be describe as:

Relative Frequency (*freq*)

$$freq_i = \frac{n_i}{N} \tag{4}$$

where n_i is the number of times that appear (i) named as *absolute frequency* in the variable and N is the total of observations or measurements.

you can get the percentage multiplying $freq_i * 100$.

Mean

Mean mathematical definition

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5}$$

Range

Range is a measure of variablity

Mathematical definition

for a X variable then:

$$Range = max(X) - min(X) \tag{6}$$

Consider that it is only a extreme measure.

Variance

Population variance mathematical definition

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \tag{7}$$

Sample variance mathematical definition

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \tag{8}$$

Standard deviation

square roof of variance

$$\sigma = \sqrt{\sigma^2} \tag{9}$$

This will be important to standardization and used some important definitions as **Chebyshev Theorem** and **Empirical Rule**.

CV

The problem with standard devaition is that not is possible determine how big is.

Variation coefficient

$$CV = \frac{S_x}{\bar{x}} \tag{10}$$

Considerations

- Is unitless for instance allow us compare meters and feets.
- Applicable on ratio variables

Quartiles

Three values that divide the data in 4 evenly spaced intervals.

Q_1, Q_2 and Q_3

- $Q_1 = P_{25}$
- $Q_2 = P_{50}$ also known as median.
- $Q_3 = P_{75}$

Interpretation

Assuming that $Q_3 = b$ means that 75% of data have a lesser value than b.

Interquantile Range and boxplot