Olivier Maugain

Reading Data

Course Notes

365 DataScience

Table of Contents

ABSTRACT

Data as a language requires us to assess its quality and accuracy in order to correctly execute a task. Much like spelling mistakes in a resume and other important personal documents can negatively affect an individual's lifestyle and career progression, poor data quality can be detrimental for a company's success.

A famous example of this is NASA losing the Mars Climate Orbiter, worth $125 million, due to the engineering teams using different systems of measurement

Therefore, before we start interpreting data and making any informed business decisions based on it, we first need to assess its quality. In these data literacy course notes on reading data, we are going to answer questions like:

- What defines the quality of our data?
- What are the consequences of poor data?
- What are the key data flaws?
- What are the methods of descriptive statistics?
- How to calculate measurements of central tendency?

Keywords: data, data literacy, data description, data quality assessment, descriptive statistics

1. Reading Data

In some areas of life, the right questions are more important than the right answers. This also applies to working with data. It will often be the case that the ultimately beneficiary of data (for example, the decision makers) is different from the person who processed these data and produced the results (or the answers).

- Although it is important to be able to trust one's own employees and partners, it never harms to get a sound understanding about what was done with these data, i.e., how these were collected, manipulated, cleaned, transformed, analyzed, and turned into insights
- One needs to be able to "read" the data, which includes the assessment of their quality and their description in statistical terms

1.1. Data quality assessment

Data quality is the *state of the information at hand*

Acceptable (or high) data quality - if the data are fit for their intended uses (e.g., planning, decision making, insights generation, statistical analysis, machine learning, etc.)

Poor data quality (e.g., if the data contain errors) - decision makers could be misled or algorithms will not learn efficiently, or perhaps even worse, learn the wrong things.

* “Garbage in, garbage out” (**GIGO**) - Without meaningful data, you cannot get meaningful results

The impact that poor data quality can be significant. Financial expenses, productivity losses, missed opportunities and reputational damage are but some of the costs that poor data quality carry.

**In 2016, IBM estimated that businesses were losing, in the US alone, $3.1 trillion every year due to poor data quality.*

Real examples of blunders (what can go wrong with data):

- A customer makes a spelling mistake when writing his address in the paper registration form
- A clerk misreads the handwriting of the customer and fills in the wrong address into the system
- The CRM system only has space for 40 characters in the “address” field and cuts every character beyond this limit
- The migration tool cannot read special characters, using placeholders instead
- A data analyst makes a mistake when joining two data sets, leading to a mismatch of attributes for all records

Key data flaws:

Incomplete data: This is when there are missing values in the data set.

- If there are too many missing values for one attribute, the latter becomes useless and needs to be discarded from the analysis

- If only a small percentage is missing, then we can eliminate the records with the missing values or make an assumption about what the values could be
- Business decision makers should ask the following questions to the people who processed the data:
- What is the proportion of missing values for each attribute?
- How were the missing values replaced?

Inaccurate data: This can happen in many ways, for examples:

- Spelling mistakes
- Inconsistent units
- A famous case is that of NASA's Mars Climate Orbiter, which in 1999 was unintentionally destroyed during a mission in 1999. The loss was due to a piece of software that used the wrong unit of impulse (pound-force seconds instead of the SI units of newton-seconds). The total cost of the mistake was estimated at $327.6 million
- Inconsistent formats
- E.g., when numbers are declared as text, which makes them unreadable for some programs
- Impossible values
  - A negative age (for a person)
  - A height of 5.27 meters for a human
  - A household size of 4.8 (for one family)
  - A future birthday (for someone already born)

  - An employer called “erufjdskdfnd”
  - “999” years as a contract duration
- Unusually high or low value does not necessarily have to be inaccurate, but could also be an outlier (i.e., a data point that differs significantly from other observations)

1.2. Data description

When communicating about or with data, it is essential to be able to describe these.

- Although it is not always possible to go through them one by one, there are ways to explore data and to get an overview about what is available.
- The objective is to single out possible issues or patterns that might be worth digging further into
- A description of the data also helps determine how useful they can be to answer the questions one is interested in

Data quality (including completeness and accuracy) are two key properties about the data that should be clarified

Beyond that, the questions that a data consumer should ask are the following:

- What do the data describe?
- Do the data cover the problems we are trying to solve?
- What does each record represent? How granular (vs. coarse) are they?
- How “fresh” are the data? Do they reflect the current situation? When were they collected?

Descriptive statistics can be used to describe and understand the basic characteristics of a data set. A descriptive statistic is a number that provides a simple summary about the sample and the measures. It allows to simplify large amounts of data in a convenient way

Descriptive statistics are broken into two basic categories:

1) Measures of central tendency

2) Measures of spread

1.3. Measures of central tendency

Measures of central tendency (or "measures of center") - focus on *the average or middle values of a data set*

They describe the most common patterns of the analyzed data set. There are three main such measures: 1) the mean, 2) the median, and 3) the mode

The mean is the average of the numbers.

- It is easy to calculate
- It is the sum of the values divided by the number of values

Example: The following data set indicates the number of chocolate bars consumed by 11 individuals in the previous week. The series has 11 values (one value per person): 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

Mean = (0 + 1 + 2 + 2 + 3 + 4 + 5 + 6 + 7 + 9 + 16) / 11 = 55 / 11 = 5

This means that any given person in this sample ate 5 chocolate bars in the previous week (on average)

The median is *the value lying in the "middle" of the data set.*

- It separates the higher half from the lower half of a data sample
- It is calculated as follows:
    - Arrange the numbers in numerical order
    - Count how many numbers there are
    - If it is an odd number, divide by 2 and round up to get the position of the median number
    - If you have an even number, divide by 2. Go to the number in that position and average it with the number in the next higher position

Example 1: Dataset with 11 values: 0, 1, 2, 2, 3, **4**, 5, 6, 7, 9, 16

Median = 4 (4 is in the middle of the dataset), Mean = 5

=> one half of the people in the sample consumed 4 or less chocolate bars and the other half consumed 4 or more bars

Example 2: Dataset with 10 values: 1, 2, 2, 3, **4, 5**, 6, 7, 9, 16

Median is 4.5 (the average of 4 and 5 in the middle of the dataset), Mean = 5.5

=> In that case, half of the people in the sample ate less than 4.5 chocolate bars and the other half ate more than 4.5 bars

Example 3: Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 93

Median = 4, Mean = 12

The median is not affected by extreme values or outliers (i.e., data points that differ significantly from other observations)

⇨ **The median is a "robust" measure of central tendency**

The mean is susceptible to outliers

⇨ **The mean is a "non-robust" measure of central tendency**

The median is often used in daily life, to represent the middle of a group.

- For example, when discussing the "average" income for a country, analyst will often use the median (rather than the mean) salary.
- As a robust measure of central tendency, the median is resistant to outliers (e.g., billionaires). It constitutes a fairer way to represent a "typical" income level.

The mode is the most commonly observed value in the data set. It is calculated as follows:

- Arrange the numbers in numerical order
- Count how many there are of each number.
- The number that appears most often is the mode.

Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

Mode = 2.

- The value "2" is represented twice, while all the other values are only represented once
- This means that 2 is the most common number of chocolate bars consumed in the previous week

Dataset with 11 values: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 16

- The series has no mode, as all values appear just once

Dataset with 11 values: 0, 1, 2, 2, 4, 4, 5, 6, 7, 9, 16

- The modes are 2 and 4, as they are both represented twice while all the other values are still only represented once
- This shows that a data set can have two (or more) modes

Dataset with 11 values:: 0, 1, 2, 2, 4, 4, 5, 6, 7, 9, 100

- The modes are 2 and 4 =>

The mode is a "robust" measure of central tendency.

The mode is not frequently used in daily or business life. However, it presents the key advantage that it can be used for nominal data (which is not possible with the mean and the median)

This means that it would be possible to calculate the mode if the data set showed the brand names of the chocolate bars purchased). Yet it makes no sense to speak of the "mean" or "median" brand

Measures of central tendency are particularly convenient when used to compare different variables or different subgroups. For example:

- To get a snapshot about the average (mean or median) income, wealth and tax rate of a population
- To compare the income, wealth and tax rate for different subgroups of a population (e.g., by region, gender, age group, or education)

N.B. A comparison of the *mean* and *median* can offer additional clues about the data

- A similar mean and median indicate that the values are quite balanced around the mean
- If, however, the median is lower than the mean, it is possible that there are outliers at the high end of the value spectrum (or "distribution" in the statistics jargon)

If median income < mean income => there is a large low- or middle-class population with a small minority with extremely high incomes (billionaires).

If median income > mean income => the economy probably consists of a large middle class and a small, extremely poor, minority.

1.4. Measures of spread

Measures of spread (also known as "measures of variability" and "measures of dispersion") describe the dispersion of data within the data set. The dispersion is the extent to which data points depart from the center and from each other

The higher the dispersion, the more "scattered" the data points are.

Main measures of spread:

- the minimum and maximum,
- the range,
- the variance and standard deviation

Minimum and maximum - the lowest, respectively the highest values of the data set

Dataset with 11 values: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16

- The minimum = 0; the maximum = 16

A minimum or maximum that appears too low, respectively too high may suggest problems with the data set. The records related to these values should be carefully examined.

Range - the difference between the maximum and the minimum

- It is easy to calculate
- It provides a quick estimate about the spread of values in the data set, but is not a very good measure of variability
- In the previous example, the range is 16 (= 16-0)

Variance - describes how far each value in the data set is from the mean (and hence from every other value in the set)

- Mathematically, it is defined as the average of the squares of the differences between the observed and the mean
- It is always positive
- Due to its squared nature, the variance is not widely used in practice

Dataset: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16 - the variance is 18.73

Standard deviation - measures the dispersion of a data set relative to its mean

- It is calculated as the square root if the variance
- It expresses by how much the values of data set differ from the mean value for that data set
- In simple terms, it can be regarded as the average distance from the mean.

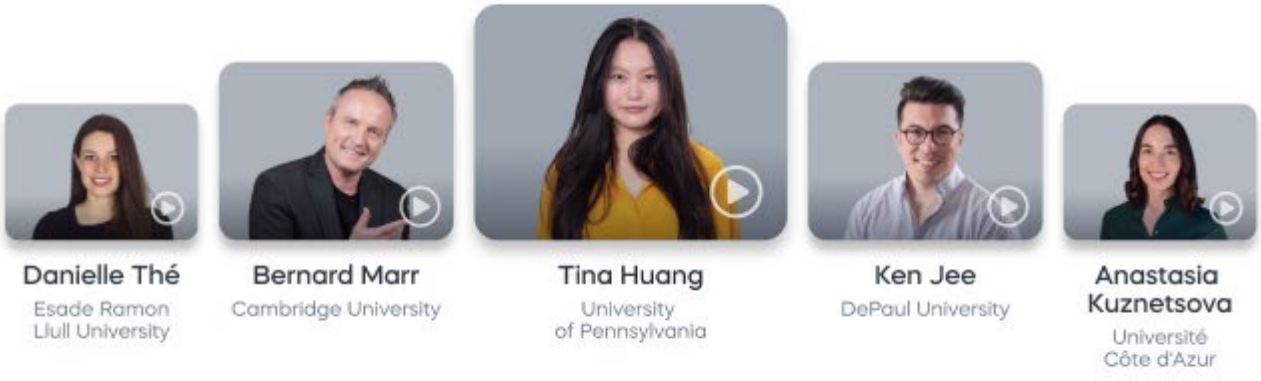
- A low standard deviation reveals that the values tend to be close to the mean of the data set
- Inversely, a high standard deviation signifies that the values are spread out over a wider range
- A useful property of the standard deviation compared to the variance is that it is expressed in the same unit as the data

Dataset: 0, 1, 2, 2, 3, 4, 5, 6, 7, 9, 16 - the standard deviation is 4.32

Learn **DATA SCIENCE** anytime, anywhere, at your own pace.

If you found this resource useful, check out our **e-learning program**. We have everything you need to succeed in data science.

Learn the most sought-after data science skills from the **best experts in the field**! Earn a **verifiable certificate** of achievement trusted by employers worldwide and future proof your car



Danielle Thé
Esade Ramon Llull University

Bernard Marr
Cambridge University

Tina Huang
University of Pennsylvania

Ken Jee
DePaul University

Anastasia Kuznetsova
Université Côte d'Azur

Comprehensive training, exams, certificates.

- ✓ 160+ hours of video
- ✓ 599+ Exercises
- ✓ Downloadables
- ✓ Exams & Certification
- ✓ Personalized support
- ✓ Resume Builder & Feedback
- ✓ Portfolio advice
- ✓ New content
- ✓ Career tracks

Join a global community of 1.8 M successful students with an annual subscription at 60% OFF with coupon code **365RESOURCES**.

~~$432~~ **$172.80**/year

Start at 60% Off

VAT may be applied

365 DataScience

Olivier Maugain

Email: team@365datascience.com