



Data Literacy

Definition

“**Data literacy** is the ability to read, understand, create, and communicate data as information. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data.”

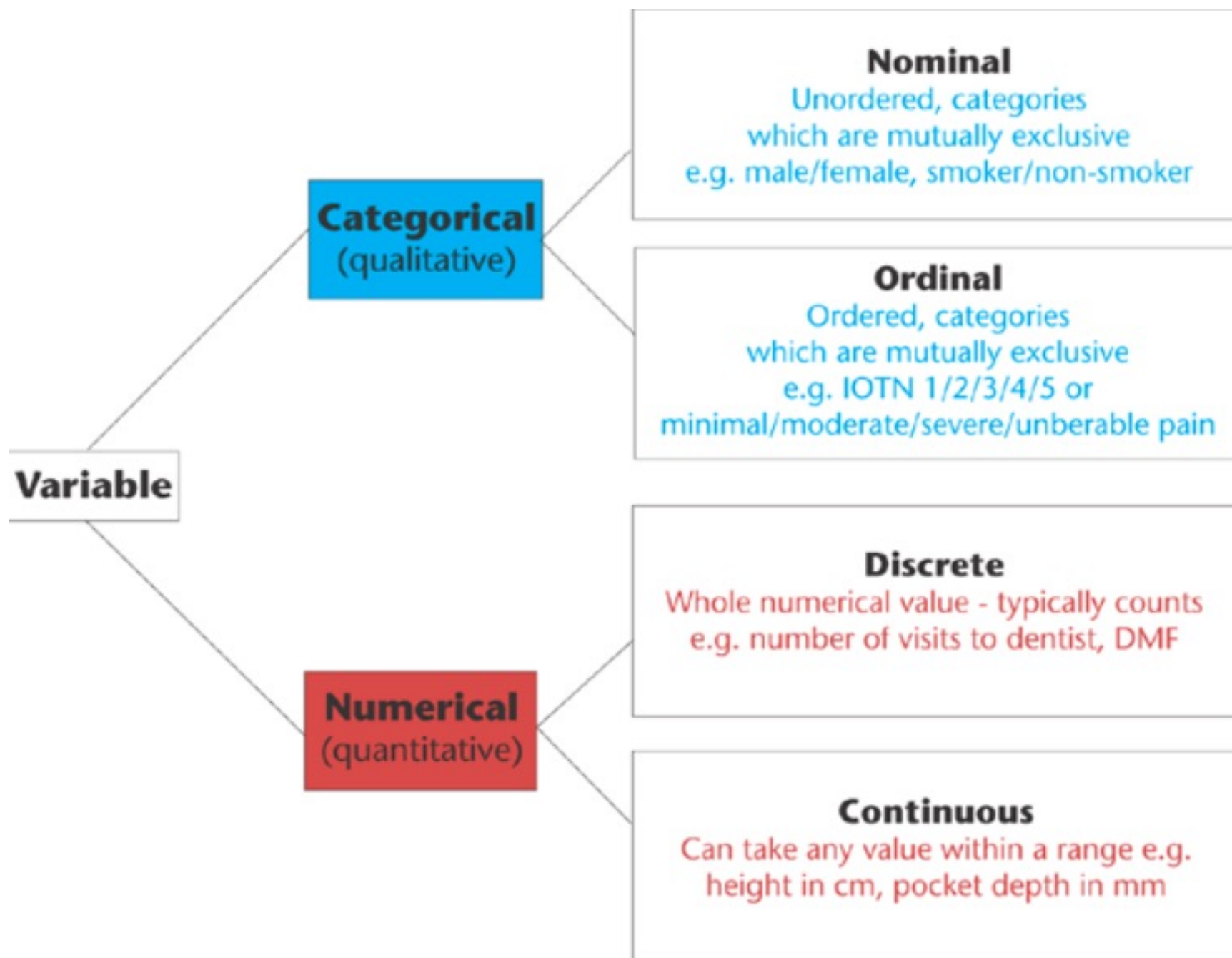
https://en.wikipedia.org/wiki/Data_literacy



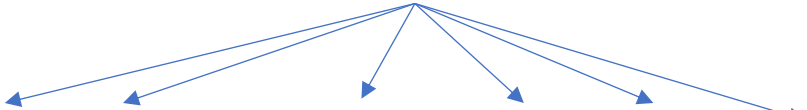
BASIC ELEMENTS OF DATA

The data literate person knows how to distinguish between different types of data, such as categorical and numerical variables, discrete versus continuous value data fields. Beyond merely being able to identify data types, however, the data literate person also understands what can and can't be done with them in analysis and visualization.

<https://www.slideshare.net/dataremixed/17-key-traits-of-data-literacy>



Variables



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

All of the variables in this data set are categorical variables – variables whose values only consist of labels/levels.

Outlook and PlayTennis are **nominal categorical** variables, the labels/level cannot be considered ordered, i.e. *Yes* \nless *No* and *No* \nless *Yes*

The remaining variables are all ordinal categorical variables – the labels/levels can be considered ordered, i.e.
Cool $<$ *Mild* $<$ *Hot*

ID	Gender	Age	Income	Rating
1	Male	28	\$50,000	4.5
2	Female	35	\$65,000	3.8
3	Male	22	\$40,000	4.2
4	Female	45	\$80,000	4.8
5	Male	31	\$55,000	3.5

- **ID:** Discrete numerical variable representing a unique identifier for each individual.
- **Gender:** Nominal categorical variable representing the gender of the individual (Male/Female).
- **Age:** Discrete numerical variable representing the age of the individual.
- **Income:** Continuous numerical variable representing the income of the individual.
- **Rating:** Continuous numerical variable representing a rating given by the individual.

Note: We see later that we will treat numerical ID variables like they appear in this table as nominal categorical variables because it makes no sense to use these identifiers as numerical values, we cannot order them or do mathematical transformations on them.



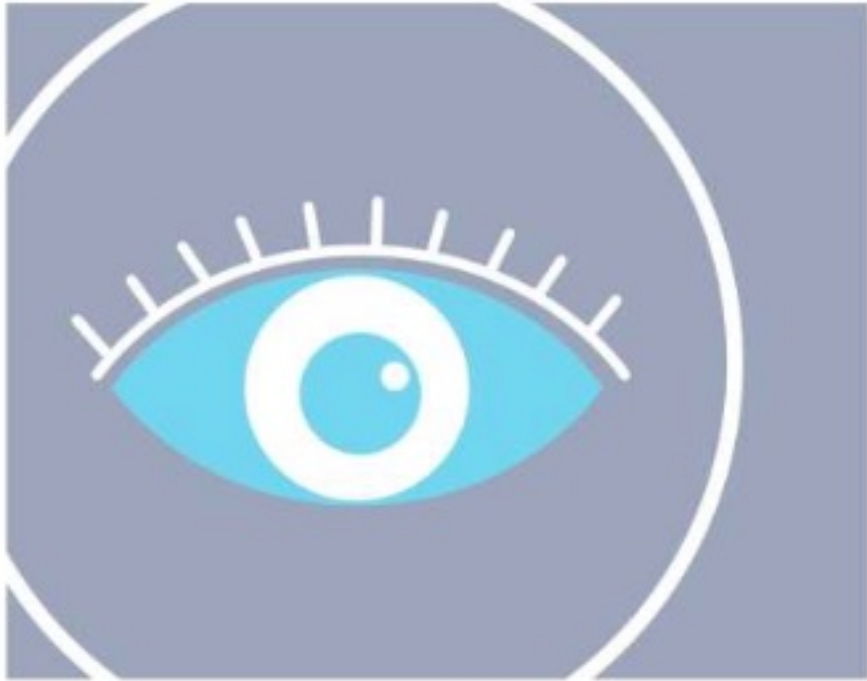
DATA STORAGE METHODS

The data literate person is familiar with ways that data is collected, structured and stored, and the attributes associated with each approach. Spreadsheets with their cells in rows and columns are seen as distinct from databases with their records arranged in relational tables or non-relational documents.



DATA ANALYSIS PRINCIPLES

Those who are data literate understand that storing data is not an end in and of itself, but rather a means of extracting valuable insight about one's environment. The data literate person must therefore grasp the fundamental principles of analysis and statistics and when they apply.



DATA VISUALIZATION RULES OF THUMB

Since the human visual system is a “very high bandwidth channel to the brain”, the data literate person understands various ways to visualize data and their respective pros and cons. The principles of cognition relating to how humans decode visual encodings such as position, length, area, and color, are well known to the data literati, as are the different chart types that make use of these encodings.

<http://www.cs.ubc.ca/labs/imager/tr/2009/VisChapter/akp-vischapter.pdf>

Channels: Expressiveness Types and Effectiveness Ranks

➔ **Magnitude Channels: Ordered** Attributes

Position on common scale



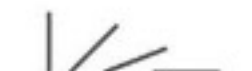
Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



Most

Effectiveness

Least

➔ **Identity Channels: Categorical** Attributes

Spatial region



Color hue



Motion

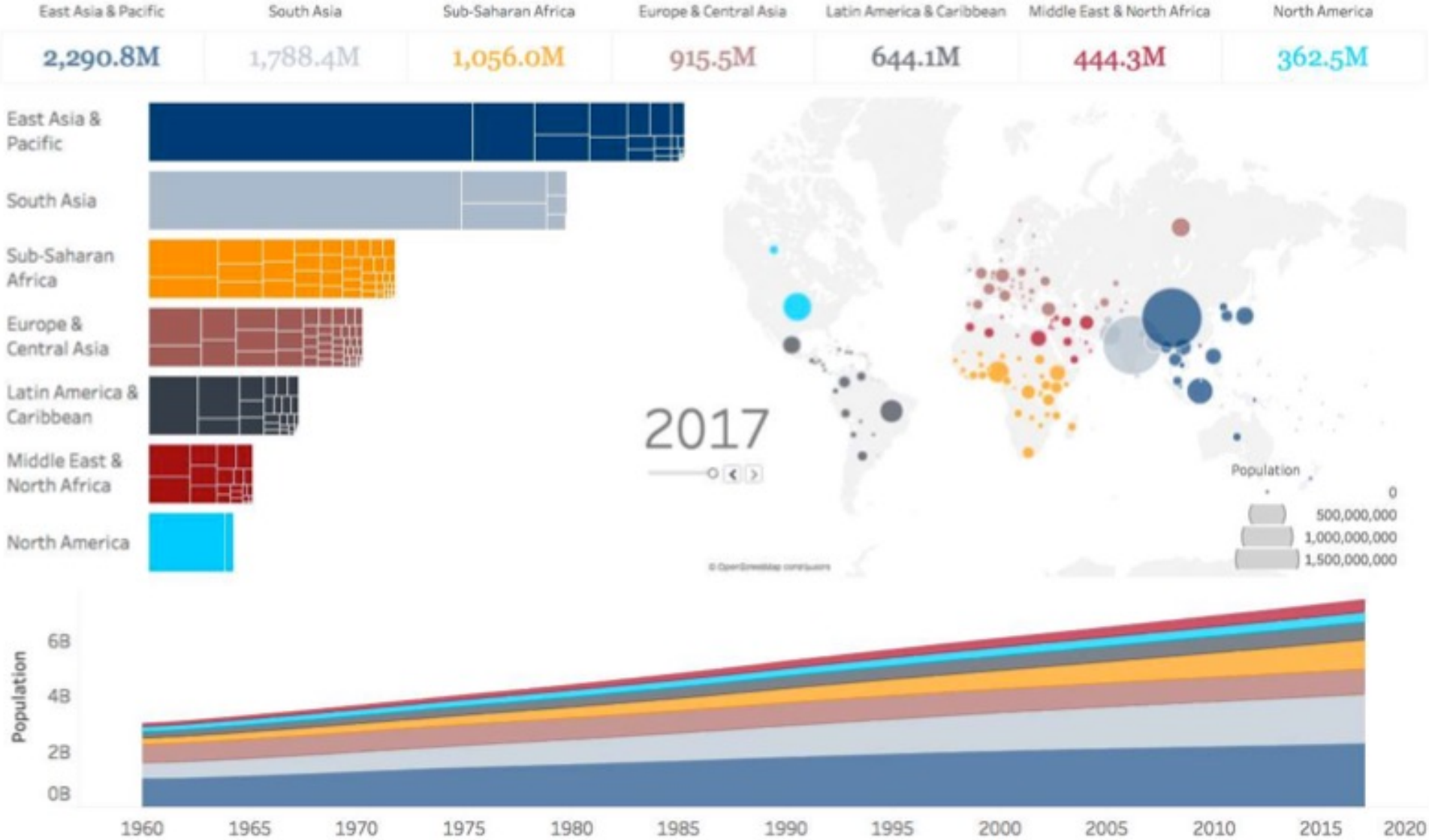


Shape



World Population Growth by Region

7,501,739,318 total population





PREPARE DATA FOR ANALYSIS

The answers to our questions can rarely be answered by one single, clean data set. Data is most often 'dirty' - full of errors and formatting issues - and relevant information is often stored in multiple places. For this reason, people who are data literate know how to clean dirty data and combine multiple data sets together for analysis.

“ Anyone who has worked with data knows that it doesn't all come in pristine form. For this reason, a data literate person needs to learn how to handle data that needs some work, or that doesn't even exist in a data form and needs to be gathered. This is often missed, but it's one of the key points in becoming data literate.”



CHERYL PHILLIPS

Lorry I. Lokey Visiting Professor
in Professional Journalism at
Stanford University
[www.comm.stanford.edu/
faculty-phillips/](http://www.comm.stanford.edu/faculty-phillips/)



Who needs spell check, anyway?

Volkswagen

Voldswagen
VOLKSWAGEN SW
Volswagen

VOLKSWAGEN

Volkswag
Volkswage
Volksqagon
Volkswagen
Volkswagen
Volkswago
Volksawen
Volks VOLKS

Volkeswagon
VOLKSWAGEN CONV
Volkswaqgon
Volkswage N

VOLKSWAGON Volks Wagen Volkswgen Voltswagen

Volkwagon
VOLKSWAGEN Voolkswagen Voikswagen
VOLKSWAGAN
Volkswagoen
VOLKSWGEN

Volks wagon
Volts Wagon
Volks wagon
VOLKSWAGEN
Volkswagon

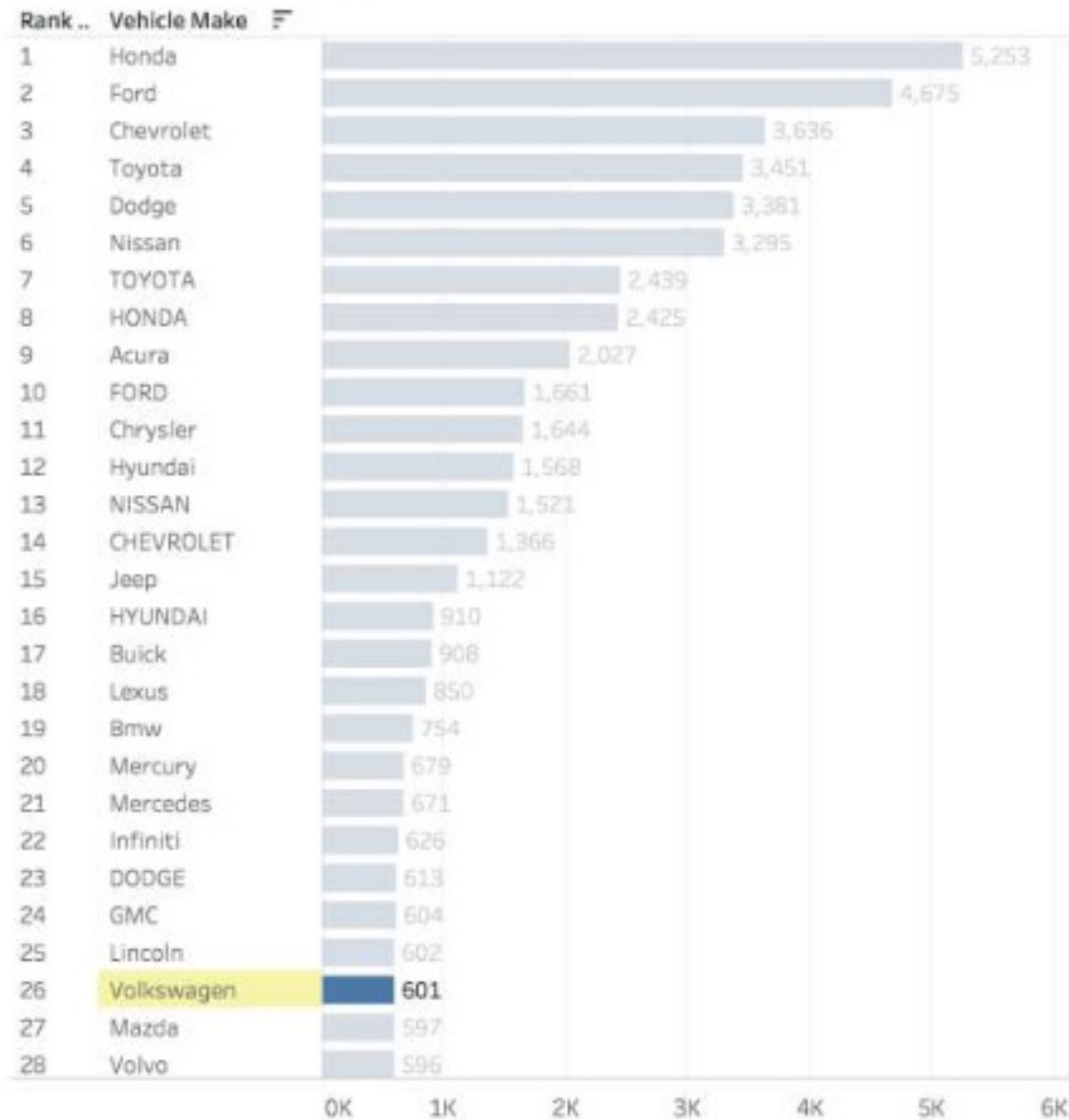
VOLKSWAGEN SW

Volkswasgen Volksawagon
Volztwagon

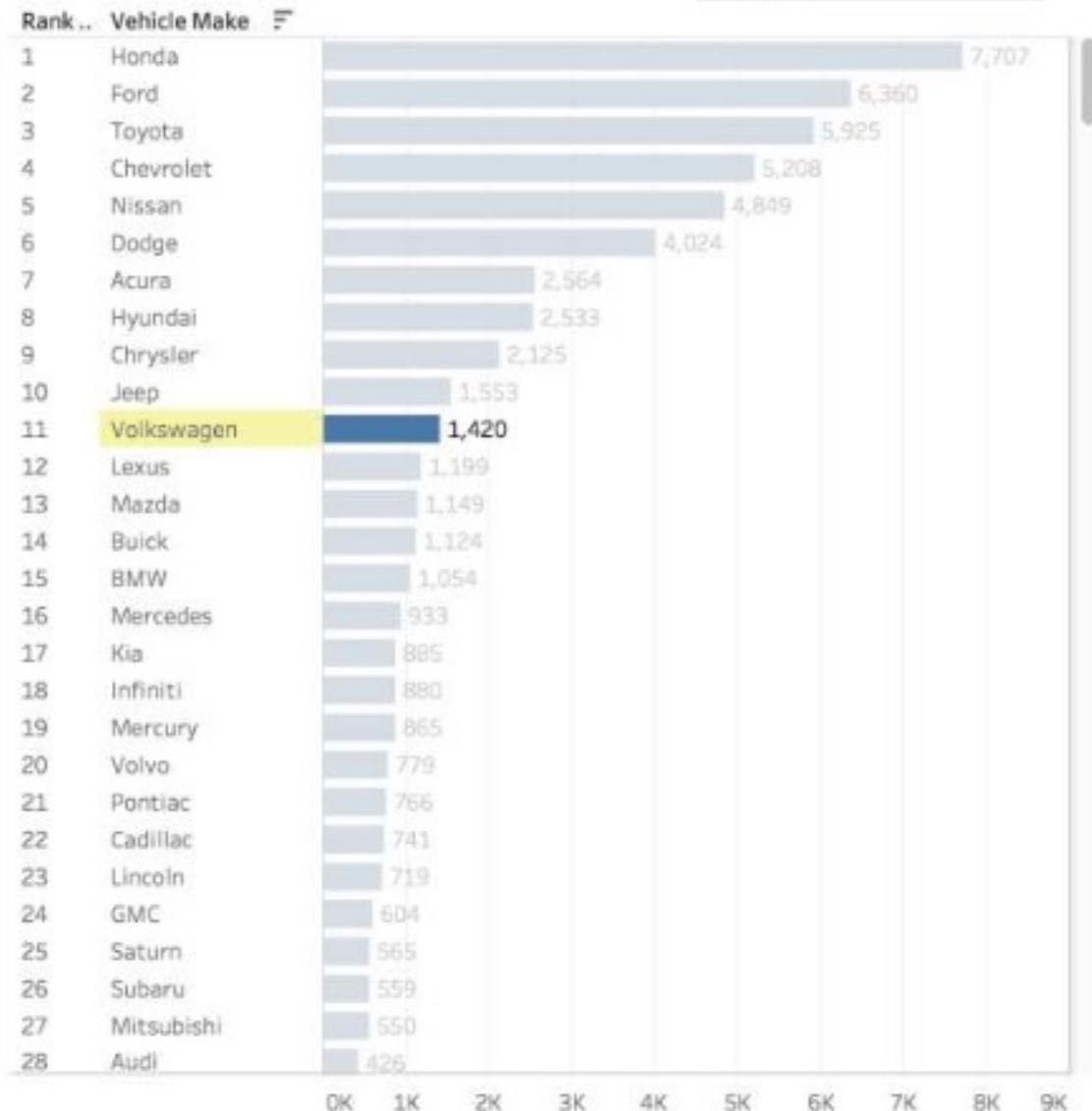
VBOLKSWAGEN Volkeswagen

What was the effect of the grouping?

Vehicle Makes - Original



Vehicle Makes - Cleaned





COMMUNICATE DATA EFFECTIVELY

Data literate people know that the true power of data is in shaping the minds and directing the decisions of their fellow human beings. For this reason, they know how to communicate effectively using data and information gleaned from it. Put another way, they speak data well.



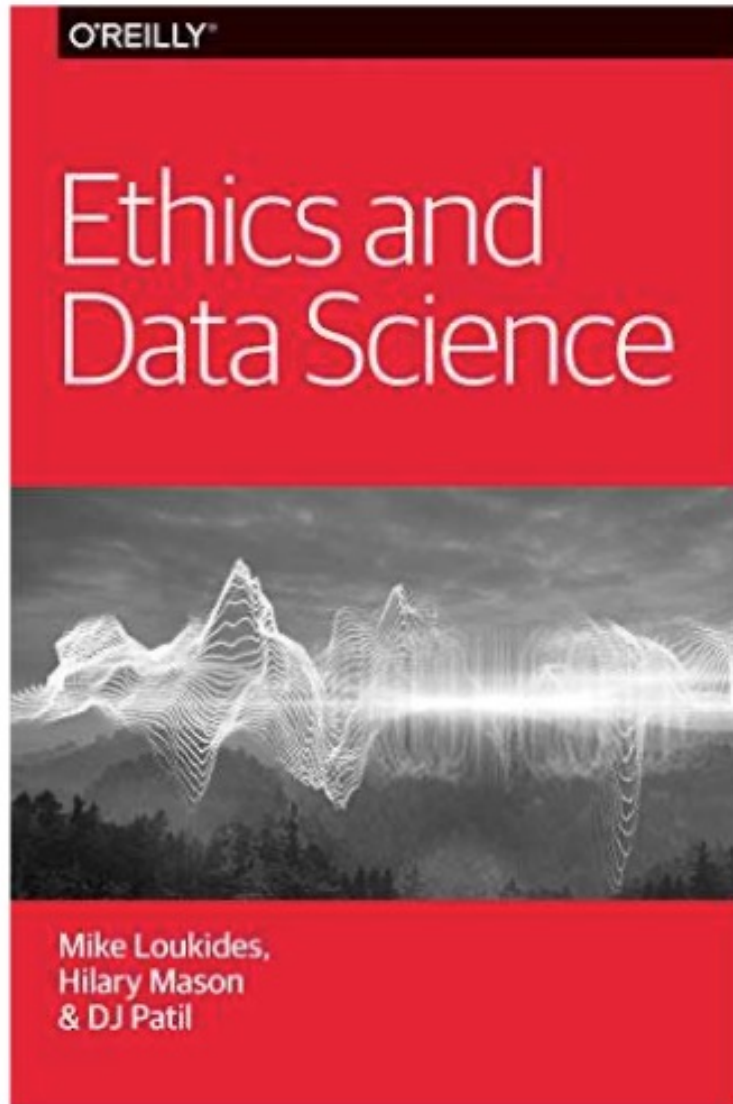
ETHICAL

Data can be used to help or to harm, and data literate people consider ethical use of data and the impact on society to be of utmost importance. Data literate people find ways to use data to help their organizations grow, but they only do so while respecting the rights and privacy of others and while seeking to improve the lives of those affected.



As data teams, we aim to...

1. Use data to improve life for our users, customers, organizations, and communities.
2. Create reproducible and extensible work.
3. Build teams with diverse ideas, backgrounds, and strengths.
4. Prioritize the continuous collection and availability of discussions and metadata.
5. Clearly identify the questions and objectives that drive each project and use to guide both planning and refinement.
6. Be open to changing our methods and conclusions in response to new knowledge.
7. Recognize and mitigate bias in ourselves and in the data we use.
8. Present our work in ways that empower others to make better-informed decisions.
9. Consider carefully the ethical implications of choices we make when using data, and the impacts of our work on individuals and society.
10. Respect and invite fair criticism while promoting the identification and open discussion of errors, risks, and unintended consequences of our work.
11. Protect the privacy and security of individuals represented in our data.
12. Help others to understand the most useful and appropriate applications of data to solve real-world problems.



Ethics and Data Science

by DJ Patil, Hilary Mason, Mike Loukides

The 5 C's:

- Consent
- Clarity
- Consistence and Trust
- Control and Transparency
- Consequences

Princeton Case Studies:

Automated healthcare app, Dynamic sound identification, Optimizing schools, Law Enforcement ChatBots, Hiring by machine, Public sector data analytics

<https://aiethics.princeton.edu/case-studies/>



UTILIZES DATA RESOURCEFULLY

A data literate person actively seeks out and creates data as a means of gathering information. If data exists that will help them make an important decision or come to a much-needed understanding about the current situation, they can be counted on to find it and make good use of it.



CONTINUOUSLY IMPROVES DATA

Knowing that analyses and their underlying data are always imperfect and incomplete to some degree, data literate individuals identify areas of improvement in the data and associated analysis. Once identified, they proactively seek to implement improvements as time and resources permit.



EFFECTIVELY ADVOCATES FOR DATA

Data literate team members advocate for the effective use of data in communication and decision-making. When data is not being utilized in important discussions and decisions, they proactively suggest ways to add a data-driven perspective, and they offer their advice or assistance to make it happen.