Data Composition and Insight Report for Visualization

The **Data Composition and Insight Report for Visualization** (Data-CIV) is a framework to consider data and its use for visualisation; to breakdown complex data into its core components, providing detailed analysis and key insights. Designed to enhance data storytelling, it prepares a developer to consider relevant information, for clear and impactful visual representation and visualisation.

By, Jonathan C. Roberts

Introduction.

The task is to thoroughly explore your dataset. Before creating a data visualization, it's essential to understand the structure and content of the data. You need to evaluate its potential and limitations—whether it contains enough information to support the visualization you have in mind. Sometimes, the dataset may not be suitable for your intended purpose or lacks the necessary details for the visual you're envisioning. For example, the data may be designed for one use, but you're attempting to illustrate something different. In such cases, the data might not align with your goals, and you may need to rethink your approach.

Data explanation
Data composition
Data use

Figure 1. Three main parts of the data-analysis. First think how to explain the data, and understand the fundamental parts. Second critically think about the composition of the data. Finally, consider how the data will be used, and if the data is fit-for-use.

The data-civ task consists of three parts.

Data Explanation. In this section, you will describe the dataset, beginning with what you know and discussing its limitations and potential. This provides a broad overview of the data, including its origin and how it was created. Include an example of the data, such as a data table, to illustrate its key components. The focus here is on presenting factual information.

Data Composition. Conduct an in-depth analysis of the data. Identify the data types and the variables, including their names and meanings. Examine how the data is organized: Is it sparse or continuous? Are there any missing values? If there are missing or erroneous data points, explain how these are indicated to the reader.

Data Use. Reflect on how the data will be utilized, who will use it, and for what purpose. Assess whether the available data is suitable for your intended objectives. Consider if you have the necessary permissions to use it and whether you possess the capability to manipulate it effectively.

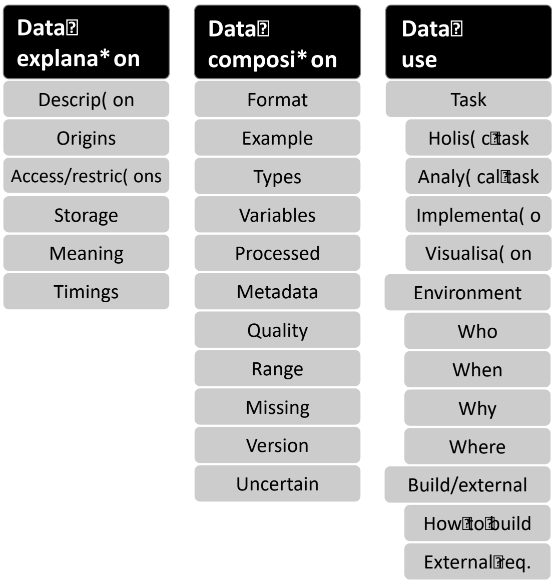


Figure 2. Think through the different parts, to perform a full data-analysis.

Data explanation

This part introduces the reader to the data; here you look at the data and explain what it is, how it is made up, and why it was created. This provides an introduction to the data.

a. **Describe** it. Describe the data. Why was it stored. What is the history about the data? What do you know about the data?
b. **Origins**. Do you know the provenance of the data? Who created it and for what purpose did they create it? What device or who created the data (machine, human or simulation)?
c. **Access/restrictions**. Can you get hold of the data? What is the license? Is it open source? Do you have permission to use it? What is the CC-BY regulation? Is it sensitive data (does it need to be stored in a secure way), or does it require GDPR consideration? Is it encrypted data? Who has access to it (e.g., secure location).
d. **Storage**. How is it stored? Locally or remote access? Stored in a single xml/xls file? Stored in a database? Secured storage? If security is needed to surround the sensor device (for creating the data) were protocols followed. When the data was stored, was it stored securely – or could it have been tampered with (eavesdropping)? If it has been transmitted, was the data securely transmitted, or has it been changed on route. Has the method of transmission (e.g., from remote sensor) changed the data or lowered the quality in any way (e.g., remote sensed from space, may send a lowered quality data, due to data transmission bandwidth). Any compression used – e.g., Lossy storage/lossy compression used, or non-lossy?
e. **Meaning/Appearance**. What does it represent? Does it represent something positive? Is it "data for good"? Is it encouraging data? Is it constructive? Would someone enjoy seeing the visualisation of this data? Would someone be interested in the data? What does it look like?
f. **Timings/temporal**. When was the data stored? Is it out of data? Is it live data (real data from the database, e.g., customer information).

Output = text descriptions of the points. At least one table (sample-data table) may include other figures, diagrams or photographs,

Demonstrate that you know the background to the data. In this section you will need to demonstrate that you understand the data. Write descriptions, follow the above sections as a guideline to help you create a meaningful and deep discussion of the data. This part needs to have a table summarising the data. Reference the table in the main text. Discuss and reference individual parts of the table to demonstrate your thoughts. Add in other pictures, diagrams and so on, to help you present this information. For example, you could add a photograph of the sensor, or add another table with information of what is stored by the sensor if it is different to the final data (for example, in remote sensing data may be stored in a remote setting in one way, and then sent to the receiver in another format.

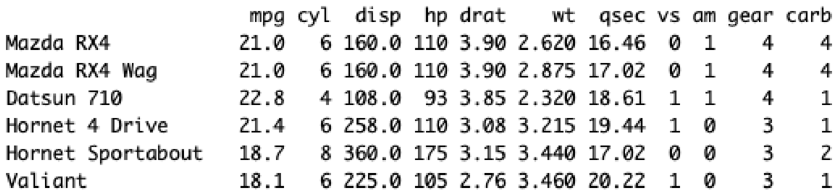
You will need to think sensibly about your data, and decide how to demonstrate you know about the information, where it comes from, how it is stored, and so on. Each data set is different, and will have different requirements.

For example, the mtcars dataset,

> "The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)."
>
> A data frame with 32 observations on 11 (numeric) variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

The data example, may include the top 6 rows of the file. This would be an indicative part of the data. But make sure that your sample (whether it is the first 6 lines, or another set of lines) is representative of the whole dataset. I.e., make sure that it contains an example of each of the cases, such as data values, missing values, errors, and so on. An example of the mtcars dataset could be the following 6 lines.



```
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
```

Figure 3. First six lines of the mtcars dataset. The rows were chosen as an representative sample of the whole dataset. It shows some of the values, and how the data is stored.

Data composition.

In this part, you think about the composition of the data. What are the components of the data? How is it formed? How is it put together? What does it look like?

a. **Format**. What does the data look like? What data types are used? Is the format correct. What formats are used for strings such as dates) e.g., US or UK, DD/MM/YY or MM/DD/YY)? Does the data need to be cleaned – eg., are some columns mixed up, is some data in the wrong column or row? Bad or inconsistent formatting?
b. **Example.** Give an example of the data. Summarise it into its main parts. Put together a table that exemplifies the data.
c. **Types**. Integer/decimal, float, string, char, Boolean, double, long,
d. **Variables.** What are the data variables? Category (nominal or ordinal (ordered categories or ranks)). Quantity (discrete or continuous). If the data is free-text, or words, can you use it? Is it already classified? What do the variables mean? If they are listed as acronyms, do you know what those acronyms mean? Can you find out, if you don't know?
    i. **Numeric variables.** Can be measured. Quantities such as numbers. Discrete or continuous.
    ii. **Categorical variables.** Describing some characteristic of a variable. Ordinal (e.g., degree classification 2.2, 2.1, 1st) or Nominal (e.g., eye colour of blue, brown, light brown etc.)
    iii. **Quantitate. It can be measured, usually stored as numbers. Sensors will probably agree. It is objective. It is not based on opinion. It is accurate.**
    iv. **Qualitive data.** It is measured. It is often non-numerical. For example it could be the result of people answering questions in the street. And in which case, people change their minds. **Is it Direct** (people asking directly) or **indirect** (observations of visual records, photos, people's recalls of the event). What are the **codes** on the data? Are there codes already? Has the data been coded? Is this **deductive** or **inductive** coding?
e. **Processed**. Has it been processed in any way? Is it averaged data? Is it the raw data? What functions have already been applied to the data, and do you know what algorithms have been used?

f. **Metadata**. What metadata is included? Origin, position, data/time of reading. Experimenter, project, address, copyright. Error-rates of the device. Data description: dimensions, coordinates, ranges, byte order (MSB, LSB) on storage (e.g., if binary stored). Encrypted key/information. Is the metadata accurate?

g. **Quality**. Is the data accurate? What decimal places – are they suitable? What is the data loss? Is it complete data – or a sample? Has the data been aggregated into bins (or categories, e.g., data range 0-10, 10-20,20-30 and so on. Or age ranges 0-5years, 6-10 years and so on). (e.g., cf William McKnight in Information Management, 2014) says different forms of data quality (entry quality, process quality, identification quality, integration quality, usage quality, aging quality, organisation quality). Depending on the origins (sensor, simulation or human):

   v. **Observational**. Is the sensor accurate enough for the purpose? How up to data is the data?

   vi. **Simulated** data. Is the simulation suitable, and modelling the right things?

   vii. **Observational** data. Has the human(s) been accurate in their recordings? How do you know? Do you know? What processes have they followed to guarantee that it is correct.

   viii. **Derived/compiled.** Results aggregated from many different sources.

   ix. **Reference or canonical.** Is the data reference data (used by many for examples, e.g., cars dataset or iris flower dataset (Fisher's Iris data). E.g., this is practice data that is often stripped of any personal information, or given away from a company to demonstrate the type of data they have.

h. **Range.** What is the frequency distribution of the data? What range? Is it sparse data (do a histogram on the data) – is the dataset continuous in its storage, or are their gaps in storing the data. Is there missing data?

i. **Missing/error/duplicates.** Is there missing data? How is this noted (e.g., blank, or a n/a or what). What happens if the sensor device had an error – how is this stored (as a blank or as missing or something else)? Is there duplicate data?

j. **Version.** What version do you have? Do you have the final version? The original data or the cleaned data? Do you know what version? Is the data out-of-date (sometimes data has a shelf-life!)

k. **Certainty/uncertain.** Is the data uncertain? What is the error-bars or error rates on the data?

Output = text descriptions of the points. At least a table demonstrating the types, variables, etc., with range information

In this section you are demonstrating that you know about the data in a deep way. This section requires you to perform critical thought, and often requires some detective work. In other words, you need to look at the written text (that often comes with the data), look at the data itself, look for papers or subsidiary and associative papers.

Similar to section 1, you will need to demonstrate through both text and diagrams/tables that you understand the information, and can argue over the individual parts. Make sure you are being **objective** in your thoughts and presentation of the information. Make sure that the information you write is correct and comprehensive. It needs to be accurate. You must be factual and truthful. Do not mislead. But do make sure that your presentations are clear and meaningful. It is not good to make a presentation, or present your data in a way that is confusing.

Make it clear that you know the data variables. The types of data, variables, components of the data, and their ranges. The best way to demonstrate this knowledge is to create a table. For example, Figure 4 shows an example table for the mtcars dataset.

Item in text	Variable name	Type	Example	Category	Quantity	Range (generate a histogram?)
Purchasing	Price	Integer	£3400		Continuous	0 to 100k
Car	Name	String	VW	Nominal		Defined finite list (lookup table)
American, Japanese, European	Origin	String	USA	Nominal		Shorter list (well defined)
MPG	MPG	Integer	15		Continuous	Short range, specific (easily find the best MPG and worst)
cylinders.	Cylinders	Integer	8		Discrete	Very short set (defined by the engine design) May not be suitable in electric cars)
horsepower	Horsepower	Integer	170		Continuous	Medium
weight.	Weight	Integer	3563		Continuous	Infinite (exact)
year.	Year	Integer	1970		Discrete	Date/time is more infinite, year more finite and even 100s
Good condition	Condition	Char	A	Ordinal		Finite short range

Figure 4. Data table for the cars dataset. Showing the name, variable types, example, categories and quantities, with their ranges.

Sometimes it may be useful to do some quick analysis of that data. Sometimes it may be useful to include a histogram. This helps to understand the distribution of the data. So for instance, you could plot the histogram over the horsepower of the mtcars dataset, which is shown in Figure 5.

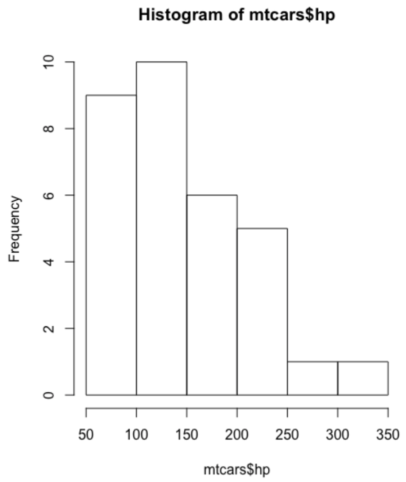


Figure 5. Histogram of the mtcars dataset, focusing on the horsepower (hp) variable. It shows that most of the cars have a horsepower between 100 and 150, and that tere are only a few cars over 250 horsepower (in this dataset).

You will need to go into detail. Make sure you look at all aspects of the data, and provide a comprehensive and detailed analysis of the data. Remember, you are demonstrating that you understand the data fully and in a deep way. And that you have investigated all aspects of the data and can express and explain the facets in a clear and well-structured way. For example, do not forget aspects of the data, such as how missing values are recorded, or what metadata is stored with the main dataset, how the information was captured, what the quality of the information is, and so on.

Data use/purpose

In this part you think about how you are going to use the data. For what purpose are you analysing the data? Do you have a story to tell (do you know one in mind already)? Do the components of the data match-in-with the purpose? Think through three parts: the task that you wish to do with the data; the environment that the task will be performed within, and

1. **Task.** What are you going to do with the data? What is the purpose of you using and requiring the data? What is your task, and the purpose that you are envisaging for the data? Does the data match the task or the story that you are trying to achieve?
    a. **Holistic task.** What is the main idea – is you task to **explore**, **discover**, **explain** or use the data to **train** something?
        i. E.g., with an exploration, the visualisation tool allows the person to try different scenarios, to locate and judge different aspects. Often a more complete and complex interface is required. The developer does not really know what is inside the dataset – it is new to them also! Sometimes the developer and the end-user are the same person; in other words they are creating an interactive tool to help them explore some data.
        ii. With Discovery, the developer has created the visualisation to allow the user to discover what the developer has already found out. E.g., they may say "see what I have already found out, and judge for yourself". Whereas with explanation visualisations, the idea is to help the person/user to understand something new, and to really deeply appreciate the data/ideas or topics from the visualisation picture.
    b. **Analytical task.** Compare, maximum, minimum, find a value. Locate, judge, value?
    c. **Implementation task:** how can you build the solution? What do you have to do to create it? Can you use something already built? How can you deliver the result?
    d. **Visualisation task**. How do you achieve your visualisation task. How do you demonstrate the values? E.g., Highlight, drag/drop, move, stick, pick, change, value, choose, slide, touch, 3d point.

2. **Environment**. Who, When, Why, Where.
    a. **Who** will be using it? What is the type of person? What are their skills? What are their background and visual literacy (are they visual literate, will they know and implicitly understand the type of visualisation mapping you are going to use?) Do they have any disabilities or special requirements? Braille, large print, colour safe, etc.
    b. **When** will they use it (what timing, do they want it now, do they want to interact with it, do they want to look at it on a static picture?)
    c. **Why** will they use it? Why do they need it? For what purpose?
    d. **Where.** Where will they use it? Will it be inside or outside? Will it be against a bright sunlight or in a darkened room?

3. **Build and External requirements.**
    a. How to build solution? Do you build it? Or someone else builds it? E.g., there is a difference in how you need to specify the tool, if you are designing it yourself and building it yourself then you can get away with a quicker simpler design specification. But if you are asking someone else, or a company to build it, then you need to explain more clearly what you wish to have built. Or use a pre-existing build, or adapt it.
    b. External requirements: memory/computer. Size of output. Size of display (mobile friendly, big screen, immersive head-mounted-display). Extensible, adaptable, software that can be used and adapted by others, or one-time-use. Skills of the builder/creator. Do you have the skills necessary and experience necessary to build it? Can you learn, or do you need to change the design specification?

Output = text descriptions of the points.

For this final section you need to demonstrate that you understand the data in a way of how it can be used. Think about the potential for the data. Is there any data that is missing. For example, if you have a dataset with different country import and export information, yet if you decide to add another dataset of the GDP, you may be able to make a better and clearer story, and compare the main values with GDP. Whatever else you do, make sure you have a clear section of Who, When, Why and Where.

There are (at least) three parts to perform a full analysis.

Task.

Think about the task that you will perform. In other words, think what work do you want to do with this data? What story do you want to tell? And so on. You should have some idea of what you want out of this data. This is the first steps in the design-study. You probably will not have a full understand of your goals (at this stage), however you should have an inkling of an idea. Some kind of vision, or imagination of what you are going to do with this data. This is the time to express that idea, and compare it against the data in front of you.

It is not good to have a dataset that will not be fit for purpose. If you want to tell one story, but your data does not contain that information, then you will not be able to tell the story with that data! You need to make sure that the data is capable of presenting the information in a way that is clear and suitable. If the data does not have enough values (it is a very small database, for instance) and yet you want to tell a story of huge-change, you will not be able to tell that story! If you want to tell a story of how the world changed over the last 100 years, but only have data for the past 5 years, you will not be able to tell this story.

Environment.

The second part is the "environment". This is the place, location or position where you will do the visualisation. Will it be presented on a big screen at a festival? Will it be placed on a poster on the side of a wall along a corridor in your work place? Will it be on a website, or an App on a phone? Again, you may not know (fully) the answers to these questions. But should have an idea of where (probably) you will display it, and for who. Who will be looking at it? Who will be using it? Who is it for? Why would they need it? What do they need to do with it? Perhaps, read it for interest; learn from the work; apply it to their own work and situation, and so on. Understanding the data, and how it can be used is very important.

The best way to consider this part is to think about Who, When, Why, Where. And structure the comments using these W's as a structure. For example, take a look at the paper " Explanatory Journeys: Visualising to Understand and Explain Administrative Justice Paths of Redress" (Roberts, et al 2021)

If you take a look at this paper, there are several sections that cover the Who, When, Why, Where aspects of the work. The project looks to explain administrative justice, particularly looking at housing and education issues. We discussed long and hard who the user would be. We started off this project with the idea that the "user" would be the expert group, and probably government officials. But after discussion with stakeholders and people who had been affected by administrative justice ideas, we changed our focus to members of the public. So the question for this data-analysis part, is did we have the data to explain administrative justice to the public. What would we need to display to the public? How is the data made up, and is it suitable to be explained? When would they need to use it? All these questions get discussed in the paper.

I your submission, you will need to consider similar issues. There may be some parts that are more certain than others. Furthermore, like the Explanatory justice project, things may change and adapt as the project continues. What you need here, is discussion of your thoughts, presentation of the things you know (now and for certain) and discuss some aspects that are less clear or certain.

Build and external requirements

The final section gets you to think about the requirements. What do you need to do now? What don't you know about your data that you need to discover? Who is going to build the solution? What are their skills and abilities? Do they have the skills, time, money and so on, to build the solution? Can they take the data and use it? There are many questions to think about there.

Next steps

The idea of the data-analysis is to demonstrate that you know your data, that you have applied yourself and done your investigation. That you can present your data in a clear way.

For information let's just quickly look at what you will be doing next! This data-analysis will then be used in the next steps. After the data-analysis has been completed, the next steps are to do a "design study". What will be done is to follow a strategy. We will follow the Five Design-Sheet method (Roberts et al, 2016). This uses 5 sheets of paper, with 5 facets to consider per page.

The first sheet is the ideas sheet, next are three separate designs (on three sheets) and then finally the realisation design, on the final sheet.

References:

Roberts, JC, Butcher, P, Sherlock, A & Nason, S 2021, 'Explanatory Journeys: Visualising to Understand and Explain Administrative Justice Paths of Redress: Explanatory Journeys', IEEE Transactions on visualization and computer graphics. <https://arxiv.org/abs/2107.14013>

J. C. Roberts, C. Headleand and P. D. Ritsos, "Sketching Designs Using the Five Design-Sheet Methodology," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 419-428, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2467271.