Data Science 1

STAT/CS 287

Jim Bagrow, UVM Dept of Math and Statistics

LECTURE 08

Last time: Tidy data

Reorganize into a **standard form** ("**tidy**"):

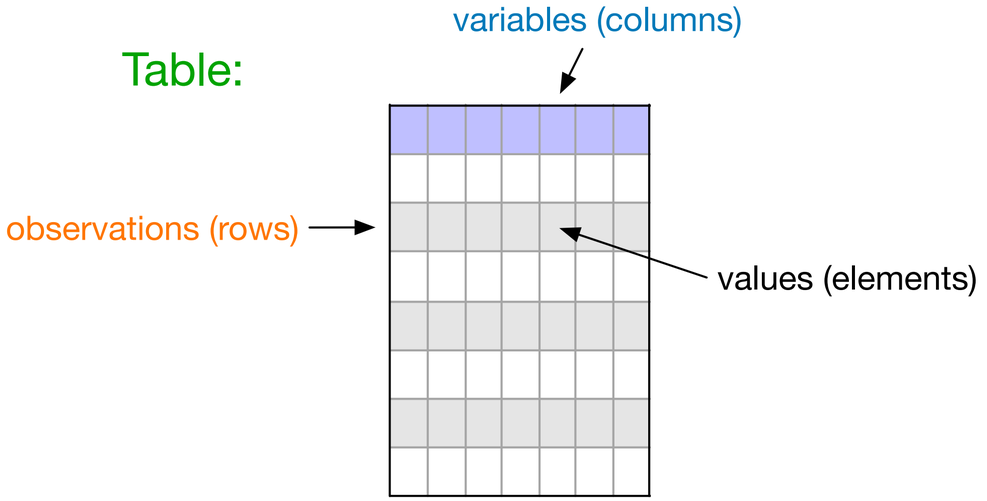
person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Make values, variables and observations *more clear*:
Dataset contains **one table** with **18 values** across **three variables** and **six observations**

Tidy data

1. Each **variable** forms a column.
2. Each **observation** forms a row.
3. Each **type of observational unit** forms a table.

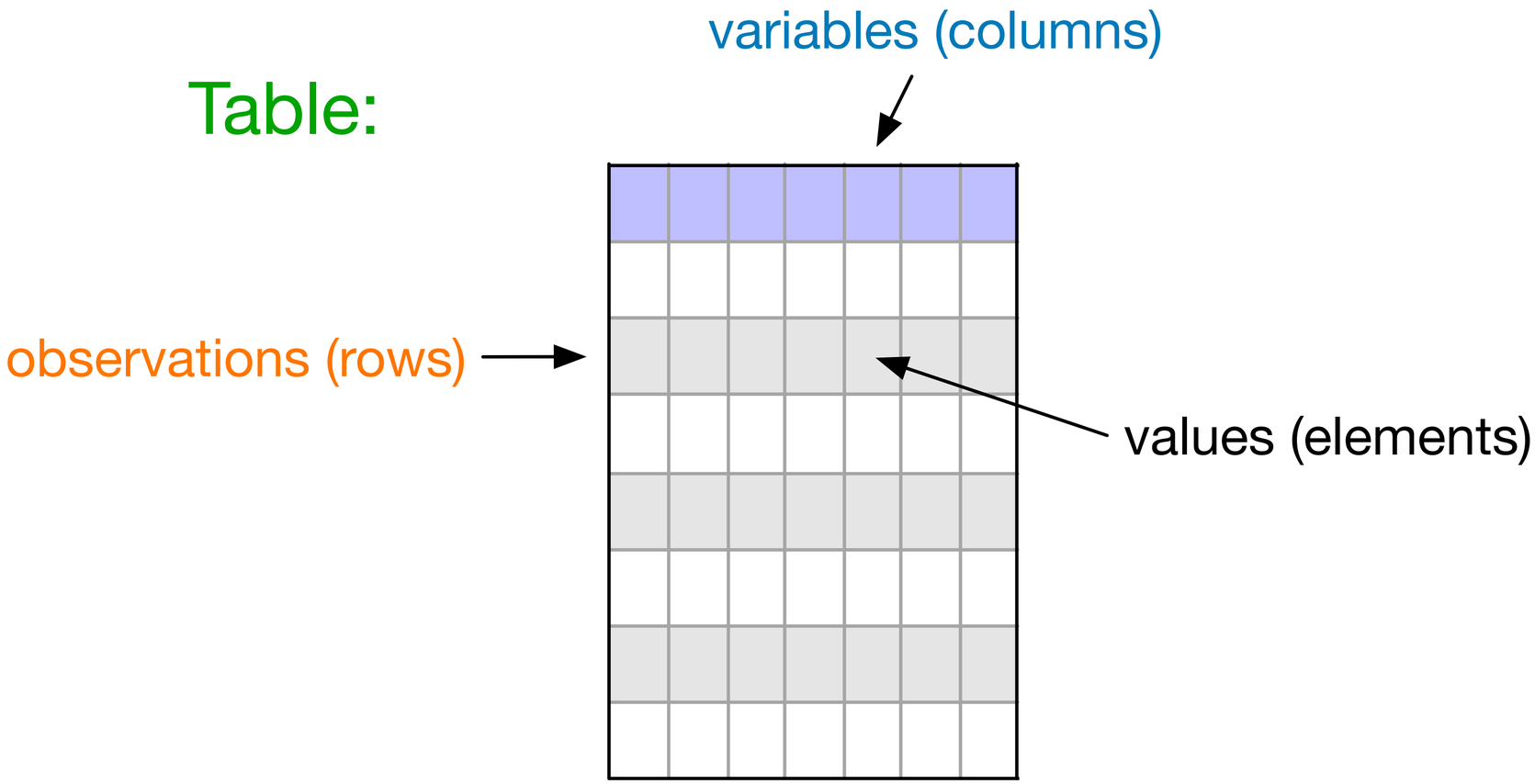
Tidy:



(data may not be *stored* in this format)

H. Wickham, 2014

Last time: Tidy data



(data may not be *stored* in this format)

H. Wickham, 2014

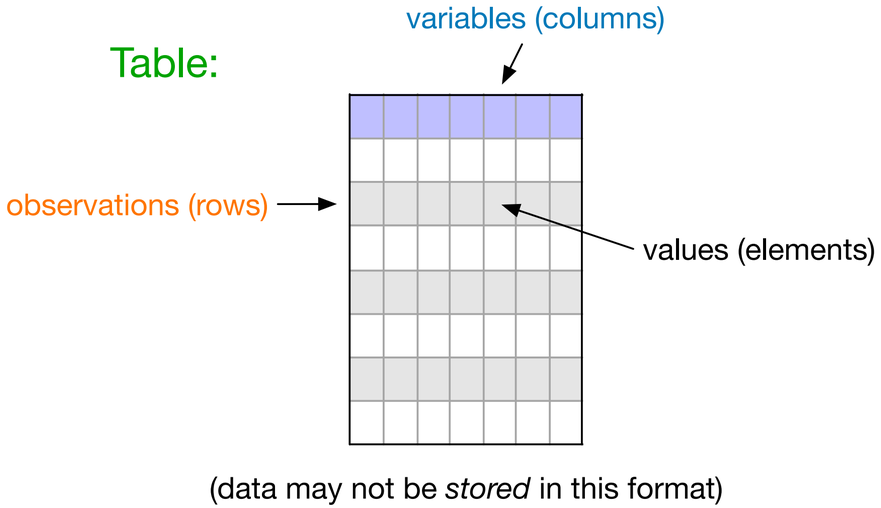
Cleaning and (pre)processing data

Question
How to get data into this tidy format?

Many steps and possible paths towards getting "raw" data into "shape"

Large amounts of "raw" data → need code to automate this process

- ✓ Code also helps with provenance (reproducibility and replicability) of data analysis



Cleaning and (pre)processing data

Question
How to get data into this tidy format?

Many steps and possible paths towards getting "raw" data into "shape"

Large amounts of "raw" data → need code to automate this process

- ✓ Code also helps with provenance (reproducibility and replicability) of data analysis

Question
Does something in the data make no sense?

Cleaning and (pre)processing data

Question
How to get data into this tidy format?

Many steps and possible paths towards getting "raw" data into "shape"

Large amounts of "raw" data → need code to automate this process

- ✓ Code also helps with provenance (reproducibility and replicability) of data analysis

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant
9	M	0
10	F	0
88	N/A	1
11	.	1
109	M	1

Cleaning and (pre)processing data

Question
How to get data into this tidy format?

Many steps and possible paths towards getting "raw" data into "shape"

Large amounts of "raw" data → need code to automate this process

- ✓ Code also helps with provenance (reproducibility and replicability) of data analysis

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant
9	M	0
10	F	0
88	N/A	1
11	.	1
109	M	1

Cleaning and (pre)processing data

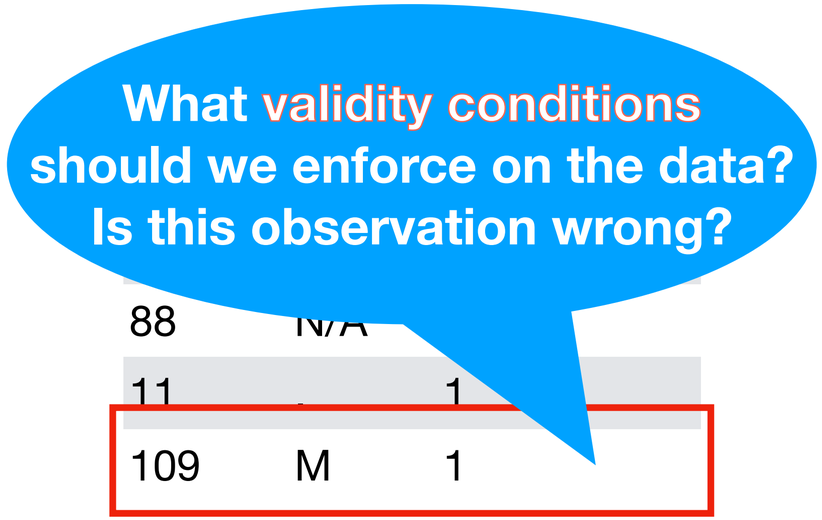
Question
How to get data into this tidy format?

Many steps and possible paths towards getting "raw" data into "shape"

Large amounts of "raw" data → need code to automate this process

- ✓ Code also helps with provenance (reproducibility and replicability) of data analysis

Question
Does something in the data make no sense?



Cleaning and (pre)processing data

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant
9	M	0
10	F	0
88	N/A	1
11	.	1
109	M	1

Cleaning and (pre)processing data

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant
9	M	0
10	F	0
88	N/A	1
11	.	1
109	M	1

Question
Should we remove this data point?

Cleaning and (pre)processing data

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant	is_bad
9	M	0	0
10	F	0	0
88	N/A	1	0
11	.	1	0
109	M	1	1

Question
Should we remove this data point?
Flag it as suspicious?

Cleaning and (pre)processing data

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant	is_bad
9	M	0	0
10	F	0	0
88	N/A	1	0
11	.	1	0
109	M	1	1

Question
Should we remove this data point?
Flag it as suspicious?
Ask data source?

Cleaning and (pre)processing data

Question
Does something in the data make no sense?

data_file.tsv

orgID	sex	is_pregnant	is_bad
9	M	0	0
10	F	0	0
88	N/A	1	0
11	.	1	0
109	M	1	1

Question
Should we remove this data point?
Flag it as suspicious?
Ask data source?

Random aside:

Step 0
Make sure you've **correctly loaded/imported** the data. I can't tell you how many times I've thrown out **hours of work** cleaning a perfectly good file just because I was reading it incorrectly.

"Raw" data

Problems even before you can get to a table

Example: bibliometrics/scientometrics data

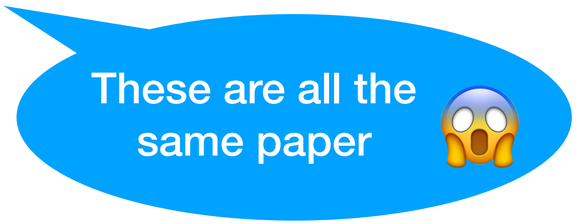
1. A.-L. Barabasi and R. Albert, Rev. Mod. Phys. 74, 47 (2002).
2. Albert, R. & Barabasi, A.-L. (2002) Rev. Mod. Phys. 74, 47–97.
3. Albert, R. & Barabasi, A.-L. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–-97 (2002).
4. R. Albert, A. Barabasi, RMP. 74, 2002.

"Raw" data

Problems even before you can get to a table

Example: bibliometrics/scientometrics data

1. A.-L. Barabasi and R. Albert, Rev. Mod. Phys. 74, 47 (2002).
2. Albert, R. & Barabasi, A.-L. (2002) Rev. Mod. Phys. 74, 47–97.
3. Albert, R. & Barabasi, A.-L. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–-97 (2002).
4. R. Albert, A. Barabasi, RMP. 74, 2002.



"Raw" data

Problems even before you can get to a table

Example: bibliometrics/scientometrics data

1. A.-L. Barabasi 02).
2. Albert, R. & E 7–97.
3. Albert, R. & E lex networks.
   Rev. Mod. Phy
4. R. Albert, A. E

This problem is **so challenging** it has **many names**:

- Record linkage
- Data **deduplication**
- Name (or record) **disambiguation**
- Identity resolution
- …

e are all the
ne paper 😱

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

The horror of natural language data 

```
'1911 - 1961'
'1958\u201386'
'1921\u201376'
'427 BC \u2013 386 BC'
'1983 \u2013 present'
'1983\u2013present'
'1991\u20132001'
'1983\{\{spaced ndash\}\}present'
'<!-- YYYY\u2013YYYY (or \u2013present) -->'
'1989\u2013present'
'1984\u20132001<br /> 2005\u2013present'
'c. 1914\u20131971'
'1960—present'
'1888---c.1920'
```

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

WARNING: Automation can hide problems:

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

WARNING: Automation can hide problems:

```
list_years = []
for s in list_dateranges:
    try:
        y1,y2 = s.split("–")
        list_years.append( (y1,y2) )
    except:
        continue
```

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

WARNING: Automation can hide problems:

```
list_years = []
for s in list_dateranges:
    try:
        y1,y2 = s.split("–")
        list_years.append( (y1,y2) )
    except:
        continue
```

```
print(len(list_years))
print(len(list_dateranges))
```

"Raw" data

Example: parsing natural language

Here's a real problem I was working on
I had a **huge list of date ranges**, for example:
1960–1980.
I want to extract the two years

No problem, you think

```
y1,y2 = s.split("–")
```

but they were stored in an arbitrary fashion, with lots of weird forms.

Maybe you can guess some ways a date range can be written, but take care and look at the data because it's very easy to be surprised:

WARNING: Automation can hide problems:

```
list_years = []
for s in list_dateranges:
    try:
        y1,y2 = s.split("–")
        list_years.append( (y1,y2) )
    except:
        continue
```

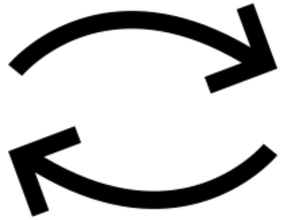
```
print(len(list_years))
print(len(list_dateranges))
```

Result: 60201 60209  **or** 34051 60209 


Need to see what
you are doing

Cleaning and exploring data

Cleaning data



Exploring data

I argue that these two are inextricably linked
You cannot clean without exploring

Cleaning and exploring data

Cleaning data

Exploring data

I argue that these two are inextricably linked
You cannot clean without exploring

Why?

- Do you have missing values?
- Are there duplicate observations?
- Does data format change halfway through?
- Any strange values (e.g. negatives for count data, letters inside zip codes)?

You need to **look** at the data to answer these questions!

Cleaning and exploring data

Cleaning data

Exploring data

I argue that these two are inextricably linked
You cannot clean without exploring

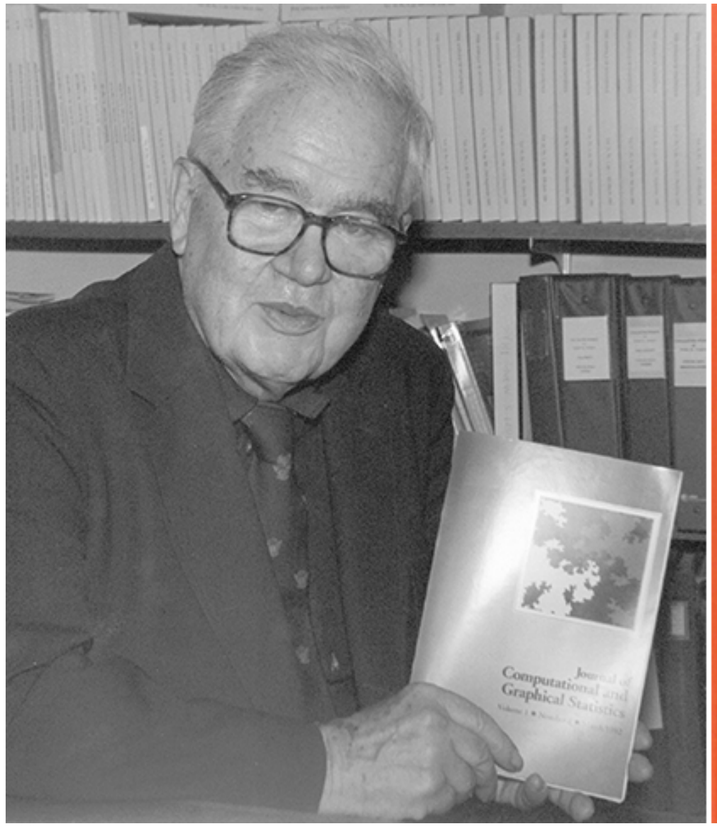
Why?

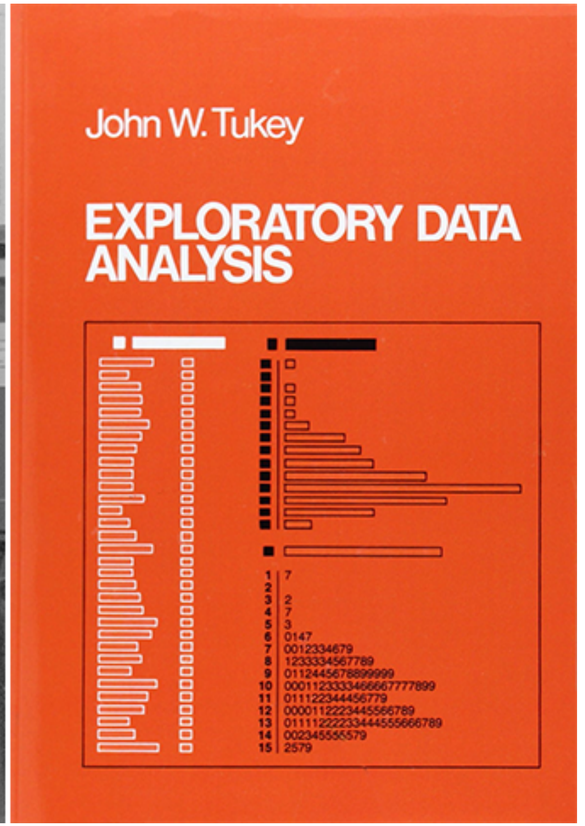
- Do you have missing values?
- Are there duplicate observations?
- Does data format change halfway through?
- Any strange values (e.g. negatives for count data, letters inside zip codes)?

You need to **look** at the data to answer these questions!

Often—especially for big data—you need code to "look" for you

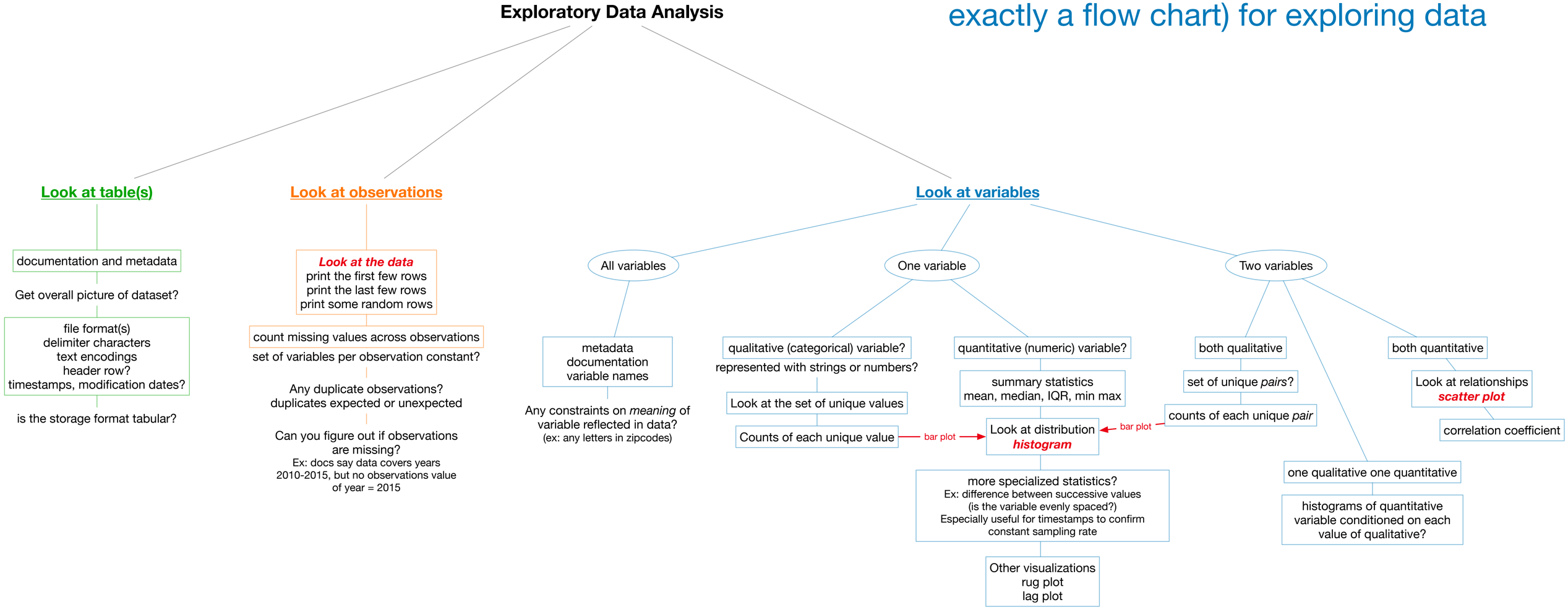
Exploring data





John Tukey (1915-2000)

A *work-in-progress* "**mind map**" (not exactly a flow chart) for exploring data



Look at table(s)

documentation and metadata

Get overall picture of dataset?

file format(s)
delimiter characters
text encodings
header row?
timestamps, modification dates?

is the storage format tabular?

Look at table(s)

cumentation and metadata

overall picture of dataset?

file format(s)
delimiter characters
text encodings
header row?
stamps, modification dates?

he storage format tabular?

Look at observations

Look at the data
print the first few rows
print the last few rows
print some random rows

count missing values across observations

set of variables per observation constant?

Any duplicate observations?
duplicates expected or unexpected

Can you figure out if observations are missing?
Ex: docs say data covers years 2010-2015, but no observations value of year = 2015

All variables

metadata
documentation
variable names

Any constraints on *meaning*
variable reflected in data?
(ex: any letters in zipcodes)

Look at observations

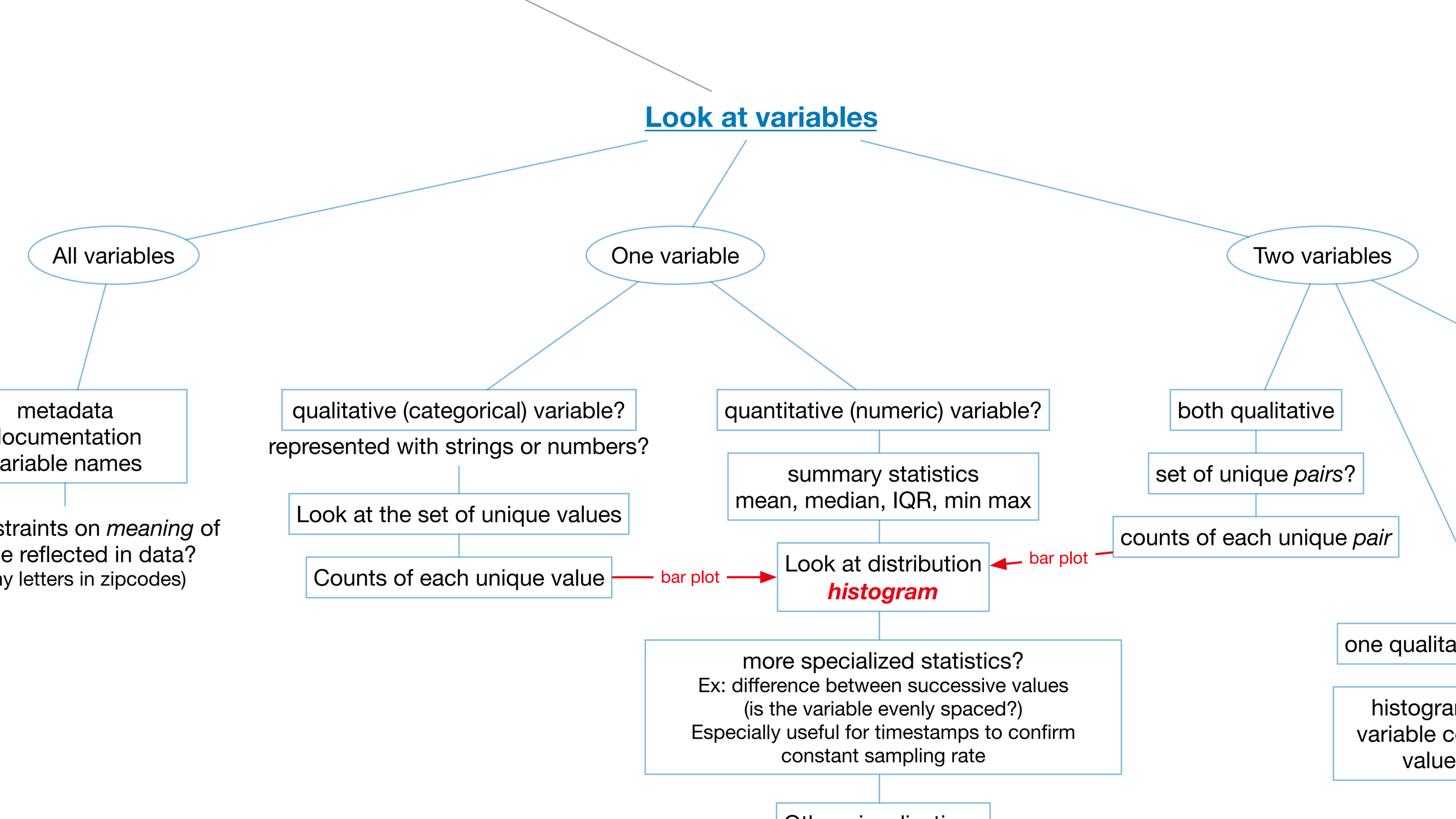
Look at the data
print the first few rows
print the last few rows
print some random rows

count missing values across observations

set of variables per observation constant?

Any duplicate observations?
duplicates expected or unexpected

Can you figure out if observations are missing?
Ex: docs say data covers years 2010-2015, but no observations value of year = 2015


Look at variables
All variables
One variable
Two variables
metadata
locumentation
ariable names
straints on *meaning* of
e reflected in data?
y letters in zipcodes)
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar plot
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at distribution
histogram
bar plot
both qualitative
set of unique *pairs*?
counts of each unique *pair*
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
one qualita
histogra
variable c
value

servations

Lo

e data
few rows
few rows
ndom rows

across observations

servation constant?

bservations?
d or unexpected

if observations
sing?
a covers years
bservations value
2015

All variables

One v

metadata
documentation
variable names

Any constraints on *meaning* of variable reflected in data?
(ex: any letters in zipcodes)

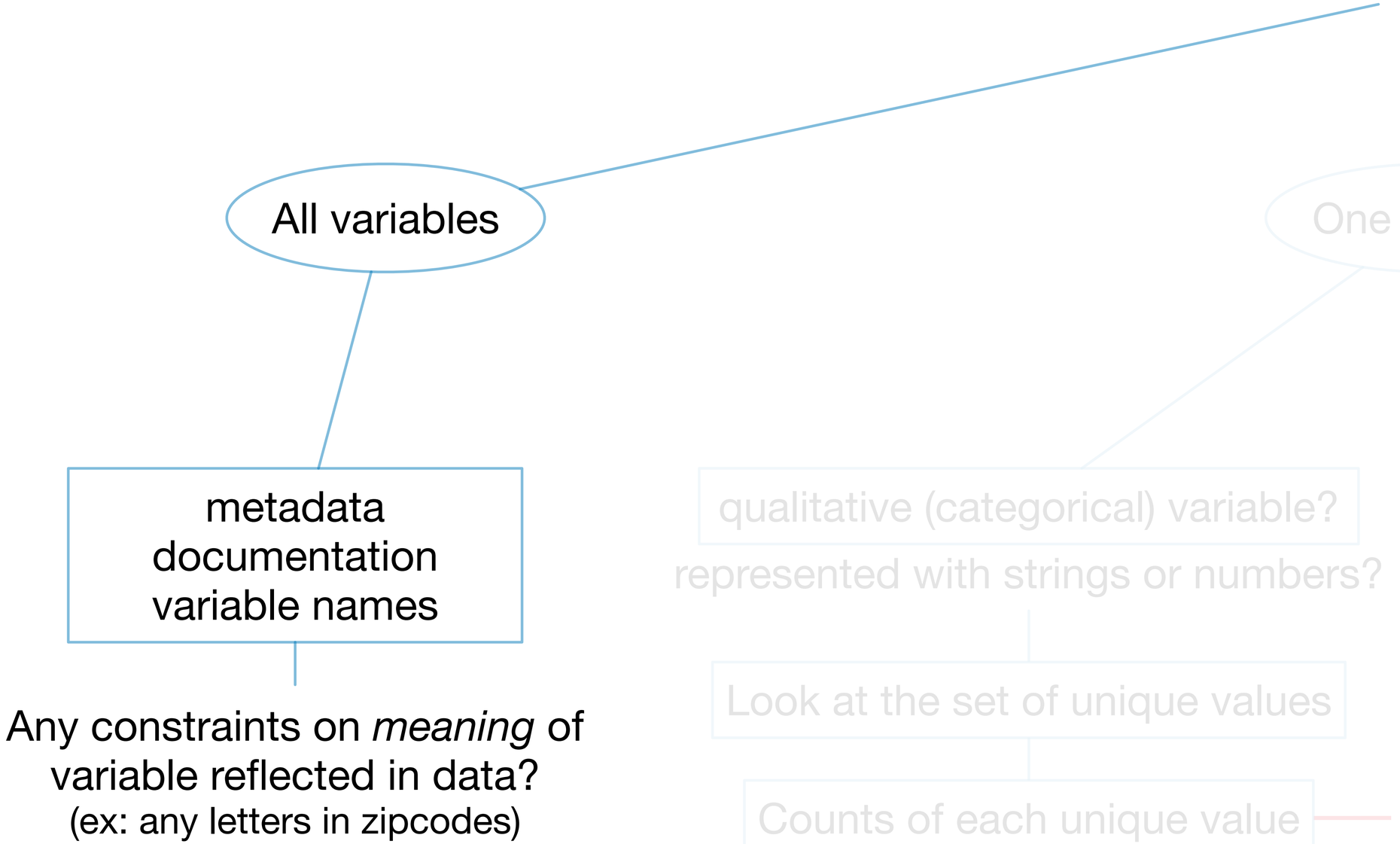
qualitative (categorical) variable?

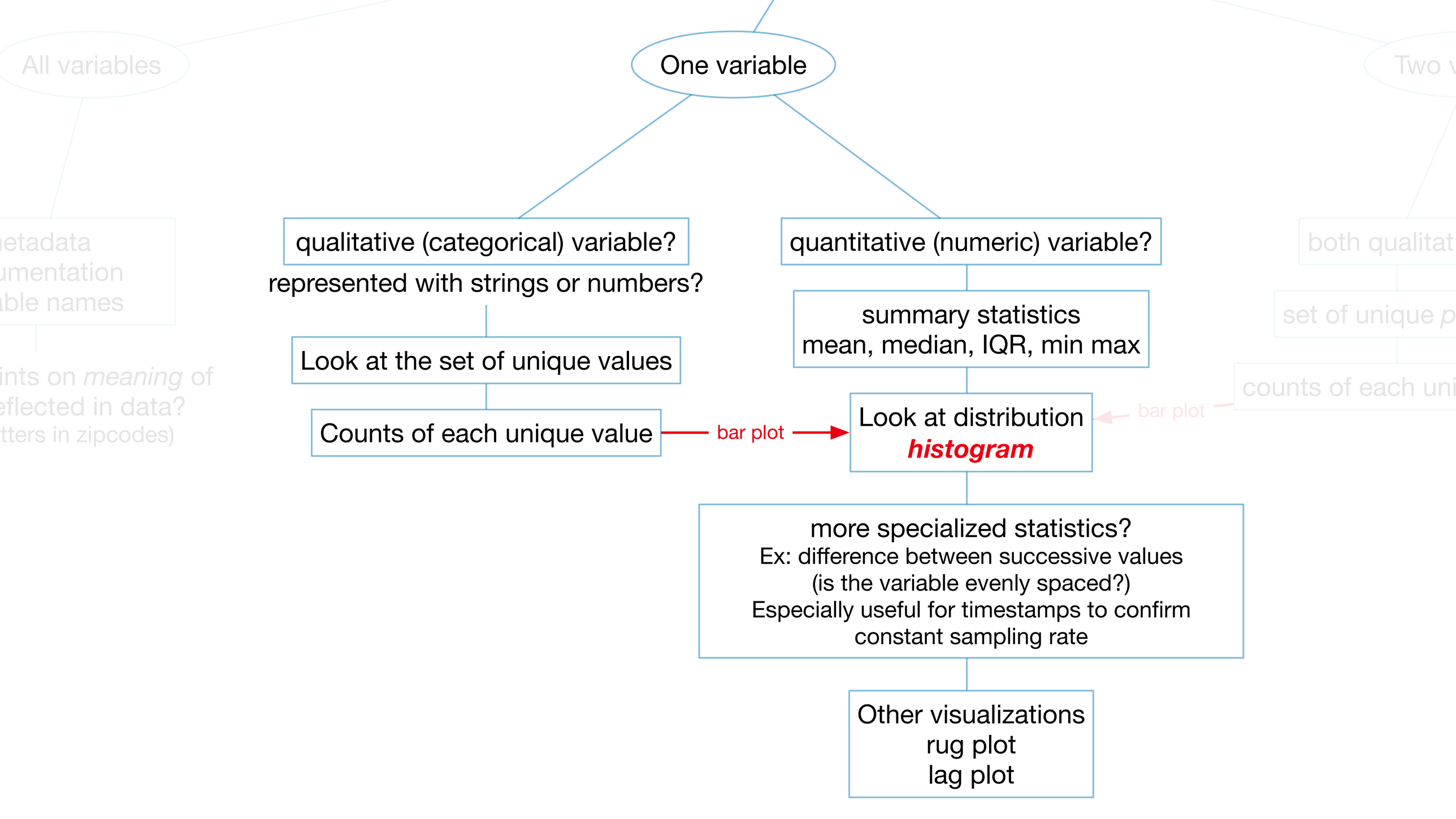
represented with strings or numbers?

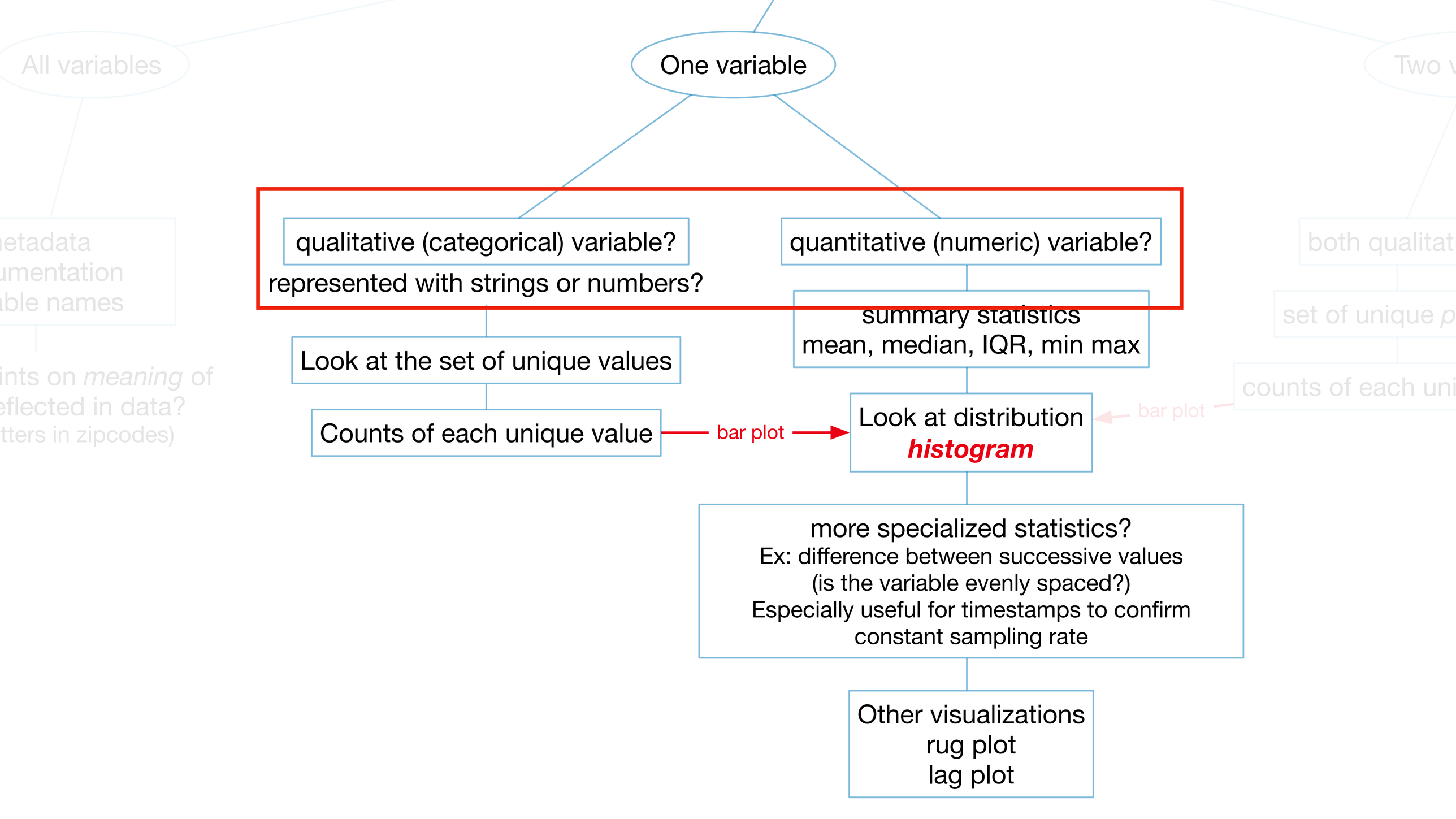
Look at the set of unique values

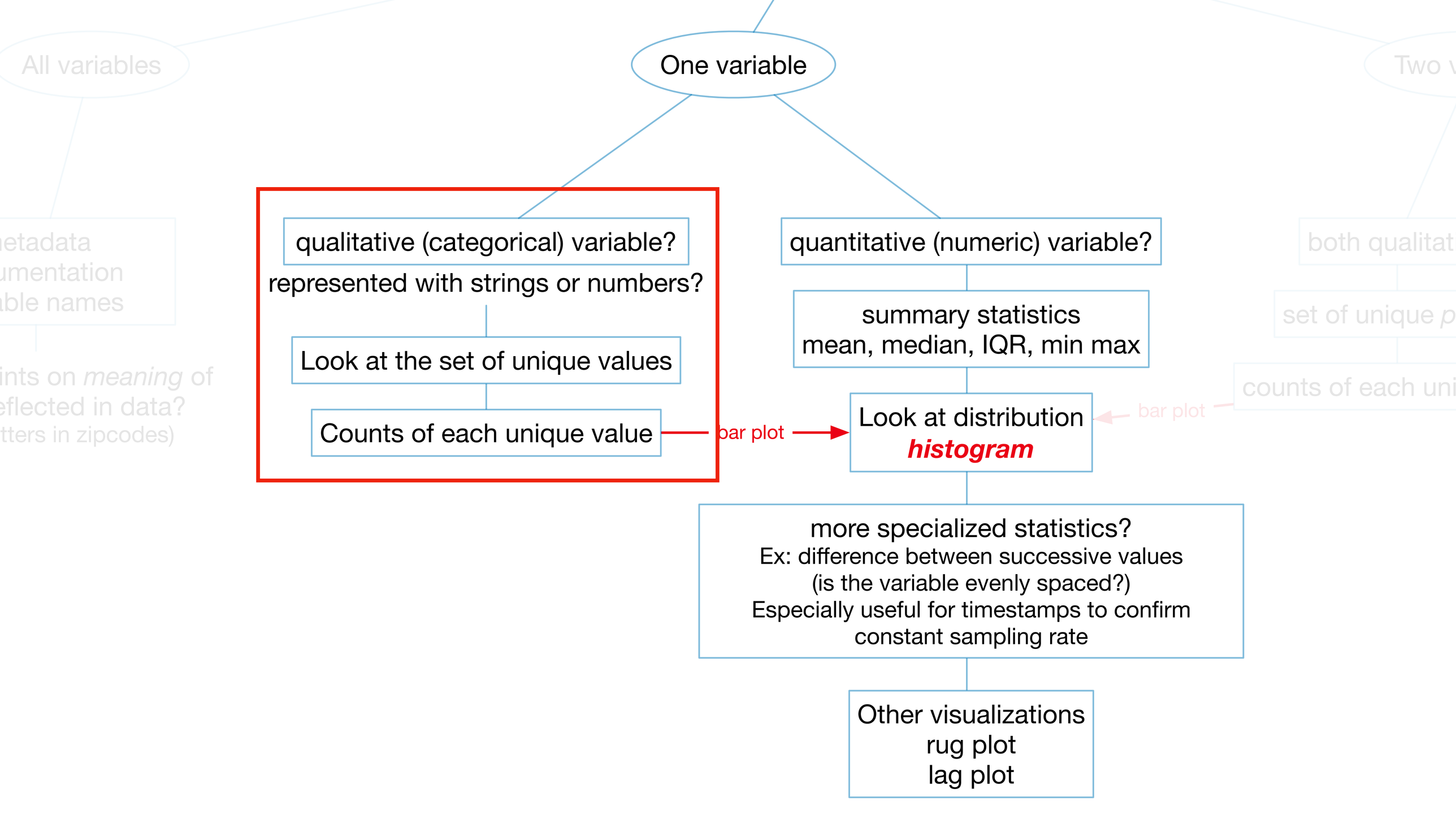
Counts of each unique value

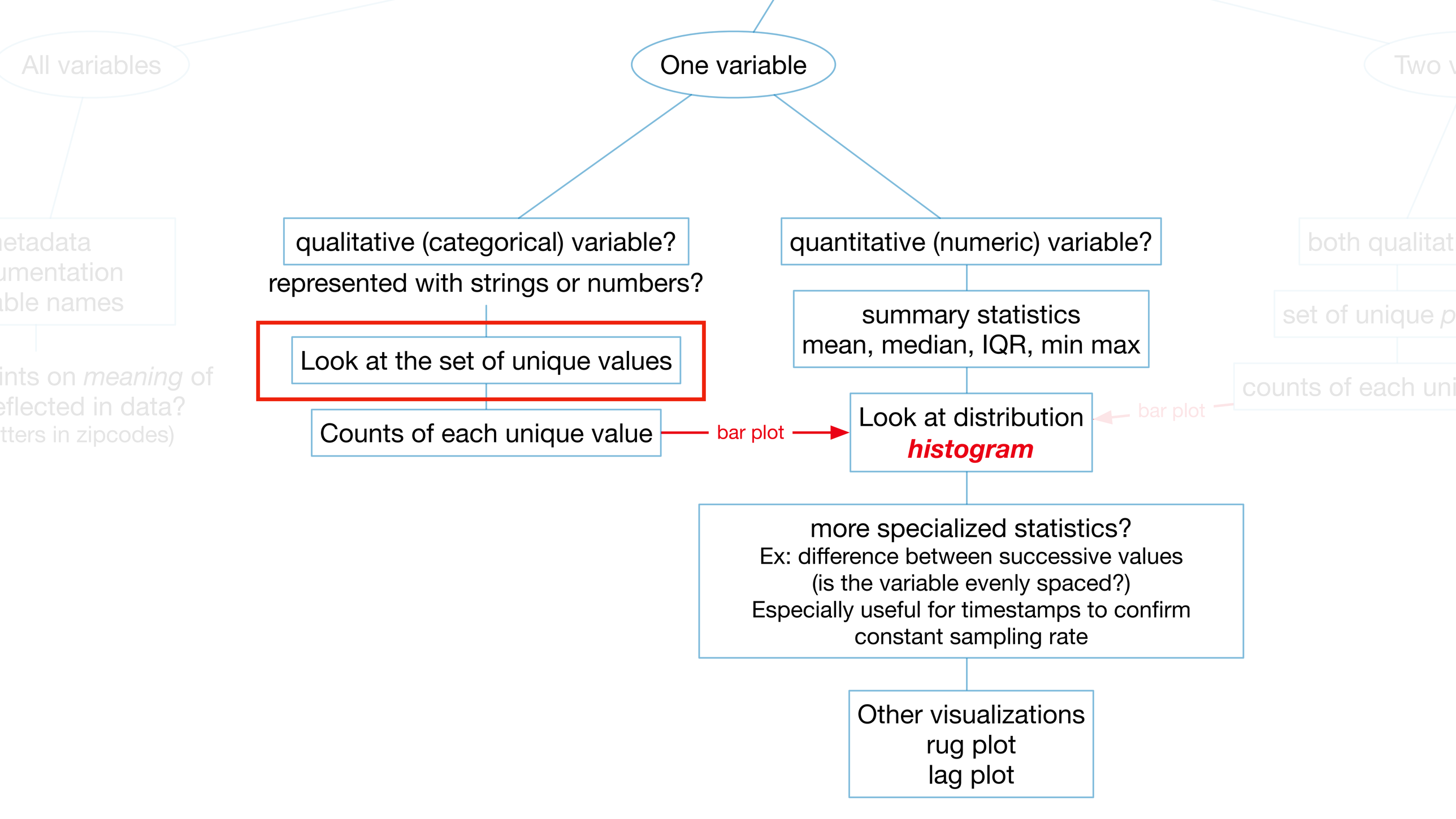
b

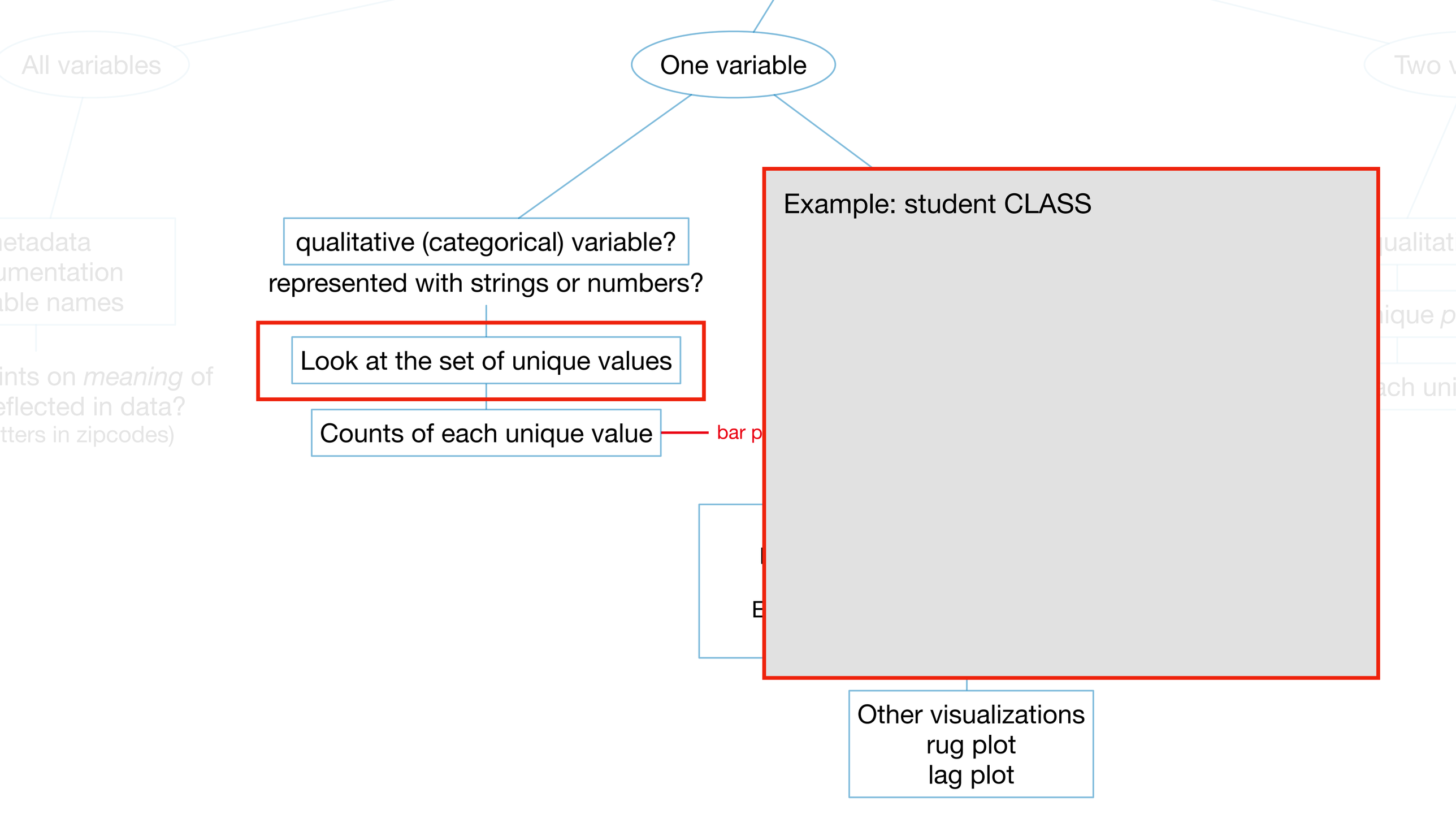

All variables
metadata
documentation
variable names
Any constraints on *meaning* of
variable reflected in data?
(ex: any letters in zipcodes)

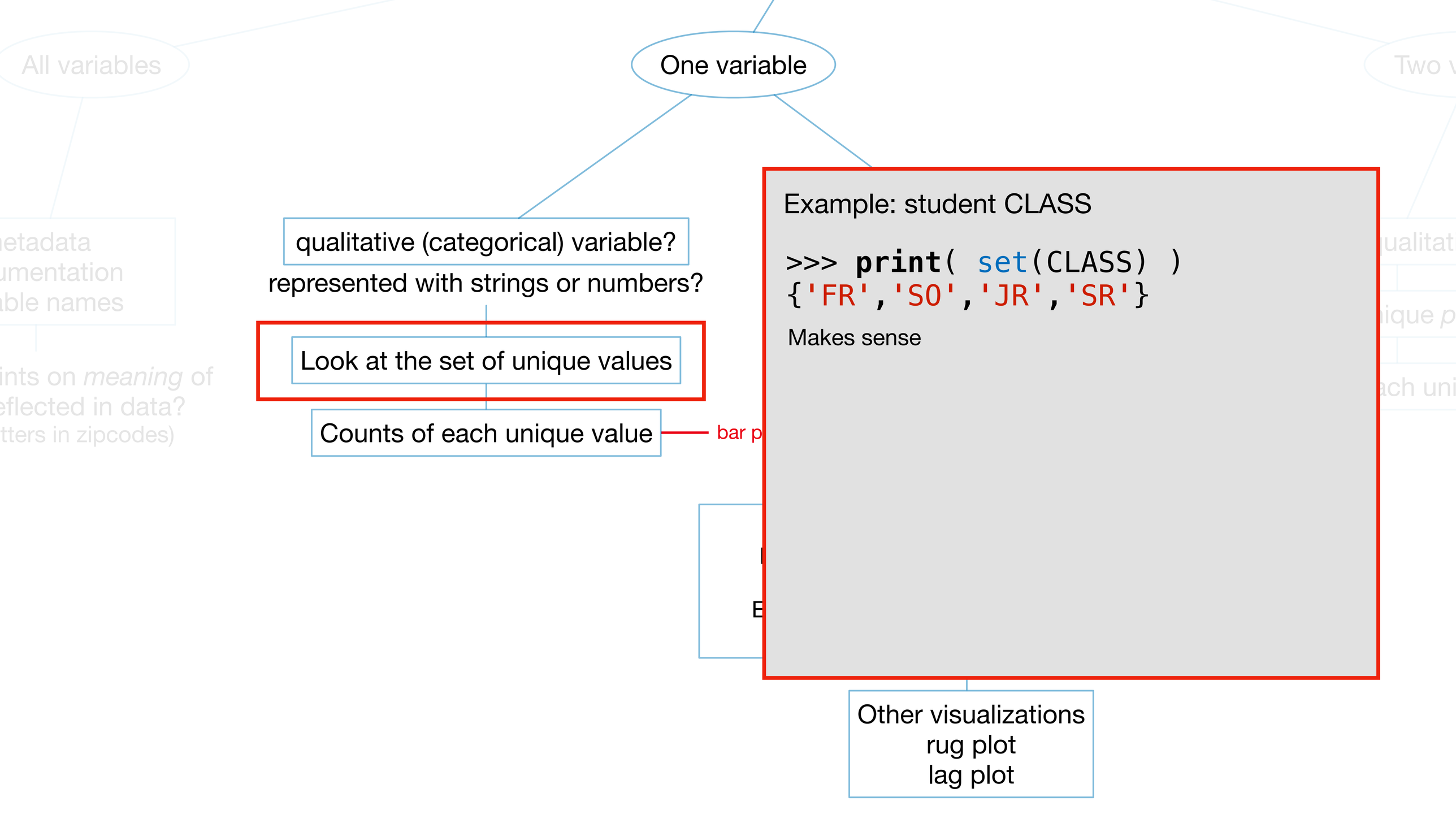

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
both qualitat
set of unique p
counts of each uni
qualitative (categorical) variable?
represented with strings or numbers?
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at the set of unique values
Counts of each unique value
bar plot
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

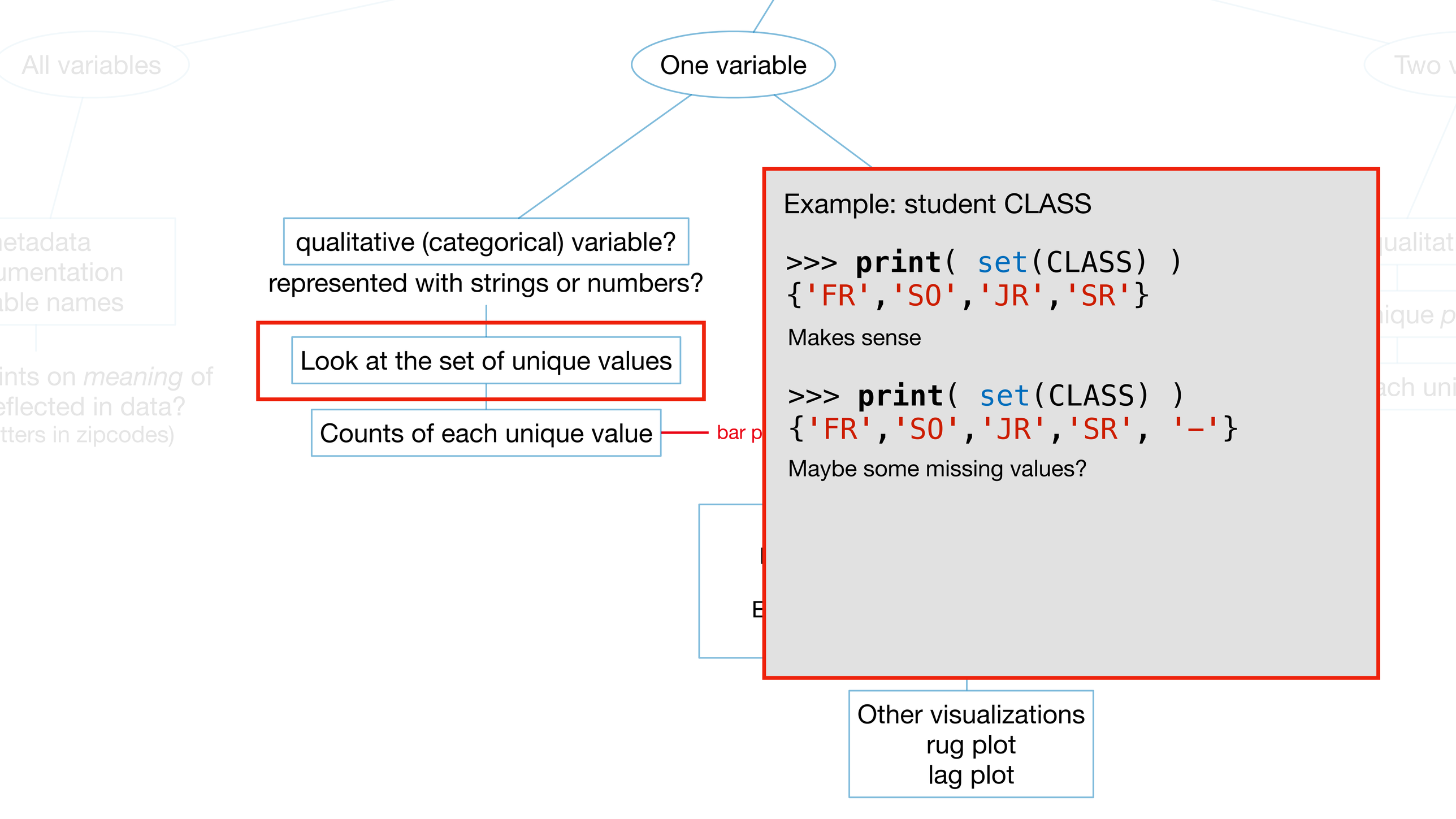

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
both qualitat
set of unique p
counts of each uni
qualitative (categorical) variable?
represented with strings or numbers?
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at the set of unique values
Counts of each unique value
bar plot
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

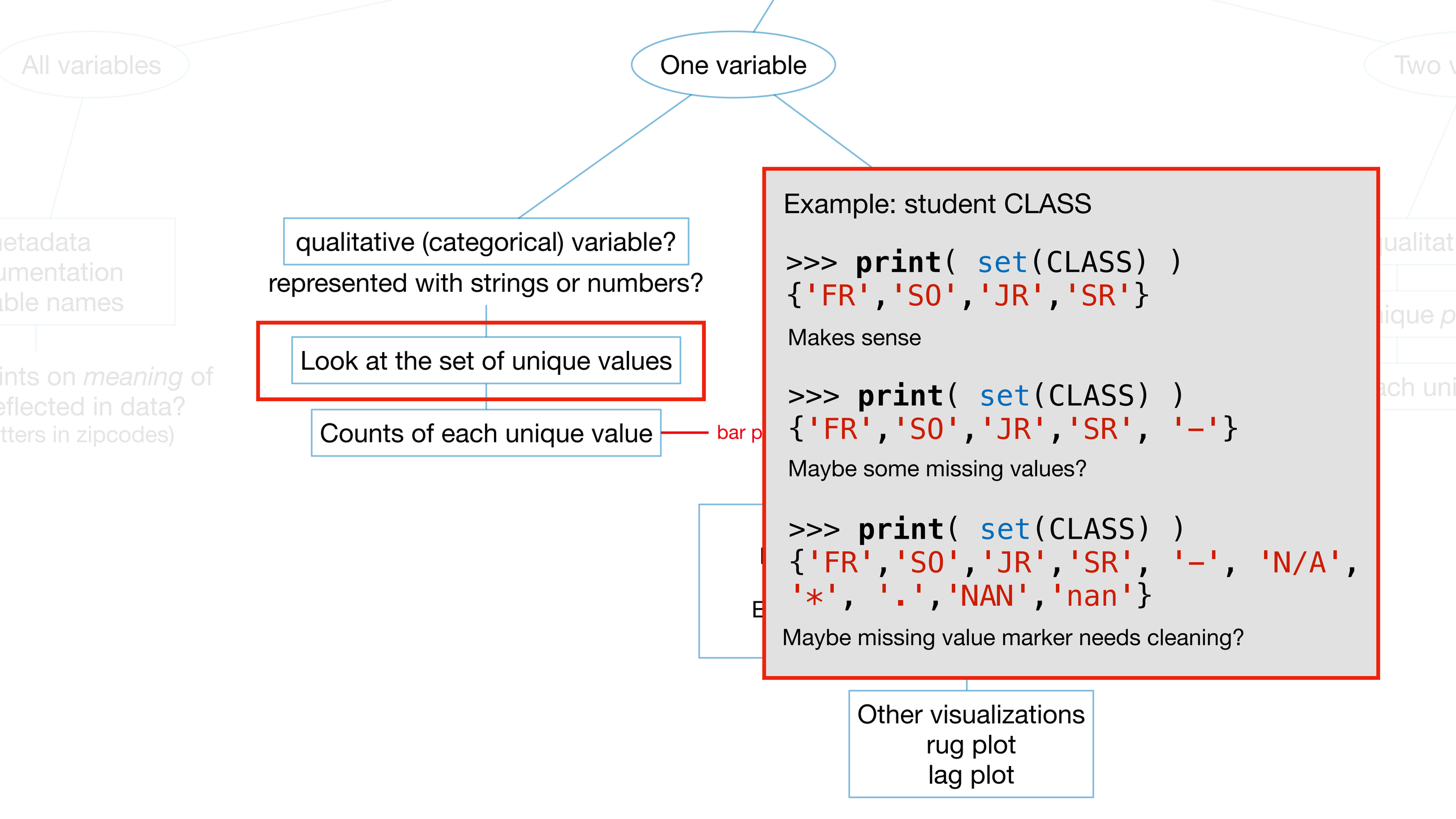

One variable
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar plot
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

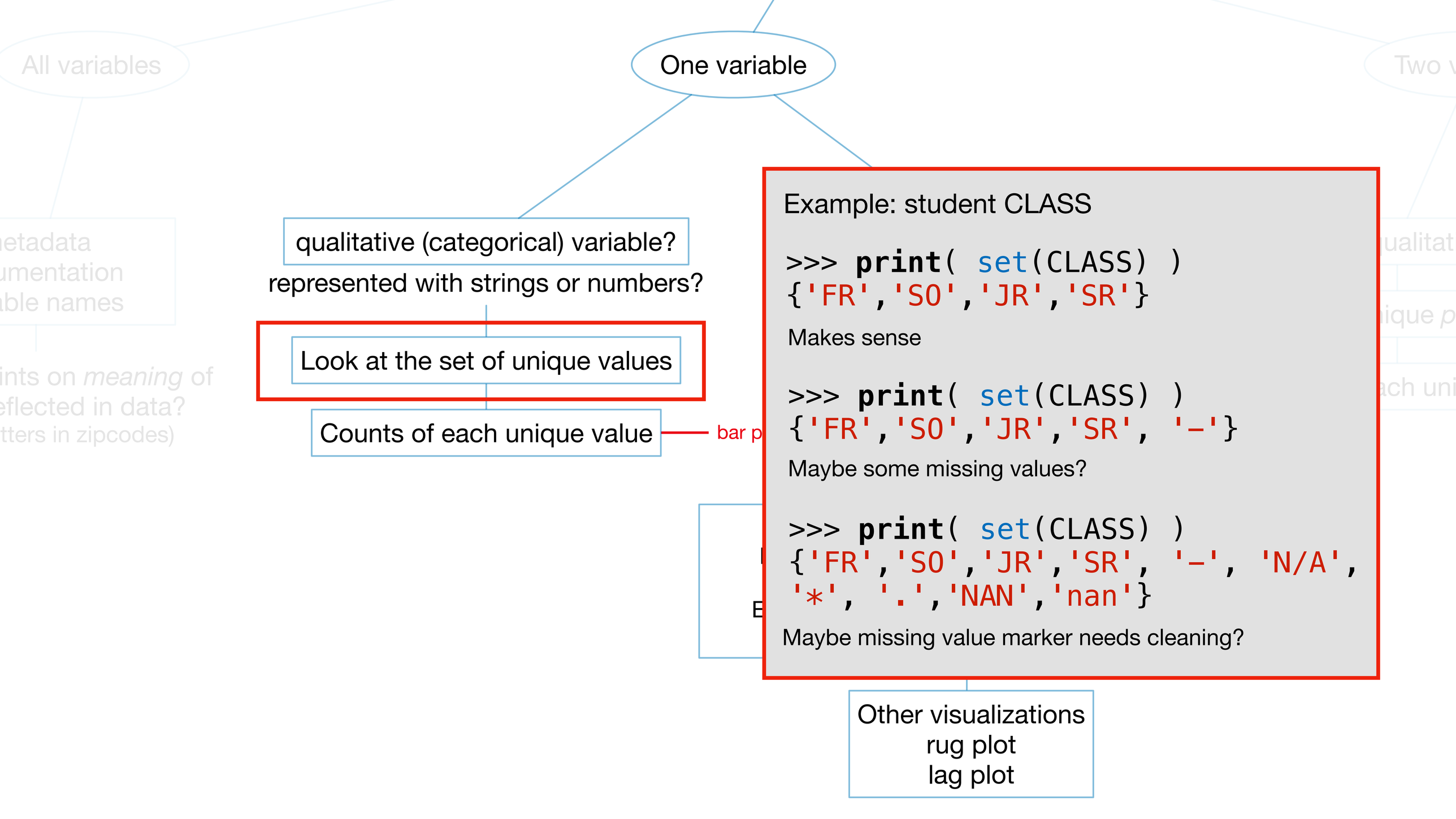

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
both qualitat
set of unique p
counts of each uni
qualitative (categorical) variable?
represented with strings or numbers?
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at the set of unique values
Counts of each unique value
bar plot
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

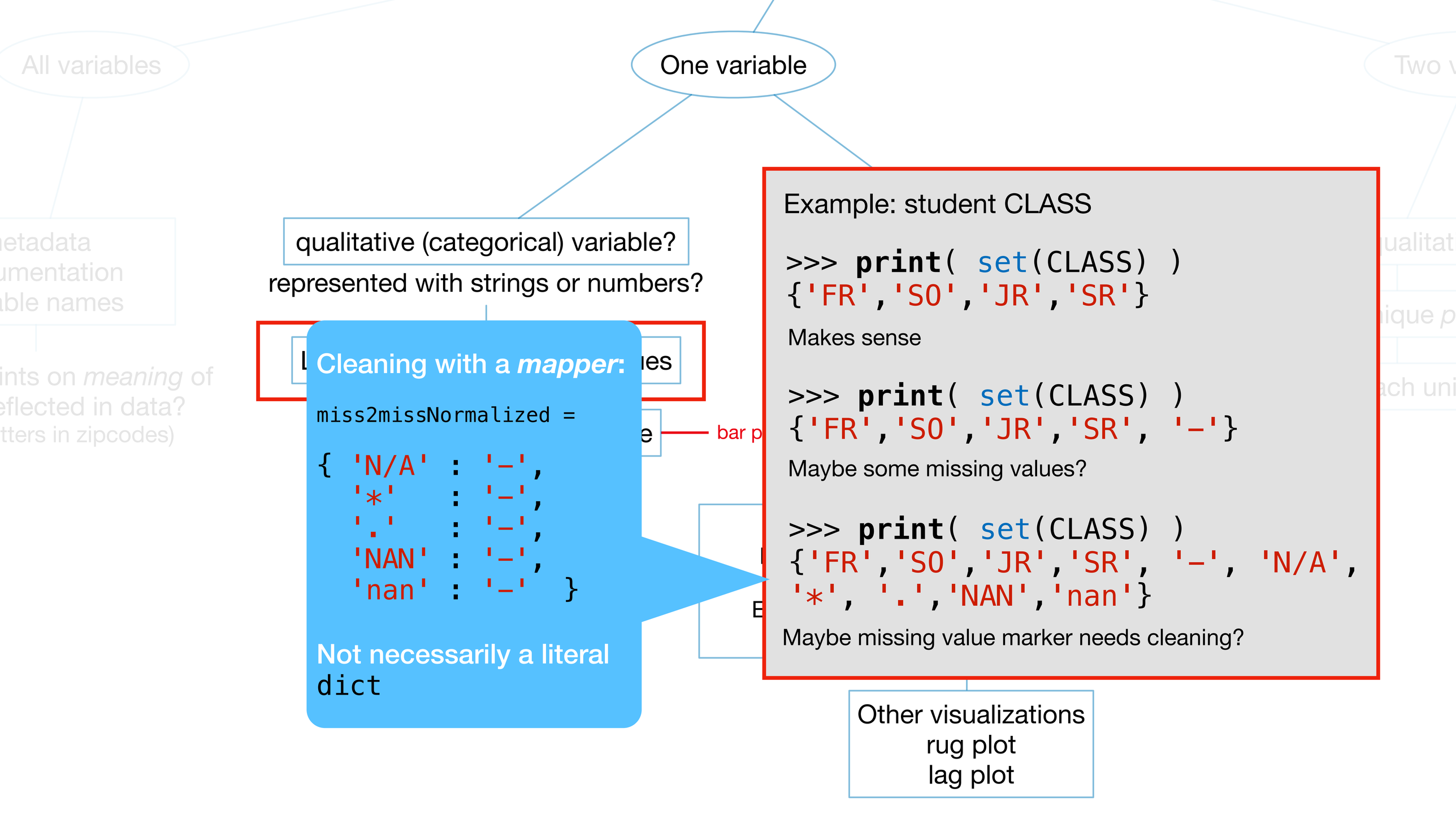

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
qualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
Other visualizations
rug plot
lag plot

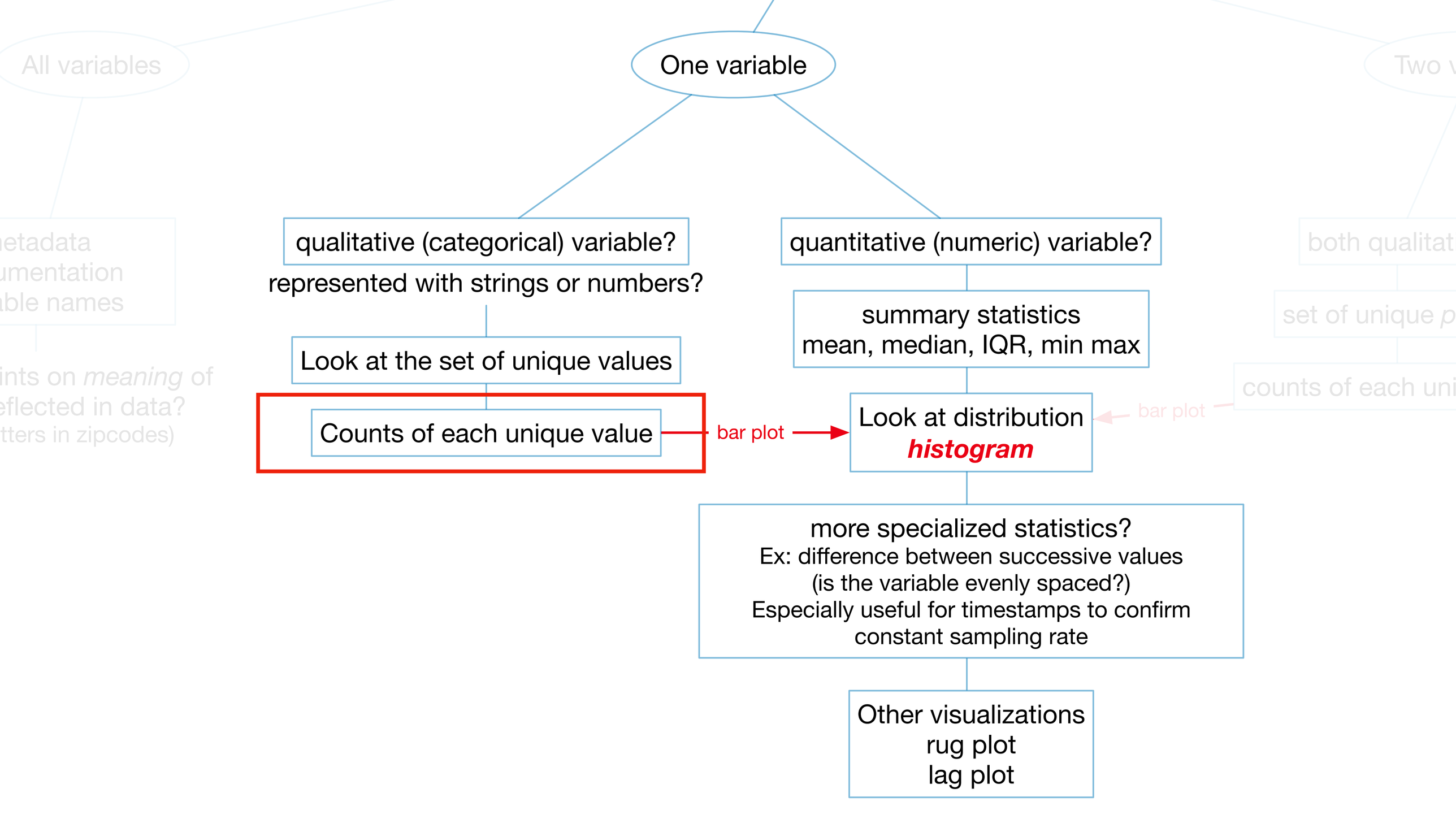

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
qualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(set(CLASS))
{'FR','SO','JR','SR'}
Makes sense
Other visualizations
rug plot
lag plot

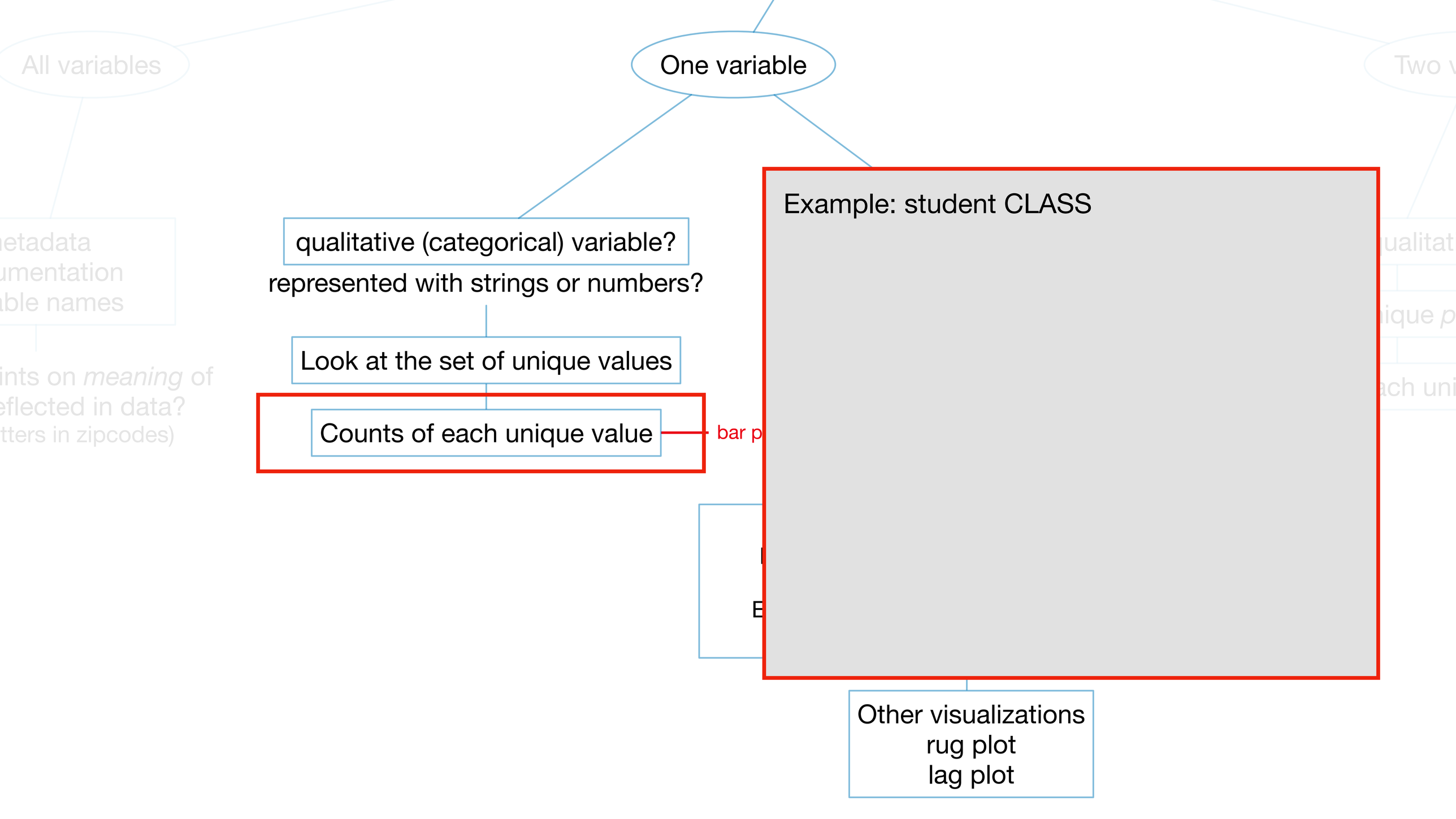

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
qualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(set(CLASS))
{'FR','SO','JR','SR'}
Makes sense
>>> print(set(CLASS))
{'FR','SO','JR','SR', '-'}
Maybe some missing values?
Other visualizations
rug plot
lag plot

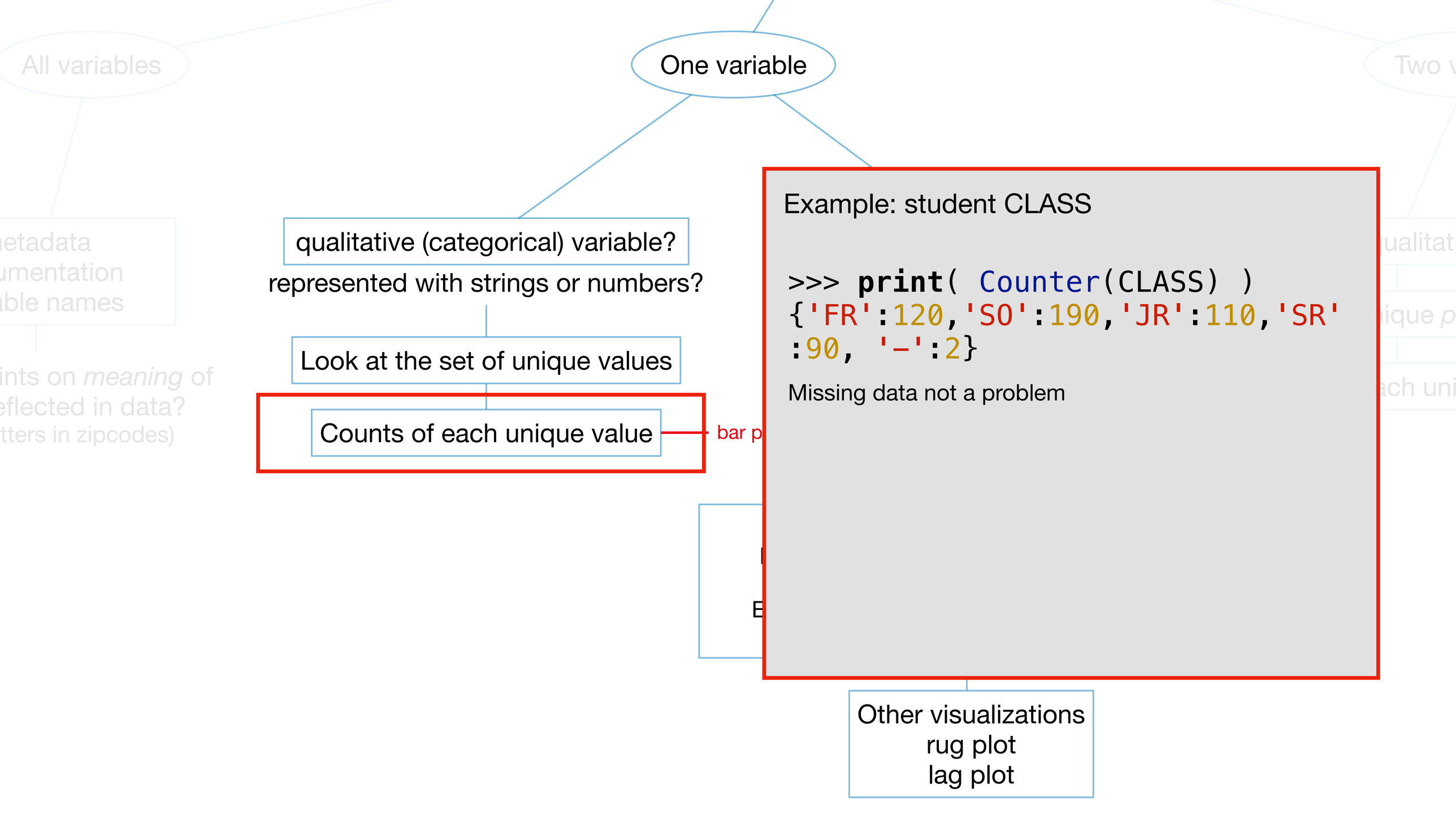

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
ualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(set(CLASS))
{'FR','SO','JR','SR'}
Makes sense
>>> print(set(CLASS))
{'FR','SO','JR','SR', '-'}
Maybe some missing values?
>>> print(set(CLASS))
{'FR','SO','JR','SR', '-', 'N/A',
'*', '.','NAN','nan'}
Maybe missing value marker needs cleaning?
Other visualizations
rug plot
lag plot

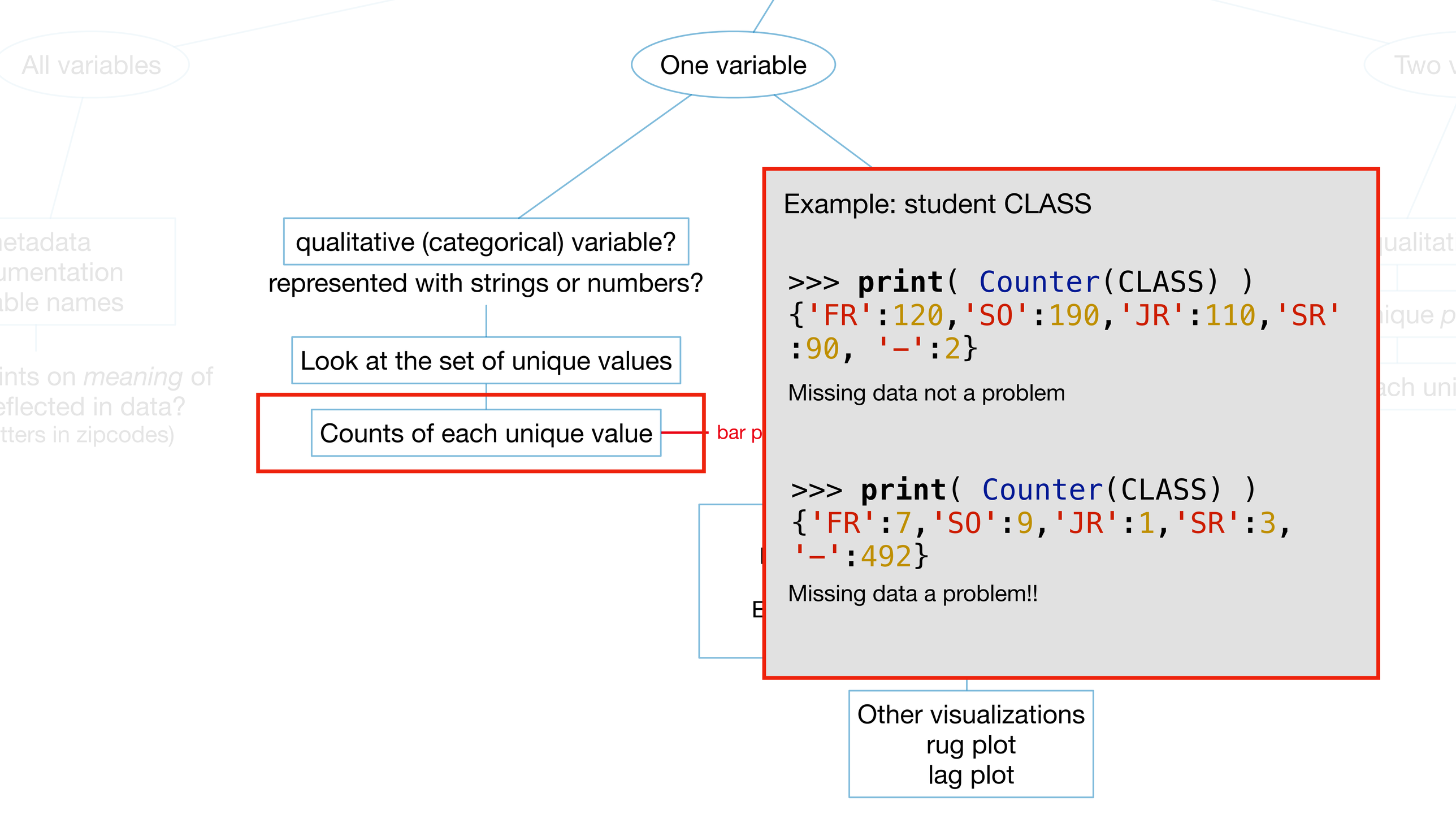

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
ualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(set(CLASS))
{'FR','SO','JR','SR'}
Makes sense
>>> print(set(CLASS))
{'FR','SO','JR','SR', '-'}
Maybe some missing values?
>>> print(set(CLASS))
{'FR','SO','JR','SR', '-', 'N/A',
'*', '.','NAN','nan'}
Maybe missing value marker needs cleaning?
Other visualizations
rug plot
lag plot

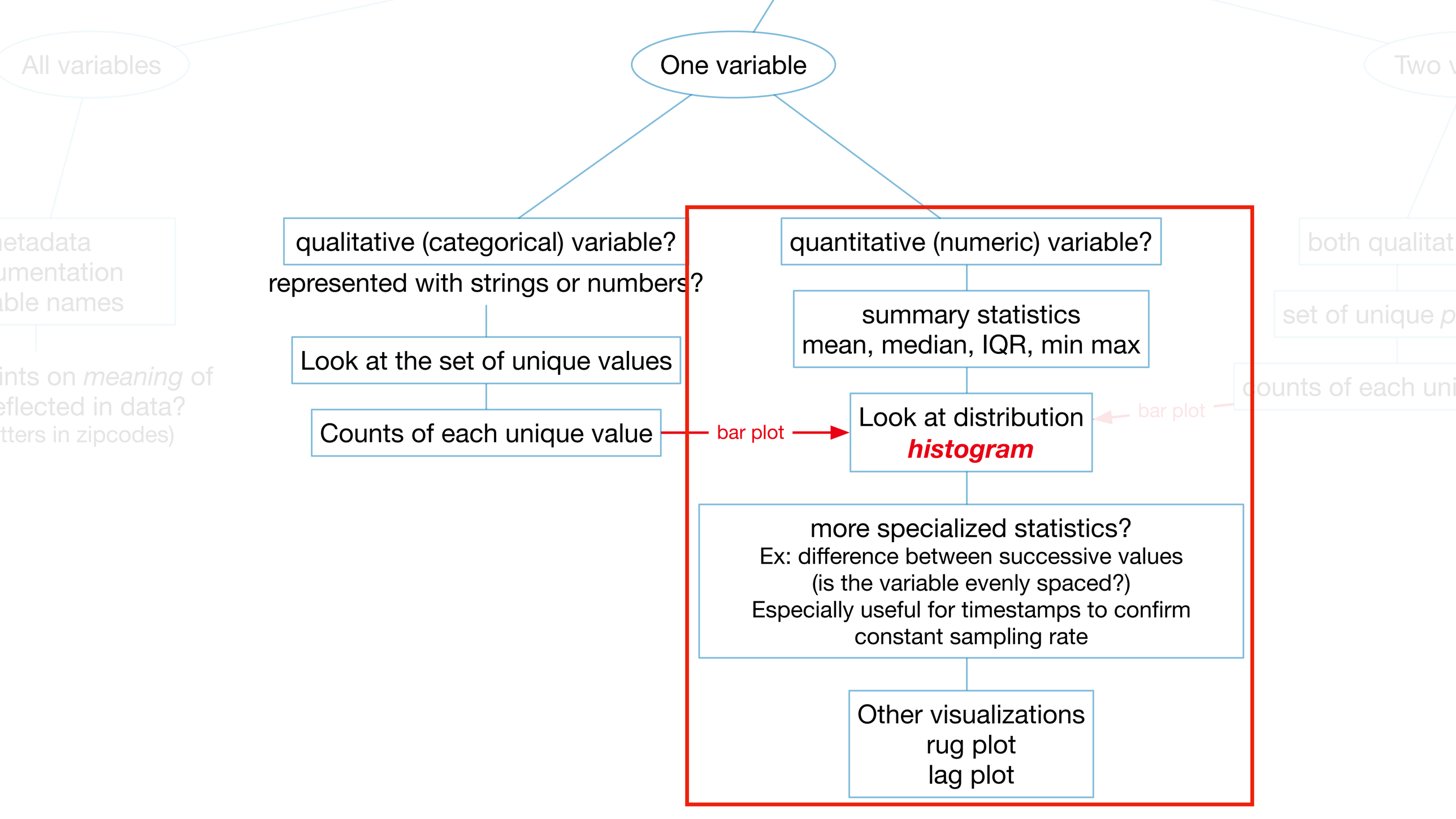

All variables
One variable
Two v
qualitative (categorical) variable?
represented with strings or numbers?
Cleaning with a *mapper*:
miss2missNormalized =
{ 'N/A' : '–',
'*' : '–',
'.' : '–',
'NAN' : '–',
'nan' : '–' }
Not necessarily a literal dict
bar p
Example: student CLASS
>>> print(set(CLASS))
{'FR','SO','JR','SR'}
Makes sense
>>> print(set(CLASS))
{'FR','SO','JR','SR', '–'}
Maybe some missing values?
>>> print(set(CLASS))
{'FR','SO','JR','SR', '–', 'N/A',
'*', '.','NAN','nan'}
Maybe missing value marker needs cleaning?
Other visualizations
rug plot
lag plot

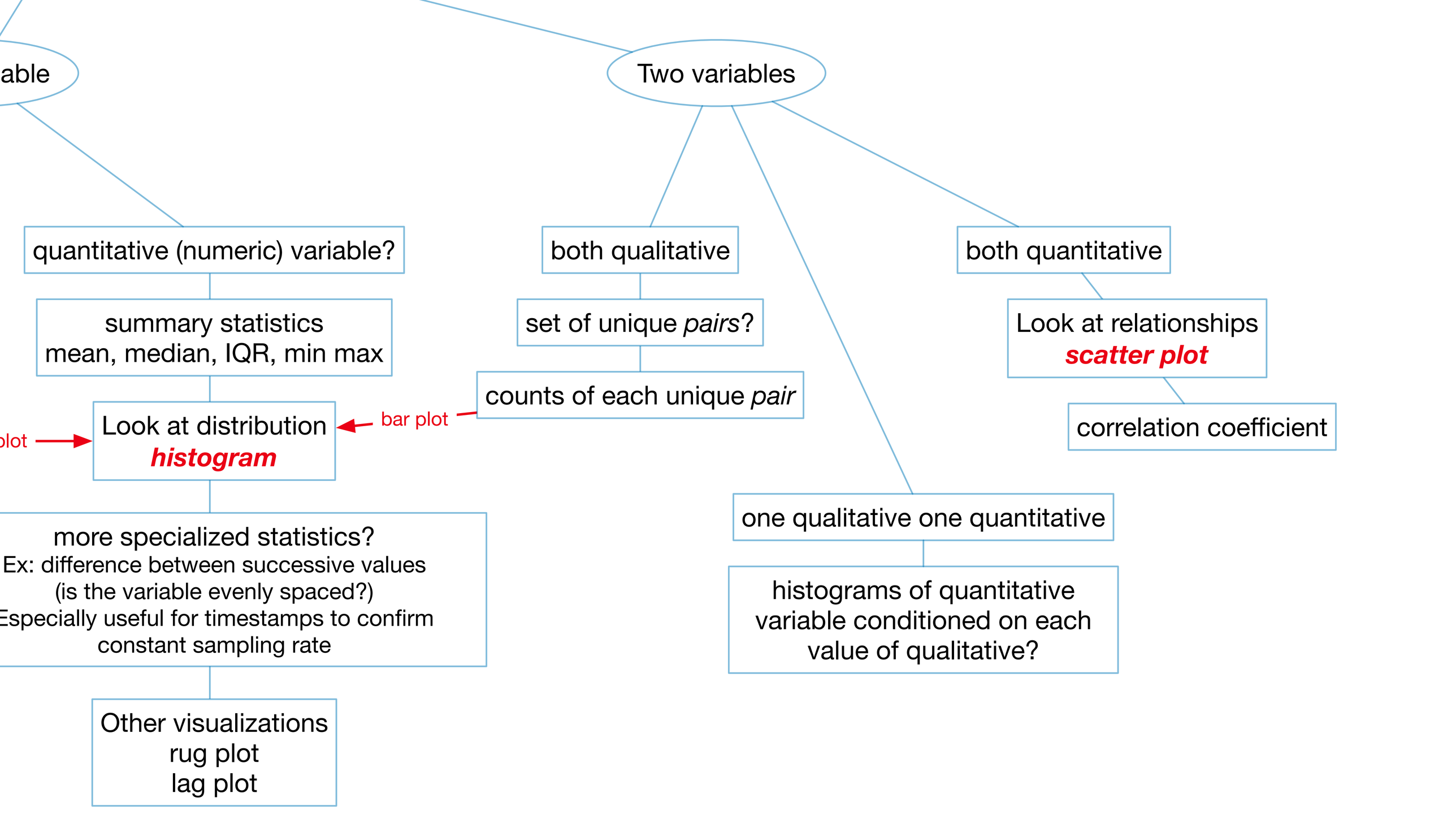

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
both qualitat
set of unique p
counts of each uni
qualitative (categorical) variable?
represented with strings or numbers?
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at the set of unique values
Counts of each unique value
bar plot
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

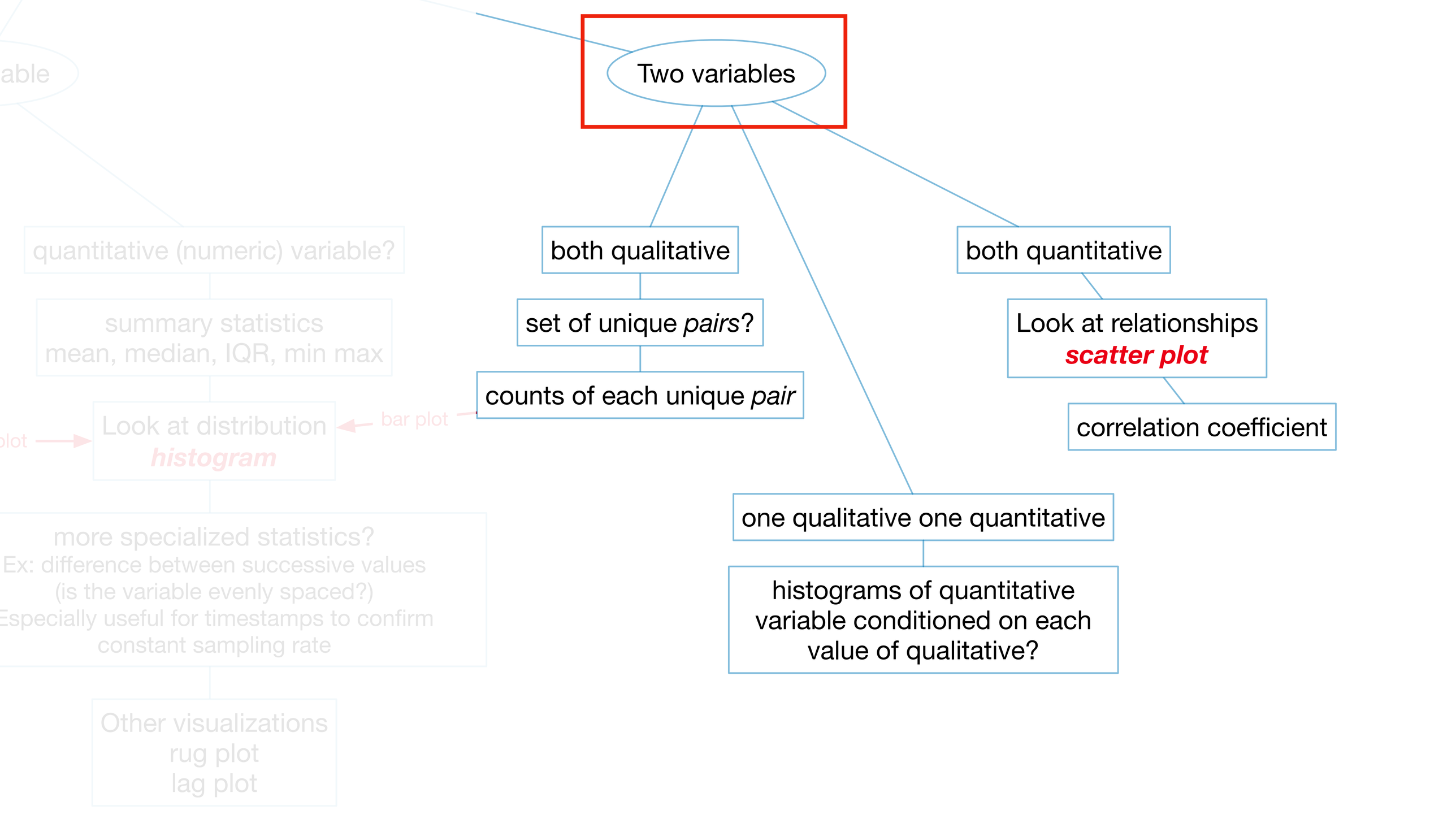

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
qualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
Other visualizations
rug plot
lag plot

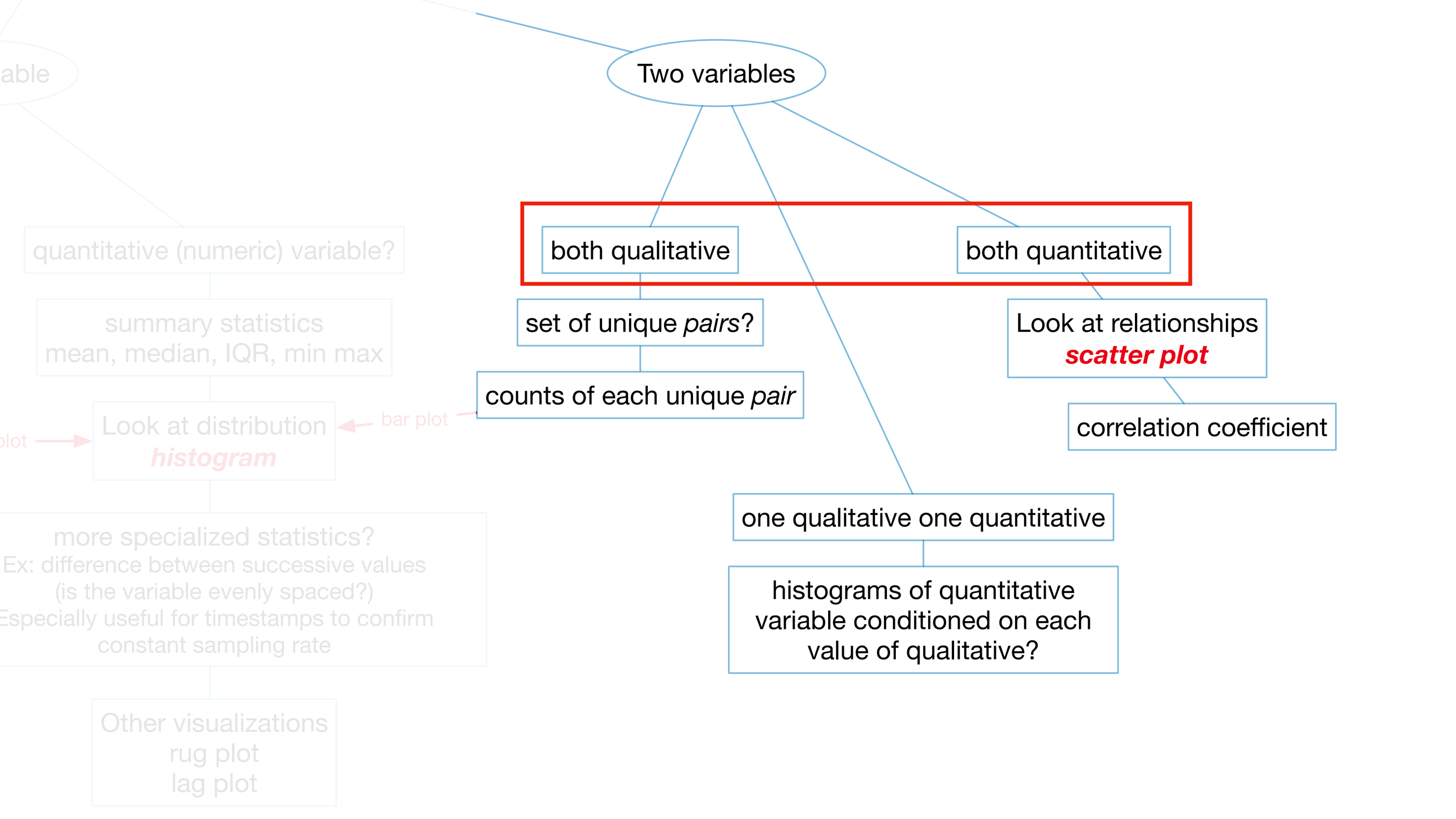

All variables
One variable
Two v
etadata
umentation
able names
ints on *meaning* of
eflected in data?
tters in zipcodes)
qualitat
ique p
ach uni
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(Counter(CLASS))
{'FR':120,'SO':190,'JR':110,'SR'
:90, '–':2}
Missing data not a problem
Other visualizations
rug plot
lag plot

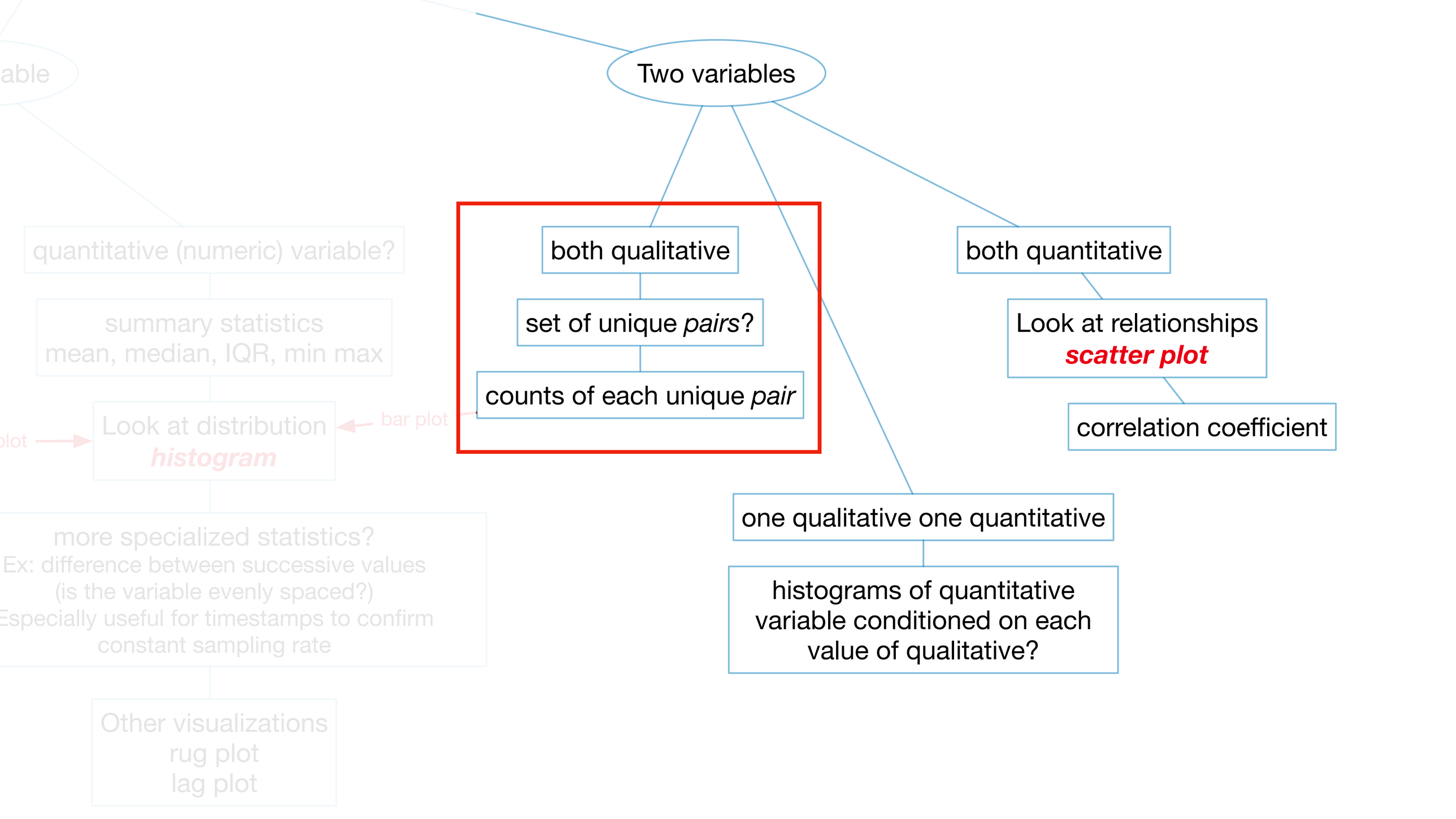

All variables
One variable
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar p
Example: student CLASS
>>> print(Counter(CLASS))
{'FR':120,'SO':190,'JR':110,'SR':90, '–':2}
Missing data not a problem
>>> print(Counter(CLASS))
{'FR':7,'SO':9,'JR':1,'SR':3, '–':492}
Missing data a problem!!
Other visualizations
rug plot
lag plot

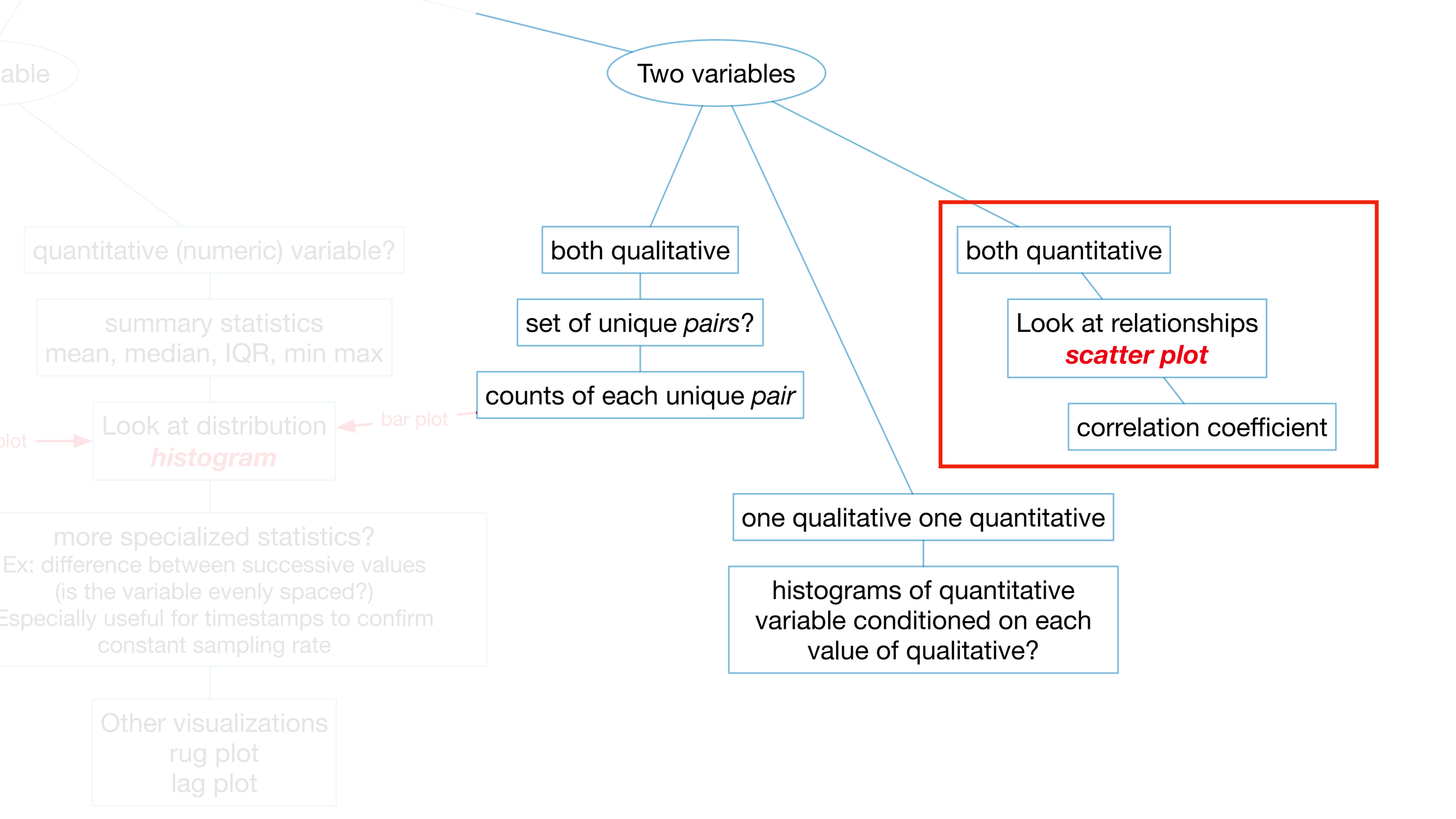

One variable
qualitative (categorical) variable?
represented with strings or numbers?
Look at the set of unique values
Counts of each unique value
bar plot
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at distribution
histogram
bar plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot


able
Two variables
quantitative (numeric) variable?
both qualitative
both quantitative
summary statistics
mean, median, IQR, min max
set of unique *pairs*?
Look at relationships
scatter plot
counts of each unique *pair*
bar plot
plot
Look at distribution
histogram
correlation coefficient
one qualitative one quantitative
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
histograms of quantitative
variable conditioned on each
value of qualitative?
Other visualizations
rug plot
lag plot

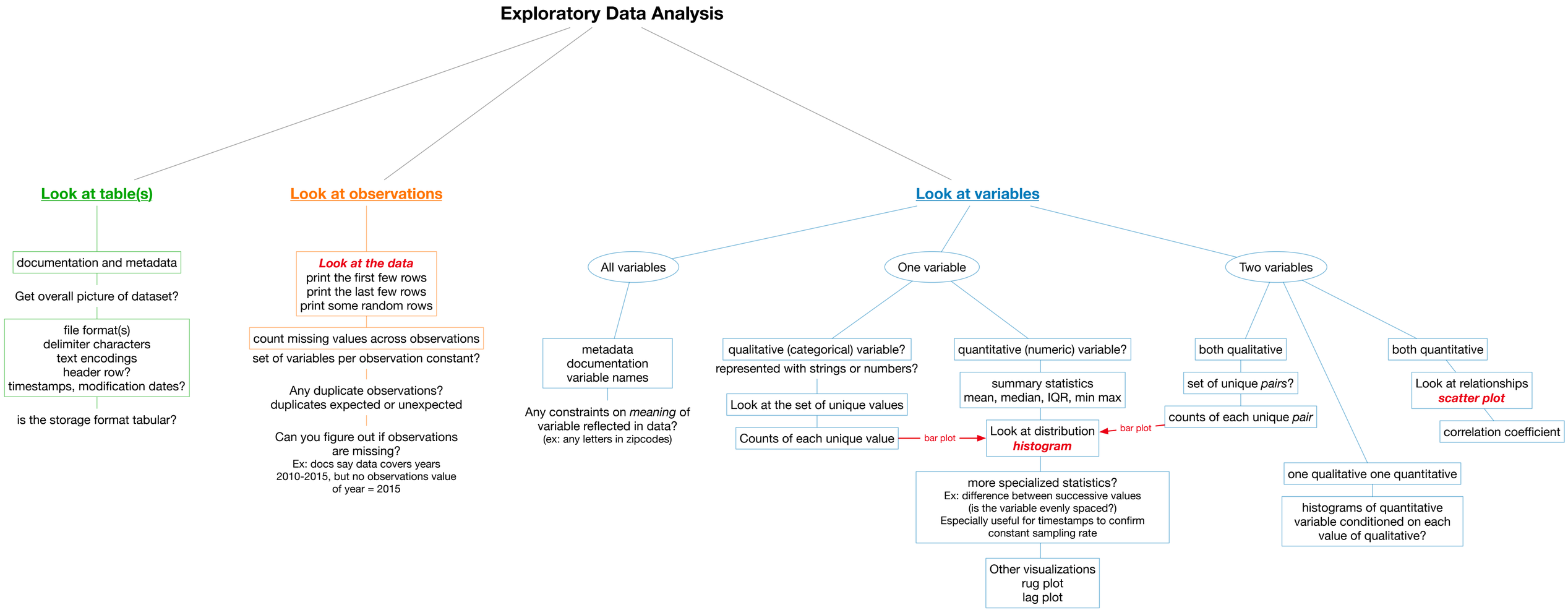

Two variables
both qualitative
set of unique *pairs*?
counts of each unique *pair*
both quantitative
Look at relationships
scatter plot
correlation coefficient
one qualitative one quantitative
histograms of quantitative variable conditioned on each value of qualitative?
bar plot
Look at distribution
histogram
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot


Two variables
both qualitative
both quantitative
set of unique *pairs*?
counts of each unique *pair*
Look at relationships
scatter plot
correlation coefficient
one qualitative one quantitative
histograms of quantitative variable conditioned on each value of qualitative?
bar plot
Look at distribution
histogram
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot

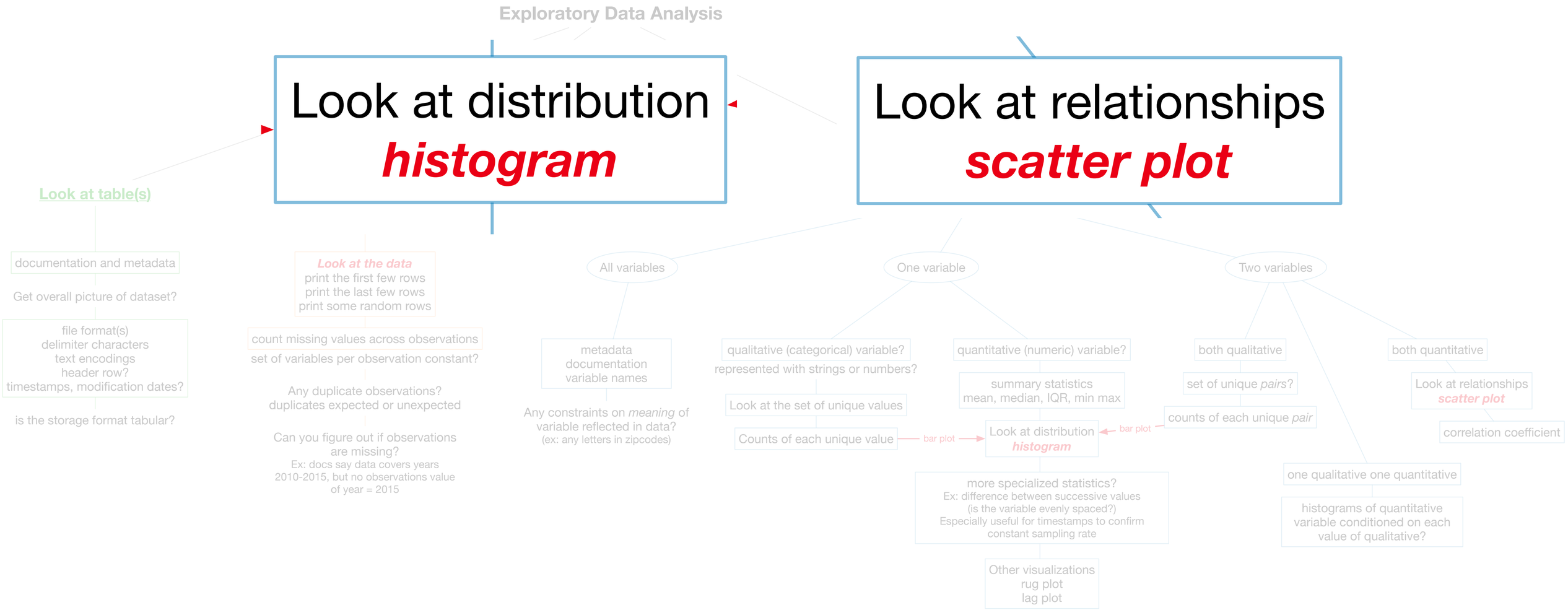

Two variables
both qualitative
set of unique *pairs*?
counts of each unique *pair*
both quantitative
Look at relationships
scatter plot
correlation coefficient
one qualitative one quantitative
histograms of quantitative variable conditioned on each value of qualitative?
bar plot
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at distribution
histogram
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot


able
Two variables
quantitative (numeric) variable?
summary statistics
mean, median, IQR, min max
Look at distribution
histogram
bar plot
plot
more specialized statistics?
Ex: difference between successive values
(is the variable evenly spaced?)
Especially useful for timestamps to confirm
constant sampling rate
Other visualizations
rug plot
lag plot
both qualitative
set of unique pairs?
counts of each unique pair
both quantitative
Look at relationships
scatter plot
correlation coefficient
one qualitative one quantitative
histograms of quantitative
variable conditioned on each
value of qualitative?

Our (new) best friends



Our (new) best friends



Our (new) best friends

Exploratory Data Analysis

Look at distribution
histogram

Look at relationships
scatter plot

Look at table(s)

documentation and metadata

Get overall picture of dataset?

file format(s)
delimiter characters
text encodings
header row?
timestamps, modification dates?

is the storage format tabular?

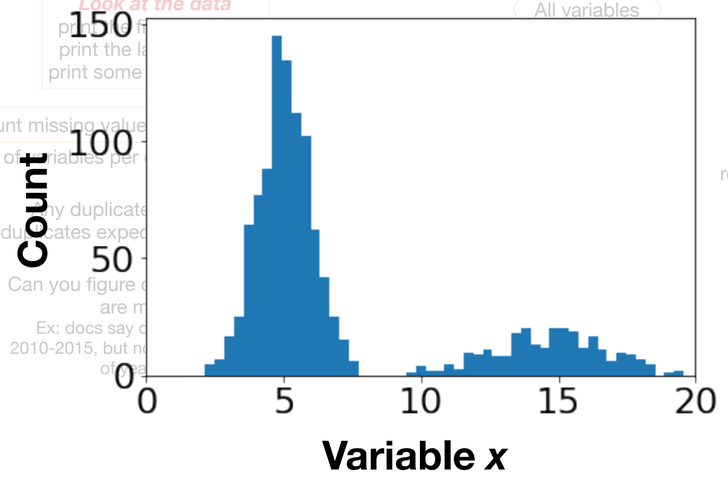
Look at the data

print the la
print some

count missing value
set of

y duplicate
du cates expec

Can you figure
are m
Ex: docs say
2010-2015, but n
of



All variables

One variable

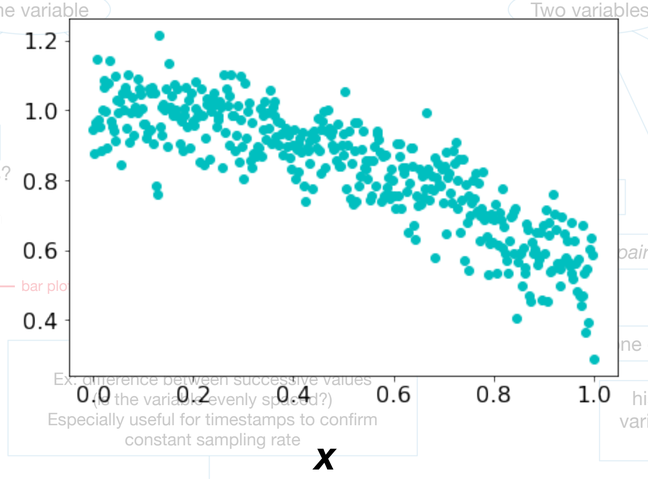
qualitative (categorical) variable?

represented with strings or numbers?

Look at the set of unique values

Counts of each unique value

bar plo



Ex: difference between successive values
(is the variable evenly sampled?)
Especially useful for timestamps to confirm
constant sampling rate

Other visualizations
rug plot
lag plot

Two variables

both quantitative

Look at relationships
scatter plot

correlation coefficient

pair

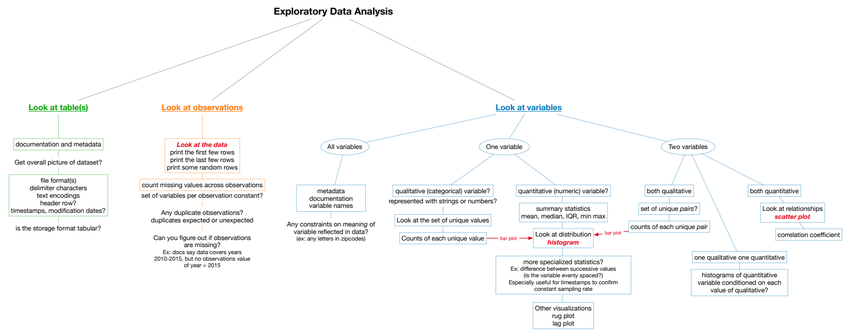
ne qualitative one quantitative

histograms of quantitative
variable conditioned on each
value of qualitative?

Summary

Cleaning data and processing data

- Many dimensions to cleaning data: problem- and domain-specific issues
- Tidy data (values, variables, observations, tables) is a great mental model for thinking about a dataset, even when the data are not literally organized in that way.
- Lots of data? Need code to help automate cleaning —but automation can hide problems
- Provenance: Keep a record of changes
- Cleaning data requires exploring data
- Need to **LOOK** at the data



Summary

Cleaning data and processing data

- Many dimensions to cleaning data: problem- and domain-specific issues
- Tidy data (values, variables, observations, tables) is a great mental model for thinking about a dataset, even when the data are not literally organized in that way.
- Lots of data? Need code to help automate cleaning —but automation can hide problems
- Provenance: Keep a record of changes
- Cleaning data requires exploring data
- Need to **LOOK** at the data