CDS501: PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

Chapter 2: Data Types and Formats

Outline

- Primary and Secondary
- Types of Data
- Data Formats

Primary Data and Secondary Data

- Primary
  - Data that is collected directly from the data source without going through any existing sources
  - Expensive and time consuming
  - Data is reliable, authentic, up to date and objective (collected with purpose)
  - Ownership – belong to the organization
  - Interview, Focus Group (Qualitative)
  - Survey/Questionnaire, Observation, Experiment (Quantitative)

Primary Data and Secondary Data

- Secondary
  - Data that has been collected in the past by someone else but made available for others to use
  - Affordable and requires very little to no cost to acquire them
  - Easily accessible (shared publicly)
  - Data may not be suited to the project needs
  - Data may not authentic – need further verification
  - Data may be outdated
  - Kaggle, UCI Repository

Types of Data

- Numerical data
- Categorical data
- Text
- Time series data
- Image data

Numerical Data

- Quantitative data
- Any data where data points are exact numbers
- Has no spatial and temporal structure

Numerical Data

- Quantitative data
- Any data where data points are exact numbers
- Has no spatial and temporal structure

Continuous

Assume any value (real numbers)

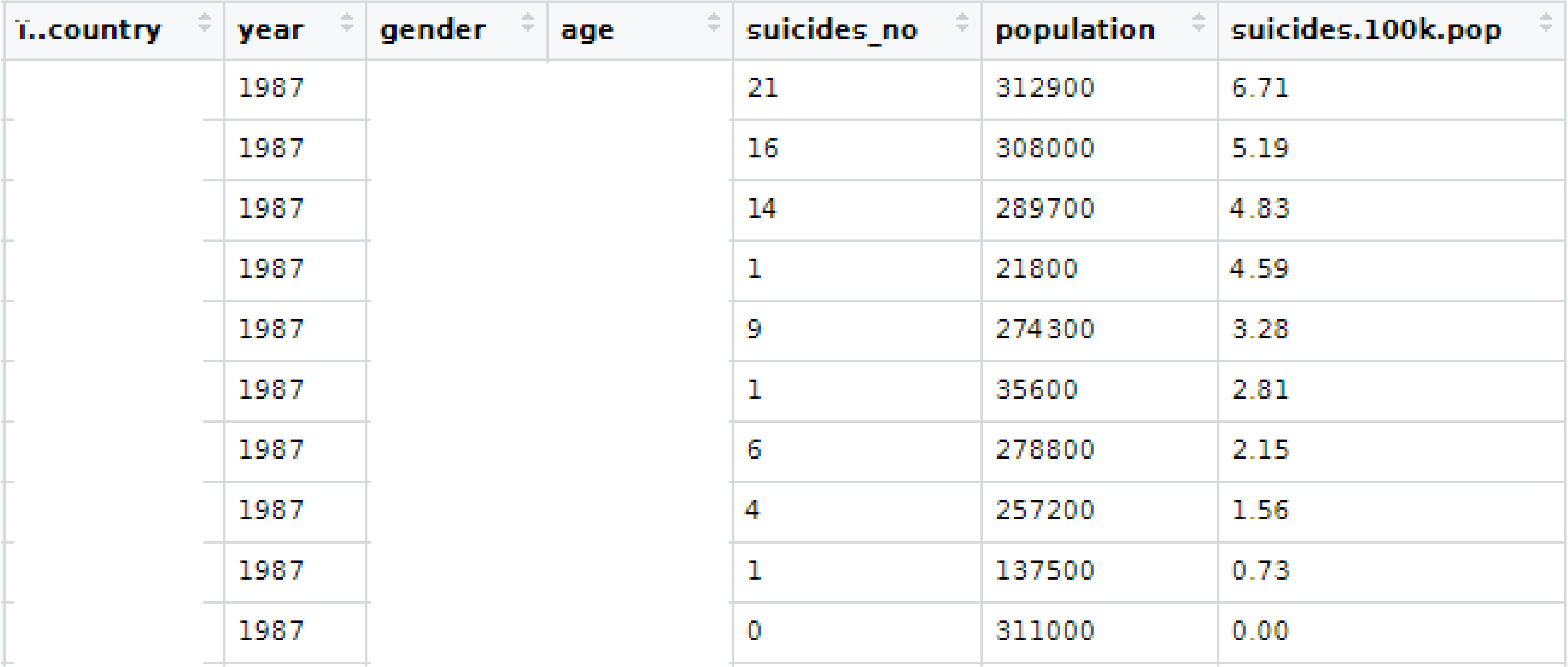
35.6, 10.0, 89.26

Discrete

Distinct values

3, 55, 10

Numerical Data



ï..country	year	gender	age	suicides_no	population	suicides.100k.pop
	1987			21	312900	6.71
	1987			16	308000	5.19
	1987			14	289700	4.83
	1987			1	21800	4.59
	1987			9	274300	3.28
	1987			1	35600	2.81
	1987			6	278800	2.15
	1987			4	257200	1.56
	1987			1	137500	0.73
	1987			0	311000	0.00

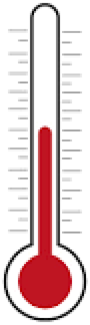
Continuous or Discrete?

salary

height

# of cars sold

# of students



body temperature

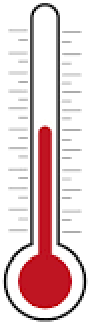
Continuous or Discrete?

salary
(continuous)

height
(continuous)

# of cars sold
(discrete)

# of students
(discrete)



body temperature
(continuous

Categorical Data

- Data that represents groups
- Can take numerical values, but the values have no meaning

Categorical Data

- Data that represents groups
- Can take numerical values, but the values have no meaning

Nominal

Categorical data without ordering

Gender, Town, Weather

Ordinal

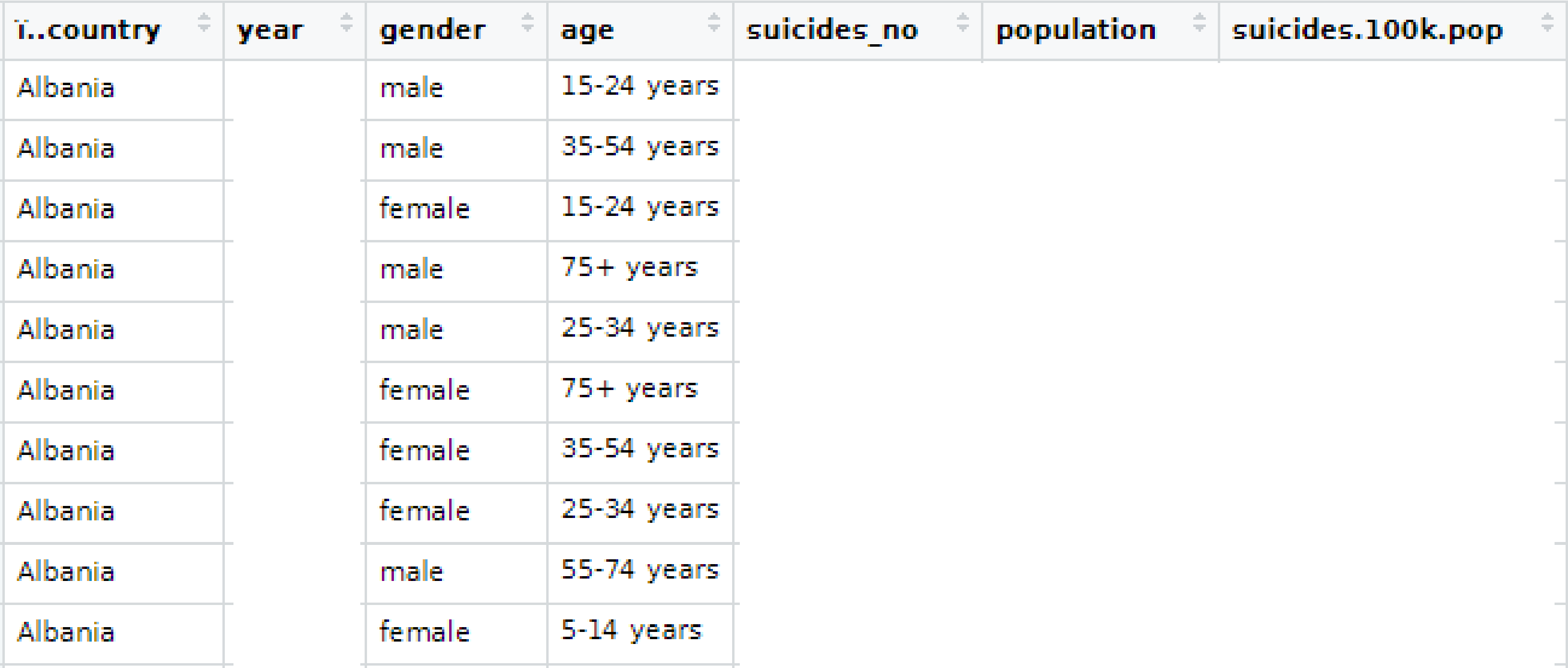
Categorical data with ordering

Size, Difficulty

Categorical Data

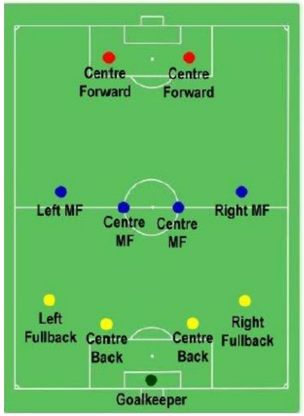
- Can take numerical values, but the values have no meaning
- Numerical data can be split into groups
  - House price
  - 0 – RM 200,000: cheap
  - RM 200,001 – RM 500,000: affordable
  - RM 500,001 – RM 1,000,000: expensive
  - RM 1,000,000 – ∞: super expensive

Categorical Data

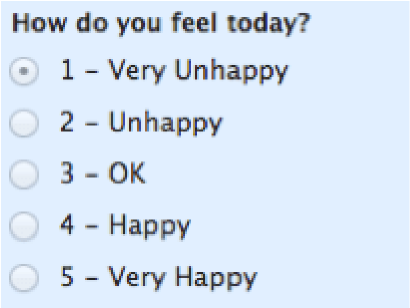


ï..country	year	gender	age	suicides_no	population	suicides.100k.pop
Albania		male	15-24 years			
Albania		male	35-54 years			
Albania		female	15-24 years			
Albania		male	75+ years			
Albania		male	25-34 years			
Albania		female	75+ years			
Albania		female	35-54 years			
Albania		female	25-34 years			
Albania		male	55-74 years			
Albania		female	5-14 years			

Nominal or Ordinal?



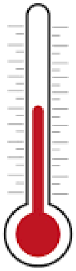
player’s position



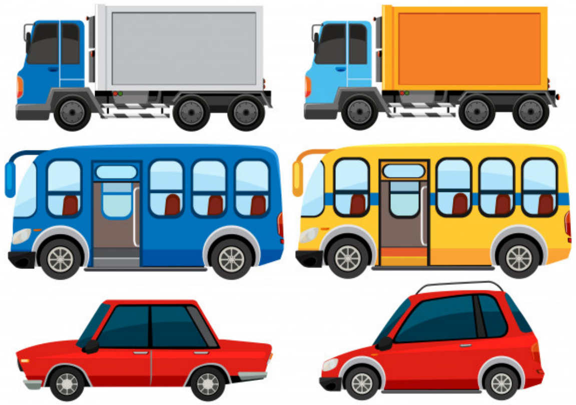
How do you feel today?

- ● 1 – Very Unhappy
- ○ 2 – Unhappy
- ○ 3 – OK
- ○ 4 – Happy
- ○ 5 – Very Happy

happiness

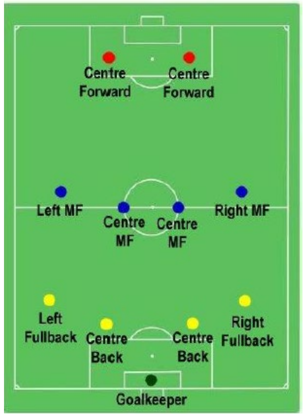


body temperature

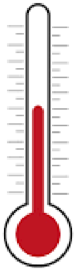


types of vehicles

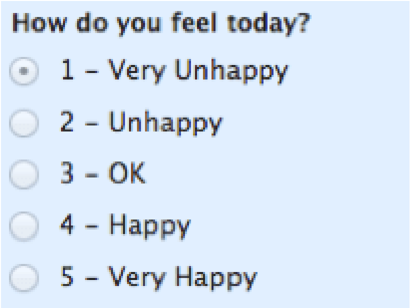
Nominal or Ordinal?



player's position
(nominal)



body temperature
(ordinal)



How do you feel today?

● 1 – Very Unhappy

○ 2 – Unhappy

○ 3 – OK

○ 4 – Happy

○ 5 – Very Happy

happiness
(ordinal)

types of vehicles
(nominal)

Text

- Words – needs to convert to a form that computers can process
- Tokenization – sentence to words
- Removing unnecessary punctuation, tags
- Removing stop words (most common words) – words that have not much semantic meaning
- Stemming & Lemmatization– reduce words to root words e.g. 'studies' to 'study'

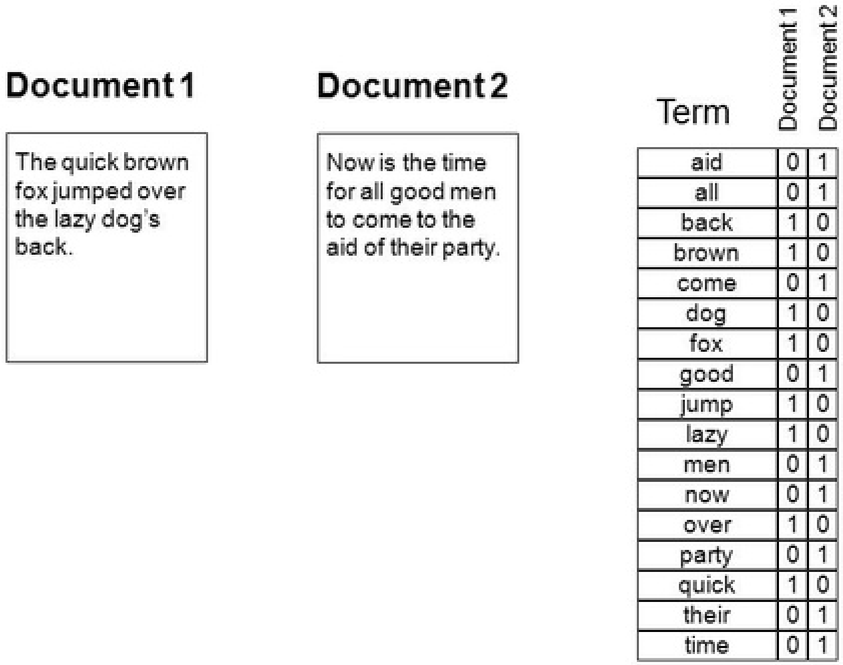
Text

- He is playing at football.
- He, is, playing, at, football, . (tokenization)
- He, is, playing, at, football (remove punctuation)
- playing, football (remove stop words)
- play, football (stemming & lemmatization)

- Represent the words using numerical representation technique such as Bag of Words (BOW), Word2Vec etc.

Text

- BOW turns each word into numbers by counting the occurrence of words



Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Time Series Data

- A sequence of values ordered by time
- The values (data points) take place in a given period of time in regular interval
- millisecond, sec, min, hour, day, week, … month, year, …
- Has temporal structure e.g. trends, seasonal, cyclic

Time Series Data

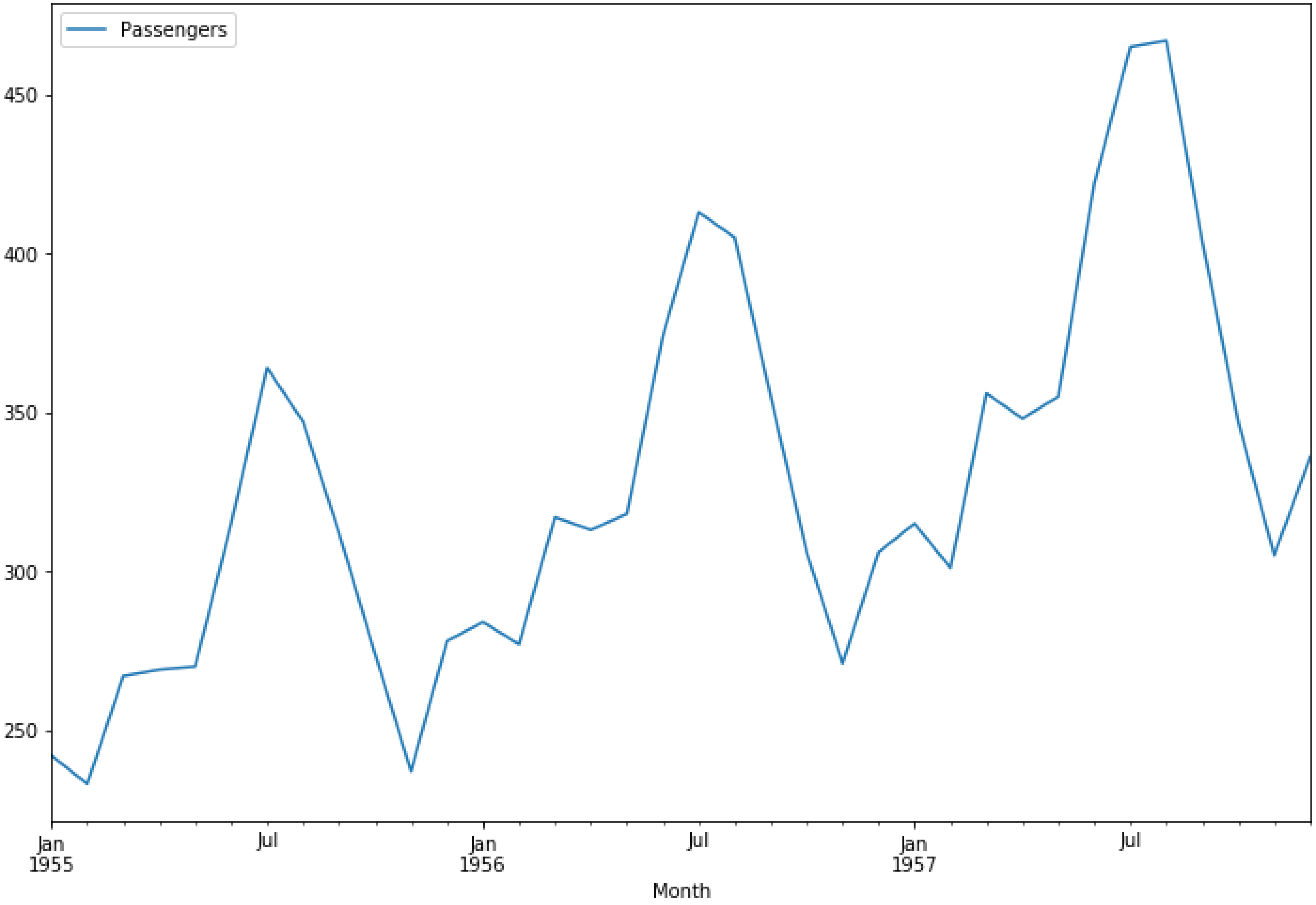
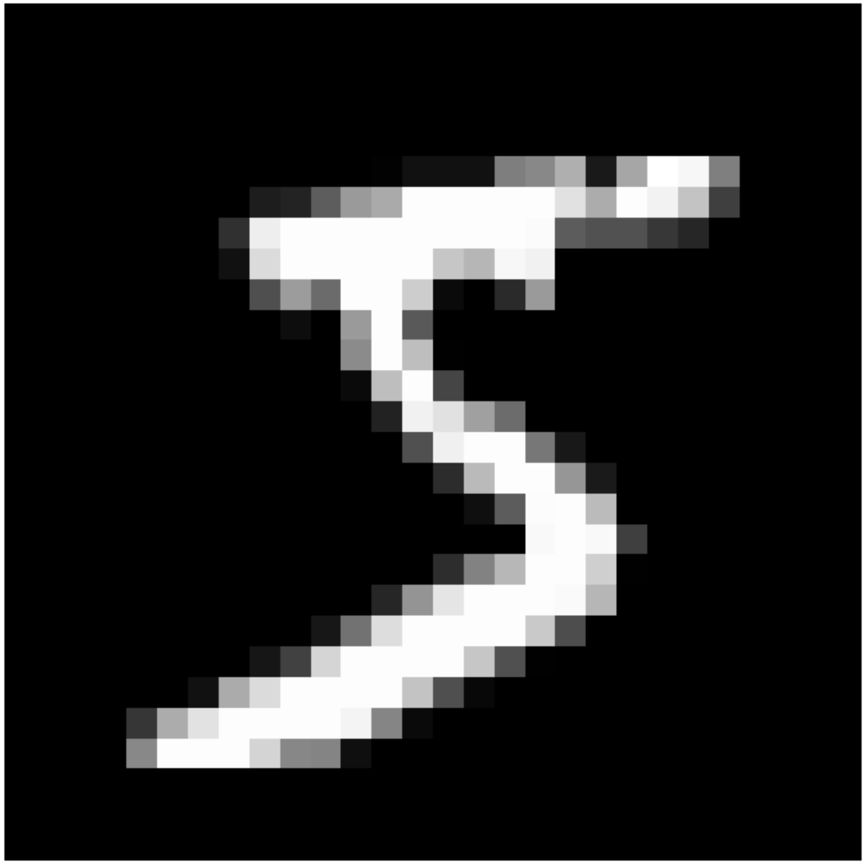


Image Data

- A set of values (pixels) which describes the intensities or the color of the pixels
- The values are arranged in an array of rows and columns that correspond to the vertical and horizontal positions of the pixels
- Has spatial structure – visual information

Image Data

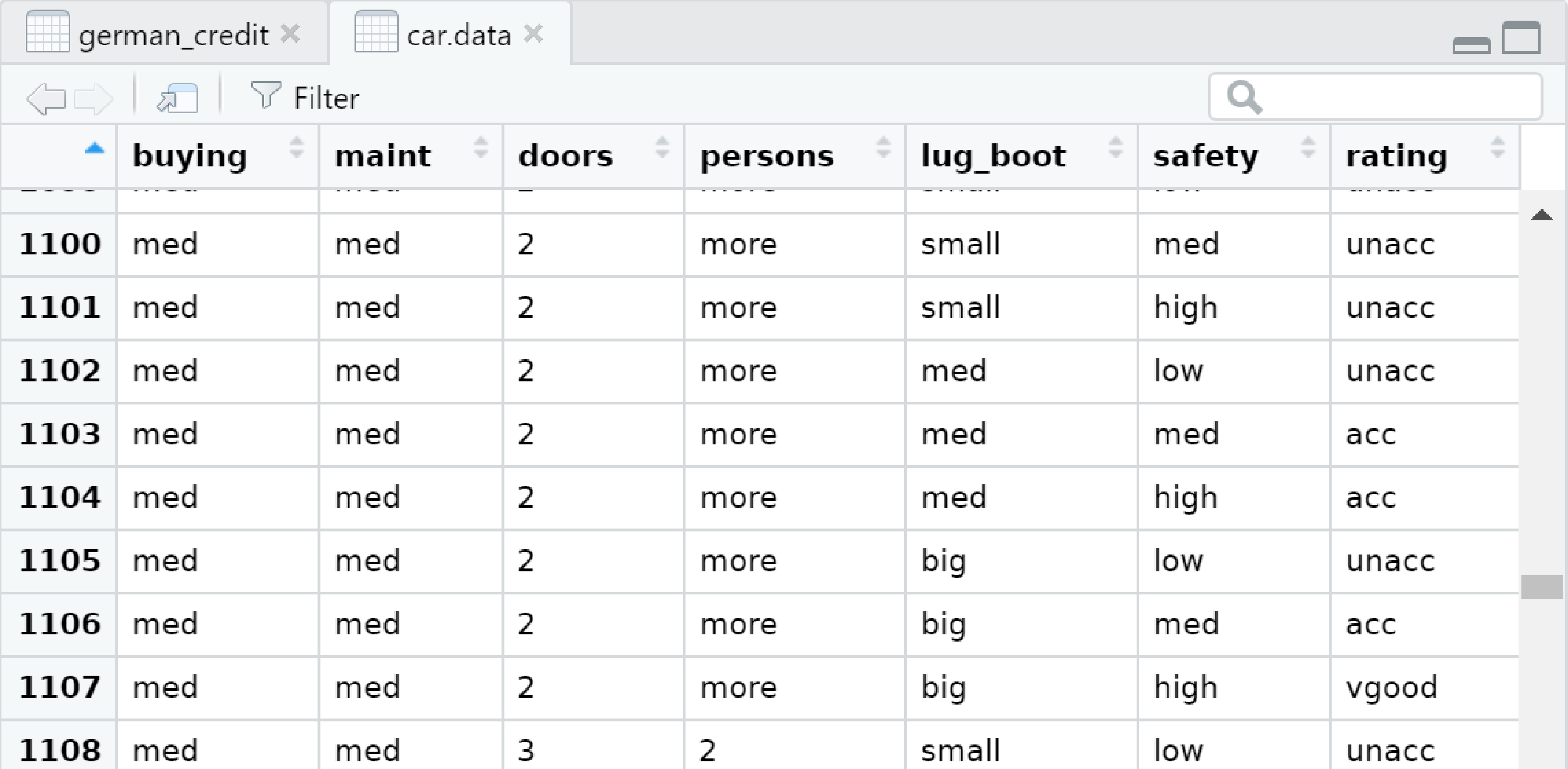


8-bit image
0 – black
255 – white

Data Formats

- Well-structured data
- Less-structured data
- Unstructured data

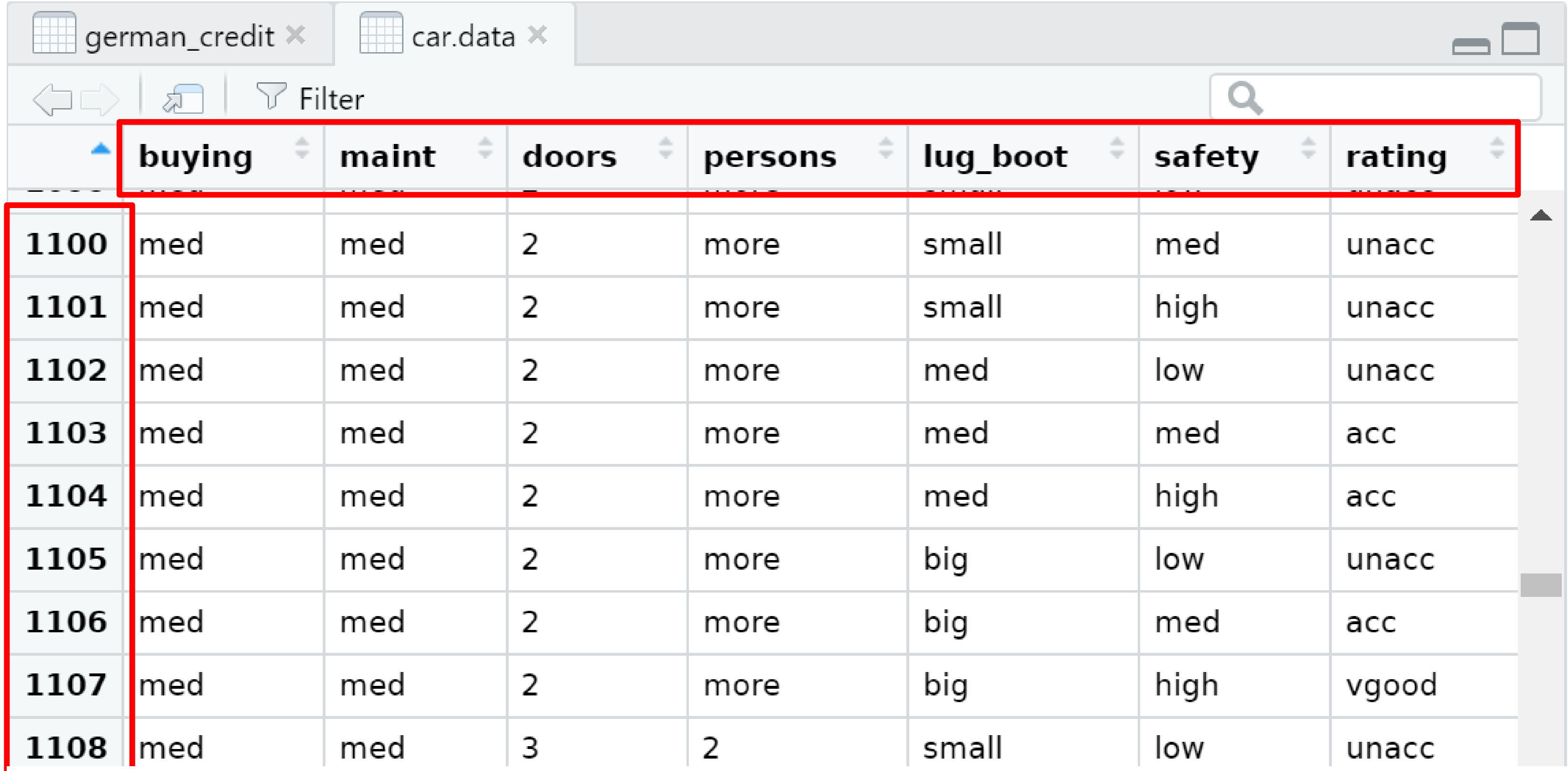
Well-structured Data



	buying	maint	doors	persons	lug_boot	safety	rating
1100	med	med	2	more	small	med	unacc
1101	med	med	2	more	small	high	unacc
1102	med	med	2	more	med	low	unacc
1103	med	med	2	more	med	med	acc
1104	med	med	2	more	med	high	acc
1105	med	med	2	more	big	low	unacc
1106	med	med	2	more	big	med	acc
1107	med	med	2	more	big	high	vgood
1108	med	med	3	2	small	low	unacc

- Table-structured data with headers – numeric or text
- Easy to search and analyze
- E.g. patient information, student information, product, real estate

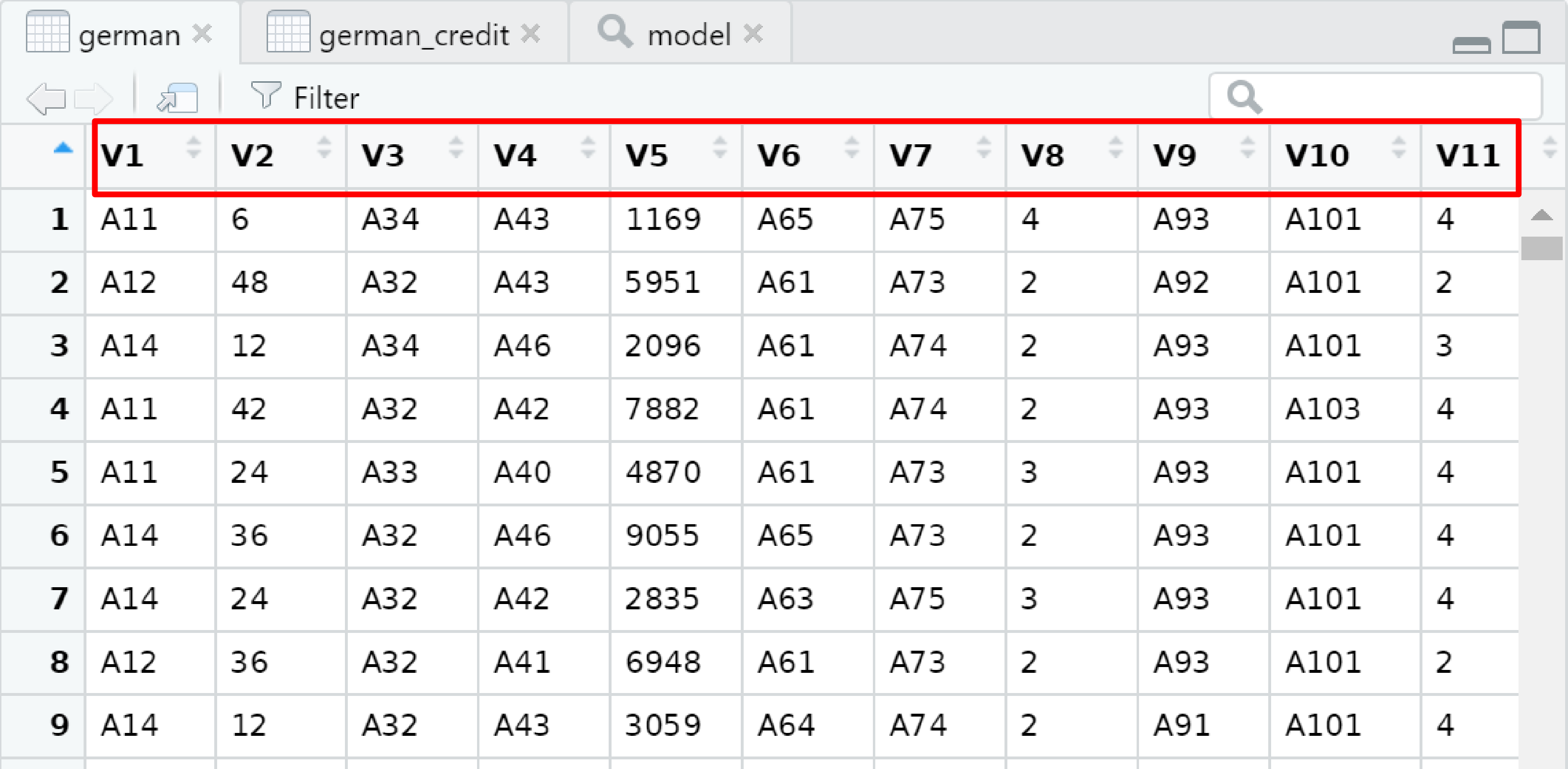
Well-structured Data



	buying	maint	doors	persons	lug_boot	safety	rating
1100	med	med	2	more	small	med	unacc
1101	med	med	2	more	small	high	unacc
1102	med	med	2	more	med	low	unacc
1103	med	med	2	more	med	med	acc
1104	med	med	2	more	med	high	acc
1105	med	med	2	more	big	low	unacc
1106	med	med	2	more	big	med	acc
1107	med	med	2	more	big	high	vgood
1108	med	med	3	2	small	low	unacc

- Rows are instances or datum about which the entity being observed
- Columns are facts or measurements (attributes or features)
- Cells are the values (data)

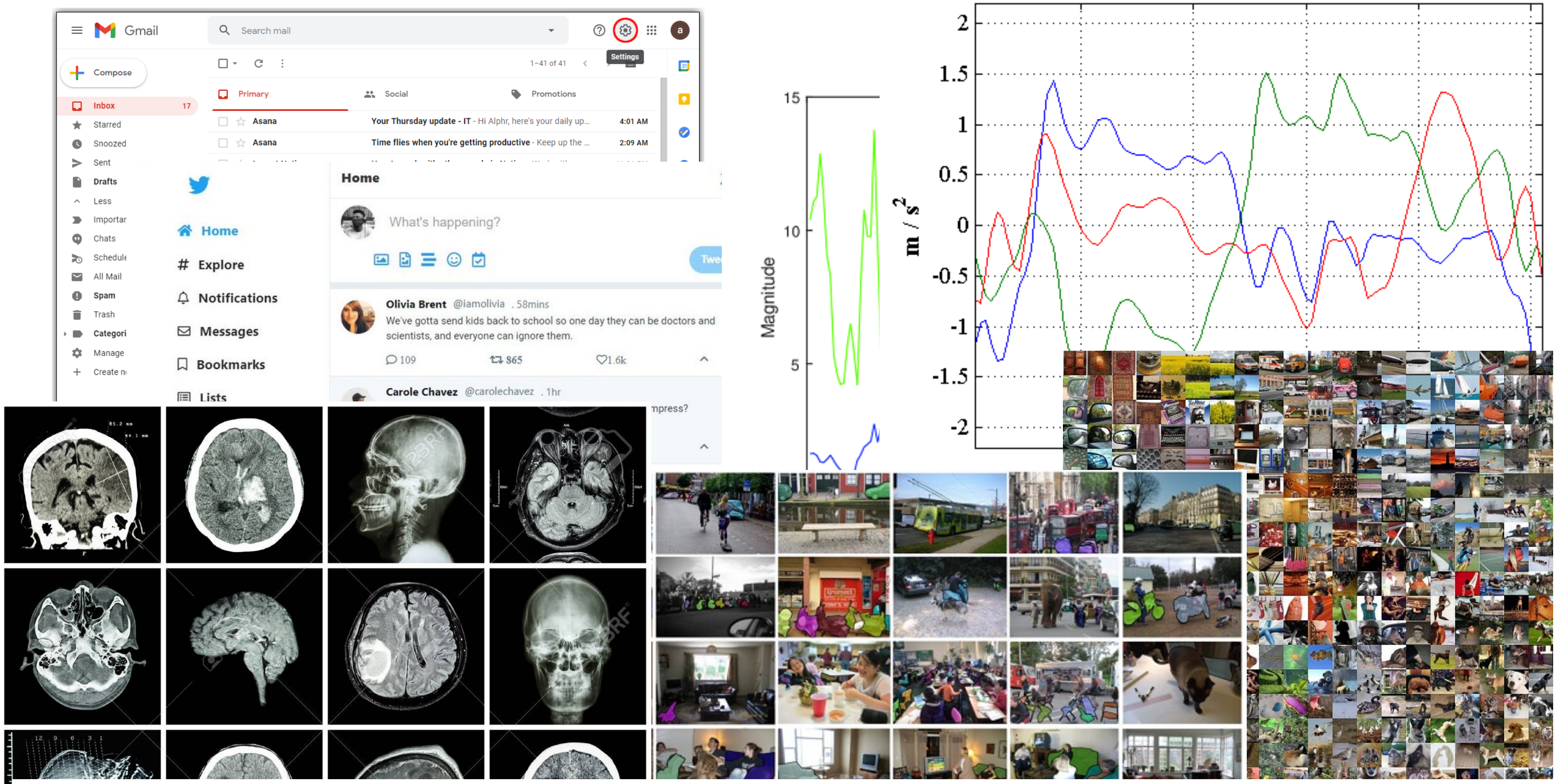
Less-Structured Data



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4
2	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2
3	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3
4	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4
5	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4
6	A14	36	A32	A46	9055	A65	A73	2	A93	A101	4
7	A14	24	A32	A42	2835	A63	A75	3	A93	A101	4
8	A12	36	A32	A41	6948	A61	A73	2	A93	A101	2
9	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4

- Table-structured data without headers or with ambiguous headers
- Not as easy to analyze
- Data is encoded value, needs to decode using the documentation

Unstructured Data



Gmail

Search mail

Settings

Compose

1–41 of 41

Primary | Social | Promotions

Inbox 17

Starred

Snoozed

Sent

Drafts

Less

Importar

Chats

Schedule

All Mail

Spam

Trash

Categori

Manage

Create n

Asana	**Your Thursday update - IT** - Hi Alphr, here's your daily up...	4:01 AM
Asana	**Time flies when you're getting productive** - Keep up the ...	2:09 AM

Home

Explore

Notifications

Messages

Bookmarks

Lists

Home

What's happening?

Twe

Olivia Brent @iamolivia . 58mins

We've gotta send kids back to school so one day they can be doctors and scientists, and everyone can ignore them.

109 865 1.6k

Carole Chavez @carolechavez . 1hr

mpress?

- Text, time series data, images
- Difficult to search and analyze
- E.g. social media, product review/rating, email, survey

Structured Data

Characteristics

- Numeric and text
- Easy to search and analyze

Resides in

- csv file
- Database
- Data warehouses

Examples

- Patient information, student information, product sales

Unstructured Data

Characteristics

- Text, time series data, images
- Difficult to search and analyze

Resides in

- Applications
- Data warehouses

Examples

- Social media, e-mails, documents, measurements

End