Transforming the Data

Transforming Quantitative Attributes

- Quantitative Attributes – Numerical, which allows us to perform mathematical computations on them
- Categorical Attributes – Often non-numerical and represent group or category memberships
- Sometimes we may want to transform a quantitative attribute to a categorical attribute
  - Numerical age to age group (child, senior) for discounts
  - Numerical weight to a weight class (flyweight, featherweight, welterweight, ...)
- Keep in mind that we lose data when transforming from a quantitative attribute to a categorical attribute

Binning

- Binning – Grouping numeric values within designated sub-ranges into "bins"
- Binning can make numeric values into categorical values
- Earlier examples (age into age groups and weight into weight classes) have predefined notions of what the groups / bins should be
- If we are unsure of what the groups should be, we can perform the following binning techniques:
  - Equal Width – The width (range) of each bin is equal
    - Bins are defined as:

$$[\min, \min + w), [\min + w, \min + 2w), \ldots, [\min + (n-1)w, \min + nw)$$

      Where, $w = (\max - \min) / (\#\text{ of bins})$
  - Equal frequency – The number of instances in each bin is equal
- We want to bin the following values into 3 groups

$$\text{Age} = [210, 15, 11, 13, 10, 35, 50, 55, 92, 72, 204, 5, 215]$$

- We need to sort this first:

$$\text{Age} = [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 210, 215]$$

- Equal Width:
  - $w = (215 - 5)/3 = 70$
  - Bin ranges $= [5, 75), [75, 145), [145, \infty)$

  - So the bins are:

$$[5, 10, 11, 13, 14, 35, 50, 55, 72]$$

$$[92]$$

$$[204, 210, 215]$$

- Equal Frequency:
  - Size of bins $= \lfloor \frac{13}{4} \rfloor = 4$
  - So the bins are:

$$[5, 10, 11, 13]$$

$$[15, 35, 50, 55]$$

$$[72, 92, 204, 210, 215]$$

  - We put the first extra in the last bin, and then the next extra in the bin before that

Exercise

- Split $[10, 20, 35, 40, 44, 60, 65, 66, 70]$ into 4 bins
- Equal Width
  - $w = (70 - 10)/4 = 15$
  - Bin Ranges $= [10, 25), [25, 40), [40, 55), [55, \infty)$
  - Equal Width Bins

$$[10, 20]$$

$$[35[$$

$$[40, 44]$$

$$[60, 65, 66, 70]$$

- Equal Frequency
  - Number per bins $\lfloor \frac{9}{4} \rfloor = 2$
  - Equal Frequency Bins

$$[10, 20]$$

$$[35, 40]$$

$$[44, 60]$$

$$[65, 66, 70]$$

Transforming Quantitative Attributes

- When visualizing data, it can be difficult to read values that have many digits such as reported income ($20,000 – $900,000)
- We scale the values to decrease the magnitude and number of digits
- Dividing each value by 1000, we get a range from (20 – 900) in thousands

- This transformation can then be used in our visualization
- Similarly, we can show 95% instead of 0.95
- We can apply other more complicated functions that transform from one quantitative attribute into another quantitative attribute
- Combining two more quantitative attributes can give us more insight from the data
  - GDP per capita (per person) $= \frac{\text{GDP}}{\text{population}}$
  - Force = Mass $\times$ Acceleration
  - Mean number of people per household
- Data transformation usually occurs after data cleaning, so we expect the data to be in "good" condition
- What type of data transformation that we apply can vary on what insights we are looking for and what our interests are
- Many of the analysis methods that we will discuss require data to be numerical
  - Similar with visualization and summary techniques
- We may want to convert text into numbers

ID	age	pet_preference	voted
0	23	Cat	Yes
1	43	Dog	Yes
2	23	None	No
3	43	Cat	Yes

↓

ID	age	prefer_cat	prefer_dog	prefer_none	voted
0	23	1	0	0	Yes
1	43	0	1	0	Yes
2	23	0	0	1	No
3	43	1	0	0	Yes

- When calculating the percent of people that prefer cats, we can simply sum over the column and divide by the number of people
- We can also represent each person by a vector of numerical values, which can be quite useful. We would replace "yes" with 1 and "no" with 0
- Encoding – The process of converting a categorical attribute into a quantitative one by assigning codes to the values of the categorical attribute
- Boolean Encoding – The process of converting a categorical attribute into a quantitative one by assigning a 1 or 0 as values

Encoding

- We can select the values that are used to transform into numeric values, but we need to be careful
- Observe the different encodings below related to the sentiment of reviews

Response	Encoding Scheme 1	Encoding Scheme 2	Encoding Scheme 3	Encoding Scheme 4
Dislike	1	-1	0	0
Neutral	2	0	0	1
Like	3	1	1	1000

  1. Encoding 1 – Often the first type considered (enumerate from 1 onward)
  2. Encoding 2 – More focused around considering the meaning of the data (sentiment)
  3. Encoding 3 – Boolean, where dislike and neutral are treated as the same
  4. Encoding 4 – Atypical, largely emphasizes the like response.

- Each encoding will change the visualization of the data and the analysis, so selecting the best one will depend on the context

Condensing the Categories

- Let's pretend that we used a survey that is asking about favorite pets and the response is a type where you write your answer
- You get the survey results back and want to start counting favorite pets
- You frequently see a few pets, but for other pets, you only see them a few times
  - How do you think that would affect visualization?
- We can condense all the infrequent answers (pets) into one category (other)
- Condensing the categories – the process of reducing the number of distinct values in a categorical attribute by merging categories into broader ones
  - In doing so, we lose some information
  - We would typically create a new attribute for our data but retain the original values for other analysis

Splitting an Attribute into Multiple Attributes

- Sometimes there is different information captured within a single attribute that we want to parse
- Example: Date: September 25th, 2025
  - We could create 2 new attributes (month and year)
  - Are there additional attributes that we can create from this date?
- Our visualization and analysis might benefit from knowing the month or year

Reshaping a Dataset

- Reshaping a dataset – Involves transforming its structure from a wide format, with extensive columns, to a long format, with fewer columns but more rows, or vice versa.
- Tidy dataset has each column representing an attribute and each row representing an observation / instance
- The table below does not meeting this criteria of a tidy dataset

Region	2019	2019	2020	2021	2022	2023
AU / NZ	29 789	30 252	30 682	30 935	31 193	31 472
Sub-S AF	1 054 046	1 081 978	1 110 090	1 138 953	1 167 589	1 196 738
E & North A.	1 118 890	1 123 194	1 127 212	1 127 305	1 127 406	1 127 487

- We can reshape the data into the table below

Region	Year	Population
Australia / New Zealand	2018	29 789
Australia / New Zealand	2019	30 252
Australia / New Zealand	2020	30 692
Australia / New Zealand	2021	30 935
Australia / New Zealand	2022	31 193
Australia / New Zealand	2023	31 472
Sub-Saharan Africa	2018	1 054 046

- The wide format is easier for humans, and the tidy, long format can be easier for some computers for analysis / visualization