Integrating the Data

Combining Datasets

- More data is often better
- Data that we might be interested in, can come from multiple sources
- We might want to combine data
- 3 different methods to combine datasets
    1. Stacking
    2. Augmenting
    3. Merging
- Stacking
    - If two datasets hae different observations / instances, but the same attributes, we can simply stack them on top of each other
        * Need to be careful about duplicated instances and remove them if needed

Dataset 1:

person_id	height	weight	team_id
1630163	6' 7"	180	1610612766
1133	6' 3"	200	1610612766
76148	6' 0"	175	1610612766
1630547	6' 4"	190	1610612766

Dataset 2:

person_id	height	weight	team_id
1077	6' 9"	224	1610612744
1108	6' 9"	244	1610612744
116	6' 9"	230	1610612744
101145	6' 3"	185	1610612744

The stacked dataset is:

person_id	height	weight	team_id
1630163	6' 7"	180	1610612766
1133	6' 3"	200	1610612766
76148	6' 0"	175	1610612766
1630547	6' 4"	190	1610612766
1077	6' 9"	224	1610612744
1108	6' 9"	244	1610612744
116	6' 9"	230	1610612744
101145	6' 3"	185	1610612744

- Augmenting

  - If two datasets have the same observations / instances but different attributes, we can combine them horizontally

person_id	height	weight	team_id
1630163	6' 7"	180	1610612766
1133	6' 3"	200	1610612766
76148	6' 0"	175	1610612766
1630547	6' 4"	190	1610612766

- Merging
  - Merging is combining datasets beyond simply stacking and augmenting
  - There are a few ways this can occur
  - We may have two datasets with different amounts of instances and attributes, but have at least one attribute in common that allows us to combine the two datasets

Dataset 1:

student_id	club_name
S1	Programming Club
S1	Drama Club
S2	Programming Club
S2	Swimming Club
S3	Math Club

Dataset 2:

student_id	major	GPA
S1	CS	3.9
S2	CS	3.6
S3	Math	3.7

Merged Dataset:

student_id	club_name	major	GPA
S1	Programming Club	CS	3.9
S1	Drama Club	CS	3.9
S2	Programming Club	CS	3.6
S2	Swimming Club	CS	3.6
S3	Math Club	Math	3.7

Augmenting a Dataset is **not** Data Augmentation

- Data augmentation involves creating new synthetic data to improve performance of machine learning models
- We will discuss data augmentation at a later time

Getting More Information from Combining Datasets

- A step back on why we care about datasets

- It allows us to gather more information
- Gives the ability to extend our analysis
- Example:
  - One dataset contains the list of food that people consume throughout a week
  - One dataset contains the nutritional information related to individual foods

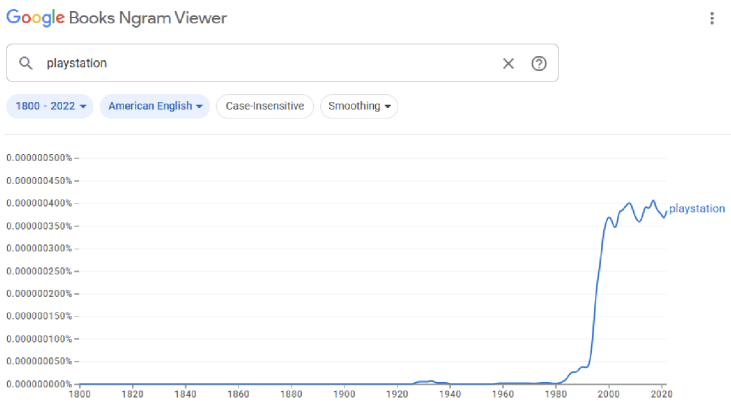
Ethics – Othering

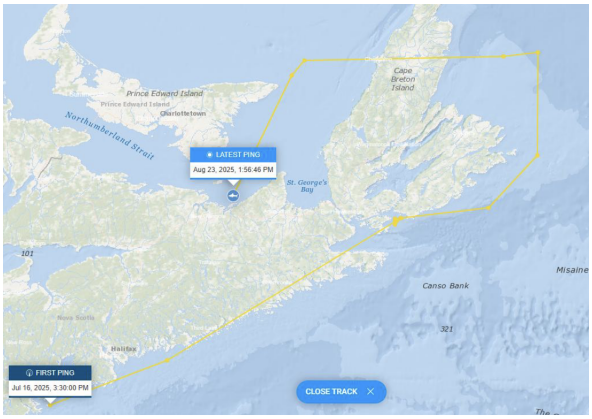
- We discussed an example of condensing values of attributes into a single category (Other) to assist with visualization and analysis
- The process of grouping responses into an unknown or other category is a version of othering
  - Can be seen as, and gives the idea that some individuals or groups are defined and labeled as not fitting within the perceived norms of a social group
- Othering should be avoided when possible and if used, it should be well documented
- For example, labeling political affiliation as Conservative, Liberal, and other is an oversimplification of the political views in Canada
- This can occur with nationality and language of individuals as well
- Othering does not provide proper representation of all those included

Data Visualization

- Data visualization – Graphical representation of information and data
- Use of charts, graphs, and maps to identify the stories in data
- Can assist us in finding patterns and trends in the data
- Can be used to deliver a story in a more impactful manner
  - The story that you want to tell will determine what visualization techniques you will use
- We will be exploring visualization techniques, latest trends in data visualization, tools for creating data visualizations, and the danger of visual misrepresentation
- Typically use the term variable instead of attribute when discussing the plotting of data

Where Data Visualization Can Help Us





Which Graphics to Use

- Selecting the correct visualization technique depends on the amount, type, and complexity of the data

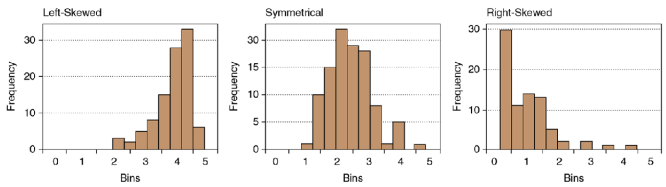
- A few effective graphics for depicting data are:
  - Histograms
  - Column Charts
  - Box Plots
  - Scatterplots
  - Line Charts
  - Density Plots
  - Violin Plots

Know Your Data

- The first step to creating an effective visualization is to know the data
  - Understand which variables / attributes are quantitative and which are categorical
  - Understand what each variable / attribute represents
- Effective data visualizations have:
  - Descriptive titles
  - Labeled axes
  - Appropriately labeled units for the variables / attributes

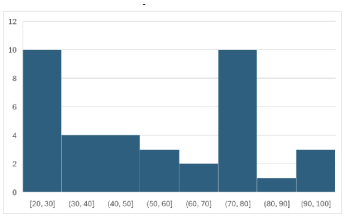
Visualization with One Quantitative Variable – Histograms

- Continuous variable that has been binned
- Evaluate the symmetry / asymmetry
- Peaks and tails
- Skewness – The extent to which the data points are spread unevenly
- A normal histogram is bell shaped, with a peak in the middle. The mean and median are roughly the same

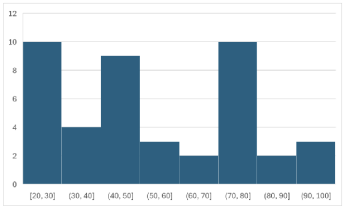


Histograms

- Bimodal – Histogram / distribution is bimodal if it has two peaks that stand out
- Multimodal – Histogram / distribution is multimodal if it has more than two peaks that stand out

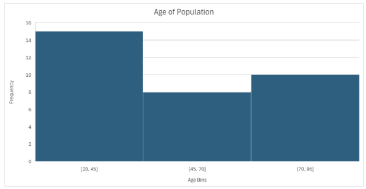


Bimodal

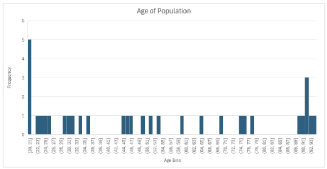


Multimodal

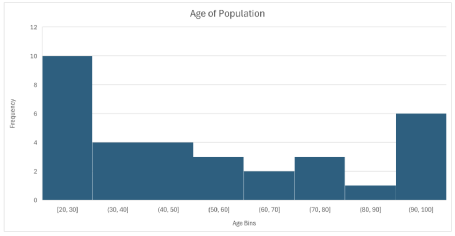
How Many Bins?



Too few



Too many



Reasonable number of bins

Notes on Histograms

- Can use proportion along the y-axis instead of raw counts
- Histograms can help us understand the distribution of data
- Histograms can assist with identifying patterns

Code for Histograms

- We will be using Plotly for data visualization

```
import plotly.express as px
fig = px.histogram(x = data_list, title = "Max Windspeeds",
    nbins = 30)
fig.show()
```