Sampling

Sampling Techniques

- Simple random sampling
  - Randomly select items from a population to be in the sample
  - 100 bats, 120 cats, 30 dogs, 40 elephants, 0 mice, ...
- Stratified sampling
  - Organize each item into groups first, then randomly select from each group
  - Ensures that there is some representation of each group
  - You can sample based on existing distribution from each group or have equal representation from each group. For example, if 20% of all the animals you are concerned about are bats, then you can sample from the bats' group 20% of the time

How would you sample?

- Population – All professional athletes
- Goal – Understand the average strength of professional athletes

1. What would be a bad sampling technique and what could be the outcome?
   - Choosing athletes from 1 sport. For example, choosing 100 powerlifters would probably give you a higher average strength than what is expected.
2. What would be a good sampling technique?
   - Choosing 10 athletes from each course.

Case Study – Selection Bias and Politics (Dewey defeats Truman)

- Selection bias – Bias introduced into the data based on the selection method
- In 1948, for the US presidential election with Thomas E. Dewey and Harry S. Truman pollsters relied on a telephone survey to predict who would win the election
- Based on the telephone survey, it was clear that Dewey was going to win and early printing of papers stated that "Dewey Defeats Truman"
- Owning telephones in 1948 was considered a luxury and more likely a wealthy household, who would more likely vote for Dewey.

Data Preparation

Data preparation has 3 components

1. Data collection
2. Data wrangling / preprocessing
3. Data management

Data Collection

- Data collection – Systematic gathering of relevant data from various sources
- Sources can include: websites, databases, surveys, logs, etc.
- We will want to avoid low-quality or irrelevant data

Data Collection – Wildlife Conservation

- Let's say that we wanted to study the movement patterns and behaviors of birds in a specific area
- To collect data we can:
  - Tag a sample of birds with a small tracking device to record GPS location
  - Install motion-sensor cameras throughout the region of interest
  - Make visual observations from viewing blinds and documenting behaviour
  - Record audio of birds in the area
  - ...
- The methods we use to collect data can vary with what we want to study / explore

Data Collection – Text

- Let's say that we want to study how people in the Maritimes use language. How would you collect data?
  - Web scraper – Use a program to automatically gather text from blogs and social media of people living in the Maritimes
  - Survey – Ask people to complete a survey and link their social media or provide text examples
  - Audio Recording – Record people speaking

Data Wrangling / Preprocessing

- Data wrangling / preprocessing – Process of converting raw data into forms necessary for analysis and storytelling
- Raw data – The original form of data when collected that has not been preprocessed for use in analysis, storytelling, or other purposes
- Wrangling data involves:

  - Cleaning data
  - Transforming data
  - Integrating data

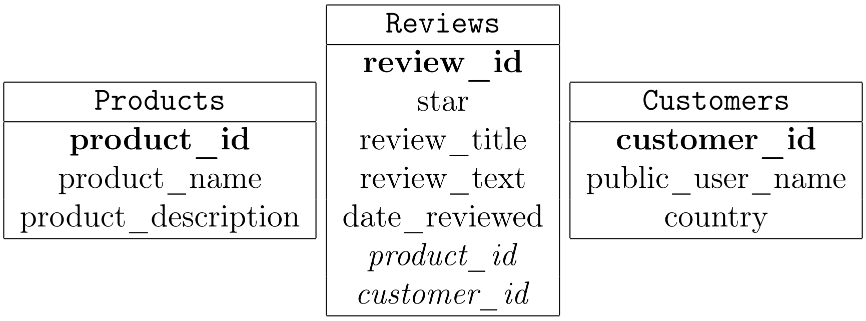
- Data wrangling is used to make data useful for analysis and storytelling

Data Wrangling – Cleaning

- Cleaning Data – Fixing or removing incomplete, abnormal, or inconsistent data to enhance quality for accurate analysis or storytelling
- Different than data tidying
  - Data cleaning improves data quality, while data tidying optimizes the structure
- The methods used for cleaning depends on the situation with the data and we will discuss specifics at a later time

Data Wrangling – Transforming

- Transforming data – Organizing and formatting data into a structure that is useful for efficient analysis
- Sometimes we go from unstructured to structured, but there can also be a reason to go from structured to a different structured format
- Example – We have thousands of amazon reviews on a few different products. To assist us in analyzing this data, we can transform the data into structured tables that are connected to each other through some selected attribute / variable



Products
product_id
product_name
product_description

Reviews
review_id
star
review_title
review_text
date_reviewed
product_id
customer_id

Customers
customer_id
public_user_name
country

Data Wrangling – Integrating

- Integrating Data – Merging multiple data sources to provide comprehensive insights that surpass what each individual source could offer alone
- After the collection and cleaning of the data, we can go into analysis, but we could potentially get more insights if we integrated data from two different datasets
- A company could integrate a public dataset with their own proprietary dataset

Data Management

- Data Management – Storing data and facilitating its access

- Often stored as tables in a database, although there are other options
- Other considerations involving data management includes
  - Security
  - Accessibility
  - Redundancy (RAID)
  - ...

Data Analysis

- Data Analysis – Process of examining datasets to uncover useful information, which can inform conclusions and support decision-making
- Involves applying statistical, logical, and mathematical techniques to sift through data, and identify patterns, trends, and relationships.
- Four main categories:
  - Descriptive Analysis – Systematically examines current and historical data, utilizing various benchmarks, to provide a detailed view of past and present outcomes
  - Diagnostic Analysis – Investigates data to determine the possible causes and reasons behind past events
  - Predictive Analysis – Uses statistical methods to predict future events based on historical data
  - Prescriptive Analysis – Provides recommendations on actions to take

Analyzing the fidget spinner

1. Descriptive Analysis (What happened)
   - We found that the average stars for our fidget is 4.7 compared to the far inferior fidget spinner which received an average rating of 2.4.
2. Diagnostic Analysis (Why it happened)
   - We group reviews based on the geolocation of the reviewer and found that many of the reviewers that gave a high rating to both fidget spinners are from the same region and our fidget spinner received many more reviews from people in that region.
3. Predictive Analysis (What is likely to happen)
   - We can predict what the ratings will be for our fidget spinner in the next month based on previous data. Perhaps the ratings always go up in October each year.
4. Prescriptive Analysis (What actions should be taken)
   - Marketers can use our provided analysis to focus their efforts. Promote our fidget spinner to the people in the region that gave higher reviews.