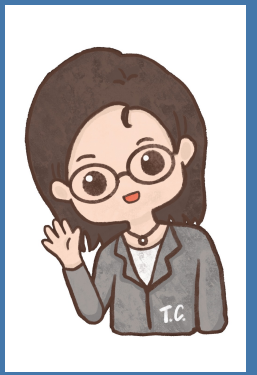
MTH 208 Exploratory Data Analysis

Lesson 02: Types of Data & Data Structures

Ying-Ju Tessa Chen, PhD

Associate Professor
Department of Mathematics
University of Dayton

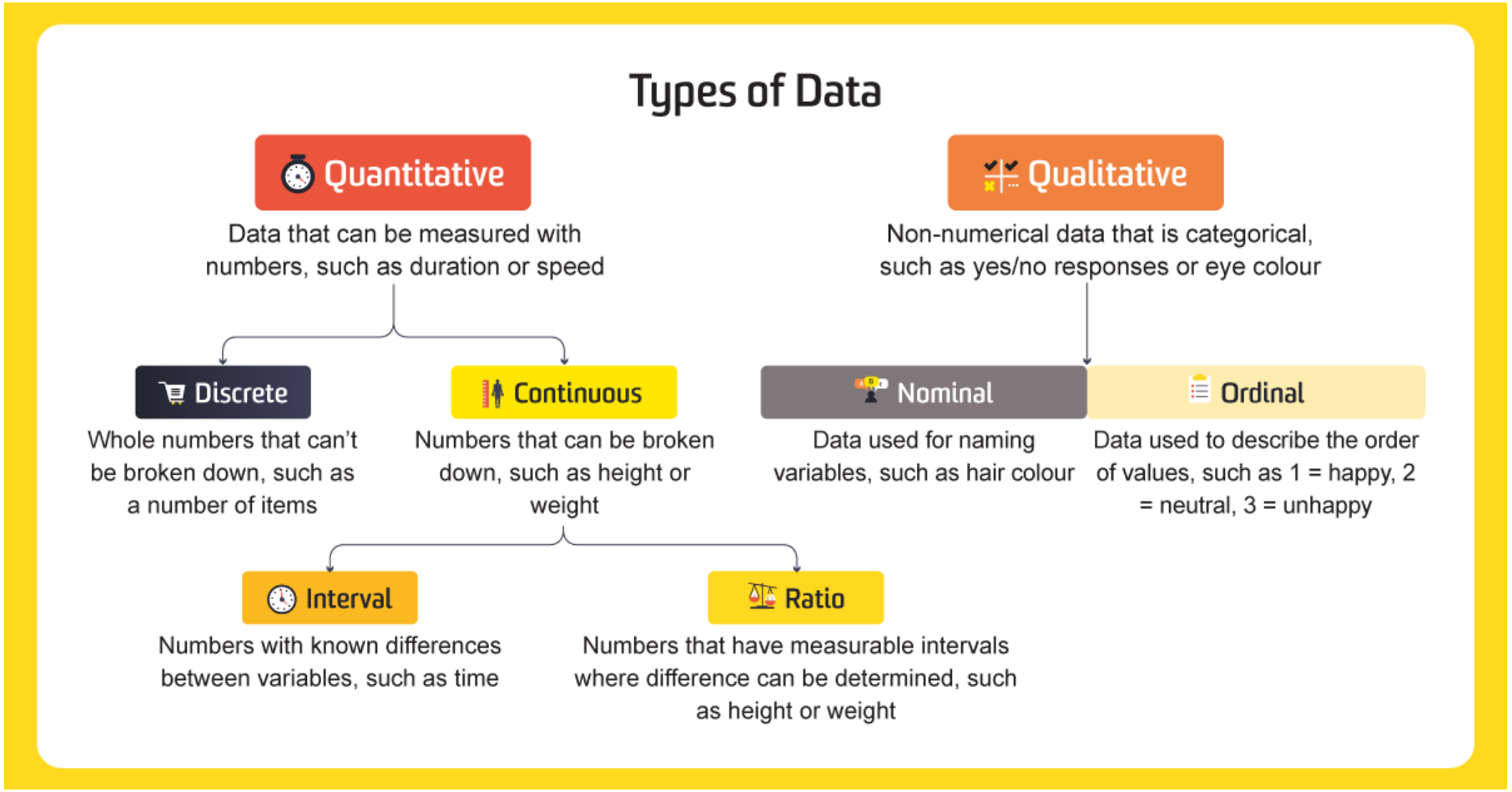
@ying-ju
ying-ju
ychen4@udayton.edu



Learning Objectives

- Introduction to Data Types
- Data Structures
- Data Collection Methods
- Data Quality
    - Data Preprocessing

Introduction to Data Types



Credit: University of New South Wales Sydney, Types of data & the scales of measurement, accessed January 7, 2024.

Introduction to Data Types (Continued)

Understanding different data types is not just a theoretical exercise; it has practical implications in every step of data analysis, from preprocessing to model building and interpretation. This foundational knowledge enables data analysts and scientists to approach data with the right tools and mindset, leading to more effective and accurate analysis.

- **Appropriate Analysis Methods**: Different data types require different analysis techniques and statistical methods. For instance, the methods used to analyze numerical data (like regression analysis) differ significantly from those used for categorical data (like chi-square tests).
- **Informed Decision-Making**: Understanding data types helps in making informed decisions about the right tools and approaches for data analysis. For example, certain data visualization techniques are more effective with numerical data (like scatter plots) than with categorical data (like bar charts).
- **Data Preprocessing**: Effective data preprocessing – such as cleaning, transforming, and encoding – relies on knowing the data types. For instance, handling missing values in numerical data might involve different strategies than in categorical data.

Introduction to Data Types (Continued)

- **Model Selection and Performance**: In predictive modeling, the choice of model and its performance is heavily dependent on the nature of the data. Models that work well with continuous numerical data may not be suitable for ordinal or nominal categorical data.
- **Interpretation Accuracy**: Accurate interpretation of analysis results requires an understanding of the data's nature. For example, the meaning of a mean (average) is clear in the context of continuous numerical data, but it’s not applicable to nominal categorical data.
- **Data Quality and Integrity**: Recognizing data types is crucial for maintaining data quality and integrity. It ensures that data is used appropriately and that any conclusions drawn are valid.
- **Compliance and Ethics**: In certain fields, especially those involving sensitive or personal information, knowing data types is essential for compliance with legal and ethical standards, such as how data is stored, processed, and shared.
- **Efficiency in Computing**: Different data types can impact the efficiency of storage and computation. For example, categorical data can often be more efficiently encoded, which can be important in large data sets.

Introduction to Data Types (Continued)

- **Qualitative (Categorical) Data**:
  - `Nominal`: Data without an inherent order (e.g., gender, ethnicity).
  - `Ordinal`: Data with a specific order or ranking (e.g., customer satisfaction ratings).
- **Quantitative (Numerical) Data**:
  - `Discrete`: Countable data (e.g., number of students in a class).
  - `Continuous`: Measurable data, can take any value within a range (e.g., height, temperature).

Note: Numerical data is always ordinal.

In statistics, there are four data measurement scales: nominal, ordinal, interval, and ratio. We will talk about their definitions, give examples, their key features, their key difference, and why it is important to know the difference between the scales.

Data Measurement Scales (Continued)

Nominal Scale

- `Definition`: A nominal scale categorizes data without any order or ranking among the categories. Each value represents a different category, but these categories cannot be logically arranged in any sequence.
- `Examples`:
  - Colors: Red, Blue, Green, etc.
  - Types of Animals: Cat, Dog, Bird, etc.
  - Genders, Ethnicities, Nationalities.
- `Key Features`:
  - Categories are mutually exclusive and collectively exhaustive.
  - There is no inherent order or ranking in the categories. Mathematical operations (like addition or subtraction) are not meaningful.
  - Often used for classification and labeling.

Data Measurement Scales (Continued)

Ordinal Scale

- `Definition`: An ordinal scale also categorizes data, but unlike the nominal scale, there is a clear order or hierarchy among the categories. However, the intervals between the ranks are not necessarily equal.
- `Examples`:
    - Educational levels: High School, Bachelor's, Master's, Ph.D.
    - Satisfaction ratings: Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied.
    - Socioeconomic status: Low, Middle, High.
- `Key Features`:
    - The categories can be ranked or ordered logically.
    - The distances between the categories are not defined or not equal.
    - We can make comparisons of more or less, but not of how much more or less.
    - Suitable for non-quantitative traits where ranking is possible.

Data Measurement Scales (Continued)

Key Differences between Nominal and Ordinal Scales

- `Ordering`: Nominal scales do not imply any order, while ordinal scales do.
- `Type of Analysis`:
  - Nominal data is typically analyzed by counting frequencies or using modes.
  - Ordinal data analysis can involve median or percentiles, along with modes.
  - Statistical Tests: Different statistical tests are used for nominal (like Chi-square test) and ordinal data (like Spearman's rank correlation).

Why Understanding the Difference is Important?

- `Appropriate Analysis`: Knowing whether our data is nominal or ordinal helps in choosing the right statistical methods. For instance, ordinal data allows for median calculation, but nominal data does not.
- `Accurate Interpretation`: Understanding the scale ensures that data is interpreted correctly. For example, knowing that a survey's response scale is ordinal prevents the mistaken application of calculations meant for interval/ratio data.
- `Research Design`: In designing surveys or experiments, distinguishing between nominal and ordinal scales can influence how questions are phrased and how responses are gathered.

Data Measurement Scales (Continued)

Interval Scale

- `Definition`: An interval scale is a numeric scale where the distance (or interval) between each value is equal, but there is no true zero point. This means you can measure the difference between values, but you cannot make meaningful ratios.
- `Examples`:
  - Temperature in Celsius or Fahrenheit: The difference between 10°C and 20°C is the same as between 20°C and 30°C. However, you cannot say that 20°C is "twice as hot" as 10°C because 0°C does not represent an absence of temperature.
  - Calendar years: The interval between years is consistent (1 year), but there is no absolute zero year.
- `Key Features`:
  - We can add or subtract values meaningfully.
  - Ratios are not meaningful because of the lack of a true zero.
  - It allows for negative values.

Data Measurement Scales (Continued)

Ratio Scale

- `Definition`: A ratio scale is similar to an interval scale, with the key difference being that it has a meaningful or absolute zero point. This allows for the comparison of ratios or proportions.
- `Examples`:
  - Weight: A weight of 0 indicates no weight, and a weight of 10 kg is twice as much as 5 kg.
  - Length or Distance: Measured in units like meters or miles, where 0 represents no length, and you can meaningfully say that 10 meters is twice as long as 5 meters.
- `Key Features`:
  - All mathematical operations (addition, subtraction, multiplication, division) are valid.
  - Ratios are meaningful - we can say that one value is twice as much or half as much as another.
  - There is an absolute zero point that indicates the absence of the quantity.

Data Measurement Scales (Continued)

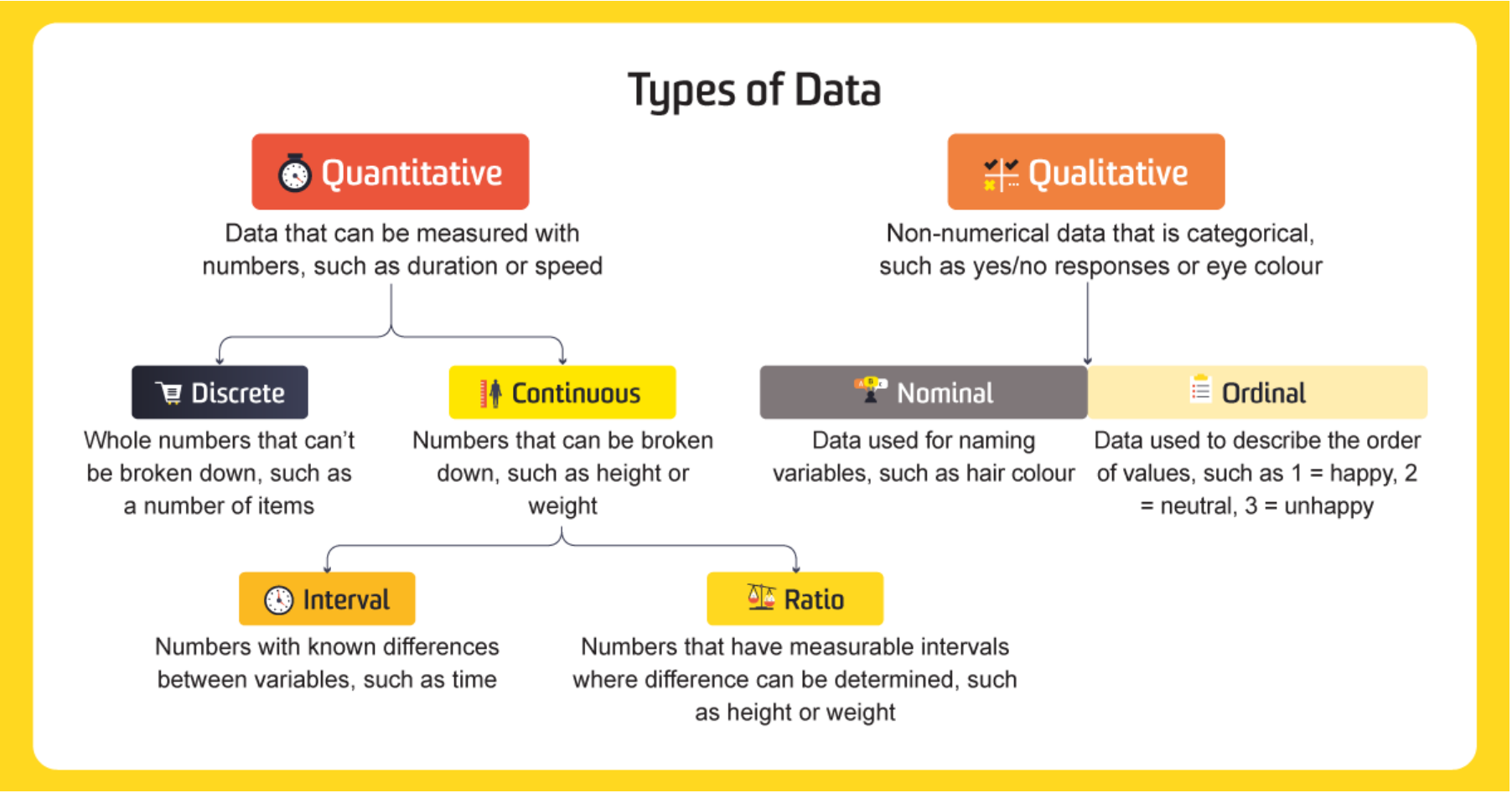
Key Differences between Interval Scale and Ratio Scale

- `Presence of a True Zero`: The most significant difference is the presence of a true zero in ratio scales, allowing for meaningful comparisons of ratios.
- `Negative Values`: Interval scales can have negative values (like temperatures below 0°C), but ratio scales cannot, as they are bounded by zero.
- `Types of Analysis`: Ratio scales allow for a wider range of statistical analyses since the data can be meaningfully multiplied or divided.

Why Understanding the Difference is Important?

- The type of scale (interval or ratio) dictates what statistical techniques are appropriate. For example, geometric mean and coefficient of variation can be used with ratio data but not with interval data.
- Misinterpreting the scale type can lead to incorrect conclusions. For instance, calculating ratios with interval data, where a true zero point doesn't exist, can be misleading.

Introduction to Data Types



Credit: University of New South Wales Sydney, Types of data & the scales of measurement, accessed January 7, 2024.

Data Structures

- **Vectors** and **Scalars**: Basics of vectors (one-dimensional arrays) and scalars (single values).
- **Matrices** and **Arrays**: Two-dimensional (matrices) and - multi-dimensional (arrays) data structures.
- **Data Frames** and **Tables**: Tabular data structures, commonly used in statistical software and programming languages like R and Python.
- **Time-Series Data**: Data collected at or ordered by time intervals (e.g., stock prices over time).
- **Cross-Sectional Data**: Data collected at a single point in time.
- **Panel Data (Longitudinal Data)**: Combines cross-sectional and time-series data, observing multiple subjects over time.

Vectors and Scalars

- **Vectors**: one-dimensional arrays that store elements of the same type (e.g., all numbers or all strings).

Example:

$$\text{My\_heartrate} = [64, 74, 79, 91, 89]$$

$$\text{Instructor} = [\text{"Tessa Chen"}, \text{"Thilini Jayasinghe"}, \text{"Gayan Warahena Liyanage"}, \text{"Matthew Wascher"}]$$

- **Scalars**: single values, which can be numbers, strings, or any other data type.

Example:

$$\text{My\_GPA} = 4.0$$

Matrices and Arrays

- **Matrices**: two-dimensional data structures with rows and columns, ideal for numerical computations and representing data tables.
- **Arrays**: multi-dimensional data structures that extend beyond two dimensions, useful in more complex data analysis scenarios involving multi-dimensional datasets.

Example:

```
##      [,1] [,2] [,3] [,4]
## [1,]   20   25    6   12
## [2,]    9   13   24    5
## [3,]    7   22   21   16
```

Example:

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,]   19    1    3
## [2,]   23   13   15
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,]   18   20   25
## [2,]    2   11   17
```

Data Frames and Tables

The terms "data frames" and "tables" are often interchangeably in the context of data analysis, but they can have distinct meanings depending on the context and the programming environment.

- **Data Frames**: key tabular data structures in R and Python (Pandas), which can hold columns of different types (e.g., numeric, character, factor).

Example (Data Frame):

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Example (Table):

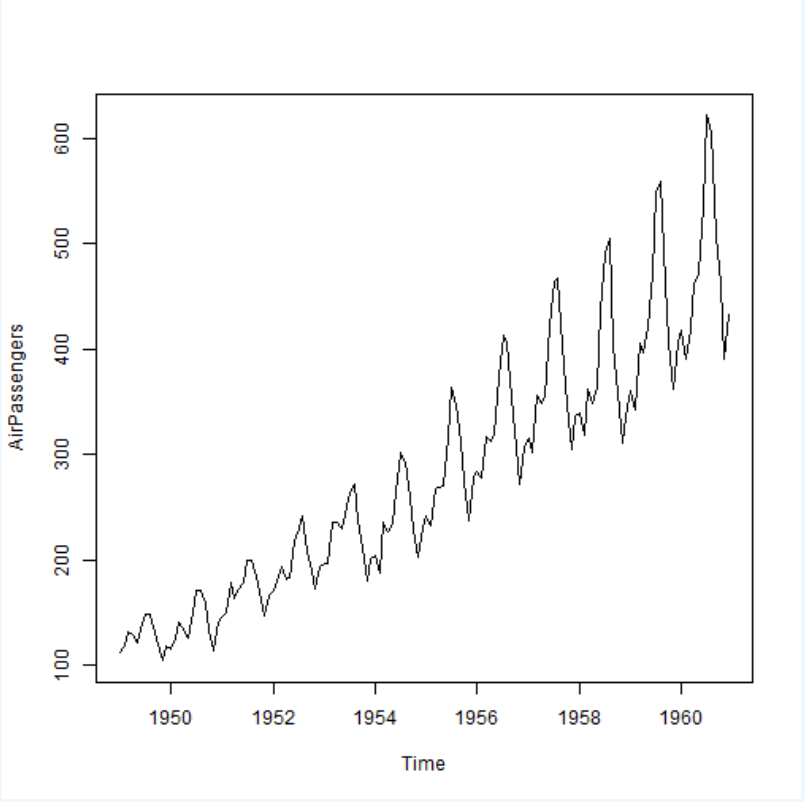
```
##
##  4  6  8
## 11  7 14
```

Note: In R, a table is a different kind of data structure. It is often the result of apply the `table()` function to factors, vectors, or a data frame. All data in a table is of the same type (usually numeric, counting frequencies).

Time-Series Data

- **Time-Series Data**: data points collected or indexed in time order, such as stock prices over time or temperature readings. The followint plot shows the monthly totals of international airlin passengers, 1949-1960.

Example:



Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1976) Time Series Analysis, Forecasting and Control. Third Edition. Holden-Day. Series G.

Cross-Sectional Data

- **Cross-Sectional Data**: data collected at a single point in time or period, providing a snapshot view. Examples include demographic data from a survey or sales data for a particular quarter.

Example:

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Panel Data (Longitudinal Data)

- **Panel Data (or Longitudinal Data)**: a combination of cross-sectional and time-series data. It involves observing multiple subjects (like individuals, companies, countries) over time. The following example shows the first a few rows of data that is related to NFL defense statistics from 2007 to 2017.

Example:

Team	Variable	yr2007	yr2008	yr2009	yr2010	yr2011	yr2012	yr2013	yr2014	yr2015	yr2016	yr2017
Buccaneers	yrswsameqb	0.000	1.000	0.000	1.000	2.000	3.000	0.000	0.000	0.000	1.000	2.000
Buccaneers	qbstexp	6.000	7.000	0.000	1.000	2.000	3.000	0.000	7.000	0.000	1.000	2.000
Buccaneers	tdinter	3.250	2.000	0.556	4.167	0.727	1.588	2.111	0.786	1.467	1.556	1.727
Buccaneers	yrdsppatt	7.500	7.200	6.400	7.300	6.500	7.300	6.300	6.700	7.600	7.200	7.900
Buccaneers	qbrating	94.600	90.200	59.800	95.900	74.600	81.600	83.900	70.500	84.200	86.100	92.200
Buccaneers	1stdwnpass%	0.577	0.617	0.602	0.597	0.650	0.625	0.584	0.616	0.598	0.647	0.690

Data Collection Methods

Here are commonly used data collection methods.

- **Surveys and Questionnaires**: Methods for collecting categorical and numerical data.
- **Observational Studies**: Collecting data without influencing the environment.
- **Experiments**: Data collected from controlled experiments.

Data Collection Methods (Continued)

Surveys and Questionnaires

- **Description**: Utilized for collecting self-reported data from respondents. They can include a variety of question types (open-ended, closed-ended, Likert scales, etc.).
- **Strengths**: Versatile and can reach a wide audience, good for collecting a broad range of information.
- **Limitations**: Subject to biases (e.g., response bias, selection bias), and the quality of data depends on the design of the survey and the honesty of respondents.
- **Applications**: Used in market research, social science, health studies, and more.

Data Collection Methods (Continued)

Observational Studies

- **Description**: Involves collecting data by observing subjects in their natural environment without manipulation or intervention.
- **Types**: Can be either prospective (observing forward in time) or retrospective (looking back at past data).
- **Strengths**: Useful for studies where intervention is not ethical or feasible, provides a more natural insight into the subject matter.
- **Limitations**: Potential for observer bias, and it can be challenging to establish causality.
- **Applications**: Common in sociology, anthropology, biology, and epidemiology.

Data Collection Methods (Continued)

Experiments

- **Description**: Controlled studies where variables are manipulated to observe the effect on the subject(s) being studied.
- **Types**: Can include randomized controlled trials, field experiments, lab experiments, etc.
- **Strengths**: Allows for the establishment of cause-and-effect relationships, high level of control over variables.
- **Limitations**: Can be expensive and time-consuming, may not always accurately represent real-world scenarios.
- **Applications**: Widely used in the natural sciences, psychology, medicine, and marketing research.

Data Quality

In the realm of data analysis, the quality of our data is as crucial as the analysis itself. This section delves into the fundamental aspects of Data Quality, a pivotal component that directly influences the reliability and validity of our analytical outcomes. We'll explore essential concepts such as accuracy and precision, understand the implications of missing and outlier values, and discuss common sources of bias and error in data collection. By grasping these core principles, we'll be equipped to identify, assess, and address data quality issues, ensuring that our analyses are based on solid, trustworthy data. This foundational understanding is vital for making informed, data-driven decisions.

Accuracy and Precision in Data Collection

- `Accuracy`: Refers to the closeness of a measurement to its actual, true value. For example, if a scale consistently reports a person's weight as it truly is, the scale is accurate.
- `Precision`: Refers to the consistency of repeated measurements. For instance, if we measure the length of the same object multiple times and get very similar results each time, our measuring process is precise, even if it's not close to the true length (not accurate).

Data Quality (Continued)

Understanding and Handling Missing Data

- **Common Reasons for Missing Data**
  - `Data Entry Errors`: Mistakes made during data input, such as skipping fields or entering incorrect information.
  - `Non-Responses in Surveys`: Survey respondents may choose not to answer certain questions or drop out of the survey entirely.
  - `Systematic Issues`: Faults in data collection systems or processes, leading to gaps in data.
  - `Unavailability or Inapplicability`: In some cases, data may be missing because it is not applicable to a particular case or simply unavailable at the time of collection.

Data Quality (Continued)

- **Impact of Missing Data on Analysis**
  - `Biased Results`: Missing data can lead to biased estimates if the missingness is related to the value itself (e.g., people with higher incomes might be less likely to disclose them).
  - `Reduced Statistical Power`: Missing data reduces the sample size, potentially weakening the statistical power of the analysis.
  - `Complications in Analysis`: Many analytical techniques assume complete data, and missing values can complicate or invalidate these methods.

Data Quality - Data Preprocessing

- **Basic Methods for Handling Missing Data**
  - **Data Deletion**:
    - `Listwise Deletion`: Removing entire records where any single value is missing.
    - `Pairwise Deletion`: Used in correlation and covariance calculations, where only the specific missing values are excluded.

  Appropriateness: Simple but can lead to biased results if the missingness is not random (Missing Completely at Random - MCAR).

Data Quality - Data Preprocessing

- **Imputation Techniques**:
  - `Mean/Median/Mode Imputation`: Replacing missing values with the mean, median, or mode of the observed values.
  - `Regression Imputation`: Estimating missing values using regression models based on other variables.
  - `Multiple Imputation`: Creating multiple imputed datasets and combining the results to account for the uncertainty of imputations.
    `Appropriateness:` More sophisticated than deletion, helps preserve data and reduce bias, especially when missingness is systematic (Missing at Random - MAR) or related to the missing value itself (Not Missing at Random - NMAR).

Choosing the Right Method

The choice of method depends on the nature and pattern of the missingness, the amount of missing data, and the specific analysis being conducted. It is crucial to understand the reasons behind missing data to choose the most appropriate handling technique.

Data Quality (Continued)

- **Understanding and Managing Outliers in Data**
  - **Definition and Causes of Outliers**
    - `What Are Outliers`: Outliers are data points that significantly differ from the majority of the dataset. They can appear as exceptionally high or low values compared to the rest of the data.
    - `Common Causes`:
      - `Measurement Error`: Mistakes or malfunctions in data collection can produce erroneous values.
      - `Natural Variation`: Inherent variability in data can lead to extreme values.
      - `Data Entry Errors`: Incorrect input or processing of data values.
      - `Sampling Issues`: Anomalies can arise from non-representative or biased sampling methods.
      - `Intentional Outliers`: In some cases, outliers might be valid extreme values, like exceptional performance in sports or unusual responses in surveys.

Data Quality - Data Preprocessing

- **Identifying Outliers**
  - `Box Plots`: Visualize the distribution of data and identify points that fall outside the interquartile range (IQR). Outliers are often depicted as dots beyond the 'whiskers' of the box plot.
  - `Standard Deviation Method`: Identify outliers as points that lie beyond a certain number of standard deviations (e.g., more than 2 or 3) from the mean.
  - `Z-Score Analysis`: Calculate the Z-score for each data point and flag those with a Z-score that exceeds a threshold as outliers.
  - `Scatter Plots`: Visually inspect data distributions for points that deviate significantly from others.

Data Quality - Data Preprocessing

- **Strategies for Handling Outliers**
  - `Exclusion`: Removing outliers from the dataset, typically when they are deemed errors or irrelevant to the analysis.
  - `Transformation`: Applying transformations (like logarithmic or square root transformations) to reduce the impact of extreme values.
  - `Imputation`: Replacing outlier values with estimates based on the rest of the data.
  - `Further Investigation`: Sometimes, outliers warrant further exploration to understand their cause or to validate their authenticity.
  - `Binning`: Grouping data into bins or categories can sometimes help mitigate the impact of outliers.
- **Making Informed Decisions**

  The approach to handling outliers depends on their nature and impact on the analysis. It is crucial to consider whether outliers contain valuable information or represent errors. .red[The decision to exclude or adjust outliers should be justified and documented, keeping in mind the objectives of the data analysis].

Data Quality (Continued)

Bias and Error in Data Collection

- **Types of Bias**:
  - `Selection Bias`: Occurs when the sample is not representative of the population. This can happen due to non-random sampling methods or excluding certain groups. It can lead to skewed results that do not accurately reflect the broader population.
  - `Response Bias`: Arises when respondents give inaccurate answers, often due to the wording of questions, social desirability, or misunderstanding. It affects the authenticity of the data collected.
  - `Measurement Bias`: Happens when data collection tools or procedures produce results that systematically deviate from the true value. This could be due to poorly calibrated instruments or biased survey questions.

Data Quality (Continued)

Understanding Measurement Error

- `Systematic Errors`: Consistent and repeatable errors that lead to a consistent deviation in measurements. These errors can often be identified and corrected.
- `Random Errors`: Irregular and unpredictable errors that arise from unpredictable factors and add variability to the data. They are harder to correct and usually require statistical methods to manage.
- `Impact on Data Quality`: Both types of errors can significantly affect the reliability and validity of the data, leading to incorrect conclusions.

Data Quality (Continued)

Strategies for Minimizing Bias and Error

- `Design of Experiments/Surveys`: Carefully designing data collection methods to ensure unbiased sampling and accurate measurements. This includes clear, unbiased questionnaires and representative sampling techniques.
- `Validation of Tools`: Regular calibration of measurement instruments and validation of survey tools to ensure they are capturing accurate data.
- `Training and Protocols`: Training data collectors to follow standardized protocols and procedures to reduce human error and variability in data collection.
- `Pilot Studies`: Conducting pilot studies to identify and address potential sources of bias and error before the main data collection.
- `Statistical Adjustments`: Employing statistical techniques to adjust for identified biases and errors, such as weighting responses or using statistical models to correct for measurement errors.

Emphasizing the Importance of Awareness

We highlight the importance of being aware of potential biases and errors in data collection and taking proactive steps to mitigate them. We should encourage critical evaluation of data sources and methodologies in research and analysis.

Conclusion and Discussion

As we wrap up this lesson on data types and structures, it's crucial to reflect on the importance of these concepts in the context of Exploratory Data Analysis (EDA). Understanding the nature of the data we're working with is not merely an academic exercise; it is fundamental to the entire process of data analysis.

- Tailoring Analysis Techniques
- Accurate Interpretation of Results
- Data Preprocessing
- Enhanced Data Visualization
- Preparing for Advanced Analysis
- Ensuring Data Quality

Final Thoughts: As we look deeper into the various aspects of EDA, remember that the foundation of any successful data analysis lies in a solid understanding of the data itself – its type, structure, and characteristics. This understanding is what enables us to transform raw data into meaningful insights and informed decisions.

References

The lectures of this course are based on the ideas from the following references.

- Exploratory Data Analysis by John W. Tukey
- A Course in Exploratory Data Analysis by Jim Albert
- The Visual Display of Quantitative Information by Edward R. Tufte
- Data Science for Business: what you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic