CGT 575/ASM 591
Data Visualization Tools & Applications

Spring 2023
Week 2 Lec 1
Tuesday & Thursday
5:30 PM – 6:45 PM
KNOY 306

01/17/23 1

Introductions

Dr. Vetria L. Byrd, PhD
Associate Professor
Computer Graphics Technology

Dr. Dharmendra Saraswat, PhD
Associate Professor of
Agricultural & Biological
Engineering

Aanis Ahmad
PhD Candidate
Electrical and Computer Engineering

Instructor	Office	Phone	Email	Office Hour *
Vetria Byrd, PhD	KNOY 371	(765) 494-6335	vbyrd@purdue.edu	Monday, 3:30 PM – 4:30 PM and by appointment
Dharmendra Saraswat, PhD	ABE 3041P	(765) 494-6335	saraswat@purdue.edu	TBD and by appointment
Aanis Ahmad	ABE 3116	(765) 775-9103	ahmad31@purdue.edu	TBD and by appointment

* Zoom meeting information provided in Brightspace

01/17/23 2

Announcements

Tableau Online Invites

Tableau Online Workspace

- You should have received an email invite
- You should have access to
  - Main class folder: CGT 575 Spring 2023
  - Public folder
  - Your work folder
- If you did NOT receive an email from me, let me know

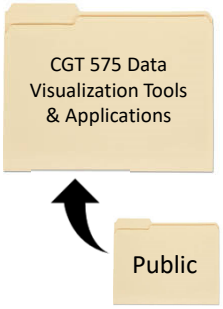
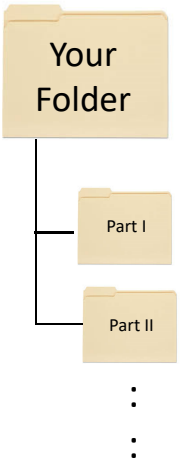


Tableau Online Workspace

- Inside your folder
  - There will be a folder for each training module
    - Part I – Part VII
  - Save your tableau workbooks to the respective folder.
    - Part I
      - Getting Started
      - The Tableau Interface
      - Distributing and Publishing
    - Part II
      - Getting Started with Data
      - Managing Extracts
      - Data Prep with Text and Excel Files

      :

      :



You've got data now what?

Describing Data

Making Sense of Data I

Chapter 2

TABLE 2.1 Data Table Showing Five Car Records Described by Nine Variables

Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	America
Buick Skylark 320	15	8	350	165	3693	11.5	70	America
Plymouth Satellite	18	8	318	150	3436	11	70	America
AMC Rebel SST	16	8	304	150	3433	12	70	America
Ford Torino	17	8	302	140	3449	10.5	70	America

TABLE 2.2 Generalized Form of a Data Table

		Variables				
Observations		x_1	x_2	x_3	$\cdots$	x_p
	o_1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1p}
	o_2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2p}
	o_3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3p}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	o_n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{np}

What would the parsing of this data look like?

TABLE 2.1 Data Table Showing Five Car Records Described by Nine Variables

Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	America
Buick Skylark 320	15	8	350	165	3693	11.5	70	America
Plymouth Satellite	18	8	318	150	3436	11	70	America
AMC Rebel SST	16	8	304	150	3433	12	70	America
Ford Torino	17	8	302	140	3449	10.5	70	America

Variable	Data type
Name	String; alphanumeric
MPG	Integer
Cylinder	Integer
Displacement	Integer
Horsepower	Integer
Weight	Integer
Acceleration	Float
Model Year	Date
Origin	String

Types of Variables

- Each variable within a data table can be examined in different ways.
- A useful initial categorization is to define each variable based on the type of values in the variable has.
- For example,
  - Fixed number is distinct values (*discrete* variable)
  - Numeric value (*continuous* variable)

Types of variables

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a _____ variable.

  A. Discrete

  B. Continuous

Types of variables

Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable.

Why?

Types of variables

Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable.

Why?

Since a fire fighter's weight could take on any value between 150 and 250 pounds.

Types of Variables

Suppose we flip a coin and count the number of heads.

- The number of heads could be any integer value between 0 and plus infinity.
- However, it could not be any number between 0 and plus infinity.
- The number of heads is an example of a _____ variable.
  - A. Discrete
  - B. Continuous

Types of Variables

Suppose we flip a coin and count the number of heads.

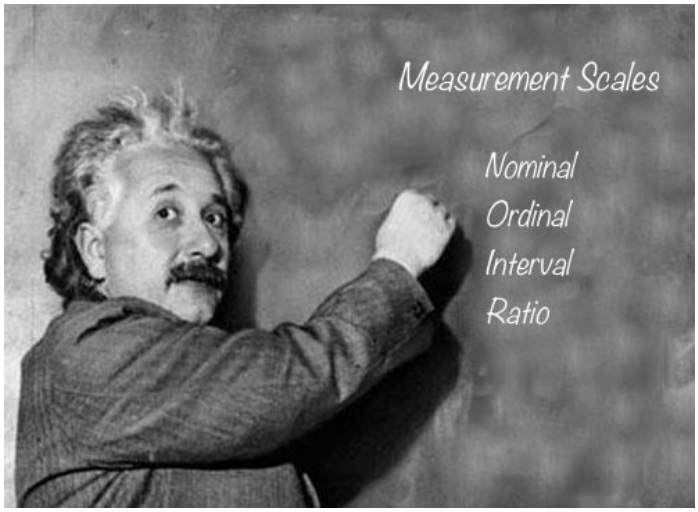
- The number of heads could be any integer value between 0 and plus infinity.
- However, it could not be any number between 0 and plus infinity.
- The number of heads is an example of a discrete variable.
  - Why?

Types of variables

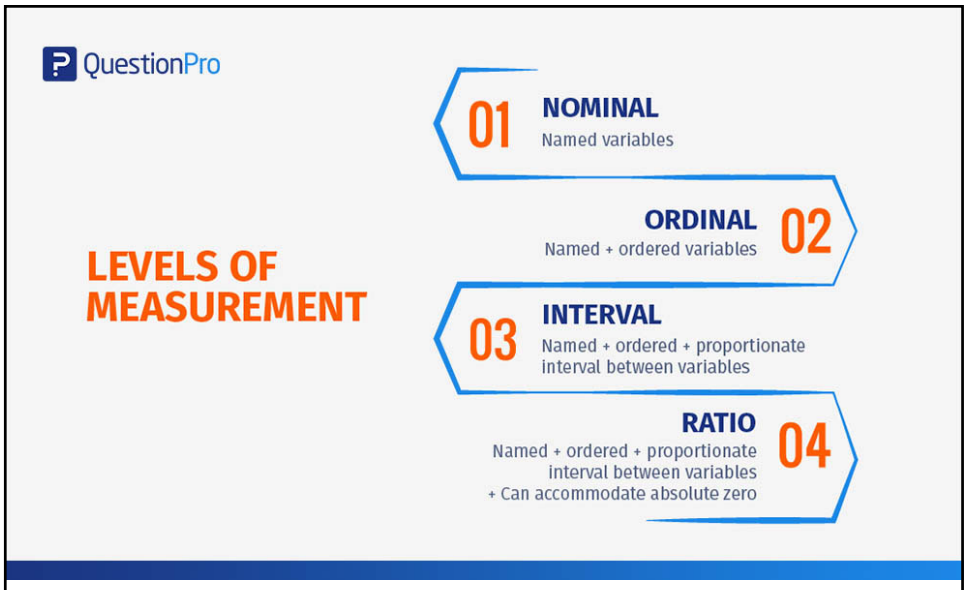
- We could not, for example, get 2.5 heads.
- Therefore, the number of heads must be a ***discrete*** variable.

Variables

- May be classified according to the *scale* on which they are measured.
- Scales help us understand the precision of an individual variable and are used to make choices about data visualization as well as methods of analysis.



https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/



https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/

Summary of data types and scale measures

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Central Tendency

Common descriptive statistical approaches for calculating the central location

Common Descriptive Statistical Approaches

- The *mode* is the most commonly reported value for a particular variable.
- The *median* is the middle value of a variable, once it has been sorted from low to high.
- The *mean*—commonly referred to as the average—is the most commonly used summary of central tendency for variables measured on the interval or ratio scales.
- The *standard deviation* is a measure of how close to the mean value actual data points are.

Visualization

- An aid to understanding the distribution of data:
  - the range of values,
  - the shape created when the values are plotted, and
  - the values called *outliers* that are found by themselves at the extremes of the range of values.

Anscombe's Quartet

Importance of Data Visualization

In-class Activity

Anscombe's Quartet

- Four data sets
- Nearly identical simple descriptive statistics
- Have very different distributions
- Appear differently when graphed

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

Anscombe's Quartet

- It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.
- There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets.

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

Anscombe's Data

Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

Anscombe's Quartet

- This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them
- Suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

In-class Activity

- Download the Anscombe' Quartet spreadsheet
- Calculate the descriptive statistics for each variable
- Generate the scatter plots with trends lines for each data set (on each tab)
- There are 4 data sets.