# Statistics - Introduction

Dr. Vinay Kulkarni

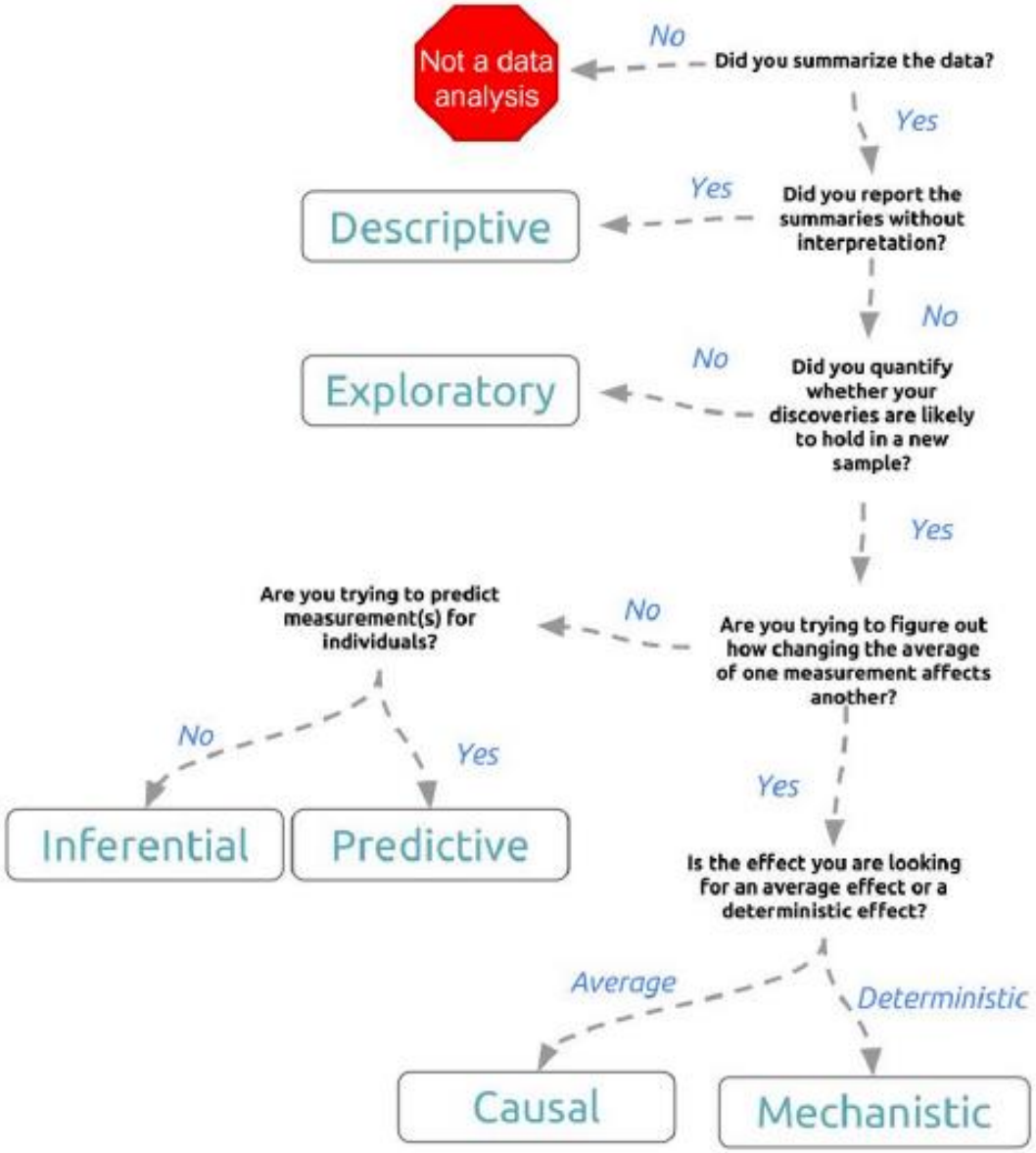
# What is ‘Statistics’

# What is ‘Statistics’

**STATISTICS**

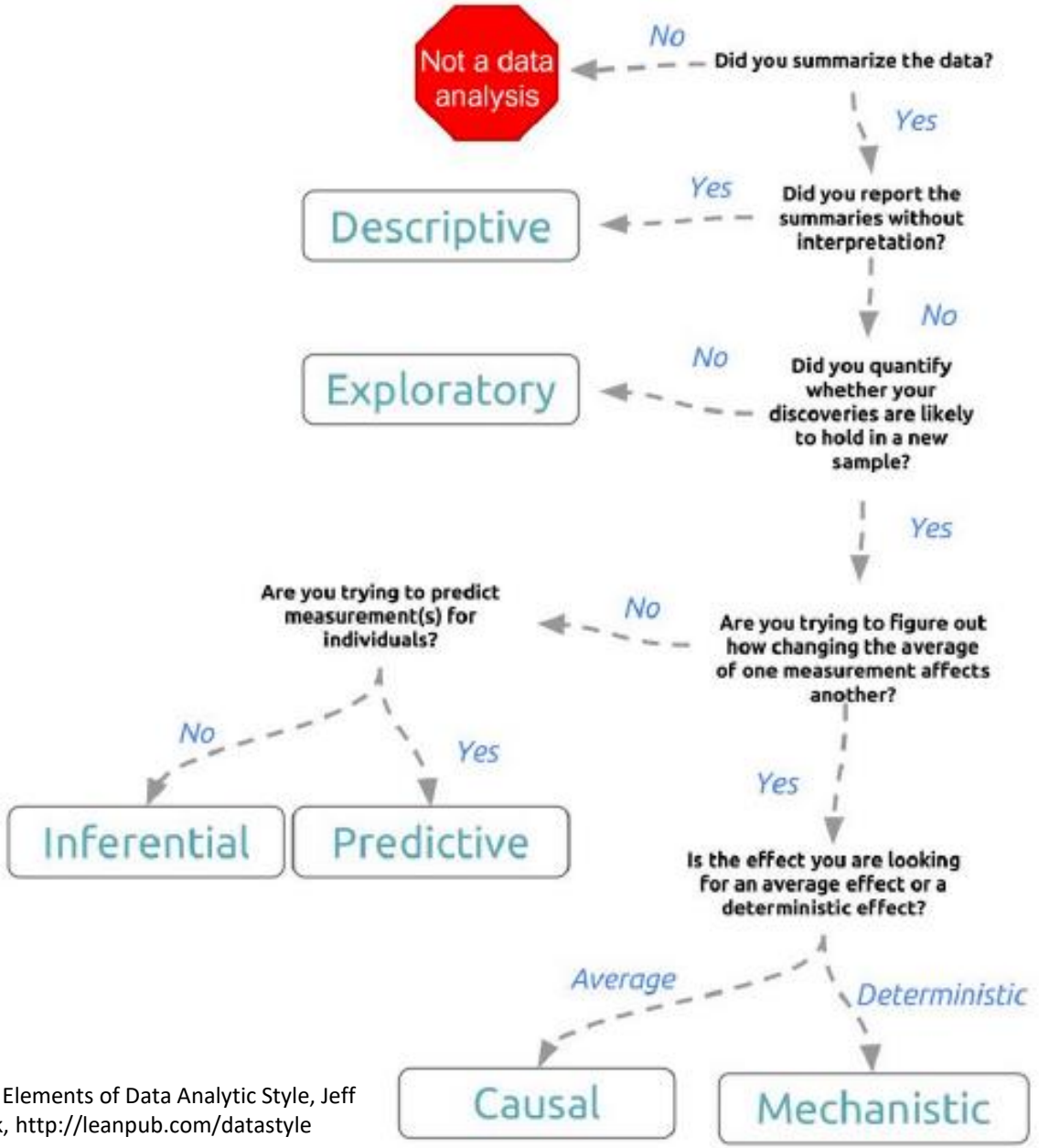
- The branch of science that deals with
  - Collecting data
  - Organizing and summarizing data
  - Analysis of data
  - Inferring / Predicting / Deciding based on the data and its analysis

# Statistics - Branches




Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

# Statistics - Branches

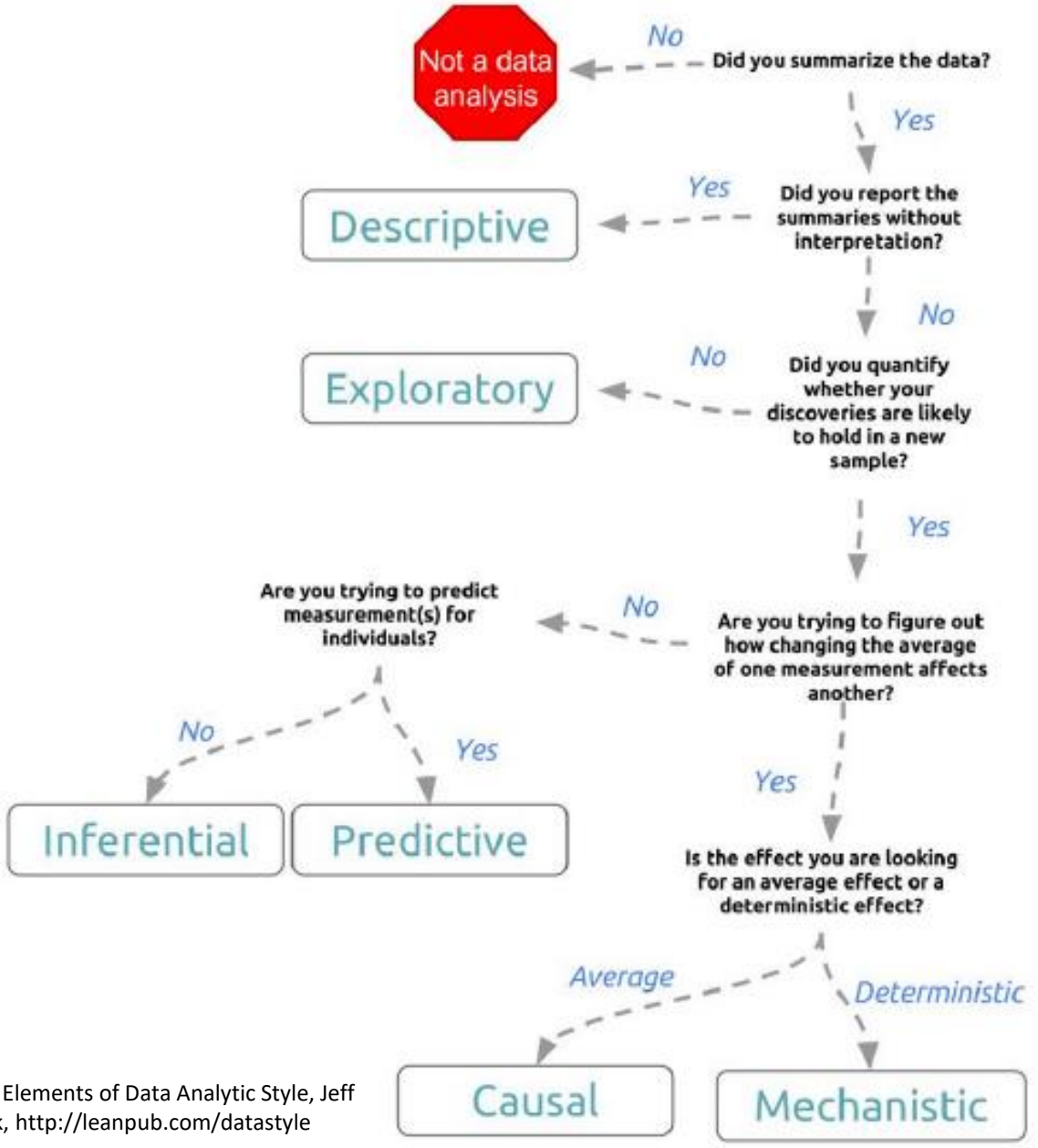



Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

**DESCRIPTIVE STATISITCS**
Seeks to summarize the measurements in a single data set without further interpretation.

**EXPLORATORY ANALYSIS**
Builds on descriptive data analysis by searching for discoveries, trends, correlations or relationships between the measurement of multiple variables to generate ideas or hypotheses.

# Statistics - Branches




Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

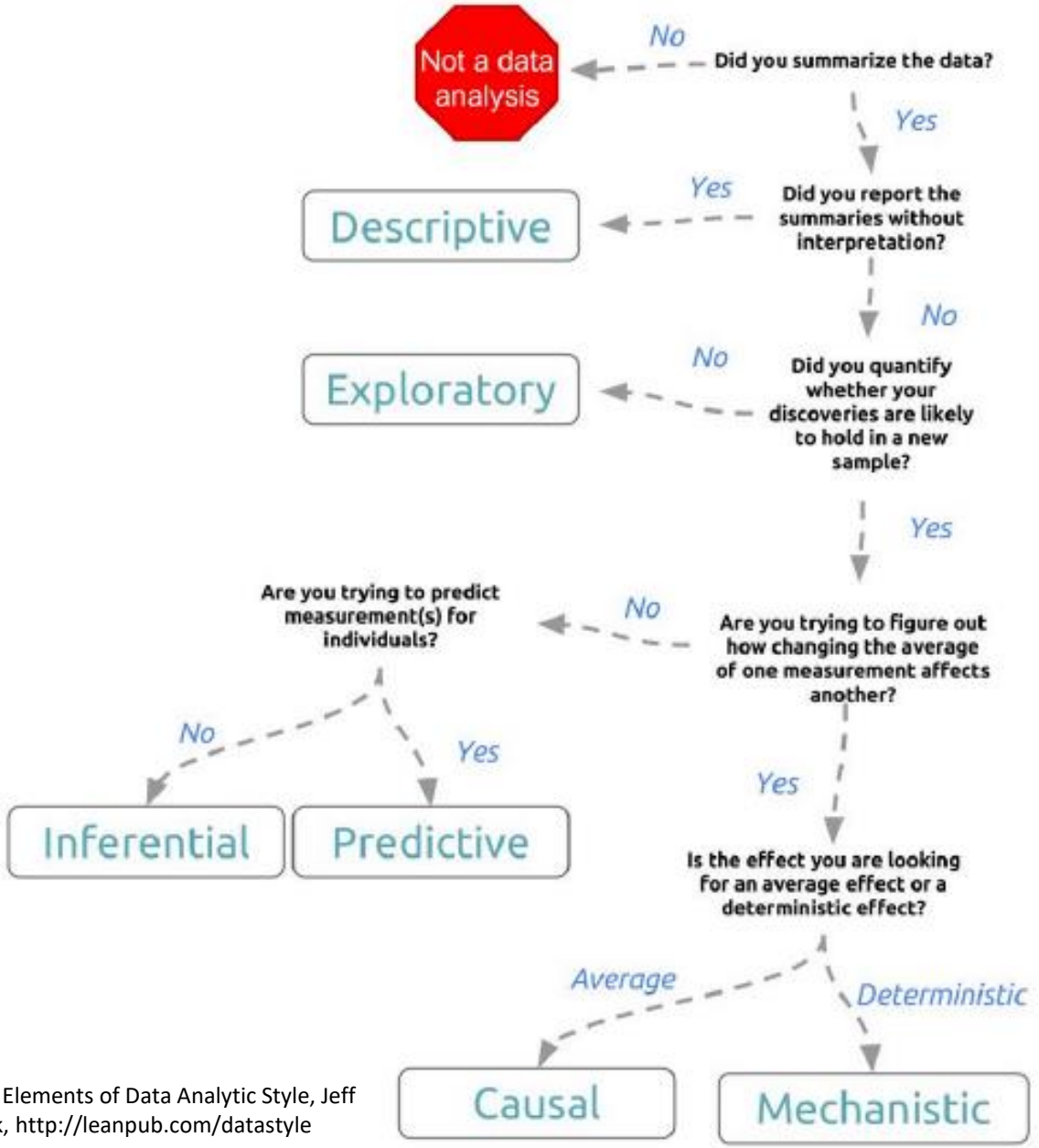
**INFERENTIAL ANALYSIS**

Goes beyond exploratory analysis by quantifying whether an observed pattern will hold beyond the data set in hand – relationships among measurements at population scale. This is the most common form of data analysis.

**PREDICTIVE ANALYSIS**

This uses a subset of measurements (features) to predict another measurement (outcome) for a person or a unit. There is however no attempt to explain why the prediction works.

# Statistics - Branches




Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

**CAUSAL ANALYSIS**

Seeks to reliably find out what happens to one measurement if you make changes to another measurement. Unlike predictive or inferential data analysis, causal analysis identifies both – magnitude and direction of relationships between variables.

**MECHANISTIC ANALYSIS**

Mechanistic analysis seeks to demonstrate that changing one measurement always and exclusively leads to a specific deterministic behaviour in another.

# The Big Picture

- Foundations
- Methods
- Tools
- Applications
- Advanced applications

# FOUNDATIONS

# What is ‘Statistics’

**STATISTICS**

- The branch of science that deals with
  - Collecting data
  - Organizing and summarizing data
  - Analysis of data
  - Inferring / Predicting / Deciding based on the data and its analysis

# Population v/s Sample

- Population
  - It is the collection of the **entire set** of objects that share a property or properties, which need to be studied
    - Population of a state
    - Population of all students in a university

- Sample
  - A smaller and a representative part, or subset, of the population
    - Important since it always impossible to know about and deal with the entire population

# Population v/s Sample: Example

- In a population comprising all students of a class, select three students to form a committee
  - Three students = Sample
  - All students of the class = Population

- Aspects related to samples:
  - Size of sample?
  - Adequacy of sample?
  - Selection of the sample?
  - Quality of results based on the sample?

# Rule: Sample v/s Population

Any set of data should be considered as a **Sample** until it is clearly specified that data is the whole **Population**

# Attributes of a Sample / Population

- Aspects that we observe about the population or sample are known as:
  - Observations
  - Measurements
  - Scores
  - Data

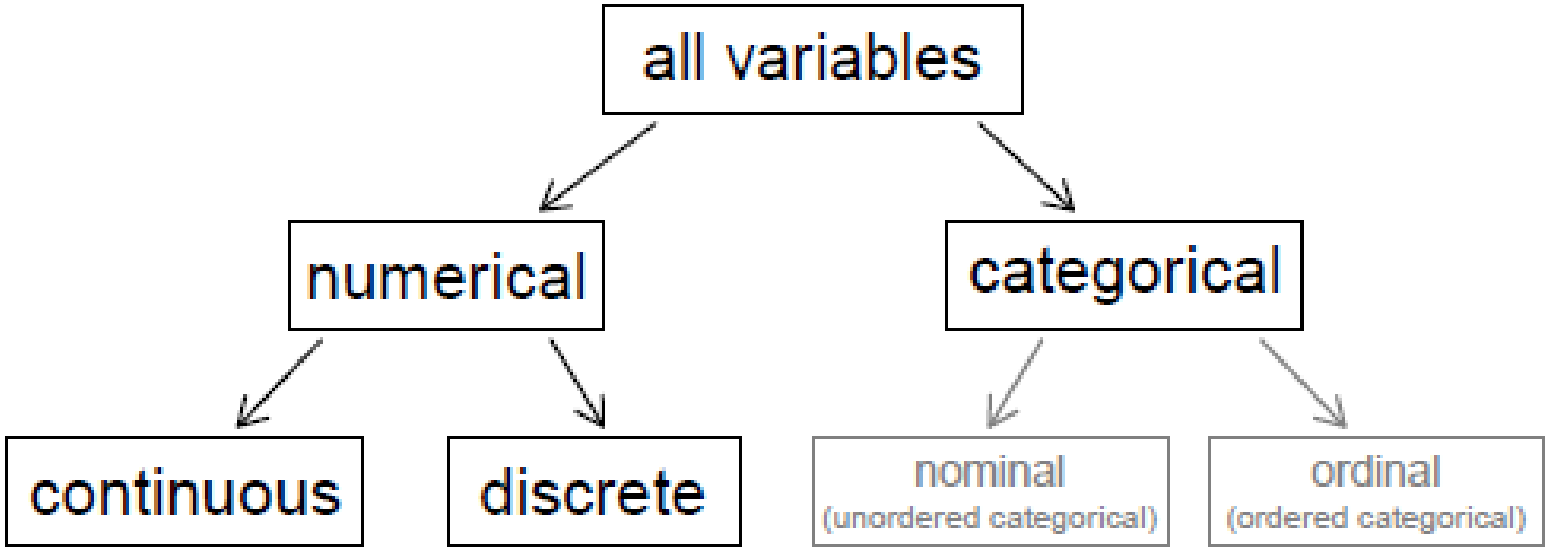
- Levels of Measurements
  - There are different types of measurements
  - Detailed in the next slide

# Types of Measurements / Data

- Nominal Data
  - No obvious ordering / no arithmetic
  - Example: Gender
- Ordinal Data
  - Can be arranged in some order / no arithmetic
  - Example: Grades in a course
- Interval Data
  - Similar to ordinal data / Subtraction possible
  - Example: Time interval; 1985 - 2014
- Ratio Data
  - Similar to interval / Division possible
  - Example: Income information

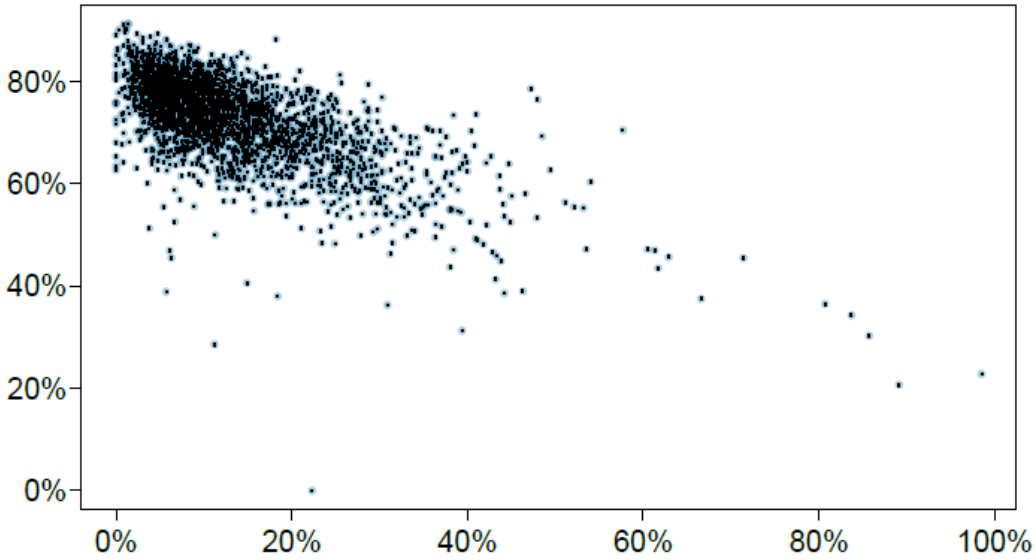
# Variables

- Variables are characteristics of individual parameters of a population
- Variables can be of following types:



# Variables

- When two variables show some connection with each other they are known as associated variables or dependent variables




## Associated or independent, not both

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

# Variables

- Explanatory variables & Response Variables
  - Also known as independent & dependent variable
- Correlation and causation
  - Confounding variable
  - A variable that is correlated with both, the explanatory and response variables