Exploratory Data Analysis

Jerry Kiely

2023-11-17

Contents

Descriptive Statistics

Measures of Central Tendency and Variation

Sources and Types of Data

Sources of Data

1. Primary Data

- the data is collected by the investigator himself / herself for a specific purpose
- direct method of data collection
- eg. data collected for research through questionnaires, interviews

2. Secondary Data

- the data is collected by someone else, but being used by the investigator for some other purpose
- an indirect method of data collection
- eg. census data being used to study the impact of education on income

Types of Data

1. Structured Data

- information is stored with a high degree of organization
- contains qualitative data, quantitative data, or a mixture of both
- eg. data arranged in an excel file, in rows and columns

2. Unstructured Data

- information that either does not have a pre-defined data model and / or is not organized in a pre-defined manner
- eq. emails, tweets, blogs, etc.

Measurement Scales

1. Nominal Scale

- the placing of data into categories without any order or structure
- no numerical relationship between categories - even if numbers are used for representation
- eg. gender, nationality, language, region, etc.

2. Ordinal Scale

- the placing of data into categories such that the order of values is meaningful, but relative degree of difference is not known
- eg. ranking the features of a product on a scale of 1 to 5
- the Likert scale - psychometric scale commonly used in questionnaires

Highly Satisfied	Dissatisfied	Neutral	Satisfied	Highly Satisfied
1	2	3	4	5

3. Interval Scale

- numeric scale in which the order as well as the relative difference between values is known
- no "true zero"
- eg. temperature can be below $0°C$

4. Ratio Scale

- numeric scale with an absolute "zero"
- addition, subtraction, multiplication, and division are all valid operations
- eg. height, weight age, etc. - always measured from 0 to a maximum value

Measures of Central Tendency

a.k.a. Measures of Central Location

- a single value that describes a set of data by identifying the central position within that set of data

The most commonly used measures of central tendency are:

- Mean
  - arithmetic mean, commonly known as average
  - sum of all values divided by the number of values
- Median
  - arrange N data elements in order
  - if N is odd take the middle value
  - if N is even take the average of the two middle values
- Mode
  - the most frequently occuring value in a data set

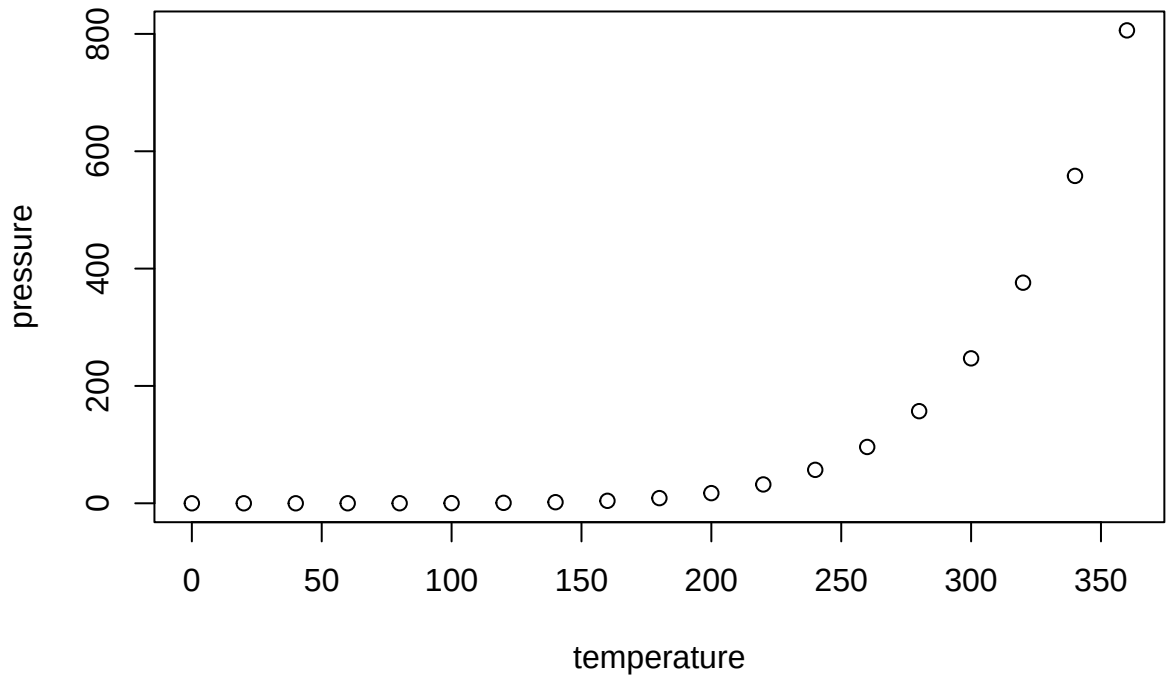
The mean, median, and mode are all valid measures of central tendency, but under different conditions, some measures are more appropriate than others.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Beyond Mean and Variance

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.