Statistics

Adejumo Ridwan Suleiman

March 22, 2005

Introduction to Statistics

What is Statistics?

- Statistics is the collection, organizing and analysing of data.

Is Data Science Statistics in Disguise?

- Unlike Statistics, Data Science is an interdisciplinary field consisting of Mathematics, Statistics, Computer Science and Domain Knowledge.

Types of Data

- Data can be classified into two types
  - Based on Measurement scale
  - Based on Time Period

Based on Measurement Scale

- Qualitative Data
  - Nominal Data e.g sex
  - Ordinal Data e.g temperature level; High, Medium and Low
- Quantitative Data
  - Ratio e.g weight
  - Interval e.g temperature in degreee celsius

Based on Time Period

- Cross-Sectional Data e.g number of viewers for different youtube genres in the year 2021
- Time Series Data e.g number of viewers for Sport channels on youtube from the year 2014-Date.

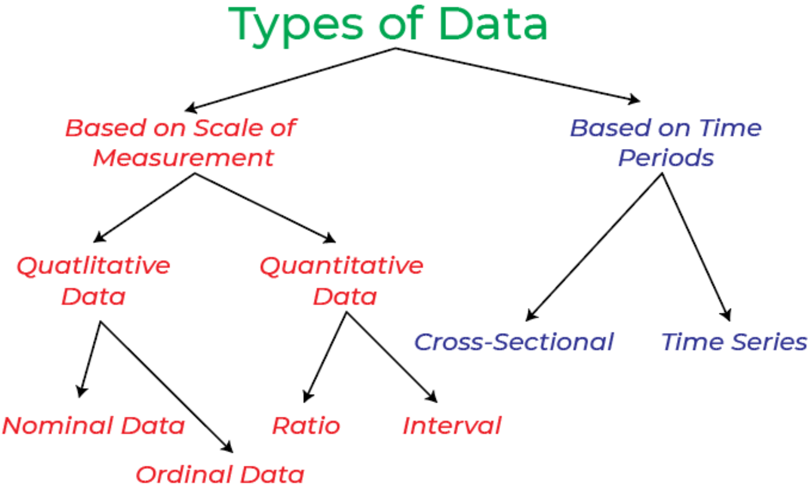
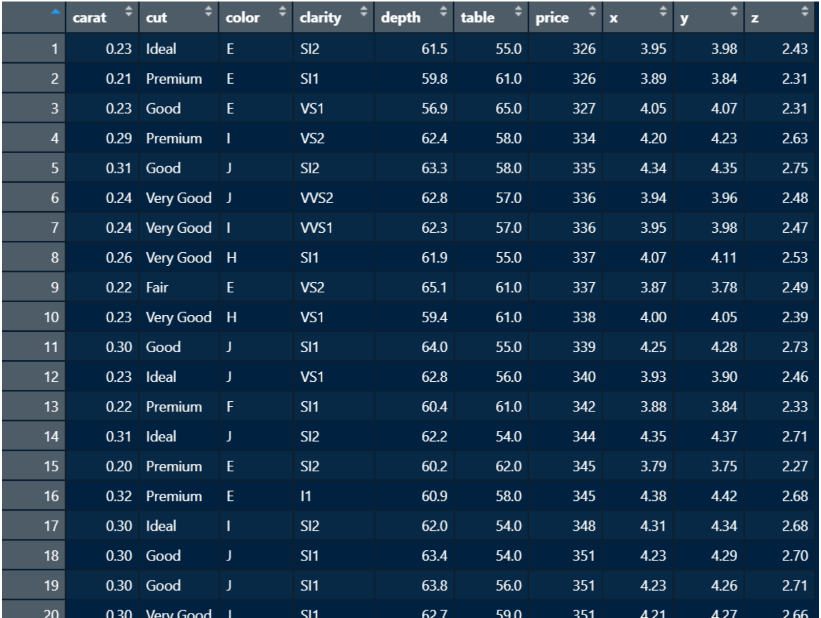


Figure 1: Types of Data

Rectangular or Structured Data



	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68
18	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29	2.70
19	0.30	Good	J	SI1	63.8	56.0	351	4.23	4.26	2.71
20	0.30	Very Good	J	SI1	62.7	59.0	351	4.21	4.27	2.66

Measures of Central Tendency

- Mean
- Median
- Mode

Mean

- Sum of all values of observations divided by the number of observations
- Mathematically denoted as:
  $$
- Sensitive to extreme or high values

Median

- Center of an ordered observations
- Also known as the middle of the observations.
- Not sensitive to extreme values