Lecture 1
Fundamentals of Descriptive Statistics

July 30

What is ‘DATA’

- For us: useful information to draw statistical conclusion.
- In this course, data = **structured format**

- Structured data types:
    - Tabular/spreadsheet data (each column can be of different type).
    - Multi-dimensional data
    - Tables of data related by key columns.
    - Time - series (evenly or unevenly spaced)
    - ……
- Large % of real world data can be transformed into structured format.

Types of ‘DATA’ we work with

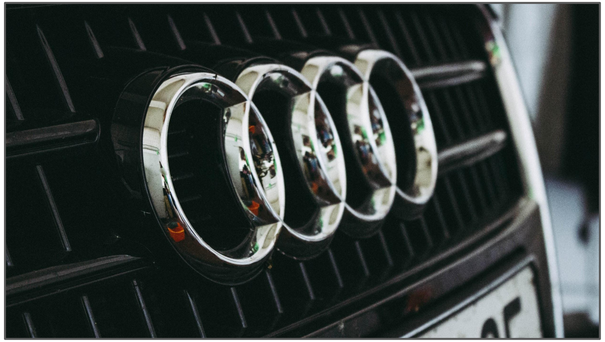
Classify data in 2 main ways:

1. Based on **type**
   a. Categorical
   b. Numerical

2. Based on **measurement level**
   a. Qualitative
   b. Quantitative

Categorical data

- Describes categories or groups.
- Car brands: Audi, BMW, Mercedes …..



- Answers to Questions (YES/ NO)
  - Do you study at McGill University?

Numerical data

- Deals with numbers.
- Further divided:
  - Discrete
  - Continuous

- Discrete data: counted in finite measure.
  - eg: Number of students in a class.
  - You can imagine each member of the data.
- Continuous data: Opposite of discrete.
  - ∞ (infinite possibilities)
  - Difficult to count
  - eg: stars in the sky, your weight

Based on Measurement

Two groups:

- Qualitative
    - Further divided:
        - **Ordinal**: The order matters.
            - eg: Grades (A, A-, B+, B-,....)
        - **Nominal**: No particular order.
            - eg: Car manufacturers (Audi, BMW, Mercedes), seasons (winter, summer, spring)
- Quantitative
    - Represented by numbers. Further divided:
        - **Ratio**: Has a true zero.
            - eg: length
        - **Interval**: Don’t have a true zero
            - eg: temperature

Visualization

- Most intuitive way to interpret the data.
- Good for a pictorial summary of the data.
- Easy to spot anomalies.
- Can identify basic patterns.
- Very helpful during presenting your work.
    - You are encouraged to use visualization tools for your assignments to both explain the data and how you used the information from visualization in your decision making.

Visualization: Categorical data

Popular methods to visualize categorical data:

- Frequency distribution tables
- Bar plots
- Pie charts
- Pareto diagrams

Visualization: Categorical data - **Frequency Tables**

- Has 2 columns

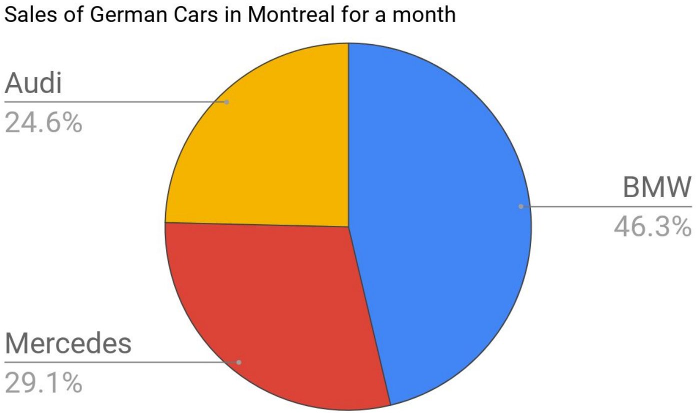
Category	**Frequency**
BMW	124
Audi	66
Mercedes	78
Total	335

Number of units sold by a car shop in a month

Visualization: Categorical data - **Pie Charts**

- Represents the same data using pie-charts.
- Need to calculate the percentage of brand (relative frequency).

Category	**Frequency**	**Rel Freq**
BMW	124	46.3%
Audi	66	24.6%
Mercedes	78	29.1%
Total	335	100%

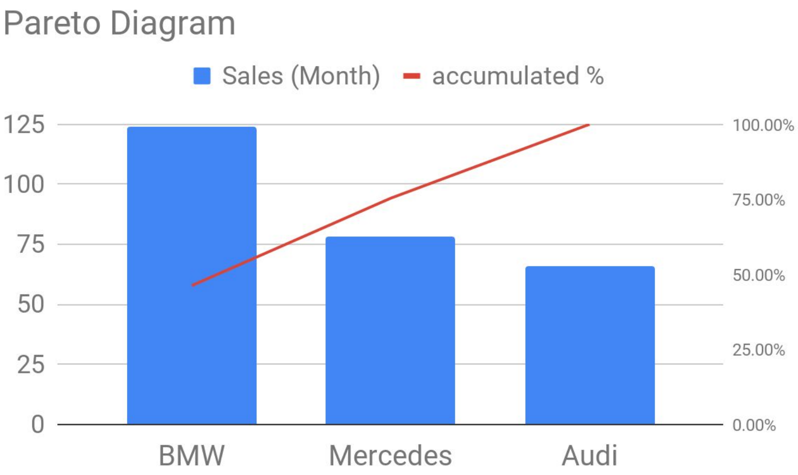


- More intuitive than Tables.
- Good for comparison & to see the share in total.

Visualization: Categorical data - **Pareto Diagram**

- Special type of bar chart where categories are in descending order (of freq).
- Need to calculate the percentage of brand (relative frequency).

Category	**Frequency**	**Rel Freq**
BMW	124	46.3%
Audi	66	24.6%
Mercedes	78	29.1%
Total	335	100%



- More intuitive than Tables.
- Good for comparison & to see the share in total.

Visualization: Numerical data

```
dataset = array([ 3, 46,  9,  4, 24, 33, 45,  1, 32, 19, 22, 49, 40, 43, 12,
24, 43, 4, 43, 23])
```

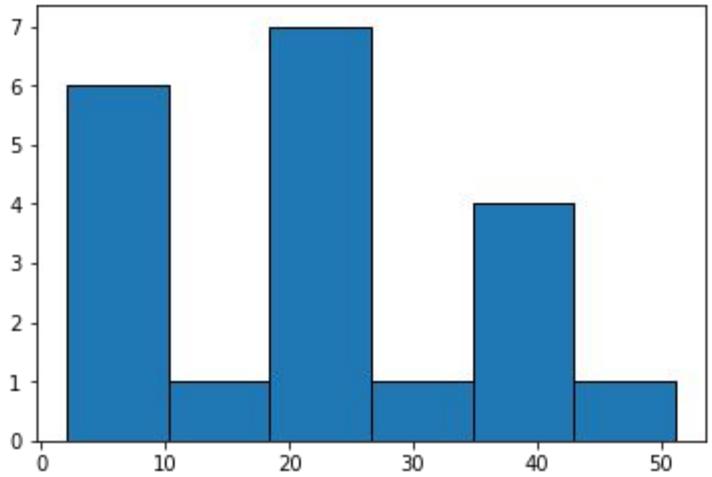
- First order the data in a table.
- Make Intervals.
  - Interval width = {Max(dataset) - Min(dataset)} / #intervals
  - For #interval = 5; interval width = 8.2
  - U can also round up this number. So 8.5.

- Histogram Charts
  - Plot of ‘frequency of occurrence’ vs ‘range of variation’ of the data.

Visualization: Numerical data - Frequency dist table

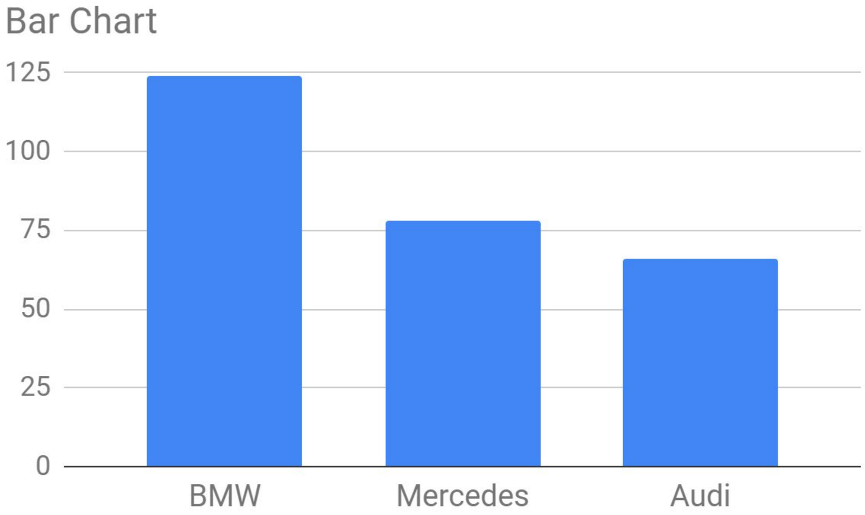
```
(array([6., 1., 7., 1., 4., 1.]),

 array([ 2. , 10.2, 18.4, 26.6, 34.8, 43. , 51.2]),
```

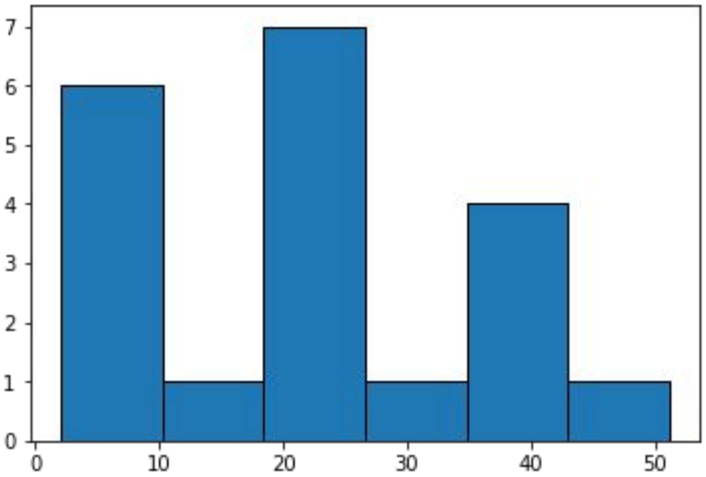
Interval Start	Interval End	Frequency
2	10.2	6
10.2	18.4	1
26.6	34.8	7
34.8	43	4
43	51.2	1



Spot the difference

Bar Chart vs Histogram





Histograms

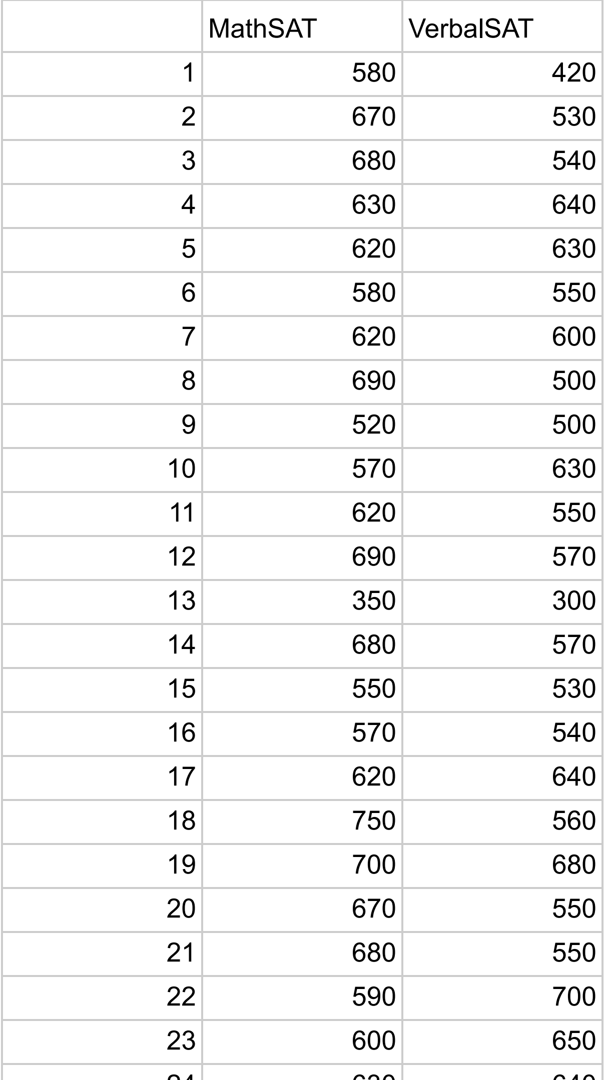
- Can have unequal intervals.
- Example: Surveys with the following options:
  - What is your age?
    - Less than 18
    - 18-35
    - 35-60
    - More than 60

Cross tables & Scatter plots

- Till now we dealt with only one variable.
- Now we will look at relationships between 2 variables

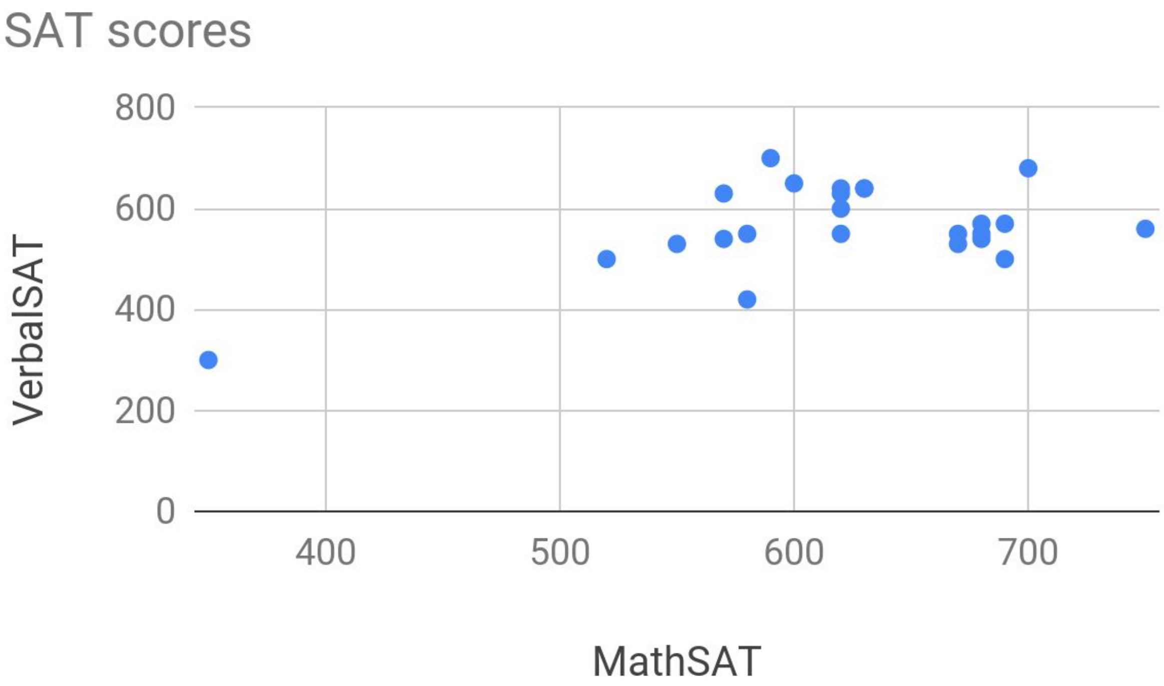
- Categorical variables: Cross tables/ Contingency tables

- Numerical data: Scatter plots



	MathSAT	VerbalSAT
1	580	420
2	670	530
3	680	540
4	630	640
5	620	630
6	580	550
7	620	600
8	690	500
9	520	500
10	570	630
11	620	550
12	690	570
13	350	300
14	680	570
15	550	530
16	570	540
17	620	640
18	750	560
19	700	680
20	670	550
21	680	550
22	590	700
23	600	650

Scatter plot


SAT scores
VerbalSAT
MathSAT