INT303 W2 Note

Data Science Process

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

Data Collection and Exploration

Types of Data

Simple or atomic:

- **Numeric**: integers, floats
- **Boolean**: binary or true false values
- **Strings**: sequence of symbols

Compound, composed of a bunch of atomic types:

- **Date and time**: compound value with a specific structure
- **Lists**: a list is a sequence of values
- **Dictionaries**: A dictionary is a collection of key-value pairs, a pair of values x : y

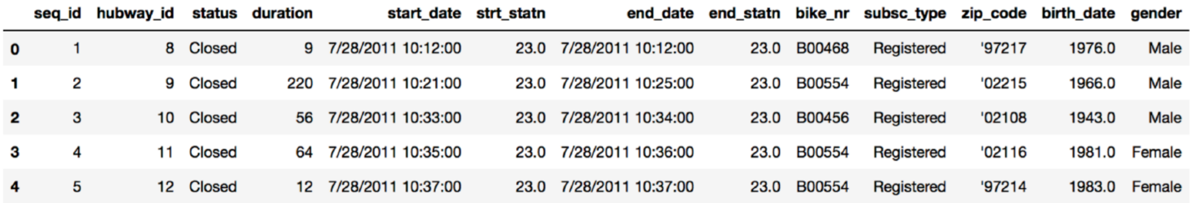
Data Storage

- **Tabular Data**: 一个二维表，其中每行通常代表单个数据记录，每个列代表一种类型的测量（csv、dat、xlsx 等）
- **Structured Data**: 数据用（复杂的）dict 的形式储存（json, xml, etc.）
- **Semistructured Data**: 并非所有记录都由同一组 keys 表示，或者某些数据记录不使用 key-value pair 结构表示

Data Format

- Textual Data
- Temporal Data
- Geolocation Data

Tabular Data



	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

每种类型的测量都称为数据的 variable 或 attribute（例如。 seq_id、status 和 duration 是 variable 或 attribute）。attribute 的数量称为 dimension。这些通常称为 features。

Types of Data

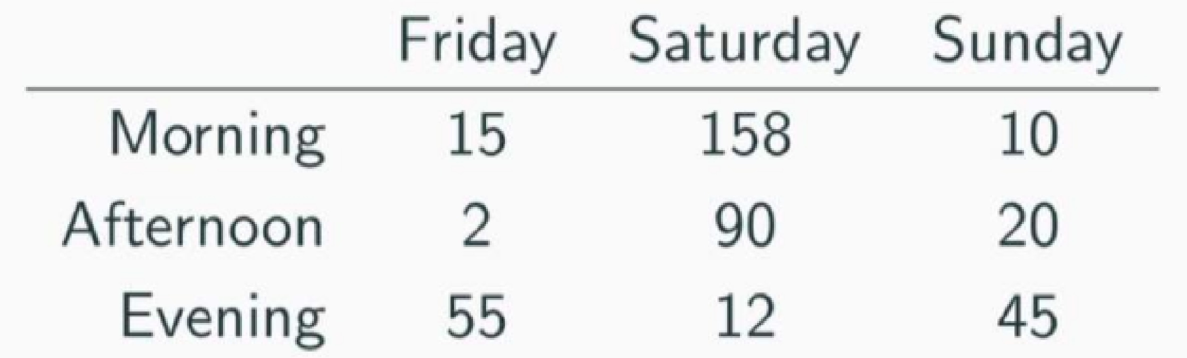
- **Quantitative variable**: is numerical and can be either:
  - **discrete**：在任何有限制的间隔内，都可能获得数量有限的值。例如："兄弟姐妹的数量"是一个离散的变量
  - **continuous**：在任何有边界的间隔内都可能具有无限数量的值。例如："高度"是一个连续变量
- **Categorical variable**：值之间没有固有的顺序，例如："你有什么样的宠物"是一个 categorical variable

Common Issues

- Missing values: how do we fill in?
- Wrong values: how can we detect and correct?
- Messy format
- Not usable: the data cannot answer the question posed

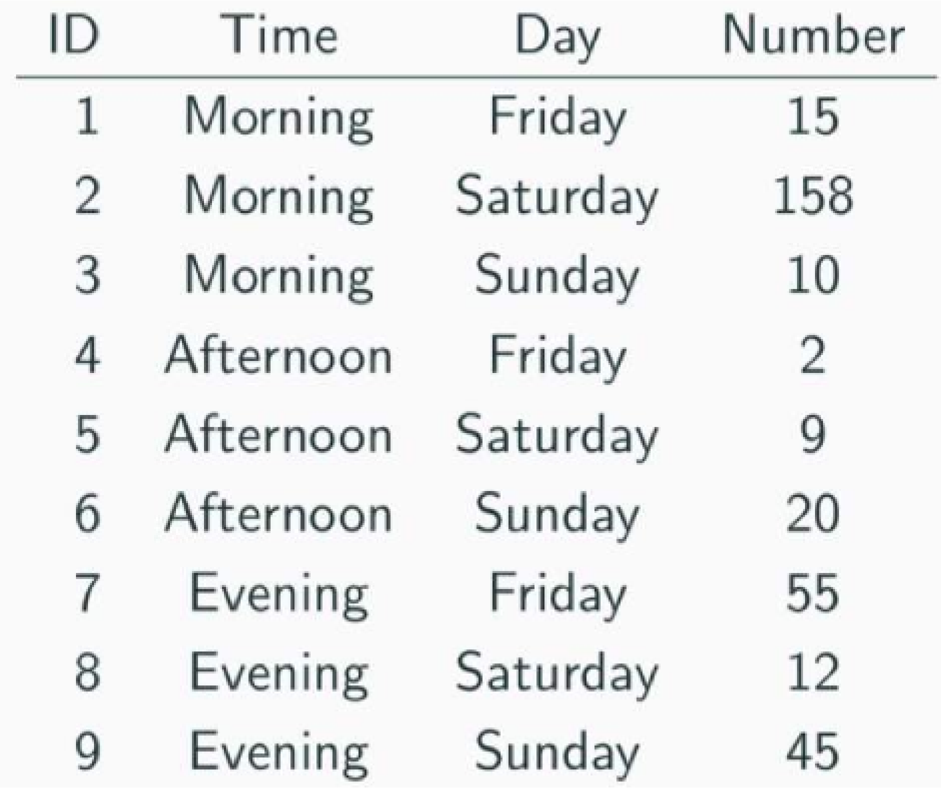
Messy Data

下表是一个周末的农产品交货数量



	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Problem：Day 和 time 都是自变量，数字是因变量。但上表的形式不利于观察和处理。应变为以下形式：



ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

Data Exploration: Descriptive Statistics

Basics of Sampling

Population versus sample:

- **population** 是正在研究的对象或事件的整个集合
- **sample** 是正在研究的对象或事件的"代表"子集（“representative” subset）

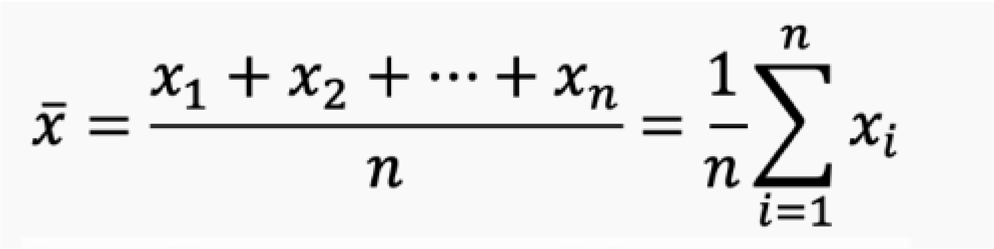
Biases in samples:

- **selection bias**：sample 中的某些项目或记录（subjects or records）更有可能被选中
- Volunteer/**nonresponse bias**：sample 中不容易获得的项目或记录将不被代表

Sample mean and median

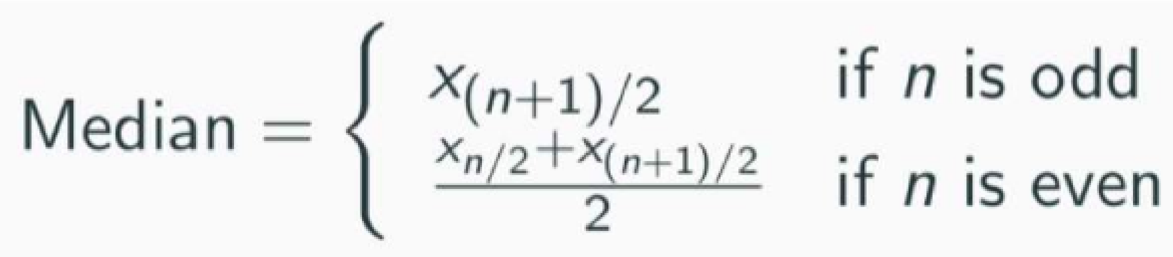
平均数：

平均值是对极端值（或离群值，outliers）的反映



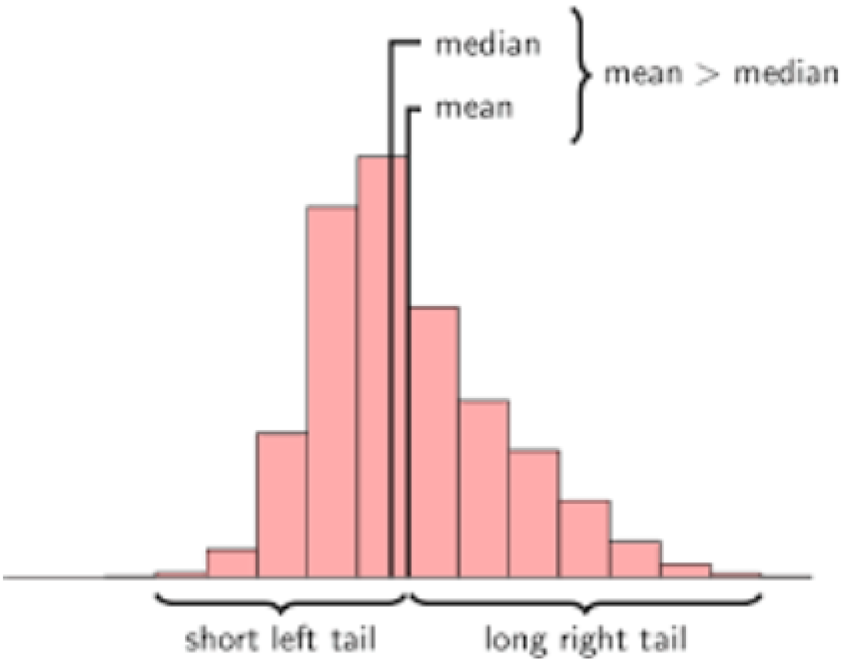
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

中位数：



$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2}+x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

Skewness



上述分布称为右偏度（right-skewed），因为平均值大于中位数。

注意：长尾（long tail）在哪边，偏度就是哪边。

Measures of spread

Range

样本的扩散程度（spread of sample）可以用 **range** 来衡量：

Range = Maximum Value - Minimum Value

Variance

方差（variance），表示为 s^2，衡量样本值平均偏离平均值的程度：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|^2$$

Standard deviation

标准差（standard deviation），表示为 s，是方差（variance）的平方根：

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|^2}$$