Basic Statistics for Data science

อาจารย์จารุมาศ แสงสว่าง

สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์

มหาวิทยาลัยราชภัฏบุรีรัมย์

Data science

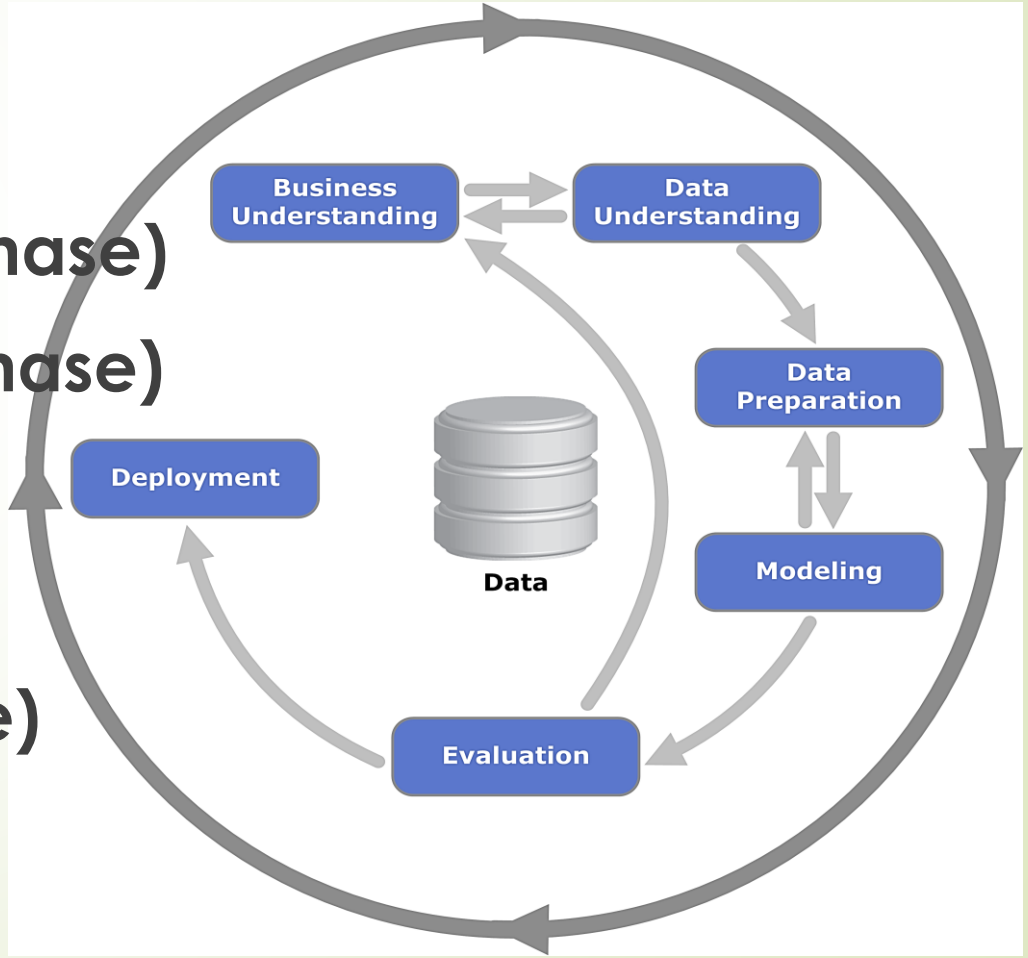
- วิทยาการข้อมูล เป็นศาสตร์ที่มีเป้าหมายในการบริหารสกัดข้อมูลเชิงลึก และวิเคราะห์ข้อมูลเพื่อทำความเข้าใจและอธิบายปรากฏการณ์ต่าง ๆ ที่เกิดขึ้น อีกทั้งมุ่งเน้นการสร้างโมเดลเพื่อวิเคราะห์ข้อมูลสำหรับการทำนายผลที่จะเกิดในอนาคตและสร้างองค์ความรู้ใหม่เพื่อช่วยในการแก้ปัญหาหรือหาช่องทางใหม่ ๆ ในการวางแผนการดำเนินงานขององค์กรจากข้อมูลที่มีอยู่ในหลายรูปแบบ

Data science

- จากการเพิ่มขึ้นของปริมาณข้อมูลในปัจจุบัน ไม่ว่าจะเป็นข้อมูลด้านการแพทย์ ข้อมูลด้านการศึกษา ข้อมูลด้านการเกษตร ข้อมูลสังคมออนไลน์ ข้อมูลการใช้งานอินเทอร์เน็ต ข้อมูลเชิงธุรกิจ ที่ทำให้เกิดวิทยาการใหม่ที่เรียกว่า **Big data** ที่มุ่งเน้นการศึกษาและวิเคราะห์ข้อมูลขนาดใหญ่ที่มีความหลากหลายและปริมาณข้อมูลมหาศาล ซึ่งหากข้อมูลเหล่านี้ไม่ได้ถูกนำมาวิเคราะห์เพื่อใช้ประโยชน์จากข้อสนเทศ จะทำให้พลาดโอกาสที่จะได้องค์ความรู้ที่มีคุณค่าต่อองค์กร ไม่ว่าจะในแง่ของการตลาดหรือการแข่งขัน แง่ของการจัดการปรับปรุง หรือพัฒนาองค์กร

กระบวนการทำงานของ Crisp-dm (Cross-Industry Standard-Process Data-Mining)

- ขั้นตอนเข้าใจธุรกิจ (Business/research understanding)
- ขั้นตอนเข้าใจข้อมูล (data understanding phase)
- ขั้นตอนการเตรียมข้อมูล (data preparation phase)
- ขั้นตอนการพัฒนาโมเดล (modeling phase)
- ขั้นตอนการประเมินโมเดล (evaluation phase)
- ขั้นตอนการนำโมเดลไปใช้ (Deployment phase)



การเรียนรู้ของเครื่อง (Machine Learning)

- **Machine Learning** เป็นศาสตร์ที่ว่าด้วยอัลกอริทึมที่ทำให้เครื่องกลเรียนรู้และเข้าใจในประเด็นที่เราสนใจจากข้อมูล โดยแบ่งการเรียนรู้เป็น 2 แบบ ดังนี้
  - การเรียนรู้แบบมีผู้สอน **(Supervised Learning)** เป็นการเรียนรู้แบบมีเป้าหมายที่ชัดเจน เป็นการเรียนรู้จากข้อมูลผลลัพธ์ที่ระบุไว้ในข้อมูลชุดข้อมูลเรียนรู้ **(Training set)** ซึ่งกระบวนการเรียนรู้นี้จะเป็นการพัฒนาตัวแบบพยากรณ์หรือตัวจำแนกเป็นหลัก
  - การเรียนรู้แบบไม่มีผู้สอน **(Unsupervised Learning)** เป็นเทคนิคการเรียนรู้เพื่อค้นหารูปแบบหรือลักษณะบางอย่างที่เหมือนกันของข้อมูลแต่ละรายการ จะไม่มีการระบุผลลัพธ์ที่ต้องการไว้ก่อน เป็นการเรียนรู้ที่พิจารณาเซตของตัวแปรสุ่มก่อนแล้วจึงสร้างโมเดลร่วมกับชุดข้อมูล

Data science skills

- Hacking Skill (สกิลเกี่ยวกับ Computer Programming, Data Base, Big data Technologies)
- Statistics & Math
- Substantive Expertise (หรือ Domain Knowledge) ความเชี่ยวชาญเฉพาะด้าน เช่น ความเชี่ยญาชาญด้านธุรกิจ

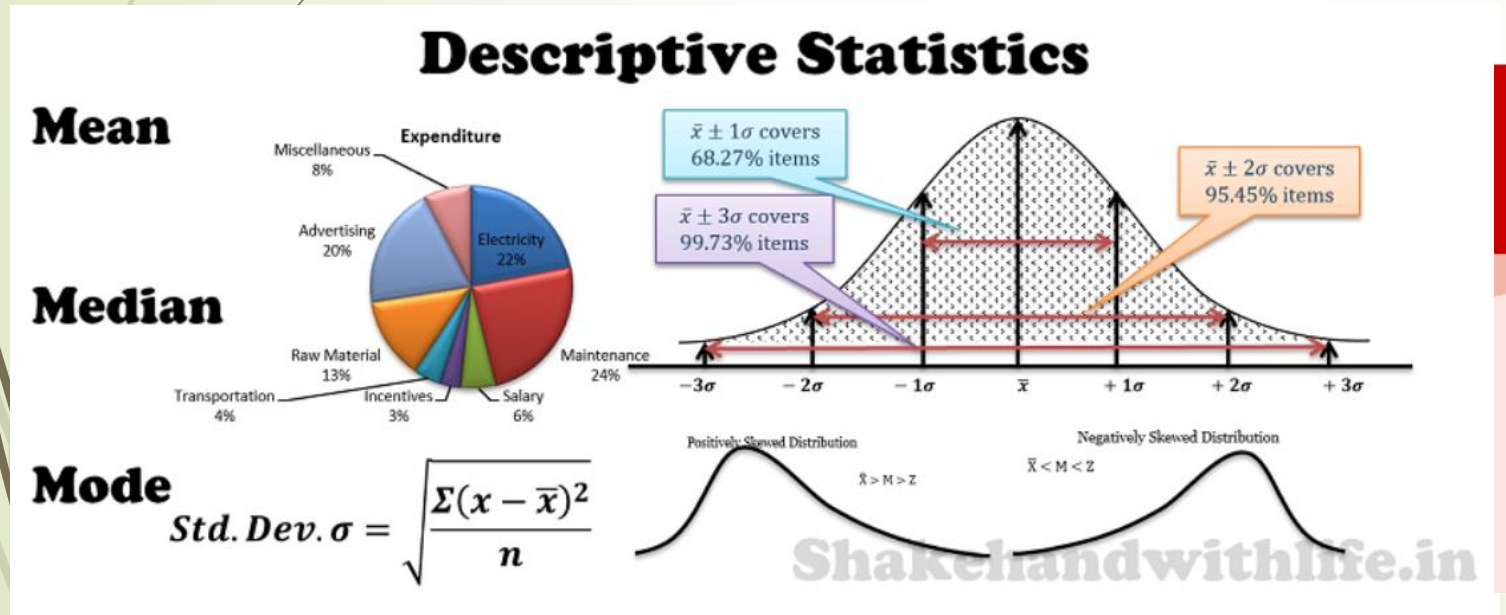
http://alldatascience.com/what-is-data-science

Statistics

- สถิติ คือ กระบวนการในการดำเนินการกับข้อมูลอย่างเป็นระบบ เพื่อให้ได้สารสนเทศประกอบการตัดสินใจภายใต้สภาวการณ์ที่ไม่แน่นอน และขจัดความอคติส่วนตัวออกจากการตัดสินใจ ซึ่งถือเป็นศาสตร์ชนิดหนึ่งที่เป็นทั้งวิทยาศาสตร์และศิลปศาสตร์ที่ว่าด้วย การเก็บรวบรวมข้อมูล การนำเสนอข้อมูล การวิเคราะห์ข้อมูล และการตีความหมาย
  - สถิติพรรณนา **(Descriptive Statistics)**
  - สถิติอ้างอิง **(Inferential Statistics)**

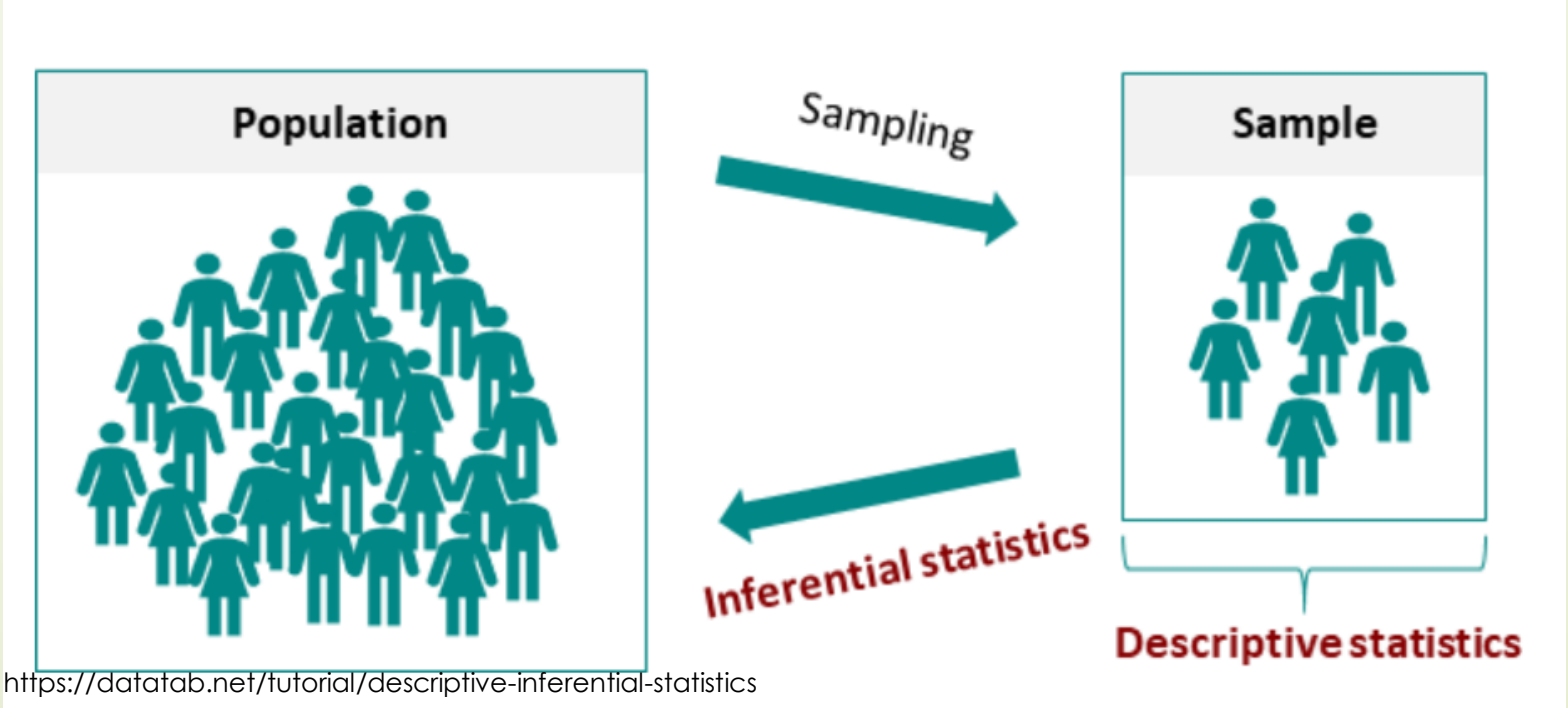
Descriptive Statistics

- สถิติพรรณนา เป็นการนำเสนอข้อมูลเพื่อสะดวกในการทำความเข้าใจในเรื่องใดเรื่องหนึ่งที่สนใจ ซึ่งอาจบรรยายข้อมูลในลักษณะกราฟ ตาราง แผนภาพ หรือค่าสถิติพื้นฐาน ได้แก่ ค่าความถี่ ค่าร้อยละ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน เป็นต้น



Inferential Statistics

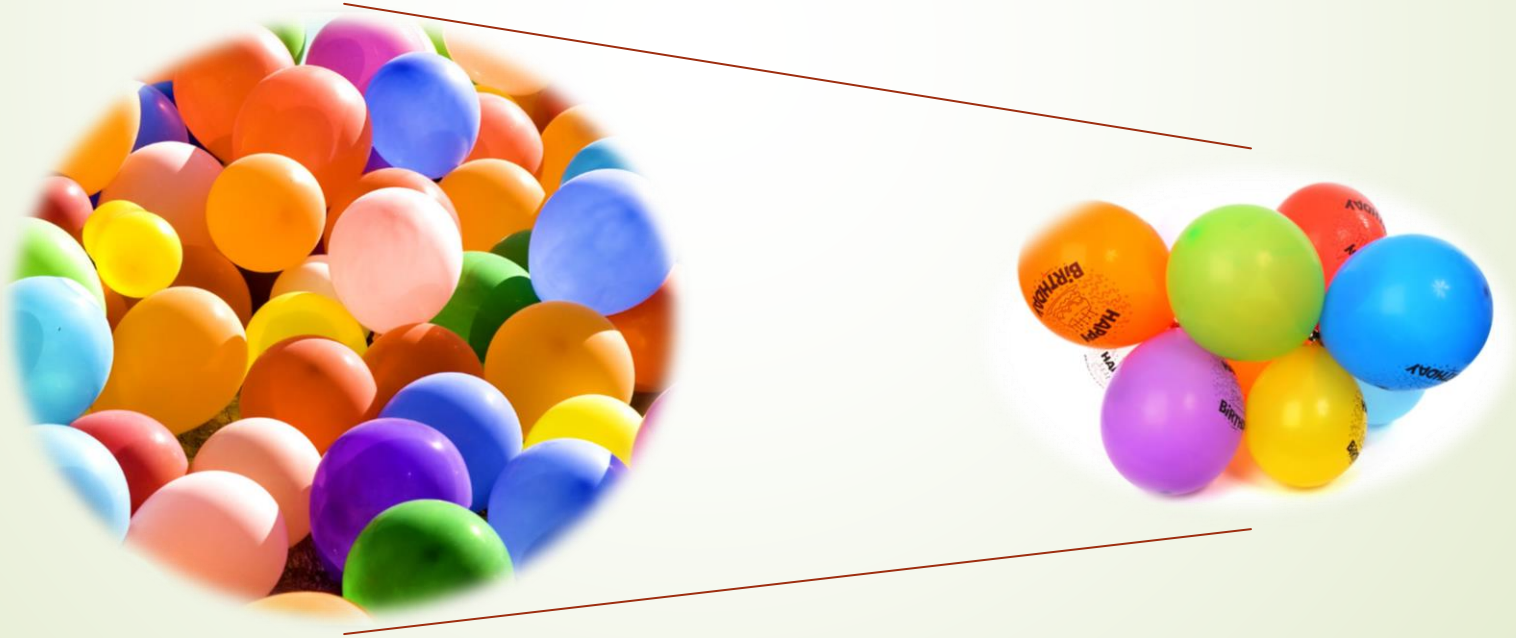
- สถิติอ้างอิง เป็นกระบวนการสรุปอ้างอิงค่าสถิติที่ได้จากกลุ่มตัวอย่างไปสู่ค่าพารามิเตอร์ของประชากรโดยอาศัยทฤษฎีความน่าจะเป็น ซึ่งประกอบด้วย เทคนิคการประมาณค่า และการทดสอบสมมติฐาน



https://datatab.net/tutorial/descriptive-inferential-statistics

Population & Sample

- ประชากร **(Population)** หน่วยทุกหน่วยที่สนใจศึกษา (ข้อมูลทั้งหมด)
- ตัวอย่าง **(Sample)** บางหน่วยที่สนใจศึกษา หรือ บางส่วนของประชากร (ข้อมูลบางส่วน)



Population **Sample**

Type of DATA

- ข้อมูลโดยทั่วไปแบ่งออกเป็น 2 ประเภท
  - ข้อมูลแบบมีโครงสร้าง **(Structured Data)** เป็นข้อมูลที่มีโครงสร้างแน่นอน มีการกำหนดตัวแปรภายในของข้อมูลอย่างชัดเจน และอยู่ในรูปแบบที่สามารถนำมาวิเคราะห์ได้โดยง่าย เช่น ข้อมูลที่อยู่ในรูประเบียน **(Record)** ตารางข้อมูล รายการข้อมูล **(Transactional Data)** เป็นต้น
  - ข้อมูลแบบไม่มีโครงสร้างแน่นอน **(Unstructured Data)** เป็นข้อมูลที่ไม่มีการกำหนดตัวแปรภายในของข้อมูลอย่างชัดเจน เช่น ข้อมูลในรูปข้อความ **(Text)** ข้อมูลภาพ เสียง วิดีโอ กราฟ เป็นต้น
- โดยส่วนใหญ่ข้อมูลที่นำมาทำการวิเคราะห์ส่วนใหญ่มักจะเป็นข้อมูลในรูปแบบของ***ข้อมูลแบบมีโครงสร้าง***

Type of DATA

- **ข้อมูลเชิงปริมาณ (Quantitative data)** เป็นข้อมูลที่สามารถวัดค่ามากหรือน้อยได้ บอกปริมาณความแตกต่างได้ โดยที่แสดงข้อมูลในรูปของตัวเลข เช่น อายุ น้ำหนัก ส่วนสูง รายได้ เป็นต้น ซึ่งข้อมูลเชิงปริมาณยังสามารถแบ่งออกได้เป็น 2 ลักษณะ ดังนี้
  - ข้อมูลเชิงปริมาณแบบต่อเนื่อง (Continues data) เป็นข้อมูลที่เป็นจำนวนจริงที่มีค่าได้ทุกค่าในช่วงที่กำหนด เช่น น้ำหนักของสินค้า ส่วนสูง อายุ คะแนนสอบ ความยาวของเส้นผม เป็นต้น
  - ข้อมูลเชิงปริมาณแบบไม่ต่อเนื่อง (Discrete data) เป็นข้อมูลที่ให้ค่าเป็นจำนวนเต็มแบบนับได้ เช่น จำนวนร้านที่ถ่ายเอกสารที่บริเวณสถานศึกษา จำนวนนักศึกษาหญิง จำนวนคนที่มีอายุมากกว่า 50 ปี เป็นต้น
- **ข้อมูลเชิงคุณภาพ (Qualitative data)** เป็นข้อมูลเชิงกลุ่มหรือจำแนกประเภท อาจจะเป็นตัวเลขหรือตัวอักษรก็ได้ โดยที่ถ้าเป็นตัวเลขจะไม่สามารถระบุค่าระดับมากน้อยหรือให้ความหมายในทางปริมาณได้ มักจะเป็นข้อความ เช่น เพศ อาชีพ ศาสนา สีผิว คุณภาพของสินค้า ความพึงพอใจ เป็นต้น

Discrete and Continuous Data

Discrete data can only take on certain individual values.

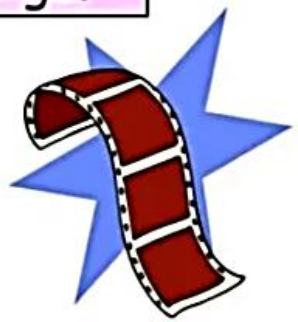
Continuous data can take on any value in a certain range.

Example 1

Number of pages in a book is a **discrete variable.**

Example 2

Length of a film is a **continuous variable.**



Example 3

Shoe size is a **Discrete variable.** E.g. 5, $5\frac{1}{2}$, 6, $6\frac{1}{2}$ etc. Not in between.

Example 4

Temperature is a **continuous variable.**

Example 5

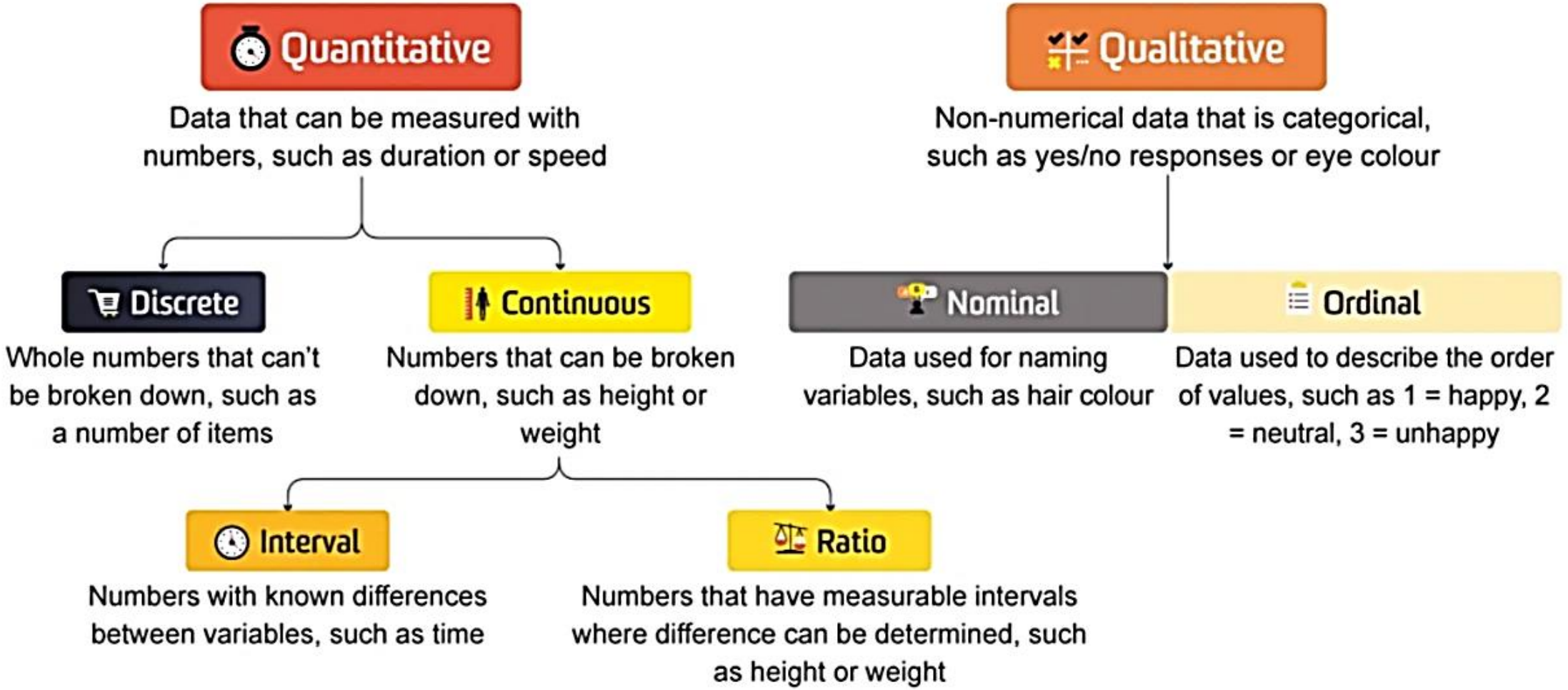
Number of people in a race is a **discrete variable.**

Example 6

Time taken to run a race is a **continuous variable**

https://medium.com/@dharmanathpatil/types-of-random-variables-1f7e17a4e3c7

Types of Data



ค่าสถิติพื้นฐานที่ใช้ใน data science

- ค่า Max Min
- ค่าวัดแนวโน้มเข้าสู่ส่วนกลาง Measures of Central Tendency
- ค่าวัดการกระจายของข้อมูล Measure of Dispersion
- ค่าผิดปกติ Outliers
- ค่าความน่าจะเป็น Probability

Descriptive Statistics

- ค่า **Max** คือ ค่าข้อมูลที่มีค่าสูงที่สุด
- ค่า **Min** คือ ค่าข้อมูลที่มีค่าน้อยที่สุด

- **EX: 9 7.8 5 10 12 7 8 7 5 9 13 27 8.5 7 15**

Measures of Central Tendency

- **Measures of Central Tendency** ค่ากลางที่ใช้ในการอธิบายข้อมูล โดยค่าที่นิยมใช้คือ ค่าเฉลี่ย มัธยฐาน และฐานนิยม
  - ค่าเฉลี่ย **(Mean)** เป็นค่าที่นิยมใช้มากสุด สามารถคำนวณได้จากผลรวมของข้อมูลหารด้วยจำนวนข้อมูล

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X}{n} \quad \text{หรือ} \quad \mu = \frac{X_1 + X_2 + \cdots + X_n}{N} = \frac{\sum X}{N}$$

  - มัธยฐาน **(Median)** เป็นค่าที่อยู่ตำแหน่งตรงกลางของข้อมูลเมื่อเรียงจากน้อยไปหามาก

    ถ้าเป็นข้อมูลจำนวนคี่ ตำแหน่งตรงกลางหาจาก $\frac{n+1}{2}$ แต่ถ้าข้อมูลจำนวนคู่นำค่าข้อมูลสองตำแหน่ง คือ $\frac{n}{2}$ และ $\frac{n}{2}+1$ มาหาค่าเฉลี่ย

  - ฐานนิยม **(Mode)** เป็นค่าข้อมูลที่เกิดบ่อยที่สุด หรือมีความถี่ หรือซ้ำสูงที่สุด

Measures of Central Tendency

(ตัวอย่าง)

จากการการประกาศขายบ้านได้สุ่มเก็บข้อมูลราคาบ้านที่ประกาศขายในเว็บแห่งหนึ่ง (หน่วย : แสนบาท) ได้ข้อมูลดังนี้

9 7.8 5 10 12 7 8 7 5 9 13 27 8.5 7 15

จากข้อมูลข้างต้นสามารถคำนวณค่าเฉลี่ย มัธยฐาน และฐานนิยมได้ดังนี้

ค่าเฉลี่ย $\bar{X} = \frac{9 + 7.8 + \cdots + 15}{15} = \frac{150.3}{15} = 10.02$

มัธยฐาน 8.5 (เรียงข้อมูล 5 5 7 7 7 7.8 8 8.5 9 9 10 12 13 15 27)

ฐานนิยม 7

Measures of Central Tendency

- ข้อเสียของค่าเฉลี่ย **(Mean)** จะอ่อนไหวกับค่าที่เป็น **outlier** หรือค่าที่สูงหรือน้อยกว่าปกติ แต่ค่ามัธยฐาน **(Median)** จะไม่อ่อนไหวกับค่า **Outlier** ตัวอย่างเช่น

บ้านที่	ราคาขาย
1	900,000
2	750,000
3	12,000,000
4	1,000,000
5	990,000
6	850,000
7	950,000

- ค่าเฉลี่ย **(Mean)**

$$\bar{X} = \frac{900000 + 750000 + \cdots + 950000}{7}$$

$$= \frac{17440000}{7} = 2,491,428.57$$

- ค่ามัธยฐาน **(Median)** เท่ากับ **950,000**

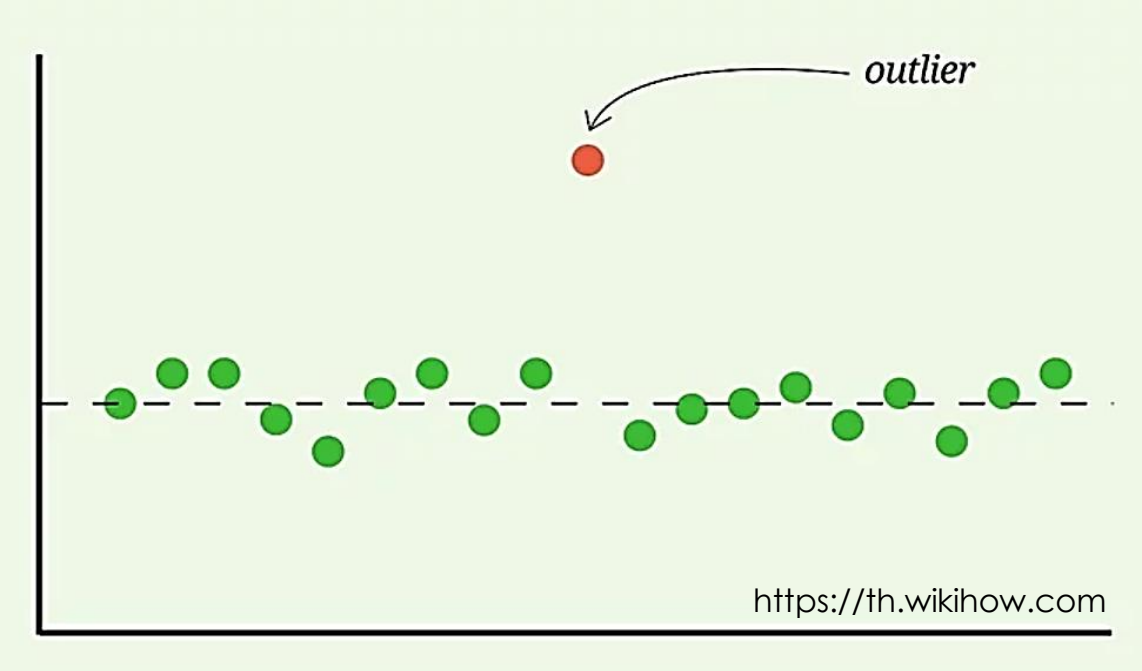
(เรียงข้อมูล >> 750,000 850,000 900,000 950,000 990,000 1,000,000 12,000,000)

Measure of Dispersion

- ค่าวัดการกระจาย **(Measure of Dispersion)** เป็นค่าที่ใช้อธิบายลักษณะการกระจายของข้อมูล ถ้าข้อข้อมูลชุดนั้น ๆ มีค่าต่างกันมาก จะหมายถึง ข้อมูลมีการกระจายมาก ถ้าข้อข้อมูลชุดนั้น ๆ มีค่าต่างกันน้อย จะหมายถึง ข้อมูลมีการกระจายน้อย ถ้าข้อมูลชุดนั้น ๆ มีค่าเท่ากันทุกค่า จะหมายถึง ข้อมูลไม่มีการกระจาย ซึ่งค่าวัดการกระจายมักนิยมใช้คู่กับค่าวัดแนวโน้มเข้าสู่ส่วนกลางเพื่อเป็นการอธิบายลักษณะของข้อมูลได้ชัดเจนมากยิ่งขึ้น *โดยวิธีที่นิยมใช้*
  - ค่าพิสัย **(Range)** คือ ค่าผลต่างระหว่างข้อมูลสูงสุดกับข้อมูลต่ำสุด
  - ค่าเบี่ยงเบนมาตรฐาน **(Standard deviation: S.D.)** รากที่สองของค่าความแปรปรวน **(Variance)**

Outliers

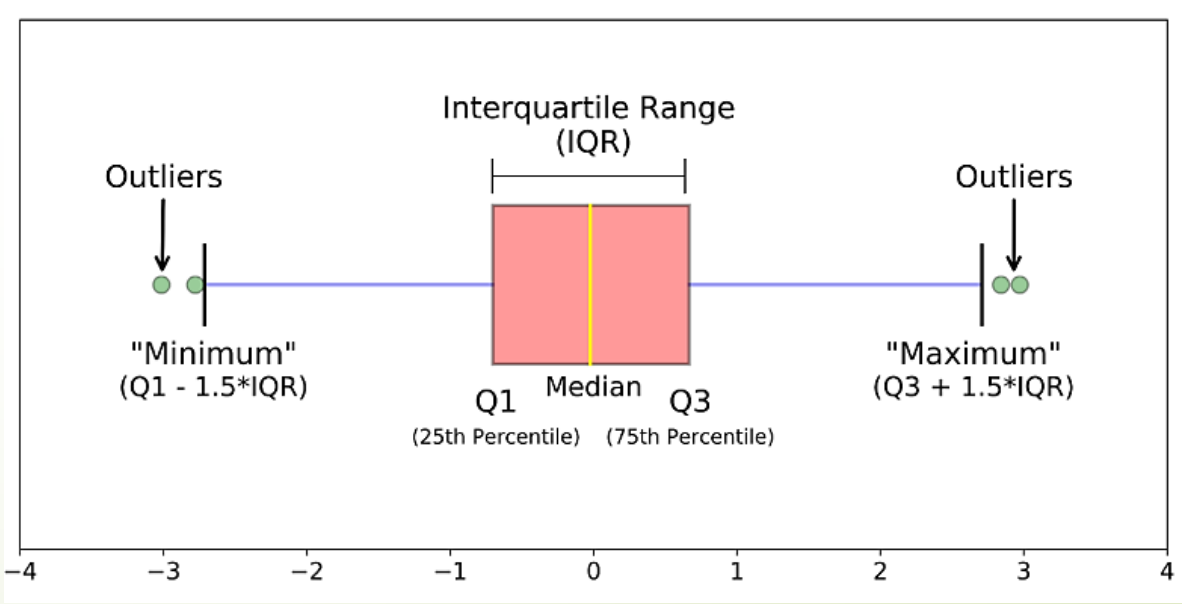
- ค่าผิดปกติ **(Outliers)** คือ ข้อมูลที่มีค่าแตกต่างทั้งมากกว่า และน้อยกว่าจากข้อมูลในชุดเดียวกันมาก ผิดปกติ จนกระทั่งทำให้สงสัยว่าเป็นข้อมูลที่ไม่อยู่ในกลุ่มเดียวกัน เป็นสาเหตุให้ผลการวัดที่ใช้เป็นตัวแทนของกลุ่มคลาดเคลื่อนไป



(จันทรัตน์ วรสรรพวิทย์, 2554)

Outliers

- ค่าผิดปกติ **(Outliers)** ในการตรวจจับข้อมูลที่ผิดปกติหรือข้อมูลที่เป็น **outliers** ในวิธีทางสถิติ หนึ่งในวิธีที่นิยมใช้คือ **“Interquartile Range (IQR)”** หรือช่วงระหว่างควอไทล์ ซึ่งเป็นวิธีที่ใช้ค่าสถิติเพื่อหาข้อมูลที่กระจายตัวหรือแตกต่างจากค่ากลางของข้อมูลโดยมีเกณฑ์ในการวัดความผิดปกติ พิจารณาจาก
  - ข้อมูลที่มีค่ามากกว่า $Q_3 + (1.5 \times IQR)$
  - ข้อมูลที่มีค่าน้อยกว่า $Q_1 - (1.5 \times IQR)$



https://medium.com/@agarwal.vishal819/outlier-detection-with-boxplots-1b6757fafa21

EX

- จากข้อมูลต่อไปนี้ 3 6 8 9 12 13 38 จงพิจารณาว่ามีค่าข้อมูลผิดปกติหรือไม่

จากข้อมูลพบว่า Q_1 = 7 Q_3 = 12.5 IQR = 5.5

Probability

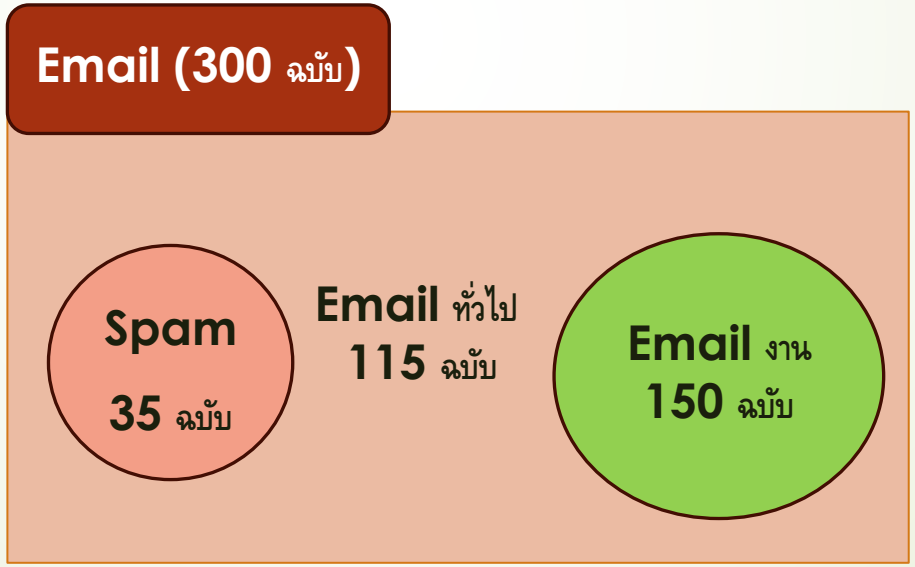
- ความน่าจะเป็น **(Probability)** เป็นค่าที่แสดงถึงโอกาสที่จะเกิดเหตุการณ์ต่าง ๆ โดยมีค่าอยู่ระหว่าง **0** ถึง **1** ซึ่งเขียนสัญลักษณ์แทน ความน่าจะเป็น คือ **P** หรือ **P(A)** ซึ่งหมายถึงความน่าจะเป็นที่จะเกิดเหตุการณ์ **A** (ในที่นี่อาจใช้ตัวอักษรอื่น ๆ แทน **A** ได้)

$$P(A) = \frac{\text{จำนวนเหตุการณ์ที่สนใจ}}{\text{จำนวนเหตุการณ์ทั้งหมด}} = \frac{n(A)}{n(S)}$$

- **Probability** ถูกนำไปประยุกต์ใช้ในการทำนายผลของ **Machine Learning** ได้หลายรูปแบบ เช่น การทำนายความน่าจะเป็นของคนสูบบุหรี่แล้วจะเป็นโรคมะเร็งปอด หรือความน่าจะเป็นที่จะเกิดอุบัติเหตุทางรถยนต์ของคนคุยโทรศัพท์ระหว่างขับรถ เป็นต้น

Probability

- โอกาสที่ Email จะเป็น spam .....
- โอกาสที่จะเป็น Email งาน .....

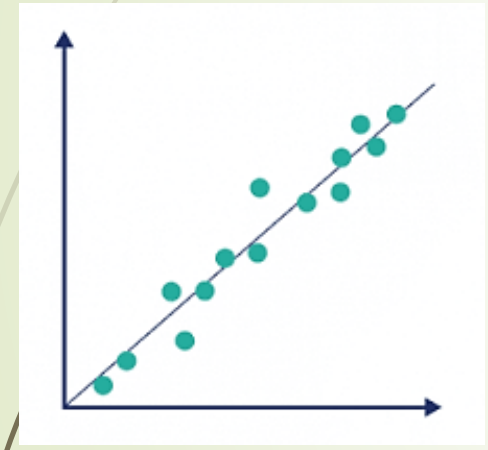


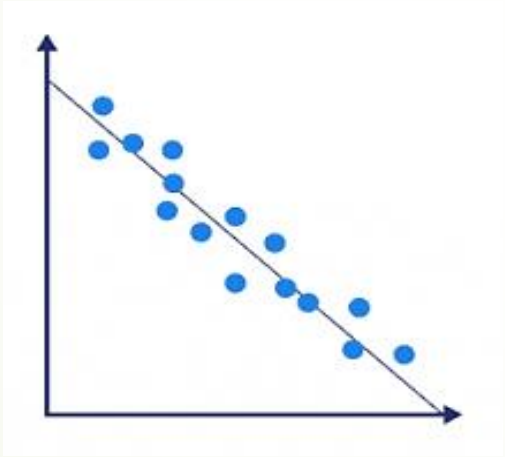
Linear Regression

- การวิเคราะห์การถดถอยเชิงเส้น หรือ Linear Regression เป็นหนึ่งในโมเดลการเรียนรู้ของเครื่อง (Machine Learning) ที่จัดอยู่ในลักษณะของการเรียนรู้แบบมีผู้สอน (Supervised Learning) ถือว่าเป็นโมเดลที่นิยมนำมาใช้ ซึ่งเป็นโมเดลที่ถูกนำมาใช้สำหรับการทำนายผลข้อมูลที่มีความสัมพันธ์กันแบบเส้นตรง
- ลักษณะที่สำคัญของ Linear Regression
  - ข้อมูลตัวอย่างที่นำมา Train Model ชนิดนี้ จะต้องมีคอลัมน์ที่เป็นผลลัพธ์ เช่น การทำนายแคลอรีที่ถูกเผาผลาญจากระยะเวลาในการออกกำลังกาย เมื่อนำข้อมูลตัวอย่างมา Train Model จะต้องประกอบด้วย ระยะเวลาในการออกกำลังกาย และ จำนวนแคลอรีที่ถูกเผาผลาญ (ผลลัพธ์)
  - ผลลัพธ์เป็นตัวเลขที่มีค่าแตกต่างกัน จึงไม่สามารถจำแนกประเภทแบบกลุ่ม (Classification) ได้ หรือกล่าวได้ว่าเป็นการทำนายผลแบบ Regression (เป็นตัวเลขที่มีค่าไม่แน่นอน)
  - แบ่งตามลักษณะของข้อมูลตัวอย่างที่นำมา Train Model หรือตามลักษณะผลลัพธ์ที่ได้ โดยจะกล่าวถึง 2 เรื่อง ได้แก่ Simple Linear Regression และ Multiple Linear Regression

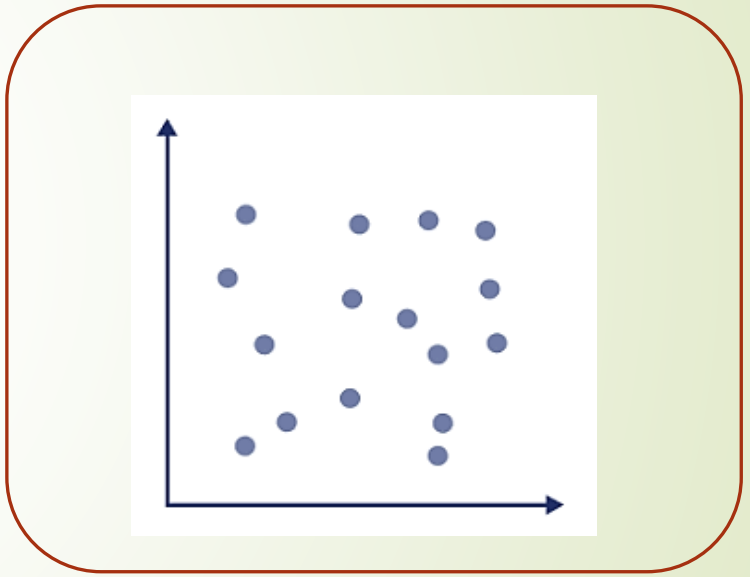
Linear Regression

- การพิจารณาเลือกใช้โมเดล Linear Regression





สามารถทำนายด้วยโมเดล Linear Regression ได้



ลักษณะเช่นนี้อาจไม่ควรทำนายด้วยโมเดล Linear Regression เพราะอาจขาดความแม่นยำได้

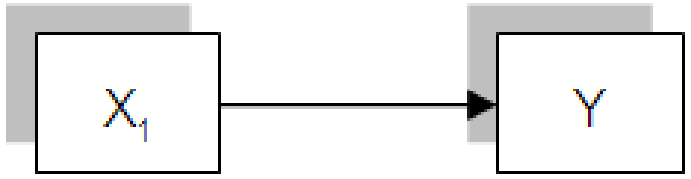
Linear Regression

- **Simple Linear Regression** หรือ การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย ใช้ในการทำนายผลสำหรับกรณีที่ข้อมูลตัวอย่างประกอบด้วยข้อมูลที่เป็น **Feature** หรือตัวแปรอิสระ (**X : Independence Variable**) เพียงตัวเดียว และข้อมูลที่เป็น **Target** หรือผลลัพธ์ หรือตัวแปรตาม (**Y : Dependence Variable**) มีค่าเป็นตัวเลขเท่านั้น โดยสามารถเขียนสมการถดถอยดังกล่าว ได้ดังต่อไปนี้

$$\hat{y} = a + bx$$

- **Multiple Linear Regression** จะมี Feature หรือตัวแปรอิสระ (X) มากกว่า 1 ตัวแปร (x_1 , x_2 , ... , x_n) ส่วนตัวแปร Target หรือผลลัพธ์ หรือตัวแปรตาม (Y) มี 1 ตัวแปรเท่านั้น ดังนั้นสมการถดถอยจึงเป็นดังนี้

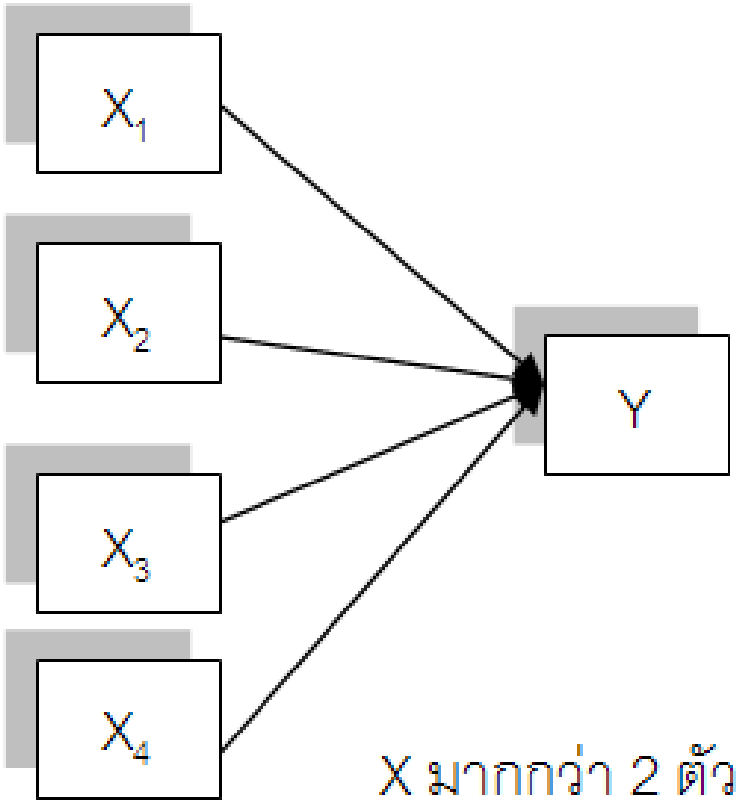
$$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$



X 1 ตัว Y 1 ตัว

การวิเคราะห์การถดถอยอย่างง่าย

(Simple regression)



X มากกว่า 2 ตัว Y 1 ตัว

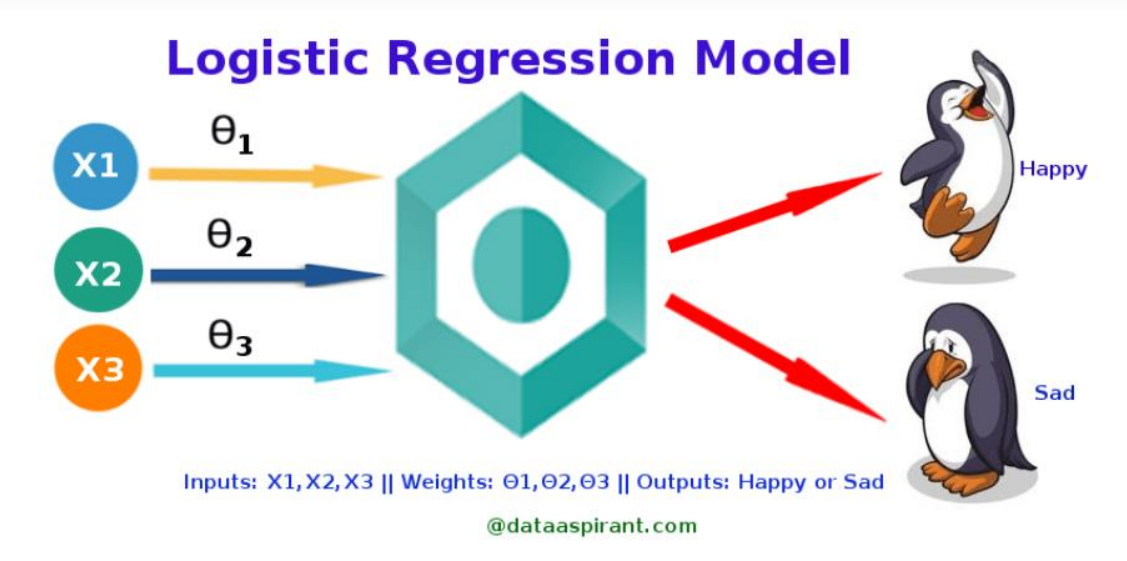
การวิเคราะห์การถดถอยพหุคูณ (Multiple regression Analysis)

Linear Regression

- ตัวอย่าง หากต้องการพยากรณ์ยอดขายสินค้า

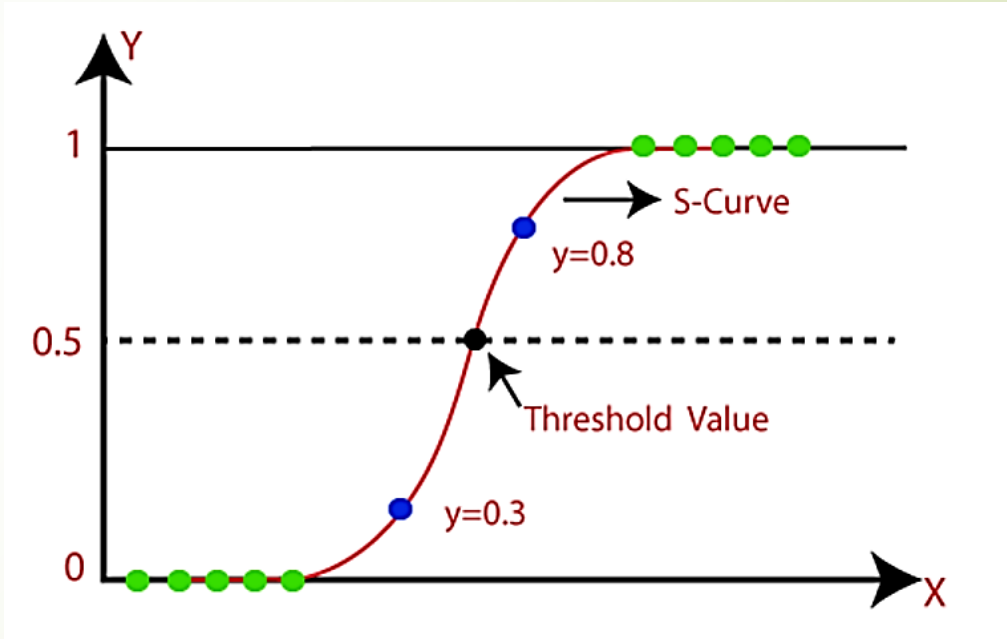
Logistic Regression in Machine Learning

- Logistic Regression คือโมเดลที่ต่อยอดมาจากสมการ Linear Equation เนื่องจากสมการเส้นตรงหรือ Linear Regression ไม่สามารถแก้ไขปัญหาบางอย่างได้
- Logistic Regression is one of **the most popular Machine Learning algorithms**, which comes under the **Supervised Learning technique**. It is used for predicting the **categorical dependent** variable using a given set of independent variables.



Logistic Regression in Machine Learning

- Logistic Regression predicts the output of a **categorical dependent variable**. Therefore the outcome must be a categorical or discrete value. It can be either **Yes or No, 0 or 1, true or False**, etc.
- Logistic Regression เป็นอัลกอริทึมสำหรับการจำแนกประเภท (Classification Algorithm) ที่ใช้สำหรับการทำนายความน่าจะเป็นของผลลัพธ์เป็นค่า discrete ซึ่งจะมีค่าอยู่ในช่วง 0 ถึง 1
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probability and classify new data using continuous and discrete datasets.



Logistic Regression in Machine Learning

- เป้าหมายในการวิเคราะห์การถดถอยโลจิสติค เพื่อทำนายโอกาส (ความน่าจะเป็น) ที่จะเกิดเหตุการณ์ที่สนใจ โดยสามารถแบ่งการวิเคราะห์การถดถอยโลจิสติคได้เป็น 2 ประเภท
  - การวิเคราะห์การถดถอยโลจิสติคทวิ หรือ Binary Logistic Regression เป็นลักษณะที่ผลลัพธ์หรือตัวแปรตามสามารถจำแนกออกได้ 2 กลุ่ม ได้แก่ เกิดเหตุการณ์ (y = 1) หรือไม่เกิดเหตุการณ์ (y = 0) เช่น การทำนายผลการสอบ (สอบผ่าน หรือ สอบตก) การพยากรณ์ฝนตก (ฝนตก หรือ ฝนไม่ตก) เป็นต้น
  - การวิเคราะห์การถดถอยโลจิสติคพหุ หรือ Multinomial Logistic Regression เป็นลักษณะที่ผลลัพธ์หรือตัวแปรตามสามารถจำแนกออกได้ได้มากกว่า 2 กลุ่ม เช่น การพยากรณ์สภาพอากาศดี (y = 3) ปานกลาง (y = 2) แย่ (y = 1) เป็นต้น
- ถ้าความน่าจะเป็นอยู่ในช่วง 0 – 0.49 หมายความว่า ไม่อยู่ในกลุ่มนั้นๆ
- ถ้าความน่าจะเป็นอยู่ในช่วง 0.5 – 1 หมายความว่า อยู่ในกลุ่มนั้นๆ