



# Intermediate Statistic

# Table of Content

## What will We Learn Today?

1. **Probabilitas**
2. **Korelasi dan sebab akibat**
3. **Statistical Plot**
  - **Histogram**
  - **Scatter plot**
  - **Boxplot**





# Probabilitas

*mengukur seberapa besar kemungkinan suatu peristiwa terjadi pada **skala 0 (tidak pernah terjadi) hingga 1 (selalu terjadi)***





# Probability in Data Science

## Probability

Probability Mass Function  
(PMF)

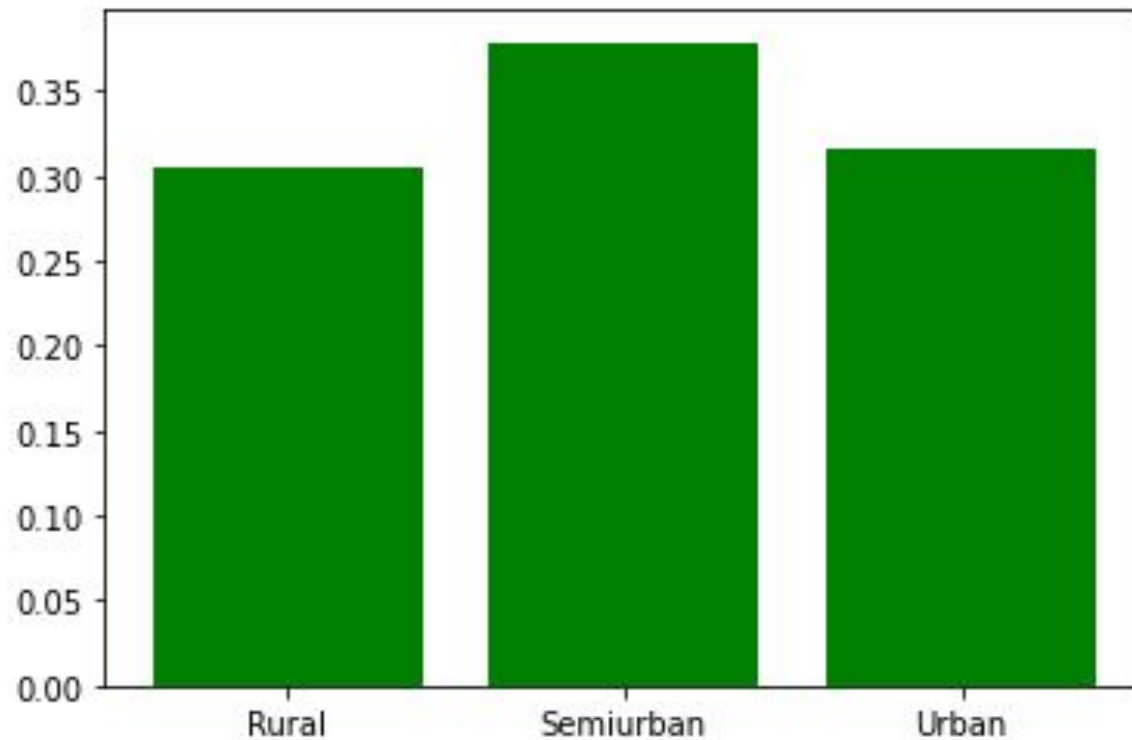
Cumulative Distribution  
Function (CDF)





# Probability Mass Function

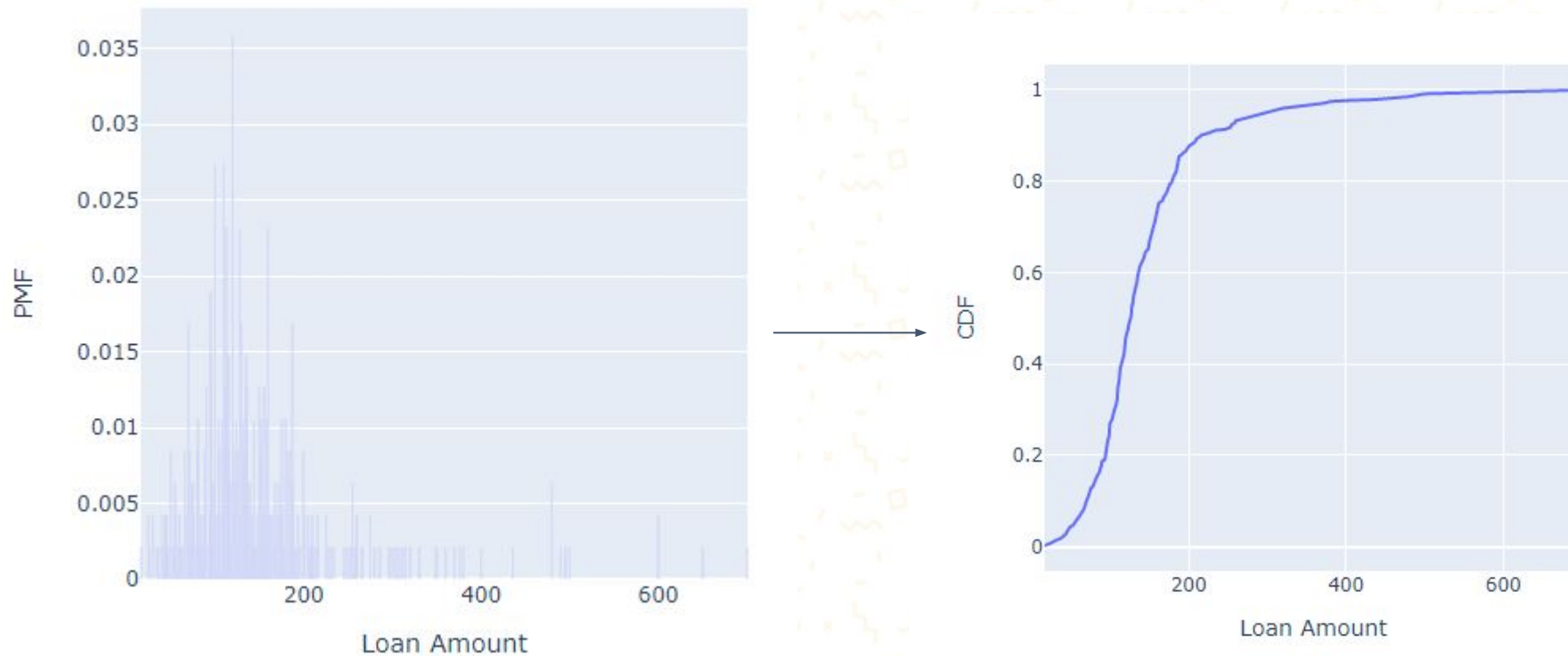
adalah frekuensi yang dinyatakan sebagai sebuah pecahan dari suatu sampel. Untuk mendapatkan probabilitas, kita membagi datanya sebanyak  $n$ , disebut **normalisasi**





# Cumulative Distribution Function

PMF bekerja dengan baik jika memiliki **variasi data yang sedikit**. Jika variasi data meningkat, probabilitas yang terkait dengan setiap nilai yang kecil akan mengalami peningkatan **random noise**





# Distribusi

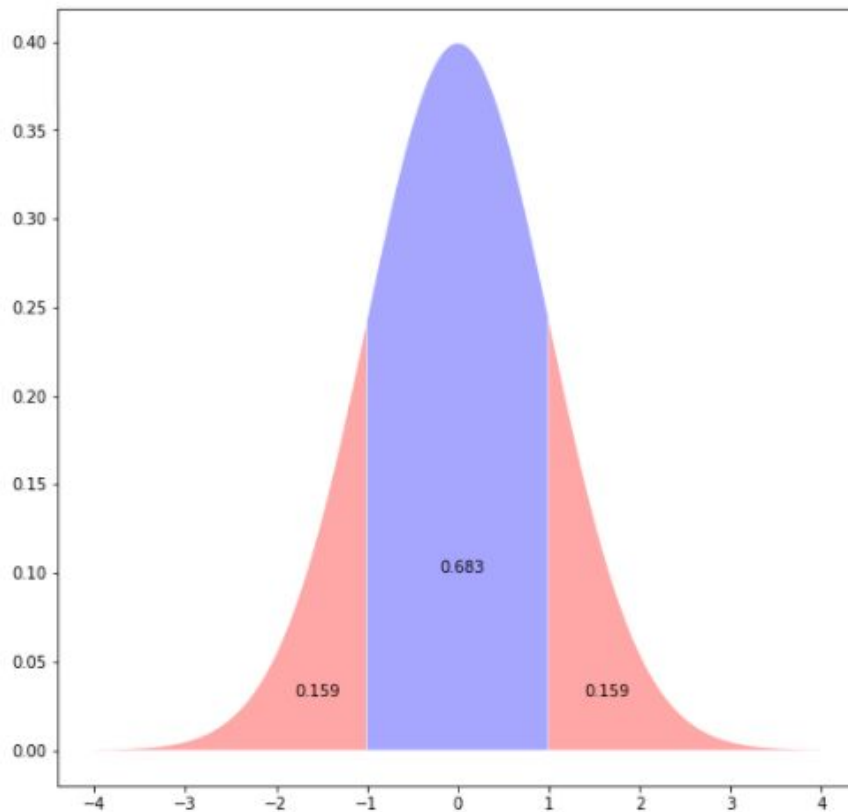
*dideskripsikan bagaimana suatu variabel tersebar secara spesifik yang **nilai mana yang paling mungkin diambil***





# Normal Distribution

adalah distribusi probabilitas kontinu yang dicirikan oleh kurva berbentuk lonceng simetris (bell-shaped curve). Terlebih lagi, distribusi normal memiliki karakter dengan di pusatnya sebagai nilai rata-rata dan penyebarannya standar deviasi

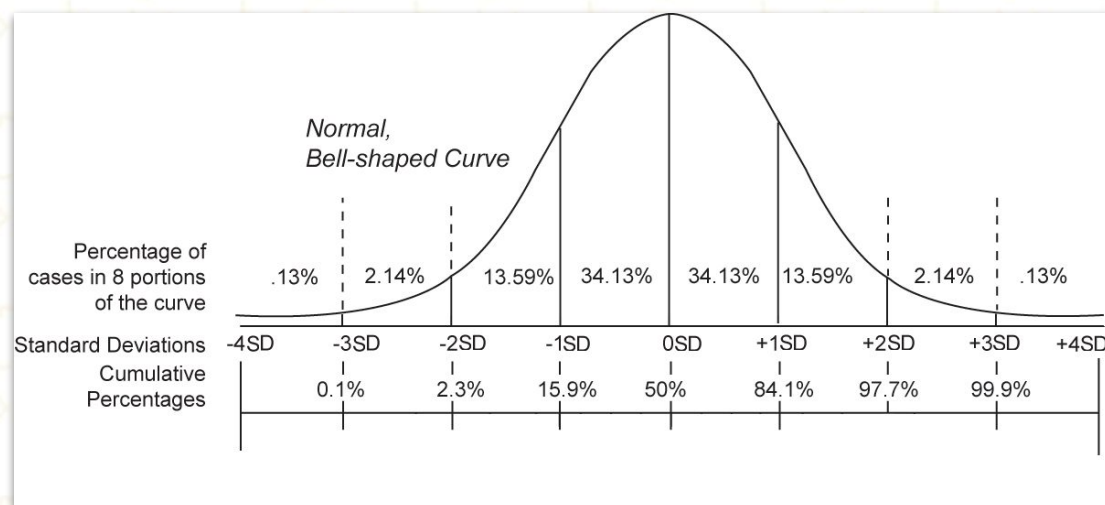






# Karakteristik dari Distribusi Normal

- Simetris jika dibagi dua dari pusatnya
- Rata-rata dan median hampir sama
- Setengah dari nilainya sama dengan yang kanan dan kiri

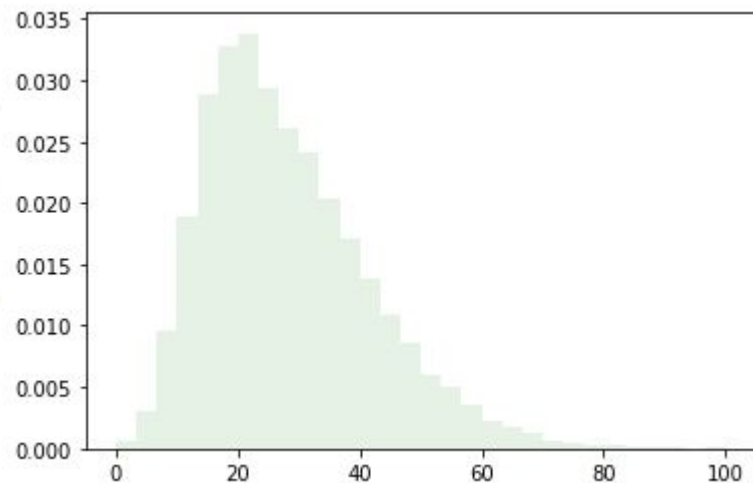




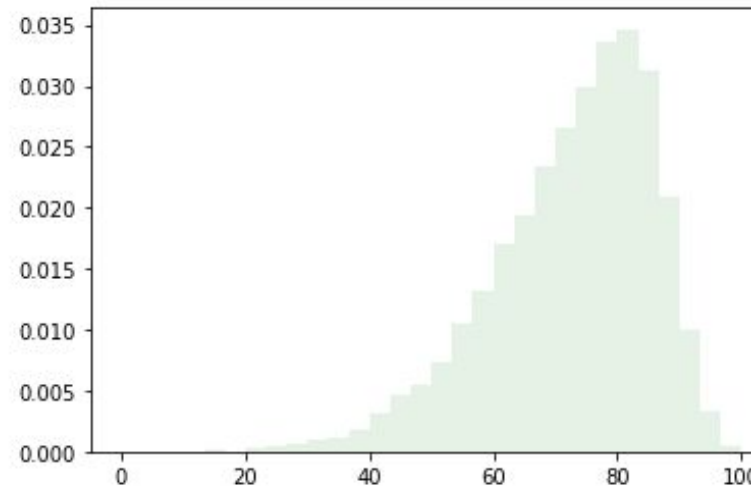
# Skewness

adalah hal yang menggambarkan bentuk dari distribusi. Ada dua hal yang harus anda ketahui:

- **Right Skewness:** ketika nilai memanjang jauh ke kanan
- **Left Skewness:** ketika nilai memanjang jauh ke kiri



**Right Skewness**



**Left Skewness**



# Korelasi dan sebab akibat



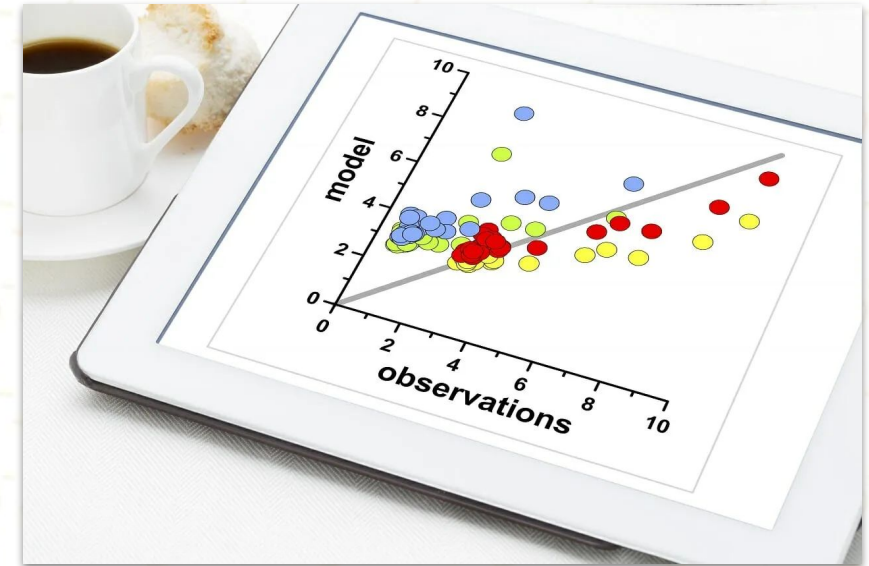


# Correlation

**Hubungan statistik** antara dua variabel between variables baik memiliki hubungan kuat ataupun lemah

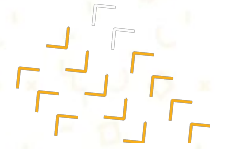
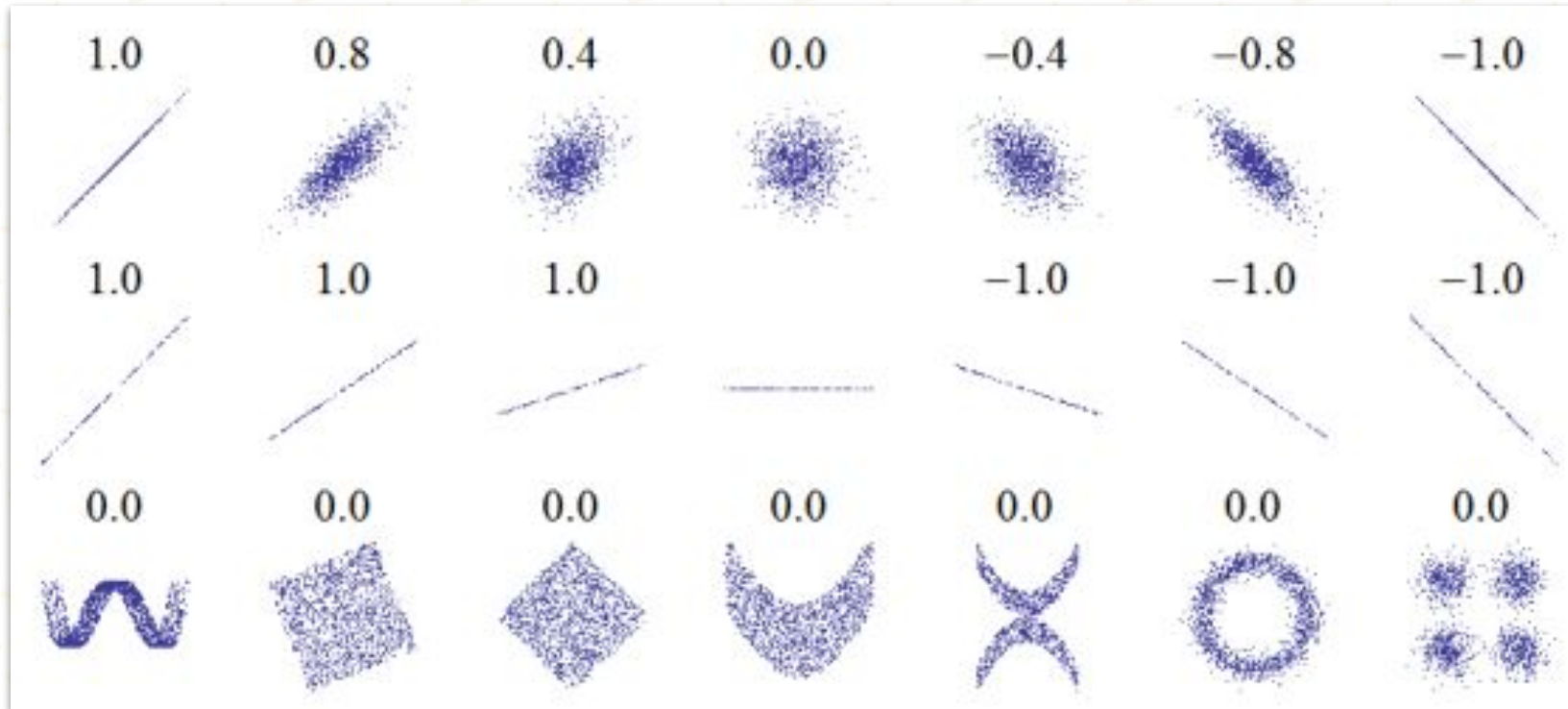
Ada tiga hal yang harus anda pahami terkait statistik:

- **Positive correlation:** kedua variabel memiliki arah yang sama
- **Neutral correlation:** tidak ada hubungan antara dua variabel
- **Negative correlation:** variabel satu dengan yang lainnya memiliki arah yang berbeda





# Beberapa contoh Korelasi



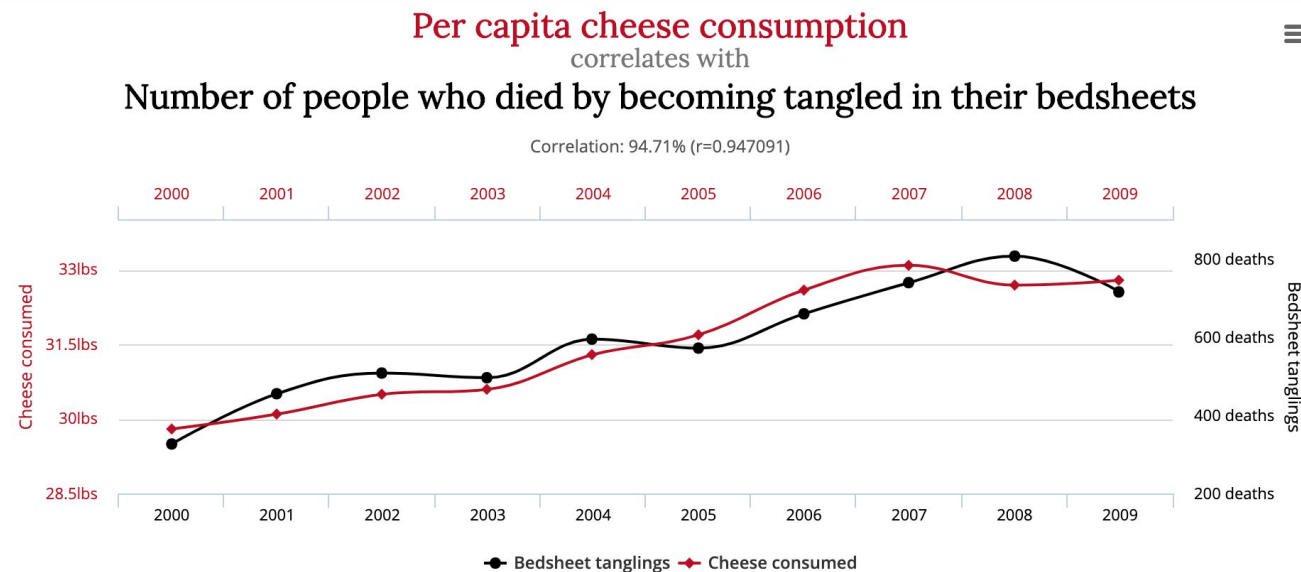


# Sebab Akibat

adalah sebuah kejadian atau proses yang berkontribusi pada kejadian yang lain

## Example:

- Setelah saya berolahraga, saya merasa lelah secara fisik.
- Memberi kucing saya makan dua kali sehari membuat kucing saya lebih gemukan



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

tylervigen.com

**Not Causation**



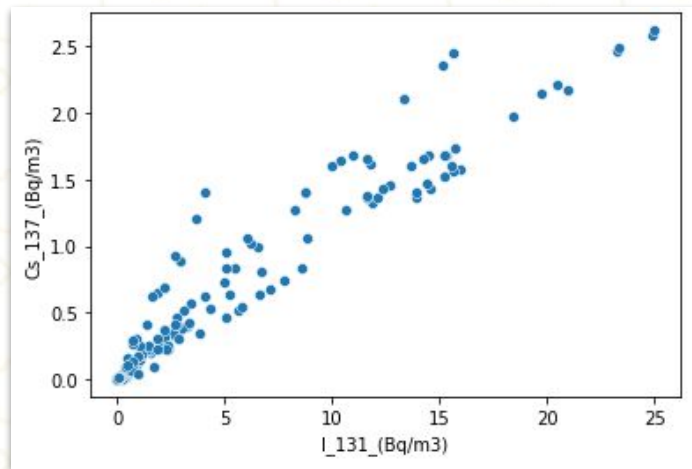
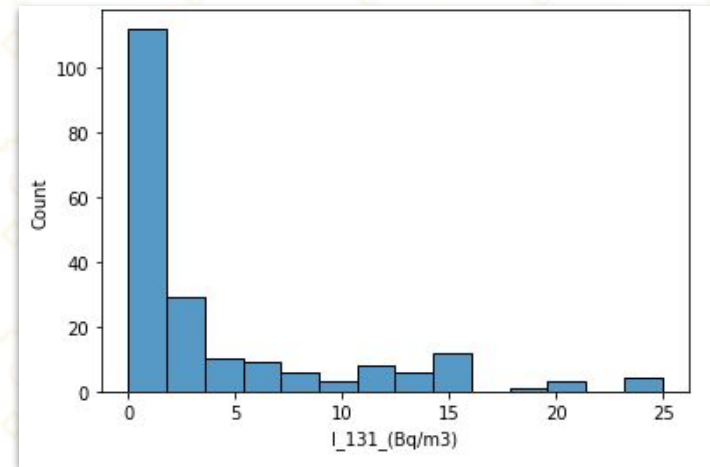
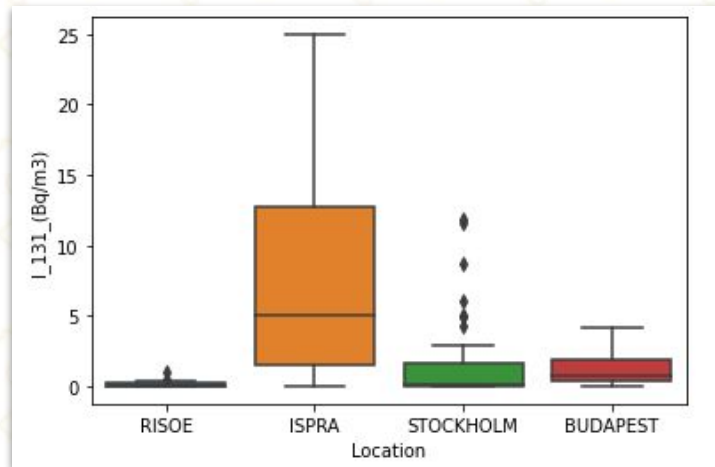
# Statistical Plot





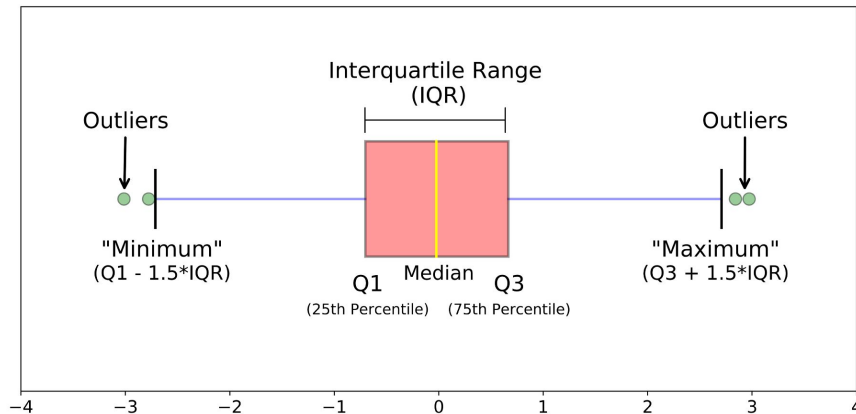
# Statistical Plots

Penyajian suatu data secara statistik dapat dipermudah dengan menampilkan data tersebut secara visualisasi statistik





# Boxplot



**Maximum:** nilai pengamatan terbesar

**Minimum:** nilai pengamatan terkecil

**First Quartile:** nilai tengah antara **minimum** dan **median**. Penjelasananya 25% data terendah ada di Q1

**Second Quartile (Median):** nilai tengah antara titik Q1 dan Q3. Penjelasananya 50% data tersebar antara titik Q1 dan titik Q3

**Third Quartile:** nilai tengah antara **maximum** dan **median**. Penjelasananya 75% data tertinggi ada di Q3

**Whisker:** is a way to show the spread between Q1 to minimum or Q3 to maximum





# Let's Hands On



# Homework

<https://www.kaggle.com/sakshigoyal7/credit-card-customers>

Pada dataset credit card customer, terdapat informasi yang anda bisa lakukan eksplorasi. Hal ini bertujuan untuk melihat data apakah memiliki pola yang nantinya dapat digunakan dalam pembuatan model. Beberapa hal yang harus dilakukan antara lain:

- Lakukan analisis PMF pada kolom Attrition\_Flag
- Pisahkan data **Existing Customer** and **Attrited customer** pada kolom Attrition\_Flag
- Lakukan analisis distribusi pada kolom Total\_Trans\_Amt dan Total\_Trans\_Ct serta tentukan skewnessnya
- Lakukan analisis hubungan/korelasi antara kolom Total\_Trans\_Amt dan Total\_Trans\_Ct dan jelaskan apakah memiliki positive, negatif atau netral korelasi



**Thank  
YOU**

