BAB II
LANDASAN TEORI

2.1 Konsep Data

2.1.1 Data

Menurut kamus internasional Cambridge data adalah kumpulan dari fakta-fakta yang berupa angka yang digunakan untuk membantu pengambilan keputusan. Menurut Joseph M. Firestone dalam buku *"Enterprise Information Portals and Knowladge Management"* data adalah kumpulan dari datum, yang mana datum adalah nilai observasi yang bisa diukur atau dihitung. Kesimpulannya, data merupakan fakta-fakta dari fenomena yang sudah ada.

2.1.2 Basis Data

Basis data adalah kumpulan data yang merepresentasikan dunia nyata didesain dan dibangun dengan tujuan tertentu [7]. Perangkat lunak yang membantu pengguna dalam mengelola dan membuat basis data disebut dengan DBMS (*Data Base Management System*) [8]. Tujuan utama dari DBMS adalah menyediakan cara untuk menyimpan dan mengambil informasi dari basis data secara lebih mudah dan efisien.

Sistem basis data didesain untuk mengelola informasi yang besar [9]. Pengelolaan data termasuk mendefinisikan struktur informasi penyimpanan dan penyediaan mekanisme untuk melakukan manipulasi informasi. Selain itu, sistem basis data juga harus bisa memastikan bahwa informasi disimpan dengan aman, baik dari *system crash* atau akses yang tidak diizinkan. Jika data memang dikelola oleh beberapa pengguna, sistem harus menghindari sebisa mungkin hasil yang anomali.

Interaksi antara pengguna dan DBMS ditentukan melalui bahasa atau sintak khusus sesuai dengan jenis DBMS yang digunakan. Pada umumnya bahasa DBMS dikelompokan ke dalam dua bentuk, yaitu DDL (*data-definition language*) dan DML (*data-manipulation language*) [8].

1. DDL (*Data Definition Language*)

DDL digunakan untuk membuat skema dan memberikan properti tambahan data, dengan kata lain DDL adalah bahasa untuk membuat tabel, membuat indeks, mengubah tabel, dan menentukan struktur penyimpanan tabel. Keluaran dari perintah DDL adalah kumpulan tabel yang disimpan dalam berkas khusus yang disebut *data dictionary*.

2. DML (*Data Manipulation Language*)

DML adalah bahasa yang memungkinkan pengguna DBMS mengakses atau melakukan manipulasi data yang diorganisasikan dengan model data yang sesuai. Tipe akses tersebut yaitu:

- Mengambil (*retrieve*) informasi dari basis data.
- Memasukan (*insert*) informasi baru ke basis data.
- Menghapus (*delete*) infromasi dari basis data.
- Modifikasi (*update*) informasi yang terdapat dari basis data.

2.1.3 Basis Data NoSQL

NoSQL diinterpretasikan dengan *"not only SQL"* (bukan SQL saja) DBMS yang tidak memakai model relasi; NoSQL tidak dibangun di atas tabel, sehingga tidak bisa memakai query SQL untuk melakukan manipulasi data. Keunggulan NoSQL dibandingkan basis data relasi ada pada skabilitas [10]. Selain skalabilitas, NoSQL memiliki keunggulan sebagai berikut [10]:

- *Scemaless data representation*: Hampir semua model NoSQL merepresentasikan *schemalees*. Administrator basis data tidak perlu memikirkan bagaimana struktur dan model basis data yang berubah setiap waktu.
- *Development time*: Query yang komplek pada basis relasi yang sudah besar, akan membuat proses penyajian data menjadi lambat (misalnya JOIN tabel pada basis data yang tersebar di banyak server). Pada NoSQL hal demikian tidak terjadi.
- *Plan ahead for scalability*: Aplikasi yang memakai model NoSQL akan sangat elastis, NoSQL dapat menangani masalah lonjakan sumber daya.

2.2 Visualisasi Data

Data yang sangat banyak, kompleks, dan tidak teratur sangat sulit untuk dimengerti sekilas pandang oleh manusia. Membaca data akan menjadi masalah ketika data secara inkremental bertambah atau berubah. Tujuan visualisasi data adalah menunjukan mana hal-hal yang penting secara jelas, melihat pola, dan menunjukan informasi dari berbagai dimensi. Visualisasi yang efektif membantu pengguna dalam melakukan analisis, mempermudah menyimpulkan data, dan membuat data yang kompleks menjadi lebih mudah untuk dimengerti [11].

Visualisasi data tidak semerta-merta mengubah data menjadi grafik visual, visualisasi data memerlukan perencanaan. Setiap jenis data memiliki teknik visualisasi berdasarkan kebutuhannya. Berdasarkan tingkat kompleksitas data, untuk menghasilkan solusi yang berharga diperlukan dari berbagai disiplin ilmu, yaitu: statistika, *data mining*, desain grafis, dan visualisasi informasi [11].

2.2.1 *Visual Encoding*

Visual encoding adalah sebuah cara memtakan data ke dalam struktur visual. Desainer membantu pembaca memahami makna data lebih cepat dengan memakai beberapa prinsip *visual encoding*. Terdapat dua tipe prinsip *visual encoding*: planar dan retina. Manusia sangat peka terhadap variabel retina [12].

Terdapat tiga tipe data dasar dalam *visual encoding*. Berikut adalah tiga tipe dasar *visual encoding* [12]:

1. Kuantitatif: Tipe data ini dibedakan berdasarkan objek yang bisa diukur. Contoh: Nilai ujian, durasi jam dalam sehari, jumlah hari dalam setahun, dll.
2. Kualitatif: Tipe data ini dibedakan berdasarkan objek yang bisa dibandingkan dan tidak memiliki ukuran yang pasti. Contoh: Baik, buruk, jelek, tidak yakin, dll.
3. *Categorical*: Tipe data ini dibedakan berdasarkan pengelompokan objek. Contoh:
    a. Buah: apel, jeruk, mangga, dll;
    b. Negara: Singapura, Indonesia, Jepang, Rusia, dll;
    c. dll.

2.2.2 Tipe-Tipe Visualisasi

Tujuan visualisasi adalah untuk membantu pemahaman manusia terhadap data dengan memaksimalkan sistem penglihatan manusia yang bisa membedakan *pattren, spot the trends,* dan identifikasi *outlier* [13]. Tantangan dari visualisasi data adalah bagaimana membuat visualisasi yang effektif, menarik, dan tepat terhadap data yang dipakai [13].

Terdapat tujuh hal yang harus dipenuhi dalam melakukan abstraksi tingkat tinggi (*high-level abstraction*), semakin banyak hal yang disembunyikan, semakin banyak juga langkah-langkah yang harus dipenuhi, langkah tersebut yaitu [14]:

1. *Overview*: Melihat gambaran dari keseluruhan data;
2. *Zoom*: Memperbesar *item* yang terlihat menarik.
3. *Filter*: Melakukan penyaringan terhadap *item* yang dirasa kurang menarik.
4. *Details-on-demand*: Pilih satu *item* dari grup tertentu dan dapat melihat detail kapan saja.
5. *Relate*: Lihat relasi dari setiap *item*.
6. *History*: Dapat mengulang kembali atau kembali ke aksi sebelumnya.
7. *Extract*: Dapat melakukan ekstraksi dari parameter yang diberikan.

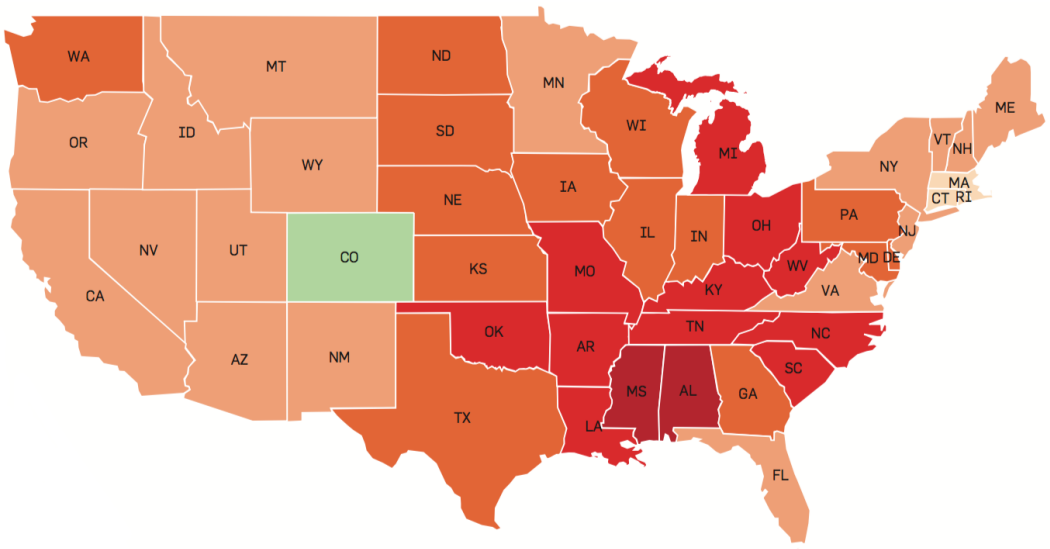
Berdasarkan taksonominya, grafik visual dibedakan menjadi [14]: 1D/Linear; 2D/Planar; 3D/Volumetric; Temporal; Multidimensional; Tree/ Hierarchical; dan Network.

2.2.2.1 1D/Linear

Grafik 1-dimensi termasuk di dalamnya adalah tipe data tekstual, kode sumber program, dan huruf alfabet. Setiap *item* yang digambarkan memiliki elemen garis . Contoh dari grafik 1D seperti kode-kode DNA, perbedaan kode sumber, dan lain-lain.

2.2.2.2 2D/Planar

Grafik 2-dimensi termasuk di dalamnya peta geografis, denah rancangan, atau *layout* koran. Setiap *item* pada grafik 2-dimensi memiliki total area dan atribut (warna, ukuran, dll).



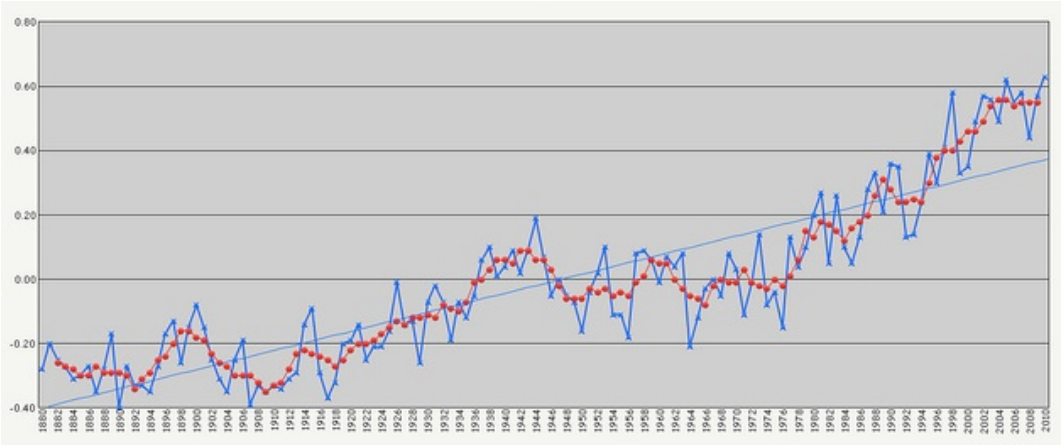
Gambar II-1 Contoh Grafik Planar/2D [13]

2.2.2.3 3D/Volumetric

Grafik 3-dimensi adalah visual yang menggambarkan objek nyata, seperti tubuh manusia, bentuk bangunan, dll. Setiap *item* pada grafik 3-dimensi memiliki volume.

2.2.2.4 Temporal

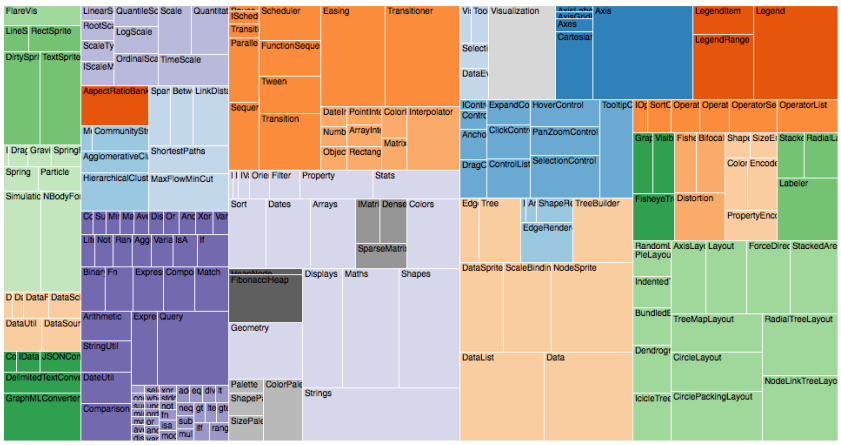
Grafik temporal adalah grafik yang berhubungan dengan waktu (*time lines*). Grafik ini menggambarkan persentasi historikal dari data 1-dimensi. Yang membedakan, grafik temporal memiliki *item* dengan waktu awal dan waktu akhir, atau periode tertentu.



Gambar II-2 Contoh Grafik Temporal [15]

2.2.2.5 Multi-dimensional

Grafik temporal didalamnya termasuk grafik-grafik yang dihasilkan dari manipulasi data dari disiplin ilmu statistika. Antarmuka representasi milti-dimensional adalah grafik 2-dimensi. Grafik multi-dimensi termasuk didalamnya grafik *pie*, histogram, *tag cloud*, *bubble cloud*, bar, *tree-map*, *scatter plot*, *bubble chart, line chart*, *step chart*, *heat-map*, *parallel sets, spider chart*, *box-plot*, *mosaic display*, *waterfall*, dan tabular.

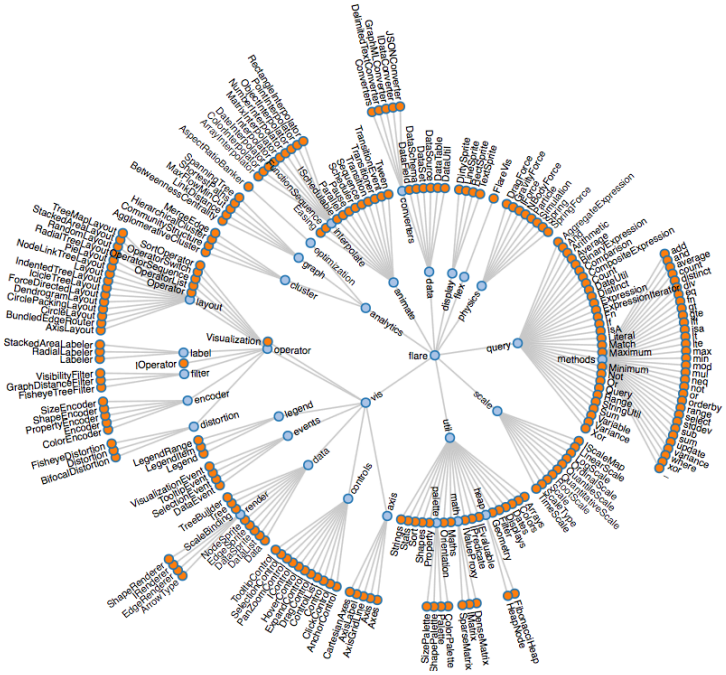


Gambar II-3 Tree-map (Multi-dimensi) [15]

2.2.2.6 Tree/Hierarchical

Tree adalah grafik herarkikal dari item-item yang memiliki hubungan satu dengan lainnya, atau yang memiliki induk (kecuali *root*). Setiap *item* antara induk

dan anak bisa memiliki banyak atribut. Grafik *tree* termasuk didalamnya grafik *tree, dendorogram, radial-tree, hyperbolic-tree, tree-map*, dan *sunburst*.



Gambar II-4 Contoh Radial-tree [15]

2.2.2.7 Network

Network adalah grafik relasi yang tidak bisa digambarkan oleh grafik *tree*, *network* memvisualisasikan relasi-relasi *item* yang banyak (semena-mena). Grafik *network* termasuk di dalamnya grafik matrix, *node-diagram*, *dependency*, *hive-plot*, *alluvial diagram*, dan *subway*.



Gambar II-5 Contoh Node-diagram [15]

2.2.3 Proses Visualisasi Data

Visualisasi data sama halnya seperti berkomunikasi, berhasil atau tidaknya komunikasi ditentukan oleh bagaimana cara pembicara menyampaikan informasi yang diberikan kepada lawan komunikasi. Terlalu meluas apa yang dibicarakan juga termasuk hal yang buruk dalam berkomunikasi. Visualisasi yang baik tentu terfokus, memberikan jawaban yang jelas, dan tidak terlalu detail [11]. Banyak metode dan cara untuk membuat visualisasi yang baik, akan tetapi pada prinsip dasarnya, visualisasi yang bagus memenuhi aspek: *overview*, *zoom* dan *filtering*, selanjutnya *details-on-demand* [14].

Proses memahami sebuah data dimulai dari beberapa pertanyaan. Selanjutnya tidak semerta-merta dijawab begitu saja, tetapi terdapat langkah-langkah dalam menjawab pertanyaan berdasarkan data. Langkah-langkah tersebut yaitu [11]:

1. *Acquire*

Tahap ini adalah tahap dimana data dikumpulkan dari berbagai sumber. Tahap ini bisa menjadi sangat rumit (contoh: mengambil data dari berkas log sistem yang besar). Tahap *acquire* hanya peduli bagaimana data didapatkan, jika produk akhir akan didistribusikan melalui internet, maka data yang ada harus memiliki struktur yang bisa disimpan dalam suatu server.

2. *Parse*

Tahap ini adalah tahap dimana data dekemas ke dalam bentuk atau format data baku dan dikelompokan ke dalam beberapa kategori. Setiap berkas data harus dapat dibaca dan bisa dibedakan dengan data lain.

3. *Filter*

Pada tahap ini beberapa data yang tidak terlalu penting dengan pertanyaan yang akan dijawab dihilangkan. Beberapa data pada berkas, mungkin perlu diterjemahkan ke dalam model matematika atau dilakukan normalisasi terlebih dahulu.

4. *Mine*

Pada tahap ini, data diaplikasikan ke dalam metode disiplin ilmu statistika atau *data mining* sebagai jalan untuk mencari pola atau dijabarkan pada konteks matematis.

5. *Represent*

Pada tahap ini data diubah dan disajikan dalam bentuk model visual dasar, seperti bar, *list*, atau *tree*. Tahap *represent* menunjukan bentuk dasar data yang akan diambil. Tahap ini adalah tahap yang sangat penting dalam membuat visualisasi data. Pemilihan model visualisasi yang tepat dapat memengaruhi bagaimana nantinya kualitas produk yang akan dihasilkan.

6. *Refine*

Pada tahap ini, model visual dibuat lebih jelas lagi dan lebih menarik. Grafik desain lebih banyak berbicara pada tahap ini. Titik-titik yang cukup penting pada visual grafik dibandingkan titik lainnya diberikan pembeda agar dapat mudah dibaca.

7. *Interact*

Pada tahap *interact*, metode untuk memanipulasi data ditambahkan, dengan arti bahwa data bisa ditampilkan sesuai kehendak pengguna. Contoh interaksi antara pengguna dan data seperti *zoom-in*, *zoom-out*, merubah rentang data, melakukan *filtering*, dll.

2.3 Statistika Deskriptif

Statistika deskriptif (*descriptive statistics*) adalah suatu metode untuk mengorganisasikan dan melihat gambaran umum dari pengukuran sampel dan populasi. Statistika deskriptif seperti namanya biasa dipakai untuk membaca data, seringkali digunakan pada sebuah sampel populasi karena sulitnya untuk mendapatkan seluruh data populasi [16]. Bila perhitungan diperoleh dari sampel

yang diambil dari populasi, maka hasil tersebut disebut statistika. Sementara bila perhitungan ditarik dari seluruh populasi, maka hasilnya disebut parameter.

Statistika deskriptif mencakup metode untuk menyusun dan mengelompokan distribusi data dan termasuk mencari satu nilai yang dapat menggambarkan semua nilai dari sampel atau populasi. Menurut Arthur M. Glenberg dan Matthew E. Andrzejewski dalam bukunya yang berjudul *"Learning from Data an Introduction to Statistical Reasoning 3rd"* terdapat tiga teknik penting pada statistika deskriptif, distribusi frekuensi dan peresentil, titik tengah dan variabilitas, dan *z* skor dan distribusi normal.

2.3.1 Distribusi Frekuensi dan Persentil

Distribusi frekuensi adalah tabulasi jumlah kemunculan dari setiap nilai. Distribusi frekuensi mengelola dan memberikan gambaran umum data, yaitu dengan cara menampilkan karakteristik terbanyak [16]. Persentil adalah persentasi skor nilai yang didapatkan dari perhitungan distribusi di bawah skor nilai. Persentil dan persen tidaklah sama. Skor nilai sendiri sering disebut sebagai persentil (*P*th), adalah skor nilai *P*% dari perhitungan di bawah distribusi [16].

2.3.2 Ukuran Gejala Pusat dan Variansi

Ukuran gejala pusat distribusi adalah nilai yang dekat dengan titik tengah dari distribusi. Biasanya digunakan sebagai representasi nilai distribusi keseluruhan [16].

2.3.2.1 Notasi Sigma

$\sum$ adalah karakter kapital dari huruf Yunani yang disebut dengan *sigma*. $\sum$ digunakan sebagai persamaan jumlah dalam bentuk lebih pendek. Contoh: terdapat sebuah data pelanggaran dari sebuah aktivitas olah raga [1], berikut adalah data tersebut:

Atlit	1	2	3	4	5	6	7	8	9
Jumlah Pelanggaran (X)	0	3	1	0	2	2	2	5	1

Dalam notasi sigma bisa ditulis sebagai berikut:

$$\sum_{i=1}^{n} X_i$$

X adalah variable yang dihitung, dalam contoh di atas adalah "jumlah pelanggaran". *i* menunjukan posisi index ke-*i*, *i*=1 di bawah ∑ menunjukan index posisi pertama yang akan dijumlahkan, *n* di atas ∑ menunjukan index posisi terakhir yang akan dihitung. Sehingga expresi sigma di atas bisa ditulis seperti:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$$

$$\sum_{i=1}^{n} X_i = 0 + 3 + 1 + 0 + 2 + 2 + 2 + 5 + 1 = 16$$

Dari data dan persamaan di atas, dapat diketahui bahwa jumlah pelanggaran dapat ditulis dengan notasi $\sum X = 16$.

2.3.2.2 Ukuran Gejala Pusat

Tedapat tiga perhitungan yang sering digunakan dalam mengukur gejala pusat, mean, median, dan modus (*mode*). Nilai pada mean, median, dan modus tidak dapat mewakili seluruh distribusi, tetapi bisa memberikan gambaran umum.

Rata-rata distribusi didefinisikan sebagai nilai penjumlahan setiap nilai yang terdapat pada distribusi (∑) dibagi dengan jumlah nilai yang terdapat pada distribusi (*n*). Formula yang dipakai untuk menghitung rata-rata adalah sebagai berikut:

$$\text{Rata-rata} = \frac{\sum X}{n}$$

Rata-rata pada sampel dan populasi dibedakan, untuk rata-rata sampel, dinotasikan sebagai *M*, sedangkan pada populasi dinotasikan sebagai μ (miu).

$$\mu = \frac{\sum X}{N} \qquad\qquad M = \frac{\sum X}{n}$$

Median distribusi adalah nilai yang berada pada titik tengah distribusi. Formula median dapat dibedakan untuk jumlah distribusi genap dan ganjil.

$$Median_{\text{ganjil}} = X_{(\frac{1}{2} * (n+1))} \qquad\qquad Median_{\text{genap}} = \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)})$$

Modus distribusi adalah nilai pada distribusi yang paling sering muncul atau yang memiliki frekuensi terbanyak.

2.3.2.3 Varian dan Standar Deviasi

Variabilitas adalah pengukuran beda nilai antara satu nilai distribusi dengan nilai distribusi pusat atau seberapa besar penyimpangan dengan nilai pusatnya. Pengukuran yang lebih menggambarkan variabilitas adalah varian, varian menghitung variabilitas berdasarkan nilai beda dengan μ [16].

Varian dari sebuah populasi dapat dihitung dengan cara [16]: (a) Hitung setiap nilai distribusi dengan μ; (b) Kuadratkan setiap nilai yang sudah dihitung; (c) Bagi jumlah dari setiap perhitungan dengan populasi, *N*.

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{SS(X)}{N}$$

Varian dari sebuah sampel untuk menggambarkan populasi, dihitung dengan formula sebagai berikut:

$$s^2 = \frac{\sum(X - \mu)^2}{n - 1} = \frac{SS(X)}{n - 1}$$

Varian dari populasi atau sampel tidak dapat dinyatakan dalam stauan ukur, karena alasan tersebut ukuran varian yang sering digunakan adalah standar deviasi. Secara matematis, standar deviasi didefinisikan sebagai akar kuadrat dari varian.

$$\sigma = \sqrt{\sigma^2} \qquad \textit{(parameter)}$$

$$s = \sqrt{s^2} \qquad \textit{(statistik)}$$

2.3.3 *z* Skor dan Distribusi Normal

z skor atau disebut juga dengan skor standar, sama seperti persentil, *z* skor juga menghitung posisi relatif dari mean. Index *z* skor adalah posisi distribusi relatif terhadap distribusi lain. *z* skor sangat berguna pada distribusi normal [16].

z skor adalah skor yang sudah dilakukan standarisasi dengan menjadikan 0 sebagai mean dari distribusi dan 1 sebagai standar deviasi [16]. Formula untuk menghitung *z* skor adalah sebagai berikut:

$$z = \frac{X - \mu}{\sigma}$$

Jarak skor dari mean berdasarkan *z* skor adalah: $+\sigma$ dan $+2\sigma$ di atas nilai mean (μ) atau : $-\sigma$ dan -2σ di bawah nilai mean (μ).

2.4 Statistika Inferensi

Tahap pertama memahami data adalah memahami sampel, selanjutnya, sampel yang diambil seringkali digunakan untuk menyimpulkan kondisi populasi sesungguhnya dimana sampel tersebut diambil. Statistika inferensi menggunakan sembarang sampel dari populasi yang diuji untuk membuat kesimpulan atau menaksir terhadap karakteristik populasi [16]. Statistika inferensi diperlukan karena terdapat dua masalah yang saling berhubungan. Pertama, skor dari populasi selalu dibutuhkan. Kedua, populasi biasanya sangat besar, menghitung skor populasi seringkali tidak bisa dilakukan.

Prosedur inferensial statistik dibedakan menjadi dua kategori, estimasi parameter dan pengujian hipotesis.

- Estimasi parameter melakukan estimasi parameter populasi dari sampel yang diambil secara acak.
- Pengujian hipotesis membandingkan dua hipotesis tentang populasi. Data dari sampel acak digunakan untuk menentukan hipotesis mana yang cenderung benar.

2.4.1 Probabilitas

Probabilitas adalah kemungkinan yang dapat terjadi dalam suatu peristiwa (*event*) tertentu. Secara matematis, probabilitas diartikan sebagai banyaknya pristiwa yang dimaksud dibagi dengan seluruh pristiwa yang mungkin [16]. Formula probabilitas yaitu:

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

dengan:

$P(A)$ = probabilitas terjadinya peristiwa A

$n(A)$ = jumlah peristiwa A

$n(S)$ = jumlah peristiwa yang mungkin

2.4.2 Distribusi Sampling

Keputusan diambil terhadap informasi yang diambil dari populasi. Informasi yang lengkap diperoleh dari seluruh populasi. Hanya saja, untuk mendapatkan seluruh data dari populasi hampir tidak mungkin dilakukan secara singkat. Sampling adalah metode yang digunakan untuk mengestimasi parameter-parameter populasi [16].

Standar deviasi dari distribusi sampling disebut dengan standar error statistik (*SE*). Perbedaan standar deviasi dan standar error adalah pada nilai, sementara *standard error* berkaitan dengan hasil perhitungan.

$$SE = \frac{\sigma}{\sqrt{n-1}}$$

Perbedaan hasil yang diperoleh dari sampel statistik dan hasil yang diperoleh dari parameter populasi disebut sampling error. Sampling error dipakai untuk menunjukan ketepatan populasi yang didasarkan pada perhitungan terhadap sampel di bawah distribusi normal. Semakin kecil sampling error, semakin besar ketepatan penaksiran.

2.4.2.1 Pendugaan Titik

Cara penaksiran skor tunggal parameter populasi mean (μ) dapat dilakukan dengan mengambil mean terhadap sampel acak populasi (*M*). Mean sampel acak tersebut disebut pendugaan titik (*point estimator* -- μ_i) [16].

2.4.2.2 Pendugaan Interval

Pendugaan interval (*interval estimation*) menunjukan interval berapa suatu parameter populasi akan berada pada dua nilai (dibatasi oleh dua nilai) karena adanya *treatment*, nilai tersebut biasanya disebut batas bawah dan batas atas [16].

$$\overline{X} - z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}} \leq \overline{X} \leq z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}} + \overline{X}$$

2.4.2.3 *Significance Level (α)*

Significant secara bahasa berarti penting, tetapi dalam statistik '*significant'* berarti probabilitas *true* atau *false*. Alpha level (α) adalah simbol yang menunjukan sebarapa besar suatu kejadian dalam sampel terjadi akibat pengambilan acak (*likely*) atau bukan acak (*unlikely*). Artinya kejadian yang muncul dari sampel apakah terjadi secara kebetulan (*random*) atau bukan kebetulan [17]. *Significance of Alpha level (α)* dibagi menjadi tiga, yaitu:

1. $\alpha = 0.05$
2. $\alpha = 0.01$
3. $\alpha = 0.001$

2.4.2.4 Nilai Statistik *t*

Nilai statistik *t* menggambarkan seberapa jauh beda rata-rata sampel antara satu sampel dengan sampel lainnya atau populasi. Nilai Statistik *t* didapatkan dari pembagian antara beda mean populasi dengan sampel dengan *standard error* [16].

$$t = \frac{\overline{x} - \mu}{SE} = \frac{\overline{x} - \mu}{S/\sqrt{n}}$$

2.5 Korelasi

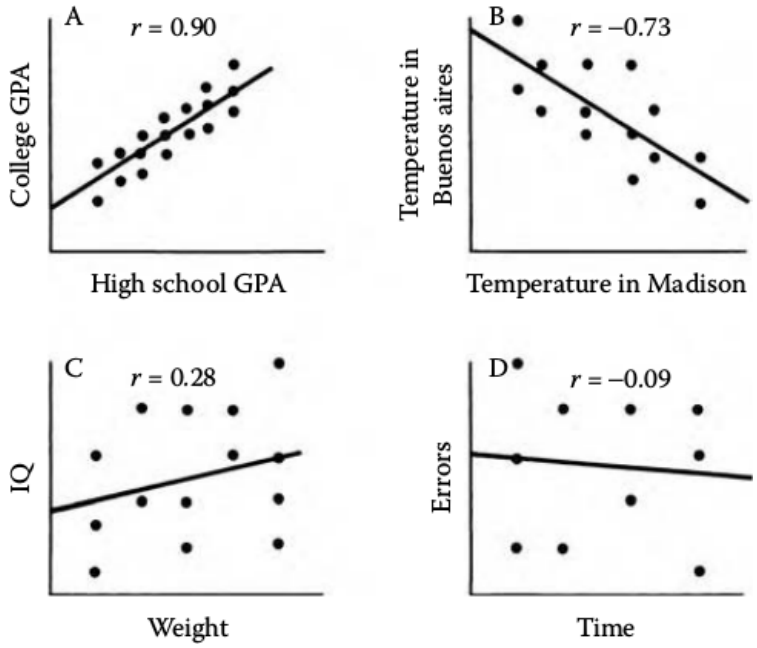
Korelasi digunakan untuk mengukur hubungan antara dua sampel, yaitu ketika pengukuran setiap variabel sampel dengan variabel lainnya harus dilibatkan. Analisis korelasi digunakan untuk membuat kesimpulan terhadap populasi dari sampel yang diukur, analisis ini bisa dilakukan hanya ketika suatu sampel yang dependen [16].

2.5.1 Korelasi *Pearson r*

Korelasi *Pearson r* atau yang sering disebut *r* saja, adalah pengukuran terhadap kekuatan garis linear antara dua variabel. Pengukuran *r* bisa diaplikasikan pada sampel atau populasi, simbol *r* dugunakan untuk mengukur statistik, sedangkan untuk parameter menggunakan simbol *Rho* (ρ) [16]. Baik *r* atau ρ memiliki properi yang sama, yaitu [16]:

1. Nilai korelasi selalu pada rentang -1 dan 1m inklusif.
2. Nilai korelasi selalu menunjukan hubungan dengan kemiringan/gradien garis regresi; garis regresi dengan kemiringan positif memiliki koefisien korelasi positif, dan sebaliknya. Garis regresi dengan gradien 0, memiliki koefisien korelasi 0.0.
3. Kekuatan relasi garis linear diindikasikan dengan nilai absolut dari korelasi. Relasi garis linear yang sempurna ditandai dengan r atau ρ yang bernilai -1 atau 1; tidak ada hubungan garis linear r atau ρ bernilai 0.

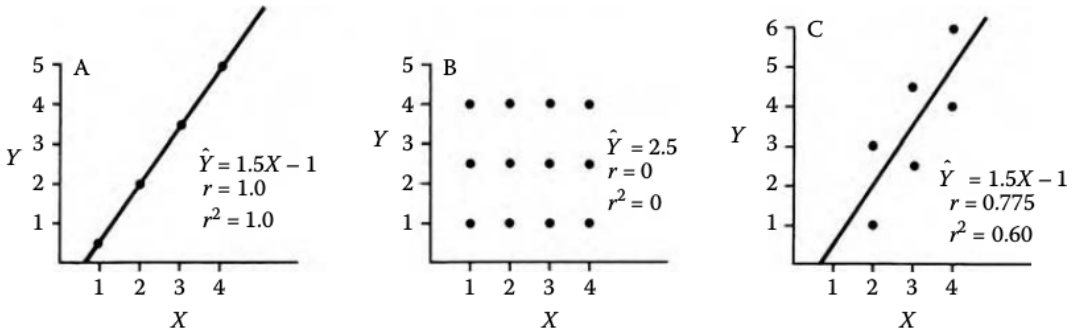
Kekuatan relasi garis linear ditandai dengan *scatter plots* yang dekat dengan garis regresi ($S_{y.x}$ kecil). Ilustrasi ini digambarkan pada Gambar II-6.



Gambar II-6 Ilustrasi Korelasi dan *Scatter plots* [16]

2.5.2 Interpretasi r^2

Kelemahan dari r atau ρ adalah tidak memiliki interpretasi langsung terhadap varian. r^2 atau ρ^2 dapat menghitung proporsi y varian terhadap x varian secara jelas [16]. Interpretasi dari r^2 digambarkan pada Gambar II-7.



Gambar II-7 Ilustrasi Korelasi r^2

Secara matematis, r2 dapat dihitung dengan mengurangi varian *x* terhadap *x* dengan total varian dari *y* dengan 1 [18].

$$r^2 = 1 - \frac{SE_{garis}}{SE_y} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum(y_i - (bx_i + a))^2}{\sum(y_i - \overline{y})^2}$$

2.5.3 Faktor-Faktor yang Memengaruhi Nilai *r*

Terdapat beberapa faktor yang harus diperhatikan ketika akan menghitung hubungan antara *x* dan *y*. Jika diabaikan, faktor-faktor tersebut akan memberikan masalah terhadap regresi [16]. Faktor-faktor tersebut yaitu [16]:

1. *Nonlinear relationship*. Korelasi linear hanya menghitung derajat hubungan kelineran. Jika *x* dan *y* berhubgan tetapi tidak linear, *r* dapat memberikan nilai kosong (meskipun terdapat hubungan antara dua variabel).
2. *Restricted range*. Pengecualian baik pada range *x* atau *y*, akan mengurangi nilai *r*.
3. *Extreme score*. Satu nilai yang sangat tinggi dapat menghasilkan bukti pada saat perhitungan korelasi ketika tidak ditemukan pada skor mayoritas.
4. *Combining griups.* Dapat terjadi tidak adanya hubungan antara grup, tetapi menggabungkan beberapa grup dapat menimbulkan kebiasan dalam korelasi linear.

2.5.4 Pendugaan Interval *Rho (ρ)*

Nilai statistik *r* dapat digunakan untuk membuat berbagai macam kesimpulan tentang *ρ*. Salah satu manfaat *r* adalah pendugaan interval terhadap *ρ*. Dua nilai *r* dapat dihitung untuk menduga nilai *ρ*. Sama seperti pendugaan interval

pada $\overline{x}$, dimana diperlukan nilai standar yaitu *z* dan α level. Pada pendugaan interval ρ juga diperlukan nilai statistik *t* dan α level [16]. Nilai statistik t untuk menduga interval ρ dapat dihitung dengan formula:

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{df}}} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{N - 2}}}$$

df adalah kependekan dari *dgree of freedom* (derajat kebebasan). Dimana *df* dapat dihitung dengan mengurangi setiap kelompok sampel dengan 1. Artinya, jika terdapat *n* sampel, maka *df* adalah *n*-(banyak sampel). Sehingga, untuk regresi linear yang memiliki sampel *x* dan sampel *y,* maka *df* adalan 2(*n*-1). Selanjutnya pada *t* statistik dinotasikan *N*-2 [16].

Selanjutnya interval ρ dapat dihitung dengan formula [19]:

$$r \pm t_{\alpha/2} SE_r$$

SE_r adalah simpangan baku dari *r*. SE_r dapat dihitung dengan formula:

$$SE_r = \sqrt{\frac{1 - r^2}{df}} = \sqrt{\frac{1 - r^2}{N - 1}}$$

2.6 Regresi Linear

Regresi linear merupakan salah satu pembelajaran mesin *supervised learning*, yaitu pembelajaran mesin yang memproyeksikan suatu nilai berdasarkan nilai lainnya [20]. Regresi linear memproyeksi nilai dengan membuat model. Model adalah represtasi dari *state*, proses, atau sistem untuk mengambil sebuah kesimpulan [20]. Regresi linear menjelaskan hubungan matematis antara dua variabel. Regresi linear memakai variabel independen *x* untuk menjelaskan variabel dependen *y* atau sering disebut juga *outcome variable* [6].

Model regresi linear yang terbentuk untuk sampel adalah $y = bx + a$, sedangkan untuk populasi adalah $y = \beta_1 x + \beta_0$. *b* adalah tingkat kemiringan untuk sampel, sedangkan β_1 adalah tingkat kemiringan (*slop*) untuk populasi. Untuk mencari tingkat kemiringan, dilakukan dengan cara mengalikan *r* dengan pembagian SE_y dengan SE_x [17].

$$\beta_1 = r\frac{SE_y}{SE_x}$$

Model regresi linear ini kemudian diplot melewati titik ($\overline{x}$,$\overline{y}$) [17].

2.7 *k*-Means

Misalkan sebuah data set *D,* data set *D* membiliki *n* buah objek pada ruang Euclidean. *k*-Means adalah algoritma yang mendistribusikan objek-objek pada data set *D* ke *cluster k, C_1, ..., C_k*, dimana $C_i \subset D$ dan $C_i \cap C_j = \varnothing$ untuk ($1 \leq i, j \leq k$) [21]. Fungsi objektif yang hendak dicapai oleh *k*-Means *clustering* adalah objek dalam *cluster* memiliki kriteria sejenis dan tidak sejenis untuk *cluster* lainnya [21].

Berikut adalah abstraksi bagaimana algoritma *k*-means bekerja [21]:

1. Secara acak, pilih *k-centroid* (titik yang akan mencadi titik tengah *cluster*) pada ruang *d* dimensi.
2. Kelompokan setiap titik ke dalam *k*-set *centroid* paling dekat.
3. Perbaharui *k-centroid* dengan rata-rata titik *k*-set atau titik-titik yang sudah dikelompokan sebelumnya.
4. Ulang langkah ke-2 dan ke-3 sampai pengelompokan *divergent*.