Statistika Bisnis

Apa itu Statistik dan Statistika?

Statistik : kumpulan angka/data yang menggambarkan suatu keadaan.

Statistika : ilmu yang mempelajari tentang statistik, yaitu bagaimana cara mengumpulkan, mengolah, menyajikan, menganalisis dan menginterpretasikan data sebagai dasar pengambilan keputusan.

Menurut Prof. Drs. Sutrisno Hadi, MA:

Statistik adalah cara untuk mengolah data dan menarik kesimpulan-kesimpulan yang teliti dan keputusan-keputusan yang logis dari pengolahan data.

Kenapa kita belajar statistika?

Pengambil keputusan menggunakan statistika untuk :

- Menyajikan dan mendeskripsikan data dan informasi bisnis dengan benar.
- Mengambil kesimpulan dari suatu populasi.
- Membuat prediksi tentang aktivitas bisnis.
- Meningkatkan proses bisnis.

Tipe Statistika

Statistika Deskriptif: Kumpulan metode untuk mengorganisir, menampilkan, dan mendeskripsikan data menggunakan grafik, tabel, dan summary measure(rata-rata, nilai tengah, distribusi data).

Statistika Inferensial: Kumpulan metode yang menggunakan sampel dari sebuah populasi untuk membantu mengambil keputusan atau Membuat sebuah prediksi tentang populasi tersebut.

Statistika Deskriptif

Contoh:

Kita mempunyai sampel data mengenai pengeluaran rumah tangga dari provinsi DKI Jakarta yang berjumlah 5000 baris. Lebih mudah bagi kita untuk mengambil kesimpulan menggunakan ringkasan data dan grafik daripada melihat original data secara langsung. Maka dari itu kita mereduksi data menjadi *manageable size* menggunakan grafik, tabel, dan *summary measure.*

Statistika Inferensial

Contoh:

Sebuah pabrik minuman ringan memproduksi 10.000 botol minuman per hari. Untuk menjaga kualitas dari minuman tersebut, beberapa sampel botol diambil untuk menguji standar kualitas minuman tersebut.

Penggunaan Statistika

1. **Akuntansi** : Perusahaan akuntan publik menggunakan statistika ketika melakukan audit terhadap kliennya.
2. **Keuangan** : Statistika digunakan untuk membantu memberikan rekomendasi investasi.
3. **Pemasaran** : Pengambilan sampel di masyarakat untuk menilai suatu produk menggunakan kaidah statistika.
4. **Ekonomi** : Para ahli ekonom menggunakan statistika untuk memprediksi kondisi ekonomi di masa mendatang.

Data dan Variabel

Data : Kumpulan keterangan atau fakta yang dikumpulkan dari suatu observasi.

Variabel : Karakteristik dari suatu objek.

1. Variabel Diskrit (terbatas)
   - Jumlah karyawan perusahaan A adalah 5000 orang.
   - Jumlah kecelakaan kerja di bulan September sebanyak 10 kasus.
2. Variabel Kontinyu (tidak terbatas)
   - Keuntungan perusahaan tahun 2019 sebesar 5,3 M.
   - Modal awal perusahaan sebesar 10,3 M.

Mengapa Data Penting?

1. Data membantu memahami suatu peristiwa/kejadian.

   Contoh: Apakah orang yang makan apel setiap hari lebih sehat daripada orang yang tidak makan apel?

2. Data membantu memprediksi perilaku pelanggan di masa depan sebagai dasar pengambilan keputusan bisnis.

   Contoh: Berdasarkan riwayat kunjungan pengguna di sosial media, kita bisa menentukan iklan apa yang tepat untuk menarik perhatian mereka.

Syarat Data yang Baik

1. **Objektif** : Data sesuai dengan keadaan yang sebenarnya.
2. **Representatif** : Data mewakili objek yang diamati.
3. **Standar eror kecil** : Nilai data mendekati nilai populasi.
4. ***Up to date*** : Data selalu ter*update* agar dapat mengambil keputusan dengan tepat.
5. **Relevan** : Data sesuai dengan masalah yang akan diselesaikan.

Jenis Data

Menurut skala pengukuran :

a. Nominal
b. Ordinal
c. Interval
d. Rasio

Skala Nominal

- Skala pengukuran yang paling sederhana dalam penelitian.
- Data dikelompokkan dalam kategori dan diberi label (bukan sebagai tingkatan).
- Kategori data bersifat *mutually exclusive* (setiap objek memiliki satu kategori).
- Data tidak bisa diurutkan.

Contoh :

1. Jenis kelamin
2. Kota kelahiran

Skala Ordinal

- Skala yang tidak hanya menunjukkan kategori, tetapi juga menunjukkan tingkatan.
- Dapat diurutkan.
- Tidak menunjukkan jarak dan interval

Contoh :

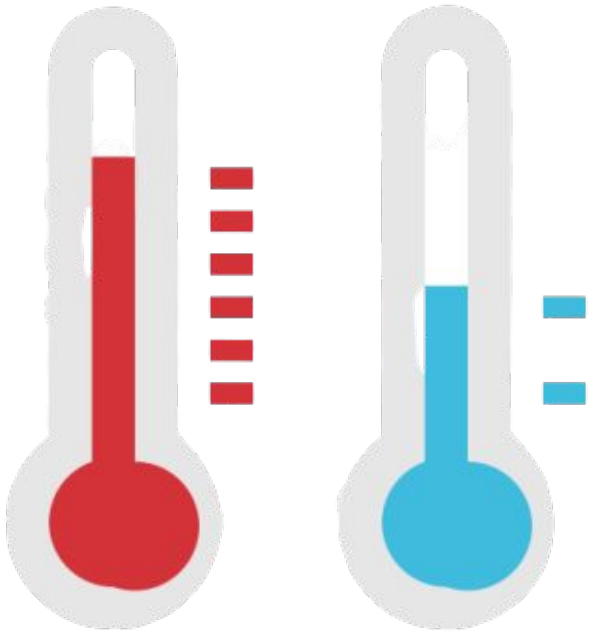
1. Tingkatan gaji pegawai (eselon I, II, III, IV).
2. Respon suatu survey (Tidak pernah, terkadang, sering, pernah)

Skala Interval

- Terdapat jarak yang sama antar tingkatan.
- Angka nol hanya menggambarkan suatu titik (tidak punya nilai nol absolut).

Contoh :

Pengukuran suhu.

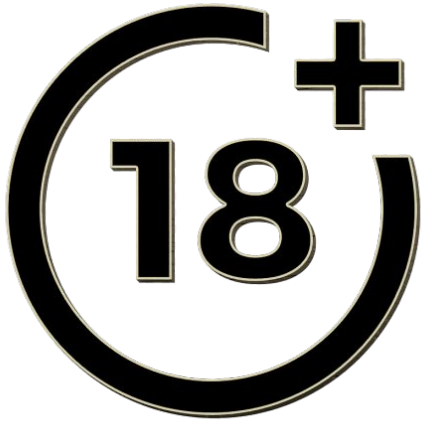


Skala Rasio

- Memiliki nilai nol mutlak.
- Terdapat semua karakteristik skala nominal, ordinal, interval.

Contoh :

1. Usia
2. Tinggi badan



Ringkasan Skala Pengukuran

Skala	Tipe pengukuran			
	Kategori	**Peringkat**	**Jarak**	**Perbandingan**
Nominal	Ya	Tidak	Tidak	Tidak
Ordinal	Ya	Ya	Tidak	Tidak
Interval	Ya	Ya	Ya	Tidak
Rasio	Ya	Ya	Ya	Ya

Jenis Data

Menurut sifat:

- Kualitatif
    - Menggunakan label
    - Skala pengukuran nominal dan ordinal
    - Data bisa numerik dan kategorik
- Kuantitatif
    - Mengindikasikan berapa banyak
    - Data selalu numerik
    - Skala pengukuran Interval dan ratio

Jenis Data

Menurut sumbernya:

1. Data Internal
2. Data Eksternal

Jenis Data

Menurut waktu pengumpulanya:

1. **Cross Sectional data**: dikumpulkan pada waktu tertentu yang sama atau hampir sama. *(Contoh: Jumlah Mahasiswa Upi Tahun 2015/2016)*
2. **Data yang dikumpulkan selama kurun waktu tertentu**. *(Contoh: pergerakan nilai tukar rupiah dalam 1 tahun)*

Jenis Data

Menurut cara memperolehnya:

1. **Data primer :** dikumpulkan dan diolah sendiri.
2. **Data Sekunder :** data yang diperoleh dalam bentuk yang sudah jadi.

Jenis Pengukuran Data

Ukuran Pemusatan Data :

- “Berapa rata - rata gaji karyawan di perusahaan A?”
- “Berapa nilai tengah data penjualan bulan April?”

Ukuran Penyebaran Data :

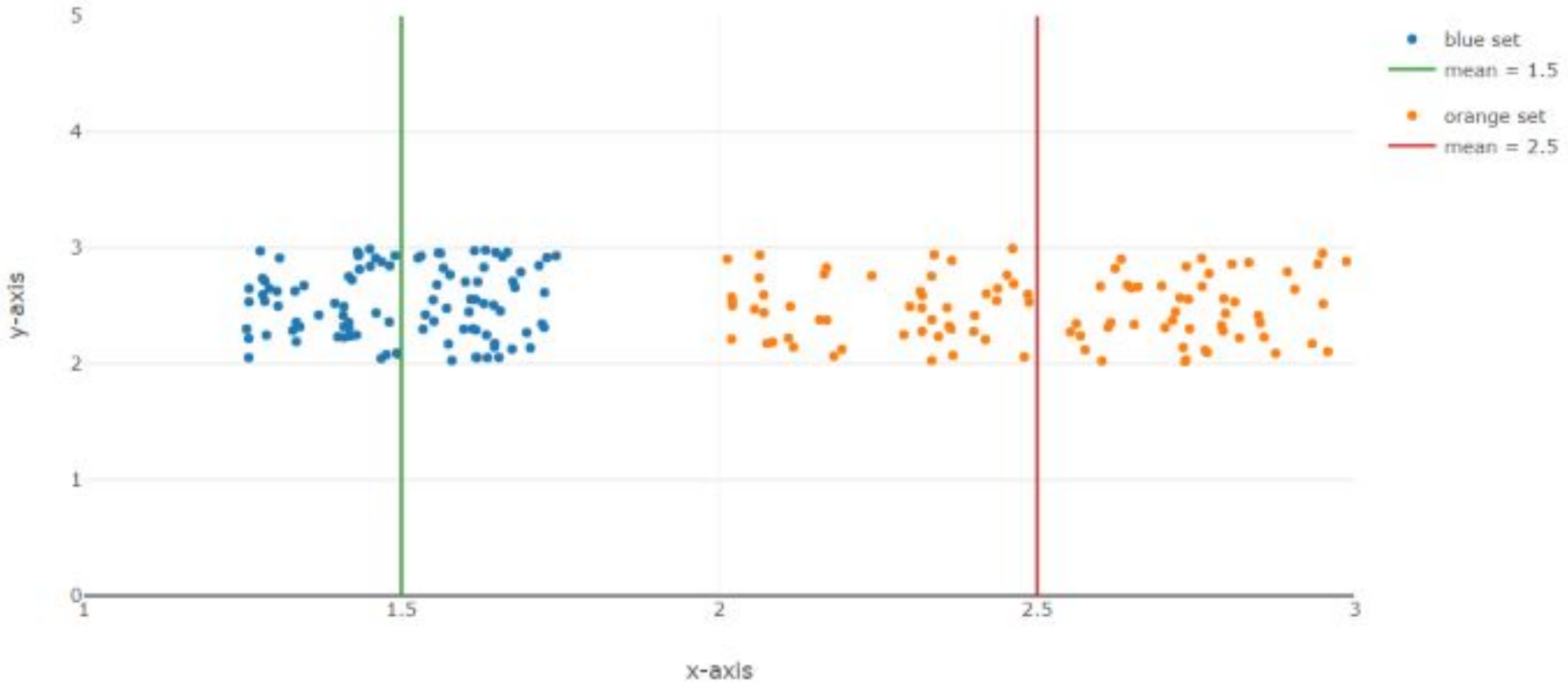
- “Seberapa jauh nilai individu menyimpang dari rata- ratanya?”

Ukuran Pemusatan Data

- Mendeskripsikan letak pusat data.
- Tidak bisa menggambarkan bentuk/persebaran data.

1. Mean :
   - Rata - rata suatu data.
   - Dipengaruhi harga ekstrim.
2. Median :
   - Nilai tengah data.
   - Tidak dipengaruhi harga ekstrim.
3. Modus :
   - Nilai data yang sering muncul.
   - Tidak dipengaruhi harga ekstrim
   - Modus bisa lebih dari satu dan bisa juga tidak ada

Mean



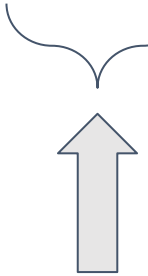
Mean hanya menentukan letak rata - rata data, namun tidak menunjukkan bagaimana penyebaran data. Mean dipengaruhi oleh harga ekstrim.

Median

10 13 13 14 16 16 16 17 18 18 18 19 19 20 23 24 25

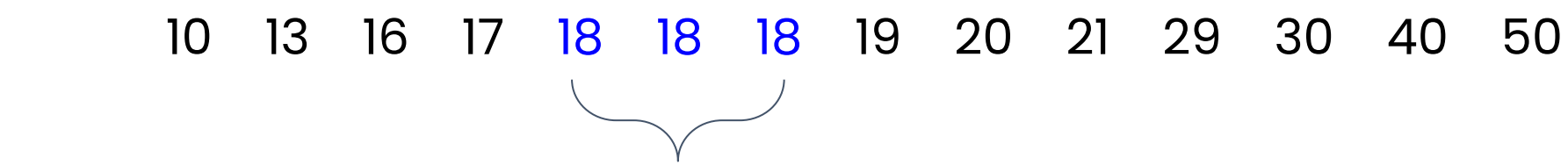
Median = 18

23 24 25 25 26 27 27 28 29 29 31 31

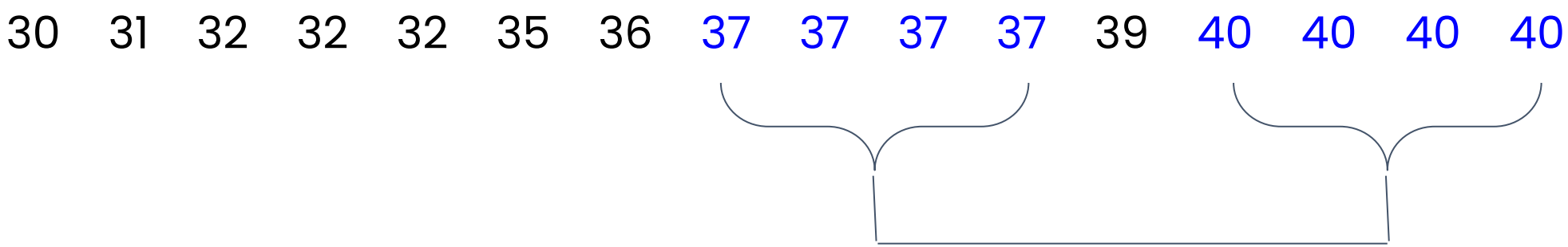


Median = (27 +27)/2
= 27

Modus



Modus = 18



Modus = 37 dan 40

Ukuran Penyebaran Data

Menggambarkan bagaimana penyebaran suatu data.

1. Range
2. Variansi
3. Standar Deviasi

Range

10 13 16 17 18 18 18 19 20 21 29 30 40 50

Range = Nilai maksimum - Nilai minimum

= 50 - 10 = 40

Kelemahan range : Ada kemungkinan melibatkan harga ekstrim sehingga besar kemungkinan mengandung kesalahan.

Variansi

- Jumlah jarak kuadrat dari setiap titik/nilai ke rata - rata.
- Variansi populasi dan sampel berbeda.

Variansi sampel :

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

Variansi populasi :

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

Variansi sampel

4 7 9 8 11

$$\bar{x} = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$ sample mean

$$s^2 = \frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5-1}$$

$$= 6.7$$ sample variance

Standar deviasi

Akar kuadrat variansi.

Standar deviasi sampel : $$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Standar deviasi populasi : $$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Standar deviasi sampel

Sample:

4	7	9	8	11

$$\bar{x} = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

sample mean

$$s = \sqrt{\frac{(4-7.8)^2 + (7-7.8)^2 + (9-7.8)^2 + (8-7.8)^2 + (11-7.8)^2}{5-1}}$$

$$= \sqrt{6.7} = 2.59$$

sample standard deviation

Standar deviasi populasi

Population:

4	7	9	8	11

$$\mu = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

population mean

$$\sigma = \sqrt{\frac{(4-7.8)^2 + (7-7.8)^2 + (9-7.8)^2 + (8-7.8)^2 + (11-7.8)^2}{5}}$$

$$= \sqrt{5.36} = 2.32$$

population standard deviation

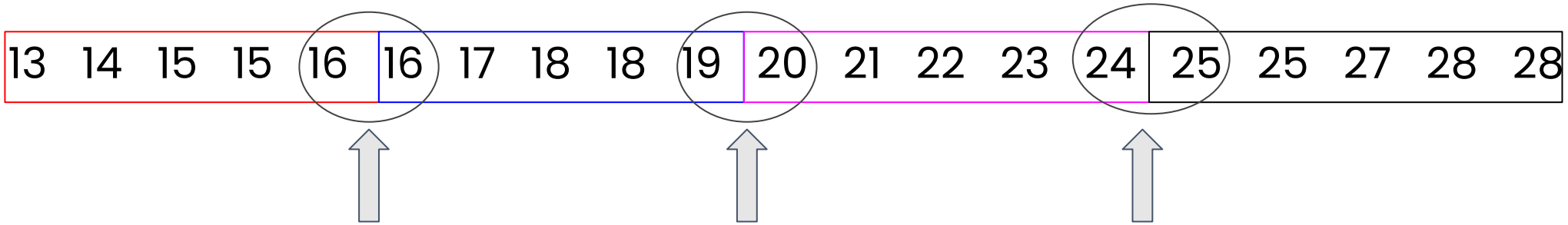
Rangkuman Ukuran Penyebaran

- Semakin data menyebar, semakin besar range, variansi, dan standar deviasi.
- Semakin data terpusat, semakin kecil range, variansi, dan standar deviasi.
- Jika nilai semua data sama (tidak ada variasi), maka ukuran penyebarannya nol.
- Tidak ada ukuran yang bernilai negatif.

Ukuran Kuartil

- Cara lain mendeskripsikan data
- Kelebihan : Semua data dipertimbangkan, namun tidak semua data masuk perhitungan.

Kuartil dan Interquartilrange (IQR)



Kuartil I = 16

Kuartil II atau Median = 19,5

Kuartil III = 24,5

IQR = Kuartil III - Kuartil I

= 8,5

Data Preprocessing

Data Preprocessing

- Menjelaskan berbagai proses yang menyiapkan data mentah untuk menjalankan prosedur lainnya.
- Tujuannya adalah mentransformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk pengguna.

Mengapa Data Preprocessing?

Data di dunia nyata itu “kotor” :

1. **Tidak lengkap** : ada atribut data yang hilang.

   Contoh : pekerjaan_ortu = “”

2. **Eror dan Noisy** : mengandung kesalahan tersembunyi.

   Contoh : gaji_ortu = “-10”

3. **Tidak konsisten** : ketidakcocokan dalam kode dan nama.

   Contoh : Umur = “40” tgl_lhr = “03/07/1997”

Pentingnya Data Preprocessing

- Data yang tidak berkualitas akan menghasilkan kualitas mining yang tidak baik.
- Data preprocessing, cleaning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining (90%).

Manfaat Data Preprocessing

- Mendapatkan hasil yang lebih akurat
- Pengurangan waktu komputasi untuk l*arge scale problem*
- Membuat nilai data menjadi lebih kecil tanpa mengubah informasi di dalamnya.

Teknik Data Preprocessing

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation

Data Cleaning

Proses pembersihan data :

- Memperkecil *noise*
- Membetulkan data yang tidak konsisten
- Mengisi atribut data yang kosong
- Mengidentifikasi atau membuang data yang tidak sesuai

Data Integration

- Data yang telah dibersihkan lalu diintegrasikan dalam satu database *datawarehouse*.
- Sumber data beragam.

Teknik :

- Tight Coupling
- Loose Coupling

Data Reduction

Data di dalam database *datawarehouse* diseleksi untuk mendapatkan hasil yang lebih akurat.

Teknik :

- *Dimensionality Reduction*
- *Numerosity Reduction*
- *Data Compression*

Data Transformation

Tujuan : Mengefisienkan proses data mining dan menghasilkan pola yang lebih mudah dipahami.

Strategi :

- *Smoothing*
- *Attribute Construction*
- *Aggregation*
- *Normalization*
- *Discretization*

Regresi

Regresi

Definisi : salah satu metode statistika yang digunakan untuk mencari hubungan antara satu variabel dependen dengan beberapa variabel independen.

Tipe Dasar Regresi :

1. Regresi linear sederhana
2. Regresi linear ganda

16 Tipe Regresi

1. Linear Regression
2. Polynomial Regression
3. Logistic Regression
4. Quantile Regression
5. Ridge Regression
6. Lasso Regression
7. Elastic Net Regression
8. Principal Components Regression (PCR)
9. Partial Least Squares (PLS) Regression
10. Support Vector Regression
11. Ordinal Regression
12. Poisson Regression
13. Negative Binomial Regression
14. Quasi Poisson Regression
15. Cox Regression
16. Tobit Regression

Manfaat Regresi

- Regresi membantu manajer keuangan dan investasi menghitung nilai aset dan memahami hubungan antara variabel, misalnya harga komoditas dan harga saham komoditas tersebut.
- Regresi dapat membantu memprediksi penjualan perusahaan berdasarkan cuaca, harga sebelumnya, GDP, dan kondisi lainnya.
- Metode Capital Asset Pricing Models (CAPM) adalah model regresi yang sering digunakan di bidang keuangan untuk menentukan harga aset dan modal.

Regresi Linear Sederhana

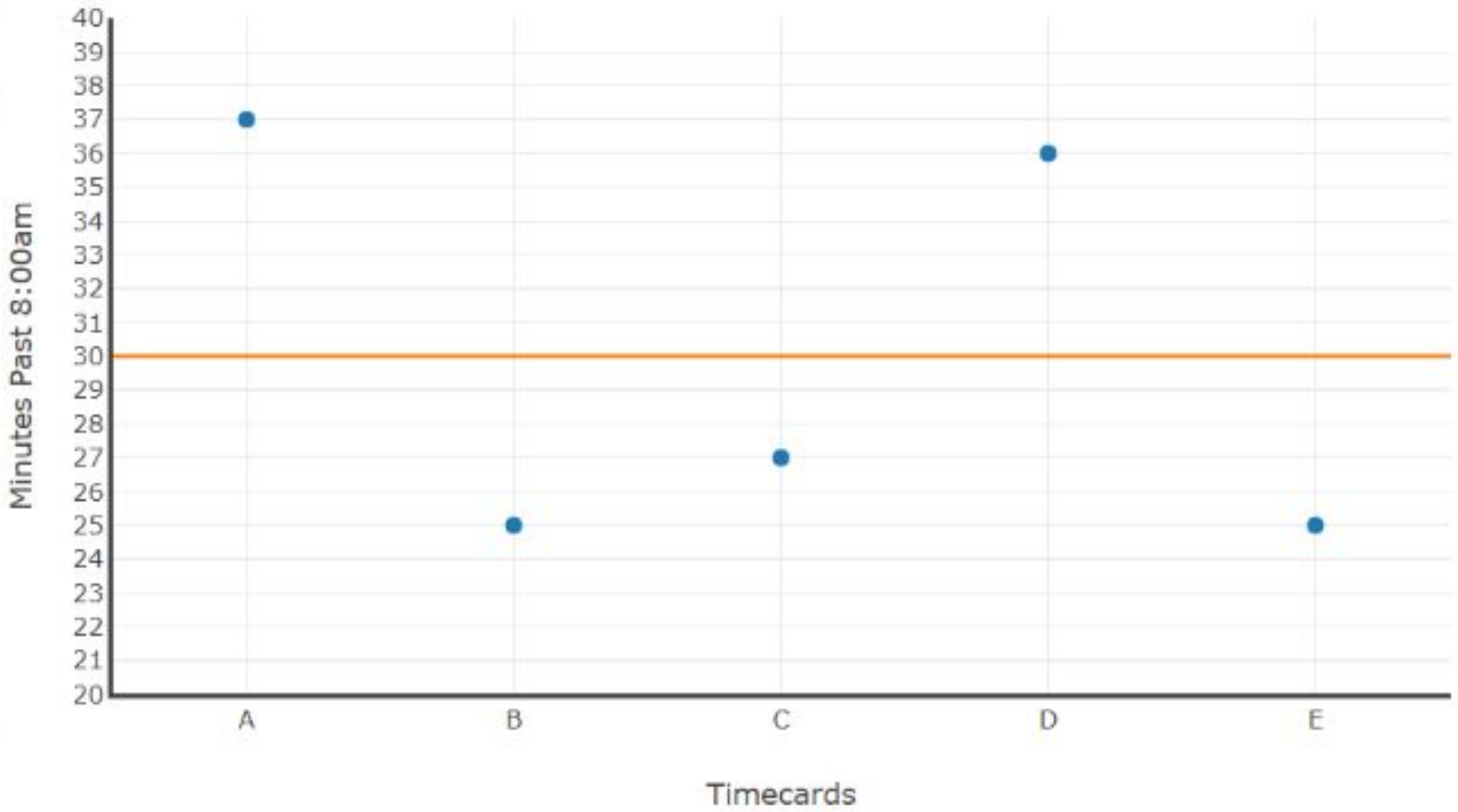
Hanya melibatkan dua variabel, satu variabel dependen dan satu variabel independen.

Menemukan garis regresi yang tepat (mewakili nilai data) :

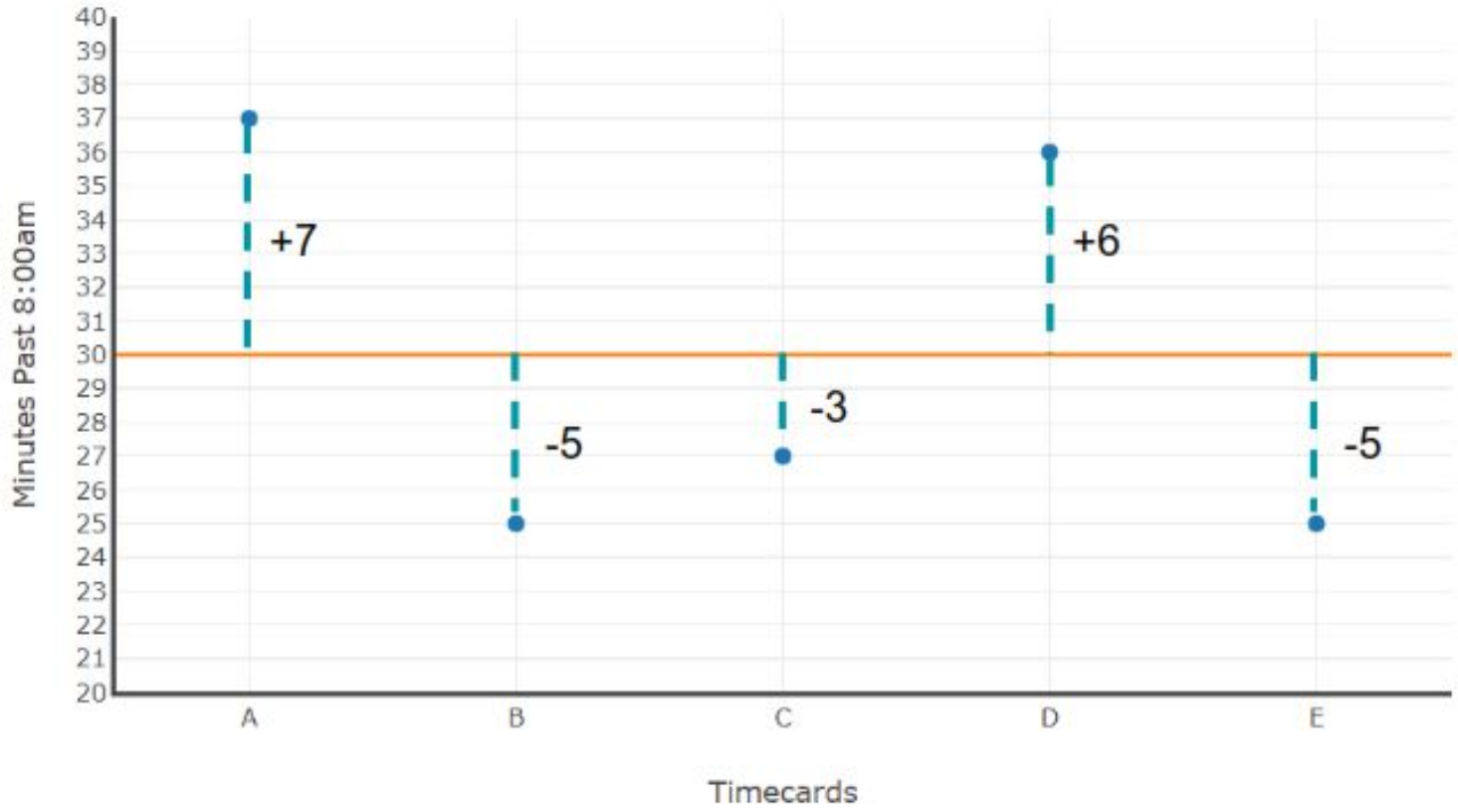
Contoh :

Seorang manajer pabrik ingin mengetahui kapan karyawannya tiba di kantor. Jam kerja dimulai pukul 8.30. Manajer mengambil 5 kartu waktu acak dan mencatat menit kedatangan karyawannya di chart.

Timecard	Minutes past 8:00am
A	37
B	25
C	27
D	36
E	25
Total:	**150**
Mean	**30**



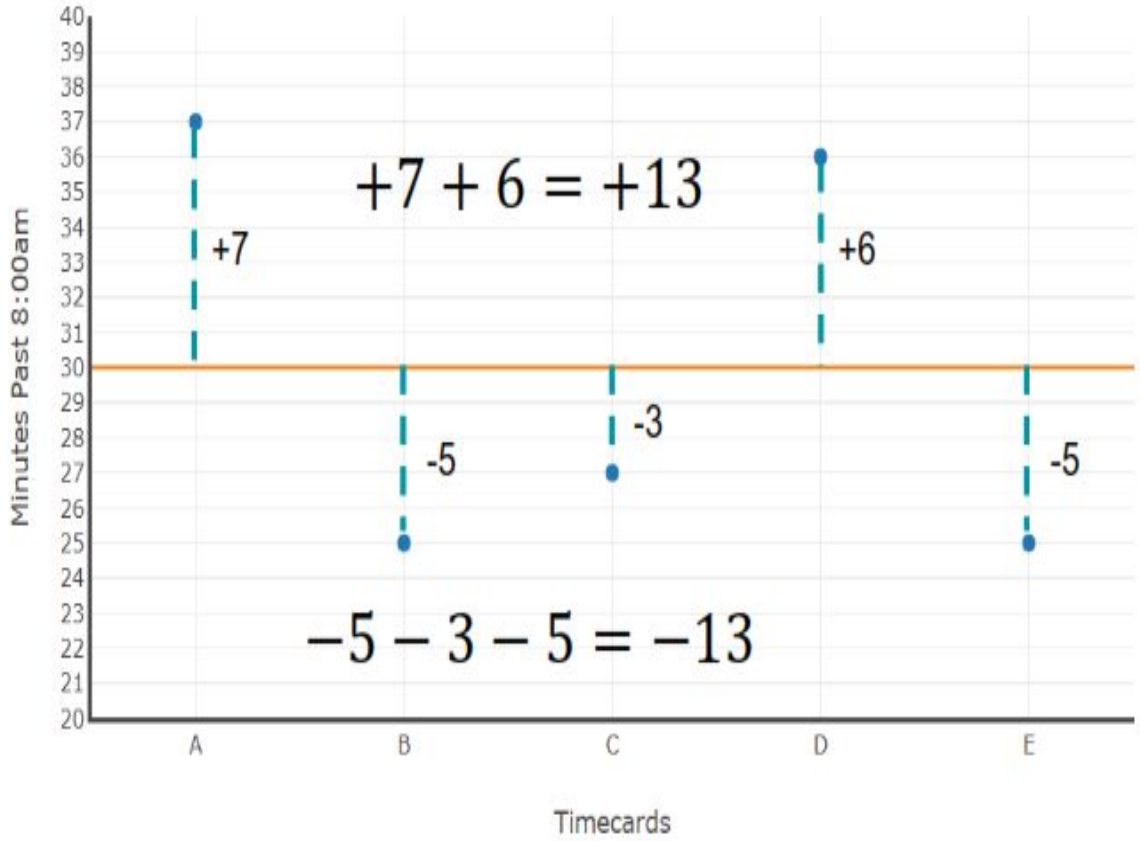
Apa yang membuat y = 30 adalah garis yang paling cocok?



Apa yang membuat y = 30 adalah garis yang paling cocok?

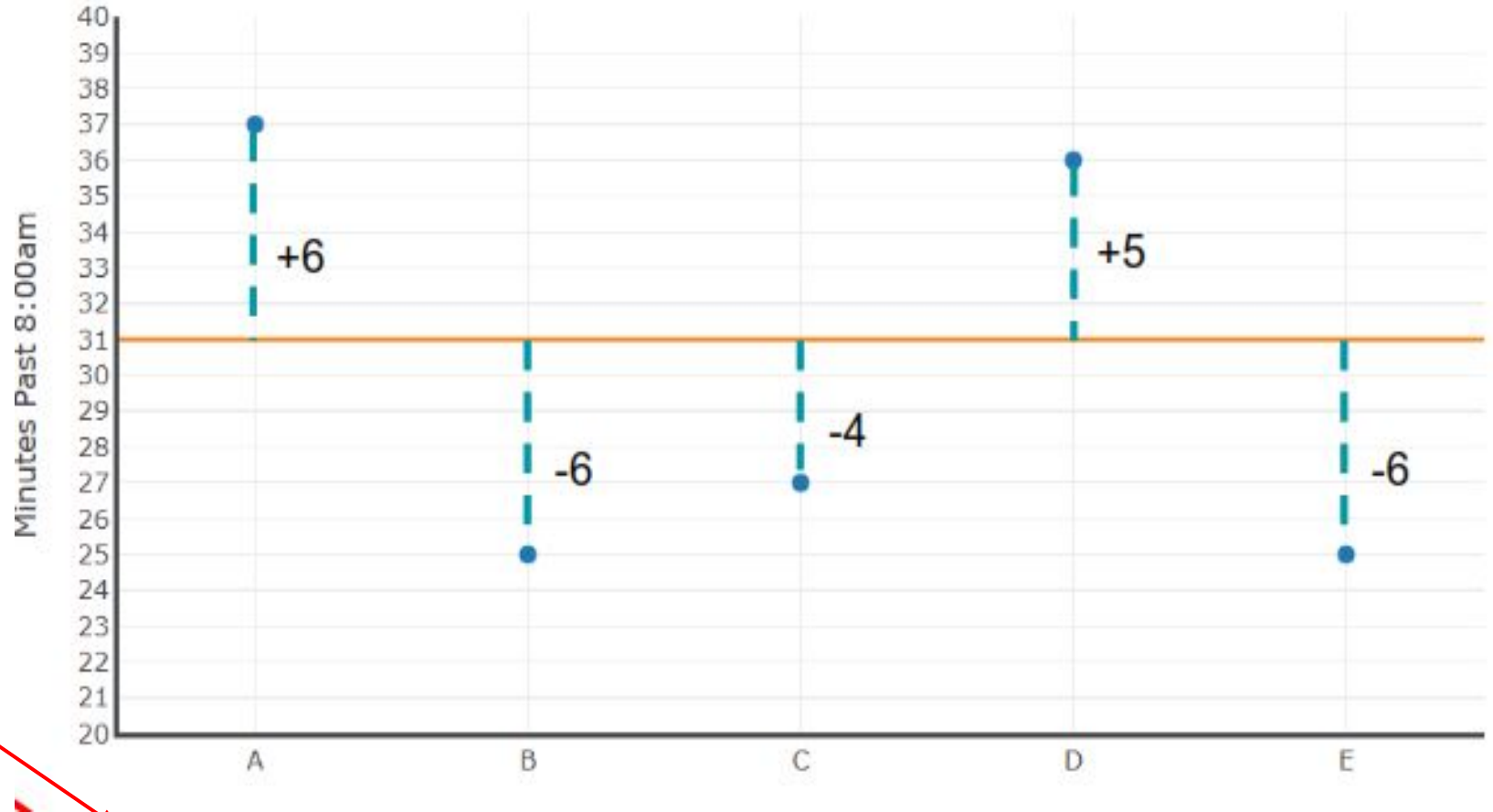
- Jumlah dari jarak di atas garis y = 30 adalah sama dengan jumlah jarak di bawah garis y = 30.
- Berusaha mencari nilai SSE sekecil mungkin.

Error (E)	Square Error (SE)
+7	49
-5	25
-3	9
+6	36
-5	25
Sum of Squares Error (SSE)	144



Apakah pergeseran garis mempengaruhi SSE ?

Error (E)		Square Error (SE)	
+7	+6	49	36
-5	-6	25	36
-3	-4	9	16
+6	+5	36	25
-5	-6	25	36
Sum of Squares Error (SSE)		144	149



Nilai SSE menjadi lebih besar

Regresi Linear Sederhana

Persamaan garis pada umumnya **y = mx + b**

- m = gradien, b = dimana garis memotong sumbu y saat x = 0

Pada regresi linear, saat ingin merumuskan hubungan antara variabel, persamaan y = mx + b menjadi $\hat{y} = b_0 + b_1 x$

Tujuan : memprediksi nilai variabel dependen (y) berdasarkan variabel independen (x).

- variabel dependen : variabel yang dipengaruhi variabel independen
- variabel independen : variabel yang mempengaruhi variabel dependen

Regresi Linear Sederhana

Memperoleh nilai b_0 dan b_1

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

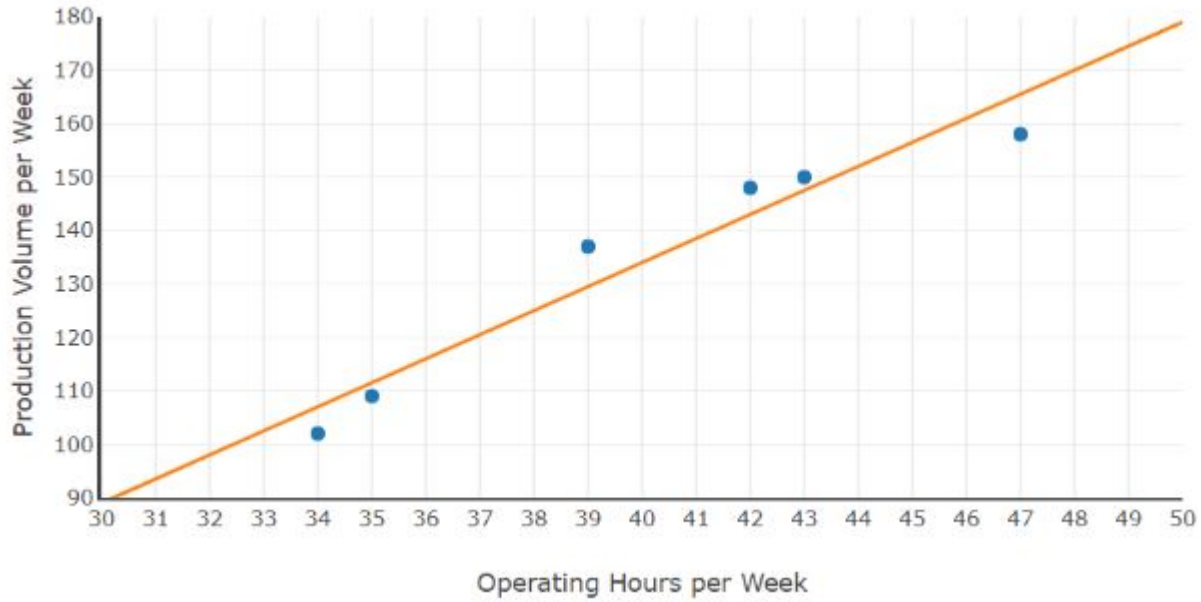
$$b_0 = \bar{y} - b_1\bar{x}$$

Contoh Regresi Linear Sederhana

Seorang manajer ingin mengetahui hubungan antara jumlah jam operasi pabrik dalam 1 minggu dengan volume produksi mingguan. Data sebagai berikut.

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

1. Menentukan variabel dependen dan independen :
   - **variabel dependen (y)** : volume produksi
   - **variabel independen (x)** : jumlah jam operasional
2. Membuat scatter plot untuk mengetahui apakah hubungan antar variabel linear atau tidak.



3. Perhitungan

	Production Hours (x)	Production Volume (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	34	102	-6	-32	192	36
	35	109	-5	-25	125	25
	39	137	-1	3	-3	1
	42	148	2	14	28	4
	43	150	3	16	48	9
	47	158	7	24	168	49
$\bar{x}, \bar{y}$	**40**	**134**		**Sum:**	**558**	**124**
					$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

3. Perhitungan

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{558}{124} = \mathbf{4.5}$$

$$b_0 = \bar{y} - b_1\bar{x} = 134 - (4.5 \times 40) = \mathbf{-46}$$

$$\hat{y} = \mathbf{-46 + 4.5x}$$

Sum:	558	124
	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

4. Kesimpulan

Persamaan regresi yang didapatkan adalah $\hat{y} = -46 + 4.5x$

Berdasarkan persamaan regresi tersebut, apabila manajer ingin pabrik menghasilkan 125 produk per minggu, maka jam operasional yang dibutuhkan adalah 38 jam per minggu.

$$\hat{y} = b_0 + b_1 x$$

$$125 = -46 + 4.5x$$

$$x = \frac{171}{4.5} = \mathbf{38}\ hours\ per\ week$$

Visualisasi Data

Visualisasi Data

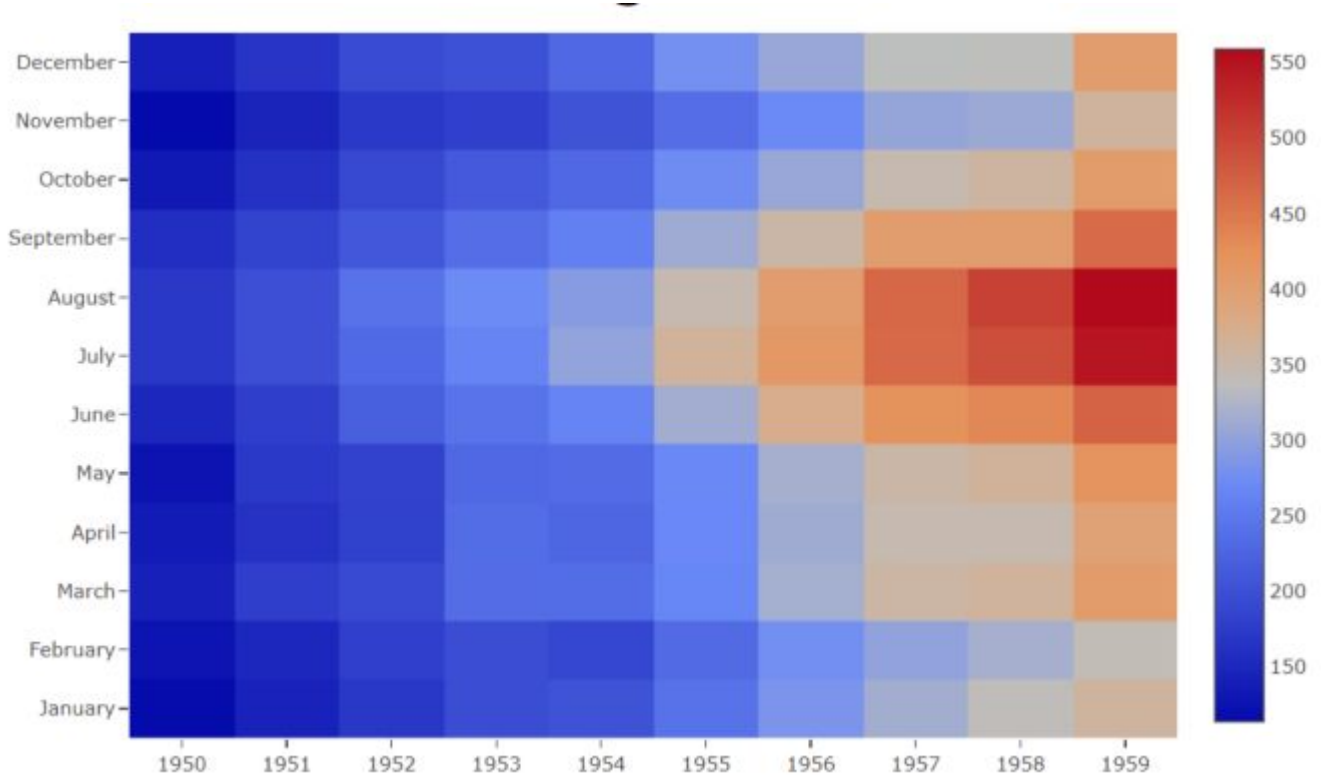
Dari data :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	year	month	passengers		year	month	passengers		year	month	passengers		year	month	passengers
2	1950	January	115		1952	July	230		1955	January	242		1957	July	465
3	1950	February	126		1952	August	242		1955	February	233		1957	August	467
4	1950	March	141		1952	September	209		1955	March	267		1957	September	404
5	1950	April	135		1952	October	191		1955	April	269		1957	October	347
6	1950	May	125		1952	November	172		1955	May	270		1957	November	305
7	1950	June	149		1952	December	194		1955	June	315		1957	December	336
8	1950	July	170		1953	January	196		1955	July	364		1958	January	340
9	1950	August	170		1953	February	196		1955	August	347		1958	February	318
10	1950	September	158		1953	March	236		1955	September	312		1958	March	362
11	1950	October	133		1953	April	235		1955	October	274		1958	April	348
12	1950	November	114		1953	May	229		1955	November	237		1958	May	363
13	1950	December	140		1953	June	243		1955	December	278		1958	June	435
14	1951	January	145		1953	July	264		1956	January	284		1958	July	491
15	1951	February	150		1953	August	272		1956	February	277		1958	August	505
16	1951	March	178		1953	September	237		1956	March	317		1958	September	404
17	1951	April	163		1953	October	211		1956	April	313		1958	October	359
18	1951	May	172		1953	November	180		1956	May	318		1958	November	310

Sulit mengambil kesimpulan dari data mentah seperti ini.

Visualisasi Data

Menjadi :



Dengan mengubah data tersebut menjadi grafik, akan lebih mudah mengambil kesimpulan.