DOLab

live class

Data Visualization and EDA

Outline

- Descriptive Statistics
- Introduction to Data Visualization
- Basic Chart and Graphs
- Other Visualizations
- Exploratory Data Analysis (EDA)

Descriptive Statistics

Descriptive Statistics

Cabang statistik yang berkaitan dengan pengumpulan, penyusunan, dan penyajian data numerik untuk memberikan gambaran ringkas tentang karakteristik dasar data

Tujuan utamanya adalah untuk menggambarkan dan merangkum data secara terorganisir agar dapat dimengerti dan dianalisis lebih lanjut.

Beberapa Teknik :

- Central tendency yang menyediakan informasi seperti mean, median, modus
- Melihat persebaran data seperti nilai min, max, standar deviasi, kuartil
- Melihat apakah data tersebar normal atau skewness
- Menampilkan grafik dan diagram secara visual untuk mempermudah pemahaman tentang pola dan karakteristik data
- Menganalisis hubungan antara dua atau lebih variable menggunakan koefisien korelasi

Mean, Median dan Modus

Data Nilai :
5, 6, 6, 5, 5, 7, 10, 9, 8, 7

Carilah mean, median dan modus

Mean adalah rata-rata dari data
Mean = (5 + 6 + 6 + 5 + 5 + 7 + 10 + 9 + 8 + 7) / 10
Mean = 6,8

Median adalah nilai tengah dari data setelah diurutkan
Data yang diurutkan 5, 5, 5, 6, 6, 7, 7, 8, 9, 10
Karena jumlah data genap (10), nilai tengahnya diambil dari data ke-5 dan ke-6 kemudian dibagi 2.
Median = (6 + 7) / 2
Median = 7,5

Modus adalah data yang paling sering muncul
Modus = 5
5 merupakan data yang paling sering muncul yaitu sebanyak 3 kali

Mean atau Median?

Untuk memahami pengaruh outlier terhadap ukuran pemusatan data antara mean atau median perhatikan contoh berikut kembali

Perusahaan DQTech ingin mengetahui rata-rata dan juga nilai tengah dari data gaji karyawannya. Data dan perhitungan adalah sebagai berikut :

Posisi	Gaji
Data Analyst	5.000.000
Quality Assurance Eng.	5.750.000
Business Intelligence Dev.	6.000.000
Data Engineer	6.500.000
Data Scientist	7.000.000

$$Rataan = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{5}(5jt + 5{,}75jt + \cdots + 7jt) = 6.050.000$$

$$Median = 6.000.000$$

Kemudian ditambahkan data Gaji Manager IT pada sampel data diatas sehingga menjadi

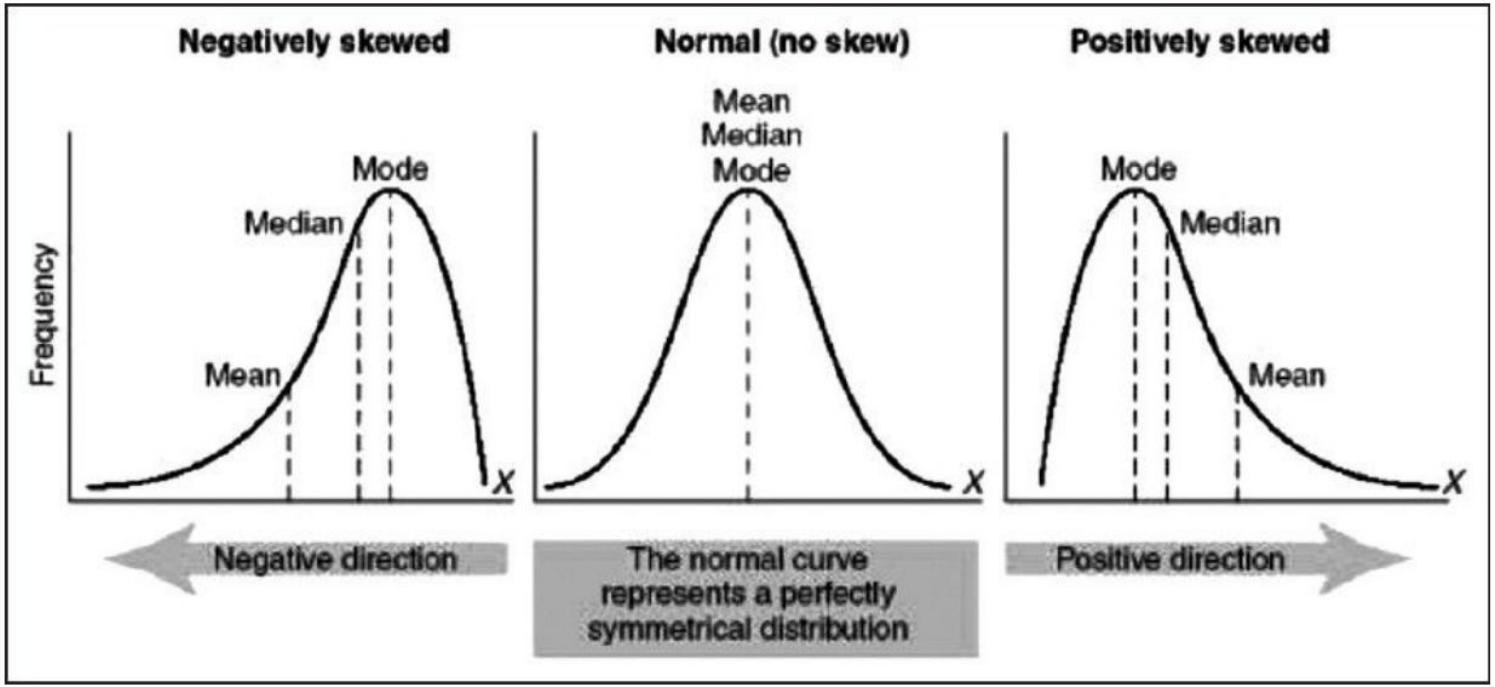
Posisi	Gaji
IT Manager	21.200.000

$$Rataan = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{6}(5jt + 5{,}75jt + \cdots + 7jt + 21{,}2jt) = 8.575.000$$

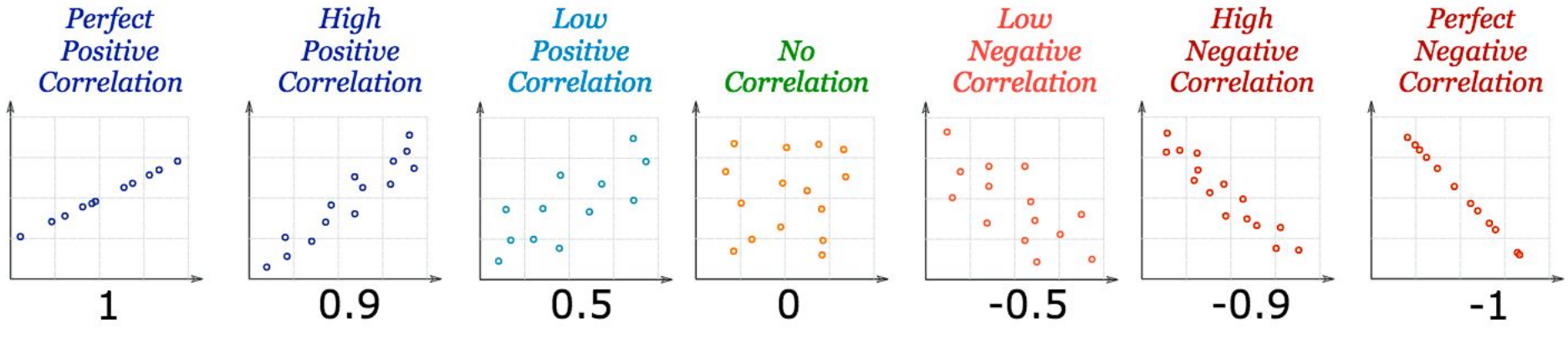
$$Median = 6.500.000$$

Dapatkah disimpulkan bahwa rata-rata gaji karyawan DQTech adalah 8.575.000?
Dalam hal ini, median lebih baik digunakan karena relatif tidak terpengaruh terhadap nilai sangat tinggi atau rendah

Tipe Distribusi



Korelasi



Semakin besar nilai korelasi berarti semakin kuat hubungan liniernya.

Rentang nilai korelasi adalah -1 hingga 1.

Introduction to Data Visualization

Data Visualization

Proses mewakili informasi atau data dengan menggunakan elemen visual seperti grafik, diagram, peta atau visual lainnya.

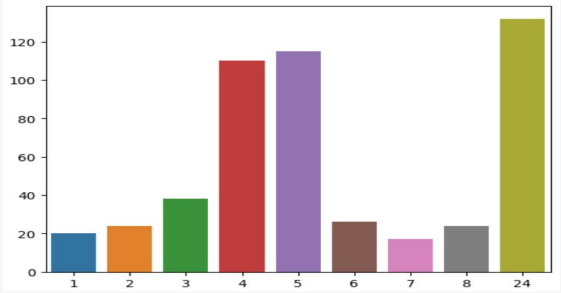
Tujuan utamanya adalah untuk menyajikan data dengan cara yang mudah dipahami dan dapat memberikan wawasan yang lebih baik bagi pengguna.

Manfaat :

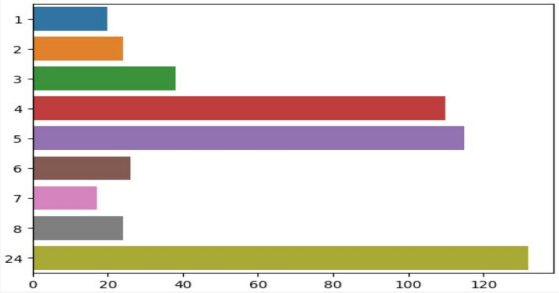
- ❑ Memudahkan pemahaman
- ❑ Mengungkapkan pola dan tren
- ❑ Melihat kejanggalan atau anomaly
- ❑ Komunikasi yang efektif

Tipe Data Visualisasi

Bar Chart



Vertical Bar



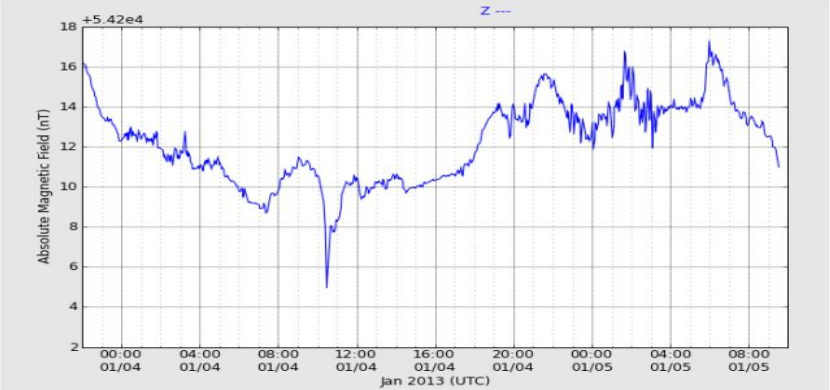
Horizontal Bar

Kapan digunakan:

- ❑ Untuk membandingkan statistical summary seperti jumlah, rata-rata, median
- ❑ Membandingkan antar kategori dengan jumlah objek yang tidak terlalu banyak
- ❑ Memvisualisasikan data kategorikal

Tipe Data Visualisasi

Line Chart

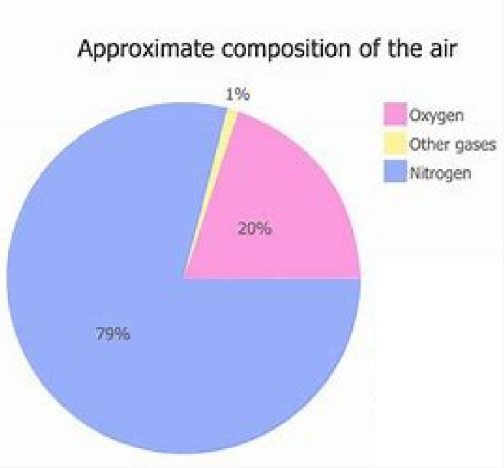


Kapan digunakan:

- Untuk membandingkan statistical summary dari waktu ke waktu
- Menampilkan pola pergerakan data seperti efek musiman, efek libur, dll

Tipe Data Visualisasi

Pie Chart

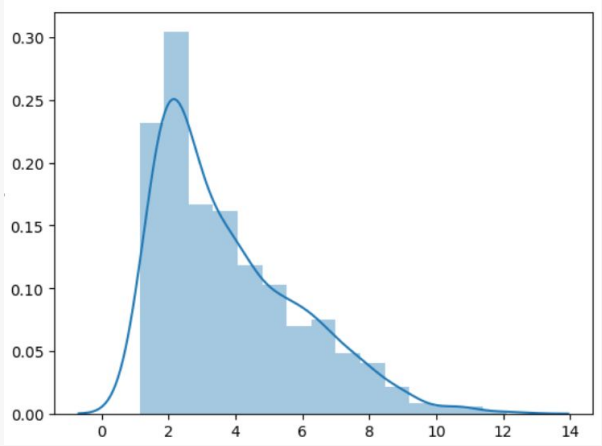


Kapan digunakan:

- ❑ Membandingkan objek mana yang memiliki porsi lebih besar secara persentase atau angka
- ❑ Membandingkan proporsi pada objek yang biasanya kurang dari 5 objek

Tipe Data Visualisasi

Histogram

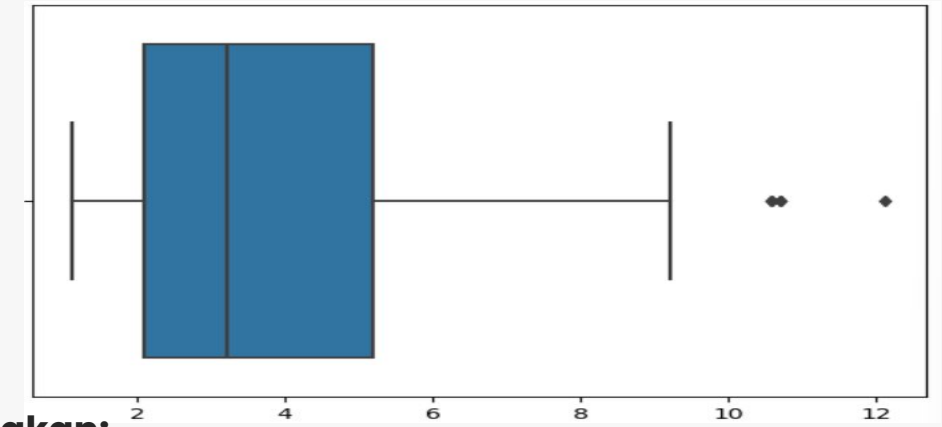


Kapan digunakan:

- Untuk mengetahui penyebaran kuantitatif pada kelompok tertentu
- Mengetahui persebaran data apakah menumpuk di suatu skala tertentu

Tipe Data Visualisasi

Box Plot

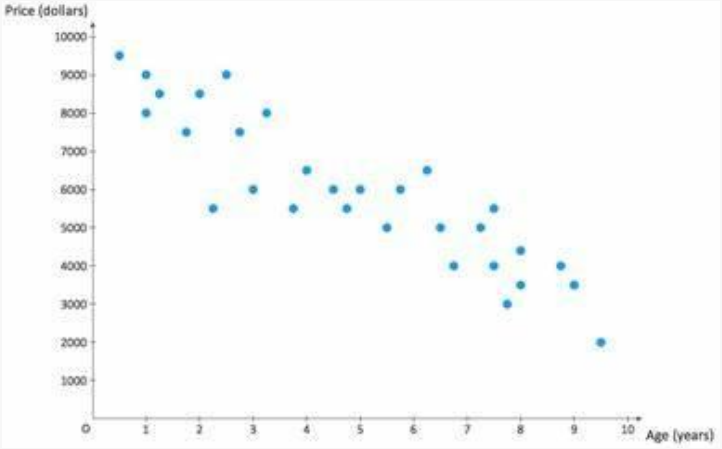


Kapan digunakan:

- Untuk mengetahui penyebaran kuantitatif secara detail (min, Q1, Q2, Q3, max)
- Untuk melihat ada tidaknya pencilan atau outlier

Tipe Data Visualisasi

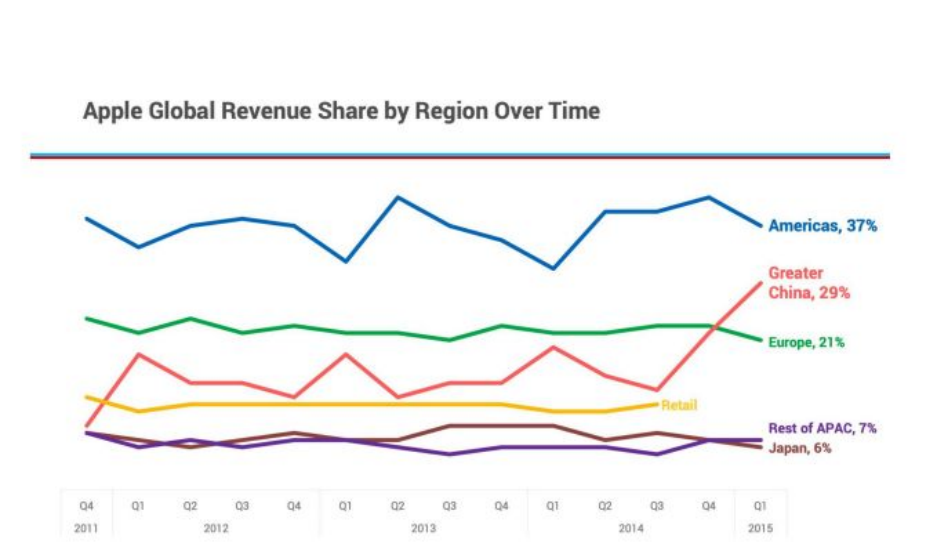
Scatter Plot

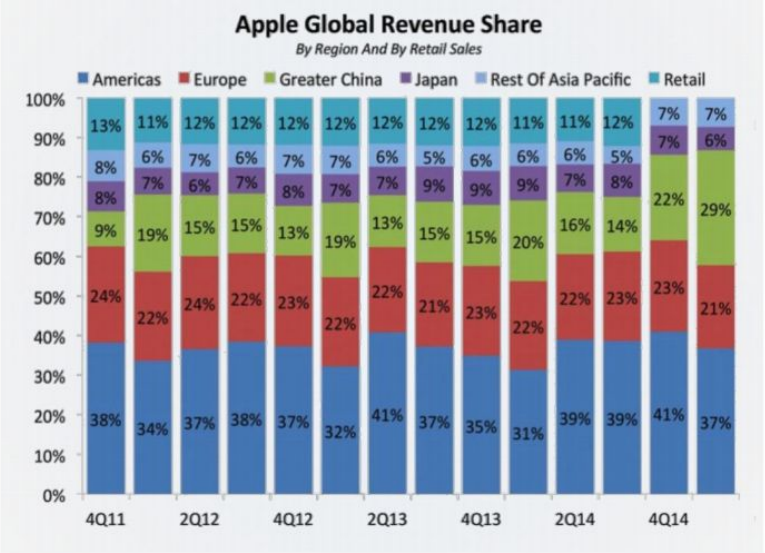


Kapan digunakan:

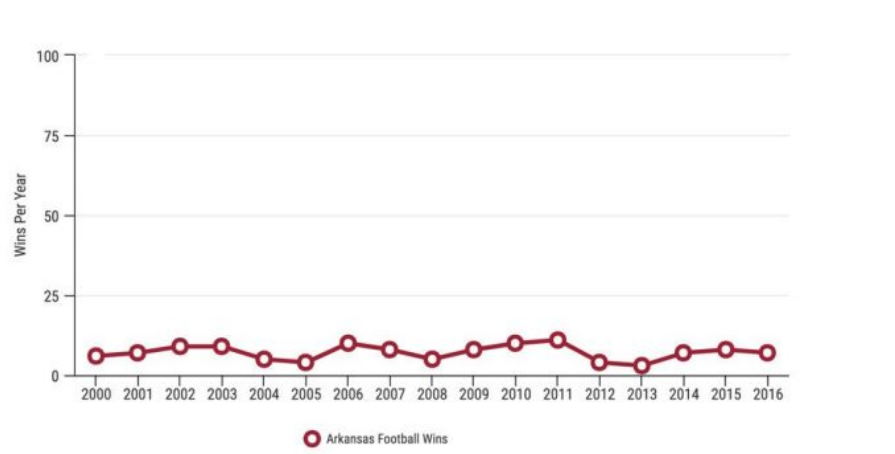
- Untuk mengetahui penyebaran kuantitatif dari dua objek kuantitatif.
- Untuk mencari korelasi atau hubungan antar dua objek

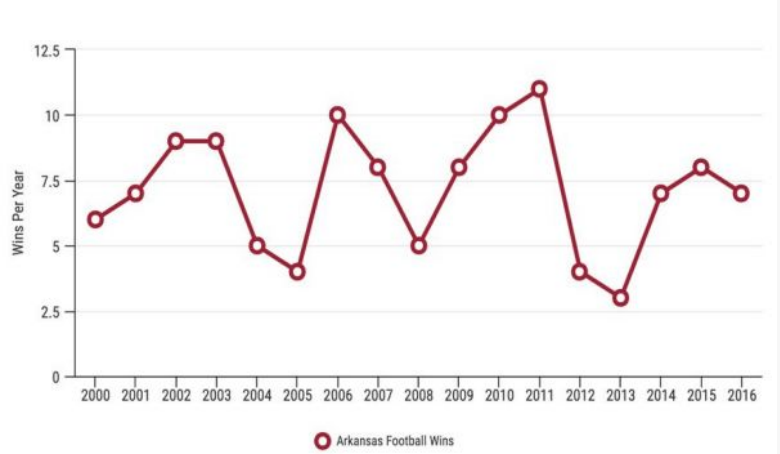
Pilih mana?


Apple Global Revenue Share by Region Over Time
Americas, 37%
Greater China, 29%
Europe, 21%
Retail
Rest of APAC, 7%
Japan, 6%
Q4 2011
Q1 Q2 Q3 Q4 2012
Q1 Q2 Q3 Q4 2013
Q1 Q2 Q3 Q4 2014
Q1 2015


Apple Global Revenue Share
By Region And By Retail Sales
Americas
Europe
Greater China
Japan
Rest Of Asia Pacific
Retail
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
4Q11
2Q12
4Q12
2Q13
4Q13
2Q14
4Q14

Pilih mana?


100
75
50
25
0
Wins Per Year
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Arkansas Football Wins


12.5
10
7.5
5
2.5
0
Wins Per Year
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Arkansas Football Wins

Introduction to Data Visualization

Data Visualization

Beberapa library yang biasa digunakan untuk visualisasi :

- ❑ Seaborn
- ❑ Matplotlib

Matplotlib

Kelebihan:

- Fleksibilitas tinggi
- Dokumentasi lengkap

Kekurangan:

- Sintaks lebih kompleks dan panjang
- Pemahaman mendalam tentang konsep visualisasi data

Seaborn

Kelebihan:

- Sintaks sederhana
- Memiliki gaya default lebih menarik dan professional

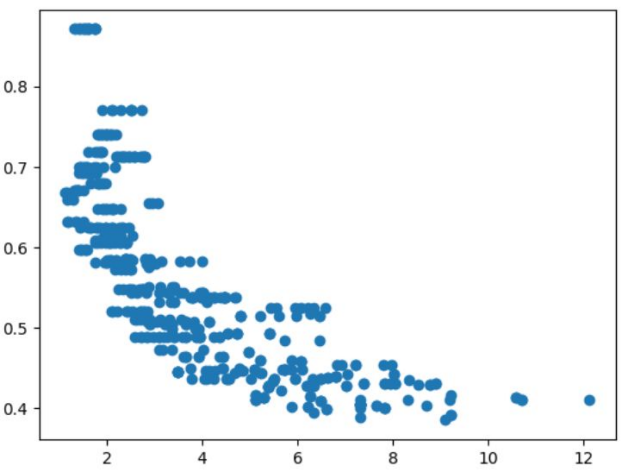
Kekurangan:

- Kurangnya fleksibilitas dalam beberapa kasus
- Dokumentasi tidak sekomprehensif seperti Matplotlib

Matplotlib vs Seaborn

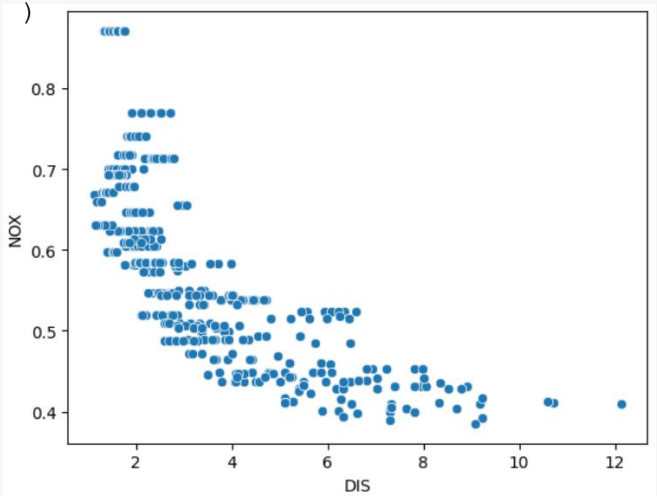
```
import matplotlib.pyplot as plt

plt.scatter(x=df['DIS'], y=df['NOX'])
```



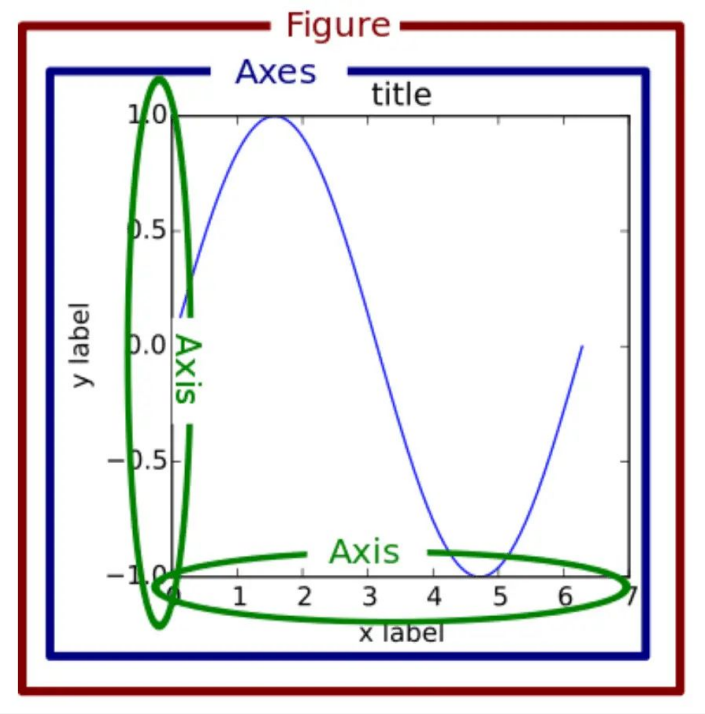

```
import seaborn as sns

sns.scatterplot(x=df['DIS'],y=df['NOX'])
```



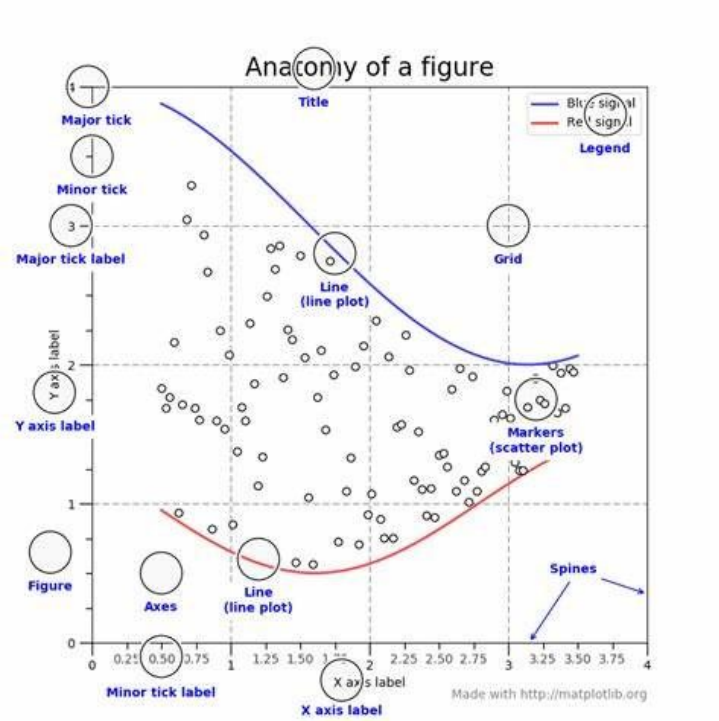
Beberapa visualisasi terlihat berbeda antara matplotlib dan seaborn.
Pada matplotlib, terlihat tidak ada label pada x-axis dan y-axis. Semuanya harus dibuat satu persatu.

Matplotlib



Anggap matplotlib sebagai figure yang mempunyai beberapa elemen yang semuanya dapat dimodifikasi.

Matplotlib Anatomy



Anatomi di samping sebenarnya dibuat menggunakan matplotlib dengan kode Python ini

Basic Charts

Basic Charts

```
import matplotlib.pyplot as plt

Bar       □ plt.bar(x, y)
Line      □ plt.plot(x, y)
Histogram □ plt.hist(data)
Box       □ plt.boxplot(data)
Scatter   □ plt.scatter(x, y)
Pie       □ plt.pie(data)
```

```
import seaborn as sns

Bar       □ sns.barplot(x=x, y=y)
Line      □ sns.lineplot(x=x, y=y)
Histogram □ sns.histplot(data)
Box       □ sns.boxplot(data)
Scatter   □ sns.scatterplot(x=x, y=y)
Pie       □
```

Seaborn tidak punya fungsi spesifik untuk membuat pie chart karena fokus pada visualisasi statistik daripada data kategori.
Namun kita bisa membuat pie chart menggunakan matplotlib yang bisa dimodifikasi.

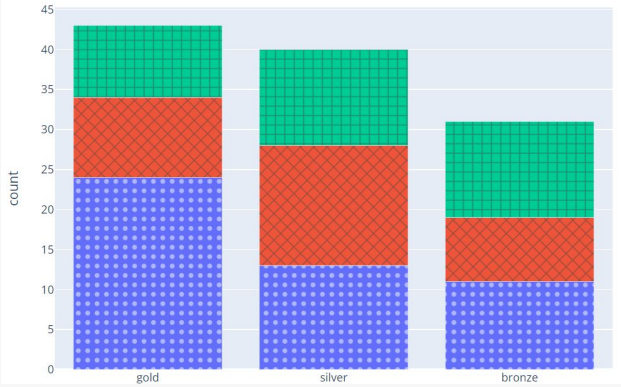
Jika butuh tampilan pie chart yang advanced, bisa menggunakan library lain seperti Plotly Express.

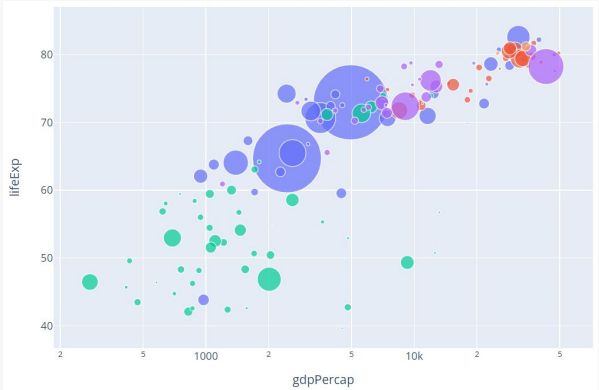
Other Visualization

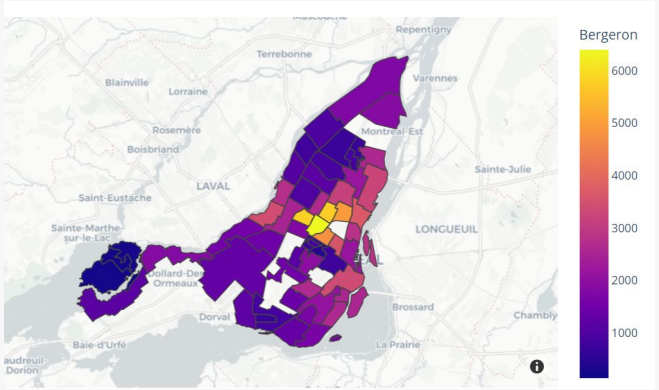
Plotly Express

```
import plotly.express as px
```

Bar	px.bar(DataFrame, x=x, y=y)
Line	px.line(DataFrame, x=x, y=y)
Histogram	px.histogram(data)
Box	px.box(data)
Scatter	px.scatter(DataFrame, x=x, y=y)
Pie	px.pie(names=labels, values=values)







Dengan plotly express, kita bisa berinteraksi langsung pada grafiknya dengan mengarahkan ke elemen tertentu

Exploratory Data Analysis

Exploratory Data Analysis

Kenapa?

Membersihkan data
- Data real itu sangat kotor
- Harus paham data yang missing dan duplikat

Mengerti data
- Eksplor data untuk dapat insights
- Lihat statistical summary
- Analisis univariate
- Analisis multivariate

Seleksi fitur untuk model
- EDA bisa deteksi fitur yang kemungkinan berpotensi untuk digunakan
- Fitur yang saling berhubungan sehingga memilih 1 fitur saja dari fitur tersebut

Statistical Summary

Hal paling pertama yang harus dilakukan adalah melihat summary statistic dari dataset.
Hal ini dilakukan untuk melihat distribusi dari tiap kolom.

- Bagaimana data min dan max?
- Distribusinya normal atau skewed?
- Bagaimana frekuensi dari data kategori

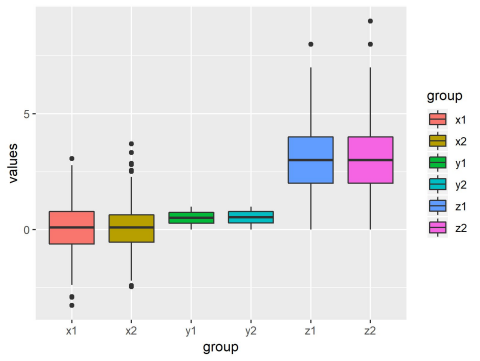
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
count	486.000000	486.000000	486.000000	486.000000	506.000000	506.000000	486.000000
mean	3.611874	11.211934	11.083992	0.069959	0.554695	6.284634	68.518519
std	8.720192	23.388876	6.835896	0.255340	0.115878	0.702617	27.999513
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000
25%	0.081900	0.000000	5.190000	0.000000	0.449000	5.885500	45.175000
50%	0.253715	0.000000	9.690000	0.000000	0.538000	6.208500	76.800000
75%	3.560263	12.500000	18.100000	0.000000	0.624000	6.623500	93.975000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000

- Jumlah data yang muncul terdapat perbedaan, berarti ada beberapa yang missing value
- Min dan max value terlihat masuk akal
- Mean >> Median berarti distribusinya skewed
- Mean ~ Median berarti distribusinya relatif simetris

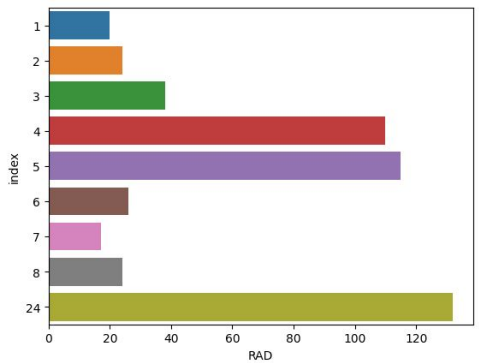
Analisis Univariate

Visualisasikan tiap kolom agar lebih memahami datanya

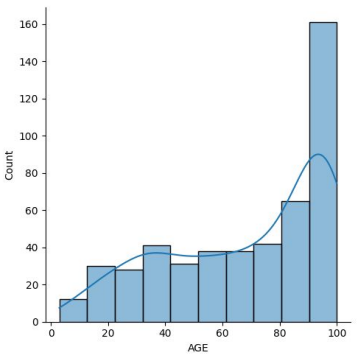
Deteksi Outlier



Melihat frekuensi



Melihat distribusi

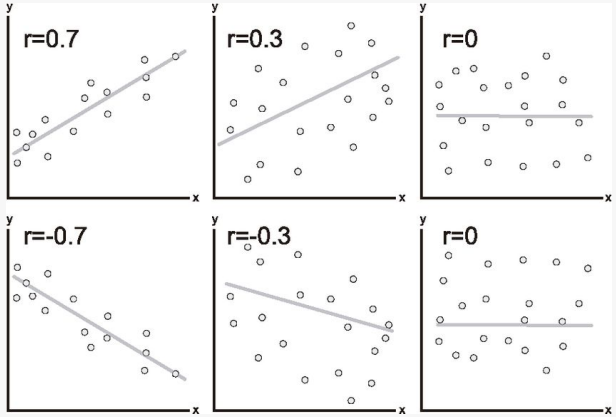


- Terdapat beberapa outlier, pahami konteks datanya untuk handling outlier tersebut
- Index 24 memiliki frekuensi terbanyak
- Distribusi AGE terlihat skewed, pahami datanya apakah perlu dilakukan transformasi.

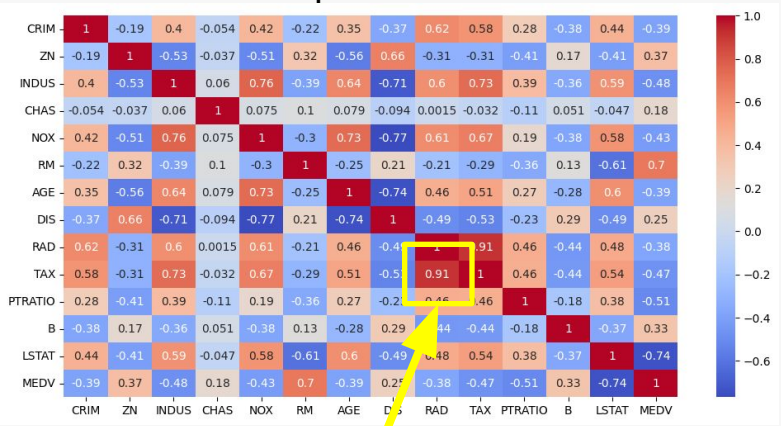
Analisis Multivariate

Lakukan analisa beberapa variable bersamaan untuk melihat hubungannya

Pearson's Correlation



Heatmap Correlation



TAX dan RAD berhubungan kuat

Deep Dive Exploration

Buat beberapa pertanyaan :

- Apakah harga rata-rata rumah pada CHAS 0 lebih besar daripada CHAS 1?
- 10 usia rumah dengan jumlah terbanyak?
- Berapa % rumah yang harganya lebih besar dari 20?
- Dan lainnya

Deep Dive Exploration

Apakah harga rata-rata rumah pada CHAS 0 lebih besar daripada CHAS 1?

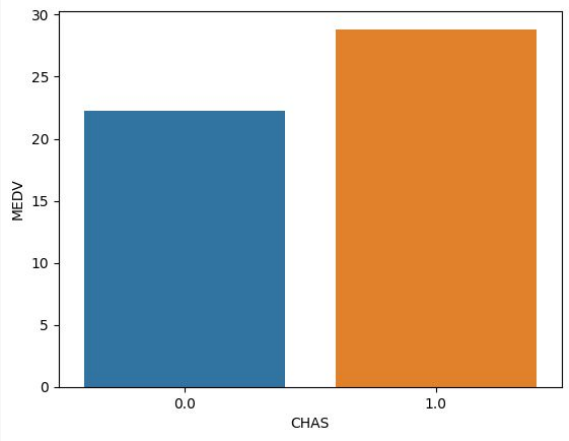
Step:

- Hitung rata-rata pada masing-masing CHAS
- Visualisasikan

Code :

- df_mean = df.groupby('CHAS')[‘MEDV’].mean().reset_index()

- sns.barplot(df_mean, x=’CHAS’, y='MEDV')



Deep Dive Exploration

10 usia rumah dengan jumlah terbanyak

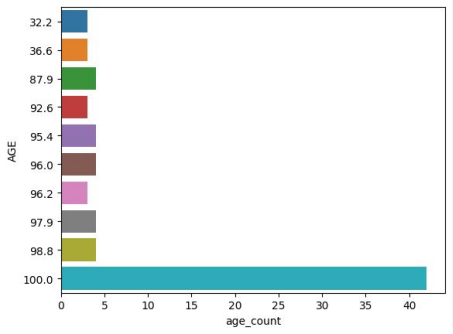
Step:

- Hitung jumlah data tiap usia rumah
- Urutkan dari umur yang mempunyai jumlah terbesar
- Ambil 10 data teratas
- Visualisasikan

Code :

```
- count_age = df.groupby('AGE').agg(age_count = ('AGE', 'count'))
- count_age = count_age.sort_values('age_count', ascending=False).reset_index()
- count_age = count_age.head(10)

- sns.barplot(count_age, y='AGE', x='age_count', orient='h')
```



Deep Dive Exploration

Berapa % rumah yang harganya lebih besar dari 20?

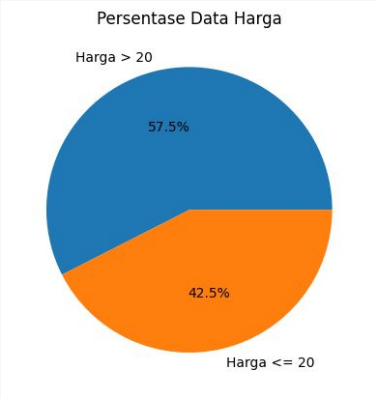
Step:

- Hitung jumlah masing-masing di atas dan di bawah 20
- Jadikan satu dataframe baru
- Visualisasikan menggunakan pie chart

Code :

```
- above_20 = df.loc[df['MEDV'] > 20, 'MEDV'].count()
- below_20 = df.loc[df['MEDV'] <= 20, 'MEDV'].count()
- result_df = pd.DataFrame({'Keterangan': ['Harga > 20', 'Harga <= 20'], 'Jumlah Data': [above_20, below_20]})

- plt.pie(result_df['Jumlah Data'], labels=result_df['Keterangan'], autopct='%1.1f%%')
- plt.title('Persentase Data Harga')
```



Terima Kasih