Data Science

Pokok Bahasan : Visualisasi Data

Dosen Pengampu : Hary Sabita, ST., MTI

05 Juli 2022

Outline :

1. Visualisasi Variabel
- Pie Chart
- Bar Chart
- Line Graphs
- Scatter Plot
- Heatmap

2. Visualisasi Statistik

- Histogram
- Correlation
- Descriptive Statistik
- Grouping (Pivot)
- ANOVA

Next - Apa tujuan mempelajari ini ??

Tujuan yang akan dicapai :

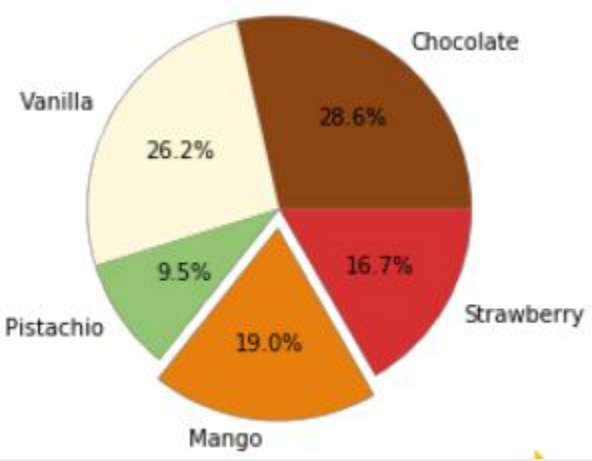
1. Mampu memahami visualisasi data
2. Akan dijelaskan dalam bentuk visualisasi variabel
3. Dapat merencanakan visualisasi data sesuai dengan perencanaan

Next - Pengertian Visualisasi

- Visualisasi berperan peran penting dalam bidang machine learning dan data science. Seringkali kita perlu menyaring informasi kunci yang ditemukan dalam sejumlah data data menjadi bentuk yang bermakna dan mudah dicerna.
- Visualisasi yang baik dapat menceritakan sebuah cerita tentang data Anda dengan cara yang tidak dapat dilakukan oleh sebuah kalimat.

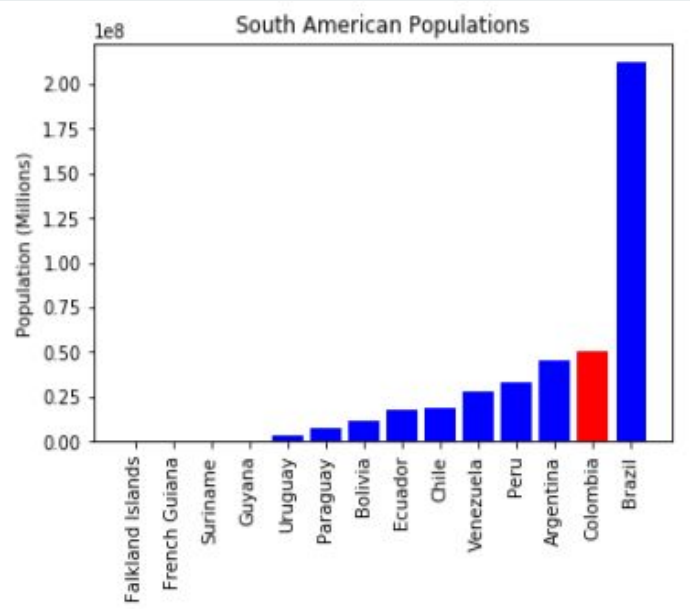
Next - Pie Chart

- Pie chart digunakan untuk menunjukkan seberapa banyak dari setiap jenis kategori dalam dataset berbanding dengan keseluruhan.
  - Variabel label berisi tupel rasa es krim
  - Variabel voting berisi tupel voting.
  - Data tersebut mewakili jumlah voting rase es krim favorit.



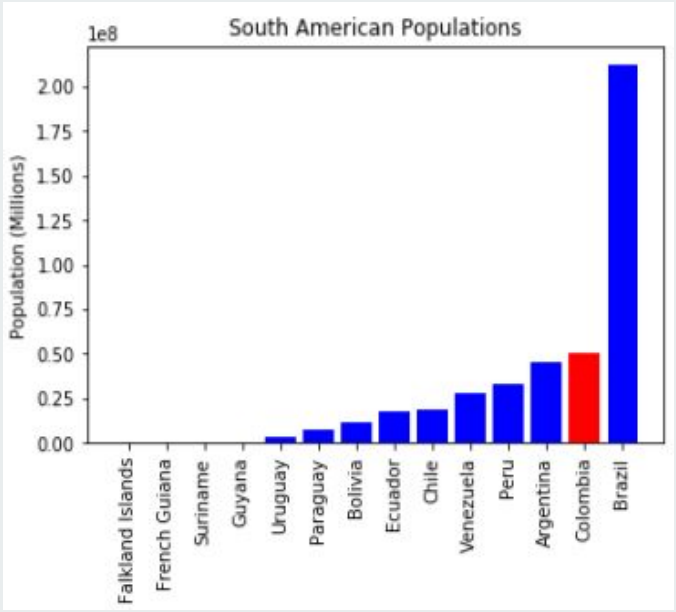
Next - Bar Chart

- Bar Chart adalah merupakan tools visualisassi yang dapat digunakan untuk membandingkan data kategorikal.
- Mirip dengan diagram lingkaran, diagram ini dapat digunakan untuk membandingkan kategori data satu sama lain.
- Diagram batang dapat menampilkan lebih banyak kategori data daripada diagram lingkaran.



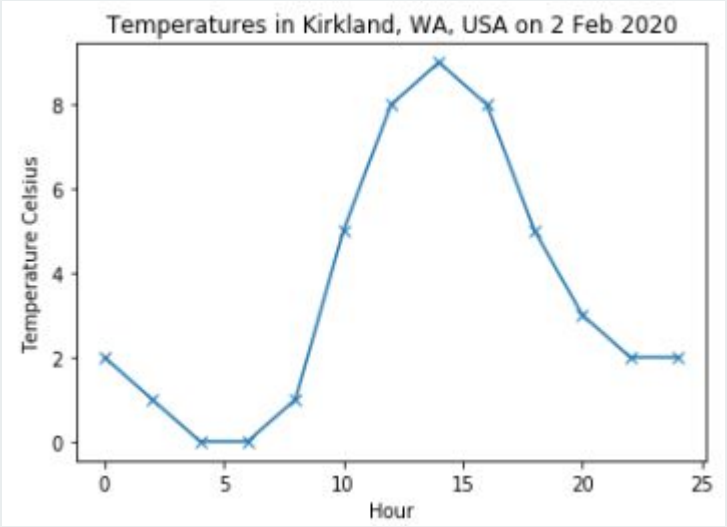
Next - Bar Chart

- Mari kita mulai dengan melihat diagram batang yang menunjukkan populasi setiap negara di Amerika Selatan.
- Visualisasi ditunjukkan dengan cara mengurutkan dari negara yang memiliki populasi terbesar ke populasi terendah.
- Highilght ditunjukkan untuk negara Colombia



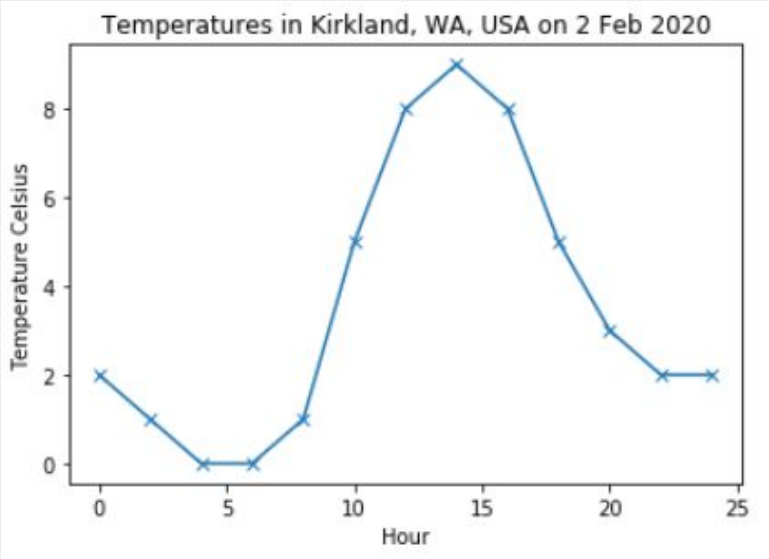
Next - Line Graph

- Line Graph adalah bentuk visualisasi lainya selain diagram lingkaran dan diagram batang.
- Diagram garis lebih berguna untuk menunjukkan bagaimana kemajuan data selama beberapa periode.
- Misalnya, grafik garis dapat berguna dalam membuat grafik temperatur dari waktu ke waktu, harga saham dari waktu ke waktu, berat menurut hari, atau metrik berkelanjutan lainnya.



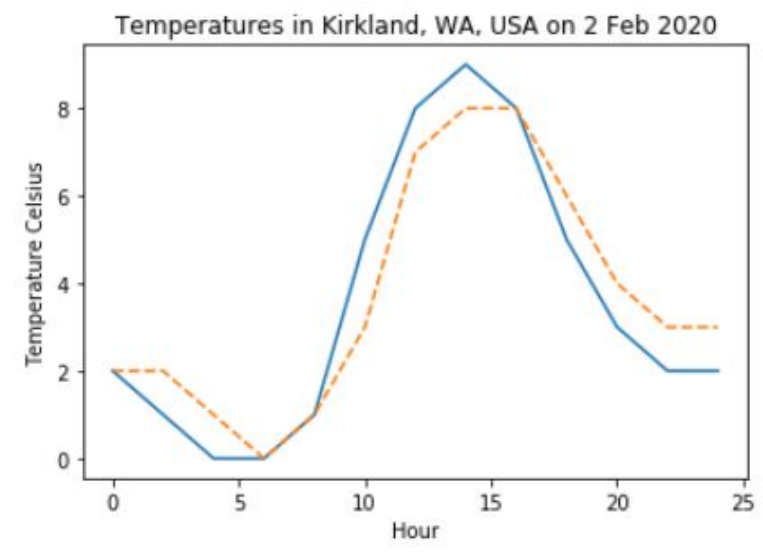
Next - Line Graph

- Kita akan membuat grafik garis yang sangat sederhana di bawah ini. Data yang kita miliki adalah suhu dalam celsius dan jam dalam sehari untuk satu hari dan lokasi.
- Anda dapat melihat bahwa untuk membuat grafik garis kita menggunakan metode plt.plot ().



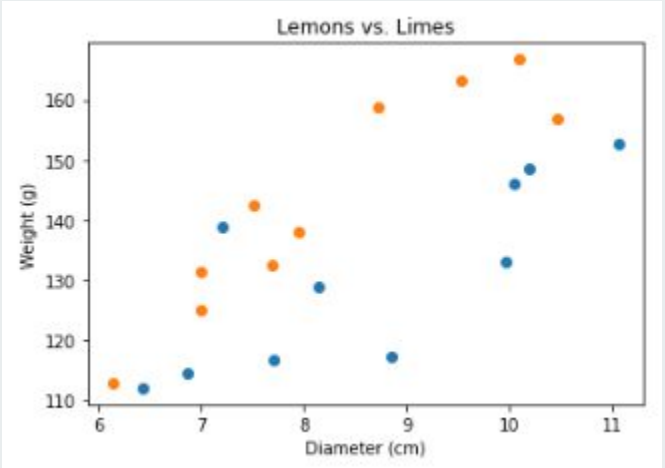
Next - Line Graph

- Kita bahkan dapat memiliki beberapa garis pada grafik yang sama didalam satu gambar
- Biasanya kita mengilustrasikan dua line graph untuk menggmbarkan dua data yitu data aktual dan data prediksi.



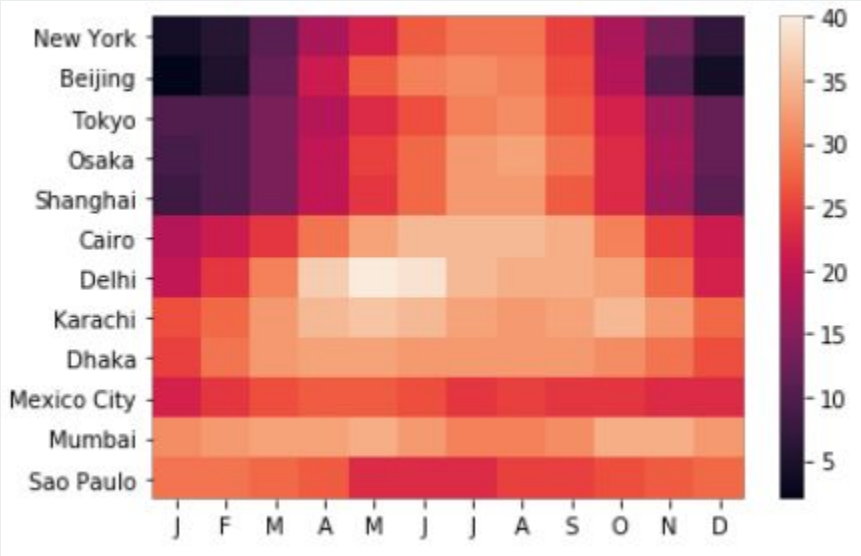
Next - Scatter Plot

- Scatter plot berfungsi baik untuk data dengan dua komponen numerik.
- Scatter plot dapat memberikan informasi yang berguna terutama mengenai pola atau pencilan.
- Pada contoh di bawah ini, kita memiliki data yang terkait dengan perbedaan lemon dan lime berdasarkan karakteristik fisiologis.
  - Berat (g)
  - Diameter (cm)



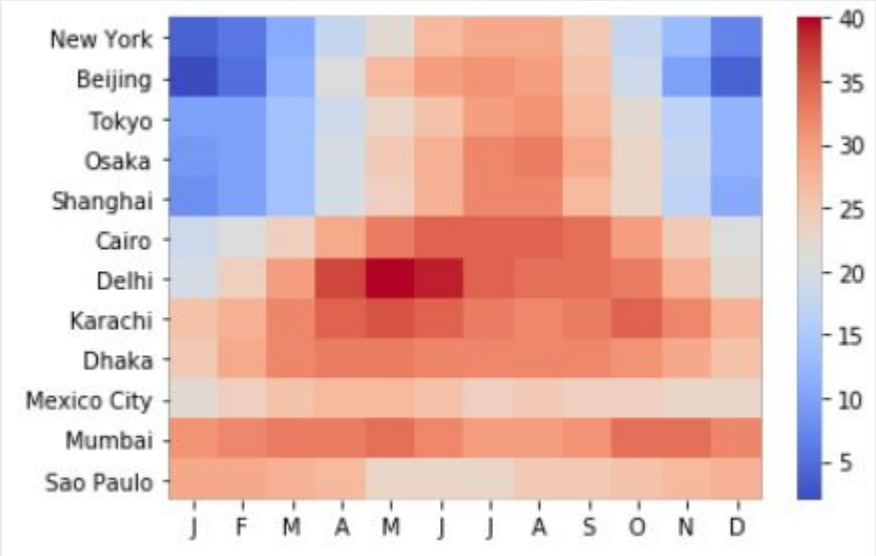
Next - Heatmap

- Heatmap adalah jenis visualisasi yang menggunakan kode warna untuk mewakili nilai / kepadatan relatif data di seluruh permukaan.
- Warna-warna ini kemudian dapat digunakan untuk memeriksa data secara visual guna menemukan kelompok dengan nilai serupa dan mendeteksi tren dalam data.



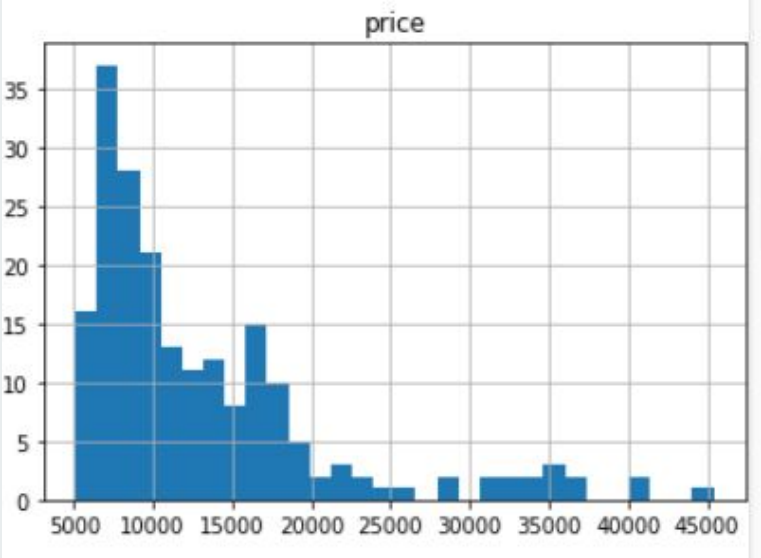
Next - Heatmap

- Kita akan bekerja dengan data tentang temperatur rata-rata setiap bulan untuk 12 kota terbesar di dunia. Untuk membuat heatmap ini, kita akan menggunakan library Seaborn.
- Seaborn adalah library visualisasi yang dibangun di atas Matplotlib.
- Library ini menyediakan antarmuka tingkat yang lebih tinggi dan dapat membuat grafik yang lebih menarik



Next - Histogram

- Histogram adalah salah satu visualisasi yang cukup penting dalam memahami distribusi pada data kita. Pandas Histogram menyediakan method yang memudahkan kita untuk membuat histogram.
- Plot histogram secara tradisional hanya membutuhkan satu dimensi data.
- Ini dimaksudkan untuk menunjukkan jumlah nilai atau kumpulan nilai secara serial.



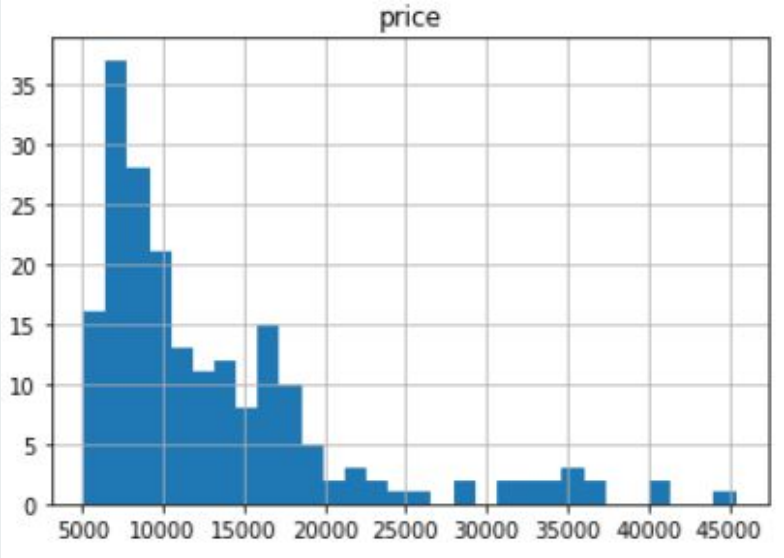
Next - Histogram

- Data yang digunakan adalah data spesifikasi mobil dari berbagai merk

	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	curb-weight	engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke	compre
0	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	0.890278	48.8	2548	dohc	four	130	mpfi	3.47	2.68	
1	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	0.811148	0.890278	48.8	2548	dohc	four	130	mpfi	3.47	2.68	
2	1	122	alfa-romero	std	two	hatchback	rwd	front	94.5	0.822681	0.909722	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	
3	2	164	audi	std	four	sedan	fwd	front	99.8	0.848630	0.919444	54.3	2337	ohc	four	109	mpfi	3.19	3.40	
4	2	164	audi	std	four	sedan	4wd	front	99.4	0.848630	0.922222	54.3	2824	ohc	five	136	mpfi	3.19	3.40	

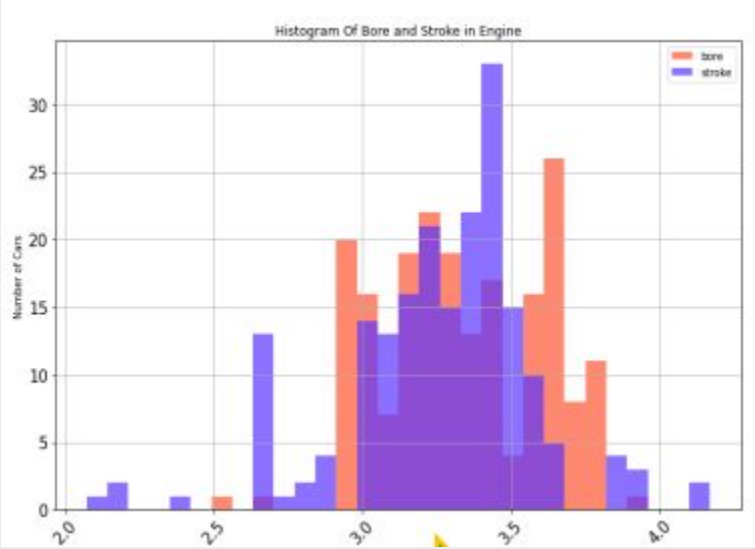
Next - Histogram

- Pandas DataFrame.hist() akan mengambil DataFrame kita dan menampilkan plot histogram yang menunjukkan distribusi nilai dalam satu seri.
- Untuk membuat histogram di panda, yang perlu kita lakukan adalah memberi tahu panda kolom mana yang ingin kita berikan datanya. Dalam hal ini, saya akan memberi tahu panda bahwa saya ingin melihat distribusi harga (histogram).



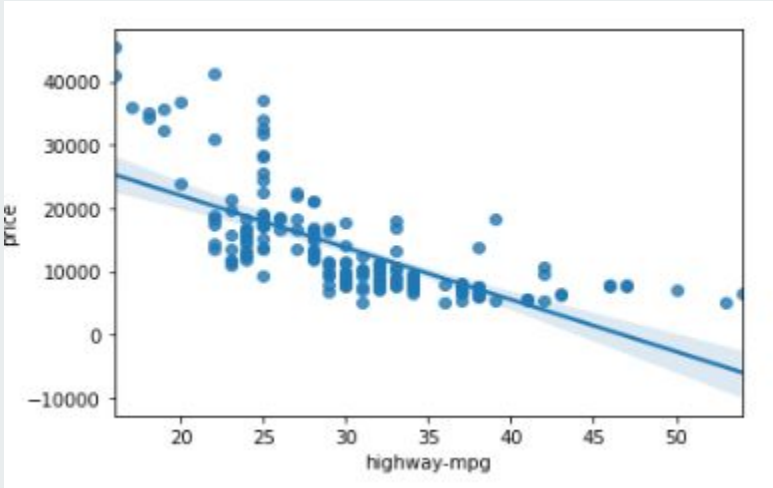
Next - Histogram

- Kita juga dapat memplot beberapa grup secara berdampingan. Di sini saya ingin melihat dua histogram, histogram price akan dikelompokkan berdasarkan roda penggerak dari kenderaan (fwd – berpenggerak roda depan, 4wd – berpenggerak 4 roda, atau rwd – pengerak belakang.



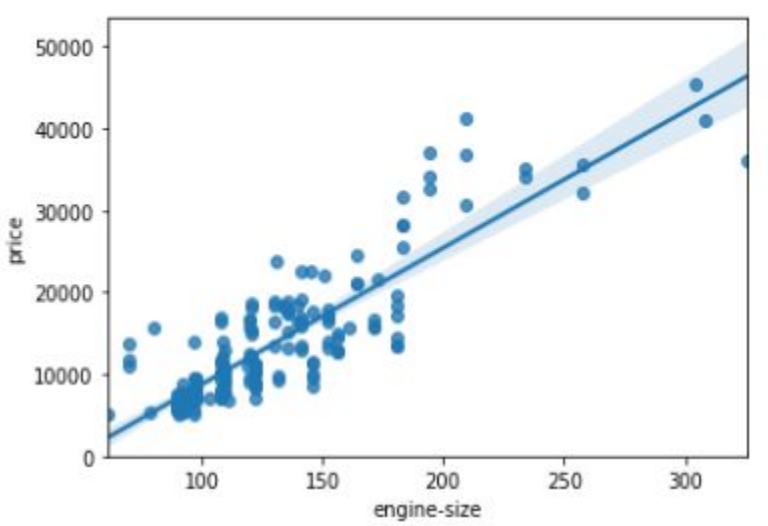
Next - Correlation & Causation

- Korelasi merupakan suatu pengukuran sejauh mana nilai saling ketergantungan antar variabel.
- Causation merupakan hubungan antara sebab dan akibat antara dua variable
- Penting untuk mengetahui perbedaan antara keduanya dan bahwa korelasi tidak mendeskripsikan sebab-akibat.
- Menentukan korelasi jauh lebih sederhana menentukan sebab memerlukan analisis lebih lanjut



Next - Correlation & Causation

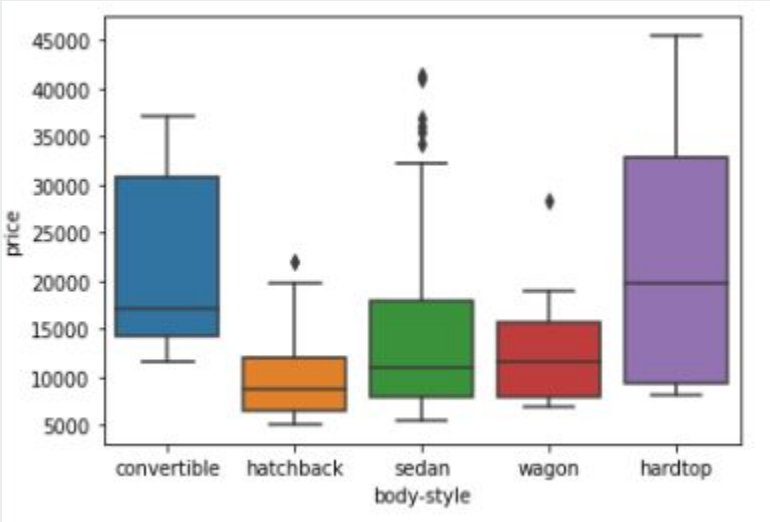
- Korelasi Pearson
- Pearson Correlation adalah metode default dari fungsi "corr". Kita dapat menghitung Korelasi Pearson dari variabel 'int64' atau 'float64'. Terkadang kita ingin mengetahui signifikansi dari estimasi korelasi, kita dapat menggunakan p-value.



- Korelasi Pearson mengukur ketergantungan linier antara dua variabel X dan Y.

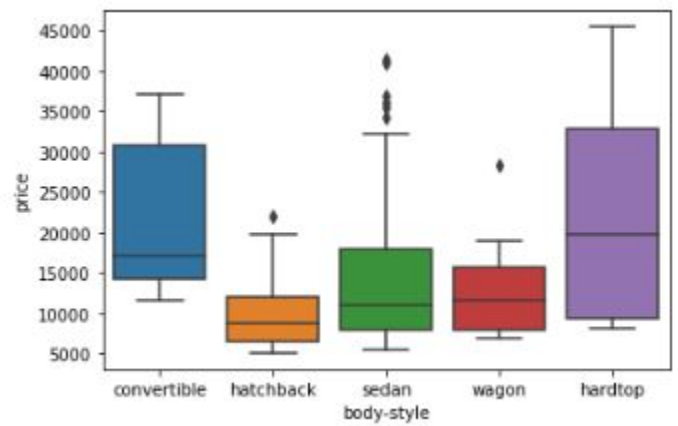
Next - Variabel Kategori Statistik

- Ini adalah variabel yang menggambarkan 'karakteristik' dari unit data, dan dipilih dari sekelompok kategori. Variabel kategori dapat memiliki tipe "objek" atau "int64". Cara yang baik untuk memvisualisasikan variabel kategori adalah dengan menggunakan boxplot.
- Boxplot menggambarkan variable variable statistic seperti quartil 1, median / quartil 2, quartil 3, nilai maksimum, nilai minimum, dan outlier.



Next - Deskriptive Statistik

- Fungsi deskripsikan secara otomatis menghitung statistik dasar untuk semua variabel kontinu.
- Analisis yang bisa kita dapatkan dari deskriptif statistik adalah
  - Jumlah variabel
  - Rata-rata
  - Standard deviasi
  - Nilai minimal
  - IQR (Interquartile Range: 25%, 50% and 75%)
  - Nilai Maximal



Next - Grouping

- Method "groupby" digunakan untuk mengelompokkan data menurut kategori yang berbeda. Data dikelompokkan berdasarkan satu atau beberapa variabel dan analisis dilakukan pada kelompok individu.
- Sebagai contoh, mari kita kelompokkan berdasarkan variabel "roda penggerak". Kita melihat bahwa ada 3 kategori roda penggerak yang berbeda.

- `df['drive-wheels'].unique()`

```
array(['rwd', 'fwd', '4wd'], dtype=object)
```

Next - Grouping

- Anda juga dapat mengelompokkan dengan beberapa variabel. Misalnya, mari kita kelompokkan berdasarkan 'roda penggerak' dan body-style'.
- Ini mengelompokkan dataframe dengan kombinasi unik 'drive-wheels' dan 'body-style'. Kita dapat menyimpan hasilnya dalam variabel 'grouped_test1'.

	drive-wheels	body-style	price
0	4wd	hatchback	7603.000000
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000
7	fwd	wagon	9997.333333
8	rwd	convertible	23949.600000
9	rwd	hardtop	24202.714286
10	rwd	hatchback	14337.777778
11	rwd	sedan	21711.833333
12	rwd	wagon	16994.222222

Next - Grouping

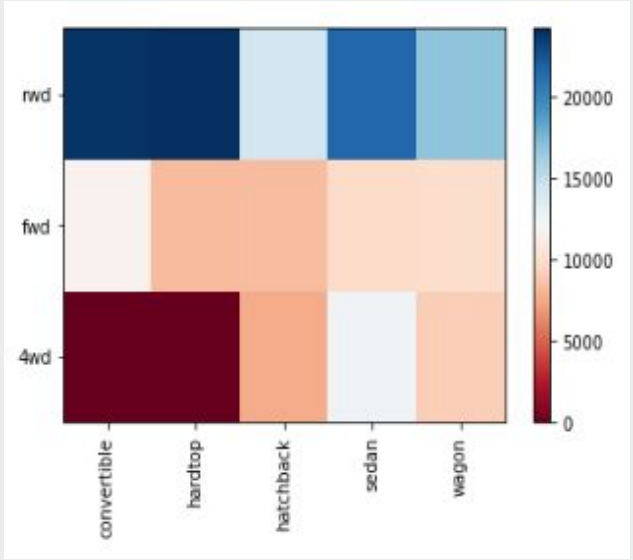
	price				
body-style	convertible	hardtop	hatchback	sedan	wagon
drive-wheels					
4wd	0.0	0.000000	7603.000000	12647.333333	9095.750000
fwd	11595.0	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.6	24202.714286	14337.777778	21711.833333	16994.222222

- Data yang dikelompokkan ini jauh lebih mudah untuk divisualisasikan ketika dibuat menjadi tabel pivot.
- Tabel pivot yang mirip seperti pada spreadsheet Excel, dengan satu variabel di sepanjang kolom dan variabel lainnya di sepanjang baris.
- Kita dapat mengonversi kerangka data menjadi tabel pivot menggunakan metode "pivot" untuk membuat tabel pivot dari grup.

Next - Grouping

- Dari table pivot kita dapat mengilustrasikan table pivot dalam bentuk heatmap.

	price				
body-style	convertible	hardtop	hatchback	sedan	wagon
drive-wheels					
4wd	0.0	0.000000	7603.000000	12647.333333	9095.750000
fwd	11595.0	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.6	24202.714286	14337.777778	21711.833333	16994.222222



Next - Anova

- Analysis of Varians (ANOVA) adalah metode statistik yang digunakan untuk menguji apakah ada perbedaan yang signifikan antara rata-rata dua kelompok atau lebih.
- ANOVA mengembalikan dua parameter
  - F-Score:
  - P-Value

- F-Score: ANOVA mengasumsikan rata-rata semua kelompok adalah sama, anova akan menghitung seberapa jauh rata-rata yang sebenarnya menyimpang dari asumsi, dan melaporkannya sebagai F-Score.
- Skor yang lebih besar berarti ada perbedaan yang lebih besar antara rata-rata.
- P-Value: Nilai-P menunjukkan seberapa signifikan secara statistik nilai skor yang dihitung.

Next - Anova

- Jika variabel harga pada dataset mobil sangat berkorelasi dengan variabel lainya, ANOVA akan mengembalikan skor F-Score yang cukup besar dan nilai-p yang kecil.
- ANOVA menganalisis perbedaan antara kelompok yang berbeda dari variabel yang sama, fungsi groupby akan berguna dalam kasus ANOVA.
- Mari kita lihat apakah jenis 'roda penggerak' mempengaruhi 'harga',

```
# grouping results
df_gptest = df[['drive-wheels','body-style','price']]
grouped_test1 = df_gptest.groupby(['drive-wheels','body-style'],as_index=False).mean()
grouped_test1
```

	drive-wheels	body-style	price
0	4wd	hatchback	7603.000000
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000
7	fwd	wagon	9997.333333
8	rwd	convertible	23949.600000
9	rwd	hardtop	24202.714286
10	rwd	hatchback	14337.777778
11	rwd	sedan	21711.833333
12	rwd	wagon	16994.222222

Next - Anova

TERIMA KASIH