3 Bab 3: Angka Diskrit dan Kontinu

Bab ini membahas angka diskrit dan kontinu.

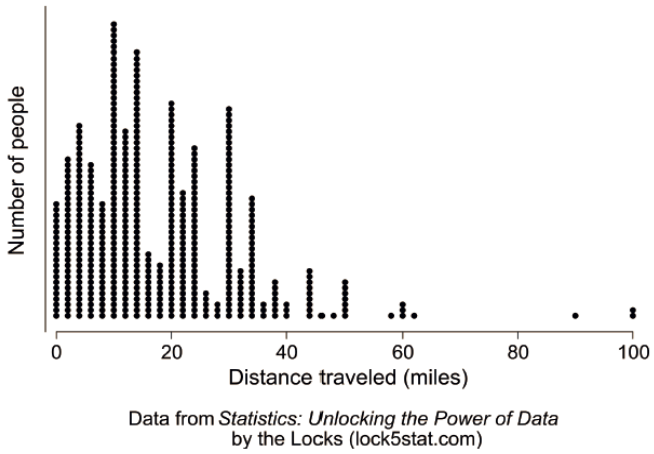
Cara terbaik untuk mempelajari visualisasi data adalah dengan mempelajari variabel kuantitatif. Variabel kuantitatif berarti memiliki rentang nilai numerik yang dapat diukur di dunia nyata. Variabel ini terdiri dari *variabel kontinu* dan *variabel diskrit*. Variabel kontinu dapat memiliki nilai yang terdiri dari banyak desimal. Variabel diskrit hanya dapat memiliki nilai tertentu, biasanya 0, 1, 2, 3 , atau disebut dalam matematika sebagai bilangan asli. Variabel yang memiliki unit pengukuran (misalnya, hektar hutan, atau jam usia pakai baterai untuk ponsel), sudah pasti kontinu atau diskrit. Tetapi bukan berarti variabel yang tidak ada unit pasti variabel kontinu. Misal rasio laki-laki terhadap perempuan di antara pelamar pekerjaan.

Bentuk data lainnya adalah kategori, yang menunjukkan bahwa masing-masing pengamatan jatuh ke dalam satu kategori. Kategori-kategori itu kadang-kadang memiliki urut-urutan alami. Urutan itu tidak menjadikan variabel itu variabel diskrit. Variabel diskrit memiliki makna numerik sebenarnya, sebagai contoh angka populasi suatu kota dapat dikurangi angka populasi kota lain yang artinya perbedaan populasi dua kota. Sebaliknya data kategori ordinal misal Sangat Tidak Setuju, Tidak Setuju, Setuju, dan Sangat Setuju, atau bahkan jika direkam sebagai angka 1, 2, 3 dan 4, angka ini tidak dapat dioperasikan secara matematis.

3.1 Satu variabel pada satu waktu

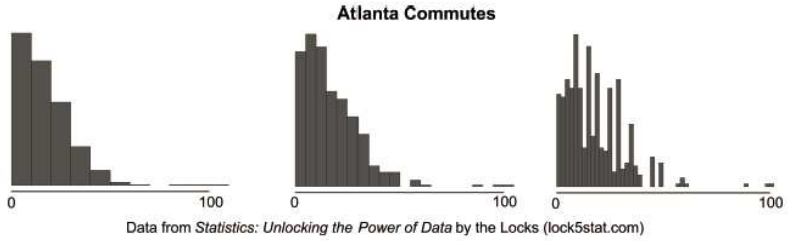
Mari kita pelajari satu variabel yang berisi data kontinu: jarak perjalanan orang bolak-balik untuk bekerja di kota Atlanta, Georgia, Amerika Serikat. Ada beberapa aspek yang dapat ditemukan. Jumlahnya sudah pasti positif, dan diduga sebagian besar dari angka berada di bawah dua puluh mil. Masuk akal untuk menyandikan jarak komuter ke sumbu horizontal, dari nol mil hingga jarak perjalanan maksimum di sebelah kanan. Jika titik untuk tiap orang dalam data ini disusun per jarak 2 mil, maka akan didapatkan Gambar 3.1. Grafik ini disebut *strip chart*, terkadang *dot plot*.

Penamaan yang membingungkan akan sering ditemukan dalam visualisasi data. Visualisasi harus selalu memberi tahu pembaca apa variabelnya, lebih baik memberi label sumbu dengan *Distance traveled* (Jarak bepergian) (mil). Satuan (mil) harus dibuat jelas. Jika ditemukan visualisasi yang membingungkan dimana tidak jelas apa mewakili parameter visual apa, maka berarti analis telah keliru. Jumlah orang di setiap jarak/spasi dibuat jelas. Dalam *strip chart / dot plot*, ketinggian setiap tumpukan titik sudah jelas maknanya. Jika kita menyandikannya dalam cara lain misal permainan warna, maka tidak akan mudah bagi pembaca membandingkan jumlah orang yang menempuh jarak perjalanan yang berbeda.



Gambar 3.1 *Strip Chart* atau *dot plot* suatu variabel kontinu.

Namun, ada keterbatasan untuk bagan dasar seperti ini. Jika datanya banyak, maka akan banyak titik pada *strip chart/ dot plot* titik-titik akan sangat tinggi atau titik-titik harus sangat kecil. Format yang lebih baik untuk penyandian yang sama dapat ditangani dengan histogram. Caranya potong variabel kontinu dalam satu wadah, cacah berapa banyak pengamatan di masing-masing wadah, lalu gambar wadah setinggi hitungannya (Gambar 3.2).



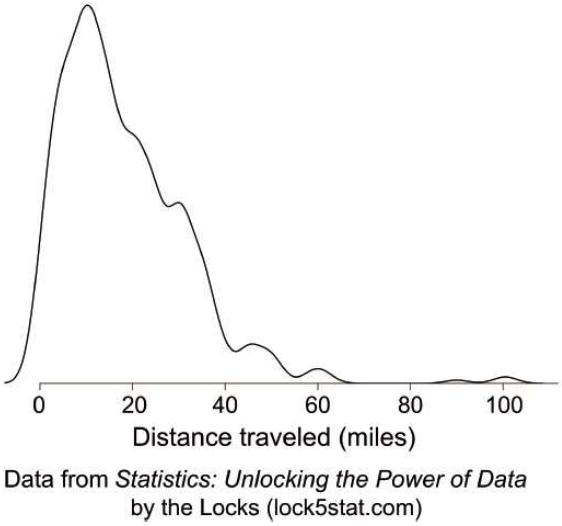
Gambar 3.2 Histogram suatu variabel kontinu. Kita dapat memilih banyak wadah yang akan digunakan untuk mencacah data. Sepuluh wadah (kiri) tampaknya kehilangan detail terlalu banyak, dua puluh wadah (tengah) terlihat baik, dan lima puluh wadah (kanan) begitu banyak sehingga mulai menunjukkan efek penumpukan dan "Noise" dalam data.

Histogram dapat menjelaskan tentang distribusi data dan banyak digunakan dalam visualisasi data serta dianggap sebagai standar dan dasar bentuk visualisasi data. Histogram ternyata juga menjadi konsep penting untuk Big Data, dan akan dibahas dalam bab selanjutnya.

Sangat mudah untuk mengetahui bagaimana histogram dihitung yaitu dengan cara menghitung data. Ada juga alternatif lain yang disebut *kernel density plot*/"plot densitas kernel" dengan cara mengganti setiap pengamatan dengan bentuk halus yang bagus (disebut kernel) seperti distribusi normal. Ketika ketinggian semua kurva kernel ditambahkan bersama di setiap titik sepanjang sumbu horisontal, dan hasilnya diplot, maka akan membentuk garis halus yang mirip bentuk seperti di histogram.

Plot densitas kernel merupakan format praktis untuk digunakan dalam visualisasi karena membutuhkan lebih sedikit "tinta" untuk menggambarnya, dan dapat dicerna oleh pembaca dengan cepat. Namun pembuatannya mengandalkan komputer untuk

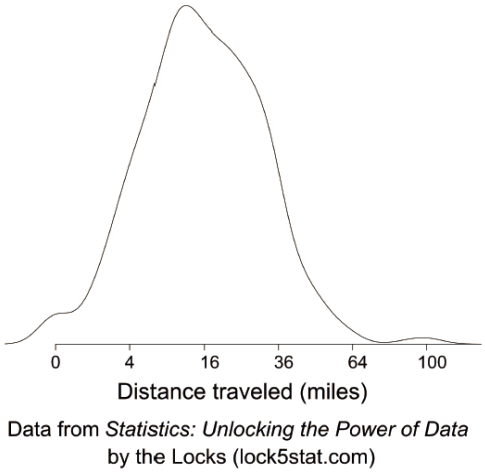
menghitungnya, dan mungkin tidak tersedia di setiap perangkat lunak.



Gambar 3.3 *Kernel Density Plot* suatu variabel kontinu.

Terkadang distribusi data sangat condong, ini dapat diolah terlebih dahulu misalnya dengan menggunakan fungsi matematika. Fungsi logaritma populer untuk hal ini, karena fungsi akan mentransformasi angka yang tinggi ke nilai yang rendah, sehingga mengurangi kemiringan positif (*positive skew*).

Bagan yang menggunakan transformasi seperti ini perlu direfleksikan pada sumbu, dan perlu dicatat dalam teksnya juga sehingga lebih jelas bagi pembaca. Dalam Gambar 3.4, data disandikan dalam akar kuadrat dari data jarak perjalanan ke sumbu horisontal, dan ini membuat grafik terlihat lebih simetris.



Gambar 3.4 Jarak komuter dalam *density plot* setelah ditransformasi akar.

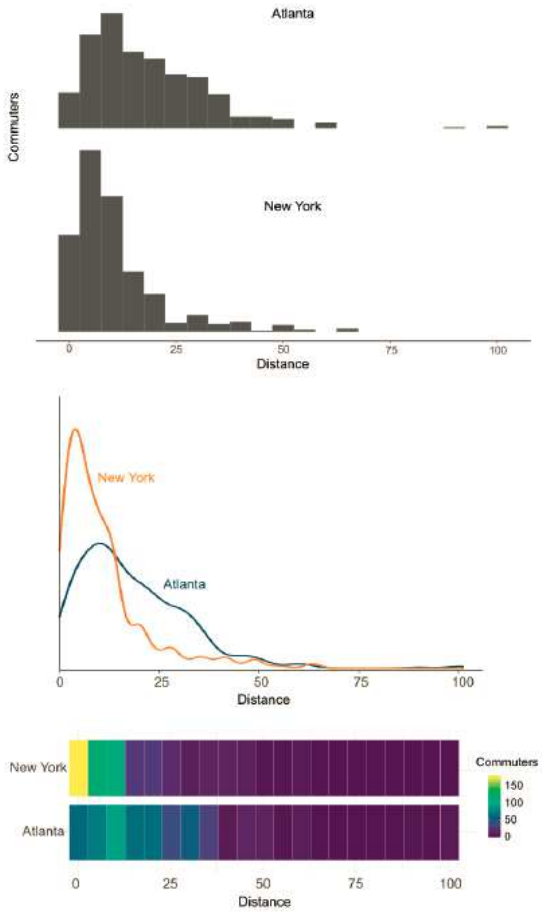
3.2 Perbandingan Data Tidak Sepadan

Misalkan terdapat beberapa data dari satu variabel dan beberapa dari variabel lainnya dan kami ingin dibandingkan. Gambar 3.5 menunjukkan tiga cara membandingkan waktu perjalanan di Atlanta dan New York. Jarak dikodekan ke lokasi horizontal, jika ingin membuat perbandingan mudah bagi pembaca, maka harus diatur beberapa visualisasi seperti histogram tumpang tindih satu sama lain, dan bukan berdampingan. *Kernel Desity* merupakan mereka kurva yang sederhana, sehingga dapat ditumpangkan secara efektif.

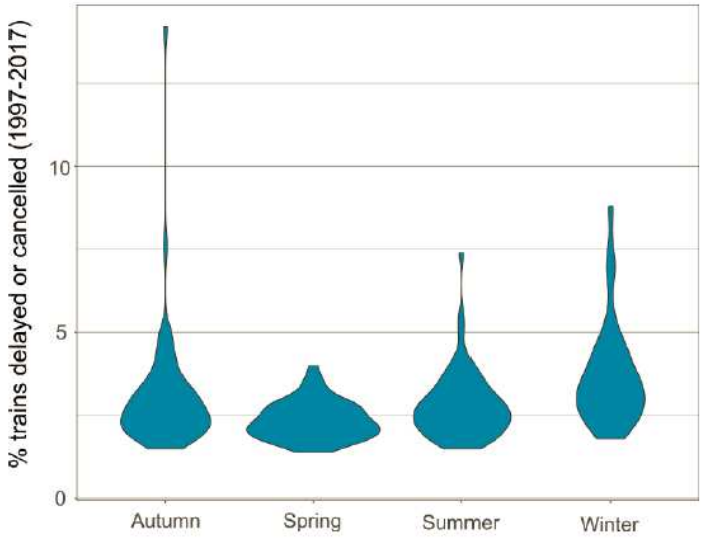
Opsi ketiga di sini adalah pita *heat map*, yang mencacah komuter dalam wadah seperti histogram, tetapi disandikan hitungan dalam warna. Cara ini ringkas dan bisa sangat menarik, tetapi warnanya tidak cocok untuk pemeriksaan yang terperinci. Dalam kasus di atas nampak sekilas warga New York bepergian untuk berkerja dengan jarak yang relatif pendek.

Hal yang sama dapat dilakukan ketika data dibagi menjadi beberapa kelompok dan kemudian dibandingkan nilai-nilai satu variabel lintas grup. Dua histogram ditumpuk di atas satu sama lain dapat menunjukkan perbandingan visual meskipun tidak sempurna. Dua *Density Plot* sebenarnya dapat ditumpangkan untuk mendapatkan hasil yang lebih jelas. Tetapi dengan bertambahnya jumlah kelompok

atau variabel untuk dibandingkan, visualisasi akan menjadi rumit. Ketika itu terjadi maka salah satu opsi yang penting adalah tidak menggambar data secara keseluruhan, dan lebih baik menggambar ringkasan statistik sebagai gantinya. Jika ada banyak variabel, patut dicoba representasi dalam dua dimensi (akan dibahas di bab selanjutnya). Untuk saat ini, varian *Kernel Density plot*, yang disebut plot biola bisa digunakan sebagai alternatif. *Kernel Density* dibalik ke dimensi vertikal, dan dicerminkan di kedua sisi garis (Gambar 3.6). Karena ini cukup kompak, maka dapat diatur cukup banyak grup atau variabel tapi dengan tidak terlalu banyak menjejali informasi ke pembaca.



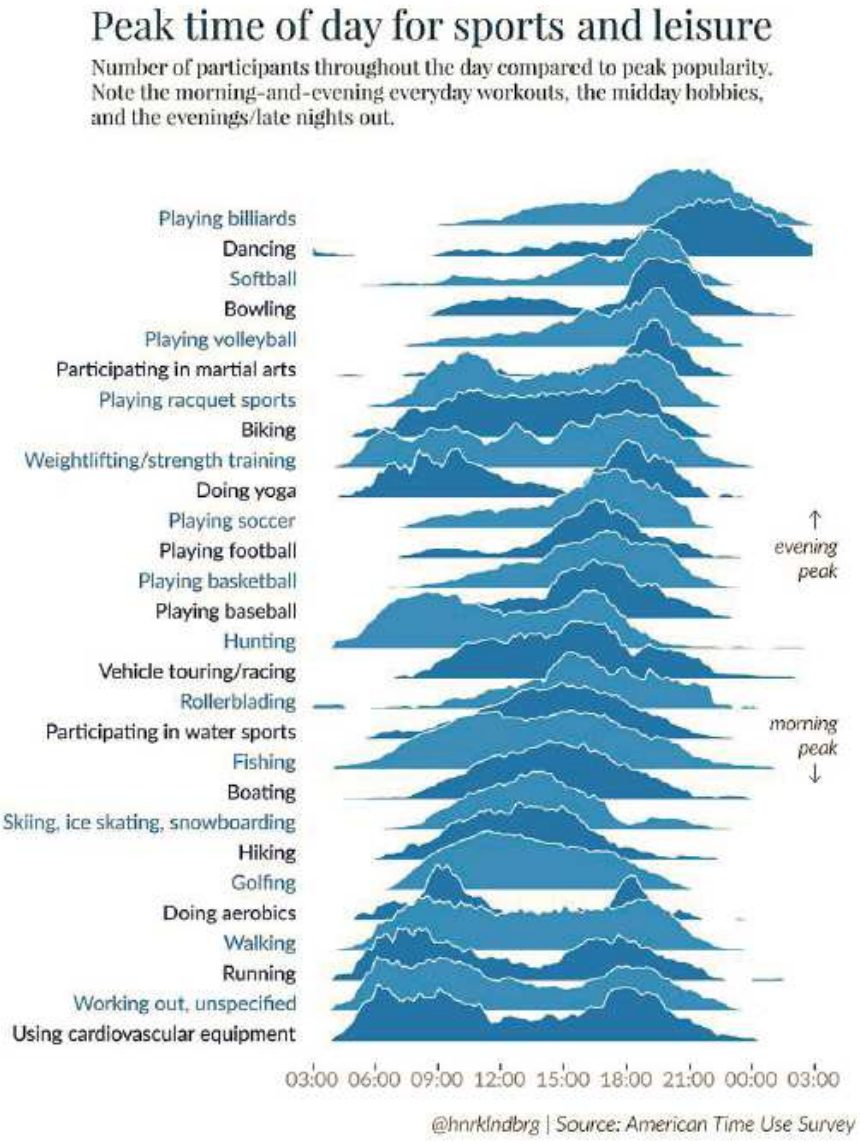
Gambar 3.5 Membandingkan data kontinu dengan histogram bertumpuk, *kernel density plot* yang ditumpangkan, dan pita *heatmap*.



Gambar 3.6 *Plot biola* data keterlambatan kereta, membandingkan empat musim. Tampak jelas tidak ada banyak perbedaan, kecuali "biola" musim dingin sedikit lebih tinggi. Dari sumbu vertikal tampak data berasal dari 20 tahun.

Visualisasi yang panjang dan tipis untuk setiap kelompok juga dapat ditumpuk di atas satu sama lain. Ini bekerja dengan baik pada pita *heatmap*, dan ketika dilakukan dengan *kernel density* itu bisa terlihat sangat jelas, seperti suatu lanskap tiga dimensi, yang membantu pembaca menyerap informasi dengan cepat (Gambar 3.7).

Jika harus dibandingkan beberapa variabel dari beberapa grup, maka akan membutuhkan banyak grafik. Evaluasi apakah pembaca akan memerlukan hal ini. Untuk contoh, orang mungkin tertarik untuk membandingkan waktu perjalanan antara Atlanta dan New York, dan demikian juga waktu kerja. Tetapi mungkin sedikit yang ingin tahu perbandingan waktu perjalanan dan waktu kerja di salah satu kota. Jadi, akan dibuat bagan waktu perjalanan di berbagai kota, dan waktu kerja lainnya di berbagai kota, dan sebagainya untuk berbagai kegiatan harian. Terkadang diperlukan kompromi dengan visualisasi data yang rumit dan berat.

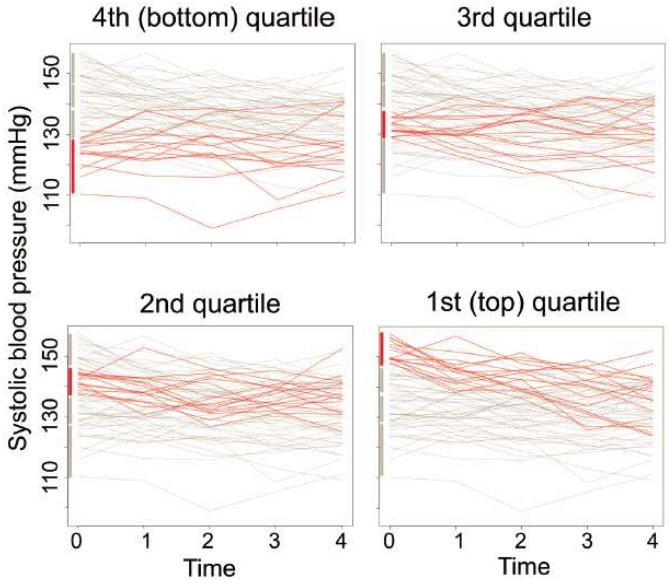


Gambar 3.7 Kepadatan waktu yang ditumpuk untuk berbagai waktu luang kegiatan (perhatikan bahwa tidak ada kernel yang digunakan, jadi tidak ada penghalusan kurva). "Joyplot" oleh Henrik Lindberg.

3.3 Perbandingan Data Sepadan

Ketika dua variabel atau dua kelompok dalam data dibandingkan, dan pengamatan individualnya juga dapat dibandingkan secara bersamaan, maka dikatakan perbandingan datanya sepadan. Contoh umum kasus misalnya data orang yang sama yang diamati pada dua titik waktu berbeda.

Salah satu cara yang baik untuk menunjukkan perubahan tingkat individu ini adalah dengan bagan garis di mana setiap pengamatan memiliki garis dan waktu sendiri dikodekan pada posisi horizontal (Gambar 3.8).



Gambar 3.8 Bagan garis digunakan untuk menunjukkan data sepadan pada kasus penanganan tekanan darah. Empat buah bagan dibuat untuk menunjukkan penekanan pada bagian quartil yang berbeda. Tampak pasien yang pada saat awal memiliki tekanan darah yang tinggi cenderung menurun tekanannya seiring waktu.

Opsi lain yang lebih jelas untuk set data besar adalah mengunakan *scatter plot*, meski kurang intuitif untuk pembaca seperti yang dilihat di Bab 2.

3.4 Asosiasi

Asosiasi antara dua variabel merupakan hal penting dalam statistik dan sain data. Misal pelanggan paling lama adalah orang-orang yang menghabiskan lebih banyak bahan makanan.

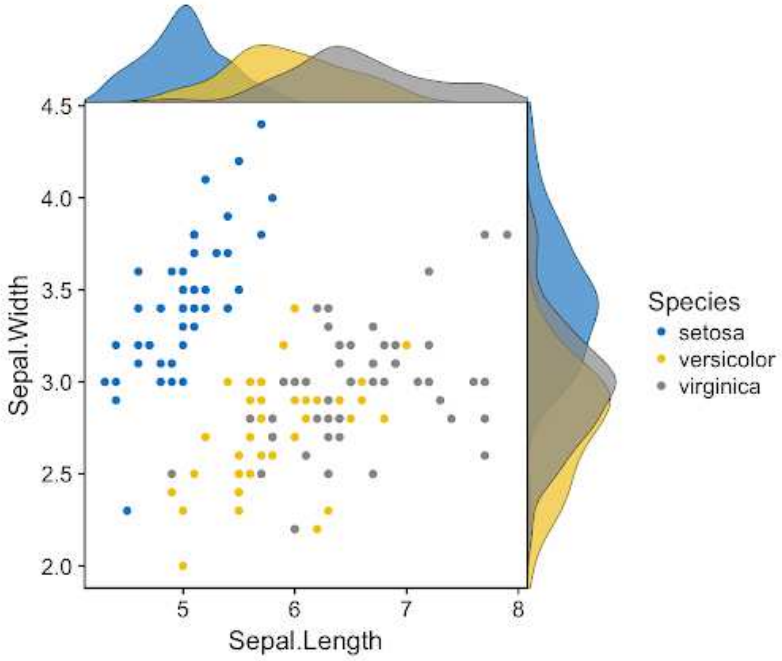
Asosiasi menunjukkan apakah nilai suatu variabel berkaitan dengan variabel lain. Misalnya variabel waktu dengan pengeluaran belanja makanan. Data seperti ini dapat diajadikan bukti apakah misalnya apakah suatu perusahaan menarik perhatian orang yang berpenghasilan lebih tinggi untuk menjadi pelanggan. Atau mungkin sebaliknya, pelanggan yang berpenghasilan rendah tidak lagi menjadi pelanggan. Fenomena seperti ini mungkin tidak menunjukkan sebab-akibat, namun dapat digunakan untuk memprediksi.

Andalan untuk visualisasi pada *Asosiasi* baik pada data diskrit atau kontinu adalah *scatter plot*. Masalah yang umum adalah *marker* mungkin bertumpuk di satu tempat. Hal ini dapat disiasati dengan membuat marker lebih tebal atau lebih gelap, tetapi tetap saja sulit bagi pembaca untuk mengetahui berapa banyak titik yang ada berdasarkan ukuran atau warna marker.

Pendekatan lain adalah dengan *jitter* yaitu dengan menambahkan angka acak kecil ke kedua variabel, yang akan membuat marker lebih tersebar.

Fitur penting yang harus dicari dalam *scatter plot* adalah apakah ada satu klaster titik atau beberapa klaster, apakah ada kemiringan ke atas atau ke bawah pada klaster titik (menunjukkan korelasi), dan apakah ada kurvatur pada *slope*. Perlu dingat bahwa korelasi bukanlah sebab-akibat. Hanya karena satu variabel naik atau turun ketika yang lain naik atau turun tidak berarti yang satu mempengaruhi yang lain.

Terkadang, variabel individu dikodekan secara horisontal dan vertikal dapat ditampilkan bersamaan, kemudian dibuatkan histogram atau *density plot* di luar *scatter plot* (Gambar 3.9).

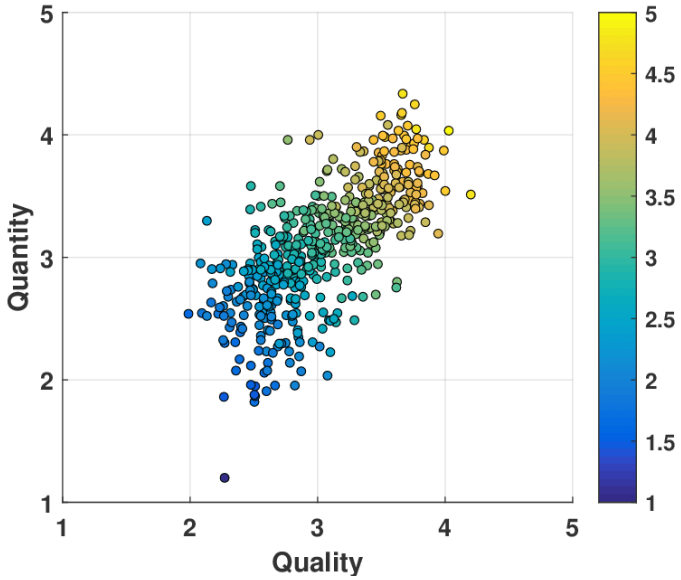


Gambar 3.9 *Scatter Plot* data spesies bunga iris dengan *kernel density plot* di bagian luarnya.

Pilihan lain adalah memiliki garis pendek di margin tempat masing-masing data titik terjadi, yang disebut "barcode marginal," atau kadang-kadang *rug plot*. Konsep ini menjadi dasar ide penting dalam visualisasi data yaitu "distribusi marginal" untuk satu variabel pada suatu waktu, dan "distribusi bersama" untuk *scatter plot* itu sendiri, dan ada juga "distribusi bersyarat" yang merupakan distribusi subset tertentu dari data (tergantung syarat kepemilikan grup itu).

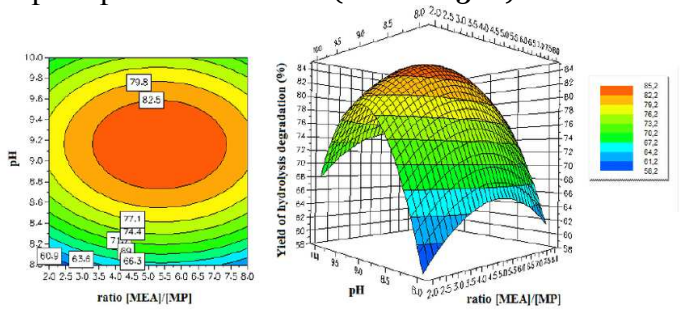
Variabel lain dapat dikodekan ke parameter yang mempengaruhi marker. Warna merupakan pilihan yang baik untuk variabel kategori, asalkan tidak terlalu banyak. Pendekatan populer lain adalah *bubble chart*, di mana ukuran *marker* merupakan variabel ketiga. Ingat, area tidak mudah dicerna, jadi harus digunakan variabel ketiga sebagai ordinal. Gelembung juga bisa saling mengaburkan jika buram. Akan dibahas di bab selanjutnya bagaimana menampilkan banyak variabel di waktu yang sama.

Jika *scatter plot* merupakan plot dua dimensi, seperti apakah tampilan histogram dan *plot density* dalam dua dimensi. Gambar 3.10 adalah suatu contoh scatter plot dalam dua dimensi.



Gambar 3.10 Contoh Scatter Plot dua dimensi.

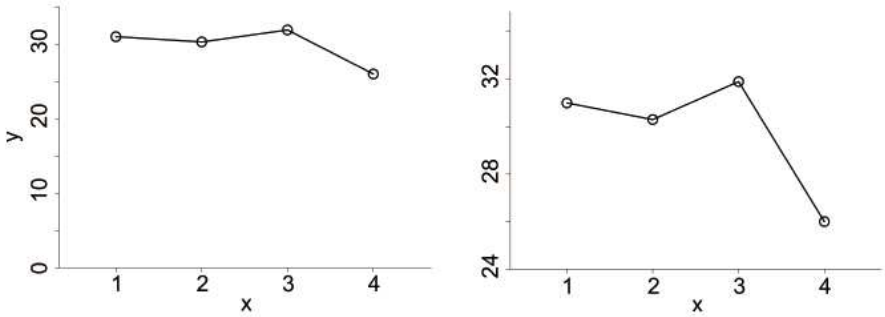
Kerapatan dapat dihaluskan pada permukaan dua dimensi dan memberi kesan seolah-olah itu adalah ketinggian yang keluar dari halaman seperti pada *Plot kontur* (Gambar 3.11).



Gambar 3.11 *Contour Plot* dalam Visualisasi 2-D dan 3-D.

Dua hal yang harus difahami. Pertama, data kontinu atau diskrit dapat dibuat agar nampak perbedaannya seakan lebih besar atau

lebih kecil dari yang sebenarnya, yaitu dengan memperbesar area yang ditempati data. Sebagai contoh pada Gambar 3.12, di mana gambar kiri menunjukkan variabel yang mencakup nilai nol dan gambar kanan yang berfokus pada kisaran data pengamatan saja.



Gambar 3.12 Data yang sama menggunakan sumbu dengan kisaran data yang berbeda akan menonjolkan informasi yang berbeda.

Peringatan kedua adalah terkadang terlihat sumbu putus seperti Gambar 3.12 (kanan) (nilai tidak dimulai dari nol). Gambar kiri dan kanan menunjukkan data yang persis sama. Mungkin sulit memperhatikan perubahan kisaran angka pada sumbu vertikal di gambar kanan, dan itu bisa dengan mudah menyesatkan pembaca. Hindari menggunakannya sumbu putus seperti ini.