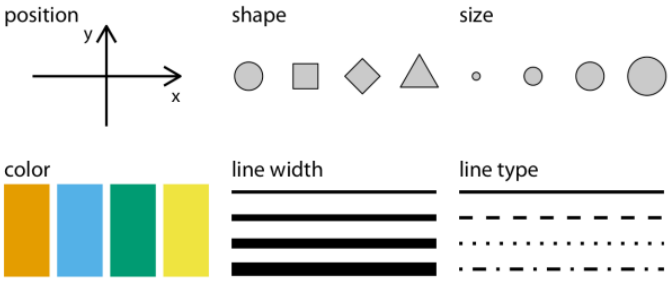
LATIFA RISKA AULIA

09010621010

FUNDAMENTALS OF DATA VISUALIZATION

1. Visualizing Data: Mapping Data onto Aesthetics

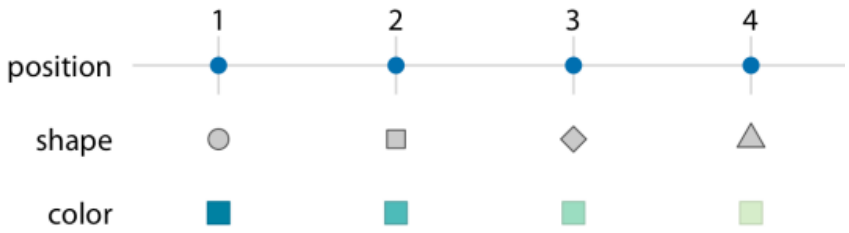
Estetika menggambarkan setiap aspek dari elemen grafis tertentu. Beberapa contohnya adalah disediakan pada Gambar 1. Komponen penting dari setiap elemen grafis tentu saja posisinya, yang menggambarkan di mana elemen tersebut berada. Dalam grafik 2D standar, kami menggambarkan posisi dengan nilai x dan y. Selanjutnya, semua elemen grafis memiliki bentuk, ukuran, dan warna. Bahkan jika kita menyajikan sebuah gambar berwarna hitam-putih, elemen grafis harus memiliki warna agar terlihat: misalnya gambar berwarna hitam memiliki latar belakang putih atau gambar berwarna putih memiliki latar belakang hitam. Dan yang terakhir kita menggunakan garis untuk memvisualisasikan data, garis ini mungkin memiliki lebar atau pola garis putus-putus yang berbeda. Selain contoh estetika yang telah di paparkan, ada banyak estetika lain yang mungkin kita temui di visualisasi data. Misalnya, menampilkan teks, kita mungkin harus menentukan font, tampilan font, dan ukuran font.



Gambar 1

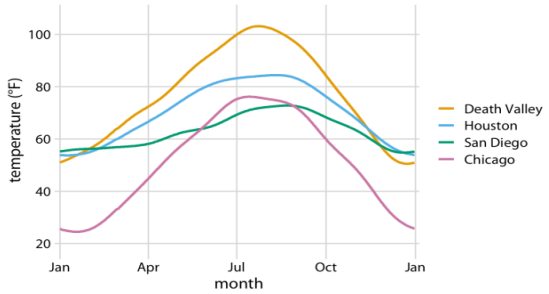
Semua estetika termasuk dalam salah satu dari dua kelompok, dua kelompok tersebut adalah kontinu dan diskrit. Nilai data kontinyu adalah nilai yang mempunyai perantara yang sewenang-wenang. Misalnya, durasi waktu adalah nilai kontinu. Antara dua durasi 50 detik dan 51 detik, terdapat banyak perantara seperti 50,5 detik, 50,51 detik, 50,50001 detik, dan seterusnya. Sebaliknya, jumlah orang di sebuah ruangan adalah nilai diskrit. Sebuah ruangan dapat menampung 5 orang atau 6 orang, tetapi tidak 5,5 orang. Pada Gambar 1, posisi, ukuran, warna, dan lebar garis dapat merepresentasikan data kontinu, tetapi bentuk dan tipe garis biasanya hanya dapat merepresentasikan data diskrit.

Untuk memetakan nilai data ke estetika, kita perlu menentukan nilai data mana yang sesuai dengan nilai estetika tertentu. Misalnya, jika grafik kita terletak di sumbu X, maka kita perlu menentukan nilai data mana yang jatuh ke posisi tertentu di sepanjang sumbu X. Demikian pula, kita mungkin perlu menentukan nilai data mana yang diwakili oleh bentuk atau warna tertentu.



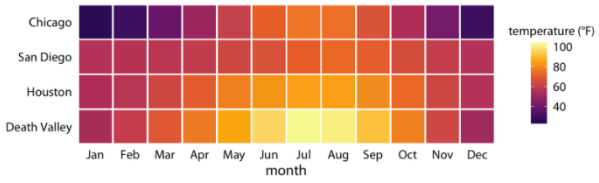
Gambar 2

Contoh nya saat kita memetakan suhu ke sumbu Y, Bulan dalam setahun ke sumbu X, dan petunjuk lokasi dengan warna, dan memvisualisasikan estetika ini dengan garis padat. Hasilnya adalah plot garis standar yang menunjukkan suhu normal di empat lokasi saat mereka berubah sepanjang tahun.



Gambar 3

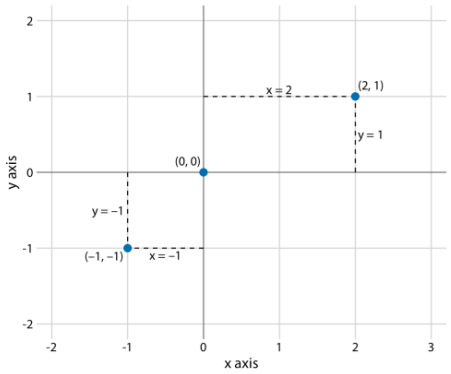
Kita coba menampilkan variabel suhu yang ditampilkan sebagai warna, kita perlu menunjukkan area berwarna yang cukup besar agar warna dapat menyampaikan informasi yang berguna.



Gambar 4

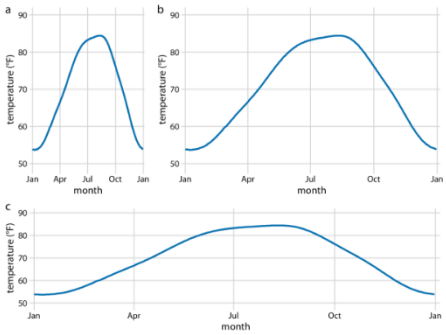
2. Coordinate Systems and Axes

Sistem koordinat yang paling banyak digunakan untuk visualisasi data adalah 2D Sistem koordinasi cartesian, di mana setiap lokasi ditentukan secara unik oleh nilai X dan Nilai Y. Sumbu X dan sumbu Y berjalan secara ortogonal satu sama lain, dan nilai data ditempatkan dalam jarak genap di sepanjang kedua sumbu. Kedua sumbu termasuk dalam skala posisi kontinu, dan keduanya dapat mewakili bilangan real positif dan negatif.



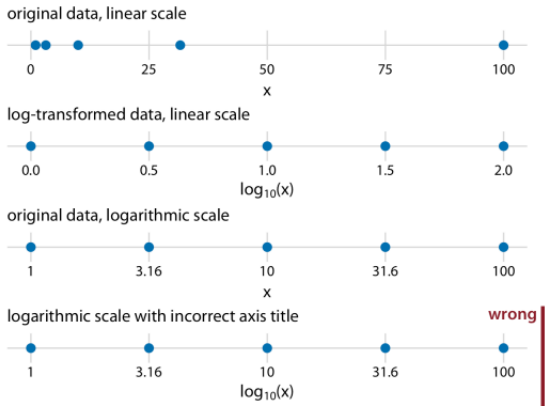
Gambar 5

Sistem koordinat Cartesian dapat memiliki dua sumbu yang mewakili dua unit yang berbeda. Sumbu Y sebagai penunjuk temperature, sumbu X sebagai penunjuk bulan.



Gambar 6

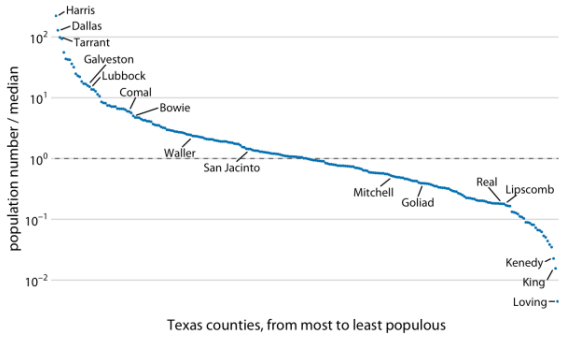
Skala nonlinear yang paling umum digunakan adalah skala logaritmik, atau skala log. Untuk membuat skala log, kita perlu melakukan transformasi log nilai data sambil mengeksponenkan angka yang ditampilkan di sepanjang garis sumbu. Proses ini ditunjukkan pada Gambar 7 , yang menunjukkan angka 1, 3.16, 10, 31.6, dan 100 ditempatkan pada skala linier dan log.



Gambar 7

Skala log sering digunakan ketika kumpulan data berisi angka dengan besaran yang sangat berbeda. Untuk wilayah Texas yang ditunjukkan pada Gambar 8, yang paling padat penduduknya

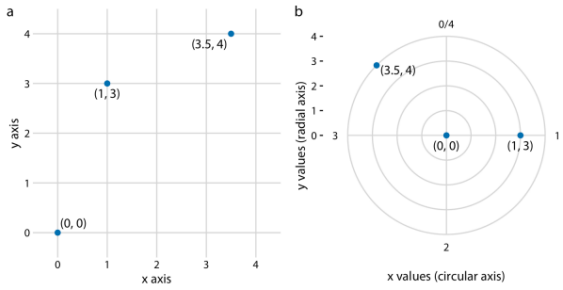
(Harris) memiliki 4.092.459 penduduk pada Sensus AS 2010 sedangkan yang paling sedikit penduduknya (Loving) memiliki 82.



Gambar 8

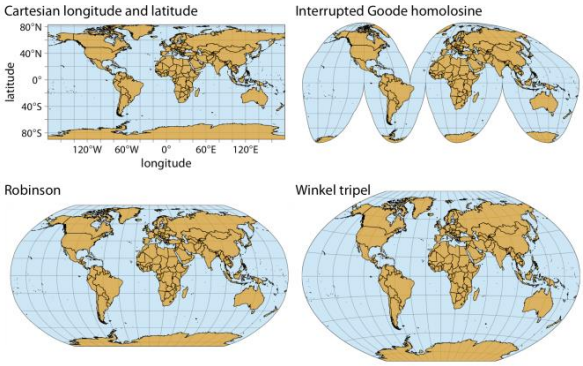
Tetapi jika ada kabupaten dengan 0 penduduk, Kabupaten ini tidak dapat ditampilkan dalam skala logaritmik, karena terletak pada minus tak terhingga.

Secara khusus, di sistem koordinat polar, kami menentukan posisi melalui sudut dan jarak radial dari titik asal, dan oleh karena itu sumbu sudutnya melingkar. Koordinat polar (kutub) dapat berguna untuk data yang bersifat periodik, sehingga nilai data di salah satu ujung skala dapat digabungkan secara logis dengan nilai data di ujung lainnya.



Gambar 9

Kita bisa menemukan sumbu melengkung dalam konteks data geospasial, yaitu peta. Lokasi di dunia ditentukan oleh bujur dan lintangnya. Tetapi karena bumi berbentuk bulat, menggambar lintang dan bujur sebagai sumbu Cartesian adalah hal yang tidak dianjurkan. Jadi peta dunia, ditampilkan dalam empat proyeksi berbeda.



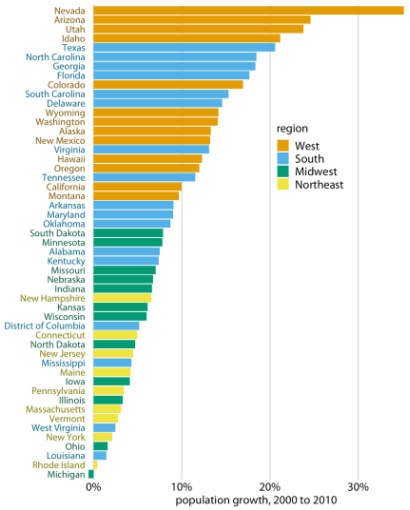
Gambar 10

3. Color Scales

Kami sering menggunakan warna sebagai sarana untuk membedakan item atau kelompok diskrit yang tidak memiliki urutan intrinsik, seperti negara yang berbeda pada peta. Dalam hal ini, kami menggunakan skala warna kualitatif. Skala seperti itu berisi kumpulan warna tertentu yang terbatas yang dipilih agar terlihat jelas berbeda satu sama lain. Kondisi kedua mensyaratkan bahwa tidak ada satu warna yang menonjol relatif terhadap yang lain. Selain itu, warna tidak boleh menimbulkan kesan keteraturan, seperti halnya urutan warna yang semakin terang. Banyak skala warna kualitatif yang sesuai sudah tersedia. Gambar 4-1menunjukkan tiga contoh yang representative, termasuk warna yang cukup terang dan cukup gelap.

Gambar 11

Sebagai contoh menunjukkan persentase pertumbuhan penduduk dari tahun 2000 hingga 2010 di negara bagian AS. Pewarnaan ini menyoroti bahwa negara bagian di wilayah yang sama mengalami pertumbuhan populasi yang serupa.

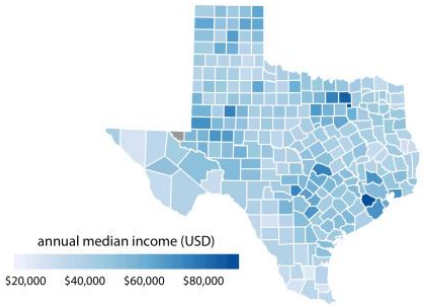


Gambar 12

Warna juga dapat digunakan untuk mewakili nilai data kuantitatif, seperti pendapatan, suhu, atau kecepatan. Dalam hal ini, kami menggunakan skala warna sekuensial. Skala seperti itu berisi urutan warna yang dengan jelas menunjukkan nilai mana yang lebih besar atau lebih kecil dari yang lain, dan seberapa jauh dua nilai spesifik satu sama lain.

Gambar 13

Dalam hal ini, kita dapat menggambar peta wilayah geografis dan mewarnainya dengan nilai data. Peta seperti itu disebut choropleth. Gambar 14 menunjukkan contoh pemetakan pendapatan rata-rata tahunan di setiap wilayah di Texas.



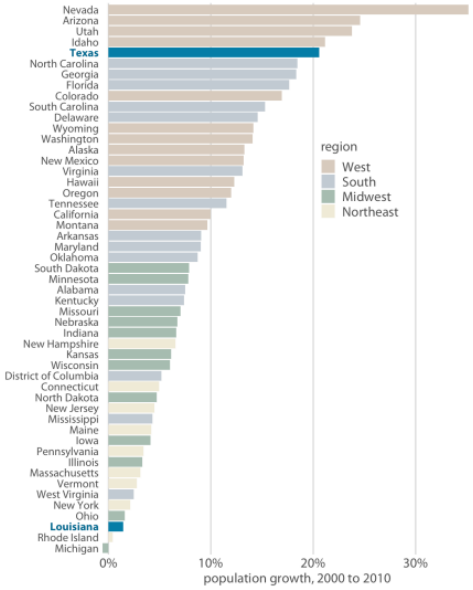
Gambar 14

Warna juga bisa menjadi alat yang efektif untuk menyorot elemen tertentu dalam data. Mungkin ada kategori atau nilai tertentu dalam kumpulan data yang membawa informasi, dan kita dapat memperkuat informasi tersebut dengan menekankan elemen figur yang relevan kepada pembaca. Cara mudah untuk mencapai penekanan ini adalah dengan mewarnai elemen gambar ini dalam warna atau kumpulan warna yang menonjol dengan jelas dibandingkan gambar lainnya.

Efek ini dapat dicapai dengan aksen skala warna, yaitu skala warna yang berisi kumpulan warna redup dan kumpulan warna yang lebih kuat, lebih gelap, dan/atau lebih jenuh.

Gambar 15

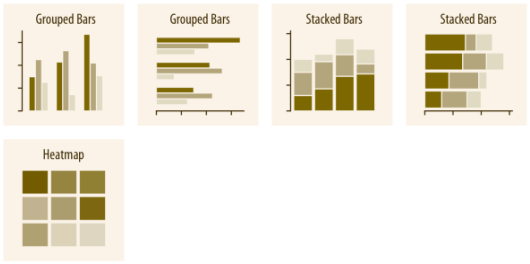
Sebagai contoh bagaimana data yang sama dapat mendukung informasi yang berbeda dengan pendekatan pewarnaan yang berbeda, dengan menyoroti dua negara bagian tertentu, Texas dan Louisiana. Kedua negara bagian berada di Selatan, Texas adalah negara bagian dengan pertumbuhan tercepat kelima di AS dari tahun 2000 hingga 2010 sedangkan Louisiana adalah negara dengan pertumbuhan terendah, dan yang lainnya adalah negara bagian dengan pertumbuhan paling lambat ketiga.



Gambar 16

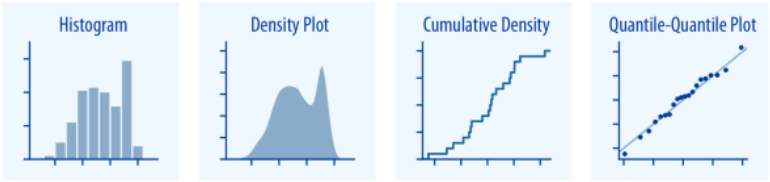
4. Directory of Visualizations

Pendekatan yang paling umum untuk memvisualisasikan jumlah/amount (nilai numerik yang ditampilkan untuk beberapa kumpulan kategori) adalah menggunakan diagram batang, baik yang disusun secara vertikal maupun horizontal.



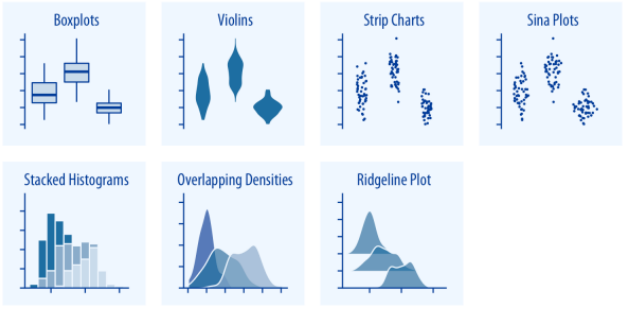
Gambar 17

Histograms and density plots memberikan visualisasi distribusi yang paling intuitif, tetapi keduanya memerlukan pilihan parameter arbitrer (sewenang-wenang) dan dapat menyesatkan. Cumulative densities and quantile-quantile (q-q) plots, selalu mewakili data dengan tepat tetapi bisa lebih sulit untuk ditafsirkan.



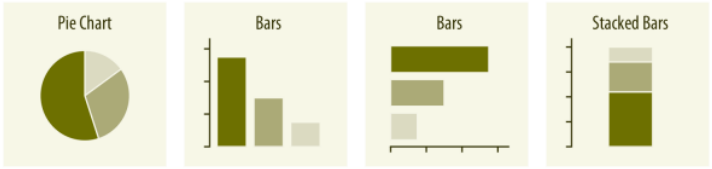
Gambar 18

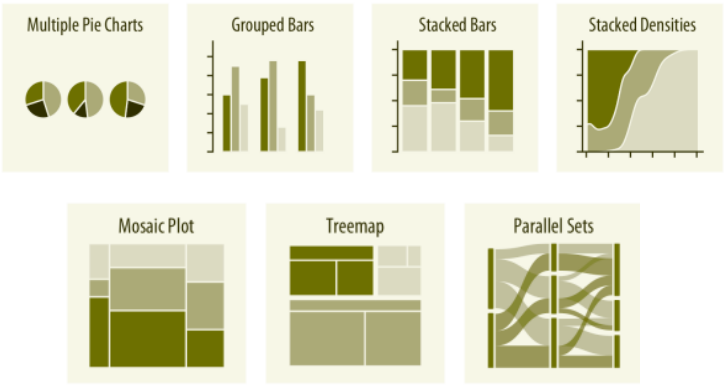
Dan distribution lainya terdapat Boxplots, violin plots, strip charts, dan sina plots, stacked histograms, overlapping densities, ridgeline plot.



Gambar 19

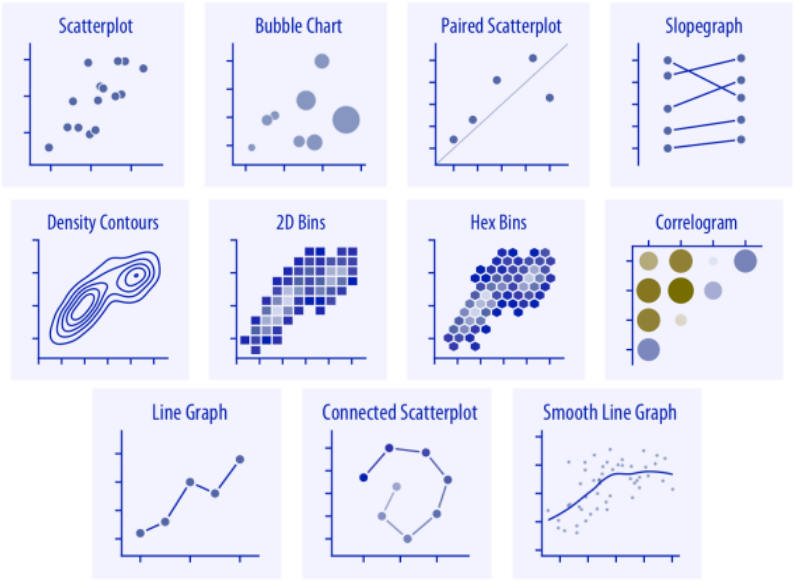
Proporsi dapat divisualisasikan sebagai pie charts, side-by-side bars, stacked bars, multiple pie charts, grouped bars, stacked bars, stacked densities, mosaic plot, treemap, parallel sets.





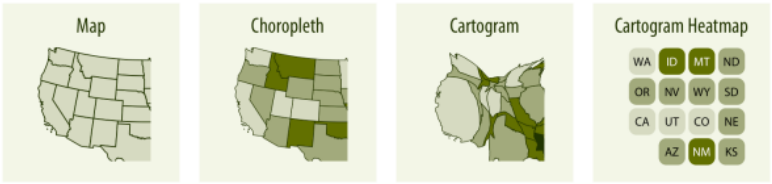
Gambar 20

X-Y relationship terdiri dari scatter plot, bubble chart, paired scatterplot, slopegraph, density contours, 2D bins, hex bins, correlgram, line graph, connected scatterplot, smooth line graph.



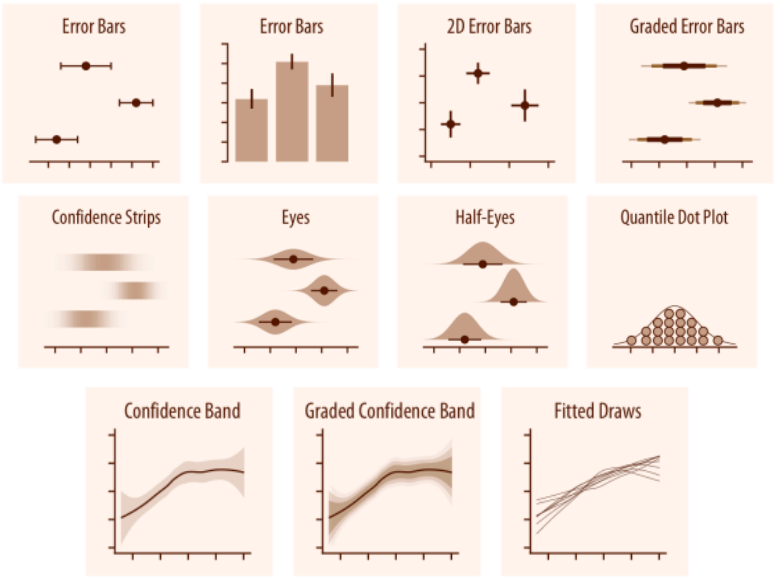
Gambar 21

Geospatial data terdiri dari Map, choropleth, cartogram, cartogram heatmap



Gambar 22

Uncertainty terdiri dari error bars, 2D error bars, graded error bars, confidence strips, eyes, half eyes, quantile dot plot, confidence band, graded confidence band, fitted draws.



Gambar 23

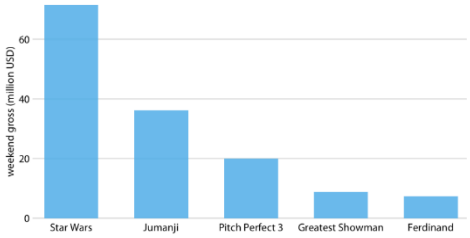
5. Visualizing Amounts

Dalam banyak skenario, kami tertarik pada besarnya beberapa rangkaian angka. Misalnya, kita mungkin ingin memvisualisasikan total volume penjualan berbagai merek mobil, atau jumlah total orang yang tinggal di kota yang berbeda. Kami menyebut kasus ini sebagai jumlah visualisasi, karena penekanan utama dalam visualisasi ini adalah besarnya nilai kuantitatif. Contoh dari memvisualisasi jumlah sebagai berikut, terdapat beberapa film paling popular pada pekan ini.

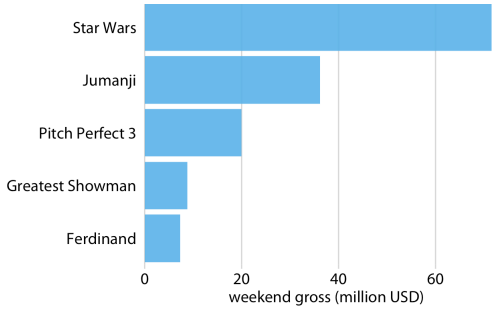
Rank	Title	Weekend gross
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

Gambar 24

Jenis data ini biasanya divisualisasikan dengan vertical bars. Pada vertical bars terdapat nama film dan weekend gross

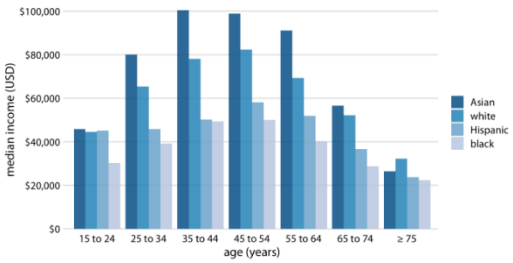


Gambar 25



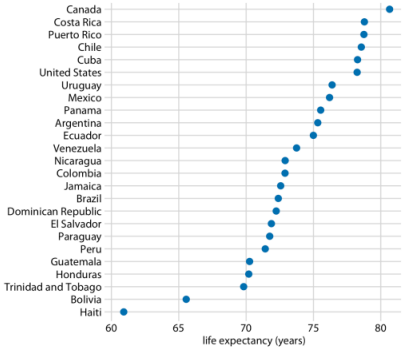
Gambar 26

Gouped bar plots yang dikelompokkan menampilkan banyak informasi sekaligus, dan dapat membingungkan. Misalnya, Rata-rata pendapatan rumah tangga tahunan AS tahun 2016 versus kelompok usia dan ras. Kelompok usia ditampilkan di sepanjang sumbu X, dan untuk setiap kelompok umur ada empat batang, sesuai dengan pendapatan rata-rata orang Asia, kulit putih, Hispanik, dan kulit hitam.



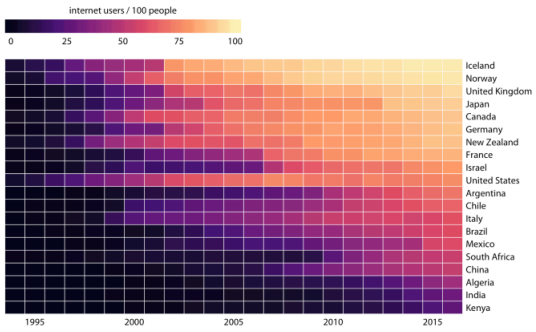
Gambar 27

Dot plot dan heatmap. Pada dot plot kita dapat menunjukkan jumlah dengan menempatkan titik-titik pada lokasi yang sesuai di sepanjang sumbu X atau sumbu Y.



Gambar 28

Sebagai alternatif untuk memetakan nilai data ke posisi melalui bar atau dot plot, kita dapat memetakan nilai data ke warna. Biasanya disebut heatmap.



Gambar 29

6. Visualizing Distributions: Histograms and Density Plots

Pada single distribusi Kita dapat mengetahui distribusi usia di antara penumpang dengan mengelompokkan semua penumpang ke dalam kotak dengan usia yang sebanding dan kemudian menghitung jumlah penumpang di setiap kategori.

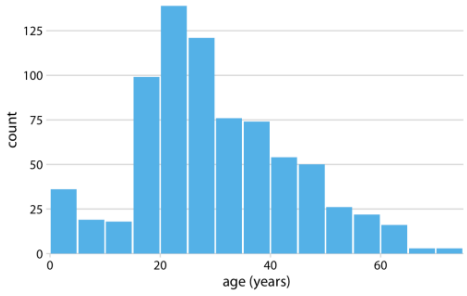
Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121

Age range	Count
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22

Age range	Count
61–65	16
66–70	3
71–75	3

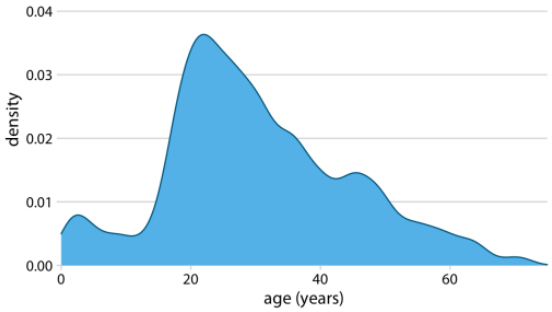
Gambar 30

Kita dapat memvisualisasikan tabel ini dengan menggambar persegi panjang yang tingginya sesuai dengan hitungan dan lebarnya sesuai dengan data usia. visualisasi seperti itu disebut histogram.



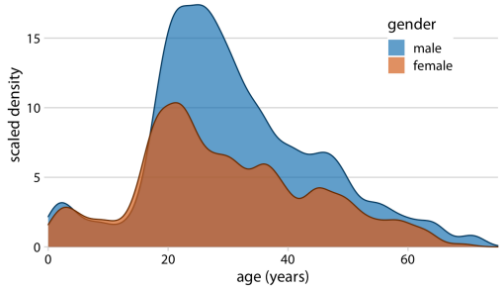
Gambar 31

Histogram telah menjadi pilihan visualisasi yang populer setidaknya sejak abad ke-18, sebagian karena mudah dibuat. Baru-baru ini, karena daya komputasi yang luas telah tersedia di perangkat sehari-hari seperti laptop dan ponsel, kami melihat mereka semakin digantikan oleh density plot. Kurva ini perlu diestimasi dari data, dan metode yang paling umum digunakan untuk prosedur estimasi ini disebut estimasi densitas kernel.Dalam estimasi tersebut, kami menggambar kurva kontinu (kernel) dengan lebar (dikendalikan oleh parameter yang disebut bandwidth) di lokasi setiap titik data, lalu kami menjumlahkan semua kurva ini untuk mendapatkan perkiraan kepadatan akhir. Kernel yang paling banyak digunakan adalah kernel Gaussian.



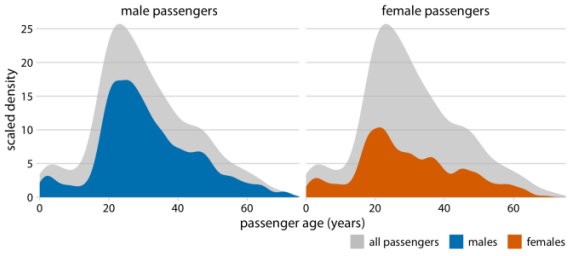
Gambar 32

Dalam banyak skenario kami memiliki banyak distribusi yang ingin kami visualisasikan secara bersamaan. Sebagai contoh, katakanlah kita ingin melihat berapa umur dari penumpang Titanic dibagi antara pria dan wanita. Apakah penumpang pria dan wanita pada umumnya memiliki usia yang sama, atau apakah ada perbedaan usia antar jenis kelamin. Berikut perbandingan Perkiraan kepadatan umur pria dan Wanita penumpang titanic. Untuk menyoroti bahwa terdapat lebih banyak penumpang pria daripada Wanita.



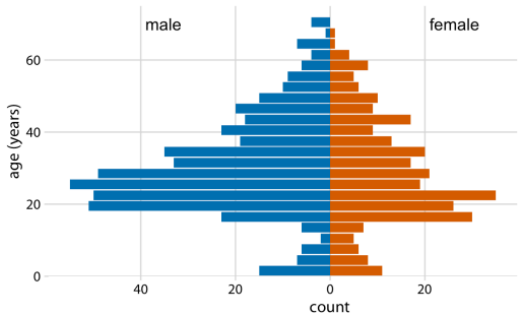
Gambar 33

Pembagian umur laki-laki dan perempuan penumpang titanic, ditampilkan sebagai proporsi dari jumlah total penumpang.



Gambar 34

ketika kita ingin memvisualisasikan tepat dua distribusi, kita juga dapat membuat dua histogram terpisah yakni age pyramid.

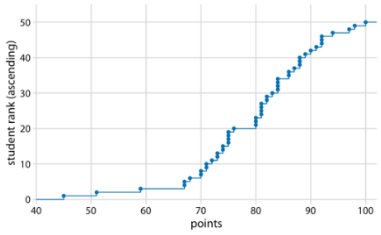


Gambar 35

7. Visualizing Distributions: Empirical Cumulative Distribution Functions and Q-Q Plots

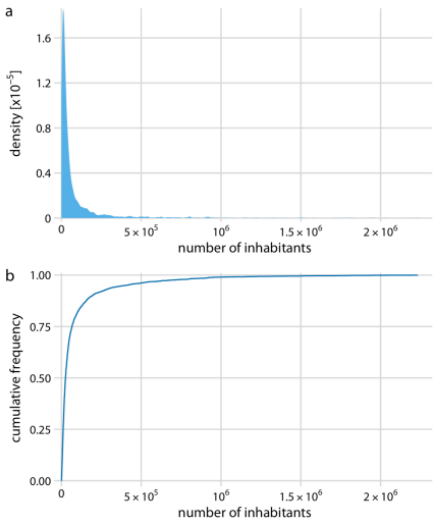
Untuk mengilustrasikan ECDF (empirical cumulative distribution function), saya akan mulai dengan contoh hipotetis sebagai berikut: kumpulan data nilai siswa. Asumsikan 1 kelas memiliki 50 siswa, dan baru saja menyelesaikan ujian di mana mereka dapat memperoleh skor antara 0 dan 100 poin. Bagaimana cara terbaik untuk memvisualisasikan ? Kita dapat memplot jumlah total siswa yang telah menerima paling banyak sejumlah poin versus poin yang didapatkan.

Kita dapat mengurutkan semua siswa berdasarkan jumlah poin yang mereka peroleh, dalam urutan paling rendah hingga paling tinggi. Hasilnya adalah fungsi distribusi kumulatif empiris



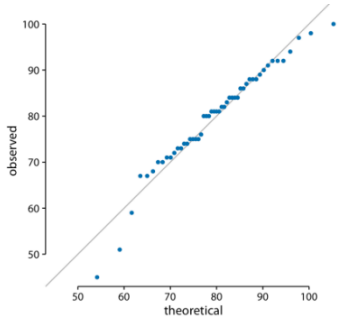
Gambar 36

Banyak kumpulan data empiris menampilkan highly skewed distributions, khususnya dengan tail yang berat di sebelah kanan, dan distribusi ini dapat menjadi tantangan untuk divisualisasikan. Contoh distribusi tersebut termasuk jumlah orang yang tinggal di kota atau kabupaten yang berbeda, jumlah kontak di jejaring sosial, frekuensi munculnya kata-kata tertentu dalam sebuah buku, jumlah makalah akademis yang ditulis oleh penulis yang berbeda, kekayaan bersih dari individu. Kali ini akan memvisualisasikan jumlah orang yang tinggal di berbagai negara bagian AS menurut Sensus AS 2010. Distribusi ini memiliki tail yang sangat panjang ke kanan. Meskipun sebagian besar kabupaten memiliki jumlah penduduk yang relatif kecil.



Gambar 37

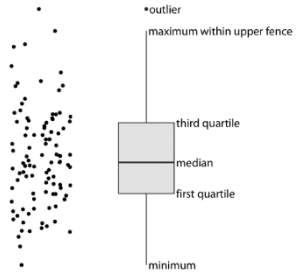
Quantile-quantile plots adalah visualisasi yang berguna ketika kita ingin menentukan sejauh mana titik data yang diamati mengikuti atau tidak mengikuti distribusi yang diberikan. Sama seperti ECDF, plot qq juga didasarkan pada peringkat data dan memvisualisasikan hubungan antara peringkat dan nilai sebenarnya.



Gambar 38

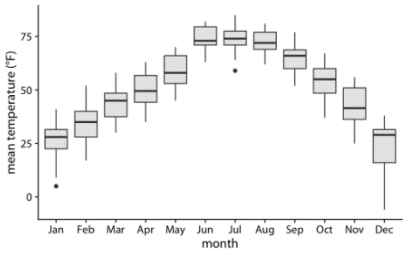
8. Visualizing Many Distributions at Once

Pendekatan paling sederhana untuk menampilkan banyak distribusi sekaligus adalah dengan menunjukkan rata-rata atau mediannya sebagai titik, dengan beberapa indikasi variasi di sekitar rata-rata atau median yang ditunjukkan oleh error bars.



Gambar 39

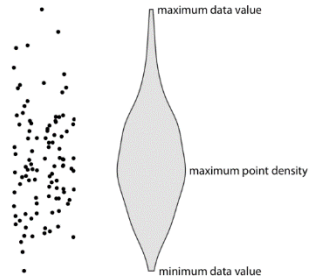
Boxplots sederhana namun informatif, dan mereka bekerja dengan baik ketika diplot bersebelahan untuk memvisualisasikan banyak distribusi sekaligus. Sebagai contoh Rata-rata suhu harian di Lincoln, NE, divisualisasikan sebagai boxplots.



Gambar 40

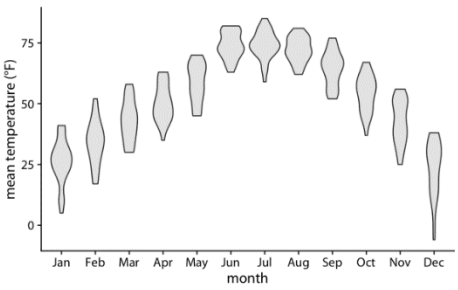
Boxplot ditemukan oleh ahli statistik John Tukey pada awal 1970-an, dan dengan cepat mendapatkan popularitas karena sangat informatif sekaligus mudah digambar dengan tangan. Namun, dengan kemampuan komputasi dan visualisasi modern, baru-baru ini kami melihat boxplots digantikan oleh violin plot.

Secara teknis, violin plot adalah perkiraan kerapatan yang diputar 90 derajat dan kemudian dicerminkan. Karena itu violin plot simetris. Violin plots dimulai dan diakhiri masing-masing pada nilai data minimum dan maksimum. Bagian paling tebal dari violin plots sesuai dengan kerapatan titik tertinggi dalam kumpulan data.



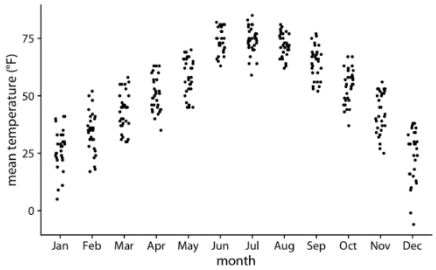
Gambar 41

Rata-rata suhu harian di Lincoln, NE, divisualisasikan sebagai violin plot.



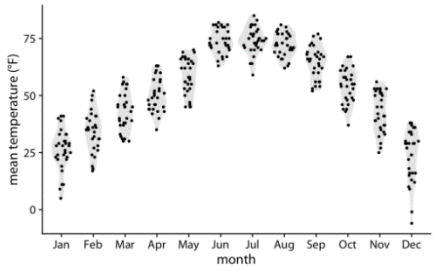
Gambar 42

Rata-rata suhu harian di Lincoln, NE, divisualisasikan sebagai grafik strip.



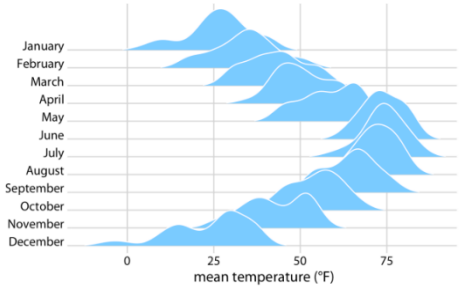
Gambar 43

Rata-rata suhu harian di Lincoln, NE, divisualisasikan sebagai plot sina (kombinasi individual points and violins)



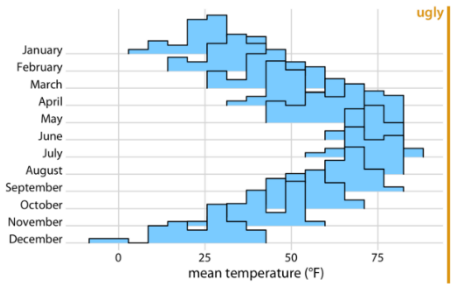
Gambar 44

kami memvisualisasikan distributions along the horizontal axis menggunakan histograms and density plots. Di sini, kami akan memperluas ide ini dengan membuat distribution plots in the vertical direction. Visualisasi yang dihasilkan disebut ridgeline plot. Sebagai contoh Temperatur di Lincoln, NE, pada tahun 2016, divisualisasikan sebagai plot ridgeline.



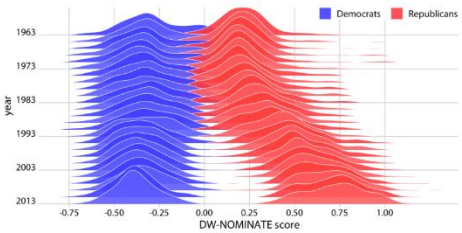
Gambar 45

Pada prinsipnya, kita dapat menggunakan histogram daripada density plots dalam visualisasi ridgeline. Namun, angka yang dihasilkan seringkali tidak terlihat bagus.



Gambar 46

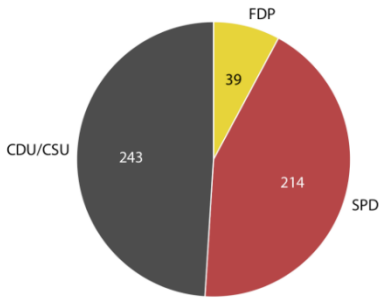
Plot Ridgeline juga bekerja dengan baik jika kita ingin membandingkan dua tren dari waktu ke waktu.



Gambar 47

9. Visualizing Proportions

Dari tahun 1961 hingga 1983, parlemen Jerman (disebut Bundestag) terdiri dari anggota dari tiga partai berbeda, CDU/CSU, SPD, dan FDP. Selama sebagian besar waktu ini, CDU/ CSU dan SPD memiliki jumlah kursi yang kira-kira sebanding, sementara FDP biasanya hanya memiliki sebagian kecil kursi. Misalnya, di Bundestag kedelapan, dari tahun 1976– 1980, CDU/CSU memegang 243 kursi, SPD 214, dan FDP 39, dengan total 496. Data parlemen semacam itu biasanya divisualisasikan sebagai diagram lingkaran.

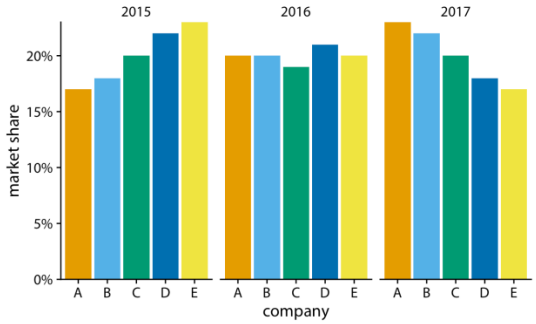


Gambar 48

Pie charts membagi lingkaran menjadi irisan sedemikian rupa sehingga luas setiap irisan sebanding dengan fraksi dari total yang diwakilinya.

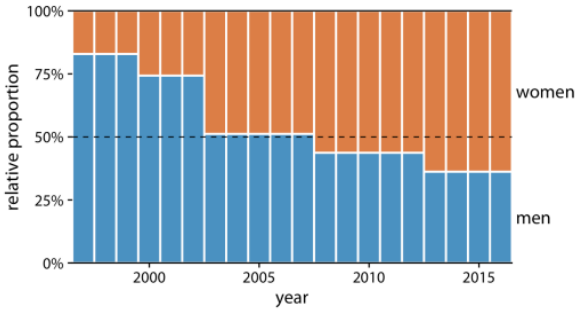
Case for side by side bars. Diberikan contoh berikut, pertimbangkan skenario hipotetis dari lima perusahaan, A, B, C, D, dan E, yang semuanya memiliki tingkat pemasaran sekitar 20%. Kumpulan data hipotetis kami mencantumkan tingkat pemasaran masing-masing perusahaan selama tiga tahun berturut-turut. Saat kami memvisualisasikan kumpulan data ini dengan diagram lingkaran, sulit untuk melihat tren tertentu.

Gambar 49



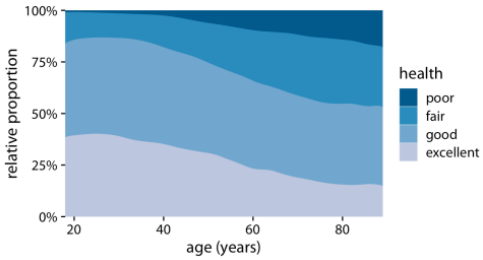
Gambar 50

Case for Stacked Bars and Stacked Densities. Sebagai contoh, perhatikan proporsi perempuan di parlemen nasional suatu negara. Kami secara khusus akan melihat negara Afrika Rwanda, yang pada tahun 2016 berada di puncak daftar negara dengan proporsi anggota parlemen perempuan tertinggi. Rwanda memiliki mayoritas parlemen perempuan sejak 2008, dan sejak 2013 hampir dua pertiga anggota parlemennya adalah perempuan. Untuk memvisualisasikan bagaimana proporsi perempuan di parlemen Rwanda telah berubah dari waktu ke waktu, kita dapat menggambar urutan stacked bar graphs.



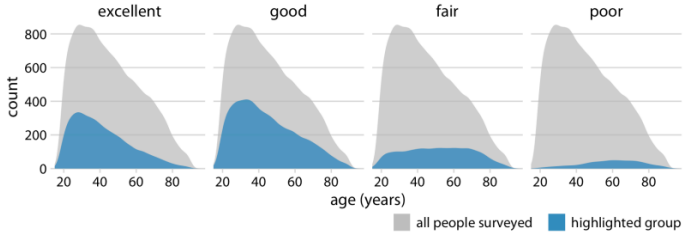
Gambar 51

Untuk memberikan contoh di mana stacked densities mungkin sesuai, pertimbangkan status kesehatan orang. Usia dapat dianggap sebagai variabel kontinu, dan memvisualisasikan data dengan cara ini cukup berhasil.



Gambar 52

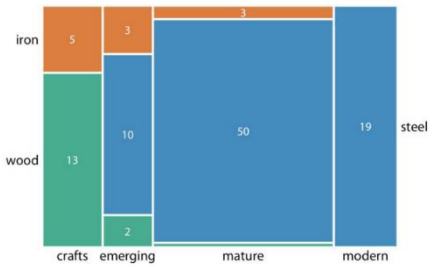
Visualizing Proportions Separately as Parts of the Total. Distribusi usia keseluruhan dalam kumpulan data ditampilkan sebagai area abuabu yang diarsir, dan distribusi usia untuk setiap status kesehatan ditampilkan dengan warna biru. Angka ini menyoroti bahwa secara absolut, jumlah orang dengan kesehatan yang sangat baik atau baik menurun melewati usia 30-40, sementara jumlah orang dengan kesehatan yang baik kira-kira tetap konstan di semua usia.



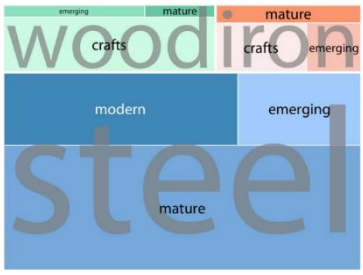
Gambar 53

10. Visualizing Nested Proportions

Mosaic plots and treemaps. Setiap kali kita memiliki kategori yang tumpang tindih, yang terbaik adalah menunjukkan secara eksplisit bagaimana keterkaitannya satu sama lain. Ini dapat dilakukan dengan plot mosaic. Sepintas, petak mozaik terlihat mirip dengan petak batang bertumpuk. Namun, tidak seperti di petak batang bertumpuk, di petak mosaik tinggi dan lebar masing-masing area yang diarsir bervariasi.



Gambar 54

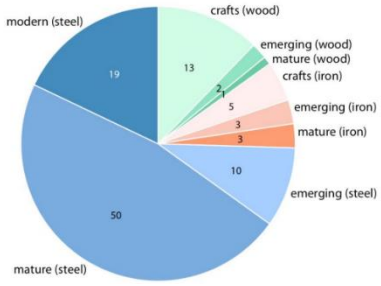


Gambar 55

Sementara plot mosaik dan peta pohon saling terkait erat, keduanya memiliki titik penekanan dan area aplikasi yang berbeda. Di sini, petak mozaik (Gambar 54) menekankan evolusi temporal dalam penggunaan bahan bangunan dari era kerajinan hingga era modern, sedangkan treemaps (Gambar 55) menekankan jumlah baja, besi, dan kayu jembatan.

Gambar 55 yakni Kerusakan jembatan di Pittsburgh berdasarkan bahan konstruksi (baja, kayu, besi) dan menurut era konstruksi (crafts, emerging, mature, modern), ditampilkan sebagai peta pohon. Luas setiap persegi panjang sebanding dengan jumlah jembatan jenis itu.

Nested pies. Sebagai alternatif, pertama-tama kita dapat membagi lingkaran menjadi potongan-potongan yang mewakili proporsi menurut satu variabel (misalnya bahan) dan kemudian membagi potongan-potongan ini lebih lanjut menurut variabel lain (era konstruksi). Dengan cara ini, sebenarnya kita membuat diagram lingkaran normal dengan sejumlah besar irisan lingkaran kecil. Namun, kita kemudian dapat menggunakan pewarnaan untuk menunjukkan sifat nested pies. Pada warna hijau melambangkan jembatan kayu, warna jingga melambangkan jembatan besi, dan warna biru melambangkan jembatan baja. Kegelapan setiap warna mewakili era konstruksi, dengan warna yang lebih gelap sesuai dengan jembatan yang baru dibangun.

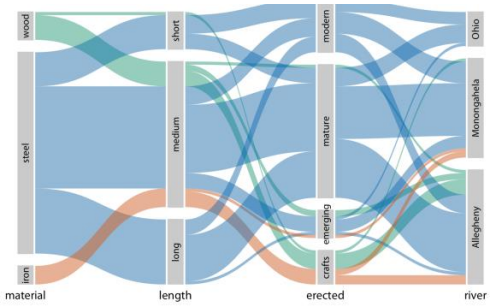


Gambar 56

Parallel sets. Dalam plot set paralel, kami menunjukkan bagaimana total dataset dipecah oleh masing-masing variabel kategori individu, dan kemudian kami menggambar pita berbayang yang menunjukkan bagaimana subkelompok berhubungan satu sama lain.

sebagai contoh. Pada gambar ini, saya telah memecah dataset jembatan berdasarkan bahan konstruksi (besi, baja, kayu), length setiap jembatan (panjang, sedang, pendek), era di mana setiap jembatan dibangun (crafts, emerging, mature, modern), dan sungai yang dibentang oleh setiap jembatan (Allegheny, Monongahela, Ohio). Pita yang menghubungkan set paralel diwarnai oleh bahan konstruksi. Ini menunjukkan, misalnya, bahwa jembatan kayu sebagian besar berukuran sedang (dengan beberapa jembatan pendek), terutama didirikan selama periode crafts (dengan beberapa jembatan berukuran sedang didirikan selama periode emerging dan mature),

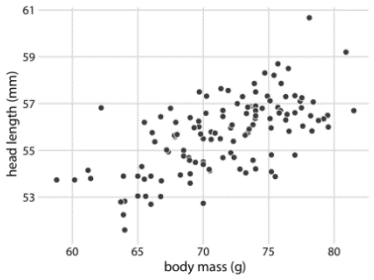
dan merentang terutama sepanjang Sungai Allegheny (dengan beberapa jembatan crafts yang membentang di sepanjang sungai Monongahela).



Gambar 57

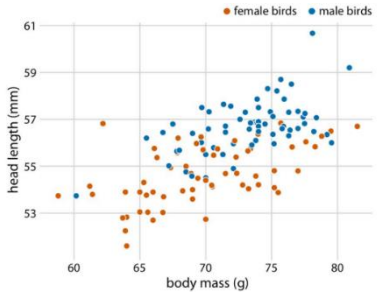
11. Visualizing Associations Among Two or More Quantitative Variables

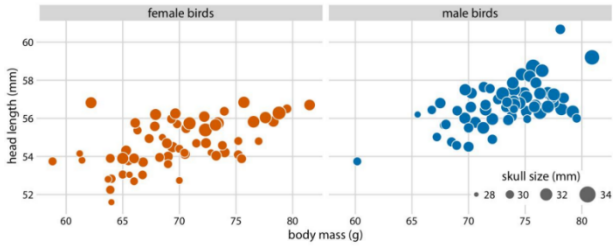
Scatterplot. Diberikan contoh plot head length terhadap body mass. Dalam plot ini, head length ditunjukkan sepanjang sumbu y dan body mass sepanjang sumbu x, dan masing-masing burung diwakili oleh satu titik. (Perhatikan terminologinya: kita mengatakan bahwa kita memplot variabel yang ditunjukkan sepanjang sumbu y terhadap variabel yang ditunjukkan sepanjang sumbu x.) Titik-titik tersebut membentuk awan yang tersebar (oleh karena itu istilah sebar), namun tidak diragukan lagi ada kecenderungan untuk burung dengan tinggi massa tubuh untuk memiliki kepala lebih panjang.



Gambar 58

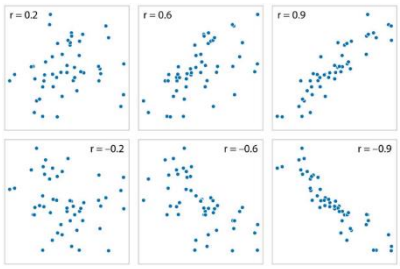
Kumpulan data blue jay berisi burung jantan dan betina, dan kita mungkin ingin mengetahui apakah keseluruhan hubungan antara panjang kepala dan massa tubuh bertahan secara terpisah untuk setiap jenis kelamin.





Gambar 59

Correlograms. Koefisien korelasi r adalah angka antara –1 dan 1 yang mengukur sejauh mana dua variabel berkovariasi. Nilai r = 0 berarti tidak ada asosiasi sama sekali, dan nilai 1 atau –1 menunjukkan asosiasi yang sempurna. Tanda koefisien korelasi menunjukkan apakah variabel berkorelasi (nilai yang lebih besar dalam satu variabel sama dengan nilai yang lebih besar pada variabel lainnya) atau antikorelasi (nilai yang lebih besar pada satu variabel sama dengan nilai yang lebih kecil pada variabel lainnya).



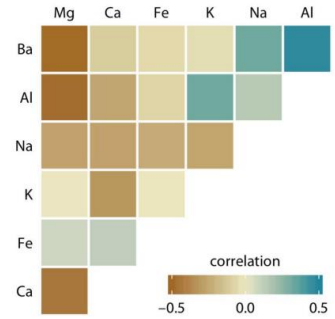
Gambar 60

Koefisien korelasi sebagai berikut:

$$r = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2} \sqrt{\Sigma_i (y_i - \bar{y})^2}}$$

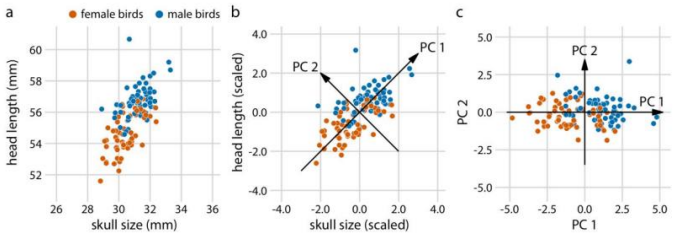
Gambar 61

Visualisasi koefisien korelasi disebut correlograms. Untuk mengilustrasikan penggunaan correlogram sebagai berikut, disediakan dataset lebih dari 200 pecahan kaca yang diperoleh selama pekerjaan forensik. Untuk setiap fragmen kaca, kami mengukur komposisinya, yang dinyatakan sebagai persen berat berbagai mineral oksida. Ada tujuh oksida berbeda yang telah kami ukur, menghasilkan total 6 + 5 + 4 + 3 + 2 + 1 = 21 korelasi berpasangan. Kita dapat menampilkan 21 korelasi ini sekaligus sebagai matriks ubin berwarna, di mana setiap ubin mewakili satu koefisien korelasi. Korelogram ini memungkinkan kami untuk dengan cepat memahami tren dalam data, seperti bahwa magnesium berkorelasi negatif dengan hampir semua oksida lainnya, dan bahwa aluminium dan barium memiliki korelasi positif yang kuat.



Gambar 62

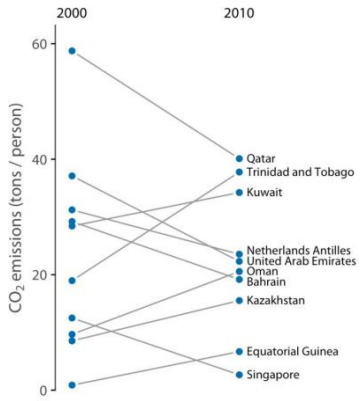
Dimension reduction. Dimension reduction (Pengurangan dimensi) bergantung pada wawasan utama sebagian besar kumpulan data berdimensi tinggi terdiri dari beberapa variabel berkorelasi yang menyampaikan informasi yang tumpang tindih. Kumpulan data tersebut dapat direduksi menjadi sejumlah dimensi yang lebih kecil tanpa kehilangan banyak informasi penting. Sebagai contoh sederhana, pertimbangkan kumpulan data dari beberapa ciri fisik orang, termasuk seperti tinggi dan berat setiap orang, panjang lengan dan kaki, lingkar pinggang, pinggul, dan dada, dll. Kita bisa pahami secara intuitif bahwa semua kuantitas ini akan berhubungan dengan ukuran keseluruhan setiap orang. Semuanya sama, orang yang lebih besar akan menjadi lebih tinggi, lebih berat, memiliki lengan dan kaki lebih panjang, dan memiliki lingkar pinggang, pinggul, dan dada yang lebih besar. Teknik dimension reduction adalah principal components analysis (PCA). PCA memperkenalkan seperangkat variabel baru, yang disebut principal components (PC).



Gambar 63

Paired data. Kasus khusus dari data kuantitatif multivariat adalah data berpasangan: data di mana terdapat dua atau lebih pengukuran besaran yang sama dalam kondisi yang sedikit berbeda. Contohnya termasuk dua pengukuran yang sebanding pada setiap subjek (misalnya, panjang lengan kanan dan kiri seseorang), ulangi pengukuran pada subjek yang sama pada titik waktu yang berbeda (misalnya, berat badan seseorang pada dua waktu yang berbeda sepanjang tahun).

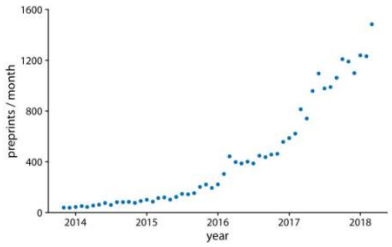
Dalam grafik kemiringan, kami menggambar pengukuran individu sebagai titik-titik yang disusun menjadi dua kolom dan menunjukkan pasangan dengan menghubungkan titik-titik yang dipasangkan dengan garis. Kemiringan setiap garis menyoroti besarnya dan arah perubahan.



Gambar 64

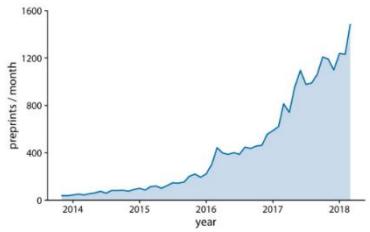
12. Visualizing Time Series and Other Functions of an Independent Variable

Individual Time Series. Sebagai demonstrasi pertama dari time series, kami akan mempertimbangkan pola pengiriman preprint bulanan dalam biologi. Preprint adalah artikel ilmiah yang diposkan oleh peneliti secara online sebelum peer review formal dan publikasi dalam jurnal ilmiah. Server preprint bioRxiv, yang didirikan pada November 2013 khusus untuk peneliti yang bekerja di bidang ilmu biologi, telah mengalami pertumbuhan substansial dalam pengiriman bulanan sejak saat itu. Kita dapat memvisualisasikan pertumbuhan ini dengan membuat bentuk scatterplot di mana kita menggambar titik-titik yang mewakili jumlah kiriman di setiap bulan.



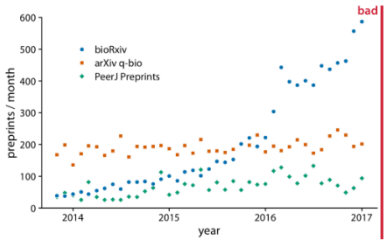
Gambar 65

Namun, ada perbedaan penting antara Gambar 65 dan diagram sebar yang dibahas di Bab sebelumnya. Pada Gambar 65, titik-titik ditempatkan secara merata sepanjang sumbu x, dan ada urutan yang ditentukan. Contoh selanjutnya, Pengajuan bulanan ke server preprint bioRxiv, ditampilkan sebagai grafik garis tanpa titik. Menghilangkan titik-titik akan menekankan tren temporal secara keseluruhan sambil mengurangi pengamatan individu pada titik waktu tertentu. Ini sangat berguna ketika titik waktu berjarak sangat padat. Kita juga bisa mengisi area di bawah kurva dengan warna solid.



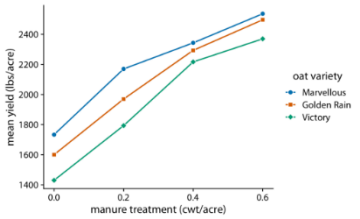
Gambar 66

Multiple Time Series and Dose–Response Curves. Kami sering memiliki banyak time courses yang ingin kami tampilkan sekaligus. Dalam hal ini, kita harus lebih berhati-hati dalam memplot data, karena angkanya bisa membingungkan atau sulit dibaca. Misalnya, jika kita ingin menampilkan pengiriman bulanan ke beberapa server preprint, scatterplot bukanlah ide yang baik, karena Individual Time Series saling bertemu.



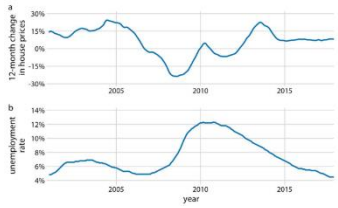
Gambar 67

Contoh selanjutnya. Kurva dosis-respons menunjukkan hasil rata-rata varietas oat setelah pemupukan dengan pupuk kandang. Kotoran berfungsi sebagai sumber nitrogen, dan hasil oat umumnya meningkat karena lebih banyak nitrogen tersedia, terlepas dari varietasnya. Di sini, aplikasi pupuk kandang diukur dalam cwt (hundredweight) per acre. hundredweight adalah satuan old imperial yang setara dengan 112 lbs atau 50,8 kg.



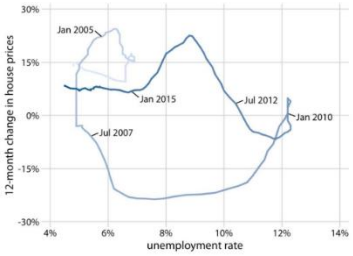
Gambar 68

Time Series of Two or More Response Variables. Sebagai contoh, kita mungkin tertarik pada perubahan harga rumah dari 12 bulan sebelumnya yang berkaitan dengan tingkat pengangguran. Kita mungkin berharap harga rumah naik ketika tingkat pengangguran rendah, dan sebaliknya.



Gambar 69

Sebagai alternatif untuk menampilkan dua grafik garis terpisah, kita dapat memplot kedua variabel satu sama lain, menggambar jalur yang mengarah dari titik waktu paling awal ke titik waktu terbaru. Visualisasi seperti itu disebut connected scatterplt, karena kita secara teknis membuat sebar dari dua variabel terhadap satu sama lain dan kemudian menghubungkan ke neighboring points.



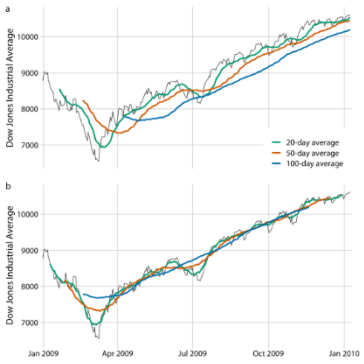
Gambar 70

13. Visualizing Trends

Smoothing. Mari kita pertimbangkan rangkaian waktu Dow Jones Industrial Average (singkatnya Dow Jones), indeks pasar saham yang mewakili harga 30 perusahaan besar AS milik publik. Secara khusus, kita akan melihat tahun 2009, tepat setelah crash 2008. Selama akhir kehancuran, dalam 3 bulan pertama tahun 2009, pasar kehilangan lebih dari 2.400 poin (~27%). Kemudian perlahan pulih untuk sisa tahun ini.

Gambar 71

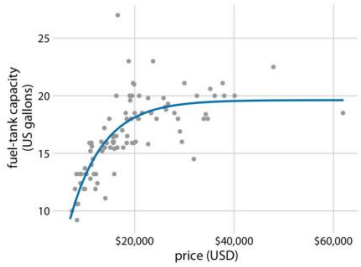
Dalam istilah statistik, kami mencari cara untuk smooth deret waktu pasar saham. Tindakan smoothing menghasilkan fungsi yang menangkap pola kunci dalam data sambil menghilangkan detail atau gangguan kecil yang tidak relevan. Analis keuangan biasanya memuluskan data pasar saham dengan menghitung moving averages. Untuk menghasilkan moving averages, kami mengambil time window, katakanlah 20 hari pertama dalam deret waktu, hitung harga rata-rata selama 20 hari ini, lalu pindahkan time window satu hari, sehingga sekarang mencakup hari ke-2 hingga ke-21. Kami kemudian menghitung rata-rata selama 20 hari ini, memindahkan jendela waktu lagi, dan seterusnya. Hasilnya adalah deret waktu baru yang terdiri dari urutan harga rata-rata.



Gambar 72

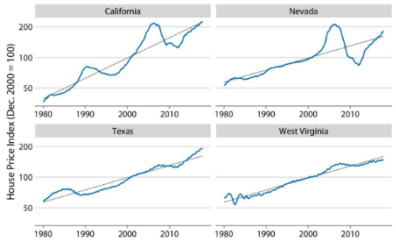
Showing Trends with a Defined Functional Form. Smoothers ini juga tidak memberikan estimasi parameter yang memiliki interpretasi yang berarti. Oleh karena itu, jika memungkinkan, lebih baik menyesuaikan kurva dengan bentuk fungsional spesifik yang sesuai untuk data dan yang menggunakan parameter dengan arti yang jelas.

Untuk data tangki bahan bakar, kita membutuhkan kurva yang awalnya naik secara linier tetapi kemudian turun pada nilai konstan. Fungsi $y = A – B \exp(–mx)$ mungkin sesuai dengan tagihan itu. Di sini, A, B, dan m adalah konstanta yang kita sesuaikan agar sesuai dengan kurva dengan data. Fungsinya kira-kira linier untuk x kecil, dengan $y \approx A – B + Bmx$; Ini mendekati nilai konstan untuk besar x, $y \approx A$, dan itu meningkat secara ketat untuk semua nilai x.



Gambar 73

Detrending and Time-Series Decomposition. Untuk deret waktu mana pun dengan tren jangka panjang yang menonjol, sebaiknya hapus tren ini untuk secara khusus menyoroti setiap penyimpangan penting. Teknik ini disebut detrending, dan saya akan mendemonstrasikannya di sini dengan harga rumah. Indeks Harga Rumah Freddie Mac dari tahun 1980 hingga 2017, untuk empat negara bagian terpilih (California, Nevada, Texas, dan Virginia Barat). Indeks Harga Rumah adalah angka tanpa unit yang melacak harga rumah relatif di wilayah geografis yang dipilih dari waktu ke waktu. Kami menurunkan harga perumahan dengan membagi indeks harga aktual pada setiap titik waktu dengan nilai masing-masing dalam tren jangka panjang. Secara visual, pembagian ini akan terlihat seperti kita mengurangkan garis abu-abu dari garis biru pada Gambar 74, karena pembagian nilai yang tidak ditransformasi setara dengan pengurangan nilai yang ditransformasi log. Indeks diskalakan secara sewenang-wenang sehingga sama dengan 100 pada bulan Desember tahun 2000. Garis biru menunjukkan fluktuasi bulanan dalam indeks dan garis abu-abu lurus menunjukkan tren harga jangka panjang di masing-masing negara bagian.



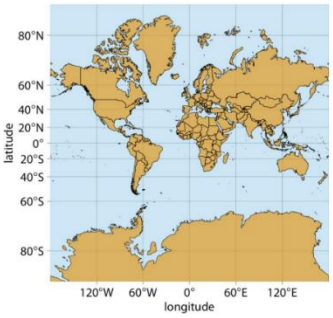
Gambar 74

14. Visualizing Geospatial Data

Projections. Bumi kira-kira berbentuk bulat dan lebih tepatnya berbentuk bulat pepat yang sedikit memipih di sepanjang sumbu rotasinya. Dua lokasi di mana sumbu rotasi berpotongan dengan spheroid disebut kutub (utara dan selatan). Untuk menentukan lokasi secara unik di bumi, kita memerlukan tiga informasi: di mana kita berada di sepanjang arah ekuator (garis bujur), seberapa dekat kita dengan salah satu kutub saat bergerak tegak lurus ke ekuator (garis lintang), dan seberapa jauh kita dari pusat bumi (ketinggian). Bujur, lintang, dan ketinggian ditentukan relatif terhadap sistem referensi yang disebut datum. Datum menentukan sifatsifat seperti bentuk dan ukuran bumi, serta lokasi nol bujur, lintang, dan ketinggian. Salah satu datum yang banyak digunakan adalah World Geodetic System (WGS) 84, yang digunakan oleh Global Positioning System (GPS).

Gambar 75

Salah satu proyeksi peta paling awal yang digunakan, proyeksi Mercator, dikembangkan pada abad ke-16 untuk navigasi bahari. Ini adalah proyeksi konformal yang secara akurat merepresentasikan bentuk tetapi memperkenalkan distorsi area yang parah di dekat kutub.

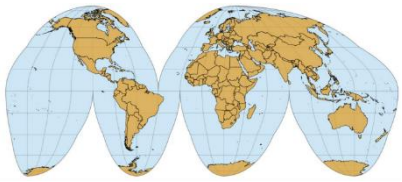


Gambar 76

Proyeksi Mercator dunia. Dalam proyeksi ini, kesejajaran adalah garis horizontal lurus dan meridian adalah garis vertikal lurus. Ini adalah proyeksi konformal yang mempertahankan sudut lokal, tetapi menimbulkan distorsi parah di area dekat kutub. Misalnya, Greenland tampak lebih

besar dari Afrika dalam proyeksi ini, padahal kenyataannya Afrika 14 kali lebih besar dari Greenland.

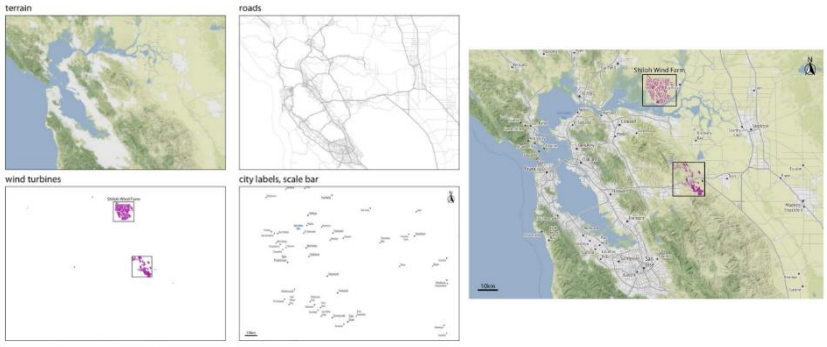
Proyeksi seluruh dunia yang secara sempurna melestarikan kawasan adalah Goode homolosine. Hal ini biasanya ditampilkan dalam bentuk terpotong, yang memiliki satu potongan di belahan bumi utara dan tiga potongan di belahan bumi selatan.



Gambar 77

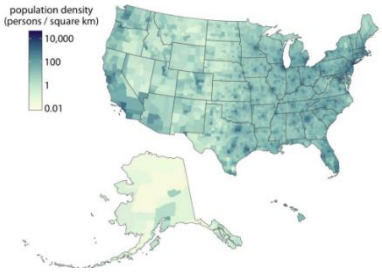
Layers. Untuk memvisualisasikan data geospasial dalam konteks yang tepat, kami biasanya membuat peta yang terdiri dari beberapa lapisan yang menunjukkan berbagai jenis informasi. Untuk mendemonstrasikan konsep ini, saya akan memvisualisasikan lokasi turbin angin di area San Francisco Bay. Di Bay Area, turbin angin dikelompokkan di dua lokasi. Satu lokasi, yang akan saya rujuk sebagai Shiloh Wind Farm, terletak di dekat Rio Vista dan yang lainnya terletak di sebelah timur Hayward dekat Tracy.

terdiri dari empat lapisan terpisah. Di bagian bawah, kita memiliki layer medan, yang menunjukkan perbukitan, lembah, dan air. Lapisan berikutnya menunjukkan jaringan jalan. Di atas lapisan jalan, saya telah menempatkan lapisan yang menunjukkan lokasi masing-masing turbin angin. Lapisan ini juga berisi dua persegi panjang yang menyoroti sebagian besar turbin angin. Terakhir, lapisan paling atas menambahkan lokasi dan nama kota. Keempat layer ini ditunjukkan secara terpisah.



Gambar 78

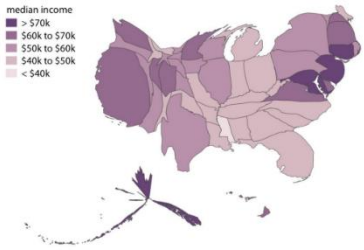
Choropleth Mapping. Kami sering ingin menunjukkan bagaimana beberapa kuantitas bervariasi di seluruh lokasi. Kita dapat melakukannya dengan mewarnai masing-masing wilayah di peta sesuai dengan dimensi data yang ingin kita tampilkan. Peta semacam itu disebut peta choropleth. Sebagai contoh Kepadatan populasi di setiap wilayah AS, ditampilkan sebagai peta choropleth. Kepadatan populasi dilaporkan sebagai orang per kilometer persegi.



Gambar 79

Cartograms. Tidak setiap visualisasi seperti peta harus akurat secara geografis agar berguna. Misalnya, masalah dengan beberapa negara bagian menempati area yang relatif luas tetapi berpenduduk jarang, sementara yang lain menempati area kecil namun memiliki jumlah penduduk yang besar. Bagaimana jika kita mengubah bentuk negara bagian sehingga ukurannya sebanding dengan jumlah penduduknya? Peta yang dimodifikasi seperti itu disebut kartogram.

Contohnya, Pendapatan median di setiap negara bagian AS, ditampilkan sebagai kartogram. Bentuk masing-masing negara bagian telah dimodifikasi sedemikian rupa sehingga luasnya sebanding dengan jumlah penduduknya.



Gambar 80

Pendapatan median di setiap negara bagian AS, ditampilkan sebagai heatmap kartogram.

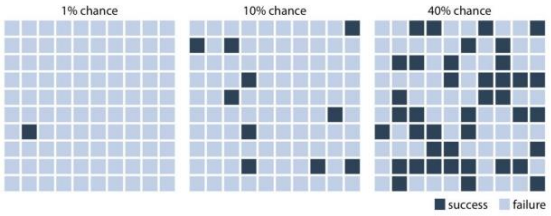
Gambar 81

15. Visualizing Uncertainty

Framing Probabilities as Frequencies. Sebelum kita dapat membahas bagaimana memvisualisasikan ketidakpastian, kita perlu mendefinisikan apa itu sebenarnya. Secara intuitif, kita dapat memahami konsep ketidakpastian dengan paling mudah dalam konteks peristiwa masa depan. Secara matematis, kita berurusan dengan ketidakpastian dengan menggunakan konsep probabilitas.

Sebagai contoh Memvisualisasikan probabilitas sebagai frekuensi. Ada 100 kotak dan setiap kotak mewakili keberhasilan atau kegagalan dalam beberapa percobaan acak. Peluang

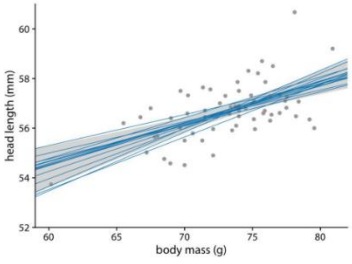
sukses 1% sesuai dengan 1 kotak gelap dan 99 kotak terang, peluang sukses 10% sesuai dengan 10 kotak gelap dan 90 kotak terang, dan peluang sukses 40% sesuai dengan 40 kotak gelap dan 60 kotak terang. Dengan menempatkan kotak gelap secara acak di antara kotak terang, kita dapat menciptakan kesan visual tentang keacakan yang menekankan ketidakpastian hasil percobaan tunggal.



Gambar 82

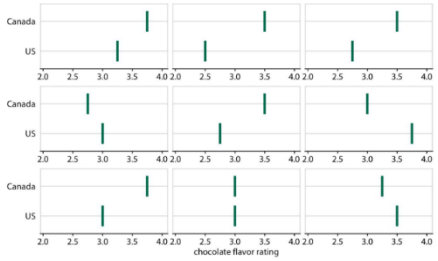
Visualizing the Uncertainty of Point Estimates. Semua ahli statistik menggunakan sampel untuk menghitung estimasi parameter dan ketidakpastiannya. Namun, mereka terbagi dalam cara mereka mendekati perhitungan ini, menjadi Bayesian dan frequentist. Bayesian berasumsi bahwa mereka memiliki pengetahuan sebelumnya tentang dunia, dan mereka menggunakan sampel untuk memperbarui pengetahuan ini. Sebaliknya, frequentist berusaha membuat pernyataan yang tepat tentang dunia tanpa memiliki pengetahuan sebelumnya. Untungnya, dalam hal memvisualisasikan ketidakpastian, Bayesian dan frequentist umumnya dapat menggunakan jenis strategi yang sama. Di sini, pertama-tama saya akan membahas pendekatan frequentist dan kemudian menjelaskan beberapa isu spesifik yang unik dalam konteks Bayesian. Frequentist paling sering memvisualisasikan ketidakpastian dengan error bars. Sementara error bars dapat berguna sebagai visualisasi ketidakpastian.

Visualizing the Uncertainty of Curve Fits. Estimasi tren ini juga memiliki ketidakpastian, dan ketidakpastian biasanya ditunjukkan dalam garis tren dengan confidence band. Confidence band memberi kita rentang garis kecocokan berbeda yang akan kompatibel dengan data. Untuk menggambar pita kepercayaan, kita perlu menentukan tingkat kepercayaan, dan seperti yang kita lihat untuk bilah kesalahan dan probabilitas posterior, akan berguna untuk menyoroti tingkat kepercayaan yang berbeda.



Gambar 83

Hypothetical Outcome Plots. Untuk mengilustrasikan konsep HOP, mari kita kembali sekali lagi ke cocholat bar rating. Saat Anda berdiri di toko kelontong berpikir untuk membeli cokelat, Anda mungkin tidak peduli dengan peringkat rasa rata-rata dan ketidakpastian yang terkait untuk kelompok cokelat batangan tertentu. Skema plot hasil hipotetis untuk cocholate bar rating dari batangan yang diproduksi di Kanada dan AS. Setiap bilah hijau vertikal mewakili peringkat untuk satu bilah, dan setiap panel menunjukkan perbandingan dua bilah yang dipilih secara acak, masing-masing dari pabrikan Kanada dan pabrikan AS.



Gambar 84