

2

Bab 2: Menerjemahkan Angka ke Gambar

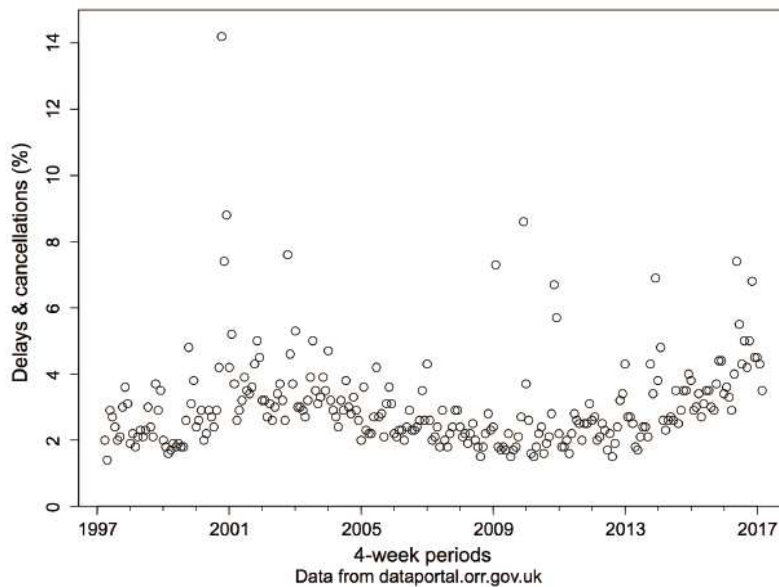
Bab ini membahas bagaimana mentranslasi angka ke gambar sehingga audien mudah mentranslasikannya ke otak untuk dicerna.

Pelopor visualisasi data William Cleveland menulis bahwa visualisasi menerjemahkan angka menjadi gambar pada halaman atau layar, sehingga pembaca dapat menerjemahkannya kembali dalam otak. Ada banyak cara bagaimana membuat angka menjadi beberapa aspek gambar. Karena itu, pilihan harus dibuat dengan hati-hati. Audien harus dapat melakukannya dengan mudah. Jika terdapat hal yang membingungkan, maka akan menghalangi audien menerima pesan dari visualisasi. Pembaca mungkin hanya memiliki waktu sedikit dan mungkin saja fobi terhadap matematika. Karena itu perlu melihat dari sudut pandang pembaca. Jika memiliki sesuatu yang penting untuk ditunjukkan, maka lakukan dengan jelas tanpa membuat pembaca berusaha keras menangkapnya. Pembaca harus dapat menerjemahkannya kembali ke angka, dan lakukan dengan cepat. Pembaca juga harus sampai pada angka yang benar dan tidak disesatkan oleh visualisasi.

2.1 Pengkodean Visual Data

Bagian ini akan membahas konsep penting visualisasi data, yaitu ketika dilakukan translasi angka ke dalam gambar, maka harus dihubungkan angka-angka dalam data dengan parameter pada gambar. Hal ini dikenal sebagai *pengkode-an visual data*. Mari kita lihat contoh sederhana berikut. Data berikut akan menjawab apakah keterlambatan kereta di South East England terburuk di musim

gugur, ketika rel menjadi licin karena adanya sampah daun yang gugur dan menutupi lintasan. Data diunduh dari orr.gov.uk dan dirapikan, (robertgrantstats.co.uk/datavizbook). Ada beberapa variabel yang harus dipakai, yaitu waktu (data tersedia untuk periode 4 minggu, dan ada sekitar 13 di masing-masing tahun) dan persentase perjalanan kereta api dicatat sebagai dibatalkan atau "tertunda secara signifikan." Waktu dibagi waktu menjadi tahun dan periode 4 minggu dalam tahun, atau dapat juga diidentifikasi musim gugur versus sisa tahun. Data mencakup dua puluh tahun sejak April 1997 hingga Maret 2017. *Scater Plot* adalah salah satu pilihan yang jelas, dengan waktu dikodekan ke lokasi horisontal, dan penundaan dikodekan ke lokasi vertikal. Seperti ditunjukkan pada Gambar 2.1, dua variabel dikodekan (waktu dan penundaan / pembatalan) ke dua parameter gambar, dan kemudian penanda digambar untuk setiap pengamatan di lokasi yang dihasilkan. Itu % keterlambatan dan pembatalan adalah nilai mendekati nol, jadi sumbu vertikal diperpanjang hingga 0%. Ini membantu untuk menyampaikan nilai-nilai penanda sebagai ketinggian obyek di grafik. Namun, ini belum dibuat untuk sumbu horizontal karena tidak ada yang bernilai nol.

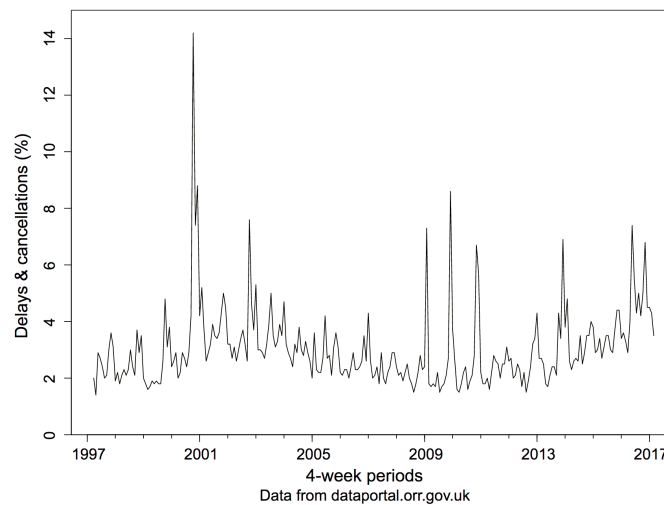


Gambar 2.1 *Scater Plot* Penundaan Perjalanan Kereta di Inggris.

Pengkodean waktu ke lokasi horizontal masuk akal secara intuitif untuk pembaca, karena pembaca membaca dari kiri ke kanan. Dari

grafik terlihat tren umum berikut: semakin buruk, lalu lebih baik, lalu lebih buruk lagi. Terlihat juga, periode buruk sebagai poin tinggi pada grafik, tetapi sulit untuk membandingkan musim gugur di tahun yang berbeda karena mereka tersebar melintasi plot.

Format alternatif dengan penyandian yang sama adalah bagan garis (*line chart*) (Gambar 2.2). Meskipun variabel yang sama dikodekan ke yang sama parameter visual, dimana data dihubungkan dengan garis. Ini memiliki efek yang berbeda. Tampak poin yang sangat tinggi lebih jelas, dan tren jangka panjang kurang begitu jelas. Poin tinggi menghasilkan garis untuk ke puncak lonjakan dan kembali turun, sehingga menghasilkan kesan garis yang lebih tebal dan menarik perhatian pembaca. Dan ini tidak terlihat dengan scatter.

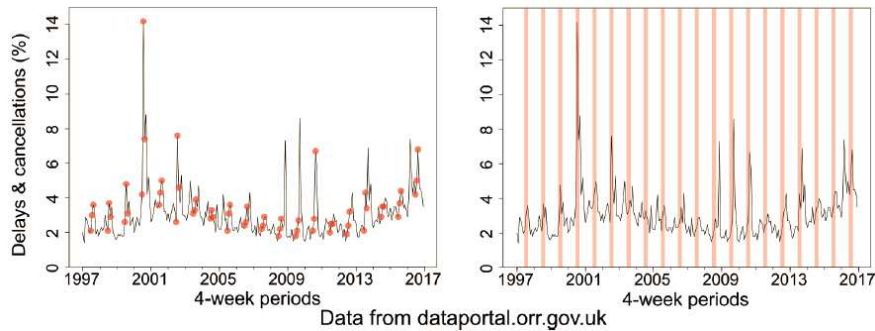


Gambar 2.2 Grafik garis keterlambatan kereta.

Secara statistik, persentase penundaan/pembatalan data cenderung positif (*skewed positive*) : sebagian besar nilainya hampir sama, beberapa jauh lebih tinggi, tetapi tidak ada nilai yang lebih rendah. Distribusi semacam ini dapat menginformasikan pilihan pengkodean dan format yang sesuai. Kita akan bahas topik distribusi data secara lebih rinci di bagian selanjutnya.

Tujuan visualisasi ini adalah membandingkan musim gugur ke waktu lain dalam setahun. Salah satu pendekatan telah dilakukan dengan dua cara pada Gambar 2.3. Di Gambar kiri, satu periode setiap tahun

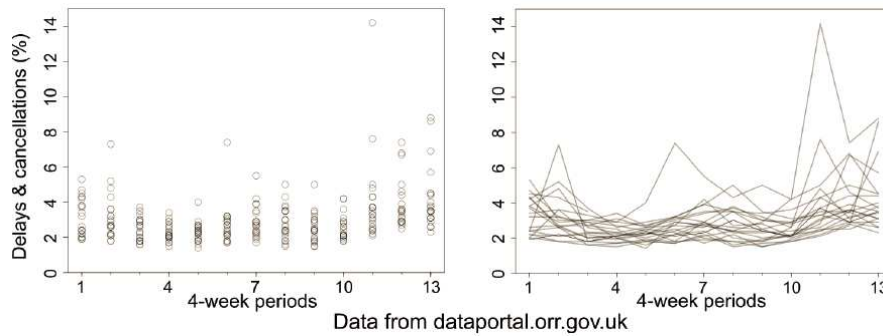
(pertengahan November hingga pertengahan Desember) dipilih dengan satu titik berwarna, dan di gambar kanan, *color shading* digunakan untuk mengidentifikasi pertengahan September hingga pertengahan Desember. Titik-titik memiliki keuntungan mampu menunjukkan di mana tepatnya data di musim gugur. Shading tidak mengganggu penanda data, tetapi efek *candy-stripe* menciptakan visualisasi yang berlebihan.



Gambar 2.3 Grafik keterlambatan kereta: bagan garis dengan *highlight*. Pada gambar kiri titik mengidentifikasi musim gugur di setiap tahun dan visualiasi mungkin berhasil, tetapi *candy-stripe* di gambar kanan mengganggu pembaca.

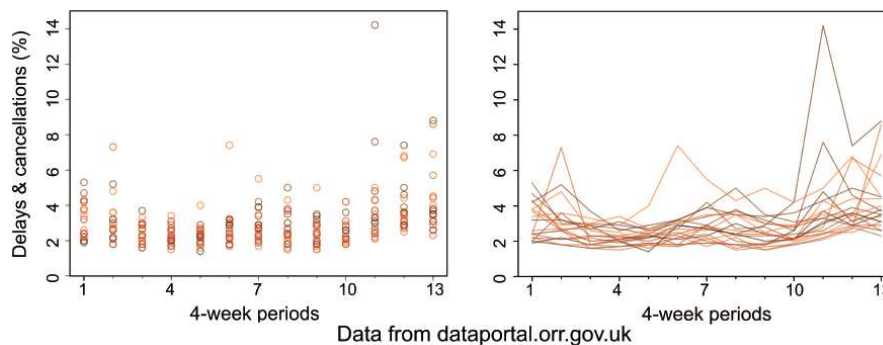
Pada tahap ini, beberapa kesimpulan baru mulai muncul, akibat penyandian dan format yang dipilih, meskipun musim gugur memang waktu terburuk untuk perjalanan kereta api dalam sebagian besar tahun hingga 2002, dari tahun 2003 dan seterusnya, periode terburuk cenderung terjadi kemudian, di musim dingin.

Pada Gambar 2.3, variabel baru dibuat untuk mengidentifikasi musim gugur. Ini hanya varian pada variabel waktu yang ada. Pendekatan lain bisa digunakan untuk membantu pembaca membandingkan musim gugur dengan sisa tahun, sekali lagi membuat varian pada variabel waktu dengan membaginya menjadi tahun dan titik-titik 4 minggu. Kemudian, periode 4 minggu ke posisi horizontal posisi (Gambar 2.4).



Gambar 2.4 Grafik keterlambatan kereta: plot sebaran dan bagan garis yang tidak baik untuk periode dalam setiap tahun.

Gambar 2.4 jelas tidak membantu! Di scatter plot, kita bisa tidak lagi tahu tahun mana titik tertentu dimiliki, dan dalam bagan garis, meskipun ada satu garis per tahun, mereka berbaur sedemikian rupa sehingga tidak mungkin untuk melacak garis apa pun di grafik. Ini karena sekarang terdapat tiga variabel yang harus dikodekan: tahun, periode 4 minggu, dan % ditunda, dan tahun diabaikan. Meski menggunakan posisi horisontal dan vertikal, tapi perlu parameter bagan lain untuk menunjukkan tahun mana masing-masing penanda atau garis terdapat. Salah satu pilihan adalah menggunakan warna yang berbeda, tetapi harus dialokasikan warna yang memiliki arti tertentu dari urutan, bukan hanya efek pelangi, jika sebaliknya maka urutan tahun akan hilang (Gambar 2.5).



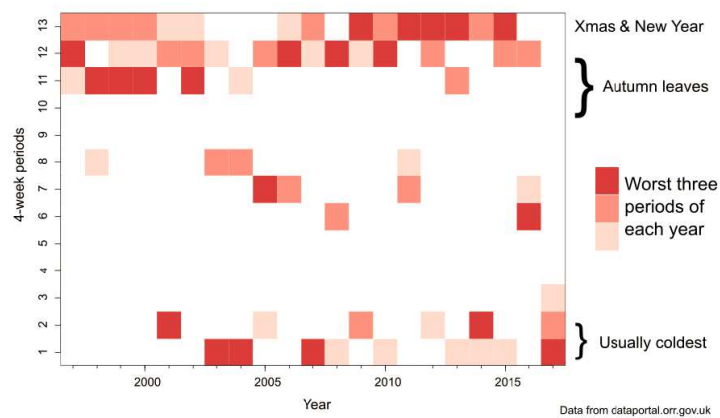
Gambar 2.5 Grafik Keterlambatan kereta: *scatter plot* dan bagan garis dengan yang lama data berwarna hitam, mulai data terbaru berwarna coklat.

Ini sedikit membantu dalam bagan garis (bandingkan garis tertinggi pada periode 10 dengan garis yang sama pada Gambar 2.4), tetapi tidak sama sekali dalam *scatter plot*. Jika dimiliki data beberapa

tahun, warnanya mungkin saja kontras. Tapi yang mendasar masalahnya di sini adalah bahwa tahun-tahun semuanya sangat mirip dan garis-garis tumpang-tindih satu sama lain. Ini harus dipertimbangkan jika ingin pesannya jelas dan tidak dibanjiri oleh *noise* yang tidak relevan.

Ada beberapa trik yang dapat diterapkan untuk membantu keluar dari masalah ini, yaitu merubah format dan akan dibahas lebih rinci di bab selanjutnya. Diantaranya adalah dengan memuluskan garis menampilkan beberapa grafik kecil, satu untuk setiap tahun, memilah data dengan analisis deret waktu statistik, atau menyajikannya secara digital sebagai animasi atau web interaktif.

Saya ingin menutup contoh ini dengan keluar dari pengkodean horizontal-vertical data keterlambatan dan waktu. Mari fokus pada waktu tahun dimana penundaan dan pembatalan benar-benar buruk. Pada Gambar 2.6, data ditampilkan sebagai kotak berwarna, dengan tahun dikodekan ke posisi horisontal, Periode 4 minggu ke posisi vertikal, dan % penundaan/pembatalan dengan warna. Untuk menghindari kelebihan warna yang kita lihat sebelumnya, hanya tiga periode terburuk dalam setiap tahun diwarnai.



Gambar 2.6 Grafik keterlambatan kereta: grid periode tahun, dengan periode 4 minggu terburuk di setiap tahun diwarnai.

Grid warna seperti ini (atau Gambar 1.4) sering disebut *heat map*. Ini memiliki keuntungan bahwa cerita yang ingin di sampaikan akan menjadi fokus. Jika ada efek daun musim gugur konsisten, itu akan muncul dengan sendirinya sebagai garis warna horizontal yang jelas.

Sebenarnya % di masing-masing periode terburuk diabaikan, dan mereka secara sederhana diwarnai menurut apakah yang terburuk, terburuk ke-2, atau ke-3 untuk setiap tahun, dan yang lainnya dibiarkan putih. Ini sengaja membuang banyak informasi di dalam data untuk memfokuskan pada kejelasan. Ini bukan keputusan yang bisa dianggap enteng.

Pada akhir contoh, harus dipahami perbedaan antara memilih format dan pengodean data ke parameter visual. Dengan contoh grafik terakhir dari keterlambatan kereta, terlihat masalah telah bergeser dari musim gugur ke bulan - bulan terdingin di musim dingin - Desember hingga Februari. disadari bahwa pilihan encoding dan format terikat dengan apa pesan yang ingin Anda komunikasikan, dan tidak ada jawaban benar yang universal untuk satu set data.

2.2 Pemilihan Parameter Visual

Ada banyak parameter yang bisa kita *encode*. Penelitian tentang persepsi visual manusia menunjukkan bahwa tidak mudah bagi pembaca untuk menerjemahkan grafik kembali ke angka. Area, misalnya, biasanya dianggap lebih mirip satu sama lain daripada yang sebenarnya. Ini akan dibahas lebih jauh di selanjutnya. Dalam Tabel 2.1, ditunjukkan parameter yang memiliki masalah seperti ini ditandai dengan tanda bintang. Tidak ideal menggunakan volume untuk menggambarkan variabel kuantitatif, karena sangat sulit untuk menilai ukuran relatif yang tepat dari setiap "objek" seperti yang digambarkan pada halaman atau layar. Volume bisa digunakan untuk data ordinal, atau parameter yang lain seperti panjang, akan lebih cocok. Namun, ada pengecualian untuk aturan di visualisasi data, dan untuk contoh memanfaatkan volume untuk menyampaikan skala data dan menarik perhatian.

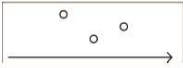
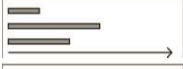

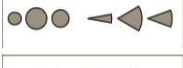






Dalam bab pendahuluan disebutkan variabel dapat bersifat kategori (dengan ordinal sebagai kasus khusus) dan kuantitatif. Beberapa parameter ini hanya cocok untuk memiliki kategori non-ordinal untuk dikodekan, sementara parameter lain lebih cocok untuk nilai-nilai ordinal.

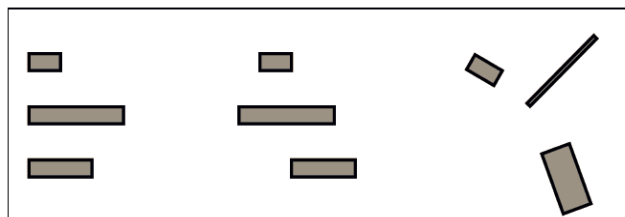
Pengukuran terhadap skala umum merupakan hal penting. Cukup mudah untuk menilai panjang relatif dari tiga baris yang dimulai pada titik yang sama, seperti pada tabel di atas, tetapi jauh lebih sulit jika tidak berdekatan. Dengan kata lain, mereka berbeda di lokasi

horizontal, sendirian ketika mereka berbeda dalam parameter lain seperti sudut atau lebar (Gambar 2.7). Ini adalah salah satu alasan mengapa diagram lingkaran sulit diterjemahkan kembali ke nilai dengan tingkat akurasi yang dapat diterima. Ada desain melingkar lainnya seperti "*radiogram histogram*" dan "*sunburst diagram*" juga mengalami masalah yang sama. Grafik tersebut mungkin *eye catching*, tetapi tidak tepat untuk komunikasi kuantitatif yang serius, dan bukan sebagai hal baru dan menarik.

Mari pelajari kasus pada Gambar 2.7 yang menunjukkan data pada rasio 1: 3: 2 ditampilkan dengan 3 cara visual. Pembaca visualisasi mungkin tidak berkonsentrasi sama sekali tetapi menyimpulkan berdasarkan kesan pertama. Karena itu sangat penting untuk menghilangkan segala peluang untuk kesalahpahaman seperti ini. Pelajaran dari contoh ini adalah, jika membandingkan beberapa nilai penting, nilai-nilai itu harus menonjol dan berdekatan.

Tabel 2.1 Parameter Visual

Parameter	Ideal data type	
Position	Quantitative	
Length	Quantitative	
Angle	Quantitative*	
Area	Quantitative*	
Volume	Quantitative*	
Color hue	Nominal or ordinal*	
Color saturation	Ordinal	
Marker shape	Nominal*	
Features	Quantitative*	
Line width	Ordinal*	



Gambar 2.7 Gambar rasio 3 buah obyek disajikan berbeda secara visual. Panjang dari tiga batang di sebelah kiri dapat dibandingkan dengan mudah, tetapi kurang untuk tiga batang di tengah, di mana posisi dasar telah bergeser, atau di sebelah kanan di mana lebar dan sudut juga berubah.

2.3 Pengertian Warna

Dari semua parameter gambar yang dapat di-*encode*, layak untuk dipelajari tentang bagaimana kita dapat mengukur warna. Ada

beberapa sistem untuk menentukan warna yang tepat, dalam buku ini sebagian besar digunakan sistem RGB (merah, hijau, biru). Hampir tujuh belas juta warna bisa dibuat dengan memberikan angka dari 0 hingga 255 dengan mencampur jumlah merah, hijau dan biru. Sebagai contoh persegi panjang paling gelap pada Gambar 2.6 memiliki nilai RGB 173, 56, 41. Komponen merah (173) adalah jumlah terbesar, diikuti dengan sedikit hijau (56) dan biru (41). Hitam pekat adalah 0, 0, 0 dan putih murni adalah 255, 255, 255.

Ini berarti warna apa pun dapat didefinisikan dengan tiga angka. Sistem notasi warna HSV menjelaskan warna dalam hal *Hue*/rona (warna dasar dari spektrum), *Saturation*/saturasi dan *Value*/nilai (tambah terang atau gelap) - istilah yang digunakan dalam Tabel 2.1. Sistem apa pun yang digunakan, yang jelas otak manusia sulit membedakan jarak antar warna dalam hal pencampuran warna merah, hijau dan biru, jadi encode tiga variabel dari data untuk warna-warna primer merupakan ide kurang baik.

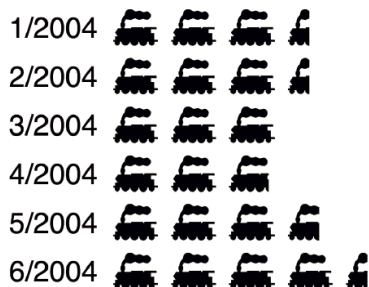
Jika warna merupakan bagian dari data, maka ide yang bagus untuk menggambar warna yang dimaksud dalam visualisasi, dan bukan tiga parameter lainnya. Namun, penggunaan warna yang lebih sederhana dapat berfungsi dengan baik, dibandingkan dengan mengkodekan keseluruhan spektrum warna, seperti opsi saturasi dalam Tabel 2.1, atau bagian kanan opsi rona.

Meskipun "jarak" antara warna tidak mampu dilihat, informasi masih bisa disampaikan melalui derajat terang/gelap (informasi ordinal) seperti pada Gambar 2.6.

2.4 Keterbatasan Bentuk Bidang

Satu masalah yang sangat umum dalam visualisasi data adalah pengkodean variabel numerik ke bentuk bidang, dimana audien sulit menterjemahkan balik dari bentuk visual ke numerik. Sebagai contoh *Bubble Chart* dalam bab ini memiliki sifat ini. Pembaca mungkin bisa mencari penanda terbesar dan terkecil, tetapi akan sulit untuk membandingkan ukuran obyek yang tidak berdekatan. Masalah yang sama mempengaruhi piktogram di mana variabel dikodekan ke ukuran relatif dari beberapa ikon. Luasnya tidak sebanding dengan ketinggian ikon atau gelembung, tapi begitulah yang terlihat. *Word clouds* memiliki masalah yang lebih dalam dimana pembaca hampir tidak mampu menilai bidang-bidang relatif kata-kata dengan

panjang berbeda, dengan karakter ukuran berbeda, dalam warna berbeda dan rotasi yang berbeda. Jika Anda ingin menggunakan ikon, ide yang lebih baik adalah dengan membaris ikon (Gambar 2.8). Ini juga disebut *pictogram*, dan kadang-kadang bagan *Isotype*. Karena di sini variabel dikodekan melalui panjangnya, maka jauh lebih mudah untuk dilihat.

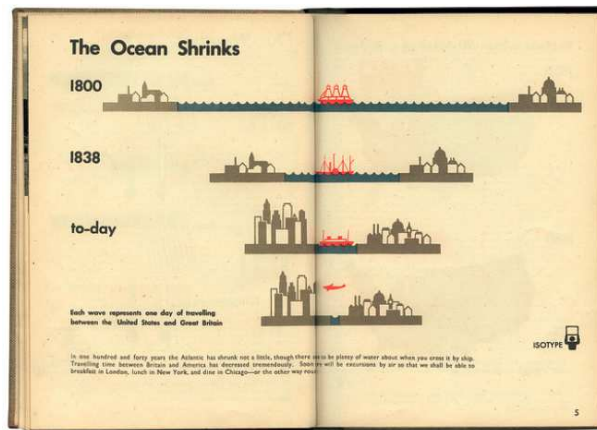


Each icon represents 1% of trains cancelled or delayed.

Gambar 2.8 Piktogram dalam Isotype.

Ada jenis piktogram khusus lain yang jarang digunakan, mungkin karena butuh banyak upaya untuk menghasilkannya, di mana ikon adalah benda nyata dalam sebuah foto. Gambar 2.10 adalah *photomontage*, menggabungkan foto-foto terpisah dari setiap pesawat lepas landas Kompleks Selatan Bandara Internasional Los Angeles dalam sehari, sementara Gambar 2.11 ditampilkan dalam kehidupan nyata. Karena gambar tersebut mengandalkan bidang, maka pembaca akan salah taksir menilai ukuran relatif dari irisan di diagram lingkaran dan grafik donat. Irisan di bagian atas atau bawah pie (atau donat) akan tampak lebih kecil dari irisan ukuran yang sama di kiri atau kanan. Pembaca biasanya menganggap lebih kecil pada sudut kurang dari 90 derajat dan melebih-lebihkan sudut yang lebih besar.

Warna dan efek 3-D yang bervariasi dapat menipu pembaca dan akan dieksplorasi lebih lanjut dalam bab selanjutnya tentang persepsi visual dan otak. Meskipun sudut proporsional dengan area (jika jari-jari tetap sama), perbedaan orientasi obyek seperti Gambar 2.7 juga akan berefek sama pada grafik pie dan donat.



Gambar 2.9 “Penyusutan Laut”, visualisasi oleh Otto dan Marrie Neurath dari Isotype Institute.



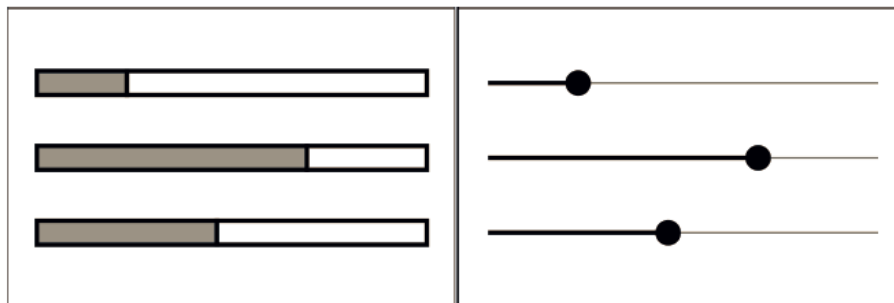
Gambar 2.10 “Wake Turbulence” oleh photographer Mike Kelley Mevisualisasi pesawat-pesawat yang take off pada satu hari dari bandara Los Angeles.

Untuk melakukan visualisasi data yang akurat dan membantu pembaca maka grafik pai dan donat harus dihindari. Dua alternatif yang baik adalah *bar chart* yang dikenal dan *wafel plot*. *bar chart* hanya mengkodekan variabel kuantitatif dengan panjang dari persegi panjang dan dapat dengan mudah dibandingkan karena memiliki

skala yang sama, seperti terlihat pada Gambar 2.7. Jika ingin disampaikan kesan tentang kontribusi setiap bilah secara keseluruhan, maka dapat dibuat semua bilah hingga tanda 100% dan warnai dalam proporsi yang mewakili setiap pengamatan dalam data.

Informasi yang sama dapat disampaikan dengan cara mengganti *bar* dengan garis, dan ditumpang-susunkan dengan lingkaran; ini disebut *dot plot*.

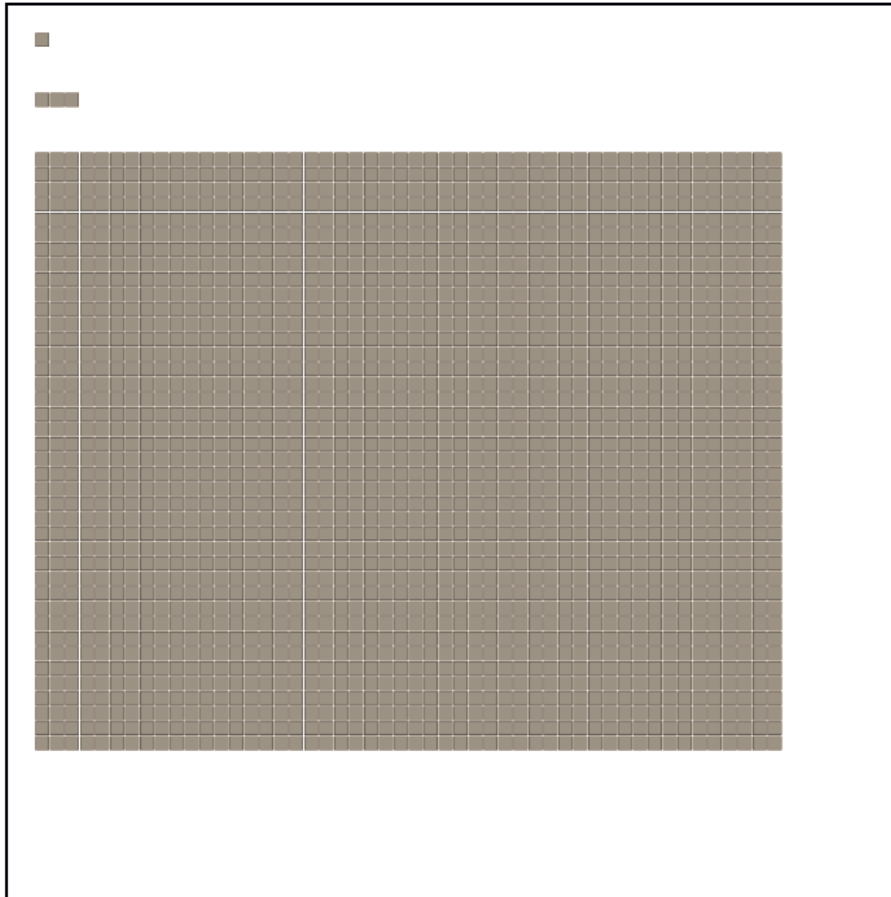
Wafel Plot adalah jenis pictogram khusus, di mana kotak kecil tersusun rapi berdekatan. Ini mirip dengan grafik batang dimaka kita bisa melihat pada panjang deretan kotak, tetapi jika nilai dalam data sangat berbeda, maka kotak bisa diatur dalam dua dimensi. Ini memiliki banyak kelebihan untuk mengakomodir nilai yang lebih besar. Tiga kategori dapat ditampilkan dalam rasio 1: 3: 2, seperti pada Gambar 2.12, tetapi juga bisa ditunjukkan dalam rasio 1: 3: 2000 (Gambar 2.13).



Gambar 2.12 Proporsi-proporsi ditampilkan dalam *bar chart* (kiri) dan dalam *dot plot* (kanan).

2.5 Anotasi

Tujuan visualisasi agar informasi dapat dipahami dengan jelas. Anotasi dengan teks dalam hal ini dapat membantu. Masalahnya jika terlalu banyak teks yang akan ditampilkan akan memberikan kesan keseluruhan yang membingungkan. Menurut Edward Tufte solusinya adalah dengan menggunakan data rasio. Aturan umum yang baik adalah dengan membayangkan visualisasi sedang dibaca terpisah dari konteks yang menyertainya: akankah disalahpahami? Memberi label sumbu dan mengatakan dari mana data berasal adalah sangat penting, dan banyak visualisasi mendapat manfaat dari judul.



Gambar 2.13 Nilai-nilai yang sangat berbeda ditampilkan berdampingan dalam *waffle plot*: 1, 3 dan 2000, jika ditampilkan dalam *bar chart* dua angka pertama tidak terlihat.

2.5 Pengujian Pengguna

Bab ini telah membahas berbagai pengkodean dan format visual. Akan ada banyak lagi yang dibahas di bab-bab selanjutnya. Apapun pendekatan yang diambil, visualisasi data yang efektif hanya dijamin jika berfungsi bagi pembacanya. Ketika melihat data orang lain, pertimbangkan apa yang disukai dan tidak disukai, dan kemudian cari cara bagaimana memperbaikinya. Dengan cara ini kita akan belajar cepat dari kesalahan orang lain.

Jika menemukan sesuatu yang membingungkan, jangan menganggap itu salah, kuncinya visualisasi harus sederhana dan intuitif. Jika Anda membuat visualisasi data, harus dipertimbangkan pengujian pengguna. Perlihatkan draft kepada beberapa orang (sampel) audiens yang dituju untuk mendapatkan pendapat mereka. Pengalaman pengguna - kemudahan penggunaan, aksesibilitas dan kenikmatan – merupakan hal penting. Jadi gambar, teks, warna skema, anotasi, dan terminologi harus seiring agar pesan dapat tersampaikan dan memungkinkan pengguna mengeksplorasi data secara visual.

Meskipun format standar dapat berkerja baik, ada juga pembaca tertarik pada kebaruan dan juga bisa karena rasa ingin tahu tentang sesuatu yang baru. Jangan menghindar dari inovasi, tapi hati-hati untuk mengujinya.