بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ ;اَلْحَمْدُ لِلّٰهِ رَبِّ الْعَالَمِيْنَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Pertemuan 4 - Mata Kuliah Data Mining - Eksplorasi Data Menggunakan Statistik -

Sub Capaian Pembelajaran:

1. Memahami tentang variasi data
2. Memahami tentang perbedaan probabilitas dan statistik
3. Memahami tentang exploratory data analysis (EDA)
4. Memahami tentang statistik deskriptif
5. Memahami tentang visualisasi data
6. Mehami tentang korelasi antar variabel

Referensi:

1. Gutman, A. J., & Goldmeier, J. (2021). Becoming a data head: How to think, speak, and understand data science, statistics, and machine learning. John Wiley & Sons. (BAB 3 dan BAB 5)

Terjemahan dan ringkasan BAB 3 dan BAB 5 buku Becoming a data head: How to think, speak, and understand data science, statistics, and machine learning. John Wiley & Sons (Gutman, A. J. & Goldmeier, J., 2021)

4. Eksplorasi Data Menggunakan Statistik

4.1. Bersiaplah untuk Berpikir Secara Statistik

4.1.1. Mengajukan Pertanyaan

Prinsip inti pemikiran statistik adalah "mengajukan pertanyaan." Banyak dari kita melakukan hal ini sampai taraf tertentu dalam kehidupan sehari-hari. Anda mungkin mempertanyakan klaim dari pengiklan (seperti "Turunkan berat badan 10 kg dalam sebulan!") dan postingan aneh di media sosial.

Contoh lainnya, setiap pemilu politik, luangkan waktu sejenak untuk berpikir serius dan jujur tentang seberapa cepat Anda curiga terhadap klaim atau angka dari partai politik. Apa yang terlintas di kepalamu? "Sumber mereka buruk. Informasi mereka salah. Mereka tidak mengerti apa yang sedang terjadi."

Contoh lebih dekat dengan anda, pikirkan tentang informasi yang Anda lihat di tempat kerja. Ketika Anda melihat data tersebar di seluruh spreadsheet dan presentasi PowerPoint. Informasi yang berdampak pada kesuksesan perusahaan Anda, kinerja pekerjaan Anda, kemungkinan bonus Anda, apakah data tersebut dipandang dengan skeptis? Berdasarkan pengalaman kami, kadang-kadang angka-angka di ruang rapat dipandang sebagai fakta yang sulit dipercaya. Akan tetapi, mungkin karena Anda tidak punya waktu untuk bertanya, atau mengumpulkan lebih banyak data. Ini adalah data yang Anda miliki, data yang harus Anda tindak lanjuti, dan data yang dapat Anda tunjukkan dan salahkan jika segala sesuatunya tidak berjalan sesuai keinginan Anda. Ketika dihadapkan pada kendala dan keterbatasan ini, skeptisisme padam. Alasan lain, yang mungkin Anda sarankan, adalah meskipun Anda memahami masalah data, atasan Anda mungkin tidak

memahaminya. Reaksi berantai terjadi ketika setiap orang berasumsi bahwa orang lain, di tingkat atas rantai manajemen atau bahkan di tingkat bawah, akan menerima begitu saja angka tersebut, dan asumsi tersebut menyebar hingga ke kita yang hanya melihat spreadsheet. Mereka akan menganggap hal itu benar, jadi kami akan bertindak seolah-olah hal itu benar.

Komentar pada "Statistical Thinking"

Kami menggunakan "statistical thinking" dalam pengertian umum sebagai istilah. Anda mungkin lebih menyukai istilah probabilistic thinking, statistical literacy, atau mathematical thinking. Apa pun frasa yang Anda pilih, semuanya berhubungan dengan evaluasi data atau bukti.

Beberapa orang mungkin bertanya-tanya mengapa pemikiran ini penting. Bisnis dan kehidupan pada umumnya akan berjalan tanpa adanya hal tersebut. Jadi kenapa sekarang? Mengapa Kepala Data harus peduli?

Dalam artikel berjudul, "Data Science: What the Educated Citizen Needs to Know," ekonom dan dokter Harvard Alan Garber menjelaskan alasannya:

Manfaat data science adalah nyata dan sangat menonjol atau penting. Prediksi yang semakin akurat akan menjadikan produk data science lebih berharga dari sebelumnya, dan akan meningkatkan minat terhadap bidang ini. Kemajuan yang dicapai juga dapat menumbuhkan rasa puas diri dan membutakan kita terhadap kelemahan. Pekerja masa depan perlu menyadari tidak hanya apa yang dilakukan data science untuk membantu mereka dalam pekerjaan mereka, namun juga di mana dan kapan kekurangannya. . . . pemahaman yang lebih dalam tentang probabilistic reasoning dan evaluasi bukti adalah keterampilan umum yang akan membantu semuanya dengan baik.

4.1.2. Ada Variasi dalam Segala Hal

Pengamatannya bervariasi, ini bukanlah berita yang menggemparkan. Pasar saham berfluktuasi setiap hari, angka jajak pendapat politik berubah tergantung minggunya, harga bahan bakar naik dan turun, dan tekanan darah Anda berubah-ubah. Bahkan perjalanan harian Anda ke tempat kerja,

jika Anda membaginya menjadi hitungan detik, akan sedikit berbeda setiap hari tergantung pada lalu lintas, cuaca, keharusan mengantar anak ke sekolah, atau berhenti untuk minum kopi. Ada variasi dalam segala hal.

Anda mungkin menerima atau setidaknya menoleransi variasi ini dalam kehidupan sehari-hari Anda dan mungkin merasa nyaman dengannya. Namun secara keseluruhan, kami memahami bahwa ada beberapa hal yang berubah karena alasan yang tidak selalu dapat kami jelaskan. Dan ketika menyangkut hal-hal seperti mengisi ban, memompa bensin, atau membayar tagihan listrik, kita hidup dengan angka-angka yang berbeda setiap kali diukur, selama angka-angka tersebut masuk akal secara intuitif.

Penjualan suatu bisnis berfluktuasi setiap hari, mingguan, bulanan, dan tahunan. Hasil survei kepuasan pelanggan dapat sangat bervariasi dari hari ke hari. Jika kita menerima kenyataan adanya variasi dalam hidup kita, kita tidak perlu menjelaskan setiap peak dan valley. Namun, inilah yang akan coba dilakukan oleh dunia usaha. Apa yang dilakukan pada minggu penjualan tinggi? kepemimpinan bertanya. Mari ulangi yang baik, kurangi yang buruk, kata mereka.

Sebenarnya ada dua jenis variasi. Salah satu jenis variasi berasal dari cara data dikumpulkan atau diukur. Ini disebut measurement variation. Yang kedua adalah keacakan yang mendasari proses itu sendiri. Ini disebut random variation. Perbedaannya mungkin tampak sepele pada awalnya, namun di sinilah statistical thinking menjadi penting. Apakah keputusan diambil sebagai respons terhadap variasi acak yang tidak dapat dikendalikan? Atau apakah variasi tersebut mencerminkan suatu proses mendasar yang sebenarnya, yang jika diungkapkan dengan benar, dapat dikendalikan?

Sederhananya, variasi menciptakan ketidakpastian. Mari kita lihat satu skenario hipotetis dan satu studi kasus historis mengenai variasi yang menyebabkan ketidakpastian.

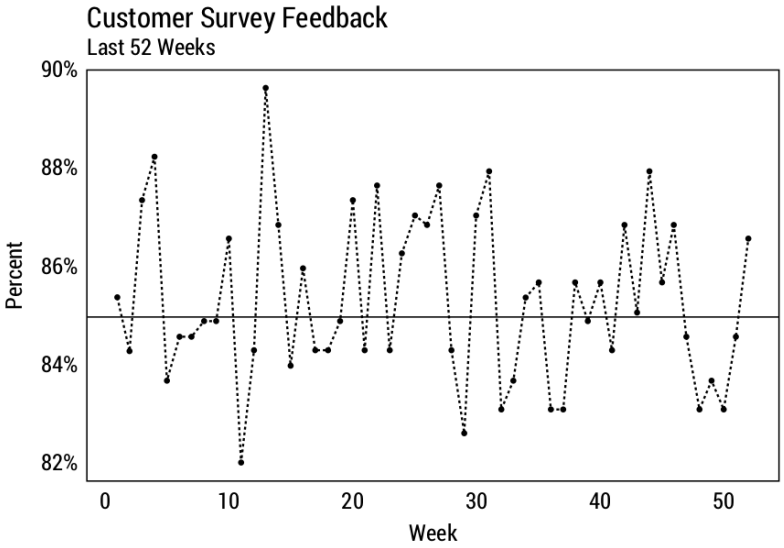
Skenario: Persepsi Pelanggan
Anda adalah manajer retail, dan kantor perusahaan memantau dengan cermat data kepuasan pelanggan dari toko Anda, yang dikumpulkan dari survey pelanggan. Survei tersebut meminta pelanggan untuk menilai kepuasan mereka pada skala 1−10, 10 berarti “sangat puas.” Banyak pertanyaan lain disertakan, tapi hanya pertanyaan pertama yang penting.

Untuk menambah taktiknya, perusahaan hanya menginginkan angka 9 dan 10. Angka-angka dikumpulkan setiap minggu dan dikirimkan kepada Anda, sebagai manajer toko, dan kantor perusahaan dalam file PDF dengan grafik berwarna. Angka-angka ini mempengaruhi bonus Anda, dan bonus atasan Anda, dan diperiksa dengan obsesif setiap minggunya saat Anda mencoba mencapai tingkat keberhasilan mingguan sebesar 85%, dihitung sebagai jumlah angka 9 dan 10 dibagi dengan jumlah total survei.

Kami akan berhenti sejenak di sini untuk membicarakan salah satu sumber variasi, bagaimana survei mengukur hasil. Menilai apa pun pada skala 1−10 sangat bermasalah. Angka 10 pada satu orang (“Mereka tidak memiliki apa yang saya cari, namun seorang karyawan membantu saya menemukan penggantinya!”) adalah angka 5 bagi orang lain (“Mereka tidak memiliki apa yang saya cari! Seorang karyawan harus membantu saya mencari penggantinya.”). Kami akan mengabaikan potensi sumber variasi lainnya seperti kehadiran karyawan yang kasar, toko yang penuh sesak, kemerosotan ekonomi yang membuat semua orang berada dalam bahaya, baik pelanggan berbelanja dengan anak-anak, dan sebagainya.

Kami tidak mengatakan bahwa survei itu sendiri harus dihapus. Sebaliknya, kami ingin menunjukkan bahwa desain (yaitu cara mengukur) data menimbulkan variasi yang sering diabaikan. Mengabaikan variasi berarti menganggap penyimpangan dari ekspektasi kita mencerminkan layanan berkualitas rendah. Namun bisnis akan berusaha mengejar angka target tinggi yang sulit dipahami (dalam kasus ini 9 dan 10) tanpa memahami bahwa pilihan mereka dalam mengukur data adalah penyebab utama variasi yang mendasarinya.

Beginilah cara hal ini bisa terjadi. Misalkan 50 orang meninggalkan ulasan setiap hari, setiap hari, selama 52 minggu. Ini menghasilkan 350 survei dalam seminggu dan 18.200 dalam setahun. Dengan partisipasi seperti itu, Anda sepertinya memiliki representasi persepsi pelanggan yang baik. Kemudian, pada akhir setiap minggu, hasilnya dihitung, perusahaan menjumlahkan angka 9 dan 10 dan membaginya dengan total mingguan, 350, dan melaporkan hasilnya dalam grafik, seperti yang ditunjukkan pada Gambar 1. Angka di atas angka 85% membuat Anda mendapat tepukan. Hasil di bawah 85%, dan Anda berkeringat.



Gambar 1. Hasil Survei Pelanggan Mingguan: Persentase Ulasan Positif. Garis horizontal pada 85% mewakili target.

Setiap hari Senin Anda mendapatkan laporan dan menelepon perusahaan tentang hasilnya. Bayangkan tekanan dari percakapan ini di minggu ke 5–9. Anda berada tepat di bawah ambang batas. Dan ketika Anda akhirnya menembus level di atas pada minggu ke 10, tidak diragukan lagi disebabkan oleh motivasi atasan Anda, minggu ke 11 datang dan memberi Anda titik terendah baru. Ini terus berlanjut.

Namun apa yang Anda lihat pada Gambar 1 adalah murni keacakan. Kami menghasilkan 18.200 nomor acak yang berupa 8, 9, atau 10—yang mewakili cara pelanggan yang berbeda memberikan suara pada layanan pelanggan yang positif, dan mengacaknya seperti setumpuk kartu. Setiap "minggu" kami mengambil 350 angka dan menghitung metriknya. Persentase rata-rata angka 9 dan 10 dalam kumpulan data adalah 85,3% (sangat mendekati nilai sebenarnya yaitu 85%), memenuhi standar perusahaan, namun setiap minggunya, hanya karena variasi acak, menyebar di sekitar ambang batas tersebut.

Tidak berpikir secara statistik menyebabkan semua orang, Anda, atasan Anda, dan kantor perusahaan, mengejar perbaikan dalam layanan untuk menaikkan angka secara sembarangan. Kami menyebut jenis aktivitas ini sebagai ilusi kuantifikasi. Ini adalah upaya untuk mendorong metrik tanpa dasar statistik yang jelas mengenai maksudnya. Apakah Anda melihat ilusi kuantifikasi di tempat kerja Anda?

Studi Kasus: Angka Kanker Ginjal

Tingkat kanker ginjal tertinggi di Amerika Serikat, diukur dengan jumlah kasus per 100.000 penduduk, terjadi di wilayah pedesaan yang tersebar di wilayah barat tengah, selatan, dan barat negara tersebut. Berhentilah sejenak untuk memikirkan mengapa hal ini bisa terjadi.

Mungkin Anda menduga, karena berada di wilayah pedesaan di pedalaman, penduduknya kurang mempunyai akses terhadap layanan kesehatan yang memadai. Atau mungkin akibat dari hidup tidak sehat yang disebabkan oleh pola makan tinggi lemak yang banyak mengandung daging dan banyak garam, atau terlalu banyak bir dan minuman beralkohol. Sangat mudah dan wajar untuk mulai membangun narasi berdasarkan fakta. Anda sudah dapat membayangkan para peneliti mulai merancang langkah-langkah remediasi untuk mengatasi masalah ini.

Namun ada fakta lain: tingkat kanker ginjal terendah di Amerika Serikat juga terjadi di wilayah pedesaan yang tersebar di wilayah barat tengah,

selatan, dan barat, seringkali bertetangga dengan wilayah dengan tingkat kanker ginjal tertinggi.

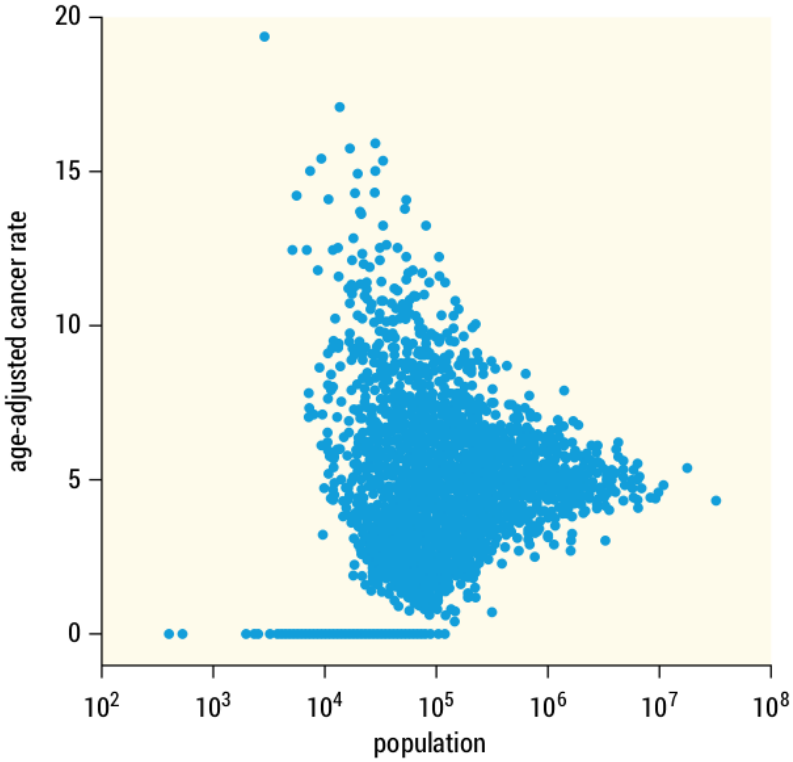
Bagaimana keduanya bisa benar? Bagaimana dua daerah dengan demografi serupa bisa mendapatkan hasil yang sangat berbeda? Setiap alasan yang mungkin Anda pikirkan untuk menjelaskan mengapa daerah pedesaan memiliki tingkat kanker ginjal yang tinggi pasti berlaku (sampai tingkat tertentu) di daerah tetangganya. Jadi, ada hal lain yang harus dilakukan.

Mari kita ambil dua kabupaten yang bertetangga di pedesaan Midwest, Kabupaten A dan Kabupaten B, dan asumsikan masing-masing kabupaten tersebut hanya memiliki 1000 penduduk. Jika Kabupaten A tidak memiliki kasus, angkanya akan menjadi 0, yang jelas merupakan kategori angka terendah. Namun jika Kabupaten B mempunyai satu kasus kanker ginjal, angkanya akan menjadi 100 kasus per 100.000 penduduk, sehingga menjadikannya angka tertinggi di negara tersebut. Rendahnya populasi di suatu daerahlah yang menyebabkan tingginya variasi, yang sekaligus menghasilkan angka tertinggi dan terendah. Sebaliknya, satu kasus tambahan di New York County (Manhattan, New York City), dengan populasi lebih dari 1,5 juta penduduk, hampir tidak memberikan dampak positif. Peningkatan dari 75 kasus menjadi 76 kasus akan mengubah jumlah kasus per 100.000 dari 5 menjadi 5,07.

Faktanya, variasi ini nyata dan diukur dalam artikel American Scientist yang berjudul, “The Most Dangerous Equation.” Gambar 2 merangkum hasil untuk negara-negara di AS. Kabupaten yang berpenduduk jarang, di sisi kiri plot, menunjukkan variasi tingkat kanker yang jauh lebih tinggi, dari 0 hingga 20, yang merupakan angka tertinggi di negara tersebut. Seiring bertambahnya populasi, bergerak dari kiri ke kanan dalam plot, variasinya mulai berkurang, sehingga menghasilkan bentuk segitiga. Variasi di sisi kanan gambar jauh lebih sedikit, yang menunjukkan bahwa daerah berpenduduk padat lebih kuat dalam menangani kasus tambahan dan menstabilkan sekitar 5 kasus per 100.000 penduduk.

Artikel tersebut membagikan contoh lain di mana angka kecil menyebabkan variasi yang tinggi. Misalnya, apakah Anda akan terkejut mengetahui bahwa

sekolah kecil mempunyai nilai ujian terbaik dan terburuk? Satu atau dua siswa yang tidak lulus ujian dapat menyebabkan perubahan besar dalam persentase keseluruhan. Jumlah kecil dapat menyebabkan hasil yang ekstrim.



Gambar 2. Cetak ulang figur Ilmuwan Amerika

4.1.3. Probabilitas dan Statistik

Sebelumnya, kami menjelaskan variasi dan membicarakan bagaimana variasi menjadi sumber ketidakpastian bagi banyak bisnis. Faktanya, ketidakpastian dapat dikelola dan di sinilah probabilitas dan statistik berperan.

Kita sering menggunakan istilah probabilitas dan statistik secara bergantian, jika tidak bersamaan, ketika mendeskripsikan hasilnya secara matematis. Namun di sini kita bisa membahas lebih dalam untuk benar-benar memahami perbedaannya.

Bayangkan sekantong besar kelereng. Di dalam, Anda tidak tahu apa warnanya. Anda tidak tahu bentuk atau ukurannya. Anda bahkan tidak tahu berapa banyak kelereng yang ada di dalam kantong, namun Anda merogoh kantong dan mengambil segenggam kelereng.

Mari kita berhenti sejenak. Anda memiliki sekantong kelereng yang belum Anda intip dan segenggam kaca menggelinding di antara jari-jari Anda yang belum Anda lihat. Anda benar-benar tidak memiliki informasi tentang apa yang ada di tangan atau di tas Anda.

Sekarang inilah perbedaannya. Probabilitas, Anda mengetahui dengan tepat apa yang ada di dalam tas, dan menggunakan informasi tersebut untuk menebak apa yang ada di tangan Anda. Dalam statistik, Anda membuka tangan dan menggunakan informasi tersebut untuk memberi tahu kami apa yang ada di dalam tas.

Penelusuran probabilitas; penelusuran statistik. Masuk akal?

Mari kita lihat dua contoh kehidupan nyata:

- **Kasino Las Vegas dibangun berdasarkan probabilitas.**
  Setiap kali Anda memainkan permainan kasino, Anda menarik dari kantong kelereng mereka, yang terdiri dari kemenangan dan kekalahan. Ada cukup kelereng pemenang di dalam tas kasino untuk membuat Anda tetap tertarik untuk bermain. Kasino memahami variasi, bahkan mereka telah mengkomersialkannya melalui pembayaran dan kerugian yang dioptimalkan untuk membuat Anda tetap tertarik dan gembira. Namun, dalam jangka panjang, kasino tahu bahwa mereka akan menghasilkan uang karena mereka menciptakan tas tempat semua kelereng ditarik, dan mereka tahu persis apa yang ada di dalamnya. Dengan setiap

taruhan yang dibuat, chip diletakkan di meja, dan tuas ditarik di sisi mesin slot, kasino mengetahui kemungkinan kesuksesan Anda. Jika Anda memikirkan berapa banyak data yang dimiliki kasino, Anda dapat melihat bahwa keduanya hidup di dunia yang bervariasi tetapi juga memiliki gambaran yang jelas tentang kemungkinan hasil.

- **Jajak pendapat politik didasarkan pada statistik.**
  Di kasino, sekantong kelereng dirancang dengan cermat dan selalu diambil sampelnya. Namun, dalam sebuah pemilu, para politisi tidak mengetahui apa yang sebenarnya ada di dalam tas tersebut sampai pada hari pemilu, ketika semua kelereng (yaitu, suara) terungkap. Politisi hanya mempunyai satu kesempatan untuk mengetahui apa yang ada di dalam tas tersebut, dan apakah tas tersebut berisi kelereng kemenangan yang cukup untuk mereka. Sebelum pemilu, politisi dan partai politik hanya mempunyai akses terhadap sejumlah kecil kelereng acak (disebut survei), dan mereka membayar banyak uang untuk akses tersebut. Dengan menggunakan sampel tersebut, mereka menyimpulkan pola di dalam tas dan menyesuaikan kampanyenya. Karena informasi mereka tidak lengkap (dan karena mereka sering menimbulkan bias dan kesalahan), mereka tidak selalu memberikan informasi yang benar. Namun ketika mereka berhasil, itulah perbedaan antara memenangkan pemilu dan tidak. Mari kita lihat sekilas beberapa konsep penting dalam probabilitas dan statistik di bagian berikut.

Mari kita lihat sekilas beberapa konsep penting dalam probabilitas dan statistik di bagian berikut.

Probabilitas vs. Intuisi

Sebelumnya, kami telah mengatakan bahwa variasi acak tidak dapat dikontrol. Tapi hal itu bisa diukur, dan probabilitas memberi kita alat untuk melakukannya.

Terkadang probabilitas sangat masuk akal bagi kita. Jika Anda telah melempar dadu atau memutar dreidel, Anda menyadari bahwa Anda

mempunyai peluang yang diketahui untuk mendarat pada angka tertentu (1 dari 6) atau huruf (1 dari 4). Permainan untung-untungan sederhana sangat masuk akal bagi kami. Probabilitas sederhana terasa intuitif. Memang, hal-hal tersebut sangat masuk akal bagi kita, sehingga sering kali mengaburkan kompleksitas yang mendasarinya. Iklan misalnya, memanfaatkan probabilitas sederhana dengan mereduksinya menjadi sesuatu yang terasa seperti kita memahaminya secara intuitif.

Anda mungkin pernah melihat iklan ini sebelumnya: “4 dari 5 dokter gigi setuju” dengan klaim iklan, X (X bisa apa saja yang Anda inginkan, seperti permen karet mengurangi gigi berlubang, atau soda kue memutihkan gigi). Sekarang, misalkan ada lima dokter gigi yang duduk di depan Anda. Mengetahui bahwa 80% dari semua dokter gigi setuju dengan X, seberapa besar kemungkinan empat dari lima dokter gigi di depan Anda setuju? 100%? 90%? atau 80%? Jawaban sebenarnya adalah 41%.

Secara intuitif, ini tampaknya terlalu rendah, tetapi ini benar. Mari kita lihat alasannya. Tabel 1 menunjukkan salah satu cara sampel yang terdiri dari lima dokter gigi dapat menyetujui X.

Tabel 1. Probabilitas Dokter Gigi Menyetujui Klaim Iklan

	Dentists				
	1	2	3	4	5
Agreement?	Yes	Yes	Yes	Yes	No
Probability	0.8	0.8	0.8	0.8	0.2

Probabilitas kombinasi: 0.8*0.8*0.8*0.8*0.2 = 0.08192
Atau, untuk singkatnya: p = 0.84*0.2 = 0.08192

Namun terdapat lima kombinasi persetujuan yang berbeda, dimana masing-masing dokter gigi dapat memilih “Tidak” seperti yang ditunjukkan pada Tabel 2. Jadi, kalikan p dengan lima: 0.08192 * 5 = 0.4096, atau disingkat 41%.

Tabel 2. Kemungkinan Kombinasi 4 dari 5 Dokter Gigi Setuju

	Dentists: Do you agree?				
Combination	1	2	3	4	5
1	Yes	Yes	Yes	Yes	**No**
2	Yes	Yes	Yes	**No**	Yes
3	Yes	Yes	**No**	Yes	Yes
4	Yes	**No**	Yes	Yes	Yes
5	**No**	Yes	Yes	Yes	Yes

Rata-rata empat dari lima dokter gigi mungkin setuju, tetapi itu tidak menjamin bahwa dalam setiap sampel dari lima dokter gigi, empat dokter gigi akan setuju untuk mengklaim X.

Kami membagikan latihan ini untuk menyoroti, sekali lagi, bagaimana variasi orang di bawah perkiraan, terutama ketika berhadapan dengan jumlah yang kecil. Apa yang diharapkan orang, berdasarkan intuisi, jarang sesuai dengan kenyataan saat kita menghitung probabilitas. Dan meremehkan variasi menyebabkan orang terlalu memperkirakan kepercayaan mereka terhadap data kecil. Hal ini disebut dengan “law of small numbers”.

Berpikir secara statistik, seperti yang seharusnya dilakukan oleh Kepala Data, berarti memperhatikan intuisi kita, menyadari bahwa hal itu dapat mempermainkan kita. Kita akan mengeksplorasi beberapa contoh dan kesalahpahaman ini di bahasan-bahasan selanjutnya.

Penemuan dengan Statistik

Statistik sering dipecah menjadi statistik deskriptif dan statistik inferensial. Anda mungkin familiar dengan statistik deskriptif meskipun Anda tidak menggunakan frasa tersebut. Statistik deskriptif adalah angka-angka yang merangkum data. Contohnya penjualan rata-rata pada kuartal terakhir, kenaikan dari tahun ke tahun, tingkat pengangguran, dll. Ukuran seperti mean, median, range, varians, dan deviasi standar merupakan statistik deskriptif dan memerlukan rumus khusus untuk menghitungnya. Statistik deskriptif adalah penyederhanaan data yang disengaja, sebuah cara untuk

menyingkat seluruh spreadsheet data penjualan perusahaan menjadi beberapa ukuran utama yang merangkum informasi utama.

Meskipun bermanfaat, kami jarang puas berhenti di sini. Kami ingin mengambil langkah ekstra dan memahami bagaimana kami dapat mengambil informasi dan membuat tebakan berprinsip untuk menyimpulkan isi umum dari keseluruhan tas. Ini adalah statistik inferensial.

Untuk saat ini, mari kita perhatikan sebuah contoh. Bayangkan bagaimana reaksi Anda saat melihat judul, “75% Orang Amerika Percaya UFO Itu Ada!” setelah mengetahui sampelnya diambil dari 20 wisatawan di Museum dan Pusat Penelitian UFO Internasional di Roswell, New Mexico. Menurut Anda, apakah Anda dapat secara akurat menyimpulkan persentase sebenarnya orang Amerika yang percaya pada UFO berdasarkan apa yang sekarang Anda ketahui tentang penelitian tersebut? Statistik ini tidak dapat dipercaya berdasarkan:

- Sampel yang bias.
  Orang-orang yang mengunjungi Roswell kemungkinan besar lebih percaya pada UFO dibandingkan masyarakat umum.
- Ukuran sampel kecil
  Anda telah mempelajari seberapa besar variasi yang ditimbulkan oleh ukuran sampel kecil. Menyimpulkan apa yang dipikirkan jutaan orang berdasarkan 20 orang tidaklah masuk akal.
- Asumsi yang mendasari
  Judulnya menyebutkan “orang Amerika” percaya pada UFO hanya karena tes tersebut dilakukan di Amerika. Namun museum ini, seperti yang mungkin Anda ingat, merupakan daya tarik internasional. Anda tidak tahu bahwa semua orang yang berpartisipasi dalam survei ini adalah orang Amerika.

Konsep seperti bias dan ukuran sampel adalah alat inferensi statistik yang membantu kita memahami apakah statistik yang kita lihat atau hitung tidak masuk akal. Dan itu adalah bagian penting dari perangkat Anda. Asumsi yang mendasarinya juga sama pentingnya untuk dipertimbangkan. Berpikir

seperti Kepala Data mengharuskan Anda untuk tidak menerima begitu saja asumsi yang disebutkan dalam kesimpulan. Jadi, saat Anda melihat data dalam pekerjaan Anda, jangan begitu saja menyetujui informasi yang Anda lihat atau bahkan intuisi yang Anda rasakan.

4.2. Eksplorasi Data Menggunakan Statistik

Proses iterasi, penemuan, dan pemeriksaan data dikenal sebagai *exploratory data analysis* (EDA). EDA dirumuskan oleh ahli statistik John Tukey pada tahun 1970-an sebagai cara untuk memahami data dengan ringkasan statistik dan visualisasi sebelum menerapkan metode yang lebih kompleks. Tukey melihat EDA sebagai pekerjaan detektif. Petunjuk tersembunyi dalam data, dan eksplorasi yang tepat akan mengungkap langkah selanjutnya yang harus diikuti. Memang, EDA adalah cara lain untuk "berdebat" dengan data Anda. Ini adalah bagian mendasar dari semua pekerjaan data yang menentukan dan mengubah arah proyek berdasarkan apa yang ditemukan.

4.2.1. EDA dan Anda

EDA bisa menjadi pemikiran yang tidak nyaman bagi sebagian orang: EDA mengungkap sifat subjektif (seni?) di balik semua pekerjaan data. Dua tim, jika diberi masalah dan data yang sama, mungkin mengambil jalur analisis yang berbeda, dan mungkin akan mendapatkan kesimpulan yang sama atau mungkin tidak. Ada terlalu banyak keputusan yang harus diambil sehingga dua tim (atau individu) mana pun dapat melakukan semuanya dengan cara yang sama. Setiap orang akan membawa latar belakang, ide, dan alatnya masing-masing untuk membuat rekomendasi tentang cara terbaik memecahkan masalah.

Oleh karena itu, dalam bahasan ini, kami menyajikan EDA sebagai proses berkelanjutan yang merupakan tanggung jawab setiap Kepala Data, baik Anda pekerja data langsung atau pemimpin bisnis di ruang rapat. Anda akan mempelajari pertanyaan untuk ditanyakan dan hal-hal yang harus diperhatikan saat menjelajahi data.

4.2.2. Mindset Eksplorasi

Lusinan alat dan bahasa pemrograman dapat membantu tim data dengan cepat dan murah mengeksplorasi data mereka dengan ringkasan statistik dan visualisasi. Namun EDA tidak boleh dianggap sebagai daftar alat atau daftar checklist. Ini lebih merupakan mentalitas yang terjalin dalam setiap fase pekerjaan data yang dapat Anda ikuti, bahkan tanpa latar belakang analitik.

Pertanyaan untuk Memandu Anda

Untuk membantu Anda menerapkan pola pikir eksploratif, kami akan memandu Anda melalui skenario singkat dengan latar belakang kumpulan data populer yang dikumpulkan untuk tujuan pendidikan: Data Perumahan Ames. Ini adalah sekilas proses EDA. Meskipun tidak ada satu jalan yang benar untuk diikuti, ada beberapa pertanyaan yang dapat Anda ajukan untuk membantu memandu tim mencapai kesimpulan yang bermakna:

- Apakah data dapat menjawab pertanyaan tersebut?
- Apakah Anda menemukan adanya hubungan/relasi?
- Apakah Anda menemukan peluang baru dalam data?

Mari kita siapkan skenarionya, lalu bahas masing-masing dari tiga pertanyaan tersebut, diskusikan mengapa pertanyaan tersebut layak untuk ditanyakan, dan bagikan tantangan yang mungkin Anda hadapi.

Pengaturan

Anda bekerja di perusahaan start-up real estat dan perlu mengarahkan trafik ke situs Anda. Namun sulit untuk menarik pengunjung dari raksasa teknologi real estat seperti Zillow.com yang berbasis di AS. Alat estimasi harga rumah yang terkenal, Zestimate®, membawa orang (dan keuntungan) ke merek Zillow. Untuk bersaing, perusahaan Anda memerlukan alat prediksinya sendiri. Jadi, Anda ditugaskan untuk membangun model yang mengambil informasi rumah sebagai masukan dan menghasilkan perkiraan harga jual sebagai keluaran.

Bos mengirimi Anda kumpulan data untuk memulai. Ini memiliki 80 kolom yang menggambarkan beberapa aspek dari ratusan rumah tempat tinggal yang dijual di Ames, Iowa dari tahun 2006 hingga 2011.

Menerima data sebanyak ini bisa sangat melelahkan. Namun, menggunakan pertanyaan-pertanyaan yang diuraikan sebelumnya dapat membantu Anda mempersempit cara mulai bekerja dengan data. Mari kita bahas.

4.2.3. Apakah Data Dapat Menjawab Pertanyaan?

Meskipun mungkin tergoda untuk memasukkan data Anda ke dalam tren algoritme saat ini (misalnya, deep learning), pertama-tama Anda perlu bertanya: “Dapatkah data menjawab pertanyaan?” Dan jawabannya seringkali ditemukan hanya dengan melihat datanya.

Tetapkan Harapan dan Gunakan Akal Sehat

Anda harus memiliki gambaran yang cukup baik tentang informasi apa yang diperlukan untuk membuat perkiraan harga jual rumah: ukuran, jumlah kamar tidur, jumlah kamar mandi, tahun dibangun, dll. Ini adalah fitur populer yang dicari pembeli rumah di situs Anda. Tanpa fitur tersebut, memperkirakan harga jual tidak akan masuk akal.

Anda dapat melihat nama kolom dan tipe data saat Anda membuka file. Fitur-fitur yang masuk akal yang Anda harapkan ada, serta data ordinal yang berguna (Kualitas rumah secara keseluruhan, 1–10, 10 berarti “Sangat Sangat Baik”), data nominal (Lingkungan), dan sejumlah fitur lainnya.

Selanjutnya, Anda mungkin akan memeriksa nilai yang diambil oleh variabel. Apakah mereka mencakup skenario yang ingin Anda analisis? Misalnya, jika Anda menemukan variabel “Jenis Bangunan: Tipe Tempat Tinggal” hanya mencakup rumah keluarga tunggal tetapi tidak mencakup apartemen, dupleks, atau kondominium, maka model Anda akan memiliki cakupan terbatas dibandingkan dengan model Zillow. Zestimate® dapat memprediksi harga jual sebuah kondominium, namun jika Anda tidak memiliki data historis kondominium, perusahaan Anda tidak dapat memprediksi harga jualnya dengan andal.

Apakah Nilai-Nilainya Masuk Akal Secara Intuitif?

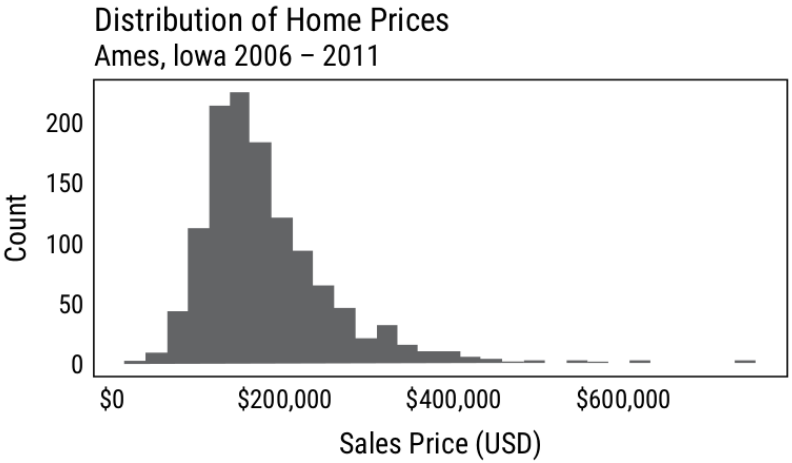
Perangkat lunak akan menghasilkan banyak ringkasan statistik untuk Anda. Tugas Anda adalah memasukkan data ke dalam konteks. Periksa apakah ringkasan statistik sesuai dengan pemahaman intuitif Anda tentang

masalahnya. Visualisasi juga merupakan komponen kunci EDA, gunakan visualisasi untuk menemukan anomali dan keanehan lainnya dalam data.

Penyegaran Visualisasi Data

Mari kita lihat beberapa contoh singkat EDA dengan histogram, box-plot, diagram batang, dan scatter-plot.

Anda dapat mempelajari bagaimana data numerik kontinu dibentuk, atau didistribusikan, dengan melihat histogram. Perhatikan histogram harga jual yang ditunjukkan pada Gambar 3. Misalnya, ada sekitar 125 rumah dengan kisaran harga $200.000, dan ekor panjang ke kanan menunjukkan rumah paling mahal. Hasil tersebut menarik harga jual rata-rata ($181.000) melewati harga median ($163.000). Sejumlah rumah mahal membuat nilai rata-ratanya lebih besar daripada nilai mediannya.

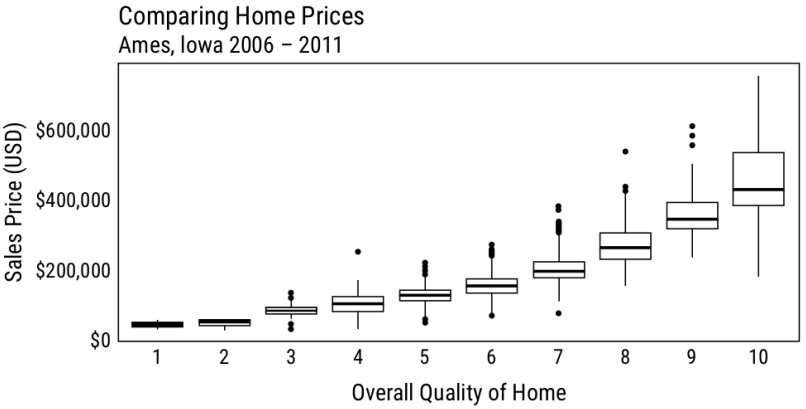


Gambar 3. Histogram yang menunjukkan bentuk harga jual

Histogram berguna untuk mengenali anomali. Jika Anda melihat nilai negatif (dibayar untuk membeli rumah?) atau kotak dengan jumlah yang sangat besar di sisi jauh Gambar 3, yang sering terjadi ketika data dibatasi

(misalnya, nilai apa pun di atas $500.000 dimasukkan sebagai $500.000), Anda mungkin ingin untuk mengajukan beberapa pertanyaan.

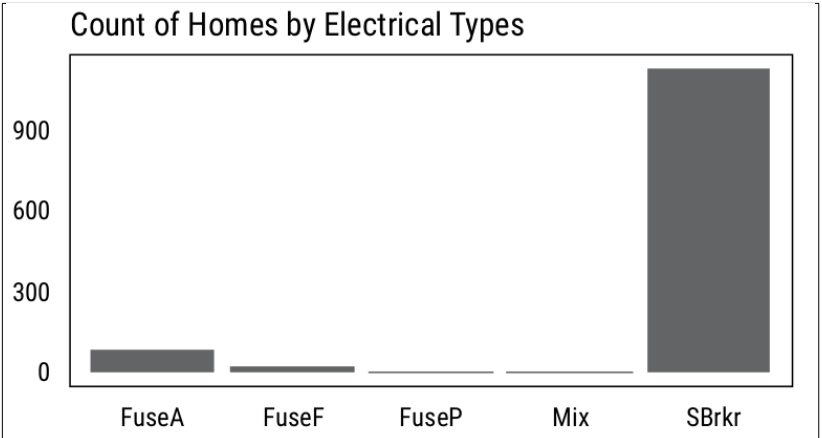
Box-plot dapat digunakan untuk membandingkan data di beberapa kelompok. Gambar 4 menunjukkan plot kotak untuk setiap peringkat kualitas rumah: 1 buruk dan 10 sangat baik.



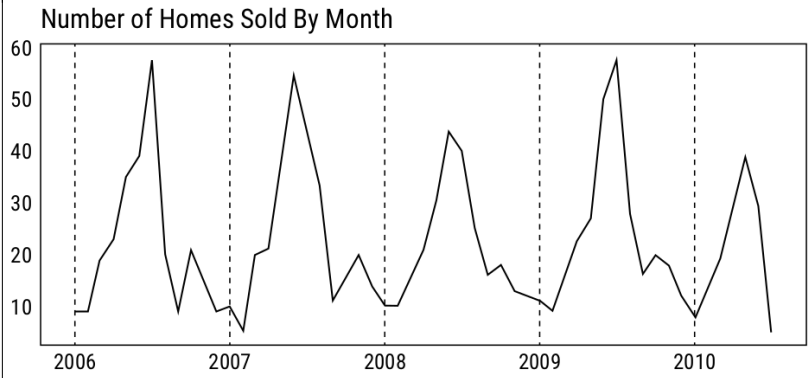
Gambar 4. Menggunakan box-plot untuk membandingkan harga jual pada peringkat kualitas yang berbeda

Di sini, hubungan antara kualitas keseluruhan dan harga rumah terasa intuitif. Rumah dengan kualitas lebih tinggi biasanya memiliki harga jual yang lebih tinggi. Kita dapat melihat rumah seharga $200.000 dengan skor kualitas keseluruhan 10 (titik terbawah), tetapi tampaknya masuk akal untuk berasumsi bahwa rumah tersebut dijual dengan harga kurang dari 10 rumah sempurna lainnya karena faktor lain. Ini adalah jenis informasi yang harus diperiksa oleh pekerja data.

Diagram batang, seperti yang ditunjukkan pada Gambar 5, menunjukkan jumlah data kategorikal.



Gambar 5. Diagram batang yang menunjukkan penghitungan berdasarkan jenis instalasi listrik

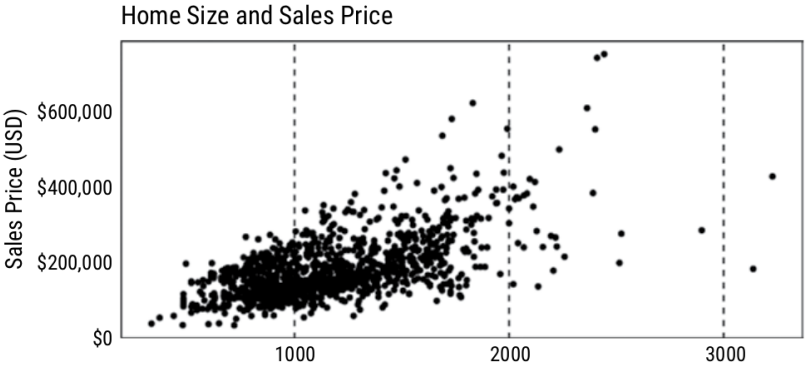


Gambar 6. Diagram garis yang menunjukkan jumlah rumah terjual pada bulan yang berbeda

Tidak semua visual menarik pada pandangan pertama. Namun, ada baiknya untuk melihat visualisasi tersebut hanya untuk memperkuat (atau mungkin menantang) pertanyaan sebelumnya, apakah datanya masuk akal secara intuitif? Gambar 5 menunjukkan bahwa hampir semua rumah memiliki nilai yang sama untuk fitur ini. Namun, untuk tugas Anda, informasi ini berguna. Karena sebagian besar rumah memiliki nilai yang sama untuk

variabel ini, kemungkinan besar variabel ini tidak akan memberikan kontribusi terhadap perbedaan yang berarti pada harga jual rumah.

Gambar 6 menunjukkan diagram garis dengan data jumlah rumah terjual per bulan. Anda dapat dengan mudah memvisualisasikan fenomena di mana penjualan rumah melonjak di musim panas dan menurun di musim dingin, sebuah contoh dari musim. Bagan garis berguna untuk mengenali tren tersebut.



Gambar 7. Scatter plot yang menunjukkan ukuran luas dan harga jual

Selanjutnya, kita dapat memeriksa scatter-plot yang menunjukkan rumah-rumah yang diplot berdasarkan ukurannya (ukuran luas lantai pertama) dan harga jual (lihat Gambar 7).

Gambar 7 menunjukkan pola intuitif. Rumah yang lebih besar umumnya dijual dengan harga lebih banyak. Tentu saja aturan tersebut tidak selalu benar. Terkadang rumah kecil harganya lebih mahal daripada rumah besar. Selalu ada variasi, tetapi tren keseluruhannya tetap ada. Dan karena kami mencoba memprediksi harga jual sebagai output, luas persegi sepertinya merupakan informasi yang bagus untuk dimiliki.

Perhatikan Outlier dan Missing Value
Setiap kumpulan data akan memiliki anomali, outlier, dan nilai yang hilang. Bagaimana Anda menangani hal-hal ini.

Misalnya, box-plot pada Gambar 4 menggunakan aturan praktis untuk menandai beberapa titik data sebagai potensi outlier. Namun hanya karena grafik data mengklasifikasikan titik-titik tertentu sebagai “outlier”, jangan matikan pemikiran kritis dan secara otomatis menghapus titik-titik tersebut dengan asumsi titik-titik tersebut tidak berguna. Anda tidak akan pernah melihat Zillow menghapus informasi berguna dari kumpulan datanya hanya karena visualisasi menggambarkannya sebagai outlier. Gunakan konteks data, rumah yang harganya jauh lebih mahal dibandingkan sebagian besar rumah lainnya merupakan fitur yang diketahui dan umum dari data real estat. Anda setidaknya harus memiliki alasan bisnis yang baik untuk menghilangkan outlier.

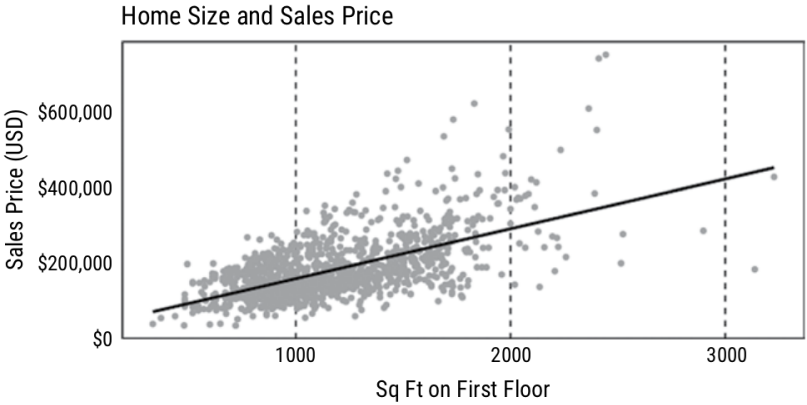
Dan bagaimana dengan missing value? Apakah nilai yang hilang pada “Ukuran Basement” berarti rumah tersebut memiliki basement dan luasnya tidak diketahui? Atau berarti tidak ada basement dan nilainya harus 0?

Pekerja data membuat ratusan keputusan kecil ini selama proyek berlangsung. Dampak kumulatifnya bisa sangat besar. Jika dibiarkan sendiri dan tanpa bimbingan ahli di bidangnya, pekerja data dapat terus mengolah data, menghilangkan kasus-kasus sulit dan tidak jelas, hingga data tersebut terlalu terlepas dari kenyataan yang ingin ditangkap agar dapat berguna. Inilah sebabnya mengapa penting bagi semua orang, termasuk manajer, untuk benar-benar memahami apa yang dilakukan tim data mereka.

4.2.4. Apakah Anda Menemukan Hubungan?
Untungnya bagi kami, data perumahan pertama dengan ringkasan statistik dan visualisasi tampak menggembirakan dan menurut Anda data tersebut memang dapat digunakan untuk membangun model prediktif harga jual, jadi Anda melanjutkan ke pertanyaan berikutnya: “Apakah Anda menemukan ada hubungan?"

Memvisualisasikan data telah memberi Anda langkah awal: kualitas keseluruhan yang lebih tinggi dan ukuran luas yang lebih besar tidak mengherankan terkait dengan harga jual yang lebih tinggi. Ini adalah umpan balik yang Anda inginkan dari data. Hubungannya masuk akal dan variabel yang Anda buat akan membantu Anda membangun model untuk memprediksi harga jual. Variabel lain apa yang mempunyai hubungan yang sama dengan harga jual?

Pada titik ini, ringkasan statistik dapat membantu mengarahkan Anda menuju pola dan hubungan yang menarik dalam data karena menghasilkan setiap plot sebar mungkin tidak praktis. Sebaliknya, hubungan yang ditemukan dalam scatter-plot dapat direduksi menjadi ringkasan korelasi statistik, yang menunjukkan (tetapi bukan bukti) hubungan antara dua variabel numerik.



Gambar 8. Luas persegi dan harga jual memiliki korelasi sebesar 0,62, yang mengukur ketatnya titik data di sekitar garis tren linier padat.

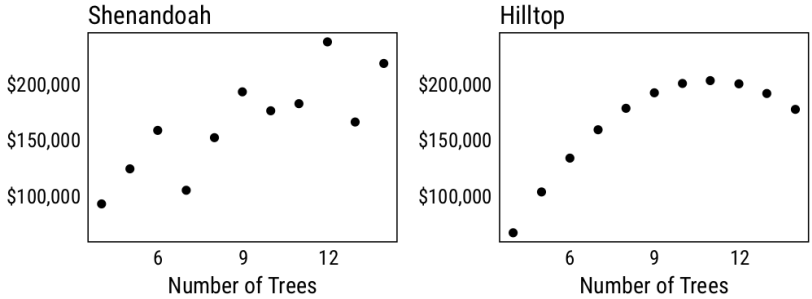
Memahami Korelasi

Korelasi adalah ukuran bagaimana dua variabel berhubungan. Jenis korelasi yang paling umum digunakan dalam bisnis adalah koefisien korelasi Pearson, statistik antara −1 dan 1 yang mengukur hubungan linier (pikirkan garis lurus sederhana) antara pasangan angka yang ditampilkan pada plot sebar. Korelasinya bisa positif, artinya peningkatan pada satu variabel

dikaitkan dengan peningkatan variabel lainnya: rumah yang lebih besar dijual dengan harga yang lebih tinggi. Atau, korelasinya bisa negatif: mobil yang lebih berat menghasilkan jarak tempuh yang lebih buruk. Sebagai referensi visual, korelasi ukuran rumah dan harga jual, ditunjukkan pada Gambar 8, adalah 0,62. Semakin “ketat” titik-titik di sekitar tren linier, semakin tinggi korelasinya.

Korelasi dapat membantu dalam dua cara. Pertama, menemukan variabel yang berkorelasi dengan harga jual akan membantu memprediksinya. Kedua, korelasi dapat membantu mengurangi redundansi dalam data Anda karena dua variabel yang berkorelasi tinggi berisi informasi yang kurang lebih sama. Bayangkan dua kolom dalam data Anda: luas rumah dalam kaki persegi dan luas dalam meter persegi. Keduanya berkorelasi sempurna; hanya satu yang diperlukan dalam analisis.

Meskipun sebagian besar dari kita memiliki pemahaman dasar tentang korelasi dan sering melaporkan metriknya, hal ini dapat menipu. Mari kita ulas bagaimana hal tersebut bisa terjadi.



Gambar 9. Dua dataset dengan korelasi 0,8

Awas: Salah Menafsirkan Korelasi

Orang sering lupa bahwa korelasi adalah ukuran tren linier, dan tidak semua tren bersifat linier. Misalkan, misalnya, Anda menganalisis dua lingkungan dalam kumpulan data perumahan, masing-masing memiliki 11 rumah. Berdasarkan beberapa statistik, terungkap bahwa jumlah pohon di sebuah properti sangat berkorelasi dengan harga jual di lingkungan tersebut.

Korelasinya kuat sebesar 0,8: properti yang memiliki lebih banyak pohon cenderung terjual dengan harga lebih tinggi.

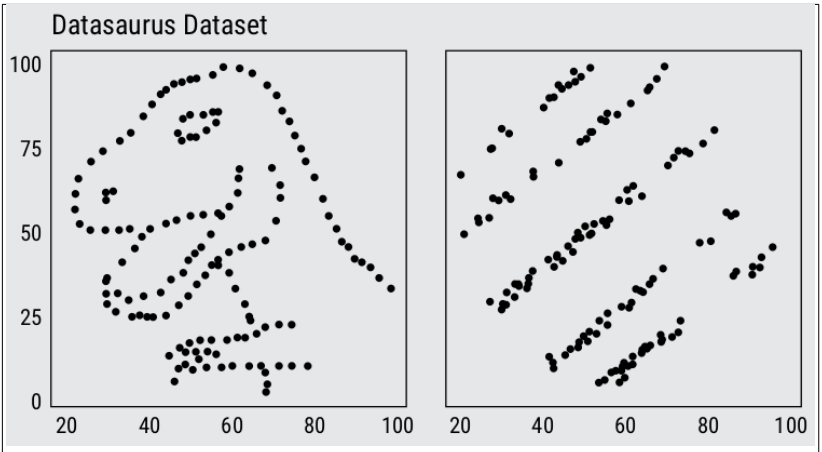
Namun pemeriksaan visual terhadap data mengungkap sesuatu yang tidak terduga. Pada Gambar 9, data untuk lingkungan di sebelah kiri menunjukkan apa yang biasanya kita harapkan dengan korelasi tinggi: tren linier dengan titik-titik data tersebar. Namun plot di sebelah kanan menunjukkan jumlah pohon dikaitkan dengan kenaikan harga jual hanya sampai satu titik (11 pohon). Setelah itu, trennya menurun. Di komunitas Hilltop, beberapa properti mungkin memiliki terlalu banyak pohon yang memenuhi halamannya.

Pengungkapan penuh: data yang ditunjukkan pada Gambar 5.7 tidak berasal dari kumpulan data Ames yang telah kami jelajahi, melainkan dari kumpulan data populer yang disebut Kuartet Anscombe,8 empat kumpulan data dengan statistik ringkasan yang identik namun visualisasinya jelas berbeda. (Kami hanya menampilkan dua dan menyesuaikan data untuk mencerminkan tema real estat.)

Pelajarannya: Gunakan visualisasi untuk memverifikasi korelasi penting dalam data Anda karena tren linier yang dapat diidentifikasi oleh korelasi mungkin tidak menjelaskan keseluruhan cerita.

Tidak Berkorelasi tapi Tetap Menarik

Gambar 10 menunjukkan dua plot, masing-masing memiliki koefisien korelasi yang identik dan mendekati nol. Jangan biarkan hal itu menipu Anda dengan berpikir tidak ada hal lucu yang terjadi. Dan meskipun Anda tidak akan menemukan banyak datasaurus seperti di plot kiri, Anda mungkin menemukan skenario di plot kanan: sebenarnya ada lima kelompok data yang berkorelasi linier, tetapi jika dilihat sebagai satu kelompok, data tersebut tidak berkorelasi linier sama sekali . Hal ini dikenal sebagai Paradoks Simpson.



Gambar 10. Datasaurus: Data dapat diunduh dan dijelajahi secara gratis. Seperti Kuartet Anscombe, kedua kumpulan data yang ditampilkan di sini memiliki ringkasan statistik yang identik.

Hati-hati: Korelasi Tidak Menyiratkan Sebab-Akibat

Kemungkinannya adalah, Anda pernah mendengar ungkapan “Korelasi tidak menyiratkan sebab-akibat” sebelumnya. Namun perlu ditegaskan kembali karena seringnya hal ini diabaikan dan bahkan disalahpahami.

Ketika dua variabel berkorelasi, bahkan berkorelasi kuat, bukan berarti variabel yang satu menyebabkan variabel yang lain. Namun orang-orang terlalu sering terjebak dalam perangkap ini, karena berusaha membangun narasi setiap kali dua variabel bergerak bersamaan. Ada beberapa contoh konyol yang digunakan para ahli statistik untuk menunjukkan bahwa korelasi tidak membuktikan sebab-akibat: Penjualan es krim berkorelasi dengan serangan hiu (keduanya melonjak pada bulan-bulan musim panas). Ukuran sepatu berkorelasi dengan kemampuan membaca (keduanya meningkat seiring waktu). Namun pernyataan bahwa mengurangi penjualan es krim akan mengurangi serangan hiu, atau bahwa membeli sepatu yang lebih besar membantu Anda membaca, jelas merupakan sebuah lelucon. Ada faktor lain yang berperan—suhu luar pada contoh es krim, usia pada contoh ukuran sepatu—yang jelas berperan dalam hubungan palsu tersebut.

Namun ketika korelasi tidak dibangun berdasarkan lelucon dan faktor penyebab sebenarnya tidak jelas, maka ungkapan “korelasi tidak berarti hubungan sebab-akibat” sering kali terlupakan.

Misalnya, dalam data real estate, Anda menemukan metrik kinerja sekolah berkorelasi dengan nilai rumah. Apakah ini berarti sekolah yang lebih baik akan meningkatkan nilai sebuah rumah? Sekolah yang bagus nampaknya membuat suatu lingkungan menjadi lebih diminati. Atau, apakah hubungan sebab-akibatnya mengarah ke arah lain: harga rumah yang lebih tinggi menyebabkan peningkatan prestasi sekolah? Mungkin peningkatan pendapatan pajak memberikan lebih banyak sumber daya kepada sekolah. Atau apakah kausalitas berjalan dua arah, sehingga menciptakan putaran umpan balik? Seringkali, kita tidak mengetahuinya. Jelas ada banyak faktor lain yang berperan, dan jarang sekali ada semua jawaban yang Anda perlukan dalam kumpulan data Anda.

Lebih aman untuk berasumsi “tidak ada kausalitas” antara dua variabel yang berkorelasi kecuali seseorang telah melakukan eksperimen yang membuktikan sebaliknya. Tapi jangan menganggap ini ekstrem. Kedua penulis telah melihat kasus-kasus di lingkungan bisnis, universitas, dan media di mana terdapat asumsi sebab-akibat, padahal sebenarnya tidak demikian. Namun ada juga kasus di mana sebuah asosiasi penting langsung diabaikan karena dianggap sebagai kesalahan sebab-akibat.

Merokok dan Kanker Paru-paru

Ronald A. Fisher, yang dianggap sebagai salah satu ahli statistik terkemuka abad ke-20, yang bahkan mengembangkan dan berkontribusi pada teknik yang dijelaskan dalam buku ini, sangat skeptis terhadap penelitian yang menghubungkan penggunaan tembakau dengan kanker.

Fisher paling prihatin dengan variabel perancu. Bagaimana jika, misalnya, beberapa orang secara genetik cenderung terkena kanker paru-paru dan ingin merokok untuk meringankan gejalanya? Menurut Fisher, penelitian awal mengenai risiko penggunaan tembakau telah melakukan “kesalahan. . .yang kuno, dalam perdebatan dari korelasi ke sebab-akibat.”

Namun kini kita tahu bahwa hubungan antara keduanya tidak dapat disangkal. Meskipun kita harus berhati-hati dalam melihat hubungan sebab akibat yang tidak ada, kita juga harus berhati-hati untuk tidak mengabaikan hubungan yang belum terbukti sebagai hubungan sebab akibat.

4.2.5. Apakah Anda Menemukan Peluang Baru Dalam Data?

EDA bukan sekadar proses untuk lebih memahami data dan menetapkan jalan ke depan untuk memecahkan masalah. Ini juga merupakan peluang untuk menemukan peluang lain dalam data; masalah yang mungkin berharga bagi organisasi Anda. Seorang data scientist mungkin menemukan sesuatu yang menarik atau aneh dalam kumpulan data dan kemudian merumuskan masalahnya.