**Memproses Data dari Kotor ke Bersih**

**Mengapa integritas data itu penting**

Analisis yang kuat tergantung pada integritas data. Integritas data adalah akurasi, kelengkapan, konsistensi, dan kepercayaan atas data di sepanjang siklusnya. Integritas data dapat terganggu dengan berbagai cara yang berbeda. Ada kemungkinan data bisa terganggu sewaktu data itu direplikasi, ditransfer, atau dimanipulasi dengan cara apa pun.

* **Replikasi data** yang membahayakan integritas data: Anda meminta rekan internasional Anda untuk memverifikasi tanggal dan tetap berpegang pada satu format. Seorang analis menyalin dataset berukuran besar untuk memeriksa tanggal-tanggalnya. Tetapi karena ada masalah memori, hanya sebagian dataset yang berhasil disalin. Analis itu akan memverifikasi dan melakukan standardisasi data yang tidak lengkap. Dataset yang tidak lengkap itu akan mendapat sertifikat kesesuaian meskipun dataset yang lebih lengkap masih berisi tanggal yang tidak diverifikasi. Dua versi dataset ini dapat memberikan hasil yang tidak konsisten. Audit akhir terhadap hasil analisis perlu dilakukan untuk mengungkapkan apa yang terjadi dan membetulkan semua tanggal yang tidak sesuai.
* **Transfer data** yang membahayakan integritas data: Analis lain memeriksa tanggal di sebuah spreadsheet dan memilih untuk mengimpor data yang divalidasi dan distandarisasi kembali ke database. Tapi anggaplah field tanggal di spreadsheet itu telah keliru diklasifikasikan sebagai sebuah teks selama proses impor (transfer) data. Sekarang banyak data tanggal dalam database itu disimpan sebagai string teks. Pada situasi ini, data itu perlu dibersihkan untuk memulihkan integritasnya.
* **Manipulasi data** yang membahayakan integritas data: Saat memeriksa tanggal, analis lain menemukan apa yang tampaknya merupakan record duplikat dalam database itu dan menghapusnya. Namun ternyata analis itu menghapus record yang unik untuk anak perusahaan dan bukan record duplikat untuk perusahaannya. Dataset Anda sekarang kehilangan data dan data itu harus dipulihkan supaya lengkap.



**Tujuan dan data yang selaras**

Anda dapat memperoleh wawasan yang kuat dan membuat kesimpulan yang akurat ketika data selaras dengan tujuan bisnis. Sebagai seorang analis data, keselarasan adalah sesuatu yang perlu Anda pertimbangkan. Penyelarasan yang baik berarti bahwa data tersebut relevan dan dapat membantu Anda memecahkan masalah bisnis atau menentukan tindakan untuk mencapai tujuan bisnis tertentu.

Dalam bacaan ini, Anda akan meninjau tujuan bisnis yang dikaitkan dengan tiga skenario. Anda akan mempelajari bagaimana data yang bersih dan tujuan bisnis yang selaras dapat membantu Anda menghasilkan kesimpulan yang akurat. Ditambah lagi, Anda akan mempelajari bagaimana variabel baru yang ditemukan selama analisis data akan membuat Anda menyiapkan batasan data sehingga Anda dapat menjaganya agar tetap selaras dengan tujuan bisnis.

**Data bersih + keselarasan dengan tujuan bisnis = kesimpulan yang akurat**

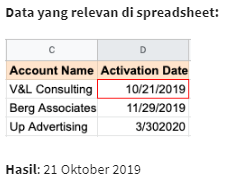
**Tujuan bisnis**

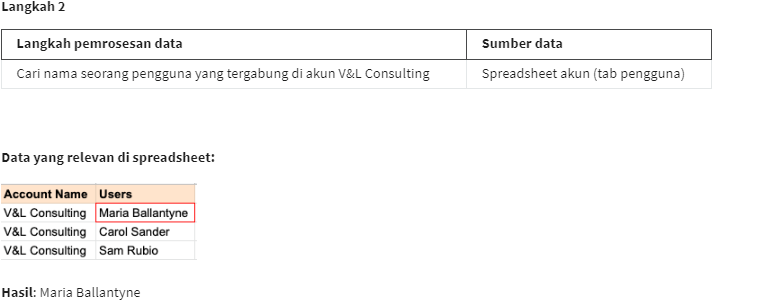
Manajer akun di Impress Me, layanan berlangganan konten online, ingin mengetahui seberapa cepat pengguna melihat konten setelah layanan langganan mereka diaktifkan.

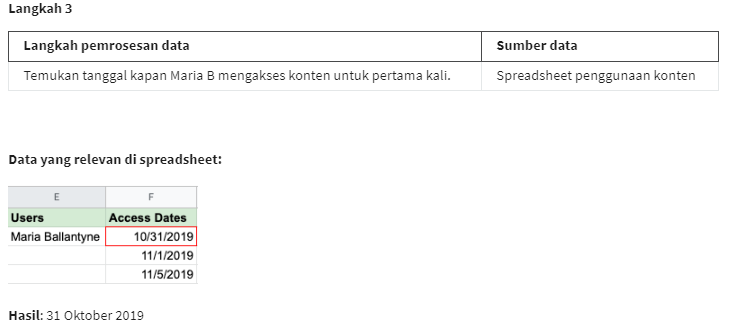
Untuk memulai, analis data memverifikasi bahwa data yang diekspor ke spreadsheet bersih dan mengonfirmasi bahwa data yang diperlukan (saat pengguna mengakses konten) tersedia. Mengetahui hal ini, analis memutuskan ada keselarasan data yang baik dengan tujuan bisnis. Yang belum ada adalah mencari tahu dengan tepat berapa lama waktu yang dibutuhkan setiap pengguna untuk melihat konten setelah langganan mereka diaktifkan.

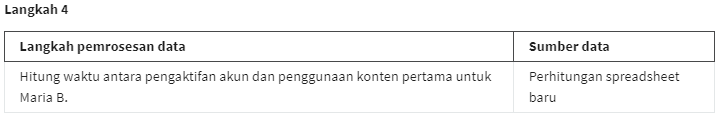
Berikut adalah langkah-langkah pemrosesan data yang dilakukan analis untuk pengguna akun bernama V&L Consulting. (Langkah-langkah ini akan diulang untuk setiap akun berlangganan, dan untuk setiap pengguna yang terkait dengan akun itu.)

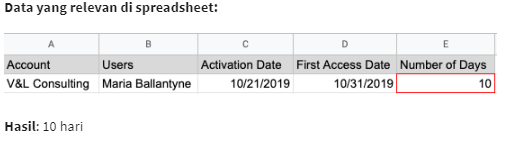












**Kiat pro 1**

Dalam proses di atas, analis dapat menggunakan VLOOKUP untuk mencari data di Langkah 1, 2, dan 3 untuk mendapatkan nilai-nilai dalam spreadsheet di Langkah 4. VLOOKUP adalah fungsi spreadsheet yang mencari nilai tertentu dalam sebuah kolom untuk menampilkan potongan informasi yang terkait. Menggunakan VLOOKUP dapat menghemat banyak waktu; tanpa itu, Anda harus mencari tanggal dan nama secara manual.

**Kiat pro 2**

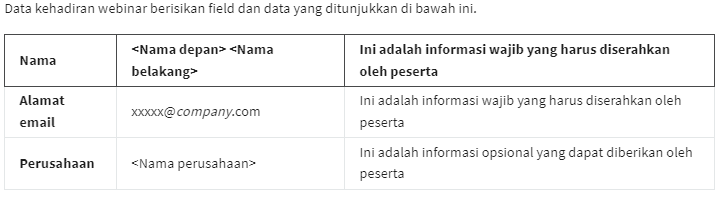
Pada Langkah 4 dari proses di atas, analis dapat menggunakan DATEDIF untuk menghitung perbedaan antara tanggal di kolom C dan kolom D secara otomatis. Fungsi ini dapat menghitung jumlah hari antara dua tanggal.

Lihat halaman Microsoft Support mengenai DATEDIF untuk cara menggunakan fungsi ini di Excel. Fungsi DAYS360 melakukan hal yang sama dalam spreadsheet akuntansi yang menggunakan standar 360 hari dalam setahun (dua belas bulan 30 hari).

**Penyelarasan dengan tujuan bisnis + pembersihan data tambahan = kesimpulan yang akurat**

**Tujuan bisnis**

Cloud Gate, sebuah perusahaan perangkat lunak, baru-baru ini menyelenggarakan serangkaian webinar publik sebagai pengenalan produk secara gratis. Analis data dan manajer program webinar ingin mengidentifikasi perusahaan yang mengirimkan lima orang atau lebih untuk menghadiri sesi ini. Mereka ingin memberikan daftar perusahaan ini kepada manajer penjualan yang dapat menindaklanjutinya sebagai potensi penjualan.



**Pembersihan data**

Data kehadiran webinar tampaknya selaras dengan tujuan bisnis. Tetapi analis data dan manajer program memutuskan bahwa perlu dilakukan pembersihan data sebelum analisis. Mereka berpikir pembersihan data diperlukan karena:

* Nama perusahaan bukan isian wajib. Jika nama perusahaan kosong, datanya mungkin bisa ditemukan dari alamat email. Misalnya, jika alamat email seseorang adalah namapengguna@google.com, ketika data dianalisis isian (field) nama perusahaan itu dapat diisi dengan Google. Langkah pembersihan data ini mengasumsikan bahwa orang-orang dengan alamat email perusahaan menghadiri webinar untuk tujuan bisnis.
* Peserta boleh memasukkan nama apa saja. Berhubung kehadiran di serangkaian webinar sedang diteliti, mereka perlu memvalidasi nama-nama dengan alamat email yang unik. Misalnya, jika Joe Cox menghadiri dua webinar tetapi masuk sebagai Joe Cox untuk satu webinar dan Joseph Cox untuk yang lain, dia akan dihitung sebagai dua orang yang berbeda. Untuk mencegah hal ini, mereka perlu memeriksa alamat emailnya yang unik untuk menentukan bahwa dia adalah orang yang sama. Setelah proses validasi, Joseph Cox dapat diubah menjadi Joe Cox agar sesuai dengan contoh lainnya.

**Penyelarasan dengan tujuan bisnis + variabel yang baru ditemukan + pembatasan = kesimpulan yang akurat**

**Tujuan bisnis**

Sebuah perusahaan bimbingan belajar, A+ Education, ingin mengetahui apakah ada jumlah jam pelajaran minimum yang diperlukan agar nilai siswa meningkat setidaknya 10%.

Analis data berpikir ada keselarasan yang baik antara data yang tersedia dan tujuan bisnis karena:

* Siswa yang masuk dan keluar untuk setiap sesi bimbingan belajar tercatat di sistem, dan jumlah jam bimbingannya dilacak
* Skor penilaian dicatat secara teratur

**Pembatasan data untuk variabel-variabel baru**

Setelah melihat data, analis data menemukan bahwa ada variabel lain yang perlu dipertimbangkan. Beberapa siswa memiliki sesi mingguan yang teratur sementara penjadwalan untuk siswa lain dilakukan secara lebih acak meskipun jumlah jam bimbingan mereka sama. Data ini juga tidak selaras dengan tujuan bisnis asli seperti yang diperkirakan sebelumnya, sehingga analis itu menambahkan pembatasan data sehingga analisis akan fokus hanya pada siswa yang konsisten memiliki sesi mingguan. Modifikasi ini membantu mendapatkan gambaran yang lebih akurat tentang masa belajar yang dibutuhkan untuk mencapai peningkatan 10% dalam skor penilaian.

**Hal penting yang bisa dipelajari**

Semoga contoh-contoh ini memberi gambaran tentang apa yang harus dicari untuk mengetahui apakah data Anda sesuai dengan tujuan bisnis Anda.

* Ketika ada data yang bersih dan keselarasan yang baik, Anda bisa mendapatkan wawasan yang akurat dan membuat kesimpulan yang didukung oleh data.
* Jika ada keselarasan yang baik tetapi data perlu dibersihkan, bersihkan data dahulu sebelum Anda melakukan analisis.
* Jika hanya ada sebagian data yang selaras dengan suatu tujuan, pikirkan bagaimana Anda dapat memodifikasi tujuan, atau menggunakan pembatasan data untuk memastikan bahwa sebagian data yang diperoleh nantinya lebih selaras dengan tujuan bisnis.

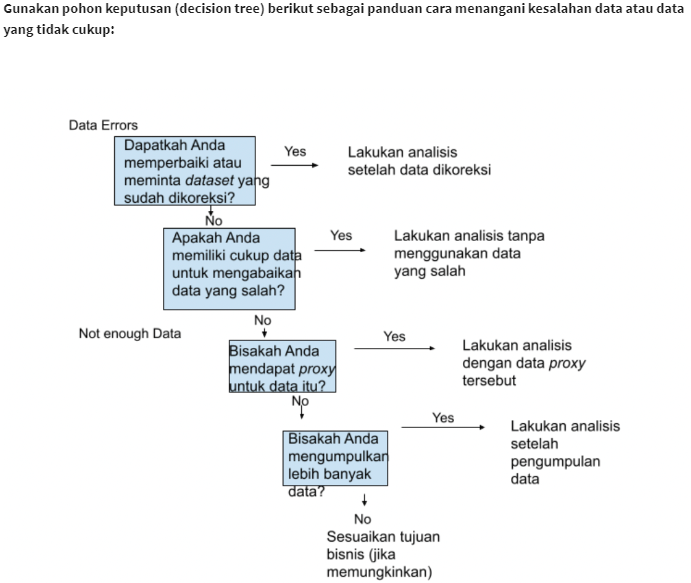
**Apa yang harus dilakukan ketika Anda menemukan masalah dengan data Anda**

Saat bersiap untuk analisis data, Anda mungkin menyadari bahwa Anda tidak memiliki data yang dibutuhkan atau Anda tidak memiliki cukup data. Pada beberapa kasus, Anda dapat menggunakan apa yang dikenal sebagai data proxy (data perwakilan) sebagai pengganti data sebenarnya. Bayangkan seperti mengganti mentega dengan minyak di suatu resep ketika saat itu Anda tidak memiliki mentega. Namun pada kasus lain, tidak ada pengganti yang bisa diterima dan satu-satunya pilihan Anda adalah mengumpulkan lebih banyak data.

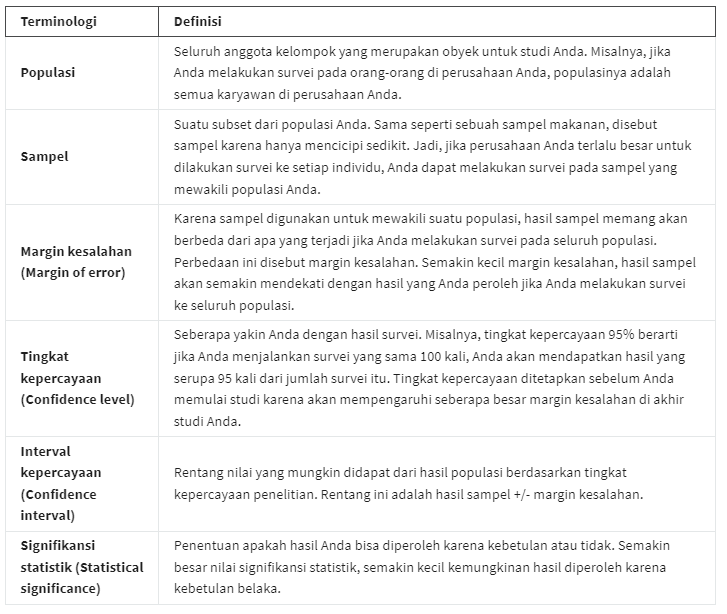
Pertimbangkan masalah data berikut dan saran-saran untuk mengatasinya.







**Menghitung ukuran sampel**



**Hal-hal yang perlu diingat saat menentukan ukuran sampel Anda**

Ketika menghitung ukuran sampel, berikut adalah hal-hal yang perlu diingat:

* Jangan menggunakan ukuran sampel kurang dari 30. Telah terbukti secara statistik bahwa 30 adalah ukuran sampel terkecil di mana hasil rata-rata sebuah sampel mulai mewakili hasil rata-rata suatu populasi.
* Tingkat kepercayaan yang paling umum digunakan adalah 95%, tetapi 90% juga dapat diterima di beberapa kasus.
* Tingkatkan ukuran sampel untuk memenuhi kebutuhan khusus pada proyek Anda:
* Untuk tingkat kepercayaan yang lebih tinggi, gunakan ukuran sampel yang lebih besar
* Untuk mengurangi margin kesalahan, gunakan ukuran sampel yang lebih besar
* Untuk signifikansi statistik yang lebih besar, gunakan ukuran sampel yang lebih besar

**Catatan:** Kalkulator ukuran sampel menggunakan rumus statistik untuk menentukan ukuran sampel. Selebihnya akan segera dibahas di materi ini! Jadi, tetaplah mengikuti materi ini.

**Mengapa ukuran sampel minimal 30?**

Rekomendasi ini didasarkan pada Teorema Limit Pusat (Central Limit Theorem, CLT) di bidang probabilitas dan statistik. Dengan bertambahnya ukuran sampel, hasilnya akan semakin mendekati bentuk distribusi normal (berbentuk lonceng) dari sampel yang berjumlah besar. Sampel 30 adalah ukuran sampel terkecil yang CLT-nya masih valid. Para peneliti yang mengandalkan analisis regresi – metode statistik untuk menentukan hubungan antara variabel terkontrol dan dependen – juga lebih menyukai sampel minimal sejumlah 30.

Masih penasaran? Tanpa terlalu banyak membahas matematika, lihat artikel ini:

* [Teorema Limit Pusat (Central Limit Theorem, CLT):](https://www.investopedia.com/terms/c/central_limit_theorem.asp) Artikel oleh Investopedia ini menjelaskan Teorema Limit Pusat (CLT) dan menggambarkan secara singkat bagaimana hal itu dapat diterapkan pada analisis indeks saham.
* [Rumus Ukuran Sampel (Sample Size Formula):](https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/sample-size-formula/) Artikel dari Statistics Solutions ini sedikit menjelaskan tentang mengapa banyak peneliti menggunakan 30 sebagai ukuran sampel minimum.

**Ukuran sampel bervariasi berdasarkan masalah bisnis**

Ukuran sampel akan bervariasi berdasarkan jenis masalah bisnis yang Anda coba selesaikan. Misalnya, jika Anda tinggal di kota dengan populasi 200.000 dan mendapatkan 180.000 orang bersedia memberi respons survei, itu adalah ukuran sampel yang besar. Tetapi bila tidak melakukan itu, seperti apa ukuran sampel yang lebih kecil dan masih dapat diterima?

Apakah 200 masih bisa diterima jika orang yang disurvei mewakili setiap distrik di kota?

Jawaban: Hal itu tergantung pada apa yang akan diteliti.

* Ukuran sampel 200 mungkin cukup besar jika masalah bisnis Anda adalah untuk mengetahui tanggapan warga tentang perpustakaan baru
* Ukuran sampel 200 mungkin tidak cukup besar jika masalah bisnis Anda adalah menentukan pilihan bagaimana warga akan mendanai perpustakaan itu

Anda mungkin dapat menerima margin kesalahan yang lebih besar dengan melakukan survei bagaimana tanggapan warta tentang perpustakaan baru versus survei pilihan warga untuk mendanai perpustakaan itu. Karena alasan itu, kemungkinan besar Anda akan menggunakan ukuran sampel yang lebih besar untuk survei pemilihan.

* **Ukuran sampel yang lebih besar memiliki biaya yang lebih tinggi**

Anda juga harus mempertimbangkan biaya dibandingkan manfaat hasil yang lebih akurat dengan ukuran sampel yang lebih besar. Seseorang yang mencoba memahami preferensi konsumen untuk suatu lini produk baru tidak akan membutuhkan ukuran sampel sebesar orang yang mencoba memahami pengaruh suatu obat baru. Untuk keamanan penggunaan obat, hasilnya jauh lebih bermanfaat dibandingkan dengan biaya untuk menggunakan ukuran sampel yang lebih besar. Sedangkan untuk preferensi konsumen, ukuran sampel yang lebih kecil dengan biaya lebih rendah dapat memberikan hasil yang cukup baik.

* **Mengetahui dasar-dasar sampel sangat membantu**

Mengetahui dasar-dasar sampel akan membantu Anda membuat pilihan ukuran sampel yang tepat. Anda selalu dapat mengutarakan kekhawatiran jika Anda menemukan ukuran sampel yang terlalu kecil. Kalkulator ukuran sampel juga merupakan alat yang hebat untuk ini. Kalkulator ukuran sampel memungkinkan Anda memasukkan tingkat kepercayaan dan margin kesalahan yang diinginkan untuk ukuran populasi tertentu. Kalkulator itu kemudian menghitung ukuran sampel yang dibutuhkan untuk mencapai hasil tersebut secara statistik.

Lihat video [Menentukan Ukuran Sampel Terbaik](https://www.coursera.org/learn/memproses-data-dari-kotor-ke-bersih/lecture/mSj5A/tentukan-ukuran-sampel-terbaik) untuk melihat demonstrasi kalkulator ukuran sampel, atau lihat bacaan [Kalkulator Ukuran Sampel (Sample Size Calculator)](https://www.coursera.org/learn/memproses-data-dari-kotor-ke-bersih/supplement/ZqcDw/kalkulator-ukuran-sampel) untuk mendapat tambahan informasi.

**Meninjau integritas data**

Sebelum analis data dapat menganalisis data, mereka harus terlebih dahulu memikirkan dan memahami data yang sedang mereka kerjakan. Menilai integritas data adalah langkah penting di proses ini. Seperti yang telah Anda pelajari di pelajaran sebelumnya, Anda harus menyelesaikan tugas-tugas berikut sebelum menganalisis data:

1. Menentukan integritas data dengan menilai akurasi, konsistensi, dan kelengkapan data secara menyeluruh.
2. Menghubungkan tujuan ke data dengan memahami bagaimana tujuan bisnis Anda dapat dicapai dengan investigasi data.
3. Mengetahui kapan pengumpulan data harus dihentikan.

Analis data melakukan aktivitas pra-pembersihan untuk menyelesaikan langkah-langkah ini. Aktivitas pra-pembersihan membantu Anda menentukan dan memelihara integritas data, yang merupakan hal penting bagi seorang analis data junior.

**Apa yang membuat data tidak mencukupi?**

Salah satu tujuan kegiatan pra-pembersihan adalah untuk mengatasi data yang tidak mencukupi. Ingat kembali pelajaran sebelumnya bahwa data mungkin tidak mencukupi karena sejumlah alasan. Berbagai permasalahan yang menyebabkan data tidak mencukupi adalah sebagai berikut:

* Data hanya berasal dari satu sumber
* Terus-menerus diperbarui dan tidak akan pernah lengkap
* Sudah usang
* Data yang terbatas secara geografis

Untuk menangani data yang tidak mencukupi, Anda dapat:

* Mengidentifikasi tren dengan data yang tersedia
* Mengumpulkan lebih banyak data jika waktu memungkinkan
* Berdiskusi dengan pemangku kepentingan dan menyesuaikan tujuan Anda
* Mencari dataset baru

**Menggunakan kekuatan statistik**

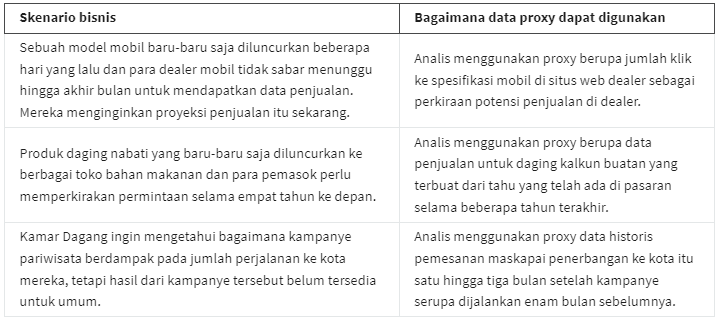
Kekuatan statistik adalah kemungkinan mendapatkan hasil yang berarti dari suatu pengujian. "Kekuatan statistik dapat dihitung dan disebutkan pada sebuah percobaan yang telah selesai untuk menyatakan sejauh mana seseorang bisa memercayai kesimpulan yang diambil dari hasil penelitian itu. Kekuatan statistik juga dapat digunakan sebagai alat untuk memperkirakan jumlah pengamatan atau ukuran sampel yang diperlukan untuk mendeteksi suatu efek dalam sebuah percobaan."

Pengujian hipotesa atau hipotesis adalah cara untuk melihat apakah suatu survei atau percobaan memiliki hasil yang berarti. Pada dasarnya, jika sebuah pengujian dinyatakan signifikan secara statistik, berarti itu adalah hasil uji yang benar dan bukan kesalahan yang disebabkan oleh kejadian acak. Biasanya, Anda membutuhkan kekuatan statistik minimal 0,8 atau 80% untuk menyatakan hasil Anda signifikan secara statistik.

**Apa yang harus dilakukan ketika tidak ada data**

**Contoh data proxy**

Terkadang data untuk mendukung tujuan bisnis tidak tersedia. Saat inilah data proxy berguna. Lihatlah skenario berikut dan di mana penggunaan data proxy untuk setiap contoh:



**Dataset terbuka (publik)**

Jika Anda adalah bagian dari suatu organisasi besar, Anda mungkin memiliki akses ke berbagai sumber data. Tetapi jika Anda mencari sesuatu yang khusus atau sedikit di luar lini bisnis, Anda juga dapat menggunakan dataset yang terbuka atau publik. (Anda dapat merujuk ke artikel Menuju Ilmu Data (Towards Data Science) sebagai penjelasan singkat tentang perbedaan antara data terbuka dan data publik.)

Berikut ini contohnya. Sebuah versi vaksin jenis nasal telah tersedia baru-baru ini. Sebuah klinik ingin mengetahui kontraindikasi apa yang mungkin muncul, tetapi baru memulai mengumpulkan data pihak pertama dari pasiennya. Kontraindikasi adalah suatu kondisi yang menyebabkan pasien tidak bisa menerima vaksin karena bahaya yang akan ditimbulkan jika vaksin itu diberikan. Untuk memperkirakan jumlah kemungkinan kontraindikasi, seorang analis data menggunakan proxy dari dataset terbuka uji coba vaksin versi injeksi. Analis memilih anggota himpunan data dengan profil pasien yang paling mendekati susunan pasien di klinik itu.

Ada banyak cara untuk berbagi dan berkolaborasi data di dalam sebuah komunitas. Kaggle (kaggle.com) yang sebelumnya kami perkenalkan, memiliki banyak dataset dalam berbagai format termasuk tipe paling sederhana, yaitu file berupa Comma Separated Values (CSV).

**CSV, J​SON, SQLite, dan BigQuery**

* CSV: Lihat dataset [pelanggan kartu kredit](https://www.kaggle.com/sakshigoyal7/credit-card-customers) ini, yang berisi informasi dari 10.000 pelanggan termasuk usia, gaji, status perkawinan, batas kartu kredit, kategori kartu kredit, dll. (CC0: Domain Publik, Sakshi Goyal).
* JSON: Lihat dataset JSON yang berisi [video-video YouTube yang sedang tren](https://www.kaggle.com/datasets/datasnaek/youtube-new) (CC0: Domain Publik, Mitchell J).
* SQLite: Lihat dataset SQLite yang berisi [data kebakaran AS](https://www.kaggle.com/rtatman/188-million-us-wildfires) (U.S. wildfire data) selama 24 tahun (CC0: Domain Publik, Rachael Tatman).
* BigQuery: Lihat dataset sampel [Google Analytics 360](https://www.kaggle.com/datasets/bigquery/google-analytics-sample) dari Google Merchandise Store CC0, Domain Publik, Google BigQuery).

**Kalkulator ukuran sampel**

Kalkulator ukuran sampel memberi tahu Anda berapa banyak orang yang perlu Anda wawancarai (atau hal-hal yang perlu Anda uji) untuk mendapatkan hasil yang mewakili populasi target. Mari kita tinjau beberapa istilah yang akan Anda temui saat menggunakan kalkulator ukuran sampel:

* Tingkat kepercayaan (Confidence level): Probabilitas bahwa ukuran sampel Anda secara akurat mencerminkan populasi yang lebih besar.
* Margin kesalahan (Margin of error): Jumlah maksimum hasil sampel yang sudah diperkirakan akan berbeda dari populasi sebenarnya
* Populasi (Population): Jumlah total yang Anda harapkan untuk diambil sampelnya.
* Sampel (Sample): Sampel adalah bagian dari populasi yang mewakili populasi tersebut.
* Estimasi tingkat tanggapan (Estimated response rate): Jika Anda menjalankan survei terhadap individu, ini adalah persentase orang yang Anda harapkan akan menyelesaikan survei Anda dari sejumlah orang yang telah menerima survei itu.

**Cara menggunakan kalkulator ukuran sampel**

Untuk menggunakan kalkulator ukuran sampel, Anda harus sudah memutuskan ukuran populasi, tingkat kepercayaan, dan margin kesalahan yang dapat diterima sehingga Anda dapat memasukkan nilainya ke kalkulator itu. Jika informasi ini sudah disiapkan, lihat kalkulator ukuran sampel di bawah ini:

* [Kalkulator ukuran sampel buatan surveymonkey.com](https://www.surveymonkey.com/mp/sample-size-calculator/)
* [Kalkulator ukuran sampel buatan raosoft.com](http://www.raosoft.com/samplesize.html)

**Apa yang harus dilakukan dengan hasilnya?**

Setelah Anda memasukkan informasi Anda ke salah satu kalkulator ini, mereka akan memberi Anda ukuran sampel yang disarankan. Perlu diingat, ukuran sampel yang dihitung adalah jumlah minimum untuk mencapai tingkat kepercayaan dan margin kesalahan yang Anda masukkan. Jika Anda mengerjakan sebuah survei, Anda perlu memikirkan perkiraan tingkat tanggapan untuk mengetahui berapa banyak survei yang perlu Anda kirimkan. Misalnya, jika Anda memerlukan ukuran sampel sebanyak 100 individu dan perkiraan tingkat tanggapannya adalah 10%, Anda perlu mengirimkan survei ke 1.000 individu untuk mendapatkan 100 tanggapan yang diperlukan untuk analisis Anda.

**Semua hal tentang margin kesalahan**

Margin kesalahan adalah jumlah maksimum hasil sampel yang sudah diperkirakan akan berbeda dari populasi sebenarnya. Secara teknis, margin kesalahan mendefinisikan kisaran nilai di bawah dan di atas hasil rata-rata sampel. Hasil rata-rata untuk seluruh populasi diharapkan berada dalam kisaran tersebut. Kita dapat lebih memahami margin kesalahan dengan menggunakan beberapa contoh di bawah ini.

* Tingkat kepercayaan (Confidence level): Persentase yang menunjukkan seberapa besar kemungkinan sampel Anda secara akurat mencerminkan populasi yang lebih besar
* Populasi (Population): Jumlah total yang Anda ambil sampelnya
* Sampel (Sample): Sampel adalah bagian dari populasi yang mewakili populasi tersebut.
* Margin kesalahan (Margin of error): Jumlah maksimum dari hasil sampel yang sudah diperkirakan berbeda dari populasi sebenarnya

Di kebanyakan kasus, umumnya digunakan tingkat kepercayaan 90% atau 95%. Namun, tergantung pada industrinya, Anda mungkin ingin menggunakan tingkat kepercayaan yang lebih ketat. Tingkat kepercayaan 99% bisa diterima di beberapa industri, seperti industri farmasi.

Setelah Anda menentukan ukuran populasi, ukuran sampel, dan tingkat kepercayaan Anda, masukkan informasi itu ke dalam kalkulator margin kesalahan seperti di bawah ini:

* [Margin of error calculator by Good Calculators (kalkulator online gratis)](https://goodcalculators.com/margin-of-error-calculator/)
* [Margin of error calculator by CheckMarket](https://www.checkmarket.com/sample-size-calculator/%23sample-size-margin-of-error-calculator)

**Margin kesalahan dalam pemasaran**

Margin kesalahan juga penting dalam pemasaran. Mari kita gunakan pengujian A/B sebagai contoh. Tes A/B (atau split test) adalah pengujian dua variasi laman web yang sama untuk menentukan laman mana yang lebih berhasil menarik lalu lintas pengguna dan menghasilkan pendapatan. Lalu lintas pengguna yang bisa diuangkan dikenal sebagai tingkat konversi (conversion rate). Tes A/B memungkinkan pemasar menguji email, iklan, dan laman arahan untuk menemukan data penyebab kampanye itu berhasil dan tidak. Pemasar menggunakan interval kepercayaan (ditentukan oleh tingkat konversi dan margin kesalahan) untuk memahami hasilnya.

Misalnya, Anda sedang melakukan tes A/B untuk membandingkan keefektifan dua baris subjek email yang berbeda untuk menarik minat orang untuk membuka email tersebut. Anda menemukan bahwa baris subjek A: “Penawaran khusus hanya untuk Anda” membuat 5% penerima email membukanya dibandingkan dengan baris subjek B: “Jangan lewatkan kesempatan ini” yang hanya menarik 3% penerima.

Apakah itu berarti baris subjek A lebih baik daripada baris subjek B? Jawabannya tergantung pada margin kesalahan Anda. Jika margin kesalahan adalah 2%, maka tingkat minat membuka email aktual atau interval kepercayaan baris subjek A berada di antara 3% dan 7%. Karena ujung bawah interval tumpang tindih dengan hasil baris subjek B sebesar 3%, Anda tidak dapat menyimpulkan bahwa ada perbedaan yang signifikan secara statistik antara baris subjek A dan B. Memeriksa margin kesalahan penting ketika membuat kesimpulan berdasarkan pengujian hasil Anda.