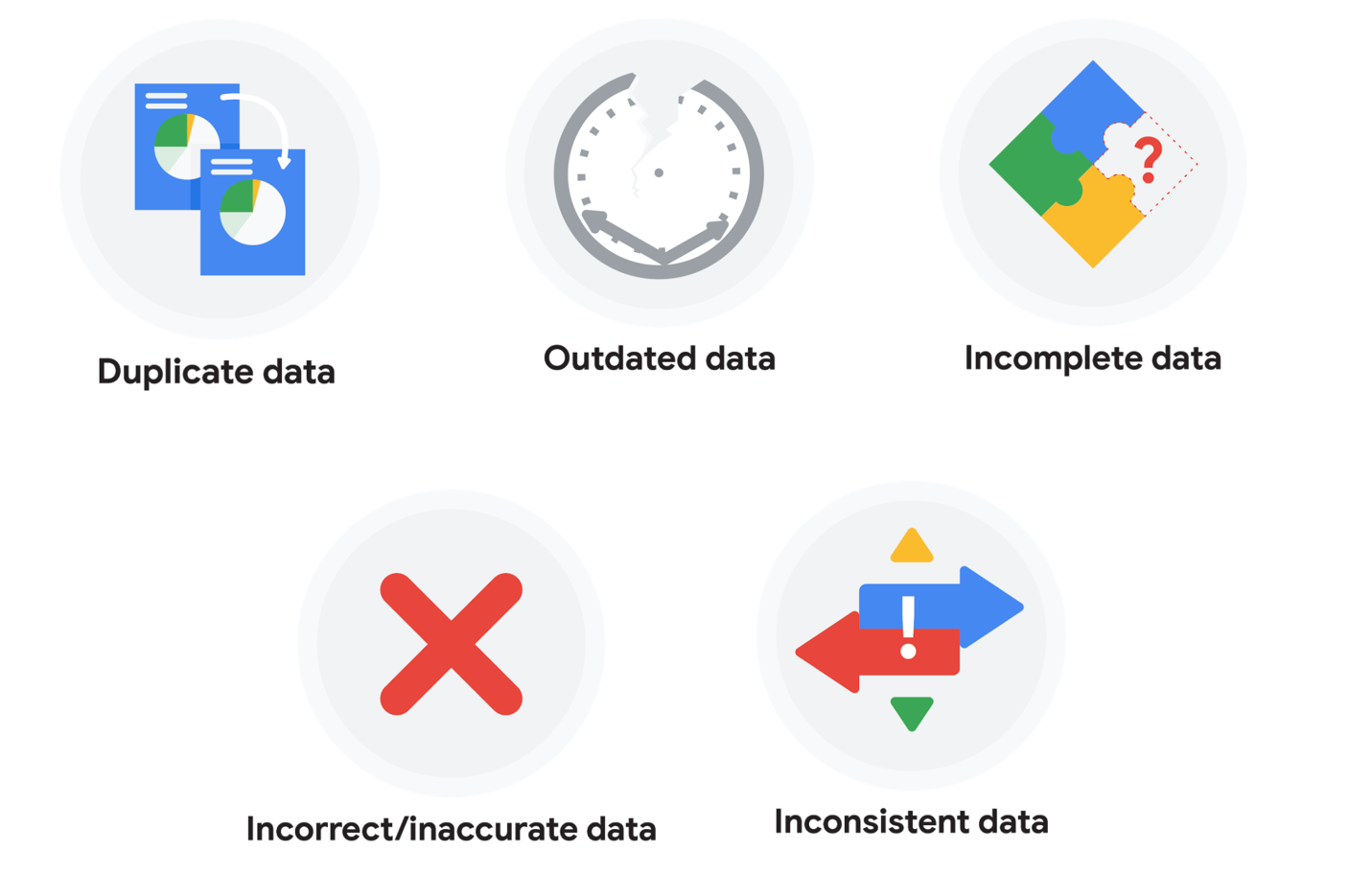
**Apa itu data kotor?**

Sebelumnya kita telah membahas bahwa **data kotor** adalah data yang tidak lengkap, salah, atau tidak relevan dengan masalah yang ingin Anda selesaikan. Bacaan ini merangkum:

* Jenis data kotor yang mungkin Anda temui
* Apa yang mungkin menyebabkan data menjadi kotor
* Betapa data kotor berbahaya bagi bisnis

**Jenis data kotor**



**Data duplikat**

| **Keterangan** | **Kemungkinan penyebab** | **Potensi kerugian bagi dunia usaha** |
| --- | --- | --- |
| Catatan data apa pun yang muncul lebih dari satu kali | Entri data manual, impor data batch, atau migrasi data | Metrik atau analisis yang tidak tepat, penghitungan atau prediksi yang berlebihan atau tidak akurat, atau kebingungan selama pengambilan data |

**Data kedaluwarsa**

| **Keterangan** | **Kemungkinan penyebab** | **Potensi kerugian bagi dunia usaha** |
| --- | --- | --- |
| Setiap data yang sudah lama sebaiknya diganti dengan informasi yang lebih baru dan akurat | Orang-orang berganti peran atau perusahaan, atau perangkat lunak dan sistem menjadi usang | Wawasan, pengambilan keputusan, dan analisis yang tidak akurat |

**Data tidak lengkap**

| **Keterangan** | **Kemungkinan penyebab** | **Potensi kerugian bagi dunia usaha** |
| --- | --- | --- |
| Data apa pun yang tidak berisi bidang penting | Pengumpulan data yang tidak benar atau entri data yang salah | Penurunan produktivitas, wawasan yang tidak akurat, atau ketidakmampuan menyelesaikan layanan penting |

**Data yang salah/tidak akurat**

| **Keterangan** | **Kemungkinan penyebab** | **Potensi kerugian bagi dunia usaha** |
| --- | --- | --- |
| Data apa pun yang lengkap tetapi tidak akurat | Kesalahan manusia dimasukkan pada saat input data, informasi palsu, atau data tiruan | Wawasan atau pengambilan keputusan yang tidak akurat berdasarkan informasi buruk yang mengakibatkan hilangnya pendapatan |

**Data yang tidak konsisten**

| **Keterangan** | **Kemungkinan penyebab** | **Potensi kerugian bagi dunia usaha** |
| --- | --- | --- |
| Data apa pun yang menggunakan format berbeda untuk mewakili hal yang sama | Data disimpan secara tidak benar atau kesalahan dimasukkan selama transfer data | Poin data yang kontradiktif menyebabkan kebingungan atau ketidakmampuan untuk mengklasifikasikan atau mengelompokkan pelanggan |

**Dampak bisnis dari data kotor**

Untuk membaca lebih lanjut mengenai dampak bisnis dari data kotor, masukkan istilah “data kotor” ke dalam bilah pencarian browser pilihan Anda untuk memunculkan banyak artikel tentang topik tersebut. Berikut adalah beberapa dampak yang dikutip untuk industri tertentu dari penelusuran sebelumnya:

* **Perbankan** : Ketidakakuratan merugikan perusahaan antara 15% dan 25% pendapatan ([sumber](https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/)).
* **Perdagangan digital:** Hingga 25% kontak database B2B mengandung ketidakakuratan ([sumber](https://www.demandgen.com/dirty-data-what-is-it-costing-you/)).
* **Pemasaran dan penjualan** : 99% perusahaan secara aktif menangani kualitas data dengan berbagai cara ([sumber](https://www.dqglobal.com/blog/why-bad-data-is-wasting-your-marketing-efforts/)).
* **Pelayanan Kesehatan** : Duplikasi catatan dapat mencapai 10% bahkan sampai 20% dari catatan kesehatan elektronik rumah sakit ([sumber](https://searchhealthit.techtarget.com/feature/Hospitals-battle-duplicate-medical-records-with-technology)).

**Kesalahan umum dalam pembersihan data**

Dalam bacaan ini, Anda akan mempelajari pentingnya pembersihan data dan cara mengidentifikasi kesalahan umum. Beberapa kesalahan yang mungkin Anda temui saat membersihkan data meliputi:



**Kesalahan umum yang harus dihindari**

* **Tidak memeriksa kesalahan ejaan** : Kesalahan ejaan bisa sesederhana kesalahan pengetikan atau masukan. Seringkali kesalahan ejaan atau kesalahan tata bahasa yang umum dapat dideteksi, namun hal ini menjadi lebih sulit jika terjadi pada hal-hal seperti nama atau alamat. Misalnya, jika Anda bekerja dengan tabel spreadsheet data pelanggan, Anda mungkin menemukan pelanggan bernama “John” yang namanya salah dimasukkan sebagai “Jon” di beberapa tempat. Pemeriksa ejaan pada spreadsheet mungkin tidak akan menandai hal ini, jadi jika Anda tidak memeriksa ulang kesalahan ejaan dan menangkapnya, analisis Anda akan memiliki kesalahan di dalamnya.
* **Lupa mendokumentasikan kesalahan** : Mendokumentasikan kesalahan dapat sangat menghemat waktu, karena membantu Anda menghindari kesalahan tersebut di masa mendatang dengan menunjukkan cara mengatasinya. Misalnya, Anda mungkin menemukan kesalahan dalam rumus di spreadsheet Anda. Anda menemukan bahwa beberapa tanggal di salah satu kolom Anda belum diformat dengan benar. Jika Anda mencatat perbaikan ini, Anda dapat mereferensikannya saat rumus Anda rusak lagi, dan memulai pemecahan masalah lebih awal. Mendokumentasikan kesalahan Anda juga membantu Anda melacak perubahan dalam pekerjaan Anda, sehingga Anda dapat menelusuri kembali jika perbaikan tidak berhasil.
* **Tidak memeriksa nilai yang salah kolom** : Nilai yang salah kolom terjadi ketika nilai dimasukkan ke kolom yang salah. Nilai-nilai ini mungkin masih diformat dengan benar, sehingga lebih sulit ditangkap jika Anda tidak hati-hati. Misalnya, Anda mungkin memiliki kumpulan data dengan kolom untuk kota dan negara. Ini adalah tipe data yang sama, sehingga mudah tercampur. Namun jika Anda mencoba menemukan semua contoh Spanyol di kolom negara, dan Spanyol salah dimasukkan ke kolom kota, Anda akan kehilangan poin data penting. Memastikan data Anda telah dimasukkan dengan benar adalah kunci analisis yang akurat dan lengkap.
* **Mengabaikan nilai yang hilang** : Nilai yang hilang dalam kumpulan data Anda dapat menimbulkan kesalahan dan memberi Anda kesimpulan yang tidak akurat. Misalnya, jika Anda mencoba mendapatkan jumlah total penjualan dari tiga bulan terakhir, namun ada satu minggu transaksi yang hilang, penghitungan Anda akan menjadi tidak akurat. Sebagai praktik terbaik, usahakan menjaga data Anda sebersih mungkin dengan menjaga kelengkapan dan konsistensi.
* **Hanya melihat sebagian data** : Penting untuk memikirkan semua data yang relevan saat Anda membersihkan. Hal ini membantu memastikan Anda memahami keseluruhan cerita yang disampaikan data, dan memperhatikan semua kemungkinan kesalahan. Misalnya, jika Anda bekerja dengan data tentang pola migrasi burung dari sumber berbeda, namun Anda hanya membersihkan satu sumber, Anda mungkin tidak menyadari bahwa beberapa data diulangi. Hal ini akan menimbulkan masalah pada analisis Anda nantinya. Jika Anda ingin menghindari kesalahan umum seperti duplikat, setiap bidang data Anda memerlukan perhatian yang sama.
* **Kehilangan tujuan bisnis** : Saat Anda membersihkan data, Anda mungkin membuat penemuan baru dan menarik tentang kumpulan data Anda-- namun Anda tidak ingin penemuan tersebut mengalihkan perhatian Anda dari tugas yang ada. Misalnya, jika Anda menggunakan data cuaca untuk menemukan jumlah rata-rata hari hujan di kota Anda, Anda mungkin juga melihat beberapa pola menarik tentang hujan salju. Itu sangat menarik, tetapi ini tidak ada hubungannya dengan pertanyaan yang ingin Anda jawab saat ini. Menjadi penasaran itu bagus! Namun cobalah untuk tidak membiarkan hal itu mengalihkan perhatian Anda dari tugas yang ada.
* **Tidak memperbaiki sumber kesalahan:** Memperbaiki kesalahan itu sendiri itu penting. Namun jika kesalahan tersebut sebenarnya adalah bagian dari masalah yang lebih besar, Anda perlu mencari sumber masalahnya. Jika tidak, Anda harus terus memperbaiki kesalahan yang sama berulang kali. Misalnya, bayangkan Anda memiliki spreadsheet tim yang melacak kemajuan setiap orang. Tabelnya terus pecah karena orang yang berbeda memasukkan nilai yang berbeda. Anda dapat terus memperbaiki semua masalah ini satu per satu, atau Anda dapat menyiapkan tabel untuk menyederhanakan entri data sehingga semua orang memiliki pemikiran yang sama. Mengatasi sumber kesalahan dalam data Anda akan menghemat banyak waktu dalam jangka panjang.
* **Tidak menganalisis sistem sebelum pembersihan data:** Jika kami ingin membersihkan data dan menghindari kesalahan di masa mendatang, kami perlu memahami akar penyebab data kotor Anda. Bayangkan Anda adalah seorang montir mobil. Anda akan menemukan penyebab masalahnya sebelum Anda mulai memperbaiki mobil, bukan? Hal yang sama berlaku untuk data. Pertama, Anda mencari tahu dari mana kesalahan itu berasal. Mungkin karena kesalahan entri data, tidak mengatur pemeriksa ejaan, format yang kurang, atau karena duplikat. Kemudian, setelah Anda memahami dari mana data buruk berasal, Anda dapat mengontrolnya dan menjaga data Anda tetap bersih.
* **Tidak mencadangkan data Anda sebelum pembersihan data** : Sebaiknya bersikap proaktif dan membuat cadangan data sebelum Anda memulai pembersihan data. Jika program Anda mogok, atau jika perubahan Anda menyebabkan masalah pada kumpulan data, Anda selalu dapat kembali ke versi yang disimpan dan memulihkannya. Prosedur sederhana untuk mencadangkan data dapat menghemat jam kerja Anda-- dan yang paling penting, membuat pusing kepala.
* **Tidak memperhitungkan pembersihan data dalam tenggat waktu/proses Anda** : Semua hal baik membutuhkan waktu, dan itu termasuk pembersihan data. Penting untuk mengingat hal itu saat menjalani proses dan melihat tenggat waktu Anda. Jika Anda menyisihkan waktu untuk pembersihan data, hal ini akan membantu Anda mendapatkan perkiraan ETA yang lebih akurat bagi pemangku kepentingan, dan dapat membantu Anda mengetahui kapan harus meminta ETA yang disesuaikan.

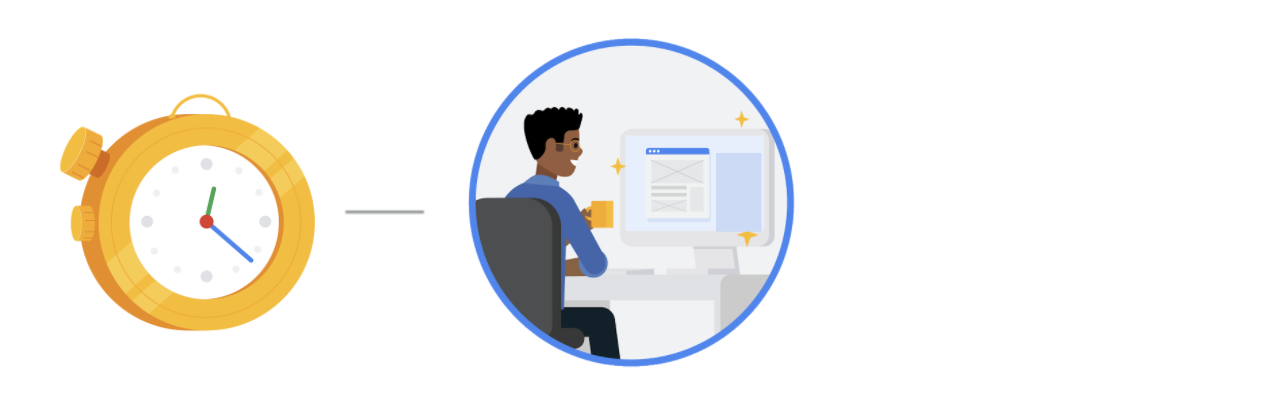
**Sumber daya tambahan**

Lihat daftar "sepuluh besar" berikut untuk pembersihan data di Microsoft Excel dan Google Spreadsheet untuk membantu Anda menghindari kesalahan paling umum:

* [Sepuluh cara teratas untuk membersihkan data Anda](https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19): Tinjau panduan tertib pembersihan data di Microsoft Excel.
* [10 tips Google Workspace untuk membersihkan data](https://support.google.com/a/users/answer/9604139?hl=en#zippy=): Pelajari praktik terbaik untuk pembersihan data di Google Spreadsheet.

**Otomatisasi alur kerja**

Dalam bacaan ini, Anda akan mempelajari tentang otomatisasi alur kerja dan bagaimana hal itu dapat membantu Anda bekerja lebih cepat dan efisien. Pada dasarnya, otomatisasi alur kerja adalah proses mengotomatiskan bagian-bagian pekerjaan Anda. Itu bisa berarti membuat pemicu peristiwa yang mengirimkan pemberitahuan ketika sistem diperbarui. Atau bisa juga berarti mengotomatisasi bagian dari proses pembersihan data. Seperti yang mungkin dapat Anda bayangkan, mengotomatiskan berbagai bagian pekerjaan Anda dapat menghemat banyak waktu, meningkatkan produktivitas, dan memberi Anda lebih banyak bandwidth untuk fokus pada aspek penting lainnya dari pekerjaan tersebut.



**Apa yang bisa diotomatisasi?**

Otomatisasi terdengar luar biasa, bukan? Namun meskipun mudah, masih ada beberapa bagian pekerjaan yang tidak dapat diotomatisasi. Mari kita lihat beberapa hal yang dapat kita otomatisasi dan beberapa hal yang tidak dapat kita otomatisasi.

| **Tugas** | **Bisakah itu diotomatisasi?** | **Mengapa?** |
| --- | --- | --- |
| Berkomunikasi dengan tim dan pemangku kepentingan Anda | TIDAK | Komunikasi adalah kunci untuk memahami kebutuhan tim dan pemangku kepentingan Anda saat Anda menyelesaikan tugas yang sedang Anda kerjakan. Tidak ada pengganti untuk komunikasi orang-ke-orang. |
| Mempresentasikan temuan Anda | TIDAK | Menyajikan data adalah bagian besar dari pekerjaan Anda sebagai analis data. Membuat data dapat diakses dan dipahami oleh pemangku kepentingan serta membuat visualisasi data tidak dapat diotomatisasi karena alasan yang sama seperti komunikasi tidak dapat diotomatisasi. |
| Mempersiapkan dan membersihkan data | Sebagian | Beberapa tugas dalam persiapan dan pembersihan data dapat diotomatisasi dengan menyiapkan proses tertentu, seperti menggunakan skrip pemrograman untuk mendeteksi nilai yang hilang secara otomatis. |
| Eksplorasi data | Sebagian | Terkadang cara terbaik untuk memahami data adalah dengan melihatnya. Untungnya, ada banyak alat yang tersedia yang dapat membantu mengotomatiskan proses visualisasi data. Tools tersebut dapat mempercepat proses visualisasi dan pemahaman data, namun eksplorasinya sendiri tetap perlu dilakukan oleh seorang data analis. |
| Memodelkan data | Ya | Pemodelan data adalah proses sulit yang melibatkan banyak faktor berbeda; untungnya ada alat yang dapat sepenuhnya mengotomatiskan berbagai tahapan. |

**Lebih lanjut tentang mengotomatiskan pembersihan data**

Salah satu cara terpenting untuk menyederhanakan pembersihan data adalah dengan membersihkan data di tempatnya. Ini akan menguntungkan seluruh tim Anda, dan ini juga berarti Anda tidak perlu mengulangi proses tersebut berulang kali. Misalnya, Anda dapat membuat skrip pemrograman yang menghitung jumlah kata di setiap file spreadsheet yang disimpan dalam folder tertentu. Menggunakan alat yang dapat digunakan di tempat penyimpanan data Anda berarti Anda tidak perlu mengulangi langkah pembersihan, sehingga menghemat waktu dan energi Anda dan tim.

**Lebih banyak sumber daya**

Ada banyak alat di luar sana yang dapat membantu mengotomatiskan proses Anda, dan alat tersebut terus meningkat seiring waktu. Berikut adalah beberapa artikel atau blog yang dapat Anda lihat jika Anda ingin mempelajari lebih lanjut tentang otomatisasi alur kerja dan berbagai alat yang dapat Anda gunakan:

* Menuju Ilmu Data[**Mengotomatiskan Analisis Data Ilmiah**](https://towardsdatascience.com/automating-scientific-data-analysis-part-1-c9979cd0817e)
* Berita MIT'[**Mengotomatiskan Analisis Big Data**](https://news.mit.edu/2016/automating-big-data-analysis-1021)
* Saran Teknologi[**10 Pilihan Terbaik untuk Perangkat Lunak Otomatisasi Alur Kerja**](https://technologyadvice.com/blog/information-technology/top-10-workflow-automation-software/)

Sebagai seorang analis data, otomatisasi dapat menghemat banyak waktu dan energi, dan membebaskan Anda untuk lebih fokus pada bagian lain dari proyek Anda. Semakin banyak analisis yang Anda lakukan, semakin banyak cara yang Anda temukan untuk membuat proses Anda lebih sederhana dan efisien.