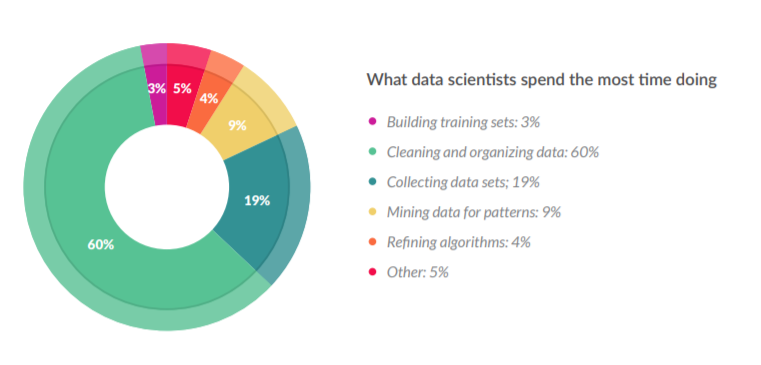
Kita mungkin berpikir pekerjaan data scientist atau machine learning engineer adalah membuat algoritma, mengeksplor data, membuat analisis, dan prediksi. Padahal faktanya, seseorang yang bekerja di bidang data membutuhkan sebagian besar waktunya untuk melakukan proses data cleaning. Hasil penelitian CrowdFlower dalam 2016 Data Science Report menyatakan bahwa 3 dari 5 data scientist yang disurvei menggunakan sebagian besar waktunya untuk membersihkan dan mengatur data.

[](https://www.dicoding.com/academies/184/tutorials/8342?from=11946)

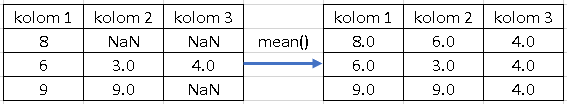
Mengapa data cleaning begitu penting sehingga sebagian besar waktu digunakan untuk menyelesaikan proses ini?

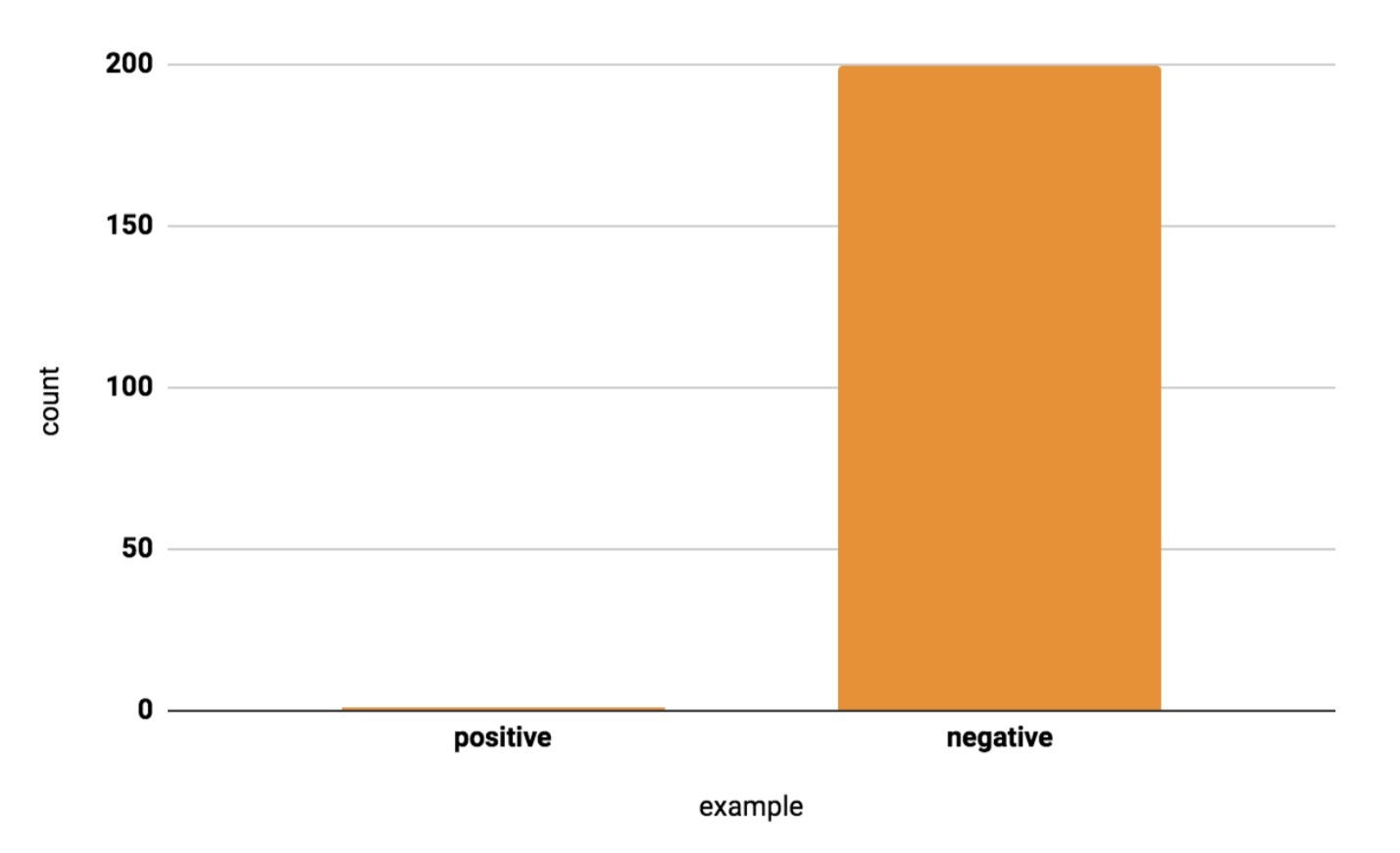
Data cleaning penting sebab proses ini meningkatkan kualitas data yang juga berpengaruh terhadap produktivitas kerja secara keseluruhan. Data yang tidak akurat bisa berpengaruh buruk terhadap akurasi dan performa model. Saat kita melakukan proses data cleaning, kita membuang data dan informasi yang tidak dibutuhkan sehingga kita akan mendapatkan data yang berkualitas. Data yang akurat dan berkualitas akan berpengaruh positif terhadap proses pengambilan keputusan. Pernahkah mendengar ungkapan “Garbage In - Garbage Out?” Dalam konteks machine learning,  jika input yang Anda masukkan itu buruk, sudah pasti hasil olah data Anda pun akan buruk.

Sampai sini paham, ya?

Data cleaning merupakan tahapan penting yang tidak boleh Anda lewatkan. Berikut adalah beberapa hal umum yang harus diperhatikan dalam proses data cleaning:

1. Konsistensi Format  
   Sebuah variabel mungkin tidak memiliki format yang konsisten seperti penulisan tanggal 10-Okt-2020 versus 10/10/20. Format jam yang berbeda seperti 17.10 versus 5.10 pm. Penulisan uang seperti 17000 versus Rp 17.000. Data dengan format berbeda tidak akan bisa diolah oleh model machine learning. Solusinya, format data harus disamakan dan dibuat konsisten terlebih dahulu.
2. Skala Data  
   Jika sebuah variabel memiliki jangka dari 1 sampai 100, pastikan tidak ada data yang lebih dari 100. Untuk data numerik, jika sebuah variabel merupakan bilangan positif, maka pastikan tidak ada bilangan negatif.
3. Duplikasi data  
   Data yang memiliki duplikat akan mempengaruhi model machine learning, apalagi jika data duplikat tersebut besar jumlahnya. Untuk itu kita harus memastikan tidak ada data yang terduplikasi.
4. Missing Value  
   Missing value terjadi ketika data dari sebuah record tidak lengkap. Missing value sangat mempengaruhi performa model machine learning. Ada 2 (dua) opsi untuk mengatasi missing value, yaitu menghilangkan data missing value atau mengganti nilai yang hilang dengan nilai lain, seperti rata-rata dari kolom tersebut (mean) atau nilai yang paling sering muncul (modus), atau nilai tengah (median).

Berikut adalah contoh handling missing value dengan nilai rata-rata (mean).  
[](https://www.dicoding.com/academies/184/tutorials/8342?from=11946)

1. Skewness Distribution  
   Skewness adalah kondisi di mana dataset cenderung memiliki distribusi data yang tidak seimbang. Skewness akan mempengaruhi data dengan menciptakan bias terhadap model. Apa itu bias? Sebuah model cenderung memprediksi sesuatu karena ia lebih sering mempelajari hal tersebut. Misalkan ada sebuah model untuk pengenalan buah di mana jumlah jeruk 92 buah dan apel 8 buah. Distribusi yang tidak imbang ini akan mengakibatkan model lebih cenderung memprediksi jeruk daripada apel.  
   [](https://www.dicoding.com/academies/184/tutorials/8342?from=11946)  
   Cara paling simpel untuk mengatasi skewness adalah dengan menyamakan proporsi kelas mayoritas dengan kelas minoritas. Untuk teknik lebih lanjut dalam mengatasi skewness atau imbalance data, Anda bisa membacanya di [tautan](https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data) ini.

Jika Anda ingin membaca lebih lanjut tentang proses data cleaning, silakan buka tautan [ini](https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4) atau tautan [berikut](https://www.kdnuggets.com/2019/06/7-steps-mastering-data-preparation-python.html).