**LAPORAN PROYEK SAIN DATA**

***“Survey Lungs cancer”***

**Dosen Pengampu : EKA MALA SARI ROCHMAN, S.Kom., M.Kom.**



Disusun oleh:

Asri Erawati (210411100152)

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOYO MADURA**

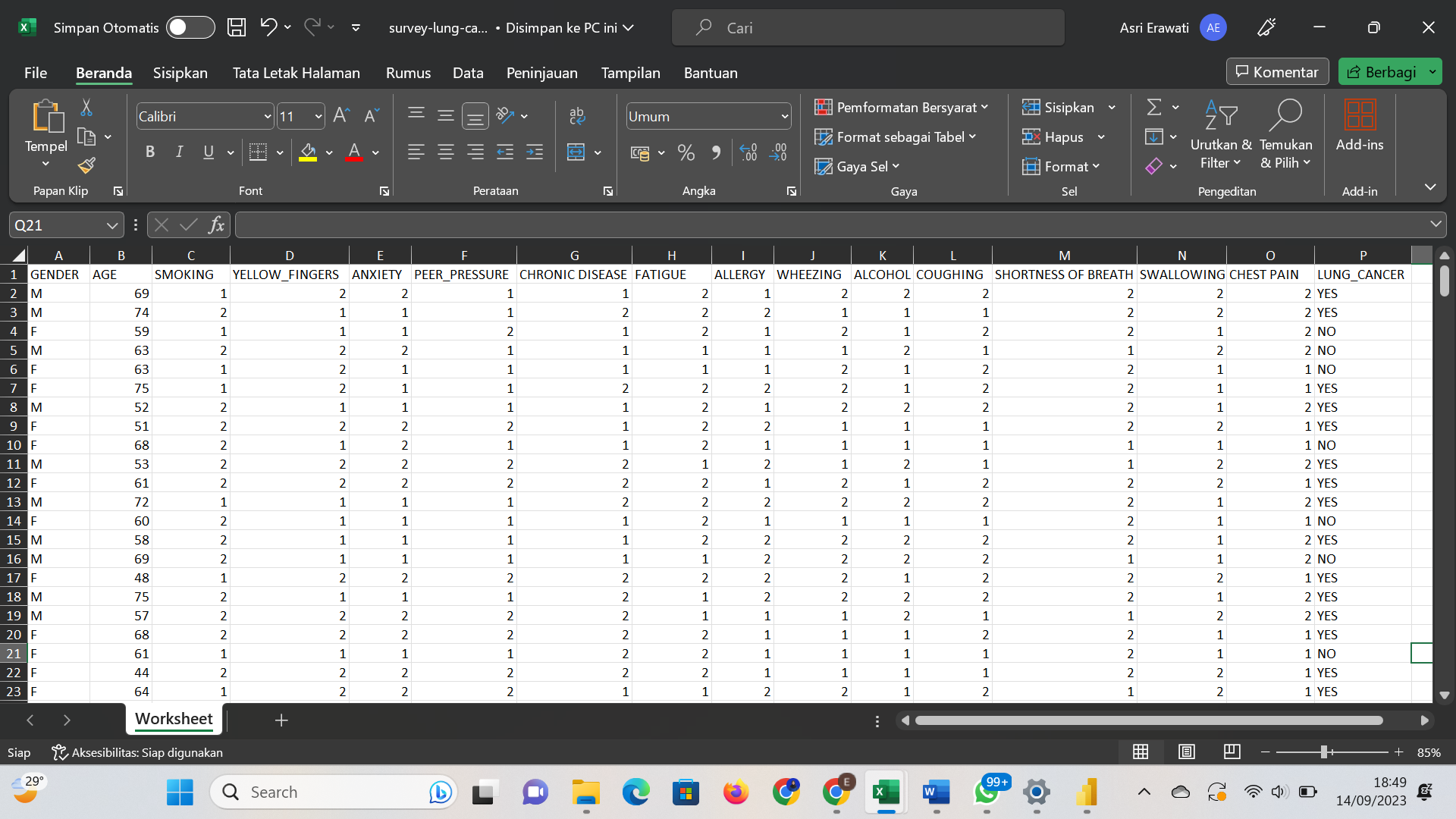
**2023**

1. Data

Data yang digunakan : <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

Data yang digunakan pada penelitian ini berupa dataset *Survey Lung Cancer* yang diperoleh dari website www.kaggle.com yang berupa file berbentuk csv. Pada data ini memiliki 16 kolom dan 309 baris, dimana untuk kolom menyatakan variabel dan untuk baris menyatakan jumlah dari data. Dari 16 kolom, terdapat satu data yang memiliki tipe numerik yaitu pada kolom AGE, sedangkan untuk 15 kolom lainnya memiliki tipe kategorik yang masing masing data memiliki 2 kategori yaitu 1 untuk NO dan 2 untuk YES. Untuk targetnya yaitu mengidap kanker paru (YES) dan tidak mengidap kanker paru (NO).

Sampel dataset



Analisis parameter

1. Gender (jenis kelamin)

Kanker paru-paru ini sebagian besar diderita oleh laki-laki dikarenakan biasanya laki-laki lebih cenderung mengkonsumsi rokok daripada perempuan, yangmana merokok ini juga merupakan salah satu penyebab terkenanya penyakit kanker paru-paru. Laki-laki juga cenderung lebih sering berada di luar rumah dibandingkan dengan Perempuan, oleh karena itu resiko laki-laki terpapar polusi juga lebih besar. Selain itu, perbedaan faktor genetik antara laki-laki dan perempuan juga dapat memengaruhi cara tubuh untuk merespon paparan zat penyebab kanker. Oleh karena itu, parameter gender ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Age (umur pasien)

Seiring dengan bertambahnya usia, maka sistem kekebalan tubuh juga akan mengalami penurunan fungsi sehingga tubuh kurang efektif dalam melakukan deteksi dan melakukan penghancuran sel-sel kanker yang muncul. Oleh karena itu, parameter Age ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Smoking (status merokok)

Salah satu faktor utama kanker adalah karena merokok. Selain dengan merokok langsung, asap rokok yang terhirup dari orang lain juga dapat meningkatkan resiko terkena penyakit kanker paru-paru. Rokok ini memiliki banyak kandungan karsinogen yang dapat menyebabkan kanker paru-paru. Oleh karena itu, parameter smoking ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Yellow finger (jari menguning)

Untuk jari menguning yang dapat dijadikan sebagai tanda terkenanya penyakit kanker paru-paru tidaklah akurat, karena kanker paru-paru sendiri adalah penyakit yang secara utama disebabkan oleh paparan karsinogen yang ditemukan dalam asap rokok, dan untuk jari menguning biasanya bukan menjadi suatu indikator yang menandakan terkenanya penyakit kanker paru-paru. Oleh karena itu, parameter yellow finger ini tidak dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru dan akan dihapus.

1. Anxiety (gangguan kecemasan)

Untuk anxiety ini tidak secara langsung menyebabkan penyakit kanker paru-paru, karena penyebab utama kanker paru-paru adalah merokok. Anxiety sebenarnya dapat mempengaruhi resiko kanker paru-paru, namun secara tidak langsung, seperti karena sedang merasa cemas dan melampiaskan dengan merokok. Sehingga, untuk anxiety ini tidak perlu digunakan untuk mendeteksi penyakit kanker paru-paru sehingga dapat dihapus.

1. Peer Pressure (tekanan lingkungan)

Salah satu penyebab penyakit kanker paru-paru adalah karena polusi udara, entah dari debu, asap rokok, dll. Jika seseorang berada dalam lingkungan yang seperti ini, maka kemungkinan untuk terkena penyakit kanker paru-paru semakin besar. Sehingga, untuk parameter peer pressure ini dapat digunakan untuk melakukan deteksi penyakit kanker paru-paru.

1. Chromic disease (penyakit kronis)

Beberapa penyakit kronis telah terbukti dapat meningkatkan risiko penyakit kanker paru-paru, misalnya penyakit paru obstruktif kronis (PPOK) dan penyakit paru kronis lainnya dapat merusak jaringan paru-paru dan meningkatkan risiko perkembangan kanker paru-paru. Sehingga, untuk parameter chromic disease ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Fatigue (kelelahan)

Apabila seseorang terkena penyakit kanker paru-paru, biasanya beberapa orang akan mengalami gejala kelelahan ini. Walaupun sedang tidak melakukan aktifitas apapun, biasanya penderita akan mengalami kelelahan. Sehingga, untuk parameter atribut fatigue ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Allergy (memiliki alergi)

Alergi adalah respons sistem kekebalan tubuh terhadap zat-zat yang biasanya tidak berbahaya, seperti serbuk sari, bulu hewan, atau alergen lainnya. Penyakit kanker paru-paru, di sisi lain adalah penyakit yang melibatkan pertumbuhan sel-sel kanker secara tidak terkendali di paru-paru. Kedua kondisi ini berbeda dan biasanya tidak ada hubungan langsung antara mereka. Sehingga untuk parameter allergy ini tidak perlu digunakan untuk mendeteksi penyakit kanker paru-paru sehingga dapat dihapus.

1. Wheezing (napas berbunyi)

Alasan utama yang menyebabkan kanker paru-paru yaitu karena merokok. Biasanya, perokok yang sudah sangat parah, saat bernafas dapat mengeluarkan suara seperti suara sesak. Sehingga untuk parameter wheezing ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Alcohol (konsumsi alkohol)

Selain merokok, faktor yang dapat menyebabkan kanker paru-paru yaitu karena mengkonsumsi alkohol. Alkohol bisa meningkatkan risiko kanker karena kandungan etanol. Etanol di dalam alkohol ini mudah merusak sel dengan mengikat DNA dan menyebabkan sel bereplikasi dengan tidak benar. Etanol nantinya juga akan memengaruhi kadar hormon yang dapat mengubah cara sel tumbuh dan membelah diri. Lalu, kandungan ini dapat menyebabkan kerusakan jaringan secara langsung dan meningkatkan penyerapan karsinogen lainnya. Sehingga untuk atribut alkohol ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Coughing(batuk)

Batuk yang tak kunjung sembuh dapat menjadi salah satu parameter yang dapat digunakan untuk mendeteksi penyakit kanker paru-paru. Seseorang bisa muncul gejala batuk ini dikarenakan kebiasaan merokok, atau terkena polusi yangmana hal ini merupakan hal yang menjadi penyebab terkena kanker paru-paru. Sehingga untuk parameter coughing ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Shortness of breath (napas pendek/tersengal)

Ketika seseorang mengidap penyakit kanker paru-paru, biasanya saluran pernapasannya terganggu, sehingga dapat mengalami napas tersengal ini. Sehingga untuk parameter shortness of breath ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Swallowing difficulty (kesulitan menelan)

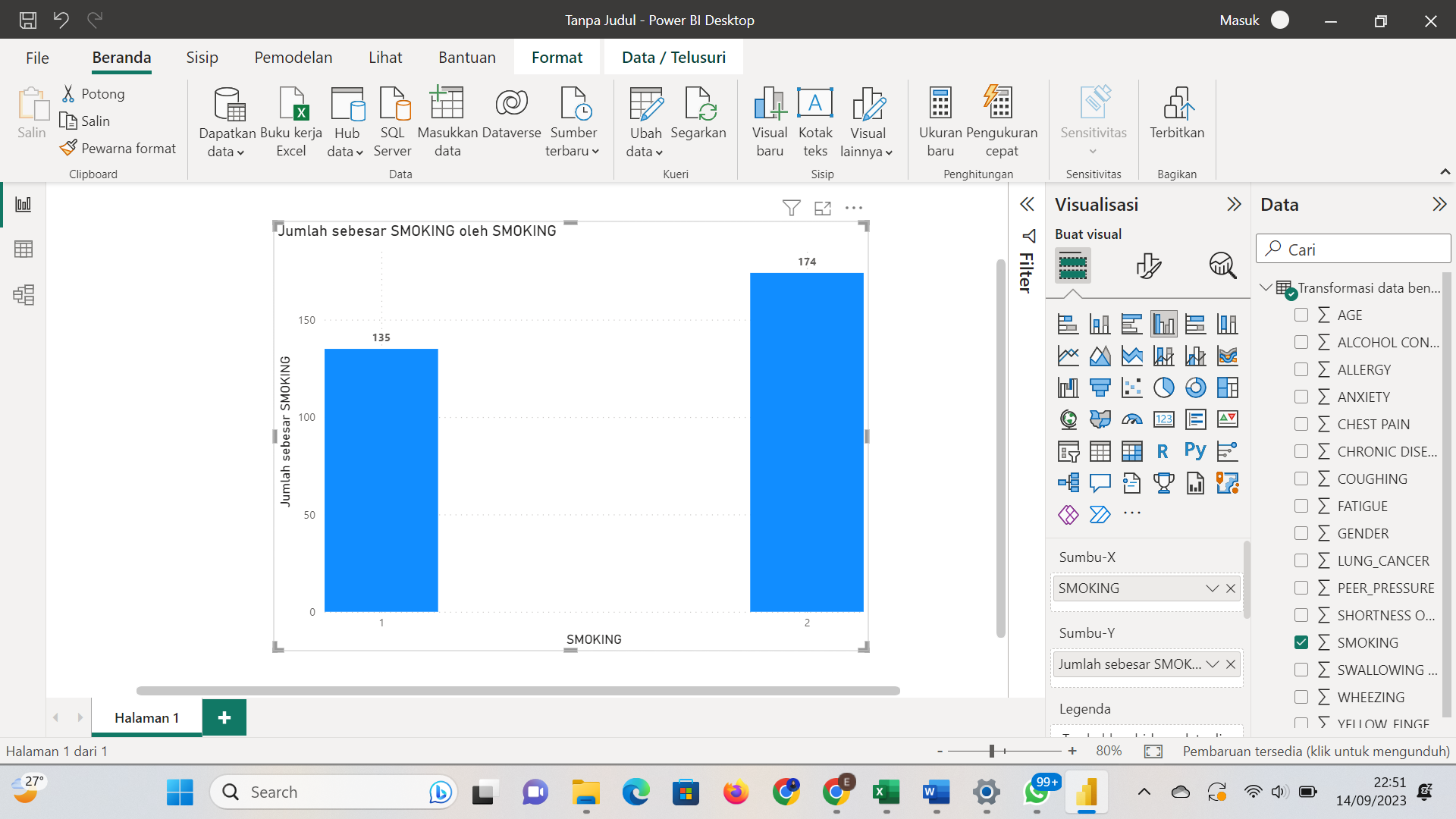
Kanker paru-paru yang telah tumbuh dengan ukuran yang signifikan dalam dada dapat mengubah anatomi dada dan memengaruhi struktur di sekitarnya, termasuk organ-organ pernapasan dan saluran pencernaan. Hal ini bisa mengganggu fungsi esofagus dan menyebabkan kesulitan menelan. Sehingga untuk parameter swallowing difficulty ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

1. Chest pain (sakit dada)

Dalam beberapa kasus, kanker paru-paru yang sudah mencapai stadium lanjut dapat menyebar (metastasis) ke organ-organ di dada, seperti tulang rusuk atau tulang belikat. Metastasis ini dapat menyebabkan rasa sakit atau ketidaknyamanan di daerah dada. Selain itu, kanker paru-paru dapat menyebabkan peradangan di jaringan di sekitarnya, termasuk pleura (lapisan tipis yang melapisi paru-paru dan dinding dada). Peradangan ini bisa menyebabkan sakit dada, yang dikenal sebagai pleuritis. Sehingga untuk atribut chest pain ini dibutuhkan untuk melakukan prediksi penyakit kanker paru-paru.

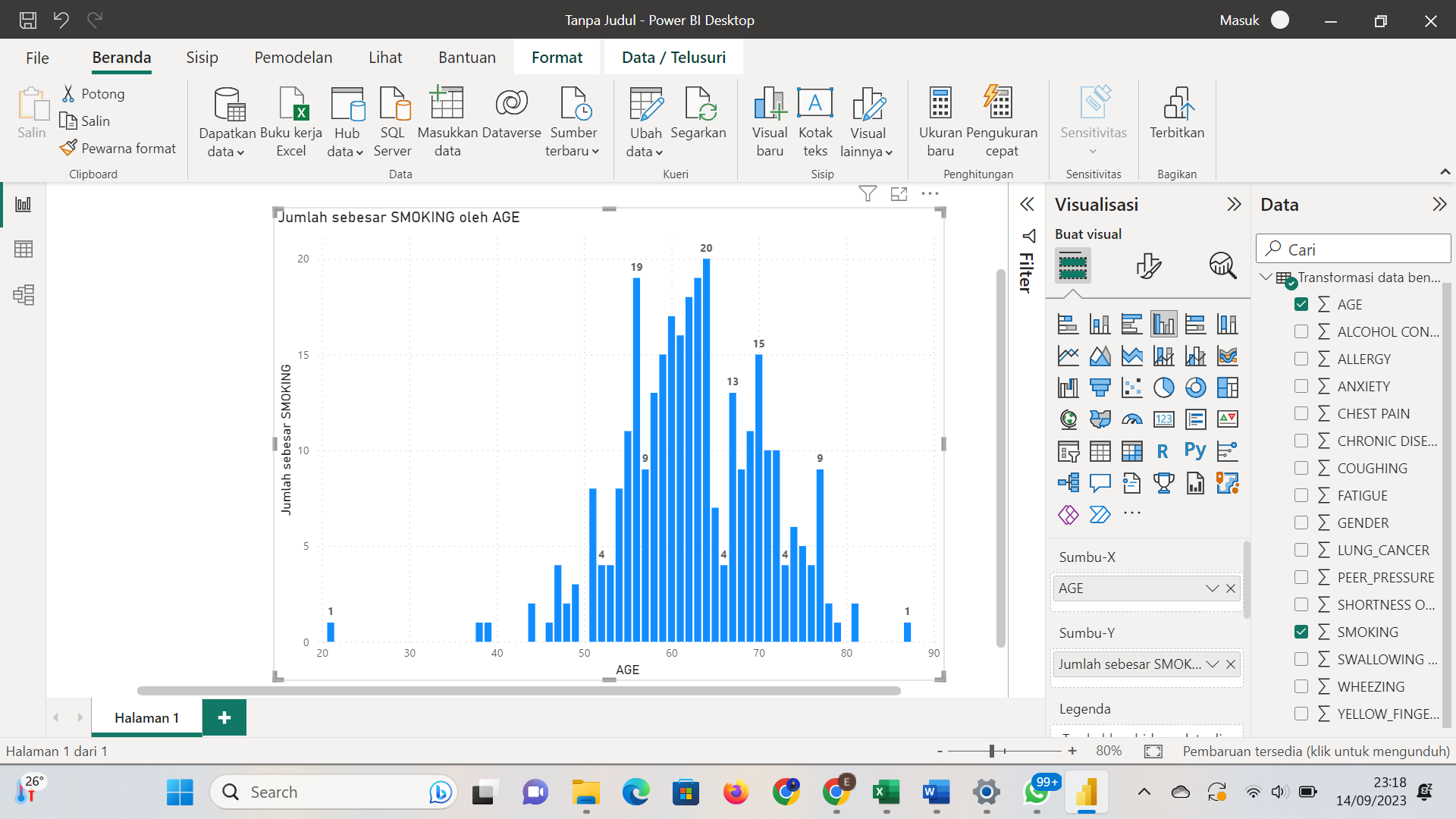
Analisis sebaran data

* Sebaran data berdasarkan atribut smoking

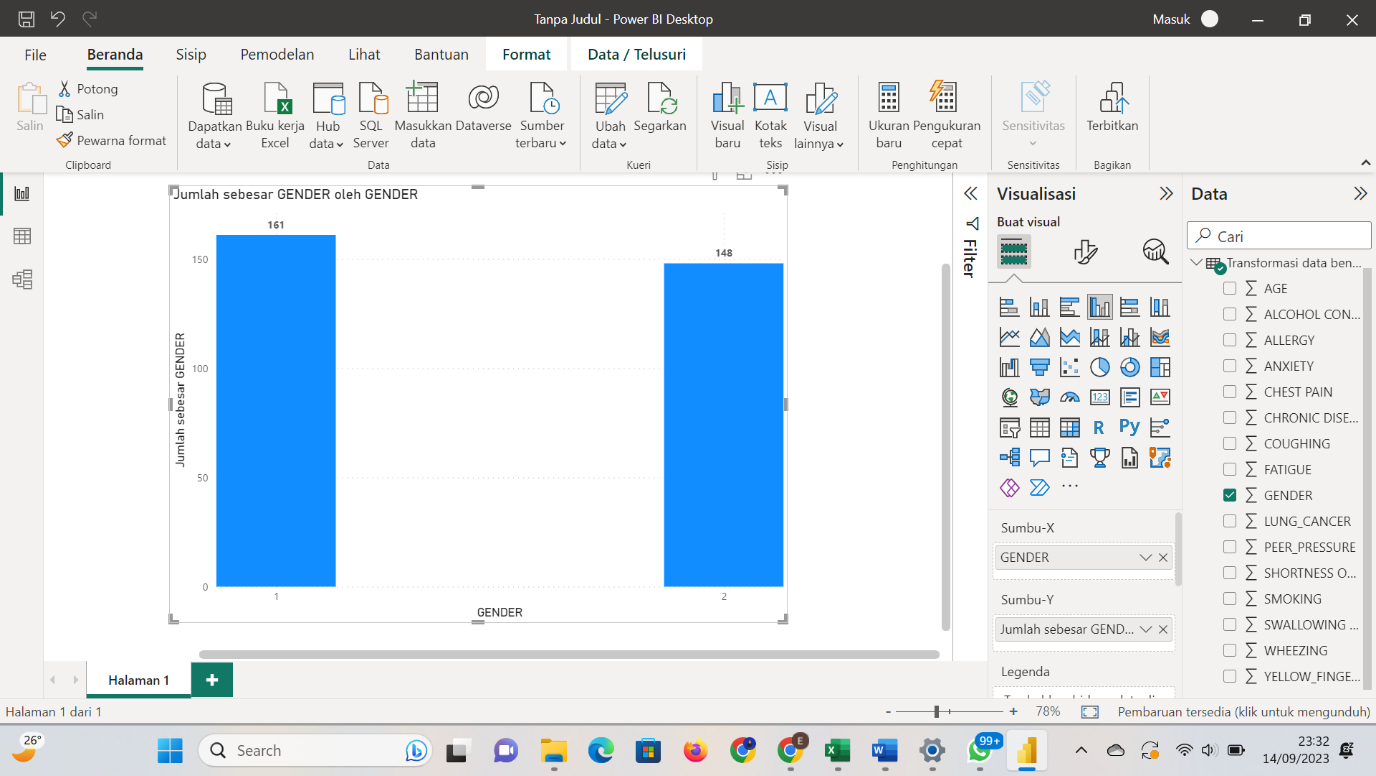


Dari analisis sebaran data pada atribut smoking. Untuk angka 1 berarti tidak merokok dan untuk angka 2 berarti merokok, berarti terdapat 135 orang yang tidak merokok dan 174 orang merokok.

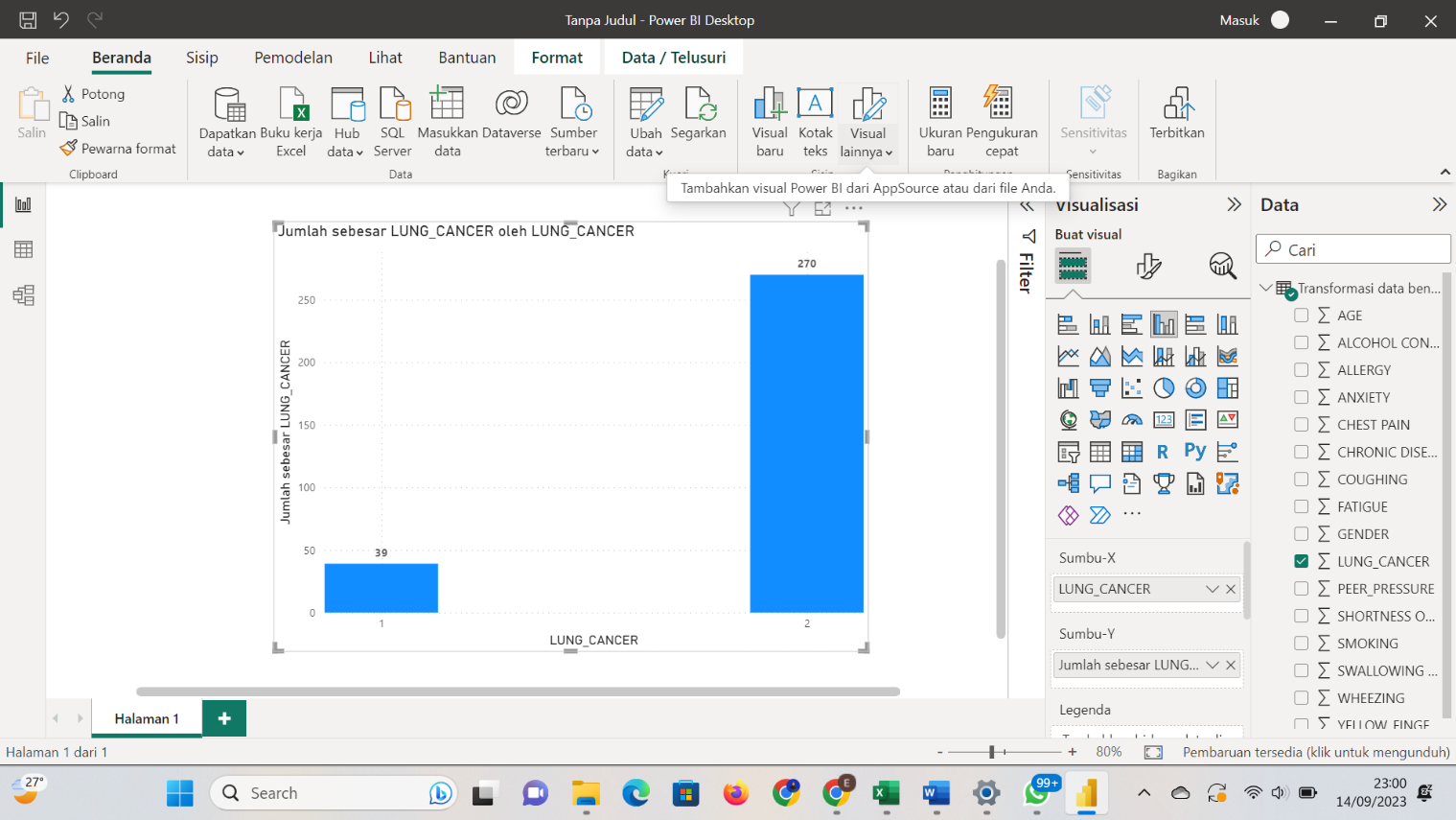
* Sebaran data berdasarkan atribut age dan smoking



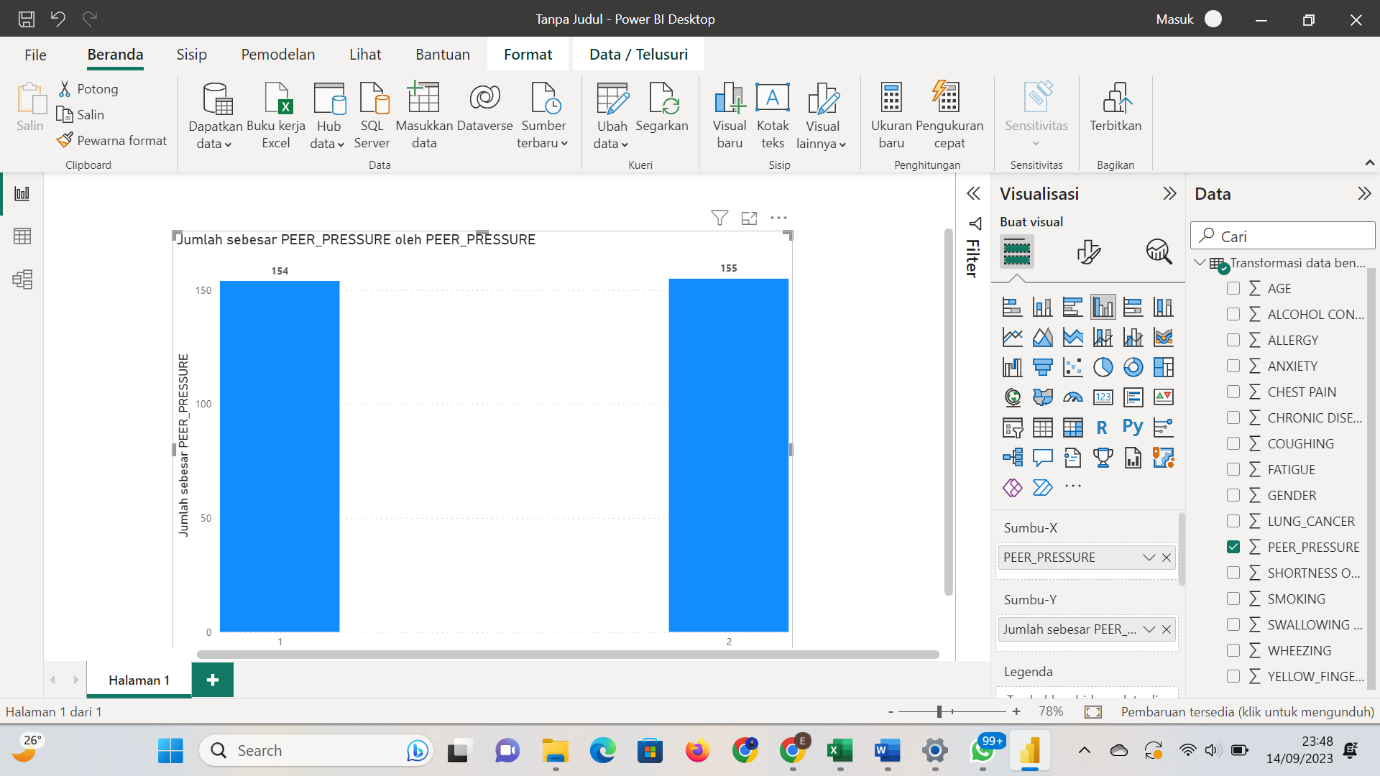
Dari analisis sebaran data pada atribut age dan smoking diatas, usia paling banyak merokok yaitu pada usia 64 tahun, kemudian dibawahnya terdapat pada usia 56 tahun. Rata-rata, banyak orang yang merokok pada usia sekitar 50 – 70 an tahun atau pada usia lansia.

* Sebaran data berdasarkan atribut gender

Dari analisis sebaran data pada atribut gender. Untuk angka 1 berarti laki-laki dan untuk angka 2 berarti perempuan, berarti dalam responden terdapat 161 orang yang berjenis kelamin laki-laki dan 148 orang berjenis kelamin Perempuan.

* Sebaran data berdasarkan atribut lung cancer

Dari analisis sebaran data pada atribut lung cancer. Untuk angka 1 berarti negatif dan untuk angka 2 berarti positif, berarti terdapat 39 orang yang negatif kanker paru-paru dan 270 orang yang positif kanker paru-paru.

* Sebaran data berdasarkan atribut peer pressure

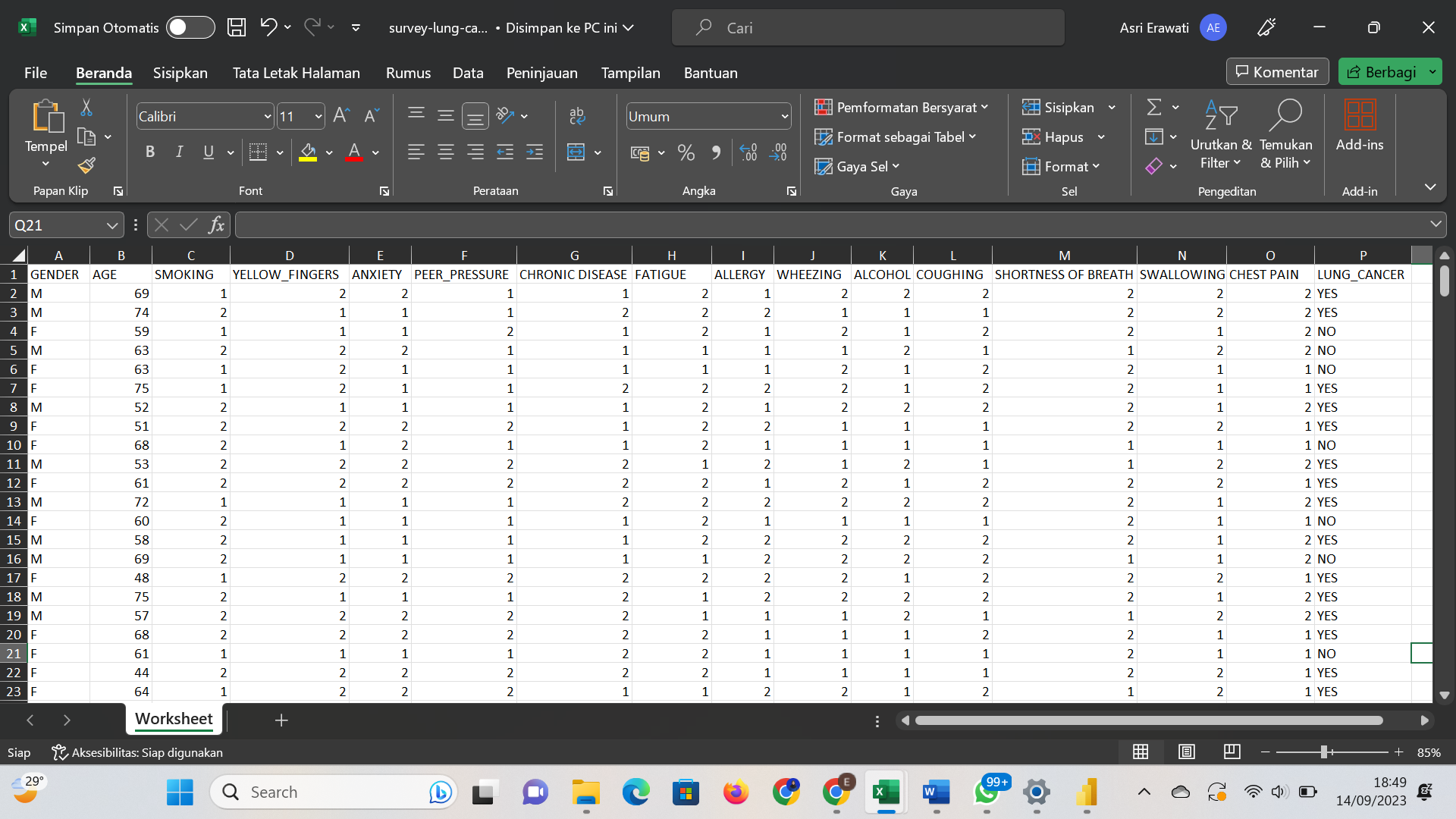
Dari analisis sebaran data pada atribut peer pressure. Untuk angka 1 berarti no dan untuk angka 2 berarti yes, berarti terdapat 154 orang yang tidak terpengaruh dari lingkungan dan 155 orang yang terpengaruh dari lingkungan.

1. Data Cleaning

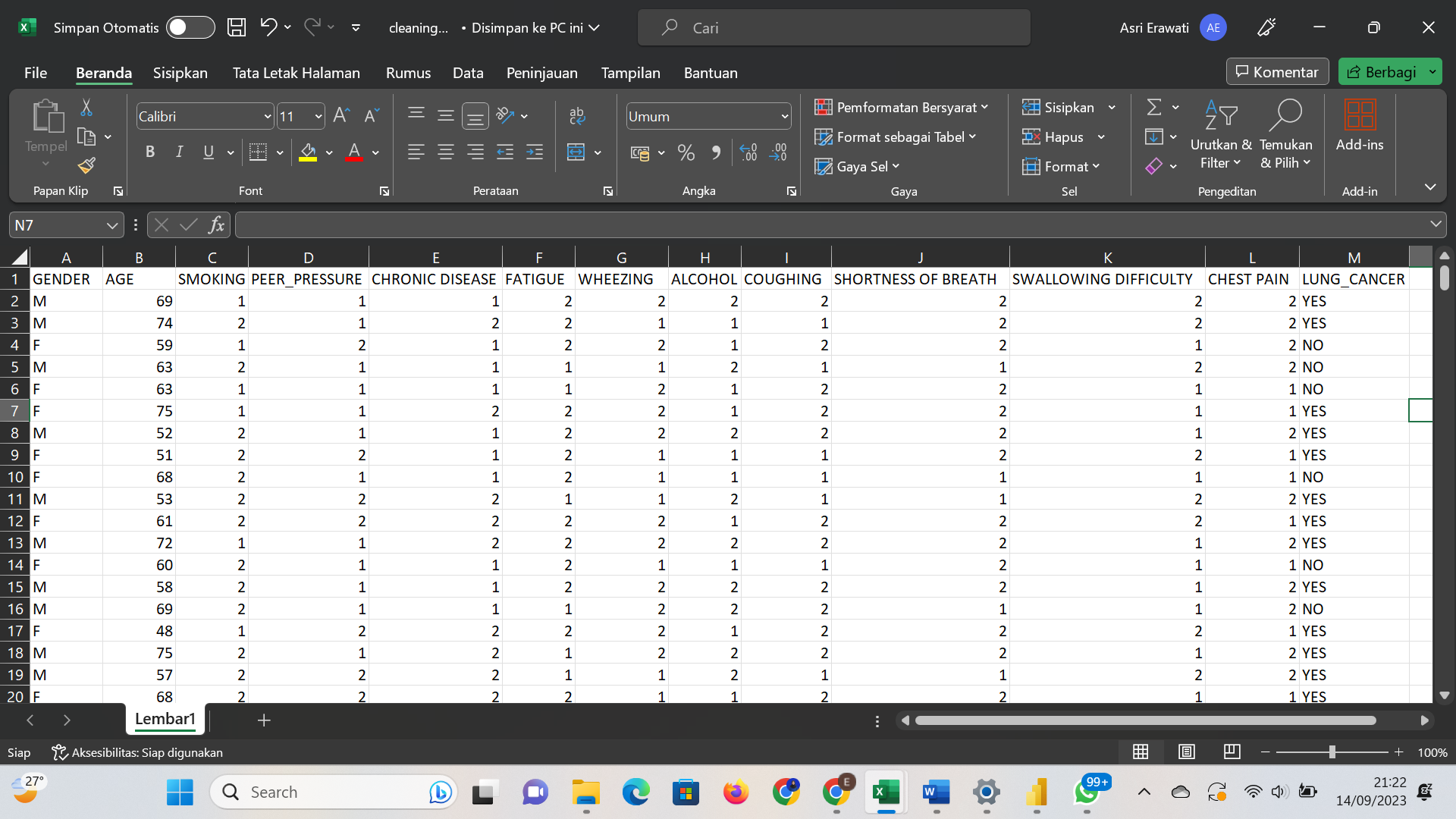
Data cleaning merupakan tahapan persiapan data untuk dilakukan analisis dengan cara menghapus atau memodifikasi data yang tidak benar, tidak lengkap, tidak relevan, adanya duplikasi data, atau dalam format yang tidak benar. Tujuan dari proses ini adalah untuk memastikan bahwa data yang digunakan dalam melakukan analisis lebih akurat dan konsisten sehingga didapatkan sebuah data yang berkualitas.

Pada dataset *Survey Lung Cancer* terdapat beberapa atribut yang tidak perlu digunakan dalam melakukan deteksi penyakit kanker paru-paru yaitu pada atribut anxiety, yellow finger, dan allergy sehingga bisa dihapus.

Sampel atribut sebelum dihapus



Sampel atribut sesudah dihapus

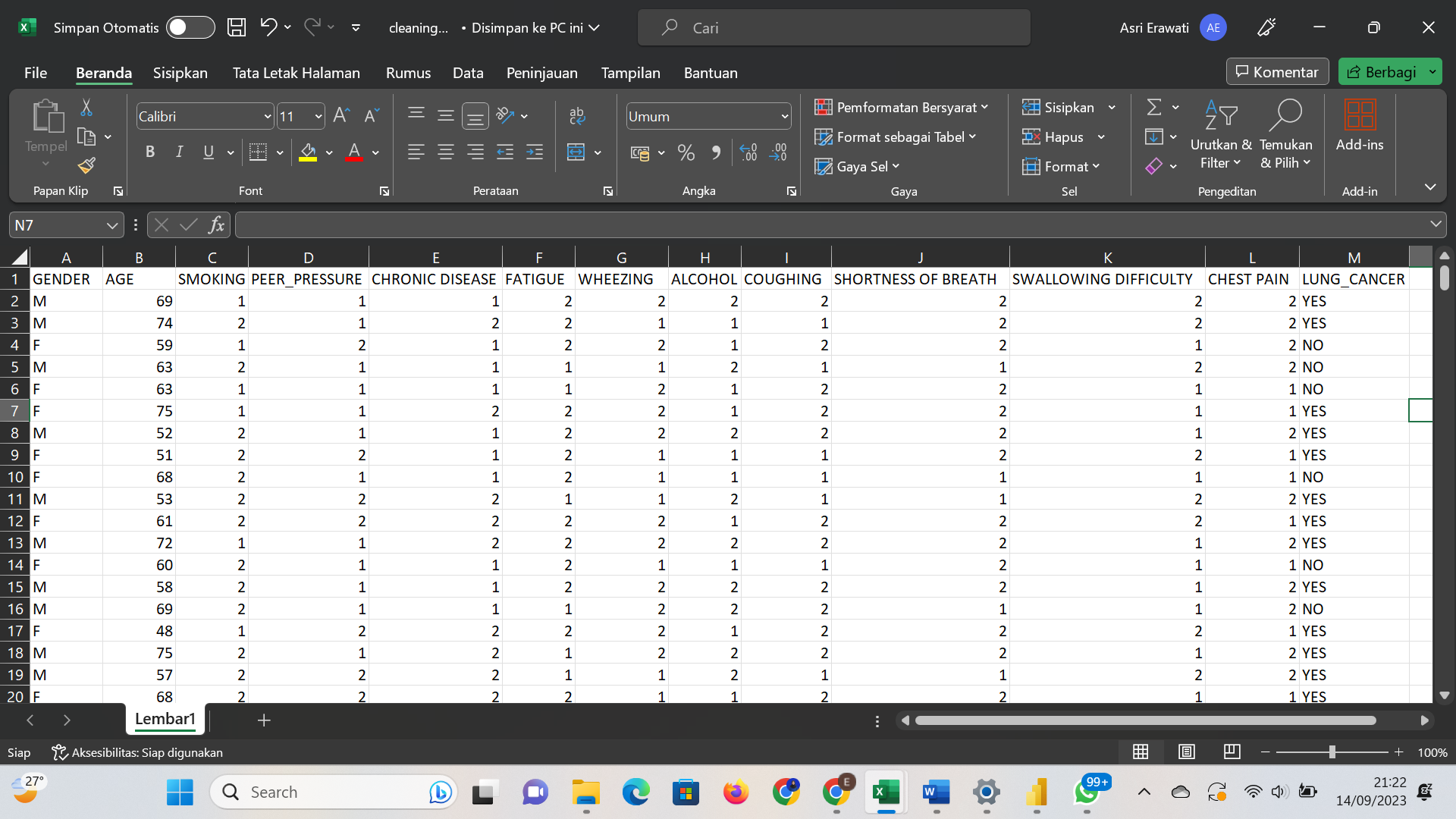


1. Transformasi data

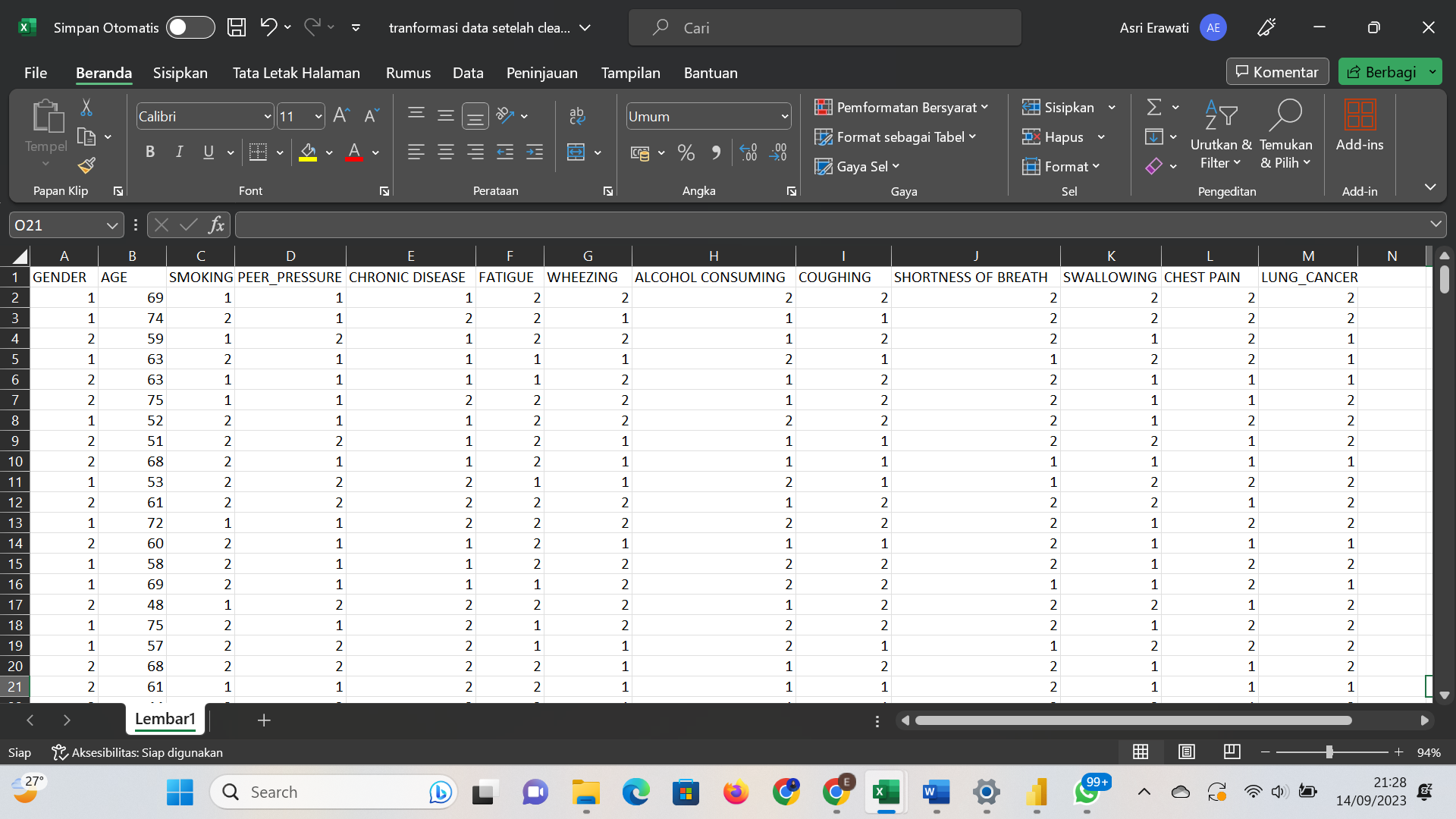
Transformasi data adalah proses mentransformasi atau mengubah data ke dalam bentuk yang sesuai, agar dapat diproses dengan lebih mudah. Biasanya data diubah dari bentuk teks ke bentuk numerik atau biner.

Pada dataset *Survey Lung Cancer,* terdapat 2 atribut yang masih berupa teks, sehingga kedua atribut ini dapat dilakukan transformasi data dengan cara merubah teks ini menjadi numerik.

Data sebelum dilakukan transformasi



Data sesudah dilakukan transformasi



1. Missing data

Missing data merupakan hilangnya sebagian informasi pada suatu dataset. Terdapat beberapa hal yang dapat mempengaruhi missing data yaitu seperti human error. Akibat dari missing data ini yaitu pendugaan parameter menjadi tidak efisien. Ukuran dari data yang berkurang dapat menyebabkan sulitnya melakukan analisis, sehingga hasil yang didapatkan menjadi tidak valid. Missing data dapat diabaikan apabila jumlahnya sedikit, dan tidak dapat diabaikan apabila jumlahnya banyak. Berikut cara mengatasi missing data:

1. Dapat diisi dengan menggunakan nilai 1 atau 0
2. Dapat mengisi dengan nilai mean/modus
3. Dapat mengisi dengan menggunakan metode machine learning

Pada dataset *Survey Lung Cancer,* tidak memiliki data kosong atau hilang sehingga tidak perlu dilakukan perbaikan untuk missing data ini.

1. Balancing data

Cara mengatasi balancing data:

1. Oversampling : yaitu dengan duplikasi atau penambahan ulang sampel dari minoritas ke kelas mayoritas sehingga jumlahnya menjadi seimbang.
2. Undersampling : yaitu dengan melibatkan pengapusan atau pengurangan sampel dari kelas mayoritas sehingga jumlahnya menjadi seimbang dengan kelas minoritas.
3. SMOTE (synthetic minority over-sampling technique) : Teknik SMOTE merupakan Teknik yang paling banyak digunakan untuk mengatasi data yang tidak seimbang. Tekniknya mirip dengan oversampling yang menduplikasi data dari kelas minoritas sehingga jumlahnya sama dengan jumlah kelas mayoritas. Tetapi, Teknik SMOTE tidak hanya dengan melakukan duplikasi data yang sama, tetapi akan membuat sampel baru yang menyerupai data asli dari kelas minoritas sehingga kelas minoritas menjadi jauh lebih beragam.