**LAPORAN TUGAS**

**ANALISIS DATA PENYAKIT ANEMIA**

**MENGGUNAKAN POWER BI**

****

**Dosen Pengampu:**

Eka Mala Sari Rochman, S.Kom., M.Kom.

**Disusun oleh:**

Rizki Ardian Samudra (210411100179)

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2023**

**PENDAHULUAN**

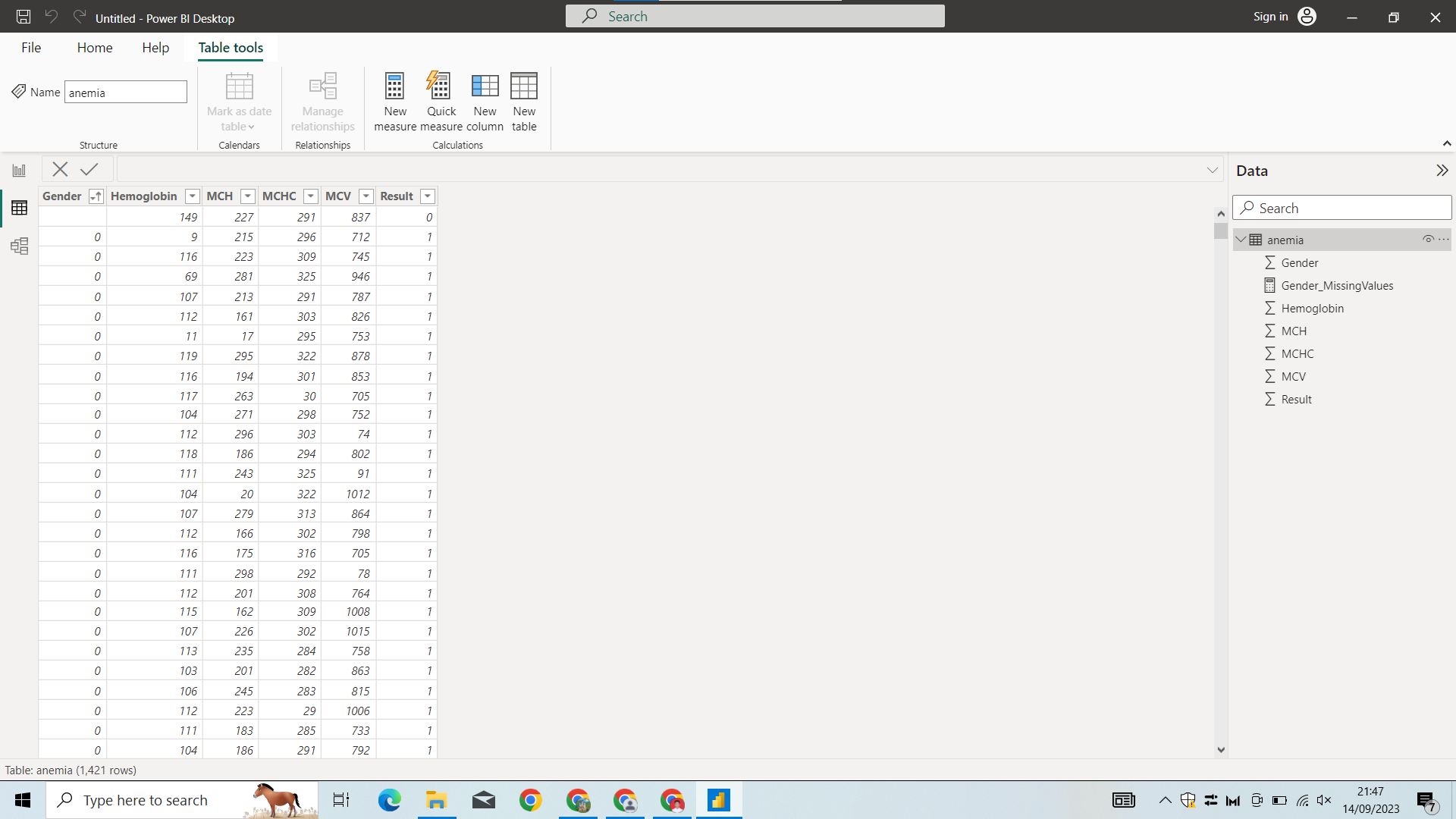
Anemia adalah kondisi medis yang ditandai oleh jumlah sel darah merah atau kadar hemoglobin yang rendah dalam tubuh, yang dapat mengakibatkan berbagai gejala seperti kelelahan, lemah, pusing, dan masalah kesehatan lainnya. Untuk mendiagnosis anemia, dokter melakukan tes darah untuk mengukur tingkat hemoglobin dan beberapa parameter lainnya dalam sel darah merah. Dataset Anemia ini dikumpulkan untuk membantu dalam memprediksi apakah seorang pasien berpotensi menderita anemia berdasarkan beberapa atribut yang telah diukur.

1. Jenis Kelamin (Gender): Atribut ini mengacu pada jenis kelamin pasien, dengan nilai 0 untuk laki-laki dan 1 untuk perempuan. Jenis kelamin dapat menjadi faktor yang relevan dalam analisis anemia karena prevalensi anemia dapat berbeda antara laki-laki dan perempuan.
2. Hemoglobin: Hemoglobin adalah protein yang terdapat dalam sel darah merah yang berperan penting dalam mengangkut oksigen dari paru-paru ke seluruh tubuh dan membawa karbon dioksida dari seluruh tubuh kembali ke paru-paru. Kadar hemoglobin yang rendah dapat menjadi indikasi adanya anemia.
3. MCH (Mean Corpuscular Hemoglobin): MCH adalah ukuran rata-rata jumlah hemoglobin yang terkandung dalam setiap sel darah merah. Informasi ini dapat memberikan wawasan tentang kualitas sel darah merah dan kemampuannya dalam membawa oksigen.
4. MCHC (Mean Corpuscular Hemoglobin Concentration): MCHC adalah ukuran konsentrasi rata-rata hemoglobin dalam satu sel darah merah. Hal ini juga berkaitan dengan kualitas sel darah merah dan dapat memberikan petunjuk tentang kemungkinan masalah seperti anemia.
5. MCV (Mean Corpuscular Volume): MCV mengukur volume rata-rata dari sel darah merah. Perubahan dalam ukuran sel darah merah dapat mengindikasikan adanya masalah dalam produksi darah dan dapat menjadi petunjuk dalam diagnosis anemia.
6. Hasil (Results): Atribut ini adalah hasil akhir dari prediksi atau diagnosis. Nilai 0 mungkin menunjukkan bahwa pasien tidak menderita anemia, sedangkan nilai 1 mungkin menunjukkan adanya anemia.

Dataset ini dapat digunakan untuk mengembangkan model prediksi yang dapat membantu dokter dalam mendiagnosis anemia lebih awal atau memantau perkembangan pasien yang sudah didiagnosis dengan anemia. Dengan memahami hubungan antara atribut-atribut ini dan hasil anemia, kita dapat mengidentifikasi faktor-faktor risiko yang berkontribusi terhadap kondisi ini dan meningkatkan upaya pencegahan serta pengobatan anemia.

**ANALISIS DAN PEMBAHASAN**

1. **Siapkan Data Anemia**

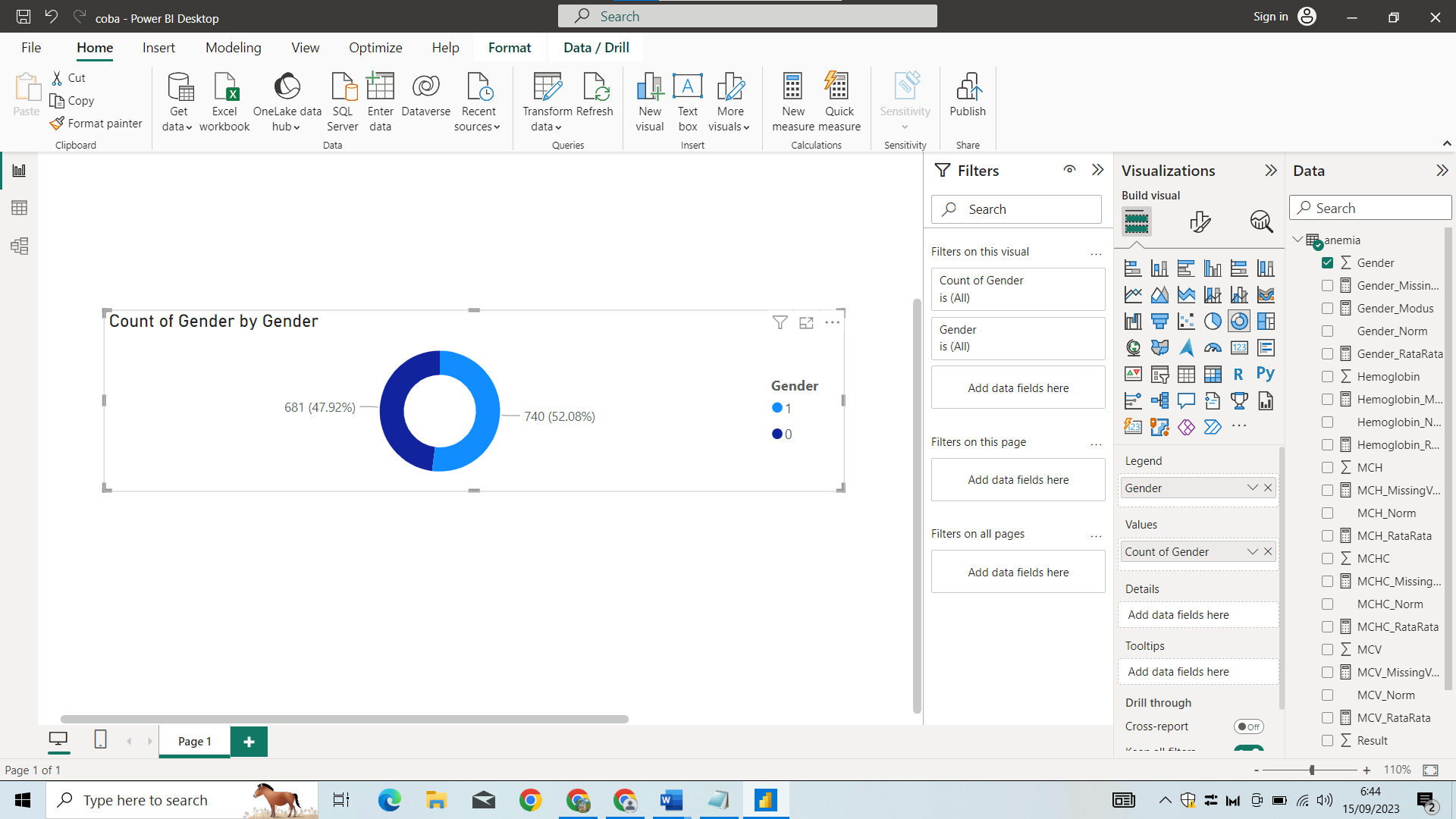


Langkah awal dalam penelitian ini adalah menyiapkan dataset anemia yang akan digunakan untuk analisis. Dataset ini dapat diperoleh dari berbagai sumber, salah satunya adalah platform Kaggle, yang merupakan salah satu sumber terkemuka untuk dataset dan kompetisi ilmiah. Kaggle menyediakan beragam dataset kesehatan dan medis, termasuk dataset terkait anemia, yang dapat diunduh dan digunakan untuk penelitian dan analisis lebih lanjut. Setelah dataset anemia diperoleh dari Kaggle, langkah selanjutnya adalah melakukan pembersihan dan eksplorasi data, tahap analisis, pengolahan data, dan pengembangan model prediksi untuk memahami lebih dalam faktor-faktor yang berkontribusi terhadap anemia.

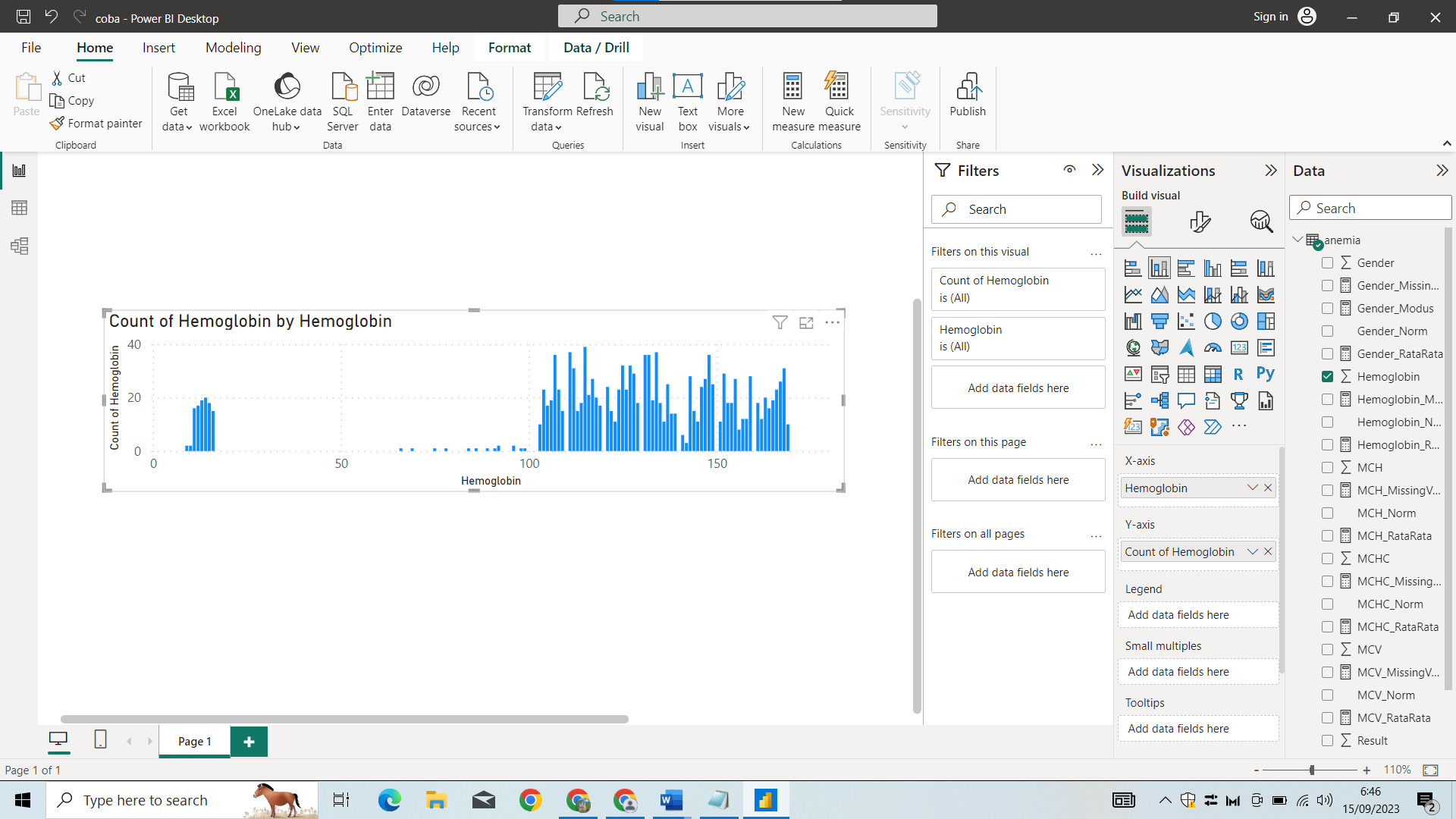
1. **Visualisasi Dataset Anemia**

Berikut ini merupakan atribut-atribut yang ada pada dataset anemia:

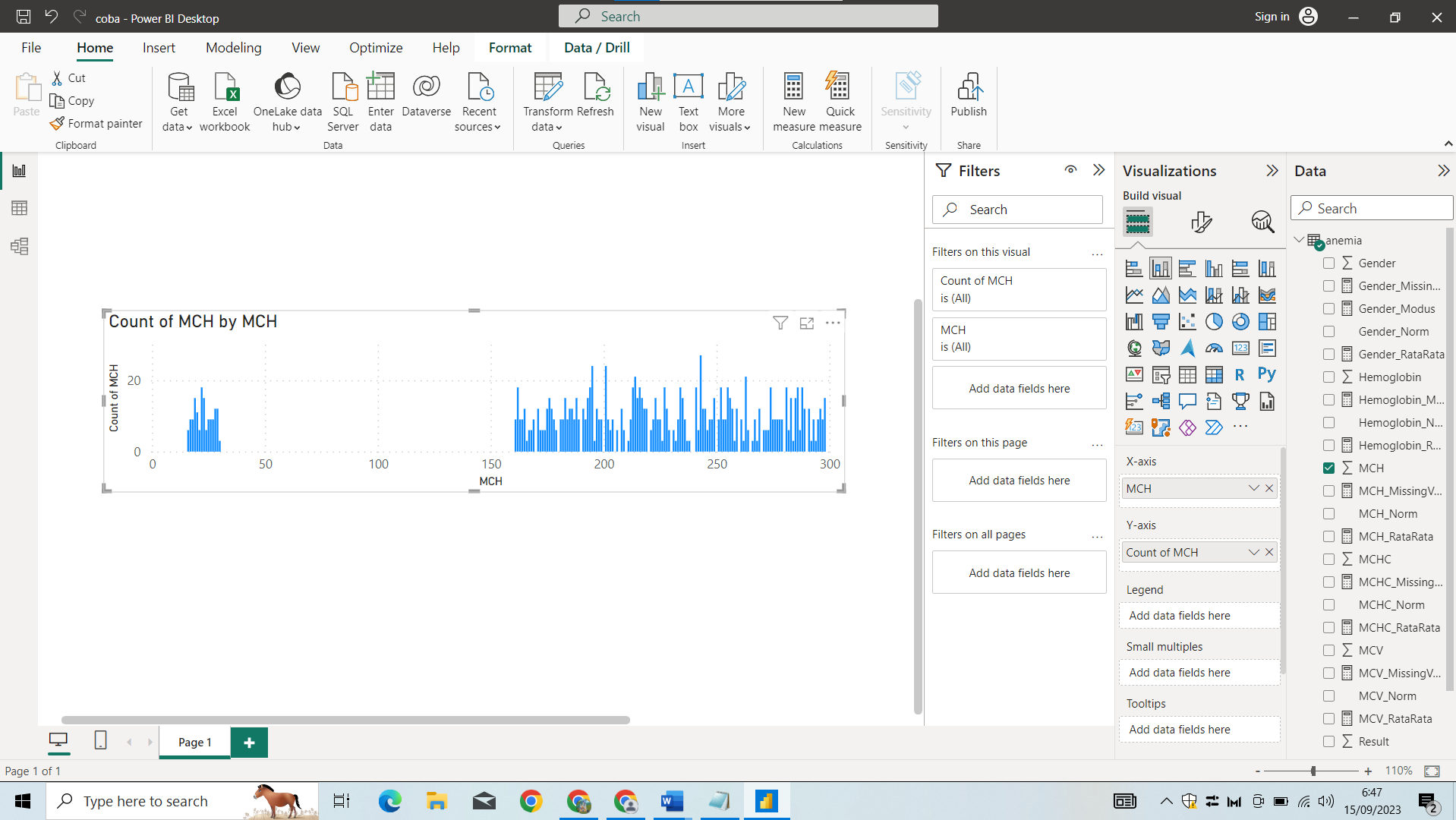
1. Jenis Kelamin (Gender): Atribut ini mengacu pada jenis kelamin pasien, dengan nilai 0 untuk laki-laki dan 1 untuk perempuan. Jenis kelamin dapat menjadi faktor yang relevan dalam analisis anemia karena prevalensi anemia dapat berbeda antara laki-laki dan perempuan.



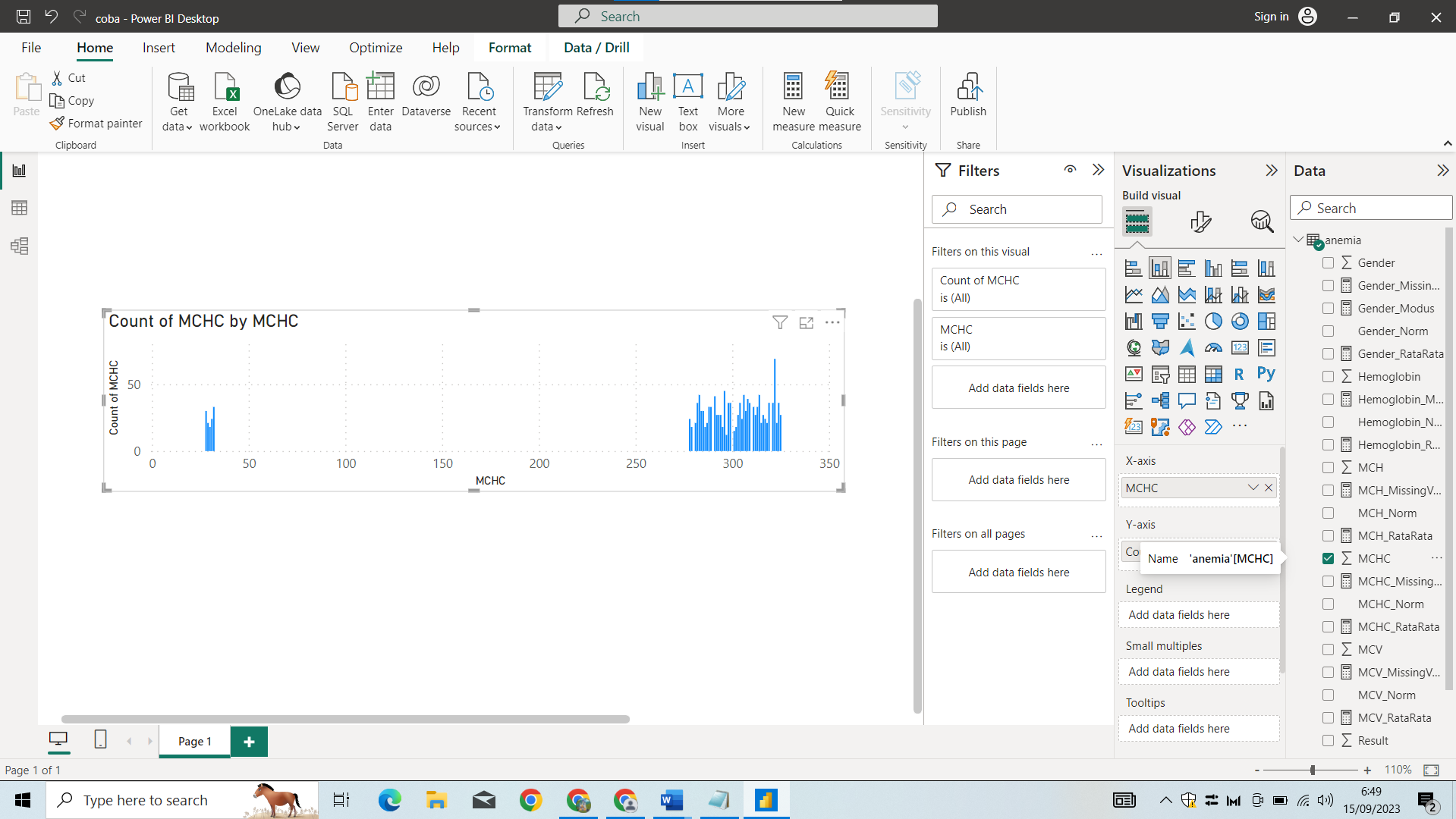
1. Hemoglobin: Hemoglobin adalah protein yang terdapat dalam sel darah merah yang berperan penting dalam mengangkut oksigen dari paru-paru ke seluruh tubuh dan membawa karbon dioksida dari seluruh tubuh kembali ke paru-paru. Kadar hemoglobin yang rendah dapat menjadi indikasi adanya anemia.



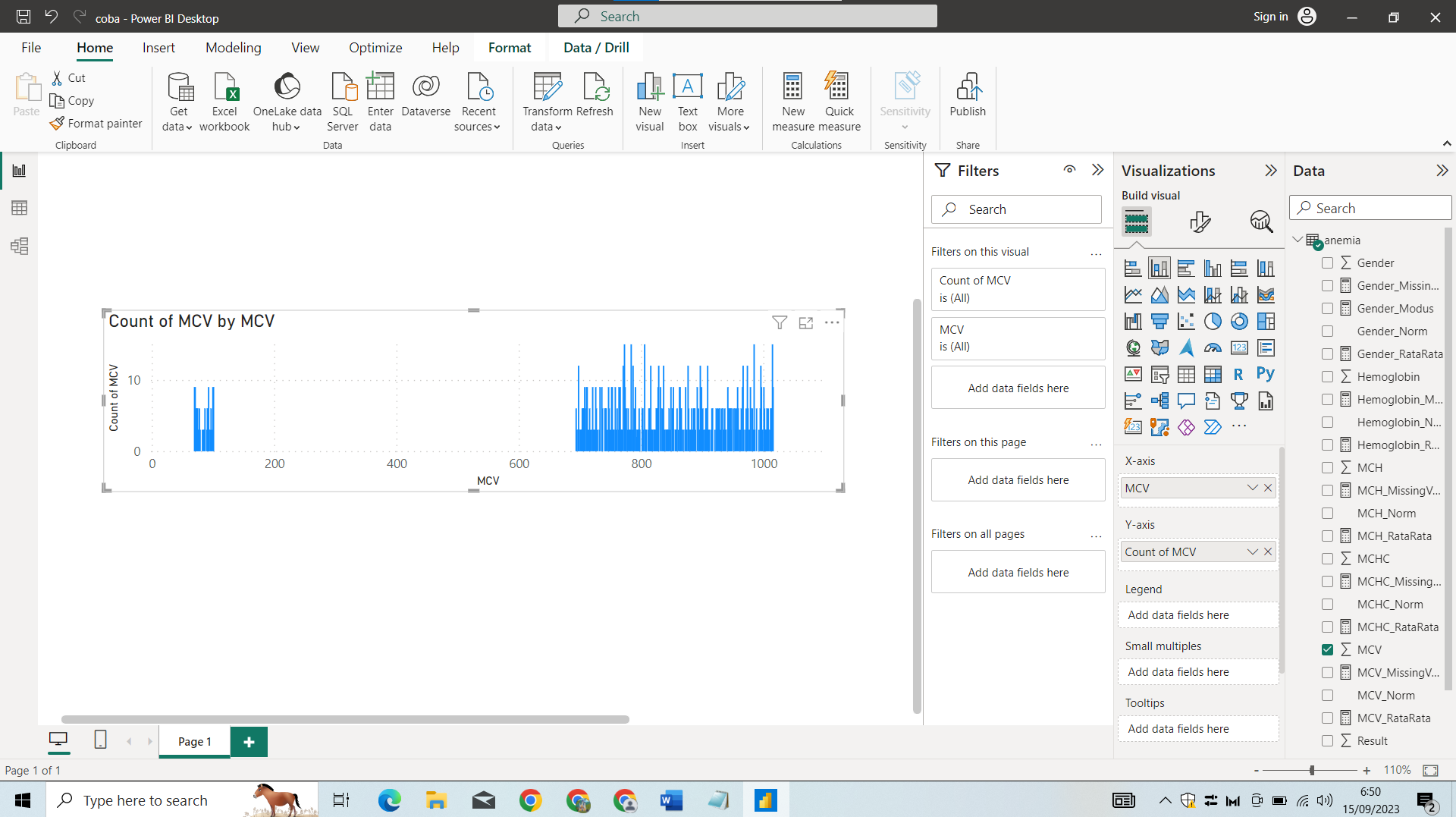
1. MCH (Mean Corpuscular Hemoglobin): MCH adalah ukuran rata-rata jumlah hemoglobin yang terkandung dalam setiap sel darah merah. Informasi ini dapat memberikan wawasan tentang kualitas sel darah merah dan kemampuannya dalam membawa oksigen.



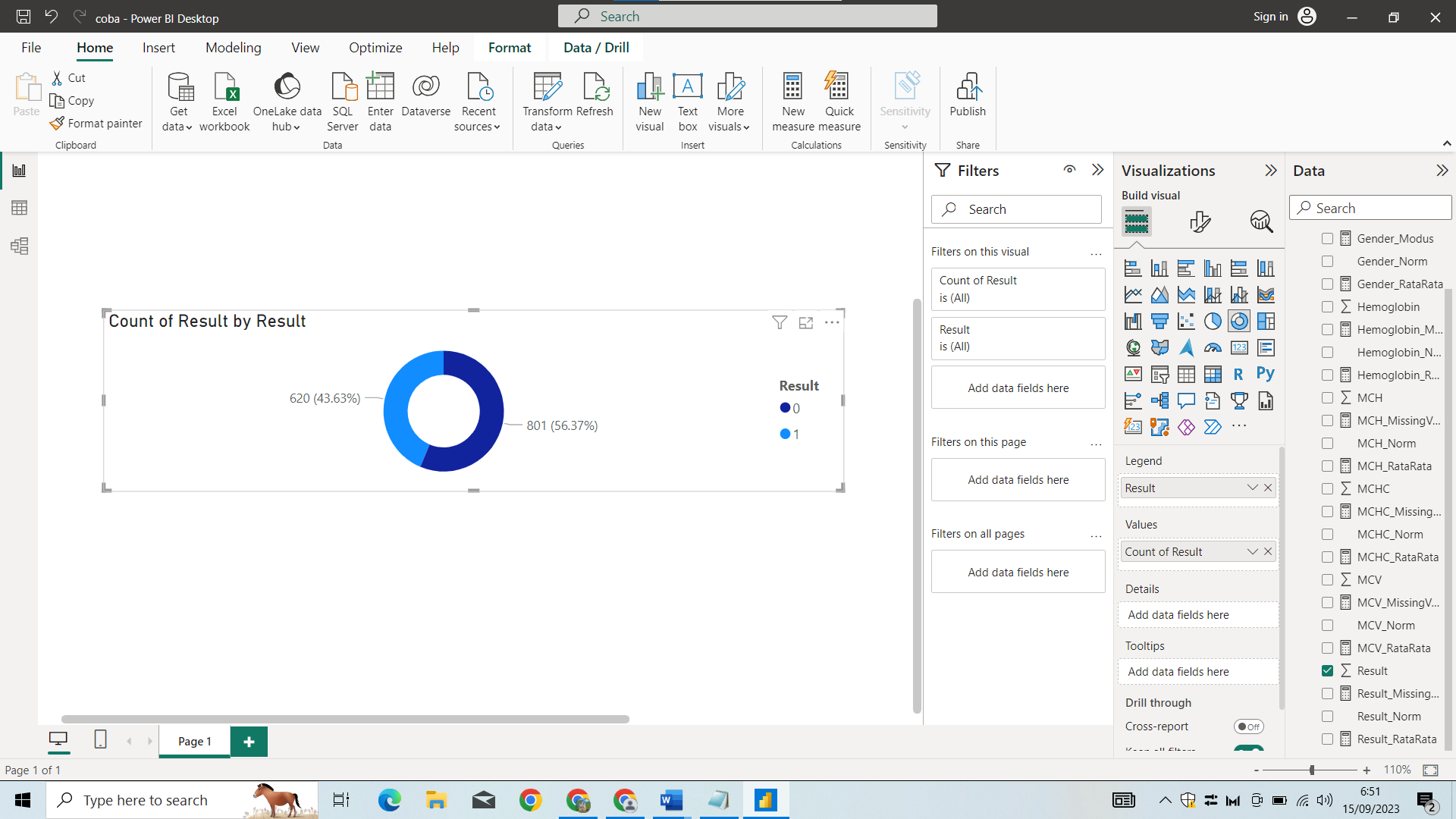
1. MCHC (Mean Corpuscular Hemoglobin Concentration): MCHC adalah ukuran konsentrasi rata-rata hemoglobin dalam satu sel darah merah. Hal ini juga berkaitan dengan kualitas sel darah merah dan dapat memberikan petunjuk tentang kemungkinan masalah seperti anemia.



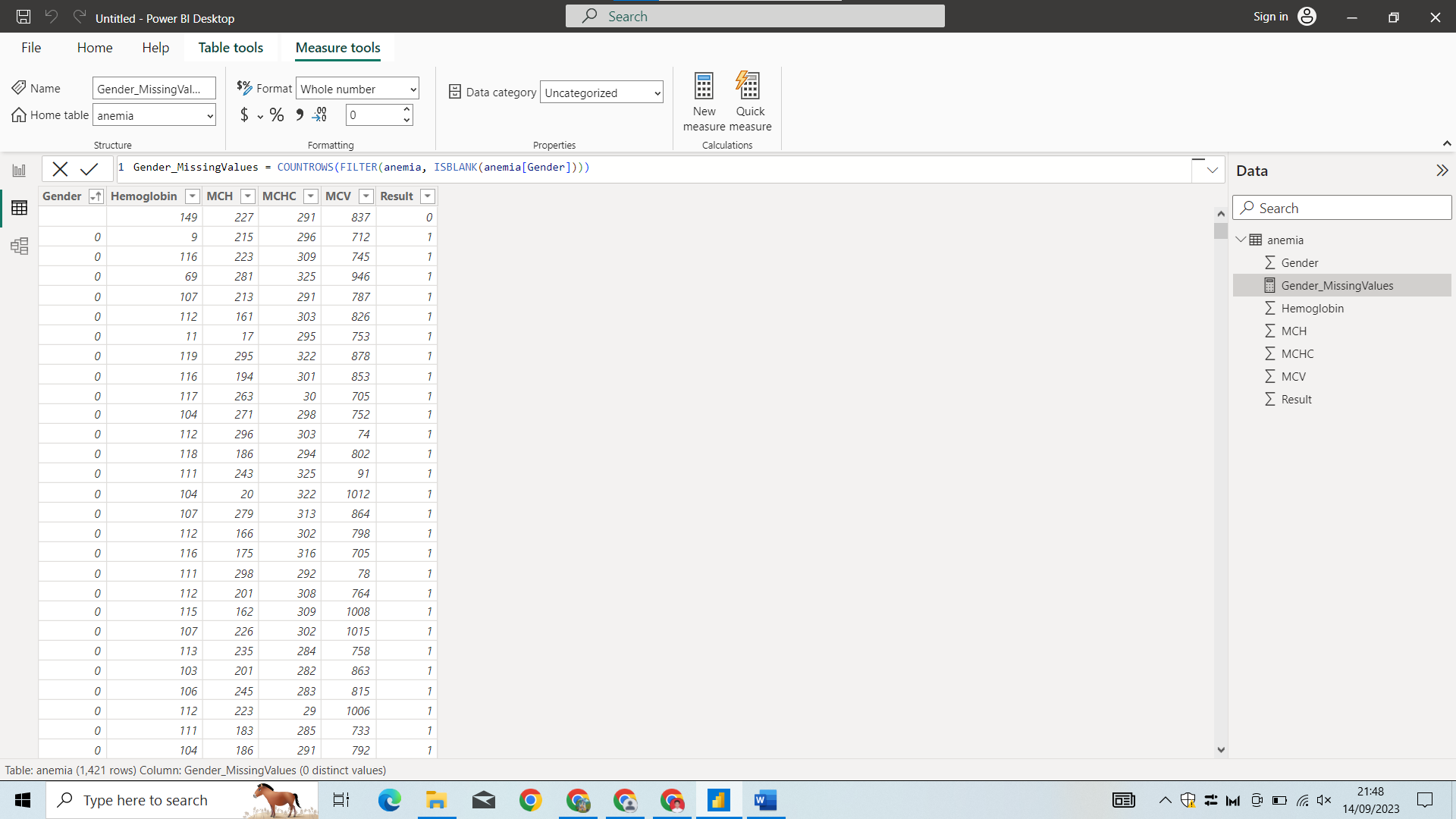
1. MCV (Mean Corpuscular Volume): MCV mengukur volume rata-rata dari sel darah merah. Perubahan dalam ukuran sel darah merah dapat mengindikasikan adanya masalah dalam produksi darah dan dapat menjadi petunjuk dalam diagnosis anemia.



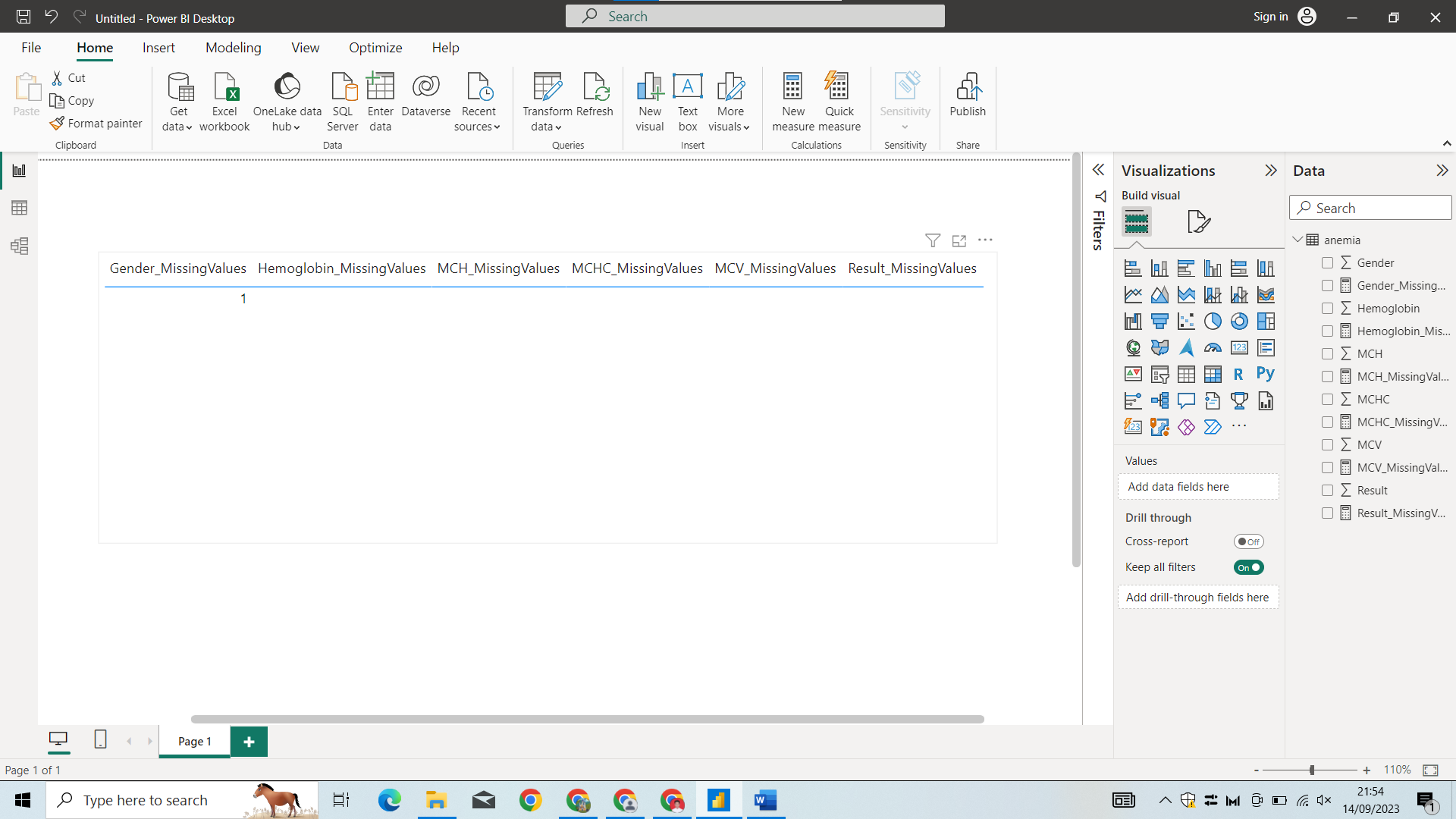
1. Hasil (Results): Atribut ini adalah hasil akhir dari prediksi atau diagnosis. Nilai 0 mungkin menunjukkan bahwa pasien tidak menderita anemia, sedangkan nilai 1 mungkin menunjukkan adanya anemia.



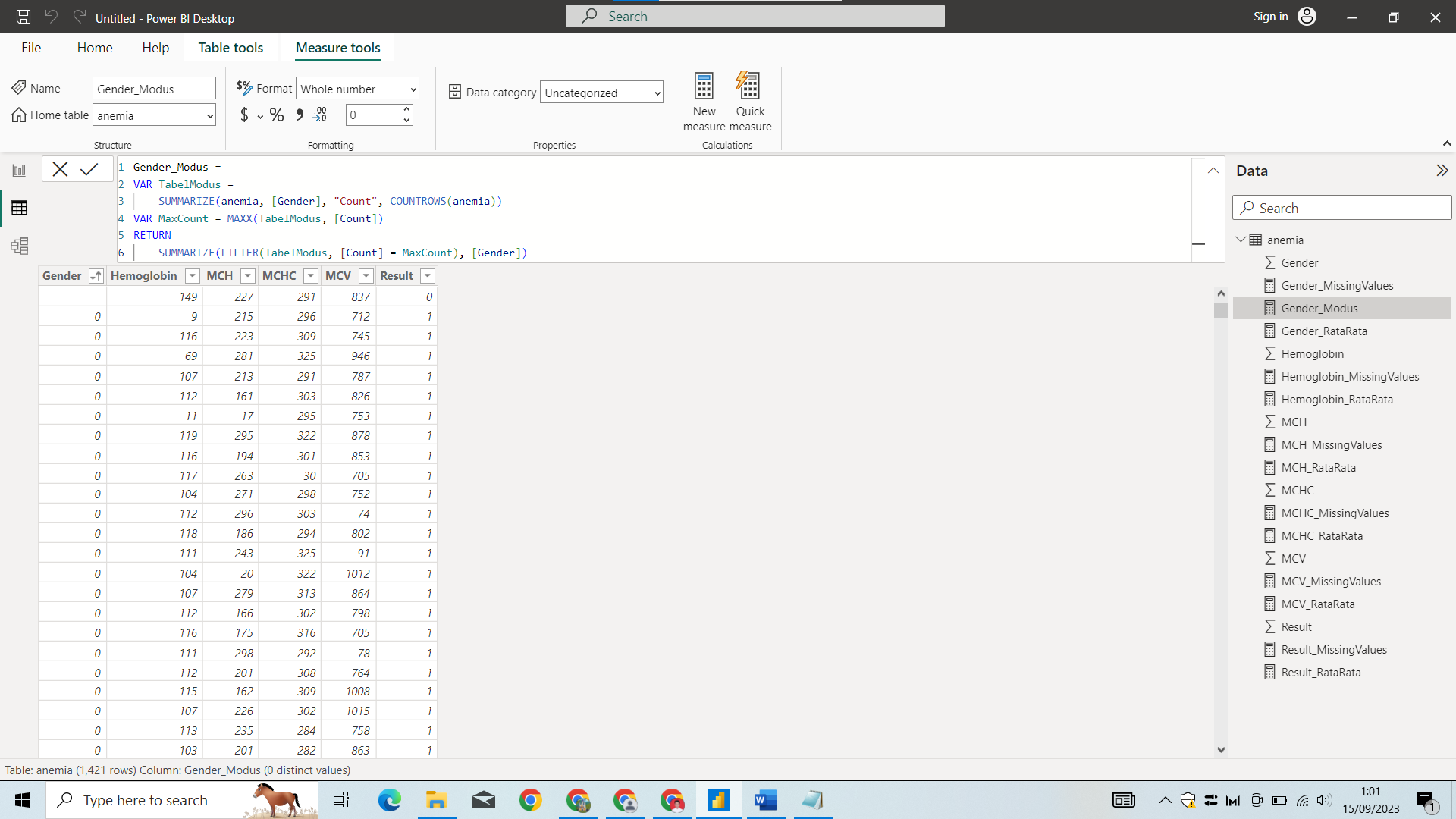
1. **Analisis Missing Values**



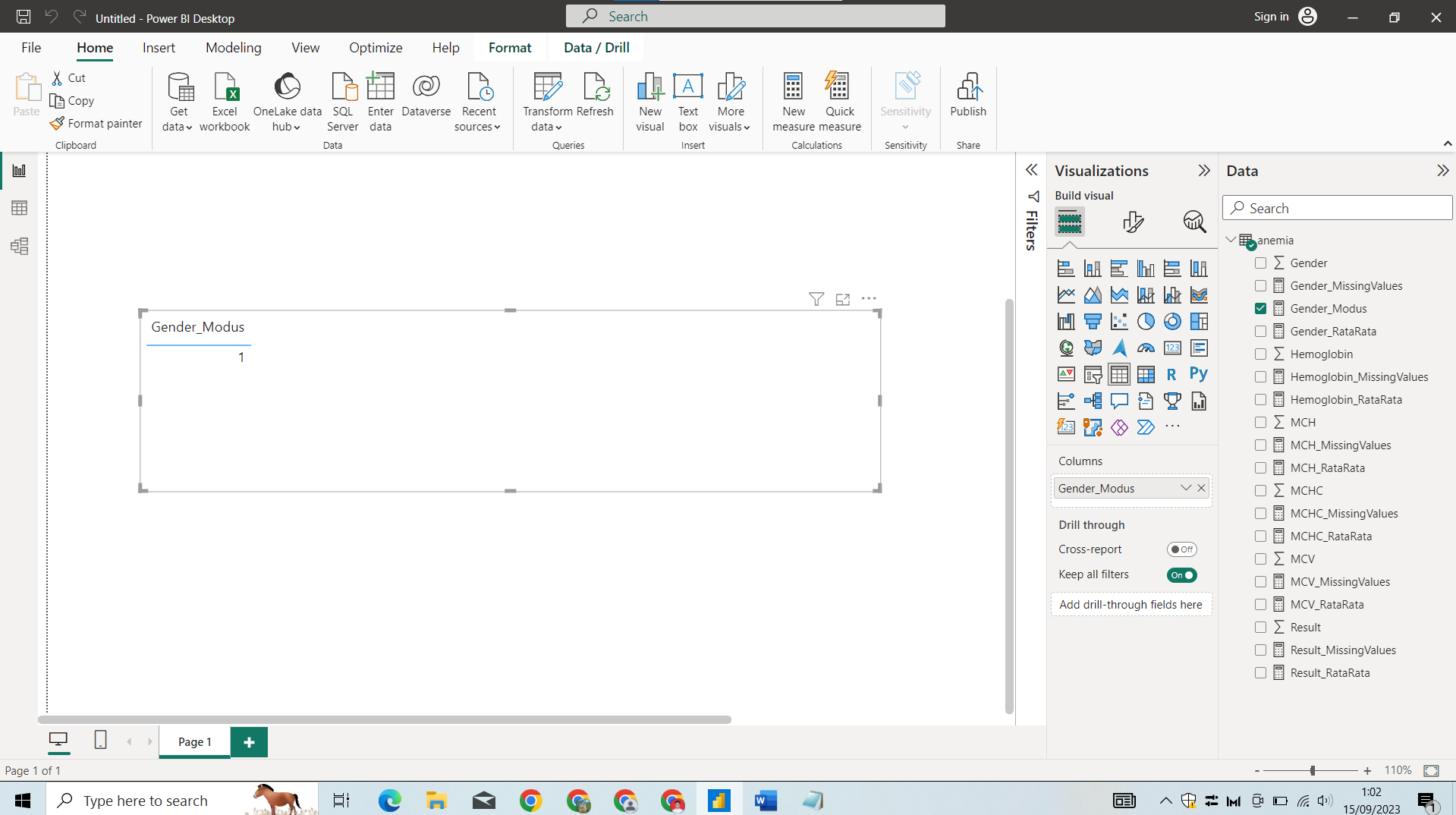
Tahap berikutnya setelah mendapatkan dataset anemia dari sumber seperti Kaggle adalah menangani nilai-nilai yang hilang atau missing values dalam dataset tersebut. Missing values adalah nilai-nilai yang tidak lengkap atau tidak ada dalam beberapa entri data, yang dapat mempengaruhi keakuratan analisis dan model yang akan dikembangkan. Proses menangani missing values melibatkan beberapa langkah, antara lain identifikasi missing values, penghapusan data yang hilang, imputasi data, pertimbangan konteks medis, dan validasi hasil.



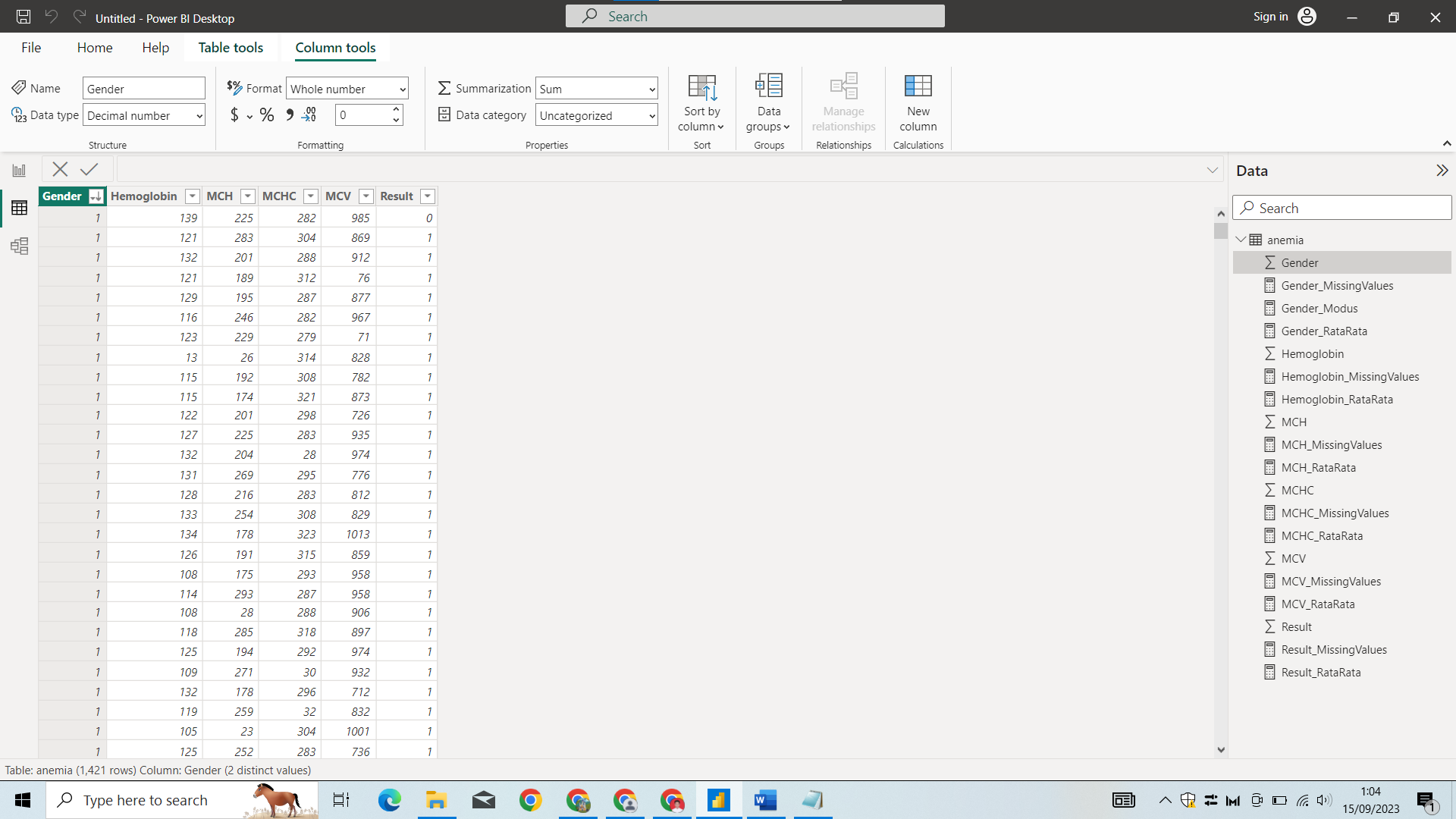
Tinjauan terhadap hasil menunjukkan bahwa satu-satunya atribut yang mengandung missing values adalah atribut "Gender" dalam dataset. Selain itu, terlihat bahwa hanya satu data yang terkena masalah missing values pada atribut "Gender". Dalam kondisi seperti ini, salah satu pendekatan yang dapat diambil adalah mengisi missing value pada atribut "Gender" menggunakan modus. Modus adalah nilai yang paling sering muncul dalam suatu atribut. Dalam konteks atribut "Gender," kita dapat mengasumsikan bahwa jenis kelamin yang paling umum adalah modus dari atribut ini. Dengan demikian, kita dapat mengisi missing value tersebut dengan jenis kelamin yang paling sering muncul dalam dataset, baik itu "laki-laki" (0) atau "perempuan" (1), tergantung pada modus yang ditemukan.



Terlihat bahwa data di atas memiliki nilai yang kosong atau tidak terisi.

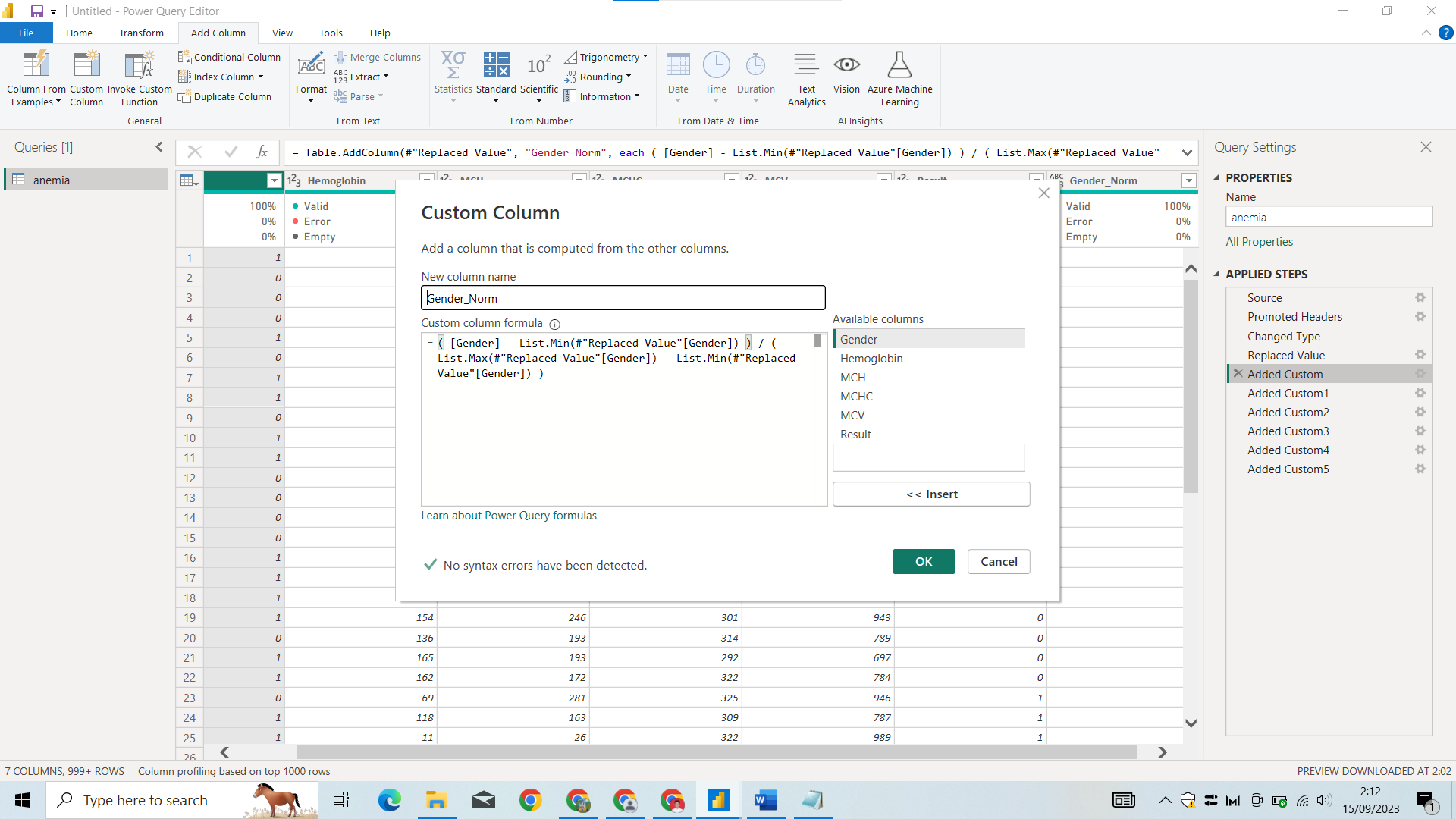


Terlihat dari hasil analisis bahwa nilai yang paling umum pada atribut "Gender" adalah 1, yang menunjukkan jenis kelamin perempuan.

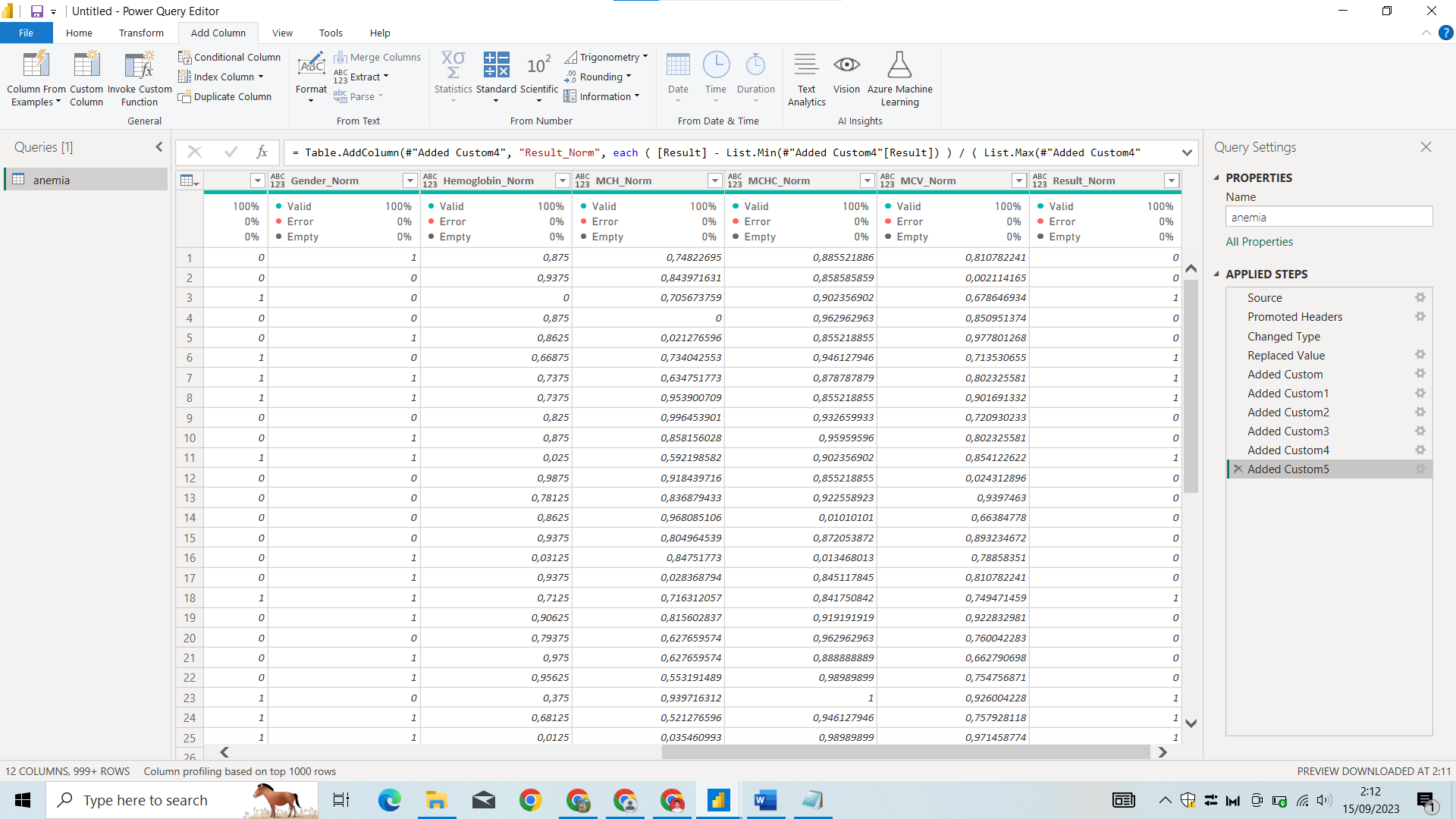


Selanjutnya, kita isikan nilai 1 pada data yang kosong tersebut.

1. **Normalisasi Data**



Tahap selanjutnya dalam analisis data adalah normalisasi data. Normalisasi adalah proses untuk mengubah skala atau rentang nilai-nilai dalam dataset sehingga semua atribut memiliki skala yang serupa atau terstandarisasi. Hal ini penting dalam analisis data karena dapat membantu model-machine learning atau algoritma statistik untuk bekerja lebih efisien dan menghasilkan hasil yang lebih baik.



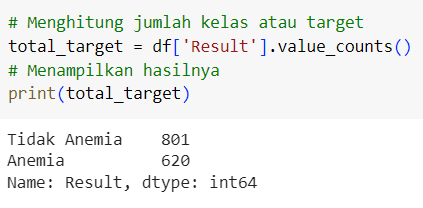
Bisa dilihat di atas, kita telah melakukan normalisasi menggunakan Min-Max Scaler agar semua atribut dalam dataset memiliki rentang nilai yang serupa, yaitu antara 0 dan 1. Normalisasi ini dilakukan untuk memastikan bahwa setiap atribut memberikan kontribusi yang setara dalam analisis data dan pengembangan model. Dengan rentang nilai yang serupa, algoritma machine learning atau analisis statistik dapat bekerja lebih efisien dan menghasilkan hasil yang lebih akurat.

1. **Balancing Data**

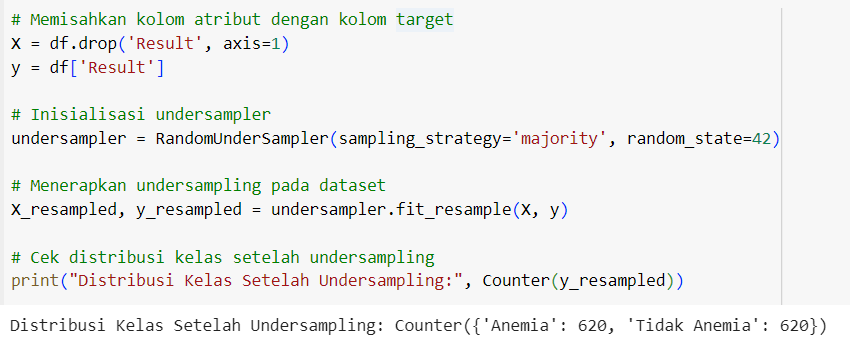
Balancing data adalah proses untuk mengatasi ketidakseimbangan antara kelas atau kelompok dalam sebuah dataset. Dalam banyak kasus, dataset mungkin memiliki distribusi yang tidak seimbang antara kelas yang berbeda, yang dapat mengakibatkan masalah dalam pemodelan atau analisis data. Ketidakseimbangan ini terutama menjadi masalah dalam tugas-tugas klasifikasi di mana ada dua atau lebih kelas yang harus diprediksi.

Dalam konteks balancing data, ada dua masalah umum yang sering terjadi:

1. Oversampling: Oversampling melibatkan peningkatan jumlah sampel dalam kelas minoritas (kelas yang jumlahnya lebih sedikit) dalam dataset. Ini dapat dilakukan dengan menggandakan atau membuat salinan data dalam kelas minoritas atau dengan menciptakan sampel sintetis yang mirip dengan sampel yang ada. Oversampling membantu menjaga keseimbangan antara kelas, sehingga model dapat mempelajari karakteristik kelas minoritas dengan lebih baik.
2. Undersampling: Undersampling melibatkan pengurangan jumlah sampel dalam kelas mayoritas (kelas yang jumlahnya lebih banyak) dalam dataset. Tujuan undersampling adalah membuat jumlah sampel dalam kedua kelas menjadi lebih seimbang. Namun, pengurangan sampel kelas mayoritas juga dapat mengurangi informasi yang tersedia untuk model.



Dari data diatas, terlihat bahwa jumlah data tidak seimbang antara dua kelas, yaitu "tidak anemia" dengan 801 sampel dan "anemia" dengan 620 sampel. Ketidakseimbangan ini adalah contoh yang umum terjadi dalam dataset klasifikasi di mana salah satu kelas memiliki jumlah sampel yang lebih sedikit dibandingkan dengan kelas lainnya.



Setelah dilakukan undersampling, jumlah sampel dalam kedua kelas menjadi sama, yaitu sebanyak 620 sampel. Tindakan ini telah menghasilkan keseimbangan antara kelas "tidak anemia" dan "anemia" dalam dataset. Dengan memiliki jumlah sampel yang sama dalam kedua kelas, model yang akan dikembangkan kemungkinan akan lebih seimbang dalam mempelajari karakteristik kedua kelas tersebut, dan hasilnya bisa lebih akurat dalam memprediksi atau mengidentifikasi kasus anemia. Balancing dataset adalah langkah yang penting untuk meningkatkan kualitas hasil analisis dan pemodelan, terutama dalam tugas-tugas klasifikasi di mana keseimbangan antara kelas sangat penting.