Nama : Yoena Rindu Perdana

NIM : 200411100145

Kelas : PSD E

Tugas 1 Analisis Data

**Dataset**

Sindrom ovarium polikistik (*Polycystic Ovary Syndrome*) adalah gangguan yang melibatkan menstruasi yang jarang, tidak teratur, atau berkepanjangan, serta tingkat hormon pria (*androgen*) yang berlebihan. Ovarium mengembangkan banyak kumpulan kecil cairan — yang disebut folikel — dan mungkin gagal melepaskan telur secara teratur.

Dataset ini berisi semua parameter fisik dan klinis untuk menentukan PCOS dan masalah infertilitas terkait. Data dikumpulkan dari 10 rumah sakit berbeda di Kerala, India. Ada total 44 parameter dalam dataset ini. Data ini diambil dari <https://www.kaggle.com/datasets/prasadbobby/pcosdata>.

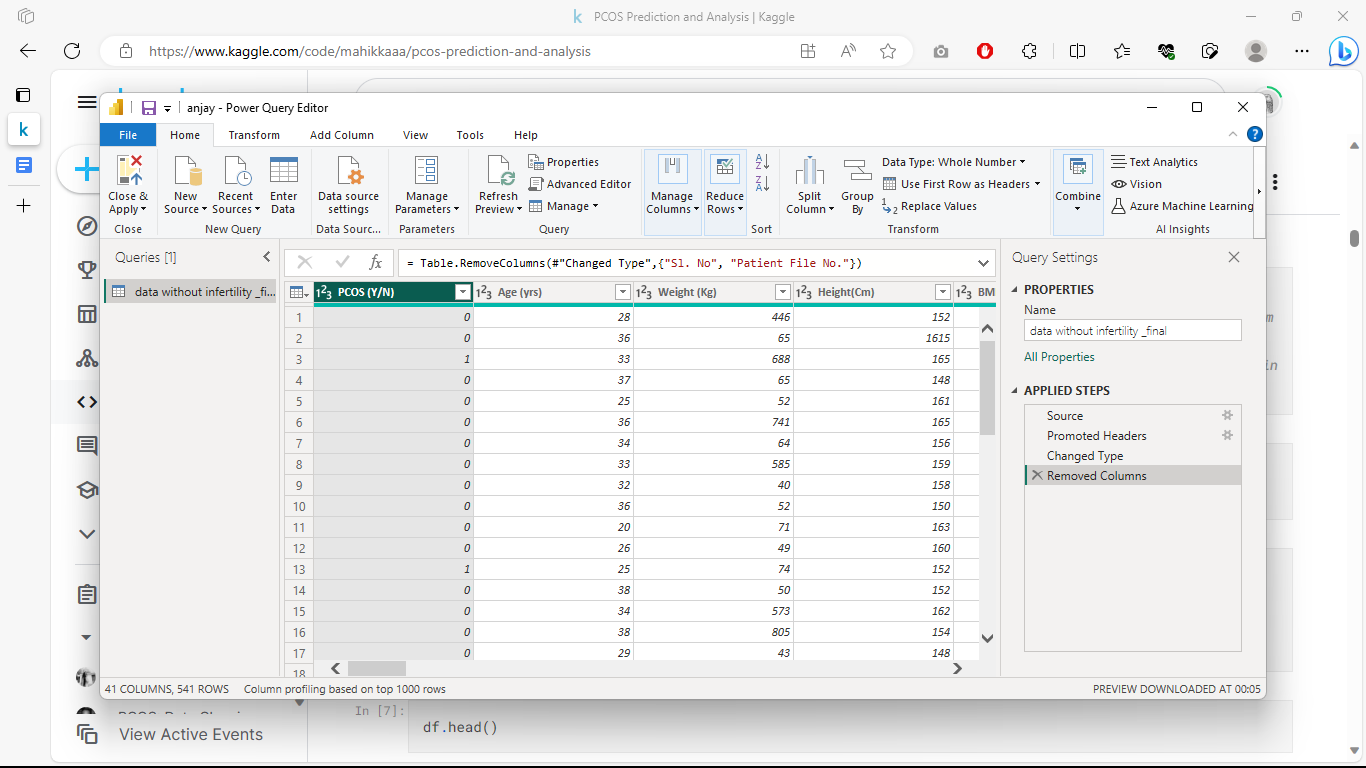
Berikut adalah 44 parameter atau feature yang ada jika diterjemahkan ke dalam Bahasa indonesia:

1. Patient File No. : Ini adalah nomor laporan yang berisi data untuk pasien tertentu.
2. PCOS: Sindrom ovarium polikistik (PCOS) adalah gangguan hormonal yang umum terjadi pada wanita usia reproduksi, kami ingin menentukan apakah pasien menderita sindrom ini atau tidak.
3. Age (yrs): Usia pasien dalam tahun.
4. Weight (Kg): Berat pasien dalam kilogram.
5. Height (Cm): Tinggi pasien dalam sentimeter.
6. BMI: Indeks massa tubuh pasien.
7. Blood group: Golongan darah pasien (A+ = 11, A- = 12, B+ = 13, B- = 14, O+ = 15, O- = 16, AB+ = 17, AB- = 18; total 8 golongan darah).
8. Pulse rate (bpm): Ini adalah denyut jantung pasien dalam denyut per menit. Denyut jantung istirahat untuk orang dewasa berkisar antara 60 hingga 100 denyut per menit.
9. RR (breaths/menit): Ini adalah tingkat pernapasan. Tingkat pernapasan normal untuk orang dewasa pada saat istirahat berkisar antara 12 hingga 16 napas per menit.
10. Hb (g/dl): Tingkat hemoglobin dalam gram per desiliter. Untuk wanita, tingkat normal berkisar antara 12,3 gm/dL dan 15,3 gm/dL.
11. Cycle (R/I): ...
12. Cycle length (hari): Ini mewakili panjang siklus menstruasi. Panjang siklus menstruasi bervariasi dari wanita ke wanita, tetapi rata-rata adalah menstruasi setiap 28 hari.
13. Marraige Status (Yrs): Tahun pernikahan.
14. Pregnant (Y/N): Jika pasien hamil.
15. No. of aborptions: Jumlah keguguran, jika ada. Ada total 541 nilai, di mana 437 pasien tidak pernah mengalami keguguran.
16. I beta-HCG (mIU/mL): Ini adalah kasus 1 beta hCG.
17. II beta-HCG (mIU/mL): Ini adalah kasus 2 beta hCG (perlu diingat: tingkat beta hCG kurang dari 5 mIU/mL dianggap negatif untuk kehamilan, dan apa pun di atas 25 mIU/mL dianggap positif untuk kehamilan) (juga unit mIU/mL adalah mili Unit Internasional per miliLiter).
18. FSH (mIU/mL): Nama lengkapnya adalah Follicle-stimulating hormone. Selama pubertas: berkisar antara 0,3 hingga 10,0 mIU/mL (0,3 hingga 10,0 IU/L) Wanita yang masih menstruasi: 4,7 hingga 21,5 mIU/mL (4,5 hingga 21,5 IU/L) Setelah menopause: 25,8 hingga 134,8 mIU/mL (25,8 hingga 134,8 IU/L).
19. LH (mIU/mL): Ini adalah Hormon Luteinisasi.
20. FSH/LH: Rasio FSH dan LH.
21. Hip (inci): Ukuran pinggul dalam inci.
22. Waist (inci): Ukuran pinggang dalam inci.
23. Waist hip ratio: Rasio pinggang dengan pinggul.
24. TSH (mIU/L): Ini adalah hormon perangsang tiroid. Nilai normal berkisar dari 0,4 hingga 4,0 mIU/L.
25. AMH (ng/mL): Ini adalah Hormon Anti-Mullerian.
26. PRL (ng/mL): Ini mewakili tingkat prolaktin.
27. Vit D3 (ng/mL): Tingkat vitamin D. Tingkat vitamin D normal dalam darah adalah 20 ng/ml atau lebih tinggi untuk orang dewasa.
28. PRG (ng/mL): Tingkat progesteron.
29. RBS (mg/dL): Nilai ini diperoleh dengan melakukan Tes Gula Darah Acak (RBS).
30. Weight gain (Y/N): Apakah terjadi peningkatan berat badan.
31. Hair growth (Y/N): Apakah terjadi pertumbuhan rambut.
32. Skin darkening (Y/N): Masalah pergelapan kulit.
33. Hair loss (Y/N): Masalah kehilangan rambut.
34. Pimples (Y/N): Masalah jerawat.
35. Fast Food (Y/N): Apakah makanan cepat saji bagian dari pola makan Anda.
36. Reg.Exercise (Y/N): Apakah Anda melakukan latihan secara rutin.
37. BP \_Systolic (mmHg): Tekanan darah sistolik, mengukur tekanan dalam arteri saat jantung Anda berdetak.
38. BP \_Diastolic (mmHg): Tekanan darah diastolik, mengukur tekanan dalam arteri saat jantung Anda istirahat di antara detakannya.
39. Follicle No. (L): Jumlah folikel di sisi kiri.
40. Follicle No. (R): Jumlah folikel di sisi kanan.
41. Avg. F size (L) (mm): Ukuran rata-rata folikel di sisi kiri dalam milimeter.
42. Avg. F size (R) (mm): Ukuran rata-rata folikel di sisi kanan dalam milimeter.
43. Endometrium (mm): Ukuran endometrium dalam milimeter.
44. Unnamed: 44

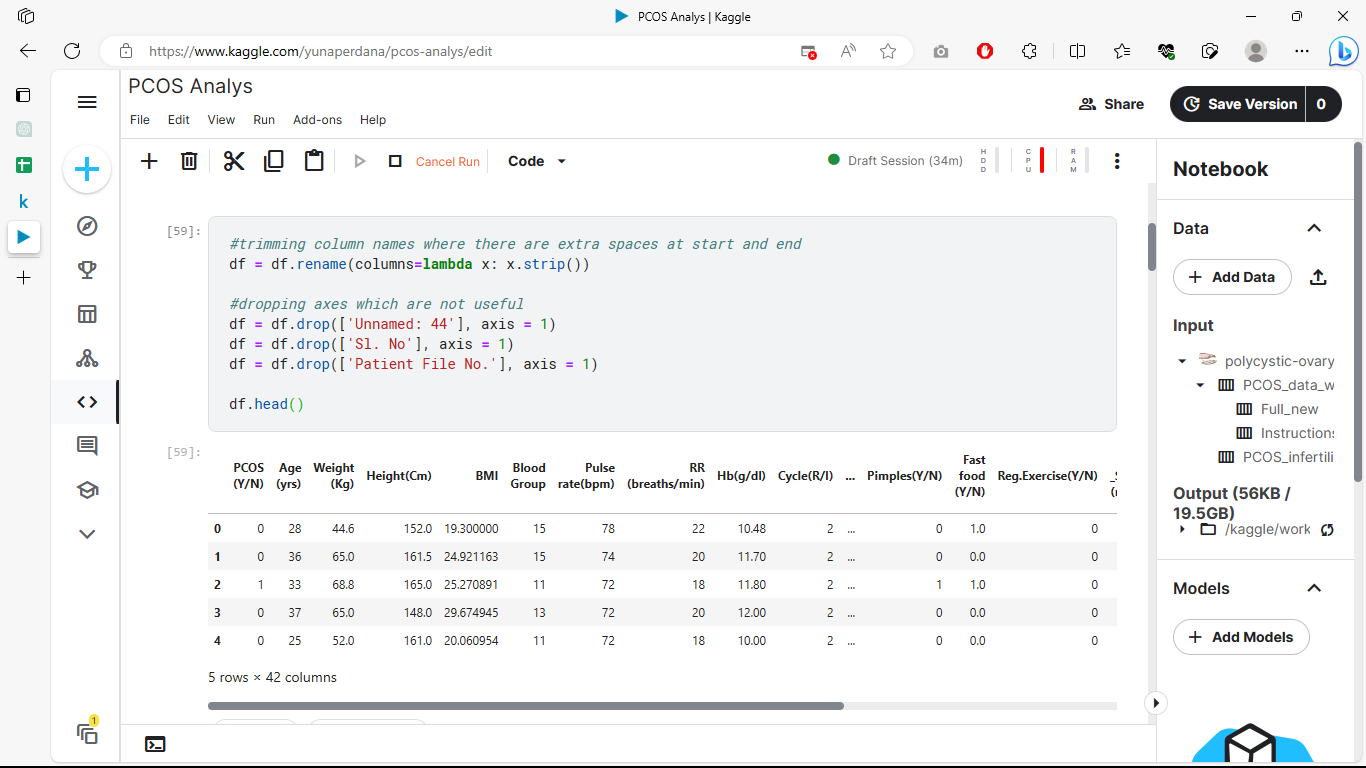
Dalam **analisis data** kali ini, penulis menggunakan software Power BI.

**Cleaning Data**

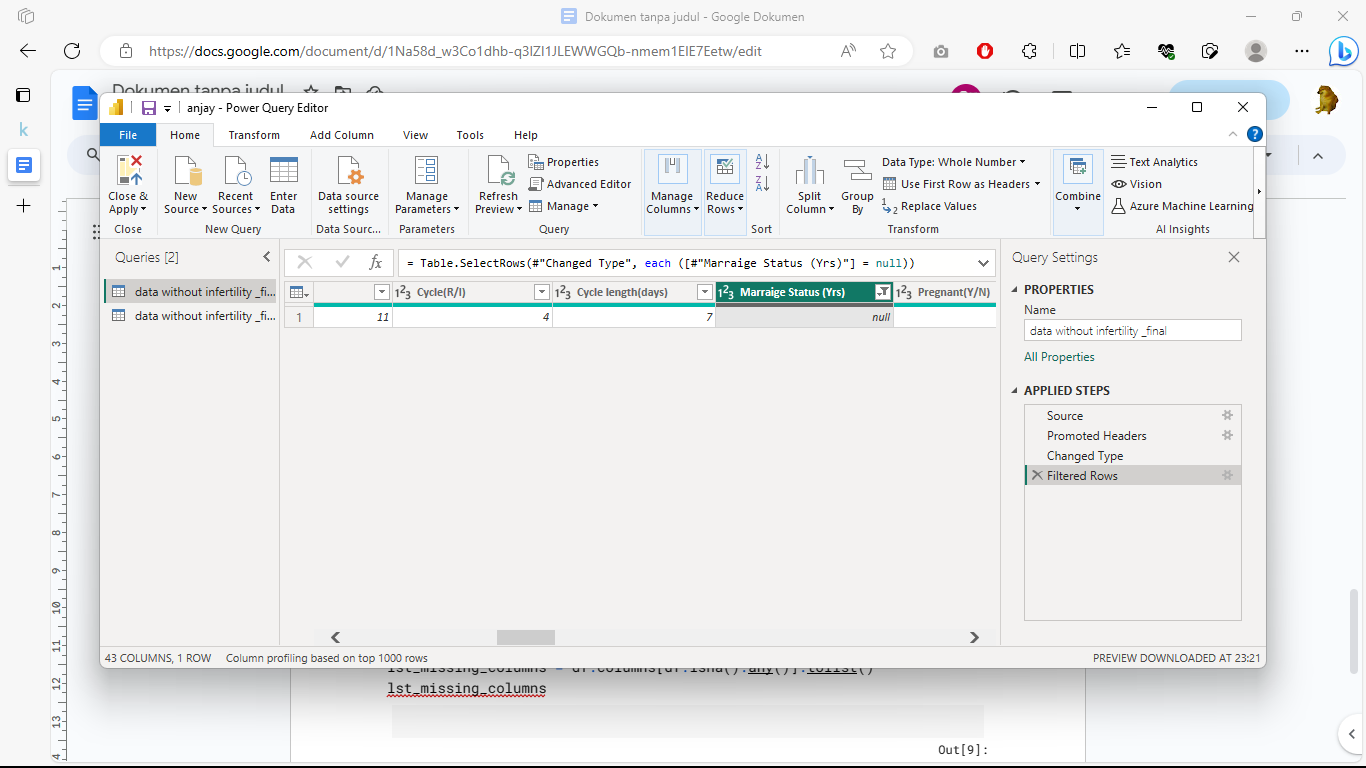
Dalam mengolah dataset ini untuk tujuan analisis terkait sindrom ovarium polikistik (PCOS) dan masalah infertilitas, penulis memutuskan untuk menghapus tiga parameter yang tidak berhubungan dengan target. Parameter pertama adalah "Patient File No," yang merupakan nomor laporan pasien dan tidak memiliki relevansi dengan karakteristik PCOS atau infertilitas. Parameter kedua adalah "S1. no.," yang tampaknya merupakan nomor urutan entri data dan juga tidak memberikan informasi yang berguna untuk analisis kami. Terakhir, penulis menghapus "column 44" karena tidak ada deskripsi atau informasi yang terkait dengan parameter ini dalam deskripsi dataset.



Atau dapat dengan memakai code di bawah ini untuk diterapkan di text editor.



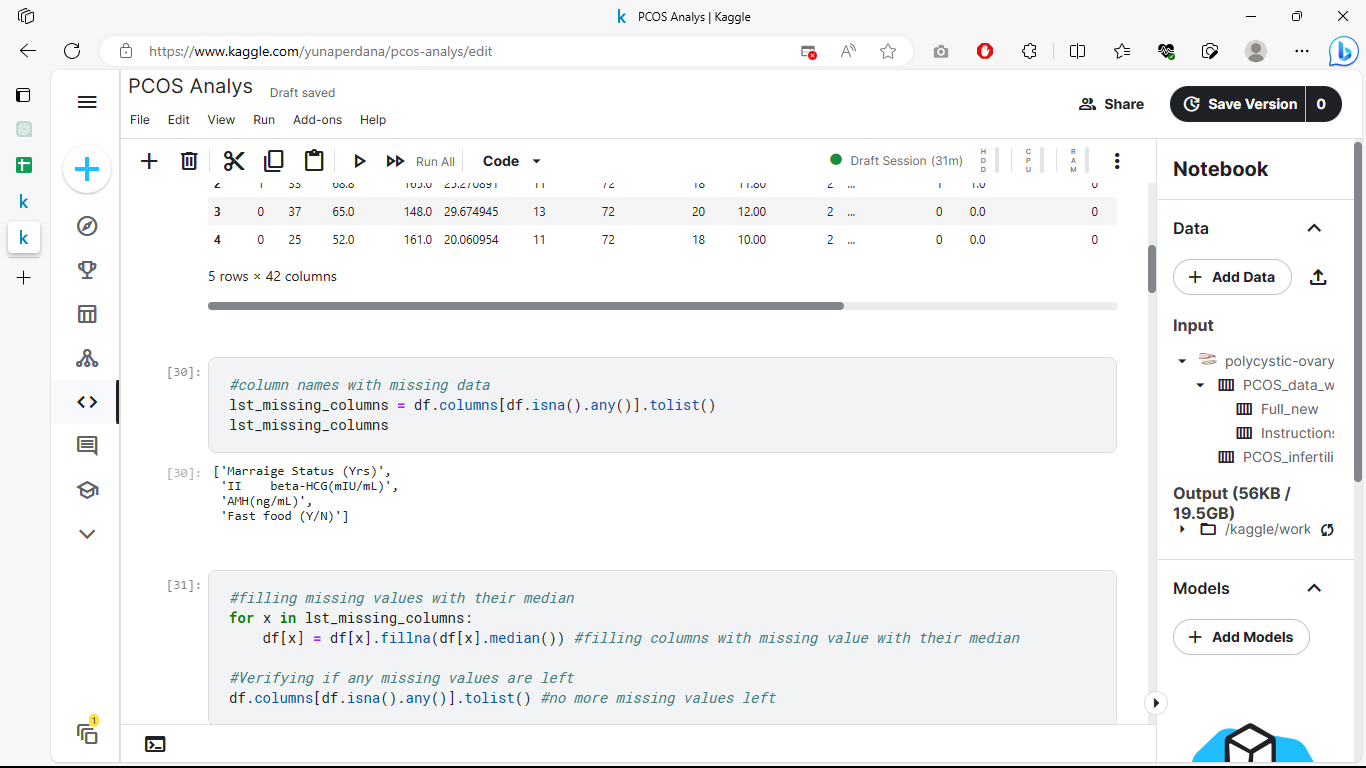
Selain itu mengindentifikasi data yang perlu dibersihkan seperti duplikasi atau missing value. penulis menggunakan fungsi filter rows untuk menampilkan atau menyaring baris yang mengandung missing value “null” dan menemukan beberapa missing values di antaranya terdapat pada parameter Marraige Status (Yrs), AMH(ng/mL), dan Fast food (Y/N). Di bawah ini merupakan bentuk fungsi filter rows untuk menemukan missing value dengan contoh di parameter Marraige Status (Yrs).



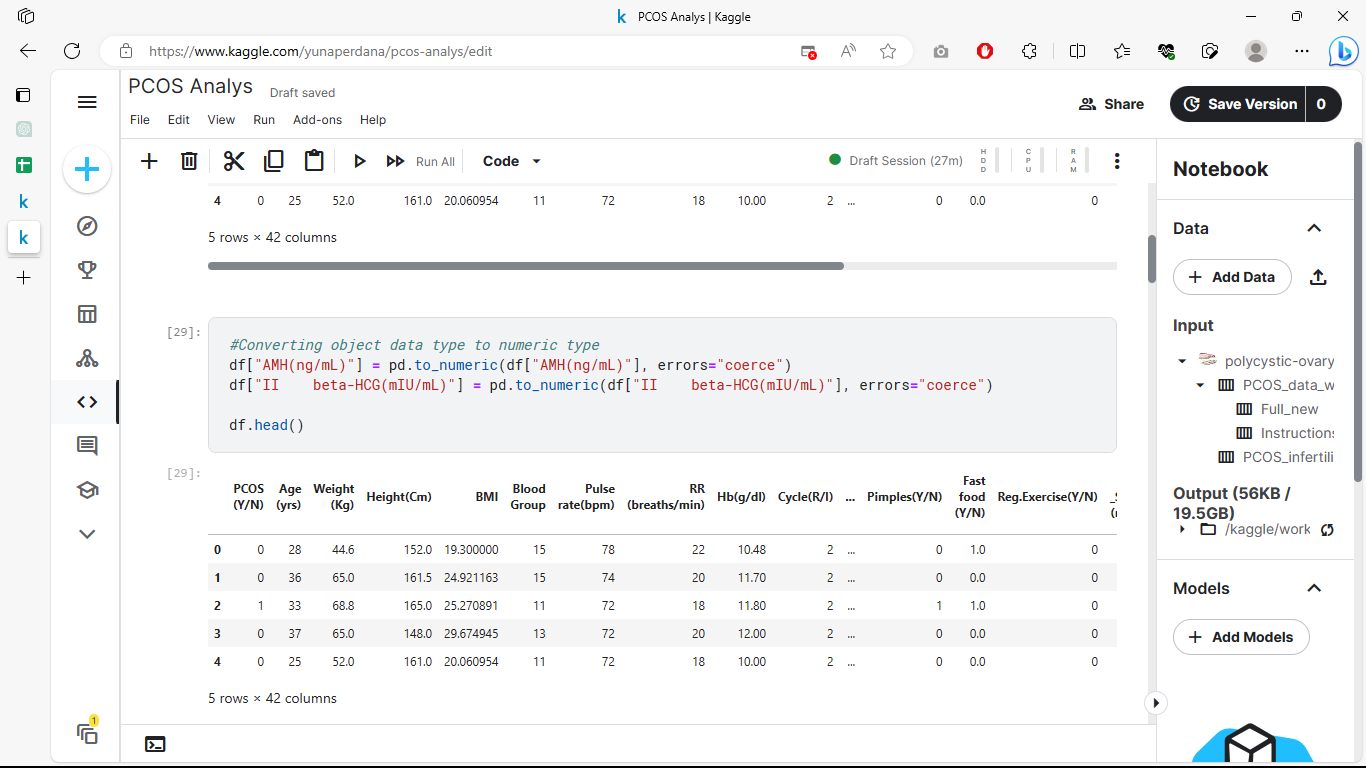
Selanjutnya dilakukan “remove rows” untuk menghapus baris yang mengandung missing value. sebagai contoh di bawah ini :



Penulis juga telah mencoba mencari nilai yang hilang (*missing value*) dengan menampilkan kolom-kolom yang mengandung nilai yang hilang tersebut melalui platform *Kaggle*, menggunakan code berikut:

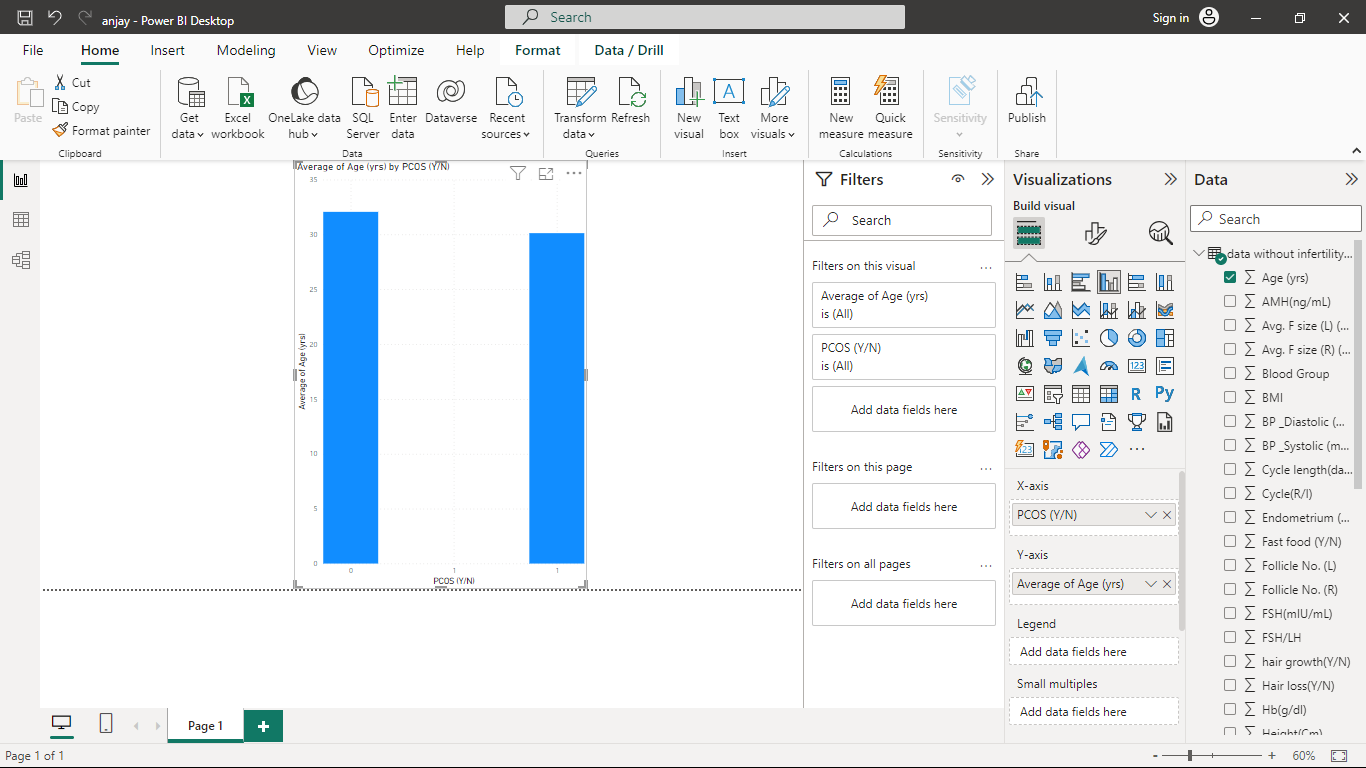
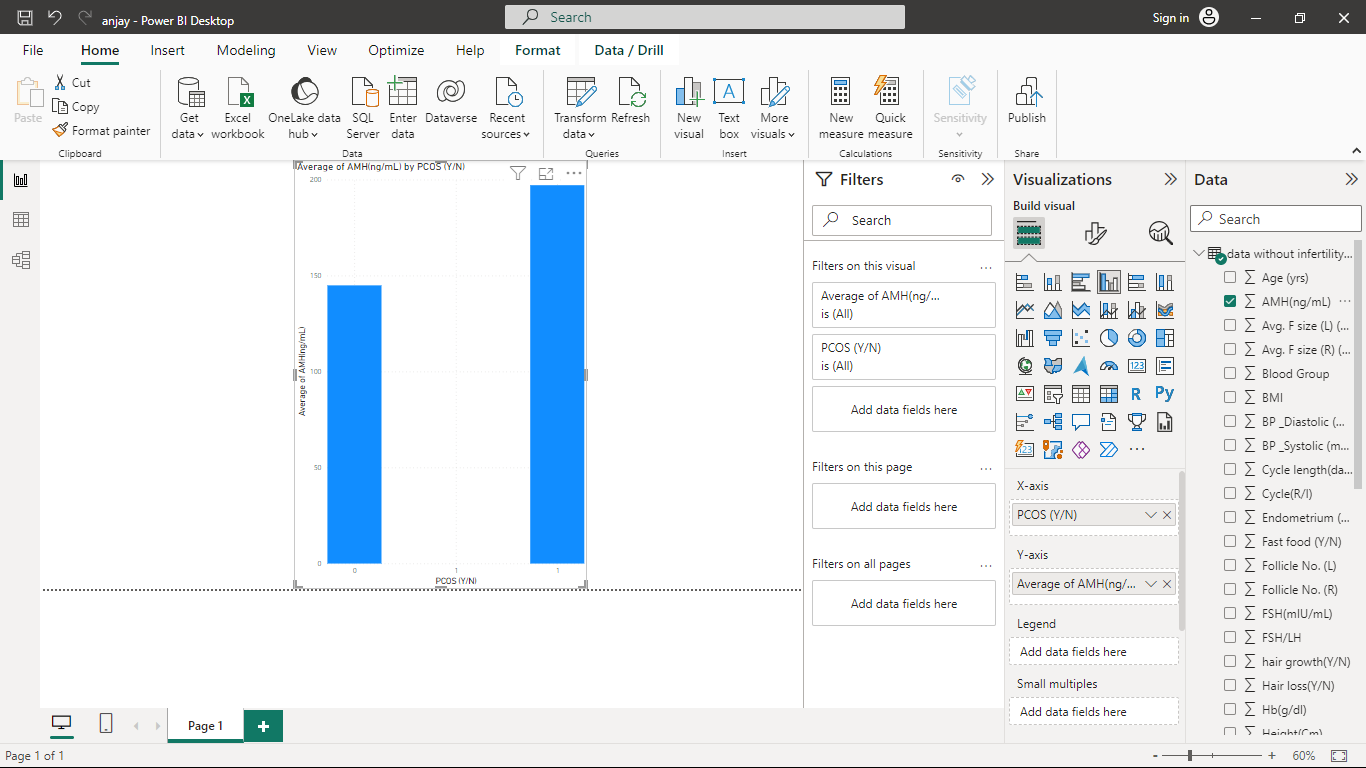
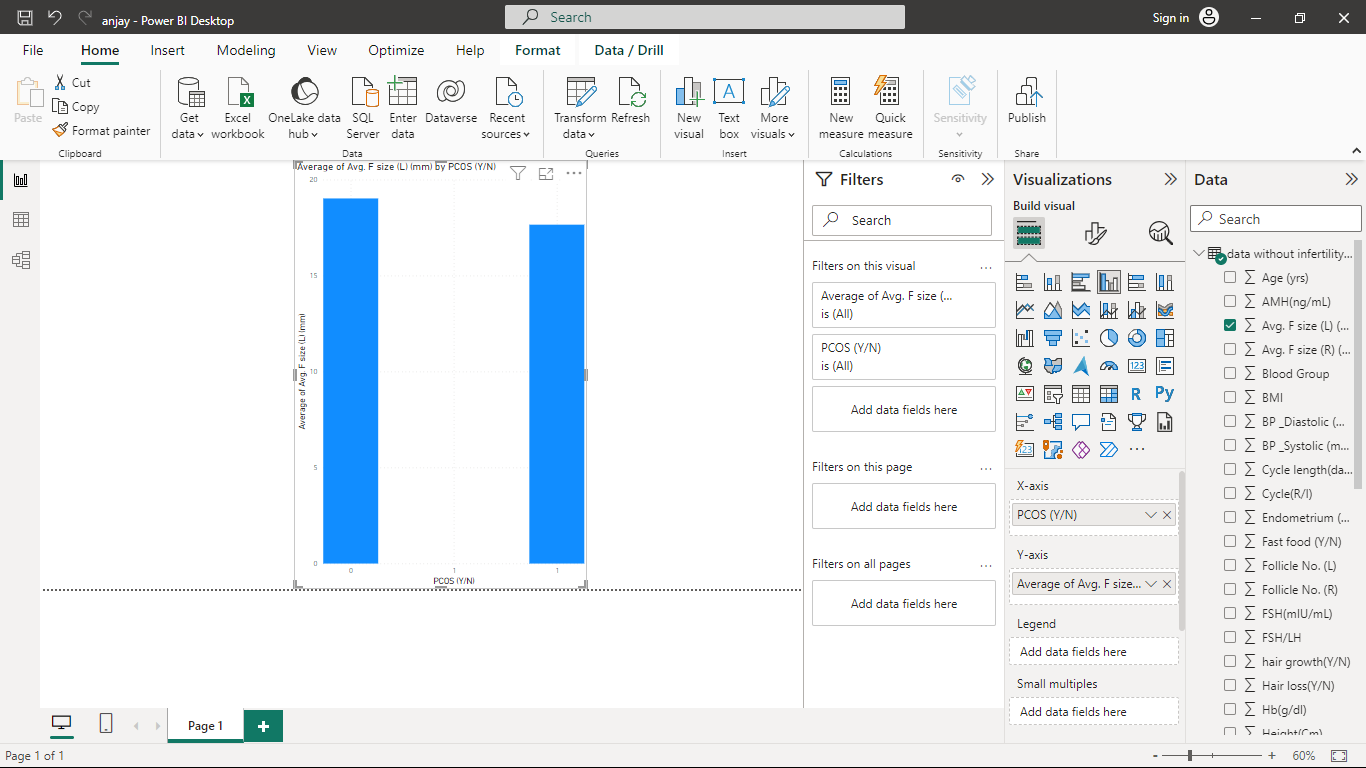
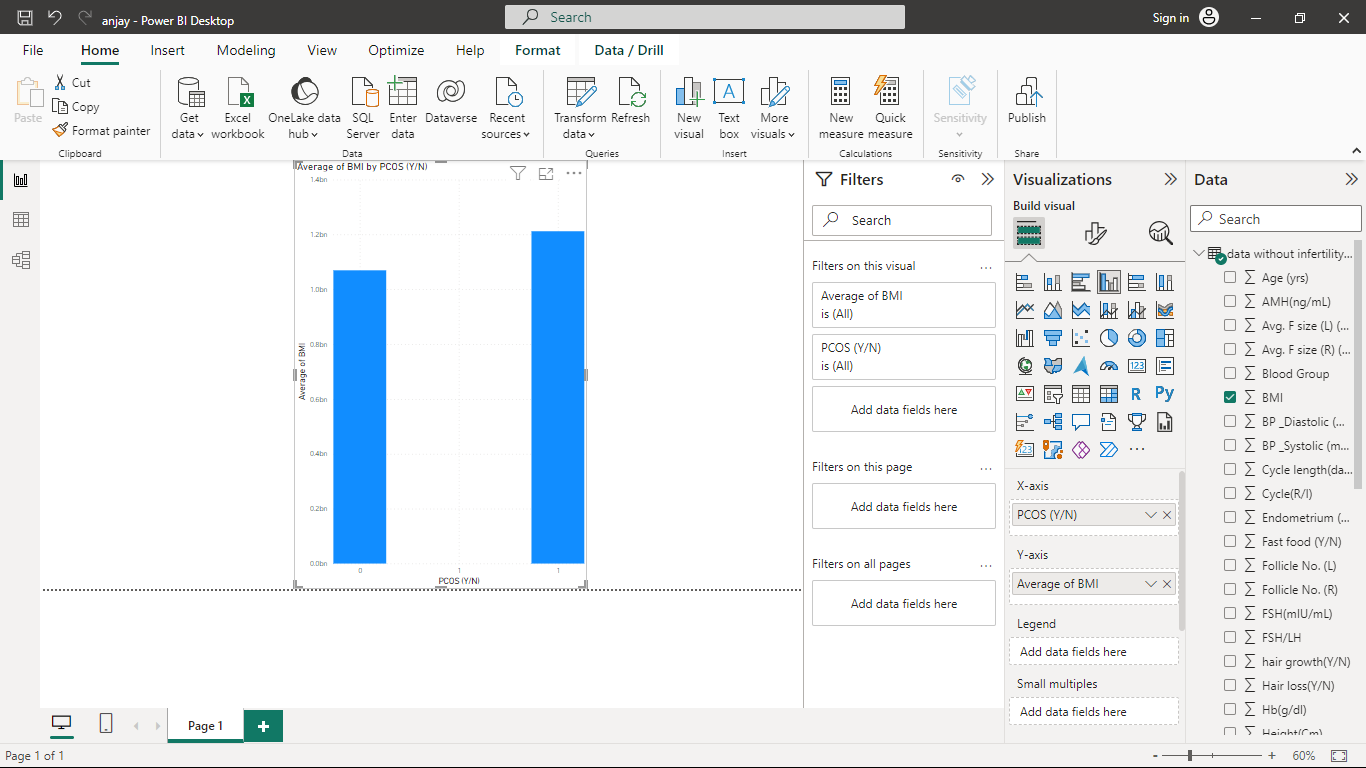
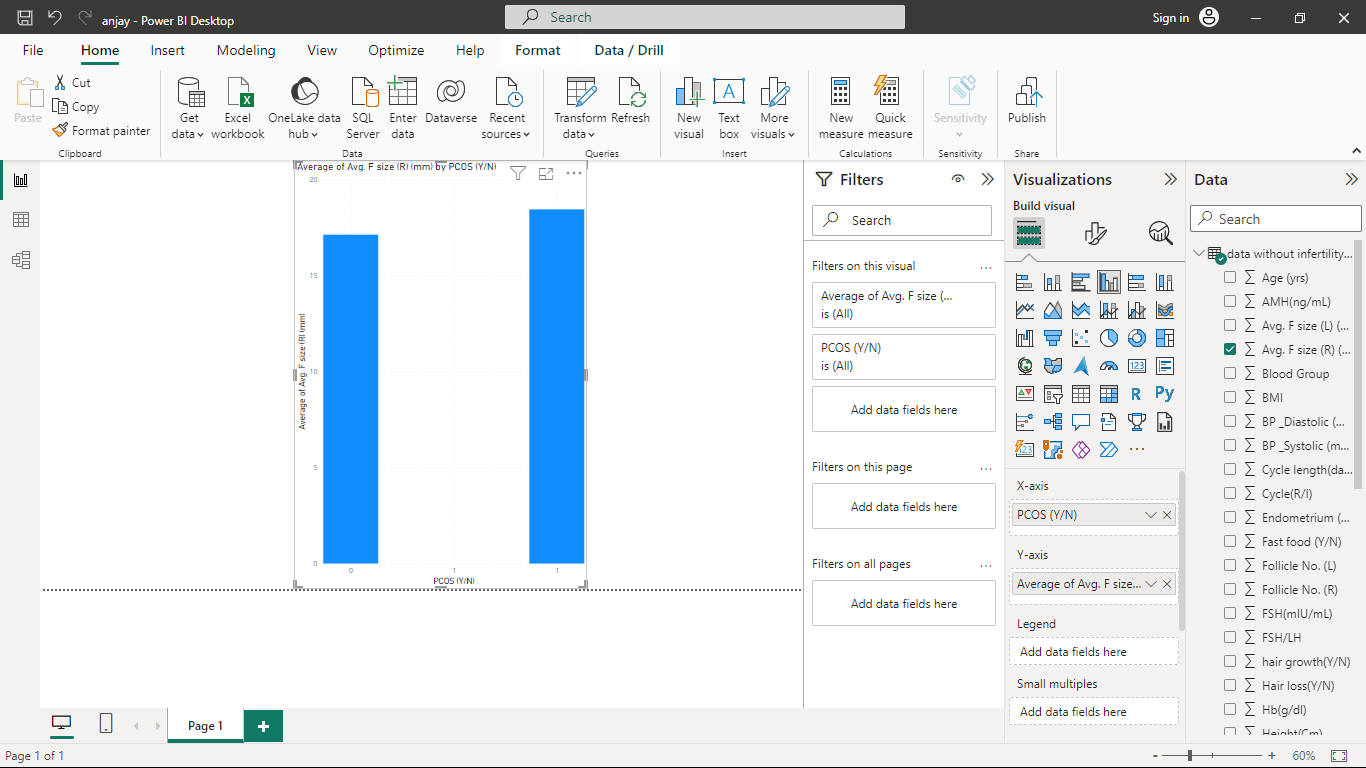
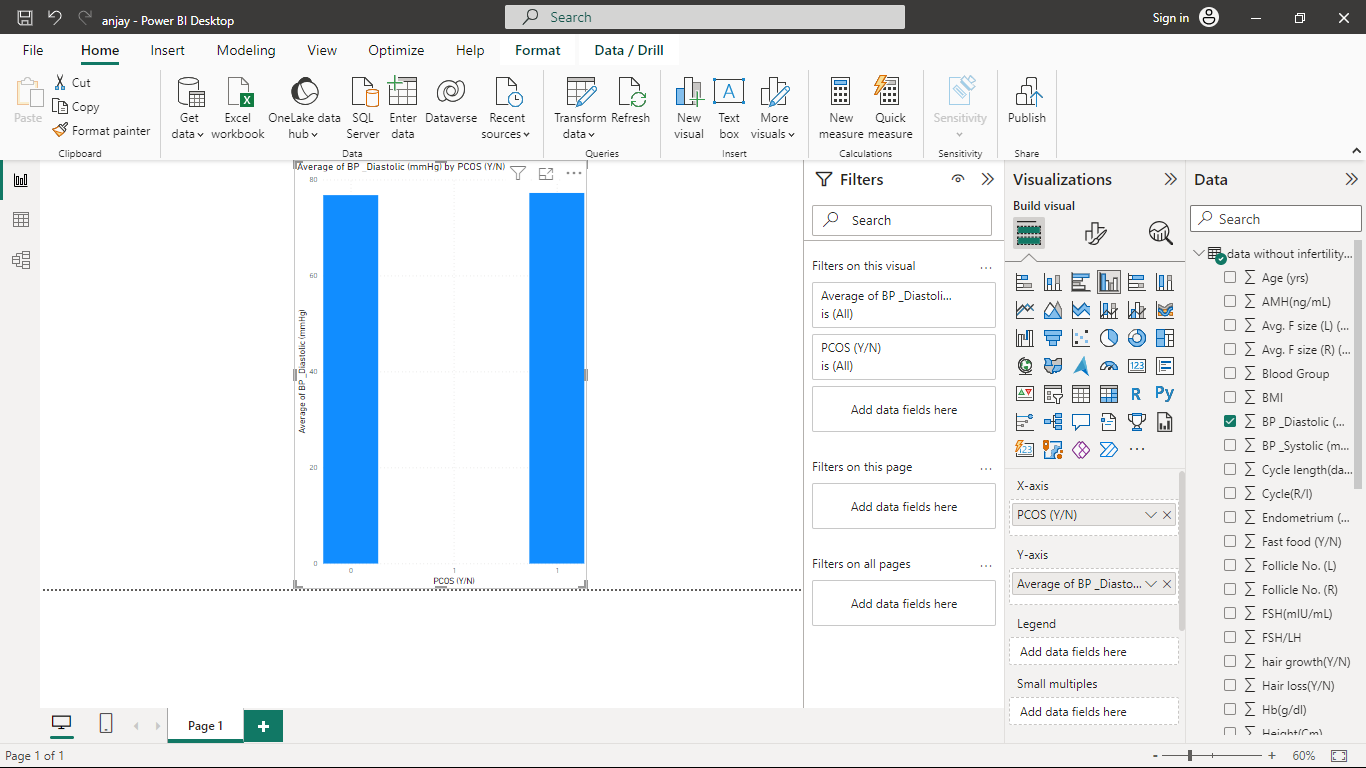
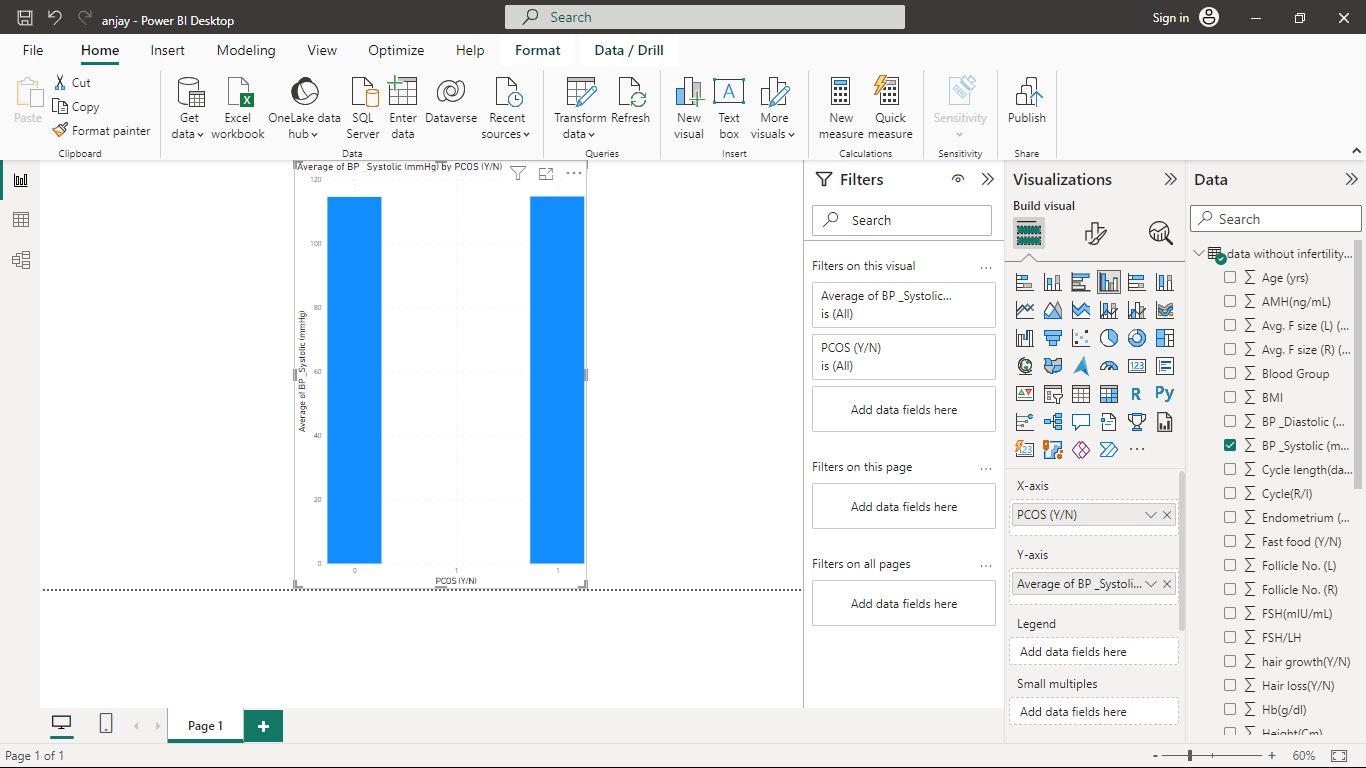
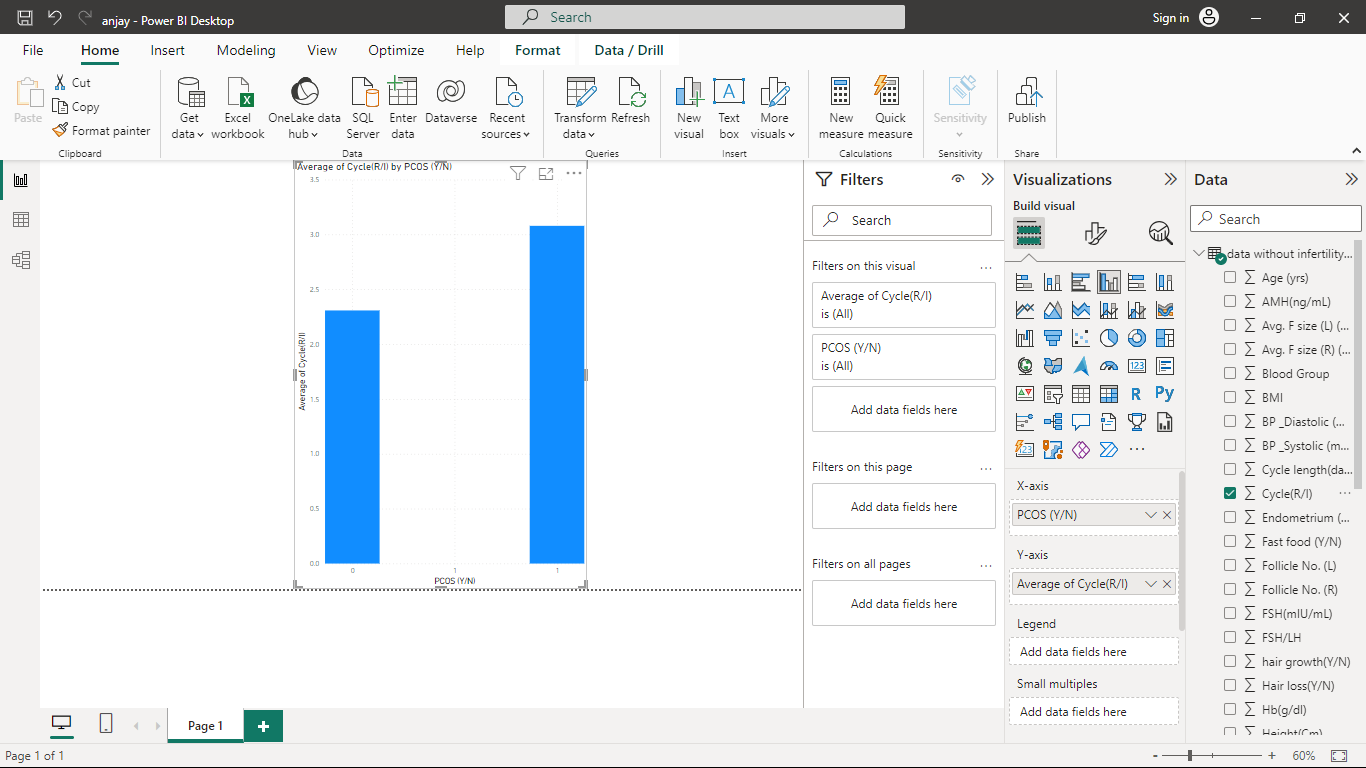
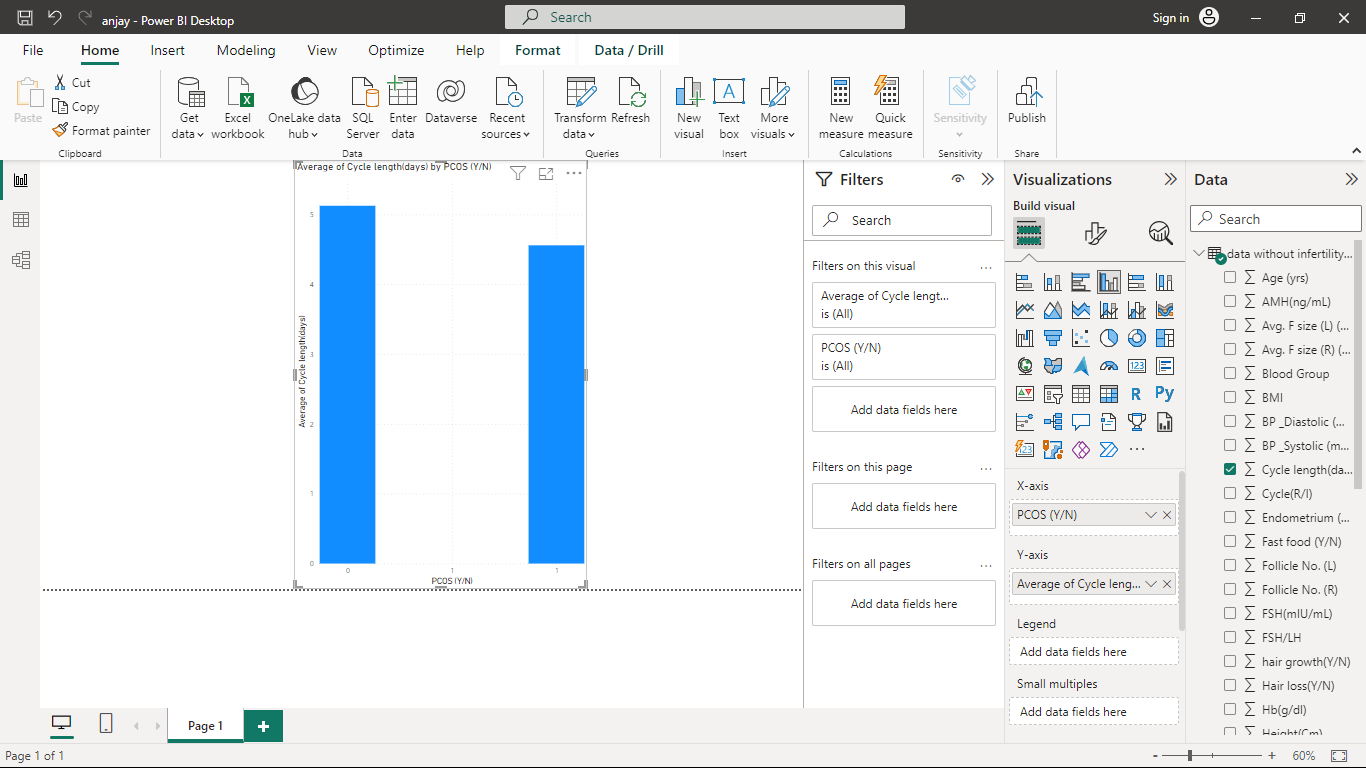
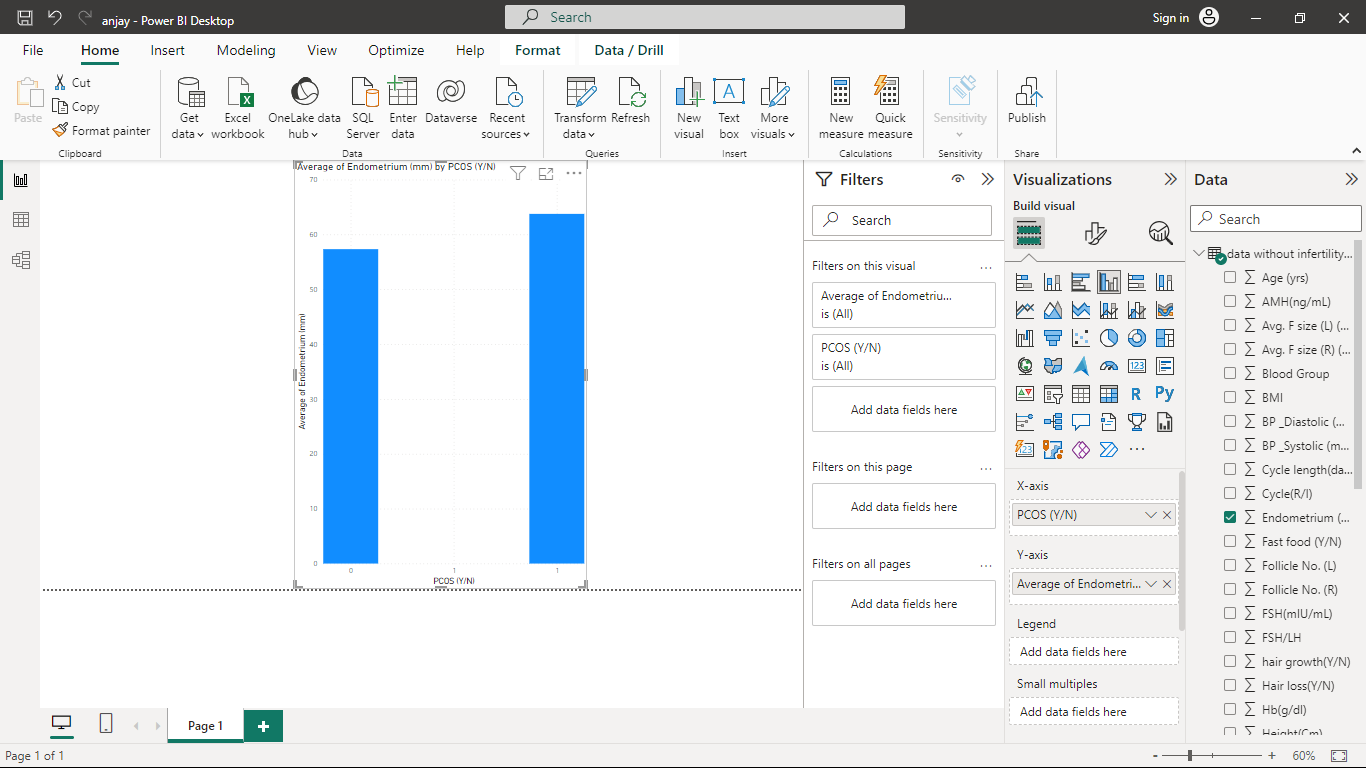
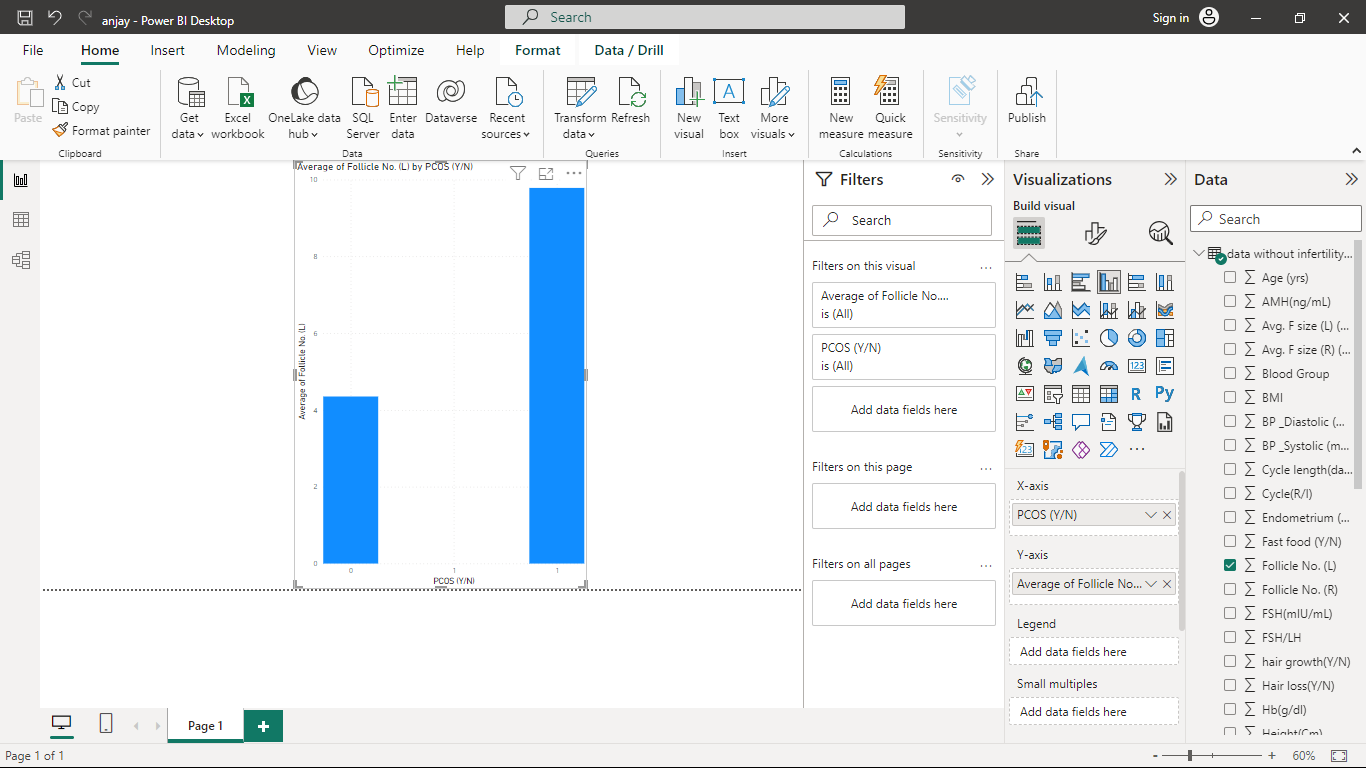
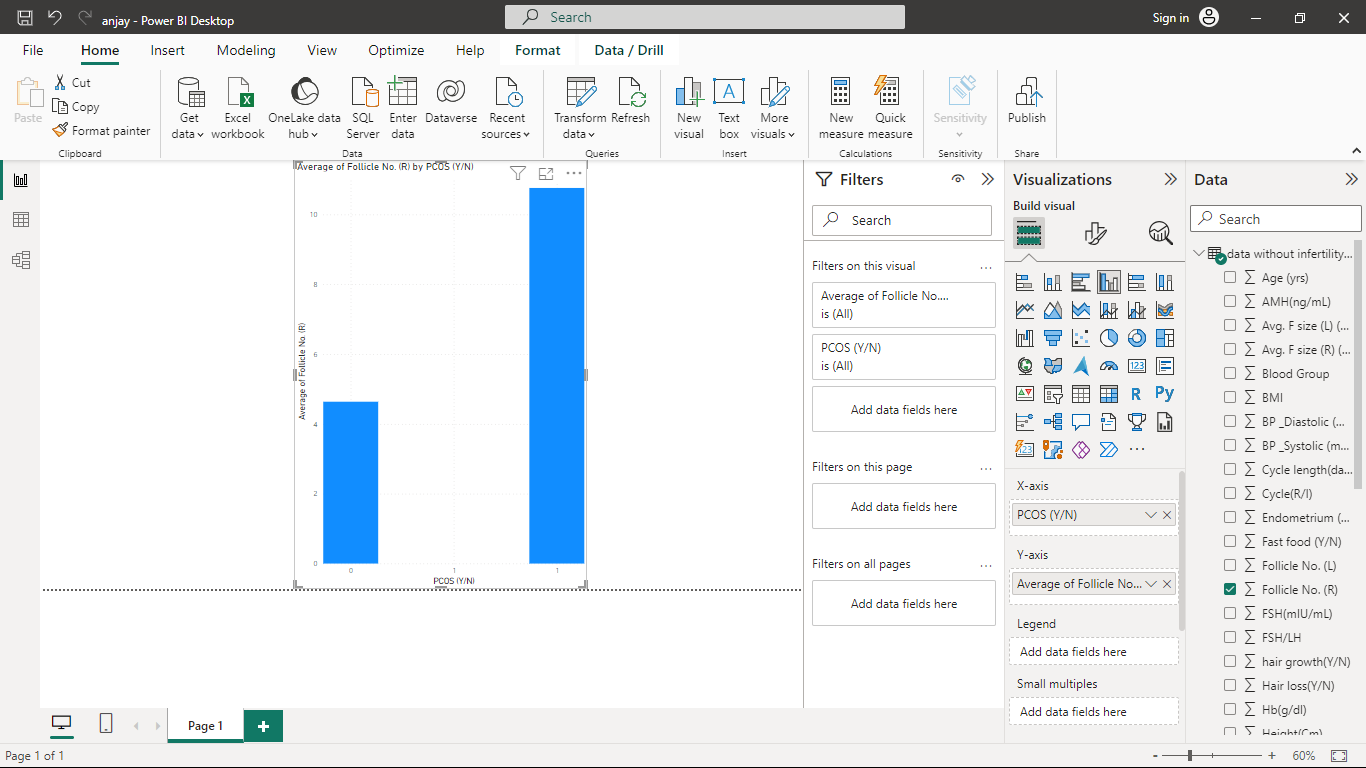
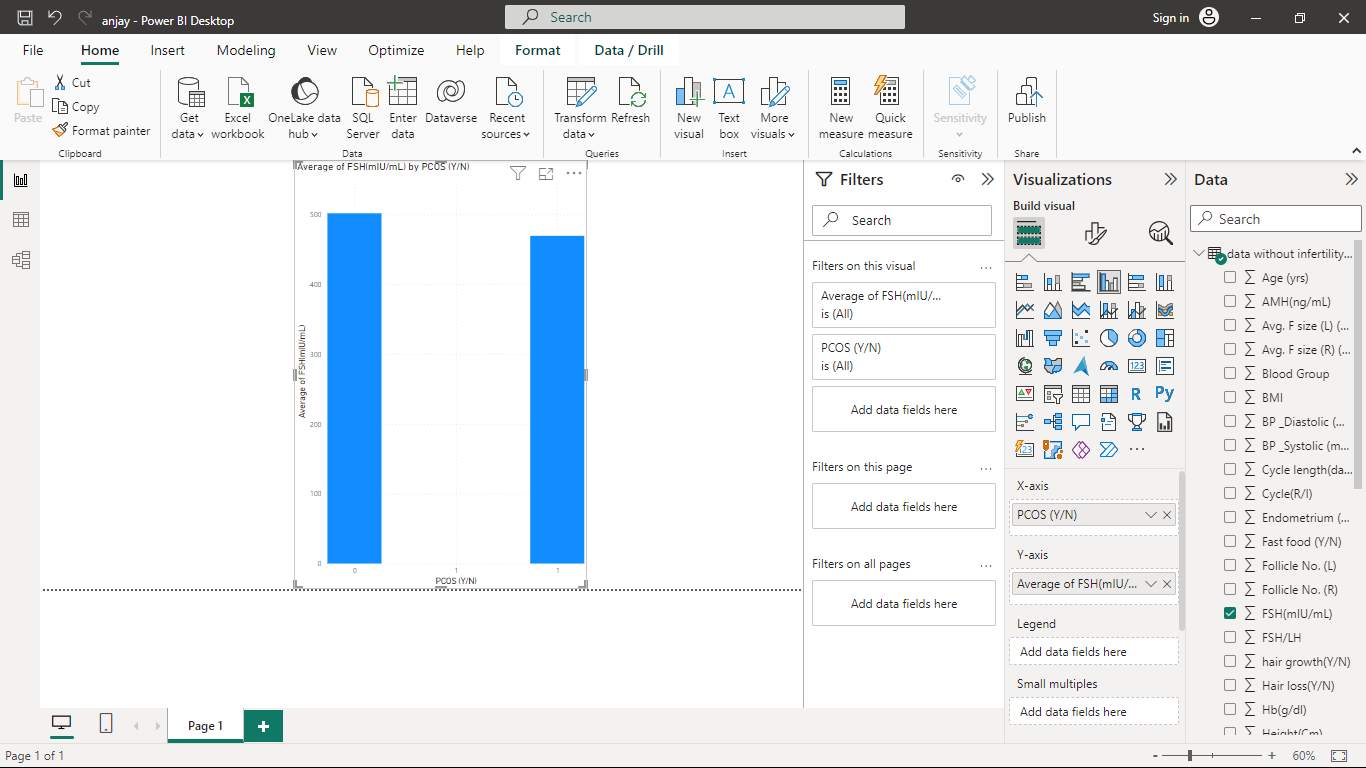
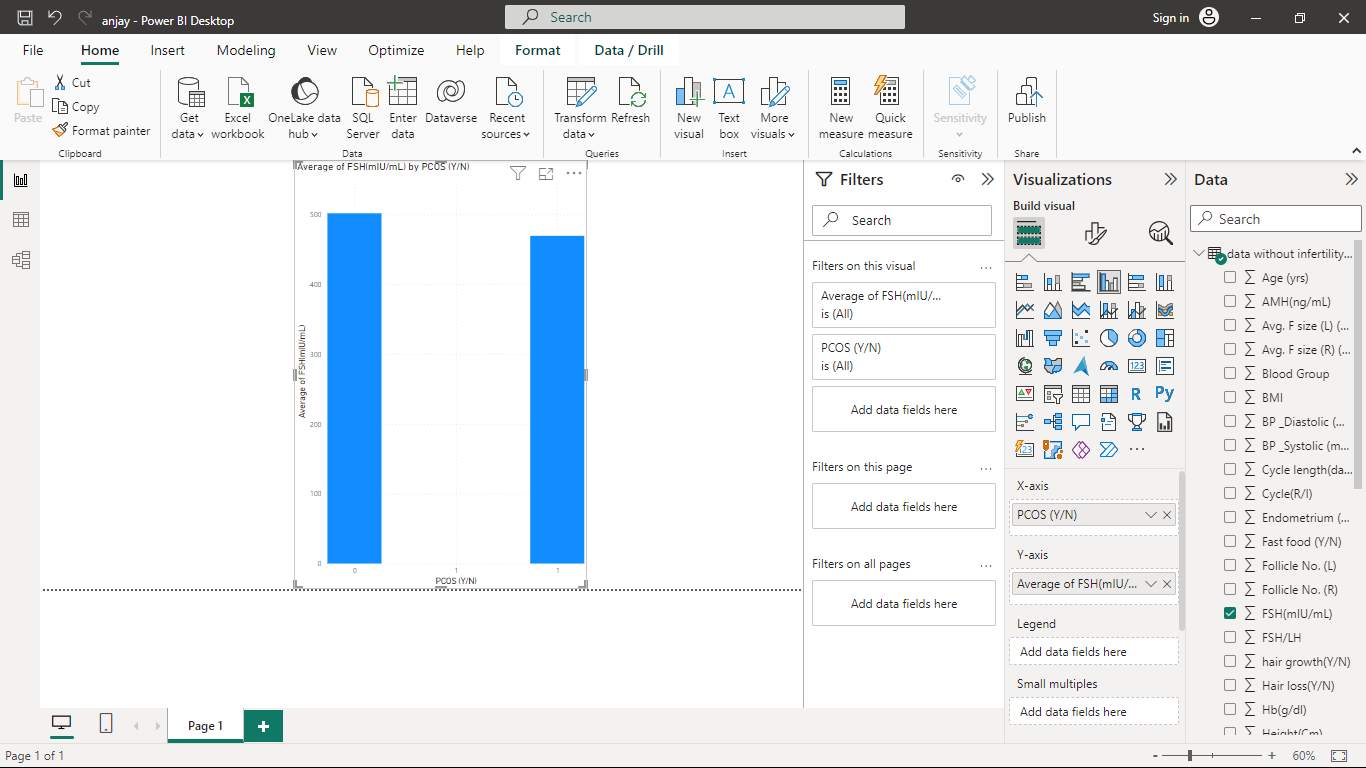
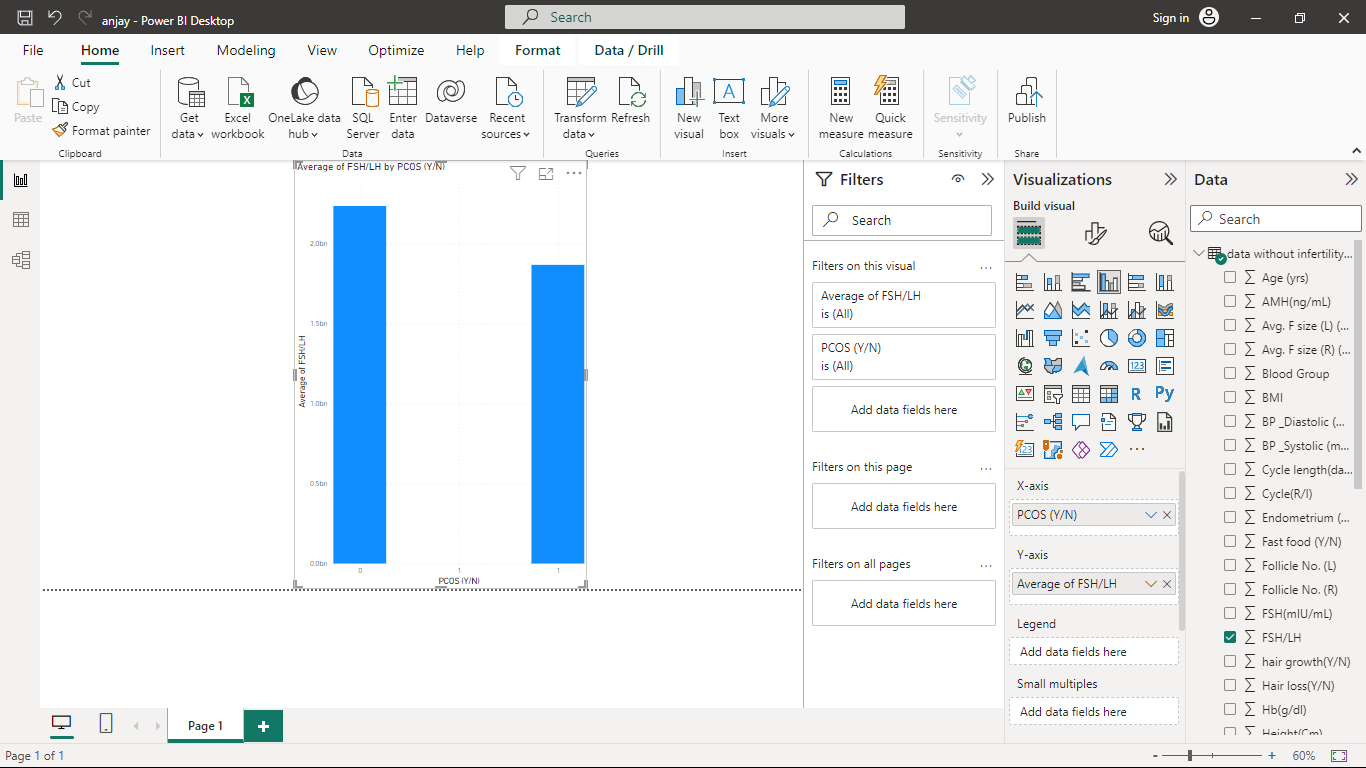
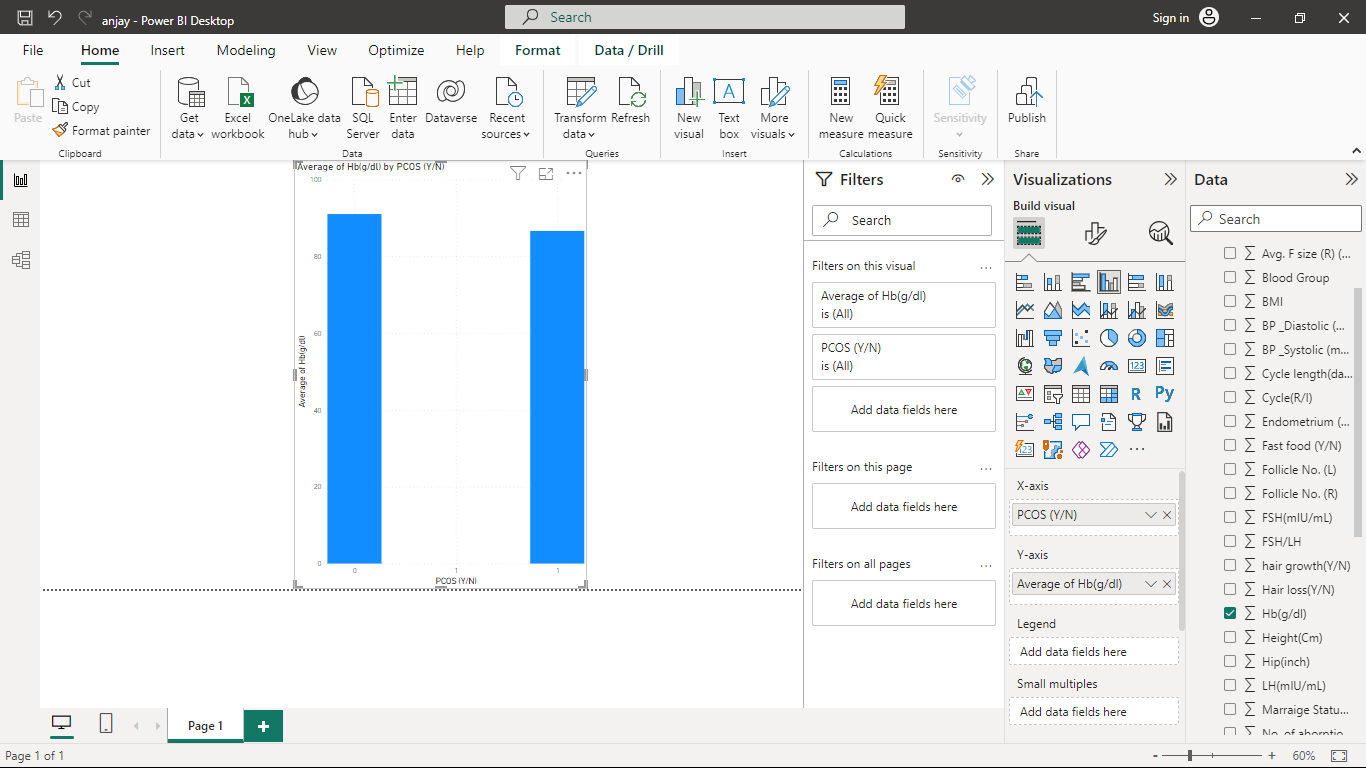
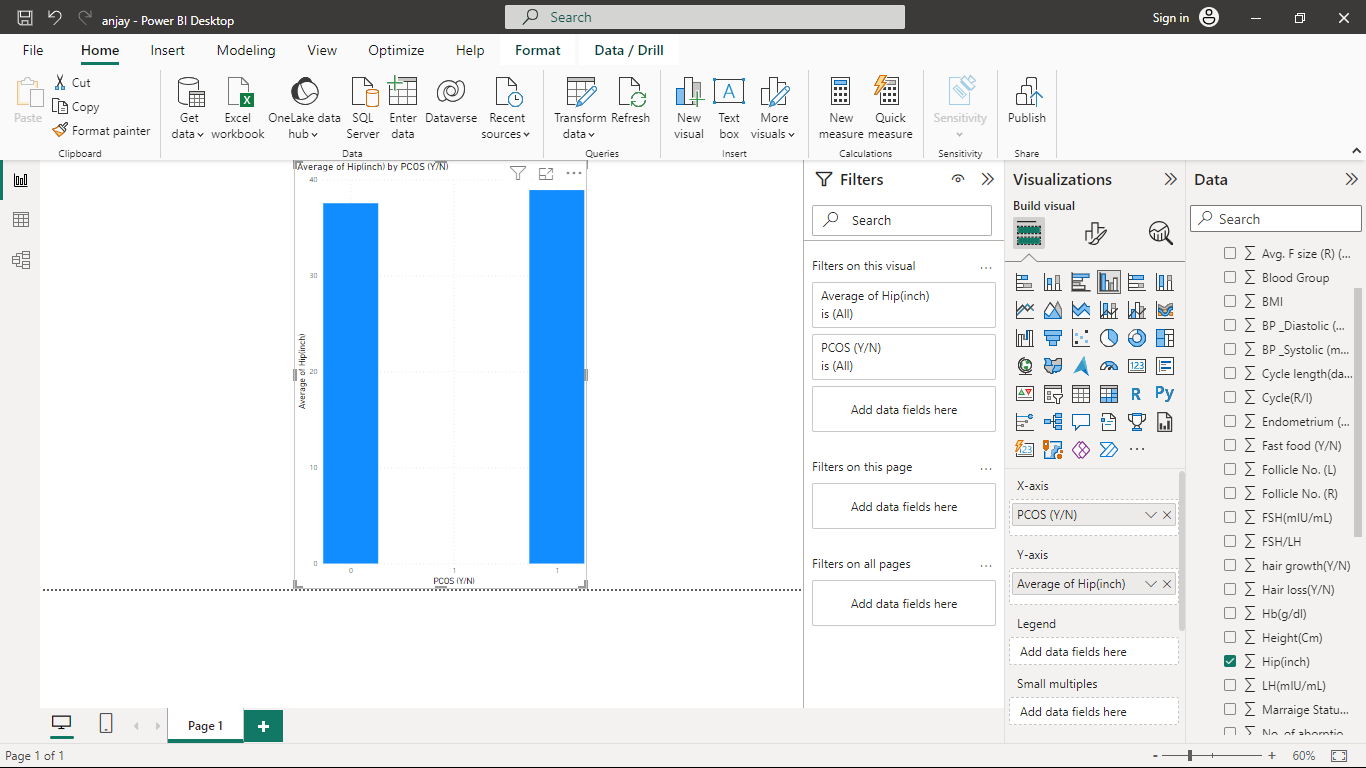
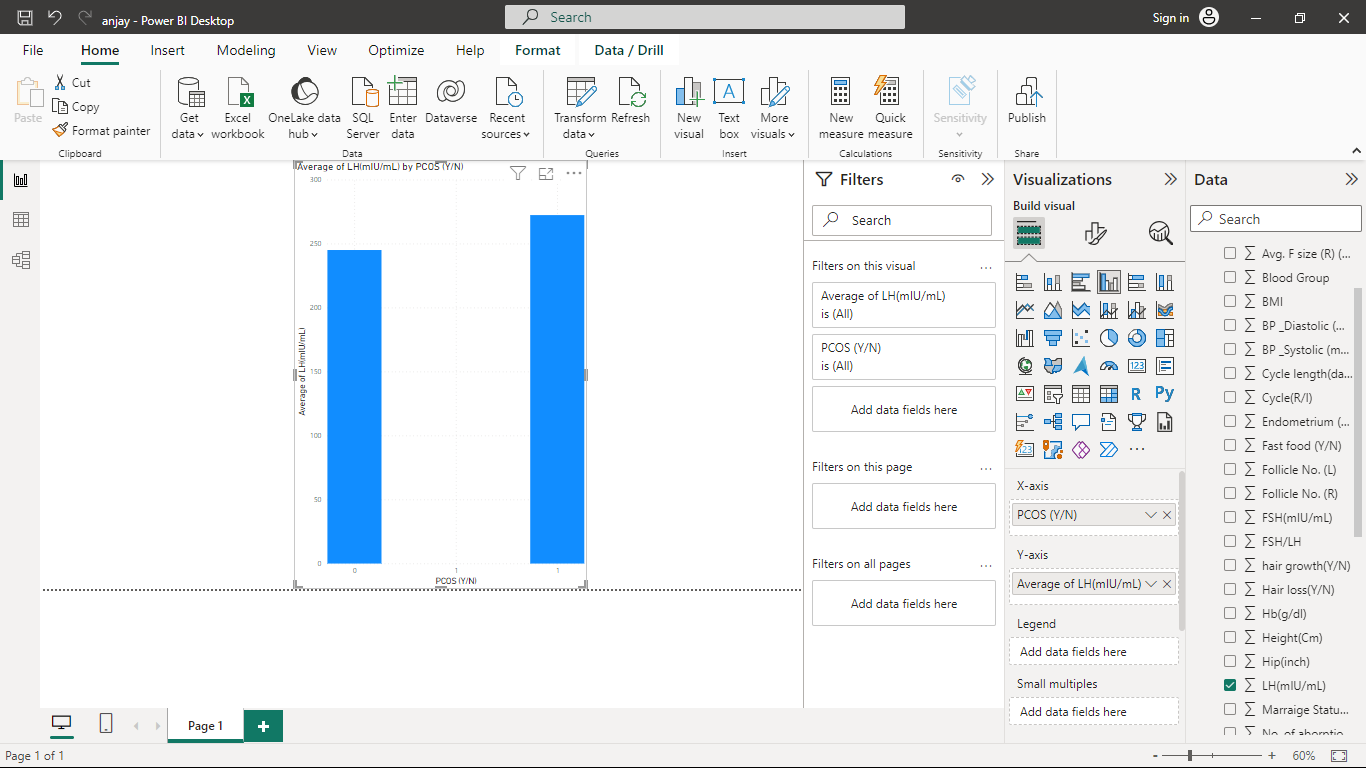
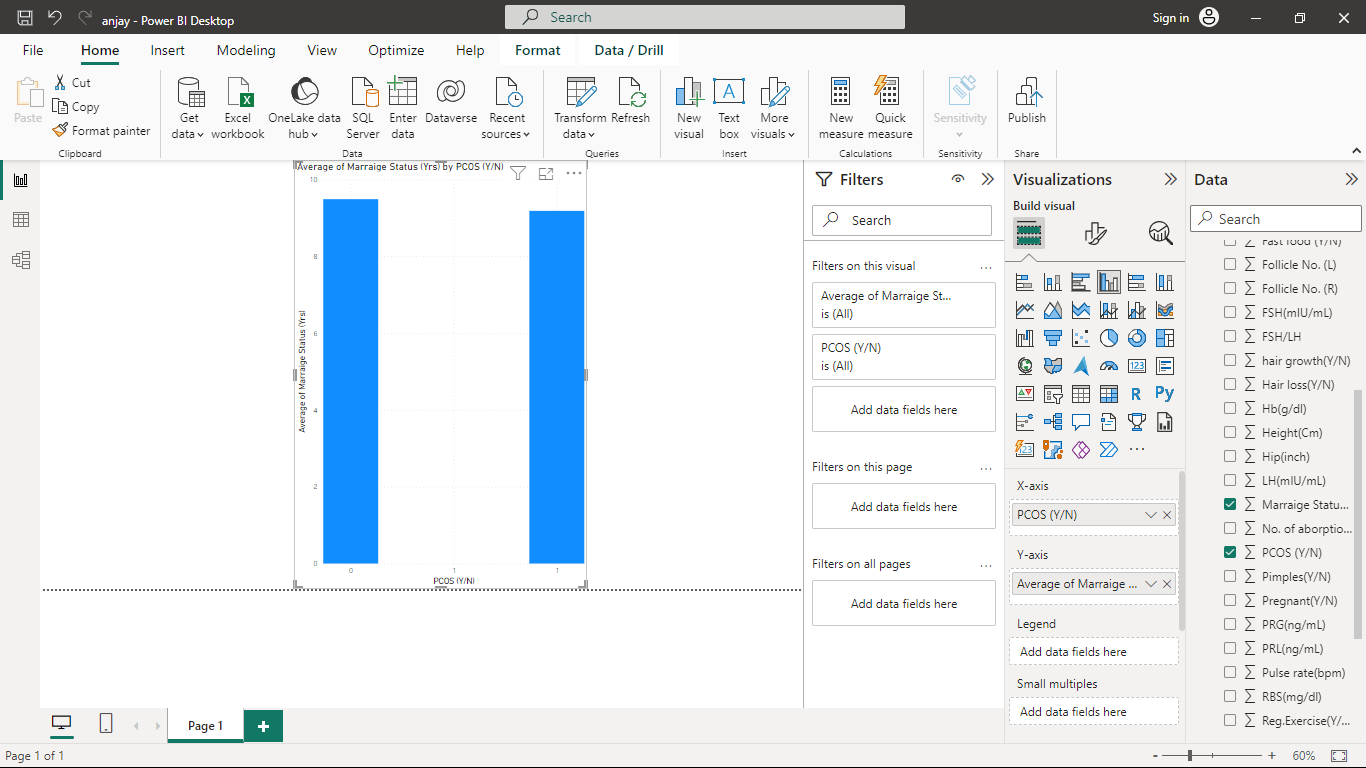
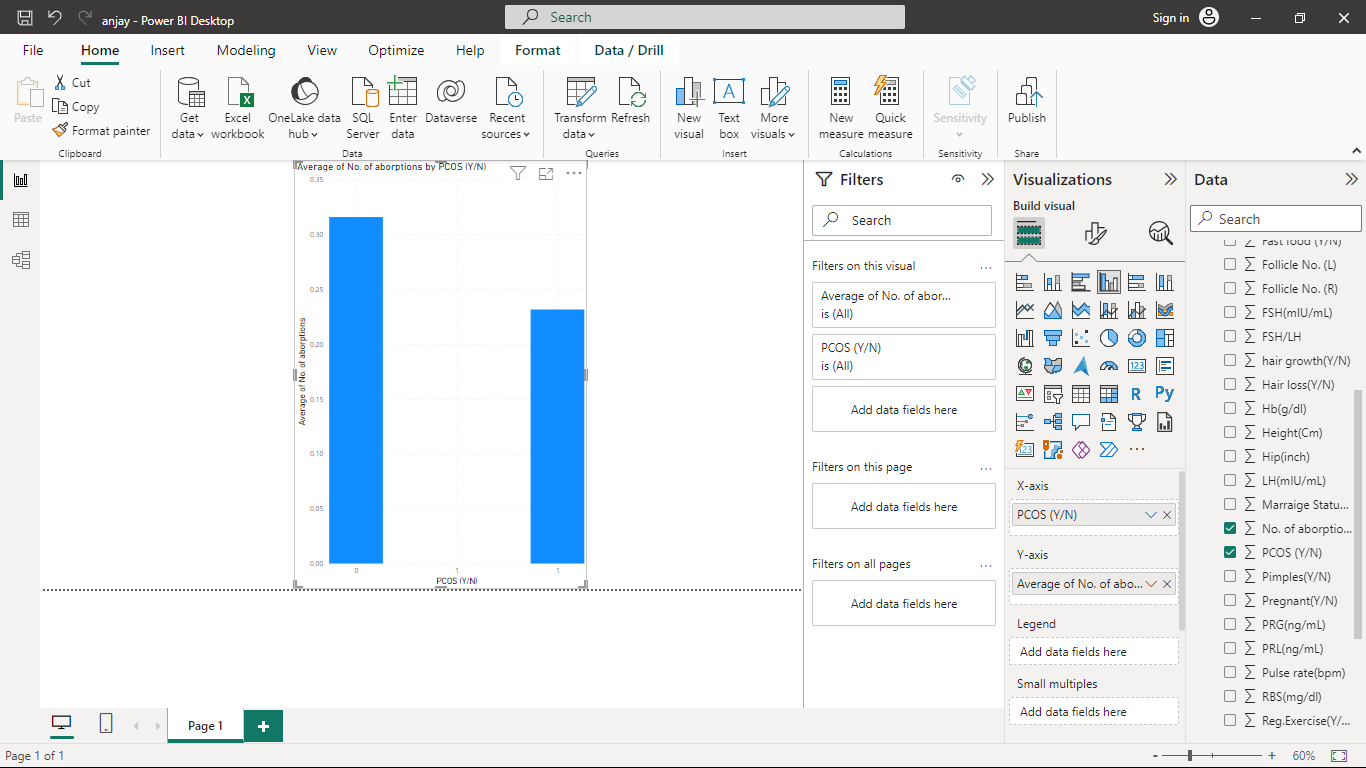
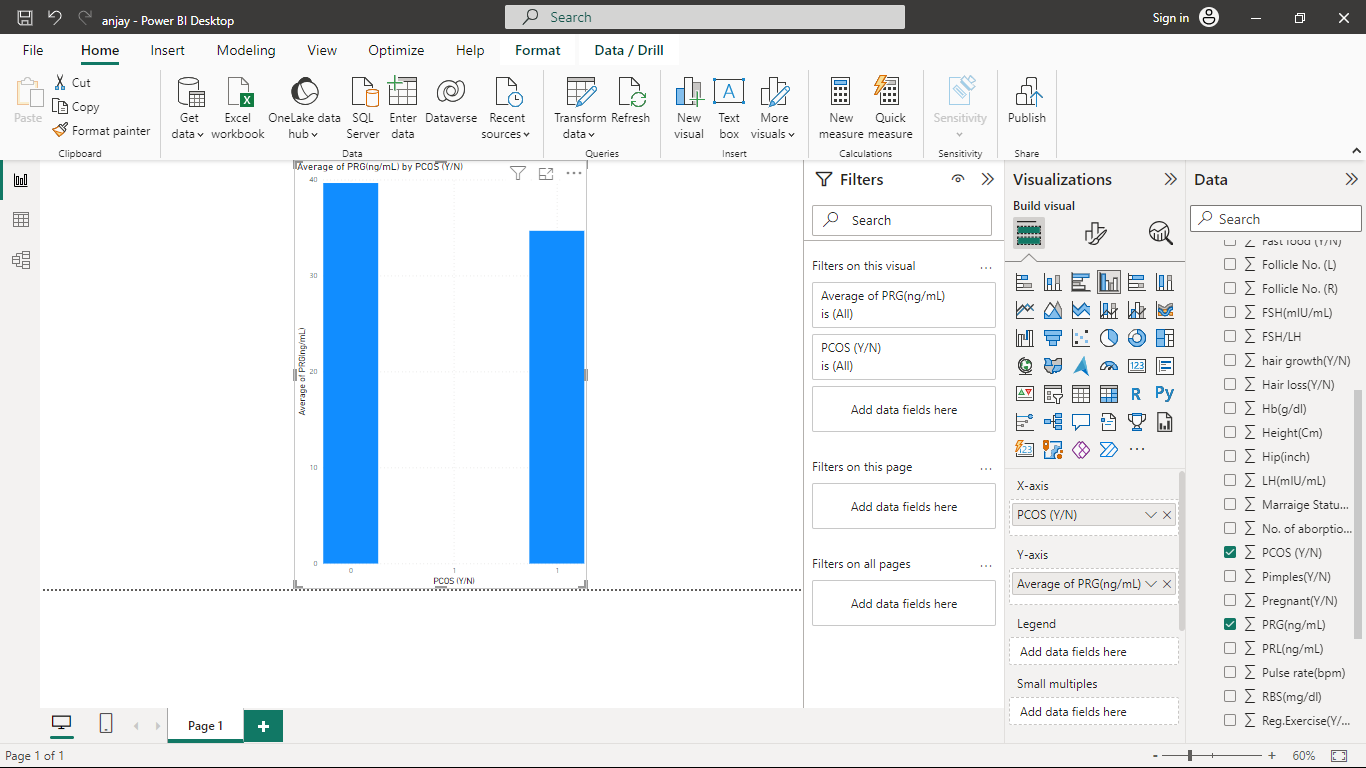
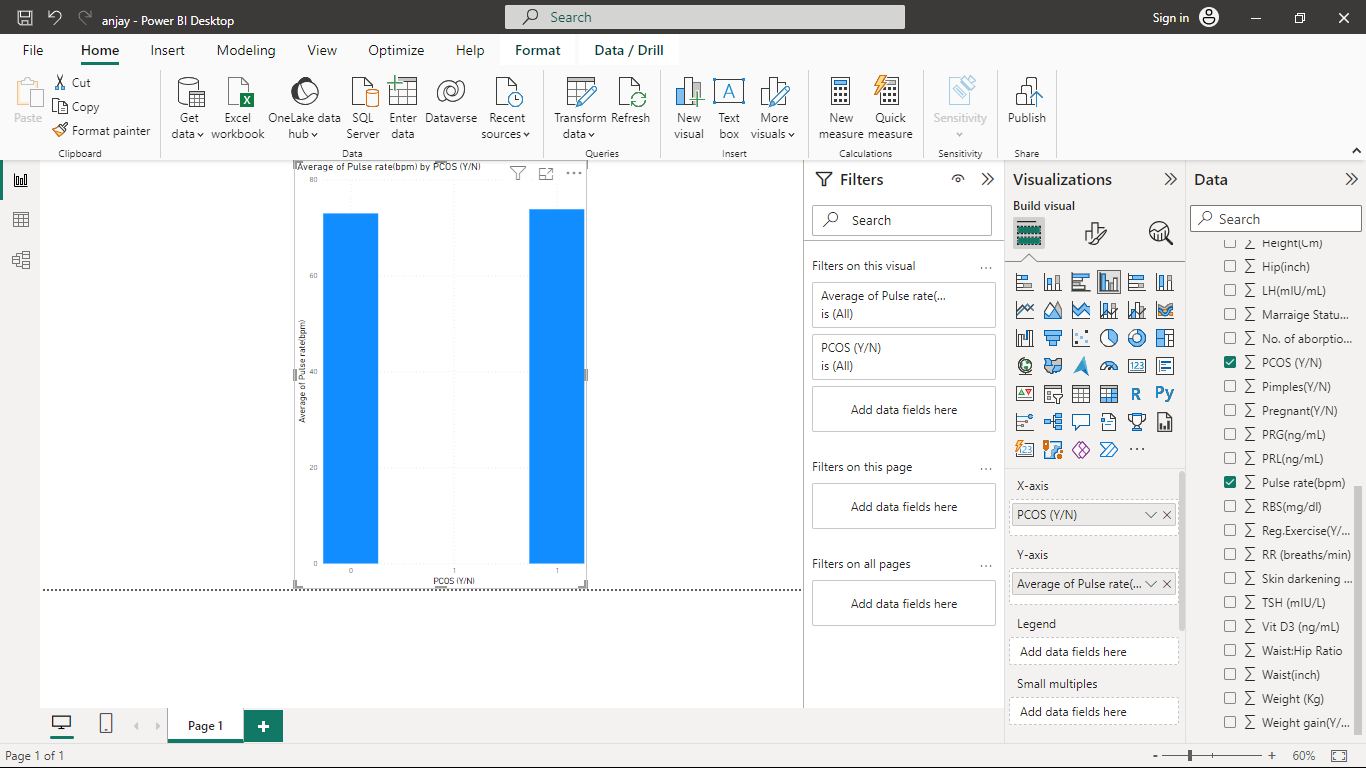
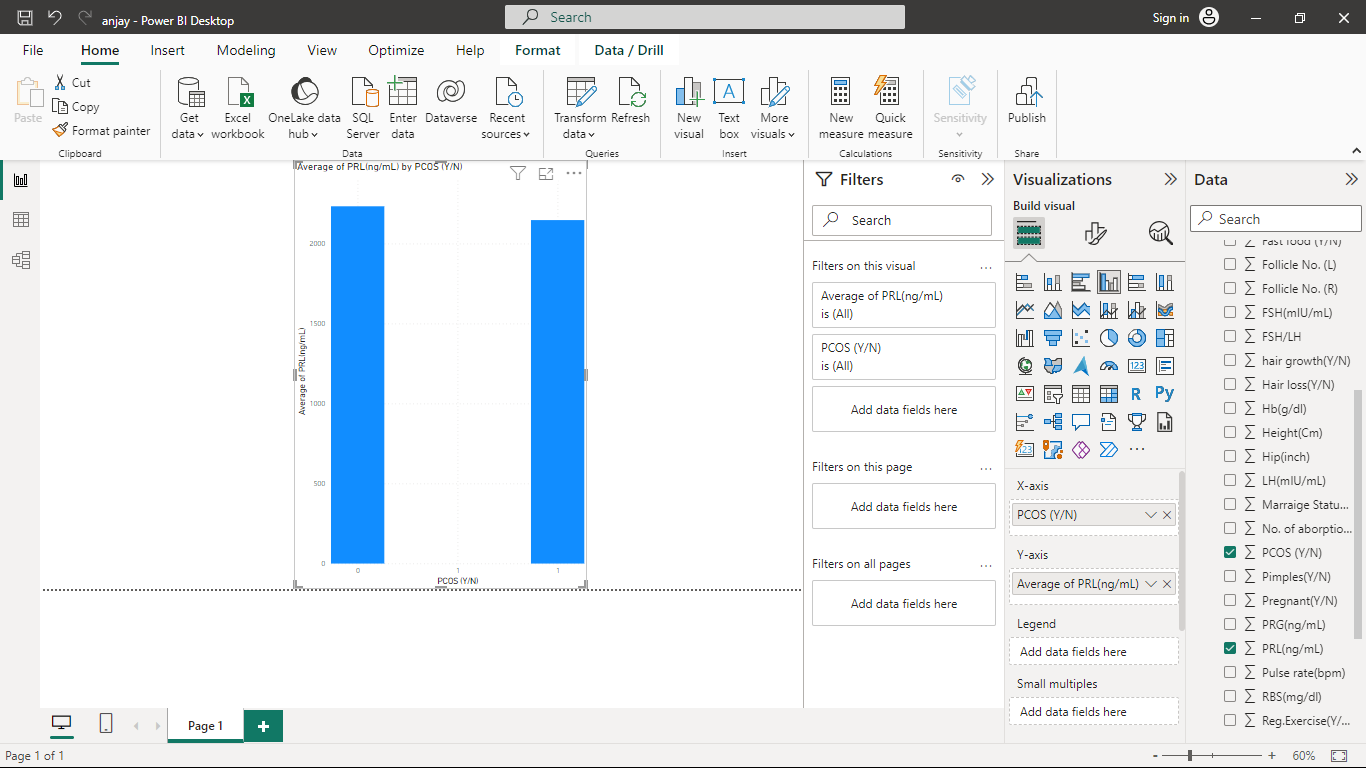
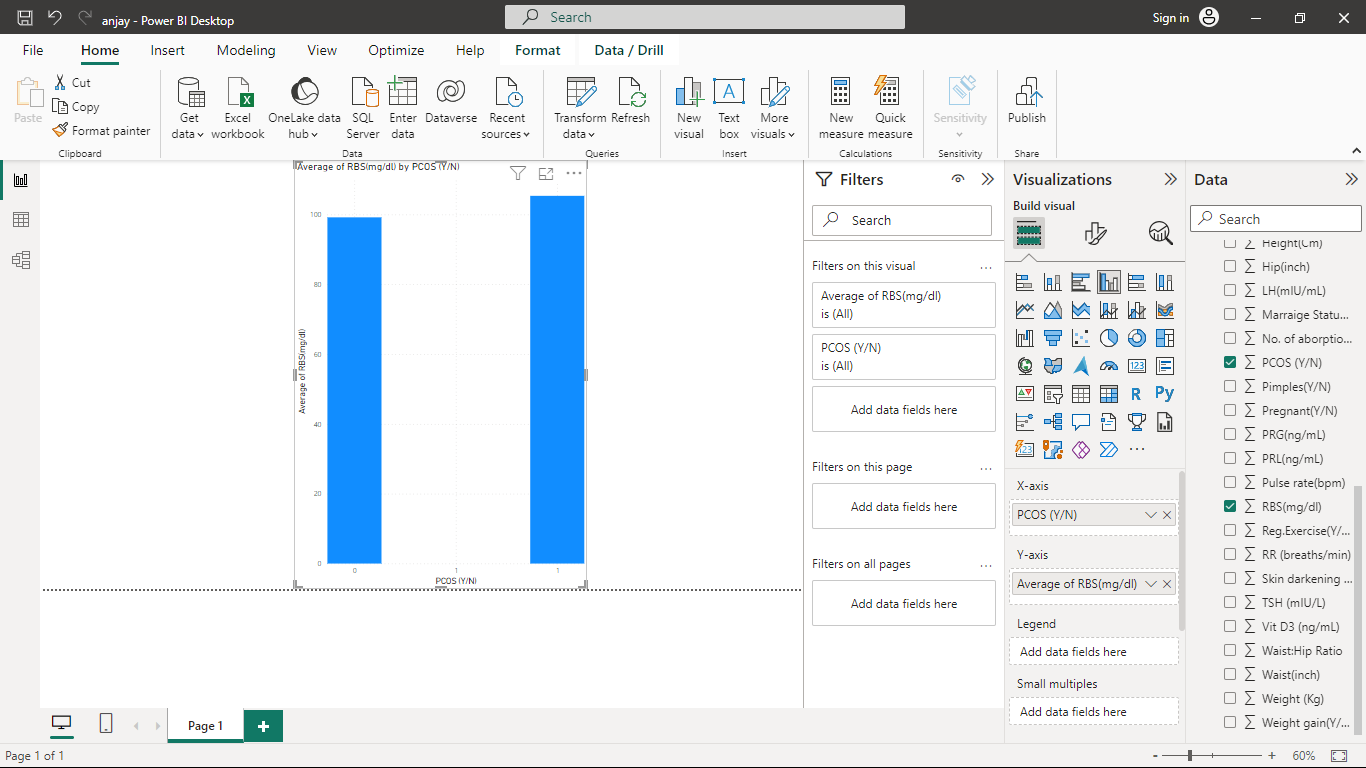
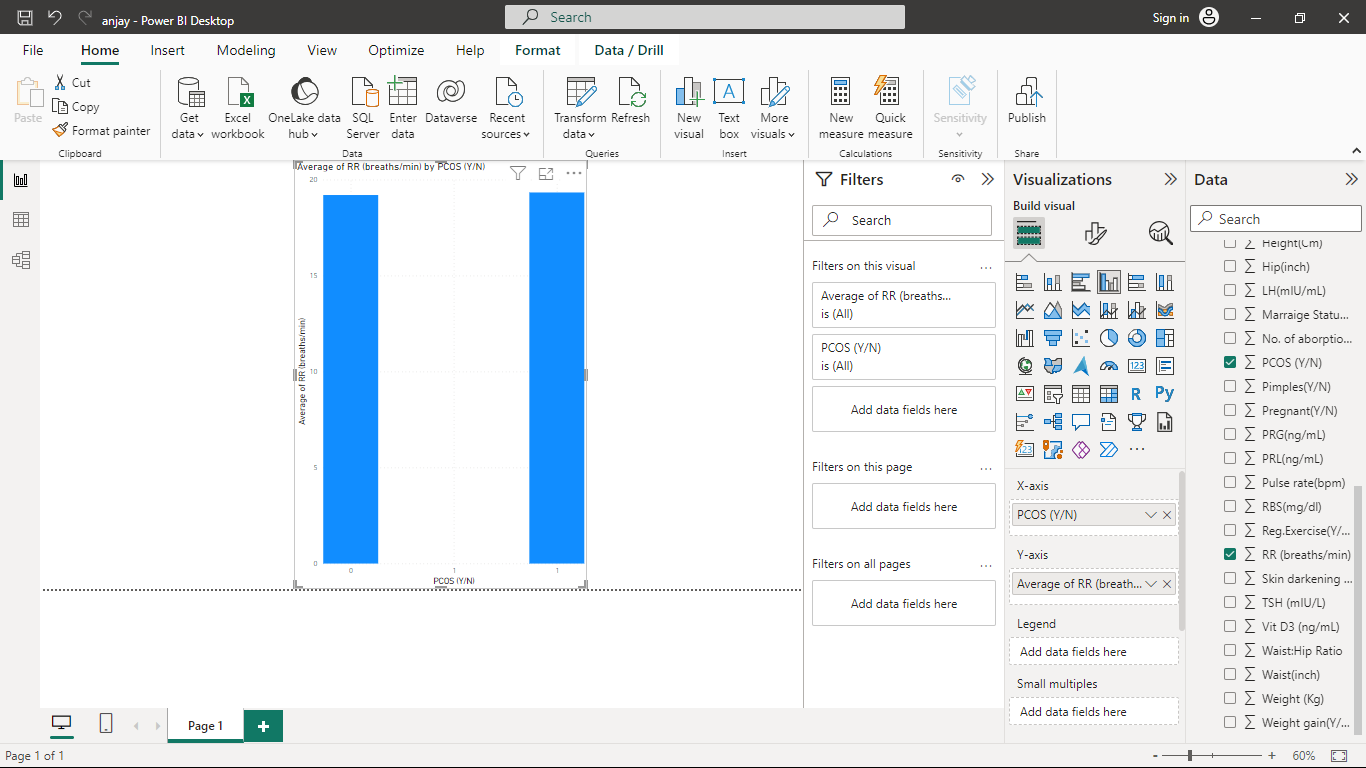
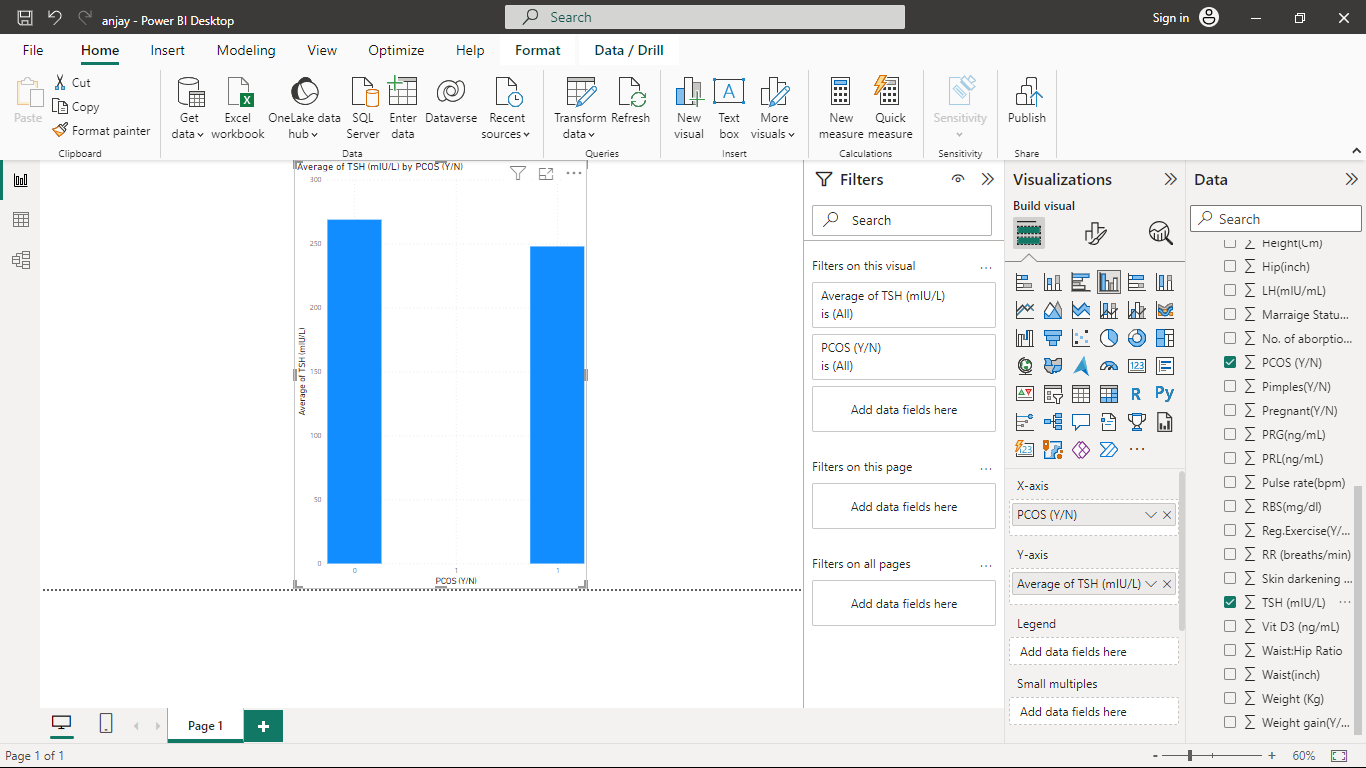
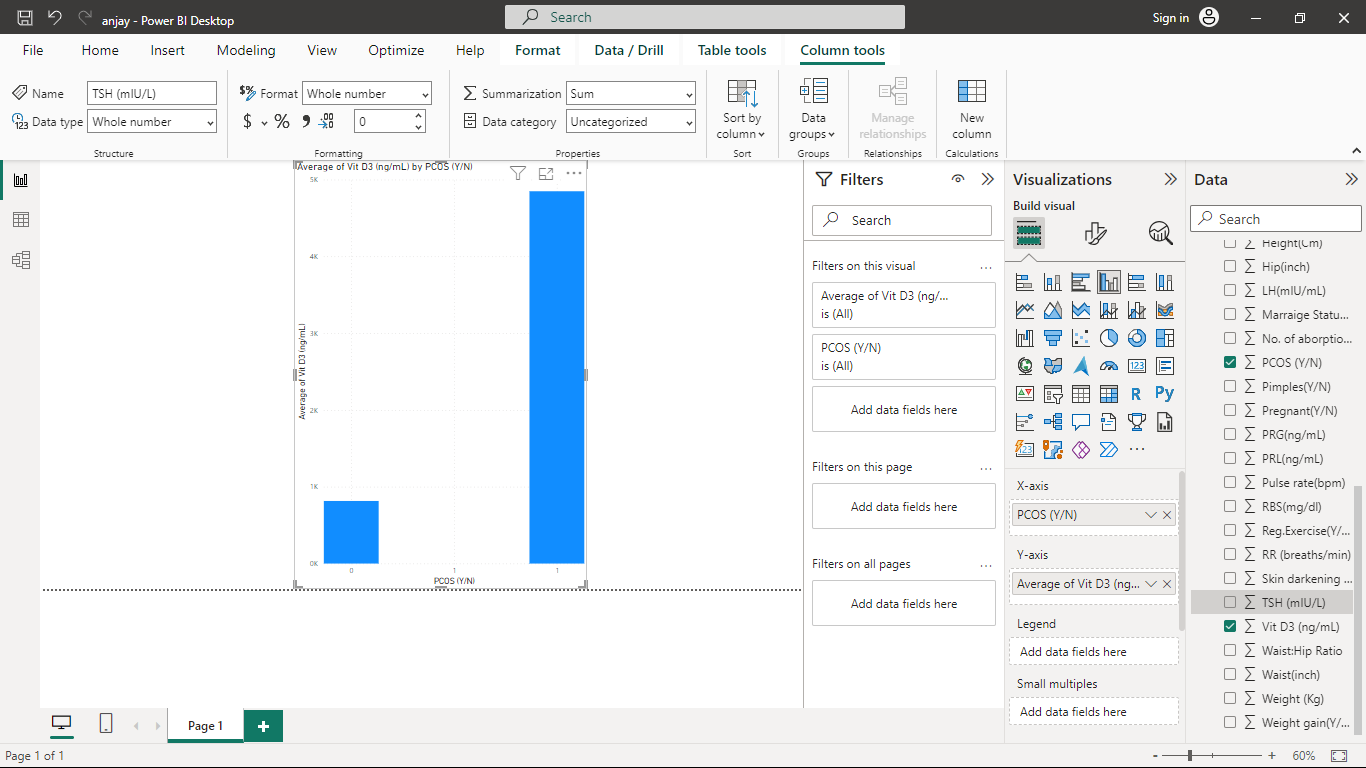
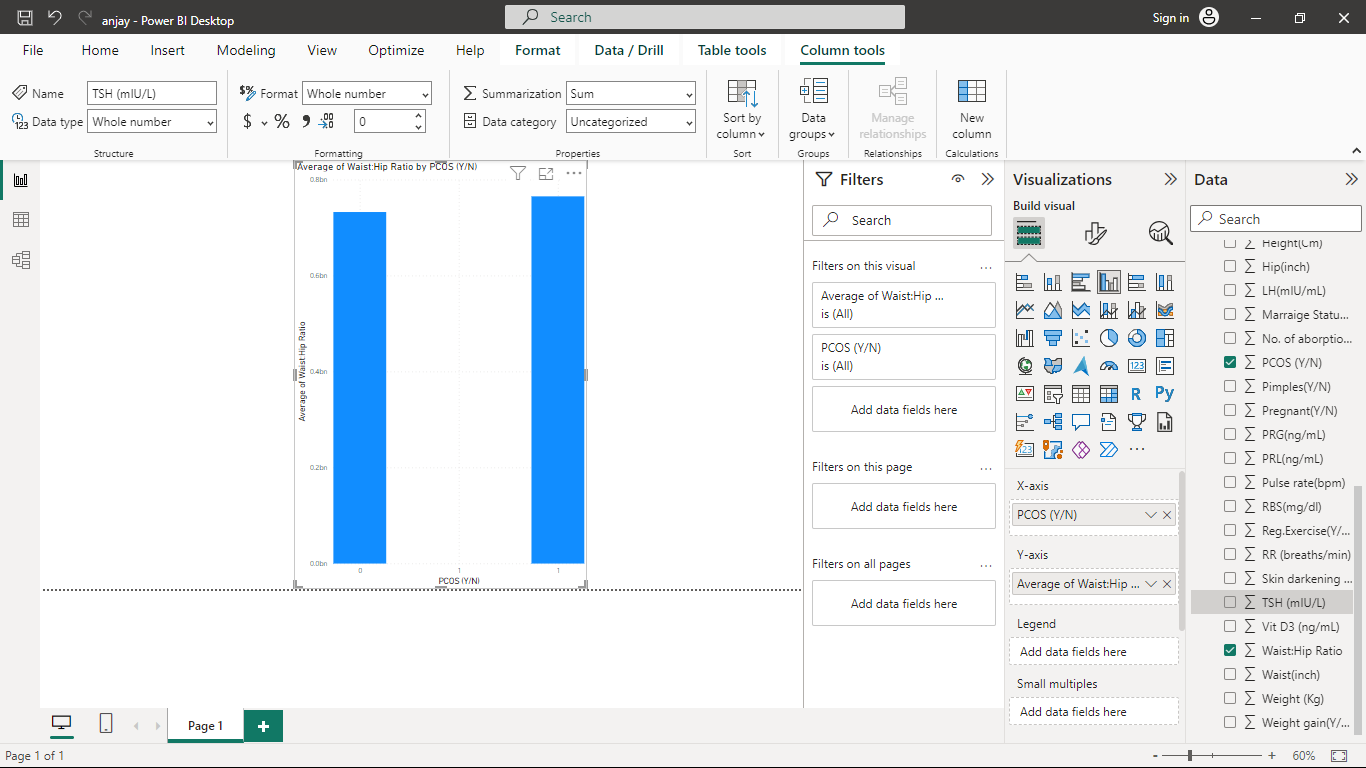
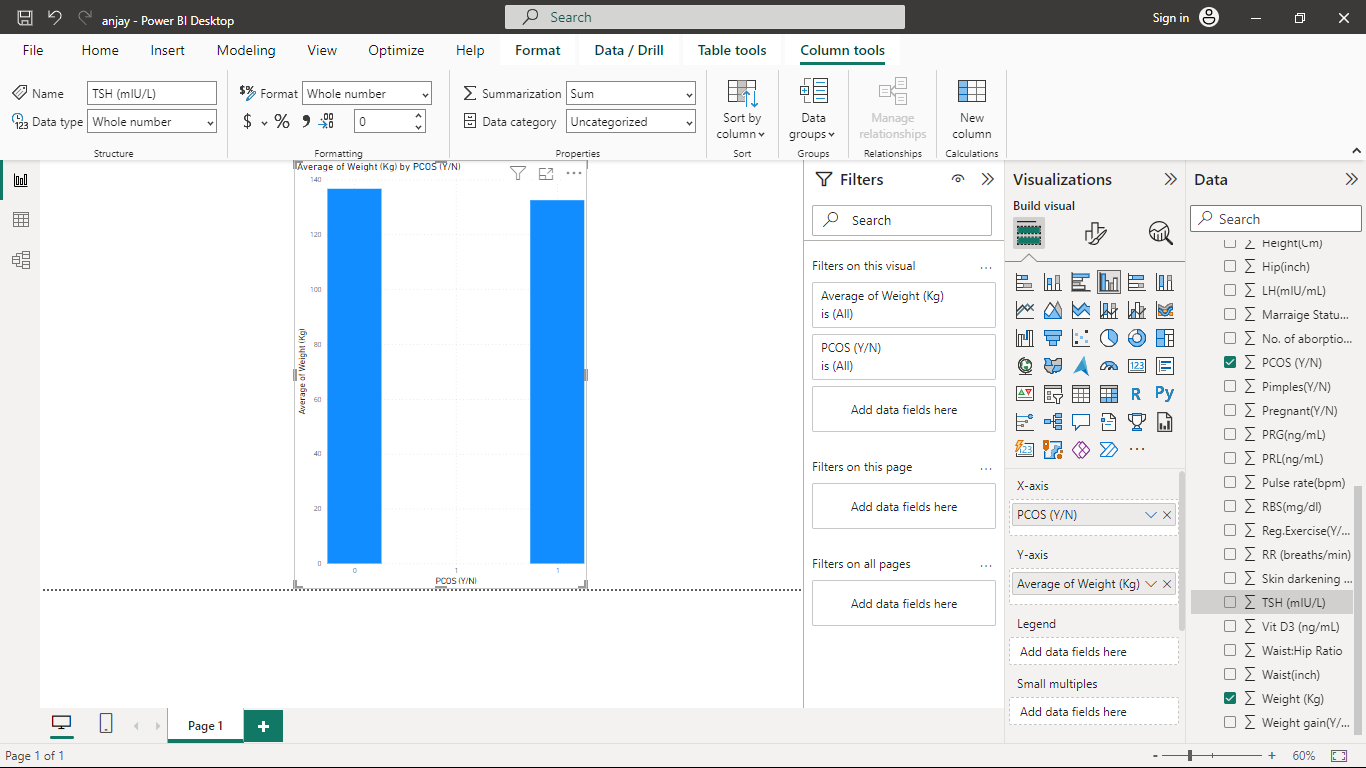
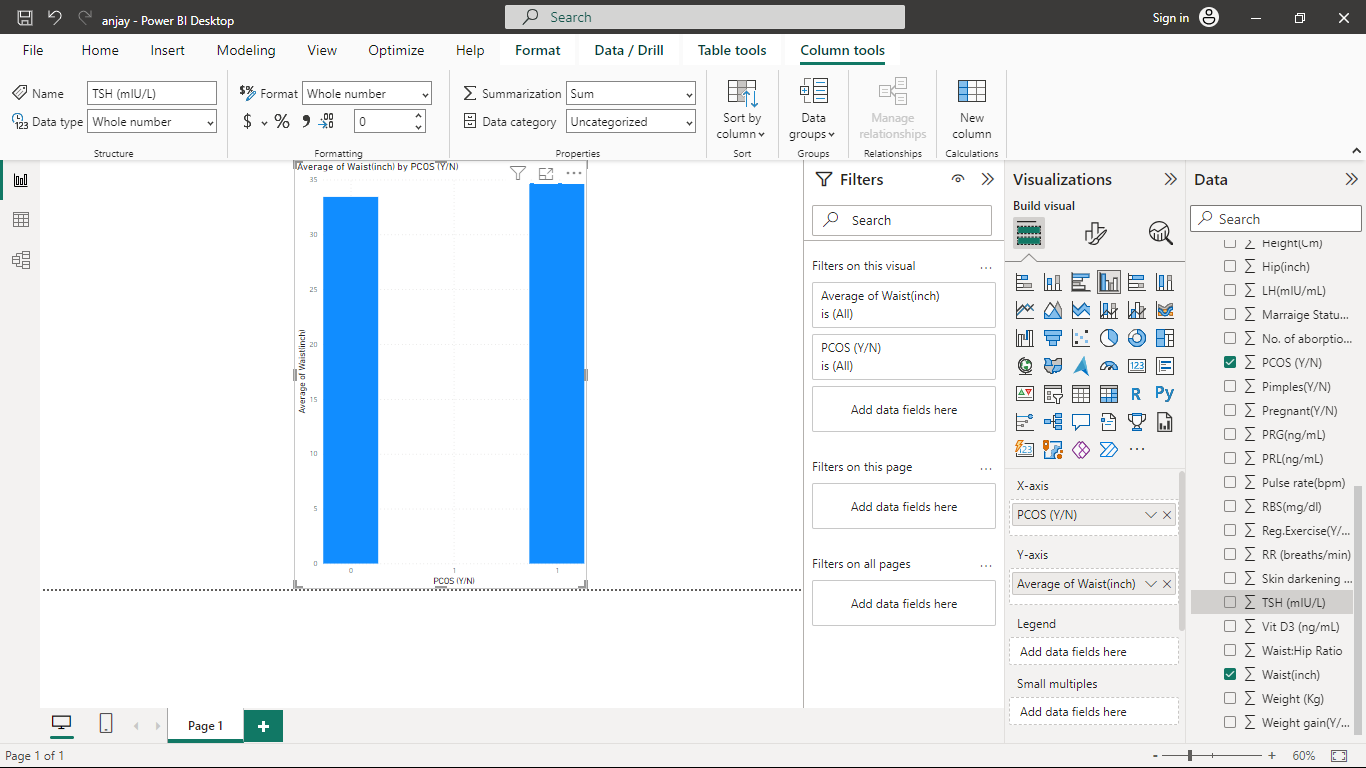


**Transformasi data**, yaitu dari tipe object menjadi numerik.



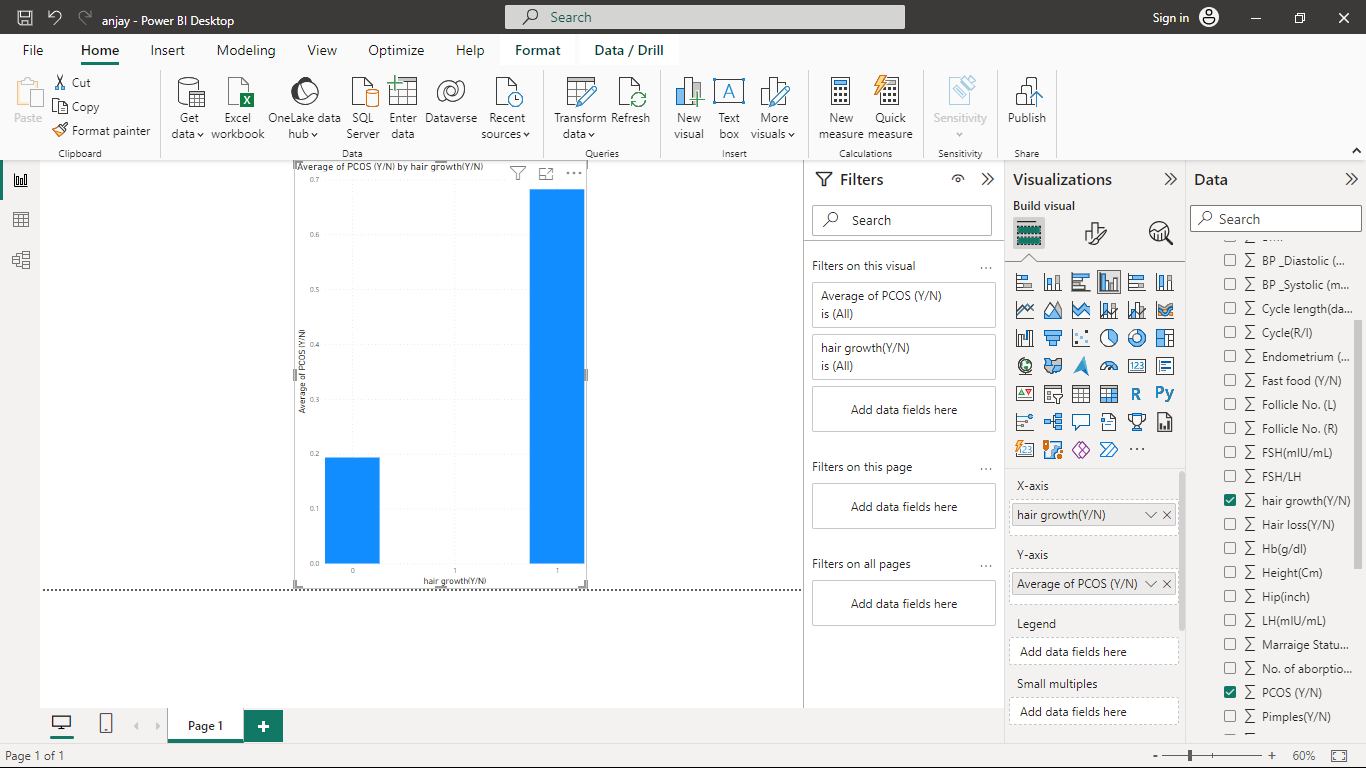
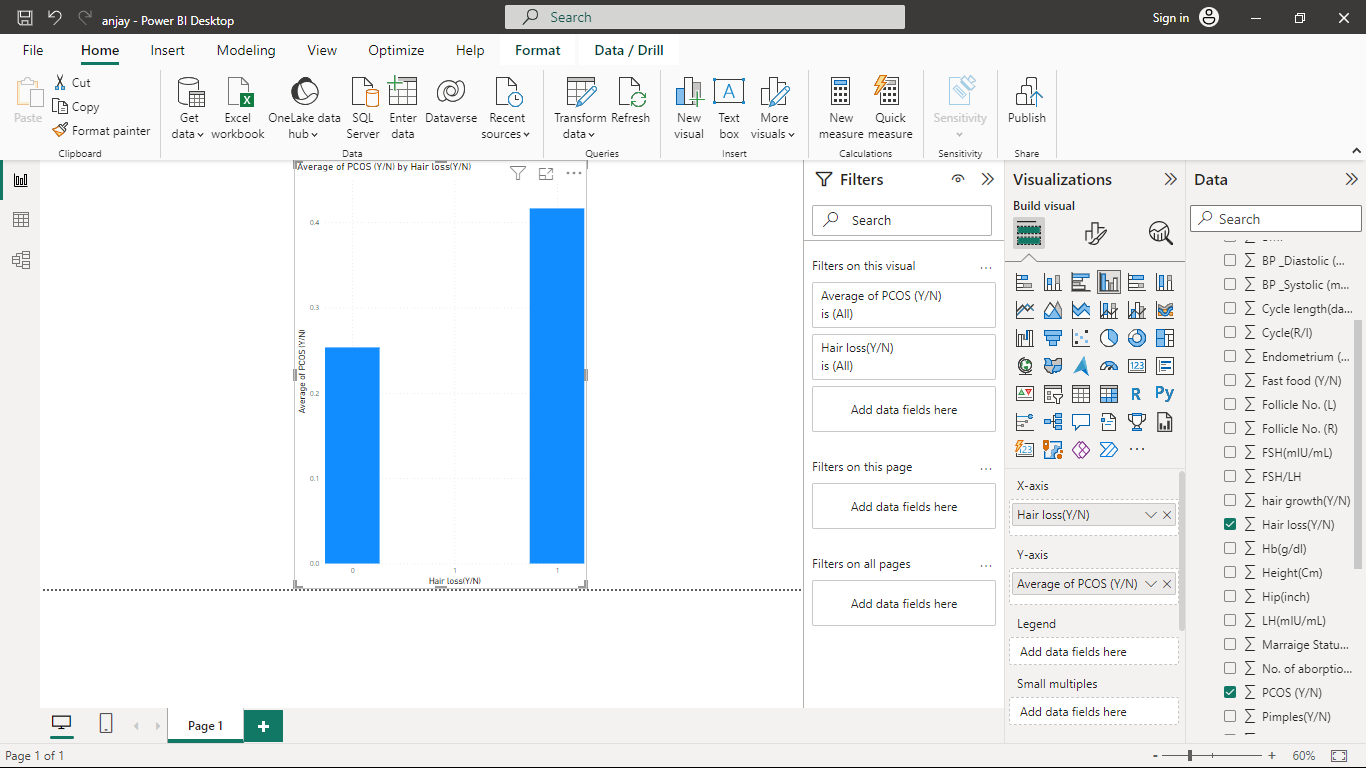
**Proporsi Sebaran Data**

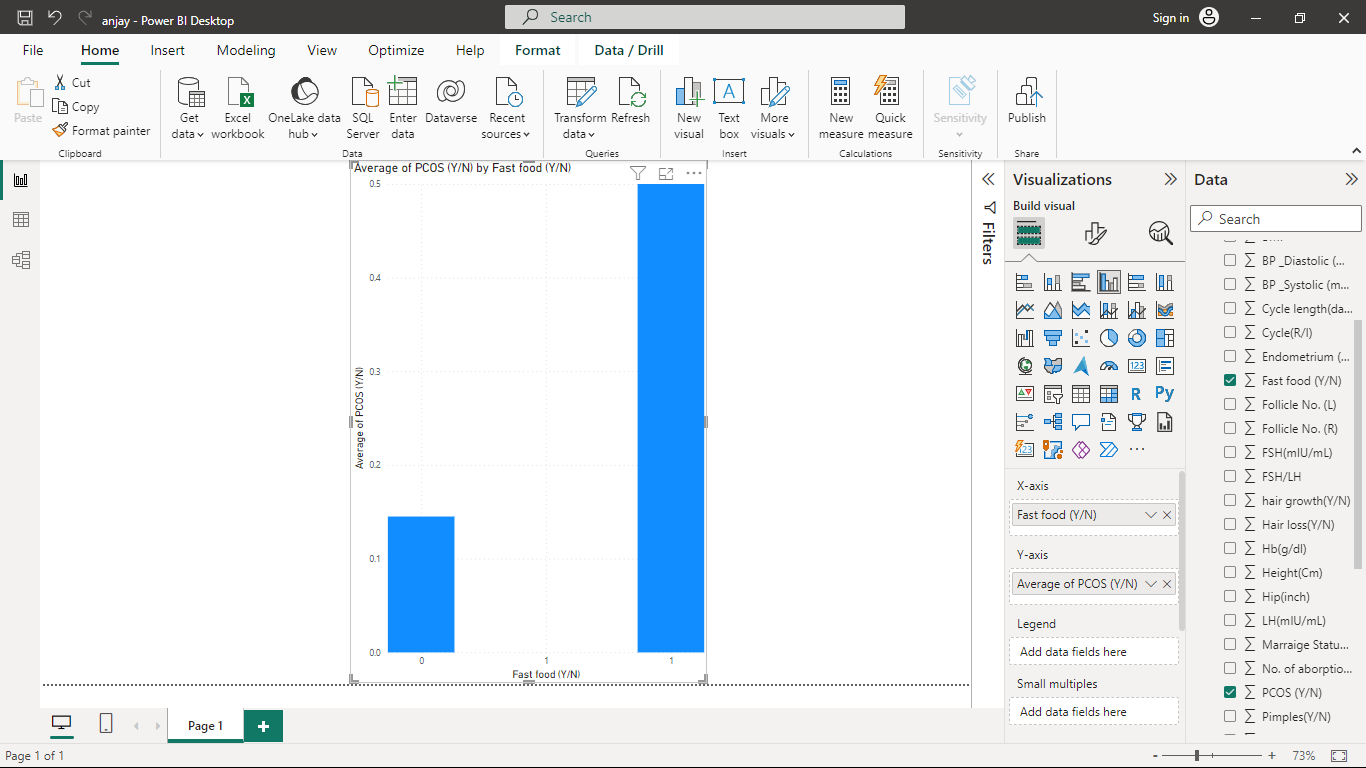
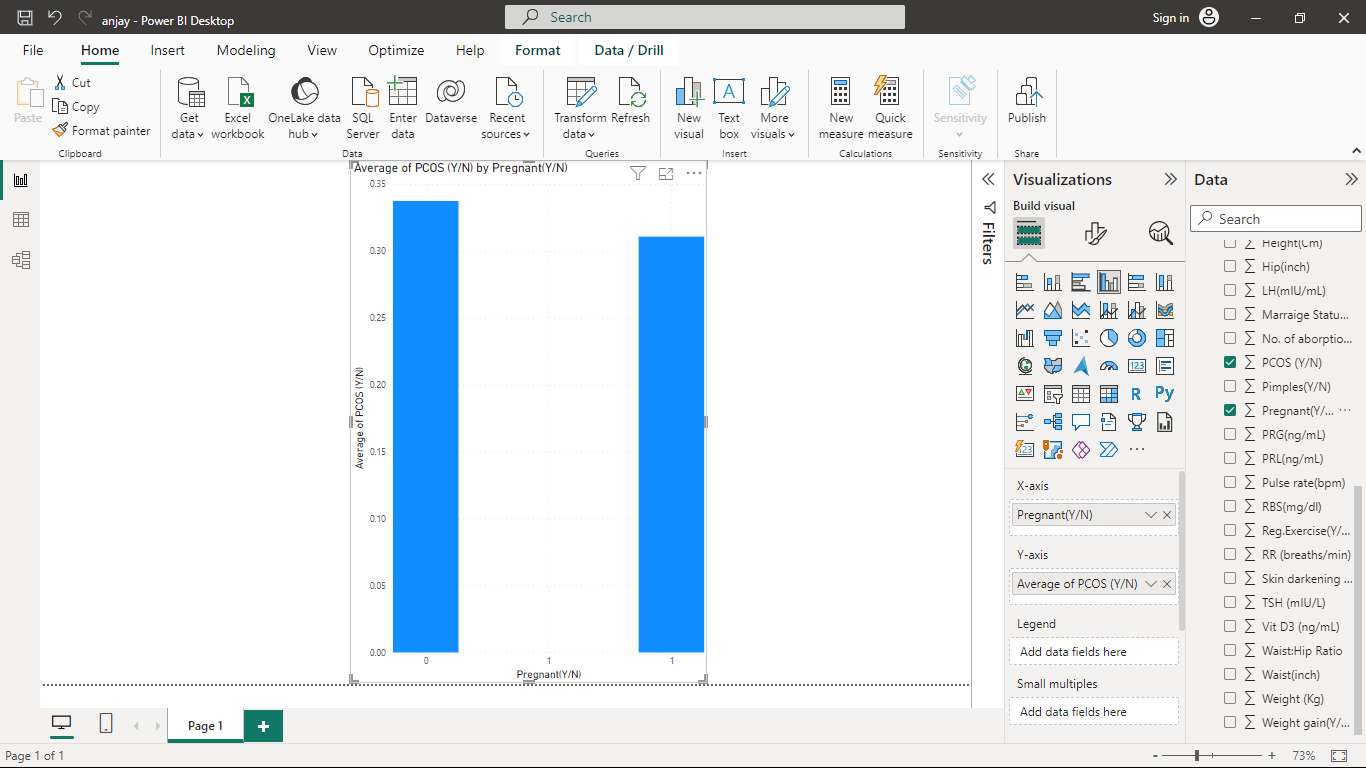
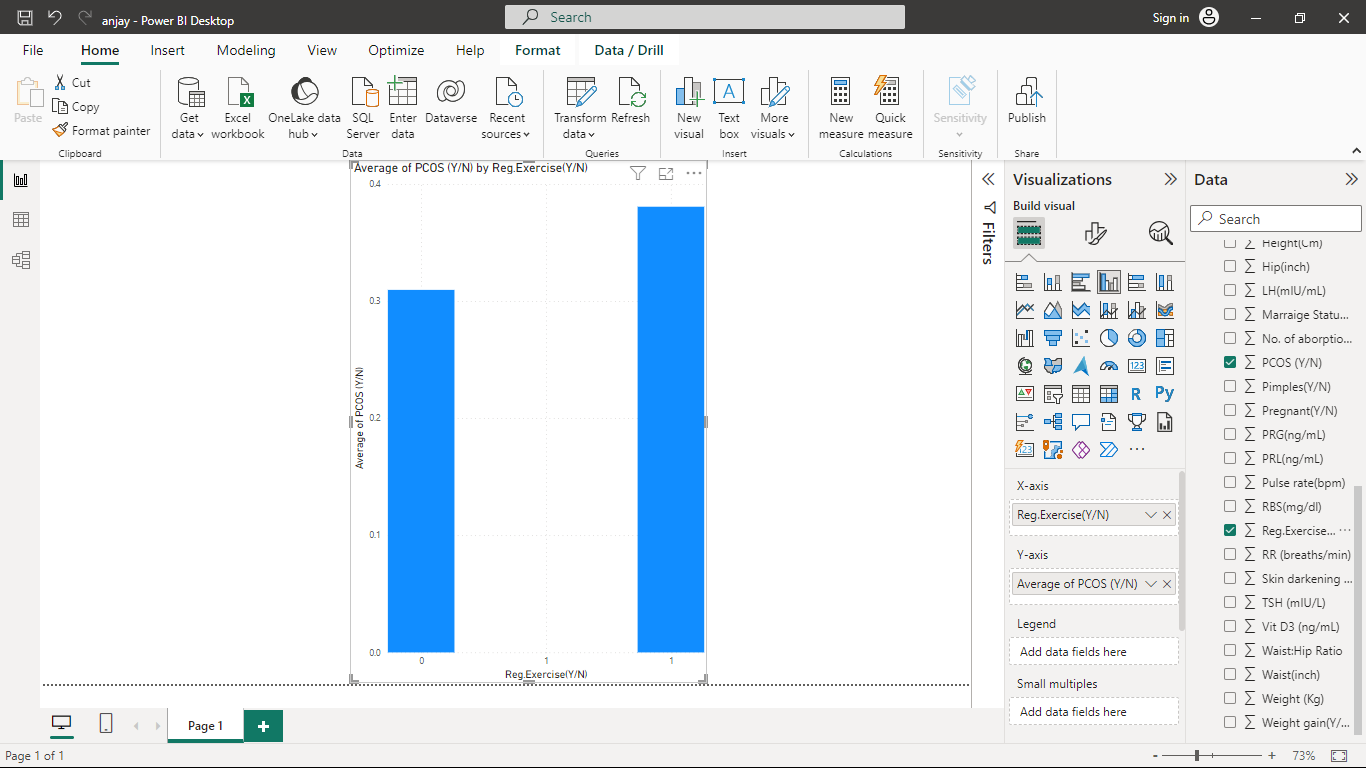
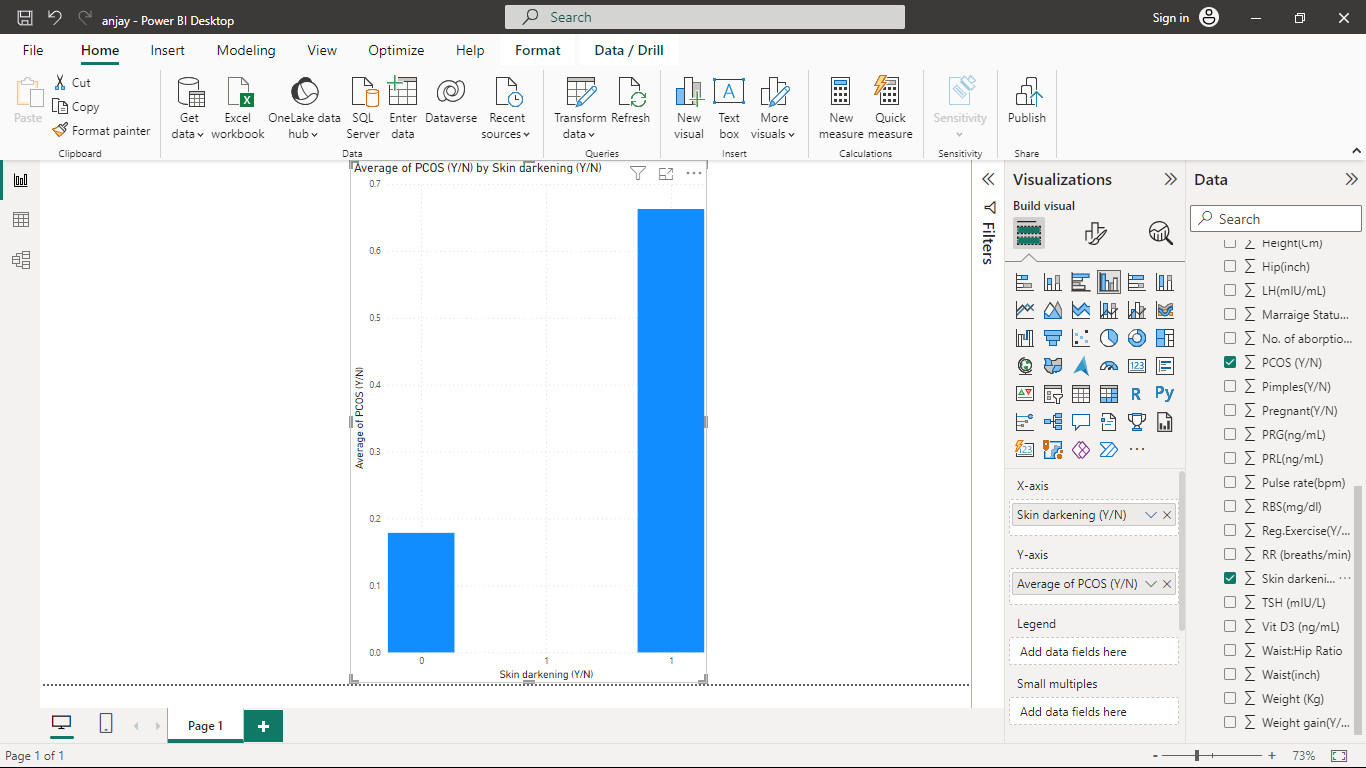
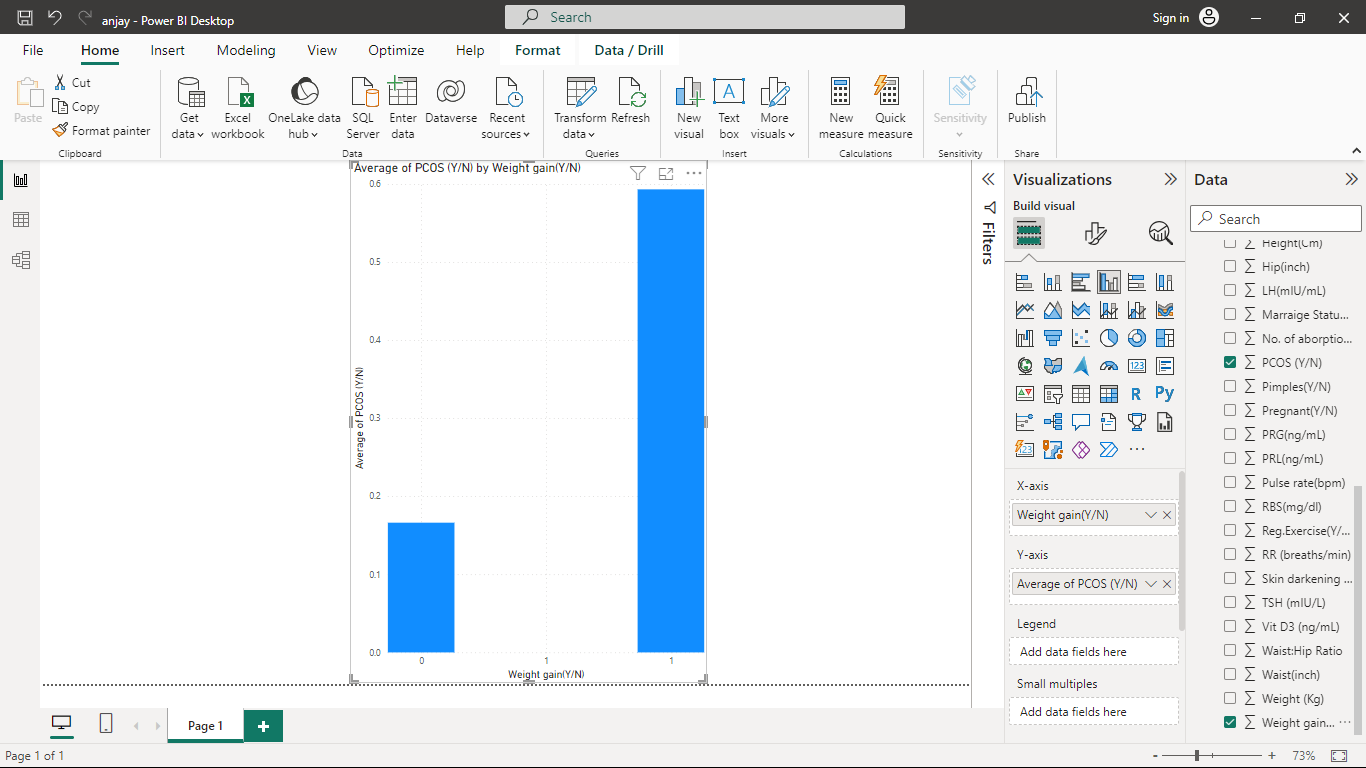
Untuk mendapatkan pemahaman visual tentang bagaimana data tersebar, kita dapat menggunakan teknik seperti plotting grafik box, histogram, atau scatter plot untuk mengamati pola dan karakteristik distribusi data. Dalam hal ini penulis menggunakan software Power Bi serta mengidentifikasi fitur numerik dan kategorikal dalam dataframe, berikut ini untuk fitur numerik :

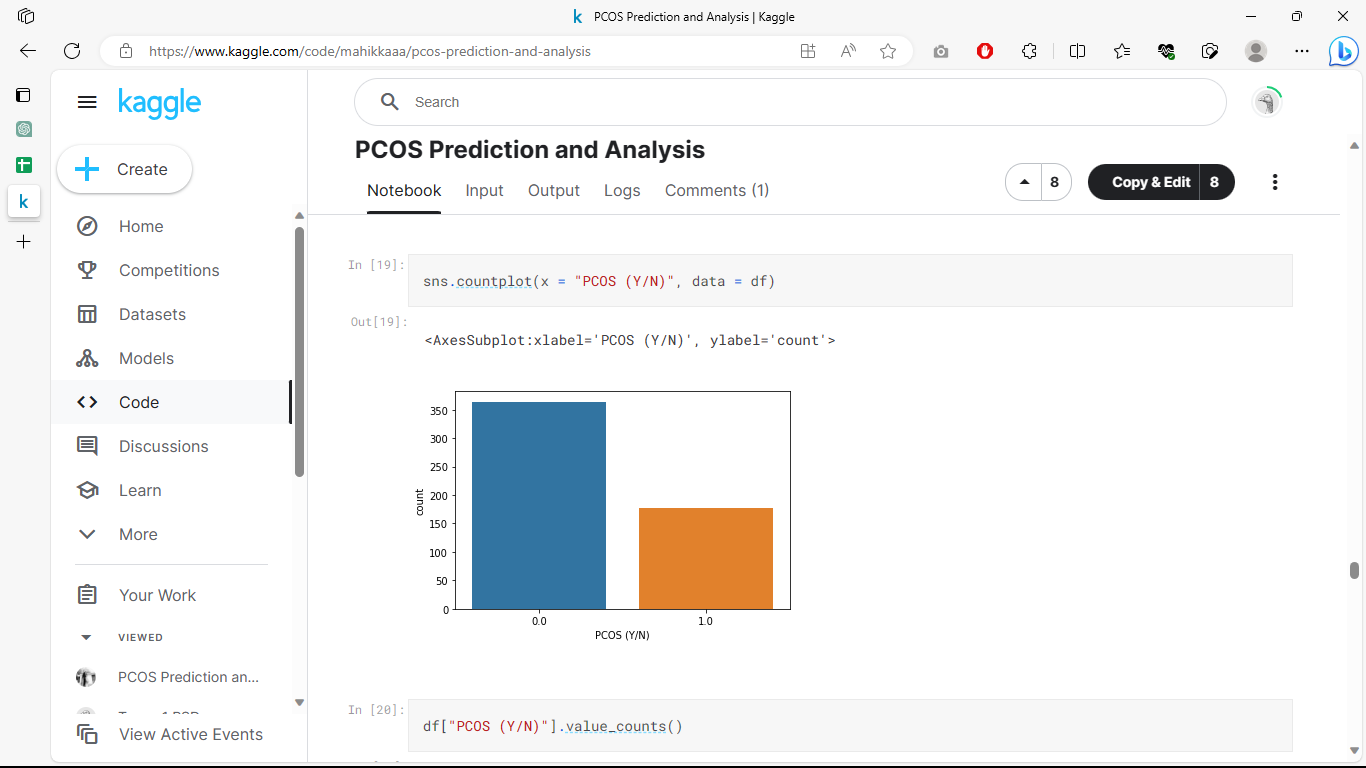
Dari pengamatan terhadap sebaran data di atas, terdapat kecenderungan bahwa semakin tinggi jumlah folikel pada ovarium, semakin tinggi kemungkinan seseorang mengalami PCOS. Kedua, pasien yang memiliki panjang siklus menstruasi yang lebih pendek cenderung memiliki kecenderungan yang lebih tinggi untuk mengalami PCOS. Ketiga, rata-rata usia pasien yang didiagnosis menderita PCOS adalah sekitar 30 tahun.

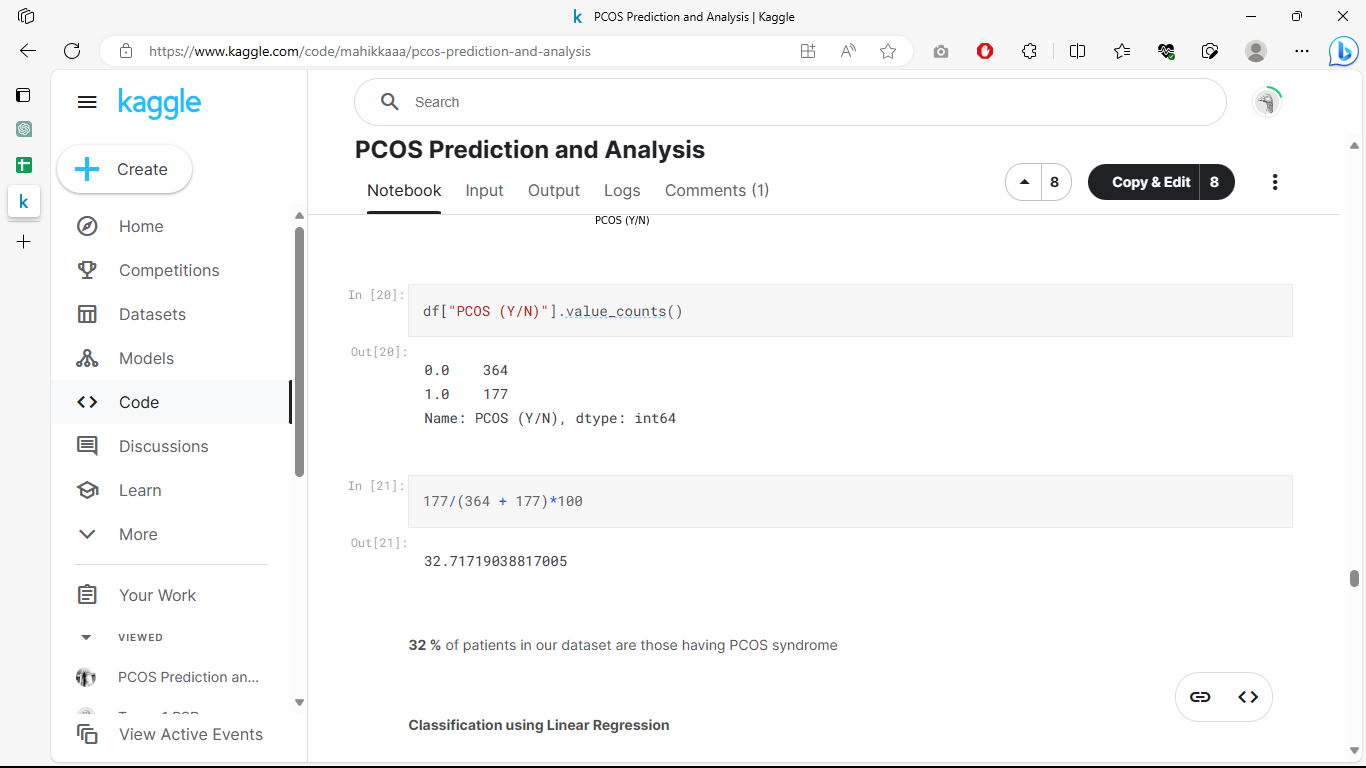
Dan di bawah ini merupakan fitur kategorial :

Dari hasil pengamatan tersebut, terlihat bahwa adanya gejala seperti peningkatan berat badan, pertumbuhan rambut yang berlebihan, jerawat, kehilangan rambut, konsumsi makanan cepat saji yang tinggi, dan perubahan warna kulit dapat berhubungan dengan peluang lebih tinggi untuk menderita sindrom ovarium polikistik (PCOS). Dengan demikian, ini juga dapat dianggap sebagai gejala-gejala yang terkait dengan PCOS.





Dari pesebaran data di atas dapat disimpulkan bahwa sebanyak 32% dari pasien dalam dataset ini mengalami sindrom PCOS.

Link Kaggle : https://www.kaggle.com/yunaperdana/pcos-analys