**DOKUMEN PROYEK**

**12S4054 - PENAMBANGAN DATA**

**Segmentasi Wilayah Peserta BPJS Tuberkulosis Berdasarkan Karakteristik Layanan dan Demografi Menggunakan K-Means Clustering**

****

**Disusun Oleh:**

| **12S22007** | **Tamara Y Sianipar** |
| --- | --- |
| **12S22018** | **Jesica A Siburian** |
| **12S22021** | **Krisnia Calysta Siahaan** |
| **12S22042** | **Ruth Septiana Simanullang** |

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INNFORMATIKA DAN TEKNIK ELEKTRO (FITE)**

**INSTITUT TEKNOLOGI DEL**

**(2024/2025)**

**BAB I**

**Business Understanding**

**1.1 Determine Business Objective**

BPJS Kesehatan merupakan lembaga yang bertanggung jawab dalam menjamin akses layanan kesehatan yang merata dan berkualitas bagi seluruh penduduk Indonesia. Seiring dengan meningkatnya jumlah peserta setiap tahun, tantangan dalam pengelolaan layanan kesehatan pun semakin kompleks, terutama untuk penyakit yang membutuhkan penanganan jangka panjang seperti Tuberkulosis (TB).

Tuberkulosis merupakan penyakit menular yang masih menjadi masalah kesehatan utama di Indonesia. Penanganan TB membutuhkan strategi layanan yang terstruktur dan alokasi sumber daya yang efektif, termasuk penyediaan fasilitas kesehatan, tenaga medis, serta edukasi masyarakat.

Untuk menjawab tantangan ini, BPJS Kesehatan memerlukan pendekatan berbasis data dalam memahami distribusi peserta yang menderita TB di berbagai wilayah administratif. Dengan informasi ini, BPJS Kesehatan dapat menyusun strategi intervensi yang lebih efektif, tepat sasaran, dan efisien dalam penggunaan sumber daya.

Tujuan bisnis utama dari proyek ini adalah:

* Meningkatkan efisiensi dan ketepatan sasaran dalam pengambilan keputusan strategis terkait penanganan Tuberkulosis.
* Mengidentifikasi karakteristik wilayah berdasarkan perilaku pelayanan (FKTP/FKRTL) dan demografi peserta Tuberkulosis
* Mendukung perencanaan distribusi pelayanan, edukasi dan intervensi berbasis klaster yang lebih relevan terhadap kondisi masing0masing wilayah.

**1.2 Determine Project Goal**

Agar tujuan bisnis dapat tercapai, proyek ini menetapkan beberapa sasaran teknis dan analitis sebagai berikut:

* Melakukan segmentasi wilayah berdasarkan karakteristik pelayanan kesehatan dan profil demografi peserta Tuberkulosis menggunakan K-Means Clustering.
* Menyusun label klaster yang menggambarkan tipologi wilayah dalam menangani TB.

**1.3 Produce Project Plan**

Untuk memastikan pelaksanaan proyek berjalan secara terstruktur, berikut adalah tahapan utama proyek beserta kebutuhan masing-masing:

1. Business Understanding

Pada tahap ini , aktivitas utama yang dilakukan adalah menentukan tujuan bisnis dan analisis kebutuhan kebijakan TB berbasis data. Kebutuhan yang digunakan di tahap ini adalah melakukan studi regulasi BPJS dan TB.

1. Data Understanding

Pada tahap ini, aktivitas utama yang dilakukan adalah melakukan eksplorasi metadata dan variabel Tuberkulosis dari file BPJS. Selain itu, melakukan identifikasi karakteristik peserta & pelayanan Tuberkulosis. Adapun kebutuhan yang digunakan dalam tahap ini adalah Data Sampel Final 2022 : Kontekstual TB.

1. Data Preparation

Pada tahap ini , aktivitas utama yang dilakukan adalah melakukan merge data, clenaing data (untuk data null dan data duplikat), feature engineering, dan normalisasi variabel numerik.

1. Modeling

Pada tahap ini , aktivitas utama yang dilakukan adalah menentukan jumlah klaster optimal, melakukan cluster menggunakan K-Means, Interpretasi hasil klaster. Kebutuhan yang diperlukan adalah data yang siap untuk dilakukan modeling.

1. Evaluation

Pada tahap ini, aktivitas utama yang dilakukan adalah melakukan evaluasi performa klaster (apakah dapat dibedakan dengan jelas?), hingga validasi domain (apakah klaster masuk akal bagi pengambil kebijakan atau keputusan?). Kebutuhan yang diperlukan untuk tahap ini adalah domain knowledge, metrik evaluasi internal.

1. Deployment

Pada tahap ini,aktivitas yang dilakukan adalah menyusun laporan akhir, visualisasi hasil segmentasi,hingga rekomendasi strategi intervensi tiap karakter.

**BAB II**

**DATA UNDERSTANDING**

Bertujuan untuk memahami struktur dan karakteristik data BPJS 2015-2021 yang akan digunakan dalam proses klasifikasi, prediksi, dan/atau klastering. Terdapat lima subset data untuk data sampe BPJS Kesehatan 2015-2021 , yaitu :

1. **Kepesertaan**

Kepesertaan data sampel BPJS Kesehatan tahun 2015-2021 sebesar 2.305.435 peserta. Variabel pada data kepesertaan terdiri dari karakteristik peserta sepert, tanggal lahir, jenis kelamin, status perkawinan, hubungan keluarga dan tempat tinggal peserta (provinsi dan kabupaten/kota). Variabel ‘Nomor Peserta’ (PSTV01) adalah nomor unik dari seluruh peserta JKN-KIS yang sudah diidentifikasi untuk menjaga kerahasiaan identitas peserta.

1. **Pelayanan Fasilitas Kesehatan Tingkat Pertama (FKTP) Kapitasi**

Total pelayanan FKTP sebesar 2.498.805 kunjungan.

1. **Pelayanan Fasilitas Kesehatan Rujukan Tingkat Lanjut (FKRTL)**

Total pelayanan FKRTL sebesar 872.201 kunjungan

1. **Pelayanan Fasilitas Kesehatan Tingkat Pertama (FKTP) Non Kapitasi**

Total pelayanan FKTP Non kapitasi tahun 2021 sebesar 95.617 tindakan

1. **Pelayanan Fasilitas Kesehatan Rujukan Tingkat Lanjut (FKRTL) Diagnosis Sekunder**

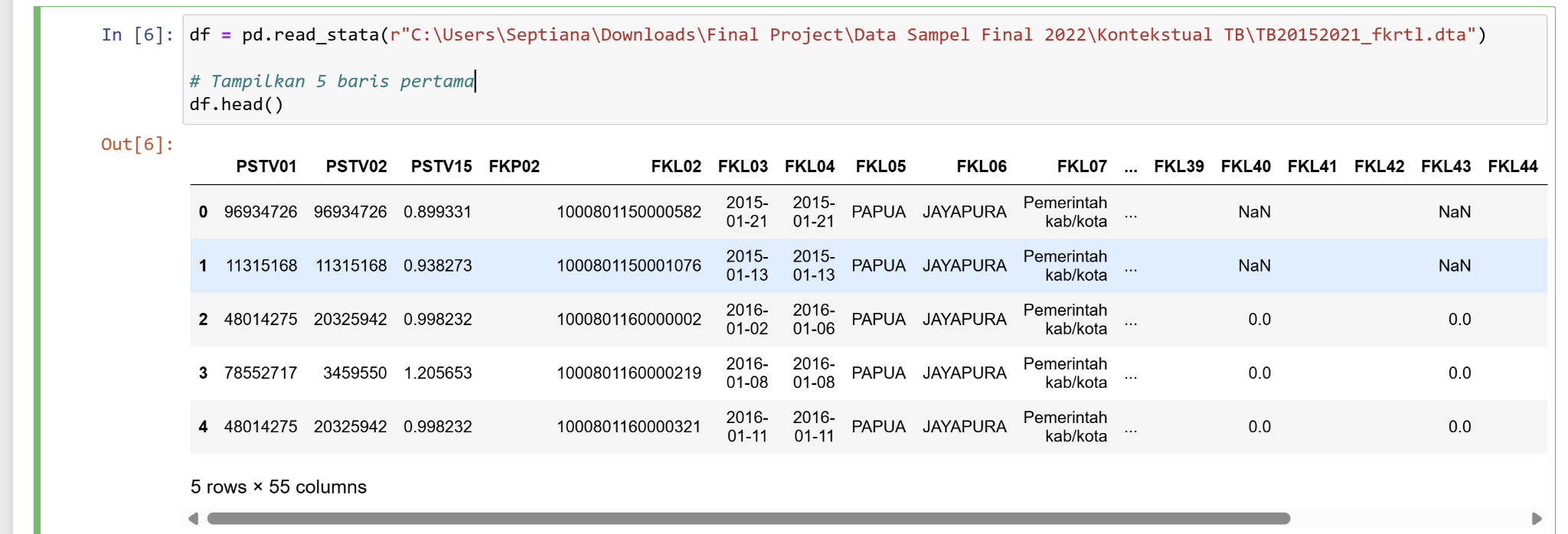
Data diagnosis sekunder sebagai bagian dari data pelayanan FKRTL dengan total observasi 925.803 diagnosis.

**2.1 Pengumpulan Data**

**2.1.1 Data Kontekstual TB**

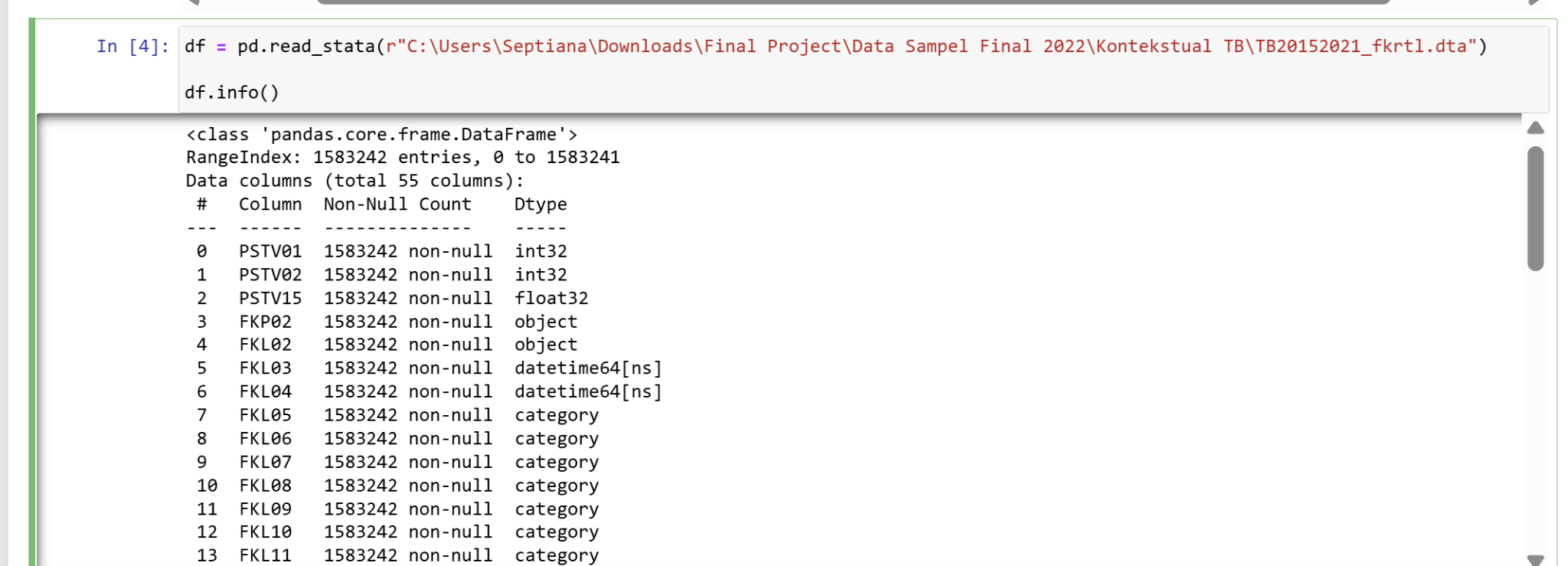
Data sampel kontekstual Tuberkulosis (TB) adalah data kepesertaan dan pelayanan dari peserta JKN-KIS melalui data warehouse BPJS Kesehatan. Peserta yang terpilih dalam data sampel kontekstual TB adalah mereka yang diidentifikasi pernah mendapatkan pelayanan FKTP maupun FKRTL dengan diagnosis TB melalui kode ICD-10 yaitu A15, A16, A17, A18 dan A19 pada tahun 2019.

1. **TB20152021\_fkrtl.dta**



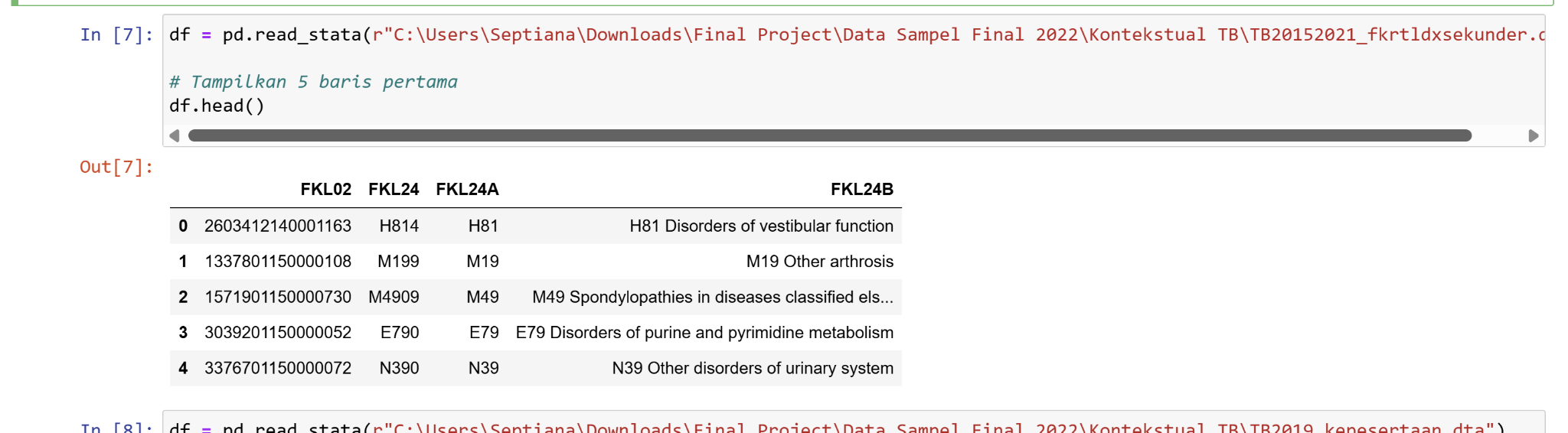
Pelayanan FKRTL adalah data pelayanan dari data sampel peserta didiagnosis TB yang melakukan kunjungan ke pelayanan FKRTL pada tahun 2015-2021 sebesar 1.583.242 kunjungan.

Detail data

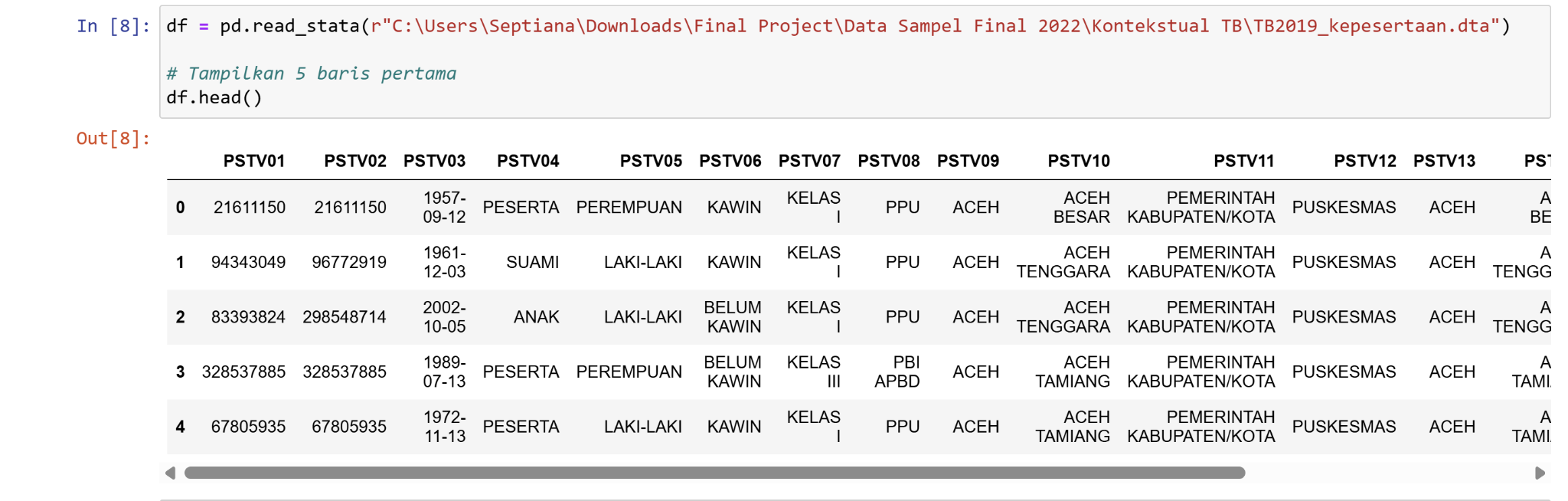


Terdapat beberapa kolom/atribut yang tersedia di dalam data FKRTL ini, diantaranya :

1. PSTV01 : Nomor Peserta
2. PSTV02 : Nomor Keluarga
3. PSTV15 : Bobot
4. FKP02 : No Asal Rujukan
5. FKL02 : ID Kunjungan
6. FKL03 : Tanggal datang kunjungan FKRTL
7. FKL04 : Tanggal pulang kunjungan FKRTL
8. FKL05 : Provinsi FKRTL
9. FKL06 : Kabupaten/Kota FKRTL
10. FKL07 : Kepemilikan FKRTL
11. FKL08 : Jenis FKRTL
12. FKL09 : Tipe FKRTL
13. FKL10 : Tingkat Pelayanan FKRTL
14. FKL11 : Jenis Poli FKRTL
15. FKL12 : Segmen Peserta saat akses layanan FKRTL
16. FKL13 : Kelas iuran premi peserta saat akses layanan FKRTL
17. FKL14 :Status pulang dari FKRTL
18. FKL15 : Kode dan nama diagnosis masuk ICD 10
19. FKL15A : Kode diagnosis masuk ICD 10
20. FKL16: Kode ICD 10 diagnosis masuk FKRTL
21. FKL16A : Nama diagnosis masuk FKRTL
22. FKL17 : Kode dan nama diagnosis primer ICD 10
23. FKL17A : Kode diagnosis primer ICD 10
24. FKL18 : Kode ICD 10 diagnosis primer FKRTL
25. FKL18A : Nama diagnosis primer FKRLT
26. FKL19 : Kode INACBGs
27. FKL19A : Deskripsi kode INACBGs
28. FKL20 : INACBGs - Kode Casemix main groups (digit ke-1)
29. FKL21 : INACBGs - Tipe kelompok kasus atau case groups (digit ke-2)
30. FKL22 :INACBGs - Spesifikasi kelompok kasus (digit ke-3)
31. FKL23 :INACBGs -Tingkat keparahan kelompok kasus (digit ke-4)
32. FKL25 : Provinsi faskes perujuk
33. FKL26 : Kabupaten/Kota faskes perujuk
34. FKL27 : Kepemilikan faskes perujuk
35. FKL28 : Jenis faskes perujuk
36. FKL29 : Tipe faskes perujuk
37. FKL30 : Jenis prosedur
38. FKL31 : Tarif regional INACBGs
39. FKL32 : Group Tarif INACBGs
40. FKL33 : Kode special sub-acute groups (SA)
41. FKL34 : Tarif special sub-acute groups (SA)
42. FKL35 : Kode special procedures (SP)
43. FKL36 : Deskripsi special procedures (SP)
44. FKL37 : Tarif special procedures (SP)
45. FKL38 : Kode special prosthesis (RR)
46. FKL39 : Deskripsi special prosthesis (RR)
47. FKL40 : Tarif special prosthesis (RR)
48. FKL41 : Kode special investigation (SI)
49. FKL42 : Deskripsi special investigation (SI)
50. FKL43 : Tarif special investigation (SI)
51. FKL44 : Kode special drugs (SD)
52. FKL45 : Deskripsi special drugs (SD)
53. FKL46 : Tarif special drugs (SD)
54. FKL47 : Biaya Tagih - oleh fasilitas kesehatan (provider)
55. FKL48 : Biaya verifikasi - BPJS Kesehatan setelah dilakukan verifikasi.
56. **TB20152021\_fkrtldxsekunder.dta**

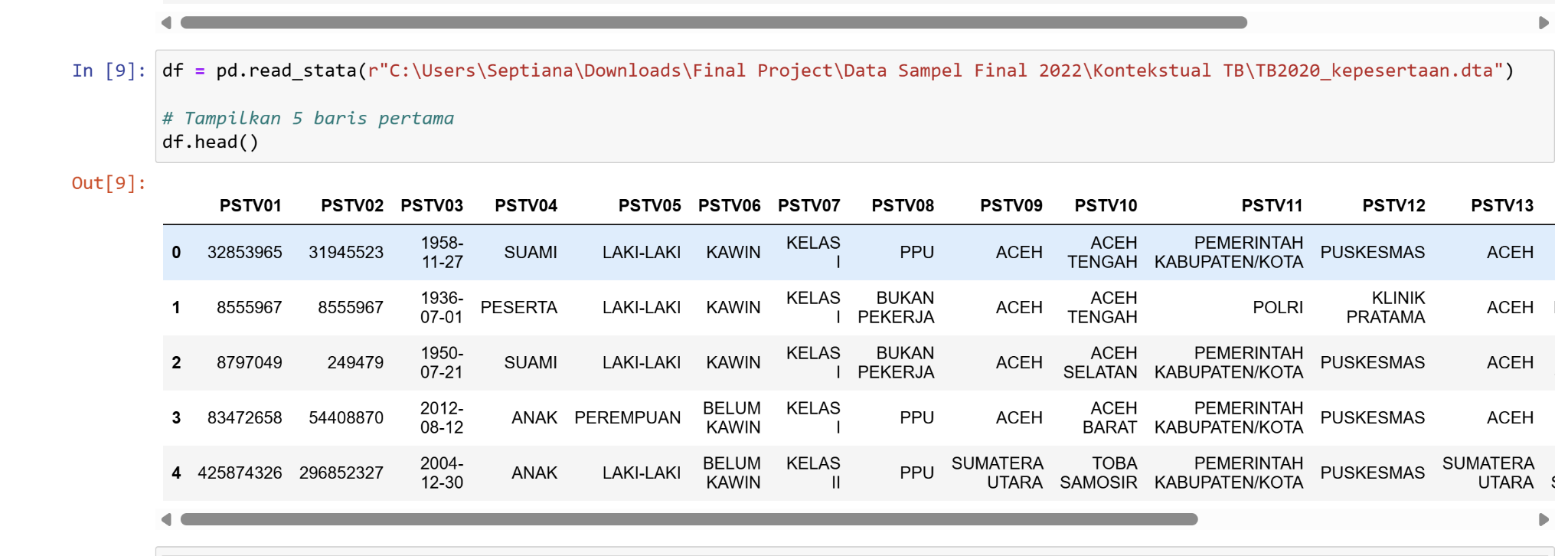


1. **TB2019\_kepesertaan.dta**



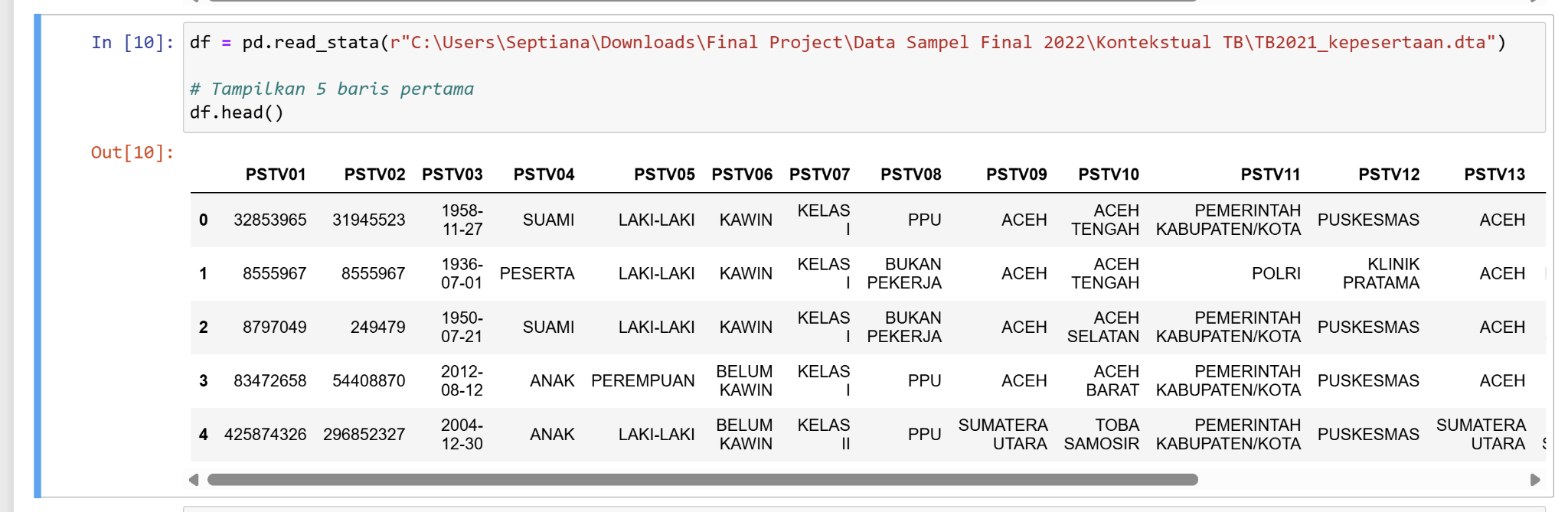
Data diatas merupakan data peserta BPJS Kesehatan tahun 2019 dengan diagnosa penyakit Tuberkulosis. Terdapat beberapa kolom/atribut yang tersedia di dalam data kepesertaan ini, diantaranya :

1. PSTV01 : Nomor Peserta
2. PSTV02 : Nomor Keluarga
3. PSTV03 : Tanggal Lahir Peserta
4. PSTV04 : Hubungan Keluarga
5. PSTV05 : Jenis Kelamin
6. PSTV06 : Status Perkawinan
7. PSTV07 : Kelas Rawat
8. PSTV08 : Segmentasi Peserta
9. PSTV09 : Provinsi Tempat Tinggal Peserta
10. PSTV10 : Kabupaten/Kota Tempat Tinggal Peserta
11. PSTV11 : Kepemilikan Faskes
12. PSTV12 : Jenis Faskes
13. PSTV13 : Provinsi Fasilitas Kesehatan Peserta Terdaftar
14. PSTV14 : Kabupaten/Kota Fasilitas Kesehatan Peserta Terdaftar
15. PSTV15 : Bobot
16. PSTV16 : Tahun Sampel
17. PSTV17 : Status Kepesertaan
18. PSTV18 : Tahun Meninggal
19. **TB2020\_kepesertaan.dta : Kepesertaan pasien TB tahun 2020**

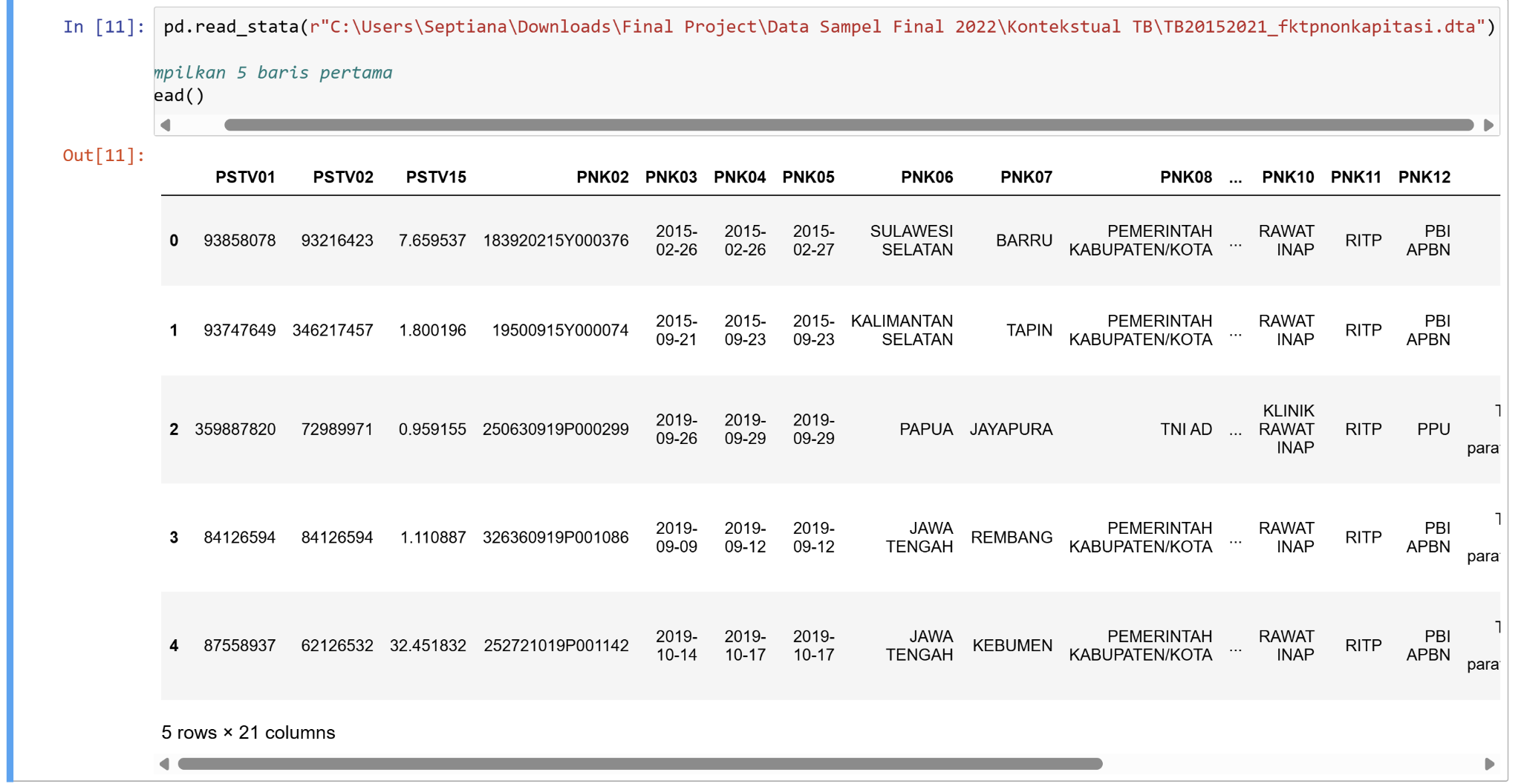


Data diatas merupakan data peserta BPJS Kesehatan tahun 2020 dengan diagnosa penyakit Tuberkulosis. Terdapat beberapa kolom/atribut yang tersedia di dalam data kepesertaan ini, diantaranya :

1. PSTV01 : Nomor Peserta
2. PSTV02 : Nomor Keluarga
3. PSTV03 : Tanggal Lahir Peserta
4. PSTV04 : Hubungan Keluarga
5. PSTV05 : Jenis Kelamin
6. PSTV06 : Status Perkawinan
7. PSTV07 : Kelas Rawat
8. PSTV08 : Segmentasi Peserta
9. PSTV09 : Provinsi Tempat Tinggal Peserta
10. PSTV10 : Kabupaten/Kota Tempat Tinggal Peserta
11. PSTV11 : Kepemilikan Faskes
12. PSTV12 : Jenis Faskes
13. PSTV13 : Provinsi Fasilitas Kesehatan Peserta Terdaftar
14. PSTV14 : Kabupaten/Kota Fasilitas Kesehatan Peserta Terdaftar
15. PSTV15 : Bobot
16. PSTV16 : Tahun Sampel
17. PSTV17 : Status Kepesertaan
18. PSTV18 : Tahun Meninggal
19. **TB2021\_kepesertaan.dta : Kepesertaan pasien TB tahun 2021**



Data diatas merupakan data peserta BPJS Kesehatan tahun 2021 dengan diagnosa penyakit Tuberkulosis. Terdapat beberapa kolom/atribut yang tersedia di dalam data kepesertaan ini, diantaranya :

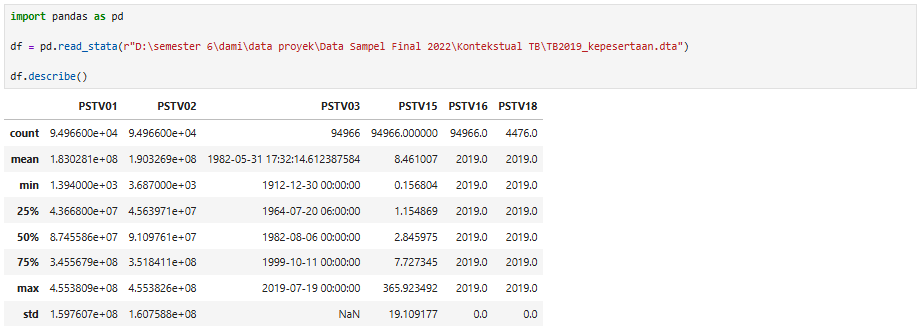
1. PSTV01 : Nomor Peserta
2. PSTV02 : Nomor Keluarga
3. PSTV03 : Tanggal Lahir Peserta
4. PSTV04 : Hubungan Keluarga
5. PSTV05 : Jenis Kelamin
6. PSTV06 : Status Perkawinan
7. PSTV07 : Kelas Rawat
8. PSTV08 : Segmentasi Peserta
9. PSTV09 : Provinsi Tempat Tinggal Peserta
10. PSTV10 : Kabupaten/Kota Tempat Tinggal Peserta
11. PSTV11 : Kepemilikan Faskes
12. PSTV12 : Jenis Faskes
13. PSTV13 : Provinsi Fasilitas Kesehatan Peserta Terdaftar
14. PSTV14 : Kabupaten/Kota Fasilitas Kesehatan Peserta Terdaftar
15. PSTV15 : Bobot
16. PSTV16 : Tahun Sampel
17. PSTV17 : Status Kepesertaan
18. PSTV18 : Tahun Meninggal
19. **TB20152021\_fktpnonkapitasi.dta** 

Terdapat beberapa kolom/atribut yang tersedia di dalam data kepesertaan ini, diantaranya:

1. PSTV01 : Nomor Peserta
2. PSTV02 : Nomor Keluarga
3. PSTV15 : Bobot
4. PNK02 : ID Kunjungan
5. PNK03 : Tanggal Kunjungan
6. PNK04 : Tanggal Tindakan
7. PNK05 : Tanggal Pulang
8. PNK06 : Provinsi Faskes
9. PNK07 : Kode Kab/Kota Faskes
10. PNK08 : Kepemilikan Faskes
11. PNK09 : Jenis Faskes
12. PNK10 : Tipe Faskes
13. PNK11 : Tingkat Layanan
14. PNK12 : Segmen Peserta
15. PNK13 : Kode dan Nama diagnosis berdasarkan ICD-10 (3 digit)
16. PNK13A : Kode diagnosis berdasarkan ICD-10 (3 digit)
17. PNK14 : Kode diagnosis (3-5 digit)
18. PNK15 : Nama diagnosis
19. PNK16 : Nama Tindakan
20. PNK17 : Biaya Tagih
21. PNK18 : Biaya Verifikasi

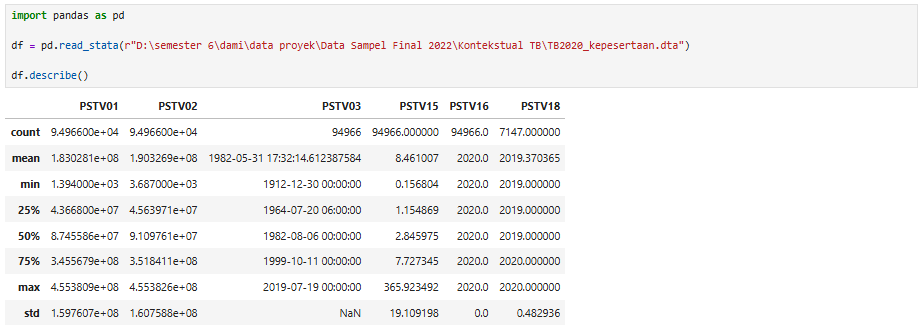
**2.2 Describe Data**

1. **TB2019\_kepesertaan.dta**

****

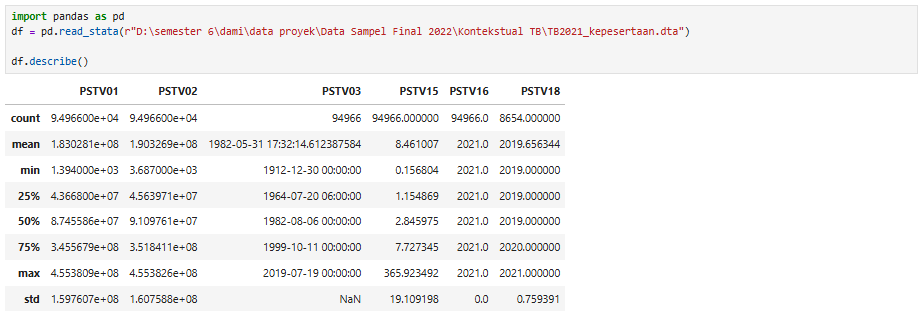
Berdasarkan output df.describe() dari data kepesertaan BPJS 2019, dapat dilihat bahwa data cukup lengkap dan rapi. Kolom PSTV01 dan PSTV02 berisi ID peserta dan keluarga tanpa nilai kosong. Tanggal lahir (PSTV03) menunjukkan rentang usia sangat luas, dari 1912 hingga 2019, dengan rata-rata sekitar tahun 1982. Bobot (PSTV15) memiliki rata-rata 8,46 dengan variasi cukup besar, yang mungkin digunakan untuk analisis representatif. Seluruh data berasal dari tahun 2019 (PSTV16), sedangkan PSTV18 hanya terisi sebagian. Secara umum, data ini siap untuk dianalisis lebih lanjut seperti segmentasi atau penghitungan usia.

1. **TB2020\_kepesertaan.dta**

****

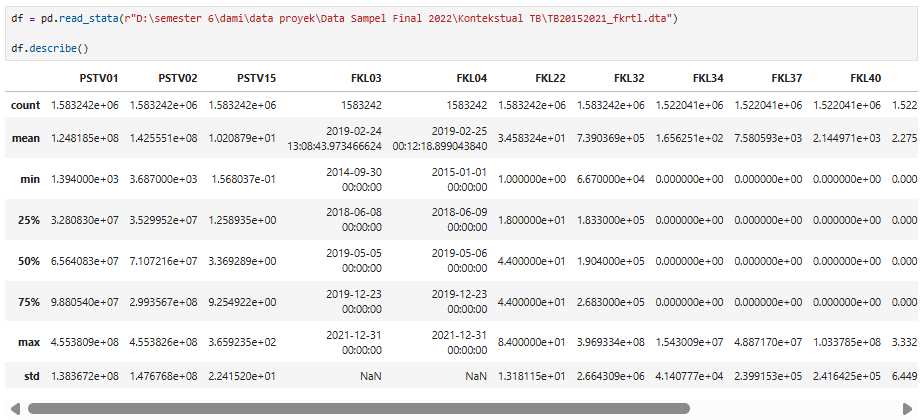
Berdasarkan hasil df.describe(), kolom PSTV01 dan PSTV02 berisi data numerik yang sangat kecil dengan nilai rata-rata mendekati nol dan standar deviasi rendah, menunjukkan penyebaran data yang sempit. Kolom PSTV03 kemungkinan berisi data tanggal karena rentangnya dari tahun 1912 hingga 2010. Kolom PSTV05 menunjukkan nilai dengan sebaran yang cukup besar, dengan rata-rata sekitar 54,86 dan maksimum hingga 365,53. Sementara itu, kolom PSTV16 tidak memiliki variasi karena seluruh nilainya adalah 2020. Kolom PSTV18 berisi data tahunan dengan dominasi nilai tahun 2020, namun hanya memiliki 7147 entri, jauh lebih sedikit dibanding kolom lain, sehingga perlu perhatian khusus dalam analisis.

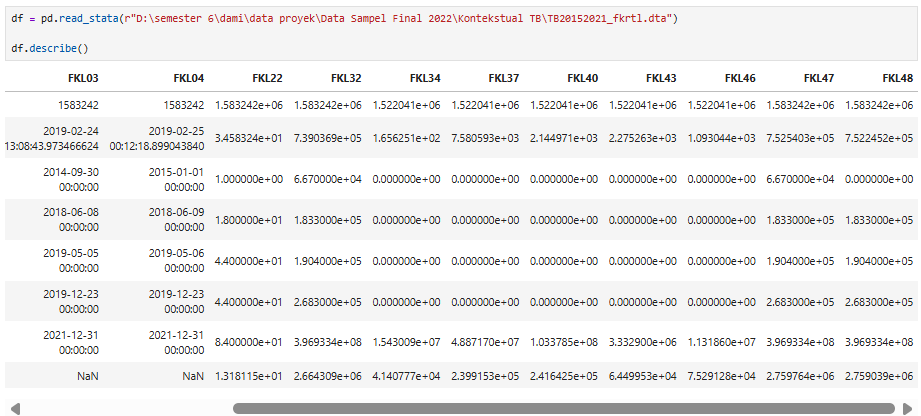
1. **TB2021\_kepesertaan.dta**

****

Berdasarkan hasil df.describe() dari data “TB2021\_kepesertaan.dta”, terlihat bahwa data ini terdiri dari sekitar 94.966 baris dan sebagian besar kolomnya terisi lengkap. Kolom PSTV01 dan PSTV02 menyimpan ID peserta dan ID keluarga tanpa nilai kosong. Tanggal lahir peserta di kolom PSTV03 menunjukkan rentang usia yang sangat luas, dari tahun 1912 hingga 2019, dengan rata-rata tahun lahir sekitar 1982. Bobot data pada PSTV15 memiliki nilai rata-rata sekitar 8,46 dan standar deviasi cukup tinggi, yang menunjukkan adanya variasi besar antar peserta. Seluruh data berasal dari tahun 2021 (PSTV16), sedangkan PSTV18 berisi data tambahan yang hanya terisi sebagian. Secara keseluruhan, data ini cukup bersih dan bisa langsung digunakan untuk analisis seperti segmentasi peserta atau studi distribusi usia dan bobot peserta BPJS.

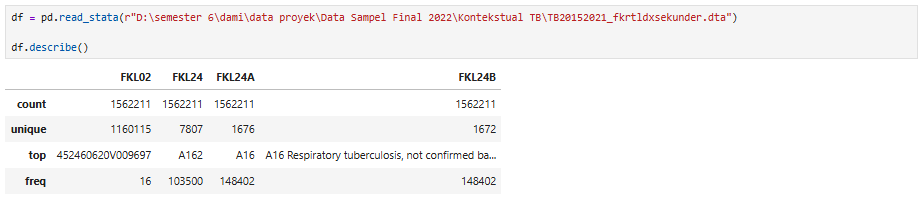
1. **TB20152021\_fkrtl.dta**

****

****

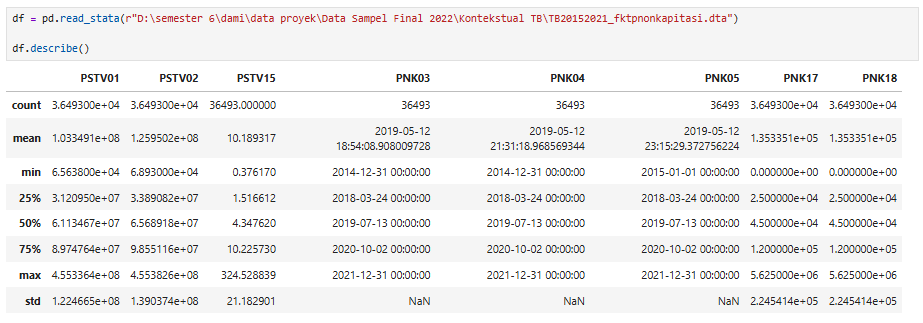
Berdasarkan hasil df.describe() dari data “TB20152021\_fkrtl.dta”, terlihat bahwa data ini sangat besar dan lengkap, dengan lebih dari 1,5 juta baris. Kolom PSTV01 dan PSTV02 berisi ID peserta dan keluarga tanpa nilai kosong. Kolom PSTV15 menunjukkan bobot data dengan rata-rata sekitar 1, dan penyebaran data cukup lebar, menandakan ada variasi dalam pembobotan sampel. Kolom tanggal seperti FKL03 dan FKL04 menunjukkan bahwa data mencakup periode yang panjang dari tahun 2015 hingga akhir 2021. Sementara itu, kolom-kolom seperti FKL22, FKL32, dan FKL34 berisi data numerik dengan nilai maksimum yang sangat besar—ini bisa berkaitan dengan jumlah kunjungan, biaya, atau unit layanan kesehatan. Beberapa kolom seperti FKL37, FKL40, hingga FKL48 memiliki nilai-nilai ekstrem yang menandakan adanya data yang sangat besar atau kasus khusus.

1. **TB20152021\_fkrtldxsekunder**

****

Berdasarkan hasil df.describe() dari data “TB20152021\_fkrtl\_dxsekunder.dta”, terlihat bahwa data ini berisi lebih dari 1,5 juta catatan yang semuanya lengkap, tanpa data kosong di kolom FKL02, FKL24, FKL24A, dan FKL24B. Kolom-kolom ini berisi informasi terkait diagnosis pasien, seperti kode penyakit dan deskripsinya. Terdapat ribuan jenis diagnosis berbeda dalam data ini, yang menunjukkan variasi penyakit yang cukup besar. Salah satu diagnosis yang paling sering muncul adalah kode A16, yaitu tuberkulosis saluran napas yang belum dikonfirmasi bakteri, dengan jumlah kemunculan lebih dari 148 ribu kali. Ini menunjukkan bahwa penyakit tersebut cukup umum dalam data dan bisa menjadi fokus penting untuk dianalisis lebih lanjut, misalnya untuk melihat tren penyakit atau kebutuhan penanganan medis.

1. **TB20152021\_fktpnonkapitasi.dta**

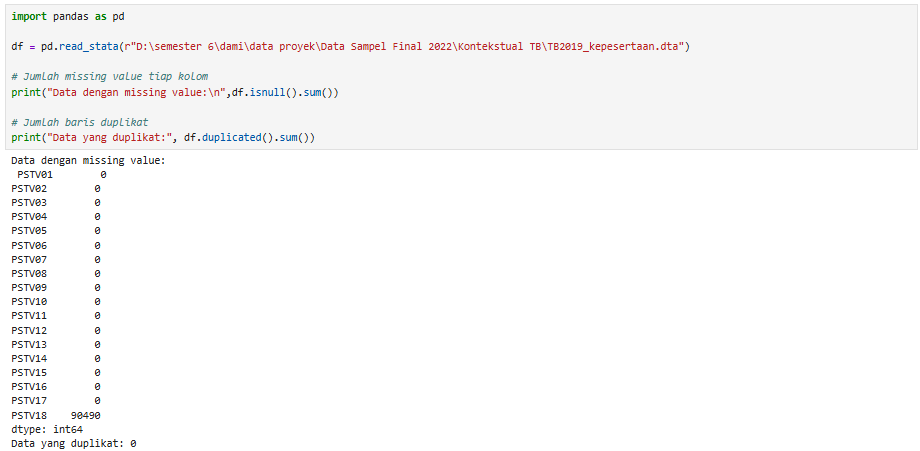
****

Berdasarkan hasil df.describe() dari data “TB20152201\_fktpnonkapitasi.dta”, total data terdiri sebanyak 36.493 baris. Kolom PSTV01 dan PSTV02 berisi ID peserta dan ID keluarga tanpa data yang hilang. Kolom PSTV15 menunjukkan bobot sampel dengan rata-rata sekitar 10, tetapi ada penyebaran data yang cukup besar, yang artinya ada perbedaan besar antar nilai bobot. Tanggal-tanggal pada kolom PNK03 sampai PNK05 menunjukkan bahwa sebagian besar data berasal dari sekitar tahun 2019, dengan rentang waktu dari 2014 sampai 2021. Sementara itu, kolom PNK17 dan PNK18 kemungkinan berkaitan dengan hal-hal seperti biaya atau data administratif lainnya, karena nilainya sangat besar dan bervariasi. Secara keseluruhan, data ini sudah siap untuk dianalisis lebih lanjut, misalnya untuk menghitung biaya atau mengevaluasi layanan kesehatan.

**2.3 Validation Data**

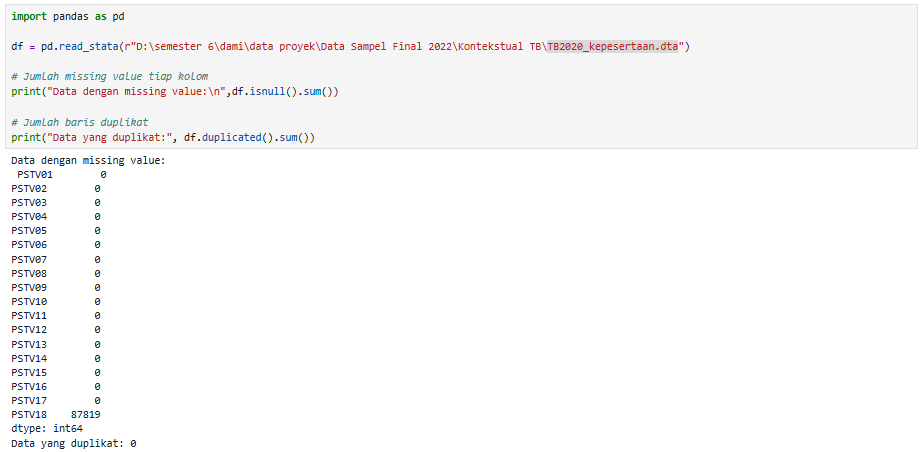
**2.3.1 Pemerikasaan Data**

1. **TB2019\_kepesertaan.dta**

****

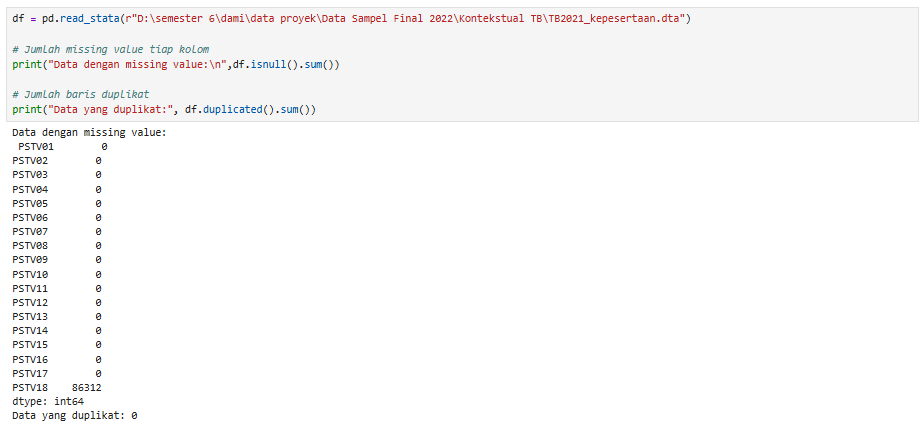
Dari hasil cek data untuk data kepesertaan tuberkulosis tahun 2019, terlihat bahwa di variabel/atribut PSTV18 memiliki 90490 missing values. Untuk data tersebut tidak memiliki data duplikat.

1. **TB2020\_kepesertaan.dta**

****

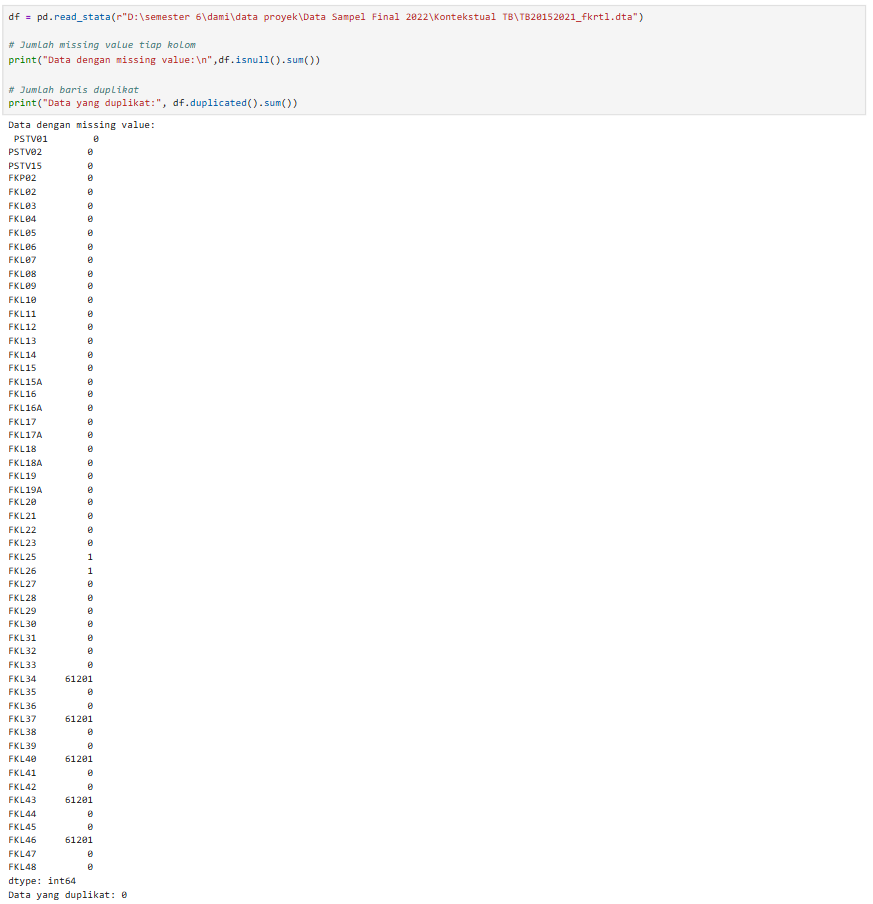
Dari hasil cek data untuk data kepesertaan tuberkulosis tahun 2020, terlihat bahwa di variabel/atribut PSTV18 memiliki 87819 missing values. Untuk data tersebut tidak memiliki data duplikat.

1. **TB2021\_kepesertaan.dta**

****

Dari hasil cek data untuk data kepesertaan tuberkulosis tahun 2021, terlihat bahwa di variabel/atribut PSTV18 memiliki 86312 missing values. Untuk data tersebut tidak memiliki data duplikat.

1. **TB20152021\_fkrtl.dta**

****

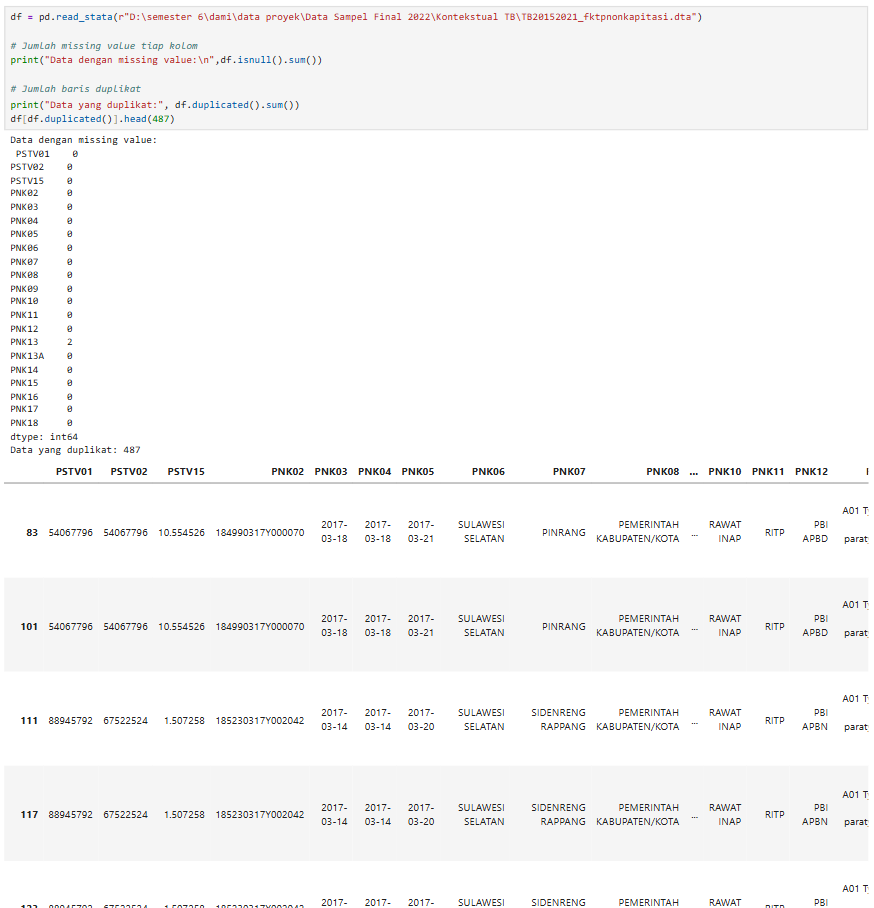
Dari hasil cek data untuk data pelayanan fkrtl , terlihat bahwa di variabel/atribut FKL34, FKL37, FKL40, FKL43, FKL46 masing-masing memiliki 61201 missing values. Untuk data tersebut tidak memiliki data duplikat.

1. **TB20152021\_fkrtldxsekunder**

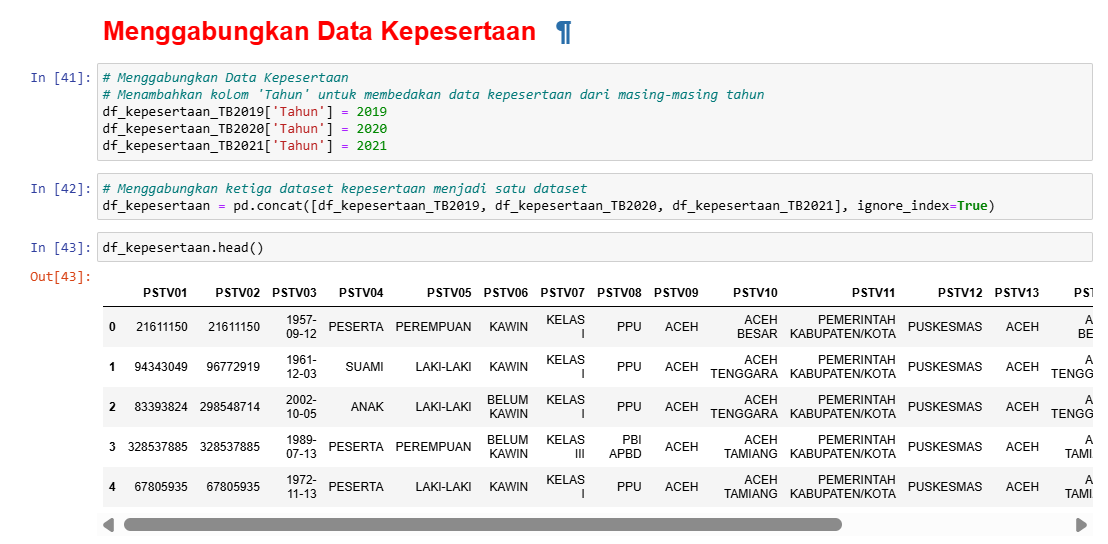
****

Dari hasil cek data untuk data pelayanan fkrtl sekunder , terlihat bahwa variabel/atribut tidka memiliki missing values. Untuk data tersebut tidak memiliki data duplikat.

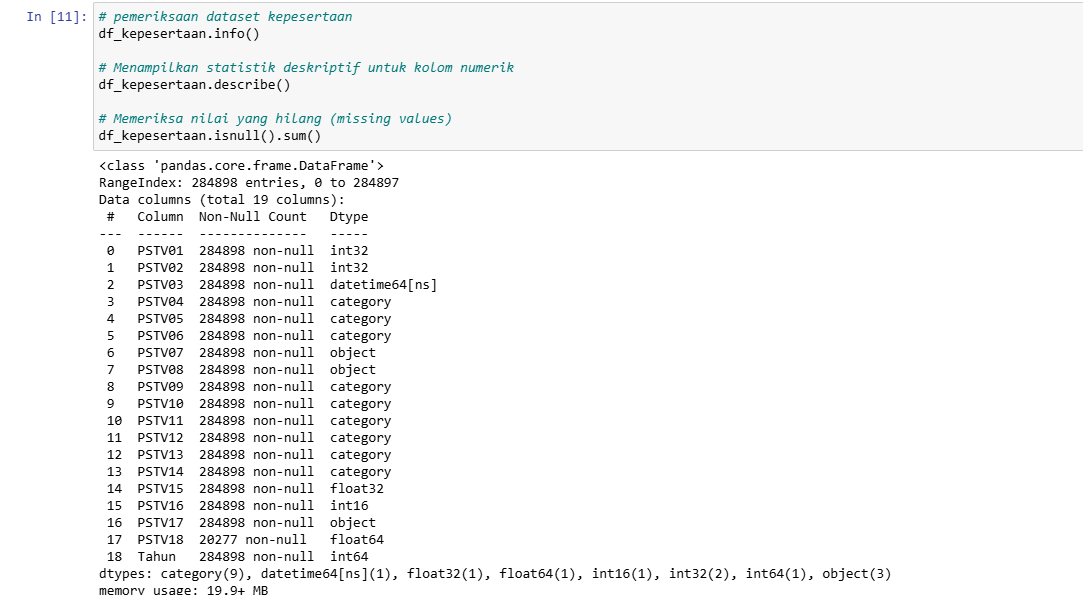
1. **TB20152021\_fktpnonkapitasi.dta**

****

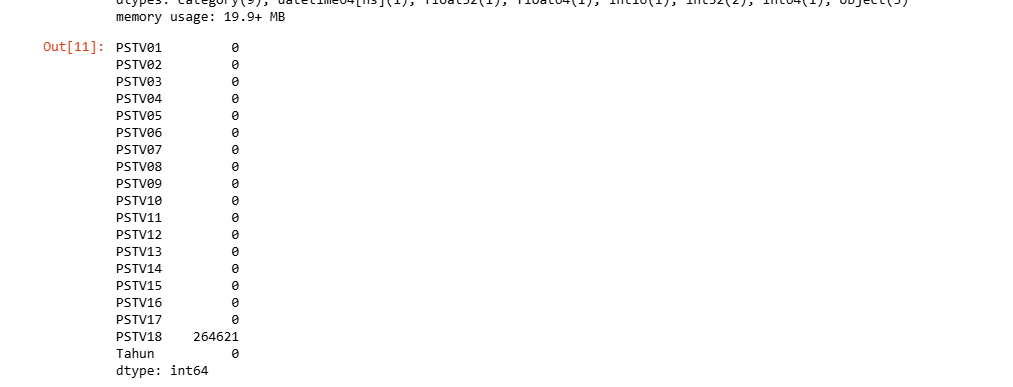
Dari hasil cek data untuk data pelayanan nonkapitasi , terlihat bahwa variabel/atribut tidak memiliki missing values. Untuk data tersebut memiliki 487 data duplikat.



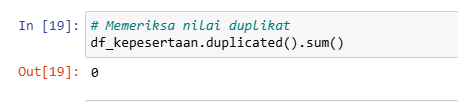
Selanjutnya, kami melakukan penggabungan data kepesertaan tahun 2019, 2020, 2021 menjadi sebuah dataset.



Selanjutnya, memeriksa dataset kepesertaan mulai dari info tipe data hingga informasi statistik dari dataset tersebut. Kami juga menambahkan satu buah kolom “Tahun” yang berisikan value untuk tahun yang digunakan (2019, 2020, 2021).



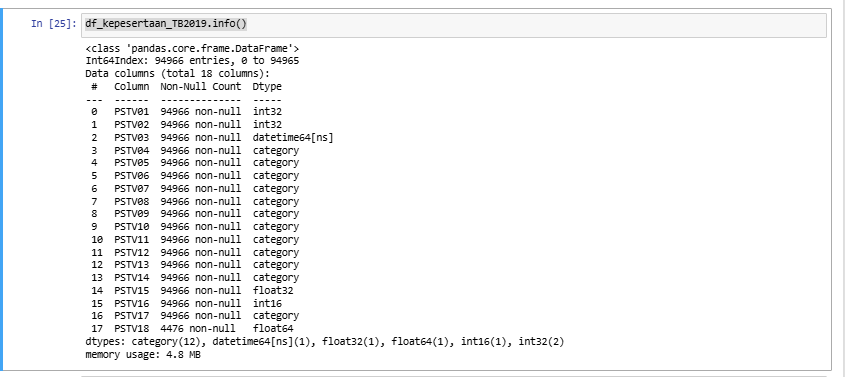
Dari informasi diatas, dapat dilihat bahwa dataset kepesertaan memiliki missing value sebanyak 264621 di atribut PSTV18.



Dan data kepesertaan ini tidak memiliki data duplikat. Selanjutnya, akan dilakukan cleaning data untuk dataset kepesertaan.

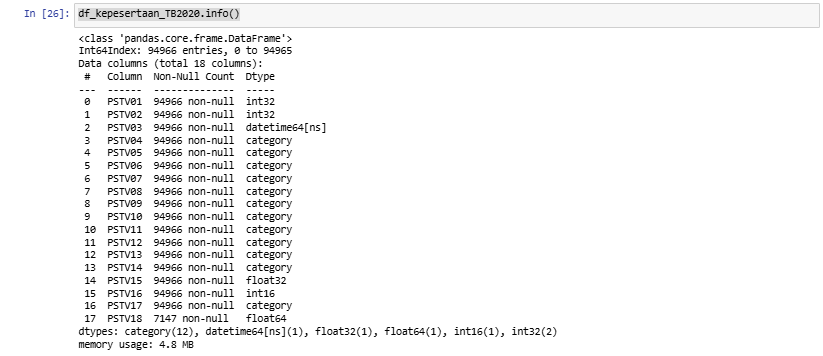
**2.3.2 Validasi Format dan Tipe Data**

1. **TB2019\_kepesertaan.dta**

****

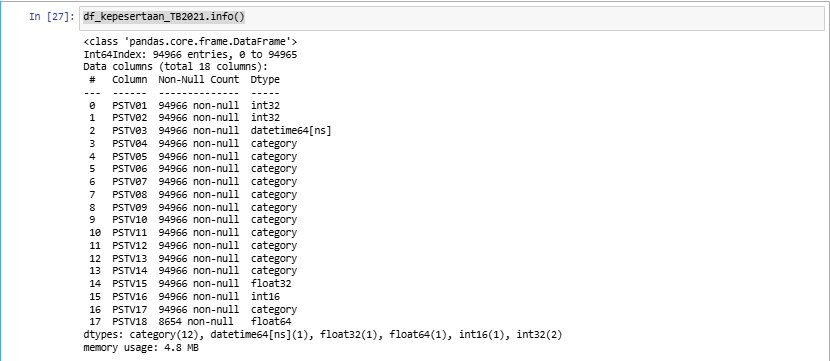
Berdasarkan output df\_kepesertaan\_TB2019.info(), dataset ini terdiri dari 94.966 baris dan 18 kolom dengan penggunaan memori 4,8 MB. Sebagian besar kolom memiliki data lengkap, kecuali PSTV18 yang hanya terisi sekitar 4,7% sehingga perlu penanganan *missing value*. Kolom PSTV01 dan PSTV02 bertipe numerik, PSTV03 bertipe tanggal, 12 kolom lainnya bertipe kategorikal (PSTV04–PSTV14 dan PSTV17), serta PSTV15 dan PSTV18 bertipe desimal. Dataset ini cukup rapi dan siap dianalisis lebih lanjut, terutama dengan fokus pada distribusi kategori, statistik numerik, dan penanganan data kosong.

1. **TB2020\_kepesertaan.dta**

****

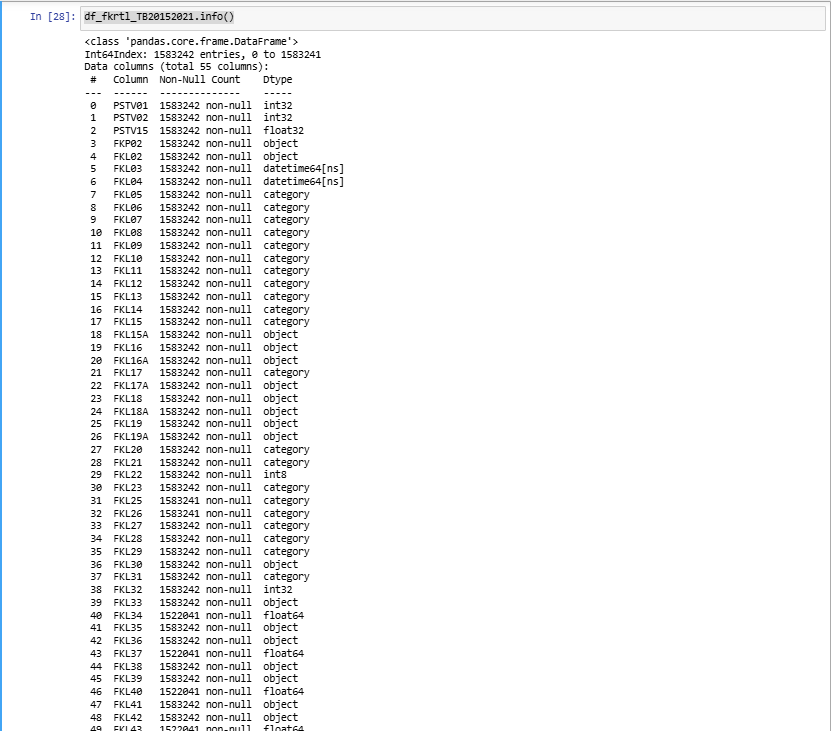
Dataset TB2020\_kepesertaan.dta berisi 94.966 baris dan 18 kolom dengan memori 4,8 MB. Semua kolom terisi penuh kecuali PSTV18 yang hanya memiliki 8.654 data (sekitar 9,1%), sehingga perlu penanganan *missing value*. Struktur data mirip dengan tahun 2019: terdapat kolom numerik, tanggal, dan 12 kolom kategorikal.

1. **TB2021\_kepesertaan.dta**

****

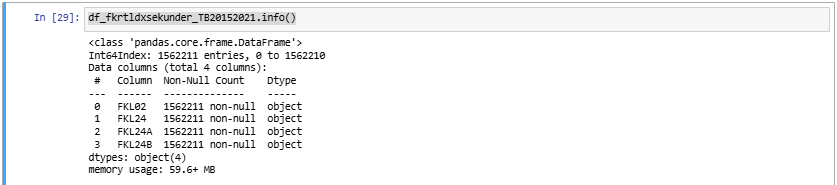
Dataset df\_kepesertaan\_TB2021 memiliki 94.966 baris dan 18 kolom dengan ukuran memori sekitar 4,8 MB. Mayoritas kolom berisi data bertipe kategorikal, sementara kolom PSTV03 berupa tanggal, dan beberapa kolom lainnya bertipe numerik seperti *int*, *float*, dan *datetime*. Sama seperti tahun 2019, 2020 hanya kolom PSTV18 yang memiliki *missing values* (sekitar 9,1%), yang perlu dibersihkan sebelum analisis lanjutan.

1. **fkrtl\_TB20152021**

****

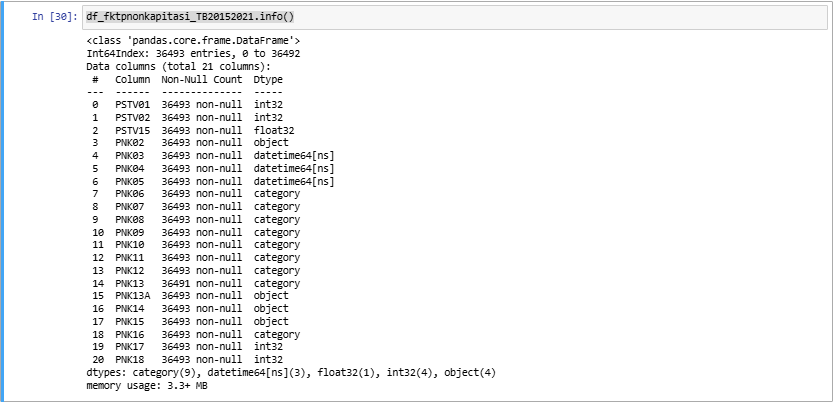
Dataset df\_fkr1\_TB20152021 terdiri dari 1.583.242 entri dan 55 kolom, dengan sebagian besar data terisi lengkap (non-null). Tipe data yang terdapat dalam dataset ini cukup beragam, meliputi tipe numerik (int32, int8, float32, float64), kategori, waktu (datetime64[ns]), serta tipe objek (teks). Beberapa kolom seperti PSTV01, PSTV02, dan PSTV03 berisi data numerik, sementara kolom seperti FKL06, FKL10, dan FKL25 menggunakan tipe kategori yang sesuai untuk kebutuhan klasifikasi. Selain itu, kolom FKL03 dan FKL04 menyimpan informasi dalam format waktu yang dapat dimanfaatkan untuk analisis temporal. Meskipun sebagian besar kolom memiliki data yang lengkap, terdapat beberapa kolom seperti FKL27, FKL40, dan FKL41 yang memiliki data kosong, sehingga memerlukan penanganan khusus pada tahap praproses data.

1. **fkrtldxsekunder\_TB20152021**

****

DataFrame ini memiliki 1562211 baris dan 4 kolom, yaitu FKL02, FKL24, FKL24B, dan FKL24C. Keempat kolom ini bertipe object, yang biasanya berarti data berupa string atau campuran (bisa juga angka yang dibaca sebagai teks). Semua kolom memiliki jumlah nilai non-null yang sama, yaitu 1562211, artinya tidak ada data yang hilang di dataset ini. Ukuran memorinya sekitar 59.6 MB. Karena semua kolom bertipe object, kemungkinan data ini berisi informasi kategori atau identitas fasilitas/instansi secara tekstual, seperti kode fasilitas, jenis layanan, atau lokasi.

1. **fktp\_nonkapitasi**

****

DataFrame ini berisi 36493 baris dan 21 kolom. Struktur data lebih kompleks dibandingkan dataset sebelumnya. Kolom seperti PSTV01, PSTV02, PSTV15, PNK17, dan PNK18 bertipe integer, menunjukkan data numerik diskrit (kemungkinan kode atau jumlah). Kolom PSTV15 bertipe float32, mengindikasikan nilai desimal (bisa jadi total biaya, skor, atau rasio). Ada juga kolom bertipe datetime64, seperti PNK03 hingga PNK06, yang merepresentasikan tanggal/waktu. Selain itu, ada banyak kolom bertipe category, yang sangat berguna untuk efisiensi memori bila data berisi kategori berulang. Data ini memakan memori sekitar 3.3 MB, jauh lebih kecil karena pemanfaatan tipe category.

**BAB III**

**DATA PREPARATION**

**3.1 Data Selection**

Pemilihan data mencakup pemilihan dataset, periode waktu, serta variabel-variabel penting yang merepresentasikan karakteristik wilayah peserta Tuberkulosis (TB) secara menyeluruh.

**3.1.1 Sumber Data**

Data yang digunakan bersumber dari dataset kontekstual Tuberkulosis milik BPJS Kesehatan Tahun 2022, yang mencakup data kepesertaan dan pelayanan kesehatan TB dalam rentang waktu tahun 2019-2021, diantaranya :

1. TB2019\_kepesertaan.dta : Data peserta Tuberkulosis tahun 2019
2. TB2020\_kepesertaan.dta : Data peserta Tuberkulosis tahun 2020
3. TB2021\_kepesertaan.dta : Data peserta Tuberkulosis tahun 2021
4. TB20152021\_fkrtl.dta : Data pelayanan Tuberkulosis di rumah sakit
5. TB20152021\_fktpnonkapitasi.dta : Data pelayanan Tuberkulosis di fasilitas kesehatan tingkat pertama (fktp) non-kapitasi

**3.1.2 Pemilihan Variabel**

**Pemilihan Variabel dari data Kepesertaan**

| Kode | Nama Variabel | Katerangan | Kategori |
| --- | --- | --- | --- |
| PSTV01 | Nomor Peserta | Digunakan sebagai kunci join antar dataset | Identitas |
| PSTV03 | Tanggal Lahir | Digunakan untuk menghitung usia peserta | Demografi |
| PSTV05 | Jenis Kelamin | Menentukan proporsi laki-laki / perempuan | Demografi |
| PSTV07 | Kelas Rawat | Indikator sosial ekonomi peserta | Sosial Ekonomi |
| PSTV08 | Segmentasi Peserta | Menunjukkan jenis kepesertaan | Sosial Ekonomi |
| PSTV09 | Provinsi Domisili | Lokasi wilayah peserta | Geografi |
| PSTV10 | Kabupaten/Koya Domisili | Unit analisis utama segmentasi wilayah | Geografi |
| PSTV12 | Jenis Faskes | FKTP atau FKRTL | Layanan |
| PSTV13, PSTV14 | Lokasi Faskes Terdaftar | Untuk validasi layanan dan domisili | Layanan |

**Pemilihan Variabel dari data FKRTL**

| **Kode** | **Nama Variabel** | **Keterangan** | **Kategori** |
| --- | --- | --- | --- |
| PSTV01 | Nomor Peserta | Untuk join dengan data kepesertaan | Identitas |
| FKL03 | Tanggal Datang | Untuk menganalisis distribusi kunjungan pelayanan kesehatan di fasilitas tingkat lanjut. |  |
| FKL04 | Tanggal Pulang | Dapat digunakan untuk menghitung durasi kunjungan dan melihat pola perawatan. |  |
| FKL06 | Provinsi FKRTL | Menunjukkan lokasi fasilitas kesehatan tingkat lanjut, membantu dalam analisis geografi pelayanan. |  |
| FKL10 | Tingkat Pelayanan FKRTL | Menunjukkan tingkat pelayanan di fasilitas kesehatan rujukan, yang memberikan informasi tentang intensitas perawatan yang diterima peserta. |  |

**Pemilihan Variabel dari data FKTP**

| **Kode** | **Nama Variabel** | **Keterangan** | **Kategori** |
| --- | --- | --- | --- |
| PSTV01 | Nomor peserta | Untuk menghubungkan data pelayanan dengan data kepesertaan. |  |
| PNK03 | Tanggal Kunjungan | Dapat digunakan untuk melihat waktu kunjungan dan analisis durasi perawatan di FKTP. |  |
| PNK04 | Tanggal Tindakan | Menyediakan informasi tentang tanggal tindakan medis yang diterima peserta. |  |

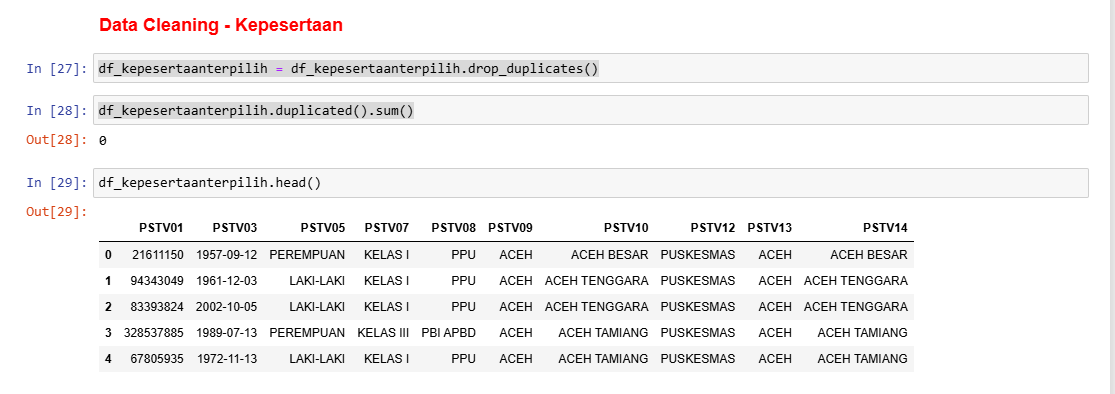
**3.2 Data Cleaning**

Dari data kepesertaan yang telah digabung sebelumnya (tahun 2019, 2020, 2021) , data fkrtl , dan data fktp non-kapitasi dengan pengimplementasian atribut-atribut yang digunakan juga, terdapat beberapa values yang duplikat diantaranya masing-masing adalah sebagai berikut :

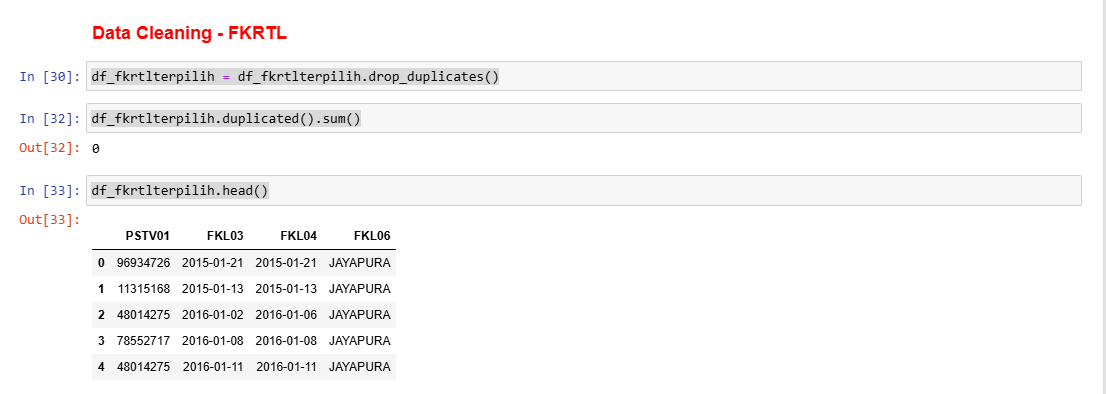
1. Kepesertaan (df\_kepesertaanterpilih) : memiliki 166989 data yang duplikat
2. FKRTL (df\_fkrtlterpilih) : memiliki 1192 data yang duplikat
3. FKTP-Non Kapitasi (df\_fktpnon) : memiliki 15880 data yang duplikat

Oleh karena itu, sebelum melakukan penggabungan tiap data, masing-masing dataset harus melakukan data cleaning guna untuk mempersiapkan dataset yang lebih berguna dan informatif.

**3.2.1 Data Cleaning - Kepesertaan (df\_kepesertaanterpilih)**

****

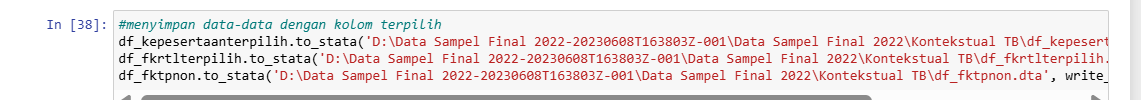
**3.2.2 Data Cleaning - FKRTL (df\_fkrtlterpilih)**

****

**3.2.3 Data Cleaning - FKTP Non-Kapitasi (df\_fktpnon)**

****

Setelah melakukan data cleaning untuk masing-masing df\_kepesertaanterpilih, df\_fkrtlterpilih, dan df\_fktpnon maka ketiga file disimpan dalam bentuk format dta. Berikut implementasi kode python nya.

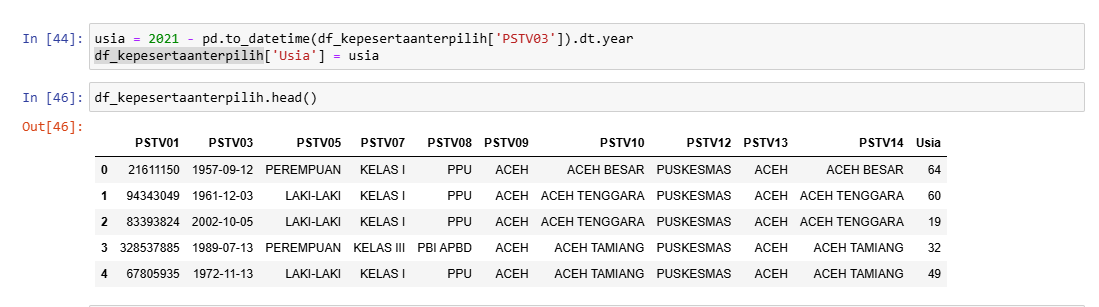


**3.3 Data Construct**

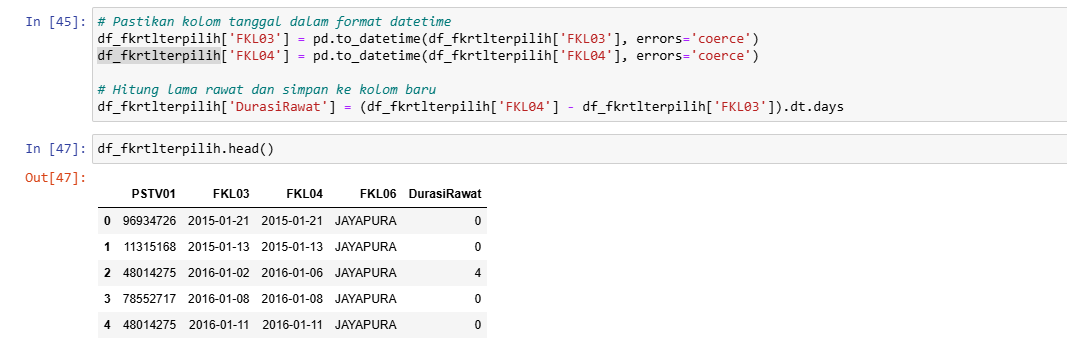
Data Construct bertujuan untuk membangun fitur atau variabel baru yang akan digunakan dalam analisis. Langkah-langkah ini penting untuk menggali lebih dalam dari data yang ada dan menciptakan representasi baru yang lebih berguna untuk model analitik K-Means Clustering.

Beberapa fitur yang akan ditambahkan dalam pengerjaan proyek ini diantaranya : usia peserta, durasi rawat,frekuensi kunjungan FKTP, frekuensi kunjungan FKRTL.

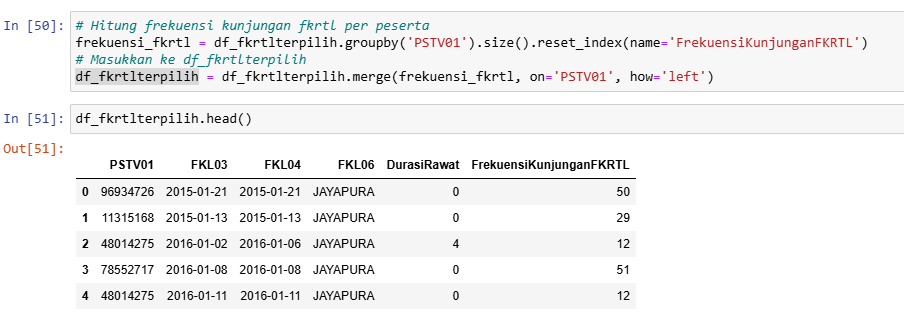
**3.3.1 Usia Peserta**

****

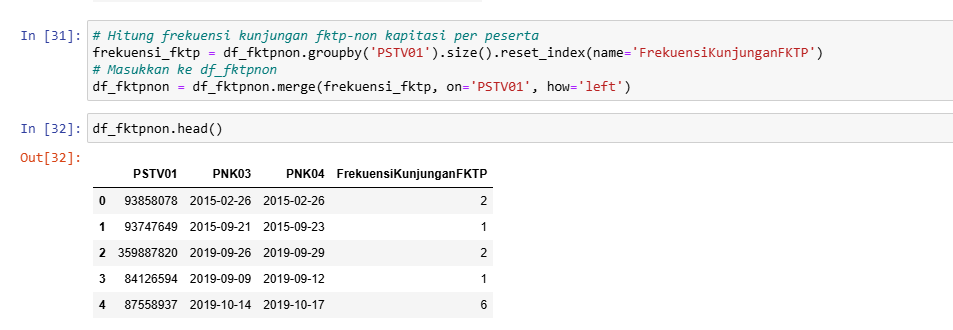
**3.3.2 Durasi Rawat FKRTL**

****

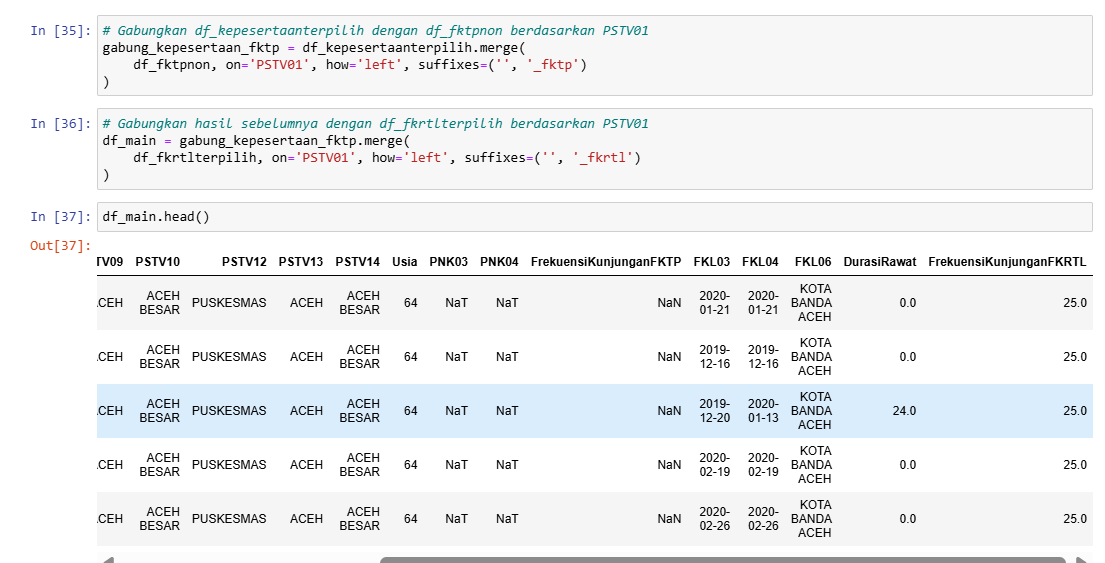
**3.3.3 Frekuensi Kunjungan FKRTL**

****

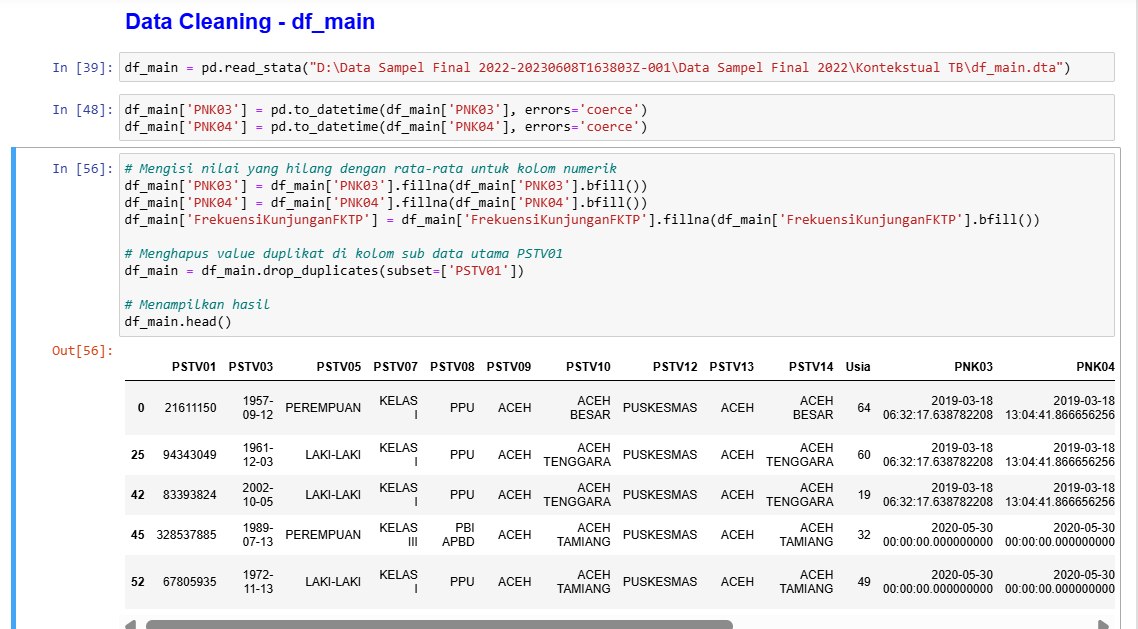
**3.3.4 Frekuensi Kunjungan FKTP Non-Peserta**

****

Selanjutnya, membuat satu dataframe yang berisikan atribut-atribut yang dibutuhkan.



Dalam DataFrame yang telah digabungkan ini, terdapat beberapa atribut/kolom yang masih menyimpan nilai kosong (NaN). Oleh karena itu, selanjutnya melakukan data cleaning untuk dataframe utama.



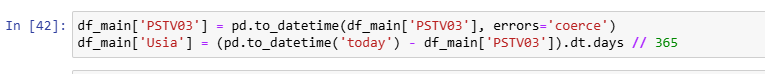
Di data cleaning df\_main ini, semua atribut/kolom telah berisikan nilai tanpa adanya nilai duplikasi di subset kolom utamanya (PSTV01).

**3.4 Labeling Data**

Ada beberapa proses labeling data yang kami gunakan untuk tugas kali ini. Masing-masing label kami gunakan dengan mengintegrasikan kolom lain sebagai sumber informasinya.

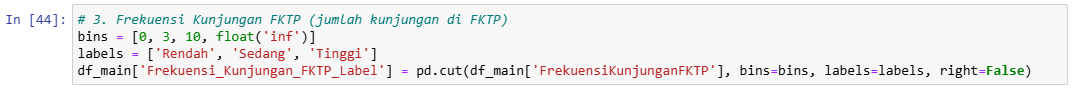
**3.4.1 Label Berdasarkan Usia**

Menghitung usia peserta berdasarkan tanggal lahir, yang dapat digunakan untuk analisis lebih lanjut (misalnya segmentasi berdasarkan usia).



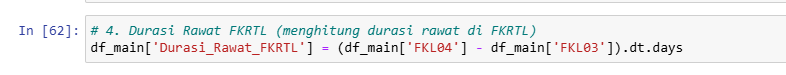
**3.4.2 Frekuensi Kunjungan FKTP**

Kolom yang Digunakan adalah PNK03 atau PNK04 (Tanggal Kunjungan FKTP). Tujuan: Menghitung berapa kali peserta mengunjungi fasilitas kesehatan tingkat pertama (FKTP).

****

**3.4.5 Durasi Rawat FKRTL**

Kolom yang Digunakan: FKL03 (Tanggal Datang FKRTL), FKL04 (Tanggal Pulang FKRTL). Tujuan: Menghitung durasi rawat di fasilitas kesehatan tingkat lanjutan (FKRTL).



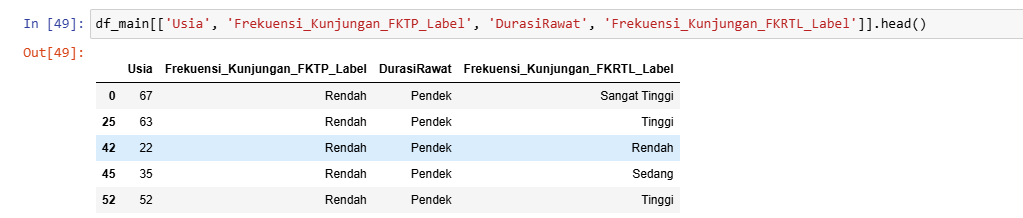
**3.4.5 Frekuensi Kunjungan FKRTL**

Kolom yang Digunakan: FKL03 (Tanggal Datang FKRTL)

Tujuan: Menghitung berapa kali peserta mengunjungi fasilitas kesehatan tingkat lanjutan (FKRTL).



Dataframe df\_main akan berisi kolom baru yang berfungsi sebagai label untuk analisis atau modeling lebih lanjut, misalnya dalam segmentasi atau prediksi, seperti dibawah ini :



**3.5 Data Integration**

Setelah memiliki dataframe utama , dimana di dalamnya sudah menggunakan atribut/kolom yang akan digunakan untuk tugas ini, selanjutnya akan dilakukan data integration.