**Rencana Tugas Mandiri Ke-2**

**Mata Kuliah Analisis Data Eksploratif**



**Mata Kuliah Analisis Data Eksploratif**

**Bentuk-bentuk Transformasi Data dan Analisis Data**

**KELAS A**

**DISUSUN OLEH :**

AHMAD ARDHY ANSYAH ( 20083010008 )

**DOSEN PENGAMPU:**

TRESNA MAULANA FAHRUDIN,/ S.S.T., MT ( 20219930501200 )

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

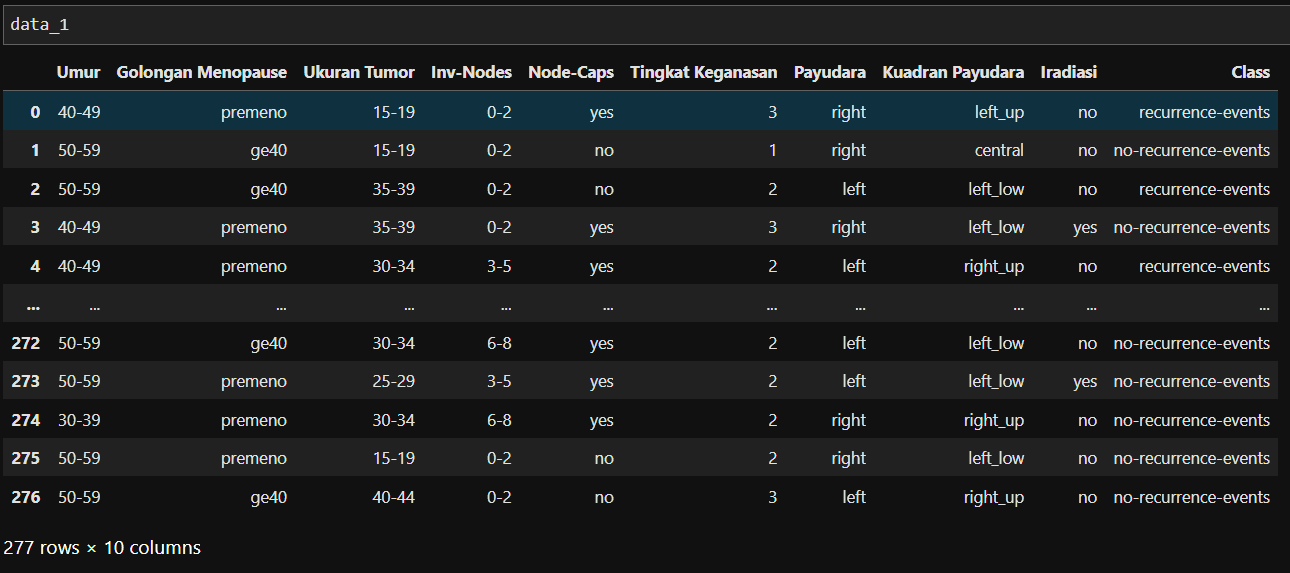
JAWA TIMUR

2022

**Rencana Tugas Mandiri ke-2**

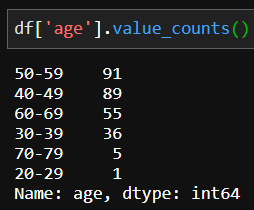
**Mata Kuliah Analisis Data Eksploratif**

1. Eksplorasi Pertanyaan Dasar (Studi Kasus: Breast Cancer Dataset)
2. Deskripsikan definisi masing-masing atribut pada dataset kanker payudara berdasarkan pengertian di dalam domain keilmuan medis dan referensi yang dapat dipertanggungjawabkan

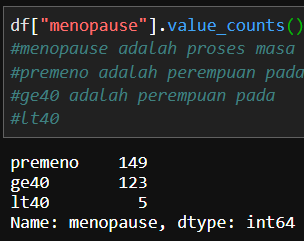


Sumber : *Using machine learning techniques to predict the recurrence of breast cancer | LinkedIn* (no date). Available at: https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva/ (Accessed: 7 October 2022).

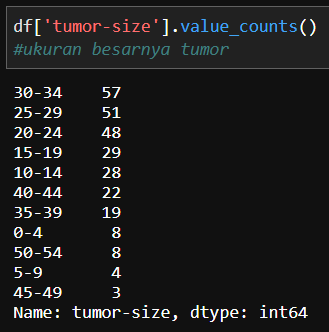
* Age : Usia pasien pada saat didiagnosis mengidap Kanker Payudara



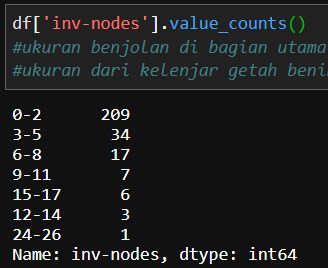
* Menopause : adalah proses masa berhentinya haid yang permanen akibat hilangnya aktivitas folikuler ovarium. Lalu isi dari data menopause ini ada 3 golongan yaitu premeno, ge40, dan lt40 pada saat didiagnosis.



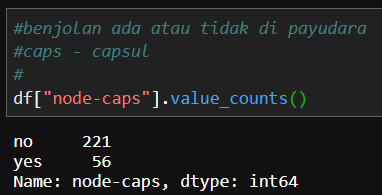
* Ukuran Tumor : Ukuran dari diameter tumor (dalam satuan mm)



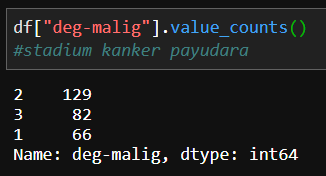
* Inv-Nodes : Jumlah (berkisar 0-39) kelenjar getah bening yang terkandung di payudara, terlihat pada saat pemeriksaan histologis.



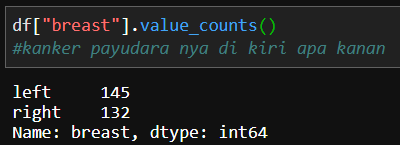
* Node-Caps : adalah kondisi kelenjar getah bening belum/tidak menembus kapsul dan sekitarnya



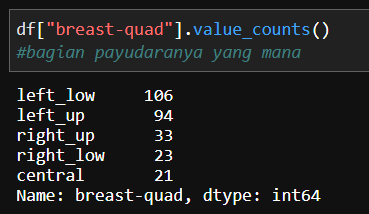
* Degmalig (tingkat keganasan) : Tingkat keganasan dari Kanker Payudara yang bernilai 1,2,3



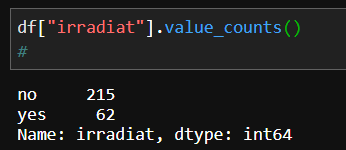
* Breast(payudara) : letak dari adanya kanker payudara, apakah di payudara kanan atau kiri



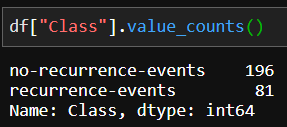
* Breast-quad(kuadran payudara) : Payudara dapat dibagi menjadi 4 kuadran, puting payudara sebagai titik pusat.



* irradiat(iradiasi) : Terapi radiasi merupakan pengobatan yang menggunakan sinar x berenergi tinggi untuk menghancurkan sel-sel kanker. Isi dari data merupakan sudah/belum dilakukan terapi iradiasi oleh pengidap kanker.

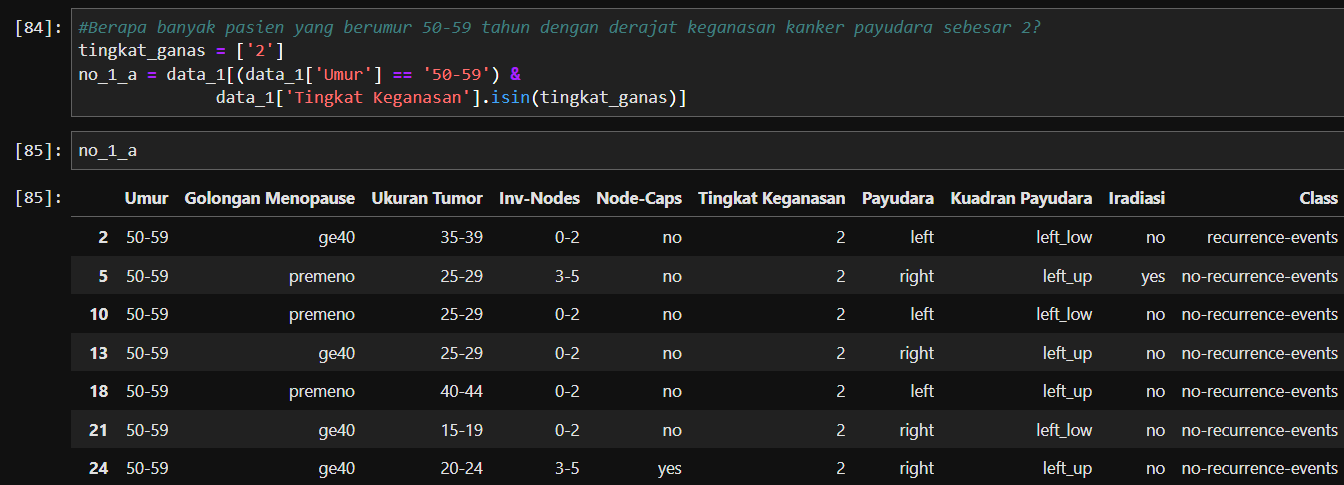


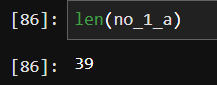
* Class : adalah klasifikasi apakah pengidap kanker mengalami kejadian berulang, apabila berulang akan masuk klasifikasi ‘recurrence-events’, apabila tidak ada kejadian berulang masuk klasifikasi no-recurrence-events.



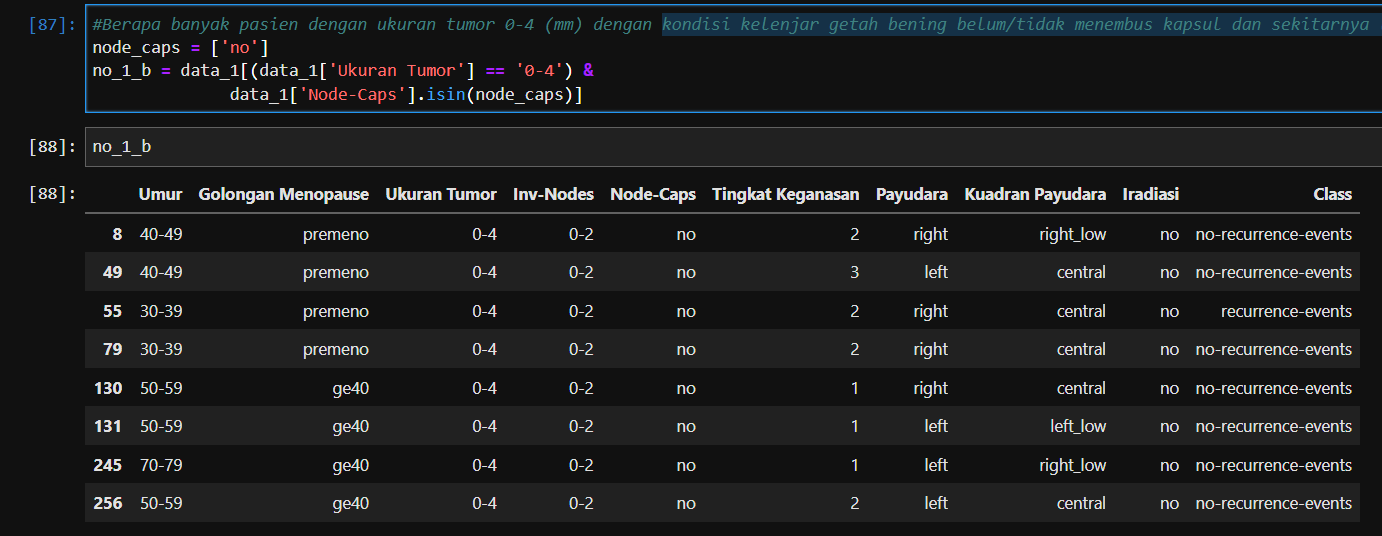
1. Lakukan eksplorasi pada dataset kanker payudara menggunakan Python dan jawablah pertanyaan sebagai berikut:

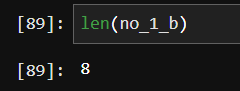
* Berapa banyak pasien yang berumur 50-59 tahun dengan derajat keganasan kanker payudara sebesar 2?



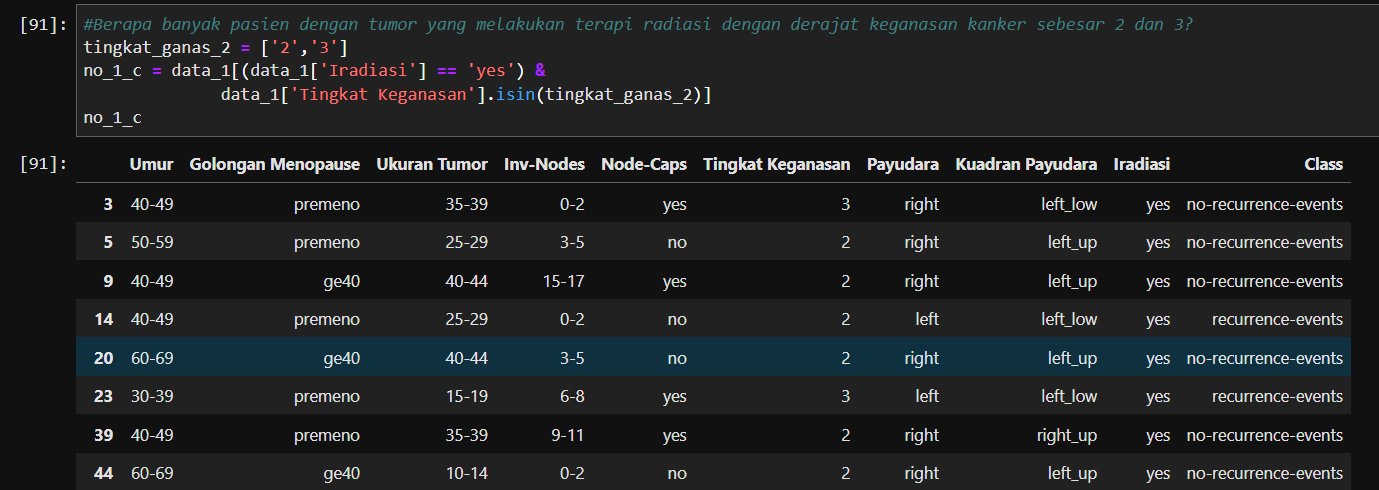


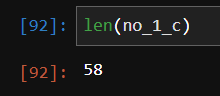
* Berapa banyak pasien dengan ukuran tumor 0-4 (mm) dengan kondisi kelenjar getah bening belum/tidak menembus kapsul dan sekitarnya (node-caps)?



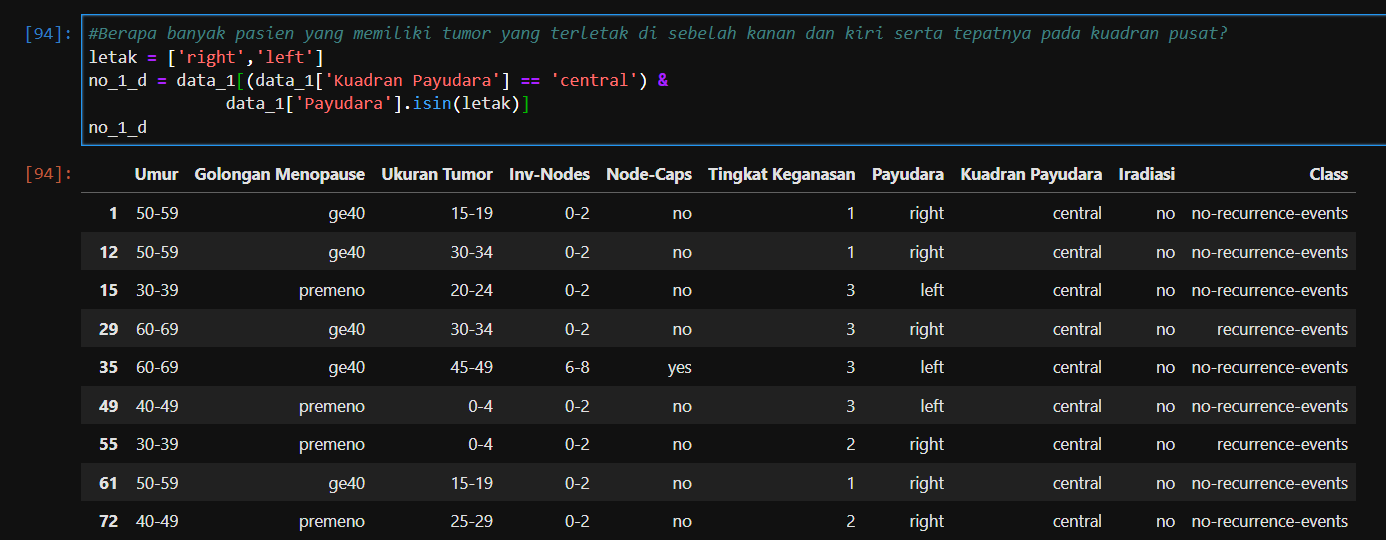


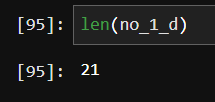
* Berapa banyak pasien dengan tumor yang melakukan terapi radiasi dengan derajat keganasan kanker sebesar 2 dan 3?



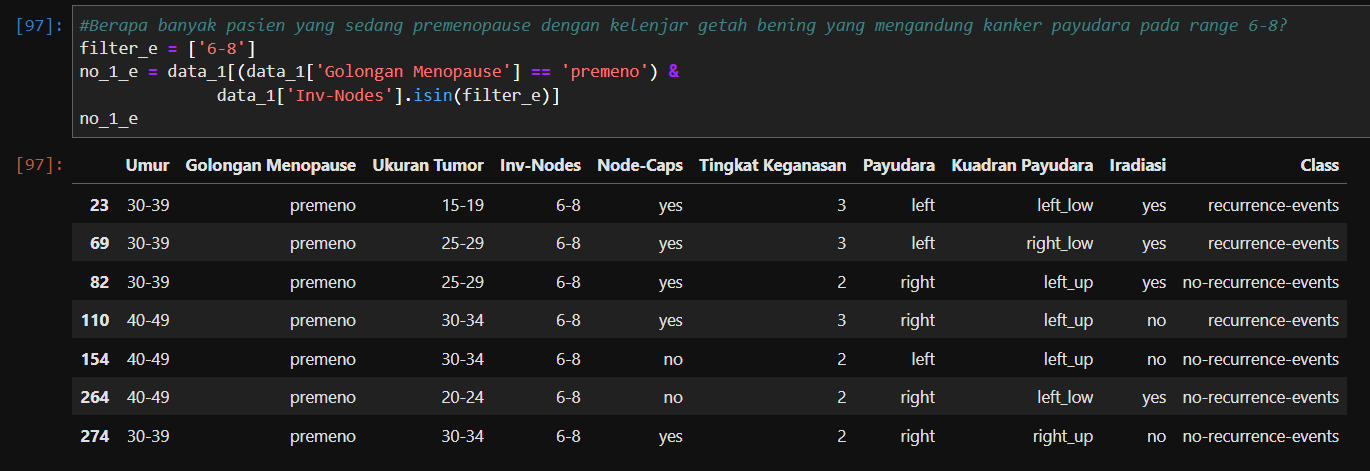


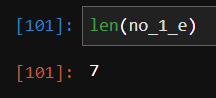
* Berapa banyak pasien yang memiliki tumor yang terletak di sebelah kanan dan kiri serta tepatnya pada kuadran pusat?



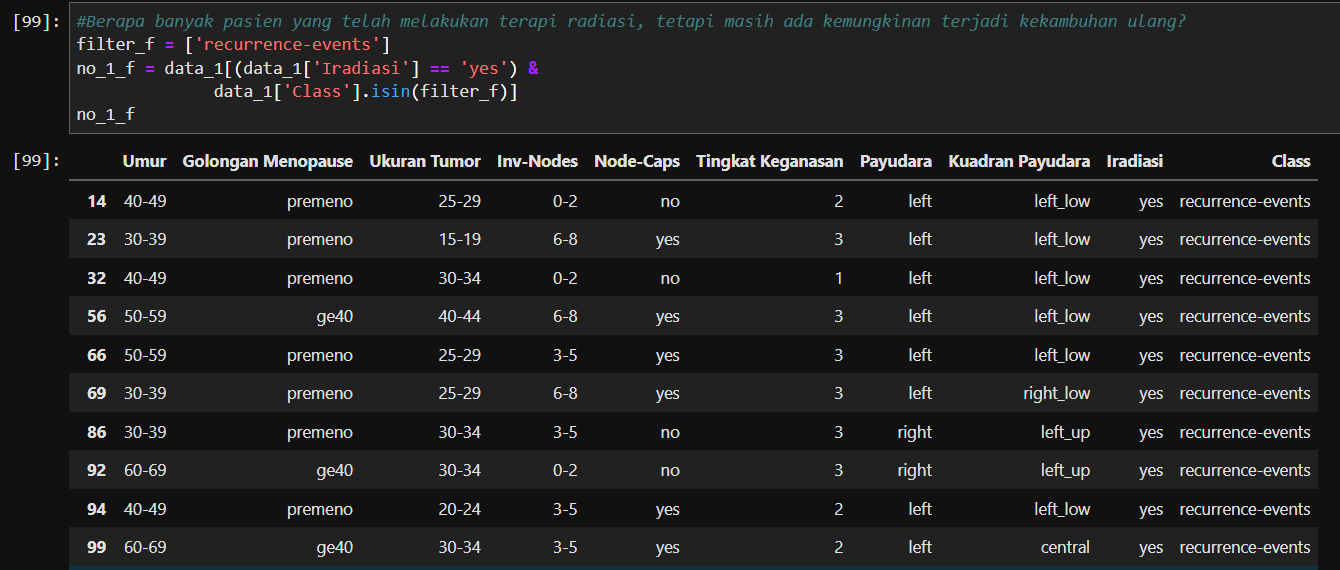


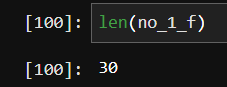
* Berapa banyak pasien yang sedang premenopause dengan kelenjar getah bening yang mengandung kanker payudara pada range 6-8?





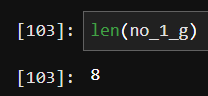
* Berapa banyak pasien yang telah melakukan terapi radiasi, tetapi masih ada kemungkinan terjadi kekambuhan ulang?





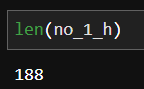
* Berapa banyak pasien yang masih berumur 30-39 tahun dengan kondisi kelenjar getah bening telah menembus kapsul dan sekitarnya (node-caps)?



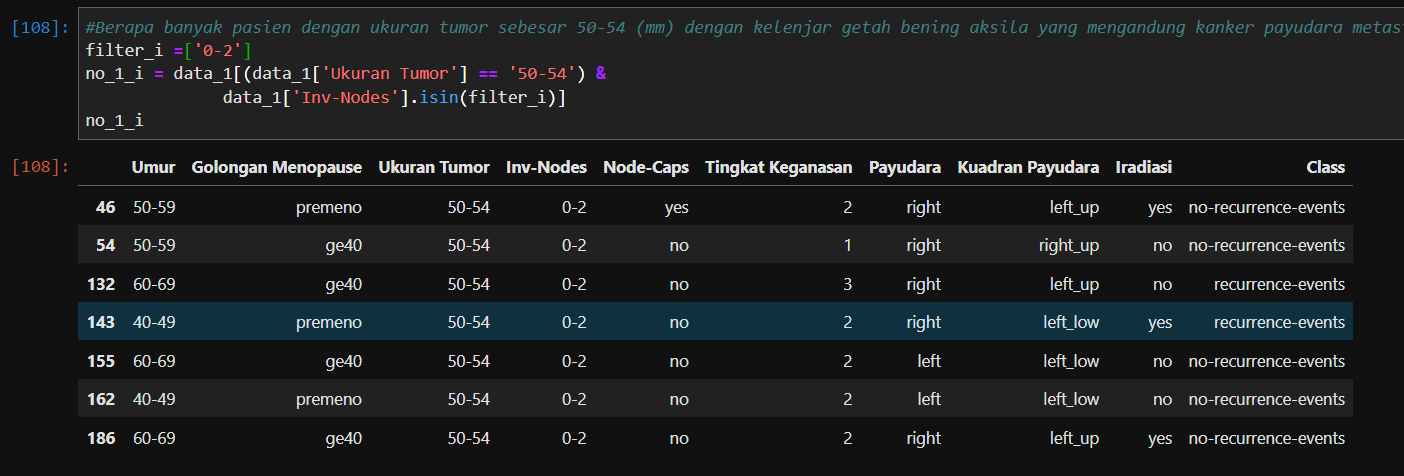


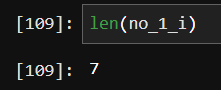
* Berapa banyak pasien yang menopause di atas umur 40 tahun, tetapi tidak melakukan terapi radiasi?



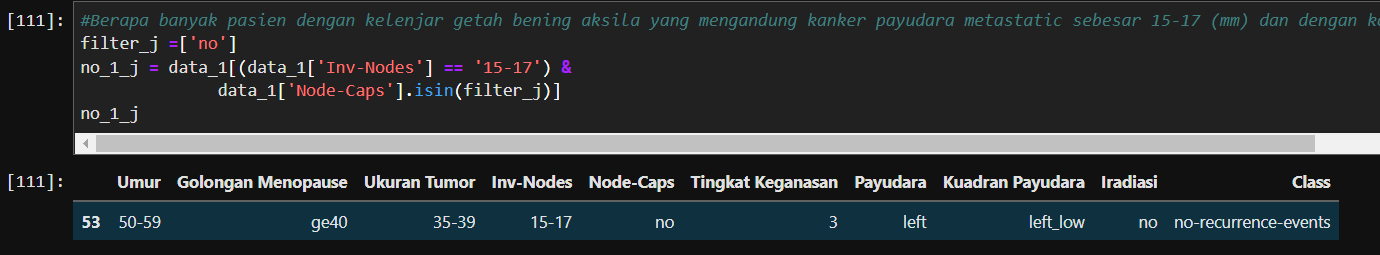


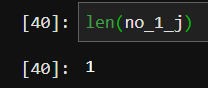
* Berapa banyak pasien dengan ukuran tumor sebesar 50-54 (mm) dengan kelenjar getah bening aksila yang mengandung kanker payudara metastatik sebesar 0-2 (mm)?





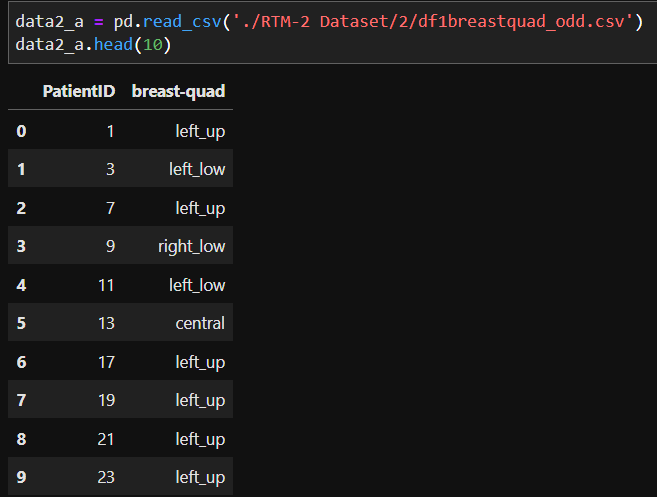
* Berapa banyak pasien dengan kelenjar getah bening aksila yang mengandung kanker payudara metastatic sebesar 15-17 (mm) dan dengan kondisi kelenjar getah bening belum/tidak menembus kapsul dan sekitarnya (*node-caps*)?

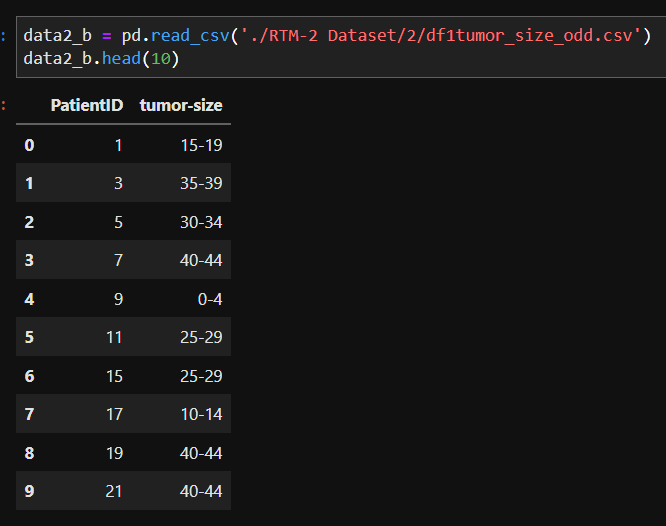


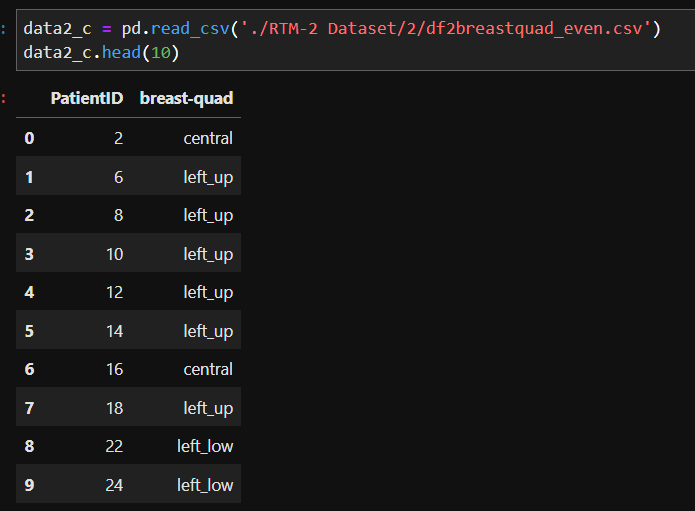


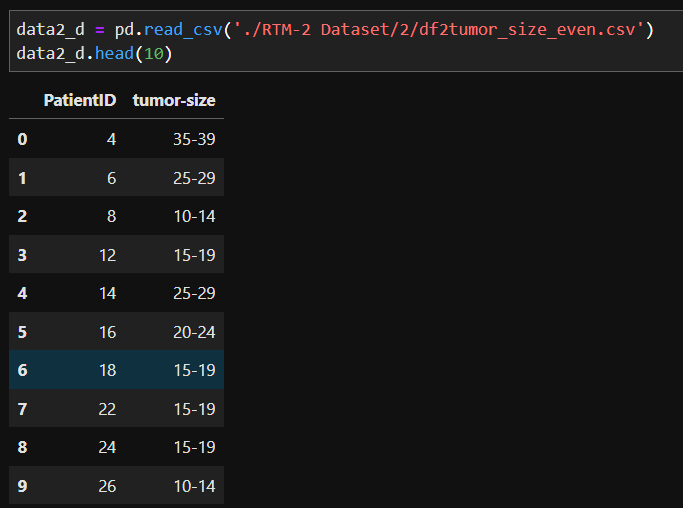
1. Penggabungan Data Menggunakan Inner Join, Left Join, Right Join, dan Outer Join (Studi Kasus: Breast Cancer Dataset)
2. Implementasikan penggabungan data (*merge*) menggunakan Inner Join, Left Join, Right Join, dan Outer Join menggunakan Python untuk menjawab pertanyaan berikut:

Import data :

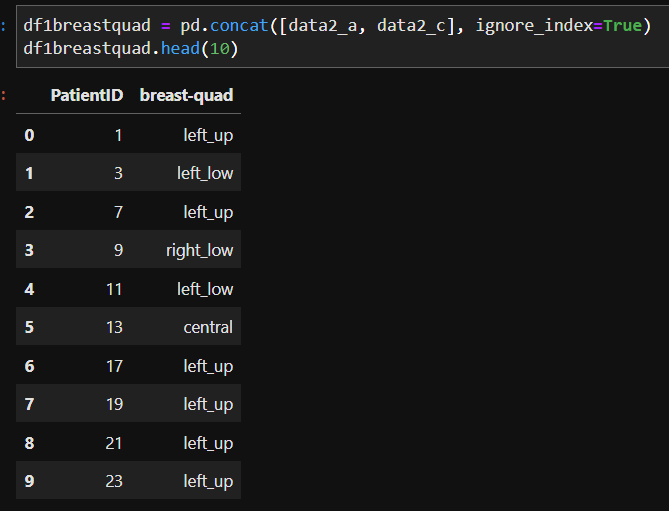


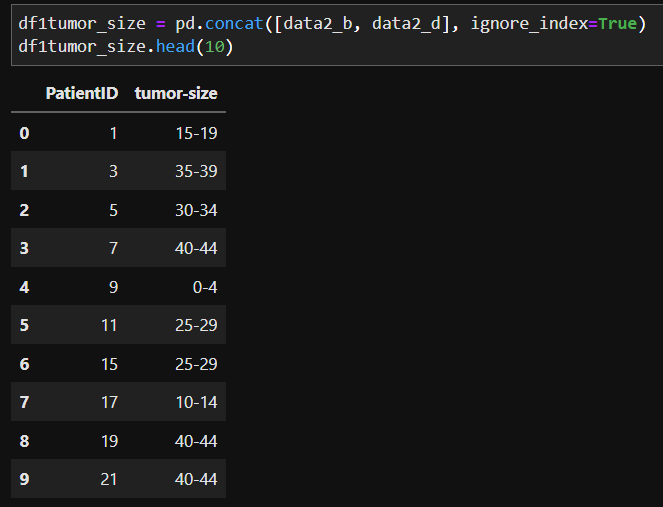






Sebelum melakukan penggabungan data dengan beberapa kriteria, perlu dilakukan penggabungan data antara data df1tumor\_size yang memiliki nilai PatientID ganjil dan genap serta df1breastquad yang memiliki nilai PatientID ganjil dan genap. Penggabungan data ini dilakukan dengan bantuan library Python Pandas Concat,

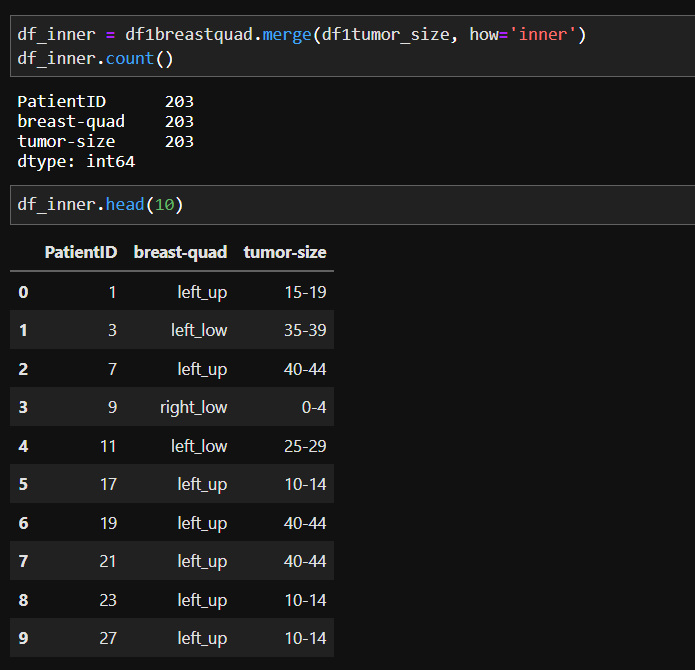




Code tersebut menggabungkan dengan menjadikan ke satu Columns

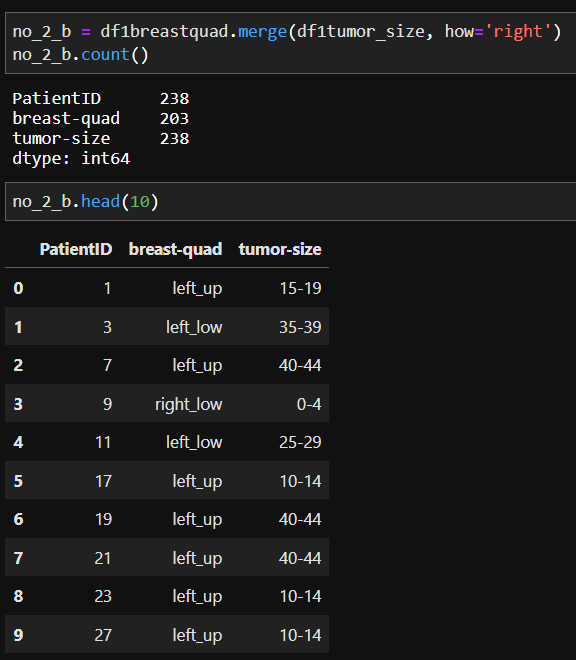
* Berapa banyak record pasien yang memiliki nilai atribut ukuran tumor dan nilai kuadran kanker payudara secara lengkap keduanya ?

Menggunakan sistem inner join

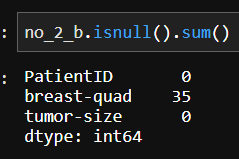


* Berapa banyak record pasien yang memiliki nilai atribut ukuran tumor, tetapi tidak memiliki nilai kuadran kanker payudara ? serta berapa banyak record nilai atribut kuadran kanker payudara yang NaN tersebut ?

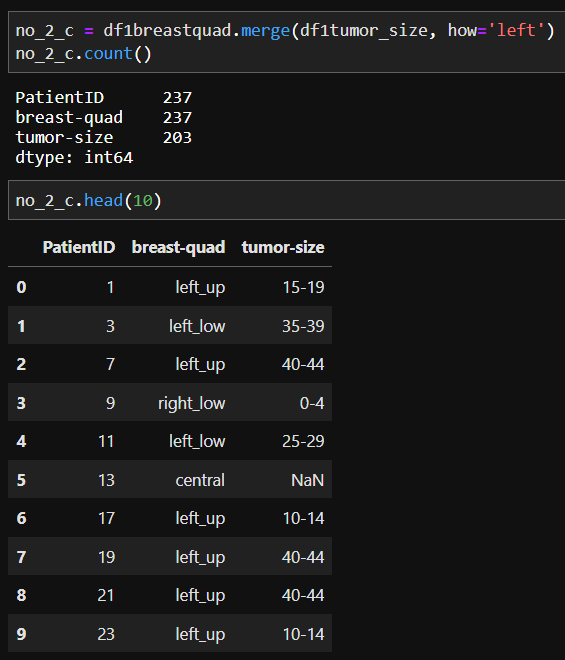
Menggunakan sistem merge right join



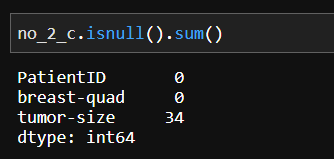
Berapa banyak record nilai atribut kuadran kanker payudara yang NaN tersebut?



* Berapa banyak record pasien yang tidak memiliki nilai atribut ukuran tumor, tetapi memiliki nilai kuadran kanker payudara ? serta berapa banyak record nilai atribut ukuran tumor yang NaN tersebut ?

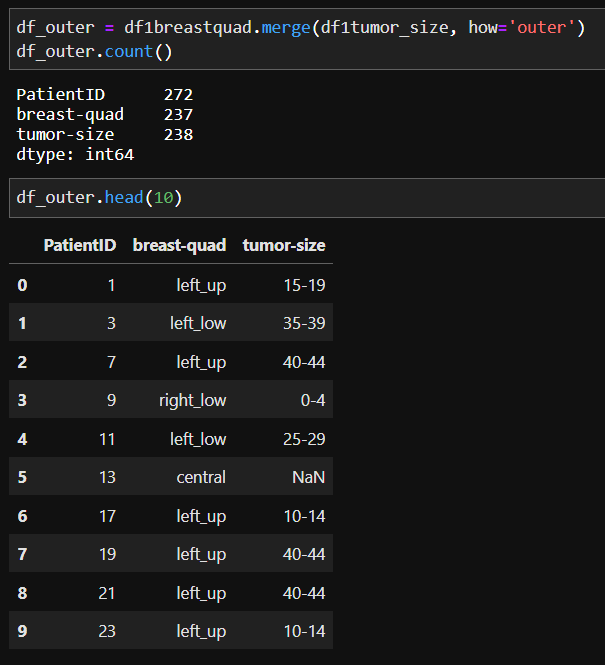


banyak record nilai atribut ukuran tumor yang NaN :

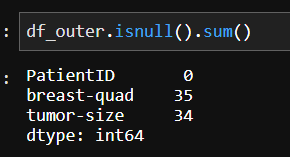


* Jika menggunakan Outer Join, Berapa banyak record pasien yang NaN pada masing-masing atribut, baik pada atribut ukuran tumor maupun kuadran kanker payudara?

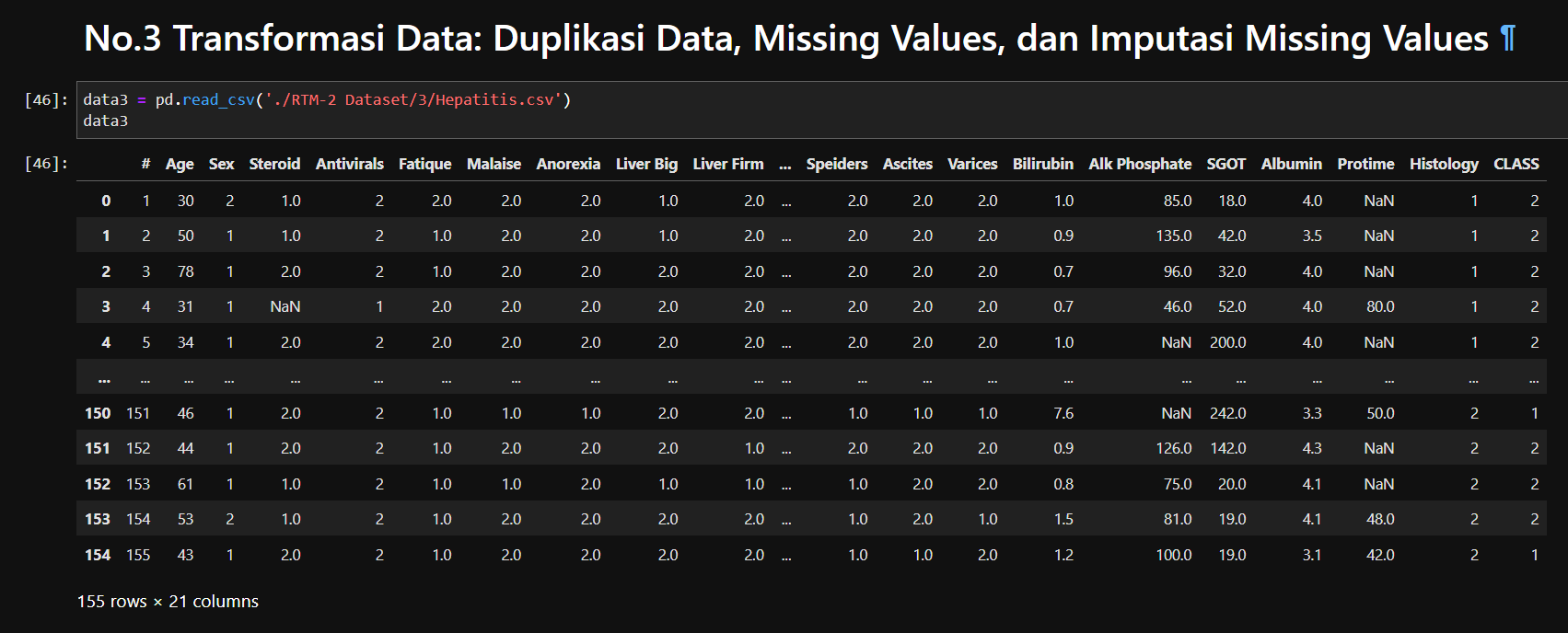
Mengunakan sistem merge outer join



Berapa banyak record pasien yang NaN pada masing-masing atribut, baik pada atribut ukuran tumor maupun kuadran kanker payudara?



1. Transformasi Data: Duplikasi Data, Missing Values, dan Imputasi Missing Values



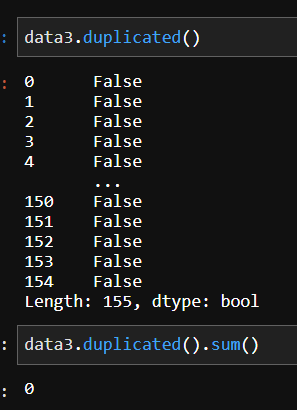
1. Deskripsikan definisi masing-masing atribut pada dataset hepatitis berdasarkan pengertian di dalam domain keilmuan medis dan referensi yang dapat dipertanggungjawabkan

Sumber : …..

* Age : adalah umur dari pengidap hepatitis
* Sex : adalah Jenis Kelamin dari pengidap hepatitis
* Steroid : adalah steroid
* Antivirals : adalah antivirus
* Fatique : adalah kelelahan
* Malaise : adalah
* Anorexia : adalah letih
* Liver Big : adalah pembesaran hati
* Liver Firm : adalah pengerasan hati
* Spleen Palpable : adalah
* Speiders : adalah
* Ascites : adalah
* Varices : adalah
* Bilirubin : adalah
* Alk Phosphate : adalah
* SGOT : adalah
* Albumin : adalah
* Protime : adalah
* Histology : adalah
* CLASS : adalah klasifikasi dari penggolongan pengidap hepatitis, apakah masuk kriteria 1 atau 0

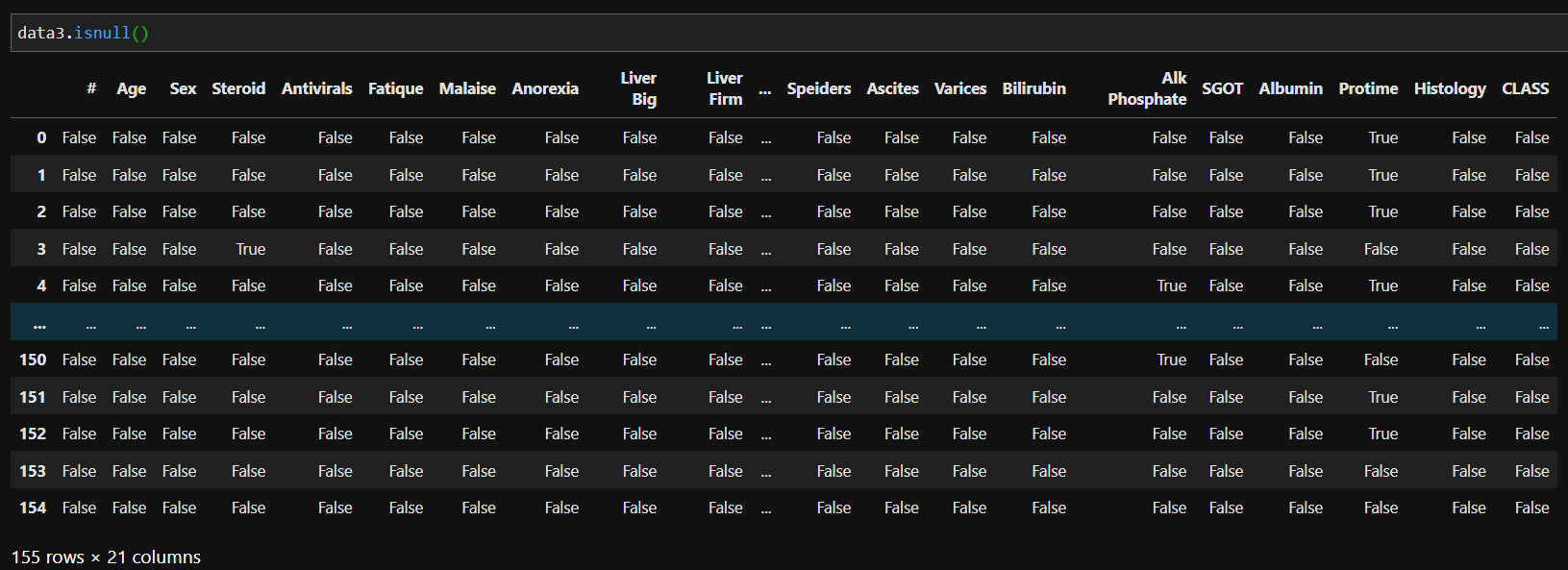
1. Lakukan transformasi data menggunakan Python dengan langkah-langkah sebagai berikut:

* Lakukan pemeriksaan apakah terdapat duplikasi data (row) pada dataset hepatitis



Dari hasil code tersebut tidak ditemukan adanya duplikasi data di data no 3 ini.

* Lakukan analisis deskriptif berapa banyak atribut yang memiliki missing values

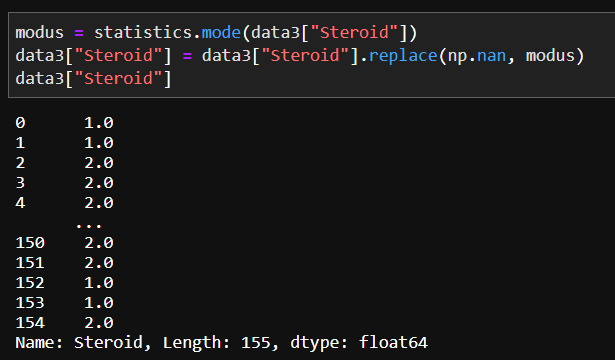




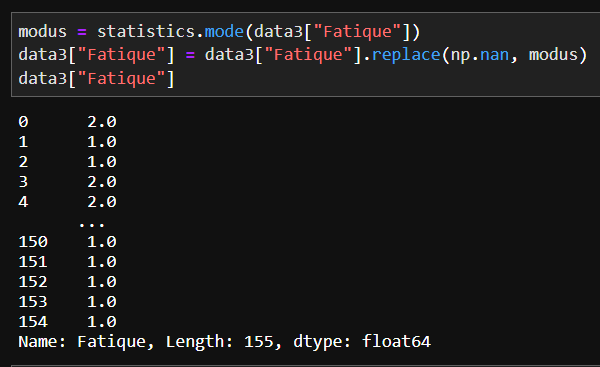
* Lakukan imputasi missing values dengan berbagai teknik yang efektif dan baik untuk mengisi atribut-atribut yang NaN, misalnya menggunakan mean, median, modus, clustering, regression, maupun metode taksiran dan prediksi lainnya.

Karena terdapat beberapa missing value pada data tersebut, sehingga perlu dilakukan penanganan missing value. Ada beberapa cara yang saya pakai dalam penanganan mising value, yaitu Mean, Modus

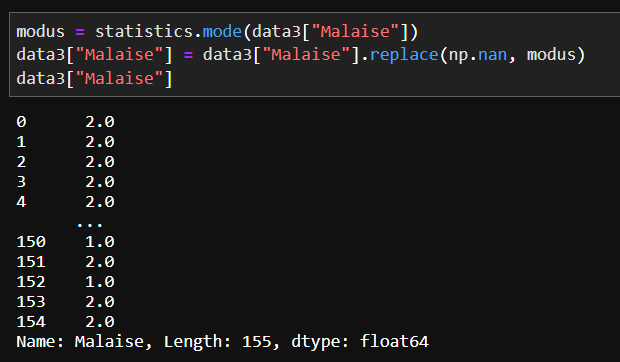
1. Penanganan Missing Value dengan Modus
2. Penanganan Missing Value pada Columns Data Steroid menggunakan modus



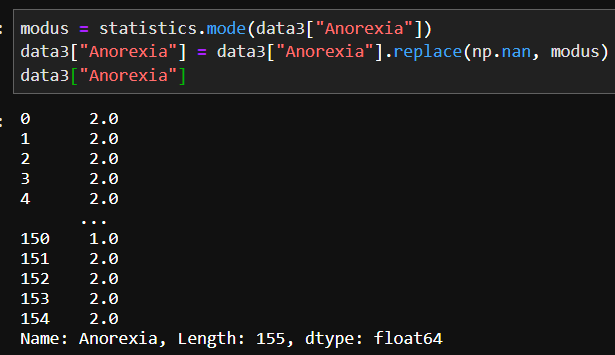
1. Penanganan Missing Value pada Columns Data Fatique menggunakan modus



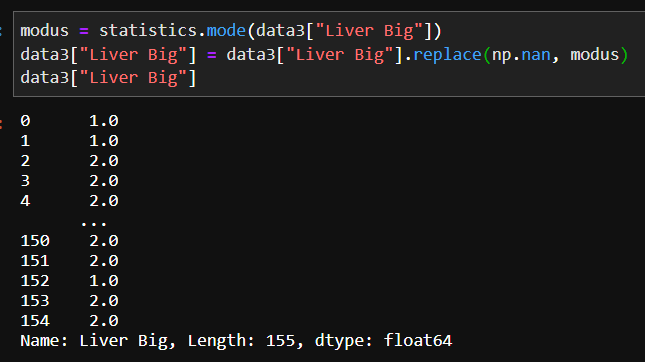
1. Penanganan Missing Value pada Columns Data Malaise menggunakan modus



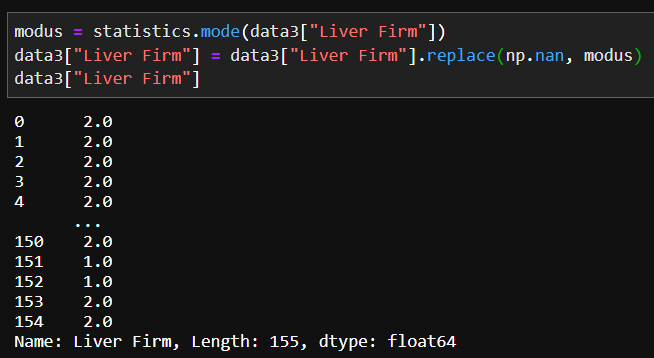
1. Penanganan Missing Value pada Columns Data Anorexia menggunakan modus



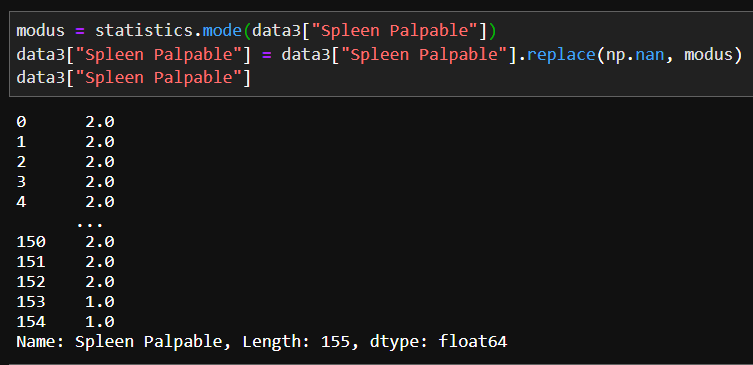
1. Penanganan Missing Value pada Columns Data Liver Big menggunakan modus



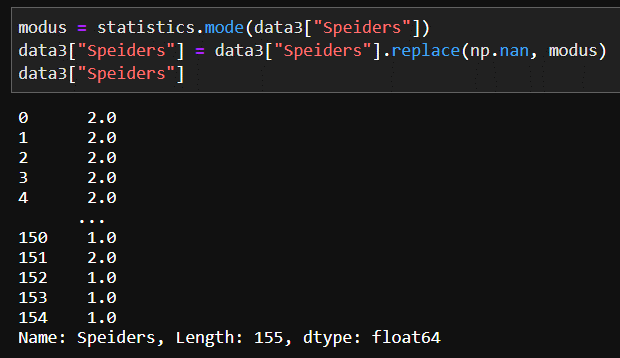
1. Penanganan Missing Value pada Columns Data Liver Firm menggunakan modus



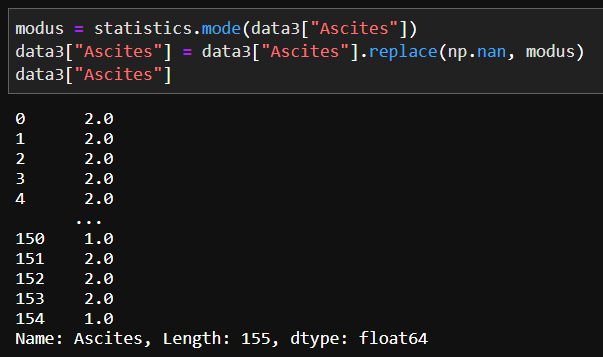
1. Penanganan Missing Value pada Columns Data Spleen Palpable menggunakan modus



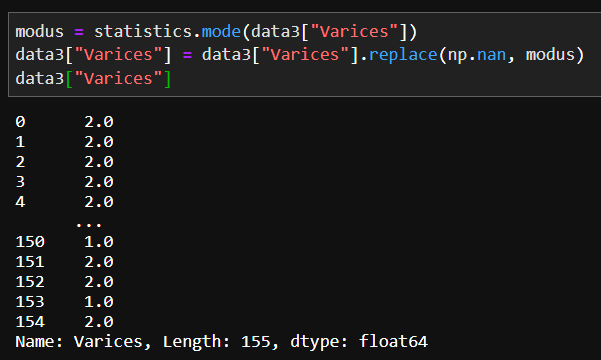
1. Penanganan Missing Value pada Columns Data Speiders menggunakan modus



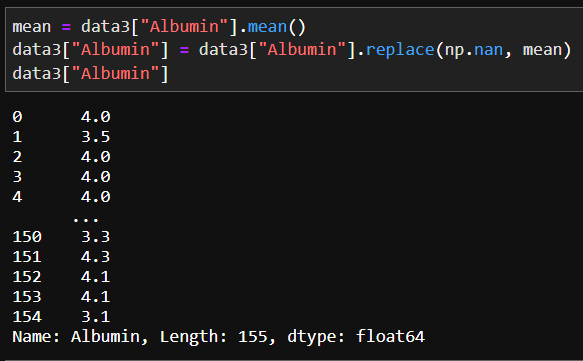
1. Penanganan Missing Value pada Columns Data Ascites menggunakan modus



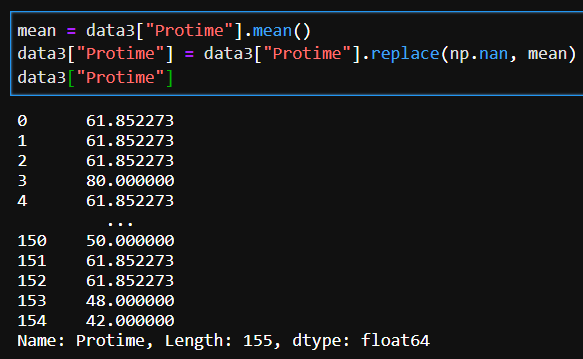
1. Penanganan Missing Value pada Columns Data Varises menggunakan modus



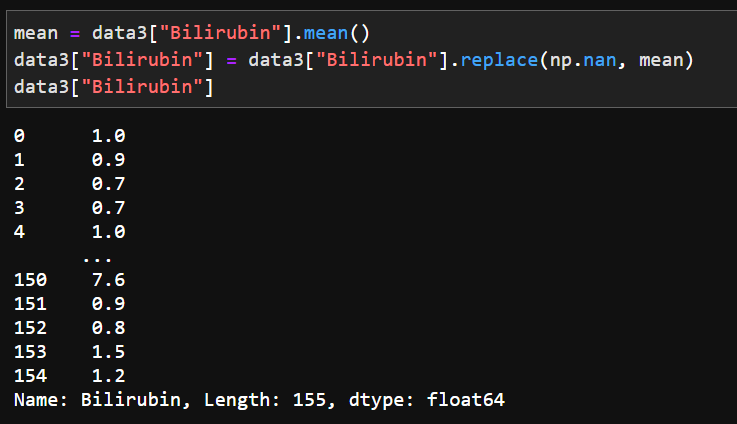
1. Penanganan Missing Value dengan Mean
2. Penanganan Missing Value pada Columns Data Albumin menggunakan mean



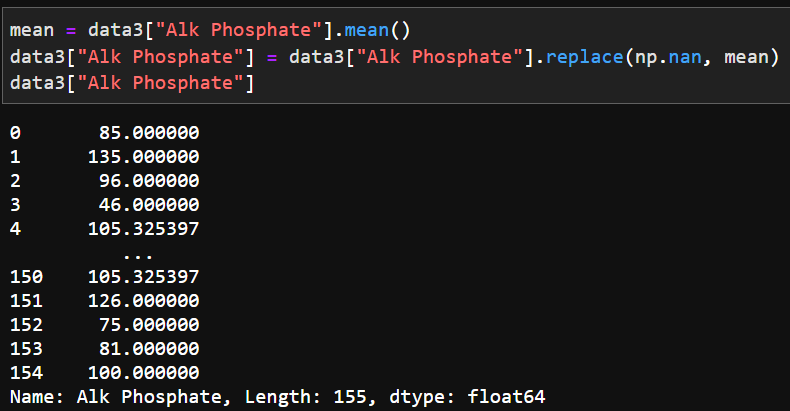
1. Penanganan Missing Value pada Columns Data Protime menggunakan mean



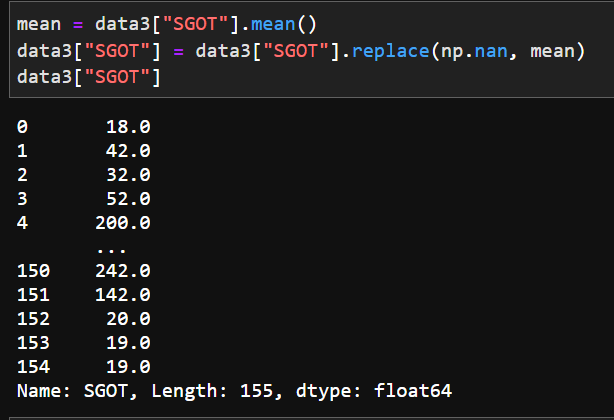
1. Penanganan Missing Value pada Columns Data Bilirubin menggunakan mean



1. Penanganan Missing Value pada Columns Data Alk Phosphate menggunakan mean



1. Penanganan Missing Value pada Columns Data SGOT menggunakan mean



1. Apabila dirasa sudah dilakukan penanganan Missing Value pada semua Columns dapat dicek kembali dengan code :



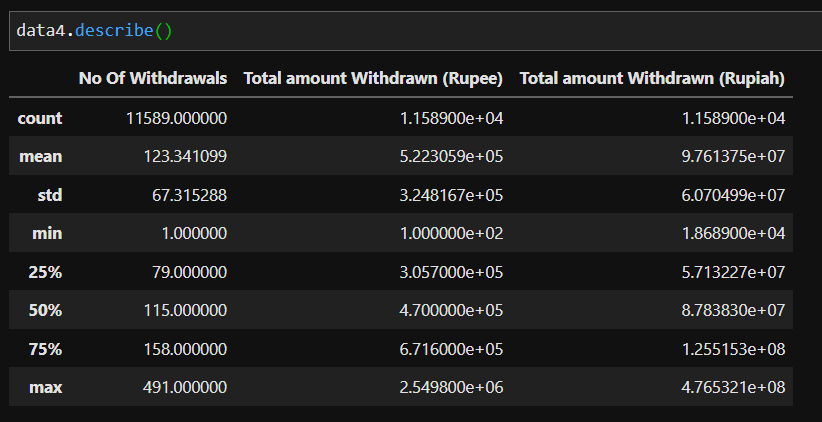
Dari beberapa proses penanganan Missing Value, akhirnya dari beberapa proses yang telah dilakukan, data sudah siap untuk dilanjutkan ke proses selanjutnya.

1. Deteksi Outlier: InterQuartile Range dan BoxPlot
2. Deskripsikan definisi masing-masing atribut pada dataset ATM Transcation berdasarkan pengertian di dalam domain keilmuan banking dan referensi yang dapat dipertanggungjawabkan

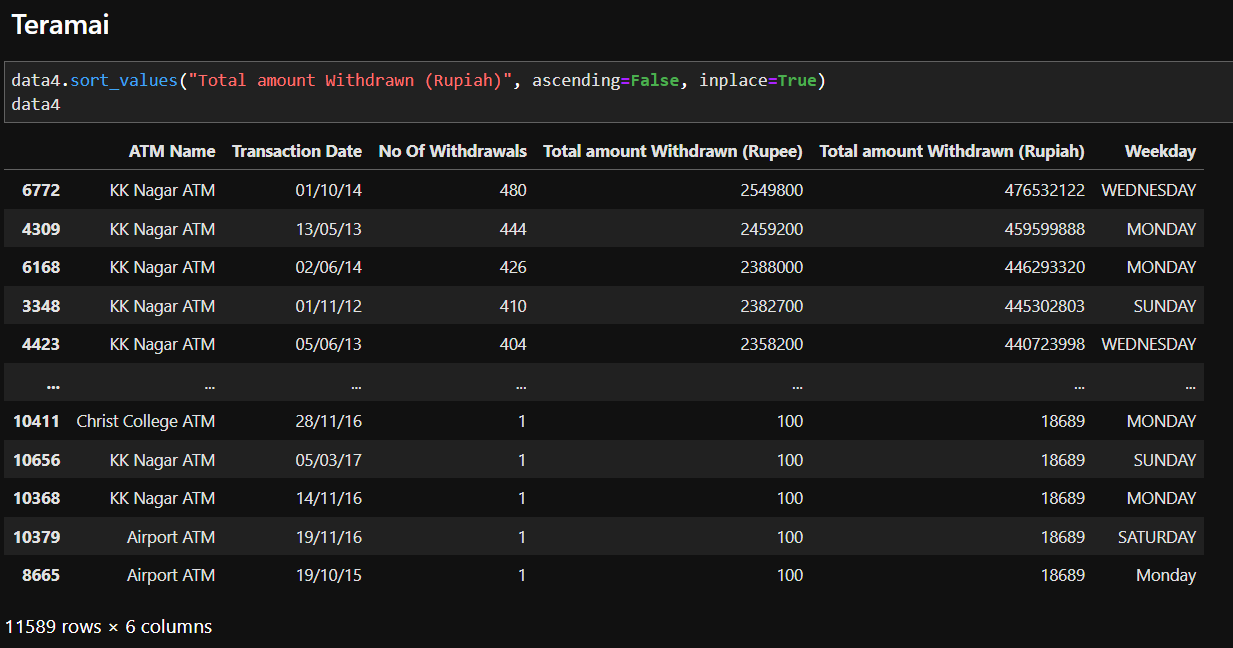
Sumber : …..

* ATM Name : adalah nama dari ATM yang dijadikan tempat transaksi
* Transaction Date : adalah data tanggal kapan transaksi dilakukan
* No Of Withdrawals : adalah jumlah penarikan uang dari transaksi yang dilakukan
* Total amount Withdrawn (Rupee) : adalah jumlah nominal penarikan uang dalam mata uang Rupee
* Total amount Withdrawn (Rupiah) : adalah jumlah nominal penarikan uang dalam mata uang Rupiah
* Weekday : adalah data hari kapan dilakukan transaksi penarikan uang

1. Lakukan analisis deskiriptif dan deteksi outlier menggunakan Python dan menjawab pertanyaan berikut:



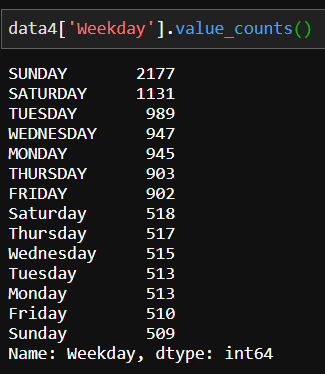
* Bank mana yang teramai dan tersepi bagi nasabah untuk melakukan penarikan uang?



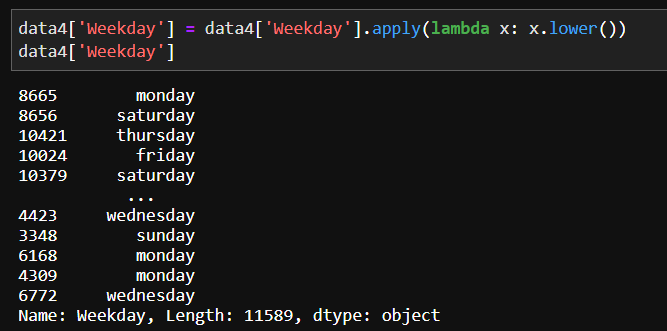
Bank yang teramai adalah : KK Nagar ATM dengan total amount withdrawn Rp476.532.122

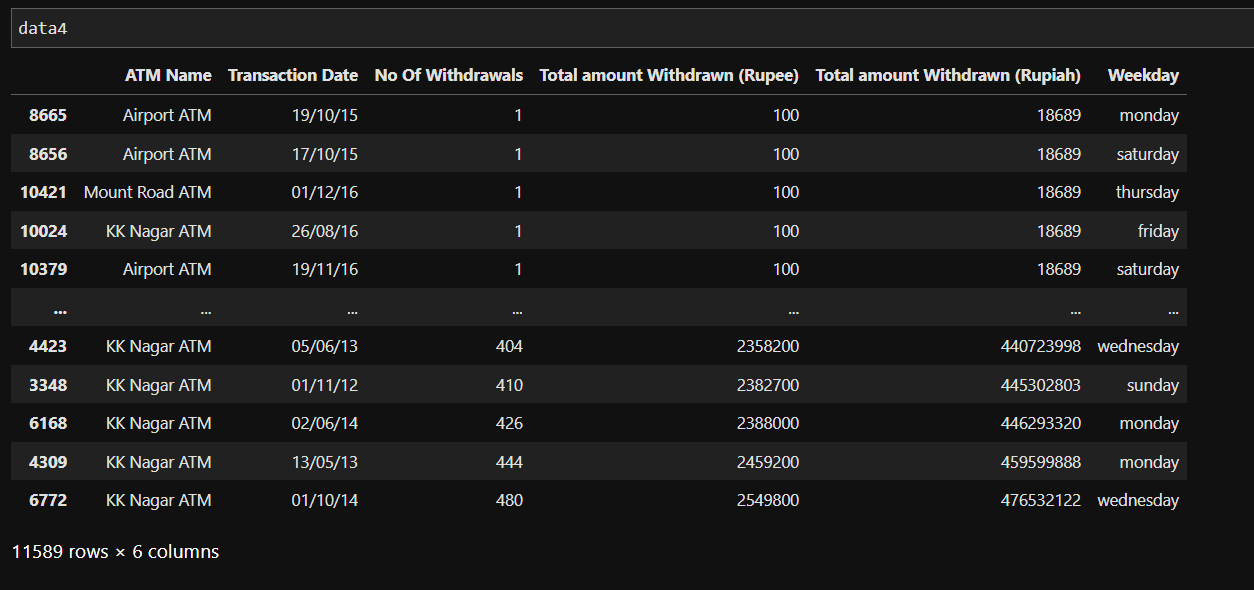
Bank yang tersepi adalah : Airport ATM dengan total amount withdrawn Rp18.689

* Pada hari apa ATM paling sering dan pada hari apa yang paling jarang dikunjungi oleh nasabah untuk melakukan penarikan uang?

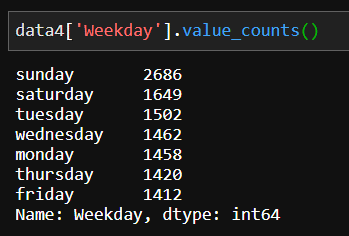


Karena terdapat nama hari yang sama karena perbedaan ada yang Uppercase dan Lowercase, perlu dilakukan preprocess data



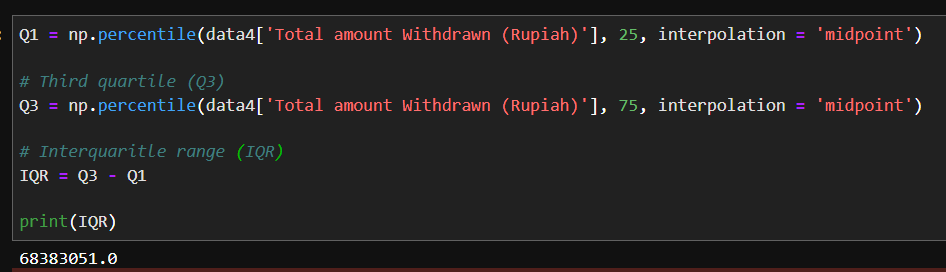


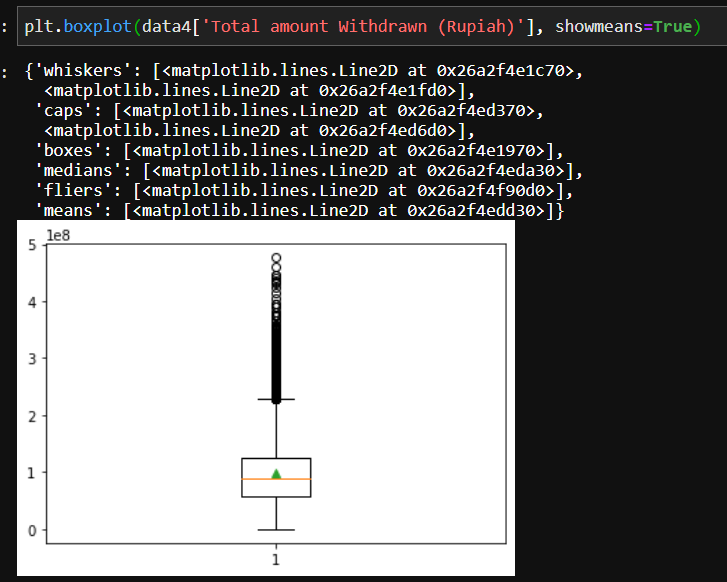
Setelah itu mari lihat ulang, hari apa yang termasuk hari paling ramai dan hari paling jarang dikunjungi oleh nasabah untuk melakukan penarikan uang



Dari hasil code tersebut, hari paling ramai adalah hari ‘Sunday’ atau hari **Minggu,** dan hari paling jarang adalah hari ‘Friday’ yaitu hari jum’at

* Lakukan deteksi *outlier* untuk melihat potensi adanya *fraud* berdasarkan atribut **Total amount Withdrawn (in Rupiah)** pada dataset ATM Transaction menggunakan metode InterQuartileRange dan visualisasikan menggunakan BoxPlot serta sebutkan nama **No Of Withdrawals** dan **nama ATM-nya** yang terdeteksi *fraud.*





Dari hasil visualisasi boxplot ini dapat disimpulkan banyak data outlier.