**THE DATA SCIENCE PROCESS**

1.Поставување на целта на истражувањето

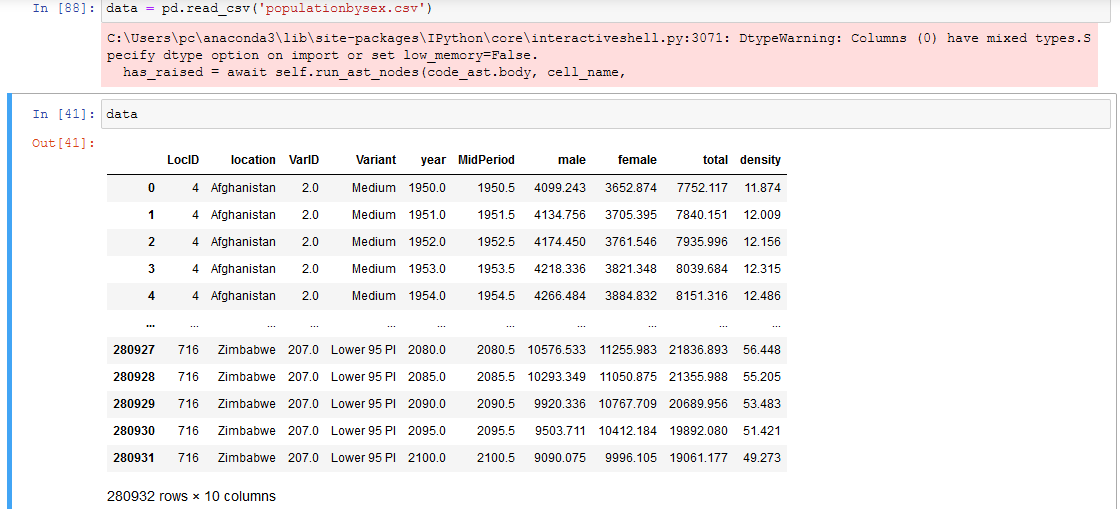
-Целта на нашиот проект е да направиме споредба на популацијата низ државите во светот, но оваа споредба да ја направиме според поделба на машкиот и женскиот пол. Исто така да го согледаме и порастот на вкупната популација во државите низ светот. Тоа истражување е направено во неколку години, да согледаме дали бројот на популацијата е зголемен или пак намален за дадена држава.

2.Превземање на податоци

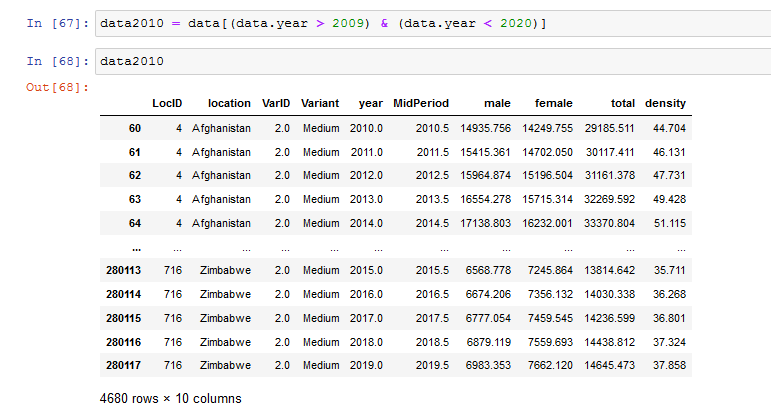
-Во овој проект користиме готово податочно множество кое го превземавме од следниот линк: <https://population.un.org/wpp/Download/Standard/CSV/>

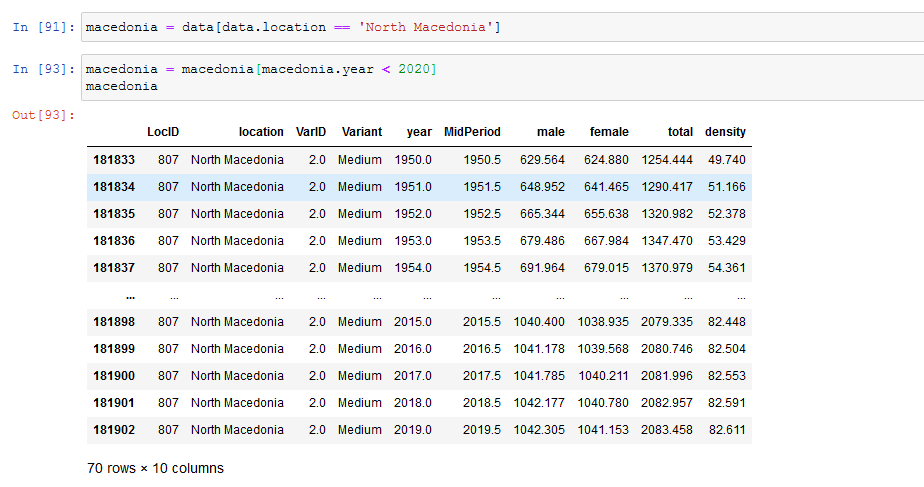
Ја користевме библиотеката pandas која ни овозможува работа со податочни множества и вршење операции над податочни множества.

За да ги добиеме информациите од даденото податочно множество го искорстивме следниот код и резултатот од истиот:



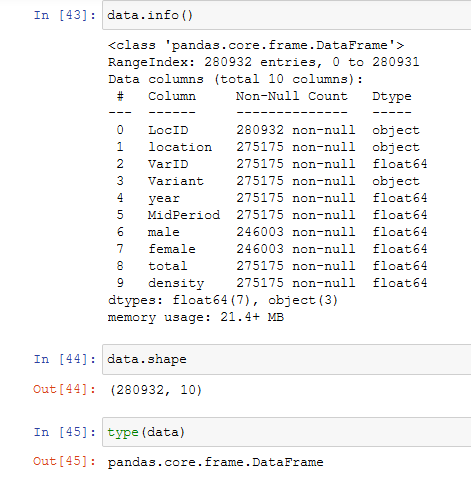
-Но поради фактот дека податочното множество кое го користиме е огромно,одлучивме да правиме анализа на податоците од 2010 до 2020 година. Податоците ги зимаме на тој начин што ја селектираме годината од податоците и ги ставаме во нова податочна рамка.

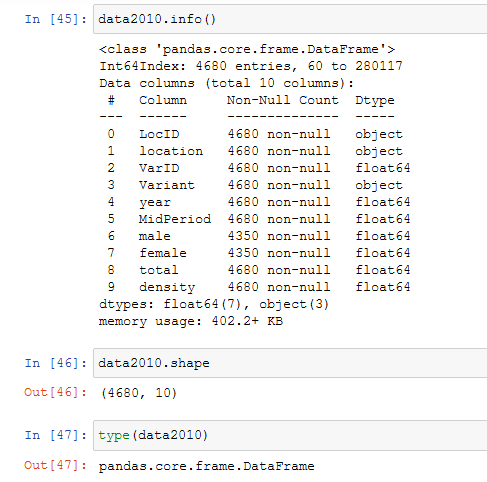
Тоа е прикажано на следната слика:   


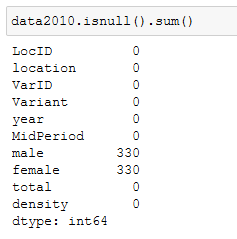
-Следно нешто е разгледување и истражување на податоците само за нашата држава:  


3.Чистење, интегрирање, и трансформирање податоци

Пред да почнеме да правиме било каква анализа, најпрво податоците треба да се средат.Чистење на податоците се врши на тој начин што проверуваме дали има празни вредности, дали има null вредности, дупликати, од каков тип се податоците.Доколку има и пренатрупаност во податоците, односно неразбирливост или нечиткост, потребно е истите да се средат така што ќе се добие еден добар, разбирлив преглед за истите.

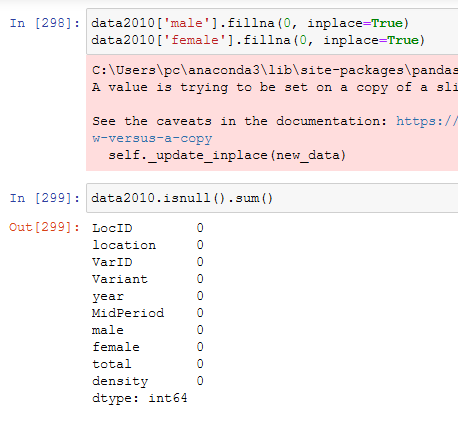
Од сликата десно се гледа дека во податочното множество има вкупно 280932 редици, а според информациите со методот info() гледаме дека има прилично голем бој на null вредности.Само во полето LocID нема null вредности.Вкупно имаме 10 колони.

Од друга страна, во податочното множество кое ние ќе го користиме,исто така гледаме дека има null вредности во колоните male i female.



Овие null вредности треба да се средат така што ќе се вметне некоја вредност за нив, или соодветно да се отстранат.Бидејќи гледаме дека вредностите се кај male i female, треба да внесеме некој број.

Пополнување на полињата со null вредности се прави со функцијата fillna() каде како параметар се става вредноста која сакаме да ја вметнеме.



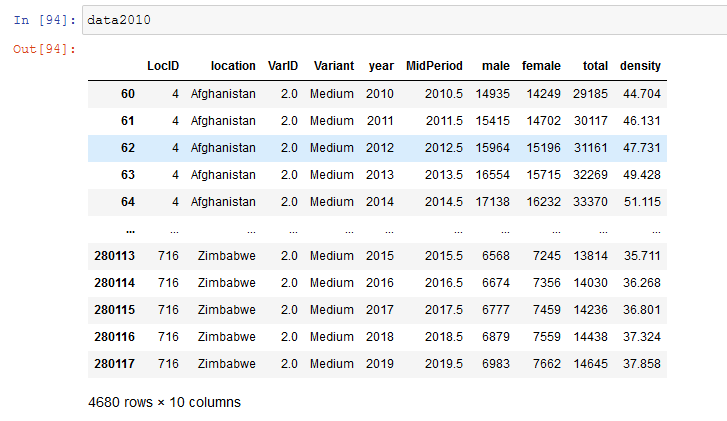
Сега гледаме дека nema null вредности во полињата.

Што се однесува до outliers, сите вредности се прилично блиску и нема некој скок на вредносите,така да сите се во еден сличен ранг.

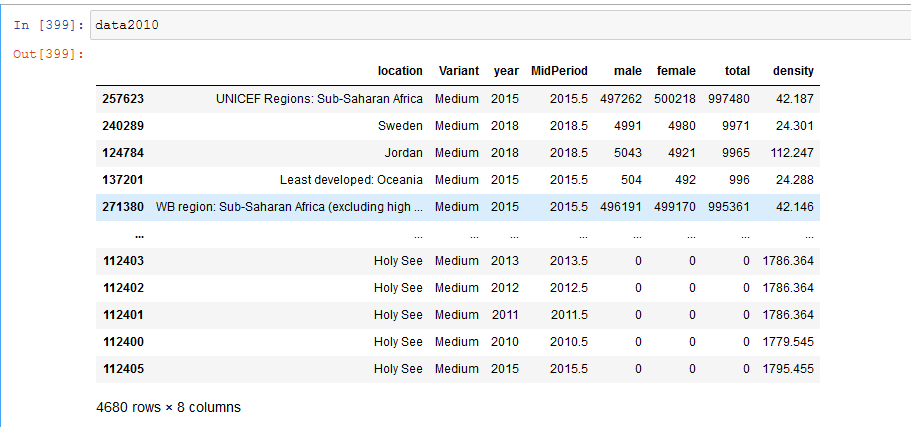
Податоците се прилично јасни, нема двосмисленост на информациитe, сметаме само дека float вредностите треба се заокружат на цели броеви.Тоа го направивме со apply и ламбда кадешто ги поделивме вредностите пред и после децималот кадешто го исфрливме децималот и добивме цел број.



После среденото, добивме подобар изглед на податоците.



Податоците ги сортиравме според големината на популацијата.



Гледаме дека во последните редови,односно регионот со најмалку жители има полиња со 0 елементи.Соодветно, треба да ги отстраниме тие редици.Со примена на одреден метод, ја средивме табелата.



