**1. Leads Analysis: You’re provided a csv file containing student information. Following**

**dataset consists of their personal information along with the colleges/stream they’ve**

**enquired(showed interest) about on our Collegedunia platform.**

**● Perform in-depth analysis on the given data and conclude trends that are**

**present.**

**● How would you treat null values that are present in the data?**

**● Can you build a predictive model that can predict the chances of a student taking**

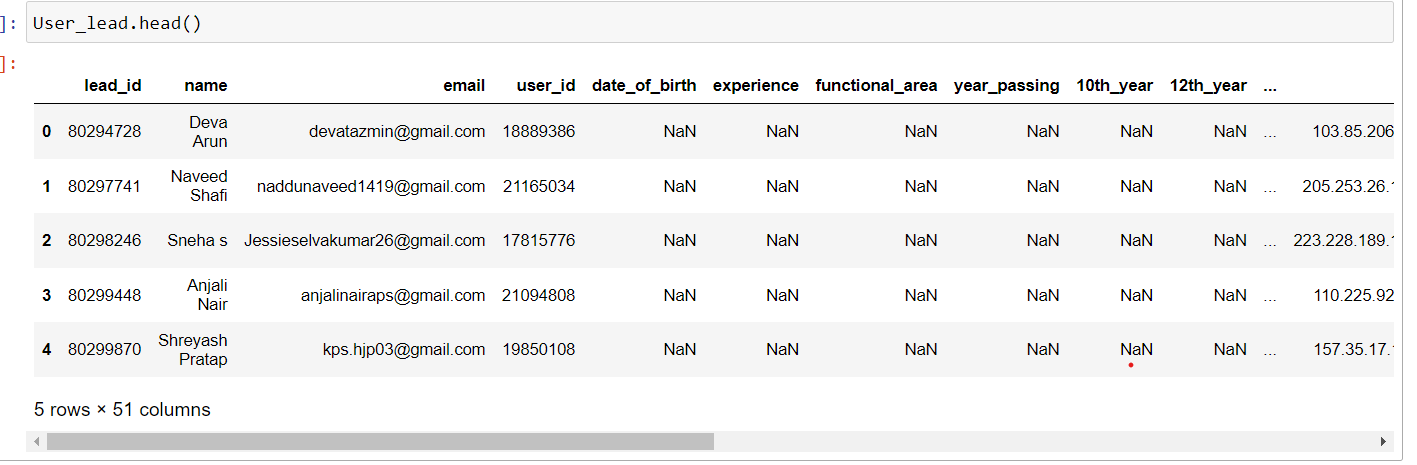
**admission in the college they’ve enquired about? If yes then how?**

**● What are the KPIs that you think should be considered to increase student**

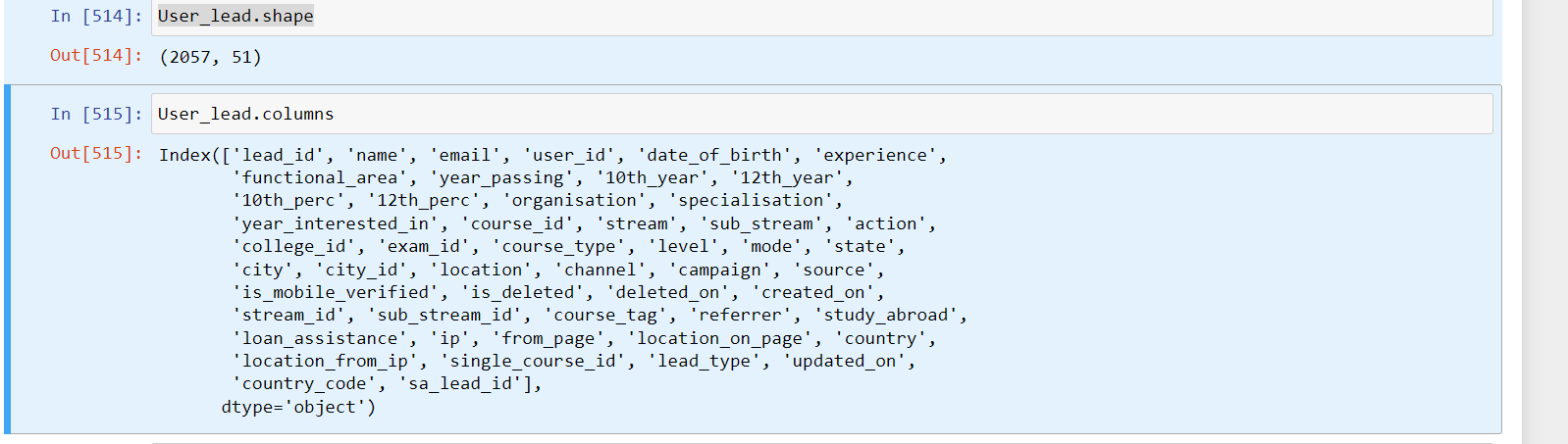
**acquisition?**

**● Visualize the data distribution for each variable.**

**Reading data from excel file:**

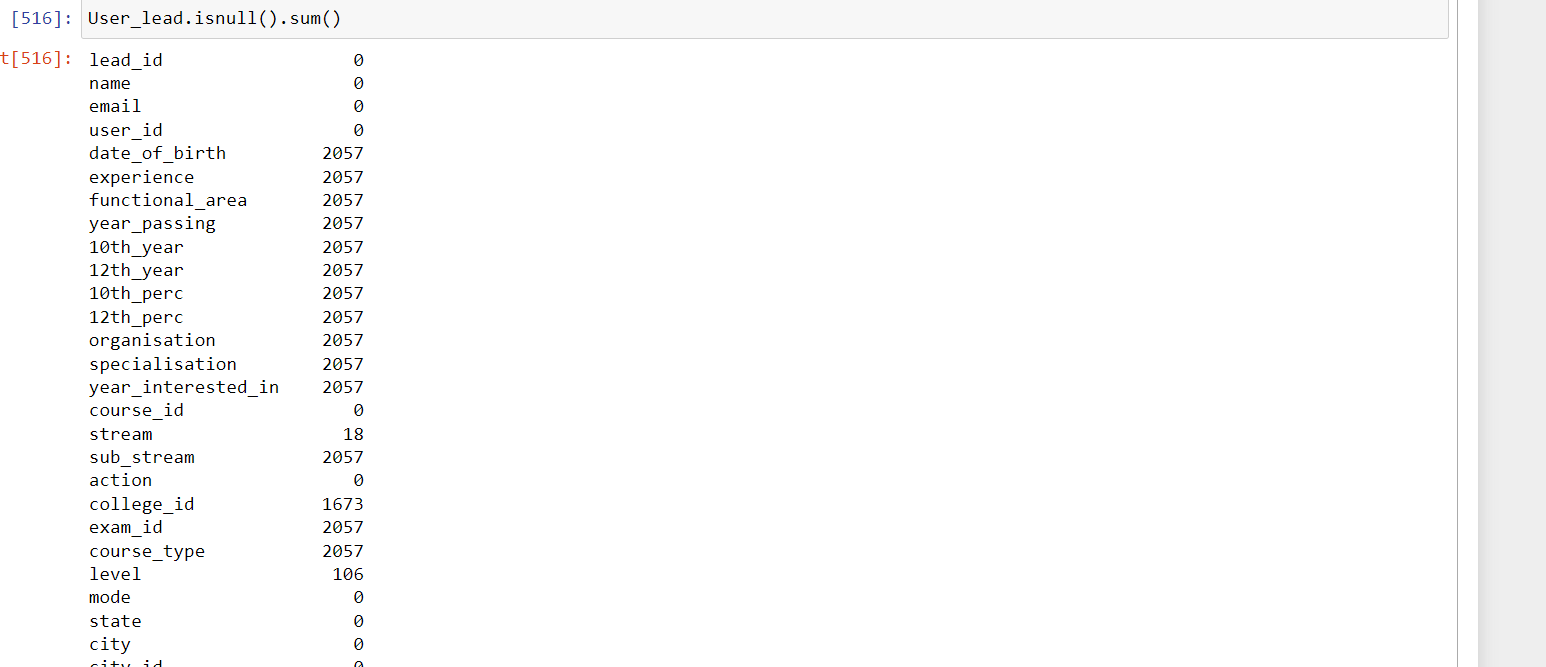


Below is the shape of data:



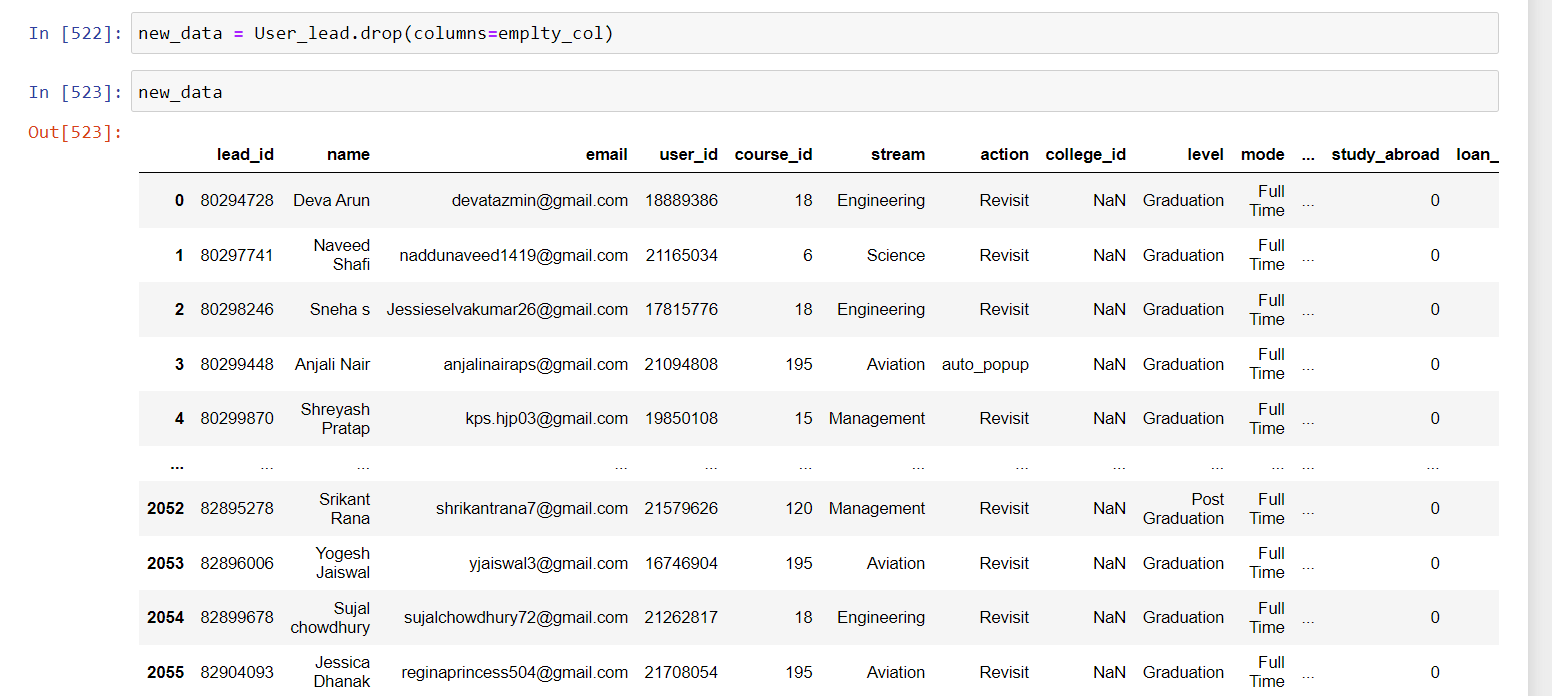
After initial analysis it has been observed that many feature columns are completely empty:

for example location Channel, campaign.

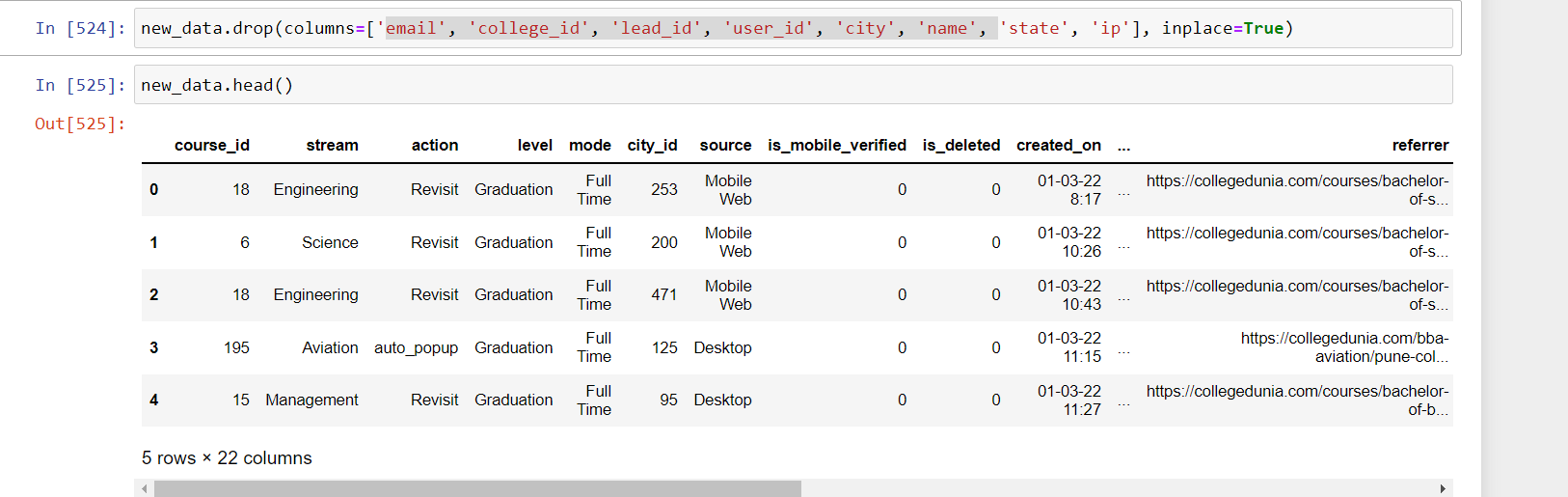


So I have decide that remove those column which has more than 1900 (out of 2057 ) NaN value.



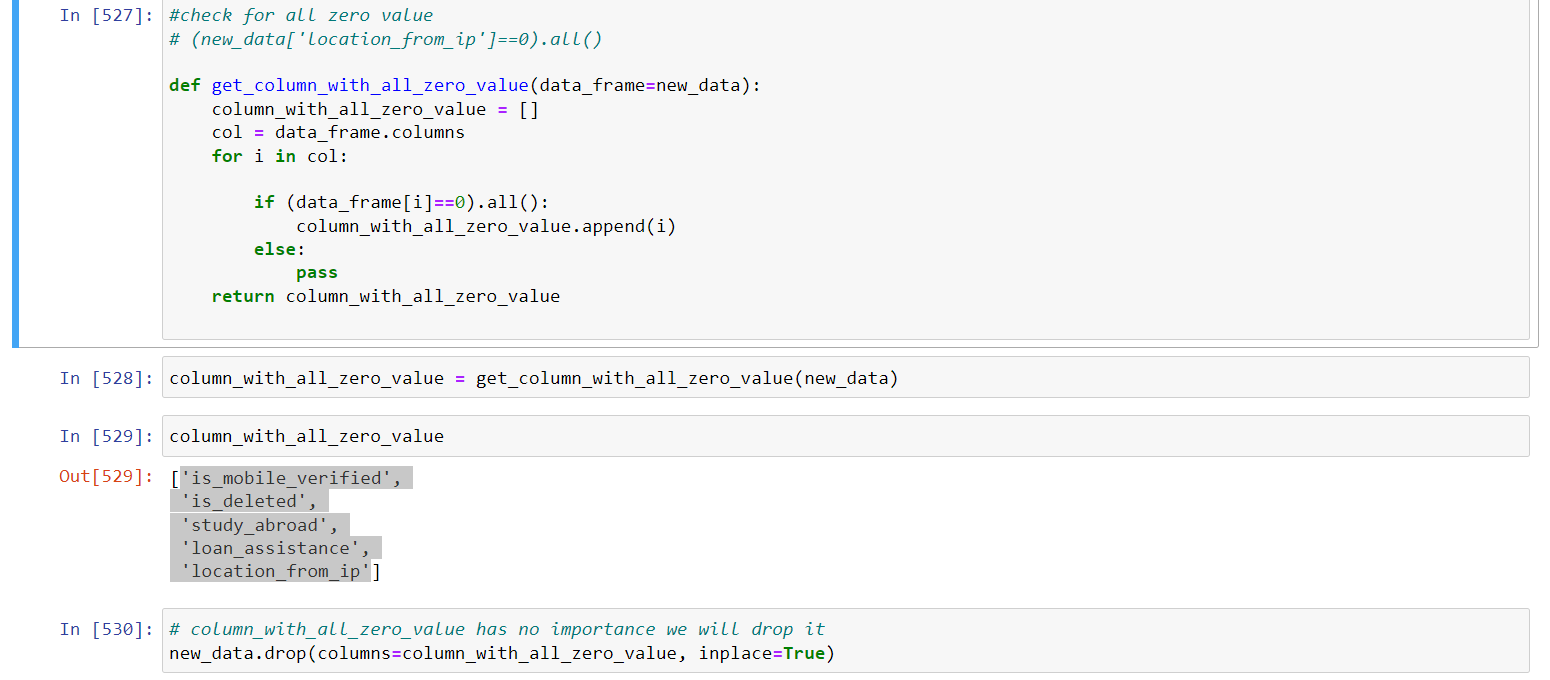


Some columns like email', 'college\_id', 'lead\_id', 'user\_id', 'city', 'name', has no significance so I removed them.

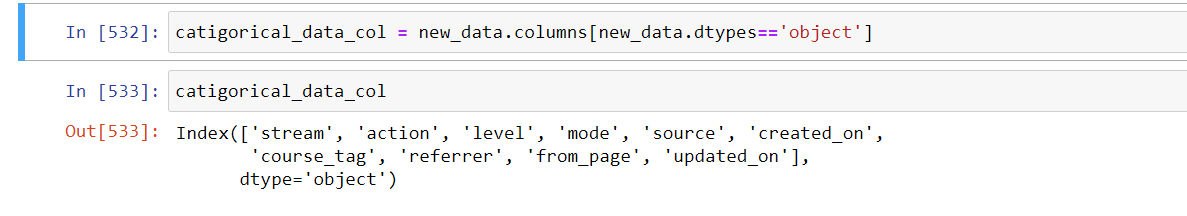


Now I have observed that many columns have only Zero value, which have no significance. I have decided to remove it.

For example 'is\_mobile\_verified','is\_deleted', 'study\_abroad', 'loan\_assistance', 'location\_from\_ip' contain only Zero.



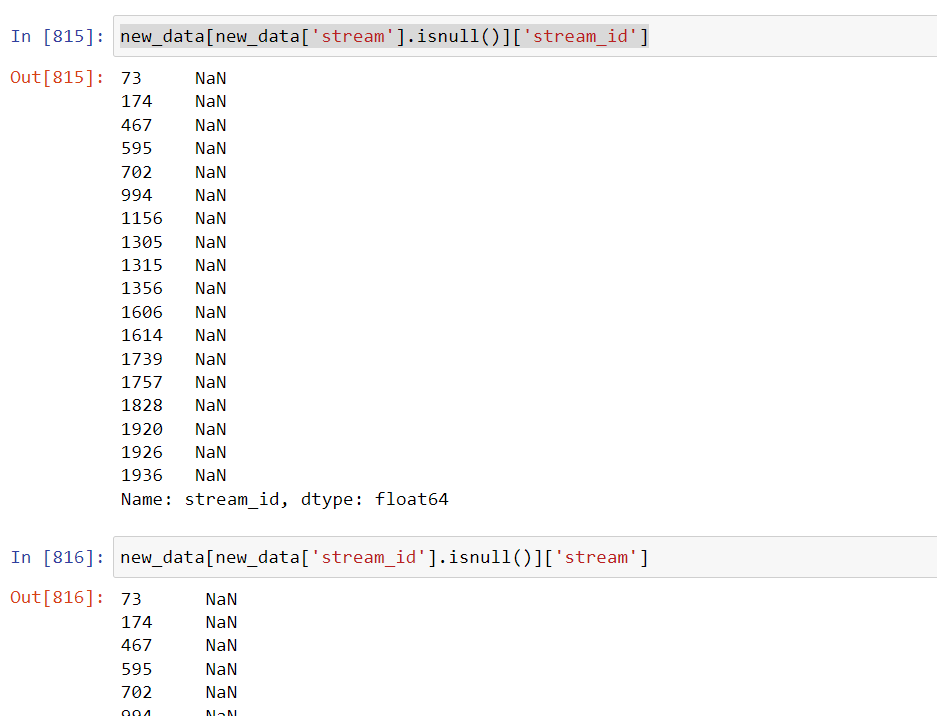
There are many categorical variables presents in dataset ex: 'stream', 'action', 'level', 'mode', 'source'



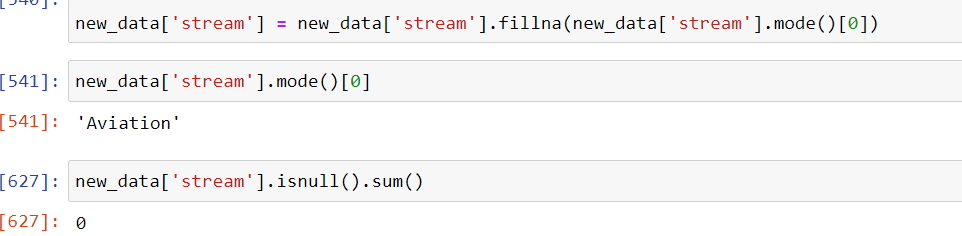
Now we will fill Null Value:

we have observed that “stream” and “stream\_id” column are corelated , and both have 18 missing value in same index

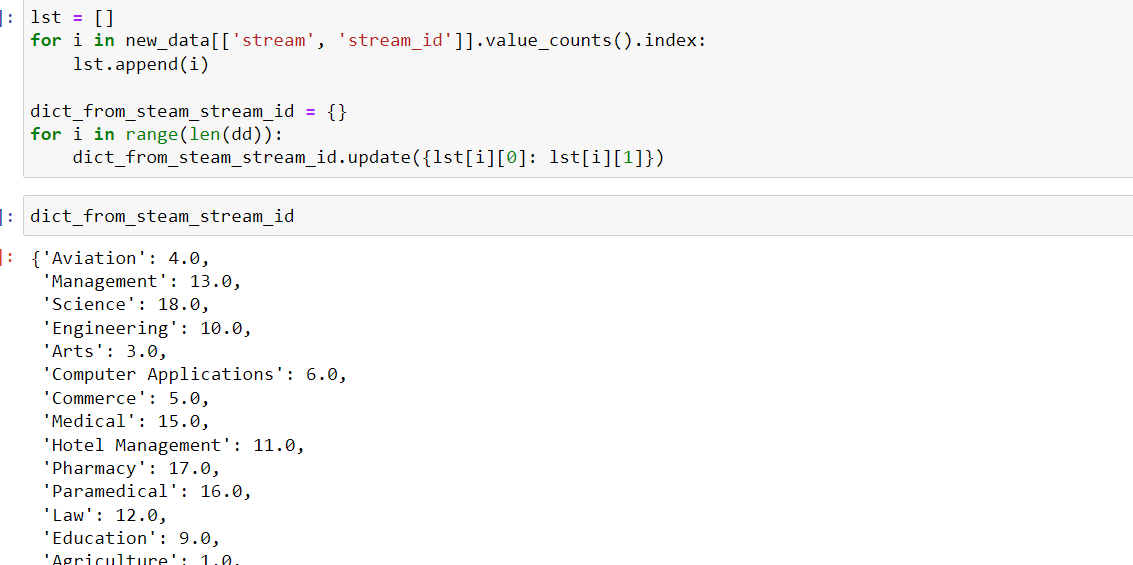
we decided that we fill null value in 'stream' column with its mode, as it is categorical data after that map these value with respective “stream\_id”



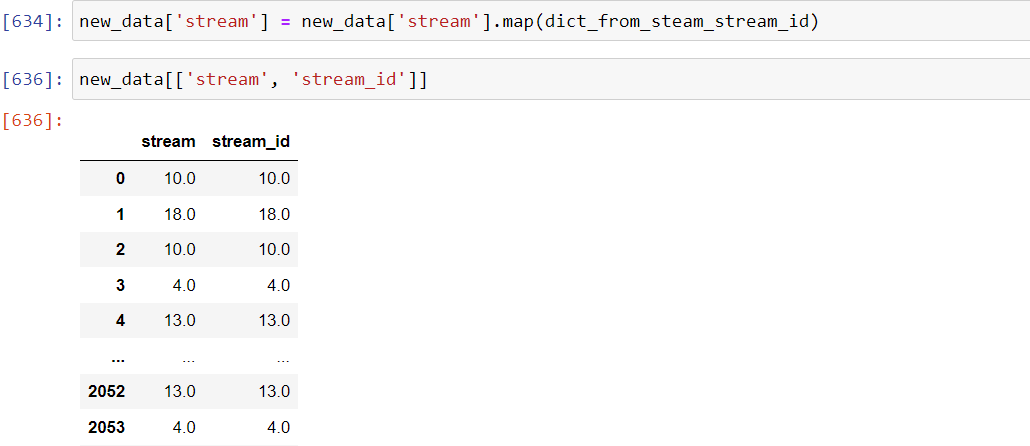
Filling stream column with mode:



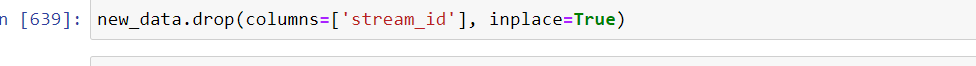
Now we are obtaining respective stream\_id for “stream” column from “stream\_id” column:



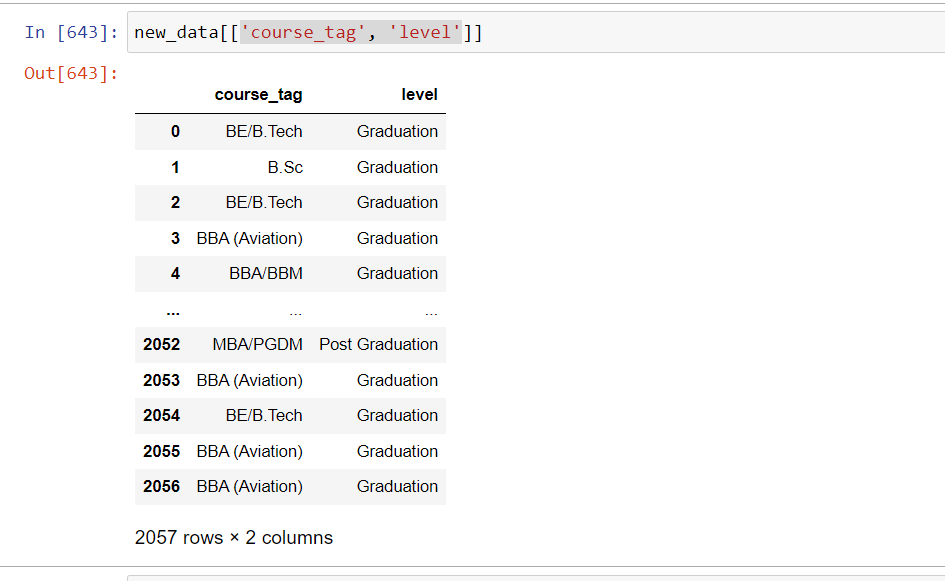
Mapping:



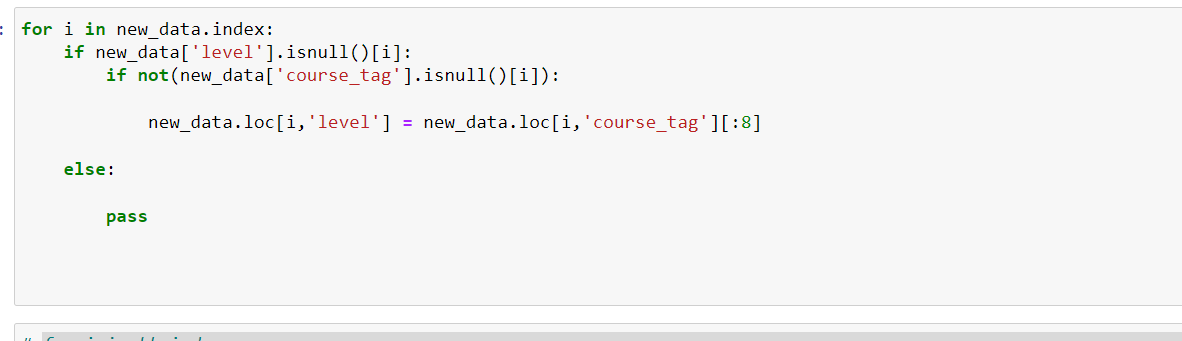
Now we will drop the stream\_id column.

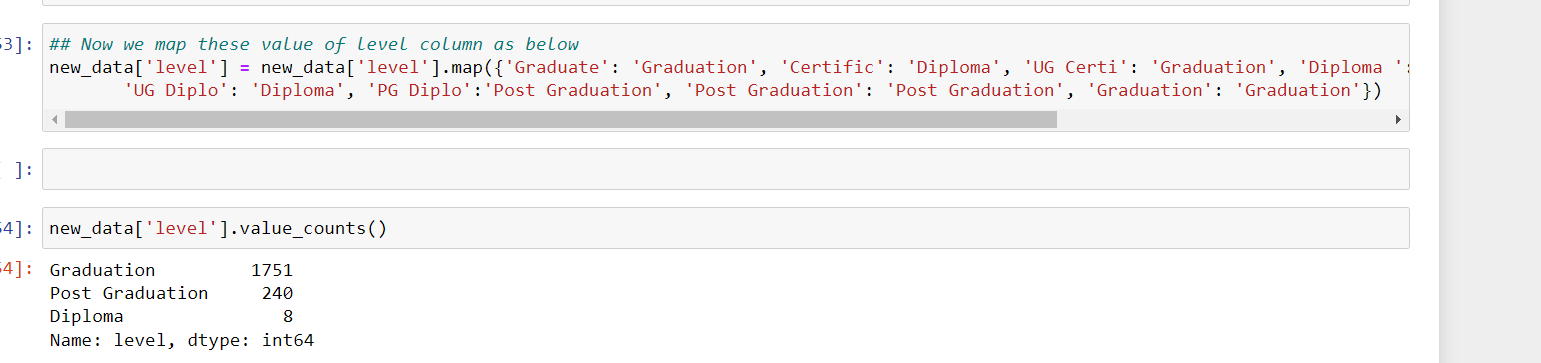


Now I have observed that 'course\_tag', 'level' column are correlated:

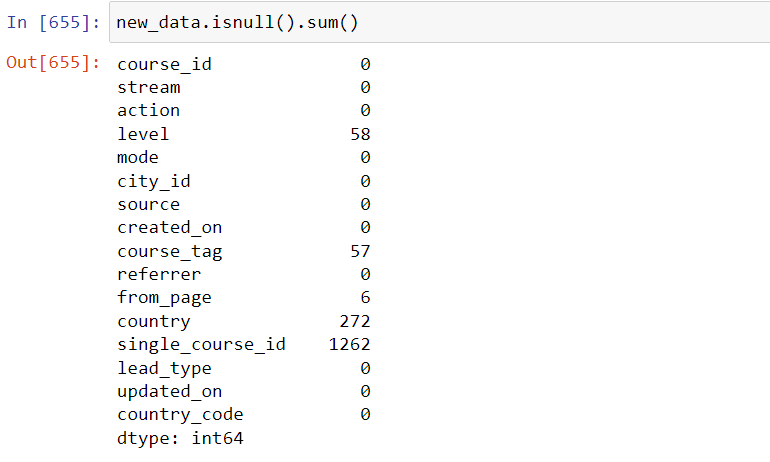


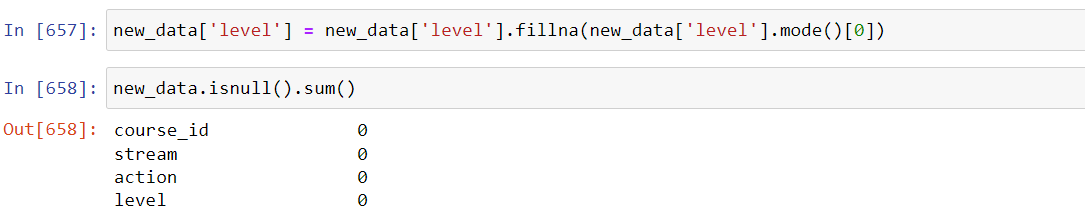
So we are filling level null value with help of “course\_tag” column:





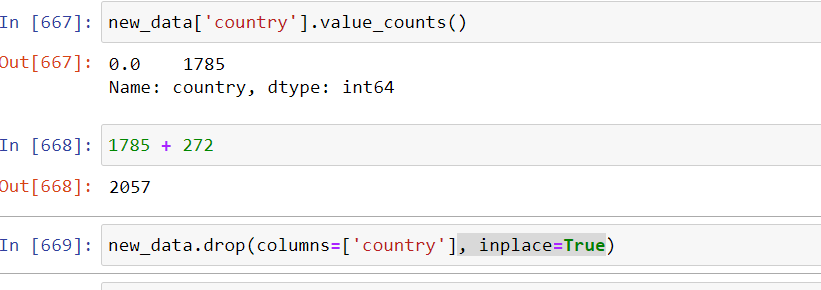
Now level column has only 58 missing value, these value filled by mode



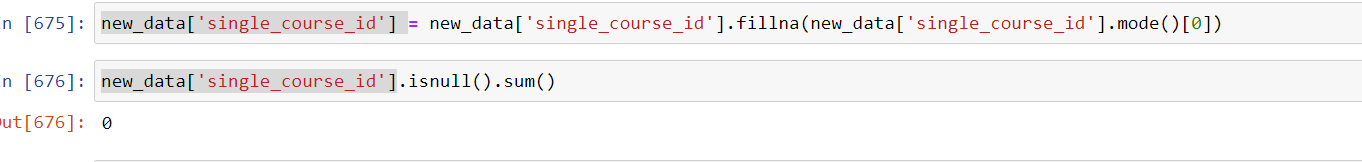


As “course\_tag” and “level” correlated so we are dropping the course\_tag column

Also country column have 1785 zeros and 272 NaN value so we can say there is no information provided by this column, we are dropping it.



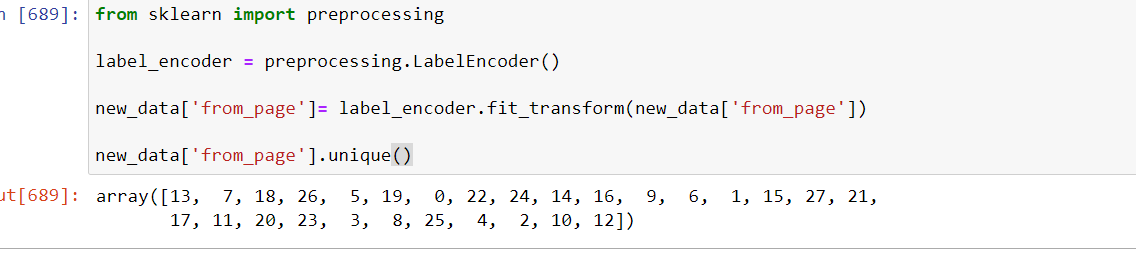
single\_course\_id is categorical feature so we are using mode to replace NaN:

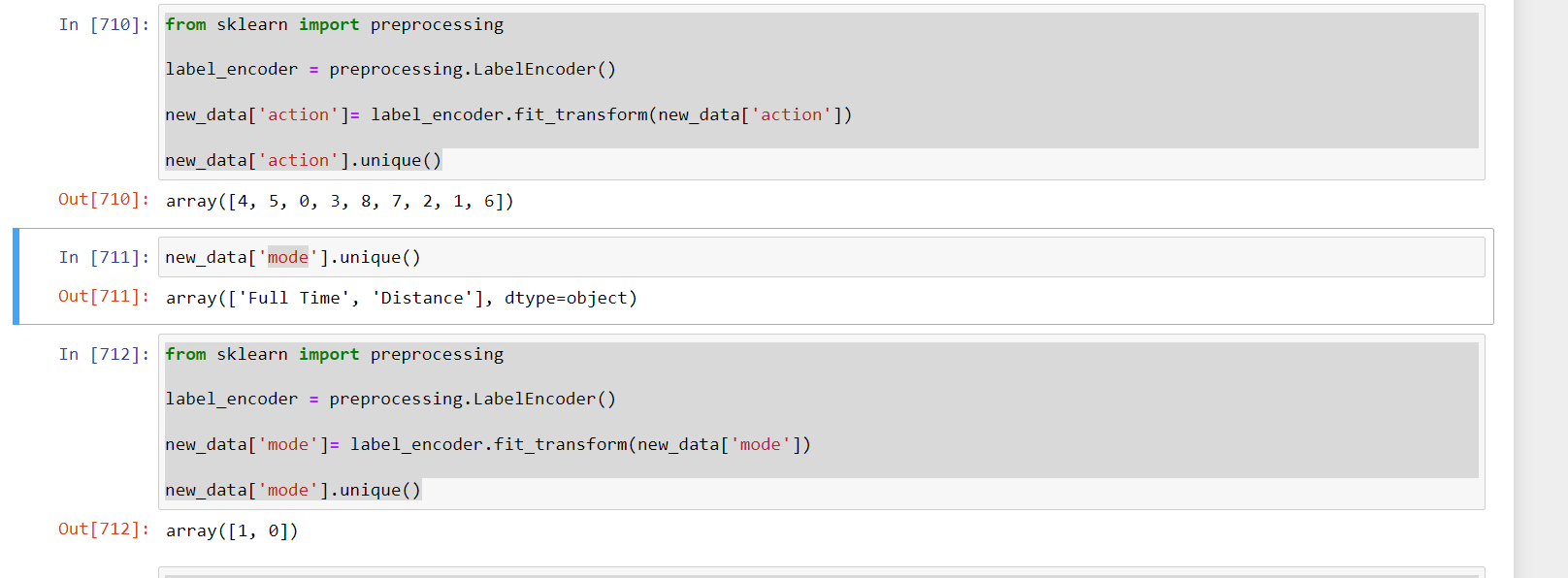


Similarly we are filling the NaN value of “from\_page” column:

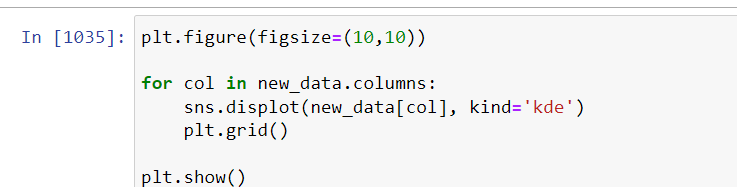


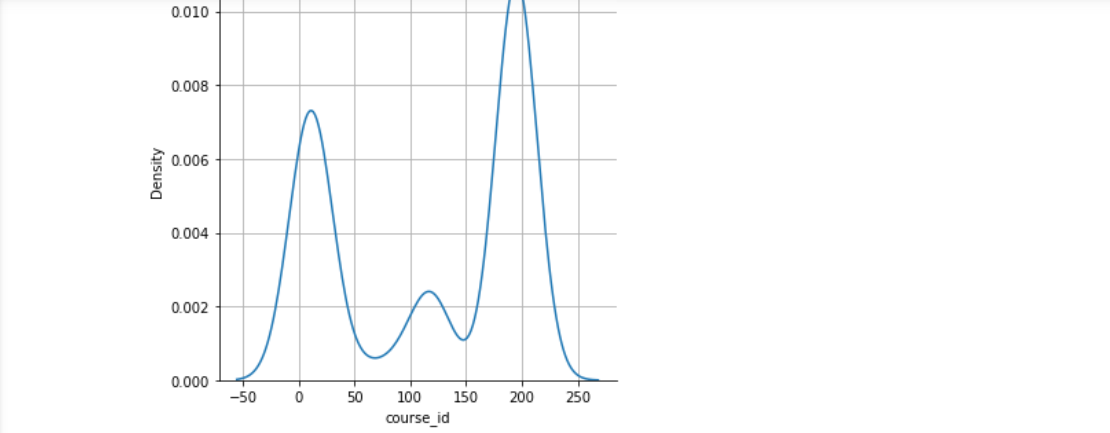
Now we are using sklearn preprocessing library to encode remaining categorical data :

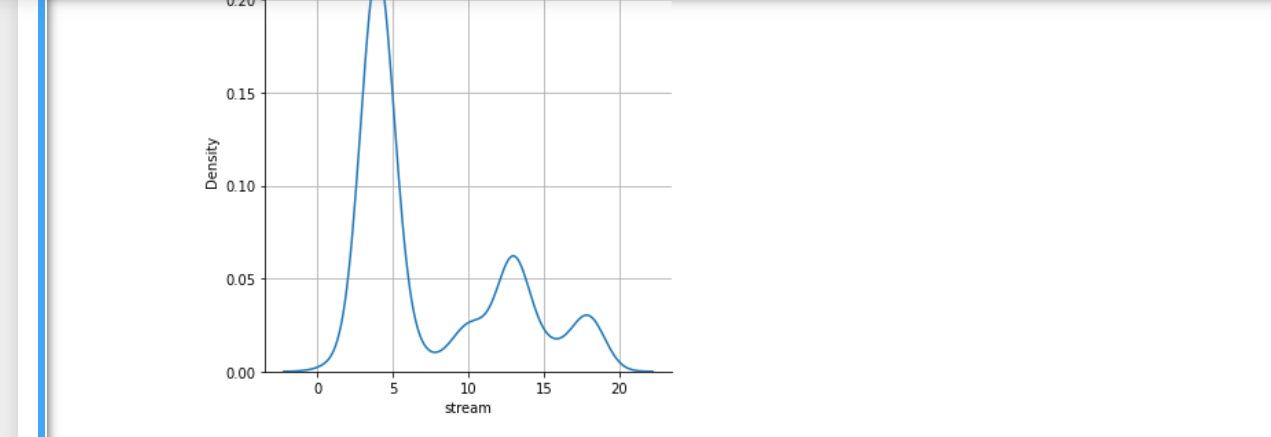


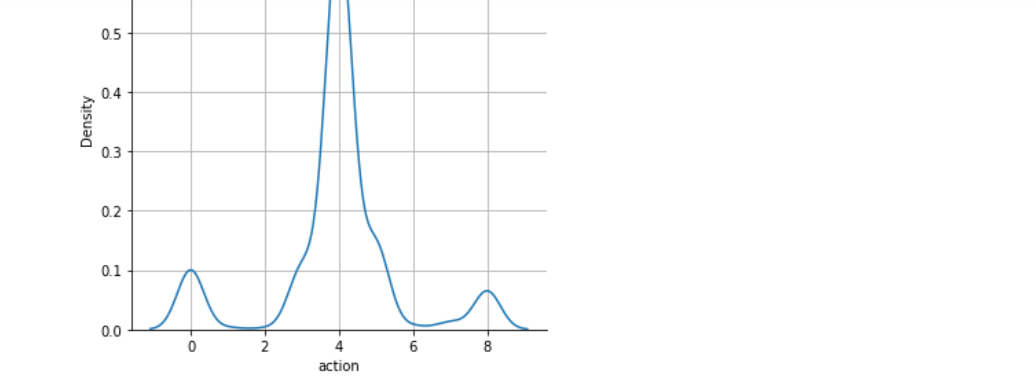


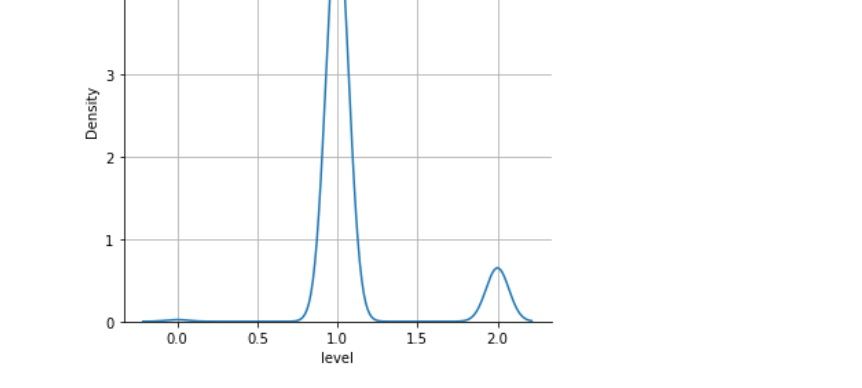
Distribution of Data:











Now our EDA is completed and I am assuming “lead\_type” is output variable.

I have made ML model using random forest to predict the “lead\_type” and it is giving more than 95% accuracy:

