**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?**

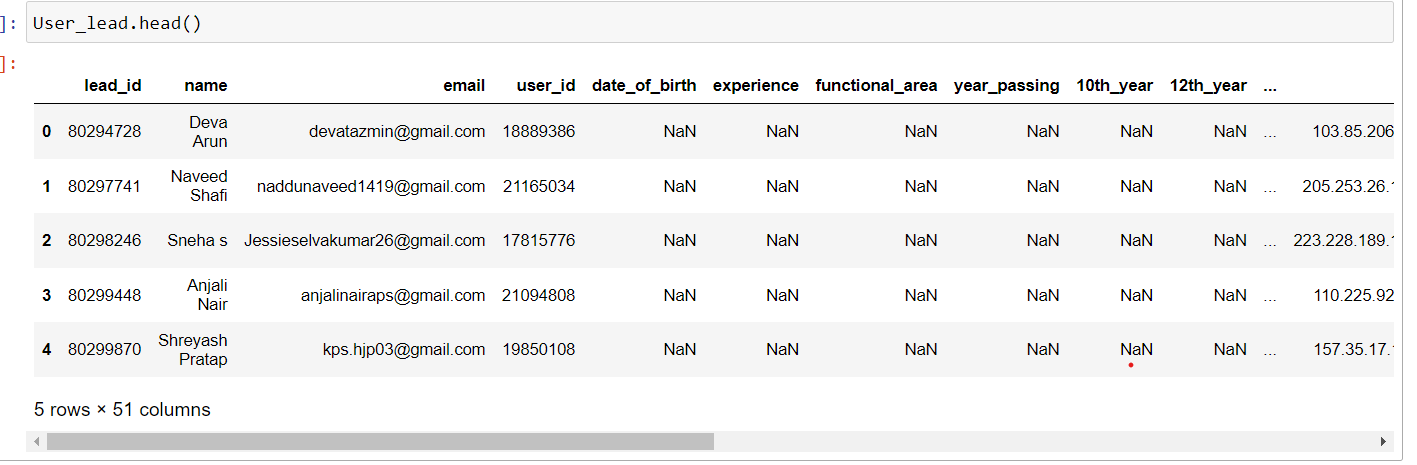
**● Visualize the data distribution for each variable.**

**Note: code is available at** [**https://github.com/neerajbafila/Collegedunia/blob/main/CaseStudy\_1.ipynb**](https://github.com/neerajbafila/Collegedunia/blob/main/CaseStudy_1.ipynb)

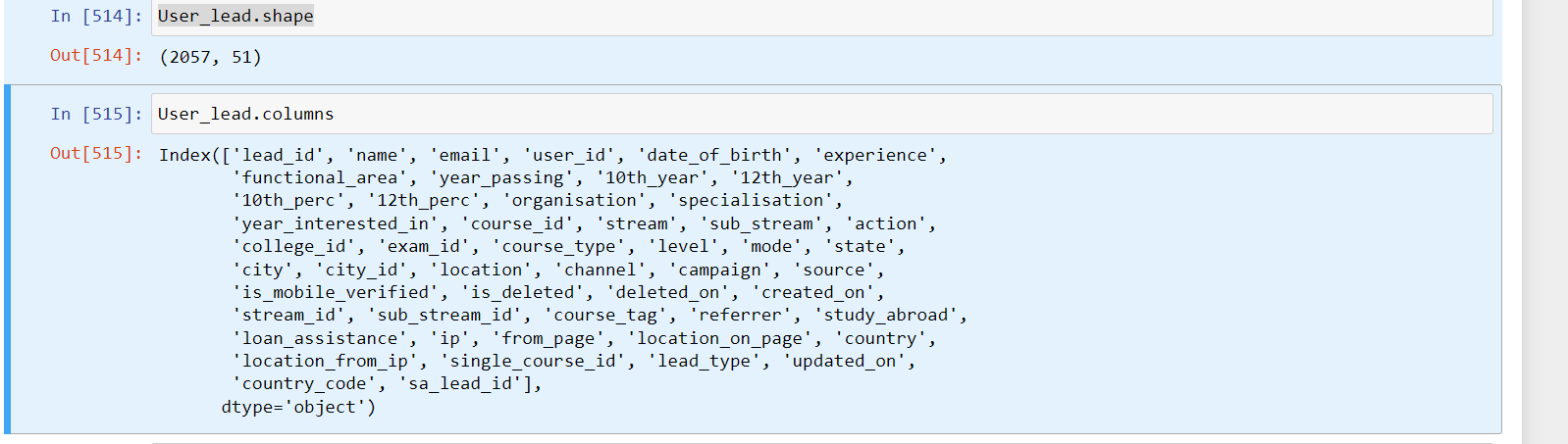
**SQL:**

https://github.com/neerajbafila/Collegedunia/blob/main/sql.ipynb

Reading data from excel file:

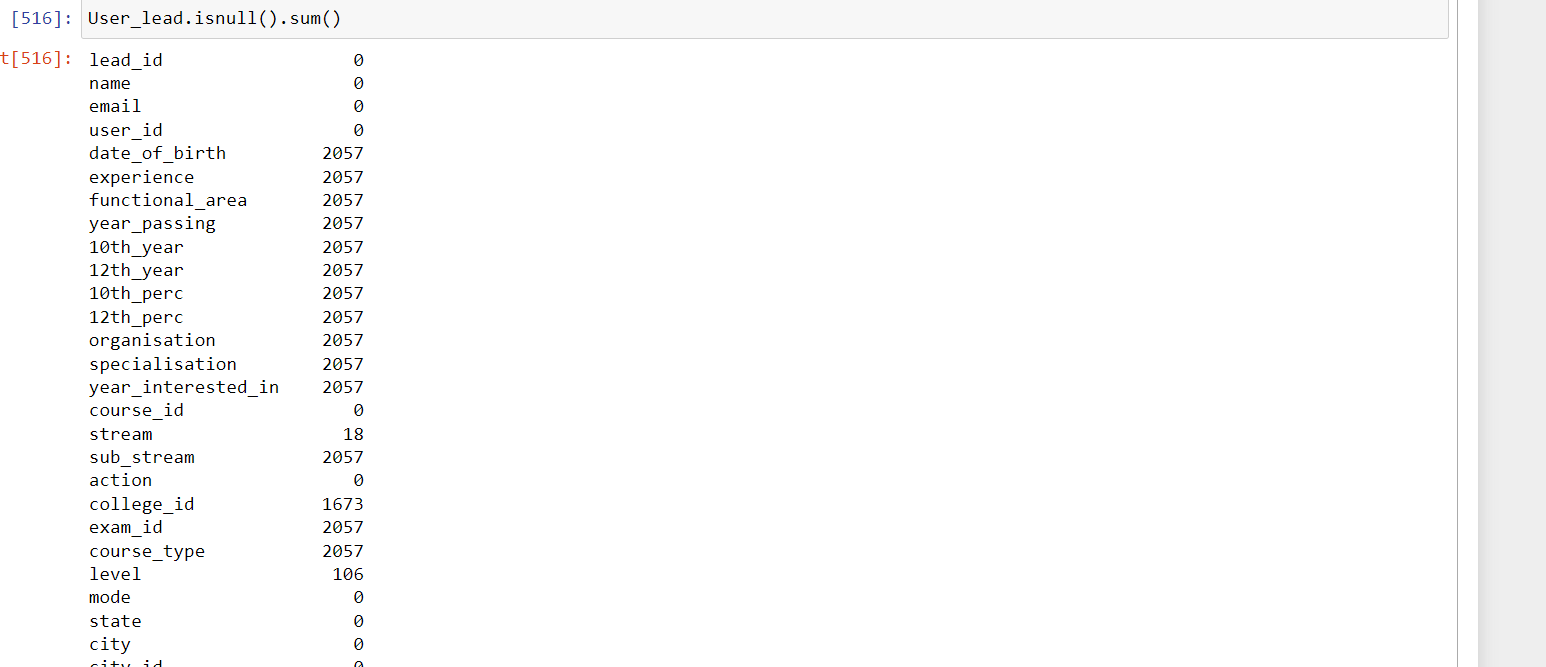


Below is the shape of data:



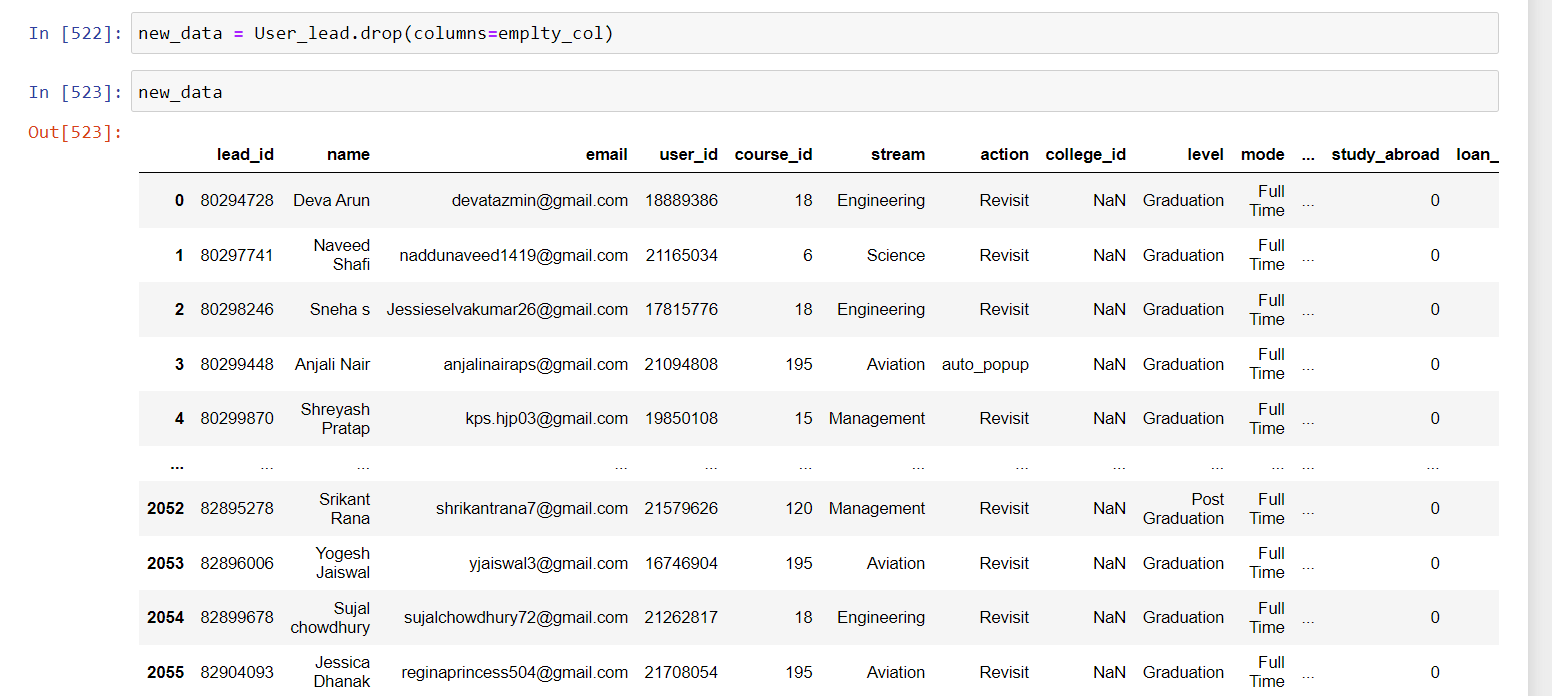
After initial analysis it has been observed that many feature columns are completely empty:

for example location Channel, campaign.

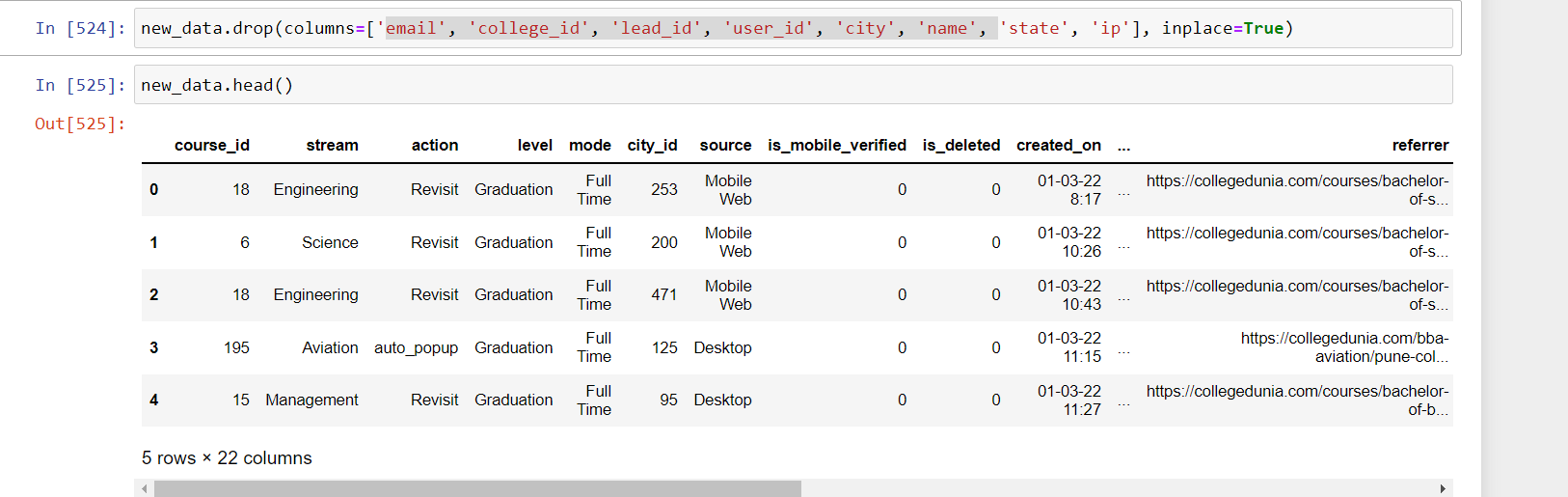


So I have decide that remove those column which has more than 1900 (out of 2057 ) NaN value.



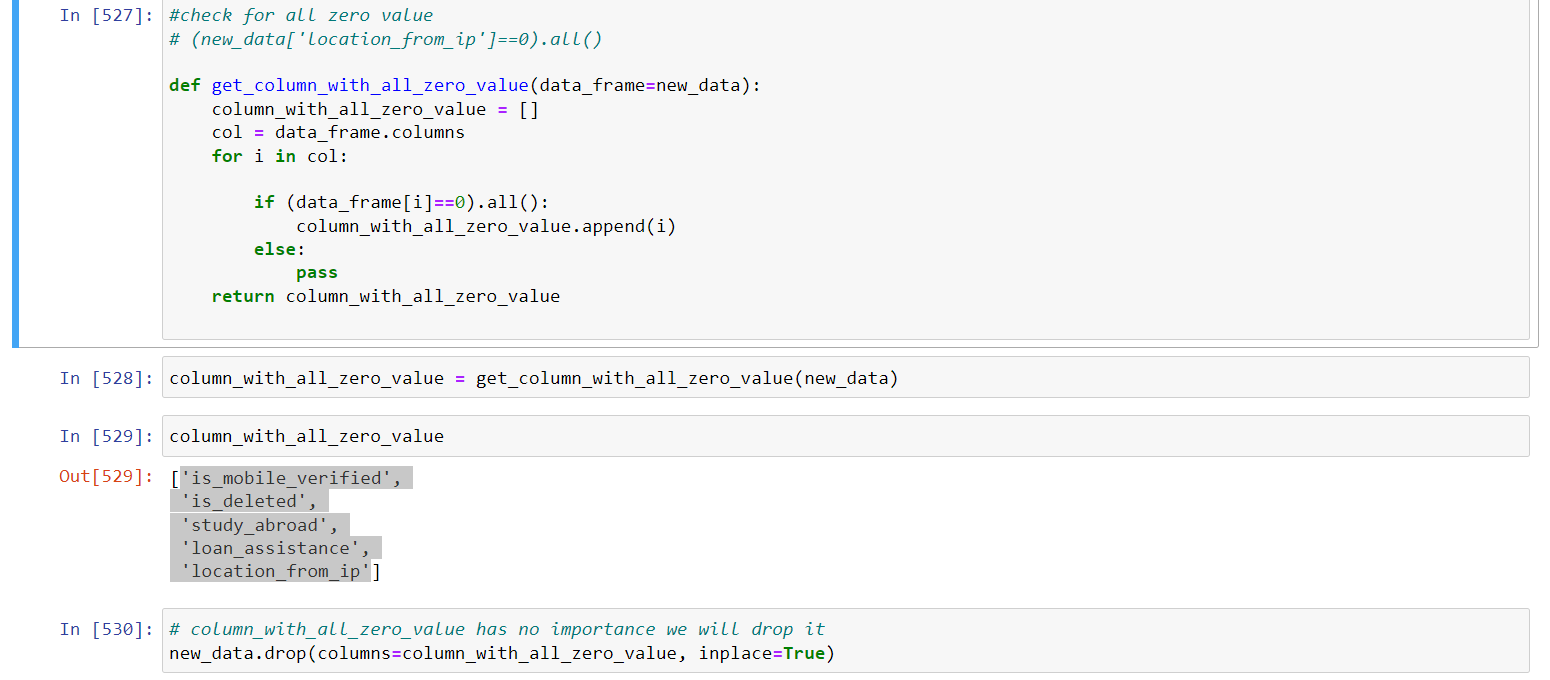


Some columns like email', 'college\_id', 'lead\_id', 'user\_id', 'city', 'name', has no significance so I removed them.

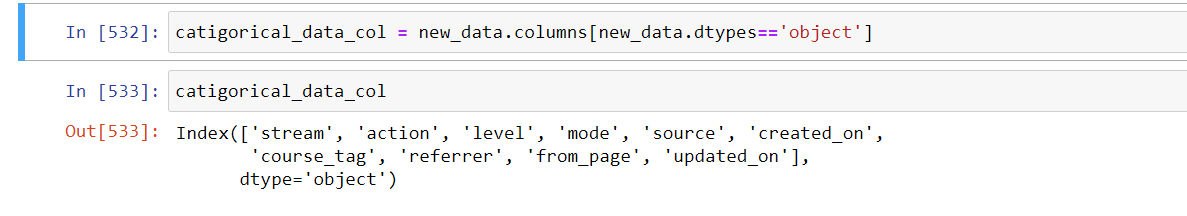


Now I have observed that many columns have only Zero value, which have no significance. I have decided to remove it.

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



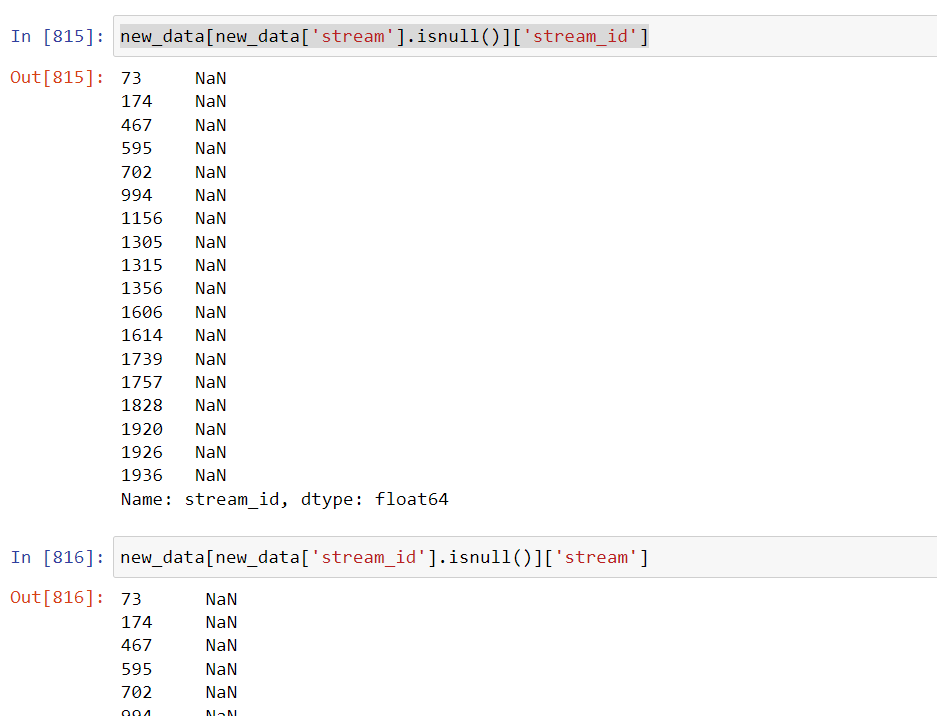
There are many categorical variables presents in dataset ex: 'stream', 'action', 'level', 'mode', 'source'



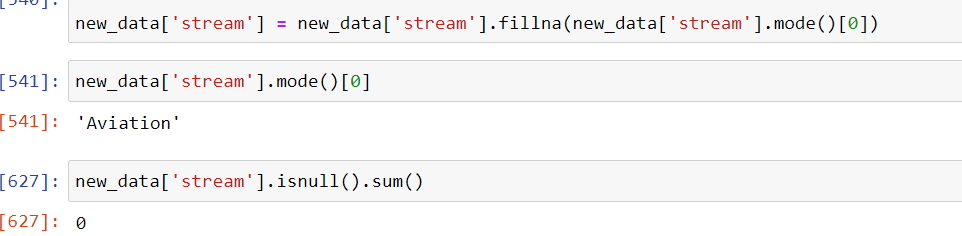
Now we will fill Null Value:

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

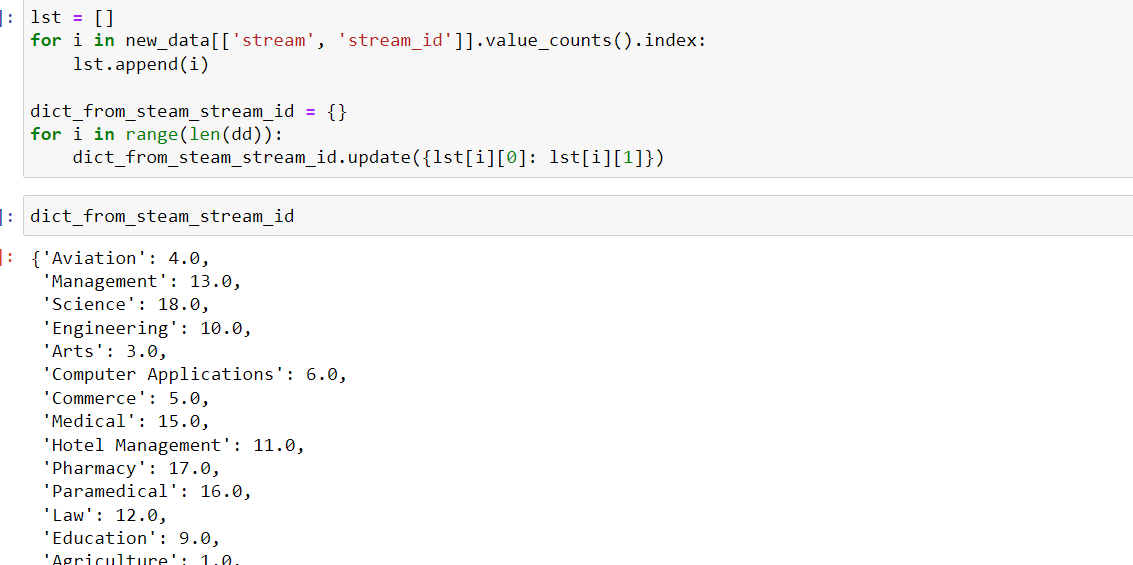
we decided that we fill null value in 'stream' column with its mode, as it is categorical data after that map these value with respective “stream\_id”



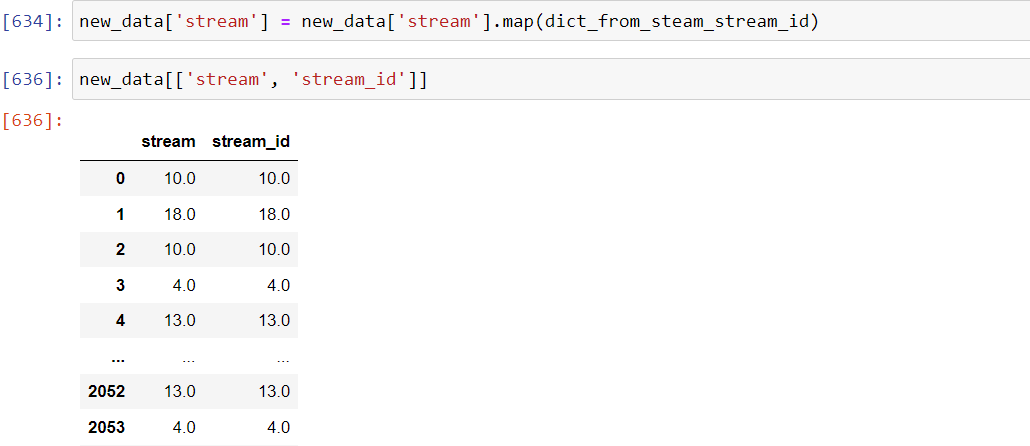
Filling stream column with mode:



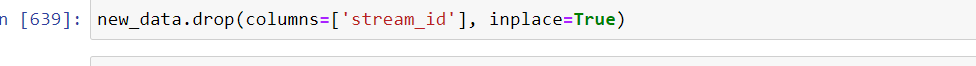
Now we are obtaining respective stream\_id for “stream” column from “stream\_id” column:



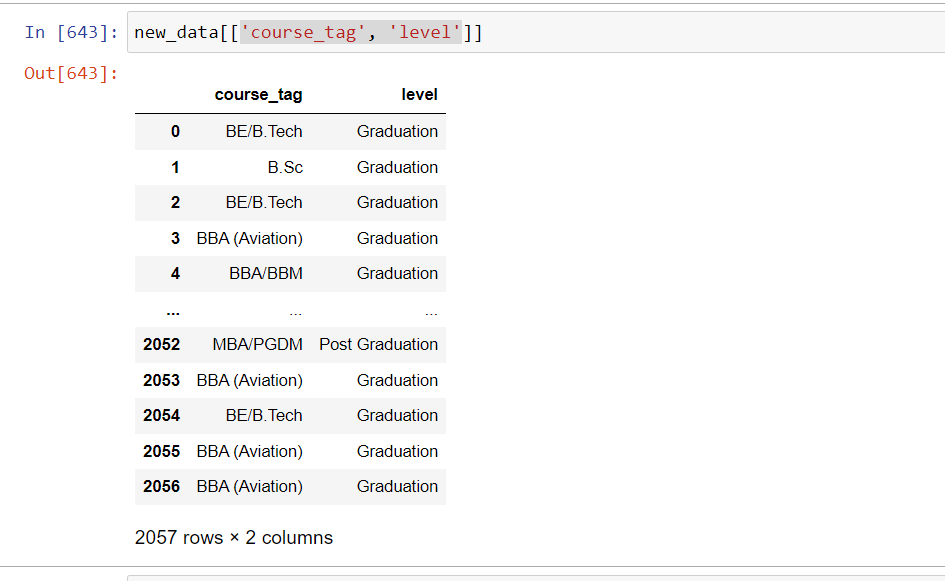
Mapping:



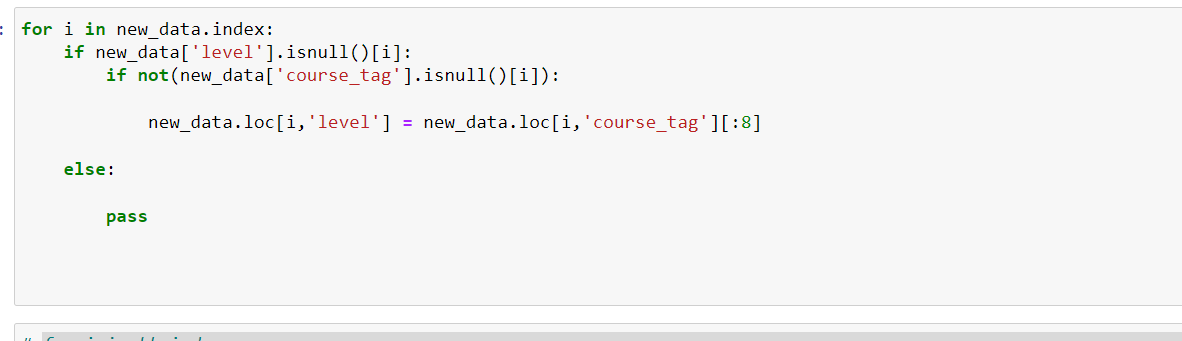
Now we will drop the stream\_id column.

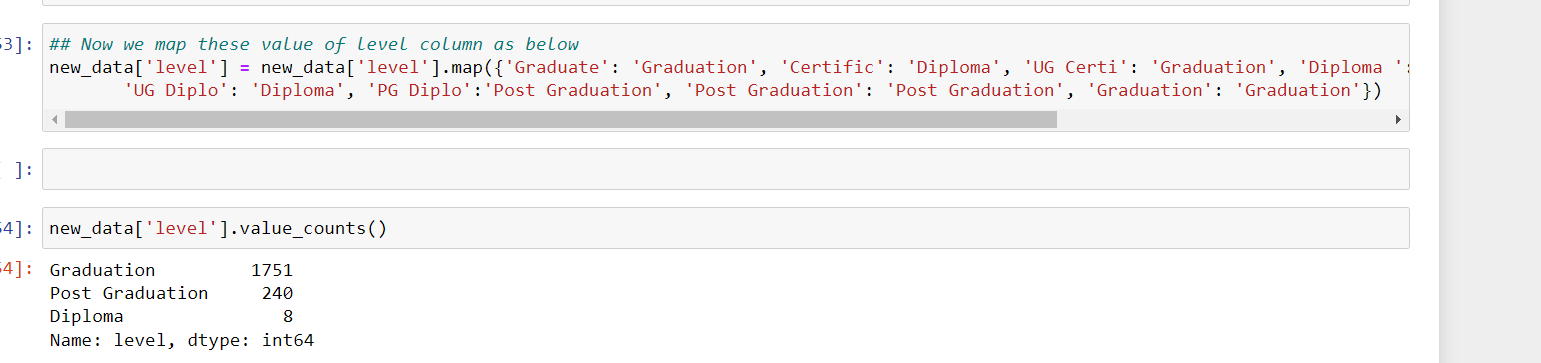


Now I have observed that 'course\_tag', 'level' column are correlated:

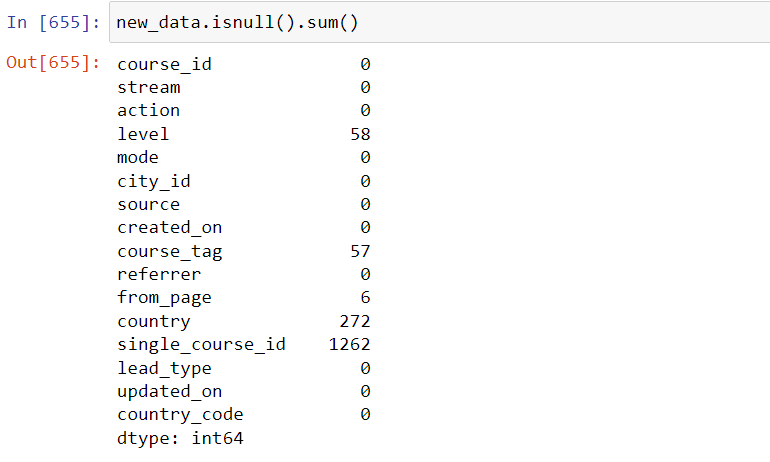


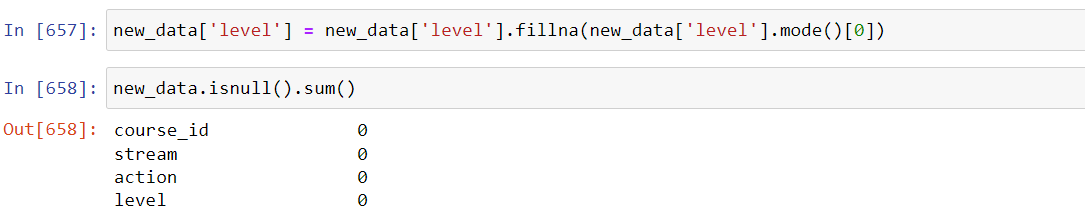
So we are filling level null value with help of “course\_tag” column:





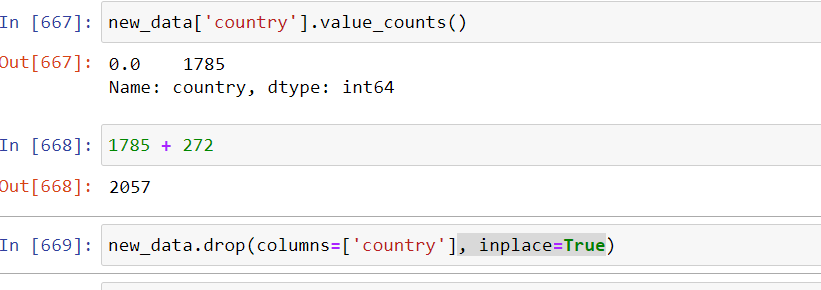
Now level column has only 58 missing value, these value filled by mode



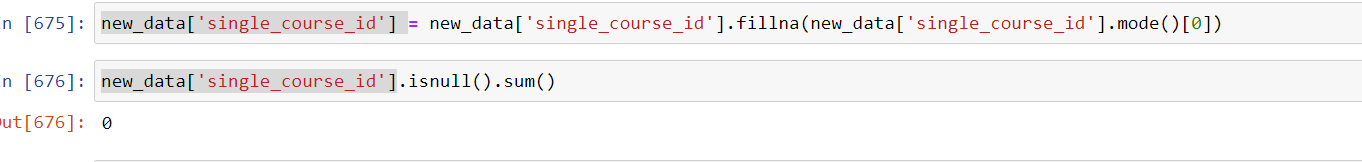


As “course\_tag” and “level” correlated so we are dropping the course\_tag column

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.



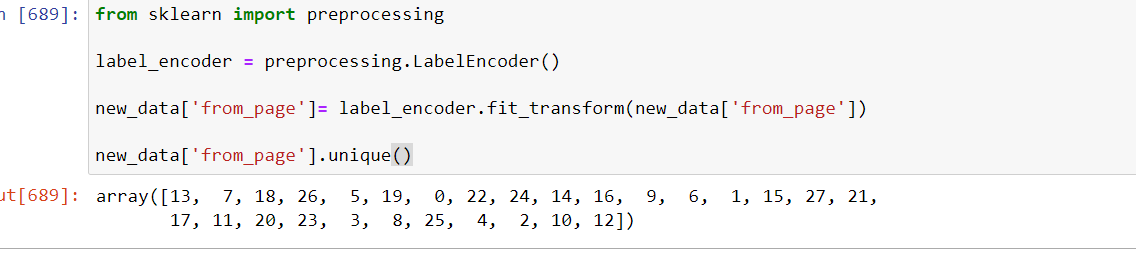
single\_course\_id is categorical feature so we are using mode to replace NaN:

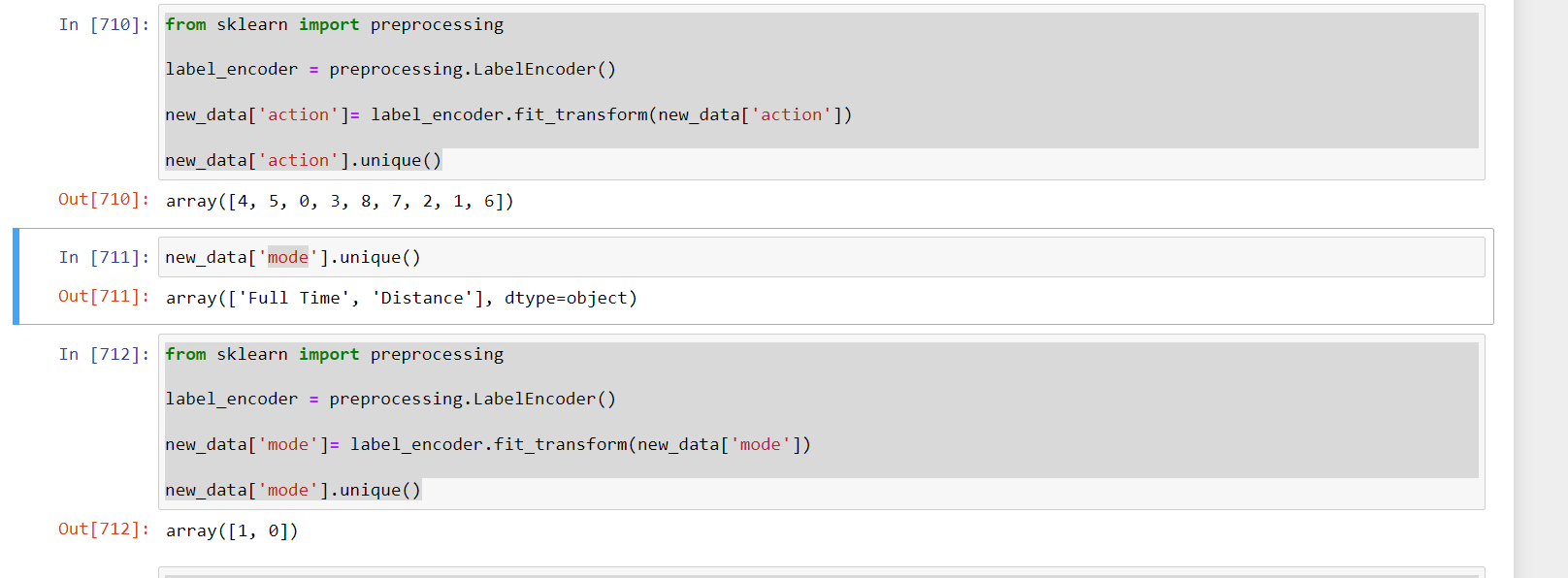


Similarly we are filling the NaN value of “from\_page” column:

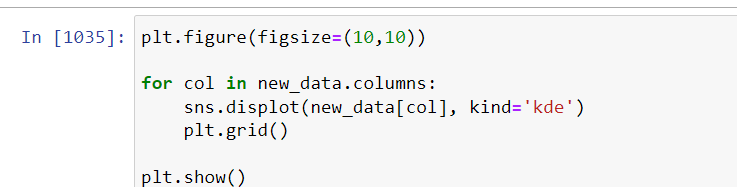


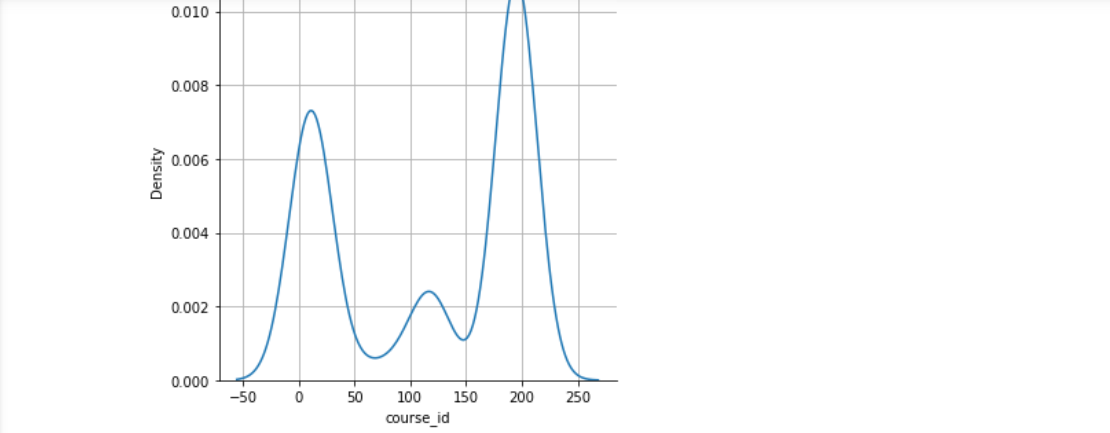
Now we are using sklearn preprocessing library to encode remaining categorical data :

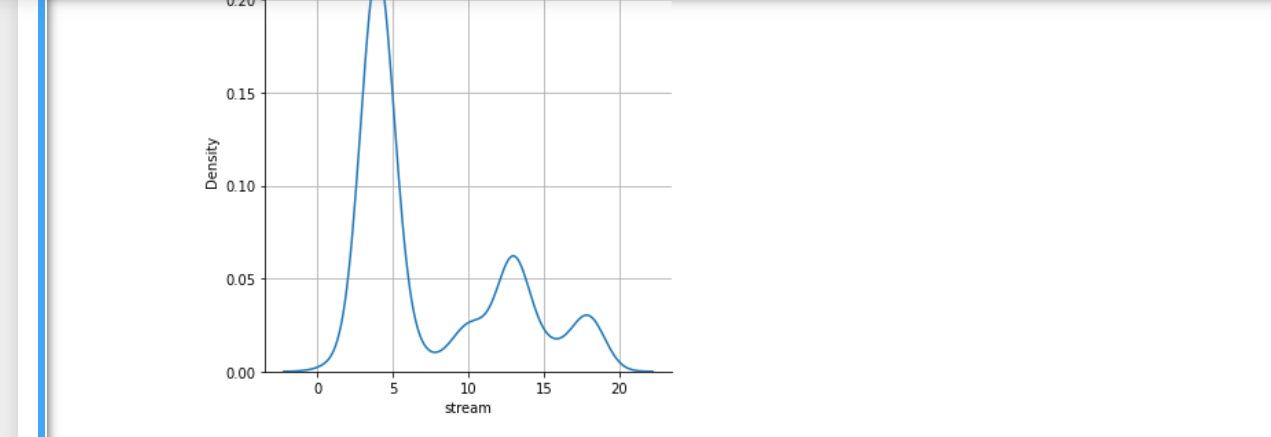


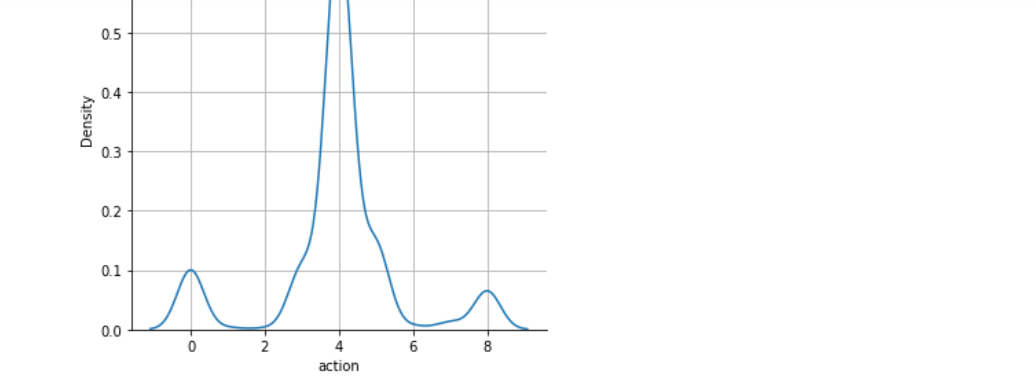


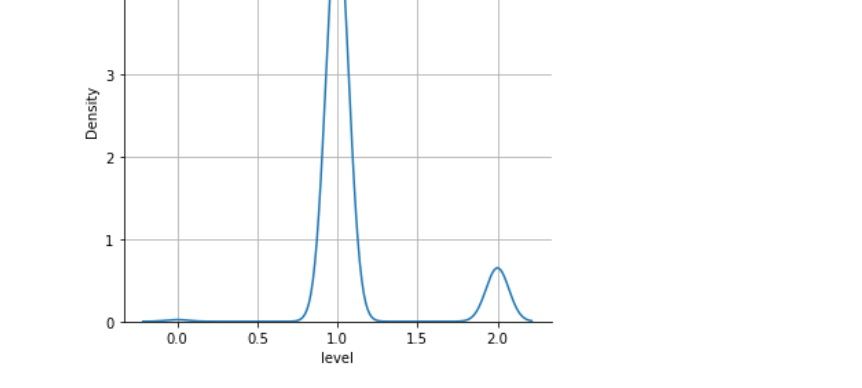
Distribution of Data:





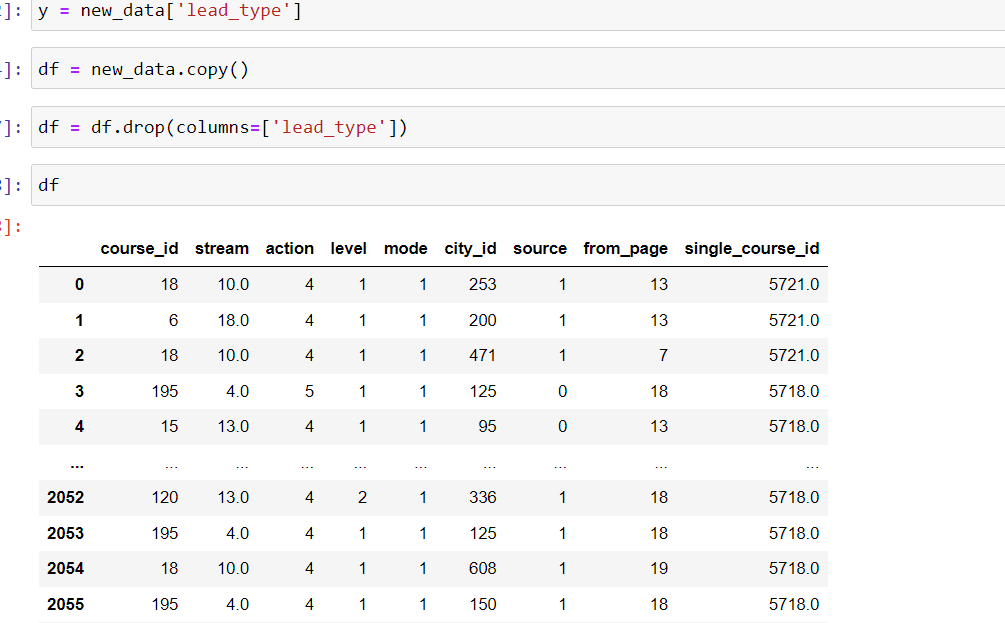


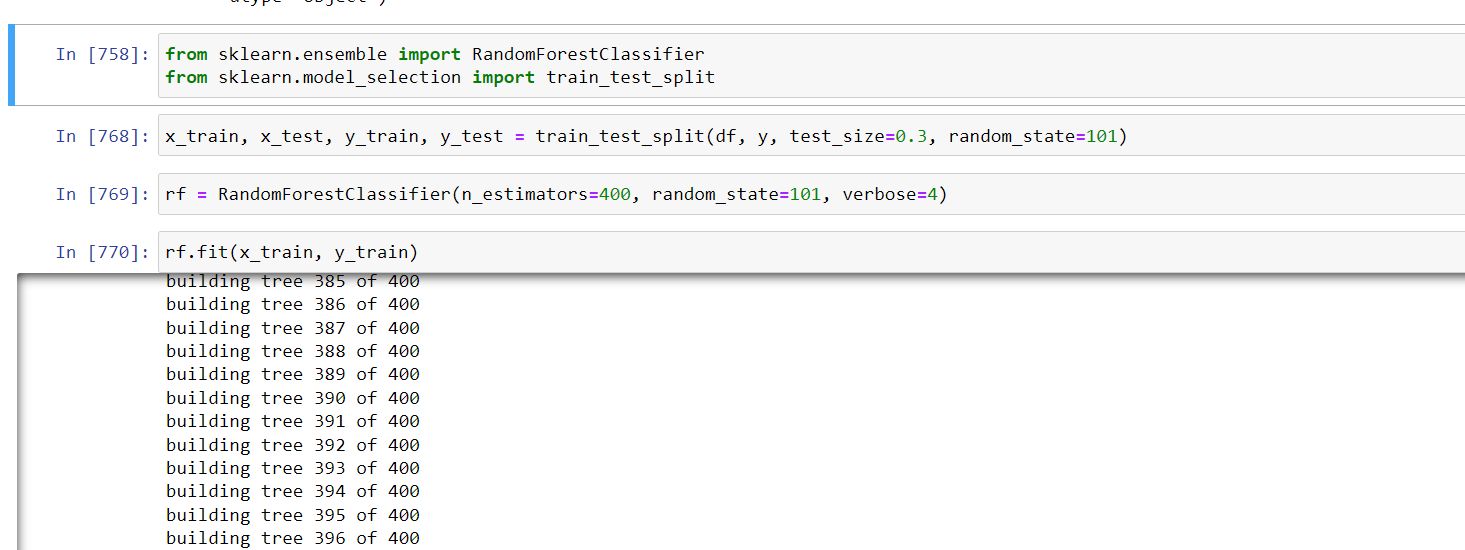


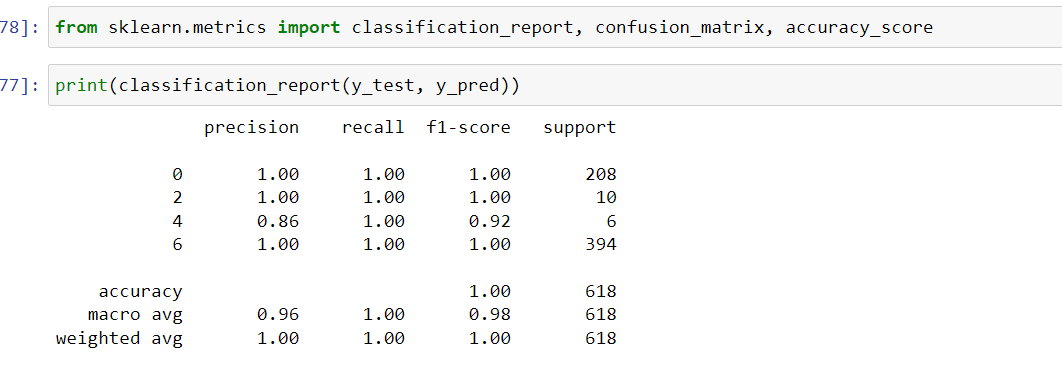


Now our EDA is completed and I am assuming “lead\_type” is output variable.

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







**2.Covid Data: The following csv contains country wise covid information of covid cases,**

**deaths and vaccinations received.**

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

**present.**

**● Which country is more prone to increment of covid cases?**

**● Which continent has received the highest number of vaccinations?**

**● Perform country wise monthly analysis on the given dataset. Compare the results**

**and draw conclusions.**

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

**Sol:**

**United States is more prone to increment of covid cases as it has most new cases of covid.**

**Asia has received the highest number of vaccinations**

Q3>

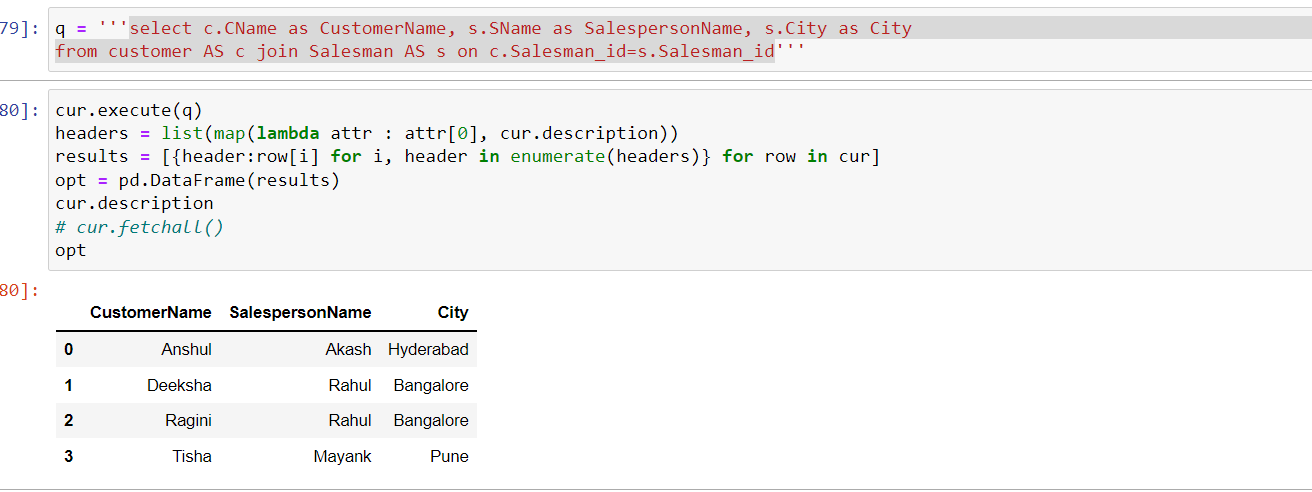
**SQL Case Studies:**

**Queries: From the following tables, write a SQL query to find the salespersons and customers who live in the same city. Return customer name, salesperson name and salesperson city.**

**Sol:**

select c.CName as CustomerName, s.SName as SalespersonName, s.City as City

from customer AS c join Salesman AS s on c.Salesman\_id=s.Salesman\_id



**● write a SQL query to find those salespeople who generated orders for their**

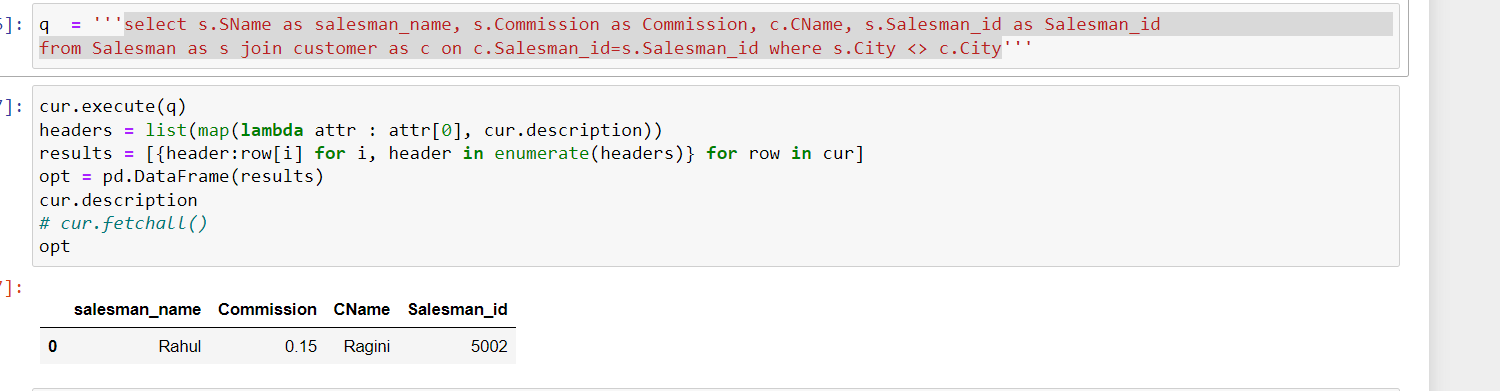
**customers but not located in the same city. Return salesman\_name,**

**commission, cust\_name, customer\_id & salesman\_id.**

**Sol:**

select s.SName as salesman\_name, s.Commission as Commission, c.CName, s.Salesman\_id as Salesman\_id

from Salesman as s join customer as c on c.Salesman\_id=s.Salesman\_id where s.City <> c.City

****

**● Create a view tagging which customers are handled by which sales person. Display Customer name, Salesman name**

Sol:

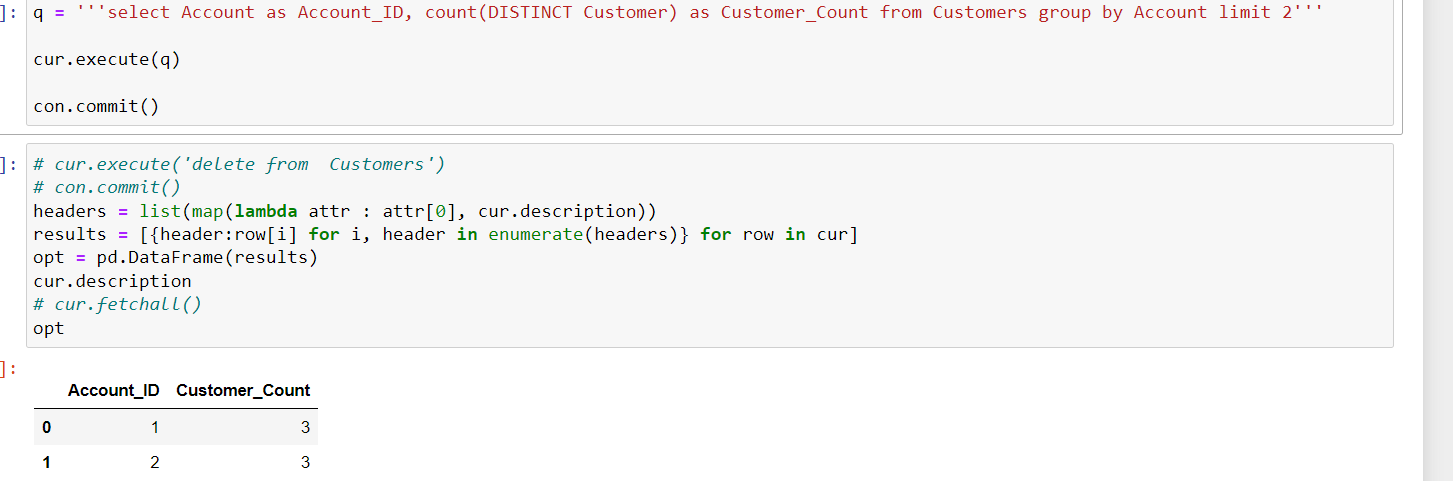
CREATE VIEW tagging As select c.CName as CustomerName from customer AS c left join Salesman AS s on c.Salesman\_id=s.Salesman\_id

union all select s.SName as SalespersonName from Salesman AS s

**2. Extract the leading two accounts having the highest total distinctive customers based on every month.**

**Sol:**

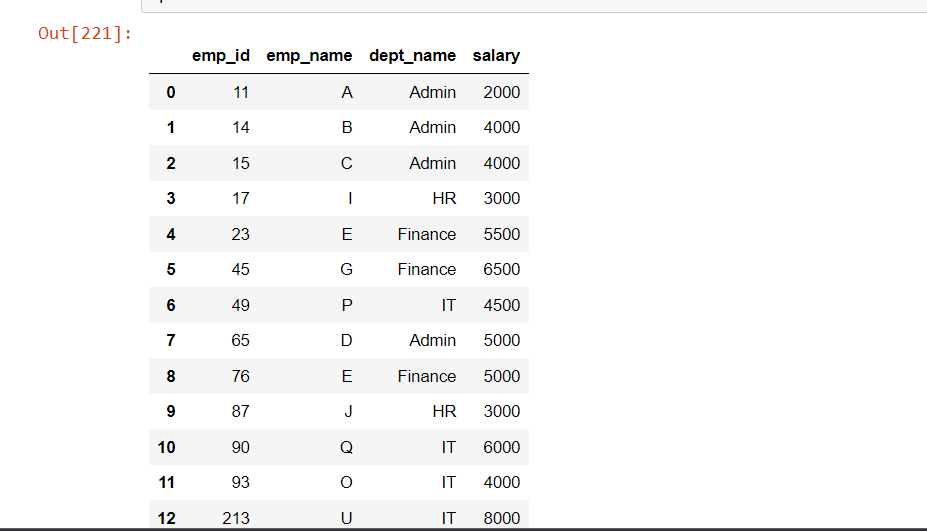
select Account as Account\_ID, count(DISTINCT Customer) as Customer\_Count from Customers group by Account limit 2'

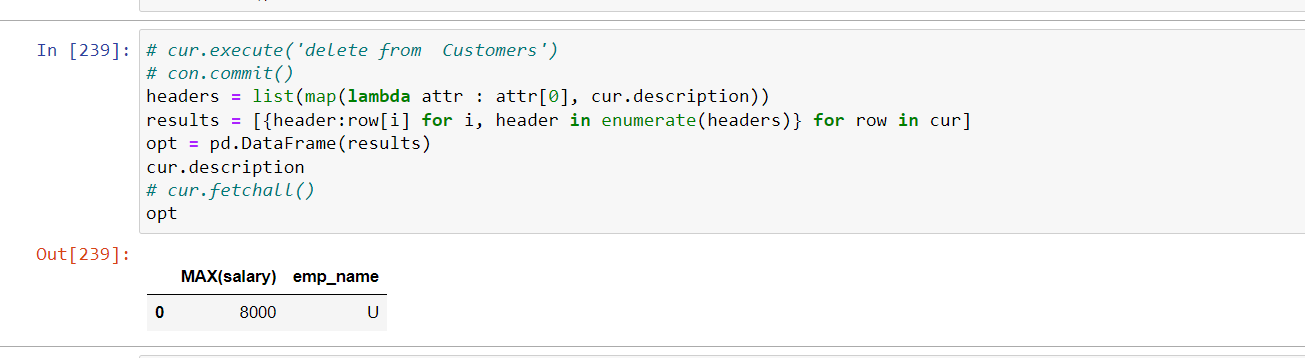


**3. Write an SQL query to display: ●**

**Employee having the highest salary ●**

select MAX(salary), emp\_name from EMPLOYEE



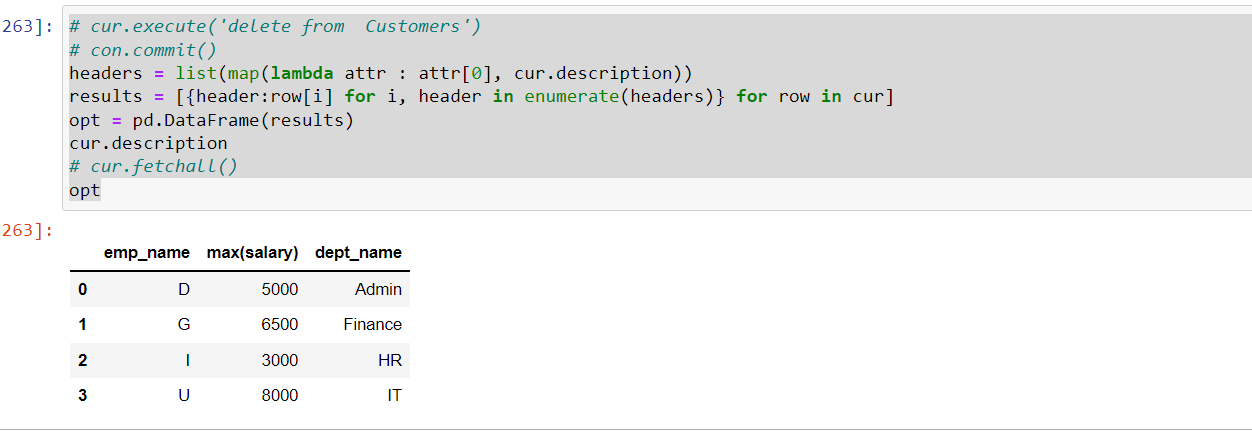


**Department having the highest total salary**

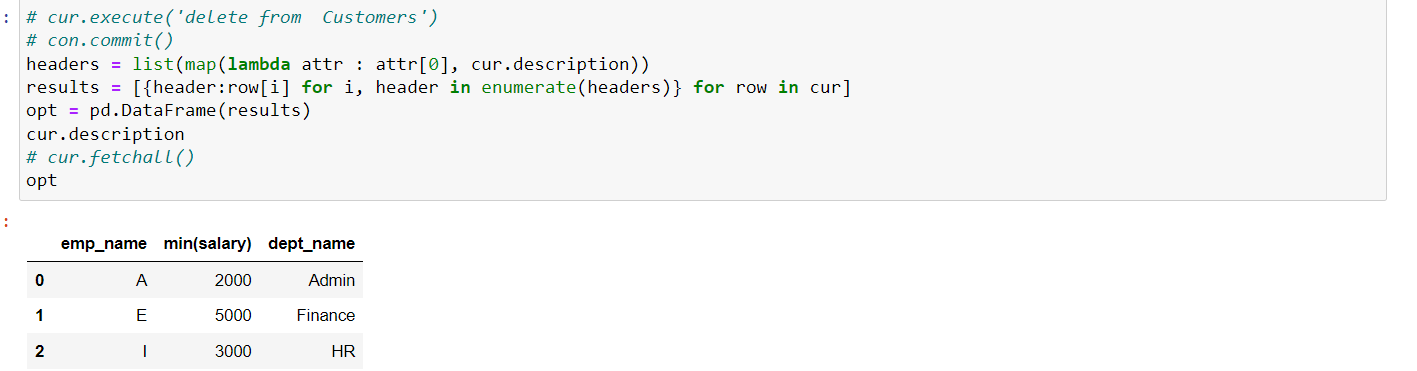
select MAX(salary), dept\_name from EMPLOYEE

**● Employees who either earn the highest salary or the lowest salary in each department from the employee table.**

select emp\_name, max(salary), dept\_name from EMPLOYEE group by dept\_name



select emp\_name, min(salary), dept\_name from EMPLOYEE group by dept\_name



**4: Display the below Output Table by writing queries using the above two tables(Table1 and Table2).**

SELECT Id FROM Source WHERE Id NOT IN (SELECT Id FROM Target)

**5. What will be the result of the following query: SELECT P.First\_Name, P.Last\_Name FROM Persons AS P LEFT OUTER JOIN Vaccinations AS V ON P.Email = V.Recipient WHERE V.Comments IS NULL;**

Ans:

It will return the First Name, Last Name from Persons table where Persons Email value matching with Vaccinations Recipient value and Vaccinations Comments value is NULL.

**6. How can a supermarket(Ex:Big Bazar), use its customer data(Ex: Name, Phone Number, Items Purchased,etc) to provide better services to its customers and improve corporate performance?**

**Ans:**

Based on data (ex: Item Purchased, no of item purchased, price, contact) we can make a cluster or do anomaly detection. By using anomaly detection, we can make separate group of users who spending more money or purchasing more items and provide them some offer based on their purchased history.

By using clustering, we can make cluster (ex: heigh purchase, medium, lower) of user based on there purchase and provide offer directly to there phone no or contact no.

For the user who are in lower side in purchasing item, our sales person needs to contact them and convince them by giving some offer on lower price item. It will increase the total customer number and also pull some new customer.

Basically, we will make clusters of different users and provide them offer as per there cluster.