**Space Server Dataset**

**Introduction:**

**1. Star –** It is a type of astronomical object consisting of a luminous spheroid of plasma which is held together by its own gravity. It is a luminous fixed point in the night sky which has a large remote incandescent body just like Sun. The nearest star to Earth is the Sun.

**2.** **Galaxy** – It is a gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter. Galaxies are categorized according to their visual morphology as elliptical, spiral, or irregular. Many galaxies are thought to have super massive black holes at their active centers. It is a system of millions or billions of stars, together with gas and dust, held together by gravitational attraction.

3. **Quasars -** A quasar is an extremely luminous active galactic nucleus, in which a super massive black hole with mass ranging from millions to billions of times the mass of the Sun is surrounded by a gaseous accretion disk. It is a massive and extremely remote celestial object, emitting exceptionally large amounts of energy and typically having a star like image in a telescope. It has been suggested that quasars contain massive black holes and may represent a stage in the evolution of some galaxies.

**1. Problem Statement:**

The data consists of 10,000 observations of space taken by the SDSS. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar. The data released by the SDSS is under public domain. It’s taken from the current data release **RD14**. The dataset offers plenty of information about space to explore. The class identifies an object to be galaxy, star or quasar.

**Dataset Description:**

The following are the descriptions of the features present in the dataset and they are as follows:

1. objid - Object Identifier

2. ra - J2000 Right Ascension (r-band)

3. dec - J2000 Declination (r-band)

4. u - better of DeV/Exp magnitude fit

5. g - better of DeV/Exp magnitude fit

6. r - better of DeV/Exp magnitude fit

7. i - better of DeV/Exp magnitude fit

8. z - better of DeV/Exp magnitude fit

9. run - Run Number

10. rerun - Rerun Number

11. camcol - Camera column

12. field - Field number

13. specobjid - Object Identifier

14. class - object class (galaxy, star or quasar object)

15. redshift - Final Redshift

16. plate - plate number

17. mjd - MJD of observation

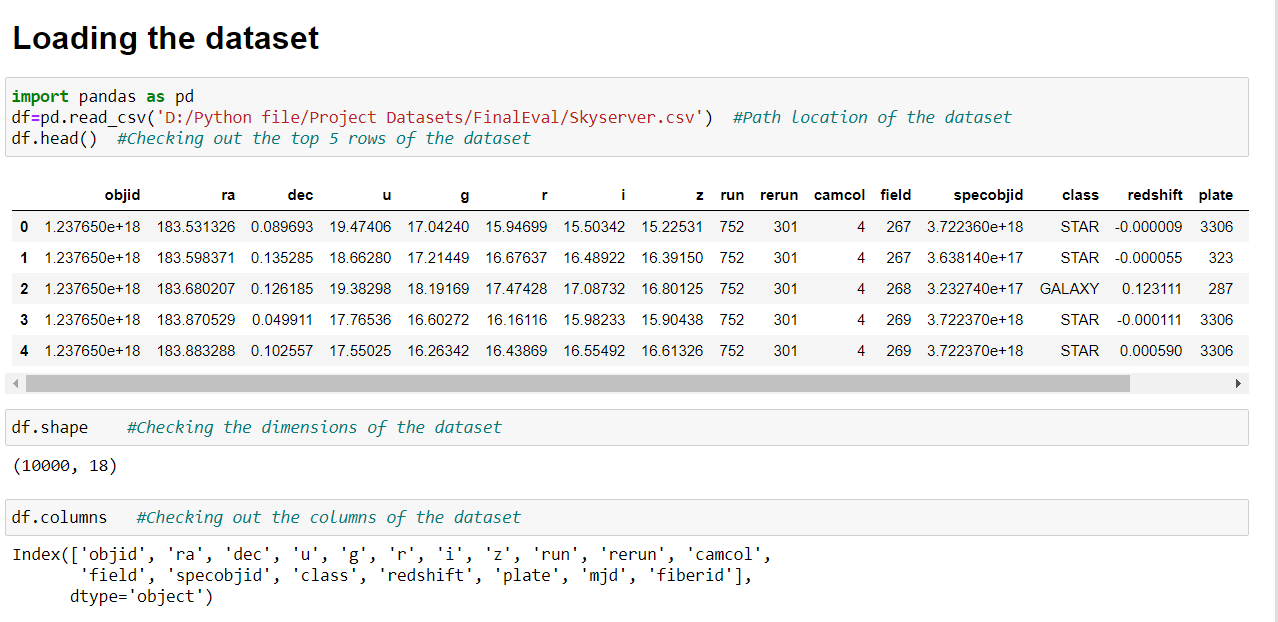
18. fiberid - fiber ID

We need to take class as the target variable and predict whether the data observation is either a star, galaxy or quasar and we are going to approach this problem statement by doing **classification analysis**.

**Importing Warning Library:**



**Load the dataset and check out its features and dimensions:**



**Checking for any null values in the dataset:**



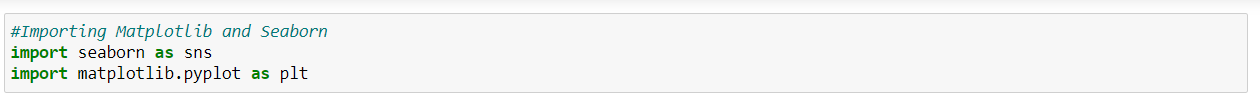
**Observations:**

1. In the Skyserver dataset, we have 10 float64 data type features, 7 int64 data type features and 1 object type feature present.
2. There are no null values present in the dataset.

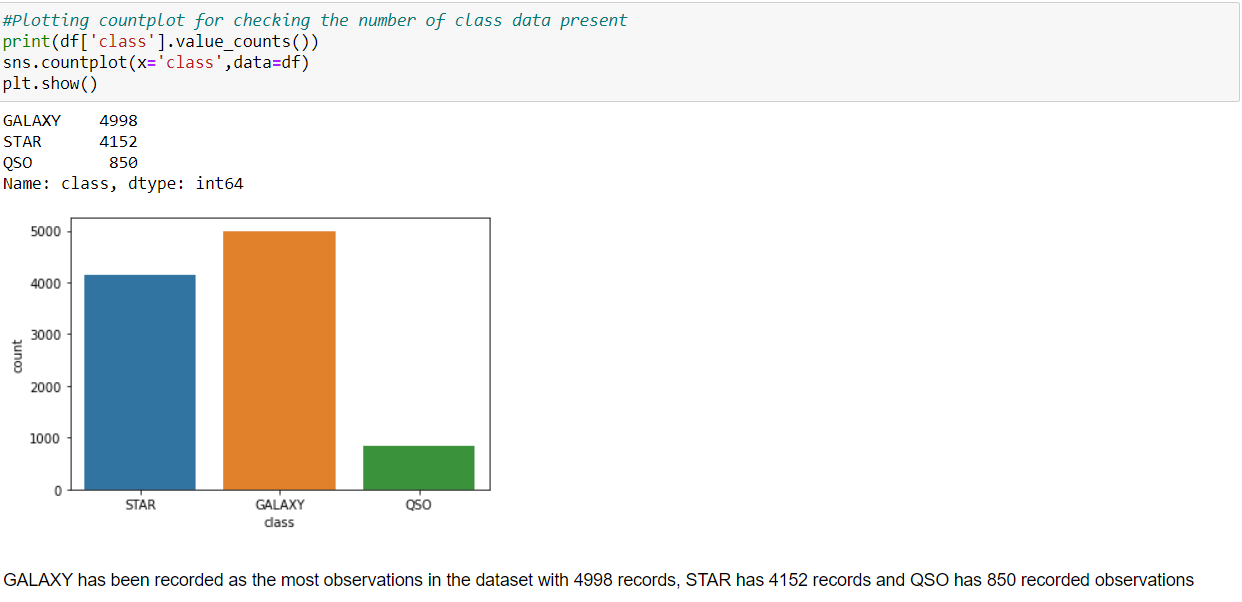
**2. Data Analysis:**

We will be using Exploratory Data Analysis (EDA) to check the relationship between features and also, we can understand the distribution of data in the dataset given. We plot graphs for the features by using data visualization libraries like matplotlib and seaborn.

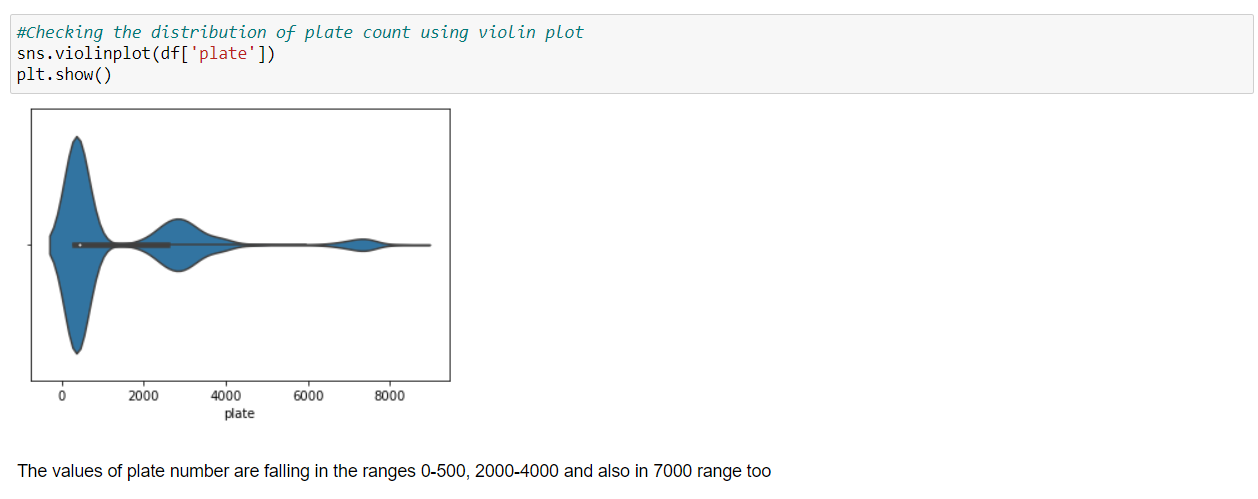
**Importing matplotlib and seaborn library:**



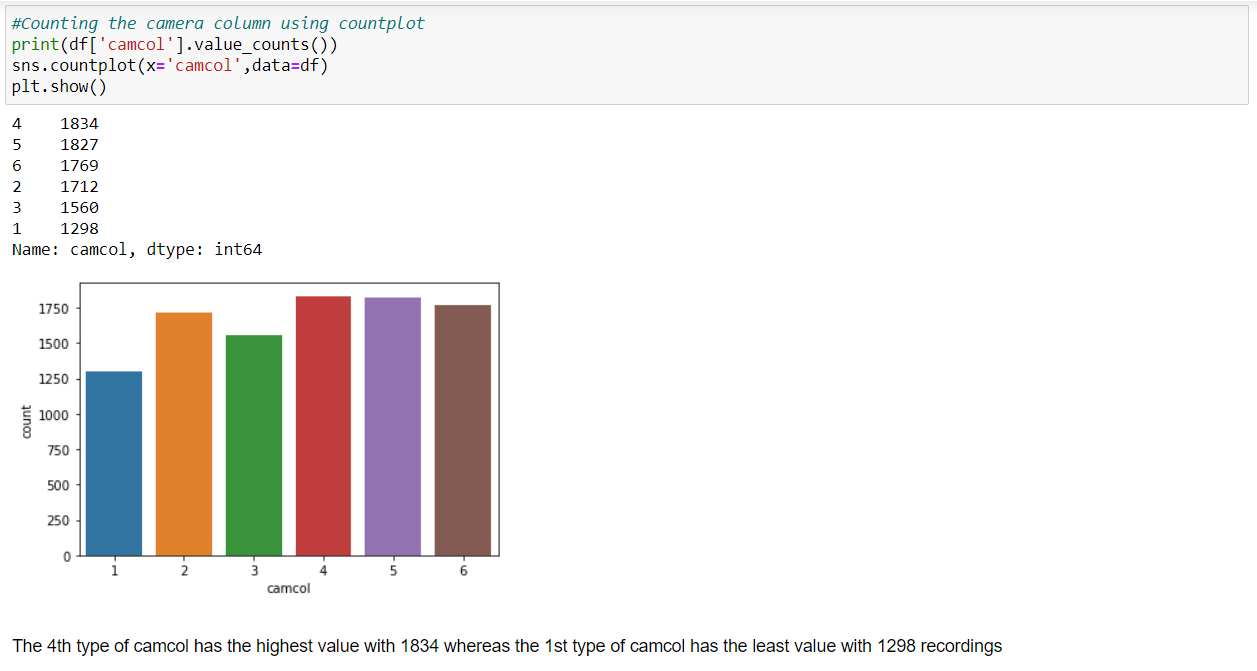
**Checking the number of class data present:**



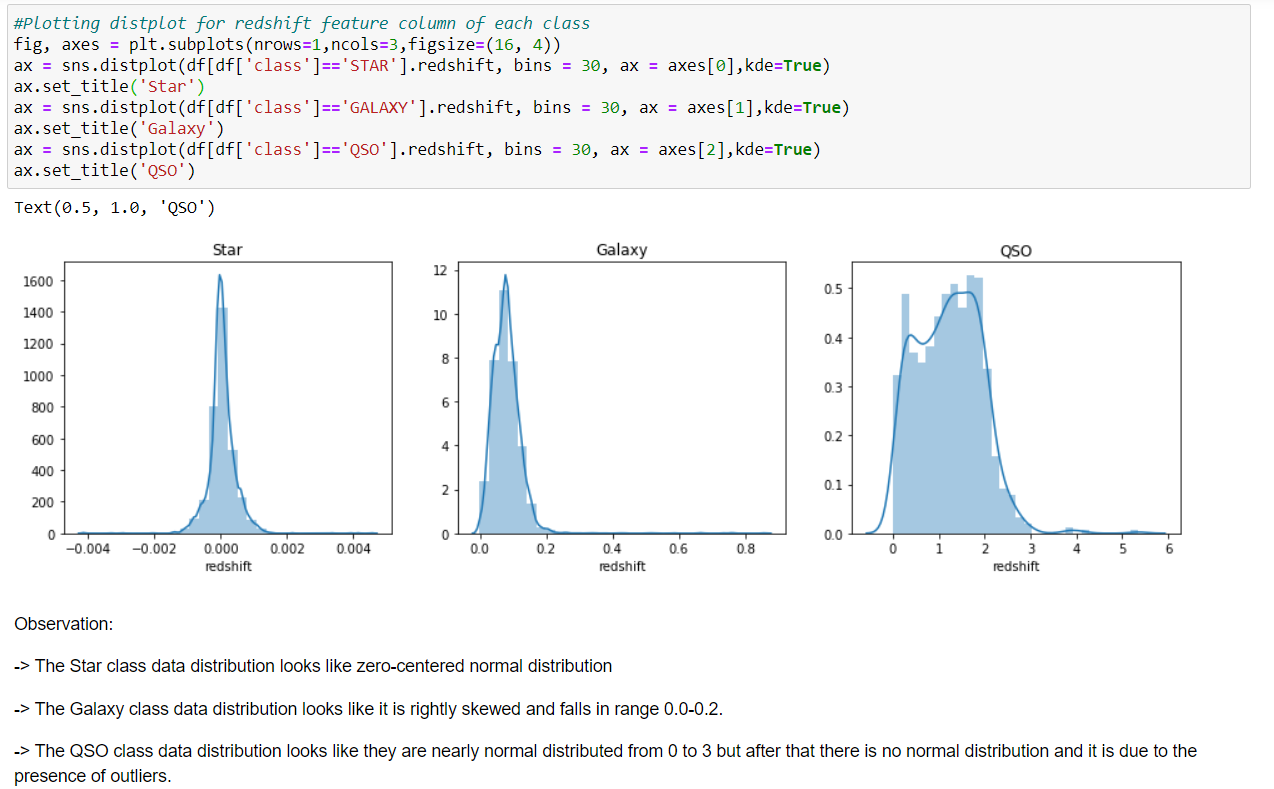
**Distribution of plate using violin plot:**



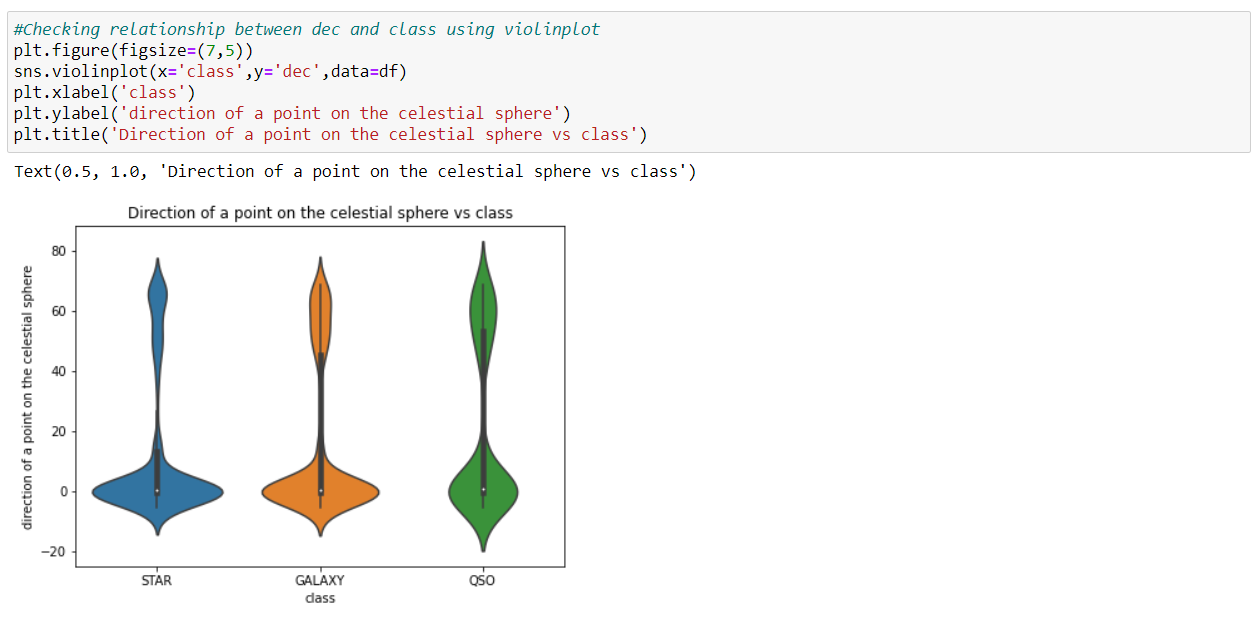
**Checking the camcol feature count:**



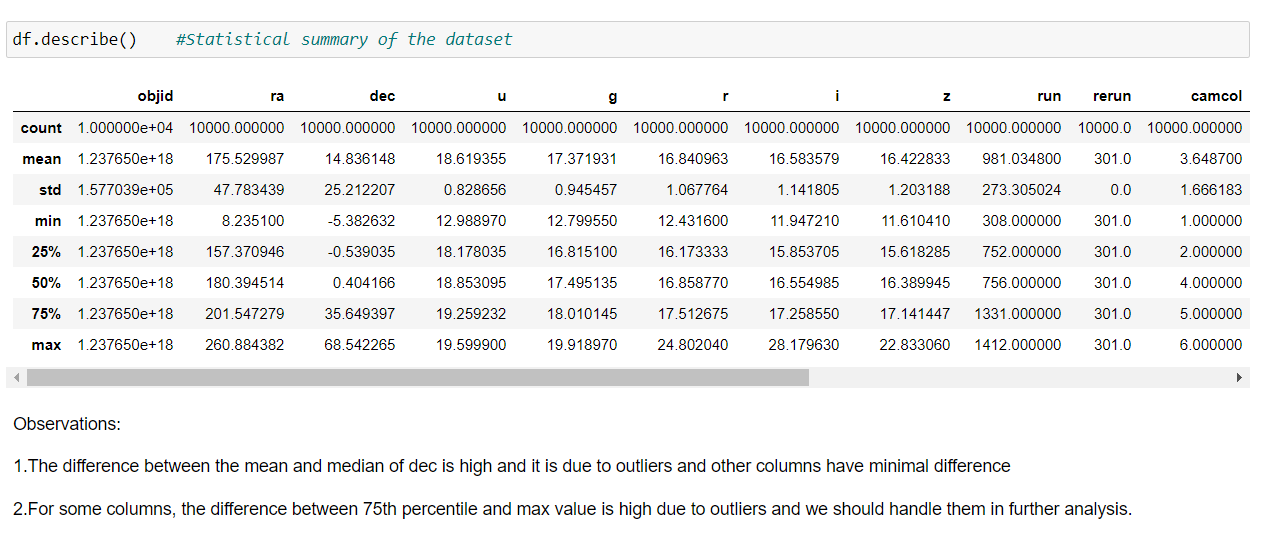
**Distplot for redshift feature column for each class:**



**Checking relationship between dec and class using violin plot:**



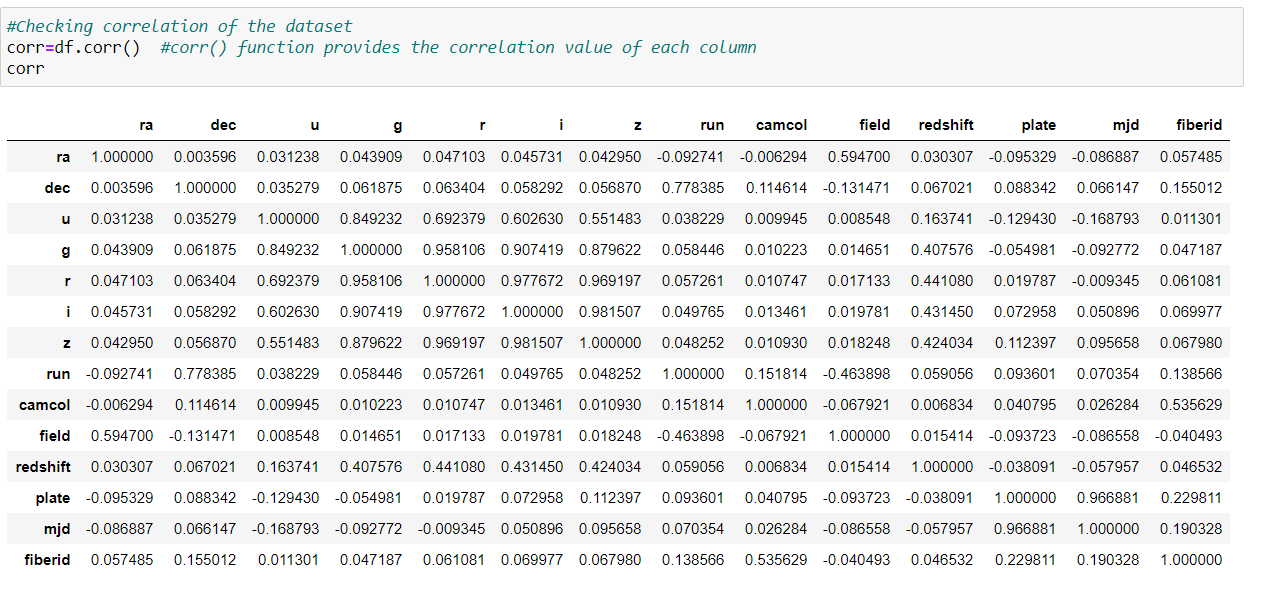
**Checking the Statistical summary of the dataset:**



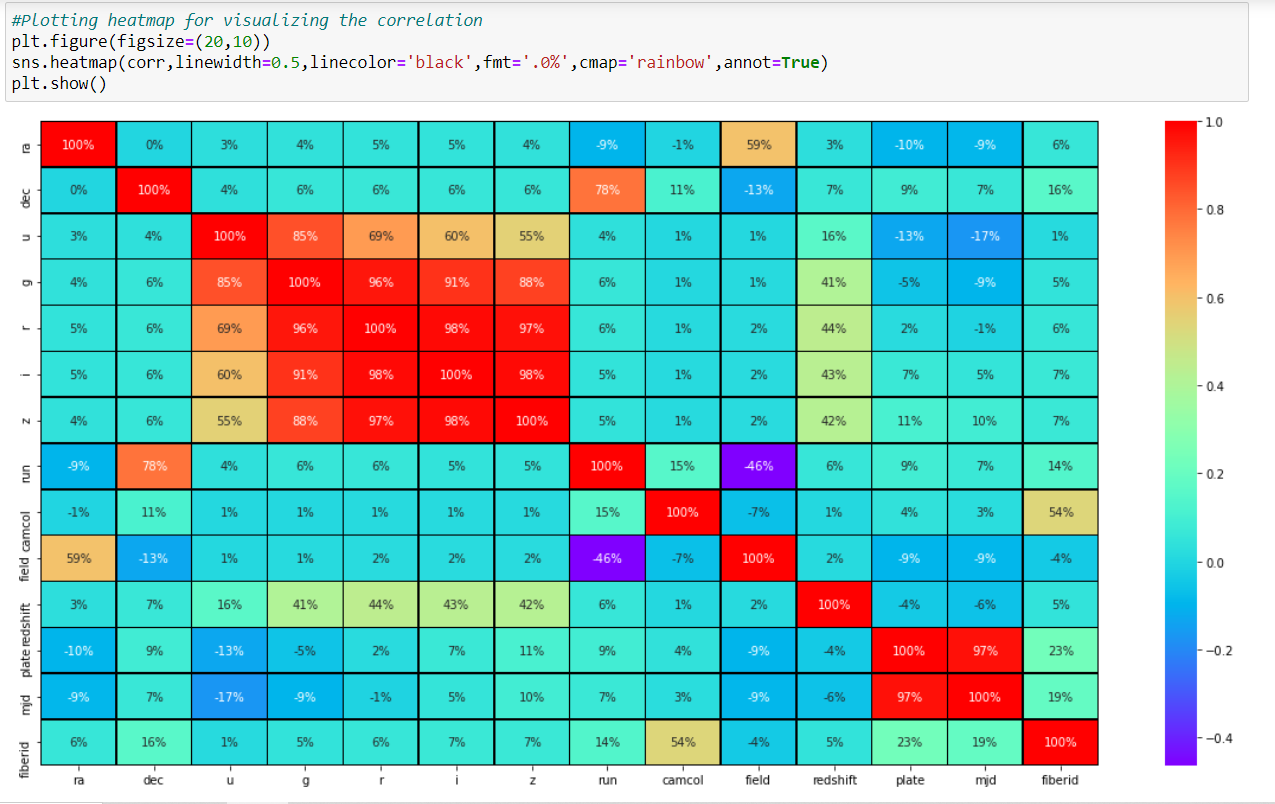
**Correlation:**

The statistical relationship between two variables is referred to as their correlation. The correlation factor represents the relation between columns in a given dataset. A correlation can be positive, meaning both variables are moving in the same direction or it can be negative, meaning that when one variable's value increasing, the other variable’s value is decreasing.

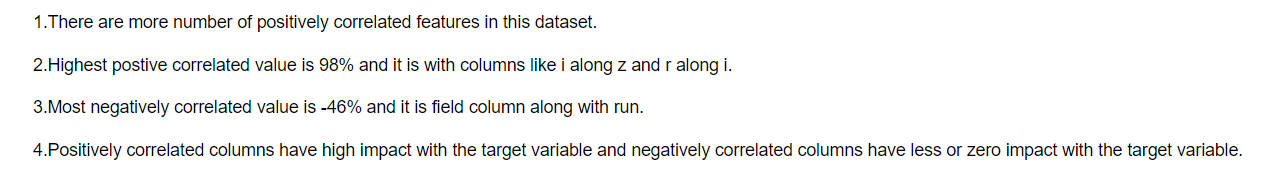
**Checking the correlation using corr:**



**Plotting heatmap for visualizing the correlation:**

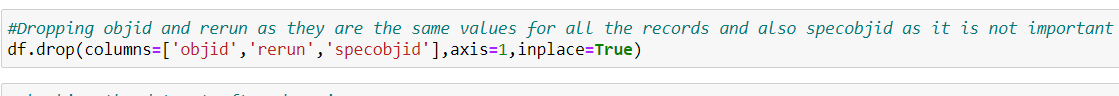


**Observations:**



We can see that rerun, objid and specobjid columns are not contributing well in the dataset as it is same across the columns and it is not correlated to any of the other columns. Considering its nil contribution towards the model building, we can drop these columns for further analysis.

**Dropping the columns:**

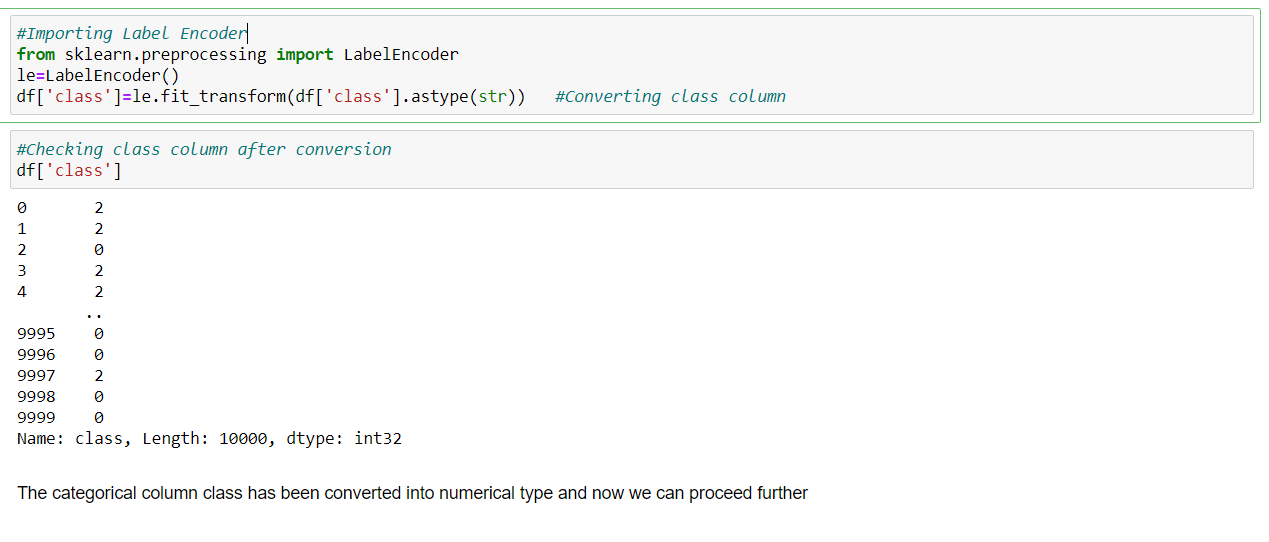


**Label Encoding:**

As class column has categorical data, we need to convert into numerical for proceeding further analysis and it can be done by using Label Encoder.

**Label Encoder** - refers to converts the labels into numeric form i.e., the columns which are in alphabetical or categorical values are assigned with numbers so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated during the process.

Import label encoder for converting categorical data (object) into numeric (int 32) data.

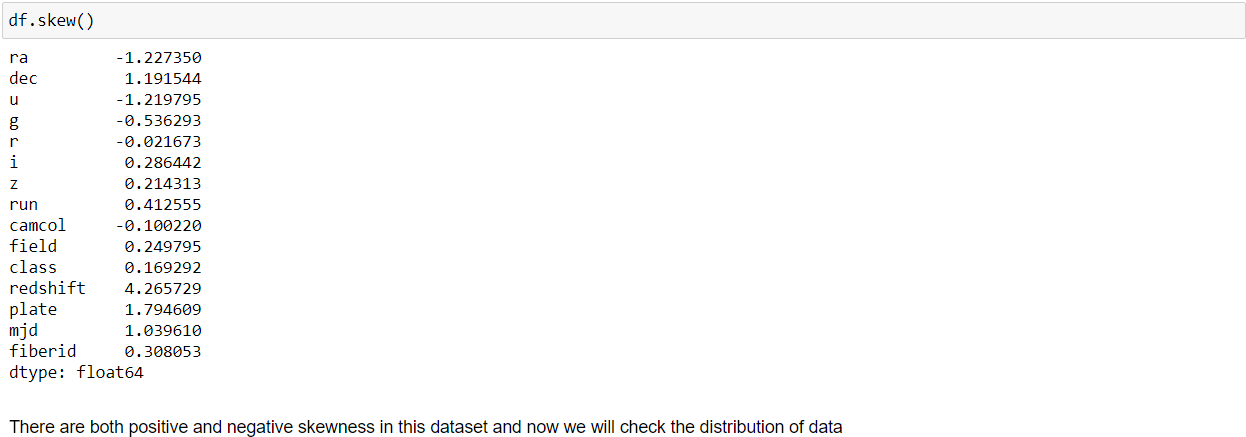


**Skewness:**

Skewness is a measure of the symmetry of a distribution. The highest point of a distribution is its mode. The mode marks the response value on the x-axis that occurs with the highest probability. A distribution is skewed if the tail on one side of the mode is fatter or longer than on the other: it is asymmetrical.

In an asymmetrical distribution a **negative skew** indicates that the tail on the left side is longer than on the right side (left-skewed), conversely a **positive skew** indicates the tail on the right side is longer than on the left (right-skewed). Asymmetric distributions occur when extreme values lead to a distortion of the normal distribution.

**Checking skewness in the dataset:**



The range of skewed data is between -0.55<df<0.55 and if the value is obtained above the range, then it is said to be skewed. We can check the distribution of the features data using distplot.

**Plotting distplot:**



If we run a loop with all the columns for plotting the distplot, we can see that all the columns would be plotted simultaneously. Below is the graphical distribution of the columns data according to the skewness of each column: