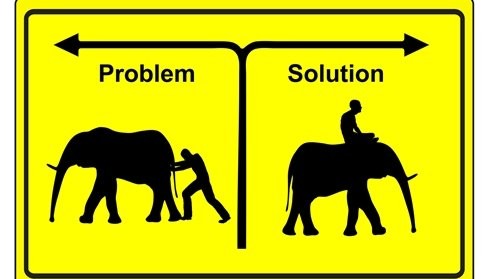
**Sloan Digital Sky Server Dataset Analysis**

Introduction:

Ours is an age of technological advancement and technology has become an integral part of life, also this technological advancement has brought a lot of changes in various fields such as Agriculture, Medical research and many other fields. This technological advanced has helped one such field i.e. Space exploration. From thousands of years humans are very much intrigued to learn about the outer space and many search explorations and researches have be done to study the outer space. The traces of such explorations can also be found in various ancient civilizations right from Mayans till ancient Indian civilization have their various findings about the outer world, in recent time the space exploration has have become very popular as due to technological advancement as well as large companies like SpaceX are investing a lot to carry out this researches and the day is not far from us when we would try to inhabit one such planet in mere future like Mars.

Today as an emerging Data Scientist we would be analysing one such dataset from Sloan Digital Sky Server which consist of different attributes about the various entities in the space and based on this attributes we have to classify them into Stars, Galaxies and Quasar. 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. This is a typical Classification problem we would be analysis the dataset by doing EDA(Exploratory Data Analysis) to analysis the various factors and the correlation between them and understand which factors would help us to class them into 3 different classes Stars, Galaxies and Quasar. After analysing the dataset we would apply various classification algorithm and try to build a Machine Learning model which would help us do prediction to the best possible accuracy.

Problem Definition:

The problem statement is the most crucial step towards solving any data analytics problem. This is usually defined by the organization. For example: "We need to increase sales of product X in category Y."

As a data scientist, you need to think of the problem statement in mathematical terms. For example: Why is product X underperforming? What data on user behaviour might help explain it?

This is where pain points come in. These are areas where the business is struggling. Some are obvious; others may be undiscovered until further on in your data analysis.

Before writing the problem statement, a data scientist needs to consider the following:

* What problem is the organization trying to solve?
* What impact does this problem have on the organization?
* What are the potential benefits of solving this problem?

The problem statement generally follows the format:

"The problem P, has the impact I, which affects B, so a good starting point would be S."

P = The problem

I = The pain points the organization is facing

B = Which parties are affected by the problem (eg: customers, suppliers, IT)

S = The proposed course of action

The next step is to translate business goals into data analysis goals so you can determine a course of action. Decide if the expected benefits are realistic and attainable from a data standpoint; for example, how long will the project take? Will you need additional data sources? Is the existing dataset accurate and adequate?

Defining the problem statement clearly and precisely helps us to understand what actually the problem is and what is the actual output that is required and helps us to build the best possible model for the same.

**Here in our SDSS dataset our problem statement or definition is” To analyse the dataset and the different attributes in the dataset and to classify the dataset in 3 different classes Star, Galaxy and Quasar”. We have to make the best possible machine learning model which analysis this dataset and which would help us do prediction by learning from the dataset. This a typical Classification problem so we would be using different classification algorithm to analyse and build the machine learning model to which would give us best possible prediction.**

Now that we have defined our problem statement we have a clear picture that we have to do i.e. we have to classify the whole dataset in three different classes star, galaxy and quasar based on their attributes, now the next step is EDA(Exploratory Data Analysis).

Exploratory Data Analysis (EDA):



The next step in any Data science project is EDA(Exploratory Data Analysis), as the name suggest we have to explore and analyse the data using different visualization tools like bar graphs, histogram, scatter plot etc which would help us to get some insight above the data and would help us to understand how the data is distributed, which are the important key attributes which helps in predicting the final output and many more insights.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Now lets try to understand our SDSS dataset and gain some insight about the data which help us to identify some patterns anomalies that are present in the dataset, but there are some important terminologies that we would like to understand that are present in the dataset.

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 table results from a query which joins two tables (actuaclly views): "PhotoObj" which contains photometric data and "SpecObj" which contains spectral data.

To ease your start with the data you can read the feature descriptions below:

***View "PhotoObj"***

* objid = Object Identifier
* ra = J2000 Right Ascension (r-band)
* dec = J2000 Declination (r-band)

Right ascension (abbreviated RA) is the angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth in question. When paired with declination (abbreviated dec), these astronomical coordinates specify the direction of a point on the celestial sphere (traditionally called in English the skies or the sky) in the equatorial coordinate system.

Source: <https://en.wikipedia.org/wiki/Right_ascension>

* u = better of DeV/Exp magnitude fit
* g = better of DeV/Exp magnitude fit
* r = better of DeV/Exp magnitude fit
* i = better of DeV/Exp magnitude fit
* z = better of DeV/Exp magnitude fit

The Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope.

Further education: <https://www.astro.umd.edu/~ssm/ASTR620/mags.html>

* run = Run Number
* rereun = Rerun Number
* camcol = Camera column
* field = Field number

Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by:

* run number, which identifies the specific scan,
* the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and
* the field number. The field number typically starts at 11 (after an initial rampup time), and can be as large as 800 for particularly long runs.
* An additional number, rerun, specifies how the image was processed.

***View "SpecObj"***

* specobjid = Object Identifier
* class = object class (galaxy, star or quasar object)

The class identifies an object to be either a galaxy, star or quasar. This will be the response variable which we will be trying to predict.

* redshift = Final Redshift
* plate = plate number
* mjd = MJD of observation
* fiberid = fiber ID

In physics, redshift happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

Each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. These fibers then feed into the spectrographs. Each plate has a unique serial number, which is called plate in views such as SpecObj in the CAS.

Modified Julian Date, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.

The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID.

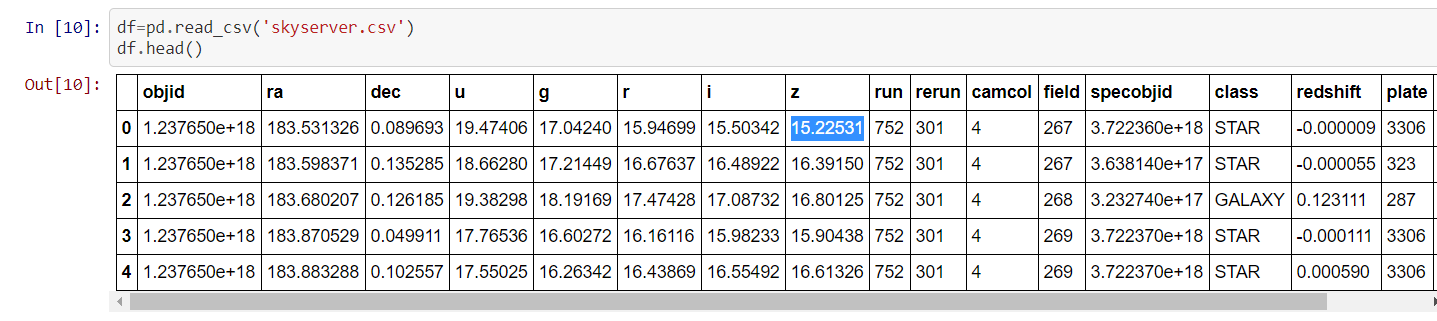
Further information on SDSS images and their attributes:

<http://www.sdss3.org/dr9/imaging/imaging_basics.php>

<http://www.sdss3.org/dr8/glossary.php>

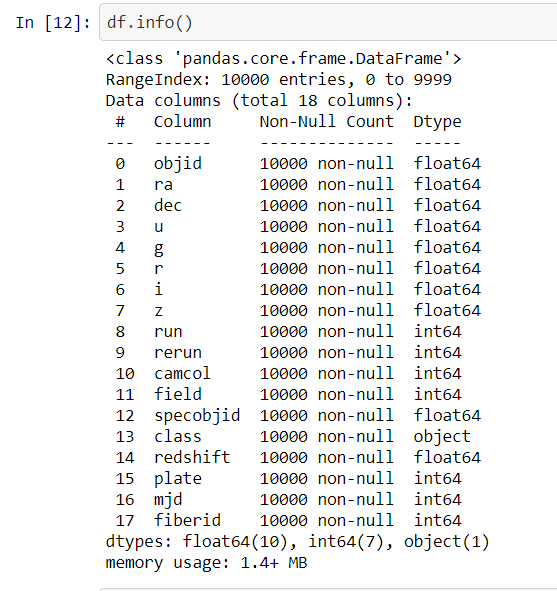
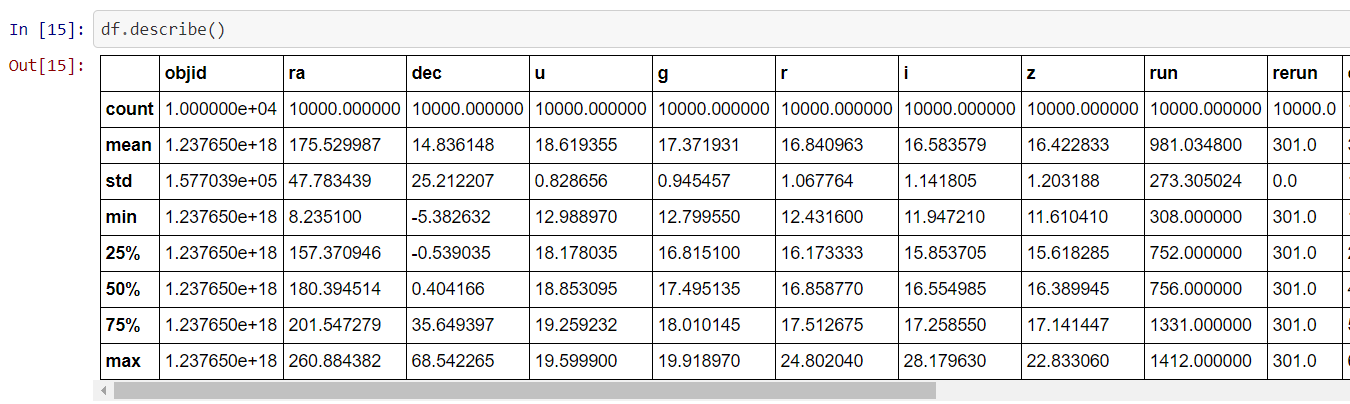
Now that we have understood all the different attributes in the dataset and what do they mean lets go ahead with our EDA process

We start off by importing all the necessary library files that we would be requiring for visualization, pre-processing of data, different model which we would be requiring to analyse the data, various metrics to check the accuracy and various parameters about the models which would help us to understand the best possible model for our data and last but not the least saving the best model for future predictions. At first we have to import all this files as the definition of this files are already present in our anaconda framework so that we can do all the necessary step from analysing to saving the model.

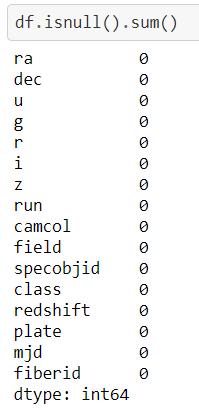
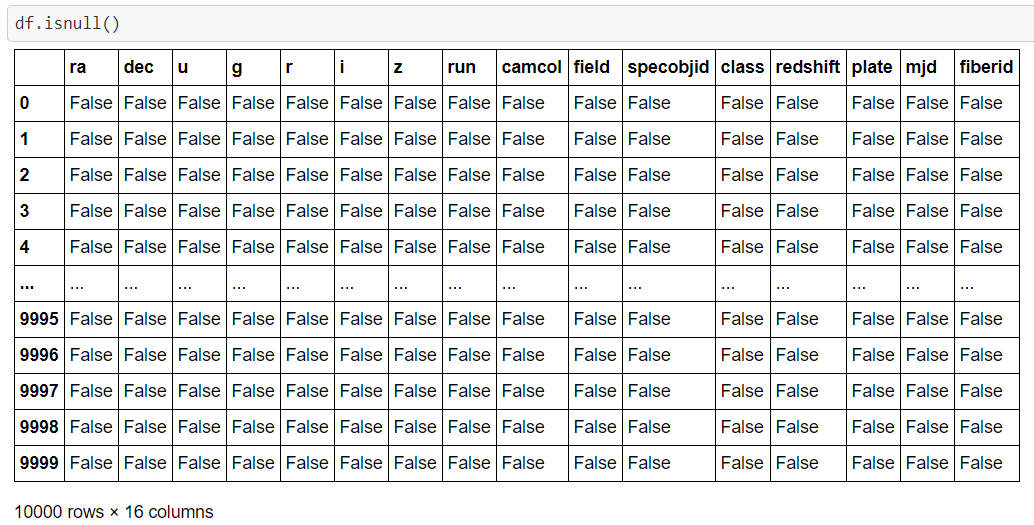
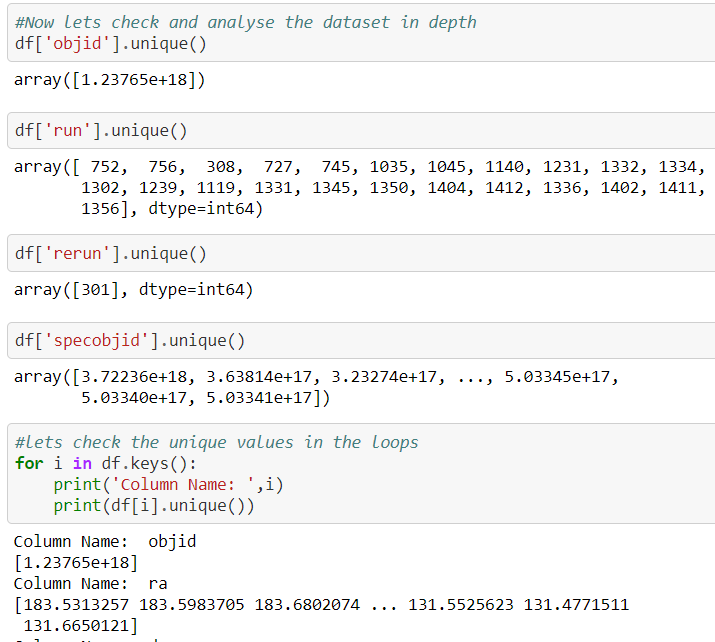
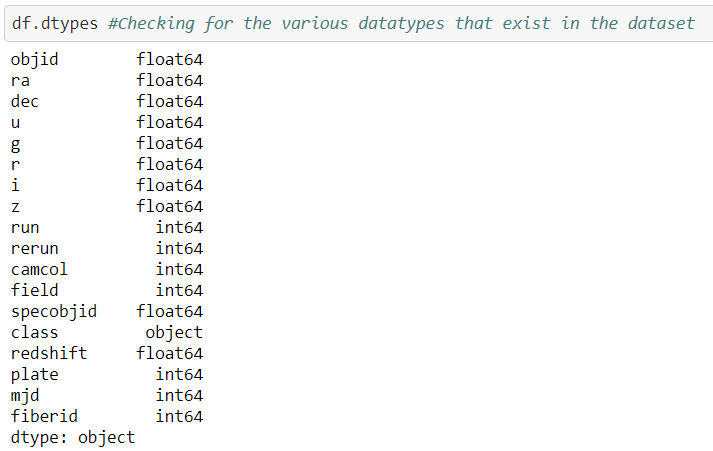


At first we would need to load the dataset in our dataframe so that we can analyse or use the data for further processing.

Now lets try to get some insight about our data with the help of inbuilt function info and describe

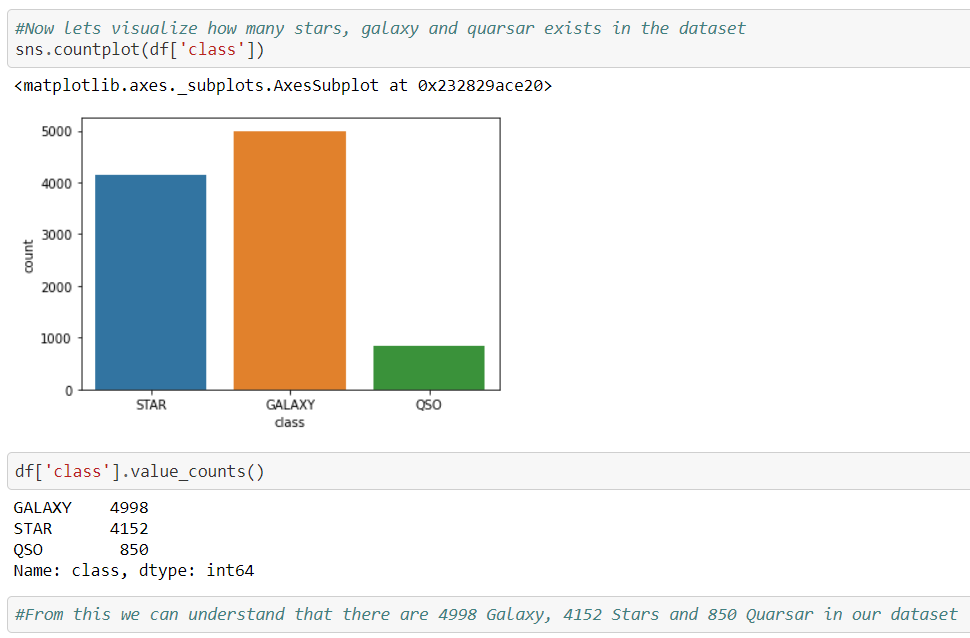
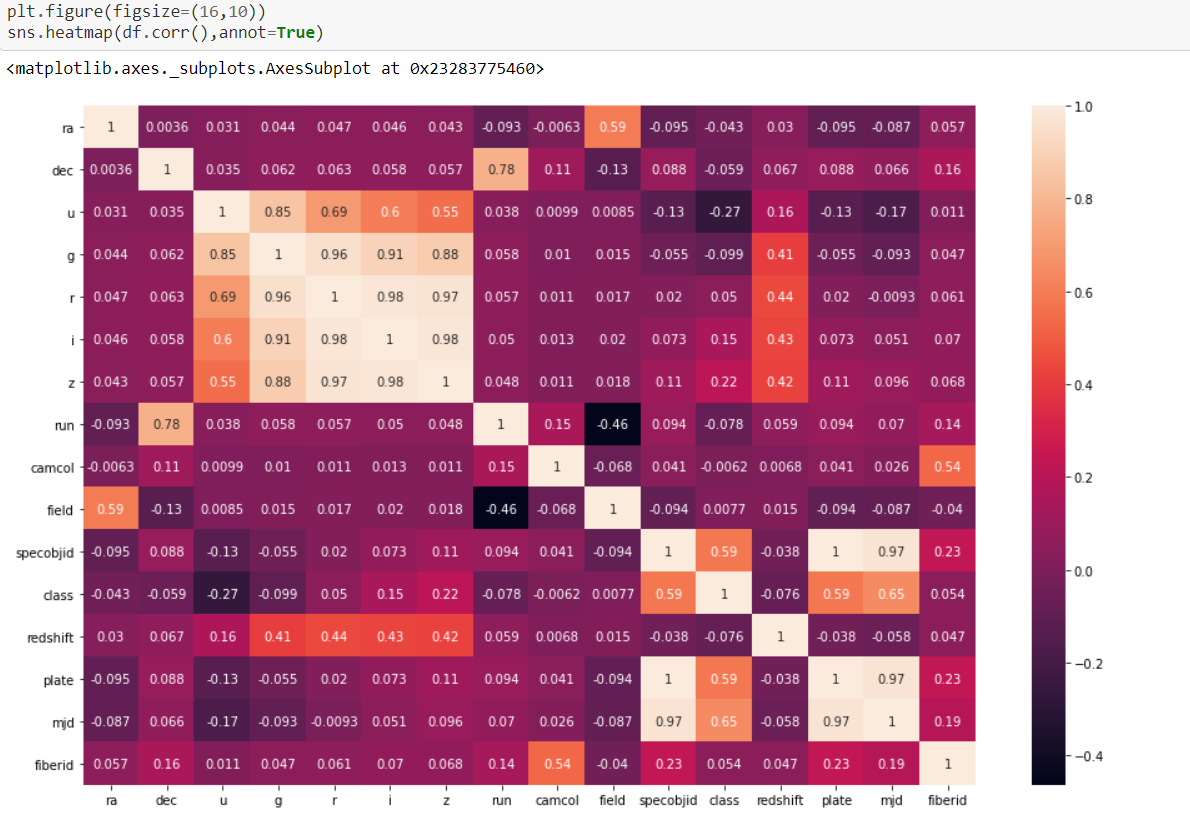


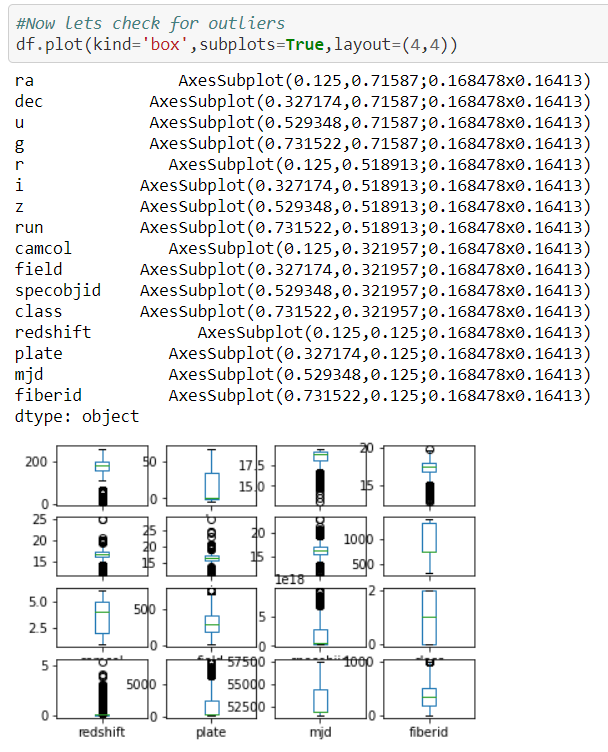
Now lets analyse the data in-depth by checking for different data types, unique values, null values in the data set.

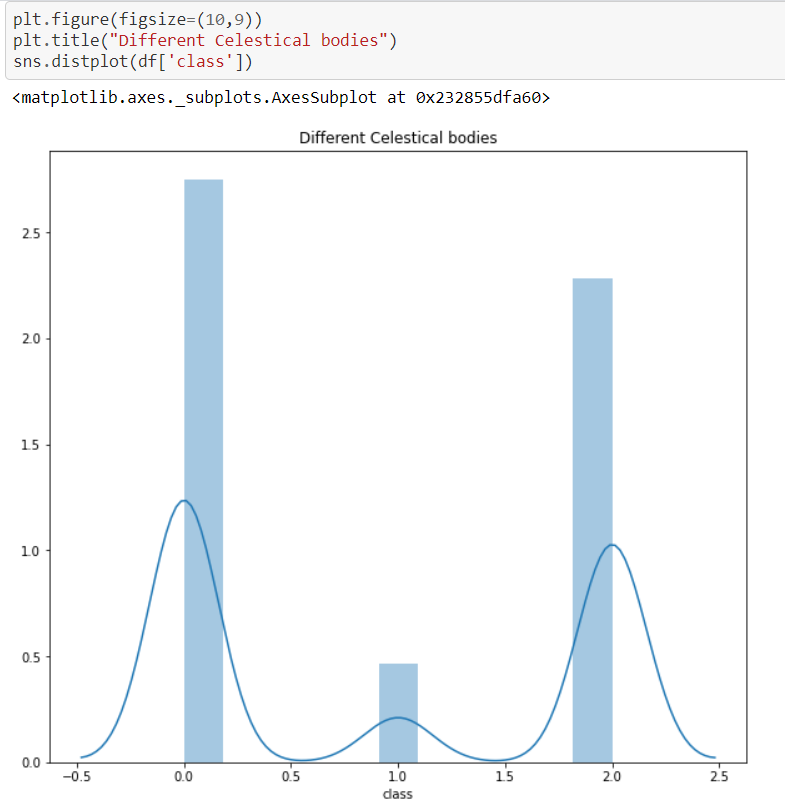


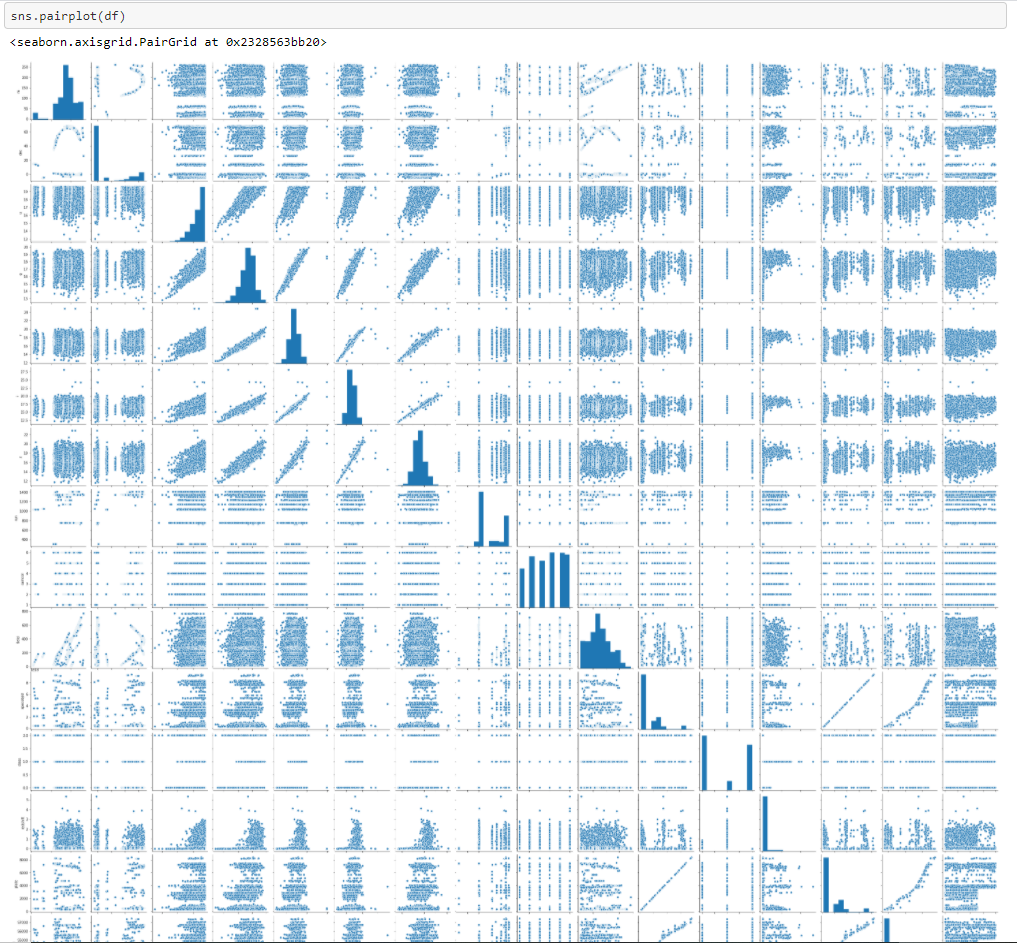
From the above analysis we can understand the different unique values that each column consists of and also we were able to trace that there are no null values present in the dataset.

In the next part we would be analysing our dataset set with the help of different dataset as it would help us to gather meaning full insights about our data

 In this part we would be using the different graphs to visualize our dataset and gain some valuable insights about our data







**EDA Findings:** Now that we have done our EDA process we can conclude that

1) There are no null values in the dataset.

2) Except the class attribute in the dataset all other attributes in the dataset are of numerical type so we have to convert the class attribute to numerical value before providing it to the model.

3) From heatmap we can see that there is a strong correlation between our target variable class and the attributes mjd, plate.

4) The columns objid and rerun are of no specific use to us so we can drop them.

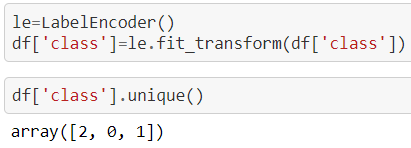
5) For few of the columns the standard deviation is quite high so we have to check for the presence of outlier.

6) Also for the few of the columns there is quite difference between the mean and the median so presence of outlier can be confirmed from the same.

Data Pre-processing:

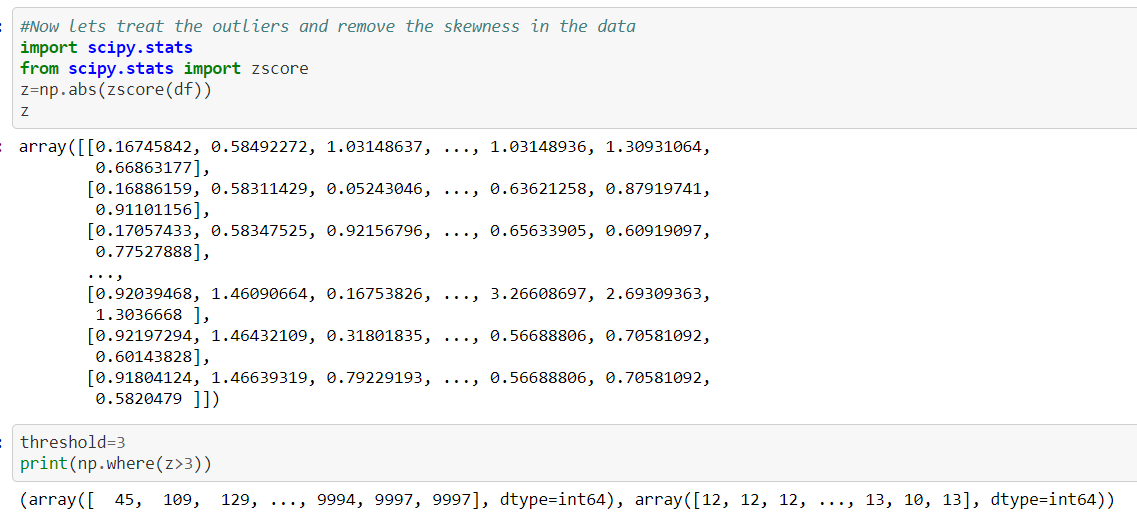


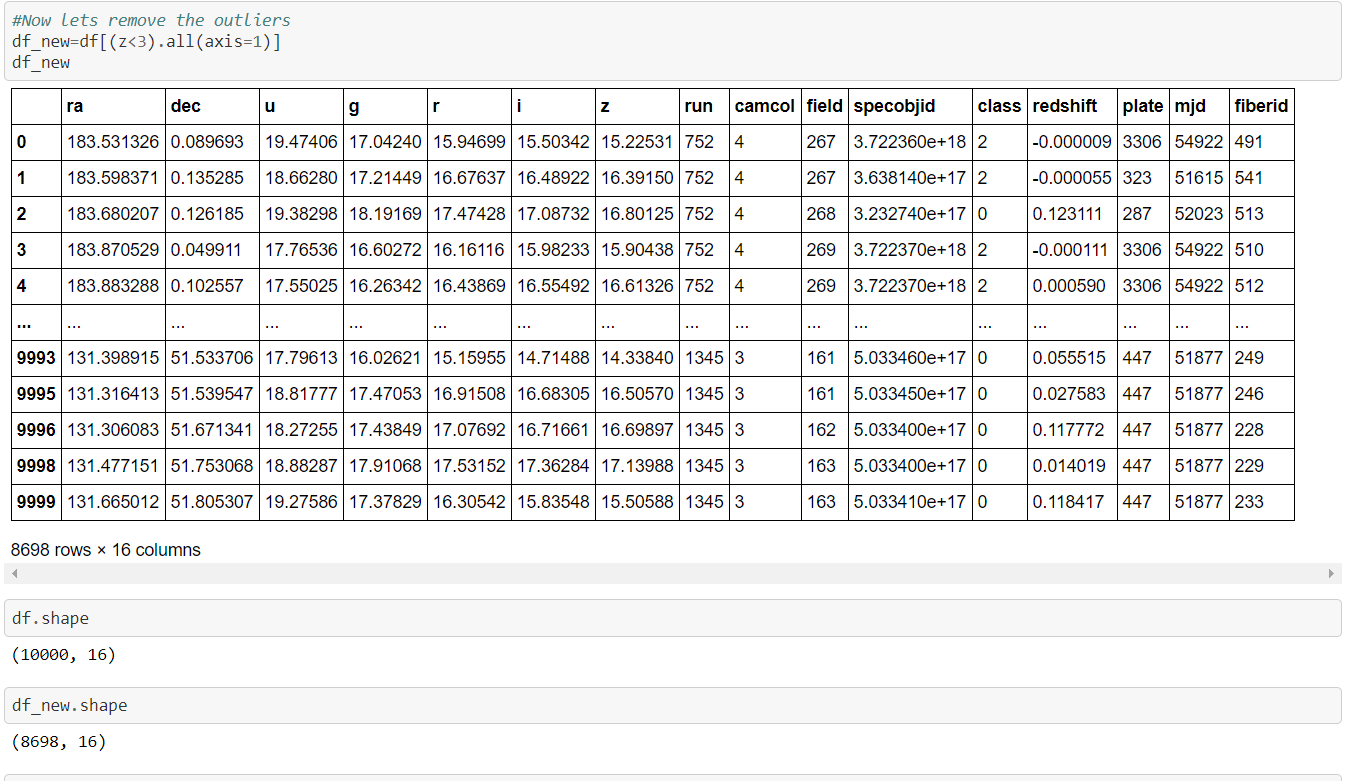
Now that we have analysed our dataset and understood that our dataset is having some anomalies such as outliers, skewness, categorical variables to have to treat this before providing this data to our model for our learning process.



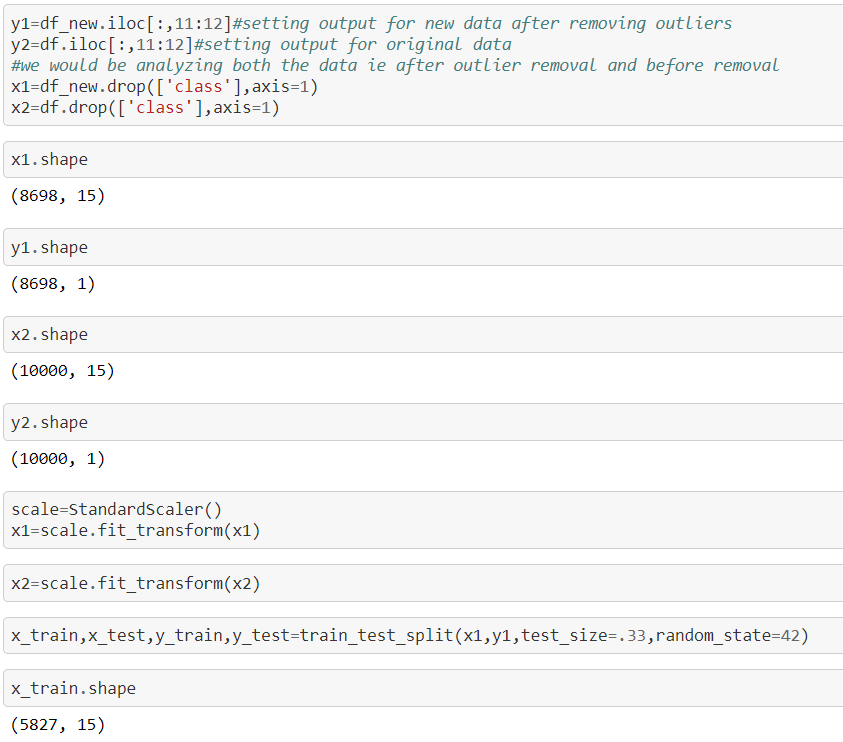
Now we have converted our target variable i.e. class which was an categorical variable into an numerical type with the help of Label encoder.

Now lets treat the outliers and the skewness in our dataset

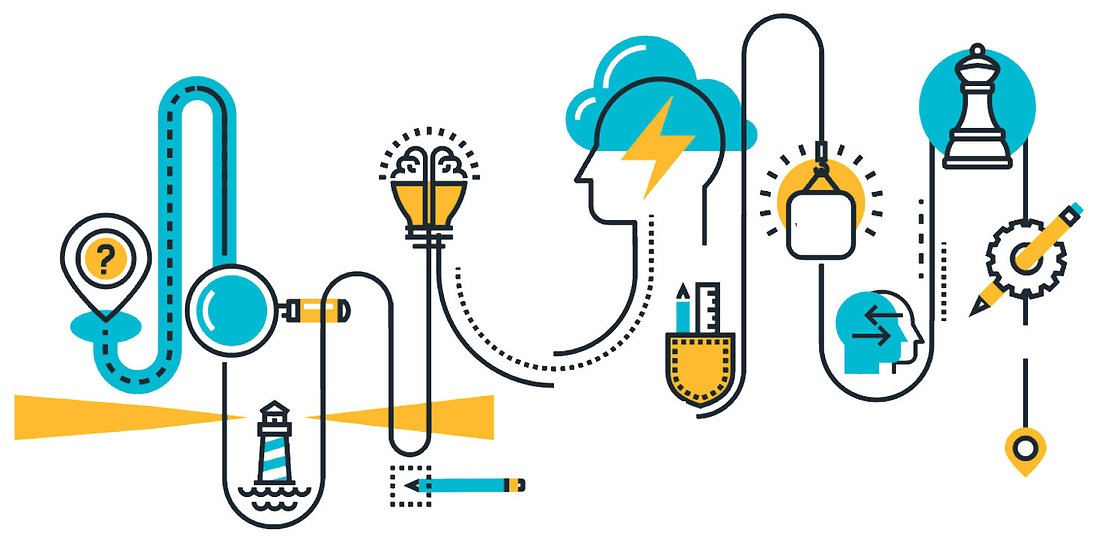




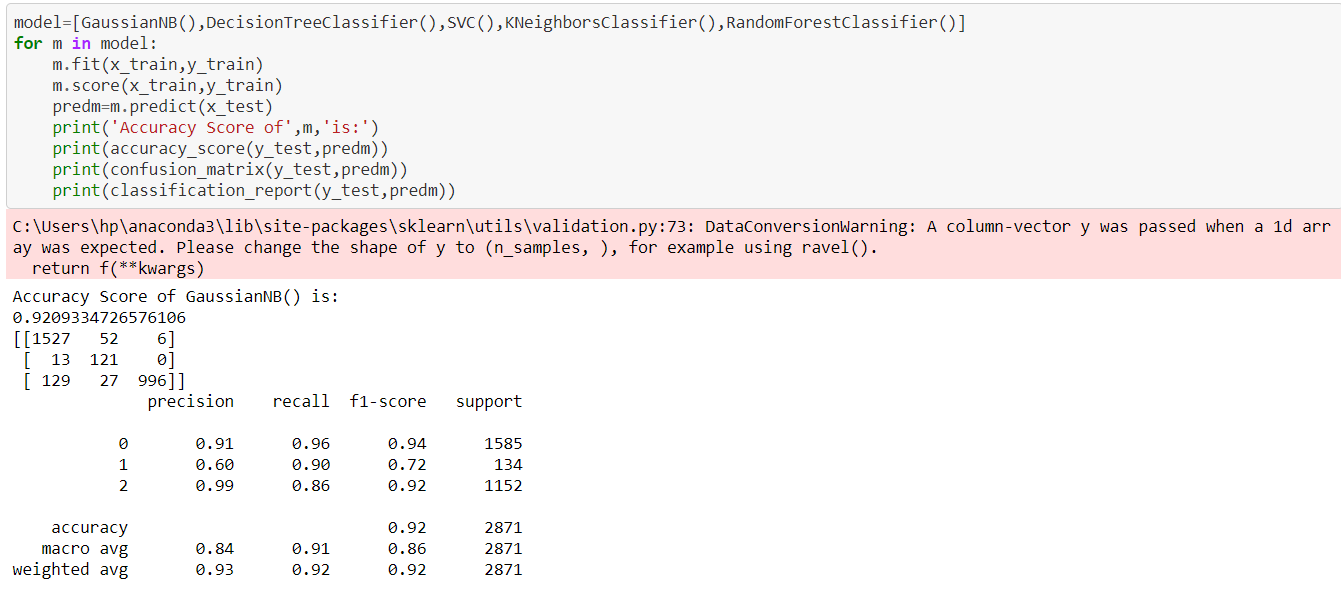
Now lets split our dataset into input and output variable to train our model and scale the variables.

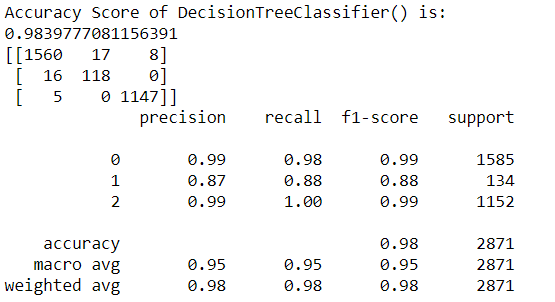


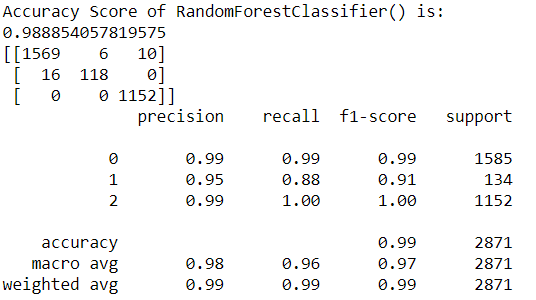
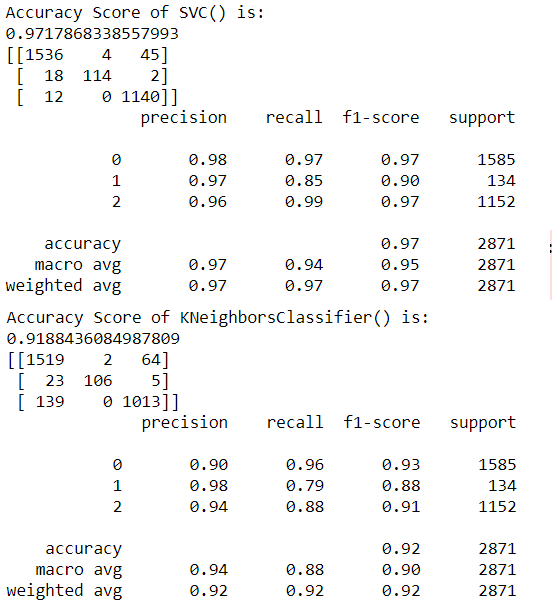
Building the Model:

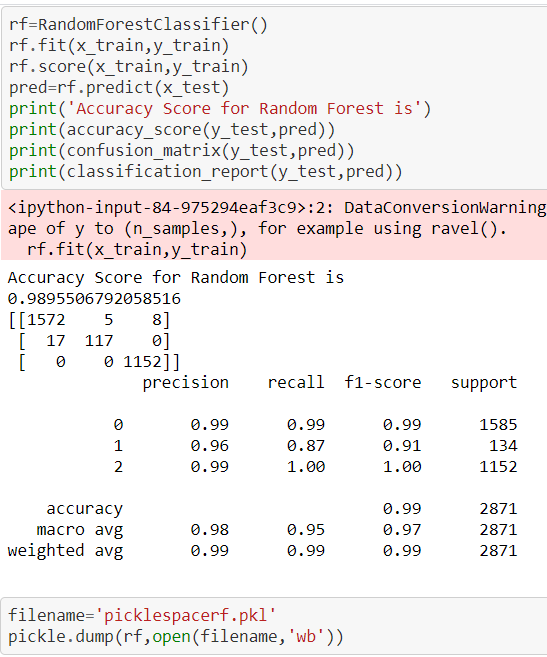


Now that we have pre-processed our data before providing it to the model we now move to the next step i.e. building and analysing different models.

 Here I have build a function which analysis the different classification algorithm on various metrics at last we would be selecting the best possible model and then tuning it.





Out of all this model the Random forest model provides us the high level of accuracy hence we would be saving that model

Conclusion:

Now that we have analysed the SDSS dataset successfully and have prepared a model with the help of Random forest algorithm with an accuracy level of more than 98 percent we can say that going a head we can use this model to predict the similar dataset when needed.

We choose Random forest algorithm as it gave us the highest level of accuracy, precision and fi-score.