**PROBLEM DEFINITION**

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.

**Inspiration:**

The aim is to try and predict/determine the whether the observed space object is a Star, galaxy or quasar.

**DATA DESCRIPTION**

Fields or columns present in the dataset :

* objid - Object Identifier - Single value
* ra - J2000 Right Ascension (r-band) - Float - Continuous

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.

* dec - J2000 Declination (r-band) - Float - Continuous
* u, g, r, i, z - Ultraviolet, Green, Red, Infrared - represent the five bands of a telescope - Float - Continuous
* run - identifies the specific scan - Integer - Continuous
* camcol - 'camera column' a number from 1 to 6, helps in identifying the scan line within the run - Integer - Continuous
* field - starts at 11 (after an initial rampup time), and can be as large as 800 - Integer - Continuous

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.

* specobjid - generated from the plate number, mjd, and fiberid - Integer - Continuous
* redshift - measure of the recession velocity of a galaxy or other sky object - Float - Continuous
* plate - unique serial number - Integer - Discrete

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.

* mjd - Modified Julian Date, used to indicate the date that a given piece of SDSS data was taken - Integer - Continuous
* fiberid - each object is assigned a corresponding fiberID - Integer - Continuous

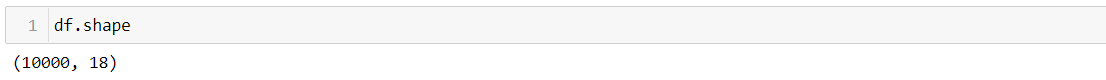
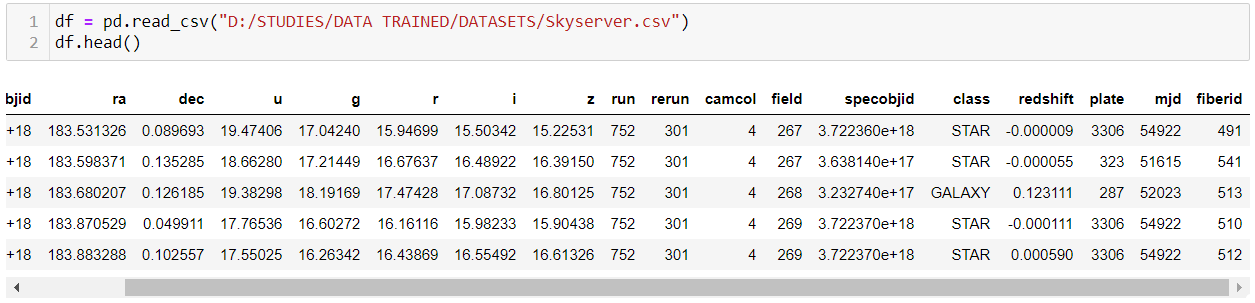
**DATA ANALYSIS & PRE-PROCESSING**

Importing Libraries:



These are all the libraries that are required at various stages - EDA, visualization, prediction and evaluation - for executing this project.

**Loading Dataset:**



I have loaded the data into ‘df’ variable. The dataset contains 10,000 rows and 18 columns.

**Statistical Data Analysis:**

